Abundance Biomass Comparison Method

RM Warwick, Plymouth Marine Laboratory, Plymouth, UK

Introduction

The `abundance biomass comparison' (ABC) method is a means of detecting the effects of anthropogenic perturbations on assemblages of organisms that is underpinned by the *r*- and *K*-selection theory (see *r*-Strategist/*K*-Strategists). Under stable conditions of infrequent disturbance the competitive dominants in the climax community are *K*-selected or conservative species with a large body size and long life span, and are usually of low abundance so that they are not dominant numerically but are dominant in terms of biomass. Frequently disturbed assemblages are kept at an early successional stage and comprise *r*-selected or opportunistic species characterized by small body size, short life span and high abundance. The ABC method exploits the fact that when an assemblage is perturbed the conservative species are less favored in comparison with the opportunists, and the distribution of biomass among species behaves differently from the distribution of numbers of individuals among species.

The Method

The ABC method as originally formulated involves the plotting of separate *k*-dominance curves (see *k*-Dominance Curves) for species abundances and species biomasses on the same graph and comparing the forms of the two curves relative to each other. The species are ranked in order of importance in terms of abundance or biomass on the *x*-axis on a logarithmic scale, with percentage dominance on the *y*-axis on a cumulative scale. Of course the species ordering is unlikely to be the same for abundance and biomass. In undisturbed assemblages a few large species are dominant in terms of biomass but not abundance, resulting in the elevation of the biomass curve relative to the abundance curve throughout its length (**Fig. 1a**). Perturbed assemblages, however, have a few species with very high abundance but small body size so that they do not dominate the biomass and the abundance curve lies above the biomass curve (**Fig. 1c**). Under moderate perturbation the large competitive dominants are eliminated but there is no population explosion of small opportunists, so that the inequality in size between the numerical and biomass dominants is reduced and the biomass and abundance curves are closely coincident and may cross over each other one or more times (**Fig. 1b**).

The contention is that these three conditions (unperturbed, moderately perturbed, or grossly perturbed) should be recognizable without reference control samples in time or space, the two curves acting as an internal control against each other and providing a snapshot of the condition of the assemblage at any one time or place. Of course, confirmatory comparisons with spatial or temporal reference samples are still highly desirable. A prerequisite of the method is adequate sample size or replication because the large biomass dominants are often rare and liable to a higher sampling error than the numerical dominants.

The evaluation of ABC curves involves their visual inspection, and can be cumbersome if many sites, times, or replicates are involved. In such cases it is convenient to reduce each plot to a single summary statistic. If the abundance (A) values are subtracted from the biomass (B) values for each species rank in the ABC curve, the sum of the $B-A$ values across the ranks will be strongly positive in the unperturbed case (**Fig. 1a**), near zero in the case where the curves are closely coincident (**Fig. 1b**), and strongly negative where the curves are transposed (**Fig. 1c**). The summation needs to be standardized to a common scale so that comparisons can be made between samples with differing numbers of species (S), the most widely used form being the W (for Warwick) statistic:

$$W = \sum_{i=1}^{S}(B_i - A_i)/[50(S-1)]$$

For replicated samples, the W statistic also provides an obvious route for hypothesis testing, using standard univariate ANOVA.

Applications

For the most part, ABC curves have been used to indicate pollution or disturbance effects on marine and estuarine macrobenthic assemblages, which are the main target for detection and monitoring programs in these habitats. For example, ABC curves for the

 doi:10.1016/B978-0-444-63768-0.00084-6

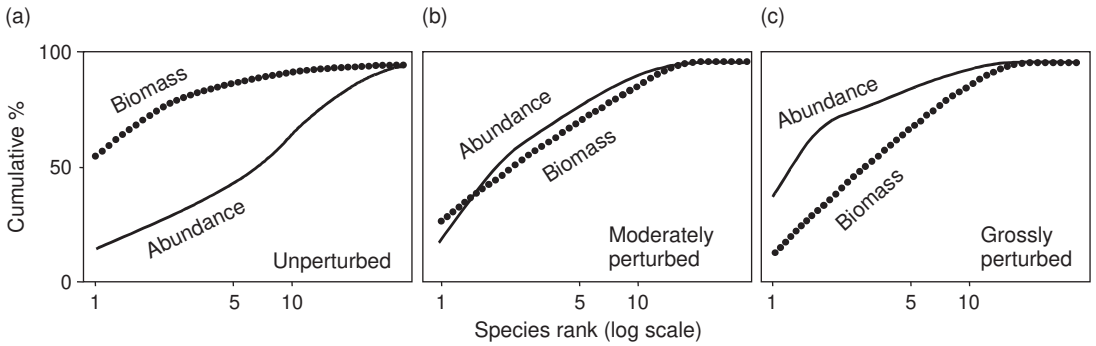


Fig. 1 Hypothetical *k*-dominance curves for species biomass and abundance, showing unperturbed, moderately perturbed, and grossly perturbed conditions.

macrobenthos in Loch Linnhe, Scotland in response to organic pollution between 1963 and 1973 are given in Fig. 2. The time course of pollution from a pulp mill, and changes in species diversity (H'), are shown top left. Moderate pollution started in 1966, and by 1968 species diversity was reduced. Prior to 1968 the ABC curves had the unpolluted configuration. From 1968 to 1970 the ABC plots indicated moderate pollution. In 1970 there was an increase in pollutant loadings and a further reduction in species diversity, reaching a minimum in 1972, and the ABC plots for 1971 and 1972 show the grossly polluted configuration. In 1972 pollution decreased and by 1973 diversity had increased, and the ABC plots again indicated the unpolluted condition. Thus, the ABC plots provide a good snapshot of the pollution status of the benthic community in any one year, without reference to the historical comparative data which would be necessary if a single species diversity measure based on the abundance distribution was used as the only criterion.

Most studies suggest that the ABC curves respond to anthropogenic perturbations but are not affected by long-term natural stresses, since the organisms living in such environments have evolved adaptations to the prevailing ecological conditions. Unperturbed ABC plots may be found, for example, in estuaries where the organisms are subjected to low and fluctuating salinities, provided there are no anthropogenic disturbances. ABC plots indicated that macrobenthic communities near an oil refinery in Trinidad were grossly to moderately stressed, while those close to the Trinidad Pitch Lake (one of the largest natural oil seeps in the world) were not. There is little evidence, however, that the method can distinguish between different types of anthropogenic disturbances. Responses to organic pollution and to physical disturbance caused by demersal trawl fisheries, for example, appear to be similar.

The method has been less well explored with respect to other components of the biota. However, it has been used successfully to indicate environmental impacts on marine phytoplankton, the cryptofauna and mollusks of rocky shores, invertebrates of freshwater streams, and fish assemblages in both marine and freshwater.

Problems and Their Solutions

Very often *k*-dominance curves approach a cumulative frequency of 100% for a large part of their length, and in highly dominated assemblages this may be after the first two or three top-ranked species. Thus, it may be difficult to distinguish between the forms of these curves. The solution to this problem is to transform the y-axis so that the cumulative values are closer to linearity, an appropriate transformation being the modified logistic transformation:

$$y_i' = \log\ [(1 + y_i)/(101 - y_i)]$$

A potentially more serious problem with the cumulative nature of ABC curves is that their form is overdependent on the single most dominant species. The unpredictable presence of large numbers of a species with small biomass, perhaps an influx of the juveniles of one species, may give a false impression of disturbance. With genuine disturbance, one might expect patterns of ABC curves to be unaffected by successive removal of the one or two most dominant species in terms of abundance or biomass, and a solution is the use of partial dominance curves, which compute the dominance of the second-ranked species over the remainder (ignoring the first-ranked species), the same with the third most dominant, etc. Thus, if a_i is the absolute (or percentage) abundance of the ith species, when ranked in decreasing abundance order, the partial dominance curve is a plot of p_i against log i ($i = 1, 2, \ldots, S-1$), where

$$p_1 = 100a_1/\sum_{j=1}^{S} a_j$$

$$p_2 = 100a_2/\sum_{j=2}^{S} a_j, \ldots, p_{S-1} = 100a_{S-1}/(a_{S-1} + a_S)$$

Earlier values can therefore never affect later points on the curve. The partial dominance curves (ABC) for undisturbed macrobenthic communities typically look like Figs. 3g and 3h, with the biomass curve (thin line) above the abundance curve (thick line) throughout its length. The abundance curve is much smoother than the biomass curve, showing a slight and steady decline before the inevitable final rise. Under polluted conditions there is still a change in position of partial dominance curves for abundance and biomass, with the abundance curve now above the biomass curve in places, and the abundance curve becoming much more variable. This implies that pollution effects are not just seen in changes to a few dominant species but are a phenomenon which pervades the

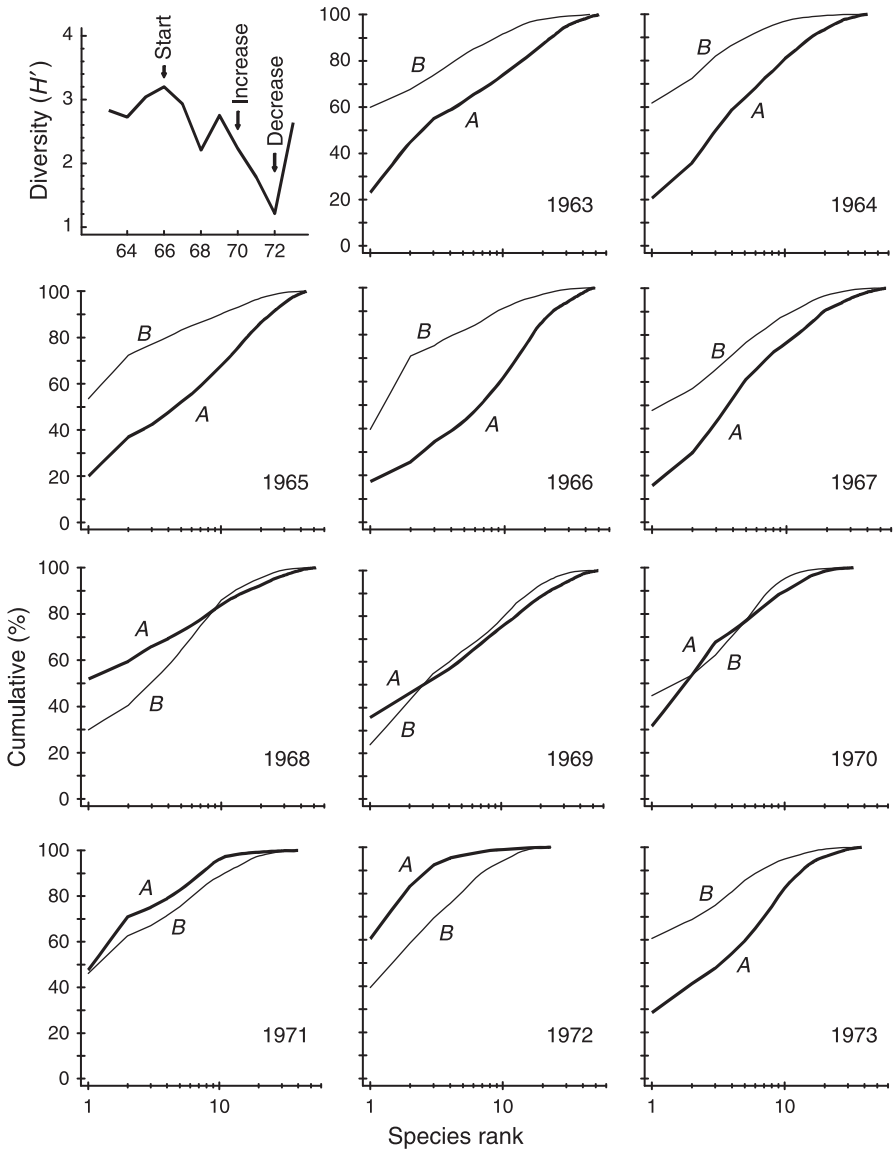


Fig. 2 Loch Linnhe macrofauna: Shannon diversity (H') and ABC plots over the 11 years, 1963 to 1973. Abundance, thick line; biomass, thin line.

complete suite of species in the community. The time series of macrobenthos data from Loch Linnhe (Fig. 3) shows that in the most polluted years, 1971 and 1972, the abundance curve is above the biomass curve for most of its length (and the abundance curve is very atypically erratic), the curves cross over in the moderately polluted years 1968 and 1970 and have an unpolluted configuration prior to the pollution impact in 1966 and 1967. Although these curves are not so smooth, and therefore not so visually appealing, as the original ABC curves, they may provide a useful alternative aid to interpretation and are certainly more robust to random fluctuations in the abundance of a small-sized, numerically dominant species.

In most cases where the presence of large numbers of small-bodied macrobenthic species in unperturbed situations has given a false impression of disturbance, those species have not been polychaetes. Prior to the Amoco Cadiz oil spill off the north coast of France in 1978, small ampeliscid amphipods (Crustacea) were present at the Pierre Noire station in relatively high abundance, and their disappearance after the spill confounded the ABC plots. The erratic presence of large numbers of small amphipods (*Corophium*) or mollusks (*Hydrobia*) also confounded these plots in the Wadden Sea. These small nonpolychaetous species are not indicative of polluted conditions. A taxonomic breakdown of the ABC response has shown that it results from (1) a shift in the proportions of different phyla present in communities, some phyla having larger-bodied species than others, and (2) a shift in the relative distributions of abundance and biomass among species within the Polychaeta but not within any of the other major phyla (Mollusca, Crustacea, Echinodermata). The shift within polychaetes reflects the substitution of larger-bodied by smaller-bodied species, and not a change in the average size of individuals within a species. In most instances the phyletic changes reinforce the trend in species substitutions within the polychaetes, to produce the overall ABC response, but in some cases they may work

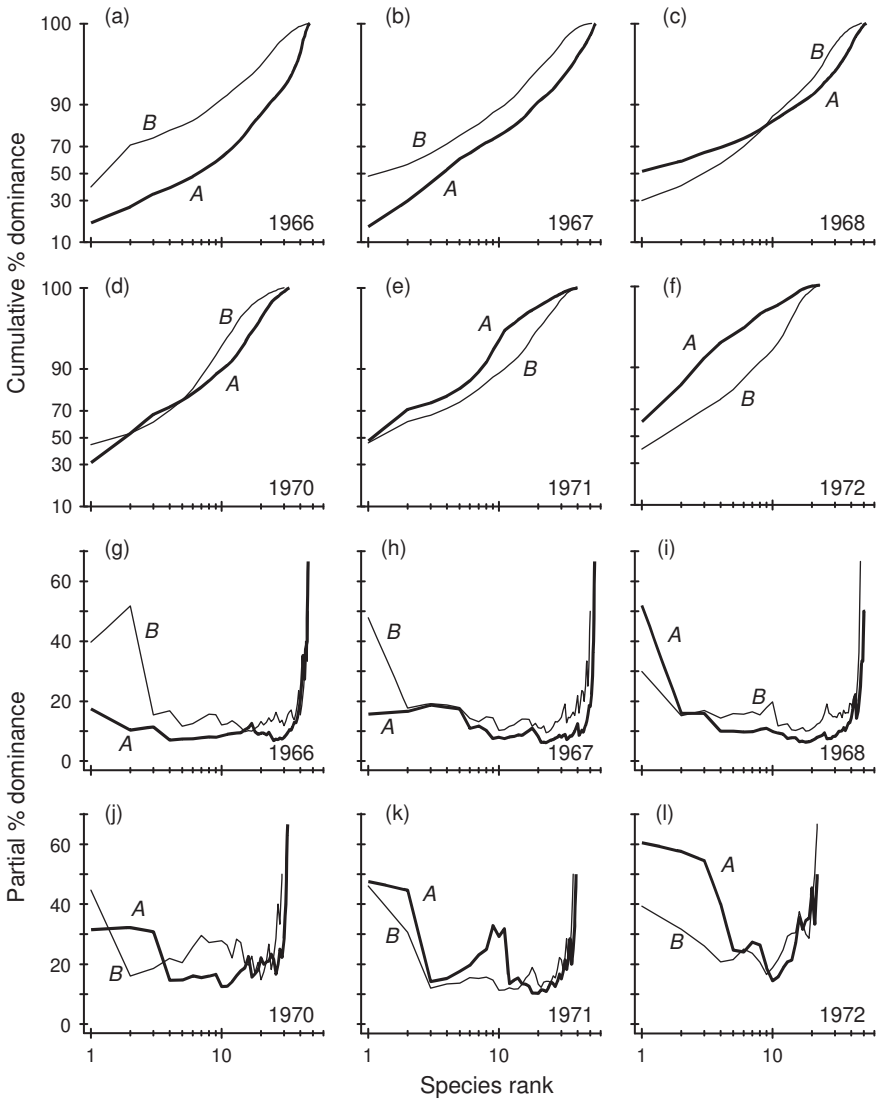


Fig. 3 Loch Linnhe macrofauna in selected years 1966–68 and 1970–72. (a–f) ABC curves (logistic transform). (g–l) Partial dominance curves for abundance (thick line) and biomass (thin line) for the same years.

against each other. Indications of pollution or disturbance for marine macrobenthos detected by this method should therefore be viewed with caution if the species responsible for the perturbed configurations are not polychaetes, and the robustness of the plots should be tested using partial dominance curves.

Finally, a practical rather than a conceptual problem with the method is that it relies on a painstaking and time-consuming (and hence costly) analysis of samples in which all the species must be separated, counted, and weighed. Several groups of marine organisms are taxonomically difficult, for example (in the macrobenthos), several families of polychaetes and amphipods; as much time can be spent in separating a few of these difficult groups into species as the entire remainder of the sample, even in Northern Europe where taxonomic keys for identification are most readily available. Many taxa really require the skills of specialists to separate them into species, and this is especially true in parts of the world where fauna is poorly described. Identification to some higher taxonomic level, for example, family rather than species, is considerably easier and quicker, and the ABC method has proved to be encouragingly robust to analysis at the family level for both macrobenthos and fish; very little information appears to be lost.

See also: Aquatic Ecology: The Estuarine Quality Paradox Concept. Conservation Ecology: k-Dominance Curves; Ecological Health Indicators. Ecosystems: Estuaries. General Ecology: Abundance; Biomass; Dominance

Further Reading

Agard, J.B.R., Gobin, J., Warwick, R.M., 1993. Analysis of marine macrobenthic community structure in relation to oil pollution, natural oil seepage, and seasonal disturbance in a tropical environment (Trinidad, West Indies). Marine Ecology Progress Series 92, 233–243.

Beukema, J.J., 1988. An evaluation of the ABC-method (abundance/biomass comparison) as applied to macrozoobenthic communities living on tidal flats in the Dutch Wadden Sea. Marine Biology 99, 425–433.

Blanchard, F., LeLoc'h, F., Hily, C., Boucher, J., 2004. Fishing effects on diversity, size, and community structure of the benthic invertebrate and fish megafauna on the Bay of Biscay coast of France. Marine Ecology Progress Series 280, 249–260.

Clarke, K.R., 1990. Comparisons of dominance curves. Journal of Experimental Marine Biology and Ecology 138, 143–157.

Dauer, D.M., Luckenbach, M.W., Rodi, A.J., 1993. Abundance biomass comparison (ABC method) – Effects of an estuarine gradient, anoxic hypoxic events and contaminated sediments. Marine Biology 116, 507–518.

Ismael, A.A., Dorgham, M.M., 2003. Ecological indices as a tool for assessing pollution in El-Dekhaila Harbour (Alexandria, Egypt). Oceanologia 45, 121–131.

Jouffre, D., Inejih, C.A., 2005. Assessing the impact of fisheries on demersal fish assemblages of the Mauritanian continental; shelf, 1987–1999, using dominance curves. ICES Journal of Marine Science 62, 380–383.

Lasiak, T., 1999. The putative impact of exploitation on rocky infratidal macrofaunal assemblages: A multiple area comparison. Journal of the Marine Biological Association of the United Kingdom 79, 23–34.

Magurran, A.E., 2004. Measuring Biological Diversity. Oxford: Blackwell.

Penczak, T., Kruk, A., 1999. Applicability of the abundance/biomass comparison method for detecting human impacts on fish populations in the Pilica River, Poland. Fisheries Research 39, 229–240.

Warwick, R.M., 1986. A new method for detecting pollution effects on marine macrobenthic communities. Marine Biology 92, 557–562.

Warwick, R.M., Clarke, K.R., 1994. Relearning the ABC: Taxonomic changes and abundance/biomass relationships in disturbed benthic communities. Marine Biology 118, 739–744.

Warwick, R.M., Pearson, T.H., Ruswahyuni, 1987. Detection of pollution effects on marine macrobenthos: Further evaluation of the species abundance/biomass method. Marine Biology 95, 193–200.

Yemane, D., Field, J.G., Leslie, R.W., 2005. Exploring the effects of fishing on fish assemblages using abundance biomass comparison (ABC) curves. ICES Journal of Marine Science 62, 374–379.

Acidification in Aquatic Systems

Morgana Tagliarolo, Ifremer, UMSR LEEISA (CNRS UG Ifremer), Cayenne, France

Glossary

Buffer A compound that limits large changes in pH in a solution. The buffer solution consists in a mixture of weak acid and bases. The buffering ability is defined as the quantity of strong acid or base that must be added to change the pH of 1 L of solution by 1 pH unit.
Chemical equilibrium State of a reversible chemical reaction when the rate of the forward and backward reaction is equal. Consequently, the concentrations of both reactant and product are stables. This equilibrium can be described by a constant (K).
Homeostasis A biological process maintaining a stable condition in an organism even if faced with external changes. An example is the ability of the human body to maintain a relatively constant body temperature independent of external temperatures.
Hypercapnia Disequilibrium in body fluids with an abnormal increase of carbon dioxide and pH drop.
Oligotrophic An environment poor in organic and inorganic nutrients.
Resilient An ecosystem or organism is defined as resilient when it is able to respond and recover from adverse situations.

Introduction

Acidification in an aquatic system is a term describing significant changes to the chemistry of freshwater, marine, and brackish systems, mostly caused by the dissolution of atmospheric carbon dioxide (CO_2). The CO_2 from the atmosphere combines with other dissolved inorganic carbon already present in the water causing several complex chemical changes. The characterization of the physicochemical properties of the carbonate system in natural waters is not straightforward since it can be described by a large number of terms and units. However, pH is the more common parameter employed for describing the acidification phenomena in ecology.

Water pH is an expression of the concentration of hydrogen ions (H^+) on a logarithmic scale, where a neutral pH of 7.00 is represented by pure water at 25°C. Surface waters in the open ocean are slightly alkaline with relatively small pH variability (average values ranging between 7.9 and 8.3). Variability ranges are wider in coastal and freshwater ecosystems where complex biogeochemical dynamics play important roles on the physical and chemical conditions of those waters. In shallow coastal areas pH can vary drastically over daily cycles and small spatial scales (~0.3–0.5 units). In natural freshwater ecosystems pH has an even wider range (between <2 and 12) depending on the region and on the water body characteristics.

Part of atmospheric CO_2 is continuously absorbed by the aquatic systems where it reacts with the water molecules to form weak acids. These acids mostly dissociate into H^+ causing pH reductions. Freshwater and seawater contain a variety of acid–base species able to react with the additional H^+ ions. The predominant ions are carbonate and bicarbonate, but other molecules can also interact. This ability of natural waters to neutralize protons is described by its total alkalinity (Fig. 1). The carbonate chemistry is significantly affected by acidification and the formation of numerous carbonate-containing minerals such as aragonite, calcite, and magnesium calcite is disrupted.

The absorption of CO_2 and the fate of hydrogen ions in water are therefore dependent on various chemical transformations and equilibrium constants. The equilibrium constants are in turn dependent on salinity, temperature, and pressure. For this reason, fresh and seawater are considerably different in the distribution of the carbonic acid fractions. Fresh and brackish water has a lower buffering capacity and thus experiences higher pH fluctuations compared with open ocean waters.

Although the acidification process has mostly been studied in the marine environment, declines in pH can also considerably affect freshwater ecosystems. Marine and freshwater systems may be acidified either from natural or man-made processes, but, while natural processes are slow (geological time frames), anthropogenic activities are accelerating and amplifying these changes.

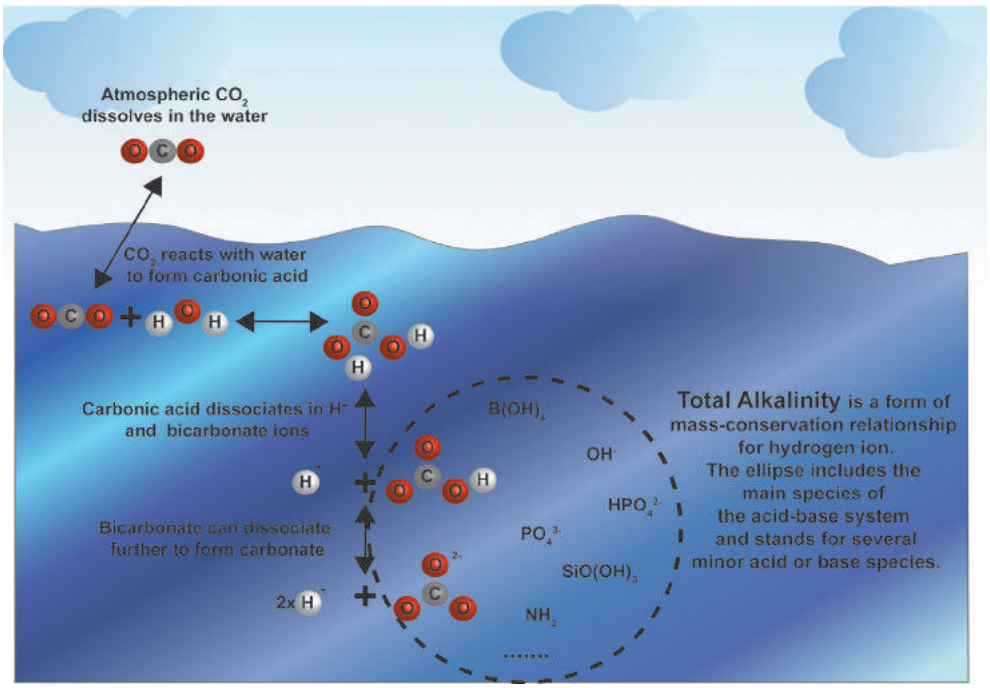


Fig. 1 General scheme on how atmospheric carbon dioxide can interfere with the aquatic chemistry.

The acidification of aquatic systems is mostly a result of the continuous CO_2 release into the atmosphere by fossil fuel combustion (coal, oil, and natural gas), deforestation, and cement production. Since the beginning of the industrial era, the pH of ocean surface water has decreased by 0.1 pH units, corresponding to a 26% increase in acidity. Moreover, pH of internal and coastal areas can also be altered by other anthropogenic inputs such as nitrogen, ammonia, and sulfur. Consequently, freshwater acidification can be a faster and larger phenomenon than ocean acidification involving notable pH drops (up to 2.5) during episodic events.

The anthropogenic CO_2 induced changes in water carbon chemistry have some direct effects on photosynthesis, calcification, and acid–base balance of aquatic organisms. The increased availability of CO_2 can potentially enhance photosynthesis when light and nutrients are available. However, the ability of micro and macro algae to utilize this excess of CO_2 appears to vary widely across taxa. A number of physiological processes can be altered in photosynthetic organisms and the final response is often a compromise between the antagonistic effect of CO_2 on photosynthesis and respiration metabolism. Low pH often causes an increase of the overall energetic costs, which in turn lead to an augmented respiration rate. Therefore, the benefits of an enhanced photosynthetic activity are generally relatively minor relative to the negative effects of acidification on respiration. Moreover, acidification has been shown to promote several metabolic pathways leading to the production and accumulation of toxic compounds in phytoplankton cells.

A decrease in pH and in carbonate ions generally causes a decline in the calcification rates producing calcium carbonate ($CaCO_3$) for shells and skeletons (Fig. 2). Calcifying organisms are extremely diverse and include many taxonomic groups and ecological niches. For instance, calcification is performed by many photosynthetic primary producers, zooplankton, mollusks, and crustaceans. Although the calcification process can be explained by a simple chemical equation, the biological mechanisms are more complex and can vary between species. The calcification process requires an energy investment for the organisms and modifications of the chemistry of the external aquatic environment can cause important perturbations of calcification rates. Although calcifying organisms are mostly negatively impacted by acidification, the growth rate of some species has been reported as insensitive to this stress or positively impacted.

Shell calcification in mollusks is performed in the extrapallial space that is isolated from the surrounding ambient water (Fig. 2). Many species have been shown to produce their own carbonate ions in this space without any interaction with seawater ions. The formation of calcium carbonate structures in this space is thus not directly inhibited by decreasing carbonate ions concentration in the external seawater. The impact of ocean acidification on shell growth is therefore a result of several interlinked physiological processes affecting metabolism and internal body pH. Furthermore, the antagonistic processes of calcification and dissolutions may also represent an important energetic cost for the organisms. For these reasons, marine mollusks have been shown to respond very differently to acidification. In a more acid ocean, on average, shells are expected to be smaller and have a modified mineralogical structure to overcome the energetic costs.

Similarly to mollusks, scleractinian corals do not precipitate their carbonate skeleton directly from seawater, but they produce it in an extracellular medium where the animal can actively manipulate the pH. Cold-water corals are particularly good in regulating

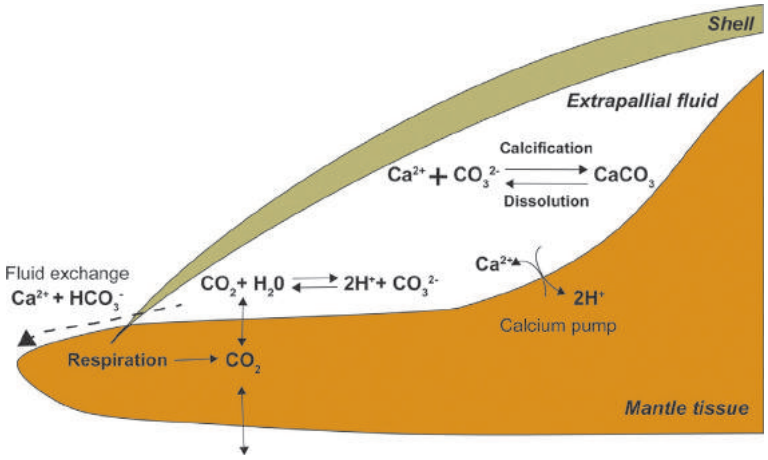


Fig. 2 Cross section of a mollusk shell showing the calcification process in the extrapallial space. Modified from McConnaughey, T.A., Gillikin, D.V. (2008). Carbon isotopes in mollusk shell carbonates. *Geo-Marine Letters* **28**, 287–299. 10.1007/S00367-008-0116-4.

their internal pH and can continue to calcify even in acidic and undersaturated waters. This important adaptive mechanism enables these deep-sea calcifiers to occupy unique niches. In contrast with these highly adapted organisms, other taxa such as the calcitic foraminifera, are not able to control their internal pH and depend entirely on seawater conditions.

The shells of marine organisms are generally made from either calcite or aragonite calcium carbonate. Aragonite is more soluble than calcite and can dissolve easier in undersaturated waters. Changes in water chemistry therefore affect calcifying organisms in different ways depending on their shells compositions. For example, one of the more abundant and important planktonic shelled group living in polar and subpolar waters (pteropods) are expected to be particularly affected by acidification since they are unable to maintain their shell in undersaturated waters with respect to aragonite. On the other hand, arctic bivalves have been shown to be generally resilient to decreasing pH.

Metabolic and biochemical processes within aquatic organisms are influenced by water pH since biological membranes are generally highly permeable to free ions and potentially affected by acid–base distresses. The internal pH of most heterotopic organisms is lower than the surrounding seawater. Active mechanisms of ion transport are constantly working in living cells to maintain pH fluctuations in body fluids within a tolerable range. For this reason, metabolic rates are linked to the homeostasis of the internal pH and to the pH gradient between the body and the external environment. Under acidification conditions some metabolic functions are depressed, oxygen transport efficiency is reduced and acid–base regulation requires higher energetic costs. The effectivity of ion transport systems differs between aquatic organisms depending on their structure and complexity. Some crustacean and echinoderms seem to be able to compensate for acidification by increasing bicarbonate concentration in their body but little is known about the sustainability of this response during long term exposures.

Although there is a general agreement about the effects of increased CO_2 concentrations on the ocean and freshwater chemistry, the magnitude and severity of the potential impact on different organisms and ecosystems is still largely unknown. The large variability in the responses of organisms to acidification is mostly due to the high biotic and abiotic patchiness of natural habitats and the multiple interactions with other existing stresses like temperature changes. This article will give an overview on the different mechanisms and responses of the ocean, coastal, and freshwater habitats and provide some wider concept on the ecological and economic impact of this phenomena.

Ocean Acidification

Oceans play an important role in mitigating atmospheric CO_2 emissions by absorbing around 25%–30% of the CO_2 added to the atmosphere by anthropogenic activities. This process causes a decrease in water's pH, in the concentration of carbonate ions and in the carbonate saturation state. These changes, together with other stressors such as temperature, are significantly affecting marine communities. Numerous research projects have focused on the effect of ocean acidification on marine organisms.

Negative effects have been recorded on most taxonomic groups, and early developmental stages are suspected to be most susceptible. Multigenerational studies on marine copepods showed that fecundity could decrease up to 29% under lower pH scenarios. Moreover, the larval nauplii stage of these taxa exhibit greater sensitivity and mortality rates with increasing CO_2 compared to the other life stages. Higher concentrations of seawater CO_2 have been shown to cause malformations and delays in the larval development of several calcifiers. Sea urchin larvae lose symmetry and their skeleton is highly deformed and corroded with decreased pH (Fig. 3). Beside invertebrates, fish embryos and young larvae are also more sensitive than adults to the effects of ocean acidification. Egg survival, hatching size, and growth rate are generally retarded despite some positive effects that have been

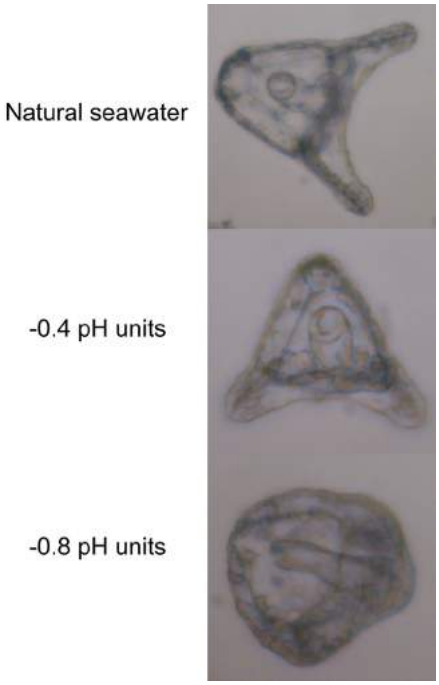


Fig. 3 Sea urchin larvae reared in control and acidified treatments. Lower pH causes asymmetrical and deformed skeletal development.

seen on few species. Even if the growth rate is not affected, other perturbations such as impaired olfactory discrimination and homing ability can still reduce fitness in the natural environment.

The consequences of ocean acidification on the biology of adult marine fish have been extensively studied. Results show a multitude of diverse responses from the death of the individual, to no effect at all. Despite this, most adult fishes can efficiently compensate for a hypercapnic acid–base disturbance, although neurosensory, growth, metabolism, and mitochondrial functions are often disrupted. Therefore, the pH compensation ability does not necessarily mean that the consequences of ocean acidification are reduced. Nonetheless, the mechanisms underneath this process are still largely unknown and predicting species responses remains difficult.

Another unexpected consequence of ocean acidification in the open ocean is the significant decrease in ocean sound absorption. Low frequency range is predicted to change significantly in the near future, mostly due to pH changes. Underwater ambient noise is thus predicted to increase, causing potential problems to human (scientific, commercial, and naval acoustic applications) and natural activities. Disturbance in animal's communication is likely to interfere with the biology and behavior of whales and other marine mammals.

Acidification is also affecting the deep ocean although the responses are delayed by the transport time of water masses between the surface and the deep areas. In those areas, the acidification can happen through a water mass movement or from the decomposition of organic matter "raining" into the deep ocean from the surface. Deep-sea ecosystems have the potential to be significantly affected by pH changes since their inhabitants are adapted to live in extremely stable conditions. The deep-sea is a cold and dark environment where metabolic activities and gas exchanges are generally reduced compared to shallow areas. Historically exceptional changes in oceans bottom pH have been documented to cause significant species extinctions and scientists are concerned that this could happen again in the near future.

The pH of many deep (below 500 m) areas is projected to decrease by 0.2 units by 2100. Carbonate-based organisms such as cold-water corals and sponge reefs are under serious threats. Animals living in the deep ocean are generally characterized by slow growth and limited recovery abilities. Acidification has been shown to negatively affect tissue functioning and membrane pumps of deep glass sponge reefs compromising their feeding efficiency. A similar reduction in feeding rates has been shown in several cold-water corals. The calcification rate of these corals is negatively affected by pH drops in laboratory experiments, but some species have been shown to be more resilient than others. The general idea is that the decrease in pH and aragonite saturation will cause the dissolution of corals skeleton. However, a number of studies suggested that deep-sea corals have some physiological mechanisms to compensate for undersaturation and several natural scleractinian reefs have been found in areas with low aragonite saturation levels.

Acidification in the Coastal Area

Intertidal areas, estuaries, coral reefs, sandy, and rocky shores are just a few examples of how diverse and variable coastal areas are. These ecosystems are submitted to continuous fluctuations due to tides, rivers inputs, waves, currents, and anthropogenic activities. Because of this variability in physical–chemical characteristics, the carbonate systems and pH experience significant fluctuations. The intense variability and the effect of multiple drivers on water pH implies that it is much more difficult to detect acidification trends in these areas compared to the open ocean.

Unlike open ocean areas, coastal ecosystems are often dominated by benthic communities and the metabolic activity of these organisms can drive important diel and seasonal pH oscillations. Coastal communities can support intense metabolic processes, including high primary production, respiration, and calcification rates. All those activities directly influence the water chemical properties such as CO_2 concentration, alkalinity, and pH. For instance, macrophytes growing in shallow coastal zones have the capacity to modify seawater pH and cause diel changes driven by primary productivity. Bivalves have been shown to be capable of raising local pH by 0.3–0.5 units during the daytime. The pH of intertidal rock pools can vary between day and night due to the antagonistic effect of respiration and photosynthesis on the water chemistry.

The presence of all these local biological processes creates a large diversity of pH niches in coastal habitats, offering areas with improved conditions for calcification. For example, calcifiers living close to macrophytes can adapt to acidification by increasing their calcification rates during daytime when CO_2 is drawn down by primary productivity. Despite the fact that this local adaptation can be valuable for adult individuals, it is possible that the pelagic larval stage, typical of many coastal calcifiers, may be critical for the species survival in a more acidic ocean.

Another important phenomena affecting coastal pH is upwelling. Upwelling is a process in which deep, cold, and nutrient rich water upwells from the bottom of the ocean to the surface. This water movement is wind driven and constantly or periodically affects several coastal areas around the word. In upwelling regions, the high concentration in nutrients promotes biological activities, but CO_2 concentrations are generally high and waters are undersaturated with respect to aragonite. The impact of acidification caused by upwelling has been demonstrated in estuarine areas where oyster larvae production was greatly reduced during strong upwelling periods.

Organisms living in coastal areas are generally tolerant to changes in pH due to the high natural variability of their habitats. However, the effect of anthropogenic CO_2 emissions is not yet fully understood for those organisms. Despite their higher tolerance to a variable environment, animals can still be affected by ocean acidification, but in a more subtle way. The physiology and development of intertidal snails are affected under acidified conditions causing a decrease in fitness and survival. Juvenile oysters show a reduction in their energy storage and a weaker shell. The immune response of intertidal mussels is disrupted by higher CO_2 conditions.

One of the more emblematic ecosystem affected by acidification worldwide are coral reefs. There is now abundant evidence that acidification and temperature changes are leading to coral bleaching and mortality. Most shallow water corals depend on an obligate relationship with a photosymbiont microalgae. The algae provides nutrients and energy to the coral facilitating his growth and calcification even in oligotrophic waters. The loss of symbionts in corals is generally called “coral bleaching” due to the impressive color change in coral tissues. When the coral bleaches, it is not dead but it becomes more vulnerable to stress and shows higher mortality rates. One of the more important causes of coral bleaching is thermal stress, but acidification is also known to favor this phenomenon. The symbiotic relationship between coral organisms and photosynthetic symbionts is often disrupted by pH changes since the symbiotic interaction is limited to a very small internal pH range.

The negative impact of acidification on corals is not limited to bleaching, but it also causes significant productivity reductions and higher rates of net dissolution of the calcified tissues. Experiments restoring preindustrial alkalinity conditions in restricted coral reef areas showed that net community calcification is highly affected by ocean chemistry. Coral calcification has been shown to be reduced of about 6%–7% since the beginning of the industrial era. In addition to coral, several other organisms living in these habitats have been shown to be particularly vulnerable to pH changes. Coral reef invertebrate recruitment and abundance are significantly lower in elevated CO_2 areas. These organisms are the key food source of several reef fish and their decline will probably entrain a cascade effect along the food chain.

Acidification in Freshwater Ecosystems

In freshwater ecosystems, the acidification process due to atmospheric CO_2 emissions is often amplified by the effect of other anthropogenic forces such as acid rain and mining drainage. CO_2 concentration in larger lakes tends to be in equilibrium with the atmosphere, but this is generally not the case for lotic systems. Biological activities, rainfalls, and stream order play an important role in regulating pH of freshwater systems. Assuming only an atmospheric CO_2 increase of about 550 μatm, the pH in Great Lakes is estimated to decline at a similar rate compared to the open ocean (about 0.19/0.09 units), but rivers and small lakes are likely to have different responses. Mining activity can cause an even higher impact with pH drops below 4 in certain systems. The release of oxidation products from the soil to the water is called acid mine drainage. This drainage causes important acidifications in hundreds of lakes and rivers situated in proximity to the mining areas.

The effects of acid rain are perhaps more flagrant. Sulfur and nitrogen oxides produced by volcanoes, industries, and other anthropic activities react with atmospheric compounds to create sulfuric and nitric acid. The acids deposit on the ground and drain into lakes and streams. North America and Europe were particularly affected by this phenomenon after the industrial revolution and new developing industrial countries such as China are now subjected to the same issue. Acidic atmospheric deposition can cause pH drops below 5 or even 4 in some freshwater systems. The effects on the aquatic ecosystems are often dramatic and the use of buffering compounds (limestone) has been employed in sensitive areas to limit the damage.

The impact of acidification on freshwater biota has been often underestimated. Only recently, some researchers focused on the effect of acidification at the species/organism level. As documented in marine systems, freshwater macrophytes and phytoplankton can be positively affected by higher CO_2 concentrations, but the response is strongly dependent on nutrient availability in the system. Studies performed on lakes historically affected by acidic atmospheric depositions showed that some freshwater zooplankton taxa are highly adaptable to acidic conditions and can survive and proliferate in a wide pH range. Following an acidification event, community diversity and structure is strongly affected but there is some evidence that ecosystems can recover when lake pH levels rise again above 6.

Another important freshwater community impacted by acidification is the bacterioplankton. Although diversity and richness were found to be generally similar between acidic and more neutral lakes, the community structure was changed. Water chemistry and acid stress have a direct influence on abundances and organization of bacterial assemblages.

In freshwater systems, mollusks, crustaceans, and many aquatic insects play the important role of breaking down the organic material to regenerate nutrients and serve as a food source for several species of fish and birds. Studies performed on several European lakes suggest that pH less than 5 might be critical for insect communities. Very little is known on the effect of pH declines on freshwater mollusks, but they are usually absent from systems with acidity below 5 probably due to an excessively high construction and maintenance cost of their calcareous structures. This is particularly true for freshwater bivalves' shells primarily made up of aragonite, since this type of shell is more soluble than the calcite one of most marine organisms. The decrease in calcium availability in acidified waters has been shown to significantly affect freshwater crustaceans that are particularly sensitive during their rapid post molt calcification of the exoskeleton. In streams, crabs, shrimps, and crayfish disappeared from anthropogenic acidified areas. Species loss among detritivorous insects and crustaceans result in a loss of the litter break down process with important consequences for the entire ecosystem.

The effect of low pH on freshwater invertebrates can be exacerbated by increased metal toxicity. Mercury concentrations are more elevated in crayfish from lower pH lakes compared to less acidified areas. Aluminum appears to be highly absorbed by invertebrates at lower pH. Nonetheless, lead toxicity is often species-specific and little information is available on potential biomagnification along the food chain.

The effect of acidified water on higher trophic levels such as fish and amphibians has received increased interest in the past few years. Under acidified conditions, the growth rate has been shown to decrease in several taxa, mostly due to a higher energetic cost of the acid–base regulation system in the fluids and tissues. The development of pink salmon during its freshwater phase has been shown to be affected by increased CO_2 levels. The growth rate and early embryonic development of this species is impaired by elevated freshwater CO_2 partial pressure. Moreover, the capacity to detect olfactory cues and avoid predators is significantly reduced under projected increases of CO_2. But, this is just an example of how acidification can induce behavioral changes in freshwater fish. Altered dial movement, behavioral changes and modified feeding patterns has been observed in several species suggesting that acidification linked behavioral changes could increase the vulnerability of these organisms to predation and food competition.

Amphibians are experiencing a general decrease in abundance due to several anthropogenic activities, one of the most important being acidification. Common species of frogs have been reported to be gradually decreasing and eventually disappearing from poorly buffered and acidified ponds and lakes. Lower pH and calcium concentration significantly limited frog reproduction and development. In dwarf newts, acidification did not inhibit females from laying eggs but the embryos were exposed to higher mortalities under low pH. A similar response has been seen on toads with up to 100% mortality in eggs maintained at a pH below 5.

All those changes in freshwater ecosystem composition and productivity have direct consequences on the higher consumers such as birds. Acidified streams have been seen to host younger and less site-faithful breeding populations of birds since they provide a lower quality habitat. Furthermore, potentially toxic metals concentration and a decreased availability of calcium, can also negatively affect birds. Piscivorous birds can suffer from strong metal bioconcentration in their prey and a reduction of calcium content in the food could affect bones and egg shells formation.

Ecological Impact

The negative effects of aquatic acidification on single species have received the most attention in recent years. However, the consequences of aquatic acidification are also extremely important at a community and ecosystem level. All organisms are embedded in a complex network of interaction between different species and populations. The impact of acidification on aquatic ecosystems is therefore inextricably linked to the impact of stressors at different tropic levels.

Predator–prey interactions can be extremely important in regulating and structuring populations and communities. Ocean acidification can lead to important indirect changes in the structure, flows, and composition of the food web. Observations from natural CO_2 enriched areas show that communities differ substantially from nearby areas with higher seawater pH. Lower diversity

and higher food web specialization are typically found in acidified areas, suggesting that acidification could induce a dramatic community shift in vulnerable systems.

Phytoplankton constitutes the foundation of aquatic food webs and regulates multiple biogeochemical processes. Acidification has been shown to have a greater impact than seawater warming or reduced nutrient supply on phytoplankton communities. About half of the global functional diversity of marine phytoplankton communities has been estimated to be altered by 2100 and this is mostly caused by ocean acidification. Freshwater phytoplankton communities are also significantly impacted by elevated CO_2 concentration since less diverse populations and smaller cells were collected in altered systems. Tolerance for low pH varied significantly between species and some authors suggested that toxic cyanobacterium strains could be more tolerant and able to proliferate under weak acidification scenarios.

Ocean acidification can also profoundly affect settlement and benthic communities. Although a wide range of species are able to settle and survive under acidified conditions, diversity is generally reduced and few taxa become dominant. This is partly due to the fact that many benthic invertebrates depend on calcified structures that are particularly sensitive to acidification. The negative impacts on the benthic community can have a cascade effect on the trophic chain and several cases of flatfish, sharks, and rays declining have been reported. Regardless of the precise conformation of the local food web, it is likely that bottom-up changes can be important. Moreover, other relatively unexplored problems such as changes in bacteria, pathogens, and parasites could also alter the functioning of the ecosystem.

Assuming that calcified structures can provide protection, function as substrate and give a wide range of other benefit to other organisms, a change in their abundance and distribution can significantly affect local communities. Coral reefs, oysters, mussels, coralline algae, and many other organisms are considered as important ecosystem engineers providing a number of ecosystem services and creating favorable habitats by modifying the environmental physical and biological conditions. Negative impacts on these organisms can cause a cascading effect on competitors and consumers. Depending on species tolerance and adaptability to acidification and climate changes, the effect on the ecosystem can be more or less important and key stone species may be replaced by more resistant ones.

Another important ecosystem alteration related to ocean acidification is the modification of carbon and nutrients biogeochemical dynamics. Variations in nutrient ratios have important effects on phytoplankton and microbial communities and this could lead to a degraded food quality for heterotrophic consumers. Changes in water chemistry could also be directly responsible for modifications in elements availability and organic matter degradation and composition. These changes can directly affect the base of the food webs and consequently the entire ecosystem.

Socioeconomical Impact

Despite the clear negative impact of acidification for the aquatic communities and ecosystem functioning, this phenomenon also has an important economic cost. The loss of economically important species such as fish and mollusks is accompanied by a decrease in aquaculture productivity and the loss of touristic and valuable habitats such as coral reefs. Moreover, any negative effect of aquatic acidification is expected to enhance the already present stressors such as overfishing and global warming.

One of the socioeconomical challenges connected to ocean acidification is the effect that this phenomenon has on seafood. Decreased pH has been shown to decrease appearance and taste of commercially important species like shrimps. Fish and mollusks are predicted to be smaller under the effects of acidification. Some commercially in demand species might decrease causing a shift in marketable goods, such as from fish to jelly fish and algae that are predicted to be more resilient to acidification.

The US oyster industry has been declining since 2005 with an average loss of $111 million per year and one of the causes is the presence of acidified waters in the area. The global economic cost of mollusk loss from ocean acidification has been estimated at about $6 billion annually. In Europe, the annual economic losses due to ocean acidification are estimated to be over $1 billion in 2100. The global value of coral reef based tourism was estimated to be $11.5 billion in 2010 and the industry is in rapid growth. The consequences of coral reefs habitat loss are expected to cost about $49–69 billion by the end of this century.

Economists estimated that the cost of the total impact of ocean acidification is in the same order of magnitude as climate change despite a high level of uncertainty still being present, due to fact that aquatic acidification has only recently begun to receive attention in international discussions. One of the main economic issues, is that ocean acidification is likely to have a greater negative impact on poorer fishing and aquaculture communities such as the small developing island states. These areas are particularly vulnerable to climate change and ocean acidification impacts, and they have fewer possibilities for alternative livelihoods.

For island nations, coral reefs loss can also constitute a substantial socioeconomic collapse. Marine natural touristic attractions such as diving, snorkeling, sightseeing, and recreational fishing will significantly suffer from ocean acidification effects. Moreover, coral reefs offer shoreline protection and support fisheries for a budget estimated at $30 billion a year.

Response Strategies to Aquatic Acidification

Limiting the effects of aquatic acidification is now critical considering the high risks of impact on natural and human systems. An effective response to aquatic acidification is likely to require a large-scale investment plan. Due to a strong connection between

aquatic acidification and the other climate change stressors, regional, national, and global strategies need to consider and manage those issues together.

The primary possible action to mitigate aquatic acidification is to reduce CO_2 emissions since no large-scale system is yet available to remove CO_2 from the atmosphere. Nonetheless, even if all CO_2 emissions were to end now, the CO_2 already released into the atmosphere will continue acidifying the ocean for centuries. It is indeed important to establish some mitigation and adaptation responses.

Ocean and freshwater acidification will not lead to the disappearance of all aquatic organisms. Some species will be able to tolerate the new conditions but ecosystems diversity and abundance are likely to change. The disappearance of economically and culturally important species may lead to extreme social forcing.

The impact of aquatic acidification on juvenile fish, food webs, and coastal habitats are likely to entrain important reductions in fishery resources. Improved managements and reduction of fishing pressure are just some examples on how we can protect and rebuild fish stocks. Other possible actions could involve the development of protection and restoration plans for particularly degraded and important areas.

In aquaculture, the impact of acidification can be limited by selecting more resilient species and employing selective breeding. The entire culture systems could also be monitored and controlled to maintain optimal water conditions for the growth of the selected species. Moreover, monitoring and response plans can be organized to warn and protect aquaculture systems from acidification and other episodic stressor events.

Nevertheless, all these actions to mitigate the effects of acidification require high financial costs and policy commitment and cannot face the extreme scenarios of global CO_2 emissions. It is therefore fundamental to take action in reducing carbon emissions to leave a greater number of effective safeguarding options to protect the marine and freshwater systems and their services to humans.

Further Reading

Dangles O, Malmqvist B, and Laudon H (2004) Naturally acid freshwater ecosystems are diverse and functional: Evidence from boreal streams. *Oikos* 104: 149–155.

Doney SC, Fabry VJ, Feely RA, and Kleypas JA (2009) Ocean acidification: The other CO_2 problem. *Annual Review of Marine Science* 1: 169–192.

Duarte CM, Hendriks IE, Moore TS, Olsen YS, Steckbauer A, Ramajo L, Carstensen J, Trotter JA, and McCulloch M (2013) Is ocean acidification an open-ocean syndrome? Understanding anthropogenic impacts on seawater pH. *Estuaries and Coasts* 36: 221–236. https://doi.org/10.1007/s12237-013-9594-3.

Hasler CT, Jeffrey JD, Schneider EVC, Hannan KD, Tix JA, and Suski CD (2018) Biological consequences of weak acidification caused by elevated carbon dioxide in freshwater ecosystems. *Hydrobiologia* 806: 1–12. https://doi.org/10.1007/s10750-017-3332-y.

Ishimatsu A, Hayashi M, and Kikkawa T (2008) Fishes in high-CO_2, acidified oceans. *Marine Ecology Progress Series* 373: 295–302.

Moiseenko TI (2005) Effects of acidification on aquatic ecosystems. *Russian Journal of Ecology* 36: 93–102. https://doi.org/10.1007/s11184-005-0017-y.

Mostofa KMG, Liu C-Q, Zhai W, Minella M, Vione D, Gao K, Minakata D, Arakaki T, Yoshioka T, Hayakawa K, Konohira E, Tanoue E, Akhand A, Chanda A, Wang B, and Sakugawa H (2016) Reviews and syntheses: Ocean acidification and its potential impacts on marine ecosystems. *Biogeosciences* 13: 1767–1786. https://doi.org/10.5194/bg-13-1767-2016.

National Research Council (2010) *Ocean acidification: A national strategy to meet the challenges of a changing ocean*. Washington, DC: National Academies Press. https://doi.org/10.17226/12904.

Phillips JC, McKinley GA, Bennington V, Bootsma HA, Pilcher DJ, Sterner RW, and Urban NR (2015a) The potential for CO_2-induced acidification in freshwater: A Great Lakes case study. *Oceanography* 28: 136–145.

Solomon S, Qin D, Manning M, Marquis M, Averyt K, Tignor M, Miller H, and Chen Z (2007) Report of the Intergovernmental Panel on climate change. In: *Climate change 2007: The physical science basis: Contribution of Working Group I to the fourth assessment report of the intergovernmental panel on climate change*. Cambrige/New York, NY: Cambridge University Press.

Tembo R (2017) The impact of ocean acidification on aquatic organisms. *Journal of Environmental & Analytical Toxicology* 07. https://doi.org/10.4172/2161-0525.1000469.

Constructed Wetlands for Wastewater Treatment[☆]

Jan Vymazal, Czech University of Life Sciences Prague, Praha, Czech Republic

Introduction

Constructed wetland treatment systems are engineered systems that have been designed and constructed to utilize the natural processes involving wetland vegetation, soils, and their associated microbial assemblages to assist in treating wastewater. Constructed wetlands must contain wetland vegetation, the treatment system with no plants are not considered constructed wetlands. They are designed to take an advantage of many of the same processes that occur in natural wetlands, but do so within a more controlled environment. Some of these systems have been designed and operated with the sole purpose of treating wastewater, while others have been implemented with multiple-use objectives in mind, such as using treated wastewater effluent as a water source for the creation and restoration of wetland habitat for wildlife use and environmental enhancement. Synonymous terms to "constructed" include man-made, engineered, and artificial wetlands.

At present, there are many different types of constructed wetlands (Fig. 1). Constructed wetlands (CWs) for wastewater treatment may be classified according to the flow regime into surface flow (SF or free water surface—FWS) and subsurface flow (SSF) systems. The FWS CWs could be further categorized according to the life form of the dominating macrophyte into systems with free floating, floating-leaved, emergent and submerged macrophytes. Within the SSF CWs it is possible to distinguish between systems with horizontal (subsurface) flow (HF or HSSF CWs) and vertical (subsurface) flow (VF or VSSF CWs). As many wastewaters are difficult to treat in a single stage system, hybrid systems that consist of various types of constructed wetlands staged in series have been introduced. In the European sense, hybrid CWs are usually formed by a combination of HF and VF systems. However, any types of CWs could be combined in order to achieve better treatment performance, especially for total nitrogen.

The first experiments aimed at the possibility of wastewater treatment by wetland plants were undertaken by Dr. Seidel in Germany in 1952 at the Max Planck Institute in Plön. However, Seidel's concept to apply macrophytes to sewage treatment was difficult to understand for sewage engineers and therefore, it was no surprise that the first full-scale free water surface (FWS) constructed wetland (CW) were built outside Germany, in the Netherlands, in the late 1960s. However, the first subsurface flow constructed wetland was built in Germany in 1974.

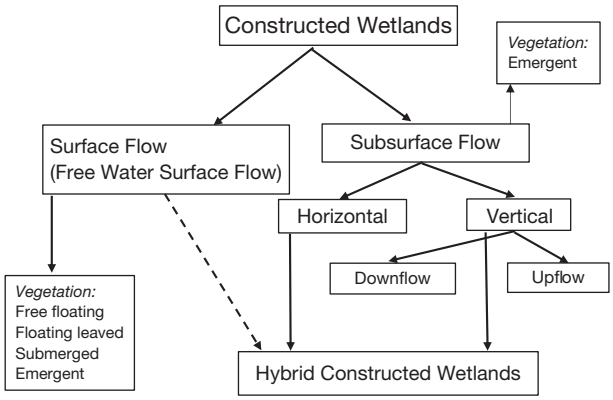


Fig. 1 Constructed wetlands classification.

[☆]*Change History:* April 2018. Vymazal updated the text and figures from the 1st edition, all Tables are new.

Free Water Surface CWs

A typical free water surface constructed wetland consists of a shallow basin constructed of soil or other medium to support the roots of vegetation (when rooting macrophytes are used) and a water control structure that maintains a shallow depth of water (Fig. 2). Flow is directed into a cell along a line comprising the inlet, upstream embankment, and is intended to proceed all portions of the wetland to one or more outlet structures. The shallow water depth, low flow velocity, and presence of the plant stalks and litter regulate water flow and, especially in long, narrow channels, ensure plug-flow conditions. FWS CWs can be classified according to the type of vegetation used (Fig. 2).

FWS CWs function as land-intensive biological treatment systems. Inflow water containing particulate and dissolved pollutants slows and spreads through a large area of shallow water. Particulates, typically measured as total suspended solids, tend to settle and are trapped due to lowered flow velocities and sheltering from wind. Most of the solids are usually filtered and settled within the first few meters beyond the inlet. While settleable organics are rapidly removed in FWS CWs by quiescent conditions, attached and suspended microbial growth is responsible for removal of soluble BOD (Biochemical Oxygen Demand, i.e., organic matter). The major oxygen source for these reactions are algae and cynobacteria growing in the water.

Nitrogen is most effectively removed in FWS systems by nitrification/denitrification. Ammonia is oxidized by nitrifying bacteria in aerobic zones, and nitrate is converted to free nitrogen or nitrous oxide in the anoxic zones near the bottom by denitrifying bacteria. Volatilization is likely as both plankton and periphyton algae grow in FWS CWs and higher pH values during the day may be favorable for ammonia loss. FWS CWs provide sustainable removal of phosphorus, but at relatively slow rates. Phosphorus removal in FWS systems occurs from adsorption, absorption, complexation and precipitation. However, precipitation with Al, Fe and Ca ions—is limited by little contact between water column and the soil. Macrophyte uptake as a removal mechanism in FWS CWs is restricted by the fact that vegetation is not regularly harvested. Also, the amount of N and P sequestered in aboveground biomass is usually quite low as compared to inflow loading (usually $<10\%$). The only exception are CWs with free floating macrophytes where harvesting in necessary for a proper function of the system. FWS CWs have been built to treat various types of wastewater (Table 1) around the world including domestic and municipal wastewater, mine drainage, urban, airport, highway and agricultural drainage, landfill leachate and variety of industrial and agricultural wastewaters.

The system with emergent vegetation is the most commonly used type of FWS CWs with *Phragmites australis* (common reed), *Typha* spp. (cattail), *Scirpus* spp. (bulrush), *Juncus* spp. (rush) and *Eleocharis* spp. (spikerush) being the most frequently used species. In other types of FWS CWs following species are commonly used:

- free floating: *Eichhornia crassipes* (water hyacinth, tropics and subtropics), *Pistia stratiotes* (water lettuce, subtropics and tropics), Lemnaceae (duckweeds, worldwide)
- floating-leaved: *Nuphar lutea* (spatterdock), *Nelumbo nucifera* (Indian lotus)
- submerged: *Ceratophyllum demersum* (coontail), *Najas guadalupensis* (southern waternymph), *Trapa natans* (water chestnut, sometimes classified as free floating), *Myriophyllum heterophyllum* (variable-leaf watermilfoil)
- floating mats: *Phragmites australis*, *Cyperus papyrus* (Papyrus), *Alternanthera philoxeroides* (Alligator weed), *Hydrocotyle umbellata* (Pennywort)

Removal of organics and suspended solids in all FWS CWs is very high while removal of nutrients is only moderate. FWS CWs provide limited contact with soil (if present) so adsorption and precipitation processes are very limited and therefore phosphorus removal mostly proceeds via soil accretion. FWS CWs provide both aerobic and anaerobic zones but neither nitrification or denitrification processes are complete. FWS CWs also provide high removal of enteric bacteria (e.g., fecal coliforms, fecal streptococci, *Clostridium perfringens*), usually in the range between one to two orders of magnitude.

Constructed Wetlands With Subsurface Flow

Horizontal Flow CWs

The most widely used concept of SSF CWs is that with horizontal subsurface flow (HF or HSSF CWs, Fig. 3). The design typically consists of a rectangular bed planted with the macrophytes and lined with an impermeable membrane. Mechanically pretreated wastewater is fed in at the inlet and passes slowly through the filtration medium under the surface of the bed in a more or less horizontal path until it reached the outlet zone where it is collected before discharge via level control arrangement at the outlet. During the passage of wastewater through the reed bed the wastewater makes contact with a network of aerobic, anoxic and anaerobic zones.

HSSF CWs require good mechanical pretreatment with suspended solids being the major target. Excessive suspended solids may cause filtration bed clogging and subsequent surface flow. Small systems for domestic sewage with low flows usually use a three-chamber septic tank. Pretreatment in systems designed for municipal sewage mostly comprise screens and Imhoff tank (sedimentation of particles, both inorganic and organic). When stormwater runoff is also treated (combined sewer system), a grit chamber (removal of inorganic particles) is included. Various types of wastewater may require different types of pretreatment. For example, landfill leachate treatment systems usually include aerated lagoons, systems for the treatment of concentrated wastewaters from agricultural operations commonly include facultative lagoons.

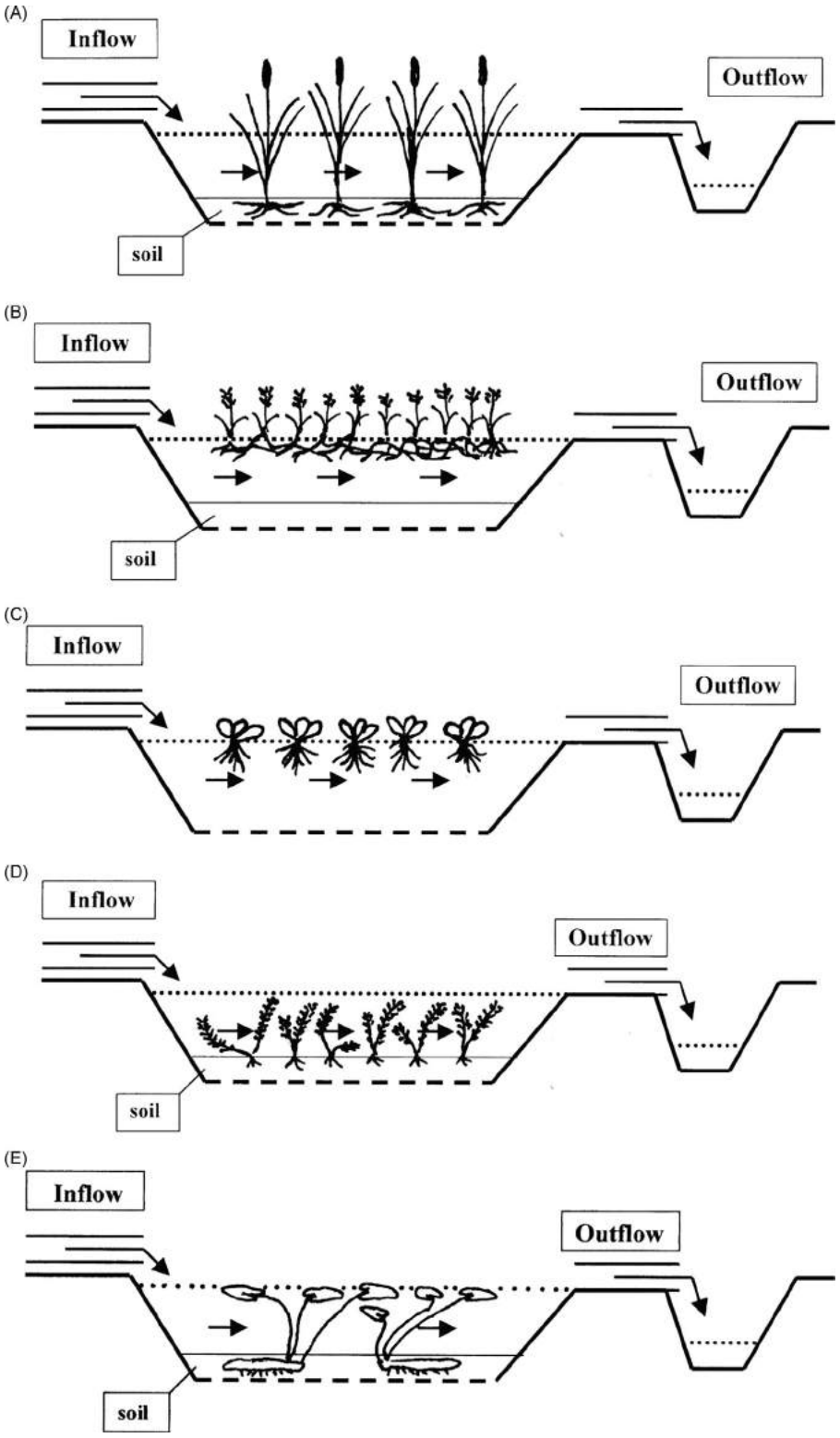


Fig. 2 Free water surface constructed wetlands. (A) With emergent vegetation, (B) with floating mats of emergent vegetation, (C) with free floating vegetation, (D) with submerged vegetation, (E) with floating-leaved vegetation (Vymazal, 2008).

Table 1 Examples of the use of FWS CWs for treatment of various types of wastewater

Vegetation	*Type of wastewater*	*Location*
FF	Municipal	United States, Cameroon, Poland, Thailand, China, Taiwan
FL	Municipal	United States, China
S	Municipal	United States, Sweden, China
	Agricultural runoff	Sweden, United States
	Urban runoff	Canada, United States
E	Municipal	All continents
	Urban stormwater	Australia, United States, United Kingdom
	Agricultural runoff	Australia, New Zealand, United States, Sweden, Denmark, Italy, Korea, Taiwan, China, Norway, Finland, Spain
	Airport runoff	Sweden, Canada
	Feedlot operations	Ireland, Canada, United States
	Mine drainage	United Kingdom, United States, Spain, South Africa, Canada, Australia, Germany, Ireland
	Refinery	United States, China, Hungary
	Pulp and paper	United States, China
	Aquaculture	United States, Taiwan
	Landfill leachate	Sweden, Norway, Canada, Poland, United States
	Food processing	Greece, Kenya, United States, Canada, Thailand, Italy, New Zealand
	Tannery	Turkey
	Woodwaste	Canada
FM	Municipal	United States, China, Turkey, Uganda, Sri Lanka, Colombia, Italy, Taiwan
	Road runoff	Belgium, United States, New Zealand
	Agricultural runoff	China
	Mine drainage	Canada
	Feedlot operations	United States

FF, free floating; *FL*, floating-leaved; *S*, submerged; *E*, emergent; *FM*, floating mats.

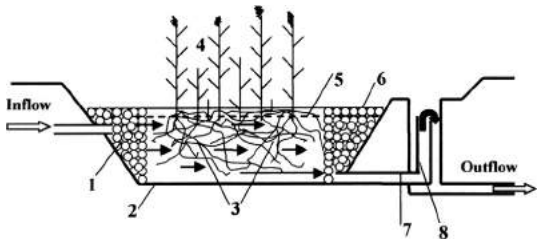


Fig. 3 Constructed wetlands with horizontal subsurface flow. 1—distribution zone filled with large stones, 2—impermeable liner, 3—filtration medium (gravel, crushed rock), 4—vegetation, 5—water level in the bed, 6—collection zone filled with large stones, 7—collection drainage pipe, 8—outlet structure for maintaining of water level in the bed. The arrows indicate only a general flow pattern. From Vymazal, J. (ed.) (2001). *Transformations of nutrients in natural and constructed wetlands*. Leiden, The Netherlands: Backhuys Publishers.

Filtration beds are filled with porous material which allows for good hydraulic conductivity in order to keep the water level below the surface as well as supports growth of macrophytes. The most commonly used filtration materials are washed pea gravel and crushed rock. The fraction size varies among countries but in general, size between 5 and 20 mm is the most common. It is recommended to use only one fraction as various fractions differ in hydraulic conductivity and short circuiting may occur. The inflow distribution and outflow collection zones are filled with large stones (ca. 50–200 mm).

The following equation, first proposed by Kickuth, is widely used for sizing of HSSF systems for domestic sewage treatment:

$$A_h = Q_d(\ln C_{in} - \ln C_{out})/K_{BOD},$$

where A_h = surface flow of bed (m^2), Q_d = average flow ($m^3\ day^{-1}$), C_{in} = influent BOD_5 ($mg\ L^{-1}$), C_{out} = effluent BOD_5 ($mg\ L^{-1}$), K_{BOD} = rate constant ($m\ day^{-1}$).

There were a lot of discussion on the K_{BOD} value. Formerly proposed value of 0.19 m day^{-1} by Kickuth resulted in too small area of the bed (about 2 m^2 per one PE, population equivalent) and consequently lower treatment effect. The field measurements in operational systems indicated that the value of K_{BOD} is usually lower (0.07–0.1 m day^{-1}). The data from 66 Danish HSSF CWs identified *K*-values for total-*N*: 0.033 m day^{-1} and total-*P*: 0.025 m day^{-1}.

The depth of vegetated beds with horizontal subsurface flow was initially based on the requirement that roots and rhizomes of the vegetation should penetrate the full depth of the bed in order to eliminate totally anaerobic zones. As the roots of the most frequently used plant, common reed (*Phragmites australis*) are capable of successful penetration to the depth of about 0.6 m and start to weaken beyond that point, the recommended bed depth was 0.6 m. Although it has been proven that oxygen transported from aboveground organs diffuses only to the thin substrate layer adjacent to the roots and rhizomes the typical depth of filtration bed is usually 0.6–0.8 m.

Constructed wetlands with horizontal subsurface flow are usually sealed in order to prevent uncontrolled water seepage into groundwater. Most of the systems use a plastic liner or membrane such as HDPE, LDPE or PVC 0.5–2 mm thick. Where local subsoil has low hydraulic conductivity (approx. 10^{-8} m s^{-1} or less), it is not necessary to use plastic liners. In order to prevent liner damage by filtration material particles, geotextile is commonly used to cover plastic liner. Also, geotextile could be also used beneath the liner.

The macrophytes growing in constructed wetlands have several properties in relation to the treatment processes that make them an essential component of the design. The most important effects of the macrophytes in HSSF CWs in relation to wastewater treatment processes are the physical effects the plant tissues give rise to such as insulation of the bed surface during the period of cold weather or provision of surface area for attached microorganisms in the filtration bed. The metabolism of the macrophytes (plant uptake, oxygen release from the roots) is of minor importance in HSSF CWs. The macrophytes for HSSF CWs should (a) be tolerant of a relatively high organic load in the wastewater, (b) have high belowground and aboveground biomass, and (c) should grow quickly in order to cover the filtration bed surface soon after planting.

The most frequently used plant around the world is *Phragmites australis*, especially in Europe, Australia, Africa and Asia. In some countries, *Phragmites* is used exclusively; for example, in the United Kingdom, where HSSF CWs are called Reed Bed Treatment Systems. On the other hand, in New Zealand or many areas of the Unites States, *Phragmites* is considered an exotic and invasive species by natural resource agencies and as a result, use of this species has been limited. Other commonly used species are *Phalaris arundinacea* (reed canarygrass), *Glyceria maxima* (sweet mannagrass) or various cattails (*Typha latifolia*, *T. angustifolia*) in Europe, *Cyperus papryrus* (papyrus) in Africa, *Typha domingensis* in South America and *Scirpus* spp. (bulrush) in North America. Weeds (plants that were not intentionally planted) occur mostly within vegetated bed margins and do not have any detrimental impact on treatment performance.

Dissolved oxygen supply is very limited in filtration beds of HSSF CWs and, therefore, anoxic and anaerobic processes usually prevail. Aerobic degradation is restricted to narrow zones adjacent to roots and rhizomes where oxygen leaks to the rhizosphere. Removal of organics (BOD_5, COD) is usually high and exceeds 85% in case of sewage. Suspended solids that are not removed in pretreatment system are effectively removed by filtration and settlement. Most suspended solids are filtered out and settled within the first few meters beyond the inlet zone. Suspended solids are removed in HSSF CWS were effectively, commonly >90%. The accumulation of trapped solids is a major threat for good performance of HSSF systems as the solids may clog the bed. Therefore, the effective pretreatment is necessary for HSSF systems. However, it has been shown that if HSSF CWs are appropriately loaded, that is, <10 g BOD_5 m^{-2} day^{-1}, <20 g COD m^{-2} day^{-1}, <10 g TSS m^{-2} day^{-1}, the partial clogging in the inflow zone occurs only after about 15 years. In addition, partial clogging has no effect on treatment performance of the system.

The major removal mechanism of nitrogen in HSSF constructed wetlands is nitrification/denitrification. Field measurements have shown that the oxygenation of the rhizosphere of HSSF constructed wetlands is insufficient and, therefore, incomplete nitrification (i.e., oxidation of ammonia to nitrate) is the major cause of limited nitrogen removal. In general, nitrification which is performed by strictly aerobic bacteria is mostly restricted to areas adjacent to roots and rhizomes where oxygen leaks to the filtration media. On the other hand prevailing anoxic and anaerobic conditions offer suitable conditions for denitrification but the supply of nitrate is limited as the major portion of nitrogen in sewage is in the form of ammonia. Adsorption and plant uptake play a much less important role in nitrogen removal in HSSF CWs. Volatilization is not effective as there is no free water surface and adsorption is greatly limited by the fact that filtration media (gravel, rock) do not provide suitable sorption sites. Plant harvesting contributes to an overall nitrogen removal only marginally (usually <10% of the inflow load) with N standing aboveground stocks in the range of 20–60 g N m^{-2}. In case of sewage, removal of ammonia and total nitrogen usually does not exceed 50%.

Phosphorus is removed primarily by adsorption and precipitation, however, media used for HSF wetlands (e.g., pea gravel, crushed stones) usually do not contain great quantities of Fe, Al or Ca and therefore, removal of phosphorus is generally low (<40% in sewage). Removal of phosphorus could be enhanced by the use of filtration media with high sorption capacity but the sorption capacity is always saturable and therefore, the media must be replaced after saturation in order to maintain high P removal. Removal via harvesting accounts usually for <5% of the inflow load with the P standing stock in the aboveground biomass in the range of 3–6 g P g m^{-2}. The removal of microbiological pollution is very seldom the primary target for constructed treatment wetlands. However, HSSF CWs are known to act as excellent biofilters through a complex of physical, chemical and biological factors which all participate in the reduction of the number of bacteria of anthropogenic origin.

HSSF CWs are used for many types of wastewater around the world (Table 2). Indeed, the most common use is for municipal and domestic sewage, however industrial and agricultural wastewaters have been successfully treated as well. Besides that, applications for various types of stormwater runoff (e.g., urban, highway, agricultural, golfcourses, nurseries, airports) and landfill leachate have been put in operation.

Table 2 Examples of the use of HF CWs for various types of wastewaters

Wastewater	*Location*
Municipal	Worldwide
Petrochemical	United States, Taiwan, China, Sudan, Oman
Pulp and paper	United States, Kenya, India
Tannery	Portugal, Tanzania
Textile	Australia, Slovenia, Tanzania
Abbatoir	Australia, Mexico, Ecuador, Uruguay, New Zealand
Food processing	Italy, Spain, Slovenia, United States, France, Lithuania, Turkey
Distillery, winery	Italy, Spain, United States, India, South Africa, Mexico, United Kingdom
Feedlot operations	Australia, China, United States, Canada, Thailand, Lithuania
Fish farms	United States, Germany, Canada
Dairy	Italy, Lithuania, Germany, United States, United Kingdom, New Zealand
Highway runoff	United Kingdom, Italy, United States
Airport runoff	United Kingdom, Switzerland, United States, Germany
Agricultural runoff	China, New Zealand
Landfill leachate	United States, Portugal, Norway, Poland, Slovenia

Vertical Flow CWs

Constructed wetlands with vertical subsurface flow (VF or VSSF CWs) usually comprise a flat bed of coarse sand or gravel planted with macrophytes (Fig. 4). The most important factors in the design of a VF CWs are: (1) to produce a bed matrix that allows the passage of the wastewater through the bed before the next dose arrives whilst at the same time holding the liquid back long enough to allow the contact with the bacteria growing on the media and achieve the required treatment. (2) To provide sufficient surface area to allow the oxygen transfer to take place and sufficient bacteria to grow. VF CWs were proposed during the 1960s but did not spread as quickly as HSSF CWs, probably because of higher operation and maintenance requirements. However, the increased demand for nitrogen and especially ammonia'nitrogen removal in the 1990s revived this type of constructed wetlands.

All VF systems are dosed intermittently, however, there is no clear recommendation how many batches per day are optimal. It is essential to achieve quick water cover of the surface in order to trap air in the interstices in the bed. In most VF systems the real distributor is the layer of carefully selected sand which first allows flooding of the surface and then gradual seepage down through the depth of the media. The vast majority of the VF systems employ a network of pipes with small holes across the surface area of the bed. The distribution pipes could be insulated by a 0.2 m layer of coarse wood chips or sea shells on the surface of the filter. It is also possible to distribute wastewater from open-ended pipes onto the bed. The area of the immediate vicinity of the discharge should be protected by some paving or tiles to prevent the wash away of the sand or gravel.

The data on maximum HLR vary widely in the literature. However, it seems that VF CWs can operate in the range of 100–1200 mm day^{-1} and usually no clogging problems occur below 800 mm day^{-1} of pretreated wastewater. The organic loading rate should not exceed 25 g COD m^{-2} day^{-1} in order to prevent clogging. The area used for VF CWs varies between 0.9 and 5 m^2 PE^{-1} (PE = population equivalent) but most systems are design with the specific area 2–3 m^2 PE^{-1} stems are usually designed as a single unit, larger systems may have several beds which are fed with wastewater in rotation. Also, some larger systems have two stages of vertical beds in operation.

The size of VF CWs is usually based on the hydraulic loading rate (HLR). The maximum HLR that can be achieved without surface flooding will be affected by many variables but is most strongly related to media size and distribution, rate of biofilm growth and hence the BOD_5/organic mass loading rate and suspended solids loading rate. The Danish guidelines recommend specific area of 3.2 m^2 per person, maximum organic loading of 18.8 g BOD_5 m^2 day^{-1} and hydraulic loading of 47 mm day^{-1}. Filtration medium is sand with a d_{10} between 0.25 and 1.2 mm, a d_{60} between 1 and 4 mm, and uniformity coefficient ($U = d_{60}/d_{10}$) should be <3.5. The contents of clay and silt (particles <0.125 mm) must be <0.5%. Austrian guidelines recommend the specific area of 4 m^2 per person and maximum loading of 20 g COD m^{-2} day^{-1}. The filtration layers should consist of 5–10 cm of gravel (4/8 or 8/16 mm) on top, followed by 50 cm of washed sand (0–4 mm), 5–10 cm transition layer (gravel 4/8 mm) and 20 cm drainage layer at the bottom (gravel 8/16 or 16/32 mm). This set-up guarantees outflow concentrations <90 mg L^{-1} COD, 25 mg L^{-1} BOD_5 and 10 mg L^{-1} $N\text{-}NH_4$ at water temperature > 12°C. In France, the VF systems consist of two beds in series with only 1.2 m^2 per person in the first stage and 0.8 m^2 per person in the second stage. The pretreatment consists only of prescreening, maximum organic load is 100 g COD m^2 day^{-1} and maximum hydraulic loading is 90 cm day^{-1}.

Plants play a very important role in VF CWs. They stabilize surface of the bed, their roots and rhizomes positively affect the hydraulic conductivity of the filter and movement of aboveground stems helps to prevent clogging. The aboveground biomass provides insulation of the bed and belowground organs provide substrate for attached bacteria growth. The oxygen transfer to the rhizosphere is limited but creates microzones where aerobic bacteria can be present. The vast majority of VF CWs is in operation in Europe and the most commonly used plant is *Phragmites australis* (common reed).

VF CWs are very effective in removal of organics, suspended solids and ammonia. The intermittent feeding allows for regular emptying the filtration bed which results in good oxygenation of the bed allowing for nitrification. Therefore, VF CWS are used in

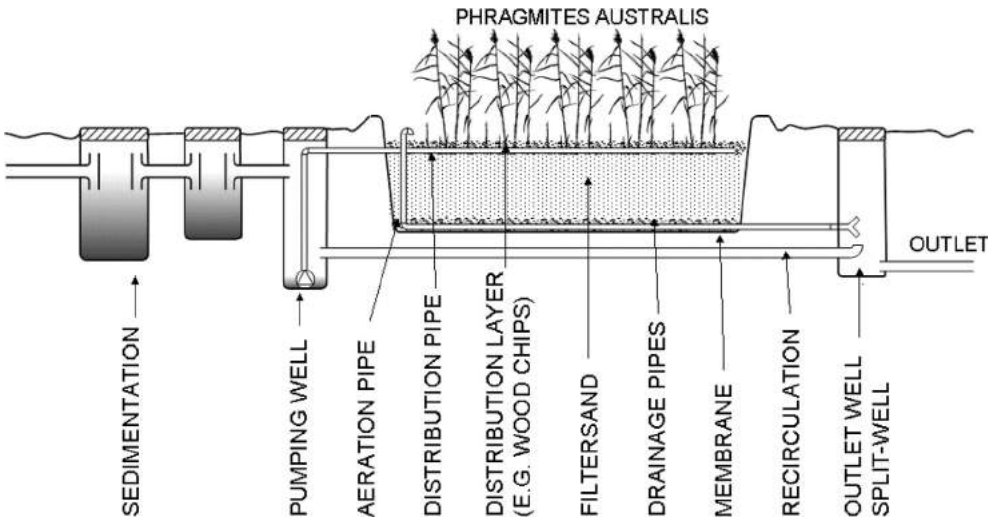


Fig. 4 Layout of a vertical flow constructed wetland system for a single household. Raw sewage is pretreated in a 2 m^3 sedimentation tank. Settled sewage is pulse-loaded onto the surface of the bed by a level-controlled pump. Treated effluent is collected in a system of drainage pipes, and half of the effluent is recirculated back to the pumping well (or to the sedimentation tank). From Brix, H. (2005). The use of vertical flow constructed wetlands for on-site treatment of domestic wastewater: New Danish guidelines. *Ecological Engineering* **25**, 491–500.

Table 3 Examples of the use of VF CWs for various types of wastewater

Wastewater	*Location*
Municipal, domestic	France, Belgium, Austria, Poland, Czech Republic, Italy
Landfill leachate	Norway, New Zealand
Abbatoir	Canada
Refinery	Pakistan
Airport runoff	United States, Canada
Textile	United Kingdom, Portugal, Japan
Aquaculture	United States, China, Canada, Vietnam
Winery	Germany
Olive mill	Turkey, Greece
Chemical industry	Portugal
Steel industry	China

case when ammonia is the target of the treatment. On the other hand, due to oxic conditions in the bed, the denitrification is limited and often missing in the system. VF CWs are commonly used for treatment of domestic and municipal sewage but the use for other types of wastewater is common (Table 3).

Hybrid Constructed Wetlands

In hybrid constructed wetlands (CWs), the advantages of various systems can be combined to complement each other. Hybrid constructed wetlands were first introduced by Seidel in Germany as early as in the 1960s. The design consisted of two stages of several parallel vertical flow (VF) beds followed by two or three horizontal flow (HF) beds in series. The VF stages were usually planted with *Phragmites australis*, whereas the HF stages contained a number of other emergent macrophytes, including *Iris*, *Schoenoplectus* (*Scirpus*), *Sparganium*, *Carex*, *Typha* and *Acorus*. The VF beds were loaded with pretreated wastewater for 1–2 days, and were then allowed to dry out for 4–8 days. In this system, nitrification, that is, oxidation of ammonia to nitrate, takes place in the VF stage and denitrification of nitrate, that is, reduction of nitrate do N_2O and N_2, proceeds in the HF stage. The oxidation of ammonia in intermittently loaded VF stage is very high but the concentration of organics in the second stage may not be high enough to support full denitrification. In the early 1980s, several hybrid systems of Seidel's type were built in France and similar system was built in 1987 in the United Kingdom. During the 1990s and the early 2000s, VF–HF systems were built in many countries in Europe, for example, Austria, Slovenia, Norway or Ireland. The VF–HF hybrid constructed wetlands were mostly designed to treat domestic or municipal wastewater where nitrified effluents were required but there were also application for other types of wastewater (Table 4).

Table 4 Examples of hybrid constructed wetlands used for various types of wastewaters

Type of hybrid system	*Location*	*Wastewater*
VF–HF	Belgium, Estonia, Tunisia, Spain, China, Italy, Brazil, France	Sewage
	Italy	Cheese production
	Slovenia	Landfill leachate
	Spain	Winery
HF–VF	Mexico, Poland, France, Italy, Nepal, China, Turkey	Sewage
VF–VF–HF	Poland	Slaughter
	France	Cheese production
	Japan	Milking parlor
VF–VF–HF–VF	Japan	Milking parlor
VF–VF–VF–HF–VF	Japan	Starch production
	Japan	Pig urine
VF–HF–VF	Morocco, Turkey	Sewage
	Bangladesh	Tannery
HF–VF–HF	Poland, New Zealand	Sewage
	Korea	Greenhouse
	Slovenia	Mixed industrial
VFd–VFu	China	Sewage
	China	Aquaculture
FWS in combination with HF and VF	China. Italy, Thailand, Mexico, Taiwan	Sewage
	Italy	Winery
	Taiwan	Aquaculture
	Thailand	Fish industry
	Canada	Landfill leachate
	Kenya	Flower plant

In the late 1990s, also HF–VF constructed wetlands were introduced. This system consisted of a large HF bed placed first and a small VF bed as the second stage. In this system, nitrification takes place in the vertical flow stage at the end of the process sequence. If nitrate removal is needed it is then necessary to recirculate the effluent back to the front end of the system where denitrification can take place in the less aerobic horizontal flow bed using the raw feed as a source of carbon needed for denitrification. The HF–VF CWs has been used exclusively for treatment of sewage so far (Table 4). Besides VF–HF and HF–VF systems other combinations of subsurface flow CWs and also combinations including FWS CWs were used in the 1990s and the early 2000s for treatment of various types of wastewater (Table 4).

Conclusions

Constructed wetlands represent an alternative treatment system to conventional treatment systems such as activated sludge process. All types of constructed wetlands exhibit high treatment efficiency for organics and suspended solids. The removal of these parameters is comparable with conventional systems. Removal of nitrogen depends on the type of constructed wetlands and nitrogen species involved. Ammonia is efficiently removed in vertical flow CWs while nitrate is removed efficiently in HF CWs. However, combination of various types of CWs (usually VF and HF CWs) can enhance removal of total nitrogen and then the efficiency is comparable with conventional systems. Removal of phosphorus is variable depending on the filtration material, however, commonly used materials do not support high phosphorus removal. The advantage of constructed wetlands as compared to conventional treatment systems is low operation and maintenance cost, in some countries also the investment cost is substantially lower than of conventional systems, ability to treat low strength wastewaters and no requirement of continuous feeding and operation.

Further Reading

Brix H (2005) The use of vertical flow constructed wetlands for on-site treatment of domestic wastewater: New Danish guidelines. *Ecological Engineering* 25: 491–500.
Hammer DA (ed.) (1989) *Constructed wetlands for wastewater treatment*, Chelsea, Michigan: Lewis Publishers.
Kadlec, R.H., Wallace, S.D. *Treatment wetlands*, 2nd edn. Boca Raton, Florida: CRC Press.
Kadlec, R.H., Knight, R.L., Vymazal, J., Brix, H., Cooper, P.F. and Haberl, R. (2000). *Constructed wetlands for pollution control. Processes, performance, design and operation*. IWA scientific and technical report no. 8. London: IWA Publishing.
Moshiri GA (ed.) (1993) *Constructed wetlands for water quality improvement*, Boca Raton, Florida: CRC Press/Lewis Publishers.
Mulamoottil G, McBean EA, and Rovers F (eds.) (1999) *Constructed wetlands for the treatment of landfill leachates*. Boca Raton, Florida: CRC Press/Lewis Publishers.
Reddy KR and Smith WH (eds.) (1987) *Aquatic plants for water treatment and resource recovery*, Orlando, Florida: Magnolia Publishing.

Reed SC, Middlebrooks EJ, and Crites RW (1988) *Natural systems for waste management and treatment*. New York: McGraw-Hill.
Seidel K (1965) Neue Wege zur Grundwasseranreicherung in Krefeld, Vol. II. Hydrobotanische Reinigungsmethode. *GWF Wasser/Abwasser* 30: 831–833.
Vymazal J (ed.) (2001) *Transformations of nutrients in natural and constructed wetlands*. Leiden, The Netherlands: Backhuys Publishers.
Vymazal J (2007) Removal of nutrients in various types of constructed wetlands. *Science of the Total Environment* 380: 48–65.
Vymazal J (2008) Constructed wetlands, subsurface flow. In: Jørgensen SE and Fath BD (eds.) *Encyclopedia of ecology*, 1st edn, pp. 749–764. Amsterdam: Elsevier B.V.
Vymazal J (2008) Constructed wetlands, surface flow. In: Jørgensen SE and Fath BD (eds.) *Encyclopedia of ecology*, 1st edn, pp. 765–777. Amsterdam: Elsevier B.V.
Vymazal J (2011) Constructed wetlands for wastewater treatment: Five decades of experience. *Environmental Science and Technology* 45(1): 61–69.
Vymazal J (2014) Constructed wetlands for treatment of industrial wastewaters: A review. *Ecological Engineering* 73: 724–751.
Vymazal J and Kröpfelová L (2008) *Wastewater treatment in constructed wetlands with horizontal sub-surface flow*. Dordrecht: Springer.
Vymazal J, Brix H, Cooper PF, Green MB, and Haberl R (eds.) (1998) *Constructed wetlands for wastewater treatment in Europe*, Leiden, The Netherlands: Backhuys Publishers.

Dead Zones: Low Oxygen in Coastal Waters

Andrew Altieri, University of Florida, Gainesville, FL, United States

Glossary

Anoxia Absence of dissolved oxygen.
Benthos/benthic The community of organisms in or near the bottom of the ocean.
Dead zone Oxygen-depleted coastal waters that are hypoxic with compromised biological communities.
Eutrophication Excess nutrients in a body of water.
Hypoxia Low dissolved oxygen. The exact threshold is context dependent, but is typically regarded as a concentration < 2.8 mg/L.
Hysteresis The existence of multiple ecological states under a given set of environmental conditions, where thresholds between states differ depending on prior state.
Residence time The average time for the water in a body of water to be renewed.
Stratification Layering of the water column into water masses with different properties.
Suspension feeders Organisms that that feed by filtering food from water.
Water column A section of water extending from the surface to bottom.

Introduction

Oxygen depletion in coastal waters is regarded by many marine biologists and oceanographers as the most pressing water pollution problem in the world because of its severe impacts and accelerating spread worldwide. In extreme cases, this hypoxia creates areas known as "dead zones" that are largely devoid of macrofauna because of mortality and emigration. Concentrations of dissolved oxygen are naturally variable, and this variation can lead to areas with low oxygen concentrations (hypoxia) that are stressful to marine life. However, the duration, severity, and number of ecosystems with hypoxia has increased due to anthropogenic factors including inputs of excess nutrients and organic matter, as well as climate change, and there are now hundreds of coastal ecosystems affected by hypoxia worldwide (Fig. 1).

This article begins by describing and defining dead zones, and then provides background on factors that drive the formation of hypoxic waters including anthropogenic inputs and natural processes. Organismal responses to low oxygen are then presented as a basis for understanding community and ecosystem impacts of hypoxia. The relevance of dead zones to human welfare are provided in the context of ecosystem services, and followed by evidence for the increasing spread of dead zones worldwide. Finally, the article provides background on emerging areas of hypoxia research and possible solutions to address the threat of dead zones.

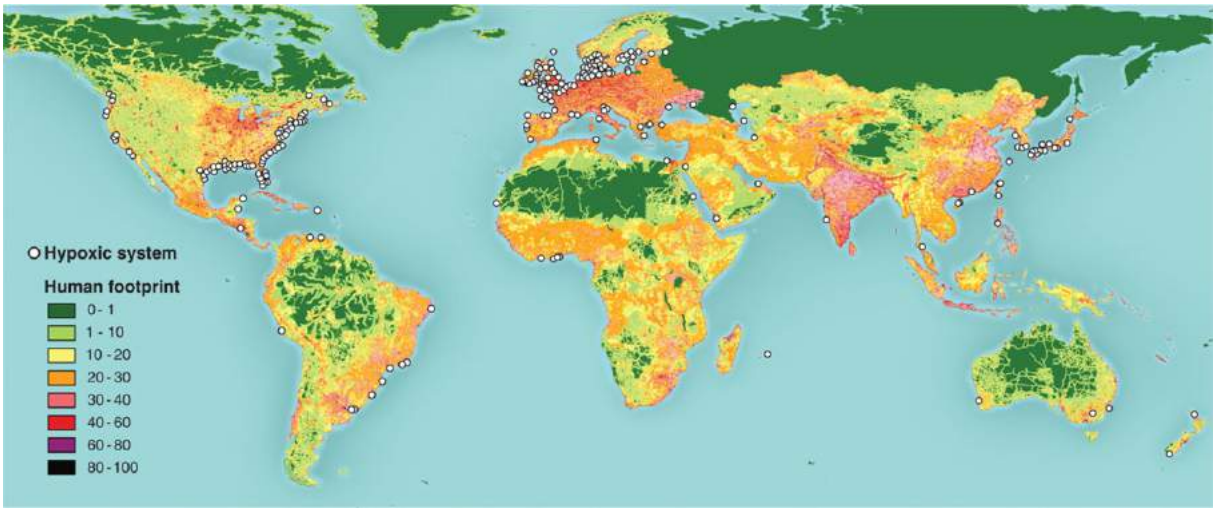


Fig. 1 Global map of hypoxic ecosystems. Dead zones are typically associated with areas of intense human activity (or "human footprint") in the coastal zone. Most known dead zones are located in the temperate regions, but recent research suggests the number of hypoxic ecosystems in the tropics are greatly underestimated. From Diaz, R.J. and Rosenberg, R. (2008). Spreading dead zones and consequences for marine ecosystems. *Science* **321**, 926–929.

 doi:10.1016/B978-0-12-409548-9.10616-5

Dead Zones Defined

The depletion of oxygen in coastal waters, such as estuaries, continental shelf areas, and coastal seas, is now a widespread phenomenon. Low oxygen typically occurs close to the sea floor and in deeper portions of stratified water columns where rates of oxygen consumption are likely to be highest and replenishment lowest. Dead zones can range in size from a small canal or channel to the basin of inland seas such as the Baltic Sea which commonly covers $>60{,}000\ km^2$ (Fig. 2). Dead zones also vary in duration. Some have become permanent features, as found in the Baltic Sea, whereas others occur seasonally, such as the Chesapeake Bay, United States. They can also vary in spatial extent and severity year-to-year based on freshwater inputs and external conditions as in the Gulf of Mexico.

Fish kills, in which dead fish and other marine organisms accumulate on the water's surface and on beaches, are one of the most visible signs of hypoxia. Another conspicuous phenomenon is the "jubilees" in Mobile Bay, United States where hypoxic bottom waters drive fish and crabs into the shallows where they are collected by local residents in a seafood bonanza. However, these events represent just a hint of hypoxia's impacts and prevalence in coastal ecosystems since oxygen measurements require specialized sensors, and the associated biological effects typically occur out of sight below the ocean's surface.

Oxygen depletion is a stress for aerobic organisms that can lead to physiological changes, behaviors responses, and eventually mortality. Hypoxia can, therefore, have a variety of direct and cascading effects on community structure and ecosystem function. In the most extreme cases, hypoxic ecosystems become largely devoid of living macrofauna, hence the name "dead zone." This term originated in the popular press which reported on hypoxic areas that fishers referred to as "dead water" for lack of any finfish or shellfish to catch. In the decades since the term "dead zone" was first coined, it has become more generally applied to any hypoxic ecosystem, and the term is used as such in this article. While there is some disagreement about whether the term "dead zone" is appropriate since microbes and a few other tolerant functional groups can persist in severely hypoxia waters, it does serve as a useful nontechnical shorthand that conveys the seriousness of the phenomenon for the living systems on which humans are so dependent.

Hypoxia in Other Ecosystems

The scope of this article is limited to hypoxia in coastal waters including bays, estuaries, coastal and inland seas, and some shelf systems. However, hypoxia also occurs in other bodies of water with important consequences for the ecology of those systems. For

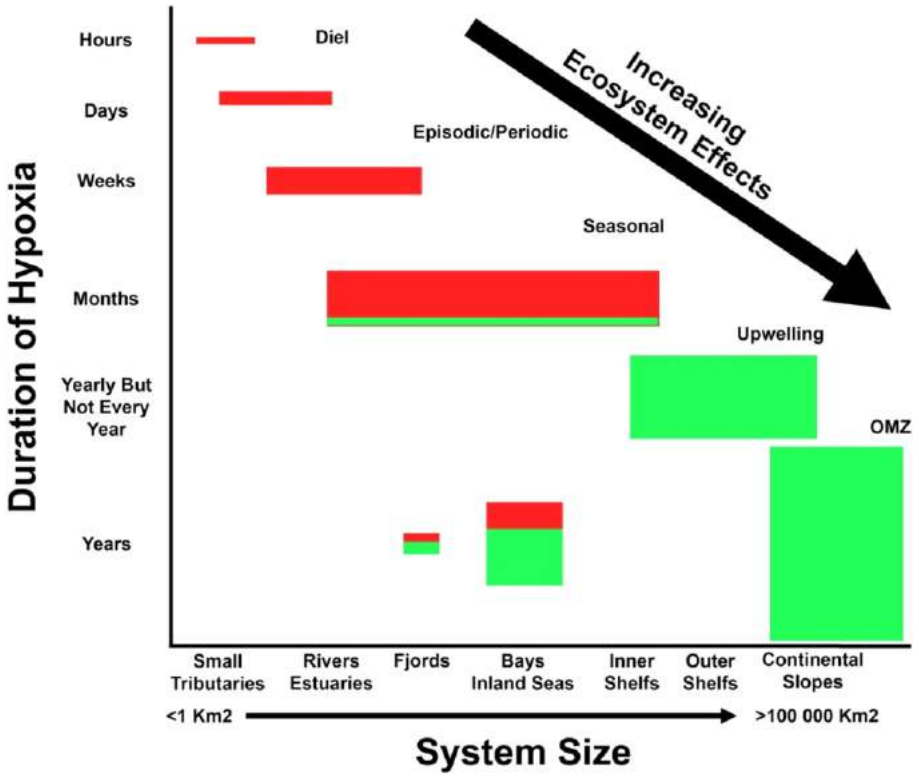


Fig. 2 Schematic representing the general relationship between ecosystem size and duration of hypoxia. The relative contribution of human factors in driving oxygen dynamics is represented in *red*, whereas natural factors are in *green*. This article focuses on those systems in *red*. From Rabalais, N.N., Diaz, R.J., Levin, L.A., Turner, R.E., Gilbert, D. and Zhang, J. (2010). Dynamics and distribution of natural and human-caused hypoxia. *Biogeosciences* **7**, 585–619.

example, freshwater lakes and ponds commonly become hypoxic because of eutrophication and/or seasonal stratification of the water column. There are also large permanently hypoxic areas in deeper ocean basins known as oxygen minimum zones (OMZ) that are a natural phenomenon with diverse adapted communities (Fig. 2). At times, hypoxic waters from the OMZ can be upwelled onto the shelf into shallow coastal waters such as along the western coasts of the United States and South Africa. There are important distinctions between OMZs and dead zones: OMZs are generally considered natural and permanent features (although their expansion appears linked to large-scale climate and oceanographic processes), whereas coastal dead zones can vary in duration and are often associated with eutrophication and other anthropogenic factors.

Quantifying Hypoxia

Hypoxia is generally defined as an oxygen concentration in water that is low relative to either the oxygen saturation at equilibrium with atmospheric concentrations or the metabolic needs of aerobic organisms. As such, dissolved oxygen levels are typically reported as either percent saturation or concentration (mg or mL of oxygen per liter of seawater). Percent saturation is often useful for understanding the drivers of hypoxia because it indicates the severity of oxygen depletion relative to that expected at equilibrium concentrations. Absolute oxygen concentrations (mg/L or mL/L) are more frequently used in studies of biological responses to hypoxia. Saturation and concentration units are not directly interchangeable because temperature, barometric pressure, and salinity affect the saturation of oxygen in water and must be known to derive percent saturation from absolute oxygen concentration.

Concentrations of dissolved oxygen vary continuously, and thresholds of tolerance to hypoxia vary among species. Nevertheless, 2.8 mg/L (equivalent to 2.0 mL/L) has been widely adopted as the oxygen concentration below which a body of water is considered hypoxic. Below that threshold, negative effects of oxygen limitation including mortality are likely to be observed across most taxa. There are proponents for elevating the threshold since fish often show sensitivities to oxygen at the 5–6 mg/L range, and many macrofauna exhibit behavioral changes when oxygen levels drop to the 3–4 mg/L range. For comparison, 100% oxygen saturation in water at 25°C and 1 atm of pressure is 8.2 mg/L.

The oxygen dynamics and ecological effects of oxygen limitation in dead zones are often measured and described in the water column and in the benthos. The benthos in many dead zones is comprised of soft sediments, and the chemical properties (including oxygen concentrations) and biological activity of those sediments are closely linked to properties of the water column. A comprehensive understanding of dead zones, therefore, incorporates characteristics of the sediments, including irrigation by burrowing organisms and the depth of the redox potential discontinuity which indicates the depth within the sediment that transitions from oxic to anoxic layers.

Causes of Dead Zones

Oxygen concentrations decline in dead zones because rates of oxygen depletion outpace rates of oxygen replenishment. This section of the article discusses the origins of low oxygen waters. First, the general processes that deplete oxygen are described with a distinction between natural and anthropogenic factors. Second, the factors that limit oxygen replenishment are outlined. Finally, a discussion is presented as to why some coastal waters are especially susceptible to oxygen depletion.

Oxygen Depletion

Decomposition and respiration

Oxygen consumption by aerobic organisms is the dominant cause of oxygen depletion in coastal dead zones. There are two main processes of biological oxygen consumption that can lead to low oxygen concentrations in coastal waters.

Decomposition by microbes is the primary process of oxygen consumption that leads to the formation of dead zones. As such, factors that increase the amount of organic matter available for decomposition and/or increase rates of microbial activity are ultimately responsible for hypoxia. Many dead zones are linked to phytoplankton blooms. The biomass of plankton that dies and sinks to the bottom fuels aerobic microbial decomposition. Macroalgal blooms can also generate organic matter that fuels hypoxic conditions. Moreover, macroalgae can further increase oxygen stress by smothering the bottom.

Allochthonous organic matter, such as detritus, sediments, and sewage, delivered to coastal waters can bypass the role of algal blooms and lead to microbial oxygen depletion. If oxygen levels drop sufficiently to cause mortality, then the biomass of dead organisms becomes a supplementary source of organic matter that can fuel further decomposition in a localized feedback loop.

A second process of oxygen depletion is respiratory oxygen consumption by primary producers such as phytoplankton and macroalgae. Oxygen consumption by primary producers will outpace their oxygen production in light limited conditions such as those produced by self-shading or at night, leading to temporary or localized oxygen depletion. One of the most conspicuous indicators of oxygen consumption by primary producers is diel cycling in which oxygen concentrations exhibit a saw tooth pattern as a system oscillates between net photosynthetic oxygen production during the day and net respiratory oxygen consumption during the night.

An additional factor leading to oxygen depletion in coastal waters is oxygen consumption by macrofauna. This generally does not lead to large or sustained dead zone formation per se, but can exacerbate the effects of existing low oxygen conditions. For

example, fish that have been shoaled by low oxygen conditions into a semienclosed area can become trapped and rapidly consume remaining oxygen, resulting in a fish kill.

Nutrients

Many studies have identified excess nutrients as the factor most responsible for the formation of dead zones worldwide. As a consequence, hypoxia is viewed as one of the most significantly negative impacts of eutrophication. Since primary productivity by phytoplankton is typically nutrient limited, inputs of nutrients can support the excess phytoplankton biomass responsible for hypoxia. Typically, primary productivity in marine ecosystems is nitrogen limited, and in lower salinity ecosystems is phosphorus limited.

There are several pathways that deliver nutrients and organic matter to coastal waters. Rivers and terrestrial run-off are the dominant sources of nutrients and organic matter for many coastal ecosystems, but atmospheric deposition, upwelled ocean water, and sewage systems play significant roles in some coastal waters. The availability of nitrogen and phosphorus is also be modified within a dead zone by microbial activity and bioturbators in the sediment. The balance between sequestration and release of nutrients from the sediment is dependent on the oxygen concentration, with the potential for feedbacks between oxygen, nutrients, and primary producers.

Inputs of nutrients and organic matter that fuel dead zone formation in coastal waters are often elevated above natural rates by human activity. While there are natural sources such as erosional processes and run-off from terrestrial ecosystems, anthropogenic sources have dramatically increased nutrients above natural background levels and contributed to the prevalence of hypoxia. For example, over the past century there was a 20-fold increase in nitrogen inputs from the Mississippi River that fuels the Gulf of Mexico dead zone. Much of these nutrients and organic matter originate from agricultural sources including excess fertilizer from fields and livestock waste. Additional anthropogenic sources include effluent from sewage treatment plants and household septic systems, urban run-off, paper mills, seafood processing plants, and aquaculture pens. The relative importance of these various sources of nutrients and organic matter is location specific, varying with land-use practices, regulations, and human densities in the watershed.

There are two overall trends that help establish the contribution of human activity to the increase in dead zone prevalence. First, hypoxia is more prevalent near major population centers and watersheds with intensive inputs of nutrients and organic material. Second, there is a temporal trend in which sewage and industrial pollution were dominant drivers of hypoxia through the mid-20th century followed by a shift toward greater run-off/riverine inputs and atmospheric deposition of nutrientsThat coincided with intensification of agricultural practices and an overall increasing rate of nutrient delivery that caused the number of new hypoxic ecosystems to increase at an accelerated rate.

Warming

Temperature has a direct effect on oxygen concentrations since warmer water has a lower capacity for dissolved gasses including oxygen. Temperature also has indirect effects on oxygen depletion since phytoplankton blooms and microbial activity are temperature dependent. This explains why hypoxia is most likely to occur in the summer, when temperature and light levels peak, and why climate change is such a concern for dead zones (as detailed below).

Limited Oxygen Replenishment

Replenishment of oxygen in hypoxic waters occurs through two primary mechanisms: vertical exchange with surface waters that are oxygenated through diffusion and wind mixing, and lateral exchange with the more oxygenated open ocean. Therefore, hypoxia and dead zones typically occur in bodies of water that are isolated by shoreline features and/or stratification.

Stratification and vertical mixing

Hypoxia typically develops at or near the bottom where the organic matter that fuels decomposition accumulates. Stratification of the water column isolates hypoxic bottom waters and prevents surface waters from re-oxygenating the entire water column. Stratification refers to layering of the water column into discrete water masses with different densities due to variation in temperature and/or salinity. Often stratification in dead zones results in two layers: a shallow low-density surface layer that may be well oxygenated from diffusion and wind-mixing of atmospheric oxygen, and a second, deeper, high-density layer where organic matter settles and oxygen is depleted by microbial respiration.

Stratification, and the potential for hypoxia to occur, is strongest when there are processes that establish or reinforce the density difference between surface and bottom waters. Freshwater inputs from rivers and/or rain can create a surface lens of freshwater which is less dense than deeper, saltier water. The sun and warm air can also heat surface waters and contribute to stratification. Processes that promote mixing of surface and bottom waters such as strong winds, tidal currents, and ocean swell can disrupt stratification.

Flushing time and horizontal mixing

Oxygen is more likely to reach a state of severe depletion in bodies of water that are semienclosed. Open ocean waters adjacent to coastal areas are generally well oxygenated due to lower levels of productivity and greater mixing by wind and waves. Limited

exchange with the ocean leads to an increase in the residence time of water (equivalent to a decrease in flushing time), which reduces both import of oxygenated waters and export of organic matter. Headlands and sills limit exchange with the ocean by restricting the width and depth of a bay's opening, respectively. In other instances, a body of water without a restriction at the mouth, such as a fjord, may have limited exchange simply because it is long and/or voluminous relative to the size of the openings through which exchange with the open ocean takes place.

The influence of limited exchange between a bay and the open ocean can be detected in gradients in oxygen concentrations which decline with distance from the open ocean (**Fig. 3**). Depth gradients in oxygen can also be apparent where a pocket of deeper hypoxic water is trapped in the basin formed by a sill.

Shoreline geomorphology is not the only factor that affects the residence time of a body of water. The flushing time of a bay can be tightly linked to freshwater input (higher inputs lead to decreased flushing times), and is therefore dependent on rainfall and water management in the watershed. Tidal amplitude will also determine the volume of water exchanged between a semienclosed body of water and the open ocean, and hypoxia is more likely to occur during or just following neap tides in highly tidal systems.

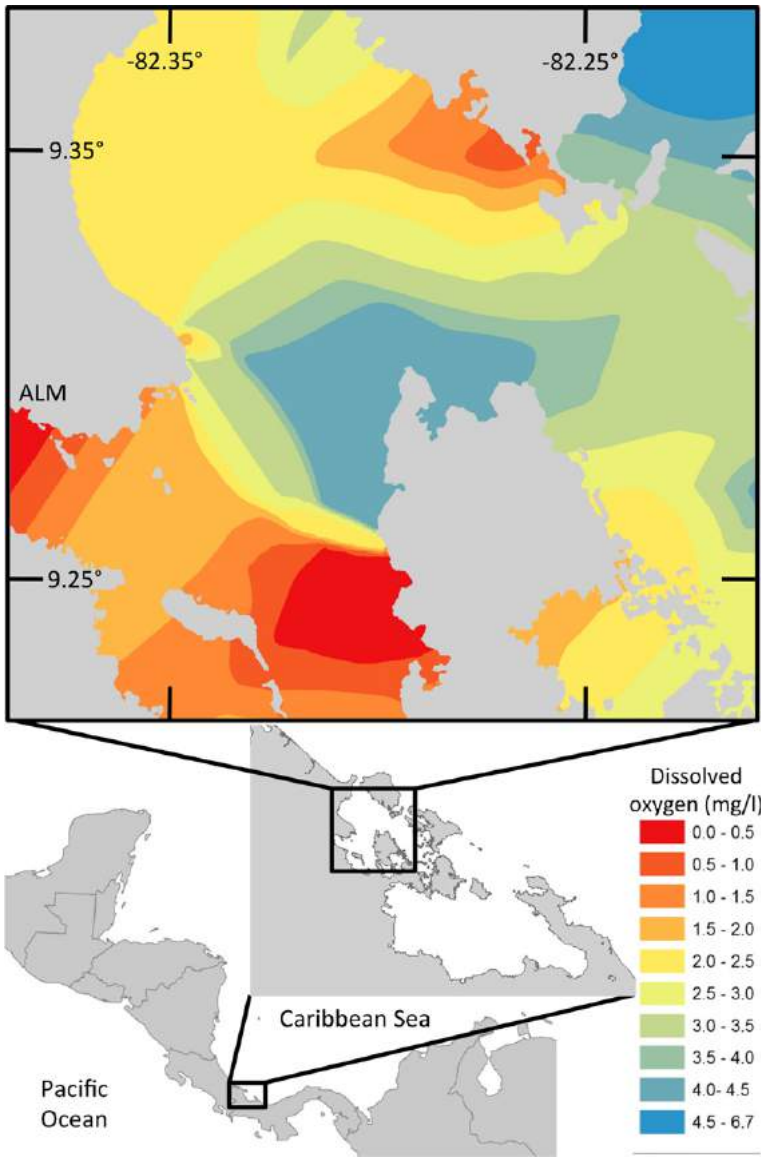


Fig. 3 Map of a dead zone at 10 m depth on the Caribbean coast of Panama in which many corals and reef organisms perished. Typical of many hypoxic ecosystems, there is a gradient in dissolved oxygen. Oxygen concentrations were lowest (*red* and *orange*) inshore, to the lower left, where nutrient inputs are greatest and water circulation is minimal, and highest (*blue* and *green*), to the upper right, where exchange with the open ocean is greatest. There is also a pocket of hypoxic water at the upper center where pollution from the nearby town is retained in an eddy on the lee side of the island. *Gray* represents land, and color represents oxygen concentration according to key at lower right. Modified from Altieri, A.H., Harrison, S.B., Seemann, J., Collin, R., Diaz, R.J. and Knowlton, N. (2017). Tropical dead zones and mass mortalities on coral reefs. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 3660–3665.

In some instances, the effectiveness of exchange with the open ocean for reducing hypoxia is short-circuited when adjacent ocean waters are themselves hypoxic. This can occur, for example, when upwelling brings hypoxic waters onto the shelf and to a bay's entrance as observed in the Saanich Inlet in Canada.

High Risk Ecosystems

Given the factors that deplete oxygen and limit re-oxygenation, there are some characteristics of coastal ecosystems that elevate the risk of oxygen depletion. Nearly all recognized dead zones are in a semienclosed bay, estuary, fjord, inland sea, atoll, or constructed harbor. These bodies of water have some degree of isolation that reduces exchange with the open ocean and the strength of wind, thereby promoting stratification and collection of nutrients and/or organic matter. Moreover, human settlements tend to concentrate on the shoreline of such sheltered bodies of water because of agricultural opportunities and natural harbors, and so agricultural run-off and sewage discharge that fuel microbial activity are often relatively high in those areas. Ecosystems with rivers and terrestrial runoff are at higher risk because they deliver nutrients and organic matter as well as the freshwater that contributes to stratification. Overharvesting of suspension feeders, such as oysters, removes a natural control of phytoplankton that also increases the likelihood that hypoxic conditions will develop. Temperate waters may be especially susceptible to hypoxia because stratification and freshwater inputs can be intensified and synchronized by seasons, and because fertilizer use historically has been higher in temperate ecosystems.

Biological Impacts of Hypoxia

Severe hypoxia that results in dead zone formation can have spectacular effects that galvanize public attention. Failed fisheries and beaches covered with dead animals are among the most dramatic and apparent effects of dead zones. Some estimates put the annual loss of biomass due to dead zones worldwide at over 9 million tons. The term “dead zone” has been earned by events where severe hypoxia rendered thousands of square kilometers of sea bottom largely lifeless. However, it is an oversimplification to view healthy ecosystems and dead zones as dichotomous states since oxygen concentrations can vary continuously. Likewise, there is a range of responses to oxygen depletion from the organism to ecosystem level that must be elucidated to develop a predictive understanding of how dead zones affect natural systems.

Organismal Responses to Hypoxia

Mortality is the most extreme consequence of exposure to hypoxia, and there are range of sublethal responses as well. Many aquatic organisms are able to tolerate a period of exposure to hypoxic or even anoxic conditions by employing strategies that include physiological and behavioral responses. These strategies fall into several broad categories: evading hypoxia, maintaining delivery of oxygen, reducing oxygen demand by conserving energy, and switching to anaerobic metabolism.

Evading hypoxia

Many animals can detect low oxygen and will move toward higher oxygen concentrations. Fish and other highly mobile organisms will flee hypoxic areas. This typically involves moving higher in the water column, into shallows, or toward the open ocean. Less mobile species such as crabs, lobsters, and sea stars will crawl to areas of higher relief and the tops of structures such as rocky outcroppings and coral colonies in an attempt to move above hypoxic bottom waters or reach higher flows with greater oxygen flux. Burrowing species, including worms and shrimp, will crawl out onto the sediment surface for the same reasons.

Maintaining delivery of oxygen

Strategies for maintaining oxygen delivery include increasing water flow over gills and other respiratory structures, and through burrows, by increased ventilation and pumping rates. Some fish are known to air gulp at the surface to attain oxygen. Other fish have the ability to increase the number of red blood cells and to increase the binding capacity of hemoglobin for enhanced oxygen uptake and internal transport. Prolonged exposure to hypoxia can lead to morphological changes such as an increased number of gill lamellae to increase diffusion capacity. Some groups such as sea cucumbers are able to deform their bodies to increase surface area to volume ratios to improve the relative efficiency of oxygen diffusion.

Reducing oxygen demand

Many species will reduce energy use in response to hypoxia to reduce metabolic oxygen demand. This can involve a decrease in overall metabolic rates as well as downregulation of protein synthesis. A hypoxia-induced binding factor can regulate multiple genes involved with this cascade of changes. This response is apparent in reduced activity and movement rates observed across a range of taxa, and changes in the relative abundance of various enzymes involved in metabolic pathways, with the onset of hypoxia.

Anaerobic metabolism

While aerobic metabolism is being downregulated, enzymes involved with anaerobic metabolism can simultaneously be upregulated. Some species tolerate periods of severe hypoxia or even anoxia through anaerobic metabolism. This ability can be so well developed that there are species of clams recognized as facultative anaerobes. However, the shift to anaerobic metabolism comes at a physiological cost since energy reserves will be depleted more rapidly due to the inefficiency of anaerobic metabolism. In addition, the accumulation of anaerobic metabolites can be harmful.

Negative effects of tolerating hypoxia

There are a number of negative physiological consequences of the stopgap measures associated with hypoxia tolerance by aerobic organisms. Reduced growth is a general response to hypoxia that has been observed in a number of taxa including fish, oysters, mussels, worms, and brittle stars. Reduced growth coincides with reduced rates of molting in crustaceans. Reductions in feeding rates, energy stores, and reproduction are also widely observed consequences of hypoxic stress. Fish, shrimp, and crabs exhibit reduced immune responses and increased susceptibility to diseases and bacterial infections when exposed to hypoxic conditions. Primary producers are also negatively affected by hypoxia. For example, hypoxia inhibits respiration and nutrient uptake in seagrass.

While many of the strategies of hypoxia described above are common across multiple taxa, there is tremendous interspecific variation in tolerance to sustained exposure to hypoxia. For example, lethal concentrations of oxygen range from 0 to 8.6 mg/L across phyla. Many species can tolerate hypoxia, but not anoxia, suggesting that survival relies on efficient use of remaining environmental oxygen (rather than an exclusive switch to anaerobic metabolism). It has been suggested that variation in tolerance among species depends on their physiological ability to use that small amount of remaining oxygen. Moreover, species-specific responses to hypoxia are dependent on the concentration, duration, and frequency of oxygen depletion, each of which can independently affect lethality.

Population Responses to Hypoxia

Mass mortality and migration associated with dead zone events have obvious effects on the population size of aquatic organisms, but there are a number of other significant ways in which hypoxia can influence population structure.

Population size structure

Hypoxia affects the size structure of populations in three ways. First, hypoxia commonly reduces individual growth rates—an effect that can be attributed to changes in metabolism and reduced feeding rates. This has broad implications for populations since body size is important for reproductive success, achieving a size refuge from predation, and storing energy for seasons with fewer resources. Even seasonal hypoxia can have a disproportionate effect given that hypoxia is most likely to occur in summer when most of the growth for a given year occurs. Second, hypoxia tolerance is often dependent on body size within a given species and can result in size selective mortality. Some studies with bivalves have found that the largest individuals of a given species are most likely to perish in low oxygen conditions, whereas other studies with crabs and fish have found that only the largest individuals persisted through stressful conditions. Small individuals can benefit from higher surface area to volume ratios for increased diffusion and passive transport of oxygen, whereas larger individuals may have higher energy stores and a better ability to detoxify hydrogen sulfide (which is a toxin that frequently co-occurs with hypoxia). Third, populations may exhibit size-selective migration in response to hypoxia. This phenomenon has been observed in benthic fish where larger individuals were more likely than small fish to flee during the intrusion of lethal hypoxic waters.

Reproduction

Inhibition of reproduction by hypoxia has been observed at all critical life history stages including hormone and gamete production, embryo development, egg survivorship, larval survivorship, larval settlement, and recruit survivorship. Vulnerabilities at any of these stages can lead to failure of population cohorts even when the reproductive adults of the population persist through hypoxic conditions.

Metapopulations

Often a spatially explicit perspective that incorporates a metapopulation framework is necessary to understand a population's response to hypoxia. A given species can be distributed in demographically linked patches across a landscape that exhibits a gradient in dissolved oxygen concentration. In such a case, hypoxia may cause local extinction of some patches, with the potential for later recolonization from remaining patches that survived the event.

Community Responses to Hypoxia

The variation among species in strategies and abilities to tolerate hypoxia has direct effects on community structure as well as indirect effects through modified species interactions.

Biodiversity

Decreasing oxygen concentrations typically lead to declines in biodiversity, as fewer species are able to tolerate increasingly severe physiological stress. Generally, sorting occurs along an oxygen gradient that reflects broader taxonomic differences in hypoxia tolerance: fish and crustaceans are typically the least tolerant, molluscs are intermediately tolerant, and worms and meiofauna are the most tolerant (Fig. 4). There are also shifts among functional groups as oxygen depletion becomes more severe: deposit feeders replace suspension feeders, shallow/surface deposit feeders replace deposit feeders deeper in the sediment, jellies (cnidarians and ctenophores) replace fish, and microbes and meiofauna replace macrofauna. Another commonly observed trend is a shift in biomass from benthic to pelagic species, particularly among harvested species. This latter shift could be explained in part by the tendency for hypoxia to be most severe at near the sea floor, but trends in water clarity and the distribution of primary productivity associated with eutrophication are likely contributing factors to this shift as well.

Predation

Hypoxia can modify predator-prey interactions in a number of ways (Fig. 5). Reductions in feeding rates are associated with the general downshift in activity levels by organisms stressed by hypoxia. Sessile and relatively sedentary species that are able to tolerate hypoxic conditions can benefit from a predation refuge as mobile predators, such as fish and crabs that are typically most sensitive to hypoxia, flee low oxygen conditions. However, when hypoxia subsides, predators can return and take advantage of prey that have remained and died, become moribund, or otherwise been rendered vulnerable as they move from burrows to the sediment surface. In some cases predators closely track the boundary of hypoxic water and will make brief forays into hypoxic waters to take advantage of vulnerable prey. Whether the net effect of hypoxia is to reduce or enhance predation rates depends on factors such as the relative tolerance of predators and prey, the spatial extent and duration of the hypoxic event, and conditions such as tides and winds that can cause movement in the boundary of hypoxic water.

Competition

The effects of hypoxia on competitive interactions have received less attention than the effects of hypoxia on predation. However, there is evidence that hypoxia can lessen competition for space and other resources by removing dominant species through selective mortality or migration. This in turn has been used to explain why diversity can be highest at the edge of a hypoxic area, as predicted by the intermediate disturbance hypothesis. Relaxed competition due to hypoxia also has the potential to increase the success of invasive species. In cases where both predation and competition for food have been analyzed, such as in omnivorous pelagic food webs, predation was found to be consistently a more important determinant of community structure than competition, regardless of oxygen concentration. However, oxygen itself can be a limiting resource, and competition for oxygen can influence community structure as observed in densely settled fouling communities.

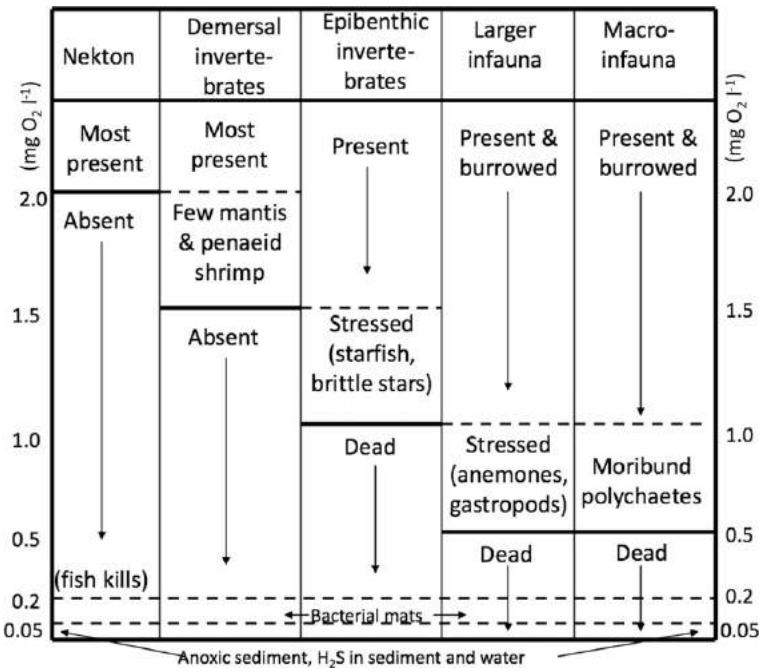


Fig. 4 The behavioral and physiological responses to hypoxia varies across taxonomic groups. As a consequence, functional diversity declines with increasing severity of hypoxia, as indicators of hypoxic stress including moribund inverts, fish kills, and mats of *Beggiatoa* bacteria and other sulfide oxidizing microbes increase. From Rabalais, N.N. and Turner, R.E. (eds.) (2001). *Coastal hypoxia: Consequences for living resources and ecosystems*. Washington, DC: American Geophysical Union.

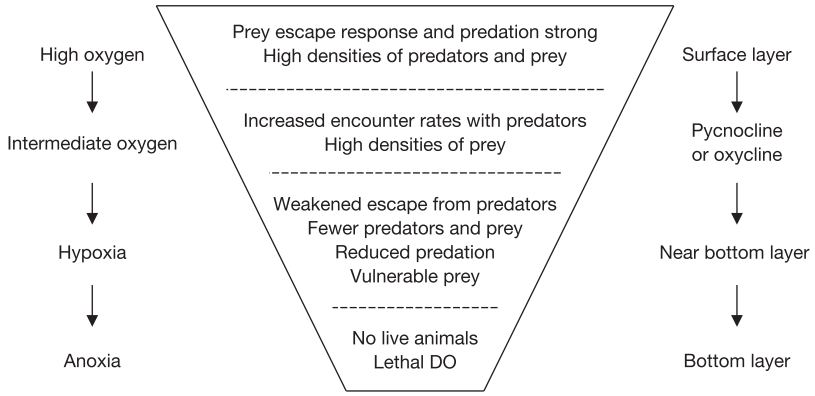


Fig. 5 Hypoxia can have a number of indirect effects on community structure by modifying species interactions such as predation. The exact consequence of hypoxia for predation is dependent on the relative tolerance of predators and prey, their mobility, and the duration and spatial structure of the hypoxic waters. Redrawn from Breitburg, D. (2002). Effects of hypoxia, and the balance between hypoxia and enrichment, on coastal fishes and fisheries. *Estuaries* **25**(4B), 767–781.

Succession

Succession describes the process of community recovery through the gradual accumulation of species and turnover in community composition following hypoxia or other disturbance. Initial recolonization will depend on supply from source populations, and so the size of a dead zone and distance from the edge can determine rates of recolonization. Often the first species to recolonize are not necessarily hypoxia tolerant species, but rather a community of opportunistic species dominated by polychaete worms and bivalves with high dispersal potential and reproductive rates. Since the return time for hypoxic events is often one or a few years whereas succession typically occurs over a longer time period (over a decade in the Black Sea), many periodically hypoxic ecosystems may never reach their natural climax structure. In ecosystems where regularly occurring hypoxia predates scientific surveys, natural or baseline reference conditions may be unknown.

Ecosystem Responses to Hypoxia

Changes in community structure combined with degraded environmental conditions associated with hypoxia can have drastic effects on ecosystem properties and dynamics including habitat provision, trophic function, biogeochemical cycling, and provision of ecosystem services.

Habitat loss

Coastal estuaries and bays are essential habitat for many aquatic species. A dead zone represents an area of critical habitat rendered unavailable or suboptimal. The result is habitat loss and possibly habitat compression in which an impacted population is forced to use only the remaining portion of their potential habitat where there may be higher risks from predation or other stressors such as high temperatures. Even temporary or small hypoxic areas can cut off migration routes or eliminate seasonally important reproductive areas.

Microbial activity

While the term "dead zone" conveys the lack of macrofauna and apparent lifelessness of a severely hypoxic area, the term is a misnomer because life, particularly microbes, can thrive in these areas. As discussed above, it is microbial decomposition of phytoplankton that commonly initiates formation of hypoxic waters. Additional microbial decomposition of fauna that succumb to hypoxia can further deplete oxygen creating a positive feedback loop in the early stages of dead zone formation. Hypoxic conditions and shallowing of the redox potential discontinuity layer in sediments is associated with increased activity of sulfate reducing bacteria that generate hydrogen sulfide. Hydrogen sulfide is toxic to metazoans on its own, and can act synergistically with hypoxia to increase stress and mortality rates. Hydrogen sulfide itself can be reduced by yet other bacteria such as *Beggiatoa*. The presence of this mat forming bacteria is commonly used as a diagnostic of an ecosystems experiencing hypoxia. The proliferation of the microbial communities in dead zones represents a shift in energy flow toward microbial pathways and away from macrofauna and higher trophic levels (**Fig. 6**).

Nutrient dynamics

Nutrient cycling is altered in hypoxic areas because of changes in biogeochemical processes associated with low oxygen conditions. The mortality and emigration of bioturbating organisms also alters the biogeochemistry of the water-sediment interface. Hypoxia leads to increased phosphorous release from sediments to the water column and decreased rates of nitrogen removal through denitrification and anaerobic ammonium oxidation (annamox). This can spark elevated water column productivity, further reinforcing hypoxic conditions through a positive feedback loop. This feedback, combined with

the nonlinear relationship between dissolved oxygen concentration and denitrification rates, explains the hysteresis in which hypoxic conditions often persist after anthropogenic nutrients inputs are reduced below concentrations that initiated oxygen depletion in the ecosystem (Fig. 7).

Ecosystem services

The effects of hypoxia on community structure, biodiversity, and ecosystem functions are detrimental to the output of ecosystem services that provide essential goods and services to humans. The filtration capacity of an ecosystem, which has the potential to mitigate eutrophication by controlling phytoplankton blooms, is reduced when hypoxia causes mortality of bivalves such as mussels and oysters. The coastal tourism industry can also be negatively impacted by hypoxia. For example, the 1976 hypoxic event in the New Jersey Bight in the northwest Atlantic resulted in losses of $16 million (2017 USD) due to impacts on recreational fishing and scuba diving activities.

Fisheries production is an ecosystem service that has been well studied in the context of hypoxia. Dead zones have catastrophic effects on some fisheries, with decreased landings, closed fisheries, recruitment failures, and displaced fishing effort. However, the net effects of hypoxia on fisheries landings are not as clear-cut as once thought. Some harvested species benefit from the predation refuge created by hypoxia, and in other systems fishery outputs have persisted by shifting to species that are more tolerant or found higher in the water column where oxygen levels remain high. When taking a broader spatial perspective across an entire ecosystem, the higher productivity associated with eutrophication beyond the hypoxic area appears on average to offset the loss of fisheries within hypoxic areas.

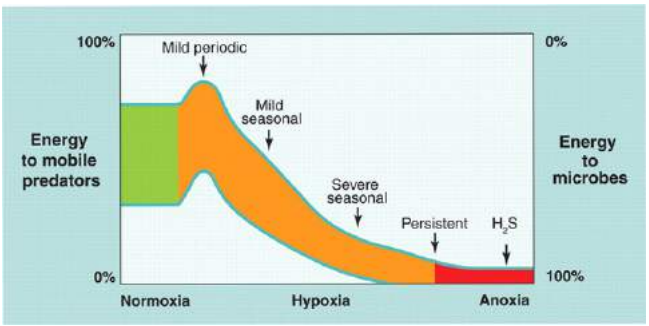


Fig. 6 Food webs shift as hypoxia becomes more severe or longer in duration. Initially there can be an increase in the transfer of energy to higher trophic levels due to predators taking advantage of increased availability of resources as prey die or become vulnerable, move to more vulnerable positions, or die. Eventually microbial pathways come to dominate as macrofauna flee or die. From Diaz, R. J. and Rosenberg, R. (2008). Spreading dead zones and consequences for marine ecosystems. *Science* **321**, 926–929.

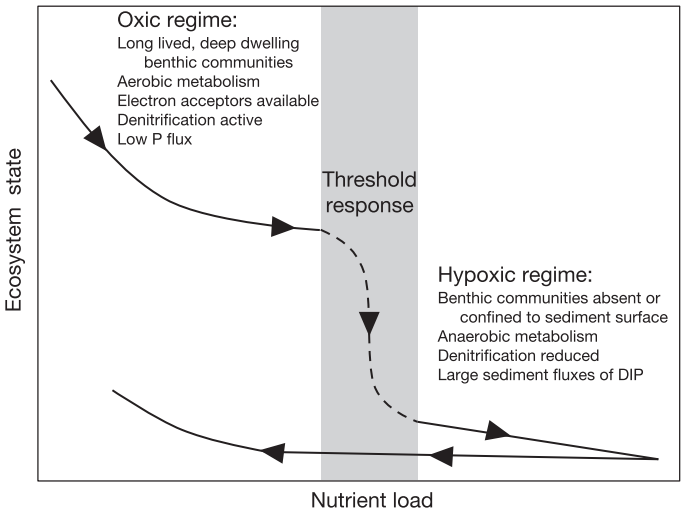


Fig. 7 Hypoxic ecosystem can exhibit nonlinear relationships between nutrient loads and severity of hypoxia. There is also evidence for hysteresis in which reversals in nutrient loading do not necessary result in concomitant reversals of hypoxia. This can result in lags between nutrient abatement and the recovery of physical conditions, community structure, and/or ecosystem function. Redrawn from Conley, D.J. et al. (2009). Ecosystem thresholds with hypoxia. *Hydrobiologia* **629**(1), 21–29.

Climate and Dead Zones

Climate modulates the severity of hypoxia and its biological impacts in a number of ways. Rising temperature is the greatest concern because of its broad impacts on hypoxia and the certainty of future increases (Fig. 8). While climate warming has likely contributed to the increasing prevalence of dead zones, predicting the response of oxygen to climate change requires an understanding of other climate factors including precipitation, cloud cover, and wind, as well as indirect links to sea level rise and land cover, that also affect the establishment of hypoxic conditions.

Effects of Temperature on Hypoxia Formation

Warming can increase the severity of hypoxia in several ways. Increased temperatures directly affect oxygen dynamics because warming reduces the solubility of oxygen in water. Temperature also has several indirect effects on oxygen concentration. Warming increases rates of microbial decomposition of organic matter which depletes oxygen. Increased temperatures also fuel the decomposition process by increasing production of organic matter in the form of phytoplankton. Warming enhances rates of nutrient cycling and therefore nutrient availability for plankton blooms, and earlier onset of spring temperatures due to climate change can extend the summer season of higher productivity.

Effects of Temperature on Organismal Responses to Hypoxia

Higher temperatures increase the susceptibility of organisms to hypoxia. Elevated temperatures lower physiological tolerances to hypoxia such that lethal thresholds may be reached at higher oxygen concentrations and shorter durations of exposure. Warming can also increase metabolic rates which leads to increased oxygen demand and depletion of energy stores that are often limiting in hypoxic conditions. On the landscape scale, low oxygen in deeper, cooler waters can drive mobile organisms into shallow, warmer water that may exceed their thermal tolerances.

Wind, Rain, and Other Climate Variables

There are numerous climate-related variables other than temperature that act on dead zones, primarily by affecting the nutrients and stratification that promote hypoxia. Precipitation can regulate the delivery of nutrients to coastal waters, and can affect the strength and depth of stratification as well as the residence time of estuarine waters. Storms bring pulses of precipitation as well as winds that can disrupt stratification. Sea level rise due to warming temperatures can affect stratification, reduce the ability of wetlands to filter nutrients entering a coastal waterbody, and increase the overall volume of water over a shelf or in a bay that is susceptible to hypoxia. Climate driven changes in large-scale ocean regimes can affect water temperatures and the concentration of oxygen and nutrients, particularly in areas of upwelling. Additional indirect links between climate and hypoxia include cloud cover, which can affect patterns of primary productivity, and changes in land-cover, which can affect nutrient loading in the

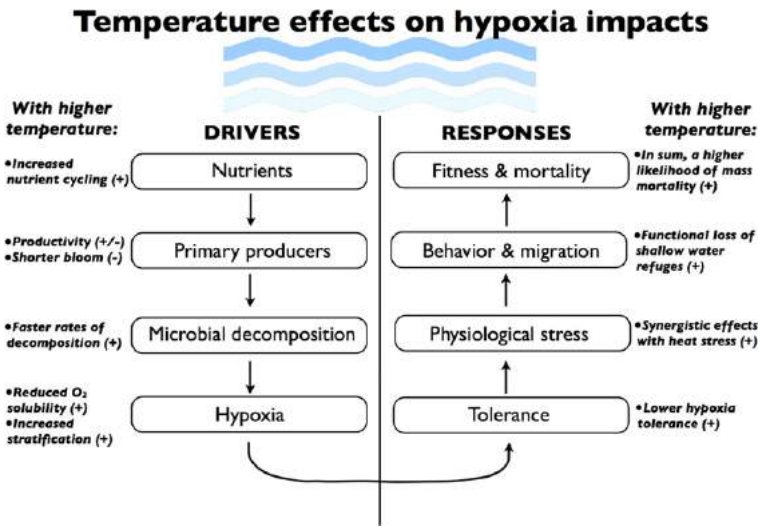


Fig. 8 Warming enhances many of physical drivers of hypoxia and exacerbates a number of negative biological consequences of hypoxia. As a consequence, the prevalence of hypoxia and its negative ecological effects are often associated with summer in temperate regions. Moreover, the warming component of climate change is expected to increase the severity of hypoxia and may have contributed to increase in the number of known dead zones in recent decades. From Altieri, A. H. and Gedan, K. B. (2015). Climate change and dead zones. *Global Change Biology* **21**, 1395–1406.

watershed. The excess carbon dioxide that is triggering climate change also affects water chemistry, and there is growing interest in the interactions between ocean acidification and hypoxic stress.

Predicted Effects of Climate Change on Dead Zones

Overall, climate change is expected to increase the severity and prevalence of hypoxia worldwide. Global temperatures are increasing, with over 90% of dead zones expected to experience a $>2^{\circ}C$ increase by the year 2100. As detailed above, warming intensifies the severity of hypoxia in several ways. Increased temperatures can also affect the duration of phytoplankton blooms and overall primary productivity rates, although the net effect of climate on oxygen concentrations through these pathways is not clear. There is also uncertainty regarding the frequency, timing, and intensity of storms and precipitation events which are expected to increase in some areas and decrease in others with climate change. Spatially explicit climate models are needed to forecast oxygen conditions in any given coastal ecosystem.

Global Spread of Dead Zones

The number and severity of dead zones identified worldwide is increasing at an alarming rate, doubling each decade over the past half century. Estimates suggest the increase has been exponential in recent decades, and a recent study identified over a hundred new sites in the Baltic Sea alone. Other work that examined latitudinal trends in numbers of dead zones and research effort suggests that there are hundreds of unidentified dead zones in the tropics.

Where long-term scientific records exist, there is evidence that ecosystems without previous signs of low oxygen, such as the northern Adriatic Sea, have become newly hypoxic in recent decades. In systems where hypoxia was previously observed, dead zones have become larger, such as in the Black Sea, Baltic Sea, and Chesapeake Bay.

Paleo indicators including foraminiferans, glauconite, and biogenic silica from cores collected in dead zones have established that hypoxia can occur naturally, but that significant increases in the number and severity of hypoxic ecosystems have been associated with developments such as the industrialization of agriculture. Much of the research to date has focused on eutrophication as the primary driver of the increased frequency of dead zones. Given the many ways in which climate affects the severity of dead zones, it is also likely that climate change has contributed to this observed increase in the hypoxia problem.

Summary and Outlook

Dead zones are a widespread phenomenon that impacts over 500 coastal ecosystems worldwide. Oxygen depletion is driven by factors including excess nutrients and climate change that are directly tied to human activities. As a consequence, the number of dead zones and their severity has increased exponentially in recent decades. Hypoxia affects the reproduction, behavior, survivorship, and diversity of marine life. The implications of hypoxia for coastal ecosystems are grave. Hypoxia compromises ecosystem functions and services including habitat provision, water filtration, biogeochemical cycling, and fisheries harvests.

This current state of knowledge should motivate efforts to confront the dead zone threat. Encouragingly, there is evidence that human action can reverse hypoxia and dead zones formation. Oxygen conditions have improved in several dozen ecosystems, largely due to reductions in nutrient inputs. Sometimes the efforts are intentional, as along the coast of Sweden where sewage treatment was enhanced, and other times incidental, as in the Black Sea where the collapse of the Soviet Union reduced fertilizer subsidies. Additional research in several areas is needed to build this positive momentum to turn the tide on dead zones, including: an increased number of ecosystems under surveillance, sustained support of long-term monitoring programs, deployment of advanced technology to detect changes in oxygen and associated water quality parameters, scientific networks for enhanced data sharing, and socio-economic data collection to better understand the economic impacts of hypoxia and inform strategies to incentivize sustainable management.

Synthetic studies that integrate across multiple study systems and time scales can establish realistic expectations for both the scope of improvement and the timeline for reaching management goals. Coastal ecosystems are highly variable, and it can take decades to verify sustained improvement of oxygen conditions. Moreover, there is an emerging concern that climate change could nullify some increases in oxygen concentration expected from reductions in eutrophication. System-specific models for predicting the effects of nutrient abatement and climate forecasts on hypoxic conditions are urgently needed to guide management. This will require studying the interactions between hypoxia and other factors such as ocean acidification, toxicity, and temperature in a multiple-stressor framework.

Although dead zones are a worldwide phenomenon, management of eutrophication and organic pollution at the scale of a single watershed can improve oxygen conditions in a given coastal ecosystem. Therefore, dead zones are an environmental problem with a solution at the level of local organizations and local management efforts when supported with robust research. Regional governance, such as the Chesapeake 2000 Agreement and the Convention on the Protection of the Marine Environment of the Baltic Sea, can further support these efforts. Individual communities, organizations, and nations have the potential to confront the global dead zone problem and contribute to our growing understanding of how hypoxia affects coastal ecosystems.

Acknowledgments

Earlier versions of this text benefited from comments by Keryn Gedan and Hannah Nelson. NOAA, EPA, the Smithsonian Institution, Sigma Xi, and the Sounds Conservancy supported the author's research on dead zones.

See also: Aquatic Ecology: Eutrophication; Microbial Communities. Conservation Ecology: Ecosystem Health Indicators. Ecological Data Analysis and Modelling: Carbon Biogeochemical Cycle and Consequences of Climate Changes. Ecological Processes: Decomposition and Mineralization. Ecosystems: Estuaries; Upwelling Ecosystems. Global Change Ecology: Sulfur Cycle; Nitrogen Cycle; Oxygen Cycle; Phosphorus Cycle

Further Reading

Altieri, A.H., Gedan, K.B., 2015. Climate change and dead zones. Global Change Biology 21, 1395–1406.
Altieri, A.H., Harrison, S.B., Seemann, J., Collin, R., Diaz, R.J., Knowlton, N., 2017. Tropical dead zones and mass mortalities on coral reefs. Proceedings of the National Academy of Sciences of the United States of America 114, 3660–3665.
Breitburg, D.L., 1992. Episodic hypoxia in Chesapeake Bay: Interacting effects of recruitment, behavior, and physical disturbance. Ecological Monographs 62, 525–546.
Breitburg, D.L., Hondorp, D.W., Davias, L.A., Diaz, R.J., 2009. Hypoxia, nitrogen, and fisheries: Integrating effects across local and Global landscapes. Annual Review of Marine Science 1, 329–349.
Conley, D.J., Björck, S., Bonsdorff, E., Carstensen, J., Destouni, G., Gustafsson, B.G., Hietanen, S., Kortekaas, M., Kuosa, H., Meier, H.E.M., Müller-Karulis, B., Nordberg, K., Norkko, A., Nürnberg, G., Pitkänen, H., Rabalais, N.N., Rosenberg, R., Savchuk, O.P., Slomp, C.P., Voss, M., Wulff, F., Zillén, L., 2009a. Hypoxia-related processes in the Baltic Sea. Environmental Science & Technology 43, 3412–3420.
Conley, D.J., Carstensen, J., Vaquer-Sunyer, R., Duarte, C.M., 2009b. Ecosystem thresholds with hypoxia. Hydrobiologia 629, 21–29.
Diaz, R.J., Rosenberg, R., 1995. Marine benthic hypoxia: A review of its ecological effects and the behavioral responses of benthic macrofauna. Oceanography and Marine Biology 33, 245–303.
Diaz, R.J., Rosenberg, R., 2008. Spreading dead zones and consequences for marine ecosystems. Science 321, 926–929.
Levin, L.A., Breitburg, D.L., 2015. Linking coasts and seas to address ocean deoxygenation. Nature Climate Change 5, 401–403.
Levin, L.A., Ekau, W., Gooday, A.J., Jorissen, F., Middelburg, J.J., Naqvi, S.W.A., Neira, C., Rabalais, N.N., Zhang, J., 2009. Effects of natural and human-induced hypoxia on coastal benthos. Biogeosciences 6, 2063–2098.
Rabalais, N.N., Turner, R.E. (Eds.), 2001. Coastal hypoxia: Consequences for living resources and ecosystems. Washington, D.C.: American Geophysical Union.
Rabalais, N.N., Turner, R.E., Wiseman, W.J., 2002. Gulf of Mexico hypoxia, aka "the dead zone". Annual Review of Ecology and Systematics 33, 235–263.
Rabalais, N.N., Díaz, R.J., Levin, L.A., Turner, R.E., Gilbert, D., Zhang, J., 2010. Dynamics and distribution of natural and human-caused hypoxia. Biogeosciences 7, 585–619.
Vaquer-Sunyer, R., Duarte, C.M., 2008. Thresholds of hypoxia for marine biodiversity. Proceedings of the National Academy of Sciences of the United States of America 105, 15452–15457.
Vaquer-Sunyer, R., Duarte, C.M., 2011. Temperature effects on oxygen thresholds for hypoxia in marine benthic organisms. Global Change Biology 17, 1788–1797.
Wu, R.S.S., 2002. Hypoxia: From molecular responses to ecosystem responses. Marine Pollution Bulletin 45, 35–45.

Relevant Websites

https://oceanservice.noaa.gov/hazards/hypoxia/
http://www.wri.org/our-work/project/eutrophication-and-hypoxia

Deep-Sea Ecology

Tracey T Sutton and Rosanna J Milligan, Nova Southeastern University, Fort Lauderdale, FL, United States

Glossary

Benthic Living on the seafloor.

Bioluminescence The production of light by an organism. The light is produced either by symbiotic luminescent bacteria or by chemical reactions within organs called photophores.

Chemosynthesis The synthesis of organic compounds by bacteria or other living organisms using energy derived from reactions involving inorganic chemicals, typically in the absence of sunlight.

Deep scattering layer A layer in the pelagic water column consisting of a variety of marine animals. It was discovered through the use of sonar, as ships found a layer that scattered the sound and was thus sometimes mistaken for the seabed.

Demersal Living close to (but not on) the seafloor.

Epifauna Animals living on the surface of the seafloor, or attached to submerged objects or aquatic animals or plants.

Metazoan Any animal that is multicellular.

Micronekton Small fishes, shrimps, and cephalopods of the open ocean; these animals are mobile enough to maintain position against local currents, but their distributions are generally confined to specific water masses.

Oxygen minimum zones Depths bands within the water column where conditions become hypoxic.

Pelagic Relating to the environment of the open ocean. This environment can further be subdivided by depth: epipelagic (0–200 m), mesopelagic (200–100 m), bathypelagic (1000 m to just above the seafloor), and benthopelagic (within 100 m above the seafloor, but not on the seafloor).

Plankton Organisms that drift or float in the sea or freshwater. Plankton can be divided into two subcategories: phytoplankton (small, single-celled plants) and zooplankton (small animals, eggs, larval stages of larger organisms, gelatinous organisms [jellyfish, tunicates, siphonophores, comb jellies]).

Introduction

The deep sea is typically considered to include all oceanic waters below 200 m depth and is the most ecologically complex cumulative ecosystem on earth, befitting its status as the largest. It contains the vast majority of the planet's metazoan organisms, the widest array of organic production (from photosynthesis to chemosynthesis), the most highly developed spectrum of living light (bioluminescence), and a bewildering array of lifespans (days to centuries) and body sizes (microns to 30 m) among its constituents. It is a fully four-dimensional ecosystem, where vertical and horizontal (distance over ground) animal movements can be tied to the time of day, the time of the year, the time of an organism's life, and/or climatological cycles occurring over decades. The ecology of the deep sea can largely be characterized as the solution to a series of significant challenges: finding enough food in a food-poor environment; finding a mate in an immense, usually dark environment; and, avoiding predators where there may be nowhere to hide. The ecology of the deep sea, despite its rigors, is also driven by its physical stability. It is by far earth's oldest cumulative habitat. Over geological timescales, terrestrial ecosystems are subject to the constant movement of the continents relative to the axis of earth's rotation, resulting in major fluctuations in their climatologies, all while the continents are being uplifted by tectonic activity and weathered by the hydrological cycle. Landmasses are connected and then broken apart, and buried in ice and then bathed in warmth (cyclically). Even the coastal seas have experienced massive shifts in position and volume as sea level rises and falls. The current formation of the Great Barrier Reef, for example, is only 6000–8000 years old. The deep-sea environment, on the other hand, has existed largely as it is now for millennia (save ephemeral habitats based on submarine volcanism). This physical stability at geologic temporal and broad spatial scales has allowed the development of organismal ecologies poorly suited to the relatively rapid-fire pace of terrestrial and coastal ecosystems, but perfectly suited to the harsh-yet-steady conditions in which coastal, "more modern" forms cannot thrive.

At its most basic, the deep sea comprises two major environment types, pelagic (water column) and benthic (seafloor), though the intersection of the two, the benthopelagic realm, has many unique characteristics in and of itself. In many ways these are separate worlds ecologically, with radically different rates of processes (lifespan, turnover, geographic distributions), but they also form a highly connected continuum of life. In this article, we will describe these two main elements, the metazoan (multicellular) life within them, and the ecological processes that drive the diverse array of assemblages that exist in each.

The Deep-Pelagic Realm: Earth's Largest-Volume Habitat

The deep-pelagic zone (waters deeper than 200 m) contains almost 95% of the ocean's volume, and thus about 94% of water on earth (the remaining 1% being ice on the continents, with the water in lakes, rivers, and atmosphere being negligible). A comparison of the spatial scales of pelagic life reveals that the vertical scale of distributions (i.e., depth, measured in meters) is finer by 3–5 orders of magnitude than the horizontal (i.e., geographic, measured in 100's of kilometers) scale. It is not surprising then that the global pelagic

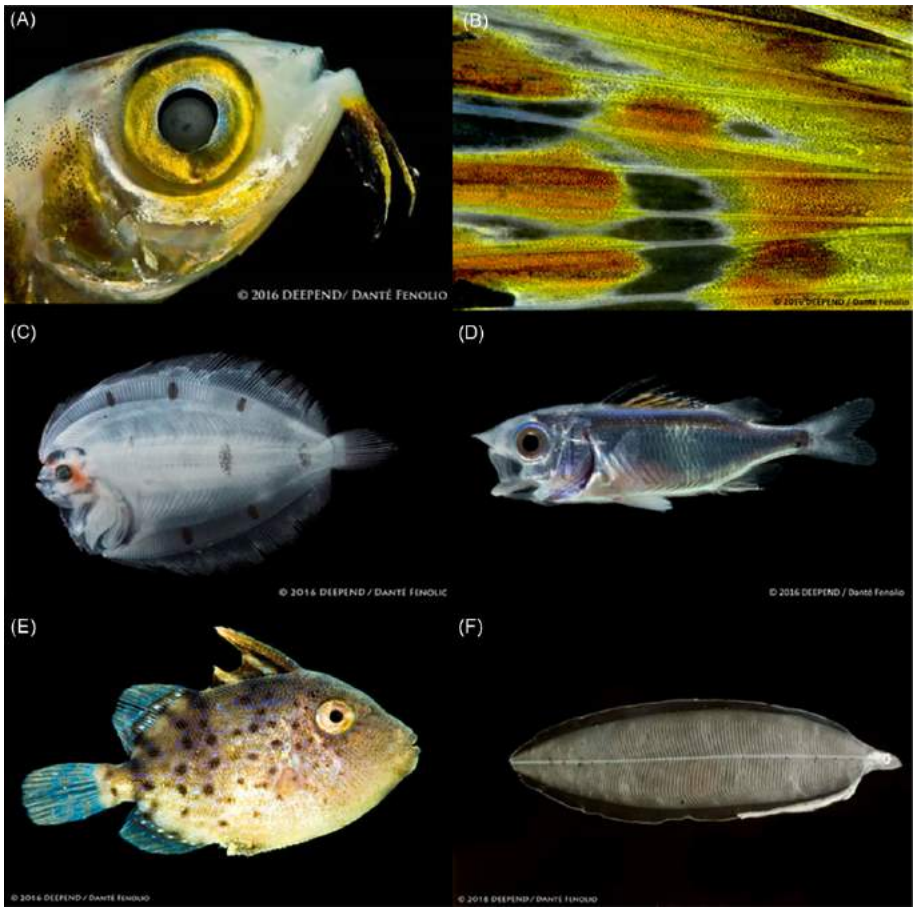


Fig. 1 Representative fishes from the epipelagic zone of the open ocean. (A) Juvenile flyingfish. (B) Close-up of the pectoral fin ("wing") of a flying fish. (C) Larval sole. (D) Juvenile squirrelfish. (E) Juvenile triggerfish. (F) Larval eel ("leptocephalus"). All images courtesy of the DEEPEND research consortium (http://www.deependconsortium.org).

habitat is structured primarily by depth, with each stratum roughly defined by the penetration of sunlight, a covariant of depth, and by the associated biota. Here we will briefly describe each stratum and the life within before considering the ecology of the assemblages.

Epipelagic Biome

The epipelagic biome extends, from the sea surface to 200 m depth and is defined as the stratum in which there is enough sunlight during daytime to support primary production in the form of phytoplankton (drifting, single-celled algae). Although the primary production per unit area is lower in the epipelagic biome than on land, and the epipelagic biome constitutes $<4\%$ of the ocean's volume, its vast surface area accounts for approximately 40% of the plant life on earth. Thus, while the epipelagic zone is not "deep," according to the previous definition of the deep-pelagic zone, it is inexorably tied to the deeper strata by providing energy (organic carbon), as well as providing nighttime habitat for deep-pelagic vertical migrators (see below). The epipelagic biome has greater variability, both physically and biologically, than deeper strata owing to the near-surface circulation patterns established by planetary forcing (e.g., wind stress, earth's rotation) and the much greater effect of seasonality on surface temperature and primary production. Because phytoplankton production occurs in the epipelagic biome, the majority of zooplankton production (small, drifting multicelled animals, the next step up in the food chain) also occurs there. In many ways the base of oceanic food webs resembles an inverted pasture, with plant life (and plant grazers) found near the surface and carnivorous animals found deeper and deeper. The epipelagic biome hosts a specialized fish fauna (Fig. 1), including flyingfishes and their relatives (needlefishes, halfbeaks), flotsam/jetsam-associated fishes (e.g., filefishes, triggerfishes, jacks, dolphinfish), "baitfishes" (e.g., herrings, anchovies), larval/juvenile fishes of astounding variety (e.g., eels, flatfishes, snappers, to name a few) and large, highly migratory fishes (e.g., tunas, billfishes, mako sharks).

Mesopelagic Biome

The mesopelagic biome (also called the "twilight zone") extends from 200 to 1000 m depth and is defined as the stratum of the ocean that receives enough sunlight that the animals within can differentiate night from day, but not enough to support

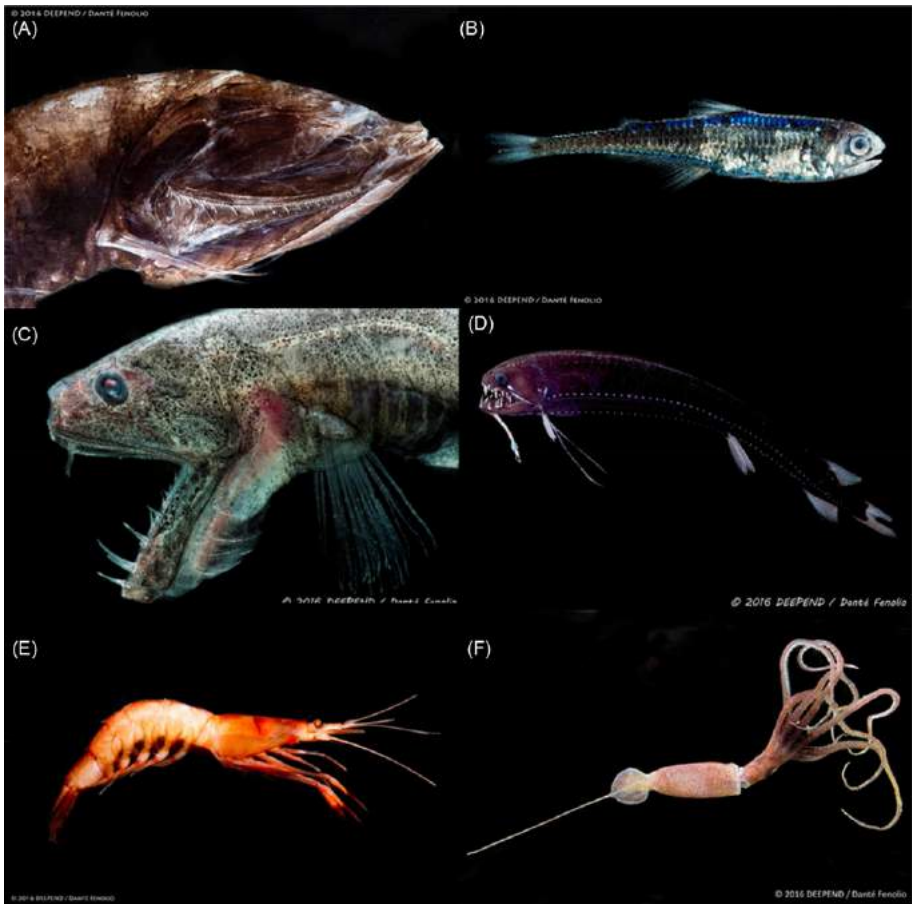


Fig. 2 Representative animals from the mesopelagic zone of the open ocean. (A) Bristlemouth. (B) Lanternfish. (C) Sabretooth. (D) Dragonfish. (E) Decapod shrimp. (F) Squid. All images courtesy of the DEEPEND research consortium (http://www.deependconsortium.org).

photosynthesis (i.e., below 1% surface light level). The highly variable conditions of the epipelagic biome are attenuated at mesopelagic depths and physical conditions become more stable over space and time. During daylight hours the mesopelagic biome hosts an array of specialized fishes from families that are absent or rare in shallow waters (e.g., lanternfishes, bristlemouths, dragonfishes, hatchetfishes), as well as highly adapted shrimps and squid (**Fig. 2**). Bioluminescence is the rule rather than the exception in this environment. Mesopelagic fishes and shrimps are the primary components of the globally ubiquitous "Deep Scattering Layers," detected using hydroacoustics (i.e., scanning sonar) and centered at around 500 m depth during the day and within the top 200 m at night. At night, these organisms vertically migrate into the epipelagic zone to feed on zooplankton, or on organisms feeding on zooplankton. This diel (daily) vertical migration is the largest animal migration on earth. The abundance of fishes in the deep-pelagic realm declines with increasing depth, but due to the enormity of the ecosystem their abundance and biomass is extremely large. Recently, global mesopelagic fish biomass has been estimated at a billion metric tons.

Bathypelagic Biome

The bathypelagic biome (also called the "midnight zone") extends from 1000 m to ~100 m above the seafloor and is defined by the absence of sunlight, and the relative invariance of temperature and salinity. There is some evidence for subdividing the bathypelagic biome at around 2500 m into the "abyssopelagic zone," characterized by dramatic decreases in fish abundance. The bathypelagic realm hosts a unique assortment of highly modified fishes (**Fig. 3**), including the oceanic anglerfishes, whalefishes, tubeshoulders, and gulper eels. Morphological adaptations of typical bathypelagic fishes, relative to mesopelagic taxa, include an increase in relative mouth size, a reduction in relative eye size, loss or reduction in photophores, replacement of reflective/silvery pigmentation with black, brown or red pigment, replacement of dense muscle mass with watery tissues with less metabolic demand, and reduction or loss of swimbladders. Sexual dimorphism is pronounced among many of the primarily bathypelagic fishes, and in some taxa (e.g., anglerfishes, whalefishes) represent the extreme example of dimorphism among all vertebrates. This is considered a form of ecological bet hedging in a food-limited environment, with taxa investing the majority of their species' biomass in females, ostensibly to facilitate greater fecundity.

Fig. 3 Representative animals from the bathypelagic zone of the open ocean. (A) Whipnose anglerfish. (B) Dreamer. (C) Headlight anglerfish. (D) Telescopefish. (E) Velvet whalefish. (F) Tubeshoulder. All images courtesy of the DEEPEND research consortium (http://www.deependconsortium.org).

Benthopelagic Biome

Numerous studies on continental slopes, mid-ocean ridges, and at seamounts have found evidence of a higher abundance and diversity of deep-pelagic organisms at the interface between the pelagic and benthic realms (the benthopelagic biome). In most cases this enhanced abundance and biomass is thought to result primarily from compression of the vertical distributions of pelagic fauna (i.e., the seafloor prevents pelagic fauna from migrating to deeper depths), although there is evidence for active station-keeping over seamounts by some pelagic fauna. Increased abundances of pelagic prey may also promote a larger number of benthopelagic fish predators (e.g., roundnose grenadiers, orange roughy), in addition to the benthic predators that would otherwise be present. Recent studies show that benthopelagic assemblages may represent a vital trophic linkage between the pelagic and demersal faunas, and are important elements of carbon sequestration in the ocean.

The Deep-Benthic Realm: Earth's Largest-Area Habitat

Where the pelagic realm is divided into vertical zones according to solar light levels, the benthic realm is typically categorized according to the topography of the seafloor. Covering over 75% of the Earth's surface, the deep seafloor begins at the continental shelf break (c. 200 m depth; 500 m around Antarctica) at the outer edge of the coastal seas and extends to about c. 11,000 m at the bottom of the Mariana Trench. Across the seafloor, myriad geological and biological features create unique, dynamic ecosystems that support a diverse benthic fauna. In this section we briefly describe each zone and the ecosystems they can contain, before considering the ecology of the fauna.

The Continental Slopes (Bathyal Zone)

The continental slopes extend gradually from the continental shelf break to the continental rise (c. 3000 m depth), with an average gradient of around 4°. Covering only 15% of the Earth's surface, the continental slopes are mostly covered by soft sediments, but are punctuated by a diverse range of geological features, such as seamounts, canyons, hydrothermal vents, cold seeps, and brine

pools. Where the physical and chemical conditions are suitable, cold-water coral reefs, sponge beds or other emergent epifauna (organisms living on the sediment surface) may generate large, complex frameworks that provide a greater diversity of habitats for a range of invertebrate and vertebrate taxa. Among the deep-sea benthic habitats, continental slopes are where the greatest variability in oceanographic conditions occurs, which can strongly affect the distribution of the benthic fauna. For example, oxygen minimum zones can lead to a reduced biodiversity of benthic animals, but an increased abundance of hypoxia-tolerant species. Similarly, varying current speeds may affect the coarseness of the benthic sediments or the deposition of "marine snow" (agglutinated flocs of dead plankton tissue and biological waste) to the seafloor, all of which may affect the abundance and diversity of the fauna living on and in the seabed.

At the upper (shallowest) end of the continental slopes, the fauna generally resembles that found in coastal environments, but as one goes deeper we see progressive shifts in morphology and life-history traits with increasing depth. Demersal fishes, for example, show a general shift towards elongate, eel-like body shapes with increasing depth, which are better suited for conserving energy in a food-poor environment. The systematic form of fishes also changes, with the more "advanced" spiny-rayed taxa (e.g., seabasses, scorpionfishes) giving way to more "basal" taxa (e.g., eels, cusk eels, and cod-like fishes). This trend is mirrored by the cartilaginous fishes, with requiem sharks (e.g., bull, tiger, and lemon sharks) and rays being replaced by basal taxa like catsharks and chimaeras.

The Abyssal Plains (Abyssal Zone)

As the continental slopes become less steep, they give way to the continental rises, and then the abyssal plains, which are vast, flat expanses covering c. 53% of the planet's surface. As we move further from the ocean surface and continental margins, animal biomass declines exponentially, and is reduced to $<1\%$ of the biomass found in coastal seas. Abyssal plains are considered energy-poor, though there is considerable regional variability in energy input depending on surface productivity—some areas can receive relatively large inputs of energy equivalent to 50%–80% of surface primary productivity. Despite their energy limitations, abyssal plains are home to a wide range of benthic organisms that are able to thrive in the high pressures of the abyss. Environmental conditions at abyssal depths are invariant compared to conditions on the upper slope and in coastal waters. There is no solar light, the waters are cold (c. 0.5–3.5°C) and well'oxygenated, and current speeds are generally low. An exception to this can occur below areas where sea-surface kinetic energy is high (e.g., beneath Gulf Stream eddies in the NW Atlantic). This energy can generate high-energy eddies at abyssal depths which cause "benthic storms." It is possible for several storms to occur per year and for those to last for several days at a time, causing broad-scale disturbances to the abyssal fauna in those regions and resuspending fine sediments into the water column. Outside of such regions, slow currents are the norm, however. Soft sediments dominate, but still support a diverse invertebrate infauna (organisms living in the sediment) and epifauna. Predatory and scavenging demersal fishes (particularly rattails; Family Macrouridae) occur in low numbers on the abyssal plains. Like other ecosystems, the distributions and diversity of benthic organisms on abyssal plains are influenced by environmental heterogeneity. At fine-scales, changes in sediment grain size, the occurrence of patches of phytodetritus, epifaunal animals themselves, animal tracks, burrows, and animal waste can all influence the occurrence and distributions of abyssal fauna. At larger scales, the sedimentary plains are interspersed by potentially hundreds of thousands of seamounts and tens of millions of abyssal hills. Abyssal hills are believed to be the most common topographic features on the planet, shed from the far larger Mid-Ocean Ridges as they spread. Similarly, despite the great depths of the abyssal plains, regional differences in overlying surface primary productivity can strongly influence the biomass and composition of abyssal animals.

Trenches (Hadal Zone)

Where tectonic plates collide, subduction zones can form and produce deep trenches in the seafloor. Trenches extend below c. 6000 m depth, and despite their great depths, harbor a unique animal fauna, even at full ocean depth (c. 11,000 m in the Mariana Trench). There are approximately 95 distinct basins deeper than 6000 m worldwide, with 16 recognized trenches. Hadal ecology is still a young and emerging field of research, but recent studies are beginning to shed light on the animals that inhabit these ecosystems and their ecology. Representatives from most marine taxa have been observed in trenches, with gastropods, polychaetes, bivalves and holothurians occurring to 11,000 m, though fishes are also present to c. 8200 m. Due to the relatively isolated nature of trenches, rates of endemism within trenches appear to be high, with different trenches harboring relatively high numbers of unique species.

Hyrdothermal Vents and Cold Seeps

Hydrothermal vents occur where there is volcanic activity and geothermal heating of the seafloor. They are characterized by buoyant plumes of heated water, the temperature of which can vary from slightly above ambient in diffuse flows to over 300°C at intense flows. The precipitation of dissolved minerals when heated water meets cool water can form columnar, chimney-like structures reaching tens of meters into the water column. These habitats are among the most ecologically unique on earth owing to the presence of high-biomass, chemosynthetically supported life. Animals that live in chemosynthetic ecosystems derive their energy from chemicals in seawater rather than from solar light, creating local areas of high animal biomass in an otherwise energy-poor deep ocean. Here chemosynthetic bacteria generate energy from sulphides ejected from volcanic fissures. Large tubeworms (*Riftia*) have evolved a symbiotic relationship with these bacteria and use them to rapidly produce large colonies at vent sites for as

long as the sites exist (~ decades). While hydrothermal vents occur worldwide, their total area is small and thus contribute only a very small fraction of total deep-sea production. Cold seeps are another example of chemosynthetic ecosystems, found in locations where hydrocarbons are expelled at the seafloor at temperatures similar to those of ambient seawater. In these cases, bacteria associated with animals such as mussels may use hydrogen sulphides or methane as their energy source.

Adaptations in the Deep Sea: The Ecological Arms Race

Relative to terrestrial, freshwater, and coastal marine ecosystems, the deep sea is a harsh environment, with organisms likely operating at or near their physiological limits. Extreme pressures, near or total darkness, low temperatures, low food availability, a medium (water) 800 × more dense than air, and in some cases, extreme oxygen limitations all constrain the biology and ecology of the deep sea. Deep-pelagic taxa live in a fully three-dimensional existence with no solid surfaces, whereas benthic and benthopelagic species are adapted to these surfaces, but in the total darkness of the abyss, probably cannot see them. Many of the bizarre-looking characteristics of deep-sea organisms is related to gaining sensory advantages relative to these constraints, and in the same fashion, to defeat the sensory capacities of both prey and predators. As these adaptations are central to the ecology of the deep-sea, we summarize the major adaptations here.

Hiding in the Deep

Many of the diagnostic characteristics of deep-sea organisms involves feeding, both finding prey and avoiding being predated upon. With respect to the latter, the low levels of light in the deep have provided opportunities for camouflage. Animals have adopted three major strategies: appearing to be something they are not (e.g., Müllerian and Batesian mimicry—imitating something inedible or noxious, respectively), being transparent, or blending in with the surrounding environment. Mimicry is largely confined to the surface ocean, where many fish larvae closely resemble inedible objects (especially bird feathers) or stinging gelatinous zooplankton (e.g., siphonophores).

Transparency and "blending in" are common in the deep oceans. Transparency is the extreme case, as it requires polarization of an organism's molecules as well as behavioral modifications to blend in. Transparency is best achieved by the pelagic invertebrates; exemplar organisms include gelatinous zooplankton (**Fig. 4**A), heteropod molluscs, arrow worms (Chaetognatha), certain amphipods, and a variety of decapod shrimps. Transparency is common among larval fishes, with the leptocephali (larvae of eels, bonefishes, tarpons, and halosaurs) being prime examples. No adult fishes are completely transparent, but many species whose body musculature is transparent "hide" their necessarily opaque parts (internal organs, eye retinae, prey in stomachs/intestines) with a silvery membrane invested with multilayered guanine crystals, meant to reflect the ambient water color. As with many cases in the ecological arms race in the deep sea, transparent organisms can be detected by predators equipped with headlight photophores and/or specialized eyes attuned to scattered light (**Fig. 4**B).

There are various ways for organisms to "blend in" with their surroundings. The reflection of a mirrored surface, adopted by many mesopelagic fishes, is an effective way to blend in with the surroundings. For example, hatchetfishes of the genus *Argyropelecus* (**Fig. 4**C) present a near vertical mirror in the water column—when viewed from almost any angle the reflected light is the color and intensity expected (lighter when viewed from below, darker when viewed from above). The bright silver reflection of these organisms when viewed from submersibles or in situ cameras is a function of artificial light introduced from the side, a phenomenon that does not occur in nature.

A "flaw" in the mirrored camouflage technique is that is does not hide an animal when viewed from directly below. To counter this vulnerability, the vast majority of mesopelagic fishes, many shrimps, and even many squids bear bioluminescent photophores (**Fig. 4**E) along their ventral (lower) margins, from which they produce light of the wavelengths and intensity of downwelling sunlight (usually blue). This ventral lighting serves to obfuscate the silhouette of an organism when viewed from directly below. An increasing number of in situ observations have reported that many predatory fishes adopt a vertical, head-up posture in the water column, ostensibly looking for prey outlined against downwelling light. Again as an arms race example, the eyes of many deep-sea predators have yellow lenses, thus allowing differentiation between natural sunlight and organismal bioluminescence.

Lastly, as one goes deeper into the water column (i.e., below 1000 m) animals become black (e.g., gulper eels, anglerfishes), dark brown (bigscales, whalefishes), or red (crustaceans, jellyfishes) (**Fig. 3**), as silvery coloration would serve to "announce" an organism when lit from the side by bioluminescence. Near the deep-sea floor, animals become monochromatically brown, gray, or semi-transparent—the lack of light precludes the need for distinctive coloration, or even hiding for that matter.

Bioluminescence—The Production and Use of Living Light

Bioluminescence is a defining feature of most animals living in the deep-pelagic realm, and even of many living near or on the seafloor. Over 90% of the metazoan species in the mesopelagic zone are bioluminescent. A lesser percentage of bathypelagic organisms are bioluminescent, though nearly all members of deep-sea anglerfishes (Suborder Ceratioidei; **Fig. 3**A–C), the most speciose fish taxon below 1000 m, produce light, as do most gelatinous taxa (especially medusoid cnidarians) and a wide variety of pelagic shrimps. Near the seafloor bioluminescence is greatly reduced among vertebrates and invertebrates alike.

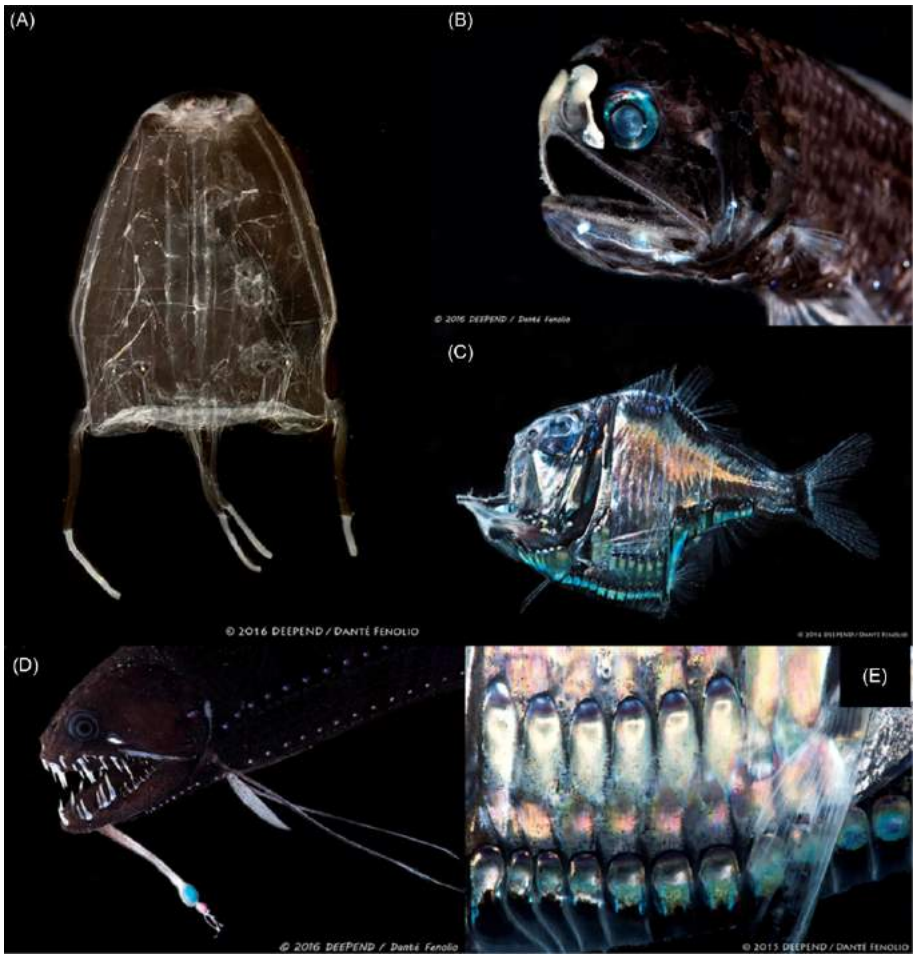


Fig. 4 Adaptations in the deep sea. (A) Transparency (box jelly). (B) Headlight photophores (lanternfish). (C) Mirrored sides (hatchetfish). (D) Chin barbel (dragonfish). (E) Ventral photophores (close-up of hatchetfish in C). All images courtesy of the DEEPEND research consortium (http://www.deependconsortium.org).

Bioluminescence among metazoans is either accomplished using luminescent bacteria cultures bound inside specialized light organs, or by intrinsic light production from photophores. In one case (e.g., female linophrynid anglerfishes), both systems are used. Intrinsic bioluminescence in the deep sea is produced from the same general chemical reactions as that of a firefly—a substrate (luciferin) is exposed to an enzyme (luciferase), which produces an unstable molecule that releases light energy to reachieve stability.

As previously mentioned, bioluminescence can be used to aid in camouflage in the water column, and also as a 'luring' mechanism (e.g., chin barbel of dragonfishes; **Fig. 2**D, **Fig. 4**D), but also has many other uses. Some animals (especially oplophorid shrimps and tubeshoulder fishes) presumably use bioluminescent displays to startle or distract predators, allowing for escape. Certain lanternfishes and dragonfishes use enlarged photophores around the eyes to illuminate prey in the water column. Bioluminescence can also be used to locate potential mates for spawning, as evidenced by the sex-specific photophore patterns of many meso- and bathypelagic fishes. This form of sexual selection and reproductive isolation may be the driver behind the exceptional diversity of the specific taxa that utilize this form of communication, the lanternfishes and dragonfishes.

Vision—Seeing in the Darkness

The importance of vision at mesopelagic and bathyal depths is manifest in the visual adaptations of the fauna. Crustaceans, cephalopods (squids and octopods), and fishes often have huge eyes relative to their body sizes, with vision enhanced by visual pigments tuned to the long, monochromatic wavelengths (470–480; blue-green light) that can penetrate through the epipelagic zone and into the mesopelagic. Fishes also possess multiple-bank retinas to further increase sensitivity and acuity in dim light. Tubular eye structures are common among predatory deep-pelagic fishes (**Fig. 3**D)—this allows binocular vision, which is an adaptive solution to the problem of sizing and ranging in an environment with no reference structures, as humans use when judging distance on land, for example. An additional adaptation among deep-sea fishes is the enlargement of the pupil to allow

light from many directions to pass through the lens onto the retina. Deep-sea crustaceans have compound eyes adapted to low light conditions as well. In many cases the pigment separating the individual ommatidia (single "eyes" of a compound eye) is reduced. This allows greater light sensitivity, but at the expense of image formation, a trade-off that is obviously successful given the large numbers and biomass of shrimps in the deep sea.

Hearing—Listening in the Darkness

Relative to the coastal and surface ocean, the deep sea is a calm and quiet environment, although anthropogenic sounds from sources like shipping, drilling and sonar are increasingly prevalent. Nonetheless, many taxa use this relative quiet to their advantage to offset the inherent visual limitations of the deep sea. All freshwater, coastal marine, and deep-sea fishes use an ear-lateral line system for acoustic sensing. "Ear stones" (otoliths) located within fishes' heads absorb sound, owing to their higher density relative to seawater, and pass this energy to a series of hair cells, which are then transmitted to the brain via nervs. The sensitivity of hearing in deep-sea fishes can be quite impressive, inferred from the observation that the complexity of the hair-cell organization of some deep-demersal fishes is the highest of any vertebrate.

Mechanoreception

The lateral line system (as the name implies) runs down the lateral sides of a fish, and is also most developed in deep-sea fishes relative to coastal forms. The form and complexity varies by taxon, but all possess a series of neuromasts (nerve endings) either exposed to seawater or imbedded in gel-filled cups. This arrangement allows mechanosensing (i.e., detection of mechanical stimuli) of the movement of water around an individual, both due to its own movement as well as movement external to the fish (water circulation and disturbance by other objects). It is not surprising that the highest development of free-standing neuromasts on any fish are found in bathypelagic anglerfishes, befitting a sit-and-wait, luring specialist in an energy-limited environment. Many deep-pelagic fishes also have extremely long fin rays, suspected to serve as extensions of the mechanosensory system (i.e., for vibration detection). Lastly, the elongate form of many fishes (oarfishes, deep-demersal cruisers such as rattails, chimaeras, cusk-eels, morid cods) may be an adaptation to allow greater extension of the lateral line systems.

Olfaction/Chemoreception

Olfaction and chemoreception are believed to be the primary methods by which mobile scavenging animals detect carrion at the seafloor. Baited camera observations (in which a time-lapse camera is used to record the arrival times of different animals reaching a bait package) suggest that this mode of feeding is quite common in the deep sea and is used by a wide range of taxa (e.g., amphipods, fishes, shrimp, sharks, crabs and squat lobsters, among the most prevalent). Chemoreception has also been implicated in the detection of mates, and is the likely method by which small male anglerfishes are able to locate the much larger females in the bathypelagic realm, where the total abundance of fishes is very low. In shallow waters, chemoreception is known to be important in larval settlement of many sessile benthic species, and may be similarly important in the deep sea.

Connecting the morphological characteristics of deep-sea animals with the increasing body of information on organismal biology is one of the most exciting fields of study in deep-sea ecology. As these characteristics are intimately tied to the first survival need in the deep—locating food—we will now discuss the fundamentals of trophic ecology in the deep sea.

Ecological Interactions in the Deep Sea

Feeding

Given that environmental variability in the deep sea is modulated by the overlying mass of water, and that for much of the deep sea (i.e., water column) there is no physical structure about which the fauna can orient, ecological interactions are particularly important as drivers of community structure. Of all interactions, trophic interactions are putatively the most important in a food-limited world. Studies of the trophic ecology of deep-sea organisms have progressed a great deal in the last two decades. Earlier studies focused on analysis of diet, morphological specializations for feeding, and the characterization of feeding guilds. These approaches have been recently expanded using biomarker approaches, revealing new food web pathways and energy sources for deep-sea communities. Recent studies have focused on the connections between the sources of food in epipelagic waters and the deep-sea animals living beneath, and on the vertical connectivity of food webs, both within the water column and between the pelagic and the benthic realms. A number of studies have also attempted to estimate feeding rates and the fluxes of energy between ecosystem components to inform spatially explicit food web modeling.

Feeding in the Pelagic Realm

In the pelagic realm, ecological interactions between species are believed to be particularly important drivers of community structure, since there is far less physical heterogeneity in the environment to drive speciation than is found on the seafloor. Where

benthic organisms can "choose" between sessile and mobile morphologies, or can adapt to different substrate types for instance, pelagic fauna can live their entire lives in the three-dimensional, fluid water column and will never encounter a surface that does not belong to another animal. Feeding strategies among metazoans in the deep-pelagic realm are constrained by a number of factors: there is no solid surface (i.e., seafloor) to accumulate organic matter and support benthic prey; ambush predation is highly reduced because of the relatively structureless environment; there is insufficient particulate organic matter to support filter feeding (a ubiquitous strategy in coastal waters); and herbivory is precluded by the lack of primary production. These constraints result in four predominant feeding modes in the deep-pelagic realm: mucus-trapping of particulate matter, zooplanktivory, micronektivory, and generalist predation. Mucus-trapping is exhibited by a relatively narrow range of gelatinous taxa, particularly appendicularins and salps, and one group of pelagic molluscs (shelled pteropods).

Zooplanktivory is by far the most ubiquitous strategy among metazoans, both in terms of individuals, species, and biomass; the most abundant species (larger zooplankton, fishes, and shrimps) consume zooplankton primarily, as do the larvae of nearly all pelagic teleosts and crustaceans. Within this feeding mode (= "guild"), the most commonly consumed prey taxa are calanoid copepods, the most energy-dense prey per unit weight known. A wide variety of deep-pelagic fishes are specialized to prey on micronekton (small fishes, shrimps, and cephalopods). While some taxa are agile swimmers able to pursue and capture prey (e.g., barracudinas), they majority of species adopt a sit-and-wait predation style, facilitated by long bodies with well-developed lateral lines (e.g., pelagic eels) or luminescent lures suspended on chin barbels (dragonfishes) or modified fin rays (deep-sea anglerfishes). Available data suggest that trophic coupling between micronektivores and their prey is extremely tight, with the former consuming the majority of the annual production of the latter. Generalist feeding, taking a wide variety of any prey that is encountered in the food-poor deep sea, is far less common than originally assumed. The best examples of generalist feeding is found within the bathypelagic zone, and include pelican eels with extremely large but weak jaws, and fangtooths, with massive jaws and fangs.

A form of zooplanktivory that has been highlighted during the last decade of research in deep-pelagic, demersal and benthic food webs is consumption of gelatinous zooplankton. In the water column, the "jelly web" can dominate intermediate trophic levels, particularly as depth increases. Gelatinous taxa are then consumed by a number of biomass-dominant deep-pelagic (e.g., deep-sea smelts) and demersal (e.g., slickheads) fishes. Occasional mass accumulations of dead or dying jellyfishes on the seafloor also provide important energy sources for demersal and benthic predators, scavengers, and benthic epi- and infauna. These findings are particularly important because it had long been assumed that jellyfish, with a high water content and low nutritional value, represented a trophic "dead end"; that is, they provided insufficient nutrition to support higher trophic levels, compared to the far more nutritious zooplankton-based foodwebs. However, these new studies suggest that jellyfishes may be an important energy source reaching the deep seafloor.

Feeding in the Benthic Realm

Where trophic specialists are common in the pelagic ocean, generalist feeders appear to be more common across the seafloor. With the exception of herbivory, all feeding types are found in the deep-benthic realm. Suspension feeding occurs when organisms feed on particles in the water column, which may be passively transported to the organism by water currents, or actively captured by moving through the water. Suspension feeding is the preferred mode for most sessile epifaunal animals, including Cnidaria (e.g., corals, anemones and hydroids), sponges, sea lilies and sea squirts, though it is also used by some errant megafauna, particularly brittle stars and sea stars. Suspension feeding is most common on the continental slopes, and at locations where elevated current speeds and particulate organic matter are found (e.g., around some seamount and canyons), but becomes less important with increasing depth.

As depth increases, deposit feeding becomes considerably more common. Where rates of particulate matter deposition are relatively high, subsurface deposit feeding by infaunal organisms tends to be more common, whereas in oligotrophic environments, surface deposit feeding by epifauna is more common. On the abyssal plains, sea cucumbers and sea urchins are common deposit-feeding epifauna that feed by ingesting sediment as they traverse its surface.

Predation is observed at all depths. At the highest trophic levels, ambush predation is exhibited by top predators (lizardfishes, sharks, and skates), and piscivory (fish consumption) predominates. Micronektivory is an important deep-demersal feeding guild, particularly in regions where Deep Scattering Layers intersect bottom topography, such as continental slopes, seamounts, mid-ocean ridges, and canyons. Members of this guild, including commercially important fishes (e.g., orange roughy, grenadiers), may venture a considerable distance above the bottom to utilize this food source. Closer to the seafloor, hyperbenthic predation on crustaceans is a major trophic guild, conducted by sit-and-wait predators (e.g., tripod fishes) and active searchers (e.g., rattails, skates, sharks, cusk-eels, morid cods). Epifaunal browsing is a feeding guild where organisms pick small benthic crustaceans and polychaetes off the seafloor without ingesting sediment (e.g., crabs, squat lobsters, spiny eels). It is interesting that sessile invertebrates (e.g., anemones, sponges, corals) are rarely consumed in the deep sea. Infaunal predation (digging into sediment) is rare in the deep sea, unlike the coastal zone where this strategy is common, and appears to be adopted by a limited fauna (eelpouts, flatfishes, amphipods).

While deep-seafloor trophic dynamics are largely driven by the flux of particulate organic matter from the sea surface, much of this material will have been consumed, egested and reworked by organisms and microbes in the water column and on the seafloor before it is consumed due to the slow sinking rates. Larger food falls sink far faster, and generally arrive at the seafloor quickly to provide a high-quality, energy-rich food source for animals that can detect them and reach them before they are consumed by

other scavengers. Unsurprisingly, scavenging and detritivory are important trophic guilds among benthic animals. The preponderance of these guilds is seen in the speed at which food falls are located and consumed, in some cases minutes to hours, depending on the size of the food fall. The largest whale falls can provide a local source of energy for scavengers for months, and even once the soft tissue has been consumed, the skeleton can provide energy and hard substrate for suspension-feeding organisms for decades. Wood falls and deposits of plant material are also readily consumed by detritivores at the seafloor when it is available. Key organisms in these guilds include sharks, rattails, crabs, hagfishes, cutthroat eels, and amphipods, the latter of which are often the first to arrive. Some fishes (e.g., snailfishes) feed directly on amphipods attracted to food falls, a modified form of hyperbenthic feeding.

Chemoautotrophy

While the majority of food reaching the deep sea can ultimately be traced back to a photosynthetic source, chemosynthesis is known to be a locally important source of energy at the seafloor, though new research suggests that bacterial chemoautotrophy may be a far more widespread source of carbon in the "dark ocean" (i.e., deep-pelagic depths) than previously recognized. Chemosynthesis occurs where bacteria are able to derive energy from chemicals in the water column, rather than using solar light.

Ecosystem Services: The Importance of Understanding the Ecology of the Deep Sea

The Functioning and Services of the Deep Sea

The deep-sea fauna serve vital functional roles in deep-sea ecosystems including carbon sequestration, biomass production, sediment bioturbation and stabilization, organic matter decomposition, and nutrient regeneration. The primary challenges in identifying deep-sea ecosystem services are the many knowledge gaps about the ecology of deep-sea ecosystems and the prevalence of intermediate services relative to final services. Intermediate services include the biologically mediated habitat, nutrient cycling, resilience and resistance, and water circulation and exchange, whereas final services include carbon storage and sequestration, food provision, genetic resources, and waste absorption and detoxification. In many places the bathypelagic and benthos represent important reservoirs of marine biodiversity. Biodiversity is highly correlated to deep-sea ecosystem functioning and increases the resilience of deep-sea ecosystems and their ability to respond to disturbance (see following *Threats* section). The deep sea absorbs ~25% of anthropogenic carbon emissions, a critical service influencing climate. The Gulf of Mexico provides a prime example of waste detoxification services by the deep sea; microbial communities degraded hydrocarbons released by the *Deepwater Horizon* oil spill. The deep sea also provides cultural services in the form of human intrigue and excitement of discovery, which often drives technological advancement (e.g., submersibles, ROVs, landers). This excitement has driven substantial economic investment in the form of physical (ships, sensors, gear) and academic infrastructure. In summary, deep-sea ecosystem services are vital to human well-being.

Threats to the Fauna and Ecology Deep Sea

The deep sea was once considered too large to be disturbed by humans. Decades of declines in deep-sea fisheries landings, along with point-source disasters and growing interest in destructive resource extraction (e.g., deep seafloor mining), has changed this notion. Given that the deep sea is a highly interconnected ecosystem tied to the processes of the surface ocean, such as major circulation patterns and primary production, it is not immune to the effects of climate change. Anticipated ecosystem changes include a more stratified ocean, with implications for the timing and quantity of plankton production, as well as range shifts as warm-water taxa expand their distributions towards the poles, presumably at the expense of polar taxa. A full treatment of climate change as a driver of ecosystem change in the future are beyond the scope of this article, so herein we will describe some of the immediate and ongoing threats to the deep sea.

The largest category of threat to deep-sea ecosystems involves resource extraction, both living and nonliving. As human populations expand exponentially in a Malthusian trajectory, so too does the need for food. As coastal fisheries become increasingly depleted due to overharvesting, fisheries have continued a decades-long trend in fishing deeper. That, combined with the increasing harvesting power of modern vessels with technological advancements (e.g., scanning sonar) and the low recovery potential of many slow-growing deep-sea fishes and habitat-forming benthic invertebrates, makes deep-sea ecosystems particularly vulnerable to trophic cascades, regime shifts, and ultimately ecosystem collapses. Historically, deep-sea fisheries have focused on relatively few species of pelagic and demersal species, with Gadiformes (cods and their allies) comprising the majority. These fisheries have been so intense in some regions (e.g., North Sea) that they have fundamentally altered the very nature of the habitat —trawl tracks have now become the dominant topographic features. Though a few of the world's smaller deep-sea fisheries are considered sustainable, other deep-sea fishes, such as orange roughy and slender armorhead, have been effectively fished to economic extinction. Mesopelagic lanternfishes likely represent the world's largest untapped fisheries resource, and interest in harvesting these fishes is rapidly increasing. If pelagic fisheries follow the same trajectories as other deep-sea fisheries, "sustainable yield" fisheries are unlikely. Fishing will simply remove individuals faster than they can replace themselves. The impact fisheries can have on deep-sea floor can be locally devastating—in the northwest Atlantic, for instance, the effect of bottom trawling below 200 m has an impact at least an order of magnitude higher than all other human activities combined.

Non-living resource extraction also has a profound effect on deep-sea ecosystems, as evidenced by the *Deepwater Horizon* oil spill in the Gulf of Mexico at 1500 m depth. While the full extent of the damage from this event may not be known for some time, the immediate effect on the deep-sea biota have been dramatic—loss of deep-sea corals, reduction in pelagic fish numbers, and mass mortality of marine mammals and seabirds. Once considered "deep," the *Deepwater Horizon* wellhead (Macondo) site now sits shallower than the average depth of deepwater drilling in the Gulf of Mexico. As the risk of accidents increases dramatically with increasing depth of drilling, it is likely that *Deepwater Horizon* will not be the last such deep-sea natural disaster. An emerging field of resource extraction is deep-sea mining, which will alter deep-benthic communities for centuries, perhaps millennia. As of 2013, the International Seabed Authority, the entity regulating deep-seafloor mining, had entered into 11 15-year contracts with several nations to mine a section of the Pacific twice the size of Germany. These activities, plus the ongoing use of the deep sea for waste disposal, call for a new perspective on deep-sea stewardship, one where we move away from a frontier mentality of exploitation and towards an ecosystem health and services management mentality.

See also: Ecological Data Analysis and Modelling: Grassland Models. Evolutionary Ecology: r-Strategists/K-Strategists. General Ecology: Biodiversity. Global Change Ecology: Emergence of Climate Change Ecology

Further Reading

General

Gage, J.D., Tyler, P.A., 1991. Deep-sea biology: A natural history of organisms at the deep-sea floor. Cambridge University Press.
Herring, P., 2002 The biology of the deep ocean. Oxford, United Kingdom: Oxford University Press.
Jamieson, A., 2015. The Hadal zone: Life in the deepest oceans. Cambridge University Press.
Rex, M.A., Etter, R.J., 2010 Deep-sea biodiversity: Pattern and scale. Harvard University Press.

Specific Topics

Danovaro, R., Snelgrove, P.V.R., Tyler, J.A., 2014. Challenging the paradigms of deep-sea ecology. Trends in Ecology & Evolution 29, 465–475.
Drazen, J.C., Sutton, T.T., 2017. Dining in the deep: The feeding ecology of deep-sea fishes. Annual Review of Marine Science 9, 337–366.
Levin, L.A., Dayton, P.K., 2009. Ecological theory and continental margins: Where shallow meets deep. Trends in Ecology & Evolution 24, 606–617.
Levin, L.A., Etter, R.J., Rex, M.A., Gooday, A.J., Smith, C.R., Pineda, J., Stuart, C.T., Hessler, R.R., Pawson, D., 2001. Environmental influences on regional deep-sea species diversity. Annual Review of Ecology and Systematics 32, 51–93.
Levin, L.A., Le Bris, N., 2015. The deep ocean under climate change. Science 350, 766–768.
McClain, C.R., Allen, A.P., Tittensor, D.P., Rex, M.A., 2012. Energetics of life on the deep seafloor. Proceedings of the National Academy of Sciences of the United States of America 109, 15366–15371.
Mengerink, K.J., Van Dover, C.L., Ardron, J., Baker, M., Escobar-Briones, E., Gjerde, K., Koslow, J.A., Ramirez-Llodra, E., Lara-Lopez, A., Squires, D., Sutton, T.T., Sweetman, A.K., Levin, L.A., 2014. A call for deep-ocean stewardship. Science 344, 696–698.
Pitcher, T.J., Morato, T., Hart, P.J.B., Clark, M.R., Haggan, N., Santos, R.S., 2007. Seamounts: Ecology, fisheries & conservation. Oxford, UK: Blackwell Publishing.
Priede, I.E., 2017. Deep-sea fishes: Biology, diversity, ecology and fisheries. Cambridge University Press.
Ramirez-Llodra, E., Brandt, A., Danovaro, R., De Mol, B., Escobar, E., German, C.R., Levin, L.A., Arbizu, P.M., Menot, L., Buhl-Mortensen, P., Narayanaswamy, B.E., Smith, C. R., Tittensor, D.P., Tyler, P.A., Vanreusel, A., Vecchione, M., 2010. Deep, diverse and definitely different: Unique attributes of the world's largest ecosystem. Biogeosciences 7, 2851–2899.
Robison, B.H., 2004. Deep pelagic biology. Journal of Experimental Marine Biology and Ecology 300, 253–272.
Smith, C.R., De Leo, F.C., Bernardino, A.F., Sweetman, A.K., Arbizu, P.M., 2008. Abyssal food limitation, ecosystem structure and climate change. Trends in Ecology & Evolution 23, 518–528.
Sutton, T.T., 2013. Vertical ecology of the pelagic ocean: Classical patterns and new perspectives. Journal of Fish Biology 83, 1508–1527.
Van Dover, C.L., German, C.R., Speer, K.G., Parson, L.M., Vrijenhoek, R.C., 2002. Marine biology—Evolution and biogeography of deep-sea vent and seep invertebrates. Science 295, 1253–1257.
Widder, E.A., 2010. Bioluminescence in the ocean: Origins of biological, chemical, and ecological diversity. Science 328, 704–708.
Woolley, S.N.C., Tittensor, D.P., Dunstan, P.K., Guillera-Arroita, G., Lahoz-Monfort, J.J., Wintle, B.A., Worm, B., O'Hara, T.D., 2016. Deep-sea diversity patterns are shaped by energy availability. Nature 533, 393–396.

Ecosystem Health Indicators—Freshwater Environments

Adam D Canning, Wellington Fish and Game Council, Palmerston North, New Zealand
Russell G Death, Massey University, Palmerston North, New Zealand

Ecosystem Health Indicators—Freshwater Environments

Waterways are akin to the lymphatic system and kidneys of the land in that they drain and process chemicals and material leaving the land. Like lymphatic systems and kidneys, if excessively stressed the health of that system will deteriorate. Anthropogenic stressors of freshwater systems include nutrient enrichment, urbanization, industrial waste, deforestation, water abstraction, flood prevention engineering, sedimentation, dam construction, climate change and invasive species (Foote et al., 2015; Dudgeon, 2014; Cucherousset and Olden, 2011; Woodward et al., 2010; Dewson et al., 2007; Dudgeon et al., 2006). If humans are unwell they will usually see a physician who will assess their "health" with a range of physical, psychological and biochemical indicators. Likewise, good environmental management practices often require that the health of an ecosystem is measured and maintained using multiple indicators (Rapport et al., 2009; Kundzewicz et al., 2007; Steedman, 1994). Though, what is meant by the notion of "healthy ecosystem" and how we measure it is currently far more challenging than for human health. As humans assessing the natural world, we often personify our examinations and in so doing we interpret our observations with a human-lens, this extends to the way we define "ecological health" (Palmer and Febria, 2012; O'Brien et al., 2016; Friberg et al., 2011; Whitfield and Harrison, 2014) In humans, Sartorius (2006) has identified three broad definitions of health:

> The first is that health is the absence of any disease or impairment. The second is that health is a state that allows the individual to adequately cope with all demands of daily life (implying also the absence of disease and impairment). The third definition states that health is a state of balance, an equilibrium that an individual has established within himself and between himself and his social and physical environment.

From these definitions, it is clear that, as humans, we associate health with the absence of stressors, the ability to resist and be resilient to stressors, and that a state of balance is achieved implying that a level of integrity is maintained despite external stressors (persistence). In freshwater ecology, O'Brien et al. (2016) found almost 200 studies directly assessing ecosystem health and over 1000 using the term in a broader context, yet there were vast differences in the way "health" was defined and interpreted. Here we do not attempt to synthesize a new definition of "ecosystem health," but adopt Costanza and Mageau's (1999) definition. This definition not only resembles broad similarity to the concepts many associate with medical health but is also sufficiently generic to apply at all levels of scale. Costanza and Mageau (1999) state that:

A healthy ecosystem is one that is sustainable—that is, it has the ability to maintain its structure (organization) and function (vigor) over time in the face of external stress (resilience).

Structure or organization measures the assembly of a community, it includes species diversity, community composition and food web topology. Whereas function or vigor is a measure of an ecosystems activity; function includes concepts such as the productivity, throughput, cycling and metabolism. The maintenance of structure and function over time in the face of external stress reflects the stability of an ecosystem. Whereby stability is a multifaceted concept that can be divided broadly into resistance (the ability to remain unchanged from stress) and resilience (the capacity and timeliness to return to preperturbation conditions). Given that the term "ecosystem" encompasses both biotic and abiotic factors, in addition to community-level measures (biotic), we also include indicators of habitat (abiotic).

If we are to halt the decline, maintain or restore freshwater ecosystem health then we need to adequately monitor change across all three components of health. Here we introduce each of the common indicator classes (both biotic and abiotic indicators) and discuss their use and interpretation; we do not discuss human-related environmental health indicators (such as those used to assess swimmability). Table 1 categorizes common measures of freshwater ecosystem health as being indicators of organization (structure), vigor (function) or resilience.

Biotic Indicators

Species diversity

Probably one of the oldest and simplest indicators of ecosystem health is species richness. It is usually assumed that a decline in species richness is associated with a decline in health and increasing anthropogenic disturbance. For example, Relyea (2005) showed a 22% reduction in zooplankton diversity in pond mesocosms following overspray of the common pesticide RoundUp^{T} (Glyphosate). However, in many instances species richness is not altered as more tolerant species replace sensitive species. Furthermore, in some situations, small to moderate disturbance can increase species diversity by reducing dominant species and opening new niches—the premise of the intermediate disturbance hypothesis (Townsend et al., 1997b; Resh et al., 1988; Ward and Stanford, 1983). For instance, Death and Zimmermann (2005) found that deforested stream sites had greater macroinvertebrate diversity than sites within the Taranaki native forest park (New Zealand), though the species composition was considerably different. Although anthropogenic disturbance increased species diversity, ecosystem health reduced as ecosystem *structure* was *not maintained* (community composition changed) in the face of *external stress* (deforestation). Thus, species richness can be a misleading indicator of ecosystem health. This has led to the development of indicators that focus on changes in community composition. In rivers, the most common richness based indicator is the number of EPT (Ephemoptera, Plecoptera and Trichoptera)

Table 1 Freshwater ecosystem health indicators and whether they measure structure, function, resilience and/or abiotic components of ecosystems

	Aspect of ecosystem health			
Indicator	*Structure*	*Function*	*Resilience*	*Abiotic*
Species diversity	X			
Community composition	X			
Algal biomass	X			
Species tolerance indices	X			
Indices of biotic integrity	X	?		
Observed/expected ratios	X			
Ecosystem metabolism		X		
Decomposition rates		X		
Primary production		X		
Secondary production		X		
δ^{15}N of consumers		X		
Eco-exergy		X		
Nutrient cycling		X	X	
Relative ascendency		X	X	
Indirect flows		X	X	
Post disturbance recovery	X	X	X	
Nutrient concentrations				X
Toxic chemicals				X
Dissolved oxygen				X
Deposited and suspended sediment				X
Temperature				X
Physical habitat				X
Flow metrics				X

taxa; species richness is also central to other indicators such as Karr et al.'s (1986) index of biotic integrity and observed/expected ratios, such as AUSRIVAS (Simpson and Norris, 2000).

Community dissimilarity

Rather than using species diversity as an indicator of health, multivariate community distance metrics (e.g., Bray-Curtis similarity Bray and Curtis, 1957), can indicate community dissimilarity from the prestressed state or a natural reference site(s). For example, Cao et al. (1996) showed, using the CY dissimilarity metric (Cao et al., 1997), a clear increase in macroinvertebrate community structure dissimilarity (from a reference site) along a sewage effluent pollution gradient in the Trent River (UK) catchment.

Dissimilarity based metrics usually perform well as metrics of deviation in ecosystem *structure* in the face of *external stress*; however, not all dissimilarity metrics are equal in their sensitivity to change in all situations. Cao and Epifanio (2010) showed that the Bray–Curtis index, CY similarity index and Canberra Metric responded more linearly than Chao's abundance-based Jaccard and Sørenson coefficients, and Morisita index (which showed strong nonlinear changes) to linear changes in stream macroinvertebrate assemblages over a simulated stress gradient. Therefore, before a dissimilarity metric is used, it must be clear that the metric is likely to respond in an ecologically sensible way for a given disturbance and community type. Furthermore, whilst dissimilarity metrics directly measure differences in community structure, they do not provide information on actual species differences that drive community differences.

Algal biomass

Nutrient enrichment, deforestation, water abstraction and increased temperature can all result in increased algal growth (Dodds, 2007) which can alter higher trophic levels through increased risk of hypoxic conditions and changes in food resources (Smith and Schindler, 2009; Bricker et al., 2008; Dodds, 2007). Algal biomass can, therefore, indicate changes in community structure. Algal biomass indicators can usually be classified as either measures of benthic cover or pigment density. Benthic cover measures usually apply to shallow rivers, streams and lake edges, where light penetrates through to gravels, and is often broken down into percent cover of filamentous algae (coarse (thick filaments), slimy (thin filaments), long and short) and algal mats (thick and thin). Algae biomass can also be estimated indirectly via the concentration of pigments as the two are strongly correlated (Hawkins et al., 1982). The most commonly assessed pigment is Chlorophyll a (green/yellow pigment), measured usually with fluorometry or spectrophotometry (Biggs and Kilroy, 2000). Pigment abundance can then be calculated as either a concentration (from pelagic samples) or as an area-density (from benthic samples). As a guideline, Welch et al. (1988) suggests that for periphyton (benthic algae) a chlorophyll *a* density of 100–150 mg m^{-2} and a filamentous algae cover of ~20% may represent a critical level for aesthetic nuisance. Dodds et al. (1998) suggest that oligotrophic streams have an annual mean (and maximum) chlorophyll *a* biomass less than 20 (60) mg m^{-2}, while eutrophic streams experience more than 70 (200) mg m^{-2}. According to the OECD (1982) trophic classification scheme for lakes, oligotrophic lakes have an annual mean (and maximum) chlorophyll *a* biomass of 2.5 (8.0) mg m^{-3} and eutrophic lakes greater than 8.0 (25) mg m^{-3}. Perhaps the biggest issue with algal monitoring is that algae have high spatial and temporal variability, even within a single stream reach. Algae can be very patchy and biomass can change considerably in the order of hours (Biggs and Kilroy, 2000).

Species tolerance indices

Species tolerance indices were one of the first uses of biological information to assess water quality, one component of ecosystem health (Rosenberg and Resh, 1993; Allan, 1984). These include indices such as the Saprobic Index in Europe (Reynoldson and Metcalfe-Smith, 1992), the Biological Monitoring Working Party (BMWP) index in the United Kingdom, the Hilsenhoff Index in the United States (Hilsenhoff, 1987), the SIGNAL (stream invertebrate grade number average level) Index in Australia (Chessman, 2003; Chessman et al., 1997) and the Macroinvertebrate Community Index (MCI) in New Zealand (Boothroyd and Stark, 2000).

The indices are designed to simplify a mix of complex community data to derive a single score as a sum of water quality tolerance scores assigned to each taxon. They are generally specific to a particular type of pollution (most often organic enrichment), a condition that limits or confuses their value as more general ecosystem indicators. In recent years they tend to have been replaced by using observed/expected ratios of species present compared with predictions of community composition for a particular waterbody type (Clarke and Murphy, 2006; Wright et al., 2000).

Body-size indicators

Since Odum (1953), average body size has been predicted to reduce with increasing stress. In lakes, increasing phytoplankton cell-size and decreasing zooplankton body-size are associated with decreasing ecological health. In rivers, diatom cell-size also increase and benthic invertebrate body-size is reduced with increasing stress (Stevenson and Pan, 1999; Townsend et al., 1997a). Havens and Hanazato (1993) reviewed a large number of studies that assess the influence of acidification and pesticide contamination on zooplankton body size in lakes (including surveys, experimental manipulations and mesocosm experiments), primarily from North America and Scandinavia. With a few exceptions, the general pattern is that as acidification and pesticide contamination increase, the macrozooplankton (e.g., large daphnids, predatory cladocerans and copepods) often become extinct and are increasingly replaced by microzooplankton (e.g., rotifers, small cladocerans and crustaceans). It is hypothesized that small bodied zooplankton are more tolerant of chemical stressors because they are r-strategists, have greater species richness (increasing probability of tolerant taxa being present), reduced predation (as *Daphnia* become extinct), fewer molts before maturity, have a lower calcium content and greater tolerance of toxic metals. However, the responses of body-size to multiple chemical stressors is not well understood (Havens and Hanazato, 1993).

Indices of biotic integrity

The index of biotic integrity (IBI) was originally created to assess the condition of riverine fish assemblages in the Illinois River (Karr et al., 1986), and has since become adopted worldwide for both rivers and lakes (Beck and Hatch, 2009; Ruaro and Gubiani, 2013), to assess fish (e.g., Zhu and Chang, 2008; Fayram et al., 2005; Joy and Death, 2004a; Lyons et al., 2000a; Kamdem Toham and Teugels, 1999; Minns et al., 1994), macroinvertebrates (e.g., Lunde and Resh, 2012; Perera et al., 2012; Masese et al., 2009; Raburu et al., 2009; Weigel et al., 2002), macrophytes (e.g.,Beck et al., 2010, 2013), and phytoplankton (e.g., Wu et al., 2012; Maulood et al., 2011; Kane et al., 2009). The term "integrity" refers to a condition that has not been anthropogenically altered. IBIs assess multiple indicators of ecological conditions relative to an un-impacted condition. Ecological metrics are often related to taxonomic richness, functional groups, and community composition. In developing an IBI, the first step is to identify measurable ecological changes over a human disturbance gradient, accounting for natural environmental drivers. The gradient should include sites in pristine condition through to highly disturbed sites. A score is applied to each metric (high scores indicate more similarity with pristine condition), the sum of scores across all chosen metrics yields the IBI score, reflecting the deviation from integrity at reference condition.

Observed/expected ratios

Predictive modeling approaches to the environmental assessment of running waters are particularly appropriate where the focus of monitoring is on overall ecosystem health, not just water quality (Norris and Thoms, 1999; Reynoldson and Metcalfe-Smith, 1992; Reynoldson et al., 1997; Bailey et al., 2004). An important reason for this is that they combine information on a variety of environmental variables and the associated macroinvertebrate or fish assemblages. RIVPACS, developed in the United Kingdom was the first in a suite of similar multivariate approaches (Australia (Hart et al., 2001a), Canada (Reynoldson et al., 2001), USA (Van Sickle et al., 2005), Ireland, New Zealand (Joy and Death, 2000, 2003, 2004b), and Indonesia (Hart et al., 2001b)).

A range of sites are sampled in "reference" or "least impacted" condition for their biological communities. A model is constructed predicting the probability of occurrence of taxa at those sites based on measurements of a suite of environmental variables that are not affected by human activity. A test site is then assessed by evaluating how many of the collected taxa at a site are predicted to be present if there is no anthropogenic disturbance and is expressed as an "observed/expected" ratio. Globally, this is probably the most widely adopted approach for assessing the structural component of ecosystem health, although to the best of our knowledge it has only been applied to running water and not still water habitats.

Ecosystem metabolism

Ecosystem metabolism represents oxygen fluxes in aquatic ecosystems, usually measured in g O_2 m^{-3} s^{-1}, and are one of the most common indicators of ecosystem function. Primary production increases during the day and photosynthesis produces oxygen, but at night those same plants respire dropping oxygen levels. At night the community becomes a net user of oxygen, causing dissolved oxygen (DO) levels to drop. Therefore, to measure ecosystem metabolism, DO, temperature, and depth (and potentially light) need to be monitored continuously for at least one full diurnal cycle. The R package streamMetabolizer (Appling et al., 2016) provides a useful platform to calculate ecosystem metabolism using several alternative calculations.

Ecosystem metabolism is most often broken down into gross primary production (GPP), ecosystem respiration (ER) and net primary production (NPP), such that NPP = GPP – ER. When GPP/ER is <1 then the community is using more energy than it fixes and is reliant on allochthonous inputs (e.g., incoming drift and leaf litter); whereas when GPP/ER is >1 the community uses less energy than it fixes, this either presents itself as increased biomass (usually algal biomass) or increased drift. Ecosystem metabolism can, therefore, impact the food base of an aquatic community. A review by Young et al. (2008) has shown that ecosystem metabolism is responsive to numerous environmental changes, these are summarized in Table 2.

Decomposition rates

Rates of leaf litter decomposition provide another indicator of ecosystem function widely used in rivers (Boyero et al., 2011; Sponseller and Benfield, 2001). Typically, this involves leaving preweighed leaf litter in mesh bags for several weeks at a site and then determining how much biomass is lost over the period. Bags are best secured close to the stream bed to reduce the influence of hydraulic conditions; bags floating in the water column can influence decay rates (Mutch et al., 1983). In many cases, leaf litter decomposition will be size-dependent, so it can be useful having multiple bags with different mesh apertures and different detritus sizes (Boulton and Boon, 1991). A key difficulty with leaf litter decomposition is that it can be difficult comparing between sites if leaf litter composition is not the same, as some leaves and twigs will decompose faster than others. An alternative technique that may circumvent that issue is using cotton strips and popsicle sticks instead of leaf litter (Tank and Winterbourn, 1996; Hildrew et al., 1984; Egglishaw, 1972). Cotton decomposition can be either measured as loss of biomass or loss of tensile strength. Cotton strips and popsicle strips are inexpensive and considerably reduce sample variability. A review by Young et al. (2008) has also shown that decomposition rates are responsive to numerous environmental changes, these are summarized in Table 3.

Secondary production

Secondary production measures the rate of biomass generation for given taxa and can indicate changes in ecosystem function. Common methods for calculating secondary production include the size-frequency method, increment-summation method and the Allen Curve method; each of these methods attempts to quantify any missing biomass from each cohort (Morin et al., 1987; Wildish and Peer, 1981). The methods quantify the eliminated individuals when the population is at equilibrium and thus the

Table 2 The responses of ecosystem metabolism indicators to specific stressors

Environmental change	*GPP*	*ER*	*GPP/ER*
Decreasing forest cover	↑	–	↑
Increasing light	↑	–	↑
Increasing temperature	↑	↑	?
Increasing fine sediment cover	↓	↑	↓
Increasing suspended sediment concentration	↓	–	↓
Reduced hyporheic connectivity	–	↓	↑
Decreasing pH	–	↓	↓
Increasing nutrient concentrations	↑	–	↑
Increasing toxic inputs	↓	↓	?
Increasing flood frequency	↓	↓	↓
Increasing river drying	↑	↑	?
Aquatic plant removal	↓	↓	?

GPP represents gross primary productivity and ER represents ecosystem respiration. Increases are indicated by *upwards arrows*; decreases are indicated *down arrows*. *Dashes* indicate no relationship, and *question marks* show uncertainty.

Table 3 The responses of decomposition rates to specific stressors

Environmental change	*Decomposition rates*
Temperature increases	↑
Increasing distance from headwaters	↓
Water velocity increases	↑
Increasing bed stability	↑
Increasing deposited sediment	↓
Decreasing pH	↓
Increasing conductivity	↑
Nutrient enrichment	↑
Increased heavy metal	↓
Increased insecticide concentration	↓
Riparian deforestation	↑
Increasing river regulation	↑

Increases are indicated by *upwards arrows*; decreases are *indicated down arrows*.

quantum eliminated is equal to the biomass fixed (i.e., the secondary production). Cross et al. (2006) examined the influence of nutrient enrichment in a detritus-based stream in North Carolina, USA. They found that secondary production in the headwaters was 1.2–3.3 times greater when the stream was experimentally enriched, compared with the preceding 15-year baseline.

δ15N of consumers

The elemental composition of organisms is dependent on dietary elemental composition, that is, "you are what you eat." The proportions of nitrogen (N) stable isotopes are dependent on the source and degree of transformation (fractionation). Anderson and Cabana (2006) assessed δ15N in invertebrates and fish from 82 reaches throughout the Saint-Lawrence Lowlands in Québec, Canada. They found that average δ15N values were highly correlated with watershed nitrogen loads ($r^2 = 0.61$), manure-borne nitrogen loads ($r^2 = 0.62$) and fertilizer-borne nitrogen loads ($r^2 = 0.45$). (Clapcott et al., 2010) found that δ15N increased with nitrogen concentration ($r^2 = 0.66$) in 84 New Zealand Streams, but they also found δ15N increased with proportion of catchment deforestation ($r^2 = 0.58$). In addition to indicating catchment activities, nitrogen and carbon isotopes can, through the use of mixing models, allow changes in dietary composition, and consequently overall food web structure and function to be identified before and after a given stressor (Nielsen et al., 2018; Layman et al., 2007; Belgrano et al., 2004). It is, however, advisable that if stable isotopes are used to compare nitrogen loads between catchments then the same species (ideally a primary consumer) should be used to avoid differences in dietary composition and trophic position from influencing the results. Stable isotope analysis is developing as a promising and affordable methodology to detect changes in land use and ecological functioning.

Eco-exergy

The eco-exergy of an organism or system is the chemical work energy including the work energy embodied in genomic information (Jørgensen and Nielson, 2014). Eco-exergy, therefore, represents the exergy of an ecological community since the time before biological evolution started, approximately 4 billion years ago (Jørgensen and Mejer, 1977; Jørgensen and Nielsen, 2014). As

Table 4 An example calculation of Eco-exergy for a hypothetical estuary

Taxa	*Biomass (g dw/L)*	*Weighting factor*	*Individual eco-exergy (MJ/L)*
Algae	100	3.4	340
Plants	50	60	3000
Jellyfish	40	144	5760
Annelids	40	287	11,480
Fish	60	344	20,640
Eco-exergy of community			*41,220*
Structural Exergy			*142.14*

Individual eco-exergy = biomass × weighting factor. The Eco-exergy of the community is the sum of the individual eco-exergies. Structural exergy is the community Eco-exergy divided by the total community biomass. The weighting factors are used to quantify the energy expended in the organization and development of living organisms and account for the chemical work energy per unit of biomass.

ecosystems mature and genomic information increases, eco-exergy is also expected to increase (Silow and Mokry, 2010; Fath et al., 2001; Patten, 1995). Eco-exergy is not directly measured but calculated by multiplying the biomass density of each taxonomic group (usually g/L or g/m^2) by the genetic information content weighting factors and then summing all individual exergies. The genetic information content conversion factors account for the energy expended in the organization and development of living organisms and account for the chemical work energy per unit of biomass. Structural exergy (relative exergy) is the total eco-exergy divided by the total biomass density—this allows comparison between ecosystems. Table 4 shows the calculation of eco-exergy for a hypothetical estuarine community; using the weighting factors from Marques et al. (1997), the eco-exergy is 41,220 MJ/L and Structural Exergy is 142.14.

The Structural Exergy provides a measure of the ability of an ecosystem to utilize available resources. An ecosystem with high Structural Exergy is hypothesized to have greater network complexity and have better niche utilization than an ecosystem with low Structural Exergy. Eco-exergy has been used to indicate on changes in ecosystem health for a variety of stressors (Jørgensen and Nielson, 2014). Ye et al. (2007) provide just one example, whereby they compared the Eco-exergy and Specific Exergy for macrophytes, phytoplankton and zooplankton from bays in Lake Taihu (China) differing in nutrient enrichment. They found a critical threshold in macrophyte exergy when total phosphorus exceeds 0.05 mg/L. There are several more examples in the literature, but depending whether the authors concentrate on exergy, structural exergy, or eco-exergy.

Network metrics

Assessment of both structural and functional aspects of ecosystem health has been notoriously difficult because of the challenges in measuring higher order emergent properties (Friberg et al., 2011; Rapport et al., 2009; Steedman, 1994). Food webs are networks that not only represent ecological community structure but also depict energy or material flows between species and provide the opportunity for more holistic assessment of energy flows and potential trophic cascades. Weighted food webs have quantified stocks and flows, and are often referred to in the literature as Ecological Networks, although technically weighted food webs are a form of Ecological Networks. Network assembly is laborious and requires data on standing biomasses of each taxonomic group, along with estimates of their productivity, consumption (both dietary links and amount of flow), assimilation efficiency, and boundary drift, a brief outline on network construction is presented in Box 1, with more elaborate details on their construction in Fath et al. (2007).

Ecological Network Analysis (ENA) provides an assortment of metrics for food web properties that can potentially serve as indicators of both structural and functional aspects of ecological health (Jørgensen et al., 2010), such as measures of energy flow distribution and efficiency, the dominance and nature of indirect flows, cycling, stability and mutualistic predator-prey interactions (Latham Ii, 2006; Fath and Borrett, 2006; Fath and Patten, 1999). These are all measures of emergent properties that cannot be readily calculated without a complete food web (Latham Ii, 2006; Fath and Borrett, 2006; Fath and Patten, 1999; Ulanowicz, 1997, 2004). The metrics, along with uncertainty analyses (Hines et al., 2018), are readily calculable in the R packages enaR (Borrett and Lau, 2014) and NetIndices (Soetaert and Kones, 2008).

Nutrient cycling

Changes in nutrient cycling can indicate changes in ecosystem function. Nutrient cycling cannot be measured directly but calculated from a quantitative network of nutrient cycling. Nutrient cycling can be calculated using a variety of methods, (Fath and Halnes, 2007; Latham Ii, 2006; Fath and Patten, 1999; Allesina and Ulanowicz, 2004; Vanni, 2002), with Finn's Cycling Index (Finn, 1976) being most common. Nutrient cycling can indicate changes in productivity and consumption, and can influence food web resilience and resistance, though to date has not been well-linked to robustness—which is more dependent on link distribution (Schaeffer et al., 1988; Canning and Death, 2017; Zhao et al., 2017; Mougi and Kondoh, 2016; Saint-Béat et al., 2015).

Relative ascendency

Relative ascendency (RA; Eq. 1) measures both the degree to which energy flows are confined to specialist pathways and the energetic size of a web. Ulanowicz (1997) hypothesized that communities have a propensity for increasing ascendency as they

Box 1 A brief summary on the methods to assemble an ecological network

To assemble an ecological network, first one needs to thoroughly sample the density of all species across all trophic levels within a community. Often the currency of weighted food webs is in terms of nitrogen, carbon or energy, in which case the researcher needs to convert count density estimates into biomass density estimates. The conversion is usually done through a combination of direct biomass measurements, or size estimates followed by conversion through published empirical size-mass relationships. Currency conversion ratios are often drawn from typical literature values (e.g., Brey et al., 2010). As flows between species are not readily quantifiable, flow quantums are estimated by mass-balancing the losses in energy/biomass through production, respiration and excretion. Production can be estimated directly through estimates of growth, such as the size-frequency method (Hamilton, 1969), mass turnover rates (Elwood and Nelson, 1972), observed growth rates (McIntire and Phinney, 1965), or indirectly through models of direct measures of production from the same or similar species (e.g., Brey, 2012; Robertson, 1979). Respiration rates are almost always estimated from chamber experiments that measure oxygen depletion under different scenarios (e.g., Elliott, 1976; Forster, 1981; McIntire and Phinney, 1965), though many species have never been assessed in a respirometer and their respiration is estimated from empirical models of similar species (Gillooly et al., 2001; Clarke and Johnston, 1999; Robinson et al., 1983). Direct measures of excretion rates include measuring fecal biomass or mass turnover rates in experimental conditions (Pandian and Marian, 1986); as with many species, excretion rates have never been estimated and are inferred from similar species. In accordance with thermodynamics, all inputs to a species should balance the outputs, that is, consumption = production + respiration + excretion. If there is data on consumption, though data on another component is missing (e.g., respiration), then the equation above can be used to estimate the missing component (Fath et al., 2007). To ascertain where the energy/biomass for each consumer enters from, the feeding links and dietary proportions need to be estimated. Dietary composition can either measured directly from gut content analysis (at the site or from previous analysis in a similar system) or indirectly from biochemical analysis or functional feeding groups (Hladyz et al., 2011; Fath et al., 2007). Once the flows between all species have been quantified, typically there are discrepancies in the data or from natural patterns that mean the food web is not at steady, that is the network inputs do not balance the outputs. Network analysis often relies on the webs being a steady-state snapshot; therefore, webs are often balanced by adjusting the flows between species and/or system imports and exports—this is usually done manually or by applying a balancing algorithm (e.g., Baird et al., 2009; Allesina and Bondavalli, 2003). Once balanced, the network structure and functioning can then be assessed and validated (Fath et al., 2007). If one wishes to simulate temporal dynamics within a food web then the webs are not balanced to steady-state, rather a differential equation (with predefined directions of control, i.e., bottom up, top down or mixed) is written for each link within a food web (such as a Lotka-Volterra equation or similar) and these are each integrated (solved) across time-steps.

mature in the absence of major disturbance. That is, in ecosystems with stable environments, the probability of prey being lost reduces, thus the need to have species with generalized, highly adaptable diets reduces. Therefore, it becomes advantageous for species to have specialized diets that are more efficient at processing prey and having greater energetic throughput. On the contrary, highly disturbed or immature communities have low ascendency which means species have multiple parallel dietary links and lower energy throughput, yielding greater stability but reduced energy efficiency.

$$RA = \frac{\mathrm{T}\sum_{i=1}^{s}\sum_{j=1}^{s}\frac{t_{ij}}{T}\log\left(\frac{t_{ij}\,T}{T_{(i+)}T_{(+j)}}\right)}{-\sum_{i,j}t_{ij}\,\log\left(\frac{t_{ij}}{T}\right)} \quad (1)$$

Where s is the number of species; t_{ij} is the flow leaving species i and entering another species j; T is the sum of all flows within the network (total system throughput); $T_{(i+)}$ is the sum of all flows leaving species I; and $T_{(+j)}$ is the sum of all flows entering species j.

Zorach and Ulanowicz (2003) found that of 48 weighted food webs assessed, relative ascendency had a central tendency about ~0.4. Ulanowicz (2009) hypothesized that this may represent a window of vitality, whereby healthy ecosystems persist via a balance between stability and efficiency. It is unclear why the webs have a similar relative ascendency, despite different environments; therefore, it may be premature to consider the "window" as goal ecosystem health or to consider an increase or decrease as good or bad. It would, therefore, be prudent to consider any change in relative ascendency (increasing or decreasing) as a change in ecosystem health.

As an example, Miehls et al. (2009) compared the relative ascendency of the Lake Quinte (Canada) carbon flow network before and after a Zebra Mussel invasion. Relative Ascendency increased by 32%, this was driven by an increase in growth and development (total Ascendency), rather than an increase in redundancy, which may reduce stability (Ulanowicz et al., 2009; Saint-Béat et al., 2015).

Indirect effects

Indirect effects occur when a change in one species affects another species via an intermediate species or pathway (Wootton, 1994). Wootton (1998) identified five ways that indirect effects may present themselves in food webs including interspecific competition, trophic cascades, apparent competition, mutual interference, and mutual exploitation. If a web has strong indirect flows, then trophic cascades can permeate further and more intensely, which can severely impact other species and network flows, thereby reducing robustness (Canning and Death, 2017; Saint-Béat et al., 2015; Zhao et al., 2016). To exemplify use, Salas and Borrett (2011) compared the indirect/direct flows (sum of indirect flows/sum of direct flows) before and after the Zebra Mussel invasion at both Lake Oneida and Lake Quinte (Canada) (Miehls et al., 2009). In both lakes, the invasion of Zebra Mussel doubled the

proportion of indirect flows (relative to the direct flows), thereby suggesting that disruptive cascades could travel further and be more destructive post Zebra Mussel invasion.

Post disturbance recovery

Resilience is the ability of an ecosystem to recover from a disturbance. Therefore, assessing the timeliness to recover and ecological distance between endpoints of disturbed ecosystems provides a direct indicator of resilience (Yount and Niemi, 1990). The indicators monitored could include, but are not limited to, any of those discussed in this article, depending on the management objectives. In many instances, particularly where comparable field cases are not available, it is more sensible to model the recovery of disturbance, rather than disturbing a system for assessment sake—especially if the disturbance is not recoverable. As an example, Pool and Olden (2015) found that following large discharge from Alamo Dam in Bill Williams River Basic (USA) fish communities took 8 days to recover to preflood assemblages.

Abiotic Indicators—Physicochemical

Nutrient concentrations

The most common nutrients of concern in freshwater ecosystems are nitrogen and phosphorus. Total nitrogen is composed of both inorganic (including nitrate'nitrogen, nitrite'nitrogen and ammonical'nitrogen) and organic forms, with inorganic forms being readily available for plant/algae uptake. Total phosphorus includes both dissolved reactive phosphorus (DRP) and sediment-bound phosphorus. DRP is the form of phosphorus that is readily usable by plants/algae. When nutrients are the limiting factor, their increase can result in increased algae and macrophyte growth. Given that nitrogen and phosphorus are often the limiting factor for algal growth in many systems, these nutrients are also those for which defined limits of degraded/not degraded are most commonly managed to. For example, the ANZECC (2000) (Australian and New Zealand Guidelines for Fresh and Marine Water Quality) guidelines for South-Australia recommend "trigger thresholds" for total nitrogen and total phosphorus in lowland rivers at 0.5 mg/L and 0.05 mg/L respectively.

Toxic chemicals

Abnormal concentrations in many bioavailable chemicals can be toxic to aquatic life. Toxic effects vary widely depending on the toxin, species and even sex of individuals. For example, van den Heuvel et al. (2012) found that female yellow perch (*Perca flavescens*) in a bitumen contaminated lake in Alberta (Canada) had markedly higher levels of testosterone levels than female yellow perch in the reference lake. The male perch hormones were unaffected by the contamination. The number of potentially toxic chemicals are vast, with the ANZECC guidelines identifying protection criteria for over 350 toxicants relevant to Australian rivers.

Dissolved oxygen (DO)

Not only can dissolved oxygen monitoring be useful for calculating ecosystem metabolism (see Ecosystem metabolism section), but dissolved oxygen concentration can also directly affect biotic communities. If DO drops too low, then fish can become stressed or asphyxiated. By way of example, moderate reductions in fish and invertebrate production occur when dissolved oxygen is <5 mg/L and 50% of New Zealand common bullies (*Gobiomorphus*) will not survive an hour below 3 mg/L (Dean and Richardson, 1999; Franklin, 2013). DO will cycle across the day as instream plants switch from producing oxygen with photosynthesis to using oxygen in respiration during the night. Thus, many fish and invertebrate species are unable to survive at a site, regardless of high oxygen concentrations recorded during the day. If DO saturation gets too high (supersaturated) then fish can suffer from gas bubble disease. This is a condition similar to decompression sickness that SCUBA divers get, whereby air embolisms occur in tissue and vessels (Doulos and Kindschi, 1990; Espmark et al., 2010; Geist et al., 2013; Mesa et al., 2000). When oxygen saturation is high, fish will try to swim deeper as the added water pressure compresses air bubbles (Boyles Law), however if water levels are also low and pools are missing then fish can suffer blistering (Fig. 1) and struggle to maintain buoyancy (Shrimpton et al., 1990). Furthermore, Mesa and Warren (1997) found that when juvenile salmonids were exposed to greater than 130% saturation for 3.5 h their predator avoidance ability was significantly reduced and they had extensive gas embolism blocks of the lateral line and gill filaments. Depending on where embolisms have occurred, they can take between 2 h and 4 days to dissipate (Hans et al., 1999).

Deposited and suspended fine sediment

Sediment is a natural component of aquatic systems, which is transported as suspended and bedload sediment, mostly at times of high river flows and floods (Clapcott et al., 2011). Small particles (<2 mm), such as clay and silt, are generally transported in suspension, whereas larger particles, such as sand and gravel, are usually transported close to the riverbed during high flow events (Death, 2008; Schwendel et al., 2011). Erosion from land use activities greatly enhances sediment supply both during low and high flow events (Lyons et al., 2000b; Scheurer et al., 2009; Fahey and Marden, 2006). Sediment levels during floods are also considerably higher in agricultural catchments than similar catchments with native vegetation (Burcher and Benfield, 2006).

Excessive deposited sediment can smother animals directly (Fig. 2) and/or motivate them to leave. It can also smother and bind with the epilithon on rock surfaces that is the food for many aquatic invertebrates and lower the nutritional quality of this food. It fills in the interstitial spaces between rocks (Fig. 2) where benthic fish and invertebrates reside or seek refuge during flood events. Stream invertebrates and many fish (e.g., eels) can live at least up to a meter under the stream bed if there are suitable interstitial spaces (Stanford and Ward, 1988; Williams and Hynes, 1974; Boulton et al., 1997; McEwan, 2009).

Fig. 1 A rainbow trout (*Oncorhynchus mykiss*) with gas embolism blisters. Image courtesy of David Palmer.

Direct impacts of excessive suspended sediment on fish include: mechanical abrasion to the body of the fish and more significantly its gill structures, death, reductions in growth rate, lowered resistance to disease, prevention of successful egg and larval development, and impediments to migration. Indirect impacts include displacing macroinvertebrate communities that provide food, and reducing visual clarity so finding prey is more difficult (Scheurer et al., 2009; Fudge et al., 2008; Herbst et al., 2012; Sternecker and Geist, 2010; Kemp et al., 2011; Collins et al., 2011).

In rivers, deposited sediment is most often measured using a bathoscope to estimate percentage cover or by using a corer to estimate fine deposited sediment concentration. In lakes, deposited sediment is usually measured using a grab-sampler or depth corer. In both rivers and lakes, suspended sediment (and clarity/turbidity) is often measured using a Secchi disk (visible depth measured), a clarity tube, or by measuring spectrophotometric absorbencies (Wren et al., 2000).

Temperature

Water temperature can affect both community composition and dissolved oxygen saturation. As most freshwater organisms are ectotherms, their productivity rates will vary with temperature. As an example, water temperature can drastically impact the success of salmonid spawning. Low temperatures (below approximately 6°C) slow egg growth dramatically and may take twice as long to develop, thus increasing their chance of being killed. On the flip side, high temperatures not only reduce a waters capacity to hold oxygen but will also increase the growth and oxygen demand of other taxa which can, in turn, cause the eggs to suffer from hypoxia. High temperatures also increase the rate at which toxins are assimilated and fatal thresholds may be reached faster. In salmonids, maximal egg survival and development has been found to occur between 8°C and 10°C, with no embryos hatching above 16°C. An upper limit of 12°C is recommended for the success of healthy spawning (Jonsson and Jonsson, 2009; Ojanguren and Brana, 2003). Ojanguren and Brana (2003) showed that embryo size at hatching decreased with increasing temperature. Whilst temperature allows embryos to mature faster, this could result in lower yolk conversion efficiency and consequently smaller body size at hatch (Kamler, 1992). Individuals incubated at low temperatures were larger because they benefited from greater yolk resorption and would consequently be larger at the onset of exogenous feeding (Ojanguren and Brana, 2003). Larger body size is key for survival and competitive ability; therefore, embryos that developed at lower temperatures would have considerable competitive advantage (Elliott, 1989; Cutts et al., 1999; Johnsson et al., 1999; Einum and Fleming, 2000).

River flow metrics

High, moderate and low river flow and frequency contribute to different aspects of ecosystem function. For example, floods are important for the natural alteration of a river's channel, adjacent vegetation, meander, clearing interstitial spaces, scouring periphyton and the repositioning of pool/riffle/run sequences. Baseflows are also essential for providing refuge during periods of no rainfall.

Flow metrics are usually derived from continuous flow gauging data (usually measured as discharge or velocity) and often attempt to quantify the magnitude and frequency of key ecological events. For example, a common statistic is the number of times within a year that flow peaks above a defined multiple of the median flow. When interpreting changes in hydrological regimes, it is important to also consider climatic changes, such as the Inter-Decadal Pacific Oscillation, as climate can cause significant interannual variability that may not be registered by short term flow monitoring. These fluctuations can cause rivers in some years to experience low flows with little variability and high flows with high variability in other years. This results in drastic changes in community composition with species who have interannual life histories being most affected (Biggs et al., 2005, 2008).

Physical habitat

The influence of geomorphological structure in determining river ecosystem structure and function often forms the underlying framework for many hypotheses in stream ecology such as the river continuum concept (Vannote et al., 1980), network dynamics

Fig. 2 New Zealand Freshwater Crayfish (*Paranephrops planifrons*) (top) and New Zealand Banded Kokopu (*Galaxias fasciatus*) (middle) smothered in fine sediment. Bottom–stream substrate with interstitial spaces partly clogged with deposited sediment.

hypothesis (Benda et al., 2004) and the riverine ecosystem synthesis (Thorp, 2008; Thorp et al., 2006). Geomorphological structure includes meanders; sinuosity; the area of pools, riffles and runs; gravel composition; riparian vegetation; and stream bed movement. The size of meanders, pools, riffles and runs can be measured directly (usually using aerial photography) or acoustic doppler profilers. Gravel composition is measured directly, often by estimating proportions of stones in each Krumbein Phi class. The area, width, length, nature and species composition of riparian vegetation is often measured directly or from aerial photography. Stream bed movement is often measured by using tracer stones of different size classes and measuring their movement over a defined period. A river with rounded stones can indicate more mobile river bed than one with angular stones too. Metrics that quantify

physical habitat change often include some or all of these parameters. Some metrics are species-specific and only measure (or place more weighting on) parameters relevant to the species being managed. Species specific habitat metrics form the basis of many physical habitat models, such as PHABSIM. In Japan, DHABSIM (Diversity Habitat Simulation) is used to calculate the Eco-Environmental Diversity (EED) Index, which attempts to quantify the diversity of habitat within a defined home-range area. It was found that EED was positively correlated with fish species richness (Sekine, 2017).

Here we have introduced numerous indicator groups of freshwater ecosystem health. No single indicator exists to measure the entirety of ecosystem health, rather each have been developed to measure a particular aspect of ecosystem health. It is often unpragmatic to measure health using indicators from all groups; typically, managers would select measures based on environmental objectives and impacts. If one were to gauge ecosystem health in its entirety, then numerous metrics that indicate on ecosystem structure, function and stability (Table 1) would be required.

References

Allan JD (1984) Hypothesis testing in ecological studies of aquatic insects. In: Resh VH and Rosenberg DM (eds.) *The ecology of aquatic insects*. New York: Praeger Publishers.

Allesina S and Bondavalli C (2003) Steady state of ecosystem flow networks: A comparison between balancing procedures. *Ecological Modelling* 165(2–3): 221–229. https://doi.org/10.1016/S0304-3800(03)00075-9.

Allesina S and Ulanowicz RE (2004) Cycling in ecological networks: Finn's index revisited. *Computational Biology and Chemistry* 28(3): 227–233. https://doi.org/10.1016/j.compbiolchem.2004.04.002.

Anderson C and Cabana G (2006) Does δ15N in river food webs reflect the intensity and origin of N loads from the watershed? *Science of The Total Environment* 367(2): 968–978. https://doi.org/10.1016/j.scitotenv.2006.01.029.

ANZECC (2000) Australian and New Zealand guidelines for fresh and marine water quality. In: *Australian and New Zealand Environment and Conservation Council and Agriculture and Resource Management Council of Australia and New Zealand, Canberra*, 1–103.

Appling, A., Hall Jr, R., Arroita, M. & Yackulic, C. 2016. streamMetabolizer: models for estimating aquatic photosynthesis and respiration. R package version 0.9. 32.

Bailey RC, Norris RH, and Reynoldson TB (2004) *Bioassessment of freshwater ecosystems: Using the reference condition approach*. Boston: Kluwer Academic Publishers.

Baird D, Fath BD, Ulanowicz RE, Asmus H, and Asmus R (2009) On the consequences of aggregation and balancing of networks on system properties derived from ecological network analysis. *Ecological Modelling* 220(23): 3465–3471. https://doi.org/10.1016/j.ecolmodel.2009.09.008.

Beck MW and Hatch LK (2009) A review of research on the development of lake indices of biotic integrity. *Environmental Reviews* 17: 21–44. https://doi.org/10.1139/A09-001.

Beck MW, Hatch LK, Vondracek B, and Valley RD (2010) Development of a macrophyte-based index of biotic integrity for Minnesota lakes. *Ecological Indicators* 10(5): 968–979. https://doi.org/10.1016/j.ecolind.2010.02.006.

Beck MW, Vondracek B, and Hatch LK (2013) Environmental clustering of lakes to evaluate performance of a macrophyte index of biotic integrity. *Aquatic Botany* 108: 16–25. https://doi.org/10.1016/j.aquabot.2013.02.003.

Belgrano A, Scharler UM, Dunne JA, and Ulanowicz R (eds.) (2004) *Aquatic food webs: An ecosystem approach*. Oxford: Oxford University Press.

Benda L, Poff NL, Miller D, Dunne T, Reeves G, Pess G, and Pollock M (2004) The network dynamics hypothesis: How channel networks structure riverine habitats. *Bioscience* 54(5): 413–427.

Biggs B and Kilroy C (2000) Stream Periphyton Monitoring Manual *Christchurch*. National Institute of Water and Atmospheric Research: New Zealand.

Biggs BJ, Ibbitt RP, and Jowett IG (2008) Determination of flow regimes for protection of in-river values in New Zealand: An overview. *Ecohydrology & Hydrobiology* 8(1): 17–29.

Biggs BJF, Nikora VI, and Snelder TH (2005) Linking scales of flow variability to lotic ecosystem structure and function. *River Research and Applications* 21(2–3): 283–298. https://doi.org/10.1002/rra.847.

Boothroyd IKG and Stark JD (2000) Use of invertebrates in monitoring. In: Collier KJ and Winterbourn MJ (eds.) *New Zealand stream invertebrates: Ecology and implications for management*. New Zealand Limnological Society: Hamilton.

Borrett SR and Lau MK (2014) enaR: An R package for Ecosystem Network Analysis. *Methods in Ecology and Evolution* 5(11): 1206–1213. https://doi.org/10.1111/2041-210x.12282.

Boulton A and Boon P (1991) A review of methodology used to measure leaf litter decomposition in lotic environments: Time to turn over an old leaf? *Marine and Freshwater Research* 42(1): 1–43. https://doi.org/10.1071/MF9910001.

Boulton AJ, Scarsbrook MR, Quinn JM, and Burrell GP (1997) Land-use effects on the hyporheic ecology of five small streams near Hamilton, New Zealand. *New Zealand Journal of Marine and Freshwater Research* 31(5): 609–622.

Boyero L, Pearson RG, Gessner MO, Barmuta LA, Ferreira V, Graça MAS, et al. (2011) A global experiment suggests climate warming will not accelerate litter decomposition in streams but might reduce carbon sequestration. *Ecology Letters* 14(3): 289–294. https://doi.org/10.1111/j.1461-0248.2010.01578.x.

Bray JR and Curtis JT (1957) An ordination of the upland forest communities of Southern Wisconsin. *Ecological Monographs* 27(4): 325–349. https://doi.org/10.2307/1942268.

Brey T (2012) A multi-parameter artificial neural network model to estimate macrobenthic invertebrate productivity and production. *Limnology and Oceanography: Methods* 10: 581–589.

Brey T, Müller-Wiegmann C, Zittier ZM, and Hagen W (2010) Body composition in aquatic organisms—A global data bank of relationships between mass, elemental composition and energy content. *Journal of Sea Research* 64(3): 334–340.

Bricker SB, Longstaff B, Dennison W, Jones A, Boicourt K, Wicks C, and Woerner J (2008) Effects of nutrient enrichment in the nation's estuaries: A decade of change. *Harmful Algae* 8(1): 21–32. https://doi.org/10.1016/j.hal.2008.08.028.

Burcher CL and Benfield EF (2006) Physical and biological responses of streams to suburbanization of historically agricultural watersheds. *Journal of the North American Benthological Society* 25(2): 356–369.

Canning AD and Death RG (2017) Trophic cascade direction and flow determine network flow stability. *Ecological Modelling* 355: 18–23. https://doi.org/10.1016/j.ecolmodel.2017.03.020.

Cao Y, Bark AW, and Williams WP (1996) Measuring the responses of macroinvertebrate communities to water pollution: A comparison of multivariate approaches, biotic and diversity indices. *Hydrobiologia* 341(1): 1–19. https://doi.org/10.1007/BF00012298.

Cao Y and Epifanio J (2010) Quantifying the responses of macroinvertebrate assemblages to simulated stress: Are more accurate similarity indices less useful? *Methods in Ecology and Evolution* 1(4): 380–388. https://doi.org/10.1111/j.2041-210X.2010.00040.x.

Cao Y, Williams WP, and Bark AW (1997) Similarity measure bias in river benthic Aufwuchs community analysis. *Water Environment Research* 69(1): 95–106. https://doi.org/10.2175/106143097X125227.

Chessman BC (2003) New sensitivity grades for Australian river macroinvertebrates. *Marine and Freshwater Research* 54(2): 95–103. https://doi.org/10.1071/mf02114.

Chessman BC, Growns JE, and Kotlash AR (1997) Objective derivation of macroinvertebrate family sensitivity grade numbers for the SIGNAL biotic index: Application to the Hunter River system, New South Wales. *Marine and Freshwater Research* 48(2): 159–172. https://doi.org/10.1071/mf96058.

Clapcott J, Young R, Harding J, Matthaei C, Quinn J, and Death R (2011) *Sediment assessment methods: Protocols and guidelines for assessing the effects of deposited fine sediment on in-stream values*. Nelson: Cawthron Institute.

Clapcott JE, Young RG, Goodwin EO, and Leathwick JR (2010) Applied issues: Exploring the response of functional indicators of stream health to land-use gradients. *Freshwater Biology* 55(10): 2181–2199. https://doi.org/10.1111/j.1365-2427.2010.02463.x.

Clarke A and Johnston NM (1999) Scaling of metabolic rate with body mass and temperature in teleost fish. *Journal of Animal Ecology* 68(5): 893–905. https://doi.org/10.1046/j.1365-2656.1999.00337.x.

Clarke RT and Murphy JF (2006) Effects of locally rare taxa on the precision and sensitivity of RIVPACS bioassessment of freshwaters. *Freshwater Biology* 51(10): 1924–1940.

Collins AL, Naden PS, Sear DA, Jones JI, Foster IDL, and Morrow K (2011) Sediment targets for informing river catchment management: International experience and prospects. *Hydrological Processes* 25: 2112–2129. https://doi.org/10.1002/hyp.7965.

Co-operation, O. f. E. & Development (1982) Eutrophication of waters: Monitoring, assessment and control. In: *Organisation for Economic Co-operation and Development*. Washington, DC: OECD Publications and Information Center. Sold by.

Costanza R and Mageau M (1999) What is a healthy ecosystem? *Aquatic Ecology* 33(1): 105–115. https://doi.org/10.1023/a:1009930313242.

Cross W, Wallace J, Rosemond A, and Eggert S (2006) Whole-system nutrient enrichment increases secondary production in a detritus-based ecosystem. *Ecology* 87(6): 1556–1565.

Cucherousset J and Olden JD (2011) Ecological impacts of nonnative freshwater fishes. *Fisheries* 36(5): 215–230. https://doi.org/10.1080/03632415.2011.574578.

Cutts CJ, Metcalfe NB, and Taylor AC (1999) Competitive Asymmetries in Territorial Juvenile Atlantic Salmon, *Salmo salar*. *Oikos* 86(3): 479–486. https://doi.org/10.2307/3546652.

Dean TL and Richardson J (1999) Responses of seven species of native freshwater fish and a shrimp to low levels of dissolved oxygen. *New Zealand Journal of Marine and Freshwater Research* 33(1): 99–106. https://doi.org/10.1080/00288330.1999.9516860.

Death RG (2008) Effects of floods on aquatic invertebrate communities. In: Lancaster J and Briers RA (eds.) *Aquatic insects: Challenges to populations*. UK: CAB International.

Death RG and Zimmermann EM (2005) Interaction between disturbance and primary productivity in determining stream invertebrate diversity. *Oikos* 111(2): 392–402.

Dewson ZS, James ABW, and Death RG (2007) A review of the consequences of decreased flow for instream habitat and macroinvertebrates. *Journal of the North American Benthological Society* 26(3): 401–415. https://doi.org/10.1899/06-110.1.

Dodds WK (2007) Trophic state, eutrophication and nutrient criteria in streams. *Trends in Ecology & Evolution* 22(12): 669–676. https://doi.org/10.1016/j.tree.2007.07.010.

Dodds WK, Jones JR, and Welch EB (1998) Suggested classification of stream trophic state: Distributions of temperate stream types by chlorophyll, total nitrogen, and phosphorus. *Water Research* 32(5): 1455–1462.

Dudgeon D (2014) Threats to freshwater biodiversity in a changing world. In: Freedman B (ed.) *Global environmental change*. Dordrecht: Springer Netherlandshttps://doi.org/10.1007/978-94-007-5784-4_108.

Dudgeon D, Arthington AH, Gessner MO, Kawabata Z-I, Knowler DJ, Lévêque C, Naiman RJ, Prieur-Richard A-H, Soto D, Stiassny MLJ, and Sullivan CA (2006) Freshwater biodiversity: Importance, threats, status and conservation challenges. *Biological Reviews* 81(2): 163–182. https://doi.org/10.1017/S1464793105006950.

Duolos SK and Kindschi GA (1990) Effects of oxygen supersaturation on the culture of cutthroat trout, Oncorhynchus clarki Richardson, and rainbow trout, Oncorhynchus mykiss Richardson. *Aquaculture Research* 21: 39–46.

Egglishaw H (1972) An experimental study of the breakdown of cellulose in fast-flowing streams. *Memorie dell'Istituto Italiano di Idrobiologia* 29: 405–428.

Einum S and Fleming IA (2000) Selection against late emergence and small offspring in Atlantic Salmon (*Salmo salar*). *Evolution* 54(2): 628–639. https://doi.org/10.1554/0014-3820(2000)054[0628:SALEAS]2.0.CO;2.

Elliott JM (1976) The Energetics of Feeding, Metabolism and Growth of Brown Trout (*Salmo trutta* L.) in Relation to Body Weight, Water Temperature and Ration Size. *Journal of Animal Ecology* 45(3): 923–948. https://doi.org/10.2307/3590.

Elliott JM (1989) The critical-period concept for juvenile survival and its relevance for population regulation in young sea trout, *Salmo trutta*. *Journal of Fish Biology* 35: 91–98. https://doi.org/10.1111/j.1095-8649.1989.tb03049.x.

Elwood JW and Nelson DJ (1972) Periphyton production and grazing rates in a stream measured with a 32P material balance method. *Oikos* 23(3): 295–303. https://doi.org/10.2307/3543167.

Espmark ÅM, Hjelde K, and Baeverfjord G (2010) Development of gas bubble disease in juvenile Atlantic salmon exposed to water supersaturated with oxygen. *Aquaculture* 306: 198–204. https://doi.org/10.1016.

Fahey B and Marden M (2006) Forestry effects on sediment yield and erosion. In: Eyles G and Fahey B (eds.) *The pakuratahi land use study*. Napier: Hawkes Bay Regional Council.

Fath BD and Borrett SR (2006) A MATLAB® function for Network Environ Analysis. *Environmental Modelling & Software* 21(3): 375–405. https://doi.org/10.1016/j.envsoft.2004.11.007.

Fath BD and Halnes G (2007) Cyclic energy pathways in ecological food webs. *Ecological Modelling* 208(1): 17–24. https://doi.org/10.1016/j.ecolmodel.2007.04.020.

Fath BD and Patten BC (1999) Review of the foundations of Network Environ Analysis. *Ecosystems* 2(2): 167–179. https://doi.org/10.1007/s100219900067.

Fath BD, Patten BC, and Choi JS (2001) Complementarity of ecological goal functions. *Journal of Theoretical Biology* 208(4): 493–506. https://doi.org/10.1006/jtbi.2000.2234.

Fath BD, Scharler UM, Ulanowicz RE, and Hannon B (2007) Ecological network analysis: Network construction. *Ecological Modelling* 208(1): 49–55. https://doi.org/10.1016/j.ecolmodel.2007.04.029.

Fayram AH, Miller MA, and Colby AC (2005) Effects of Stream Order and Ecoregion on Variability in Coldwater Fish Index of Biotic Integrity Scores within Streams in Wisconsin. *Journal of Freshwater Ecology* 20(1): 17–25. https://doi.org/10.1080/02705060.2005.9664932.

Finn JT (1976) Measures of ecosystem structure and function derived from analysis of flows. *Journal of Theoretical Biology* 56(2): 363–380. https://doi.org/10.1016/S0022-5193(76)80080-X.

Foote KJ, Joy MK, and Death RG (2015) New Zealand dairy farming: Milking our environment for all it's worth. *Environmental Management* 56(3): 709–720. https://doi.org/10.1007/s00267-015-0517-x.

Forster ME (1981) Oxygen consumption and apnoea in the shortfin eel, *Anguilla australis* schmidtii. *New Zealand journal of marine and freshwater research* 15(1): 85–90. https://doi.org/10.1080/00288330.1981.9515900.

Franklin PA (2013) Dissolved oxygen criteria for freshwater fish in New Zealand: A revised approach. *New Zealand Journal of Marine and Freshwater Research* 48(1): 112–126. https://doi.org/10.1080/00288330.2013.827123.

Friberg N, Bonada N, Bradley DC, Dunbar MJ, Edwards FK, Grey J, Hayes RB, Hildrew AG, Lamouroux N, and Trimmer M (2011) Biomonitoring of human impacts in freshwater ecosystems: The good, the bad and the ugly. *Advances in Ecological Research* 44: 1–68.

Fudge TS, Wautier KG, Evans RE, and Palace VP (2008) Effect of different levels of fine-sediment loading on the escapement success of rainbow trout fry from artificial redds. *North American Journal of Fisheries Management* 28: 758–765. https://doi.org/10.1577/M07-084.1.

Geist DR, Linley TJ, Cullinan V, and Deng Z (2013) The effects of total dissolved gas on chum salmon fry survival, growth, gas bubble disease, and seawater tolerance. *North American Journal of Fisheries Management* 33: 200–215. https://doi.org/10.1080/02755947.2012.750634.

Gillooly JF, Brown JH, West GB, Savage VM, and Charnov EL (2001) Effects of size and temperature on metabolic rate. *Science* 293(5538): 2248–2251. https://doi.org/10.1126/science.1061967.

Hamilton AL (1969) On estimating annual production. *Limnology and Oceanography* 14(5): 771–781.

Hans KM, Mesa MG, and Maule AG (1999) Rate of disappearance of gas bubble trauma signs in juvenile salmonids. *Journal of Aquatic Animal Health* 11: 383–390.

Hart BD, Davies PE, Humphrey CL, Norris RH, Sudaryanti S, and Trihadiningrum Y (2001a) Application of the Australian river bioaassessment system (AUSRIVAS) in the Brabtas River, East Java, Indonesia. *Journal of Environmental Management* 62: 93–100.

Hart BT, Davies PE, Humphrey CL, Norris RN, Sudaryanti S, and Trihadiningrum Y (2001b) Application of the Australian river bioassessment system (AUSRIVAS) in the Brantas River, east java, Indonesia. *Journal of Environmental Management* 62(1): 93–100. https://doi.org/10.1006/jema.2001.0424.

Havens KE and Hanazato T (1993) Zooplankton community responses to chemical stressors: A comparison of results from acidification and pesticide contamination research. *Environmental Pollution* 82(3): 277–288. https://doi.org/10.1016/0269-7491(93)90130-G.

Hawkins CP, Murphy ML, and Anderson NH (1982) Effects of canopy, substrate composition, and gradient on the structure of macroinvertebrate communities in cascade range streams of oregon. *Ecology* 63(6): 1840–1856. https://doi.org/10.2307/1940125.

Herbst D, Bogan M, Roll S, and Safford H (2012) Effects of livestock exclusion on in-stream habitat and benthic invertebrate assemblages in montane streams. *Freshwater Biology* 57: 204–217. https://doi.org/10.1111/j.1365-2427.2011.02706.x.

Hildrew AG, Townsend CR, Francis J, and Finch K (1984) Cellulolytic decomposition in streams of contrasting pH and its relationship with invertebrate community structure. *Freshwater Biology* 14(3): 323–328. https://doi.org/10.1111/j.1365-2427.1984.tb00045.x.

Hilsenhoff WL (1987) An improved biotic index of organic stream pollution. *Great Lakes Entomologist* 20(1): 31–39.

Hines DE, Ray S, and Borrett SR (2018) Uncertainty analyses for Ecological Network Analysis enable stronger inferences. *Environmental Modelling & Software* 101: 117–127. https://doi.org/10.1016/j.envsoft.2017.12.011.

Hladyz S, Åbjörnsson K, Chauvet E, Dobson M, Elosegi A, Ferreira V, Fleituch T, Gessner MO, Giller PS, and Gulis V (2011) Stream ecosystem functioning in an agricultural landscape: The importance of terrestrial-aquatic linkages. *Advances in Ecological Research* 44(3): 211–276.

Johnsson JI, Nöbbelin F, and Bohlin T (1999) Territorial competition among wild brown trout fry: effects of ownership and body size. *Journal of Fish Biology* 54(2): 469–472. https://doi.org/10.1111/j.1095-8649.1999.tb00846.x.

Jonsson B and Jonsson N (2009) A review of the likely effects of climate change on anadromous Atlantic salmon Salmo salar and brown trout Salmo trutta, with particular reference to water temperature and flow. *Journal of Fish Biology* 75(10): 2381–2447.

Jørgensen SE and Mejer H (1977) Ecological buffer capacity. *Ecological Modelling* 3(1): 39–61. https://doi.org/10.1016/0304-3800(77)90023-0.

Jørgensen SE and Nielsen SN (2014) Use of eco-exergy in ecological networks. *Ecological Modelling* 293: 202–209. https://doi.org/10.1016/j.ecolmodel.2014.05.007.

Jørgensen SE, Xu F-L, and Costanza R (2010) *Handbook of ecological indicators for assessment of ecosystem health*. CRC press.

Joy MK and Death RG (2000) Development and application of a predictive model of riverine fish community assemblages in the Taranaki region, of the North Island, New Zealand. *New Zealand Journal of Marine and Freshwater Research* 34: 243–254.

Joy MK and Death RG (2003) Biological assessment of rivers in the Manawatu-Wanganui region of New Zealand using a predictive macroinvertebrate model. *New Zealand Journal of Marine and Freshwater Research* 37: 367–379.

Joy MK and Death RG (2004a) Application of the Index of Biotic Integrity Methodology to New Zealand Freshwater Fish Communities. *Environmental management* 34(3): 415–428. https://doi.org/10.1007/s00267-004-0083-0.

Joy MK and Death RG (2004b) Predictive modelling and spatial mapping of freshwater fish and decapod assemblages using GIS and neural networks. *Freshwater Biology* 49: 1036–1052.

Kamdem Toham A and Teugels GG (1999) First data on an Index of Biotic Integrity (IBI) based on fish assemblages for the assessment of the impact of deforestation in a tropical West African river system. *Hydrobiologia* 397: 29–38. https://doi.org/10.1023/a:1003605801875.

Kamler E (1992) *Early life history of fish: An energetics approach*. Springer Science & Business Media.

Kane DD, Gordon SI, Munawar M, Charlton MN, and Culver DA (2009) The planktonic index of biotic integrity (P-IBI): An approach for assessing lake ecosystem health. *Ecological Indicators* 9(6): 1234–1247. https://doi.org/10.1016/j.ecolind.2009.03.014.

Karr JR, Fausch KD, Angermeier PL, Yant PR, and Schlosser IJ (1986) *Assessing biological integrity in running waters*. A method and its rationale. Illinois Natural History Survey, Champaign, Special Publication, 5.

Kemp P, Sear D, Collins A, Naden P, and Jones I (2011) The impacts of fine sediment on riverine fish. *Hydrological Processes* 25: 1800–1821. https://doi.org/10.1002/hyp.7940.

Kundzewicz ZW, Mata LJ, Arnell NW, Doll P, Kabat P, Jimenez B, Miller K, Oki T, Zekai S, and Shiklomanov I (2007) Freshwater resources and their management. In: Parry ML, Canziani OF, Palutikof JP, van der Linden PJ, and Hanson CE (eds.) *Climate change 2007: Impacts, adaptation and vulnerability. Contribution of Working Group II to the Fourth assessment report of the Intergovernmental Panel on Climate Change*, pp. 173–210. Cambridge University Press. ISBN: 9780521880091.

Latham Ii LG (2006) Network flow analysis algorithms. *Ecological Modelling* 192(3–4): 586–600. https://doi.org/10.1016/j.ecolmodel.2005.07.029.

Layman CA, Arrington DA, Montaña CG, and Post DM (2007) Can stable isotope ratios provide for community-wide measures of trophic structure? *Ecology* 88(1): 42–48.

Lunde KB and Resh VH (2012) Development and validation of a macroinvertebrate index of biotic integrity (IBI) for assessing urban impacts to Northern California freshwater wetlands. *Environmental Monitoring and Assessment* 184(6): 3653–3674. https://doi.org/10.1007/s10661-011-2214-4.

Lyons J, Gutiérrez-Hernández A, Díaz-Pardo E, Soto-Galera E, Medina-Nava M, and Pineda-López R (2000a) Development of a preliminary index of biotic integrity (IBI) based on fish assemblages to assess ecosystem condition in the lakes of central Mexico. *Hydrobiologia* 418(1): 57–72. https://doi.org/10.1023/a:1003888032756.

Lyons J, Weigel BM, Paine LK, and Undersander DJ (2000b) Influence of intensive rotational grazing on bank erosion, fish habitat quality, and fish communities in southwestern Wisconsin trout streams. *Journal of Soil and Water Conservation* 55(3): 271–276.

Marques JC, Pardal MÂ, Nielsen SN, and Jørgensen SE (1997) Analysis of the properties of exergy and biodiversity along an estuarine gradient of eutrophication. *Ecological Modelling* 102(1): 155–167. https://doi.org/10.1016/S0304-3800(97)00099-9.

Masese FO, Raburu PO, and Muchiri M (2009) A preliminary benthic macroinvertebrate index of biotic integrity (B-IBI) for monitoring the Moiben River, Lake Victoria Basin, Kenya. *African Journal of Aquatic Science* 34(1): 1–14. https://doi.org/10.2989/AJAS.2009.34.1.1.726.

Maulood BK, Alobaidy AHMJ, Alsaboonchi A, Abid HS, and Alobaidy GS (2011) Phytoplankton index of biological integrity (P-IBI) in several marshes, southern Iraq. *Journal of Environmental Protection* 2(04): 387.

McEwan AJ (2009) *Fine scale spatial behaviour of indigenous riverine fish in a small New Zealand stream*. MSc, Massey University.

McIntire CD and Phinney HK (1965) Laboratory studies of periphyton production and community metabolism in lotic environments. *Ecological Monographs* 35(3): 238–258. https://doi.org/10.2307/1942138.

Mesa MG and Warren JJ (1997) Predator avoidance ability of juvenile chinook salmon (Oncorhynchus tshawytscha) subjected to sublethal exposures of gas-supersaturated water. *Canadian Journal of Fisheries and Aquatic Sciences* 54: 757–764. https://doi.org/10.1139/f96-326.

Mesa MG, Weiland LK, and Maule AG (2000) Progression and severity of gas bubble trauma in juvenile salmonids. *Transactions of the American Fisheries Society* 129: 174–185.

Miehls ALJ, Mason DM, Frank KA, Krause AE, Peacor SD, and Taylor WW (2009) Invasive species impacts on ecosystem structure and function: A comparison of the bay of Quinte, Canada, and Oneida Lake, USA, before and after zebra mussel invasion. *Ecological Modelling* 220(22): 3182–3193.

Minns CK, Cairns VW, Randall RG, and Moore JE (1994) An Index of Biotic Integrity (IBI) for Fish Assemblages in the Littoral Zone of Great Lakes' Areas of Concern. *Canadian Journal of Fisheries and Aquatic Sciences* 51(8): 1804–1822. https://doi.org/10.1139/f94-183.

Morin A, Mousseau TA, and Roff DA (1987) Accuracy and precision of secondary production estimates1. *Limnology and Oceanography* 32(6): 1342–1352. https://doi.org/10.4319/lo.1987.32.6.1342.

Mougi A and Kondoh M (2016) Food-web complexity, meta-community complexity and community stability. *Scientific Reports* 6(24478). https://doi.org/10.1038/srep24478http://www.nature.com/articles/srep24478#supplementary-information.

Mutch R, Steedman R, Berte S, and Pritchard G (1983) Leaf breakdown in a mountain stream: A comparison of methods. *Archiv fur hydrobiology* 97: 89–108.

Nielsen JM, Clare EL, Hayden B, Brett MT, and Kratina P (2018) Diet tracing in ecology: Method comparison and selection. *Methods in Ecology and Evolution* 9(2): 278–291. https://doi.org/10.1111/2041-210X.12869.

Norris RH and Thoms MC (1999) What is river health? *Freshwater Biology* 41(2): 197–209. https://doi.org/10.1046/j.1365-2427.1999.00425.x.

O'Brien A, Townsend K, Hale R, Sharley D, and Pettigrove V (2016) How is ecosystem health defined and measured? A critical review of freshwater and estuarine studies. *Ecological Indicators* 69: 722–729. https://doi.org/10.1016/j.ecolind.2016.05.004.

Odum EP (1953) *Fundamentals of ecology*. Saunders Company: Philadelphia, W. B.
Ojanguren A and Brana F (2003) Thermal dependence of embryonic growth and development in brown trout. *Journal of Fish Biology* 62(3): 580–590.
Palmer MA and Febria CM (2012) The heartbeat of ecosystems. *Science* 336(6087): 1393.
Pandian TJ and Marian MP (1986) An indirect procedure for the estimation of assimilation efficiency of aquatic insects. *Freshwater Biology* 16(1): 93–98. https://doi.org/10.1111/j.1365-2427.1986.tb00950.x.
Patten BC (1995) Network integration of ecological extremal principles: Exergy, emergy, power, ascendency, and indirect effects. *Ecological Modelling* 79(1–3): 75–84. https://doi.org/10.1016/0304-3800(94)00037-I.
Perera R, Wattavidanage J, and Nilakarawasam N (2012) Development of a macroinvertebrate-based index of biotic integrity (M-IBI) for Colombo-Sri Jayawardhanapura canal system (a new approach to assess stream/wetland health). *Journal of Tropical Forestry and Environment* 2(1).
Pool TK and Olden JD (2015) Assessing long-term fish responses and short-term solutions to flow regulation in a dryland river basin. *Ecology of Freshwater Fish* 24(1): 56–66. https://doi.org/10.1111/eff.12125.
Raburu PO, Masese FO, and Mulanda CA (2009) Macroinvertebrate Index of Biotic Integrity (M-IBI) for monitoring rivers in the upper catchment of Lake Victoria Basin, Kenya. *Aquatic Ecosystem Health & Management* 12(2): 197–205. https://doi.org/10.1080/14634980902907763.
Rapport DJ, Gaudet CL, Constanza R, Epstein P, and Levins R (2009) *Ecosystem health: Principles and practice*. John Wiley & Sons.
Relyea RA (2005) The impact of insecticides and herbicide on the biodiversity and productivity of aquatic communities. *Ecological Applications* 15(2): 618–627. https://doi.org/10.1890/03-5342.
Resh VH, Brown AV, Covich AP, Gurtz ME, Li HW, Minshall GW, Reice SR, Sheldon AL, Wallace JB, and Wissmar RC (1988) The Role of Disturbance in Stream Ecology. *Journal of the North American Benthological Society* 7(4): 433–455. https://doi.org/10.2307/1467300.
Reynoldson TB and Metcalfe-Smith JL (1992) An overview of the assessment of aquatic ecosystem health using benthic invertebrates. *Journal of Aquatic Ecosystem Health* 1(4): 295–308.
Reynoldson TB, Norris RH, Resh VH, Day KE, and Rosenberg DM (1997) The reference condition: A comparison of multimetric and multivariate appraoches to assess water-quality impairment using benthic macroinvertebrates. *Journal of the North American Benthological Society* 16: 833–852.
Reynoldson TB, Rosenberg DM, and Resh VH (2001) Comparison of models predicting invertebrate assemblages for biomonitoring in the Fraser River catchment, British Columbia. *Canadian Journal of Fisheries and Aquatic Sciences* 58(7): 1395–1410.
Robertson A (1979) The relationship between annual production: Biomass ratios and lifespans for marine macrobenthos. *Oecologia* 38(2): 193–202. https://doi.org/10.1007/bf00346563.
Robinson WR, Peters RH, and Zimmermann J (1983) The effects of body size and temperature on metabolic rate of organisms. *Canadian Journal of Zoology* 61(2): 281–288. https://doi.org/10.1139/z83-037.
Rosenberg DM and Resh VH (eds.) (1993) *Freshwater biomonitoring and benthic macroinvertebrates*. New York: Chapman & Hall.
Ruaro R and Gubiani ÉA (2013) A scientometric assessment of 30 years of the index of biotic integrity in aquatic ecosystems: Applications and main flaws. *Ecological Indicators* 29: 105–110. https://doi.org/10.1016/j.ecolind.2012.12.016.
Saint-Béat B, Baird D, Asmus H, Asmus R, Bacher C, Pacella SR, Johnson GA, David V, Vézina AF, and Niquil N (2015) Trophic networks: How do theories link ecosystem structure and functioning to stability properties? A review. *Ecological Indicators* 52: 458–471. https://doi.org/10.1016/j.ecolind.2014.12.017.
Salas AK and Borrett SR (2011) Evidence for the dominance of indirect effects in 50 trophic ecosystem networks. *Ecological Modelling* 222(5): 1192–1204. https://doi.org/10.1016/j.ecolmodel.2010.12.002.
Sartorius N (2006) The meanings of health and its promotion. *Croatian Medical Journal* 47(4): 662–664.
Schaeffer DJ, Herricks EE, and Kerster HW (1988) Ecosystem health: I. Measuring ecosystem health. *Environmental Management* 12(4): 445–455. https://doi.org/10.1007/bf01873258.
Scheurer K, Alewell C, Banninger D, and Burkhardt-Holm P (2009) Climate and land-use changes affecting river sediment and brown trout in alpine countries-a review. *Environmental Science and Pollution Research* 16(2): 232–242. https://doi.org/10.1007/s11356-008-0075-3.
Schwendel AC, Death RG, Fuller IC, and Joy MK (2011) Linking disturbance and stream invertebrate communities: How best to measure bed stability. *Journal of the North American Benthological Society* 30(1): 11–24. https://doi.org/10.1899/09-172.1.
Sekine M (2017) *DHABSIM Solver Manual*. Japan, iRIC Software.
Shrimpton JM, Randall DJ, and Fidler LE (1990) Assessing the effects of positive buoyancy on rainbow trout (Oncorhynchus mykiss) held in gas supersaturated water. *Canadian Journal of Zoology* 68: 969–973. https://doi.org/10.1139/z90-139.
Silow EA and Mokry AV (2010) Exergy as a tool for ecosystem health assessment. *Entropy* 12(4): 902.
Simpson JC and Norris RH (2000) *Biological assessment of river quality: Development of AUSRIVAS models and outputs*. Freshwater Biological Association (FBA): Ambleside.
Smith VH and Schindler DW (2009) Eutrophication science: Where do we go from here? *Trends in Ecology & Evolution* 24(4): 201–207. https://doi.org/10.1016/j.tree.2008.11.009.
Soetaert K and Kones J (2008) *NetIndices: Estimating network indices, including trophic structure of foodwebs in R*. R package version.
Sponseller RA and Benfield EF (2001) Influences of land use on leaf breakdown in southern Appalachian headwater streams: A multiple-scale analysis. *Journal of the North American Benthological Society* 20: 44–59. https://doi.org/10.2307/1468187.
Stanford JA and Ward JV (1988) The Hyporheic habitat of river ecosystems. *Nature* 335: 64–66.
Steedman RJ (1994) Ecosystem health as a management goal. *Journal of the North American Benthological Society* 13(4): 605–610. https://doi.org/10.2307/1467856.
Sternecker K and Geist J (2010) The effects of stream substratum composition on the emergence of salmonid fry. *Ecology of Freshwater Fish* 19: 537–544. https://doi.org/10.1111/j.1600-0633.2010.00432.x.
Stevenson RJ and Pan Y (1999) Assessing environmental conditions in rivers and streams with diatoms. *The Diatoms: Applications for the Environmental and Earth Sciences* 1(4).
Tank JL and Winterbourn MJ (1996) Microbial activity and invertebrate colonisation of wood in a New Zealand forest stream. *New Zealand Journal of Marine and Freshwater Research* 30(2): 271–280. https://doi.org/10.1080/00288330.1996.9516714.
Thorp JH (2008) The riverine ecosystem synthesis: toward conceptual cohesiveness in river science, *Amsterdam*. Boston: Academic Press.
Thorp JH, Thoms MC, and Delong MD (2006) The riverine ecosystem synthesis: Biocomplexity in river networks across space and time. *River Research and Applications* 22(2): 123–147.
Townsend C, DolÉDec S, and Scarsbrook M (1997a) Species traits in relation to temporal and spatial heterogeneity in streams: a test of habitat templet theory. *Freshwater Biology* 37(2): 367–387. https://doi.org/10.1046/j.1365-2427.1997.00166.x.
Townsend CR, Scarsbrook MR, and Dolédec S (1997b) The intermediate disturbance hypothesis, refugia, and biodiversity in streams. *Limnology and Oceanography* 42(5): 938–949.
Ulanowicz RE (1997) *Ecology*. The Ascendent Perspective: Columbia University Press.
Ulanowicz RE (2004) Quantitative methods for ecological network analysis. *Computational Biology and Chemistry* 28(5): 321–339. https://doi.org/10.1016/j.compbiolchem.2004.09.001.
Ulanowicz RE (2009) The dual nature of ecosystem dynamics. *Ecological Modelling* 220(16): 1886–1892.
Ulanowicz RE, Goerner SJ, Lietaer B, and Gomez R (2009) Quantifying sustainability: Resilience, efficiency and the return of information theory. *Ecological Complexity* 6(1): 27–36. https://doi.org/10.1016/j.ecocom.2008.10.005.
van den Heuvel MR, Hogan NS, Roloson SD, and Van Der Kraak GJ (2012) Reproductive development of yellow perch (*Perca flavescens*) exposed to oil sands–affected waters. *Environmental Toxicology and Chemistry* 31(3): 654–662. https://doi.org/10.1002/etc.1732.

Van Sickle J, Hawkins CP, Larsen DP, and Herlihy AT (2005) A null model for the expected macroinvertebrate assemblage in streams. *Journal of the North American Benthological Society* 24(1): 178–191.
Vanni MJ (2002) Nutrient cycling by animals in freshwater ecosystems. *Annual Review of Ecology and Systematics:* 341–370.
Vannote RL, Minshall GW, Cummins KW, Sedell JR, and Cushing CE (1980) The river continuum concept. *Canadian Journal of Fisheries and Aquatic Sciences* 37: 130–137.
Ward J and Stanford J (1983) Intermediate-disturbance hypothesis: An explanation for biotic diversity patterns in lotic ecosystems. *Dynamics of Lotic Systems, Ann Arbor Science, Ann Arbor MI* 347–356. 2 fig, 35 ref.
Weigel BM, Henne LJ, and Martínez-Rivera LM (2002) Macroinvertebrate-based index of biotic integrity for protection of streams in west-central Mexico. *Journal of the North American Benthological Society* 21(4): 686–700. https://doi.org/10.2307/1468439.
Welch EB, Jacoby JM, Horner RR, and Seeley MR (1988) Nuisance biomass levels of periphytic algae in streams. *Hydrobiologia* 157(2): 161–168. https://doi.org/10.1007/bf00006968.
Whitfield AK and Harrison TD (2014) Fishes as indicators of estuarine health. In: *Reference Module in Earth Systems and Environmental Sciences*. Elsevierhttps://doi.org/10.1016/B978-0-12-409548-9.09062-X.
Wildish DJ and Peer D (1981) Methods for estimating secondary production in marine Amphipoda. *Canadian Journal of Fisheries and Aquatic Sciences* 38(9): 1019–1026. https://doi.org/10.1139/f81-140.
Williams DD and Hynes HBN (1974) The occurrence of benthos deep in the substratum of a stream. *Freshwater Biology* 4: 233–256.
Woodward G, Perkins DM, and Brown LE (2010) Climate change and freshwater ecosystems: Impacts across multiple levels of organization. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365(1549): 2093–2106. https://doi.org/10.1098/rstb.2010.0055.
Wootton JT (1994) The nature and consequences of indirect effects in ecological communities. *Annual Review of Ecology and Systematics* 25: 443–466.
Wootton JT (1998) Effects of Disturbance on Species Diversity: A Multitrophic Perspective. *The American Naturalist* 152(6): 803–825. https://doi.org/10.1086/286210.
Wren D, Barkdoll B, Kuhnle R, and Derrow R (2000) Field techniques for suspended-sediment measurement. *Journal of Hydraulic Engineering* 126(2): 97–104.
Wright JF, Sutcliffe DW, and Furse MT (2000) Assessing the biological quality of fresh waters: RIVPACS and other techniques. In: *Assessing the biological quality of fresh waters: RIVPACS and other techniques*. Freshwater Biological: Association.
Wu N, Schmalz B, and Fohrer N (2012) Development and testing of a phytoplankton index of biotic integrity (P-IBI) for a German lowland river. *Ecological Indicators* 13(1): 158–167. https://doi.org/10.1016/j.ecolind.2011.05.022.
Ye C, Xu Q, Kong H, Shen Z, and Yan C (2007) Eutrophication conditions and ecological status in typical bays of Lake Taihu in China. *Environmental Monitoring and Assessment* 135(1): 217–225. https://doi.org/10.1007/s10661-007-9644-z.
Young RG, Matthaei CD, and Townsend CR (2008) Organic matter breakdown and ecosystem metabolism: functional indicators for assessing river ecosystem health. *Journal of the North American Benthological Society* 27(3): 605–625. https://doi.org/10.1899/07-121.1.
Yount JD and Niemi GJ (1990) Recovery of lotic communities and ecosystems from disturbance—A narrative review of case studies. *Environmental Management* 14(5): 547–569.
Zhao L, Zhang H, O'Gorman EJ, Tian W, Ma A, Moore JC, Borrett SR, and Woodward G (2016) Weighting and indirect effects identify keystone species in food webs. *Ecology Letters* 19(9): 1032–1040. https://doi.org/10.1111/ele.12638.
Zhao L, Zhang H, Tian W, Li R, and Xu X (2017) Viewing the effects of species loss in complex ecological networks. *Mathematical Biosciences* 285: 55–60. https://doi.org/10.1016/j.mbs.2016.12.006.
Zhu D and Chang J (2008) Annual variations of biotic integrity in the upper Yangtze River using an adapted index of biotic integrity (IBI). *Ecological Indicators* 8(5): 564–572. https://doi.org/10.1016/j.ecolind.2007.07.004.
Zorach AC and Ulanowicz RE (2003) Quantifying the complexity of flow networks: How many roles are there? *Complexity* 8(3): 68–76. https://doi.org/10.1002/cplx.10075.

Further Reading

Leontief WW (1936) Quantitative input and output relations in the economic systems of the United States. *The Review of Economic Statistics* 105–125.

Equilibrium Concept in Phytoplankton Communities

A Basset, Scienze e Tecnologie Biologiche ed Ambientali – Lecce, Lecce, Italy
GC Carrada, Universitá degli Studi di Napoli Federico 11, Napoli, Italy
M Fedele and L Sabetta, Universitá del Salento, Lecce, Italy

Background

In various fields of science `equilibrium' refers to a balance between opposing forces: positive and negative values in an equation, reactants and products in a chemical reaction, production and respiration in an ecosystem, and supply and demand in a market. An equilibrium will be stable or unstable, steady state or dynamic, depending on both the stability of the interacting forces and the control and retro-control mechanisms of the interactions. Since forces are spatially and temporally defined, equilibrium is also spatially and temporally defined, or scale dependent.

The concept of equilibrium was introduced in those fields that are now known as community and ecosystem ecology in the late nineteenth/early twentieth centuries by Stephen A. Forbes and F. E. Clements. These authors had defined ecological communities as organisms interacting with each other (i.e., predator–prey, competitor–competitor, etc.), and had described community evolution toward an ordered stage or condition, which was formalized by Clements through the climax concept. A climax was defined as a community equilibrium stage that is persistent, self-sustaining, and at which further `change' is limited if at all possible. It has a mass balance component that is described by the thermodynamic equilibrium between gross production and respiration at the ecosystem level, and a food web stability component, which is described by the interacting population equilibrium between gains and losses of individuals at the community level. Stable or dynamic equilibrium conditions correspond to the assumption of perfect fitness equivalence among coexisting species in a certain time interval.

Numerical (or biomass) abundance and taxonomic composition are two major components of every guild and community, which, taking into account average individual body size and metabolic rates, give information on the overall energy and matter flow and on its apportionment among species. However, since the early definitions of climax and community persistency, community ecology has mainly used the term equilibrium to refer to a situation characterized by little variability in species composition over time or by the indefinite coexistence of a single species in a mixture.

Focusing on primary producers in terrestrial ecosystems, where producers are large and long-living trees, the temporal scale of the equilibrium conditions can be measured in terms of hundreds of years: the primary succession established on the Lake Michigan sand dunes spans over 11 000 years. In contrast, in aquatic and deep-water ecosystems, where producers are microscopic phytoplankton species, dynamic equilibrium conditions can be described on temporal scales of days, weeks, or months.

Phytoplankton are phototrophs, able to reproduce and build up populations utilizing sources of CO_2, inorganic nitrogen, sulfur and phosphorous compounds, and a number of other elements (Na, K, Mg, Ca, Si, Fe, Mn, B, Cl, Cu, Zn, Mo, Co, and V), most of which are required in small concentrations and not all of which are known to be required by all groups. In addition, several phytoplankton species are known to require vitamins, namely, thiamine, the cobalamines, and biotin.

When light is available, the processes of absorbing light and nutrients to build up biomass and reproduce are carried out at a very high rate by the small phytoplankton cells, measuring between 0.5 μm and *c*. 200 μm as linear dimension. In oligotrophic conditions, ranging from atoll lagoons to coastal marine and open ocean environments, phytoplankton biomass has an average turnover rate of 1.4–2.0 d^{-1}; in eutrophic conditions phytoplankton biomass has a much higher turnover rate, up to 10 times per day in shallow, enclosed, brackish ecosystems (e.g., Mediterranean lagoon ecosystems – such as Lake Alimini Grande, Puglia, Italy).

The high turnover rate of phytoplankton cells has different implications for the equilibrium concept of phytoplankton communities depending on whether it is applied to numerical and biomass abundance or to species interaction and community taxonomic composition. As regards numerical and/or biomass densities, a high turnover rate confers high resilience on phytoplankton communities with respect to spatial and temporal variation of those forces determining phytoplankton production, respiration, and predatory loss. Spatial and temporal patterns of numerical and/or biomass densities in phytoplankton communities are commonly observed at the ecosystem level in lake, lagoon, and marine ecosystems as a result of the equilibrium between phytoplankton requirements and both abiotic and biotic conditions. As regards taxonomic richness and species composition, a well-known tradeoff occurs between turnover rates and population stability, which also holds for phytoplankton communities. The dynamics of phytoplankton populations and the coexistence of a number of species on a limited amount of inorganic and organic resources distributed in a relatively isotropic and unstructured environment has challenged phytoplankton community ecologists over the last 50 years, representing a classic paradox of ecology: that is, the plankton paradox.

The dichotomy between (1) the balance between phytoplankton (numbers and biomass) and limiting factors and (2) the apparent lack of competitive equilibria among species has long been the subject of debate on the equilibrium concept in phytoplankton communities. The critical aspect of phytoplankton communities, which is difficult to explain in the context of the

equilibrium concept, is their taxonomic richness, the number of species being much higher than the number of factors limiting phytoplankton numerical and biomass densities.

Numerical and Biomass Equilibrium in Phytoplankton Guilds (Patterns and Limiting Factors)

Common patterns of numerical and/or biomass phytoplankton variations in space and time reveal the balance between nutrient and light availability and phytoplankton requirements and losses. Numerical and biomass equilibrium in phytoplankton communities is achieved (and can be observed at different scales) as a result of the phytoplankton's high potential resilience, enabling local communities to attenuate unbalanced conditions, and to adapt to fluctuations of limiting factors, which occur on a range of scales from hours to centuries, as well as from meters to hundred of kilometres.

Large-scale patterns of phytoplankton numerical and biomass variation have received much attention. However, evidence from coastal marine ecosystems shows that phytoplankton species manage to balance environmental conditions even while settling along the water column, resulting in deterministic phytoplankton size structures which are perturbed when storm events disrupt water column stability, determining mixing conditions. Since scales are measures of how organisms perceive their environment, and what is small and fast for an organism may be large and slow for another one, variance cascading of ecological processes, from large and slow to small and fast, is a point that deserves the utmost attention.

Large-Scale Patterns of Phytoplankton Numerical and Biomass Variation

Phytoplankton biomass in relatively large bodies of water, for example, coastal marine ecosystems and large and/or deep lakes, shows highly auto-correlated spatial and temporal patterns (Fig. 1), revealing a balance between nutrient and light availability and phytoplankton biomass; descriptions of spatial and temporal patterns and manipulative enrichment experiments have shown that when light conditions allow phytoplankton growth, natural waters are an environment of striking nutrient deficiency, so that competition is likely to be extremely severe. Therefore, a dynamic equilibrium characterizes the auto-correlated phytoplankton biomass patterns, where nutrient and light availability, phytoplankton absorption and reproduction, and zooplankton grazing represent the contrasting forces whose balance varies predictably in space and time at certain spatial and temporal scales.

As regards space, in large aquatic ecosystems (i.e., lakes and sea), three main spatial scales can be detected: one, up to roughly 1 km, which is dominated by horizontal turbulent diffusion; the second, between 1 and 5 km, at which the effect of turbulent diffusion is overridden by growth, grazing, and vertical mixing; the third, at scales larger than 5 km and up to 100 km, at which phytoplankton distribution is controlled mainly by advection, eddies, and upwelling. Diffusion processes imply turbulent diffusion caused by a wide range of movements in three dimensions, whereas advective processes are related to large-scale horizontal and vertical movements of water (and the suspended organisms).

Spatial patterns can be detected as chlorophyll density variation by means of satellite images. Strong auto-correlated spatial patterns occur along transects from the coastline to open waters (Figs. 1a (Adriatic Sea–Mediterranean Sea) and 1b (Lake Michigan – North America), across fronts and up-welling zones, even at the largest spatial scale of the oceanic transects (e.g., the Atlantic transect, Fig. 2), across coast-open sea gradients, eddies, and temperature gradients. In large and deep lakes of North America, 60%–80% of the phytoplankton spatial variation can be accounted for in terms of depth and light, temperature, nutrients and zooplankton density.

As regards time, seasonal variation dominates temporal patterns of phytoplankton densities, particularly in large and deep lakes and marine ecosystems. In temperate lakes, phytoplankton density has two main pulses of increase, in early spring and late summer (Fig. 1c). Early spring pulses are generally dominated by diatoms, which benefit from the availability of silicates, while late summer pulses are generally dominated by dynoflagellates. Larger-scale patterns, following nutrient enrichment from the catchment area of the waters, are also commonly observed in lakes and relatively enclosed seas.

Both temporal and spatial patterns of numerical and biomass densities in phytoplankton communities are related to directional and predictable variations of light limitation, nutrient limitation, and zooplankton grazing. Light and nutrient limitations have independent effects on phytoplankton communities but they also have a synergic effect mediated by water column stability. These relationships will be described in the following sections.

Body size of phytoplankton cells has a deterministic influence on light and nutrient limitations as perceived by phytoplankton cells and on the risk of predation. Light absorption (A) for a spherical cell is directly proportional to pigment-specific absorption cross section (a^*; m^2 mg pigment^{-1}), intracellular pigment concentration (c_i), and cell volume (V):

$$A = a^* c_i V \quad [1]$$

Nutrient requirement U depends on cell mass (BS):

$$U = b\mathrm{BS}^{0.75} \quad [2]$$

where b represents the nutrient uptake of a unitary cell mass. Nutrient uptake per unit of cell mass or volume (V) depends, in turn, on nutrient diffusion toward the cell surface:

$$U/V = 3D\Delta C r^{-2} \quad [3]$$

where r is the cell radius, D is the substrate diffusion coefficient, and ΔC the nutrient concentration gradient from the cell surface to the bulk media. As a consequence of the body size dependency of light absorption, nutrient uptake, and diffusion, the three

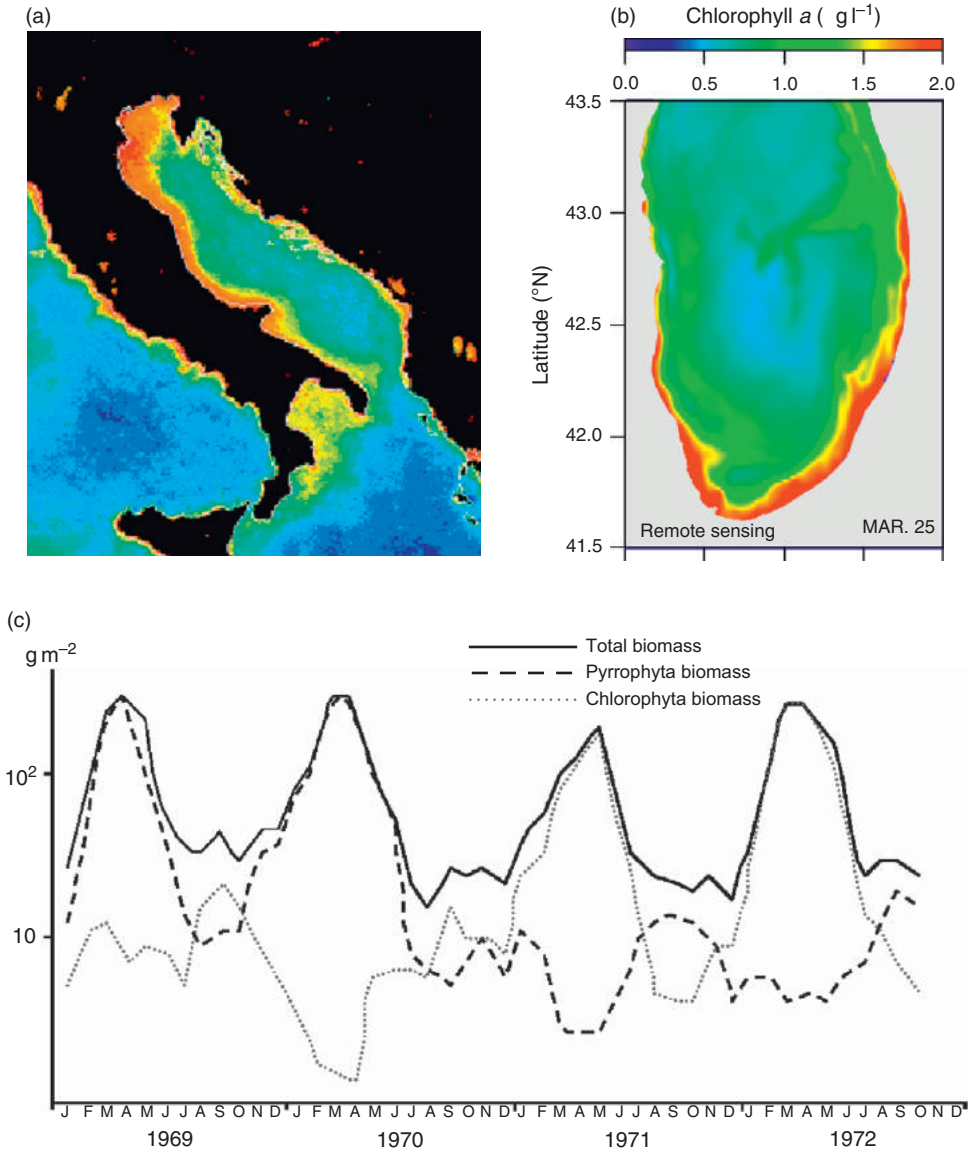


Fig. 1 Auto-correlated spatial and temporal patterns of phytoplankton biomass in natural waters. (a) Spatial distribution of Chl *a* in the Adriatic coastal marine areas – Mediterranean Sea. (b) Spatial distribution of Chl *a* in the Lake Michigan – North America. (c) Seasonal variations in phytoplankton fresh weight biomass 1969–72 (g m^{-1}) in trophogenic layer of Lake Kinneret, Israel. (b) Reproduced from website of Marine Ecosystem Dynamics Modeling Laboratory at School of Marine Science and Technology, University of Massachusetts – Dartmouth (http://fvcom.smast.umassd.edu/, with permission. (c) Reproduced from Berman, T., Pollingher, U., 1974. Annual and seasonal variations of phytoplankton, chlorophyll, and photosynthesis in Lake Kinneret. Limnology and Oceanography 19, 31–54, with permission; copyright (2008) by the American Society of Limnology and oceanography, Inc.

phytoplankton size fractions (i.e., pico-phytoplankton (cell diameter <2 μm), nano-phytoplankton (cell diameter between 2 and 20 μm), and micro-phytoplankton (cell diameter >20 μm)), together with the phytoplankton size spectra, show deterministic spatial patterns which are relevant to understanding phytoplankton community organization, although they are not treated specifically in this section.

Light Limitation of Phytoplankton Biomass

Light is a major resource potentially limiting phytoplankton communities. Absorption of light in water is approximately logarithmic:

$$I_z = I_0 e^{-kz} \quad [4]$$

where I_0 is the incident light intensity, z is the water depth, I_z is the light intensity at depth z, and k is the light extinction coefficient. The light extinction coefficient is affected by both solutes and suspended particles, including phytoplankton cells. Even in pure

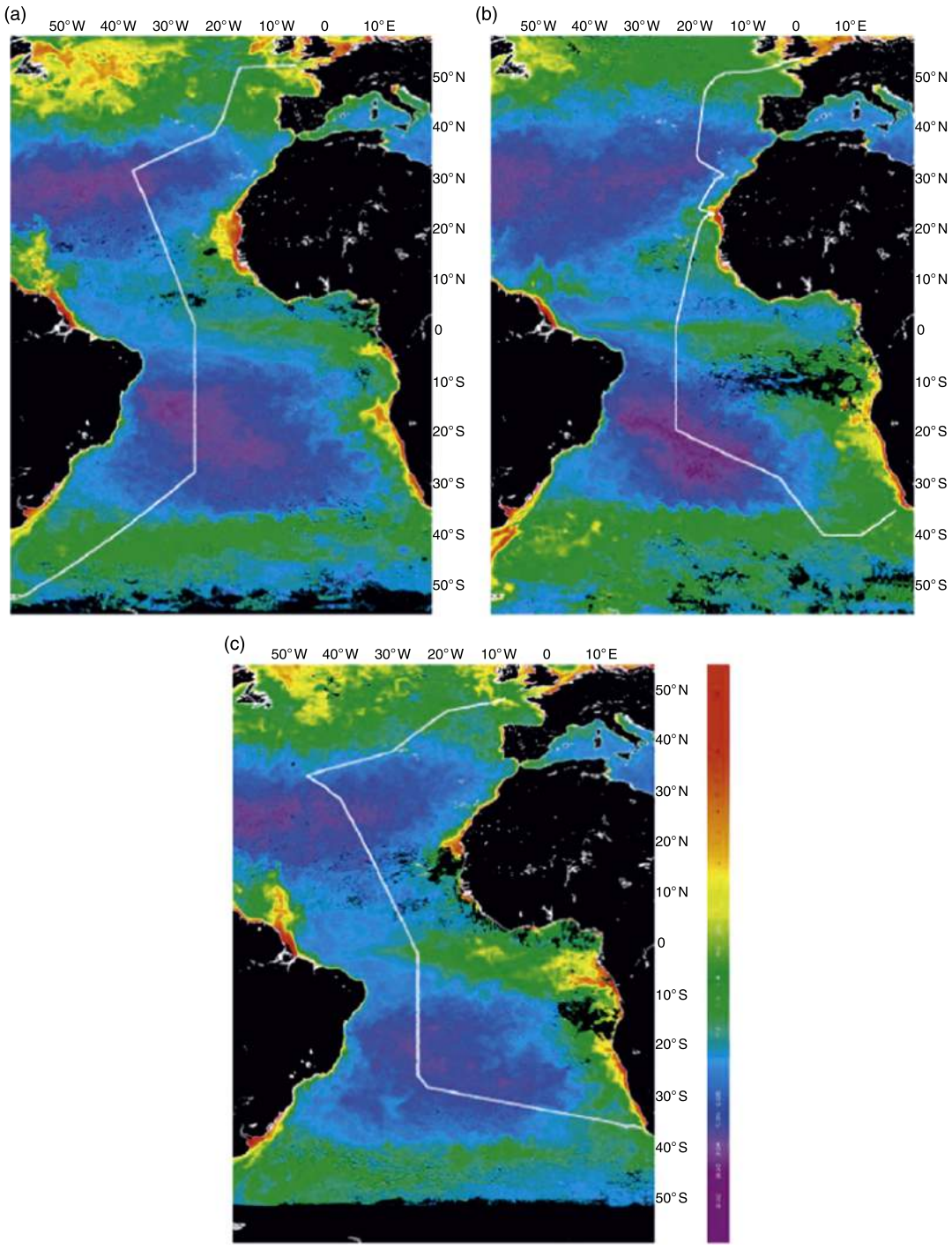


Fig. 2 Composite chlorophyll *a* images of the Atlantic Ocean (551° N–551° S by 601° W–201° E) for the months of (a) May 2004 (AMT14), (b) October 2004 (AMT15), and (c) June 2005 (AMT16). The tracks of each cruise are shown as a white line and are the same on each cruise between the Equator and 201S along the 251W meridian. Reproduced from Robinson, C., Poulton, A.J., Holligan, P.M., et al., 2006. The Atlantic Meridional Transect (AMT) programme: A contextual view 1995–2005. Deep-Sea Research II 53, 1485–1515, with permission from Elsevier.

(distilled) water, only about one half of incident visible light energy remains at a depth of 10 m and only 1% at 130 m. The thickness of the layer which benefits from more than 1% of the incident light (i.e., the euphotic zone) is very restricted in most oceans and lake ecosystems and only in very shallow and clear waters does the euphotic zone reach the bottom. Although many phytoplankton species are adapted to photosynthesizing at relatively low light intensities, the attenuation of light along the water column ultimately limits primary production.

Phytoplankton biomass growth rates are a saturating function of incident irradiance:

$$dB/dt = B_i[\mu_i I_0/(I_0 + K_i) - m_i] \quad [5]$$

where B_i is the biomass of species i, I_0 is the incident irradiance, μ_i is the maximum growth rate per unit biomass, K_i the half-saturation constant of a Monod-type function, and m_i is the phytoplankton biomass loss as a result of respiration, grazing, sedimentation, and washout.

Residence within the euphotic zone is essential for autotrophic organisms. However, since most phytoplankton species, with the major exception of Cianophyceae, cannot regulate their buoyancy and sink in the water column at a speed depending on their size, shape and appendices, as well as on the difference between the density of the cell and that of the water, their permanence in the euphotic zone is closely related to water column stability and mixing depth. Turbulent water movement influences the light to which phytoplankton are exposed. If the mixing depth (d_m) significantly exceeds the euphotic depth (d_e), then the phytoplankton is alternately carried into and out of the illuminated layers; primary production and phytoplankton biomass are strongly limited under these conditions. If $d_e > d_m$, phytoplankton can survive outside the mixing layer, provided they have some means of regulating their vertical position within the euphotic layer. The predominance of flagellates in deep, transparent tropical seas and in the metalimnia of stratificated lakes is made possible largely by their motility. Extensive vertical mixing in lakes and seas generally favors the dominance of diatoms. Apart from the morphological adaptation which enables diatoms to achieve suspension in turbulent eddies, they are adapted physiologically to contend with fluctuating irradiance conditions. Average photic conditions (related to the interaction between spatial and temporal fluctuation in water movements) and the suspension adaptations of phytoplankton strongly affect the rate of biological production, and account for the scarce winter development of phytoplankton at high latitudes.

Nutrient Limitation of Phytoplankton Biomass

Coast-open sea (or open lake) spatial patterns and manipulative enrichment experiments have shown that phytoplankton can become limited by the availability of nutrients when light and temperature are adequate and loss rates are not excessive. Phosphorus (P) and nitrogen (N), which are required in the largest amounts, are the two most common limiting nutrients. Ecological stoichiometry describes the relative requirements of phytoplankton groups with respect to the different nutrients entering the cell's anabolism. As a general average for marine phytoplankton, the uptake ratio of C, N, P, and S is 106:16:1:0.8, respectively. Therefore, the balance between the requirement and the availability of limiting nutrients, whose spatial and temporal variations determine the observed patterns of biomass density, depends on lateral nutrient inputs, on deep water nutrient inputs, and on nutrient turnover rates within the water column. Silicates, for example, strongly limit diatom growth, mainly because of their low turnover rates within the euphotic layer; diatoms are limited at silicate concentrations of below 1.4 mg l^{-1} – which are rapidly reached during the spring blooms. On the other hand, P has a much faster turnover rate than N in the euphotic zone, partly due to inorganic excretion by zooplankton grazers, allowing phytoplankton growth even at very low P concentrations. Since almost all nutrients have sedimentary biogeochemical cycles and terrestrial origins, the coast-open sea(lake) patterns arise from the progressive dilution of terrestrial inputs in aquatic ecosystems.

Taxonomic Composition and Richness Equilibria in Phytoplankton Communities

The equilibrium concept in phytoplankton communities refers to the steady state of a set of phytoplankton populations which are coexisting and hence `sampled together'. This means that while growth and loss processes occur simultaneously in the populations of a guild, it is possible that the result of these processes is the persistence of composition over time.

Persistence can be measured on the timescale of significant environmental variability as well as the turnover and generation times of individual species. In phytoplankton communities, where generation times are measured in hours and days, a few weeks of similar densities and community composition can be considered indicative of equilibrium or steady-state persistence. Cases of steady-state persistence are: the co-dominance of five species (*Planktothrix agardhii*, *Limnothrix redekei*, *Dictytosphaerium* sp., *Cyclotella meneghiniana*, and *Cryptomonas erosa*) over nine unperturbed weeks in El Porcal Lake (a gravel pit in Central Spain); the dominance of a number of species in 31 fluctuating sites of a wetland (La Safor, Mediterranean Spanish coast); and the persistence of some nondominant species (*Peridinium willei* and *Planktonema lauterbornii*) over more than 3 weeks in the water column mixing period in Las Madres Lake (Central Spain).

The observation of persistence within phytoplankton communities is also related to the spatial and temporal scale of the observer. From a practical point of view, it is a rule for researchers to sample at least weekly. This is because the time generation of microalgae is from 0.3 to 3 days, so a week is a representative timescale of population response. From a statistical point of view, a similar assemblage must be found in at least three successive samplings in order for it to be considered stable. Therefore, the minimum timescale commonly utilized to evaluate persistency may cover more than 50 generations of the smallest species.

The equilibrium concept in phytoplankton communities is supported by evidence of nutrient and resource limitation, of interspecific competitive coexistence and of trait-specific competitiveness ability among phytoplankton species. The high resilience of phytoplankton communities also supports the achievement of equilibrium conditions among species pairs or species groups on a

given time and spatial scale. The relatively small number of dominant species in phytoplankton communities supports this simplification.

However, the equilibrium concept in phytoplankton communities is not supported by the large number of coexisting species in a medium which can be considered relatively isotropic and unstructured; nor is it supported by the well-known tradeoff between population growth rate and stability. These considerations, which were first raised by Hutchinson, led to the famous `paradox of the plankton' and disequilibrium or nonequilibrium theories.

Therefore, the two branches of theories (i.e., equilibrium vs. nonequilibrium) address two different aspects of phytoplankton community structure: the relationship between pairs or small groups of species on the one hand, and the organization of phytoplankton communities as a whole, on the other hand.

Equilibrium Theories

The original description of persistent or steady-state phytoplankton communities was focused on dynamic features of temporal changes in phytoplankton and was based on resource-partitioning concepts, and more specifically on competition. The Tilman extension of the Volterra equilibrium model to primary producers represented the basis of equilibrium theory in phytoplankton communities.

To explain the `plankton paradox', equilibrium theories propose that several species of phytoplankton can coexist in true competitive equilibrium if they are collectively restricted from further growth by different nutrients. The relative abundance of the coexisting species can thus be controlled by the ratio of limiting resources. The hypothesis assumes that (1) several nutrients are in relatively short supply; (2) growth of each species is restricted by a single nutrient or a unique combination of several nutrients, according to the Liebig's law principle that the growth of each population occurs at the rate permitted by the most limiting factor; and (3) different species have different uptake capacities for the various nutrients. Equilibrium approaches to the phytoplankton paradox are deduced theoretically and verified in chemostat experiments. Equilibrium conditions between species pairs on two limiting resources can be described in terms of the competitive ability of each species with respect to each resource, as described by the resource requirement at equilibrium:

$$R^* = DK_i/(r_i - D) \qquad [6]$$

where R^* is the equilibrium extracellular resource density, r_i is the maximum growth rate, K_i is the resource concentration at which growth is half of the maximum growth rate and D is the dilution rate, taking account of resource concentrations at the start and resource inputs; the result of the equilibrium can be also derived graphically (**Fig. 3**).

Laboratory competition experiments have supported the relevance of resource competition as a process that determines the species composition of phytoplankton communities. Under chemostat-type steady-state conditions phytoplankton species limited by different resources can coexist in equilibrium; and the relative abundance of coexisting species is controlled by the ratio of limiting resources. Tilman has shown this with 76 competition experiments between the two species of freshwater algae *Asterionella formosa* and *Cyclolella meneghiana* under a wide range of Si:P ratios.

As results of these experiments *Asterionella formosa* was observed to be competitively dominant when both species were phosphate limited; *Cyclolella meneghiana* was dominant when both species were silicate limited; and both species stably coexisted when each species was growth rate limited by a different resource. Extension of this approach from a two-species system to multispecies experiments has supported the equilibrium theories; however, they are able to explain only a small proportion of the diversity and species richness of natural phytoplankton.

The Plankton Paradox, or 'Supersaturated' Communities

The remarkable diversity of phytoplankton communities in aquatic ecosystems is recognized as a 'paradox', because many more species with similar requirements (to be satisfied from the surrounding environment, apparently isotropic or unstructured) co-occur than is expected in a competitive equilibrium. The high level of richness of phytoplankton communities in the same water body has been called by G. E. Hutchinson the `plankton paradox' and the resulting communities are defined as `supersaturated'.

The plankton paradox was based on the assumption that: (1) phytoplankton guilds are assembled on the basis of differential growth rates among species, determined by the availability of inorganic nutrients; (2) species interact through competition for mineral nutrients, and (3) pelagic habitats are closed homogeneous systems. However, pelagic habitats are open systems, intraspecific competition is generally stronger than interspecific competition, and nutrients (as well as some other limiting factors) vary in time and space at different scales.

Therefore, in general terms, competitive equilibrium is never expected when a virtually complete competitive replacement of one species by another occurs in a time (t_c) of the same order as the time (t_e) taken for a significant seasonal change in the environment which reverses the competitive ability of the interacting species. Thus, ideally there are classes of cases:

1. $t_c \ll t_e$, competitive exclusion at equilibrium complete before the environment changes significantly;
2. $t_c \cong t_e$, no equilibrium achieved; and
3. $t_c \gg t_e$, competitive exclusion occurring in a changing environment to the full range of which individual competitors would have to be adapted to live alone.

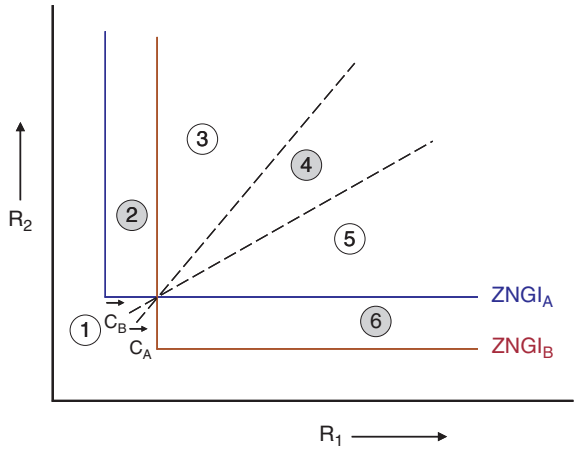


Fig. 3 Potential coexistence of two species for two limiting resources. ZNGI*i* (*i*=A and B) curve represents the amounts of the two resources which must be available for the species *i* to maintain an equilibrium population. Dashed lines are consumption vectors (CA and CB) of the species A and B, which represent the total consumption rates of resources by the species at equilibrium. ZNGI's curves cross at a two-species equilibrium point. ZNGI's curves and consumption vectors define six regions. Habitats which have resources points within the region 4 will have both species coexisting, while species A will dominate in habitat 2 and 3, and species B will dominate in habitat 5 and 6. Reproduced from Tilman, David; *Resource Competition and Community Structure*. © 1982 Princeton University Press. Reprinted by permission of Princeton University Press.

Case 2 gives rise to a number of nonequilibrium theories concerning phytoplankton communities. A nonequilibrium model assumes that fluctuations or disturbances occur with sufficient frequency to disrupt the course of competitive exclusion, but not so frequently as to force species to adapt to the overall variability. High biodiversity is attained when fluctuations or disturbances keep the competing populations far from equilibrium, thus allowing more species to coexist.

Two types of fluctuation can be distinguished.

1. Fluctuations caused by internal nonequilibrium dynamics generated by competitive interactions.
2. Fluctuations caused by external factors, such as oscillation in light and nutrient supply due to the seasonal cycle or less predictable factors such as changes in weather and hydraulic conditions. Moreover, spatial heterogeneity is an important reason to expect coexistence of species. Even in seemingly homogeneous environments, such as the open ocean, meso-scale vortices and fronts generate barriers preventing complete mixing and competitive exclusion.

Nonequilibrium Theories

Fluctuations determined by internal feedback factors

It is usually thought that in the absence of any externally imposed environmental fluctuation, phytoplankton communities will approach competitive exclusion and ecological equilibrium. However, phytoplankton continue to fluctuate erratically even when the environment is completely constant and uniform. This is because the high turnover rates and feedbacks in the phytoplankton system itself generate complex dynamics preventing the system from coming to equilibrium, the resulting community being referred to as 'supersaturated'.

It has been shown that competition for limiting resources leads to chaotic dynamics if multiple species compete for at least three resources. Theoretical models have shown that these dynamics depend crucially on the relationship between the resource requirements and the resource consumption characteristics of the species. Competition generates: (1) stable coexistence, if species consume most of the resources for which they have high requirements; (2) oscillations and chaos, if species consume most of the resources for which they have intermediate requirements; and (3) competitive exclusion, with a winner that depends on the initial conditions, if species consume most of the resources for which they have low requirements.

Fluctuations determined by external factors

Environmental disturbances that occur so frequently as to preclude competitive exclusion in phytoplankton species lead to a nonequilibrium, promoting coexistence and enhancing diversity. General analytic models show that resource supply in a situation of nonequilibrium can allow coexistence of species competing for the same fluctuating limiting resources. Theoretical results showing that resource fluctuations allow coexistence have been confirmed experimentally for phytoplankton competing for key nutrients (phosphorous and ammonia) and light. The coexistence-promoting mechanism believed to operate in variable environments is the creation of temporal niche opportunities that allows competitors to utilize the scarce resources at different times.

The coexistence of two or more species under variable resource supply conditions is possible when species exhibit a gleaner-opportunist tradeoff that entails a tradeoff between a low minimum resource requirement and a high maximum growth rate. A

gleaner species grows better at low resource levels as a result of a low minimum resource requirement. An opportunistic species is one that can take advantage of high resource levels.

Environmental fluctuations such as a pulsed nutrient supply can increase phytoplankton species diversity, facilitating coexistence. In equilibrium theory, the gleaner or `affinity specialist' is able to exclude the other competitors if the system tends toward the equilibrium resource level of the minimum R^* value. Fluctuations in resource levels can prevent this. If nutrients are sinusoidally fluctuating, then a species with a high growth rate, able to take higher advantage of high resources levels (i.e., opportunistic or velocity specialist), might coexist with a slower-growing but more efficient (gleaner) population. These results can be generalized to a wide class of growth functions and to pulsed variability of nutrient supply simulating periodic upwelling events.

Light fluctuations may also have an effect on the outcome of competition between phytoplankton species, leading to nonequilibrium coexistence. Light is never homogeneously distributed, even in a well-mixed water column; the light intensity at the surface always exceeds the light intensities at the bottom and a spatial gradient in light distribution with depth, coupled with diffusion of algal cells through the water column, may allow coexistence of many species of phytoplankton as well as their vertical segregation. Temporal variability in light supply can also have significant effects on the outcome of competition between phytoplankton species when species exhibit the gleaner-opportunist tradeoff. Rapid light fluctuations can reverse competitive dominance from a gleaner to an opportunist; slow light fluctuations can change the identity of the dominant competitor and also lead to the stable coexistence of competitors. Coexistence is easiest between species that are highly differentiated along the gleaner-opportunistic tradeoff.

Summary

The concept of 'equilibrium', as a balance between contrasting forces, may resemble to be not appropriate to describe structure and dynamics of phytoplankton communities, where species show extremely high population growth rates and high temporal and spatial variability. However, the small body size of phytoplankton individuals, which strongly affects individual energetics and population potential to growth, actually enable populations to adapt rapidly to varying environmental conditions, increasing population resilience and facilitating the achievement of equilibrium conditions. In fact, in terms of biomass, phytoplankton communities emphasize patterns of equilibrium with the environmental limiting factors both in space and in time. Most of the debate on the occurrence of some kind of equilibrium in phytoplankton communities deals with the taxonomic composition and richness of phytoplankton communities, being synthesized by the `plankton paradox' concept. Hutchinson introduced the term `plankton paradox' to describe the coexistence of so many phytoplankton species under limiting and relatively isotropic conditions, suggesting that whenever rates of competitive exclusion and of environmental changes are very close, a `dynamic equilibrium' with supersaturated communities can be observed. The `plankton paradox' is still unsolved; however, factors determining competitive abilities of phytoplankton species, potentially leading to `stable equilibrium conditions', as well as factors contributing to the maintenance of supersaturated communities under `dynamic equilibrium' conditions, have been emphasized. They include the intrinsic variability of phytoplankton populations and the occurrence of external factors of disturbance contributing to increase the patchiness in the apparently isotropic and unstructured water column.

See also: Aquatic Ecology: Eutrophication. Ecosystems: Estuaries; Lagoons. Evolutionary Ecology: Metacommunities. General Ecology: Ecological Stoichiometry: Overview; Ecophysiology; Biodiversity

Further Reading

Berman, T., Pollingher, U., 1974. Annual and seasonal variations of phytoplankton, chlorophyll, and photosynthesis in Lake Kinneret. Limnology and Oceanography 19, 31–54.
Harris, G.P., 1986. Phytoplankton Ecology: Structure, Function and Fluctuation. London: Chapman and Hall.
Huisman, J., Weissing, F.J., 1999. Biodiversity of plankton by species oscillations and chaos. Nature 402, 407–410.
Hutchinson, G.E., 1961. The paradox of the plankton. American Naturalist 95, 137–145.
Robinson, C., Poulton, A.J., Holligan, P.M., *et al.*, 2006. The Atlantic Meridional Transect (AMT) programme: A contextual view 1995–2005. Deep-Sea Research II 53, 1485–1515.
Tilman, D., 1982. Resource Competition and Community Structure. Princeton: Princeton University Press.

Relevant Website

http://fvcom.smast.umassd.edu–The Marine Ecosystem Dynamics Modeling Laboratory (MEDML).

Estuarine Ecohydrology

E Wolanski, Australian Institute of Marine Science, Townsville MC, QLD, Australia
L Chicharo and MA Chicharo, Universidade do Algarve, Faro, Portugal

The Failure of Present Estuarine Management From Ignoring Ecohydrology

Throughout the world, estuaries and coastal waters have experienced environmental degradation. Present proposed remedial measures based on engineering and technological fix have been unable to restore the ecological processes of a healthy, robust estuary and, as such, will not reinstate the full beneficial functions of the estuary ecosystem. The successful management of estuaries and coastal waters requires ecological engineering, that is, an ecohydrology-based, basin-wide, approach. Ecohydrology is the science that relates hydrological processes to the biological dynamics of ecosystems at various spatial and temporal scales. The ecohydrology concept was developed in the framework of UNESCO's International Hydrological Programme (IHP). It hypothesized and empirically confirmed in a number of demonstration sites that the ecological services of rivers and lakes can be restored by using hydrology to regulate biota dynamics and vice versa. The synergic integration at the basin scale of various ecohydrological measures based on ecological needs provides the scientific background for twinning ecosystem variables in order to enhance the carrying capacity and the resilience of ecosystems while promoting positive socioeconomic feedbacks.

Implementing the ecohydrology approach necessitates getting away from present management practices based on regulation focused on the geography (e.g., individual municipalities or counties) or on individual, specific activities (e.g., farming and fisheries, water resources, urbanization, shipping) without integrating among localities and users so as to consider the ecosystem. Without this change in thinking and management concept, estuaries and coastal waters will continue to degrade, whatever integrated coastal management plans are implemented because these ignore the basic ecology fact that the land, the river, the estuary, and the coastal waters are connected by being in the same ecosystem.

Can science-based management save estuaries and coastal waters? About environmental management the main thing that can be reliably managed is the human behavior and practices. The main thing that ecological engineers can do is to highlight the role of ecohydrology in offering a robust, science-based way to quantify both the present human impact on the degradation of estuarine and coastal ecosystems and the likelihood of success of various remediation measures in improving the health of these ecosystems. Remediation measures include changing human behavior and practices in using the land and water resources. These measures also include boosting the ecosystem robustness by manipulating the ecosystems so as to reinforce the beneficial ecological feedbacks in these ecosystems. The eventual success of these measures relies on adopting an ecohydrological approach.

Estuaries and coastal areas are traditionally among the most highly impacted waters. For centuries human populations have settled and have benefited from the services provided from these highly productive ecosystems. During the last century, the human population worldwide has increased by a factor of about 10 in coastal areas of many developing countries and probably a factor of 4 for most developed countries, as a result of natural population growth and migration. As a result the carrying capacity of coastal ecosystems is exceeded and this has commonly led to a serious environmental degradation of estuaries and coastal waters worldwide. This is best demonstrated for the megacities and harbors in the Asia Pacific region where the best of engineering practices have failed to provide a healthy environment to 100 million people. In most sub-Saharan African countries, migration to the coast is still increasing the coastal population yearly at typically 5%–8% because coastal areas provide free access to food (fisheries) and timber (mangrove trees) and hope for jobs in harbors and coastal cities.

In view of this ever-increasing human impact on coastal ecosystems, corrective or preventive measures based on an ecohydrological approach should be considered instead of just engineering measures as has been the general approach so far. This does require changing the legislation that is a quagmire for the management of estuarine and coastal waters. The successful implementation of an ecohydrology approach to managing estuaries and coastal waters must be based on a sound scientific knowledge of the ecosystems functions and dynamics.

The ecohydrological approach has been tested successfully in various locations and systems (salt marsh estuary, mangrove estuary, and coral reefs) around the world, particularly in the Guadiana Estuary in temperate Portugal and Darwin Harbor and the Great Barrier Reef in tropical Australia. The health of these ecosystems was demonstrated to depend on the connectivity between estuarine and coastal waters, on the links between physical and biological processes, on the drainage basin hydrology, and on the disturbance caused by human activities. It was possible at these sites to quantify the human impact on the ecosystem health, and the impact of remedial activities.

Major observed impacts in estuarine and coastal ecosystems are the increasing eutrophication risk, toxic algal bloom events, muddiness and siltation of estuaries and coastal erosion, and also modifications in biodiversity resulting in the loss of traditional ecosystem services (e.g., coastal fisheries) and having negative socioeconomic impacts (e.g., the loss of income and employment for coastal communities). The degradation is most dramatic for coastal coral reefs located within the estuary and in its coastal waters (Fig. 1). Coral reefs possess the highest diversity of any marine and most terrestrial ecosystems and they greatly benefit humanity by building islands and atolls, by protecting shorelines from coastal erosion, and supporting fisheries and diving-related tourism. Coastal reefs are being destroyed at an accelerating and unsustainable rate worldwide (e.g., up to 50% in the last 15 years

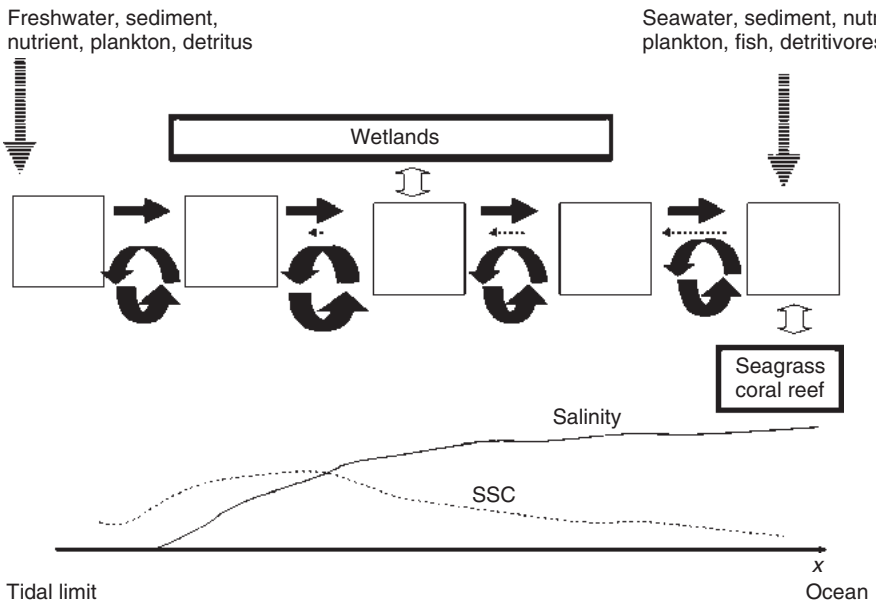


Fig. 1 The physical submodel schematizes the estuary as a series of cells connected to each other by advection driven by both the river discharge (straight thick arrows) and the salinity-induced currents (broken thin arrows) and by tidal mixing (curved arrows). The open boundary conditions are located at the tidal limit where a riverine flux of water and waterborne particles is imposed, and at the mouth where an oceanic flux is also imposed. The estuary also exchanges water and waterborne particles with the wetlands (perimarine wetlands, mangroves, and salt marshes) and the seagrass and the coral reefs. The suspended solid concentration (SSC) commonly has its largest value in the turbidity maximum zone. The salinity increases in the estuary toward the mouth.

in some Asian countries) by human activities that can be devastating (e.g., mining for limestone, fishing with explosives and cyanide, infilling for urbanization) or threatening but possibly manageable (e.g., increased runoff of mud, nutrients, and pesticides from adjacent river catchments, overfishing, and global warming). The present coral reef strategy principle relies on drawing a line around coral reefs and protecting the corals inside that line. It is politically convenient and it ignores the fact that the land and the reef are ecologically connected through the rivers. It thus invariably fails worldwide wherever the reefs are impacted by runoff that is modified in quantity and quality by human activities on land.

Intensive agriculture practices, urban sewage, and manure from pig farms and cattle feedlots provide quantities of nutrients that reach the estuary by runoff or as groundwater and can cause eutrophication to a level dependent on the robustness of the estuary. The riparian vegetation upstream and the vegetation in salt marshes and/or mangrove swamps downstream play a bottom-up role as buffers that retain nutrients and reduce the load remaining in the estuary. Also, filter-feeders and grazers (e.g., fish and bivalves) can exert a top-down control on primary producers' biomass (**Fig. 2**) and improve water quality. Indeed, bivalve suspension feeders have also been suggested as a means to control algal blooms, in both the marine and freshwater regions of the estuary. However, bivalve excretion increases the pool of nutrients available for other primary producers as the macroalgae like *Ulva* sp. or the colony-forming *Phaeocystis* sp.

Studies also provide evidence that changes in the river discharge impact on the structure of estuarine plankton, fish, and bivalves communities. In fact, the response to low-flow periods has been recognized as one of the most important factors in structuring biota, both in the estuary and coastal waters. During low-flow periods, a decrease in the concentrations of those nutrients reaching the sea may occur, especially in dammed rivers, since the silicates (Si) trapped in the dam is not reintroduced downstream in the system, as may occur with N and P (as consequence of the use of fertilizers in agriculture). In such situation, Si can be limiting to the growth of diatoms, and may contribute to the occurrence of toxic algal blooms. The reduction of low flows by dams and irrigation is thus a threat to the health of estuarine ecosystem.

Dams reduce river flows. These high river discharges are ecologically important because they 'feed' the coastal waters with sediments and nutrients, ensuring the adequate nutrients Si:N:P ratios and promoting coastal waters productivity There is increasing evidence that coastal fisheries landings are related to the high river flows discharge and not climate factors, as has been demonstrated for South Portugal, the East Mediterranean coast, the Black Sea, and the Gulf of Carpentaria. For instance in the Guadiana Estuary and coastal waters following high-discharge periods, catches are dominated by planktivorous fish like the anchovy (*Engraulis encrasicolus*) and the sardine (*Sardina pilchardus*). This is because high river discharges promote primary and bacterial production from increased nutrient loading and organic matter loading. They also increase larval and juvenile survival because higher turbidity associated with greater sediment loads reduces predation. They facilitate the retention of early life-history stages and increase the survival. They provide an environmental clue (salinity gradients) for shrimps and fish to migrate towards the estuary where tidal wetlands (salt marshes and mangroves) are used as a nursery for the larvae and juveniles. In areas where the continental shelf is narrow, the effects of freshwater runoff can reach the upper continental slope, especially during upwelling events, increasing larval physiological

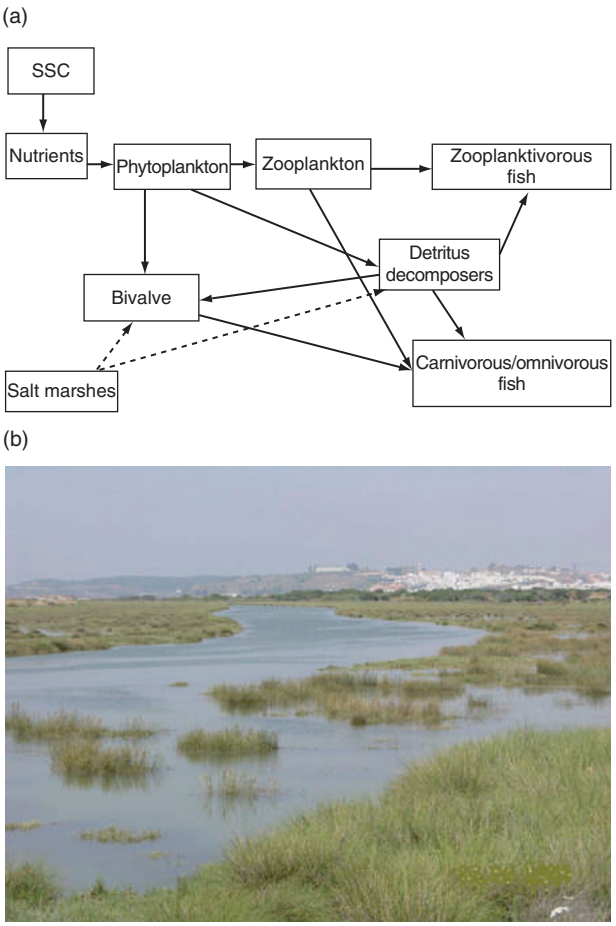


Fig. 2 (a) The biological submodel for the Guadiana Estuary, Portugal. The salt marsh is modeled as a source of detritus and juvenile bivalves (broken arrows). At death all organisms become detritus (not shown). SSC, suspended solid concentration. (b) Photograph of Guadiana Estuary salt marshes and encroaching urbanization in the background.

condition and consequently decreasing larval mortality. These larvae can find abundant food in winter because during this rainy period the freshwater discharge is usually high, promoting salinity stratification and the development of a shallow surface mixed layer where phytoplankton that can bloom long before the spring phytoplankton blooms occur.

These examples show that the regulation of estuarine and coastal biota processes is highly dependent on the river discharge. This offers politicians and water resources managers the option to try to improve estuarine ecosystem health on a case to case basis by modifying mostly one variable.

The estuarine ecosystem modeler is faced with complex processes and feedback processes between the physics and the biology, which in practice cannot be fully quantified because the data are inadequate. For this reason, models should only be used as tools that help to explain the reality, steer field researchers towards studying key processes that most control estuarine health, and help quantify the impact of human activities on the health of estuarine ecosystems. Models should not be seen as able to replace reality and the need for field observations. Governments often want to cut budgets for field research and instead promote the cheaper option of modeling. Estuarine ecosystem models are simply unable to provide reliable answers without field studies.

Estuarine ecohydrology models are based on the knowledge and quantification of physical and biological processes controlling the ecosystem. They can quantify the likely effectiveness of ecohydrology solutions for eutrophication by two opposite ecological approaches, namely top-down control on vegetation by herbivores, or bottom-up control of nutrients loads. They can also quantify the likely effectiveness of dealing with eutrophication by regulating the river flow, for example, allowing freshets. The control of nutrient input into the system can be obtained by creating or restoring wetlands, which will act as buffers and trap nutrients. These wetlands, when located near the coast, also protect the coast from erosion, thus protecting inland infrastructures and coastal populations. The cumulative effect of wetland creation and controlling the riverine nutrient loads through changes in land-use practices invariably lead to an efficient decrease in nutrients in the estuary, often larger than the sum of the two impacts individually. For each of these solutions and for each case, models should be used before field implementation of the remediation activities to assess likely improvements in the health of the ecosystems and to compare these with observations, so as to be able to separate human from climate effects.

A top-down approach can be considered and modeled for control of eutrophication, that requires the knowledge about the ecological food web in the area: who eats who, when, and how much. The quantification of these processes and relations is basic to the selection of the most effective grazer species. In some cases manipulation of the food web could be necessary, for example, (1) controlling the density of the predator species; (2) allowing the herbivore species – that will control the vegetation growth – to become more abundant, and (3) introducing bivalves to filter the water and reduce the risk of toxic algae blooms. Such is the case in San Francisco Bay where eutrophication is inhibited because half of the water is filtered daily by bivalves that were accidentally introduced.

Toxic algal blooms are frequent in nutrient-enriched estuaries and, in particular, in dammed estuaries, and estuaries with reduced flushing and very long residence times. For instance, Tokyo Bay in Japan and the Pearl River estuary in China have respectively about 100 and 200 days of toxic algae blooms every year in various areas. Bivalves have been successfully tested to control phytoplankton blooms and reduce eutrophication and toxic algal blooms. Another way to control algal blooms is to generate freshets (i.e., freshwater discharge pulses). During freshets, with the duration of typically few days, the amount of nutrients available for phytoplankton growth increases, causing a decrease in competitive exclusion mechanisms and the consequent increase in phytoplankton diversity. In turn this promotes the growth and the diversity of zooplanktonic species. Consequently, more zooplankton species may play a regulatory role by controlling phytoplankton density and 'avoiding' the phytoplankton growth of just few species, therefore reducing the risk of toxic algal blooms.

The use of freshets to control algal blooms requires a profound knowledge of the system on a case to case basis, in order to determine the timing, the magnitude, and the duration of the freshet. A poor timing of freshets could negatively impact species that use the estuaries and its tidal wetlands as a nursery ground, if the freshet interferes with the natural recruitment of these species, for instance by flushing out the eggs and larvae.

Estuarine ecohydrology models have been developed and verified for few ecosystems, recently for the Guadiana Estuary in Portugal and Darwin Harbor and the Great Barrier Reef in Australia in a program supported by UNESCO-ROSTE, NOAA, AIMS, and the University of Algarve/CCMAR. The explicit aim was to offer science-based solutions to management. The ecohydrology models due to its holistic approach is a tool that aims to facilitate an interaction between scientists, economists, the public, and decision-makers to promote the ecologically sustainable development of an estuary and its coastal waters.

An Estuarine Ecohydrology Model

This model is based on the dominant ecohydrological processes in tidal estuaries. The model is best suited to estuaries that are fairly well mixed vertically. In practice this constraint is adequate for many applications because critical conditions in estuaries commonly arise during low-flow conditions brought upon by dams and water extraction. These conditions are exacerbated by excess nutrients from, typically, sewage discharge, effluent from feedlots, and fertilizers from farming. Thus the physical submodel, sketched in **Fig. 1**, views the estuary as a series of connecting cells that exchange water by advection as result of the river runoff and the saltwater inflow, and by tidal mixing. The upstream cell receives freshwater, sediment, nutrient, and freshwater plankton. The downstream cell receives seawater and marine sediment, nutrients, and plankton. Cells can also exchange water and particulates with fringing wetlands both in the fresh and the saline region of the estuary. The physical processes and the open boundary conditions control the salinity distribution and the suspended solid concentration (SSC; both are sketched in **Fig. 1**). An estuarine turbidity maximum (ETM) commonly prevails. The salinity and SSC distribution vary both spatially and temporally.

This physical submodel model is linked to an ecological submodel that is adapted to local conditions and is based on the results of field studies. This submodel is also a simplification of the processes in estuaries, focusing on the dominant processes in terms of mass transfer. The resulting ecological submodel for the Guadiana Estuary is shown in **Fig. 2** and that for Darwin Harbor in **Fig. 3**. **Fig. 2** and **3** show the minimum level of complexity of the model that is necessary to capture the ecology of these estuaries. There are important features that are found in both temperate and tropical estuaries, namely (1) the importance in turbid waters with a very long residence time of the suspended solid in releasing for biological productivity particulate nutrients, (2) the dominance of the microbial loop especially near the ETM zone, (3) the role of detritivores and bivalves, and (4) the role of wetlands as a source of detritus as well as mainly a nursery ground. These tidal wetlands are made up of mangroves in the tropics and salt marshes in the temperate climates. A key commonality also is the importance of river floods, even short-lived ones, in attracting coastal fish to migrate up-estuary by kinesis or taxis by swimming following environmental clues, primarily salinity gradients.

The model has been successfully verified against field data for these two estuaries.

Estuarine Ecohydrology Applications

For the Guadiana Estuary the model is particularly useful in quantifying the importance of short freshets in enhancing plankton and fish diversity (**Fig. 4**). It is also useful in quantifying the importance of tidal wetlands (salt marshes) in reducing eutrophi-

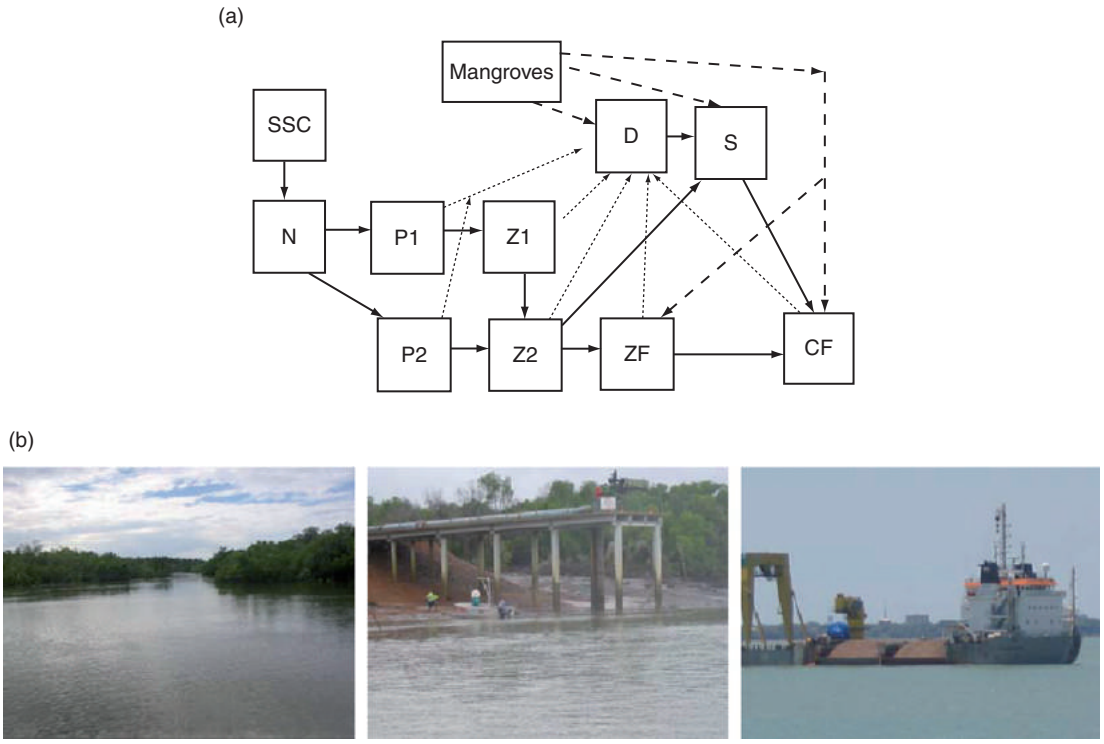


Fig. 3 (a) The ecology submodel for tropical Darwin Harbour, Australia. SSC, suspended solid concentration; N, nutrients; P, phytoplankton (two dominant species with different turnover rates); Z, zooplankton (two dominant species with different preys and turnover rates); D, detritus; S, detritivores; ZF, zooplanktivorous fish; CF, carnivorous fish. The mangrove swamp is modeled as a source of detritus as well as a source of young detritivores and fish (thick broken arrows). At death all organisms become detritus (thin broken arrows). (b) Photographs in Darwin Harbour of pristine mangroves in traditional Aboriginal land on the west bank (left), aquaculture industries encroaching in mangroves in the southern region (middle), urbanization and port development on the east bank removing all natural habitats (right).

cation (Fig. 5). This demonstrates clearly two management strategies that can be used to alleviate eutrophication and promote biodiversity and estuary robustness, that is, create freshets from the dam and creating or restoring tidal wetlands. The ecology submodel also highlights the key role of bivalves in filtering the water, suggesting another strategy of manipulating bivalves to reduce eutrophication.

Similarly for Darwin Harbor the model is useful to assess to what degree the estuarine ecosystem health may degrade – and what level of human disturbances in the drainage area is admissible to maintain a reasonable ecosystem health – as a result of future human activities in the catchment, particularly the urbanization of Darwin, the impact on the estuarine health of nutrient enrichment from sewage discharges, and the destruction of tidal wetlands (mangroves) for shipping and industry.

Coral Reef Ecohydrology Model

The sustainable development of coastal waters and coral reefs is dependent on development policies for land and water resources. To quantify this dependency, it is necessary to understand the key biological and oceanographic processes governing the health of these coastal ecosystems. These processes are then incorporated in a coral reef ecohydrological model that can predict reef health. The immediate use and role of this model is to help answer two key questions for managers. These two key questions are as follows: (1) to what extent do changes in quality and quantity of terrestrial runoff lead to reef degradation by generating phase shifts – the process by which areas formerly dominated by corals are overgrown by algae; and (2) is the reef capable of sustaining or rebuilding its biodiversity by self-seeding if remediation measures are implemented on land and in rivers to moderate the human impact?

The estuarine ecohydrology model has been modified and applied to coastal coral reefs that are subject to human impacts from (1) land runoff resulting in an increase in suspended sediment, increased water turbidity, and increased nutrient concentration, and (2) from global warming resulting in increased bleaching events in summer. The coral reef physical submodel (Fig. 6) is more complex than that of estuaries because it considers also river floods and tropical cyclones that both negatively affect coral reefs, and the oceanographic connectivity between reefs that enable self-seeding as well as the connectivity, that is, the exchange of coral planulae between reefs. The biological submodel (Fig. 7) is based on the competition for hard substrate space between the algae

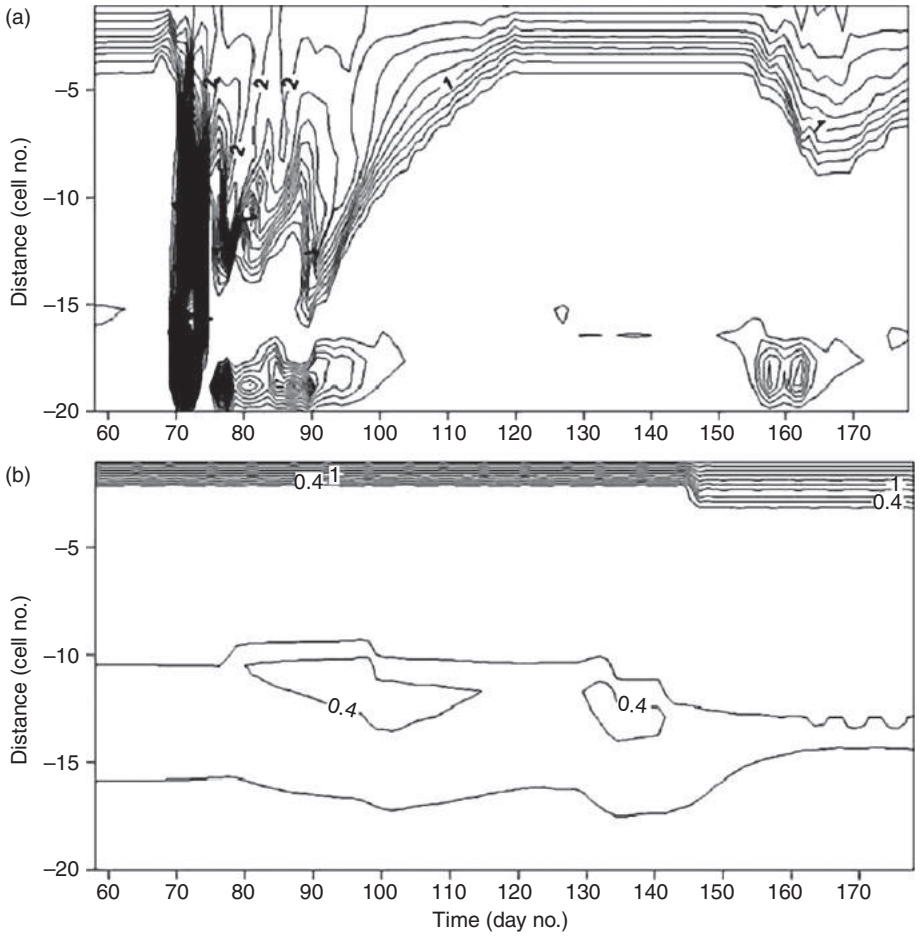


Fig. 4 Predictions of the along-channel distribution (cell 20, river mouth; cell 1, tidal limit = 60 km) of zooplanktivorous fish biomass in the Guadiana Estuary, Portugal, in 2004, (a) without and (b) with the Alqueva dam. The dam construction was completed in 2003 and in 2004 the dam suppressed all freshets. The 'without dam' calculation assumes natural river hydrographs that were calculated from rainfall data. The 'with dam' calculation uses the observed river runoff as the open boundary condition at the tidal limit. To convert biomass to concentration for fish $2.8 \approx 2.87\ g\ cm^{-2}$.

and the coral. The coral is preyed upon by crown-of-thorns starfish, whose population dynamics is also modeled. Algae are preyed upon by herbivorous fish that in turn is preyed upon by carnivorous fish that is harvested by people. Suspended sediment concentration (turbidity) and nutrients modulate all these processes. Additionally the success of recruitment of juvenile coral decreases with increasing algal cover on the hard substrate. Global warming results in an increased mortality of adult corals.

Coral Reef Ecohydrology Applications

This model was developed for the Great Barrier Reef and successfully verified against an extensive data set. It has also been applied to reefs in Micronesia (Guam, Palau, and Pohnpei).

The people of Micronesia have centuries, as opposed to decades in the Great Barrier Reef, of experience in dealing with coral reefs upon which their livelihood depends. They have traditional management policies that highlight the need to manage human activities that affect coral reef ecosystems. In many islands, the people have direct ownership of coral reefs and the fisheries they support. For Micronesia, the model highlighted the beneficial role of mangroves, and this has resulted in a legislative protection of mangroves in at least one state (Palau). For Guam, Palau, and Pohnpei, the study demonstrated the need for integrating land-use and coral reef management. In islands where some form of traditional leadership still exists, this model has measurably helped in improving local environmental planning because these traditional leaders take into account the long-term, multigenerational impacts of activities in the development of environmental policies. Thus ecohydrology and ecological engineering are accepted and becoming a powerful tool in such islands.

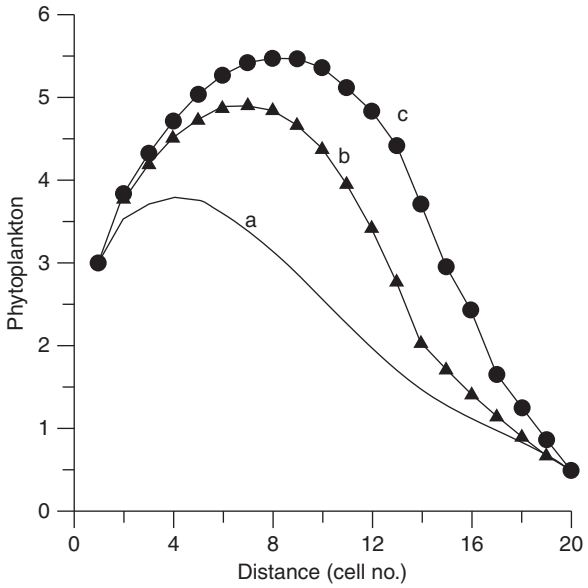


Fig. 5 Predictions of the along-channel distribution (cell 20 = river mouth; cell 1 = tidal limit = 60 km) of phytoplankton biomass for (a) low-flow conditions (2 $m^3 s^{-1}$) in summer in the Guadiana Estuary, Portugal, for (b) a hypothetical doubling of the riverine nutrient inflow as a result of planned irrigation farming using Alqueva dam water, and (c) with in addition the removal of the salt marshes. To convert biomass to chlorophyll *a* concentration 3.5 $\approx$ 7.8 $\mu g\ l^{-1}$.

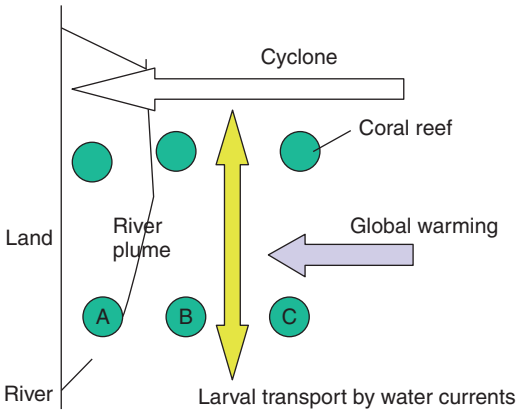


Fig. 6 Sketch of the dominant physical processes in the coral reef ecohydrology model. A, B, and C represent coral reefs with increasing distance from the coast.

For the Great Barrier Reef of Australia, the model suggests that land use has contributed to the degradation of the health of the Great Barrier Reef and to an increased frequency and intensity of crown-of-thorns starfish infestations. The model also predicts that the health of the Great Barrier Reef will significantly worsen by the year 2050 as a result of global warming. The model demonstrates to managers and politicians that it is worth improving land-use practices to recover the health of the Great Barrier Reef ecosystem. Indeed the model suggests that much-improved land-use practices will enable some regions of the Great Barrier Reef to recover, even with global warming. However, in the longer-term the situation is more gloomy, because the model suggests that if global warming proceeds unchecked only biological adaptation can prevent a collapse of the Great Barrier Reef health by the year 2100.

This ecohydrology model can be used to quantify the effectiveness of remedial measures on land. In theory thus ecological engineers can offer to economists and politicians the hard science data on ecosystem health that are needed to develop management policies that integrate socioeconomics and ecosystem health, as a first step towards planning an ecologically sustainable development. In practice however for the Great Barrier Reef and most corals reefs worldwide outside of a few islands in Micronesia, ecological engineering may have little impact because of two phenomena. Firstly there is the 'tragedy of the commons' where few

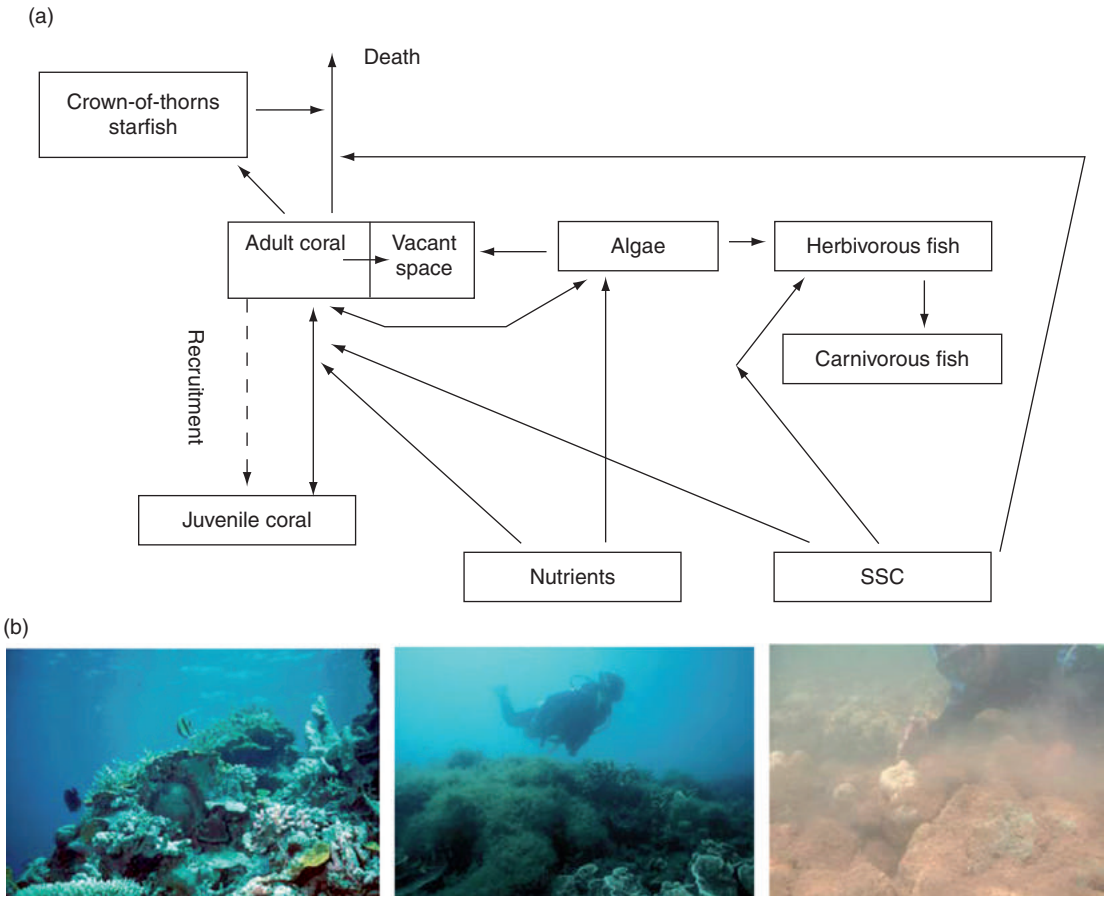


Fig. 7 (a) Sketch of the dominant biological processes in the coral reef ecohydrology model where SSC indicates the suspended sediment concentration that is a measure of turbidity. (b) Underwater photographs of a healthy coral reef (left), a coral reef overgrown by algae – this is a stable ecological state as long as poor water quality prevails (middle), a coral reef smothered and killed by mud from eroded soil from the adjoining river catchment (right).

take any responsibility but everyone has ownership – this is the same problem which has resulted in the collapse of fisheries worldwide. Secondly there is uncertainty in the science of cause and effects of reef degradation – some of that uncertainty is inherent to science, much is purposely manufactured – and this uncertainty helps politicians and decision-makers to justify ignoring the problem and implement no remedial measures.

Conclusions

Estuarine and coastal habitats are increasingly degraded worldwide. Ecohydrology demonstrates unambiguously that the land, the river, the estuary, and coastal waters are components of the same ecosystem. They are connected through a number of physical and biological processes that determine environmental health. Traditionally, the estuarine management strategy relies on technology and engineering fixes and it neglects ecohydrology principles; these strategies invariably fail to maintain ecosystem health and the ecological services that these ecosystems provide. Coastal coral reef management strategies worldwide also neglect ecohydrology science and also invariably fail. Ecohydrology science offers a number of solutions, including top-down and bottom-up ecological manipulation as well as the use of created or restored wetlands to help restore the health of estuarine and coastal waters. Combined with some technological fixes, such as the creation of freshets and smarter land-use, ecohydrology science offers an ecologically sustainable management strategy for estuaries and coral refs. Worldwide the implementation of this science-based strategy will most likely stall, and estuaries and coastal waters will continue to degrade, until a political solution is found to the quagmire (i.e., the present estuarine and coastal management framework) which basically ignores ecosystem ecology.

Estuarine and coastal areas suffer an increasing pressure from anthropogenic activities. Modifications of physical and chemical parameters affect biodiversity and hamper the traditional uses and services by local populations. Estuarine and coastal management relied, traditionally, on technology and engineering fixes, neglecting the ecological equilibrium of the systems.

Alternatively, ecohydrology science, based on the interplay between hydrology and ecology, offers a number of sustainable and long-lasting solutions to increase the robustness and to restore the health of estuarine and coastal waters. Estuarine ecohydrology models aiming to facilitate the interaction between scientists, economists, the public, and decision-makers have been developed and verified for a few ecosystems, recently for the Guadiana Estuary in Portugal and Darwin Harbor and the Great Barrier Reef in Australia, under a scientific program supported by UNESCO-ROSTE, NOAA, AIMS, and the University of Algarve/CCMAR.

See also: Aquatic Ecology: Eutrophication; The Estuarine Quality Paradox Concept. Ecological Processes: Physical Transport Processes in Ecology: Advection, Diffusion, and Dispersion. Ecosystems: Estuaries; Floodplains; Lagoons. General Ecology: Salinity

Further Reading

Chícharo, L., 2004. Estuarine and coastal areas: How to prevent degradation and restore. In: Zalewski, M., Wagner-Lotkowska, I. (Eds.), Integrated Watershed Management – Ecohydrology and Phytotechnology. Japan, IHP: UNEP, UNESCO, pp. 202–208. ch. 13.

Chícharo, M.A., Chícharo, L., 2004. Estuarine and coastal areas: How and what to measure. In: Zalewski, M., Wagner-Lotkowska, I. (Eds.), Integrated Watershed Management – Ecohydrology and Phytotechnology. Japan, IHP: UNEP, UNESCO, pp. 124–131. ch. 8.

Dyer, K.R., 1997. Estuaries. A Physical Introduction. Chichester, UK: Wiley.

Erzini, K., 2005. Trends in NE Atlantic landings (southern Portugal): Identifying the relative importance of fisheries and environmental variables. Fisheries Oceanography 14 (3), 195–209.

Hutchinson, G.E., 1961. The paradox of the plankton. The American Naturalist 95, 137–147.

Kimmerer, W.J., 2002. Physical, biological and management responses to variable freshwater flow into the San Francisco estuary. Estuaries 25, 1275–1290.

Lindeboom, H., 2002. The coastal zone: An ecosystem under pressure. In: Field, J.G., Hempel, G., Summerhayes, C. (Eds.), Oceans 2020: Science, Trends, and the Challenge of Sustainability. Washington, USA: Island Press, pp. 49–84.

Lloret, J., Palomera, I., Salat, J., Solé, I., 2004. Impact of freshwater input and wind on landings of anchovy (Engraulis encrasicolus) and sardine (Sardina pilchardus) in shelf waters surrounding the Ebre River delta (northwestern Mediterranean). Fisheries Oceanography 13 (2), 102–110.

McLusky, D.S., Elliott, M., 2004. The Estuarine Ecosystem. In Ecology, Threats and Management. Oxford, UK: University Press.

Roelke, D.L., 2000. Copepod food quality threshold as a mechanism influencing phytoplankton succession and accumulation of biomass, and secondary productivity: A modelling study with management implications. Ecological Modelling 134, 245–274.

Syvitski, J.P.M., Harvey, N., Wolanski, E., *et al.*, 2005. Dynamics of the coastal zone. In: Crossland, C.J., Kremer, H.H., Lindeboom, H.J., *et al.* (Eds.), Coastal Fluxes in the Anthropocene. Berlin: Springer.

Wolanski, E., 2001. Oceanographic Processes of Coral Reefs: Physical and Biological Links in the Great Barrier Reef. Boca Raton, FL: CRC Press.

Wolanski, E., 2006. The Environment in Asia Pacific Harbours. Dordrecht, The Netherlands: Springer.

Wolanski, E., Boorman, L.A., Chicharo, L., *et al.*, 2004. Ecohydrology as a new tool for sustainable management of estuaries and coastal waters. Wetlands Ecology and Management 12, 235–276.

Wolanski, E., De'ath, G., 2005. Predicting the present and future human impact on the Great Barrier Reef. Estuarine, Coastal and Shelf Science 64, 504–508.

Wolanski, E., Richmond, R., McCook, L., Sweatman, H., 2003. Mud, marine snow and coral reefs. American Scientist 91, 44–51.

Wolanski, E., Sarsenski, J., 1997. Larvae dispersion in mangroves and coral reefs. American Scientist 85, 236–243.

Zalewski, M., 2002. Ecohydrology – The use of ecological and hydrological processes for sustainable management of water resources. Hydrological Sciences Bulletin 47, 823–832.

The Estuarine Quality Paradox Concept☆

Michael Elliott, University of Hull, Hull, United Kingdom
Victor Quintino, University of Aveiro, Aveiro, Portugal

Introduction

Estuaries are naturally stressed, highly variable ecosystems and at the same time they are exposed to high degrees of anthropogenic stress (Elliott and Whitfield 2011). In marine and estuarine areas those stressful conditions are manifest as a suite of symptoms which relate to the structure and functioning of biological communities (Pearson and Rosenberg, 1978; Gray and Elliott, 2009). This article describes the similarity between the features of organisms and assemblages in both estuaries and anthropogenically-stressed areas and hence it emphasizes the difficulty of distinguishing natural from human-induced stress in estuaries - this concept was named the *Estuarine Quality Paradox* by Elliott and Quintino (2007). The features of coping with stress, whether natural or anthropogenic, relate to both areas. The analysis of such estuarine features requires extensive data relating to the structure and functioning of the ecological communities and their response to natural and anthropogenic perturbations (Quintino *et al.*, 2006).

Estuaries are regarded as stressed ecosystems because of their highly variable salinities which are also, by definition, usually lower than the adjacent coastal areas (Elliott and McLusky, 2002; McLusky and Elliott, 2007; Elliott and Whitfield, 2011). The other physico-chemical elements such as hydrographic patterns (current speed and direction, density stratification, riverine and tidal inputs, etc.), temperature, nutrient levels, pH, etc. are also more variable than in corresponding coastal and marine sites and so this variability will also contribute to stress on the biota if they are not able to withstand it (Whitfield *et al.*, 2012). Although the estuarine fauna and flora has freshwater elements, such as tubificid oligochaetes, which are of course able to tolerate very low salinities but are stressed by any salinity greater than 0.5, most estuarine biota are marine-derivatives in which the assumption is that therefore they will be stressed in salinities below 30 (McLusky and Elliott, 2004). Hence typical estuarine species, that is, those such as certain nereid polychaetes which occur mainly in estuaries, do not appear be stressed by variable salinity and indeed appear to thrive because of it.

Salinity is the predominant environmental master factor in estuaries in which the balance between the river flow and marine ingression, especially in tidal systems, creates the salinity gradient (Smyth and Elliott, 2016). Salinity then has a large physiological effect on marine organisms which may encounter estuarine conditions (Smyth *et al.*, 2014). This in turn produces a set of ecological patterns and gradients called the Remane curve and ecotones where systems merge (Whitfield *et al.*, 2012; Basset *et al.*, 2013b). This article aims to consider these features in the estuarine biota but places more emphasis on the macrobenthic invertebrates given their often sessile nature and an inability to move rapidly if the conditions are unsuitable. However, it emphasizes that the patterns also are shown in other organisms, such as opportunistic macroalgae and fish.

Anthropogenically and Naturally Stressed Areas

Environmental impact assessments in the coastal and estuarine environments have by necessity focused on the faunal and floral community structure, and especially the analysis of soft sediment benthic infaunal communities, given its sessile nature (McLusky and Elliott, 2004; Gray and Elliott, 2009). This large body of information has led to conceptual models against which change is measured. This is typified by the Pearson-Rosenberg paradigm (Pearson and Rosenberg, 1978), an extensively-used conceptual model and similar to the Rhoads-Germano model cited in North America (Rhoads and Germano, 1982). It has long formed the basis of approaches and indices used to detect and explain anthropogenic stress, especially that for various forms of organic enrichment. The latter includes many causes of environmental damage such as pulp and paper mill waste discharge, urban sewage effluent discharge, the later stages of oil pollution on beaches and the seafloor, and the disposal of organic dredged sediments (Gray and Elliott, 2009).

The Pearson-Rosenberg model and others shows the well-accepted features of anthropogenically-stressed benthic infaunal communities, which are well known (**Table 1**) (Elliott and Quintino, 2007; Gray and Elliott, 2009). They are characterized by small organisms, high abundances of few species, low diversity and low individual biomass organisms with the potential to produce high biomasses. In addition, they have a high turnover and biological productivity (as shown by an increase in the production to average biomass ratio, P/B) and a dominance by oligochaetes and oligochaetiform polychaetes. This is summarized in ecological terms by the increase in r-strategists and the replacement of k-strategists, that is, the dominance of fast breeding, poor competitors over the long-lived, good competitors in ecological terms. Furthermore, given the inputs of organic matter (from sewage, etc.) and/or fine sediment, which also accumulates organic matter, in areas of low hydrodynamic energy such as estuaries, there is dominance by detritus and deposit feeding organisms in which food rather than space is the limiting factor (McLusky and Elliott, 2004). The assemblage is tolerant of adverse environmental conditions such as low oxygen and low and variable salinity,

☆*Change History:* March 2018, M. Elliott made minor changes to the text, figures and references.

 doi:10.1016/B978-0-12-409548-9.11054-1

Table 1 Conceptual basis and assumptions inherent in macrobenthic impact studies

(A) Natural state
(1) A natural macrobenthic assemblage either tends towards or is in an equilibrium state;
(2) Under nonimpacted conditions, there are well-defined relationships (which therefore may be modeled) between faunal and environmental (abiotic) variables;
(3) In approaching the normal equilibrium state, the biomass becomes dominated by a few species characterized by low abundance but large individual size and weight;
(4) Numerical dominance is of species with moderately small individuals, this produces among the species a more even distribution of abundance than biomass;
(5) The species are predominantly K-selected strategists;
(B) Moderate pollution
(6) With moderate pollution (stress), the larger (biomass) dominants are eliminated, thus producing a greater similarity in evenness in terms of abundance and biomass;
(7) Also, with moderate pollution, diversity may increase temporarily through the influx of transition species;
(C) Severe pollution
(8) Under severe pollution or disturbance, communities become numerically dominated by a few species with very small individuals;
(9) Those small individuals are often of opportunist, pollution-tolerant species which have r-selected strategies;
(10) Under severe pollution, any large species that remain will contribute proportionally more to the total biomass relative to their abundance than will the numerical dominants;
(11) Thus, under severe pollution, the biomass may be more evenly distributed among species than is abundance;
(12) However, under severe pollution, species with large individuals may be so rare as to be not taken with normal sampling;
(13) The change in assemblage structure with increasing disturbance is predictable, follows the conceptual models and is amenable to modeling and significance testing;
(D) Recovery
(14) Opportunists are inherently poor competitors and may thus be out-competed by transition species and K-strategists if conditions improve;

McManus and Pauly (1990) also consider that under normal conditions:
(1) The biomass-dominants will approach a state of equilibrium with available resources;
(2) The smaller species are out of equilibrium with available resources;
(3) The abundances of the smaller species are subject to more stochastically controlled variation than the larger species.

especially where many polluting discharges have a freshwater component. However, it is especially of note that these characteristics have been seen in many stressed ecosystems not only those found in the marine and estuarine environment (Odum, 1969, 1985) (**Table 2**).

The estuarine stress response has been observed in the macroflora (especially seaweeds) as well as the macrofaunal benthic community (e.g., Wilkinson *et al.*, 2007) and it relates to individual (physiological) stress (Smyth and Elliott, 2016). Polluted estuarine areas, especially those influenced by organic discharges, sewage run-off and industrial effluent which are noted for high levels of nutrients, become dominated by opportunistic green algae, occasionally forming large mats which then adversely affect the estuarine functioning such as the ability of wading birds to feed on the infaunal prey. Large concentrations of ephemeral green filamentous algae naturally occur in transitional waters bodies which usually have large nutrient inputs from riverine and anthropogenic sources and, depending on the relative riverine and tidal influences, retain these nutrients (Wilkinson *et al.*, 1995). Estuaries naturally show the transition from a highly diverse marine flora, with many red and brown macroalgae as well as green macroalgae in the outer regions, to an upper estuarine algal flora dominated by the Chlorophyceae. Under high organic and nutrient loading the features produce macroalgal mats, with the latter often displacing seagrass beds (de Jonge and Elliott, 2002).

Not only is such an estuarine stress detected at the community level but also at the physiological level of biological organization (Smyth and Elliott, 2016; Solan and Whiteley, 2016). Methods such as Scope-for-Growth (SFG) have long been used to good effect as an indication of anthropogenic stress in marine and estuarine areas (e.g., Widdows and Johnson, 1988; Mazik *et al.*, 2013). However, Navarro (1988) and Guerin and Stickle (1992) both indicate the way in which salinity stress, through natural freshwater inputs, reduces energetic budgets. Therefore, as SFG shows reduced physiological fitness due to salinity stress, it is difficult to use the technique for detecting and separating anthropogenic stress from natural stress.

Methods for detecting anthropogenic stress include those centred on the primary community structural variables (abundance, species richness, and biomass) and derived community structural variables (such as diversity indices, abundance (A/S) and biomass (B/A) ratios, evenness indices) (see Quintino *et al.*, 2006; McLusky and Elliott, 2004; Gray and Elliott, 2009; Borja *et al.*, 2011; Ducrotoy *et al.*, 2011 for references and details). They also include functional analyses such as those involving feeding guilds (as in the Infaunal Trophic Index, ITI, by Word, 1990) and their responses to elevated organic levels (as in the AZTI Marine Biotic Index, AMBI, by Borja *et al.*, 2000 and the Benthic Quality Index, BQI, by Rosenberg *et al.*, 2004, among others). For example, detritus and deposit-feeding dominance is reflected in any assessment of trophic analysis. There are many numerical and graphical methods which aim to detect and illustrate stress in benthic communities (e.g., see the 27 families of methods in Gray and Elliott, 2009, with references). For example: Species-Abundance-Biomass curves (Pearson & Rosenberg, 1978), Abundance-Biomass-Comparison curves (Warwick, 1986), the AMBI and diversity indices, etc. (e.g., Borja *et al.*, 2007; Rosenberg *et al.*, 2004). Despite this, these methods detect naturally as well as anthropogenically-stressed areas. Given that all impact assessments are required to

Table 2 Trends expected in stressed ecosystems: the estuarine features as applied to topics summarized by Odum (1985)

Feature	*Odum (1985)*	*Estuarine feature*	
Energetics	1. Community respiration increases	Yes, in general:	Higher respiration in larger populations of small organisms and organic rich sediments; possibly with osmoregulatory stress caused by salinity change
	2. P/R (production/respiration) becomes unbalanced	Unknown	Possibly due to higher respiration caused by salinity stress
	3. P/B and R/B (maintenance: biomass structure) increase	Yes, in general:	Higher P/B in smaller and shorter-lived organisms, for example, dominance by oligochaetes and small polychaetes; high turnover organisms
	4. Importance of auxiliary energy increases	Depends on meaning:	Increase in allochthonous energy input as well as relatively high autochothonous production
	5. Exported or unused primary production increases	Depends on meaning:	Export of material to adjacent sea areas but also import from catchment
Nutrient cycling	6. Nutrient turnover increases	Yes, but:	Because of physical characteristics - high nutrient inputs and use
	7. Horizontal transport increases and vertical cycling of nutrients decreases (cycling index decreases)	Partly the case:	Both horizontal and vertical cycling is high, depending on flushing characteristics and residence time; importance of material movement from pelagic to benthic system
	8. Nutrient loss increases	Yes, but:	because of the physical characteristics—high nutrient loss through flushing and export through predators
Community structure	9. Proportion of r-strategists increases	Yes:	High abundances of few, short-lived stress-tolerant species
	10. Size of organisms decreases	Yes:	High abundances of small organisms dominate in benthos; low megafaunal populations
	11. Lifespans of organisms decreases	Yes, in general:	On average, benthic and planktonic community composed of short-lived organisms; planktonic organisms adapted to prevent flushing of populations
	12. Food chains shorten because of reduced energy flow at higher trophic levels and/or greater sensitivity of predators to stress	Not necessarily:	Food chains can be very short (macrophytes-herbivorous ducks) but also very long because of the opportunistic nature of many predators; while marine predators (stenohaline marine fishes) may be reduced there are many other fish and bird predators
	13. Species diversity decreases and dominance increases; if original diversity is low, the reverse may occur; at the ecosystem level, redundancy of parallel processes theoretically declines	Yes (first part); unknown (second part):	Classic estuarine community in all components of few species; exacerbated with distance landward in the estuary; competition between species may be less than competition within species
General system-level trends	14. Ecosystem becomes more open (i.e., input and output environments become more important as internal cycling is reduced)	Not necessarily so;	Internal cycling is important even though nutrients and organic matter are delivered from external sources
	15. Autogenic successional trends reverse (succession reverts to earlier stages)	Unknown:	
	16. Efficiency of resource use decreases	Not necessarily so:	While there may be an excess of organic resources, leading to export, much is used within the system to support high predator populations
	17. Parasitism and other negative interactions increase, and mutualism and other positive interactions decrease	Not shown:	
	18. Functional properties (e.g., community metabolism) are more robust (homeostatic-resistant to stressors) than are species composition and other structural properties	Yes:	Ability of the system to withstand stressor-effects without adverse impacts

Box 1 The Estuarine Quality Paradox (modified and expanded from Elliott and Quintino, 2007)

This states that the dominant estuarine faunal and floral community is adapted to and reflects high spatial and temporal variability in naturally highly stressed areas such as estuaries and other transitional water bodies. Despite this, the community has ecological features very similar to those found in anthropogenically-stressed areas thus making it difficult to detect anthropogenically-induced stress in estuaries. Furthermore, as estuaries are naturally organically-rich areas then the biota has similarities to anthropogenically-organic rich areas such as due to urban and industrial organic effluents. Stress therefore becomes a subsidy which benefits those organisms able to tolerate the conditions. Because of this, there is the danger that any impact assessment methods which are based on those features and used to plan environmental improvements are inherently flawed.

detect or predict anthropogenic change against a background of natural variability, the so-called signal-to-noise ratio, then this becomes particularly difficult in highly variable estuaries.

As suggested above, the natural (non-anthropogenic) characteristics of estuaries, and many other transitional waters are highly influenced by freshwater inputs leading to a reduction in salinity, high organic production and organic inputs in the estuary and from the adjoining catchment and wetlands (the allochthonous and autochthonous production). This produces a low diversity of all components but often high abundances of those species which can tolerate the variable and (assumed to be) environmentally stressful conditions (e.g., McLusky and Elliott, 2004, 2007). The features of anthropogenic stress coincide with those for natural stress (**Tables 1** and **2**) (from Elliott and Quintino, 2007). This makes it particularly difficult to determine the effects of human activities in estuaries and other highly variable but organically-enriched transitional waters: that is, the estuarine benthic communities have many of the same characteristics as areas suffering from human-induced stress. This is the so-called "*Estuarine Quality Paradox*." This is especially the case in the mid to upper estuarine regions which receive organic matter from natural autochthonous and allochthonous sources and where the Freshwater-Seawater Interface is highly variable in space and time and thus has naturally stressful conditions (Day *et al.*, 2012).

In ecological theory terms, in having the characteristics described above, it is questioned whether the estuarine community is a climax one or whether it is just held as a subclimax level, that is, whether it has reached a final, stable assemblage or is attempting to attain such a final state. The estuarine fauna and flora do not show recovery to maintain a full k-strategist complement, and large individuals (both fauna and flora) are not present, hence there is a naturally lower Biomass/Abundance ratio and higher Abundance/Species Richness ratio, and the trophic system is dominated by organic/detritus-responsive invertebrates and nutrient-reflective algae. This then suggests that the variable state is paradoxically stable.

The Estuarine Quality Paradox

These features, in particular the difficulty of separating natural and anthropogenic stress in estuaries, produced the concept of the "Estuarine Quality Paradox." Since being introduced in the paper by Elliott and Quintino (2007) and mentioned in Dauvin (2007), this concept has been cited and used on many occasions (e.g., Dauvin and Ruellet, 2009). It is defined in Box 1.

There are many systems globally which aim the assess changes to the natural status because of human activities and pressures (Borja *et al.*, 2016). The Estuarine Quality Paradox has repercussions for the implementation of all environmental management systems which all rely on an ability to detect a change in estuarine flora and fauna from a defined reference condition. These systems include the European Union Water Framework Directive (Borja *et al.*, 2000, 2007, 2010b; Apitz *et al.*, 2006; Breine *et al.*, 2007; Coates *et al.*, 2007; Hering *et al.*, 2010), the National Land and Water Resources Audit in Australia (Heap *et al.*, 2001), the Clean Water Act in the US (USEPA, 2002) and the 1998 Water Act in South Africa (Adams *et al.*, 2002). In some of these quality assessments, which require environmental managers to determine a pass–fail criterion for good status, it is suggested that the Estuarine Quality Paradox will make it more difficult to determine that criterion and hence make any boundaries more uncertain (see Alvarez *et al.*, 2013; Basset *et al.*, 2013a).

In turn, the conclusion that a usual estuarine condition is similar to a stressed and polluted condition has repercussions for indications of what constitutes health and the criteria showing poor health, for example, see Tett *et al.* (2013). Hence the presence of symptoms such as opportunistic species should not necessarily be taken as a sign of anthropogenic stress. Similarly, this affects the ability to determine when recovery is taking place and whether this is also against the background of variability (Verdonschot *et al.*, 2013; Elliott *et al.*, 2007; Borja *et al.*, 2010a, 2011a,b; Duarte *et al.*, 2015).

As this paradox has now been accepted, as indicated by a large number of citations, by relying on the presumption that estuaries are stressed areas then there is the need to break out of this circle. This requires science and environmental management either to fully quantify and explain the natural variability and stress and subtract this from measures of the anthropogenic stress or alternatively by having an alternate set of methods which can detect anthropogenic stress against a background of natural stress. This may involve relying not on the structure of the system (the features at one time such as species richness or diversity) but to consider whether an estuary is still functioning as an estuary. Hence, we recommend that assessments should use functional symptoms as well as structural ones (see also de Jonge *et al.*, 2006; Hooper *et al.*, 2005). This agrees with Odum (1985), who

suggested that functional properties may be more robust than structural ones (de Jonge *et al.*, 2006). However, this contrasts with management recommendations and procedures such as the EU Water Framework Directive, which relies heavily on structural features such as taxonomic richness, diversity and abundance. Because of this, we may need different methods for the open coast than for the estuaries.

Estuarine Resilience, Environmental Homeostasis and the Stress-Subsidy Continuum

With regard to organismal physiological response and stress tolerance, Margalef (1981) emphasized that stress leads to organisms implementing homeostasis as a stress-reduction or stress-avoidance mechanism (see also Costanza *et al.*, 1992). Odum (1985) then regarded stress as a detrimental or disorganizing influence and McLusky and Elliott (2004) considered that with regard to polluting effects, stress reduces the fitness-for-survival. Individual organisms have an ability, termed homeostasis, to adapt to and withstand changes in environmental conditions that may be outside their optimal ranges. This is expanded here to the term *environmental homeostasis* to emphasize the ability of each level of biological organization, be it at the individual, population, community or ecosystem level, to withstand/tolerate/adjust/adapt to stressors. Thus, environmental homeostasis here is taken to be the ability of the estuarine organisms to achieve a stable state by compensating for the estuarine environmental physico-chemical variability; this may also be regarded as ecological robustness in which the resistance of the ecosystem to change and the resilience of the ecosystem to recover from change become emergent properties (Borja *et al.*, 2010a; Duarte *et al.*, 2015; Tett *et al.*, 2013). Hence it is concluded here that homeostasis can operate at any level of biological organization: individual (physiological) homeostasis, population homeostasis, community homeostasis and ecosystem homeostasis. It is hypothesized that in estuaries, the high natural variability and environmental homeostasis may increase resistance and resilience and an ability to withstand stress, both natural and anthropogenic. Therefore, the background of high estuarine variability (i.e., noise) increases the difficulty of detecting anthropogenic perturbation signals.

Given the above and using Odum's (1985) framework of symptoms in naturally-stressed areas, in estuaries salinity decrease is not a stress but a subsidy (Costanza *et al.*, 1992). Whitfield *et al.* (2012) show the dominant effect of salinity, through the so-called Remane diagram, and Basset *et al.* (2013b) show the effects of such environmental variables which create estuaries as the sites of multiple ecotones, each spreading across the range of environmental tolerances of organisms (Smyth and Elliott, 2016; Solan and Whiteley, 2016). If an estuary had high levels of natural stress, irrespective of anthropogenic stress, then there would be severe adverse consequences whereas there are not and therefore the system copes with that stress. Hence instead of considering estuaries as naturally stressed areas, they should be considered as areas with a subsidy rather than a stress, that is, as a perturbation with a positive effect on the system (Costanza *et al.*, 1992). The positive effect being the ability of estuarine organisms to tolerate the adverse and variable environmental conditions by capitalizing on the lack of inter-specific competition which leads to high population densities. Hence the natural estuarine system is maintained by providing a benefit for those species adapted to the inherently variable conditions. Therefore, those estuarine environmental managers who require to detect change need to determine whether natural stress is occurring, in particular whether salinity decrease is a stress or, for a brackish community, the stress would be in not experiencing a decrease in salinity. Thus reduced and highly varying salinity may only be a stress for a marine-dominated or marine-derived estuarine community. In essence, if a species is typically estuarine then the conditions are a subsidy, generating benefits whereas if a species in not adapted to estuarine conditions then these are a stressor.

Because of the many quality assessments linked to the EU Water Framework Directive (Borja *et al.*, 2010b; Hering *et al.*, 2010), there are numerous studies which define and quantify the way in which the estuaries respond to human activities. As shown by Elliott and Whitfield (2011), the estuarine functioning, such as the ability to support high predator populations of fishes and overwintering birds, does not rely on a high biodiversity per se. The biodiversity-ecosystem functioning (BEF) debate, that is, that a high diversity is required for successful functioning and vice versa, seems to be well developed and agreed for terrestrial, freshwater and microbial systems (e.g., Strong *et al.*, 2015; Loreau *et al.*, 2002, and papers therein), but little considered for estuarine and other transitional waters. Estuaries and other transitional waters thus become an anomaly in this in that they function successfully precisely because they have a low biodiversity. Therefore, analyses of ecosystem structure (which often rely on diversity measures of various types) related to human impacts are not sufficient and so ecosystem function has to be given more importance. This then has to be incorporated into the conceptual models, such as by Costanza and Mageau (1999), which aim to assess estuarine ecosystem vigor (based on function) together with organization (i.e., structure).

Concluding Comments

The acceptance of the Estuarine Quality Paradox depends on two primary concepts related to what constitutes stress and the ability to detect stress/show stress. Stress can be regarded as a perturbation with a negative effect on an area and thus resulting from a pressure which will reduce the ability to survive (and function) of a level of biological organization (cell, individual, population, community or ecosystem). In relation to human activities, a pressure is regarded as the mechanism of change in the natural status of a system which in turn adversely impacts the human system and welfare (Elliott *et al.*, 2017). As the stress can be caused by

natural or anthropogenic factors, managers therefore need to determine the presence and cause of that stress and separate the various influences. If, however, an area has such a high degree of natural variability that it is difficult detecting such a change (i.e., a low signal to noise ratio), then this makes it more difficult to detect anthropogenic change. Therefore, following the logic described here regarding natural estuarine variability stress as a subsidy then signs of that subsidy (natural estuarine biological characteristics) need to be separated from the assessment of stress per se (anthropogenic estuarine biological characteristics).

The Estuarine Quality Paradox provides an example of the subsidy-stress gradient (or more correctly a continuum) (Rapport *et al.*, 1985). Hence whereas in open marine systems it is easier to detect an anthropogenically-affected area along that continuum, this is more of a challenge in estuaries. The currently-used assessment methods indicate that detecting this difference between natural and anthropogenic stressors is difficult unless the anthropogenic stressor action is severe, such as adjacent to a sewage outfall, fish cage or oil spill area. Hence, we have to question whether we are using methods based on a false paradigm and how we can break out of this difficulty. For example, the continued use of ecological structural elements instead of a combination of structure and functioning is unlikely to indicate whether an estuarine area is anthropogenically-affected.

Secondly, as indicated here, we have to reassess whether we regard transitional waters (including estuaries) as naturally environmentally-stressed areas and we conclude that natural environmental variability only constitutes stress for those organisms not able to tolerate it. Thirdly, we have to further assess estuaries within the framework of the biodiversity-ecosystem functioning debate and consider that, in such transitional habitats, successful ecosystem functioning does not require a high biodiversity. Finally, there is the need to test the hypothesis that environmentally-variable areas can better withstand, absorb or mitigate anthropogenic perturbations (environmental homeostasis) which again makes more difficult the detection of anthropogenic signals against background (natural) environmental noise.

Acknowledgements

Thanks are due, for the financial support, to CESAM (UID/AMB/50017/2013), supported through national funds by FCT/MCTES and the co-funding by the FEDER (POCI-01-0145-FEDER-007638), within the PT2020 Partnership Agreement and Compete 2020.

See also: Aquatic Ecology: Estuarine Ecohydrology; Abundance Biomass Comparison Method; Eutrophication. Conservation Ecology: Biodiversity Indices. Conservation Ecology: Ecological Health Indicators; Ecosystem Health Indicators. Ecological Processes: Succession and Colonization. Ecosystems: Estuaries. Evolutionary Ecology: r-Strategists/K-Strategists. General Ecology: Biomass; Salinity; Biodiversity

References

Adams, J.B., Bate, G.C., Harrison, T.D., Huizinga, P., Taljaard, S., van Niekerk, L., Plumstead, E.E., Whitfield, A.K., Wooldridge, T.H., 2002. A method to assess the freshwater inflow requirements of estuaries and application to the Mtata estuary, South Africa. Estuaries 25, 1382–1393.

Alvarez, M.C., Franco, A., Pérez-Domínguez, R., Elliott, M., 2013. Sensitivity analysis to explore responsiveness and dynamic range of multimetric fish-based indices for assessing the ecological status of estuaries and lagoons. Hydrobiologia 704 (1), 347–362.

Apitz, S.E., Elliott, M., Fountain, M., Galloway, T.S., 2006. European environmental management: Moving to an ecosystem approach. Integrated Environmental Assessment and Management 2, 80–85.

Basset, A., Barbone, E., Borja, A., Elliott, M., Lasinio, G.J., Marques, J.-C., Mazik, K., Muxika, I., Neto, J.M., Reizopoulou, S., Rosati, I., Teixeira, H., 2013a. Natural variability and reference conditions: Setting type-specific classification boundaries for lagoon macroinvertebrates in the Mediterranean and Black Seas. Hydrobiologia 704 (1), 325–345.

Basset, A., Barbone, E., Elliott, M., Li, B.-L., Jorgensen, S.E., Lucena-Moya, P., Pardo, I., Mouillot, D., 2013b. A unifying approach to understanding transitional waters: Fundamental properties emerging from ecotone ecosystems. Estuarine, Coastal & Shelf Science 132, 5–16.

Borja, A., Basset, A., Bricker, S., Dauvin, J.-C., Elliott, M., Harrison, T., Marques, J.-C., Weisberg, S.B., West, R., 2011a. Chapter 1.08: Classifying ecological quality and integrity of estuaries. In: Simenstad, C., Yanagi, T. (Eds.), Classification of estuarine and nearshore coastal ecosystems,Wolanski, E., McLusky, D.S. (Eds.), Treatise on estuarine & coastal science, vol. 1. Amsterdam: Elsevier, pp. 125–162.

Borja, Á., Dauer, D.M., Elliott, M., Simenstad, C.A., 2010a. Medium- and long-term recovery of estuarine and coastal ecosystems: Patterns, rates and restoration effectiveness. Estuaries and Coasts 33, 1249–1260.

Borja, A., Barbone, E., Basset, A., Borgersen, G., Brkljacic, M., Elliott, M., Garmendia, J.M., Marques, J.C., Mazik, K., Muxika, I., Neto, J.M., Norling, K., Rodríguez, J.G., Rosati, I., Rygg, B., Teixeira, H., Trayanova, A., 2011b. Response of single benthic metrics and multi-metric methods to anthropogenic pressure gradients, in five distinct European coastal and transitional ecosystems. Marine Pollution Bulletin 62, 499–513.

Borja, A., Elliott, M., Andersen, J.H., Berg, T., Carstensen, J., Halpern, B.S., Heiskanen, A.-S., Korpinen, S., Lowndes, J.S.S., Martin, G., Rodriguez-Ezpeleta, N., 2016. Integrative assessment of marine systems: The ecosystem approach in practice. Frontiers in Marine Science 3.Article 20. https://doi.org/10.3389/fmars.2016.00020.

Borja, Á., Elliott, M., Carstensen, J., Heiskanen, A.-S., van de Bund, W., 2010b. Marine management—Towards an integrated implementation of the European marine strategy framework and the water framework directives. Marine Pollution Bulletin 60, 2175–2186.

Borja, Á., Franco, J., Pérez, V., 2000. A marine biotic index to establish the ecological quality of soft bottom benthos within European estuarine and coastal environments. Marine Pollution Bulletin 40, 1100–1114.

Borja, Á., Josefson, A.B., Miles, A., Muxika, I., Olsgard, F., Phillips, G., Rodríguez, J.G., Rygg, B., 2007. An approach to the intercalibration of benthic ecological status assessment in the North Atlantic ecoregion, according to the European water framework directive. Marine Pollution Bulletin 55, 42–52.

Breine, J., Maes, J., Quataert, P., Van den Bergh, E., Simoens, I., Van Thuyne, G., Belpaire, C., 2007. A fish-based assessment tool for the ecological quality of the brackish Schelde estuary in Flanders (Belgium). Hydrobiologia 575, 141–159.

Coates, S., Waugh, A., Anwar, A., Robson, M., 2007. Efficacy of a multimetric fish index as an analysis tool for the transitional fish component of the water framework directive. Marine Pollution Bulletin 55, 225–240.

Costanza, R., Mageau, M., 1999. What is a healthy ecosystem? Aquatic Ecology 33 (1), 105–115.
Costanza, R., Norton, B.G., Haskell, B.D., 1992. Ecosystem health: New goals for environmental management. Washington, DC, USA: Island Press.
Dauvin, J.-C., 2007. Paradox of estuarine quality: Benthic indicators and indices, consensus or debate for the future. Marine Pollution Bulletin 55, 271–281.
Dauvin, J.-C., Ruellet, T., 2009. The estuarine quality paradox: Is it possible to define an ecological quality status for specific modified and naturally stressed estuarine ecosystems? Marine Pollution Bulletin 59 (2009), 38–47.
Day Jr., J.W., Kemp, W.M., Yáñez-Arancibia, A., Crump, B.C. (Eds.), 2012. Estuarine ecology, 2nd edn, Wiley-Blackwell978-0-471-75567-8. 568 pp.
de Jonge, V.N., Elliott, M., 2002. Causes, historical development, effects and future challenges of a common environmental problem: Eutrophication. Hydrobiologia 475/476, 1–19.
de Jonge, V.N., Elliott, M., Brauer, V.S., 2006. Marine monitoring: Its shortcomings and mismatch with the EU water framework directive's objectives. Marine Pollution Bulletin 53, 5–19.
Duarte, C.M., Borja, A., Carstensen, J., Elliott, M., Krause-Jensen, D., Marbà, N., 2015. Paradigms in the recovery of estuarine and coastal ecosystems. Estuaries and Coasts 38 (4), 1202–1212. electronic 2013. https://doi.org/10.1007/s12237-013-9750-9.
Ducrotoy, J.P., Mazik, K., Elliott, M., 2011. Bio-sedimentary indicators for estuaries: A critical review. Paris: Union des océanographes de France978-2-9510625-2-8, pp. 1–77.
Elliott, M., Whitfield, A., 2011. Challenging paradigms in estuarine ecology and management. Estuarine, Coastal & Shelf Science 94, 306–314.
Elliott, M., Quintino, V., 2007. The estuarine quality paradox, environmental homeostasis and the difficulty of detecting anthropogenic stress in naturally stressed areas. Marine Pollution Bulletin 54, 640–645.
Elliott, M., Burdon, D., Hemingway, K.L., Apitz, S., 2007. Estuarine, coastal and marine ecosystem restoration: Confusing management and science – A revision of concepts. Estuarine, Coastal & Shelf Science 74, 349–366.
Elliott, M., Burdon, D., Atkins, J.P., Borja, A., Cormier, R., de Jonge, V.N., Turner, R.K., 2017. "And DPSIR begat DAPSI(W)R(M)!"—A unifying framework for marine environmental management. Marine Pollution Bulletin 118 (1–2), 27–40.
Elliott, M., McLusky, D.S., 2002. The need for definitions in understanding estuaries. Estuarine, Coastal & Shelf Science 55, 815–827.
Gray, J.S., Elliott, M., 2009. Ecology of marine sediments: Science to management. Oxford: OUP., 260 pp.
Guerin, J.L., Stickle, W.B., 1992. Effects of salinity gradients on the tolerance and bioenergetics of juvenile blue crabs (*Callinectes sapidus*) from waters of different environmental salinities. Marine Biology 114, 391–396.
Heap, A., Bryce, S., Ryan, D., Radke, L., Smith, C., Smith, R., Harris, P., Heggie, D., 2001. Australian estuaries & coastal waterways: A geoscience perspective for improved and integrated resource management. A report to the national land & water resources audit. Theme 7: Ecosystem health. Australian Geological Survey Organisation. Record 2001/07.
Hering, D., Borja, A., Carstensen, J., Carvalho, L., Elliott, M., Feld, C.K., Heiskanen, A.S., Johnson, R.K., Moe, J., Pont, D., Solheim, A.L., van de Bund, W., 2010. The European water framework directive at the age of 10: A critical review of the achievements with recommendations for the future. Science of the Total Environment 408 (19), 4007–4019.
Hooper, D.U., Chapin III, F.S., Ewel, J.J., Hector, A., Inchausti, P., Lavorel, S., Lawton, J.H., Lodge, D.M., Loreau, M., Naeem, S., Schmid, B., Setälä, H., Symstad, A.J., Vandermeer, J., Wardle, D.A., 2005. Effects of biodiversity on ecosystem functioning: A consensus of current knowledge. Ecological Monographs 75, 3–35.
Loreau, M., Naeem, S., Inchausti, P. (Eds.), 2002. Biodiversity and ecosystem functioning: Synthesis and perspectives. Oxford: Oxford University Press.
Margalef, R., 1981. Stress in ecosystems: A future approach. In: Barrett, G.W., Rosenberg, R. (Eds.), Stress on natural ecosystems. New York: John Wiley & Sons, pp. 281–289.
Mazik, K., Hitchman, N., Quintino, V., Taylor, C.J.L., Butterfield, J., Elliott, M., 2013. Sublethal effects of a chlorinated and heated effluent on the physiology of the mussel, *Mytilus edulis* L.: A reduction in fitness for survival? Marine Pollution Bulletin 77 (2013), 123–131.
McLusky, D.S., Elliott, M., 2004. The estuarine ecosystem; ecology, threats and management, 3rd edn. Oxford: Oxford University Press, 216 pp.
McLusky, D.S., Elliott, M., 2007. Transitional waters: A new approach, semantics or just muddying the waters? Estuarine, Coastal & Shelf Science 71, 359–363.
McManus, J.W., Pauly, D., 1990. Measuring ecological stress: Variations on a theme by R.M. Warwick. Marine Biology 106, 305–308.
Navarro, J.M., 1988. The effects of salinity on the physiological ecology of *Choromytilus chorus* (Molina, 1782) (Bivalvia: Mytilidae). Journal of Experimental Marine Biology and Ecology 122, 19–33.
Odum, E.P., 1969. The strategy of ecosystem development. Science 164, 262–270.
Odum, E.P., 1985. Trends expected in stressed ecosystems. Bioscience 35, 419–422.
Pearson, T.H., Rosenberg, R., 1978. Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. Oceanography and Marine Biology Annual Review 16, 229–311.
Quintino, V., Elliott, M., Rodrigues, A.M., 2006. The derivation, performance and role of univariate and multivariate indicators of benthic change: Case studies at differing spatial scales. Journal of Experimental Marine Biology and Ecology 330, 368–382.
Rapport, D.J., Regier, H.A., Hutchinson, T.C., 1985. Ecosystem behavior under stress. The American Naturalist 125 (5), 617–640.
Rhoads, D.C., Germano, J.D., 1982. Characterization of organism-sediment relations using sediment profile imaging: An efficient method of remote ecological monitoring of the seafloor (Remots™ system). Marine Ecology Progress Series 8, 115–128.
Rosenberg, R., Blomqvist, M., Nilsson, H.C., Cederwall, H., Dimming, A., 2004. Marine quality assessment by use of benthic species-abundance distributions: A proposed new protocol within the European Union water framework directive. Marine Pollution Bulletin 49, 728–739.
Smyth, K., Elliott, M., 2016. Chapter 10: Effects of changing salinity on the ecology of the marine environment. In: Solan, M., Whiteley, N. (Eds.), Stressors in the marine environment: Physiological and ecological responses: Societal implications. Oxford: OUP9780198718826, pp. 161–174. Hardback.
Smyth, K., Mazik, K., Elliott, M., 2014. Behavioural effects of hypersaline exposure on the lobster *Homarus gammarus* (L.) and the crab *Cancer pagurus* (L.). Journal of Experimental Marine Biology and Ecology 457, 208–214.
Solan, M., Whiteley, N. (Eds.), 2016. Stressors in the marine environment: Physiological and ecological responses: Societal implications. Oxford: OUP9780198718826. Hardback.
Strong, J.A., Andonegi, E., Bizsel, K.C., Danovaro, R., Elliott, M., Franco, A., Garces, E., Little, S., Mazik, K., Moncheva, S., Papadopoulou, N., Patrício, J., Queirós, A.M., Smith, C., Stefanova, K., Solaun, O., 2015. Marine biodiversity and ecosystem function relationships: The potential for practical monitoring applications. Estuarine, Coastal and Shelf Science 161, 46–64.
Tett, P., Gowen, R., Painting, S., Elliott, M., Forster, R., Mills, D., Bresnan, E., Capuzzo, E., Fernandes, T., Foden, J., Geider, R., Gilpin, L., Huxham, M., McQuatters-Gollop, A., Malcolm, S., Saux-Picart, S., Platt, T., Racault, M.-F., Sathyendranath, S., Molen, J.v.d., Wilkinson, M., 2013. Framework for understanding marine ecosystem health. Marine Ecology Progress Series 494, 1–27. + suppl. material.
United States Environmental Protection Agency (USEPA), 2002. Federal water pollution control act (as amended through P.L. 107–303, November 27, 2002). (33, U.S.C. 1251 et seq.), 230 pp. http://www.epa.gov/region5/water/cwa.htm.
Verdonschot, P.F.M., Spears, B., Feld, C.K., Brucet, S., Keizer-Vlek, H., Gunn, I., May, L., Meis, S., Borja, A., Elliott, M., Kernan, M., Johnson, R., 2013. A comparative review of recovery processes in rivers, lakes, estuarine and coastal waters. Hydrobiologia 704 (1), 453–474.
Warwick, R.M., 1986. A new method for detecting pollution effects on marine macrobenthic communities. Marine Biology 92, 557–562.
Whitfield, A.K., Elliott, M., Basset, A., Blaber, S.J.M., West, R.J., 2012. Paradigms in estuarine ecology—The Remane diagram with a suggested revised model for estuaries: A review. Estuarine, Coastal & Shelf Science 97, 78–90.
Widdows, J., Johnson, D., 1988. Physiological energetics of *Mytilus edulis*: Scope for growth. Marine Ecology – Progress Series 46, 113–121.

Wilkinson, M., Telfer, T.C., Grundy, S., 1995. Geographical variations in the distribution of macroalgae in estuaries. Netherlands Journal of Aquatic Ecology 29, 359–368.
Wilkinson, M., Wood, P., Wells, E., Scanlan, C., 2007. Using attached macroalgae to assess ecological status of British estuaries for the water framework directive. Marine Pollution Bulletin 55, 136–150.
Word, J.Q., 1990. The infaunal trophic index: A functional approach to benthic community analyses. (PhD thesis).University of Washington. 297 pp.

Further Reading

Borja, A., Elliott, M., 2007. What does 'good ecological potential' mean, within the European water framework directive? Marine Pollution Bulletin 54, 1559–1564.
Rhoads, D.C., Germano, J.D., 1986. Interpreting long-term changes in benthic community structure: A new protocol. Hydrobiologia 142, 291–308.
Weisberg, S.B., Ranasinghe, J.A., Dauer, D.M., Schaffner, L.C., Diaz, R.J., Frithsen, J.B., 1997. An estuarine benthic index of biotic integrity (B-IBI) for Chesapeake Bay. Estuaries 20, 149–158.

Eutrophication

Daniel A Lemley and Janine B Adams, Nelson Mandela University, Port Elizabeth, South Africa

Definition

Concern over eutrophication is a relatively recent development in the scientific literature, with the earliest recollection dating back to the 1950s. Furthermore, a working definition of what this phenomenon entails was only provided in the mid-1990s, where Nixon (1995) described it as "*an increase in the rate of supply of organic matter to an ecosystem.*" This was a crucial step as it recognized eutrophication as a process rather than confusing it with cause or consequence (i.e., a trophic state). Despite this, however, the definition leaves considerable room for interpretation, particularly from a management standpoint. With this in mind, subsequent reworkings of what constitutes eutrophication were required to meet both scientific and legal requirements. As a result, eutrophication has been defined as "*the enrichment of water by nutrients causing an accelerated growth of algae and higher forms of plant life to produce an undesirable disturbance to the balance of organisms present in the water and to the quality of the water concerned*" (OSPAR, 2003). The key feature of this definition is founded in the prerequisite for deleterious consequences to occur as an explicit response to anthropogenic nutrient loading for it to be considered eutrophication. This is an important distinction to make, as symptoms generally associated with eutrophication can also occur in pristine ecosystems, with little or no human impacts, due to natural features such as increased water temperature and residence times (Human et al., 2018).

Coastal Eutrophication

The "*coastal eutrophication problem*" was initially brought to light in an important study by Cloern (2001), where three evolving conceptual models were put forward, encompassing the early, contemporary and future vision for addressing this issue. The early model (Phase I), with its foundations in the field of limnology, was based on a signal–response relationship, whereby phytoplankton biomass was assumed to respond proportionately to changes in nutrient loading with subsequent increases in decomposition/organic matter loading and oxygen depletion. However, the difference in the responses of freshwater and coastal-estuarine environments to nutrient enrichment necessitated the inclusion of three fundamental advancements into a more robust contemporary conceptual model (Phase II). These included the recognition of (1) a suite of direct and indirect responses to anthropogenic nutrient enrichment, (2) interacting system-specific features that act as a filter to modulate these responses (i.e., each coastal system has varying sensitivity to nutrient enrichment), and (3) the potential for successful reversal of direct/indirect responses through implementation of appropriate management actions. Despite being an advancement, Cloern (2001) acknowledged the "*narrow and limited*" nature of this contemporary model as it lacks a broad ecosystem-perspective that compensates for the complex interactions that exist between nutrient enrichment and other stressors (e.g., toxic contaminants, freshwater manipulation and climate change). As such, the final model (Phase III) suggests an intersystem comparative approach aimed at garnering a mechanistic understanding of eutrophication, with the ultimate objective of producing a set of tools that can support effective management of these ecosystems.

Anthropogenic manipulation of freshwater affects the physical and biogeochemical balance of estuaries by altering the input, transport, and assimilation of water (i.e., quantity and quality), inorganic nutrients, dissolved and particulate organic matter, toxic metals and organopollutants. The ecological significance of altering the quantity and quality of freshwater inputs into estuaries include the risk of mouth closure, resistance to scouring events, accelerated rates of sedimentation, a shift to or loss of marine-dominated systems, restructuring of food webs and an increased incidence of eutrophication (Paerl et al., 2014; Cloern et al., 2016). With a particular emphasis on eutrophication, the natural progression of symptoms generally entails an initial surge in microalgal, macroalgal and/or epiphyte biomass related to anthropogenic nutrient loading (i.e., "primary symptoms"), followed by more severe and developed impacts such as a loss of submerged aquatic vegetation, oxygen depletion, proliferations of harmful algal blooms (HAB), imbalanced food webs, reduced biodiversity, altered biogeochemical cycling, fish-kills and the formation of "dead zones" (i.e., "secondary symptoms") (Bricker et al., 2003; Conley et al., 2009; Ferreira et al., 2011).

Estuarine and coastal habitats are important for ecological sustainability, water quality and nutrient cycling, as well as recreation, making their susceptibility to climate change and eutrophication a multidisciplinary, modern problem. Embedded in the definition

of eutrophication, the intensification of anthropogenic nutrient loading—particularly nitrogen (N) and phosphorus (P)—to aquatic ecosystems is regarded as a key culprit responsible for facilitating eutrophic conditions. Management usually aims to mitigate those factors which might exacerbate or disrupt these processes, such as controlling nutrient input or maintaining hydrological integrity. However, whilst P reductions have served as the core focus of management interventions aimed at preventing eutrophic symptoms globally, particularly in freshwater ecosystems, the recent proliferation of non-N_2 fixing HAB taxa suggests that elevated N loads are equally important to address, that is, the "N problem" (Paerl and Scott, 2010). By only controlling P inputs to freshwaters the uptake of N by algae is reduced, thus facilitating increased transport of N to downstream coastal waters (i.e., generally N-limited environments), potentially resulting in exacerbated eutrophic symptoms (Conley et al., 2009; Glibert, 2017). As such, it is imperative that nutrient reduction strategies, aimed at thwarting estuarine eutrophication, consider both N and P. However, because nutrients alone are not an obligatory indicator of eutrophication (i.e., manifesting itself in various forms), this phenomenon has proven to be a far subtler issue than originally anticipated. Therefore, frameworks and accompanying methodologies should adopt a multimetric and adaptive approach, incorporating both pressure and response variables, to obtain robust assessments of eutrophic condition in coastal environments.

Coastal Eutrophication: "Where, What, When"

The "*Where*"

The extent of eutrophication is widespread globally due to the synergistic relationship between ever-increasing human populations and the concomitant increase in anthropogenic nutrient loading to aquatic ecosystems. As such, coastal eutrophication issues have been documented worldwide, including the United States (US), Europe, Australasia, Latin America, Asia and Africa (Selman et al., 2008). For example, Bricker et al. (2008) indicated that 65% of estuaries in the US exhibit moderate to high eutrophication problems. Similarly, a study by Andersen et al. (2011) classified 93% of the assessment units (i.e., coastal and open water areas) within the Baltic Sea—one of the largest brackish-water basins in the world—as being "affected by eutrophication". Recent technological advances related to the identification and reporting of eutrophic conditions has contributed to the steadily increasing awareness of eutrophication. However, in regions such as Asia, Latin America, Africa and the Caribbean, a lack of explicit and systematic monitoring of coastal eutrophication, means that the true global extent and prevalence of this phenomenon is almost certainly underrepresented by the 415 affected areas identified by Selman et al. (2008). Socio-economic and political landscapes further compound the likelihood of eutrophication in these regions, where accelerated population growth coupled with rapid economic development (e.g., China) or, alternatively, limited resources (e.g., Africa) increase the amount of nutrients entering coastal waters. This is evidenced by studies in China (Strokal et al., 2014), Africa (Yasin et al., 2010) and South America (van der Struijk and Kroeze, 2010) where model results indicated a significant increase in total nutrient export by rivers for the period 1970–2000.

The "What and When"

Over recent decades the frequency, magnitude and extent of HABs have increased globally in estuarine and coastal waters (Anderson et al., 2002; Glibert et al., 2008; Glibert, 2017). Embedded in the definition of eutrophication, anthropogenic nutrient loading (N and P) is the key culprit responsible for the increased incidence of HABs. The primary concerns arising from such HAB events is founded in the array of possible consequences, including direct toxic effects on higher trophic levels (e.g., shellfish poisoning and bioaccumulation), bottom-water oxygen depletion related to bloom decay processes (i.e., decomposition and microbial respiration), mechanical interference and suffocation of faunal communities (e.g., mucilage production), and habitat destruction through shading of submerged aquatic vegetation. Such impacts can have severe social (e.g., esthetic degradation and loss of subsistence resources), economic (e.g., loss of tourism and collapse of fisheries) and human health implications. Broadly speaking, Dinophyceae are the most common HAB-forming phytoplankton group in coastal waters. Perhaps one of the most relevant HAB species from a eutrophication perspective is the high-biomass, potentially toxic dinoflagellate *Prorocentrum minimum*, due to its pan-global distribution and known association with regions of high dissolved inorganic and organic N and P river exports (Glibert et al., 2008). Other dinoflagellate HAB species indicative of eutrophic conditions include: *Akashiwo sanguinea*, *Alexandrium catenella*, *A. minutum*, *A. tamarense*, *Gymnodinium catenatum*, *Heterocapsa rotundata*, *H. triquetra*, *Karenia brevis*, *K. mikimotoi*, *Karlodinium veneficum*, *Lingulodinium polyedrum* and *Pyrodinium bahamense* var. *compressum* (Burkholder et al., 2008 and references therein; Lemley et al., 2018). Additionally, the incidence of coastal HABs species belonging to other phytoplankton classes are also on the rise globally, such as Raphidophyceae (e.g., *Heterosigma akashiwo*, *Chattonella* spp.), Prymnesiophyceae (e.g., *Prymnesium parvum*, *Phaeocystis* spp.), Bacillariophyceae (e.g., *Pseudo-nitzschia* spp.) and Cyanophyceae (e.g., *Microcystis aeruginosa*).

Macroalgal proliferations are another common symptom of eutrophication, with the incidence of bloom conditions increasing globally in response to increased nutrient availability. At a glance, macroalgal blooms have similar consequences to those imposed by HABs (e.g., oxygen depletion, loss of submerged vegetation). However, the ability of macroalgal blooms to persist in the absence of a natural disturbance (e.g., flushing event, high turbulence) (Valiela et al., 1997) facilitates an additional suite of possible impacts, including the production of foul odors, clogging of waterways, and drastic restructuring of trophic pathways (i.e., collapse of microalgal and grazer communities). All three macroalgal clades (Chlorophyta, Rhodophyta and Phaeophyceae) possess bloom-forming species, yet the most commonly occurring macroalgal blooms are composed of Chlorophyta species, known as "green

tides." Blooms of species belonging to the genera *Ulva* and *Cladophora* are particularly widespread examples of this phenomenon, and have been documented in North America, Europe, South America, Africa, Asia and Australia (Teichberg et al., 2010; Thornber et al., 2017).

An important distinction to make regarding the interpretation of these two acute symptoms of eutrophication (i.e., HABs and macroalgal proliferations) is that both can occur under natural, oligotrophic conditions. Therefore, to be considered a symptom of eutrophication, it is essential that such events are (1) related to increased anthropogenic nutrient loading, (2) persistent (i.e., not stochastic) and recurrent (i.e., seasonal), and (3) induce deleterious consequences on the environment (e.g., oxygen depletion, loss of ecosystem structure and function). For example, with regards to the latter, a study by Diaz and Rosenberg (2008) reported that eutrophication-induced hypoxia (i.e., episodic, periodic, seasonal or persistent) is responsible for approximately half of the 400-plus known dead zones globally. Some of the potential consequences of dead zones include mass mortalities of benthic fauna, reduction in demersal fish predators, release of poisonous microbially generated hydrogen sulfide under anoxic conditions, and ultimately a loss of nursery and recruitment areas. This is of concern, as such occurrences are an indication that a system has reached a critical point of eutrophication where ecosystem resilience is severely impaired, or even lost.

Eutrophication Assessment Methods

Numerous mandates have been formulated globally to protect coastal ecosystems from continued degradation arising from human-related impacts. The United States (US) and European Union (EU) are at the forefront of designing and implementing such legislation aimed at addressing these issues. In the US, the Clean Water Act (PL 92-500, 1972), together with the Coastal Zone Management Act (PL 92-583, 1972) and Harmful Algal Bloom and Hypoxia Research and Control Act (PL 105-383, reauthorized in 2004), provides the core impetus for the development and application of eutrophication assessment efforts, such as the Assessment of Estuarine Trophic Status (ASSETS) and Environmental Protection Agency National Coastal Assessment (EPA NCA) methodologies. In the EU, the assessment of eutrophication in marine and coastal waters is encompassed under umbrella regulations including the Water Framework Directive (WFD; 2000/60/EC), Marine Strategy Framework Directive (MSFD; 2008/56/EC), Nitrates Directive (ND; 1991/676/EC) and Urban Wastewater Treatment Directive (UWWTD; 1991/271/EC). The MSFD was an important development in the EU, as it explicitly includes eutrophication as 1 of the 11 holistic quality descriptors that together enable the assessment of environmental status of marine waters. Like the US, a variety of assessment methodologies have been designed to fulfill the requirements of these directives. Some examples of these include the Trophic Index (TRIX), Transitional Water Quality Index (TWQI), Oslo Paris Convention for Protection of North Sea Comprehensive Procedure (OSPAR COMPP), French Research Institute for Exploration of the Sea (IFREMER), and Helsinki Convention Eutrophication Assessment Tool (HEAT).

Other countries that have implemented preventative legislative actions, include South Africa (National Water Act of 1998), Australia (Oceans Policy of 1998), China (Law on Prevention and Control of Water Pollution of 1984) and Canada (Oceans Act of 1997). However, the overarching objective of these policies is to promote sustainable development whilst protecting and/or restoring coastal ecosystems. As such, eutrophication is not explicitly incorporated as a prerequisite component which needs to be addressed; thus, resulting in a lack of regionally specific eutrophication assessment methodologies. This is generally overcome by either adopting existing methodologies (e.g., China; Xiao et al., 2007) or developing an assessment method outside the bounds of a specific legislative framework (e.g., South Africa; Lemley et al., 2015). In the case of the latter, this is a crucial step toward providing a baseline from which the eutrophic condition of coastal waters can be monitored—particularly in regions where such assessment frameworks are lacking.

Eutrophication assessment methodologies serve as the basis for routine monitoring and establishment of ecological objectives worldwide. As such, it is imperative that the process involved with the formulation, implementation and interpretation of these methodologies are continuously improved to ensure results that are useful for management. A common thread that exists between eutrophication assessment methods is the integration of biological (micro- and macroalgae, marine angiosperms, zooplankton, macroinvertebrates and fishes) and physico-chemical (nutrients and dissolved oxygen) indicators that provide information at a desired level of confidence. Furthermore, selected indicators should reflect a gradient in responses that correspond with the level of human-induced impacts (i.e., nutrient loads). Many assessment frameworks utilize similar indicators, yet discrepancies arise due to differences in (1) temporal and spatial scales of analysis, (2) the characteristics included in indicator metrics (e.g., concentration, frequency, spatial coverage), (3) the statistical measures used to assess each indicator parameter (e.g., percentile, mean, absence/presence), (4) the determination of reference conditions and (5) the manner in which multiple lines of evidence are combined into a final status rating (Devlin et al., 2011; Ferreira et al., 2011).

As such, the application of different frameworks to the same coastal water body can culminate in confounding assessment ratings. For example, a study by McLaughlin et al. (2014) assessed the eutrophic condition of estuaries in the Southern California Bight (US) using multiple assessment frameworks (i.e., ASSETS, IFREMER and UK-WFD), and highlighted the varying applicability of specific indicators and threshold values on an individual estuary level. Similar observations have been made in Europe, where Garmendia et al. (2012) indicated the suppressed relevance of benthos and macroalgae indicators in Spanish estuaries where hydromorphological conditions are not conducive to the development of these communities. These overall sentiments were echoed in a comparative study by Devlin et al. (2011), where five assessment methodologies (i.e., UK-WFD, OSPAR COMPP, TRIX, ASSETS and EPA NCA) were applied to two estuaries in the United Kingdom. Whilst the various approaches indicated similar overall ratings, the authors suggested (1) sufficient timeframes of analysis (i.e., annual data), (2) use of frequency, spatial extent and duration in

indicator metrics, (3) inclusion of "secondary symptoms" as indicators, and (4) a multicategory rating scale to ensure a more representative assessment (i.e., Phase II approach sensu, Cloern, 2001).

Management Insights From Selected Case Studies

The implementation of proactive integrated coastal management, including both regulatory and cooperative approaches, provides the foundation for successful rehabilitation of eutrophic systems. The Tampa Bay Estuary, US provides a good example of where such an approach has been successfully adopted (Lewis III et al., 1998; Greening and Janicki, 2006). In summary, the Tampa Bay Estuary began exhibiting eutrophic conditions by the late 1970s (i.e., phytoplankton/macroalgae blooms and seagrass losses of ~50%) due to accelerated human settlement and associated nutrient loading. The restoration of seagrass habitat became the key management goal, with ensuing targets and actions for nutrient load (particularly N) and chlorophyll-*a* (related to light availability) reductions adopted to achieve this. Management actions taken to reduce N loads, and subsequently chlorophyll-*a* levels, included the requirement for advanced wastewater treatment of discharges from municipal plants, stormwater regulations and upgrades, implementation of agricultural best management practices, and habitat restoration projects. The responses to these actions included significant reductions in N loads and chlorophyll-*a* concentrations, increased light availability, and a recovery of seagrass coverage by approximately 25%. The continued reversal of eutrophic conditions in the system is largely a result of the multientity (i.e., Tampa Bay Estuary Program), integrated catchment management process that facilitates and coordinates these efforts. This case study demonstrates an effective and cooperative approach, following a logical sequence of (1) identifying the problem, (2) setting system-specific quantitative management goals, (3) establishing the environmental requirements necessary to achieve these goals, (4) implementing pertinent management actions and plans, and (5) continuous monitoring to assess progress toward meeting established goals.

The Blackwater Estuary in Ireland provides a EU example of where management actions, driven by national regulations (e.g., Good Agricultural Practice for Protection of Waters S.I. No. 101 of 2009) and European Directives (i.e., WFD, MSFD, ND and UWWTD), have proved successful in improving deteriorating water quality and biological trends (Ní Longphuirt et al., 2015). Recent improvements in farming practices, due to fertilizer reduction measures, have culminated in significant reductions in nutrient loads to coastal waters. Continuous monitoring of the Blackwater Estuary has taken place to assess the efficacy of implemented measures in achieving water quality improvements. Over a 20-year monitoring period, results indicated a decrease in N (17%) and P (20%) river loads, together with downstream estuarine reductions in P concentrations and chlorophyll-*a*. An interesting observation emanating from this case study was the decoupling between N load reductions and downstream response times, that is, N concentrations increased over time in the estuary. This paradoxical occurrence in the estuary was attributed to internal nutrient pools of N released via sediment remineralization (i.e., "legacy nutrients") and the suppressed N assimilation by primary producers (i.e., lower chlorophyll-*a*). As such, this example highlights the importance of considering the response times of system-specific processes and the implications of imbalanced nutrient reductions when assessing the effectiveness of eutrophication mitigation measures.

Finally, a study by Iho et al. (2015) compared trends in water protection and water policies in the Chesapeake Bay and Baltic Sea. These two coastal systems share similar characteristics, being large bodies of shallow and brackish waters that have been intensively managed throughout the past 40 years due to the increased incidence of eutrophic conditions. Despite successful reductions in point source (PS) nutrient loads and region-specific success stories, management measures have largely been unable to achieve desired water quality targets in both. At the forefront of the difficulties encountered from a policy and program implementation perspective is the multitude of stakeholders and political jurisdictions involved in coordinating the protection of Chesapeake Bay (US EPA and six states) and the Baltic Sea (EU and nine littoral countries), each with their own environmental laws and policies. Each of these systems are managed under a specific overarching regulatory framework, which set clear objectives for protection. For example, the Clean Water Act (CWA) makes provisions for a Total Maximum Daily Load (TMDL) to Chesapeake Bay, while the Helsinki Convention—facilitated by the Helsinki Commission (HELCOM)—aims to restore good ecological condition in the Baltic Sea using the Baltic Sea Action Plan (BSAP). In the case of Chesapeake Bay, the primary reason for shortfalls in achieving the desired water quality improvements is largely due to significant and unregulated discharges from nonpoint sources (NPS). Whilst the CWA has successfully enforced PS control via a permitting process (i.e., compliance with water-quality-based effluent limits), the same federal regulations do not apply to NPS, instead shifting responsibility to the states thus culminating in a mosaic of state and local initiatives which adopt a voluntary compliance strategy toward NPS control (i.e., nonbinding). A similar scenario exists in the Baltic Sea, where HELCOM have provided country- and pollution-specific (N and P) recommendations for pollution abatement through the BSAP; however, these limits are nonbinding as there are no means of enforcement due to the multinational political climate. Therefore, a key obstacle hindering the successful reversal of eutrophication in the Baltic Sea is founded in the difficulty of integrating the various EU directives into the national legislation of HELCOM contracting parties. Ultimately, this study highlighted that an "ideal policy," in terms of efficiently controlling pollutants, is achieved when (1) mitigation costs are equal across dischargers, and (2) mitigation costs equal the benefits obtained from improving environmental health (Iho et al., 2015). However, because the benefits of such actions are often unknown and financial/institutional constraints are commonplace globally, it is necessary for policies to be cost-effective and to adopt innovative ways (e.g., performance-based instruments) of enticing a positive attitude toward managing eutrophication.

References

Anderson DM, Glibert PM, and Burkholder JM (2002) Harmful algal blooms and eutrophication: Nutrient sources, composition, and consequences. *Estuaries* 25: 704–726.

Andersen JH, Axe P, Backer H, Carstensen J, Claussen U, Fleming-Lehtinen V, Järvinen M, Kaartokallio H, Knuuttila S, Korpinen S, Kubiliute A, Laamanen M, Lysiak-Pastuszak E, Martin G, Murray C, Møhlenberg F, Nausch G, Norkko A, and Villnäs A (2011) Getting the measure of eutrophication in the Baltic Sea: Towards improved assessment principles and methods. *Biogeochemistry* 106: 137–156.

Bricker SB, Ferreira JG, and Simas T (2003) An integrated methodology for assessment of estuarine trophic status. *Ecological Modelling* 169: 39–60.

Bricker SB, Longstaff B, Dennison W, Jones A, Boicourt K, Wicks C, and Woerner J (2008) Effects of nutrient enrichment in the nation's estuaries: A decade of change. *Harmful Algae* 8: 21–32.

Burkholder JM, Glibert PM, and Skelton HM (2008) Mixotrophy, a major mode of nutrition for harmful algal species in eutrophic waters. *Harmful Algae* 8: 77–93.

Cloern JE (2001) Our evolving conceptual model of the coastal eutrophication problem. *Marine Ecology Progress Series* 210: 223–253.

Cloern JE, Abreu PC, Carstensen J, Chauvaud L, Elmgren R, Grall J, Greening H, Johansson JOR, Kahru M, Sherwood ET, Xu J, and Yin K (2016) Human activities and climate variability drive fast-paced change across the world's estuarine-coastal ecosystems. *Global Change Biology* 22: 513–529.

Conley DJ, Paerl HW, Howarth RW, Boesch DF, Seitzinger SP, Havens KE, Lancelot C, and Likens GE (2009) Controlling eutrophication: Nitrogen and phosphorus. *Science* 323: 1014–1015.

Devlin M, Bricker S, and Painting S (2011) Comparison of five methods for assessing impacts of nutrient enrichment using estuarine case studies. *Biogeochemistry* 106: 177–205.

Diaz RJ and Rosenberg R (2008) Spreading dead zones and consequences for marine ecosystems. *Science* 321: 926–929.

Ferreira JG, Andersen JH, Borja A, Bricker SB, Camp J, da Silva MC, Garcés E, Heiskanen A, Humborg C, Ignatiades L, Lancelot C, Menesguen A, Tett P, Hoepffner N, and Claussen U (2011) Overview of eutrophication indicators to assess environmental status within the European Marine Strategy Framework Directive. *Estuarine, Coastal and Shelf Science* 93: 117–131.

Garmendia M, Bricker S, Revilla M, Borja Á, Franco J, Bald J, and Valencia V (2012) Eutrophication assessment in Basque estuaries: Comparing a North American and a European method. *Estuaries and Coasts* 35: 991–1006.

Glibert PM (2017) Eutrophication, harmful algae and biodiversity—Challenging paradigms in a world of complex nutrient changes. *Marine Pollution Bulletin* 124: 591–606.

Glibert PM, Mayorga E, and Seitzinger S (2008) *Prorocentrum minimum* tracks anthropogenic nitrogen and phosphorus inputs on a global basis: Application of spatially explicit nutrient export models. *Harmful Algae* 8: 33–38.

Greening H and Janicki A (2006) Toward reversal of eutrophic conditions in a subtropical estuary: Water quality and seagrass response to nitrogen loading reductions in Tampa Bay, Florida, USA. *Environmental Management* 38: 163–178.

Human LRD, Magoro ML, Dalu T, Perissinotto R, Whitfield AK, Adams JB, Deyzel SHP, and Rishworth GM (2018) Natural nutrient enrichment and algal responses in near pristine micro-estuaries and micro-outlets. *Science of the Total Environment* 624: 945–954.

Iho A, Ribaudo M, and Hyytiäinen K (2015) Water protection in the Baltic Sea and the Chesapeake Bay: Institutions, policies and efficiency. *Marine Pollution Bulletin* 93: 81–93.

Lemley DA, Adams JB, Taljaard S, and Strydom NA (2015) Towards the classification of eutrophic condition in estuaries. *Estuarine, Coastal and Shelf Science* 164: 221–232.

Lemley DA, Adams JB, and Rishworth GM (2018) Unwinding a tangled web: A fine-scale approach towards understanding the drivers of harmful algal bloom species in a eutrophic South African estuary. *Estuaries and Coasts*. https://doi.org/10.1007/s12237-018-0380-0.

Lewis RR III, Clark PA, Fehring WK, Greening HS, Johansson RO, and Paul RT (1998) The rehabilitation of the Tampa Bay Estuary, Florida, USA, as an example of successful integrated coastal management. *Marine Pollution Bulletin* 37: 468–473.

McLaughlin K, Sutula M, Busse L, Anderson S, Crooks J, Dagit R, Gibson D, Johnston K, and Stratton L (2014) A regional survey of the extent and magnitude of eutrophication in Mediterranean estuaries of southern California, USA. *Estuaries and Coasts* 37: 259–278.

Ní Longphuirt S, O'Boyle S, and Stengel DB (2015) Environmental response of an Irish estuary to changing land management practices. *Science of the Total Environment* 521–522: 388–399.

Nixon SW (1995) Coastal marine eutrophication: A definition, social causes, and future concerns. *Ophelia* 41: 199–219.

OSPAR (2003) In: *Strategies of the OSPAR commission for the protection of the marine environment of the north-east Atlantic (reference number: 2003e21), OSPAR Convention for the Protection of the Marine Environment of the Northeast Atlantic: Ministerial Meeting of the OSPAR Commission, Bremen, 25 June 2003, vol. Annex 31 (Ref. B-4.2)*.

Paerl HW and Scott JT (2010) Throwing fuel on the fire: Synergistic effects of excessive nitrogen inputs and global warming on harmful algal blooms. *Environmental Science & Technology* 44: 7756–7758.

Paerl HW, Hall NS, Peierls BL, and Rossignol KL (2014) Evolving paradigms and challenges in estuarine and coastal eutrophication dynamics in a culturally and climatically stressed world. *Estuaries and Coasts* 37: 243–258.

van der Struijk LP and Kroeze C (2010) Future trends in nutrient export to the coastal waters of South America: Implications for occurrence of eutrophication. *Global Biogeochemical Cycles* 24: GB0A09.

Selman M, Greenhalgh S, Diaz R, and Sugg Z (2008) Eutrophication and hypoxia in coastal areas: A global assessment of the state of knowledge. In: *WRI Policy Note 1, Water Quality: Eutrophication and hypoxia*. Washington, DC: World Resources Institute.

Strokal M, Yang H, Zhang Y, Kroeze C, Li L, Luan S, Wang H, Yang S, and Zhang Y (2014) Increasing eutrophication in coastal seas of China from 1970 to 2050. *Marine Pollution Bulletin* 85: 123–140.

Teichberg M, Fox SE, Olsen YS, Valiela I, Martinetto P, Iribarne O, Muto EY, Petti MAV, Corbisier TN, Soto-Jiménez M, Páez-Osuna F, Castro P, Freitas H, Zitelli A, Cardinaletti M, and Tagliapietra D (2010) Eutrophication and macroalgal blooms in temperate and tropical coastal waters: nutrient enrichment experiments with *Ulva* spp.. *Global Change Biology* 16: 2624–2637.

Thornber CS, Guidone M, Deacutis C, Green L, Ramsay CN, and Palmisciano M (2017) Spatial and temporal variability in macroalgal blooms in a eutrophied coastal estuary. *Harmful Algae* 68: 82–96.

Valiela I, McClelland J, Hauxwell J, Behr PJ, Hersh D, and Foreman K (1997) Macroalgal blooms in shallow estuaries: Controls and ecophysiological and ecosystem consequences. *Limnology and Oceanography* 42: 1105–1118.

Xiao Y, Ferreira JG, Bricker SB, Nunes JP, Zhu M, and Zhang X (2007) Trophic assessment in Chinese coastal systems—Review of methods and application to the Changjiang (Yangtze) Estuary and Jiaozhou Bay. *Estuaries and Coasts* 30: 901–918.

Yasin JA, Kroeze C, and Mayorga E (2010) Nutrient export by rivers to the coastal waters of Africa: Past and future trends. *Global Biogeochemical Cycles* 24: GB0A07.

Freshwater Aquaculture

James H Tidwell and Leigh A Bright, Kentucky State University, Frankfort, KY, United States

Introduction

Aquaculture is the culture of aquatic plants and animals under controlled or semicontrolled conditions. In its simplest terms, aquaculture is underwater agriculture. While long ago man transitioned to agriculture for its land sourced foods, in the area of aquatic foods, mankind has largely remained at the hunter and gatherer stage until recent years. Aquaculture is now the world's fastest growing food producing sector. Aquaculture possesses inherent characteristics which allow it to be one of the most efficient and low environmental impact methods of producing high quality protein for the human population.

History

Most agree that aquaculture began in China 5000–7000 years ago. It is believed that early efforts involved the common carp (*Cyprinus carpio*) and evolved from catching wild fish, holding those fish in baskets until market, and then feeding them to grow them to larger sizes. Carp culture was spread from China to various parts of Southeast Asia by Chinese immigrants. The Romans reportedly introduced carp culture into Europe. Egyptian hieroglyphics appear to show pond culture and tilapia (*Oreochromis niloticus*) are identifiable in the paintings (Stickney and Treece, 2012).

In what is now the United States, native Hawaiians have been growing fish in ponds for over 700 years. They allow seawater and the juvenile fish to enter the ponds with the rising tide, and then block off the openings before the tide recedes. The trapped juveniles are then allowed to grow to harvest size. Trout were among the early fishes to be cultured in Europe (as early as the 14th century). Initially the European brown trout (*Salmo trutta*) though the rainbow trout (*Oncorhynchus mykiss*) later became more popular. The British introduced trout to some of their colonies in Asia and Africa. In the 1800s, interest in recreational fishing and the need for supplemental stocking led to the development of private and public hatcheries in Europe and North America. It was not until the mid-twentieth century that the art of aquaculture developed into what exists today as a complex multidisciplinary science (Stickney and Treece, 2012).

Importance and Current Status

Fish is a vital component of the human food supply and man's most important source of high quality animal protein. (As used here, the general term "fish" includes fish, mollusks, and crustaceans consumed by humans). It is estimated that in 213 fish provided more than 3.1 billion people world-wide with at least 20% of their average per capita animal protein intake (FAO, 2016). It is a particularly important protein source in regions where high-quality protein from terrestrial livestock is relatively scarce. For example, in 2013 fish supplied only 11.7% of animal protein consumed in developed countries, but 20% of animal protein in developing countries (FAO, 2016).

Consumption of food fish from all sources is increasing, having risen from 40 MMT in 1970 to 146 MMT in 2014 (FAO, 2016). Global per capita fish consumption has increased over the past four decades, rising from 9.0 kg/person in 1961 to an estimated

20 kg/person in 2014 (FAO, 2016). Based on projected increases in consumption rates alone (assuming no increase in the human population) it is estimated that the demand for seafood will increase by more than 10 million tonnes per year by 2020 (Diana, 2009). While increases in per capita consumption account for a portion of the increase in total demand, it is human population growth that is the main driving force for this steadily increasing demand for food fish.

In 2014, the total world supply of fish from all sources was about 167 MMT (FAO, 2016). Capture fisheries primarily from the ocean, produced about 93 MMT of which about 21 MMT was destined for nonfood uses, primarily as fish meal in animal feeds. The other 87% of total fishery production was for human food (FAO, 2016).

Fish is the only important human food source where a large portion is still gathered from the wild and we once thought that the oceans, which cover ¾ of the earth's surface, contained an unlimited source of seafood. However, while demand for fish as food increases >10 million tons each year, sustainable harvests of wild fish are not able to expand significantly. For marine capture fisheries, FAO reports that in 2013 only 10% of the stock groups were under exploited while 58% were fully exploited (FAO, 2016). Global marine capture fisheries production has been at best stagnant for over 25 years. Global marine capture fisheries increased only 10% between 1992 and 2014 (FAO, 2016). The maximum yield fisheries to be captured from the world's oceans have likely been reached. In fact, by some estimates, current ocean harvests may already be *greater* than levels considered sustainable (Coll et al., 2008).

Status of Aquaculture

As we look to the future we see the demand for food fish increases each year while the supply from wild harvest is not expected to increase. So where do we get our fish in the future? The only other source for food fish is aquaculture and global aquaculture growth has been extraordinary. In the 1970s aquaculture contributed less than 4% of total seafood production but by 2014 contributed more than 50%. Aquaculture has now passed capture fisheries as the leading source of food fish for the human population and that proportion will continue to increase each year from here forward.

These projections only bolster the case that a prudent development of aquaculture is essential. In 2008, total aquaculture production of food fish was 53 million tonnes (FAO, 2008). It is anticipated that to keep pace with demand, aquaculture production of food fish will need to increase to 102 million tonnes (>90% increase) by 2025 (FAO, 2016).

Centers of Production

So where is aquaculture production occurring? Currently, Asia dominates the industry. In 2009, Asia accounted for 89% of world aquaculture production by quantity and 79% by value (FAO, 2008). China alone produces more than 60% of the world's aquaculture by volume (FAO, 2016). Of the top ten countries in aquaculture production in 2014, only two (Chile and Norway) were not in the Asian region and they account for approximately 3% of world production. However, there are rapid increases in production occurring in some countries outside of Asia.

Differences From Terrestrial Agriculture

Terrestrial animal agriculture relies on only a few species. In cattle, milk and meat production both utilize one species. In pigs, all commercial production is based on one species. In poultry, we have hundreds of varieties of chickens but they are all actually one species, and then we also have the turkey. These animals are *all* warm-blooded and differ at only the genus or class level. However, in aquaculture we raise over 400 species (Duarte et al., 2009), all are cold blooded, and many differ at class or even phylum level.

Even more than terrestrial farm animals, aquatic animals are captives of their environment. All terrestrial livestock are homeotherms (warm-blooded). That means they are able to regulate their body temperatures to stay within the narrow range needed for proper function. However, aquacultured animals are primarily poikilothermic (cold-blooded), so their body temperature is basically the temperature of their environment.

Terrestrial animals live in a relatively high oxygen gaseous environment (>21% by weight) and use lungs for gas exchange. Aquatic animals have evolved diverse structures to utilize the oxygen which is dissolved in the water. However, oxygen is much scarcer in water (0.00001% by weight) making an adequate oxygen supply a bigger consideration for the aquatic animal. Also, to complicate matters even more, the effects of temperature on the fish and the availability of oxygen in the water operate in conflict. As water temperatures increases, the metabolism of the animal (and its oxygen demand) *increases*. However, as temperature increases the solubility of oxygen in the water *decreases*, meaning less is available just when the animal needs it most. Nature played a cruel joke on the aquaculturist (Timmons et al., 2002).

Differences in Freshwater and Marine

For terrestrial animals air is pretty much air but for aquatic animals freshwater and saltwater represent very different environments and physiological challenges. Virtually all aquatic animals have to work against their external environment to maintain their internal environment. Most fish require an internal osmotic concentration of 250–500 m Osmol/kg. However, freshwater is <0.1 m Osmol/kg and seawater is approximately 1000 m Osmol/kg (Evans and Claiborne, 2008). That means that in marine fish, water is constantly trying to leave the fish and in freshwater species, water is constantly trying to move into the fish. Some species, known as euryhaline species, have mechanisms which allow them to adapt to a range of salinities. Others are known as stenohaline are strictly confined to one environment or the other.

Classifications of Culture Species

One way of classifying fish is based on tier optimal water temperature. Fish are often characterized as coldwater, coolwater or warmwater species. Some descriptions add a fourth category of tropical species. The characteristics, requirements, and especially tolerances of the fish within these groups are largely controlled by enzyme functions and efficiencies. Many enzymes only operate within a limited temperature band, which in the case of a poikilothermic animal means that the animal itself also operates efficiently within a narrow temperature band (Somero and Hochachka, 1971), so it is important that the culture system provide that proper temperature.

Finfish and invertebrates whose thermal optimum for growth is below 20°C are classified as coldwater species. Important aquaculture species within this group include the Atlantic salmon (*Salmo salar*), the rainbow trout (*Oncorhynchus mykiss*), and the Pacific oyster (*Crassostrea gigas*). Their optimum temperature for growth is about 10°C–16°C. They require relatively high oxygen levels (>5 mg/L) and tolerate only low levels of ammonia (<0.0125 mg/L un-ionized).

Species whose optimum temperature is around 20°C are considered coolwater species. Currently there are fewer commercially important aquaculture species in this category. The striped bass (*Morone saxatilis*) and yellow perch (*Perca flavescens*) and European perch (*Perca flaviatilis*) are examples.

Many important aquaculture species are considered warmwater species, with an optimum temperature around 30°C. Among crustaceans, they would include the Pacific white shrimp (*Litopenaeus vannamei*), the tiger shrimp (*Penaeus mondon*), and the freshwater prawn (*Macrobrachium rosenbergii*). Among mollusks are the American oyster (*Crassostrea virginica*), Northern quahog (*Mercenaria mercenaria*), and blue mussel (*Mytilus edulis*). Among finfish we would include the common carp (*Cyprinus carpio*), channel catfish (*Ictalurus punctatus*), and sea bass (*Dicentrachus labrax*). Compared to coldwater species, warmwater species in general tend to have a greater tolerance for lower dissolved oxygen (DO) levels and a greater tolerance for ammonia concentrations.

As stated earlier some classifications add a fourth category. Tropical species would be those whose optimum temperature would be >30°C. You may also add a characteristic of having a minimal lethal temperature of ≤10°C–15°C. Examples would be tilapia, with an optimum temperature of 29°C–31°C (Popma and Masser, 1999). They also tend to be very tolerant of low oxygen concentrations and high concentrations of nitrogenous waste products.

Ecological Services Provided by Culture Systems

Aquatic animals are very much "captives" of their environment. In aquaculture, we do not so much manage the animals as much as we manage their environment. That is also largely the function of the different aquaculture systems—to manage the animals' environment. Not only must the environment be maintained to support life but in the case of aquaculture, it needs to be maintained in such a way as to support maximum growth rate, with maximum efficiency, and a minimum of waste. A few environmental variables are fundamental and a discussion of the ways that aquaculture systems control them is the unifying theme of this chapter, how the specific system being examined provides the cultured animal with the, (1) proper temperature for growth, (2) sufficient oxygen to breathe, (3) removes the inevitable waste products and in some cases, (4) provides some or all of the animals' food needs.

Types of Production Systems

There is a vast range of intensities of production in freshwater aquaculture systems. For reservoir ranching systems, there may be 10 kg of animals in a hectare (10 kg/ha) while in an intensive recirculating system we may be have 10 kg of animals in *one square meter* (or 100,000 kg/ha). Within this continuum of increasing intensity we have three major categories or classifications for aquaculture production systems. These groupings are primarily based on the amount of control or intervention the aquaculturist provides in terms of the three basic functions or ecosystem services, each system must provide (proper temperature, adequate oxygen, and waste removal). However, the demarcations between these systems are not always clear and distinct. There are even hybrids among and between these categories

Open Systems

Production systems within this category rely entirely on natural ecological processes to address the three major functions. These systems are normally natural bodies of water that are now being stocked for commercial production. Many of these systems could be considered stock enhancement rather than aquaculture. Biomass densities are usually low enough that natural processes can provide sufficient oxygen for the biomass being supported. The oxygen can be sourced from diffusion, photosynthesis by natural algal communities, or both. Waste products are also removed by natural processes within the systems, operating at natural rates. Bacterial breakdown of solid wastes is by heterotrophic bacteria and fungi. Nitrogenous waste products, such as ammonia excreted by the animal, are either flushed away or processed into less toxic forms by the chemoautotrophic components of the natural nitrogen cycle (nitrification) or assimilated by algae. Water temperatures in these systems are ambient. In open systems, site selection is the major control factor the producer has for all environmental services.

Cage culture is an open system production method. It basically represents a "fencing-off" of a portion of the natural aquatic habitat. Some cages are literally fenced compounds in shallow water. The bottom of the cage is the mud bottom of the bay or lake. In others, the cages have net bottoms and are suspended off of the bottom by flotation. These can be small cages (1–4 m^3) floating in shallow freshwater ponds. Larger cages used in marine environments are usually referred to as net pens.

Semiclosed Systems

Within this category we still rely largely on nature to provide the three basic ecological services of proper temperature, sufficient oxygen, and waste removal. However, within the semiclosed category the production units themselves are now largely manmade and includes ponds and raceways. Within the production units we now have the ability to add or remove water. Also, there is more management input in these systems and the first steps toward supplementing or enhancing natural processes.

In semiclosed systems water is taken from a natural source such as rainfall, springs, streams, or rivers. The water is then gravity flowed or pumped into specially designed and constructed production units. The water can be used once and discharged or constantly cleaned and re'oxygenated by natural processes. Compared to open systems, semiclosed systems have several advantages. One is much higher production rates, as much as 1000 times the productivity of an open system. This is due to the greater control and inputs into these systems and the fact that their physical parameters can be maximized for greater productivity.

Other advantages include easier and more efficient use of prepared feeds, control over water depth or water replacement, mechanical aeration can be practical and cost-effective, poaching and predation are more easily controlled, competitors and predators can be eliminated, water quality deterioration and diseases can be more easily detected and rectified, and there is some potential for temperature control. However, there are also negatives. Construction and equipment costs can be significant, there are more management demands monitoring and intervention, energy and feed inputs are higher, and a greater likelihood that water quality issues and diseases will occur.

Raceways are semiclosed systems and are basically large manmade earthen or concrete troughs. A typical length/width/depth ratio in linear raceways is 30:3:1 (Stickney, 2000). High quality water flows into and through the trough bringing in needed oxygen and flushing away wastes. Water sources are usually ground waters coming to the surface in the form of springs or surface water from snow melt or rain runoff from higher elevations. The water can often be reused several times as the water flows through multiple raceways in series

In raceways the oxygen is provided by the incoming water. It must come into the raceway saturated with oxygen. Wastes produced in these systems are passed on for processing further downstream in the receiving waters, or onsite in designed treatment units. Temperatures in raceway systems reflect their water source. Retention time is low so temperature changes little within the system.

Ponds are also examples of semiclosed production systems and there are several types. The simplest and easiest to construct is to build a watershed or impoundment pond. These are constructed by building a dam across a natural waterway to retain the rain runoff at a level set by the dam. The pond's shape is largely dictated by the land's topography.

A purpose built pond for aquaculture is usually a leveed pond. They commonly have a 2:1 length to width ratio. Ponds used in commercial aquaculture production vary widely in size. In ponds most of the oxygen budget is based on oxygen production by photosynthetic phytoplankton. In the past this was the limiting factor for production within these systems. Without supplemental feeding the natural carrying capacity of a pond is around 250–500 kg/ha. With supplemental feeding this can be increased to about 1500 kg/ha. However, at this biomass density and the accompanying feeding rate of 30–40 kg/ha/day, the chance of low oxygen periods during the night or early morning starts to become unacceptably high (Boyd, 1979). In most commercial scale pond production systems man has intervened by providing mechanical aeration. With this change, feed rates can be increased to about 100 kg/ha/day and production can be increased over three fold to >4500/kg/ha.

Ponds still rely on natural processes to remove waste products. Again, solid wastes are broken down primarily by heterotrophic bacteria and detritivores on the pond bottom. Ammonia (NH_4^+) excreted by fish or shrimp is directly assimilated by algae or converted to less toxic nitrite (NO_2^-) by *Nitrosamonas* bacteria then on to nitrate (NO_3^-) by *Nitrobacter* bacteria. The nitrate form can then be assimilated by algae. Since the advent of mechanical aeration, the efficiency of this nitrogen removal system is now the primary bottleneck in further pond production intensification. Water temperatures in these systems are basically ambient. They lag

behind, but reflect, the mean air temperature in the region. Near the equator there is little seasonal fluctuation. As you move further from the equator seasonal fluctuations become more pronounced and can either affect growth or even produce mortality. There has been some work in using waste heat from power plants and other industries to warm pond waters. However, fish production is usually a secondary consideration. The needs of the primary industry usually take priority and created problems during shut down for repairs or maintenance.

In ponds, a significant portion of the food for the culture organism can also be generated internally. If this is the primary food for the system it is said to be an extensive pond system. The natural carry capacity of an unfed pond is in the range of 250 kg/ha. This carrying capacity can be increased by adding nutrients to the system. In most extensive or low-input systems these nutrients are supplied in the form agricultural by-products, or animal (or human) waste. This is known as organic fertilization. If needed nutrients are supplied in their purely chemical form (often derived from petrochemicals) they are knows as inorganic fertilizers. Again there are positives and negatives to each.

Positives of organic fertilizers include low costs, slow release of nutrients, and sustainability aspects of reuse. Negatives include the need to handle bulk and sometimes wet (and heavy) materials for small amounts of nutrients. To be made available to the phytoplankton and food web these materials must first be decomposed by microbes, which can be fairly slow and this is an oxygen consuming process These products can also directly deteriorate water quality (by yielding ammonia etc. . .) if misapplied. Inorganic fertilizers have the positive aspects of acting quickly, without deteriorating water quality through added nitrogenous wastes. However, inorganics can be more expensive and can actually work "too well" if misapplied. A slight over application of phosphorous can cause of a rapid phytoplankton bloom which can die off just as rapidly. This can result in oxygen depletion as the phytoplankton decomposes.

Closed Systems

In closed systems you are reusing the same water within a manmade culture system. Also, there is human intervention of some type and at some level in *all* of the basic processes. The major advantage of closed systems is that they provide the operator complete control over all of the environmental variables in the culture system. The major *disadvantage* of closed systems is that the operator now has complete *responsibility* for all aspects of the animals' environment.

In closed systems water temperature can be maintained very near the optimum growing temperature for the cultured animal. This can have a tremendous positive impact on not only growth rate but also efficiency, both of which are highly important in these systems. Because of this temperature control, we can now raise tropical animals in temperate zones, if that is where the markets are. Waste heat from industrial processes can provide economic advantages if the schedules and proximities of the systems are compatible.

With closed systems water can be constantly disinfected with ultraviolet (UV) lights or ozone to crop down pathogenic organisms. Predators and poachers can be completely eliminated. External environmental events like floods or cold snaps are no longer a problem. Feed can be efficiently administered and consumption and conversion accurately monitored. Water supply volumes become less of a concern. However, for large systems their loss of >5% per day can still become substantial.

Recirculating aquaculture systems (RAS) are examples of closed systems and are also known as closed loop systems, recycle systems, and intensive recycle systems. As these names imply, as opposed to trout raceways which have a constant inflow of new water, these systems use the same water over and over. They do this by constantly adding air or oxygen to the water and removing the waste products produced by the fish. If aeration is used, production is limited to about 40 kg/m^3 (0.33 lb/gal). With the use of pure oxygen (oxygenation) production can be increased to approximately120 kg/m^3 (1.0 lb/gal) (Timmons et al., 2002). To remove waste products most systems rely on mechanical filters to remove solid wastes then nitrogenous wastes, such as ammonias are detoxified to nitrite then nitrate, using the same nitrifying bacteria discussed in ponds. However, these bacteria are now cultured at very high densities inside containers known as biofilters. The nitrifying bacteria in the biofilter need a surface to attach to some special materials, known as media, are packed or suspended inside the biofilter to provide surface area for the bacteria to grown on. However, it is also important for the biofilter media to have sufficient open areas for the water to flow through and wash over the bacteria so that they can "consume" the inorganic waste and excrete less toxic versions.

In recent years, another type of closed or recycle system has been developed. Instead of relying on chemoautotrophic nitrifying bacteria, which utilize inorganic compounds such as ammonia for energy, these systems are colonized with heterotrophic bacteria which consume the organic wastes. These bacteria are not confined in biofilters but live suspended in the culture vessel along with the animal being cultured. Once this bacterial population is established and stabilized, very high production rates can be achieved (>5 kg/m^3). However, these systems also have very high oxygen demand and relatively little research has been conducted on their complex microbial ecology. These systems can quickly remove nitrogenous waste products from the animals by directly converting it to bacterial biomass. In traditional RAS quick removal of solids is important to their function. In heterotrophic or biofloc systems these solids are largely retained in the culture tanks and became colonized with heterotrophic bacteria, fungi, and protozoans into suspended particles called bioflocs. The bioflocs also recycle the wastes as they can be grazed by the animals and directly consumed as high protein forage.

Many new approaches in recent years are blurring the lines between the different production systems and even major categories (open, semiclosed, and closed). They take aspects of different systems and combine them in new ways to overcome the

shortcomings of one, capitalize on the positives of another, or break a system into its functional components so they can be individually manipulated.

In one hybrid system, termed aquaponics, the basic components of a recirculating system are utilized. However, the biofilter has been replaced by plants which assimilate the nitrogenous waste products and then turn it into saleable plant products. Aquaponics is increasingly of interest in areas where water availability is limited. Much of their early development has occurred on small islands where the supply of fresh fish and freshwater, fresh produce are all limited. There is increasing interest in evaluating these systems in arid countries, such as those in the Middle East. Now there are also efforts to apply these technologies to generating fresh fish and vegetables in or near major urban areas. These are meant to address the phenomenon known as "food deserts." These are urban areas where city dwellers do not have ready access to healthy foods at reasonable prices (Ford and Dzewaltowski, 2008). This makes these populations, especially certain ethnic groups, even more susceptible to health problems such as Type II diabetes and obesity. Urban aquaponics might represent an "oasis" of healthy foods in or near the food deserts.

Sustainability Advantages of Aquaculture

While it has been popular among certain groups and in the popular press to criticize aquaculture, an objective evaluation shows that it is, and can continue to be, one of the most eco-friendly methods to produce high quality protein for human consumption. Fish are inherently more efficient than other farm animals. Much of this is based on the fact that fish are cold-blooded (poikilothermic) animals. This means that they do not expend any energy maintaining their internal body temperature. They also do not expand energy fighting gravity (also giving them less investment in skeleton). Aquatic animals also excrete waste products more efficiently than terrestrial animals. These add up to fish converting feeds to flesh much more efficiently than other animals. Better conversion efficiencies also mean less waste produced.

The Future and the Challenge

As we have seen, the demand for fish increases each year. To even maintain the current level of per capita consumption, the fish supply will have to almost *double* in the next 20 years. That translates into almost 40 million tonnes of additional supply per year and basically, all of it has to come from aquaculture. As Melba Reantso of FAO described it, "aquaculture is *now* known as the emerging new agriculture, the catalyst of the "blue revolution," the answer to the world's future fish supply, the fastest growing food producing sector, and the future of fisheries." Still, the task ahead is daunting. Aquaculture is expected to supply global seafood security, nutritional well-being, poverty reduction and economic development by meeting all of these demands, but also accomplishing this with a minimum impact on the environment and maximum benefit to society.

References

Boyd CE (1979) *Water quality in warmwater fish ponds*. Auburn, Alabama: University of Alabama Press.
Coll M, Libralato S, Tudela S, Palomera I, and Pranovi F (2008) Ecosystem overfishing in the ocean. *PLoS One* 3(12). e3881.
Diana JS (2009) Aquaculture production and biodiversity conservation. *Bioscience* 59(1): 27–38.
Duarte CM, Holmer M, Olsen Y, Soto D, Marbà N, Guiu J, Black K, and Karakassis I (2009) Will the oceans help feed humanity? *Bioscience* 59(11): 967–976.
Evans DH and Claiborne JB (2008) Osmotic and iconic regulation in fishes. In: Eraus DH (ed.) *Osmotic and ionic regulation: Cells and animals*, 1st edn., pp. 295–366. Boca Raton: CRC Press.
FAO (2008) *State of the world fisheries and aquaculture*. Rome: FAO.
FAO (2016) *State of the world fisheries and aquaculture*. Rome: FAO.
Ford PB and Dzewaltowski DA (2008) Disparities in obesity prevalence due to variation in the retail food environment: Three testable hypotheses. *Nutrition Reviews* 66(4): 216–228.
Popma T and Masser M (1999) Tilapia: Life History and Biology. In: *SRAC Publication No. 283*. Stoneville, Mississippi: Southern Regional Aquaculture Center.
Somero GN and Hochachka PW (1971) Biochemical adaptation to the environment. *American Zoologist* 11(1): 159–167.
Stickney RR (2000) *Encyclopedia of aquaculture*. New York: John Wiley and Sons, Inc.
Stickney RR and Treece GD (2012) History of aquaculture. In: Tidwell JH (ed.) *Aquaculture Production Systems*, 1st edn., pp. 15–50. Ames, Iowa: Wiley-Blackwell.
Timmons MB, Ebeling JM, Wheaton FW, Summerfelt ST, and Vinci BJ (2002) *Recirculating aquaculture systems*. Cayuga Aqua Ventures, Ithaca: Northeastern Regional Aquaculture Center.

Intertidal Zonation☆

Maya C Pfaff, Department of Environmental Affairs: Oceans and Coasts, Cape Town, South Africa
Ronel Nel, Nelson Mandela University, Port Elizabeth, South Africa

Glossary

Backwash The return flow of water down the beach face following a swash.
Biotic Of biological (and not physical) nature.
Biogeographic Referring to the distribution of species and ecosystems across large geographic scales and through geological time.
Desiccation Loss of water from organisms caused by exposure to air, which leads to physiological stress.
Intertidal zone The area on the coast that lies between the low-tide and the high-tide marks.
Littoral Coastal; but also with specific reference to the intertidal zone.
Macrofauna Benthic invertebrates that live on or in sediment and are >1 mm in length; this size-based grouping contains a diversity of taxonomic group.
Meiofauna Small benthic invertebrates that live in both marine and fresh water environments; this grouping is defined by size (45–500 μm) and contains a diversity of taxonomic groups.
Morphodynamic state A classification scheme for beaches based on sediment grain size, wave action, tidal amplitude, and resultant beach slope. Beaches are classified along a gradient from reflective to dissipative states.
Semidiurnal Having a recurrence period of half a day and a frequency of twice a day.
Swash The flow of water from a broken wave up the beach face.
Stress A situation that results in an organism's reduction of biological performance, such as growth, reproduction, or survival.
Suspension feeder An organism that feeds on particular organic matter suspended in the water column.
Tidal amplitude The difference in vertical height between high and low tide.

The Intertidal Zone

At the interface between land and sea exists a narrow band where tides and waves generate some of the most extreme environmental gradients on Earth. This is the intertidal zone, so called because it is the area between the low- and high-water marks. The intertidal zone is characterized by considerable fluctuations in moisture and temperature between high tide when it is submerged under the sea, and low tide when it is exposed to air and sun. It is also the place where ocean waves collide with the land, introducing an along-shore gradient of pounding forces and splash between wave-exposed headlands and wave-sheltered bays. Despite these extreme conditions, the intertidal zone features an abundance of organisms that are adapted for survival in various niches of this harsh environment. On rocky shores, high densities of sessile or sedentary organisms form characteristic horizontal bands that show remarkable similarities around the world (Fig. 1). This pronounced intertidal zonation has inspired generations of naturalists and scientists to decipher the ecological processes and mechanisms that produce such distinct biotic patterns. Rocky shores have consequently served a role as outdoor laboratories in which seminal ecological theories were forged and tested through a wealth of field experiments. Other intertidal habitats such as sandy beaches are also subjected to stratified environmental conditions and display biotic zonation patterns resembling those on rocky shores, although less obvious to the eye. Hence, intertidal zonation both of rocky shores and sandy beaches will be described in this article.

Rocky Shore Zonation

Rocky intertidal communities exhibit the most pronounced zonation patterns of all intertidal habitats. Successive horizontal bands of organisms are a unifying theme of rocky shores; however, various factors can modify these widespread patterns. The confinement of a species' distribution to a discrete level on the shore was originally attributed to its physiological limits, but several ground-breaking studies have since demonstrated that biological interactions can determine lower limits, and sometimes also upper limits, of biotic zones.

Environmental Gradients Imposed by Tides and Waves

Most shores experience semidiurnal tides, which correspond to two high tides and two low tides per day. Thus, twice a day conditions in the intertidal zone alternate between being more or less terrestrial to wholly marine, inducing a steep vertical

☆*Change History:* March 2018. MC Pfaff and R Nel updated the article. It has been completely re-written, and thus all sections have essentially been updated. This is an update of C. Robles, Intertidal Zonation, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 2003–2010.

Fig. 1 Examples of intertidal zonation on rocky shores around the world: Near the Cape of Good Hope in South Africa (*top left panel*; 34.2°S 18.5°E), the low shore is occupied by pink coralline algae and the territorial limpet *Scutellastra cochlear*, the mid shore by a mixture of algae and the introduced mussel *Mytilus galloprovincialis*, and the high shore by various barnacles; off Vancouver Island in Canada (*top right panel*; 48.9°N 125.3°W), the predatory starfish *Pisaster ochraceus* lives low on the shore preventing the downward expansion of mussels, and barnacles and algae inhabit the zones above; in northern Chile (*bottom left panel*; 29.0°S, 71.5°W), characteristic bands of algal growth are interspersed by a mid-shore belt where grazing limpets and bare rock dominate; and on the sub-Antarctic Marion island (*bottom right panel*; 46.8°S 37.7°E) encrusting coralline algae and the kelp *Durvillaea antarctica* dominate the lower part of the shore and cushions of green algae host a number of invertebrate species on the high shore. Photographs by M. Pfaff (South Africa, Canada, Marion Island) and M. Aguilera (Chile).

gradient of environmental factors (**Fig. 2**). Intertidal organisms need to be extremely well adapted and hardy to tolerate such contrasting environments. Species that live high up on the shore spend much longer periods emerged and exposed to air and sun than those lower down on the shore. They must be able to withstand desiccation, thermal stress from extremely high (or low) air temperatures, the damaging effects of UV radiation, and osmotic stresses due to changes in salinity through evaporation or precipitation. Abiotic stresses therefore increase the higher up on the shore an organism lives.

Many marine organisms draw vital resources from seawater. Algae, for instance, absorb nutrients directly from seawater, and many invertebrates feed on suspended particles in the water column. Organisms found on the high shore have less time available to access these resources because they spend more time emerged. Resource limitation therefore adds to the abiotic stresses incurred by the organisms high up on the shore. Growth rates, reproductive outputs, and body sizes of organisms are inversely related to physiological stress and resource limitation, and tend to increase toward the lower extent of the intertidal zone. Moreover, biotic cover, species richness, and biomass also increase down the vertical slope.

Another environmental gradient that fundamentally influences the ecology of intertidal shores is induced by the gradual decline of wave action between exposed headlands and sheltered bays. Wave splash ameliorates the physical stresses associated with emersion during low tide and thereby markedly alters intertidal conditions, especially high up on wave-exposed shores. Furthermore, waves continuously supply rocky shores with particulate food for suspension feeders and with planktonic larvae,

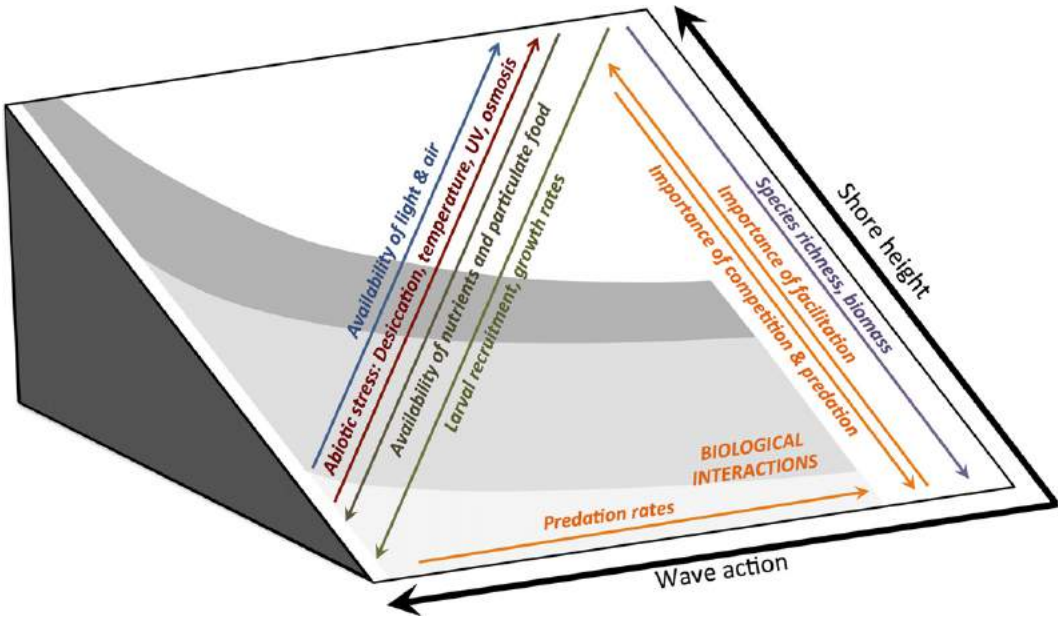


Fig. 2 Horizontal and vertical gradients of the abiotic and biotic factors that shape zonation patterns on intertidal rocky shores. The extent and alongshore compression of the three intertidal zones from wave-exposed toward sheltered shores is shown in gray shades (dark = high shore, medium = mid shore, light = low shore). Certain factors are more important for some species than others, and their importance might vary through time. Zonation patterns are thus often modified through complex dynamics and interactions of various processes.

which are important for the persistence of populations and communities. Last but not least, strong wave forces favor sessile or sedentary species that are firmly attached to the substratum and have streamlined body shapes. Along-shore differences in wave exposure gradient that exist between rocky shores of headlands and bays therefore shape characteristic along-shore patterns of biotic communities. These provide the backdrop upon which vertical zonation patterns are superimposed, and are therefore an integral part of descriptions of universal zonation patterns on rocky shores.

Zonation of Biota on Temperate Rocky Shores

Over the past two centuries, various zonation schemes have been developed for rocky shores. Zones have been defined by reference to dominant organisms, or in relation to particular levels of the tide, such as the mean high (or low) water level during neap and spring tides. Definitions based on physical criteria were initially considered to be more objective and were therefore more widely supported. However, they proved to be problematic in practice because tides are extremely variable and theoretical tidal levels do not necessarily have a clear manifestation on the shore. Moreover, zonation patterns are also influenced by factors other than tides, such as wave run-up, splash and spray, which rendered any classification based wholly on tides unsuitable for describing biological zones. Today, the most widely accepted framework is one that was proposed in 1949 by Anne and Allan Stephenson. They developed a classification scheme based on common patterns in the relative positions of certain community types along the intertidal gradient, which they had recorded during their extensive travels around the world. This so-called universal classification scheme for the zonation of rocky shores describes three zones, namely the "supralittoral," "midlittoral," and "infralittoral" zones (Fig. 3). While this scheme mostly holds true for patterns of temperate shores with moderate wave exposure, it is not always applicable since many factors modify and obscure the pattern, as discussed in further detail below (see "Modifications of the "Universal" Zonation Scheme" section).

Supralittoral fringe or high shore

Highest up on the shore, this zone is exposed to air most of the time and is only occasionally soaked by waves, spray, and spring high tides. Among the few organisms that can endure the long emergence times are encrusting lichens (e.g., *Verrucaria*), cyano-bacteria (blue-green algae), and large aggregations of tiny *Littorina* snails or other related genera (hence the occasional use of the term "Littorina zone" for this zone). In some locations, grapsid crabs, insect larvae, and isopods are common fauna of the high shore.

Midlittoral or mid-shore

The broad mid-shore zone encompasses the majority of the tidal gradient and is therefore usually divided into two: the "Upper Balanoid zone," where acorn barnacles (e.g., *Chthamalus*) dominate, as well as limpets (e.g., *Collisella*, *Patella*, *Scutellastra*); and the

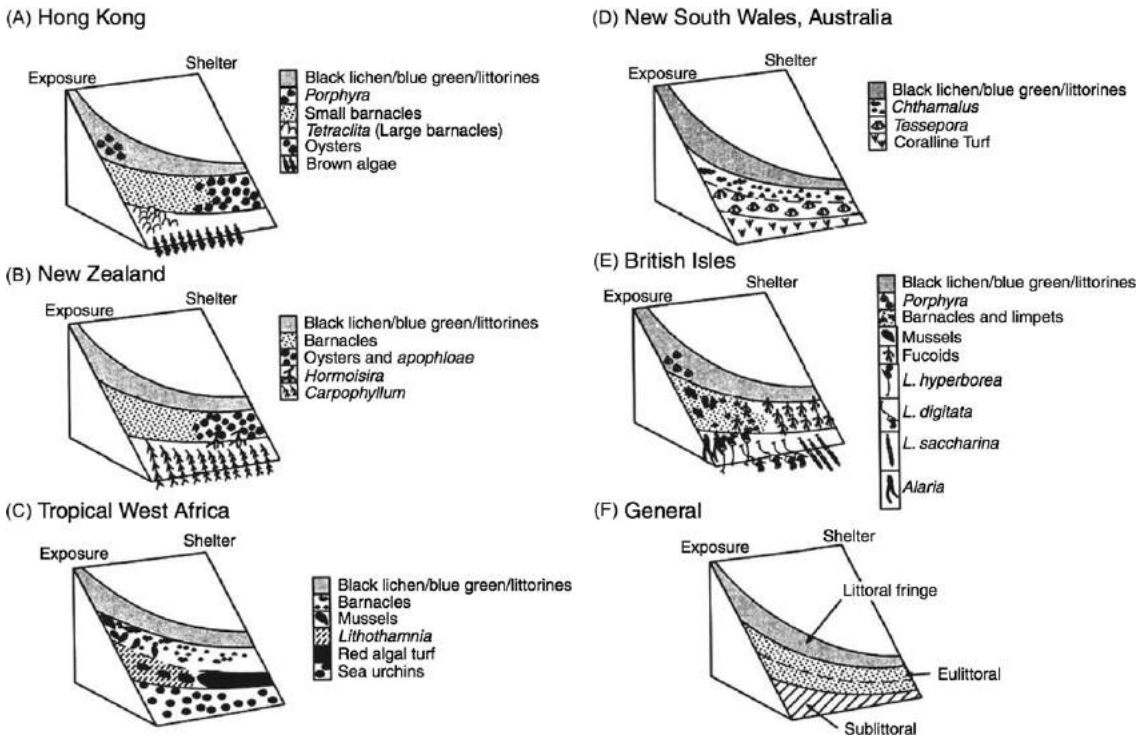


Fig. 3 Diagrammatic representation of intertidal zonation patterns in different parts of the world, and their modulation from exposed to sheltered extremes of wave exposure. Superficially, different regions show similar patterns; however, idiosyncrasies exist in each region depending on biogeographic context. Adapted from Raffaelli, D. and Hawkins, S.J. (1996). *Intertidal Ecology*. London: Chapman and Hall.

"Lower Balanoid zone," where suspension feeding mussels (e.g., *Mytilus*, *Aulocomya*), barnacles (e.g., *Semibalanus*), or oysters (e.g., *Crassostrea*) attain high densities. Depending on the geographic location, various species of macroalgae may also dominate in this zone.

Infralittoral or low shore

This narrower band at the bottom of the intertidal zone is only exposed to air during spring low tide. Red algae, including the pink encrusting and articulated coralline forms, dominate here. Where they exist, kelps (e.g., *Laminaria*, *Ecklonia*, *Durvillea*) or large tunicates can also prevail (e.g., *Pyura*).

Modifications of the "Universal" Zonation Scheme

A universal zonation scheme system is an attractive concept and provides a useful reference, but patterns in nature rarely adhere strictly to any one system of classification. Variations in zonation patterns exist across space and time and span a wide range of scales.

Spatial factors

At the broadest spatial scale, zonation patterns are modified by latitude. In the tropics (and also in high latitudes) extreme air temperatures are common, which makes life on the high shore taxing. In addition, tropical oceans are characterized by oligotrophic conditions and high grazing rates. Hence, productivity and biomass of algae on these shores tend to be very low.

At the scale of 1000s of km, marked differences in zonation patterns may exist between adjacent biogeographic provinces. A case in point is the shoreline of South Africa, which spans several biogeographic regions over 3000 km. On the west coast of the country, upwelling generates cold, nutrient-rich conditions, facilitating fast-growing giant kelps and high densities of the limpets *Scutellastra argenvillei* and *Cymbula cochlear* (**Fig. 1**A) in the low shore. The introduced mussel *Mytilus galloprovincialis* dominates the mid-shore and barnacles are scarce. On the warm-temperate south coast, low-shore kelps are replaced by the tunicate *Pyura* and a diversity of red and brown algae, while on wave-exposed shores the indigenous mussel *Perna perna* occupies the lower mid-shore, *M. galloprovincialis*, the upper mid-shore, and barnacles are very common on the high shore. Further north on the subtropical east coast, upright coralline algae, sponges, and mixed algal turfs characterize the low shore, various species of zoanthids cover the mid shore rocks, while on the high shore, a belt of oysters (*Saccostrea cucullata*) represents a distinctive feature of this region. Along this

remarkable biogeographic gradient, species richness and diversity increase from the cold west to the warm east coast, while biomass shows the inverse trend, increasing toward the west.

At the scale of 100s–1000s of meters, differences occur between wave-exposed and sheltered rocky shores of all regions. Most notably, the vertical range of intertidal species on wave-exposed shores is greater than at sheltered sites, with upper limits of some species extending considerably higher than the highest tide level (**Fig. 3**). This is a direct consequence of wave run-up, splash, and spray, which reduce the emersion time of the high shore during low tides and consequently reduce thermal stress and desiccation. Furthermore, suspension feeders benefit from food particles supplied by waves. The mid-shore of wave-exposed sites is therefore often dominated by dense mussel beds (lower mid shore) and barnacles (upper mid shore), both of which are suspension feeders. With decreasing wave exposure, mussels tend to be replaced by dense stands of algae while barnacles tend to decrease in density and co-occur with limpets, periwinkles, and algae.

Various processes cause significant small-scale heterogeneity in the vertical distribution of biota, which is reflected in the irregular rather than distinctly banded intertidal communities that occur at many locations. This patchiness is generated by physical and biological factors, such as slope, rock type, salinity, localized currents, sand inundation, and small-scale variability in settlement, recruitment, competition, and predation. The realization that localized processes commonly outweigh the effect of tides and waves has challenged the generality of zonation patterns and the usefulness of a universal classification scheme. Instead, it might be more valuable to recognize that multiple factors affect the spatial distribution of populations and communities, and that their relative contributions to vertical zonation patterns depend on scale and local context.

Temporal factors

Zonation patterns are driven by dynamic processes and are therefore also variable through time. Seasonal differences in the vertical distributions of mobile species and ephemeral algae are common and usually attributed to seasonal variability of one or more physical stressors, such as temperature and desiccation.

Marked shifts in zonation patterns have been observed over decadal time-scales, some of which reflect human impacts. For example, a fundamental transformation of many intertidal shorelines has occurred due to the invasion of introduced species. A well-known case is that of the Mediterranean mussel *M. galloprovincialis*, which has become spatially dominant on numerous temperate rocky shores around the world, displacing indigenous species and generating a wide mussel band where there was no such feature (**Fig. 4**). Long-term datasets that demonstrate the effects of climate change on rocky shore zonation are scarce. If, as predicted, the sea level rises due to ice melt and thermal expansion of the oceans, intertidal zones will most likely shift upward over time. However, if air temperatures continue to rise because of greenhouse gas emissions, the thermal stress gradient across intertidal zones will get steeper, forcing the spatial upper limit of most species downward. While accurate predictions of shifts in zonation patterns are difficult, if not impossible, it is likely that some species would expand horizontally or vertically, while others may contract or disappear in some regions.

Fig. 4 Changes in zonation patterns at Froggy Pond in False Bay, South Africa (34.21°S 18.46°E) between 1980 (*left*) and 2015 (*right*). A band of the introduced mussel *Mytilus galloprovincialis* has become a prominent feature of South African intertidal zones replacing indigenous species, such as algae and limpets. Photographs by G. Branch, adapted from Branch, G.M. and Branch, M. (2018). *Living Shores*. Cape Town: Struik Nature.

Causes of Zonation on Rocky Shores

Critical tidal levels

Initial explanations of intertidal zonation patterns were primarily based on observations of species distributions and their correlations with physical environmental gradients. The Critical Tidal Factors hypothesis, which was widely subscribed to during the first half of the 20th century, stated that the vertical limits of species distributions coincide with those sections on the shore where the gradient of physical stresses is particularly steep. At these critical levels, the physiological limits of species are exceeded and they are replaced by a different set of species, which are adapted to the changed conditions. Such critical levels were thought to exist because of the manner in which the tidal range varies over the course of a fortnightly lunar cycle (Fig. 5). During neap tides (around half moon), the highest levels of the shore remain exposed to air and sun but as the tidal cycle shifts to the spring tides (around full/new moon), tidal amplitudes become high and the whole intertidal slope gets submerged twice daily by the elevated high tides. Physical stress therefore changes in a stepwise fashion between critical tidal levels that can be defined according to how much time they are emerged for. Examples are the mean low-water mark during spring tides and the mean high-water mark during neap tides, both of which were thought to coincide with lower/upper species limits. Results of numerous laboratory experiments showing that high-shore species are more tolerant of aerial exposure than those found lower on the shore were taken as support of this theory.

A series of field experiments and novel quantitative analyses conducted during the 1960s and 1970s, however, conclusively demonstrated that the Critical Tidal Levels hypothesis was not supported by either physical or biological evidence. Since then, the focus has shifted away from correlative inferences of zonation patterns based on physical variables, toward experimental investigations of their causes. This important conceptual shift, in the manner that intertidal ecologists started linking patterns with processes and mechanisms, led to the understanding that biological factors, such as competition and predation, are in fact instrumental in determining the upper and lower limits of zonation patterns.

Causes of upper limits

There seems to be widespread acceptance that the upper limits of most intertidal species are determined by physical stresses, such as thermal stress and desiccation. Since most intertidal species are of marine origin, they require periodic wetting to fulfill some primary functions such as feeding, reproduction, and respiration, to survive. Furthermore, temperatures in the sea are less variable or extreme than those on land so that organisms occurring higher up on the shore are more likely to exceed their capacity to tolerate thermal extremes and perish. Young life stages tend to be more vulnerable to desiccation, and episodes of extreme temperatures periodically eliminate recruits of algae and sessile animals high up on the shore, thereby limiting the upper ranges of many species. Even lower down, extremely hot conditions on temperate shores have been known to cause bleaching and mortalities of red algae, indicating that their upper limits are periodically reset by physical extremes. However, physical stress does not have to be lethal to limit species distributions. Physiological adaptations may enable organisms to tolerate extreme conditions, but because of the associated high energetic costs, mobile organisms may rather select to settle on habitat where conditions are more moderate to minimize costs of maintenance and production. Moreover, the shorter emersion times higher up on the shore correspond to less time for marine larvae that are suspended in seawater to settle on the shore. The smaller likelihood of larvae settling high up on the shore further imposes upper limits to species distributions.

That the upper limits of species distributions are influenced by physical stresses has also been confirmed through several field experiments. Sessile organisms, such as algae or barnacles, when transplanted further up on the shore, showed decreased growth

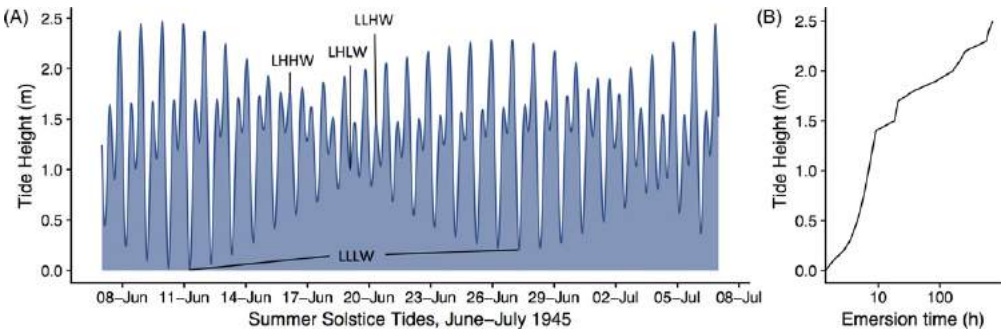


Fig. 5 (A) Tidal variation over the course of a month at Monterey Bay, California, showing mixed semidiurnal tides characterized by two different high- and low-tide amplitudes each day. Periods with greater tidal amplitudes (e.g., 7–14 June, 22–29 June) are referred to as spring tides and those with lower amplitude as neap tides (e.g., 15–21 June). The time period and region shown correspond with those for which the Critical Tidal Factors hypothesis was proposed (Doty, 1946). Levels considered as critical tidal factors are labeled according to the diurnal and spring neap variation in the tides. For example, LLLW refers to lower low low water, which is the spring (extreme) low of the lower of the two diurnal low tides. Other levels shown are lower high high water (LHHW), lower high low water (LHLW), and lower low high water. Intertidal zonation patterns were originally thought to correspond to discontinuities in tidal emersion (i.e., the time a particular shore level spends out of the water) shown in (B). However, this theory was later falsified by experimental proof of the importance of biological interactions in setting upper and lower limits of intertidal zones.

and eventually perished. Natural transplant experiments at the scale of whole communities have taken place when rocky shore ledges were lifted by earthquakes, as reported from Chile and Mexico. The communities that were displaced upward perished entirely and zonation patterns eventually re-established further down on the shore. Other experiments have manipulated moisture levels in high-shore zones and reported that species at the top of the shore shifted further upward, given enough moisture. However, to determine the exact mechanisms whereby environmental factors impose upper limits is difficult. There are various physical stress gradients that occur in the intertidal zone (**Fig. 2**) and the specific factor or combination of factors that determine the upper limits of a species may vary with location or time. It is therefore not always clear as to whether upper limits are in fact determined by physical factors, and alternative mechanisms have also been proposed, such as effects of food or nutrient limitation. Such factors can affect a species' performance and survival in a similar way to physical stress. Larval habitat selection can also be an important driver of zonation. Larvae of sessile species, such as barnacle cyprids, are known to carefully choose the substratum on which they settle, based on chemical cues, resulting in occupancy of a specific, preferred intertidal zone by adults. Vertical limits of some species are therefore the result of behavioral mechanisms, and not physical stresses.

In some cases, the upper limits of species are determined by biological interactions among species. In the British Isles, experimental removal of different species of fucoid algae, which naturally occur in discrete horizontal belts, has led to shifts of fucoids from lower zones into the higher, vacated zone, out of their natural range. These and similar experiments suggest that several mid and low-shore species are able to tolerate conditions higher up on the shore, but that their upper limits are set by competition with higher shore species. Grazing is another important biological factor responsible for upper distribution limits on rocky shores. In various temperate shores of the world, experimental removal of grazers, such as limpets, has allowed algae to extend further up the shore than where they occurred under grazed conditions.

Causes of lower limits

Since physical stress levels decrease further down the shore, lower limits of intertidal organisms are rarely set by physical factors, but are instead a product of biological interactions. However, as for upper limits, many different factors (as well as complex interactions between them) can determine lower limits, the most prominent being competition and predation.

A pioneering study that highlighted the importance of competition for zonation patterns focused on barnacles on the west coast of Scotland. There, the high shore is characterized by a distinct band of *Chthamalus* above a zone of the larger, faster-growing *Semibalanus*. Transplant experiments indicated that *Chthamalus* was able to settle and survive the physical conditions below its natural zone, but that it would then be crushed and smothered by *Semibalanus*, which in turn could not tolerate conditions higher up on the shore. Removal of the latter facilitated a downward shift of *Chthamalus*, convincingly showing that the lower limits of this species are determined by competitive interactions with *Semibalanus*.

Another influential experiment from the north-west coast of the United States proved the importance of predation for setting the lower limits of sessile invertebrates. On this shore, the mussel *Mytilus* forms dense stands along a clearly defined band in the intertidal zone, below which there is a zone dominated by the predatory mussel-consuming starfish *Pisaster*. For a 5-year period, starfish were cleared from the low shore at two sites, which facilitated a downward expansion of *Mytilus* to two meters below its natural lower limit. From this study emanated the concept of a keystone species. A keystone predator fulfills a central role in maintaining species diversity by limiting the abundance of a spatially dominant prey species, in this case *Mytilus*.

The lower limits of zone-forming algae, which are common features of rocky shores, tend to be imposed by competition or grazing. Several species of fucoid algae occur in different intertidal zones of Scotland. Experimental removal of a lower shore species allowed a higher shore species to expand downward. The competitive hierarchy of fucoids was found to be correlated with their growth rates, such that faster growing species dominate the lower shores while slower growing species are displaced up the shore. Low-shore kelps of cold-temperate shores, such as *Laminaria*, *Lessonia*, and *Durvillea*, have extremely fast growth rates and are therefore known to push other seaweeds upward.

Multi-factorial approaches

From the above it is clear that there are multiple biological and physical factors that can interact dynamically along intertidal gradients. This results in a highly complex web of interactions that are difficult to disentangle by experiments alone. With the advancement of numerical modeling, intertidal zonation patterns have been revisited within a multifactorial framework that is based on information from an extensive body of empirical studies, some of which are described above. These models have the advantage that physical stresses and vertical trends in recruitment and growth rates imposed by tides can be combined with biotic processes such as variable settlement, indeterminate growth, size-dependent competition, grazing, and predation. Several such models have successfully reproduced intertidal zonation patterns based on multifactor interactions. For example, an individual-based model of fucoid algae in the British Isles, for which inputs were derived from laboratory and field studies, was able to recreate and confirm a mechanism for algal zonation that was proposed more than a century ago, namely that the zonation was a result of the trade-off between growth rate and tolerance to desiccation. Another example, where the mechanisms underlying zonation patterns across a wave-exposure gradient were confirmed using a model, is the predation-based model of the *Mytilus–Pisaster* system.

Both the above models showed that zonation patterns can be shaped by complex interactions between physical gradients and various biological characteristics of a population, as opposed to simple direct effects of single physical or biological factors. According to this view, the upper and lower limits of a species are determined by multiple factors that are dynamic in time and

space. This means that the vertical extent of a species will vary in space and time and that where or when its lower and upper limits converge, conditions are such that the species cannot exist. This can come about at the range limit of a species, or represent a localized extinction due to unfavorable conditions.

Sandy Beach Zonation

At first glance, sandy beaches appear to be devoid of life, other than perhaps for crowds of tourists. While keen observers may recognize a few resident bird species, ghost crabs, or plants on vegetated dunes, it might seem unlikely to them that beaches are in themselves ecosystems with unique biota distributed in specific zones across the shore. However, pronounced physical gradients similar to those described for rocky shores also exist on sandy beaches, and faunal assemblages mirror them in their vertical and horizontal distribution patterns.

Dependence of Beach Slope on Particle Size, Waves, and Tides

Like rocky shores, sandy beaches are characterized by wave action and fluctuations of the tides, which continuously reshape the unconsolidated sediments that these shores are composed of. This creates a dynamic, three-dimensional environment in which most animals are buried below the sand surface during low tide. Three "super parameters," namely sediment particle size, wave action, and tidal range, interact in predictable ways to create beaches with a variety of slopes and different morphodynamic states. When a breaking wave rolls up the beach, the "swash" moves sediment up the shore, and the "backwash" carries it back down. If the beach consists of coarse sand, water drains quickly into the permeable substrate on the upwash, accreting the beach and forming a steep beach face... Fine sand, on the other hand, gets waterlogged due to its low permeability, resulting in equally strong upwash and backwash processes. Thereby sand particles suspended in the swash are removed, the beach gets eroded and the beach face flattens. Thus, given constant wave action, the beach slope steepens with an increase in grain size.

Wave action and tidal range furthermore affect the beach slope. Stronger wave action on exposed shores is associated with increased swash action and more waterlogged beaches. The stronger swash processes cause erosion of the beach face and thereby create flatter beaches. Beaches with a large tide range, where large amounts of water move up and down the intertidal zone during a tidal cycle removing particles, tend to be wide and flat.

Beach Morphodynamic States

Based on beach slope, particle size, and wave action, beaches can be classified according to various morphodynamic states. At the extremes of the spectrum are "reflective" or "dissipative" beaches. Reflective beaches are characterized by small waves, fast swashes and coarse sand. They are steep, with narrow or no surf zones, and the wave energy arriving on the beach face is reflected back out to sea. Reflective beaches are considered physically harsh for intertidal life and are characterised by depauperate communities. By contrast, dissipative beaches have large waves with long and slow swashes, and fine sand. They have flat slopes and wide surf zones, where most energy is dissipated and little of this energy reaches the beach face. Thus, conditions on dissipative beaches are benign and they feature relatively high species richness and abundances. Most beaches are intermediate between these two extreme states. Where fine sand and large waves occur in combination with tidal ranges exceeding two metres, ultra-dissipative conditions develop.

Physical Zonation Schemes

The most widely recognized zonation scheme for sandy beaches is that developed by the French ecologist Salvat in 1964. In contrast to the universal zonation scheme for rocky shores, which is based on biotic patterns, this sandy beach zonation classification is based on physical features. Four zones are defined, based on vertical changes in the hydrodynamic properties of the intertidal zone, namely the zones of saturation, resurgence, retention, and drying (**Fig. 6**). The zones are most clearly identifiable during spring low tide; however, due to the dynamic nature of beaches and lack of a permanent reference points, zones are not fixed and may vary between spring and neap tides. The saturation zone is the area where the sediment is permanently saturated with water, or waterlogged. Oxygen levels in the sediments of this zone are frequently low due to high bacterial activity and low exchange rate of water, especially on fine-grained beaches. The resurgence zone is the compact zone extending from the low tide mark upward and is characterized by wet sand during low tide. Moisture is retained through capillary action and exchanged throughout a tidal cycle. The resurgence zone is difficult to distinguish visibly and it blends into the neighboring retention zone. The retention zone is located on the upper shore, and is characterized by damp (not wet) sand. Physically it extends from about 20–30 cm above the low tide water table (which is the upward extent of capillary action) to the neap high-tide level of the shore, and water permeates into the substrate during incoming tides. Due to frequent replenishment and contact with air, oxygen levels are highest in this intertidal zone. The highest level of the shore is the drying zone. It is only under water during spring high tides, and mostly constitutes dry, loose sand. For convenience, the zones on sandy beaches are commonly referred to as "supralittoral" (drying zone), "littoral" (retention and resurgence zones), and "sublittoral" (saturation zone). In further sections, these terms will be used to describe biotic zonation patterns.

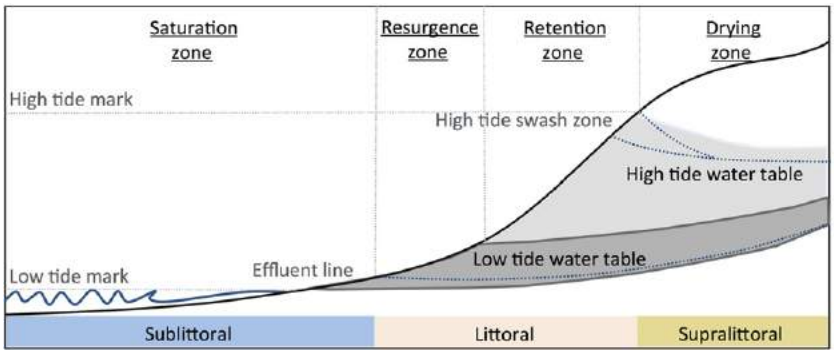


Fig. 6 Diagram showing the sandy beach zonation scheme of Salvat (1964). Four vertical zones (Saturation, Resurgence, Retention, and Drying zones) are defined based on physical characteristics, such as the wetness of the substratum. Adapted from McLachlan, A. and Defeo, O. (2018). *Ecology of Sandy Shores*. London: Academic Press.

Faunal Zonation Patterns

Macrofauna

Characteristic macrofaunal assemblages are associated with the physically defined vertical zones of sandy beaches; however, biotic zonation patterns also vary horizontally along the gradient of morphodynamic beach states (**Fig. 7**). On reflective beaches, only two biotic zones can be distinguished: a littoral zone of marine species on the lower shore and a supralittoral zone of air breathers higher up. On intermediate and dissipative beaches, these biotic zones are found higher up and a third sublittoral zone exists low down on the shore. Nearer to the ultra-dissipative extreme, the sublittoral saturation zone widens and a fourth biotic band can be distinguished (**Fig. 4**).

The physical harshness of reflective beaches allows for only highly robust organisms to persist. Under the most extreme conditions on steep reflective beaches only supralittoral oniscid pill bugs such as *Tylos*, wrack feeding sand hopper amphipods (*Talitrus* or *Orchestia*), and ocypodid ghost crabs survive in the supralittoral zone. In the mid intertidal, or littoral zone, a number of crustaceans are commonly found, especially generalists such as cirolanid isopods (*Excirolana* or *Eurydice* species). Lowest on the shore, just beyond the lower boundary of the intertidal zone by definition, are fast burrowing hippid crabs (*Hippa* or *Emerita*) and donacid bivalves. The more benign conditions that occur lower down on the shore are associated with a greater availability of niches; this is reflected in the greater number of specialist species on intermediate and dissipative shores. Polychaetes (*Scolelepis*) and amphipods (*Urothoe*) are more abundant here and scavenging whelks, such as *Bullia* or *Olivella*, thrive. The saturation zone can be considered an upward extension of the surf zone, and it is characterized by filter-feeding and deposit-feeding crustaceans, polychaetes, and donacids, as well as bentho-planktonic forms such as mysid shrimps (e.g., *Gastrosaccus*).

Species richness increases linearly along the spectrum from reflective to dissipative beaches, whereas abundance and biomass increase exponentially. With an increase in the density and diversity of biota, biological interactions such as predation and competition are more likely to have a structuring effect on communities. However, manipulative experiments to confirm interactions and mechanisms are difficult in this highly dynamic environment.

In comparison to rocky shores, the physical and macrofaunal zonation patterns on sandy beaches are more dynamic and their boundaries shift with the tides. Also, species found on sandy beaches are not tightly restricted to any one zone; they can be distributed over multiple zones and across a range of physical conditions. To cope with these changing conditions, sandy beaches are dominated by mobile species, in contrast to rocky shores that are dominated by sessile or sedentary organisms. The ability to run, burrow, or swash-ride makes it possible for animals to escape unfavorable conditions across the beach and move with the changing tide. In this three-dimensional habitat, biological interactions are less important in structuring communities, and intertidal across-shore zonation is more tightly linked with physical gradients. The exceptions are ultra-dissipative beaches, where species occur in higher densities and biological interactions may become more influential in structuring the zonation of intertidal communities.

Meiofauna

On a scale that is virtually invisible to the naked eye, sandy beaches host an astonishing diversity of life forms, known as meiofauna. Almost all known metazoan phyla are present in the meiofauna. Whereas macrofauna are adapted to displace sediments with their robust exoskeletons, most meiofauna live in the moist spaces between the sand grains, and therefore typically have slender or worm-like body shapes. Meiofaunal communities are also characterized by intertidal zonation patterns, which closely reflect water saturation levels of the sediment. The supralittoral dry zone on the high shore is dominated by small nematodes and oligochaetes. Lower down, in the littoral zone where the sand is moist and oxygen saturation is consistently high ($>50\%$), a variety of species coexist, including harpacticoid copepods, nematodes, mystacocarids, Oligochaetes, and turbellarians. In the sublittoral zone, where oxygen levels tend to be depleted through bacterial decay of organic matter, low densities of

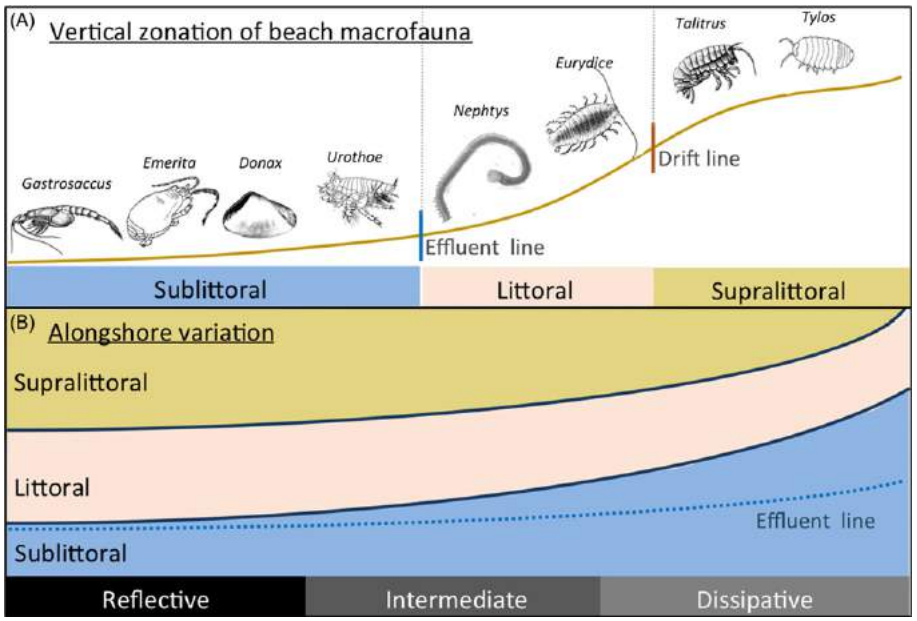


Fig. 7 Zonation scheme for sandy beach macrofauna with (A) the typical organisms occupying different levels across the shore (and names of most common genera) and (B) the alongshore modulation of the scheme according to morphodynamic beach state. Two zones get exposed during low tides on reflective beaches, three on intermediate beaches, and four on dissipative beaches. Modified from McLachlan, A. and Defeo, O. (2018). *Ecology of Sandy Shores*. London: Academic Press.

nematodes are found. The abundance of meiofauna increases with an increase of filtration rates from dissipative to reflective beaches, which is contrary to the pattern observed for macrofauna.

Impacts of Zonation Research

The ubiquitous zonation patterns that exist along the intertidal stress gradient have inspired numerous ecologists to investigate their underlying mechanisms. Several of these studies were instrumental in pushing the envelope of community ecology by introducing novel ideas of community regulation through species interactions (see "Causes of Zonation on Rocky Shores" section). Field experiments on intertidal rocky shores, in particular, have made important contributions toward understanding the roles of predation, competition, grazing, facilitation, and the interactive effects of physical and biological factors on structuring communities. Despite extensive research, intertidal zonation remains a field of current interest. New challenges include understanding how the impacts of human-induced pressures, such as introduced species, pollution, exploitation, and climate change, will affect the interplay of physical and biological drivers that shape intertidal zonation patterns.

See also: Behavioral Ecology: Thermoregulation in Animals: Some Fundamentals of Thermal Biology. Conservation Ecology: Biodiversity Indices. Ecological Processes: Waves as an Ecological Process. Ecosystems: Rocky Intertidal Zone. Evolutionary Ecology: Colonization. General Ecology: Carrying Capacity; Ecophysiology; Succession; The Intermediate Disturbance Hypothesis; Epiflora and Epifauna; Temperature Regulation. Global Change Ecology: Emergence of Climate Change Ecology

Further Reading

Connell, J.H., 1961. The influence of interspecific competition and other factors on the distribution of the barnacle *Chthamalus stellatus*. Ecology 42, 710–723.
Connell, J.H., 1972. Community interactions on the marine rocky intertidal shores. Annual Review of Ecology and Systematics 3, 169–192.
Doty, M.S., 1946. Critical tide factors that are correlated with the vertical distribution of marine algae and other organisms along the Pacific Coast. Ecology 27 (4), 315–328.
Harley, C., 2009. Zonation. In: Denny, M.W., Gaines, S.D. (Eds.), Encyclopedia of tide pools and rocky shores. Berkeley: University of California Press, pp. 647–653.
Lewis, J.R., 1964. The ecology of rocky shores. London: English University Press.
McLachlan, A., Defeo, O., 2018. Ecology of sandy shores. London: Academic Press.
Menge, B.A., Branch, G.M., 2001. Rocky intertidal communities. In: Bertness, M.D., Gaines, S.D., Hay, M.E. (Eds.), Marine community ecology. Sunderland: Sinauer Associates, Inc., pp. 221–252.
Paine, R.T., 1974. Intertidal community structure: Experimental studies on the relationship between a dominant competitor and its principal predator. Oecologia 15, 93–120.
Raffaelli, D., Hawkins, S.J., 1996. Intertidal Ecology. London: Chapman and Hall.

Salvat, B., 1964. Les conditions hydrodynamiques interstitielles de sediments meubles intertidaux et la rapartition verticale de la fauna endogee. Comptes Rendus de l'Académie des Sciences 259, 1576–1579.

Robles, C.D., Desharnais, R., 2002. History and current development of a paradigm of predation in rocky intertidal communities. Ecology 83, 1521–1536.

Stephenson, T.A., Stephenson, A., 1972. Life between tide marks on rocky shores. San Francisco, CA: W.H. Freeman.

Underwood, A.J., 1978. A refutation of critical tidal levels as determinants of the structure of intertidal communities on British shores. Journal of Experimental Marine Biology and Ecology 33, 261–276.

Underwood, A.J., Denley, E.J., 1984. Paradigms, explanations, and generalizations in models of the structure of intertidal communities on rocky shores. In: Strong, D., Simberloff, D., Abele, L.G., Thistle, A.B. (Eds.), Ecological communities: Conceptual issues and the evidence. Princeton, NJ: Princeton University Press, pp. 151–180.

Maximum Sustainable Yield

Athanassios C Tsikliras, Aristotle University of Thessaloniki, Thessaloniki, Greece
Rainer Froese, GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany

Introduction

Although the term "over-fishing" was coined already in the mid-1850s (Cleghorn, 1854), the overexploitation of marine fisheries resources was only realized in the early 1900s, when the first study (Garstang, 1900) and articles on overfishing (Petersen, 1903; Kyle, 1905) were published. By that time, the need for simple and easy to understand guidance on catch limits emerged. The maximum catch that a population can support seemed to be an excellent reference point for fisheries management.

Maximum Sustainable Yield (MSY) is the most well-known acronym in fisheries science and, as a concept, has a history of around 100 years (Baranov, 1918). It was formulated in the 1930s when mathematical models were introduced in population ecology (Hjort et al., 1933) and bloomed in the 1950s with the development of surplus production models.

Today, MSY has been adopted by the vast majority of regional management bodies (Hilborn and Walters, 1992; Quinn and Deriso, 1999; Mace, 2001; Hart and Reynolds, 2002; EC, 2009; Pauly and Froese, 2014) and it is therefore widely used as a reference point in the assessment of exploited populations (stocks) around the world.

The MSY Concept

Definition of MSY

MSY (also called maximum surplus production, maximum equilibrium catch, maximum constant yield, maximum sustained yield, sustainable catch: Ricker, 1975; Hilborn and Walters, 1992; Quinn and Deriso, 1999; Mace, 2001) is the highest theoretical equilibrium yield that can be continuously taken from a stock under existing (average) environmental conditions (FAO, 2001). It is the highest catch that still allows the population to sustain itself indefinitely through somatic growth, spawning, and recruitment (Graham, 1943; FAO, 2001).

MSY was formally introduced by Milner Schaefer (Schaefer, 1954) who developed the model named after him based on the logistic curve of population growth (Fig. 1). Plotting the first derivative (= the slope) of that curve over the corresponding biomass (the collective weight of the individuals at a certain time) shows the increase in biomass (termed surplus production or yield) with time, in the form of a parabolic curve (Fig. 2). The interpretation of the parabola is easy: at the left end there is zero biomass and therefore zero yield. At the opposite end, where the population is at carrying capacity of the ecosystem for this stock, there is no surplus production by definition and thus again zero yield. In other words, initially the population grows exponentially, unrestricted by environmental conditions. But as population size approaches carrying capacity, growth slows down and eventually seizes. Because of the symmetric shape of the logistic curve, maximum surplus production or yield is reached at half of maximum population size in the Schaefer model (Fig. 2; see "Methods to Estimate MSY section"). Taking away this maximum surplus production by fishing prevents the population from growing any further, basically keeping it at half of maximum population size, producing maximum surplus forever; hence this is the point of MSY.

In this simple model, the rate of population increase r is a linear function of biomass, maximum at zero population size, and zero at carrying capacity (Hart and Reynolds, 2002; Quinn and Deriso, 1999).

Various explanations have been offered for the typical S-shape of population growth, such as improved somatic growth at low population size versus increased intraspecific competition at high densities (Hart and Reynolds, 2002). Carrying capacity (K) has

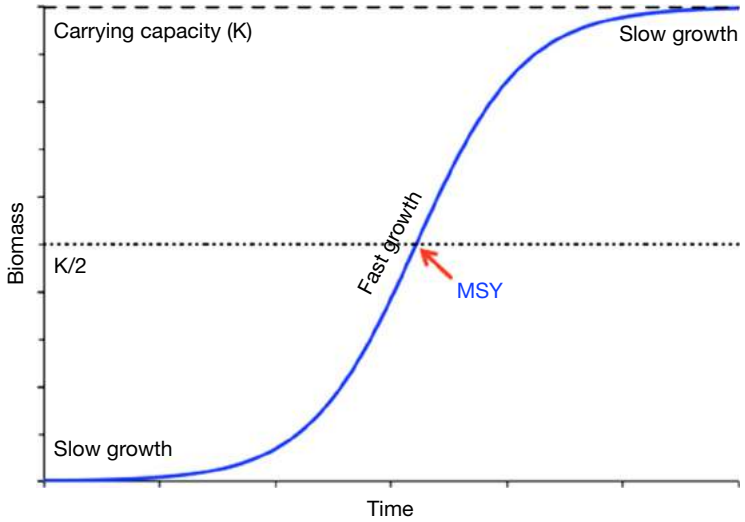


Fig. 1 The logistic (sigmoid) curve of population growth over time. The carrying capacity (K) and MSY (=K/2) are indicated along with phases of slow and fast population growth.

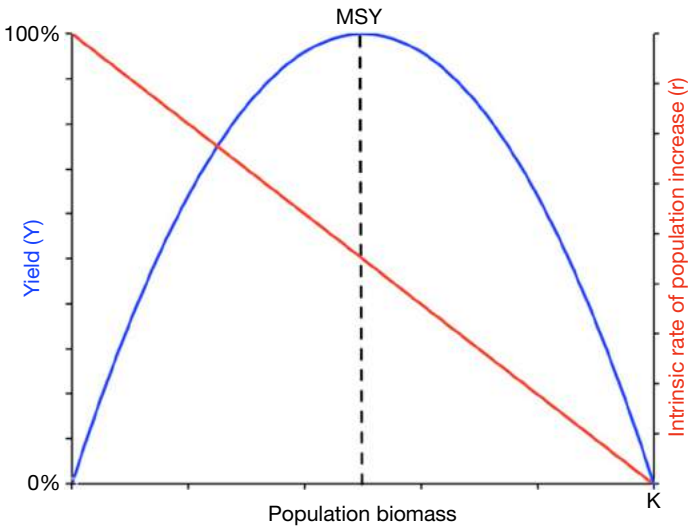


Fig. 2 The parabolic curve of surplus production or equilibrium yield (as % of MSY) as a function of population biomass (as % of carrying capacity, *K*). The second *Y*-axis shows the intrinsic rate of population increase r, which in the Schaefer model is maximum at zero biomass and declines linearly to zero at carrying capacity. Note that MSY is produced at half of r_{max}.

been interpreted in Malthus' terms as being caused by limited food availability (Seidl and Tisdell, 1999). However, today most ecologists agree that the main driver of population growth is the interplay between reproductive success and mortality (Charnov, 1993; Sibly et al., 2012): once the number of new individuals equals the number of deaths population growth seizes. At low population sizes, new individuals exceed the number of deaths and population growth is exponential. But while the number of deaths remains proportional to population size, the production of new individuals slows and reaches a more or less constant value once the population has grown beyond about a quarter of carrying capacity. As a result, the exponential growth slows to a linear growth at about half of carrying capacity and declines thereafter, approaching carrying capacity in an asymptotic curve (Fig. 3). Density effects causing death by starvation are thought to apply mostly to early life stages (Houde, 1987; Bailey and Houde, 1989; Hüssy et al., 1997) and cause a limit to reproductive success, as indicated by the green curve shown in Fig. 3.

Today, the MSY definition most widely used is the one proposed by Ricker (1975). According to "the green book" of Ricker, MSY is defined as the largest average catch or yield that can continuously be taken from a stock under existing environmental conditions but for species with fluctuating recruitment the maximum catch may be obtained by taking fewer fish in some years than in others (Ricker, 1975).

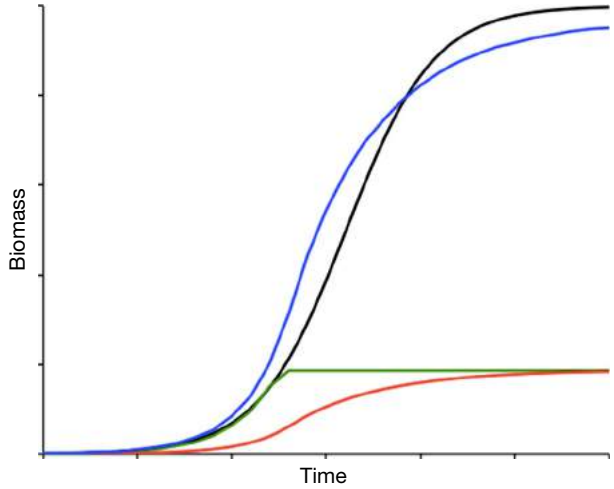


Fig. 3 The theoretical logistic curve of population growth over time (*black line*) compared with a hypothetical curve (*blue line*) resulting solely from the number of deaths (mortality: *red line*) and the number of replacements through recruitment (reproductive success: *green line*), which is more or less constant above about a quarter of carrying capacity.

Related Biological Reference Points

Current fisheries management is based on fishing mortality and biomass reference points that correspond to MSY, although MSY itself is rarely used as a reference point. Two related reference points are applied: one is the fishing mortality or fishing pressure (F_{MSY}) that, if applied over a time span similar to generation time, will eventually result in a catch equal to MSY (Fig. 4) where F describes the part of the total mortality rate that is caused by fishing. For example, $F = 0.6$ means that about 60% of the fish that are there on average over the year will be killed by fishing. The other related reference point is the biomass at MSY, B_{MSY}, which is the smallest stock size that can support catches equal to MSY, and which is the biomass corresponding to the peak in Fig. 2.

In age-structured assessment models, the fishing mortality that results in the maximum yield per recruit (F_{MAX}) is close to F_{MSY} if the yield per recruit versus F curve has a well-defined peak. However, if that peak is less well defined, as in Fig. 4, then F_{MAX} may be substantially larger than F_{MSY} (Longhurst, 2006).

Methods to Estimate MSY

MSY, $F_{MSY,}$ and B_{MSY} can be estimated from surplus production models, which require catch and effort or an index of biomass or relative abundance (e.g., catch per unit of effort) as input. Alternatively, these or similar reference points can be obtained from age-structured models, which are, however, more data demanding.

Surplus Production Models

Surplus production models are used to assess stock status and exploitation in data-limited areas where reliable information on age and length structure and natural mortality are not available (Beverton and Holt, 1957; Punt, 2003). They are applied not only to stocks with available commercial catch data and some index of exploitable biomass, such as catch per unit of effort (CPUE) derived from scientific surveys, but also to migratory stocks and crustaceans that are difficult to age (Polacheck et al., 1993). They assume that sustainable catch is a simple function of population biomass, regardless of the size and age composition of that biomass (Holt, 2014).

The most widely used surplus production model is the one developed by Schaefer (1954):

$$B_{t+1} = B_t + r_{max}B_t\left(1 - \frac{B_t}{K}\right) - C_t$$

where B_t is the biomass of the stock at time t and $t + 1$, r_{max} is the maximum intrinsic rate of population increase, K is a parameter which corresponds to the unfished equilibrium stock size or carrying capacity, and C is the catch per unit of time (usually a year).

Surplus production or yield (Y) is calculated as:

$$Y = r_{max}B_t\left(1 - \frac{B_t}{K}\right)$$

MSY is calculated as:

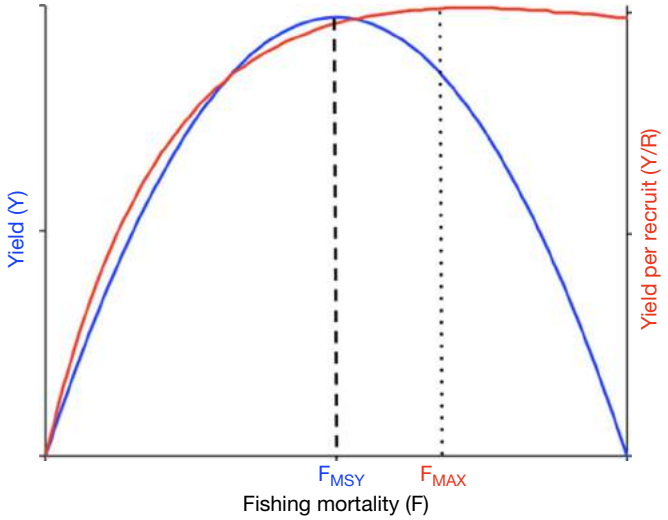


Fig. 4 Difference between the position of F_{MSY}, the fishing mortality expected to yield MSY in a surplus production model (*blue line*) and those of F_{MAX}, the mortality from yield-per-recruit curves (*red line*).

$$MSY = \frac{r_{max}}{2} \times \frac{K}{2} = \frac{r_{max}K}{4}$$

where $r_{\max}$ is the maximum intrinsic rate of population increase, B_t is the population biomass at time t, and K is the carrying capacity of the ecosystem for this population (Schaefer, 1954).

Another surplus production model was developed (Fox, 1970) that assumes a logarithmic relation between biomass and catch:

$$B_{t+1} = B_t + r_{max}B_t\left(1 - \frac{\log(B_t)}{\log(K)}\right) - C_t$$

where variables and parameters are as defined above.

In the Fox model surplus production or yield is calculated as:

$$Y = r_{max}B_t\left(1 - \frac{\log(B_t)}{\log(K)}\right)$$

MSY is calculated as:

$$MSY = \frac{r_{max}K}{e}$$

In the Schaefer model, maximum yield (MSY) is obtained at 50% of carrying capacity and in the Fox model at 37% of carrying capacity. Pella and Tomlinson (Pella and Tomlinson, 1969) proposed a model with a third parameter p that determines the shape of the yield curve and allows maximum production to occur at any biomass.

The Pella-Tomlinson model is:

$$B_{t+1} = B_t + \frac{r_{max}}{p}B_t\left(1 - \frac{B_t}{K}\right)^p - C_t$$

where variables and parameters are as defined above and p is a shape parameter that results in the Schaefer curve if $p = 1$ and approximates the Fox curve if p approaches 0.

Surplus production or yield is calculated as:

$$Y = \frac{r}{p}B_t\left(1 - \left(\frac{B_t}{K}\right)^p\right.$$

MSY is calculated as:

$$MSY = r_{max}K\left(\frac{1}{1+p}\right)^{\left(\frac{1}{p}+1\right)}$$

The Schaefer surplus production model is the one most commonly used in fisheries management because of its simplicity and applicability in data-poor stocks.

Age-Structured Models

In cases where age and length data are available, surplus production models have been replaced by age-structured models that also provide estimates of MSY and relevant reference points but these models are data demanding (Hilborn and Walters, 1992; Mace, 2001) and require population age and length, growth parameters, mortality and maturity, as well as selectivity of the main gears. Estimates of MSY, F_{MSY}, and B_{MSY} are typically obtained from stochastic simulations.

Age-structured models are widely used in assessing stocks and require estimates of mortality, maturity, catch, and abundance per age group, but these models are not suitable when only catch and biomass data are available.

Economic Considerations

Only 150 years ago, the advisor on fisheries to the British Government (Huxley, 1884) declared that humans were unable to overexploit marine fish stocks. The subsequent advent of steam trawlers and the collapse of North Sea herring (Dickey-Collas et al., 2010) proved him wrong and for over 75 years it was well understood that fisheries need to be regulated to sustain fish stocks and profitable fisheries (Graham, 1943). This can be easily demonstrated by adding cost of fishing (which increases about linear with effort) and profits (the difference between the value of the catch and the cost of fishing) to the parabola graph of the relation between yield and effort (Fig. 5). In most fisheries the cost of fishing at the MSY level is less than the value of the maximum sustainable catch and maximum profit or maximum economic yield (MEY) is actually obtained with even less fishing effort, simply because the linear decline in cost is steeper than the corresponding decline in catch near the peak of the parabolic curve (Fig. 5). Unfortunately, effort in most fisheries in the world is far above the MSY level resulting in low catches and economic loss (Costello et al., 2012; Froese et al., 2017). This is possible because governments give handouts (= subsidies) to the fishers, which lower the cost of otherwise economically unsustainable overfishing.

MEY is the antidote to the illusion of most fishers (and some politicians) that higher fishing effort results in higher profits. In reality, profits decline once effort exceeds the MEY level and catches decline once effort exceeds the MSY level (Fig. 5).

History and Legal Status of MSY

MSY is based on the classical ecological concept of logistic population growth that was developed in the 1830s (Verhulst, 1838), continuing the earlier work of Robert Malthus on demographics (Malthus, 1798). The first application of the logistic model on

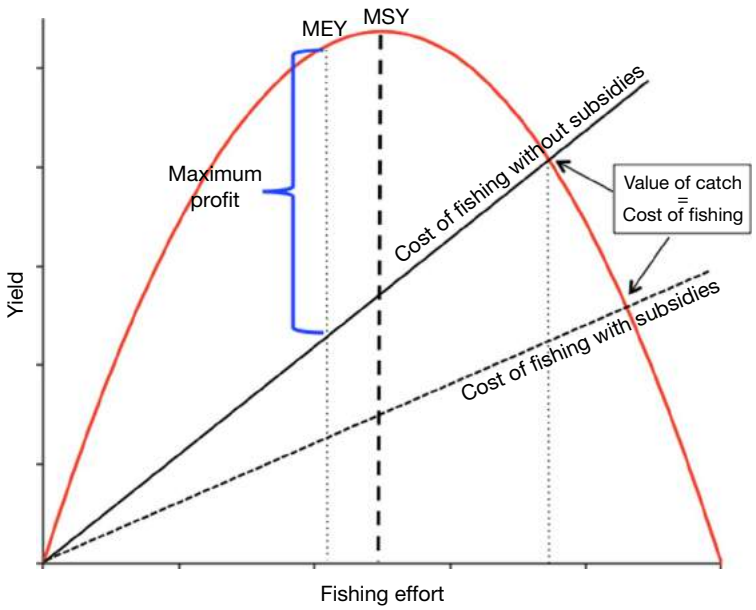


Fig. 5 Yield (*red line*) as a function of fishing effort, with cost of fishing with (*dashed black line*) and without subsidies (*continuous black line*) assumed to increase about linearly with effort. Profit is the difference between the value of the yield and the cost of fishing, with a maximum economic yield (MEY, *dotted vertical line*) obtained from effort and catches below the MSY level (*dashed vertical line*). If costs of fishing are lowered by subsidies, fishing can continue beyond the break-even point, where the value of the catch equals the cost of fishing (second *dotted vertical line*).

marine species was by Johan Hjort and his colleagues in the 1930s, who studied the blue whale fishery, based on mortality and reproduction data (Hjort et al., 1933). They developed the notion of optimal catch, which occurred at intermediate exploitation levels based on their observations on fin whales in Iceland, cod and herring in Norway, and plaice in the southern North Sea (Holt, 2014). Hjort et al. (1933) showed that the greatest rate of population growth increase occurs when the population size is about half its ultimate size and that there was a maximum catch that could be sustained, later termed MSY (Hart and Reynolds, 2002).

Shortly after Hjort's work, Michael Graham further developed the logistic population growth equation and applied it to fisheries data (Graham, 1935, 1943). He identified slow and fast population growth phases, with fast growth, low density and younger fish at low population sizes and slow growth, high density and many old fish at large populations near carrying capacity (Graham, 1935). Based on his observations from several species, he argued that "*a lower fishing rate would give as great a yield when the stock became stabilized at that rate*" (Graham, 1935). Later, Graham concluded that "After a certain point the total yield of a fishery does not increase any more, whatever the fishermen do" and clearly linked exploitation to economics when he wrote that "the benefit of efficient exploitation lies more in economy of effort than in increase of the yield" (Graham, 1943).

Schaefer (1954) developed the eponymous Schaefer model using the logistic growth curve on Californian sardine. He replaced population numbers with biomass and defined surplus production as yield. Thus, he formally introduced the concept of MSY, then termed maximum equilibrium catch (Schaefer, 1954). Schaefer preferred the expression maximum equilibrium catch to optimum catch "as being more descriptive of exactly what is meant" (Schaefer, 1954). Schaefer, who was working with tuna, had to ignore age composition because there was not, at the time, a way of determining the age of an individual tuna (Holt, 2014).

In fact, MSY was in use as a theoretical concept a few years before the publication of Schaefer's surplus production model, when it was adopted as the scientific foundation of the US High Seas Policy, in 1949 (Chapman, 1949; Finley, 2011). MSY adoption was largely based on Graham's theoretical analysis (Graham, 1943; see the section "History and Legal Status of MSY") and his conclusion that less fishing can provide in some cases more fish. It was later that MSY was quantified with the Schaefer's surplus production model, and then, in 1955, adopted as the goal of international fisheries policy at the Rome Conference on fisheries problems (Smith, 1994).

After the late 1950s, MSY has been adopted as the primary management goal by several international organizations (IWC, IATTC, ICCAT, ICNAF) and countries (Mace, 2001; Froese et al., 2011). The United Nations Convention on the Law of the Sea (UNCLOS, 1982) made the MSY approach mandatory for fisheries in the exclusive economic zones (EEZs) of its signatories, which were obliged to include the MSY concept into national or regional fisheries legislation (Mace, 2001). All 39 then existing regional fisheries organizations (RFMOs) agreed to manage their mandated stocks such that they were capable of producing MSY (Longhurst, 2006). The follow-up conference on the United Nations Fish Stocks Agreement (UN, 1995) clarified in its Appendix II that during a phase of reducing excessive fishing effort, the one associated with MSY could be used as a target, but that once that target had been reached, MSY had to be treated as a limit, that is, fishing effort should be less than the one resulting in MSY.

For example, in the MSA (2006) the goal of fisheries management is defined as "optimal yield," which is "prescribed on the basis of the MSY from the fishery, as reduced by any relevant economic, social, or ecological factor" (Froese et al., 2011). In addition, fisheries management based on MSY has been formally implemented in New Zealand (MFNZ, 2008), Australia (DAFF, 2007), and Europe (EC, 2013). In most of these areas MSY-based policies have been quite successful in rebuilding stock biomass (Hilborn, 2007a; Mesnil, 2012).

Critique of MSY

The implementation of MSY as a catch that can be taken continuously independent of recruitment, stock size, stock structure, and environmental conditions has been questioned and criticized early on (e.g., Beverton and Holt, 1957; Larkin, 1977; Sissenwine, 1978). Its assumptions, uncertainties, limitations, and misapplications have been repeatedly pointed out (Hilborn and Walters, 1992; Caddy and Mahon, 1998; Punt and Smith, 2001; Hilborn, 2007b; Kesteven, 1997; Holt, 2009). For example, MSY cannot be determined for a stock unless this stock is overexploited, that is, the top of the parabola (=MSY) needs to be well surpassed for it to be determined. Therefore, MSY and optimum fishing effort cannot be predicted in early stages of developing fisheries and stock assessments should focus on detecting it as rapidly as possible (Hilborn and Walters, 1992). Once MSY is detected, fishing effort should be reduced by up to 30% in order to achieve sustainability (Hilborn and Walters, 1992). Also, the assumption of average recruitment and average environmental conditions may lead to wrong advice in highly fluctuating stocks (Hilborn and Walters, 1992). The dependence of MSY on size at first capture and age structure in the stock is ignored (Longhurst, 2006; Holt, 2009; Anderson et al., 2008). MSY is achieved by setting limits on fishing mortality but, as different fishing gears select and target different composition of species and some species are not targeted at all, MSY is rarely attained simultaneously for all species within an area (Maunder, 2002). The social aspects mentioned in UNCLOS (1982) have often been misunderstood as allowing for temporary overfishing to secure employment. The recovery potential of depleted stocks is overestimated by the simple parabola (Quinn and Deriso, 1999; Hutchings and Reynolds, 2004).

It is argued that surplus production models are too simple to fully describe the dynamics of populations subject to variability in recruitment, interactions with other species, catchability, selectivity, environmental conditions, and changing climate (Pella and Tomlinson, 1969) and they require a good contrast between fishing effort and stock abundance (Hilborn and Walters, 1992).

Improving MSY

The epitaph for MSY of Larkin (1977) was rather premature (Barber, 1988) as it was referring to the early, simplistic application of MSY that was considered a viable fishing target with constant catch removal. MSY has been conceptually transformed through time and improved (Kesteven, 1997; Mace, 2001) to become a limit that should be avoided (*target reference point* refers to a desirable state at which management should aim while *limit reference point* refers to an undesirable state which management should avoid: Caddy and Mahon, 1998), which brings the MSY concept in line with contemporary scientific views (Mace, 2001; Mesnil, 2012). Concerning the social issues, there is no conceivable scenario where overfishing is good for society because it results in subsequent lower catches and food supply, and lower future profits and fewer jobs in the sector.

The wide-spread critique that MSY ignores environmental conditions and species interactions is actually overstated (Froese et al., 2017), because the key parameter r_{max}, the maximum intrinsic rate of population growth, summarizes in a single value the interplay of natural mortality (caused mostly by predation), somatic growth (driven by food availability), and recruitment (strongly determined by environmental conditions). In other words, environmental and climatic effects are summarized in their impact on the survival of adults, that is, natural mortality (M), the availability and nutritional value of food, and the effort associated with acquiring it are summarized in somatic growth (k), and the interannual variability in environmental conditions that determine the survival of eggs and larvae are summarized in recruitment (i.e., the number of individuals surviving to join the exploited population) (Pauly and Froese, 2014). In other words, varying food availability, interspecific relationships, environmental/climate changes, and selectivity of the fishing gear are all incorporated in r_{max}. Increasing size at first capture will increase MSY, overfishing of prey species will decrease MSY.

Because of species interactions such as competition for resources and predator–prey relationships it is not possible for all populations to deliver MSY at the same time (Walters et al., 2005). But achieving, for example, 90% of MSY for all commercial fish and shellfish will already result in a substantial overall reduction of anthropogenic mortality for most target and nontarget species and will restore their biomasses to levels that should allow them to fulfill their roles as prey and predators in the ecosystem (Mace, 2001).

Forage fish (anchovies, herrings, sardines, and sand eels) are the crucial link between lower and upper trophic levels in the food web because they transport energy from millimeter-sized phytoplankton and zooplankton to the larger fish eaters of the food web (Baxter, 1997; Pikitch et al., 2012). For that reason they must be fished less and should be used for human consumption rather than for animal feed (Froese et al., 2016a).

Conclusion

It is now well established that fisheries management failed to preserve fish populations and some scientists have blamed it on the MSY concept (e.g., Mesnil, 2012). But is it a matter of science or a matter of administration and policy if stocks are in bad shape? So far, MSY has not been proven wrong as a concept but its estimation was not always correct and the administrative measures taken for its adoption were often inadequate or inappropriate (Kesteven, 1997). After its reform and continuous update, MSY remains a useful concept and a realistic approach to fisheries management and administration (Kesteven, 1997) and according to an anecdotal quote attributed to John Gulland "MSY is the most important concept in fisheries management" (Mangel et al., 2002). On top of that MSY carries a simple message that appears sensible to politicians and stimulates support by the public (Mesnil, 2012) and for that reason it is still widely used in assessing stock status and exploitation. It can be easily improved by considering size structure and setting catch length (L_C) close to optimum length (L_{OPT}) (Froese et al., 2016b). If $F < F_{MSY}$ and L_C is close to L_{OPT}, "pretty good" catches below but close to MSY are possible, with minimized impact on stock and environment. Pretty good yield (PGY) is a term introduced by Alec MacCall (National Marine Fisheries Service, Santa Cruz, CA, United States, retired) in 2000, proposing catches of about 80% of MSY as a meaningful and realistic target.

Acknowledgment

The authors would like to thank Daniel Pauly, Kostas Stergiou, and Henning Winker for their helpful suggestions and comments.

References

Anderson CNK, Hsieh C-h, Sandin SA, Hewitt R, Hollowed A, Beddington J, May RM, and Sugihara G (2008) Why fishing magnifies fluctuations in fish abundance. *Nature* 452: 835–839.

Bailey KM and Houde ED (1989) Predation on eggs and larvae of marine fishes and the recruitment problem. *Advances in Marine Biology* 25: 1–83.

Baranov FI (1918) On the question of the biological basis of fisheries. *Nauchnyĭ issledovatelskiĭ iktiologisheskiĭ Institut, Izvestiia* 1: 81–128.

Barber WE (1988) Maximum sustainable yield lives on. *North American Journal of Fisheries Management* 8: 153–157.

Baxter BS (ed.) (1997) *Forage fishes in marine ecosystems. Proceedings of the international symposium on the role of forage fishes in marine ecosystems, 13-16 November 1996.* University of Alaska Sea Grant Report. 97-01.

Beverton RJH and Holt SJ (1957) On the dynamics of exploited fish populations. *Fisheries Investigation II*, **XIX**: 1–238.

Caddy JF and Mahon R (1998) Reference points for fisheries management. *FAO Fisheries Technical Paper* 347: 1–83.

Chapman WM (1949) United States Policy on high seas fisheries. *Department of State Bulletin XX* 498: 67–80.

Charnov EL (1993) *Life history invariants*. Oxford: Oxford University Press.

Cleghorn J (1854) On the fluctuations in the herring fisheries. *British Association for Advancement of Science* 24: 124.

Costello C, Ovando D, Hilborn R, Gaines SD, Descenes O, and Lester SE (2012) Status and solutions for the world's unassessed fisheries. *Science* 338: 517–520.

DAFF (2007) *Commonwealth fisheries harvest strategy: Policy and guidelines*. Australian Government, Department of Agriculture, Fisheries and Forestry. 55 p.

Dickey-Collas M, Nash RDM, Brunel T, van Damme CJG, Marshall CT, Payne MR, Corten A, Geffen AJ, Peck MA, Hatfield EMC, Hintzen NT, Enberg K, Kell LT, and Simmonds EJ (2010) Lessons learned from stock collapse and recovery of North Sea herring: A review. *ICES Journal of Marine Science* 67: 1875–1886.

EC (2009) *Green paper: Reform of the common fisheries policy*. Brussels: EC. COM 163, http://ec.europa.eu/fisheries/reform.

EC (2013) *Common fisheries policy, (CFP), "regulation (EU) no 1380/2013 of the European Parliament and of the council of 11 December 2013 on the common fisheries policy, amending council regulations (EC) no 1954/2003 and (EC) no 1224/2009 and repealing council regulations (EC) no 2371/2002 and (EC) no 639/2004 and council decision 2004/585/EC"*. OJ L **354** (2013).

FAO (2001) *FAO Fisheries glossary*. http://www.fao.org/fi/glossary/default.asp.

Finley C (2011) *All the fish in the sea: Maximum sustainable yield and the failure of fisheries management*. Chicago: The University of Chicago Press.

Fox WW (1970) An exponential surplus-yield model for optimizing exploited fish populations. *Transactions of the American Fisheries Society* 99: 80–88.

Froese R, Branch TA, Proelß A, Quaas M, Sainsbury K, and Zimmermann C (2011) Generic harvest control rules for European fisheries. *Fish and Fisheries* 12: 340–351.

Froese R, Walters C, Pauly D, Winker H, Weyl OLF, Demirel N, Tsikliras AC, and Holt SJ (2016a) A critique of the balanced harvesting approach to fishing. *ICES Journal of Marine Science* 73: 1640–1650.

Froese R, Winker H, Gascuel D, Sumaila UR, and Pauly D (2016b) Minimizing the impact of fishing. *Fish and Fisheries* 17: 785–802.

Froese R, Demirel N, Coro G, Kleisner KM, and Winker H (2017) Estimating fisheries reference points from catch and resilience. *Fish and Fisheries* 18: 506–526.

Garstang W (1900) The impoverishment of the sea. *Journal of the Marine Biological Association of the UK* 6: 1–69.

Graham M (1935) Modern theory of exploiting a fishery, and application to North Sea trawling. *Journal de Conseil International pour l'Exploration de la Mer* 10: 264–274.

Graham M (1943) *The fish gate*. London: Faber and Faber.

Hart PJB and Reynolds JD (eds.) (2002) *Handbook of fish biology and fisheries, Vol. 2: Fisheries*. UK: Blackwell Publishing.

Hilborn R (2007a) Moving to sustainability by learning from successful fisheries. *Ambio* 36: 296–303.

Hilborn R (2007b) Defining success in fisheries and conflicts in objectives. *Marine Policy* 31: 153–158.

Hilborn R and Walters C (1992) *Quantitative fisheries stock assessment: Choice, dynamics and uncertainty*. Dordrecht: Springer Science.

Hjort J, Jahn G, and Ottestad P (1933) The optimum catch. *Hvalradets Skrifter* 7: 92–127.

Holt SJ (2009) Sunken billions—But how many? *Fisheries Research* 97: 3–10.

Holt SJ (2014) The graceful sigmoid: Johan Hjort's contribution to the theory of rational fishing. *ICES Journal of Marine Science* 71: 2008–2011.

Houde ED (1987) Fish early life dynamics and recruitment variability. *American Fisheries Society Symposium* 2: 17–29.

Hüssy K, St. John MA, and Böttcher U (1997) Food resource utilization by juvenile Baltic cod *Gadus morhua*: A mechanism potentially influencing recruitment success at the demersal juvenile stage? *Marine Ecology Progress Series* 155: 199–208.

Hutchings JA and Reynolds JD (2004) Marine fish population collapses: Consequences for recovery and extinction risk. *Bioscience* 54: 297–309.

Huxley TH (1884) Inaugural address. *Fisheries Exhibition Literature* 4: 1–22.

Kesteven GL (1997) MSY revisited. *Marine Policy* 21: 73–82.

Kyle HM (1905) Statistics of the North Sea fisheries. Part II. Summary of the available fisheries statistics and their value for the solution of the problem of overfishing. *Rapports, Conseil Permanent International pour l'Exploration de la Mer* **3**.

Larkin PA (1977) An epitaph for the concept of maximum sustained yield. *Transactions of the American Fisheries Society* 106: 1–11.

Longhurst A (2006) *Mismanagement of marine fisheries*. Cambridge: Cambridge University Press.

Mace PM (2001) A new role for MSY in single-species and ecosystem approaches to fisheries stock assessment and management. *Fish and Fisheries* 2: 2–32.

Malthus TR (1798) *An essay on the principle of population*. London: J. Johnson, in St. Paul's Church-yard.

Mangel M, Marinovic B, Pomeroy C, and Croll D (2002) Requiem for Ricker: Unpacking MSY. *Bulletin of Marine Science* 70: 763–781.

Maunder MN (2002) The relationship between fishing methods, fisheries management and the estimation of maximum sustainable yield. *Fish and Fisheries* 3: 251–260.

Mesnil B (2012) The hesitant emergence of maximum sustainable yield (MSY) in fisheries policies in Europe. *Marine Policy* 36: 473–480.

MFNZ (2008) *Harvest strategy standard for New Zealand fisheries*. Wellington, New Zealand: Ministry of Fisheries. 27 p, www.fish.govt.nz.

MSA (2006) Magnuson-Stevens Fishery Conservation and Management Reauthorized Act. In: *Public Law*, pp. 109–479. www.nero.noaa.gov/sfd/MSA_amended_20070112_FINAL.pdf.

Pauly D and Froese R (2014) Fisheries Management. In: *eLS*. Chichester: John Wiley & Sons.

Pella JJ and Tomlinson PK (1969) A generalized stock production model. *Bulletin of the Inter-American Tropical Tuna Commission* 13: 421–458.

Petersen CGJ (1903) What is overfishing? *Journal of the Marine Biological Association* 6: 587–594.

Pikitch E, Boersma PD, Boyd IL, Conover DO, Cury P, Essington T, Heppell SS, Houde ED, Mangel M, Pauly D, Plagányi É, Sainsbury K, and Steneck RS (2012) *Little fish, big impact: Managing a crucial link in ocean food webs*. Washington, DC: Lenfest Ocean Program. 108 pp.

Polacheck T, Hilborn R, and Punt AE (1993) Fitting surplus production models: Comparing methods and measuring uncertainty. *Canadian Journal of Fisheries and Aquatic Science* 50: 2597–2607.

Punt AE (2003) Extending production models to include process error in the population dynamics. *Canadian Journal of Fisheries and Aquatic Sciences* 60: 1217–1228.

Punt AE and Smith ADM (2001) The gospel of maximum sustainable yield in fisheries management: Birth, crucifixion and reincarnation. In: Reynolds JD (ed.) *Conservation of exploited species*, pp. 41–66. Cambridge: Cambridge University Press.

Quinn TJ and Deriso RB (1999) *Quantitative fish dynamics*. New York: Oxford University Press.

Ricker WE (1975) Computation and interpretation of biological statistics of fish populations. *Bulletin of the Fisheries Research Board of Canada* 191: 1–382.

Schaefer MB (1954) Some aspects of the dynamics of populations important to the management of the commercial marine fisheries. *Bulletin of the Inter-American Tropical Tuna Commission* 1: 25–56.

Seidl I and Tisdell CI (1999) Carrying capacity reconsidered: From Malthus' population theory to cultural carrying capacity. *Ecological Economics* 31: 395–408.

Sibly RM, Brown JH, and Kodric-Brown A (2012) *Metabolic ecology: A scaling approach*. UK: Wiley-Blackwell.

Sissenwine MP (1978) Is MSY an adequate foundation for optimum yield? *Fisheries* 3: 22–42.

Smith T (1994) *Scaling fisheries: The science of measuring the effects of fishing, 1855–1955*. Cambridge: Cambridge University Press.

UN (1995) *Agreement for the implementation of the provisions of the United Nations convention on the law of the sea of 10 December 1982, relating to the conservation and management of straddling fish stocks and highly migratory fish stocks. United Nations conference on straddling fish stocks and highly migratory fish stocks*. New York: United Nations. 37 p.

UNCLOS (1982) *The law of the sea. Official text of the United Nations Convention on the Law of the Sea with Annexes and tables*. New York: United Nations. 224 p.

Verhulst P-F (1838) Notice sur la loi que la population suit dans son accroissement. *Correspondance mathématique et physique* 10: 113–121.

Walters CJ, Christensen V, Martell SJ, and Kitchell JF (2005) Possible ecosystem impacts of applying MSY policies from single-species assessment. *ICES Journal of Marine Science* 62: 558–568.

Micro- and Macroplastics in Aquatic Ecosystems

Imogen E Napper and Richard C Thompson, Plymouth University, Plymouth, United Kingdom

Glossary

Contaminant A foreign substance or impurity that is introduced into an environment where it is not normally found.
Ghost fishing This results from abandoned, lost or discarded fishing gear (ALDFG) in the environment. This gear continues to fish, trap and potentially kill animals, smother habitats, and act as a hazard to navigation.
Marine litter Marine litter is human-created waste that has been deliberately discarded, unintentionally lost or transported by wind and rivers into the sea and on beaches. It mainly consists of plastics, wood, metals, glass, rubber, clothing, and paper. Land-based sources account for up to 80% of marine litter—these include tourism, sewage, and illegal or poorly managed landfills. The main sea-based sources are shipping and fishing.
Microplastic Microplastics are small plastic pieces less than 5 mL in diameter.
Persistent organic pollutants Hazardous organic chemical compounds that are resistant to biodegradation and therefore remain in the environment for a long period of time.
Plastic Plastics are synthetic polymers made from organic molecules. They are typically derived from fossil oil and gas reserves, but they can also be made from renewable sources such as plants. Plastics are light weight, inexpensive, durable and versatile materials that can be used for many different applications.
Pollutant A harmful or hazardous substance, which can cause an adverse effect in the environment.

Introduction

Plastics are synthetic or semi-synthetic organic polymers. They are typically lightweight, strong, inexpensive, durable, and corrosion-resistant (Derraik, 2002; Thompson et al., 2009b). Most plastic items are composed of hydrocarbons derived from fossil oil or gas feedstocks (American Chemistry Council, 2015). During the conversion from resin to product, a wide variety of additives (such as fillers, plasticizers, flame retardants, thermal stabilizers, antimicrobial agents, and coloring agents) may be added to enhance performance and appearance (Andrady and Neal, 2009). As a consequence, plastic materials can take many forms including rigid items together with more flexible films, adhesives, foams, and fibers. The most commonly used polymers are high-density polyethylene (HDPE), low-density polyethylene (LDPE), polyvinyl chloride (PVC), polystyrene (PS), polypropylene (PP), and polyethylene terephthalate (PET), which cumulatively account for approximately 90% of total plastic production. These plastics bring a wide range of societal benefits in healthcare, agriculture, transport, construction, and packaging (PlasticsEurope, 2016). The versatility of plastic materials has resulted in a substantial increase in their use from 5 million tons globally in the 1950s to over 300 million tons today (Andrady and Neal, 2009; PlasticsEurope, 2015).

Despite the durability of plastics, the main uses are in relatively short-lived applications, such as packaging, which accounts for around 40% of all production. While packaging can help protect food, drink, and other items (thus reducing damage and wastage of products), it also results in rapid accumulation of persistent plastic waste. This has led to one of the most ubiquitous and long-lasting recent changes to the surface of our planet; the accumulation and fragmentation of plastic debris (Barnes et al.,

2009). Plastics represent a substantial fraction of the municipal waste stream and around 75% of all marine litter is plastic. This debris is widely reported in the environment where it has accumulated at the sea surface (Law et al., 2010), on shorelines of even the most remote islands (Barnes, 2005), in the deep sea (Bergmann and Klages, 2012; Woodall et al., 2014), and in arctic sea ice (Obbard et al., 2014). There is also increasing awareness of the accumulation of plastic litter on land as well as in freshwater habitats (Eerkes-Medrano et al., 2015). The accumulation of this litter presents a range of negative economic and environmental consequences (Werner et al., 2016).

Once in the environment, exposure to ultra-violet radiation, heat and oxygen can cause plastics to become brittle and physical action can then cause them to break down into smaller pieces, including microplastics. The timescale for degradation of discarded plastics is not known with certainty and will depend on the chemical nature of the material, the characteristics of the environment in which they persist and the manner in which degradation is measured (Andrady and Neal, 2009). However, some polymer chemists suggest that all of the conventional plastic that has ever been produced, with the exception of any material that has been incinerated, still persists in the environment in a form too large to be biodegraded (Thompson et al., 2005).

Definitions of Plastic Litter

Plastic debris can be defined and described in a variety of ways including by shape, color, polymer type, origin and original usage (e.g., packaging). Plastics enter the aquatic environment in a wide range of sizes (Cole et al., 2011; Hidalgo-Ruz et al., 2012) and have been reported hundreds of meters in length to microns in diameter. There are three categories that are typically used to describe the size of plastic contamination; macroplastic (>20 mm diameter), mesoplastic (5–20 mm), and microplastic (<5 mm) (Barnes et al., 2009; Thompson et al., 2009a). Although there is uncertainty about absolute quantities of plastic in the environment, or the ultimate sinks for this debris, there is evidence of increasing quantities over time (Jambeck et al., 2015; Thompson et al., 2004).

Macro and Mesoplastic

Macroplastic refers to plastic items larger than 20 mm. Due to its high visibility, contamination of the environment by macroplastic may be perceived as one of the most concerning forms of plastic pollution. The accumulation of macroplastic has been reported in a wide range of habitats (Browne, 2015; Ryan et al., 2009). Clean-up campaigns typically focus on these larger items and there is wide geographical variability in abundance, which increases the difficulty of analyzing potential trends. However, due to the size of this debris, it is often possible to categorize items according to their original usage; for example, packaging, fishing or sewage related debris. Plastic debris that is larger than 5 mm but smaller than 20 mm is termed mesoplastic.

Microplastic

Microplastic is used as a collective term to describe a heterogeneous range of small plastic particles and fibers. In 2008, the National Oceanographic and Atmospheric Agency (NOAA) of the United States hosted the first International Microplastics Workshop and as part of this meeting formulated a working definition to include all plastic particles less than 5 mm in diameter (Arthur et al., 2009). The lower size limit is typically set by the capacity of capture methodology or analytical identification equipment and is currently around 20 μm. However, it is widely believed that plastic debris is present in the environment in the nano-size range (Mattsson et al., 2015). As with macroplastics, microplastics can differ in specific density, chemical composition, and shape (Fig. 1B) (Duis and

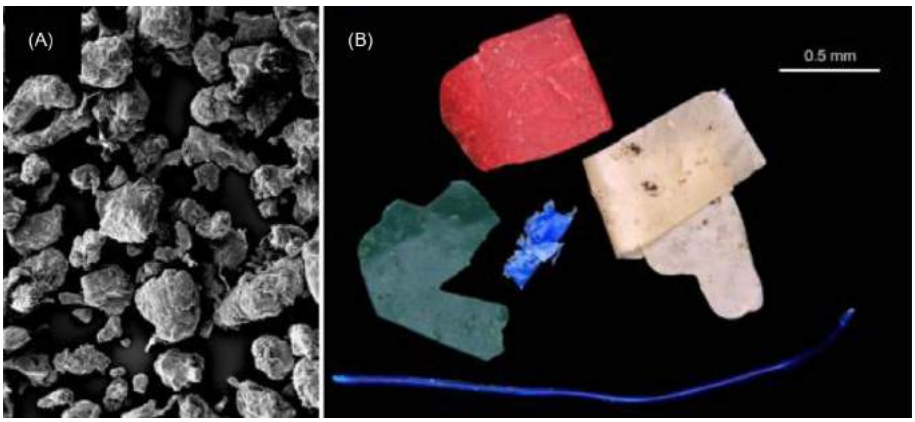


Fig. 1 Pictures showing typical microplastic samples: (A) Polyethylene particles extracted from a cosmetic product. (B) Fragments of microplastic collected from a shoreline near to Plymouth, UK. *Note*: Scale bar applies to both pictures. (A) Napper & Thompson, Plymouth University Electron Microscopy Suite. (B) Thompson, Plymouth University.

Coors, 2016; Law and Thompson, 2014). Microplastic can then be further divided into categories based on their origin; primary and secondary microplastics.

Primary microplastic

Particles that directly enter the environment in the microplastic size (<5 mm in diameter) are described as primary microplastics. Primary microplastics are produced through extrusion or grinding, either as a feed stock for manufacture of products (Turner and Holmes, 2015) or for direct use (Browne, 2015); for example in cleaning products (Cole et al., 2011; Derraik, 2002), cosmetics (Fig. 1A) (Napper et al., 2015; Zitko and Hanlon, 1991), and as air-blasting media (Gregory, 1996).

Secondary microplastic

Secondary microplastics are those formed in the environment from the fragmentation of larger items of plastic debris (Cole et al., 2011; Law and Thompson, 2014). This degradation occurs as a consequence of ultra-violet (UV) radiation and oxidation, which overtime can reduce the structural integrity of the plastic, resulting in fragmentation. This can be facilitated by physical forces from abrasion, wave-action, and turbulence (Barnes et al., 2009; Browne et al., 2007). The process is ongoing, with fragments becoming smaller and smaller over time (Cole et al., 2011; Galgani et al., 2010). Even if emissions of larger items of plastic to the environment were to cease with immediate effect, it is likely that we would still see an increase in the quantity of microplastic as a consequence of the fragmentation of larger items that are already in the environment. Secondary microplastic can also be generated as a consequence of wear during the use of a product. For example, fibers generated from the laundering of clothes or from the wear of tyres.

Sources

The majority of plastic in the sea originates from inland sources and is emitted to the oceans from coastlines or by rivers (Jambeck et al., 2015). In addition, quantities are released from ocean-based sources such as shipping and aquaculture (Andrady, 2011; GESAMP, 2015). Smaller particles may also be carried in the air and deposited at sea (Dris et al., 2015). It has been estimated that on a global scale, the input of plastic into the oceans from land based sources is in the region of 6.4 million tons per annum (Jambeck et al., 2015). Furthermore, assuming there are no improvements in waste management infrastructure, the cumulative quantity of plastic waste available to enter the marine environment from land could increase by approximately three times over the next decade (Fig. 2) (Jambeck et al., 2015). However, a more precise estimate will require direct measurement of the input rates of plastic waste by wind, tidal and ocean wave transport. It will also require a methodical measurement of waste generation, collection rates, classification, and waste disposal methods for rural areas and urban centers in countries around the world (Law, 2017). Microplastics have been detected at very high levels globally in rivers and lakes which could further add to this estimation (Auta et al., 2017; Free et al., 2014; McCormick et al., 2016).

Sources of Macroplastic

Much of the litter in aquatic environments enters as macroplastic from land-based actions such as general littering, dumping of waste and loss during waste collection as well as that from inappropriately managed landfill sites (Jambeck et al., 2015; Mehlhart

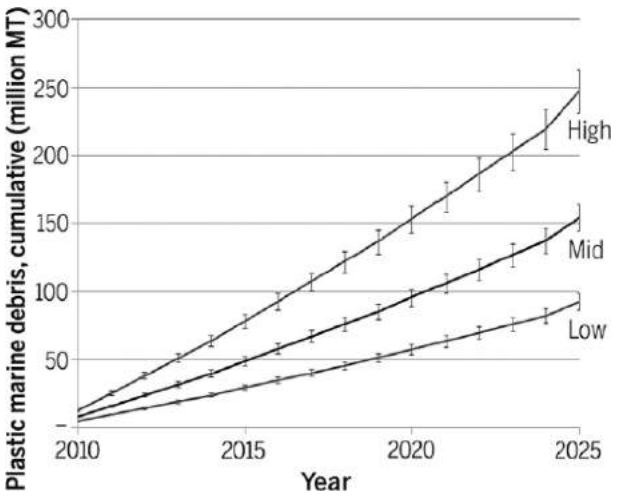


Fig. 2 The estimated mass of mismanaged plastic waste (millions of metric tons) input into the ocean by populations living within 50 km of a coast in 192 countries, plotted as a cumulative sum from 2010 to 2025 (Jambeck et al., 2015).

and Blepp, 2012). Plastic waste is collected, and then contained in a waste management framework which is designed to help minimize loss to the environment. From these land-based sources, plastic litter then has the potential to end up in municipal wastewater and freshwater systems (Cole et al., 2011; Leslie et al., 2013). This can result from windblown litter escaping into the wider environment (Barnes et al., 2009; Mehlhart and Blepp, 2012; Pruter, 1987). In industrialized countries, waste that is deposited in landfills is usually covered regularly with soil or a synthetic material, and the landfill is cordoned by a fence to prevent any debris accidentally leaving. However, in developing regions this is often not the case (Barnes et al., 2009; Jambeck et al., 2015). The residues from plastic recycling could also unintentionally escape into the environment (Moore, 2008).

Some items of plastic litter are also released directly from industrial activities, such as commercial fishing. Studies have indicated a significant relationship between the number of ocean-based plastic items found on beaches and the level of commercial fishing activity (Cunningham and Wilson, 2003; Ribic et al., 2010). In 2010, the amount of fishing gear lost to the environment was estimated at around 640,000 tons per year (Good et al., 2010). Discarded fishing items, which include monofilament lines and nylon netting, can float at a variety of depths and result in "ghost fishing" and entanglement of aquatic organisms (Good et al., 2010). Furthermore, unintentional loss of in-service macroplastic products can occur when catastrophic events, such as tsunamis, hurricanes, or floods, carry large amounts of material of all kinds into the marine environment (Law, 2017).

Sources of Microplastic

Primary microplastic can result from spillage/mishandling of industrial preproduction plastics or from the use of cosmetics (Duis and Coors, 2016; Law, 2017; Napper et al., 2015). Plastic microbeads from facial scrubs are an example of a cosmetic use source. After their intended use, these microbeads are likely to enter household wastewater and some will escape the waste water treatment system into the environment (Murphy et al., 2016). It has been estimated that 94,500 microbeads could be released from an exfoliant in a single use, accumulating to the United Kingdom alone to be emitting 16–86 ton per year (Napper and Thompson, 2016). Other potentially important sources of microplastics are from microplastic used in medicines, drilling fluids for oil/gas exploration and in industrial abrasives (i.e., for air-blasting to remove paint from metal surfaces) (Derraik, 2002; Duis and Coors, 2016; Gregory, 1996; Mintenig et al., 2014; Sundt et al., 2014).

Further sources of microplastic to the marine environment occur as a consequence of the breakdown of larger plastic debris (secondary microplastic). These can then enter the marine environment through two different pathways; a "direct" (sewage or storm water) or an "indirect" (fragmentation of existing plastic debris) source. Washing of clothes made from synthetic materials is an example of a direct secondary microplastic source. Again, this microplastic can enter the environment via wastewater after the release of fibres from a washed garment. Some fabrics release fibres more readily than others; research by Napper and Thompson (2016) reported that a wash load of 6 kg of acrylic clothing could release over 700,000 fibres.

For primary and secondary microplastic pieces larger than around 20 μm, it is possible to identify what type of plastic polymer a particular piece of marine debris is made out of (Thompson et al. 2004). However, it is extremely difficult to trace back the debris to its origin.

Waste Water Treatment Plants

For any plastic that enters waste water treatment plants, the efficiency of capture (i.e., before the effluent is discharged into the environment) depends on the particular treatment process. However, there is limited information on the efficiency of waste water treatment plants to capture plastic; particularly microplastic. Some studies indicate extremely high capture rates (>95%) of plastic particles (Carr et al., 2016; Murphy et al., 2016). Given the large volume of influent daily, even low loss rates could result in detectable concentrations of these plastic particles in the environment (Browne et al., 2011; Eriksen et al., 2013). Murphy et al. (2016) predicted that waste water treatment plants could release 65 million microplastics every day (Murphy et al., 2016). Wastewater and any plastic debris therein can also bypass treatment as a consequence of sewage overflows. Even if microplastic is intercepted during wastewater treatment, the resultant sewage sludge is often returned to the land as a fertilizer, hence plastic is still released to the environment (Kirchmann et al., 2017).

Source Trends

The trends of production, consumer-use and demographics all point to a further increase of plastic use in the future (Auta et al., 2017; Sutherland et al., 2010). Hence there are considerable concerns that the problems of plastic pollution will escalate unless disposal practices change. Given the large amount of macroplastics entering the environment, it is generally assumed that most microplastics have arisen from the fragmentation of larger items (Andrady, 2011; Hidalgo-Ruz et al., 2012). The fragmentation rates of macroplastics are largely unknown, and as a result little quantitative information is available on the relative contribution of primary and secondary microplastics to the overall amount in the environment (Koelmans et al., 2014; Law and Thompson, 2014; Sundt et al., 2014). It is also clear that substantial quantities of fibres have accumulated in the environment (Dris et al., 2015; Mathalon and Hill, 2014). These are widely reported in microplastic samples and likely originate from abrasion of textiles during use or as a consequence of laundering (Browne et al., 2011; Mathalon and Hill, 2014; Napper and Thompson, 2016).

Distribution

Plastic debris has been reported at the sea surface (Cózar et al., 2014), suspended in the water column (Lattin et al., 2004) and in sediments, including those in the deep sea (Fischer et al., 2015; Van Cauwenberghe et al., 2013; Woodall et al., 2014). Airborne transport may also be relevant for very small microplastics; for example, microplastics escaping from uncovered landfills (Rillig, 2012), or the dispersal of particles formed by wear in service such as textile (Napper and Thompson, 2016) and tyre wear (Kole et al., 2017).

Plastic debris of all sizes can accumulate in the oceans (Thompson et al., 2004), estuaries (Browne et al., 2010), and even in remote locations such as in arctic ice (Obbard et al., 2014). Plastic debris has also been found in freshwater environments showing that this is an aquatic system issue, not just limited to the marine environment (Ivleva et al., 2017; Mani et al., 2016; Wagner et al., 2014). However, there are no definitive estimates of the total quantity of litter in the environment. Estimates based on counts at sea suggest between 7000 and >250,000 tons of plastic are now present in the open-ocean surface (Cózar et al., 2014; Eriksen et al., 2014). Per annum it has been predicted that 6.4 million tons enters the ocean from land based sources (Jambeck et al., 2015).

Rivers can transport considerable quantities of plastic (micro–macro-size) to the oceans and some of this debris can travel from locations far inland. In rivers, microplastic concentrations have been found to vary along and across the river, reflecting various sources such as waste water treatment plants, tributaries, and weirs (Claessens et al., 2011; Klein et al., 2015; Mani et al., 2016). Once released into the environment, and due to their durability, plastic debris has the potential to become widely dispersed via wind and currents (Faure et al., 2015; Lambert et al., 2014; Ryan et al., 2009).

At the water surface, smaller pieces of plastic present lower rise velocities and are more susceptible to vertical transport (Reisser et al., 2015; Song et al., 2014). Some polymers such as polyvinyl chloride (PVC), and polyethylene terephthalate (PET), are denser than water and are more likely to sink, while polyethylene (PE), polypropylene (PP), and polystyrene (PS) are more likely to float. However, microplastics typically accumulate fouling from microorganisms as well as sediment particles on their surface. Over time this increases their apparent density causing even some of the less dense polymers to sink (Zettler et al., 2013). Hence, the sea bed could be the most likely long-term place for the accumulation of plastic debris.

Impacts

There is a reasonably extensive evidence base relating to the harm caused by marine litter. This can have a range of negative impacts on commercial fisheries, maritime industries, and infrastructure. It has also been found to affect a wide range of marine organisms; as a consequence of entanglement and ingestion (Gall and Thompson, 2015; Sutherland et al., 2010; Wang et al., 2016). Over 700 species of marine organisms have been reported to encounter plastic debris, which can result in severe physical harm and death, or more subtle effects on behavior and ecological interactions (e.g., the ability to escape from predators or migrate) (Gall and Thompson, 2015). It is likely that there are also a range of sublethal effects that have not yet been recognized.

Impacts of plastic vary according to the type and size of the debris, and can occur at different levels of biological organization in a wide range of habitats (Browne, 2015). Encounters between plastic litter and organisms can negatively affect individuals and a substantial proportion of some populations can be contaminated with plastics; for example, over 40% of sperm whales beached on North Sea coasts had marine litter including, ropes, foils, and packaging material found in their gastrointestinal tract (Unger et al., 2016), while over 95% of the population of northern fulmars (*Fulmar glacialis*) may contain plastic litter in some European waters (Van Franeker et al., 2011).

Over 330 species (50 marine mammals) have been found to become entangled or ingested plastic debris (Kühn et al., 2015). However, evidence of harm from entanglement is easier to observe and report than ingestion. This is because ingestion typically only becomes apparent when the carcass of an animal opens; either as a result of dissection or decomposition. It has been reported from UK surveys that there is an incidence rate of entanglement between 2% and 9% for some populations of seabirds and marine mammals (Werner et al., 2016).

The ingestion of meso- or macroplastic litter has been reviewed for numerous marine species; particularly mammals, birds, and turtles (Derraik, 2002; Gall and Thompson, 2015; Gregory, 2009; Kühn et al., 2015; Laist, 1997). Studies have also shown that both freshwater invertebrates and fish ingest microplastic (Imhof et al., 2013; Phillips and Bonner, 2015). The potential for ingestion is greater with pieces in the microplastic size range since their small size makes them readily accessible for ingestion by a wide range of organisms (Davison and Asch, 2011) including whales, fishes, mussels, oysters, shrimps, copepods, and lugworms (Cole et al., 2013; Ferreira et al., 2016; Lusher et al., 2015a,b).

However, ingestion also depends on properties other than size including shape, density and color. For instance, low-density (i.e., buoyant) microplastics are potentially more likely to be ingested by pelagic feeders and high-density microplastics by benthic feeders (Scherer et al., 2017; Wright et al., 2013). With very small particles, including those in the nano-size range, there is also the potential for uptake across cell membranes, but little is known about any associated impacts (Koelmans et al., 2014). Organisms at lower trophic levels can ingest and accumulate microplastic particles, and it has been shown that microplastics can transfer between trophic levels in the food-web (Watts et al., 2014). Ingestion has shown to lead to physical effects that include physiological stress responses and even signs of tumor formation (Rochman et al., 2013).

There are also concerns about the potential for ingestion to facilitate the transfer of chemicals to marine life (Bakir et al., 2014). Hydrophobic organic pollutants readily sorb onto plastics, and can accumulate at concentrations several orders of magnitude higher than in seawater (Mato et al., 2001). These chemicals can then be released to organisms upon ingestion. However, modelling estimates indicate the amount of chemical transfer from water to organisms via plastic is probably not a major pathway leading to harm (Bakir et al., 2016).

Additive chemicals incorporated into plastic products at the time of manufacture may also transfer to marine organisms upon ingestion (Tanaka et al., 2013). These chemicals are intentionally added during the manufacture or processing; for example to enhance the plastics durability and corrosion resistance (Andrady, 2016) or act as stabilizers, plasticizers or flame retardants. Some additives, such as plasticizers, are used at high concentrations (10%–50%) to ensure the functionality of the product. Degradation of plastic containing these additives may result in the additives leaching out and becoming bioavailable to organisms (Oehlmann et al., 2009; Talsness et al., 2009). Although, there is currently little evidence of harmful effects associated with the release of additives from plastic litter in the environment.

A further source of concern is colonization of organisms on plastic debris; species found on plastic debris can differ from the free-floating microbial communities in the oceans (Zettler et al., 2013). For example, microplastics collected in the surface waters of the North Atlantic were colonized by a variety of organisms including bacteria, cyanobacteria, diatoms, ciliates, and radiolaria (Zettler et al., 2013). As plastics have been reported to travel over long distances, they may contribute to the dispersal of alien or invasive species (Barnes, 2002).

Contamination of the marine environment with plastic debris can also have negative economic consequences on tourism, aquaculture, navigation, and fisheries. With fisheries, plastic litter can reduce or damage catches and vessels. For example, the total cost of removing litter of all types from 34 UK harbors was estimated at approximately £236,000. Based on this, it was estimated that marine litter costs the ports and harbor industry in the United Kingdom approximately £2.1 million each year (Mouat et al., 2010). Estimates indicate cleaning of UK beaches costs local authorities around £15.5 million (Mouat et al., 2010). Tourism can be affected as visitors to the marine environment consider litter annoying and hence litter can influence the locations they chose to visit (Brouwer et al., 2015). There is also emerging evidence that even small quantities of litter on beaches can have a negative effect on human well-being (Wyles et al., 2016).

Solutions

The potential threats to aquatic ecosystems presented by plastic debris have been identified as a major global conservation issue and a key priority for research (Sutherland et al., 2010). To fully understand the sources and scale of this contamination would require an internationally coordinated effort with comparable sampling and microplastic extraction techniques, as well as standardized recording methodologies to map and evaluate distribution (Waller et al., 2017). However it is clear that substantial quantities of litter are entering the aquatic habitats daily (Jambeck et al., 2015; Mani et al., 2016). Critically, it must be recognized that the accumulation of plastic in the oceans is actually a symptom of a wider more systemic problem of linear use of materials and the rapid accumulation of waste, including end of life plastic. Hence the overarching solutions to the problem of marine litter lie on land. Even in the absence of complete information on distribution and impacts, it is clear that the key action must be to reduce the quantity of litter entering the oceans from the land.

There are some management strategies and policies in place to reduce plastic contamination (GESAMP, 2015). Banning microbeads in cosmetics is an example of such legislation (Hirst and Baker, 2017). However, based on the levels of concern and the scale of problems outlined in this article it would appear that the measures currently in place are insufficient. In some cases, there are difficulties associated with enforcement; for example, the regulation of dumping at sea (MARPOL) is extremely difficult to enforce. Even in highly developed countries with robust waste management infrastructure, there are unnecessary obstacles to recycling, including the lack of availability of collection points, contamination of recycling feedstock, and the limited marketability of some recycled material (Andrady, 2005; Law, 2017).

Benefits of citizen focused activities such as beach cleaning are well recognized for their educational value as well as in terms of the litter removed (Nelms et al., 2016). Annual clean-up operations are now organized in many countries (Barnes et al., 2009) and are often run by voluntary organizations. They can remove substantial quantities of litter from beaches and the coastline. Volunteer involvement in two of the largest clean up schemes in the United Kingdom (Marine Conservation Society Beach Watch and Keep Scotland Beautiful National Spring Clean) has been estimated to provide a value of approximately £119,500 in term of cleaning, which suggests that the total cost of voluntary action to remove marine litter is considerable (Mouat et al., 2010).

Current rates of entry for litter into the marine environment far exceed the potential for removal by clean-up. Additionally, while awareness is being raised by opportunities associated with relatively low-cost beach cleaning activities, there are concerns about the viability and efficiency of large scale mechanical clean-up operations at sea. If clean-up is seen as a substantive solution, it must be acknowledged that there will be a need for such clean-up in perpetuity. Therefore, the main priority must be to focus on preventing litter entering the oceans in the first place and a better understanding in the behaviors that lead to littering, as well as those that lead to engagement in recycling (Pahl and Wyles, 2017). Most plastics are inherently recyclable, yet many single-use items are not compatible with recycling. A key challenge therefore is to ensure end-of-life disposal via recycling is appropriately considered at the design stage.

There are also some potential distractions to the key solutions; such as altering the carbon source used to make plastics by utilizing plant base carbon rather than fossil carbon from oil and gas. While this utilizes a renewable and hence a more sustainable carbon source, it will not reduce the generation of waste nor the accumulation of litter. Biodegradable plastics are another potential distraction; while products that have been designed to degrade rapidly may reduce the amount of highly visible macroplastic waste, many of these items merely fragment compromising the potential for product reuse and accelerating the production of microplastic fragments (Thompson et al., 2009b). Biodegradable or compostable plastics only present a solution in very specific settings where the associated waste collection is specifically managed, provides conditions suitable for degradation and products are labelled accordingly to facilitate appropriate disposal.

Education, outreach and awareness are effective ways to promote change in limiting indiscriminate disposal. However, in the past, approaches to address marine litter have mostly focused on end-of-pipe measures; in order to develop long term sustainable solutions there needs to be education and change in behavior right along the supply chain and this could be facilitated by greater dialogue between the various stakeholders from design, through production and use, to disposal. In short, what is needed is a much better stewardship so that the benefits of plastic can be realized without the accumulation of unnecessary waste in managed systems and in the environment.

Summary

While the problems of this litter are especially evident in aquatic ecosystems, most litter originates from land-based sources. Therefore, to help manage and reduce emissions it is essential to better understand the relative importance of these sources and to assess regional variation. Once in the ocean, plastic debris has the potential to travel considerable distances away from its original location. The debris can be micro to macroplastic in size; depending on its original size, rate of degradation, and source. It can have implications for wildlife and economic consequences on tourism, aquaculture, navigation, and fisheries. Education, outreach, and awareness will be effective in limiting the indiscriminate disposal of plastic waste. However, strategies and policies should also be in place to minimize the amount of plastic waste produced and promote appropriate waste management; for example, via recycling. For effective long-term change, it will be particularly important to apply these solutions internationally to minimize inputs of plastic litter into the ocean.

References

American Chemistry Council (2015) *Resin review*. Washington, DC: American Chemistry Council.

Andrady AL (2005) In: *Plastics in marine environment: A technical perspective*. Proceedings of the Plastic Debris Rivers to Sea Conference, California: Algalita Marine Research Foundation.

Andrady A (2011) Microplastics in the marine environment. *Marine Pollution Bulletin* 62: 1596–1605. https://doi.org/10.1016/j.marpolbul.2011.05.030.

Andrady AL (2016) *Environmental sustainability of plastics*. Chichester: John Wiley.

Andrady AL and Neal MA (2009) Applications and societal benefits of plastics. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 364: 1977–1984. https://doi.org/10.1098/rstb.2008.0304.

Arthur, C., Baker, J., Bamford, H., 2009. Proceedings of the International Research Workshop on the Occurrence, Effects And Fate Of Microplastic Marine Debris, September 9–11, 2008. NOAA Tech. Memo. NOS-OR&R30. University of Washington Tacoma, Tacoma, WA.

Auta HS, Emenike C, and Fauziah S (2017) Distribution and importance of microplastics in the marine environment: A review of the sources, fate, effects, and potential solutions. *Environment International* 102: 165–176. https://doi.org/10.1016/j.envint.2017.02.013.

Bakir A, Rowland SJ, and Thompson RC (2014) Enhanced desorption of persistent organic pollutants from microplastics under simulated physiological conditions. *Environmental Pollution* 185: 16–23. https://doi.org/10.1016/j.envpol.2013.10.007.

Bakir A, O'Connor IA, Rowland SJ, Hendriks AJ, and Thompson RC (2016) Relative importance of microplastics as a pathway for the transfer of hydrophobic organic chemicals to marine life. *Environmental Pollution* 219: 56–65. https://doi.org/10.1016/j.envpol.2016.09.046.

Barnes D (2002) Invasions by marine life on plastic debris. *Nature* 416: 808–809. https://doi.org/10.1038/416808a.

Barnes D (2005) Remote islands reveal rapid rise of southern Hemisphere Sea debris. *Scientific World Journal* 5: 915–921. https://doi.org/10.1100/tsw.2005.120.

Barnes D, Galgani F, Thompson RC, and Barlaz M (2009) Accumulation and fragmentation of plastic debris in global environments. *Philosophical Transactions of the Royal Society B* 364: 1985–1998. https://doi.org/10.1098/rstb.2008.0205.

Bergmann M and Klages M (2012) Increase of litter at the Arctic deep-sea observatory HAUSGARTEN. *Marine Pollution Bulletin* 64: 2734–2741. https://doi.org/10.1016/j.marpolbul.2012.09.018.

Brouwer, R., Galantucci, M., Hadzhiyska, D., Ioakeimidis, C., Leermakers, A., Ouderdorp, H., Boteler, B., Fernandez, P., 2015. Socio-economic assessment of the costs of marine litter, D 4.11. Developed under CleanSea project co-funded by the European Union Seventh Framework Programme under grant agreement no. 308370.

Browne MA (2015) Sources and pathways of microplastics to habitats. In: *Marine anthropogenic litter*, pp. 229–244. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-16510-3_9.

Browne MA, Galloway T, and Thompson R (2007) Microplastic—An emerging contaminant of potential concern? *Integrated Environmental Assessment and Management* 3: 559–561. https://doi.org/10.1002/ieam.5630030412.

Browne MA, Galloway TS, and Thompson RC (2010) Spatial patterns of plastic debris along estuarine shorelines. *Environmental Science & Technology* 44: 3404–3409. https://doi.org/10.1021/es903784e.

Browne MA, Crump P, Niven SJ, Teuten E, Tonkin A, Galloway T, and Thompson R (2011) Accumulation of microplastic on shorelines woldwide: Sources and sinks. *Environmental Science & Technology* 45: 9175–9179. https://doi.org/10.1021/es201811s.

Carr SA, Liu J, and Tesoro AG (2016) Transport and fate of microplastic particles in wastewater treatment plants. *Water Research* 91: 174–182. https://doi.org/10.1016/j.watres.2016.01.002.

Claessens M, De Meester S, Van Landuyt L, De Clerck K, and Janssen CR (2011) Occurrence and distribution of microplastics in marine sediments along the Belgian coast. *Marine Pollution Bulletin* 1: 2199–2204. https://doi.org/10.1016/j.marpolbul.2011.06.030.

Cole M, Lindeque P, Halsband C, and Galloway TS (2011) Microplastics as contaminants in the marine environment: A review. *Marine Pollution Bulletin*. https://doi.org/10.1016/j.marpolbul.2011.09.025.

Cole M, Lindeque P, Fileman E, Halsband C, Goodhead R, Moger J, and Galloway TS (2013) Microplastic ingestion by zooplankton. *Environmental Science & Technology* 47: 6646–6655. https://doi.org/10.1021/es400663f.

Cózar A, Echevarría F, González-Gordillo JI, Irigoien X, Ubeda B, Hernández-León S, Palma AT, Navarro S, García-de-Lomas J, Ruiz A, Fernández-de-Puelles ML, Duarte CM, Echevarria F, Gonzalez-Gordillo JI, Irigoien X, Ubeda B, Hernandez-Leon S, Palma AT, Navarro S, Garcia-de-Lomas J, Ruiz A, Fernandez-de-Puelles ML, and Duarte CM (2014) Plastic debris in the open ocean. *Proceedings of the National Academy of Sciences of the United States of America* 111: 10239–10244. https://doi.org/10.1073/pnas.1314705111.

Cunningham DJ and Wilson SP (2003) Marine debris on beaches of the greater Sydney region. *Journal of Coastal Research* 19: 421–430. https://doi.org/10.2307/4299182.

Davison P and Asch RG (2011) Plastic ingestion by mesopelagic fishes in the North Pacific Subtropical Gyre. *Marine Ecology Progress Series* 432: 173–180. https://doi.org/10.3354/meps09142.

Derraik JGB (2002) The pollution of the marine environment by plastic debris: A review. *Marine Pollution Bulletin* 44: 842–852. https://doi.org/10.1016/S0025-326X(02)00220-5.

Dris R, Gasperi J, Rocher V, Saad M, Renault N, and Tassin B (2015) Microplastic contamination in an urban area: A case study in greater Paris. *Environment and Chemistry* 12: 592–599. https://doi.org/10.1071/EN14167.

Duis K and Coors A (2016) Microplastics in the aquatic and terrestrial environment: Sources (with a specific focus on personal care products), fate and effects. *Environmental Sciences Europe* 28. 2https://doi.org/10.1186/s12302-015-0069-y.

Eerkes-Medrano D, Thompson RC, and Aldridge DC (2015) Microplastics in freshwater systems: A review of the emerging threats, identification of knowledge gaps and prioritisation of research needs. *Water Research* 75: 63–82. https://doi.org/10.1016/j.watres.2015.02.012.

Eriksen M, Mason S, Wilson S, Box C, Zellers A, Edwards W, Farley H, and Amato S (2013) Microplastic pollution in the surface waters of the Laurentian Great Lakes. *Marine Pollution Bulletin* 77: 177–182. https://doi.org/10.1016/j.marpolbul.2013.10.007.

Eriksen M, Lebreton LCM, Carson HS, Thiel M, Moore CJ, Borerro JC, Galgani F, Ryan PG, and Reisser J (2014) Plastic pollution in the World's oceans: More than 5 trillion plastic pieces weighing over 250,000 tons afloat at sea. *PLoS One* 9: e111913. https://doi.org/10.1371/journal.pone.0111913.

Faure F, Saini C, Potter G, Galgani F, de Alencastro LF, and Hagmann P (2015) An evaluation of surface micro- and mesoplastic pollution in pelagic ecosystems of the western Mediterranean Sea. *Environmental Science and Pollution Research* 22: 12190–12197. https://doi.org/10.1007/s11356-015-4453-3.

Ferreira P, Fonte E, Soares ME, Carvalho F, and Guilhermino L (2016) Effects of multi-stressors on juveniles of the marine fish *Pomatoschistus microps*: Gold nanoparticles, microplastics and temperature. *Aquatic Toxicology* 170. https://doi.org/10.1016/j.aquatox.2015.11.011.

Fischer V, Elsner NO, Brenke N, Schwabe E, and Brandt A (2015) Plastic pollution of the kuril-Kamchatka trench area (NW pacific). *Deep Sea Research Part II: Topical Studies in Oceanography* 111: 399–405. https://doi.org/10.1016/j.dsr2.2014.08.012.

Free CM, Jensen OP, Mason SA, Eriksen M, Williamson NJ, and Boldgiv B (2014) High-levels of microplastic pollution in a large, remote, mountain lake. *Marine Pollution Bulletin* 85: 156–163. https://doi.org/10.1016/j.marpolbul.2014.06.001.

Galgani F, Fleet D, Van Franeker J, Katsanevakis S, Maes T, Mouat J, Oosterbaan L, Poitou I, Hanke G, Thompson R, Amato E, Birkun A, and Janssen C (2010) *International Council for the Exploration of the sea for exploitation of the sea Conseil international pour l'Exploration de la Mer*. Copenhagen: ICES.

Gall SC and Thompson RC (2015) The impact of debris on marine life. *Marine Pollution Bulletin* 92: 170–179. https://doi.org/10.1016/j.marpolbul.2014.12.041.

GESAMP, 2015. Sources, fate and effects of microplastics in the marine environment: A global assessment, In Kershaw, P.J. (ed.), (IMO/FAO/UNESCO-IOC/UNIDO/WMO/IAEA/UN/UNEP/UNDP Joint Group of Experts on the Scientific Aspects of Marine Environmental Protection). GESAMP No. 90, p. 96.

Good TP, June JA, Etnier MA, and Broadhurst G (2010) Derelict fishing nets in Puget sound and the Northwest Straits: Patterns and threats to marine fauna. *Marine Pollution Bulletin* 60: 39–50. https://doi.org/10.1016/j.marpolbul.2009.09.005.

Gregory MR (1996) Plastic scrubbers' in hand cleansers: A further (and minor) source for marine pollution identified. *Marine Pollution Bulletin* 32: 867–871. https://doi.org/10.1016/S0025-326X(96)00047-1.

Gregory MR (2009) Environmental implications of plastic debris in marine settings: entanglement, ingestion, smothering, hangers-on, hitch-hiking and alien invasions. *Philosophical Transactions of the Royal Society B* 364: 2013–2025. https://doi.org/10.1098/rstb.2008.0265.

Hidalgo-Ruz V, Gutow L, Thompson RC, and Thiel M (2012) Microplastics in the marine environment: A review of the methods used for identification and quantification. *Science and Technology* 46: 3060–3075. https://doi.org/10.1021/es2031505.

Hirst, D., Baker, J., 2017. Proposed ban on microbeads; Number CDP 2017/0073, London.

Imhof HK, Ivleva NP, Schmid J, Niessner R, and Laforsch C (2013) Contamination of beach sediments of a subalpine lake with microplastic particles. *Current Biology* 23: R867–R868. https://doi.org/10.1016/j.cub.2013.09.001.

Ivleva NP, Wiesheu AC, and Niessner R (2017) Microplastic in aquatic ecosystems. *Angewandte Chemie International Edition* 56: 1720–1739. https://doi.org/10.1002/anie.201606957.

Jambeck JR, Geyer R, Wilcox C, Siegler TR, Perryman M, Andrady A, Narayan R, and Law KL (2015) Microplastic in aquatic ecosystems. *Science* 347: 768–771. https://doi.org/10.1126/science.1260352.

Kirchmann H, Börjesson G, Kätterer T, and Cohen Y (2017) From agricultural use of sewage sludge to nutrient extraction: A soil science outlook. *Ambio* 46: 143–154. https://doi.org/10.1007/s13280-016-0816-3.

Klein S, Worch E, and Knepper TP (2015) Occurrence and spatial distribution of microplastics in river shore sediments of the Rhine-Main Area in Germany. *Environmental Science & Technology* 49: 6070–6076. https://doi.org/10.1021/acs.est.5b00492.

Koelmans A, Gouin T, Thompson R, Wallace N, and Arthur C (2014) Plastics in the marine environment. *Environmental Toxicology and Chemistry* 33: 5–10. https://doi.org/10.1002/etc.2426.

Kole P, Löhr AJ, Van Belleghem FGAJ, and Ragas AMJ (2017) Wear and tear of tyres: A stealthy source of microplastics in the environment. *International Journal of Environmental Research and Public Health*. https://doi.org/10.3390/ijerph14101265.

Kühn, S., Bravo Rebolledo, E.L., Van Franeker, J.A., 2015. Deleterious effects of litter on marine life, in: Marine Anthropogenic Litter. p. 75–116. doi:https://doi.org/10.1007/978-3-319-16510-3_4.

Laist DW (1997) Impacts of marine debris: Entanglement of marine life in marine debris including a comprehensive list of species with entanglement and ingestion records marine debris. In: Coe JM and Rogers DB (eds.) *Marine debris—Sources, impacts, and solutions*, pp. 99–135. New York: Springer.

Lambert S, Sinclair C, and Boxall A (2014) Occurrence, degradation, and effect of polymer-based materials in the environment. *Reviews of Environmental Contamination and Toxicology* 227: 1–53. https://doi.org/10.1007/978-3-319-01327-5_1.

Lattin GL, Moore CJ, Zellers AF, Moore SL, and Weisberg SB (2004) A comparison of neustonic plastic and zooplankton at different depths near the southern California shore. *Marine Pollution Bulletin* 49: 291–294. https://doi.org/10.1016/j.marpolbul.2004.01.020.

Law KL (2017) Plastics in the marine environment. *Annual Review of Marine Science* 9: 205–229. https://doi.org/10.1146/annurev-marine-010816-060409.

Law KL and Thompson RC (2014) Microplastics in the seas. *Science* 345: 144–145. https://doi.org/10.1126/science.1254065.

Law KL, Moret-Ferguson S, Maximenko NA, Proskurowski G, Peacock EE, Hafner J, Reddy CM, Morét-Ferguson S, Maximenko NA, Proskurowski G, Peacock EE, Hafner J, and Reddy CM (2010) Plastic accumulation in the North Atlantic Subtropical Gyre. *Science* 329: 1185–1188. https://doi.org/10.1126/science.1192321.

Leslie HA, Van Velzen MJM, and Vethaak AD (2013) Microplastic survey of the Dutch environment novel data set of microplastics in North Sea sediments, treated wastewater effluents and marine biota. In: *IVM Institute for Environmental Studies*. Amsterdam: VU University.

Lusher AL, Hernandez-Milian G, O'Brien J, Berrow S, O'Connor I, and Officer R (2015a) Microplastic and macroplastic ingestion by a deep diving, oceanic cetacean: The True's beaked whale *Mesoplodon mirus*. *Environmental Pollution* 199: 185–191. https://doi.org/10.1016/j.envpol.2015.01.023.

Lusher AL, Tirelli V, O'Connor I, Officer R, Halsband C, and Galloway TS (2015b) Microplastics in Arctic polar waters: The first reported values of particles in surface and sub-surface samples. *Scientific Reports* 5: 14947. https://doi.org/10.1038/srep14947.

Mani T, Hauk A, Walter U, and Burkhardt-Holm P (2016) Microplastics profile along the Rhine River. *Scientific Reports* 5: 17988. https://doi.org/10.1038/srep17988.

Mathalon A and Hill P (2014) Microplastic fibers in the intertidal ecosystem surrounding Halifax Harbor, Nova Scotia. *Marine Pollution Bulletin* 81: 69–79. https://doi.org/10.1016/j.marpolbul.2014.02.018.

Mato Y, Isobe T, Takada H, Kanehiro H, Ohtake C, and Kaminuma T (2001) Plastic resin pellets as a transport medium for toxic chemicals in the marine environment. *Environmental Science & Technology* 35: 318–324. https://doi.org/10.1021/es0010498.

Mattsson K, Hansson L-A, and Cedervall T (2015) Nano-plastics in the aquatic environment. *Environmental Science: Processes & Impacts* 17: 1712–1721. https://doi.org/10.1039/C5EM00227C.

McCormick AR, Hoellein TJ, London MG, Hittie J, Scott JW, and Kelly JJ (2016) Microplastic in surface waters of urban rivers: Concentration, sources, and associated bacterial assemblages. *Ecosphere* 7. https://doi.org/10.1002/ecs2.1556.

Mehlhart G and Blepp M (2012) Review of sources and literature in the context of the initiative of the declaration of the global. In: *Study on land-sourced litter (LSL) in the marine environment*. Darmstadt/Freibg: Öko-Institut e.V. Google Sch.

Mintenig, S., Int-Veen, I., Löder, M. G., 2014. Gerdts Mikroplastik in ausgewählten Kläranlagen des Oldenburgisch-Ostfriesischen Wasserverbandes (OOWV) in Niedersachsen Abschlussbericht des Alfred-Wegener-Instituts (AWI) im Auftrag des Oldenburgisch-Ostfriesischen Wasserverbandes (OOWV) und des Nieder. Landesbetriebs für Wasserwirtschaft, Naturschutz (NLWKN), Küsten- und.

Moore CJ (2008) Synthetic polymers in the marine environment: A rapidly increasing, long-term threat. *Environmental Research* 108: 131–139. https://doi.org/10.1016/j.envres.2008.07.025.

Mouat, J., Lopez Lozano, R., Bateson, H., 2010. Economic impacts of marine litter: Kommunenes Internasjonale Miljoorganisasjon (KIMO International) [WWW Document]. URL eeloket.nl/images/Economic impacts of marine litter_1290.pdf (accessed 8.1.17).

Murphy F, Ewins C, Carbonnier F, and Quinn B (2016) Wastewater treatment works (WwTW) as a source of microplastics in the aquatic environment. *Environmental Science & Technology* 50: 5800–5808. https://doi.org/10.1021/acs.est.5b05416.

Napper IE and Thompson RC (2016) Release of synthetic microplastic plastic fibres from domestic washing machines: Effects of fabric type and washing conditions. *Marine Pollution Bulletin* 112: 39–45. https://doi.org/10.1016/j.marpolbul.2016.09.025.

Napper IE, Bakir A, Rowland SJ, and Thompson RC (2015) Characterization, quantity and sorptive properties of microplastics extracted from cosmetics. *Marine Pollution Bulletin* 99: 178–185. https://doi.org/10.1016/j.marpolbul.2015.07.029.

Nelms S, Coombes C, Foster L, Galloway T, Godley B, Lindeque P, and Witt M (2016) Marine anthropogenic litter on British beaches: A 10-year nationwide assessment using citizen science data. *Science of the Total Environment* 579: 1399–1409. https://doi.org/10.1016/j.scitotenv.2016.11.137.

Obbard RW, Sadri S, Wong YQ, Khitun AA, Baker I, and Thompson RC (2014) Global warming releases microplastic legacy frozen in Arctic Sea ice. *Earth's Future* 2: 315–320. https://doi.org/10.1002/2014EF000240.Abstract.

Oehlmann J, Schulte-Oehlmann U, Kloas W, Jagnytsch O, Lutz I, Kusk KO, Wollenberger L, Santos EM, Paull GC, Van Look KJW, and Tyler CR (2009) A critical analysis of the biological impacts of plasticizers on wildlife. *Philosophical Transactions of the Royal Society B* 364(1526): 2047–2062.

Pahl S and Wyles KJ (2017) The human dimension: How social and behavioral research methods can help address microplastics in the environment. *Analytical Methods* 9: 1404–1411. https://doi.org/10.1039/C6AY02647H.

Phillips MB and Bonner TH (2015) Occurrence and amount of microplastic ingested by fishes in watersheds of the Gulf of Mexico. *Marine Pollution Bulletin* 100: 264–269. https://doi.org/10.1016/j.marpolbul.2015.08.041.

PlasticsEurope, 2015. Plastics—The facts. [WWW Document]. PlasticsEurope. URL http://www.plasticseurope.org/Document/plastics--the-facts-2015.aspx (accessed 6.30.17).

PlasticsEurope, 2016. Plastics—The facts [WWW Document]. Plast. Shape Futur. URL http://www.plasticseurope.org/Document/plastics--the-facts-2016-15787.aspx?Page=DOCUMENT&FolID=2 (accessed 11.4.16).

Pruter AT (1987) Sources, quantities and distribution of persistent plastics in the marine environment. *Marine Pollution Bulletin* 18: 305–310. https://doi.org/10.1016/S0025-326X(87)80016-4.

Reisser J, Slat B, Noble K, Du Plessis K, Epp M, Proietti M, De Sonneville J, Becker T, and Pattiaratchi C (2015) The vertical distribution of buoyant plastics at sea: An observational study in the North Atlantic Gyre. *Biogeosciences* 12: 1249–1256. https://doi.org/10.5194/bg-12-1249-2015.

Ribic CA, Sheavly SB, Rugg DJ, and Erdmann ES (2010) Trends and drivers of marine debris on the Atlantic coast of the United States 1997–2007. *Marine Pollution Bulletin* 60: 1231–1242. https://doi.org/10.1016/j.marpolbul.2010.03.021.

Rillig MC (2012) Microplastic in terrestrial ecosystems and the soil? *Environmental Science and Technology* 46(12): 6453–6454. https://doi.org/10.1021/es302011r.

Rochman CM, Hoh E, Kurobe T, and Teh SJ (2013) Ingested plastic transfers hazardous chemicals to fish and induces hepatic stress. *Scientific Reports* 3. 3263https://doi.org/10.1038/srep03263.

Ryan PG, Moore CJ, van Franeker JA, and Moloney CL (2009) Monitoring the abundance of plastic debris in the marine environment. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 364: 1999–2012. https://doi.org/10.1098/rstb.2008.0207.

Scherer C, Brennholt N, Reifferscheid G, and Wagner M (2017) Feeding type and development drive the ingestion of microplastics by freshwater invertebrates. *Scientific Reports* 7: 17006. https://doi.org/10.1038/s41598-017-17191-7.

Song YK, Hong SH, Jang M, Kang J-H, Kwon OY, Han GM, and Shim WJ (2014) Large accumulation of micro-sized synthetic polymer particles in the sea surface microlayer. *Environmental Science & Technology* 48: 9014–9021. https://doi.org/10.1021/es501757s.

Sundt, P., Schultze, P.-E., Syversen, F., 2014. Sources of microplastic—Pollution to the marine environment. M-321 | 2015. Mepex Norwegian Environment Agency 1–108.

Sutherland WJ, Clout M, Côté IM, Daszak P, Depledge MH, Fellman L, Fleishman E, Garthwaite R, Gibbons DW, De Lurio J, Impey AJ, Lickorish F, Lindenmayer D, Madgwick J, Margerison C, Maynard T, Peck LS, Pretty J, Prior S, Redford KH, Scharlemann JPW, Spalding M, and Watkinson AR (2010) A horizon scan of global conservation issues for 2010. *Trends in Ecology & Evolution* 25: 1–7. https://doi.org/10.1016/j.tree.2009.10.003.

Talsness CE, Andrade AJM, Kuriyama SN, Taylor JA, and vom Saal FS (2009) Components of plastic: Experimental studies in animals and relevance for human health. *Philosophical Transactions of the Royal Society B* 364(1526): 2079–2096.

Tanaka K, Takada H, Yamashita R, Mizukawa K, Fukuwaka Ma, and Watanuki Y (2013) Accumulation of plastic-derived chemicals in tissues of seabirds ingesting marine plastics. *Marine Pollution Bulletin* 69: 219–222. https://doi.org/10.1016/j.marpolbul.2012.12.010.

Thompson RC, Olsen Y, Mitchell RP, Davis A, Rowland SJ, John AWG, McGonigle D, and Russell AE (2004) Lost at sea: Where is all the plastic? *Science* 304: 838. https://doi.org/10.1126/science.1094559.

Thompson R, Moore C, Andrady A, Gregory M, Takada H, and Weisberg S (2005) New directions in plastic debris. *Science* 310(5751): 1117.

Thompson RC, Moore CJ, vom Saal FS, and Swan SH (2009a) Plastics, the environment and human health: Current consensus and future trends. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 364: 2153–2166. https://doi.org/10.1098/rstb.2009.0053.

Thompson RC, Swan SH, Moore CJ, and vom Saal FS (2009b) Our plastic age. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 364: 1973–1976. https://doi.org/10.1098/rstb.2009.0054.

Turner A and Holmes LA (2015) Adsorption of trace metals by microplastic pellets in fresh water. *Environment and Chemistry* 12: 600–610. https://doi.org/10.1071/EN14143.

Unger B, Rebolledo ELB, Deaville R, Gröne A, IJsseldijk LL, Leopold MF, Siebert U, Spitz J, Wohlsein P, and Herr H (2016) Large amounts of marine debris found in sperm whales stranded along the North Sea coast in early 2016. *Marine Pollution Bulletin* 112: 134–141. https://doi.org/10.1016/j.marpolbul.2016.08.027.

Van Cauwenberghe L, Vanreusel A, Mees J, and Janssen CR (2013) Microplastic pollution in deep-sea sediments. *Environmental Pollution* 182: 495–499. https://doi.org/10.1016/j.envpol.2013.08.013.

Van Franeker JA, Blaize C, Danielsen J, Fairclough K, Gollan J, Guse N, Hansen PL, Heubeck M, Jensen JK, Le Guillou G, Olsen B, Olsen KO, Pedersen J, Stienen EWM, and Turner DM (2011) Monitoring plastic ingestion by the northern fulmar *Fulmarus glacialis* in the North Sea. *Environmental Pollution* 159: 2609–2615. https://doi.org/10.1016/j.envpol.2011.06.008.

Wagner M, Scherer C, Alvarez-Munoz D, Brennholt N, Bourrain X, Buchinger S, Fries E, Grosbois C, Klasmeier J, Marti T, Rodriguez-Mozaz S, Urbatzka R, Vethaak A, Winther-Nielsen M, and Reifferscheid G (2014) Microplastics in freshwater ecosystems: What we know and what we need to know. *Environmental Sciences Europe* 26: 12. https://doi.org/10.1186/s12302-014-0012-7.

Waller CL, Griffiths HJ, Waluda CM, Thorpe SE, Loaiza I, Moreno B, Pacherres CO, and Hughes KA (2017) Microplastics in the Antarctic marine system: An emerging area of research. *Science of the Total Environment* 598: 220–227. https://doi.org/10.1016/j.scitotenv.2017.03.283.

Wang J, Tan Z, Peng J, Qiu Q, and Li M (2016) The behaviors of microplastics in the marine environment. *Marine Environmental Research* 113: 7–17. https://doi.org/10.1016/j.marenvres.2015.10.014.

Watts AJR, Lewis C, Goodhead RM, Beckett SJ, Moger J, Tyler CR, and Galloway TS (2014) Uptake and retention of microplastics by the shore crab *Carcinus maenas*. *Environmental Science & Technology* 48: 8823–8830. https://doi.org/10.1021/es501090e.

Werner, S., Budziak, A., Van Franeker, J.A., Galgani, F., Hanke, G., Maes, T., Matiddi, M., Nilsson, P., Oosterbaan, L., Priestland, E., Thompson, R., Veiga Joana, M., Vlachogianni, T., 2016. Harm caused by marine litter—European Commission, JRC Technical Report. https://doi.org/10.2788/690366.

Woodall LC, Sanchez-Vidal A, Canals M, Paterson GLJ, Coppock R, Sleight V, Calafat A, Rogers AD, Narayanaswamy BE, and Thompson RC (2014) The deep sea is a major sink for microplastic debris. *Royal Society Open Science* 1: 140317. https://doi.org/10.1098/rsos.140317.

Wright SL, Thompson RC, and Galloway TS (2013) The physical impacts of microplastics on marine organisms: A review. *Environmental Pollution* 178: 483–492. https://doi.org/10.1016/j.envpol.2013.02.031.

Wyles KJ, Pahl S, Thomas K, and Thompson RC (2016) Factors that can undermine the psychological benefits of coastal environments. *Environment and Behavior* 48: 1095–1126. https://doi.org/10.1177/0013916515592177.

Zettler ER, Mincer TJ, and Amaral-Zettler LA (2013) Life in the "plastisphere": Microbial communities on plastic marine debris. *Environmental Science & Technology* 1: 7137–7146.

Zitko V and Hanlon M (1991) Another source of pollution by plastics—Skin cleaners with plastic scrubbers. *Marine Pollution Bulletin* 1: 41–42.

Microbial Communities

Cristiana Callieri, Ester M Eckert, and Andrea Di Cesare, National Research Council, Verbania, Italy
Filippo Bertoni, ASCA—University of Amsterdam, Amsterdam, The Netherlands

Introduction

Aquatic microbial communities encompass all the organisms living in (marine, fresh, and brackish) water habitats that are not visible to the naked eye (conventionally, less than 1 mm), thus including representatives from all three domains of life—*Archaea*, *Bacteria*, and *Eukarya*. Viruses, together with viroids, plasmids and other non-cellular life forms, are excluded from the three domains system, but, given their impact on microbial populations, are usually included in microbial community studies. Unlike a population—which includes only individuals of the same "species"—or a guild—which groups metabolically related populations—a community is not defined by phylogeny or functional traits but by its habitat. Thus, an aquatic microbial community can be defined as an *assemblage* of co-occurring, and potentially interacting, microscopic "species," present in a defined habitat in space and time.

Despite the small size of these organisms—and the related challenges in studying them—microorganisms are key to the ecological dynamics of the biosphere. As it reaches even the most extreme environments, microbial life's sheer biomass is consequential to the entire planet: in the oceans alone, microbes contribute to as much as 90% of the total biomass. And while oceans cover over 70% of the World's surface (1.3×10^9 km^3), meaning that most water on Earth (~97%) is seawater, inland waters—including lakes, ponds, rivers, streams, wetlands, and groundwater—and ice confined in polar caps and glaciers only account for ~1% and ~2% of the hydrosphere respectively, but have a similarly crucial role for life in the biosphere. All these aquatic environments are dominated by microbial communities. But microorganisms are not only the most diffused life form; they are also characterized by an incredible functional and genetic diversity, contributing to most, if not all, biogeochemical processes on Earth. As such, microbial communities represent a crucial component of global dynamics, and an important repository of genetic diversity.

In aquatic environments their abundance is significant, with concentrations—even excluding microbial eukaryotes—around approximately 10^6 cells per milliliter of marine or fresh water. They are responsible for the ongoing production and recycling of organic matter and are involved in fundamental energy flows, exhibiting the ability to perform a range of biogeochemical transformations.

The planetary impact of microbial communities and of their diversity (Fig. 1)—all the more evident when considering the crucial contributions they offered to the history of life on this planet, like the Great Oxygenation Event—depends on the accumulation of smaller causes into effects at a larger scale. But, given the vast nature of the environments the microbes inhabit, microbial life is most commonly approached at a large scale and through its cumulative effects. This is why, in this overview, we will focus especially on planktonic communities and on large-scale interactions and dynamics. In the conclusions we will get back to the importance of microscale dynamics and their interactions with macrobiota, to briefly consider the possibilities they open for novel research.

Microbial Lifestyles in Aquatic Habitats

In aquatic habitats there is a major distinction in lifestyles that structures research on microbial communities: this is the division between microbial organisms living suspended in the water column and therefore drifting with water currents, known as *plankton* (from the Greek term for vagabond), and those living attached to a substrate and in the sediment deposited at the bottom of water bodies, known as *benthos* (Greek for depth).

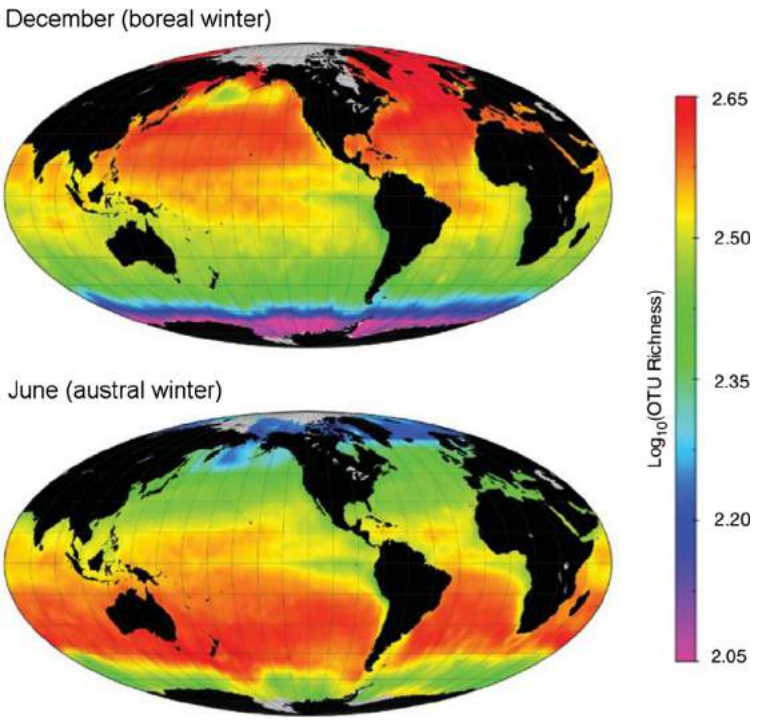


Fig. 1 Maps of predicted global marine bacterial diversity (modified from Ladau, J., et al. (2013). Global marine bacterial diversity peaks at high latitudes in winter. *ISME Journal* **7**, 1669–1677). Color scale shows relative richness of marine surface waters as predicted by Species Distribution Modeling (SDM). In December, the Operational Taxonomic Unit (OTU) richness peaks in temperate and higher latitudes in the Boreal Hemisphere. In June, OTU richness peaks in temperate latitudes in the Austral Hemisphere. Predicted richness during the spring and fall is intermediate.

Planktonic microorganisms are transported by water currents or simply by the turbulence created by water layers at different temperatures (i.e., density). Their growth is therefore strongly driven by physical forces that can transport them in layers enriched with nutrients or, conversely, can segregate them to deep waters (called hypolimnetic, in lakes and bathypelagic, in oceans). Despite this general dependence on water dynamics, some planktonic organisms possess locomotion appendages like flagella or cilia, or internal gas vacuoles by which the cells can regulate their buoyancy. Therefore, even at slow velocity and in calm water conditions, the motile cells can regulate their position. Usually considered as free-living cells, these organisms can also form aggregates of various kinds in different environmental conditions.

Benthonic microorganisms live attached to a fixed substrate or at the interface between water and sediment. Collecting organic material settling from above, the benthos takes part in the sedimentation processes and is usually more concentrated than the plankton. These organisms usually form aggregates embedded in an extracellular matrix composed of extracellular polymeric substances (EPS), called biofilms, which facilitate transport and exchange of chemicals, nutrients, and signaling compounds, while helping in protecting against predation. Early mats of similar biofilms are thought to have had an important role in the evolution of life (see Box 1). Within these aggregates, bacterial community composition varies based on the substrate, the broader habitat, and the species included. For example, the *Cytophaga-Flavobacteria-Bacteroidetes* (CFB) group mostly dominate the biofilm on macrophytes, followed by alpha- and betaproteobacteria. The presence of *Planctomycetes* on plant biofilms depends strongly on the nutrient content and plant age, as these bacteria can be affected by organic compounds produced by plants.

Box 1 Living at the interface: *the phycosphere*

Far from capturing the complexity of microbial lifestyles, this foundational division sometimes gets in the way of studying more detailed, micro-scale interactions and habitats. Indeed, as it is often true in ecology, even at the microscale the interface between two different environments tends to be more productive and rich in interactions. Microhabitats like marine snow particles, or the phycosphere of algae—the halo of organic compounds excreted by phytoplanktonic cells—are crucial to microbial communities, especially in environments otherwise characterized by low nutrient concentrations. In the phycosphere, bacteria or other microorganisms can be found as free-living single cells near algal cells or directly attached on their surface, in biofilms (Fig. 2). The study of the phycosphere at single-cell level has shown how this microenvironment is a hot-spot of bacteria–phytoplankton interactions, where the exchange of infochemicals and metabolites is very high. One of these substances is dimethylsulfoniopropionate (DMSP), a source of sulfur and carbon for heterotrophic bacteria, which is actively exudated by algae and produced during cell lysis. The lifestyle of bacteria in aggregates has the advantage to protect them against small flagellate predation (hetero- and mixotrophic nano flagellates), to increase the sedimentation rate and to offer a nutrient richer microenvironment. While challenging our categorizations of microbial life, microenvironments and their microdynamics remind us of the importance of paying attention to life at any scale—as even the smallest niche might reveal key to the history of life on Earth.

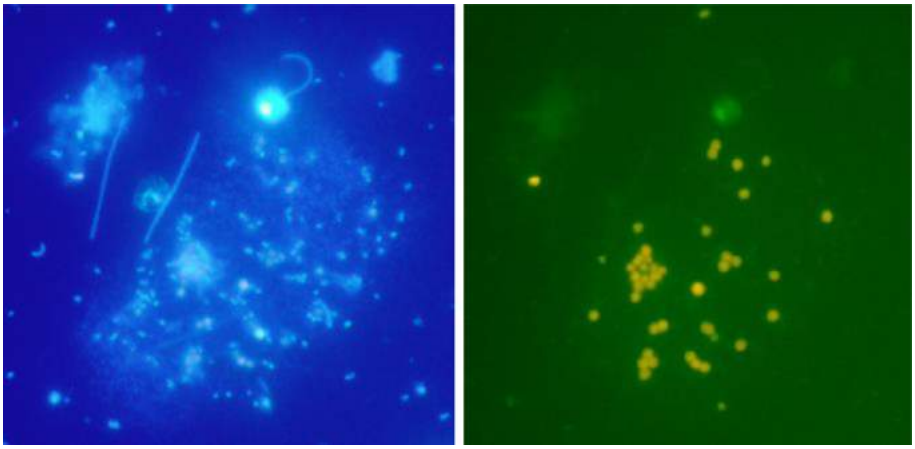


Fig. 2 An example of phycosphere: a microenvironment where bacteria, cyanobacteria and eukaryotes meet. *Left*: DAPI colored and visualized at epifluorescence; *Right*: the same field without DAPI where only autofluorescence is visible (1250 ×).

Microbial Diversity of Planktonic Communities

Prokaryotes

Prokaryotes are defined by their cell structure, which lacks membrane-bound organelles and nucleus. They include the whole domains *Bacteria* and *Archaea* and can perform autotrophic and heterotrophic metabolisms. An important subdivision in this group is that between Gram-positive bacteria, bounded by a single-unit lipid membrane and a thick layer of peptidoglycan, and Gram-negative bacteria, with an outer membrane, a lipopolysaccharide layer, but a thin peptidoglycan layer. This rough classification, based on the result of a staining procedure, remains still valid, despite the changes brought about by molecular biology in the classification of microorganisms. This is because the composition of the outer part of the cells, which is identified by the Gram stain, strongly determines the capacity of cells to resist to stressing environmental conditions or to communicate with other microorganisms, making it a key for cell taxonomy (see Box 2).

The marine bacterium *Pelagibacter ubique*, belonging to the SAR11-cluster of *Alphaproteobacteria*, is the most abundant microorganism on earth. *Actinobacteria*, *Alphaproteobacteria* and *Betaproteobacteria*, *Flavo-* and *Sphingobacteriales* numerically dominate the freshwater bacterioplankton. The acI clade of *Actinobacteria* and the Beta I and II clades of *Betaproteobacteria* comprise 30%–50% of the total bacterial cells in the water column. Similarly, the ultramicrobacteria of freshwater LD12 lineage, sister group of the marine SAR11, previously considered rare in lacustrine ecosystems, have been recently found to reach abundances comparable with those of the oceans. This discovery has opened a new view on the dispersal of microorganisms between marine and freshwater habitats.

Another important phylum of photosynthetic bacteria is composed by *Cyanobacteria*. These microorganisms come in a variety of forms (including both coccoid and filamentous forms) and lifestyles (from single free-living cells to colonial aggregates) and are ubiquitous in aquatic environments. This is because, thanks to their long evolutionary history, they possess a variety of metabolic

Box 2 Estimating Microbial Diversity

The definition of species for Prokaryotes is hampered by various traits, including their poorly developed morphology and the fact that the vast majority of these organisms remains uncultured, and thus little is known about their physiology and behaviour. The main source of information on bacterial and archaeal diversity derives from environmental surveys of DNA sequences. Species are often defined through similarity thresholds of marker sequences, such as the small-subunit rRNA (16S and 18S rRNA gene for prokaryotes and eukaryotes, respectively). Taxonomic grouping of the sequences is considered operational, meaning that it is based on the dataset of a specific study, and the term species is replaced by *Operational Taxonomic Unit* (OTU). However, even such molecular microbial diversity is far from being fully explored, and recent estimates of the global number of OTUs of Archaea and Bacteria differ as much as between millions and billions. This is because many taxa are very low in abundance, as in many sequencing campaigns a large part of the sequences is presented only once in the whole dataset (Fig. 3). Such sequences can be related to active species that are very low in number, to dormant or dead cells, to free DNA or simply due to sequencing errors. Thus, the actual richness of this so-called rare-biosphere is difficult to estimate. The upper pelagic zone of the ocean alone is estimated to harbor more than 35×10^3 prokaryotic species, whereas richness in the deep sea is around an order of magnitude poorer. In general, temperature seems to be a driving factor of richness in the oceans. Estimates of freshwater diversity are also in the range of tens of thousands and terrestrial ecosystems are an important source of diversity of freshwater ecosystems. Therein sediments are particularly rich in OTUs. Molecular richness of microbial eukaryotes found in aquatic environments is predicted to be more than 100×10^3 with similar numbers of OTUs in the sea and in freshwaters. These numbers far exceed the numbers of species determined morphologically (morpho species). This is, on the one hand, due to the fact that some species look the same but are genetically different (many fungi, e.g., but also diatoms), in which case morphological identification underestimates species diversity. On the other hand, some species have more than one copy of the 18S rRNA gene in their genome, which can be very polymorph (e.g., in some Ciliate species over 40 polymorphic sites of the 18S rRNA gene were found) and thus diversity is vastly overestimated by molecular technics.

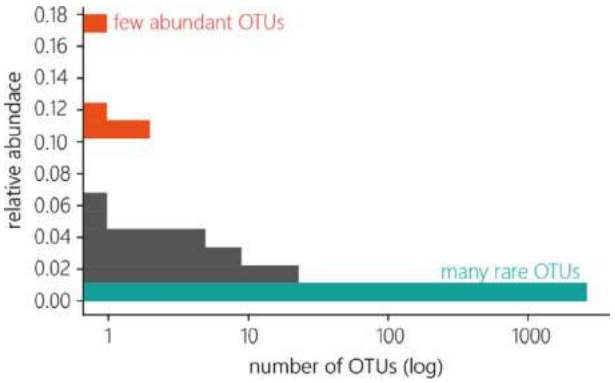


Fig. 3 Potential outcome of a meta-sequencing study of the 16S rDNA gene diversity clustered into Operational Taxonomic Units (OTUs). Most of the OTUs are very low in abundance in the dataset and therefore rare, only very few OTUs have high relative abundances and are considered abundant.

pathways that give them competitive advantage and facilitate their adaptation to diverse environmental conditions, from the most eutrophic to pristine oligotrophic waters. Because of this adaptability and plasticity, they are particularly successful in changing, adverse conditions: hence the current increase in occurrence of toxic cyanobacterial harmful algal blooms (CHABs), facilitated by anthropogenic climate disruptions on a planetary scale—which associated cyanobacteria to environmental distress.

One of the most enigmatic microbial groups in the oceans is the Marine Group I (MGI.1a) of Archaea: the *Thaumarchaeota* (formerly 1.1a Crenarchaeota). These Archaea are commonly found in oceanic and freshwater plankton, mainly in mesopelagic and hypolimnetic waters, where they comprise up to about 40% of the microbial community. Members of the MGI.1a clade have also been found in high-altitude lakes during stratification, whereas the clade SAGMGC-1 was retrieved in high densities in surface water in winter.

Microbial Eukaryotes

Microbial eukaryotes, whose cells are recognized by the membrane-bound nuclei and organelles, are massively abundant; in the euphotic zones of the world's oceans alone, there are 10^2–10^4 protist cells per mL. The definition of eukaryotic microbes is usually arbitrarily based on size, including everything smaller than 3 μm. Thus, representatives of the whole domain *Eukarya*, except for vertebrates, are included in this category, like many unicellular algae (e.g., Chlorophyta, Diatoms, Dinoflagellates). Functional classification is mainly based on their metabolism, meaning whether they sequester carbon through primary production, heterotrophy or mixotrophy—in which the same cell can fix CO_2, as well as assimilate it from dissolved or particulate organic carbon. Considering that many species are capable of this versatile metabolism, comparably little attention has been devoted to mixotrophy, mainly due to technical difficulties in identifying whether a pigmented cell is currently following a purely autotrophic lifestyle or not.

The most common eukaryotic freshwater primary producers, referred to as *phytoplankton*, include members of *Chlorophyta*, *Haptophyceae*, *Bacillariophyceae*, *Cryptophyta*, *Chrysophyceae* and *Dinoflagellata*. In the euphotic zone of the open ocean *Dinophyceae*, *Dictochophyceae*, *Bacillariophyceae*, *Pelagophyceae*, *Synurophyceae*, and *Chrysophyceae* are particularly abundant. Unpigmented microbes include but are not limited to small aquatic fungi, choanoflagellates, cercozoans, ciliates, and *Chrysophyceae*. In addition to these widespread groups, a large portion of eukaryotic plankton diversity seems to reside in heterotrophic protistan groups, particularly those known to be parasites or symbiotic hosts. To date many eukaryotes remain uncultured, such as the abundant flagellated bacterial predators classified in the marine *Stramenopiles* groups 4 and 1 (MAST-4 and MAST-1C), which include *Bacillariophyceae* and *Chrysophyceae*. On marine snow (sinking organic particles) fungi and *Labyrinthulomycetes* seem to dominate the biomass of the community. Despite this, and their likely importance in aquatic environments, aquatic fungi are still poorly understood and deserve more attention, since they are highly diverse in terms of function and phylogenetic diversity.

Virioplankton

Despite their still controversial classification, viruses are abundant and crucial to ocean life. It has been estimated that ocean waters harbor about 4×10^{30} viruses, most of them infecting heterotrophic and autotrophic bacteria. The ratio between viruses and bacteria is comprised between 3 and 10. Through the water column, viruses mainly occupy the euphotic layers and their abundance drops as depth increases; contextually, virioplankton is more abundant near the coast and decreases in the open ocean. Although generally viruses are small acellular organisms, composed almost exclusively by a genome coding for few proteins, the diversity of the mechanisms employed in their replication and biochemistry is higher than that observed for cellular organisms. Moreover, viruses harbor high genetic and biological diversity, and seem to have played—and still play—a key role in the evolution of life.

Microbial Functional Pathways in Aquatic Systems

Aquatic microbial communities show a large variation of metabolic activities, ranging from primary production to the heterotrophic utilization of organic compounds. The conversion of inorganic carbon (or other electron acceptors) into organic living material—which characterizes both photoautotrophic (relying on the energy of light) and chemoautotrophic (dependent on chemical energy in minerals) microorganisms—is performed by the producers that are at the base of the trophic chain. Autotrophic picoplankton (microorganisms smaller than 2 μm, including picocyanobacteria and picoeukaryotes) contribute substantially to primary production, both in marine and freshwaters. In the oceans, *Prochlorococcus* and *Synechococcus* contribute around 75% to CO_2 fixation, but in lakes the contribution of *Synechococcus* ranges from 5% to 65% of total phytoplankton production (Lake Constance, Germany), with higher contribution measured in ultra-oligotrophic lakes (80% in Lake Baikal). These estimates have been obtained by the use of radioactive compounds that are good tracers even when measuring low activities.

With the diffusion of molecular biology techniques functional pathways became easier to study, as it is now possible to quantify the presence of specific functional genes, and to better characterize the transcription products, thus demonstrating the functionality of genes. Alternative heterotrophic metabolisms performed by *Bacteria* and *Archaea* in addition to the well-known glycolysis (crucial to sugar metabolism), have been studied in details. Important among them is the pentose phosphate pathway to generate NADPH and pentose, studied through the presence of genes codifying for different enzymes in the pathway.

Another important metabolic pathway in aquatic systems is nitrification, the two-step metabolism, which oxidizes the ammonium to nitrate and then to nitrite. The microorganisms involved in the first step are nitrifying bacteria that include ammonia-oxidizing bacteria (AOB) like *Nitrosomonas* and *Nitrosoccoccus* and ammonia-oxidizing archaea (AOA) like Thaumarchaeota. The *amo*A gene, encoding for the alfa-subunit of the ammonia monooxygenase enzyme, which catalyzes an important step of bacteria nitrification, is associated with an archaeal metagenomic fragment. This discovery underlines the ecological importance of Archaea in aquatic ecosystems, by emphasizing their role in the first step of nitrification and their possible competitions and relations with nitrifying Bacteria.

Another pathway that recently received more attention is the dark inorganic carbon assimilation that can be performed by chemoautotrophs affiliated with Bacteria and Archaea. Chemolithoautotrophic bacteria fix CO_2 in the dark through a variety of carboxylation reactions to satisfy metabolic requirements such as the synthesis of fatty acids, nucleotides and amino acids or anaplerotic demands. Thaumarchaeota too can perform dark assimilation, but they use the hydroxypropionate–hydroxybutyrate carbon assimilation pathway.

The dichotomy between autotrophic and heterotrophic organisms—also thanks to molecular biology—became increasingly insufficient to describe the scaffold of microbial functional pathways. Like for macrobial eukaryotes, also the microbial world is rich in cells combining different metabolic strategies. Despite this richness, the combination of autotrophic and heterotrophic metabolisms to sustain growth and maintainance can be an energetic cost for the cell. For example, the light-harvesting apparatus needs a high investment in energy terms, but can be a resource in case of organic matter limitation in some oligotrophic systems. Even the opposite strategy can prove successful, as in the case of *Synechococcus*, a picocyanobacteria abundant in marine and freshwaters, that consumes organic sulfur compounds, playing an important role in the cycle of dimethylsulfopropionate and methanethiol. The ability of *Synechococcus* and *Prochlorococcus* to take up amino acids and urea has been demonstrated in axenic cultures, showing how these cyanobacteria can occupy different trophic levels, also contributing to the secondary production in aquatic environments.

Some bacteria can use chromophores like proteorhodopsin, which works as a light-driven proton pump, therefore allowing these organisms to act as a photo-organoheterotrophic bacteria. Important groups like SAR11 (*Pelagibacter ubique*) and some *Flavobacteria* are among the bacterial group containing proteorhodopsin. Another illustration of the metabolic complexity of microbial communities is the discovery of a group of anoxygenic photosynthetic bacteria that use bacteriochlorophyll *a* (Bchl-*a*) for photosynthesis in the absence of oxygen, thus being able to also survive in aerobic conditions. The unexpected abundance of new metabolic pathways in microbial communities we briefly described here is allowing microbiologists to revise our understanding of nutrient and energy flows in aquatic ecosystems.

Extrinsic and Intrinsic Drivers Structuring the Microbial Community Assembly

When considering microbial community assembly in aquatic habitats, the Baas-Becking hypothesis ("everything is everywhere, but the environment selects") is a useful heuristic selection on the part of the environment—and its variables and (micro) dynamics—that can help understanding how communities form and are organized. In addition, microbial dispersal is a passive process but not an entirely stochastic one, so that microbial biogeographical patterns are affected by the organisms' fitness to the environment (or specific niche) and by microbial life history traits. The forces that influence the structure and function of aquatic microbial communities can be extrinsic—meaning they act at a broader, regional scale—and intrinsic—such as food-web interactions that are more site-specific. The former ones impart a sort of synchronization to different populations in the community by the action of a local array of factors like temperature or meteorology. Intrinsic factors further affect the pattern of synchronized microbial population, resulting in fluctuations in the community that are harder to predict.

Extrinsic Drivers

The extrinsic drivers regulating microbial communities can be broadly grouped in physical ones, like temperature, and currents, and chemical ones, like salinity, dissolved oxygen, nutrients, metals, and antibiotics. Considering the current climatic changes brought about by CO_2 and other greenhouse gases emissions, and the subsequent prediction of a global rise in ocean temperatures in the order of 2–4°C, temperature as a physical extrinsic driver has been receiving much attention—especially considering its profound impact on biological processes. Recent studies directly linked temperatures to changes in richness and diversity of microbial communities. Moreover, temperature was also documented as one of the extrinsic driver associated with the increase of the relative abundance of *Vibrio* spp. (a genus comprising different pathogens for humans and animals) resulting in a possible threat for human and animal health.

In addition to temperature, in aquatic environments currents are important drivers because of their effect on the geographic dispersal of bacterial communities, which can strongly affect their composition. In particular, planktonic microbial communities drifting in the ocean experience a temperature variability up to 10°C greater than the estimated seasonal fluctuations, resulting in a selection of bacteria more tolerant to this thermal exposure.

Among the chemical drivers affecting microbial communities, salinity is certainly one of the main factors responsible for community assemblage composition, sometimes even more important than temperature, informing the many differences between marine and freshwaters. In addition to salinity, dissolved oxygen (DO) also occupies a central role in community dynamics because of its influence on biotic interactions and on nutrient flows within aquatic ecosystem. In particular, it has been found that higher DO concentrations, positively correlates with *Alphaproteobacteria* biomass while negatively affecting bacterial diversity, that dramatically decreases in anoxic waters.

Nutrient availability determines a "bottom-up" pressure in the regulation of aquatic microbial communities. It is generally considered that the availability of phosphorus is limiting the growth of planktonic communities in most freshwaters and nitrogen and iron in many marine systems. Additionally, a shift in nutrient availability will strongly impact community composition, especially when coupled with temperature. However, given the complexity and interlinked nature of these dynamics, it is not always easy to identify a direct effect of nutrients on the whole community examined; this might result in smaller effects, for example a change limited to the composition of variable taxa only. The challenges of understanding the dynamics of the aquatic nutrient pool are worsened by anthropogenic impacts on nutrients concentrations, and even by particular climatic events like strong winds (possibly associated with dust dispersion in water) and volcanic eruptions (characterized by the release of ash and pumice in water altering the carbon/phosphorous ratio).

In addition to these extrinsic drivers, two other factors have a strong impact on microbial communities and are often linked to anthropogenic pollution: antibiotics and metals. Originally, antibiotics are natural compounds released by microorganisms to kill other microbial competitors, and that have an important role as signal molecules, affecting the structure of microbial communities. The artificial isolation, synthesis and overuse of antibiotics in human and animal medicine and aquaculture, constitutes a massive and growing source of pollution, as these compounds are released through excreta into aquatic environments and accumulate, with unpredictable outcomes on microbial life. This is because antibiotics, regardless of their concentration, can exert a selective pressure, altering the structure of microbial communities in terms of both phenotypic distribution and composition, and promoting the spread of antibiotic resistance genes, thus constituting a major threat for human health.

While some metals (i.e., copper, zinc, nickel, lead, cadmium, mercury, silver, gold, and chromium) are necessary in small concentrations to microbial metabolisms for several cellular functions, in high concentrations they can be toxic for microbial life. Other metals, like uranium and antimony, are always toxic for microorganisms. Given their importance in contemporary human industries (mercury in electronics, dental and health care industry; copper and zinc as growth promoters in husbandry), metals are released and eventually accumulated in aquatic environment at an alarming rate. This anthropogenic contamination shapes microbial communities, engendering a strong reduction of microbial diversity, affecting metabolic pathways, and negatively influencing the expression of enzymes.

Intrinsic Drivers

The intrinsic (site-specific and biotic) drivers that structure microbial communities are those related to the interactions between species within the same environment. Interactions between microorganisms can be categorized in five broad groups: (1) *predation*, when one organism feeds on another one; (2) *commensalism*, when one organism benefits from another one without affecting it; (3) *mutualism*, when both organisms benefit each other; (4) *parasitism*, when the parasite benefits and the host is harmed; and (5) *competition*, when the organisms do not directly interact with each other but compete for the same resources. Some of these interactions can act directly on species composition, as for example predation, or indirectly by the production of allelopathic substances with beneficial or detrimental effects.

The classical, and highly simplified, food chain would comprise primary producers that are eaten by primary consumers, which in turn are eaten by secondary consumers and so on up to the top predators. However, the photosynthesis performed by phytoplankton is not 100% efficient in terms of carbon fixation into biomass, and an estimated 13% of dissolved organic carbon is directly released into the water. These high quality substrates, such as sugars, are readily taken up by aquatic bacteria, which then themselves are consumed by primary consumers. These organisms are in turn eaten by zooplankton species and thereby the organic carbon is channeled back into the classical food chain, reaching higher organisms. Since this pathway is like a loop attached to the

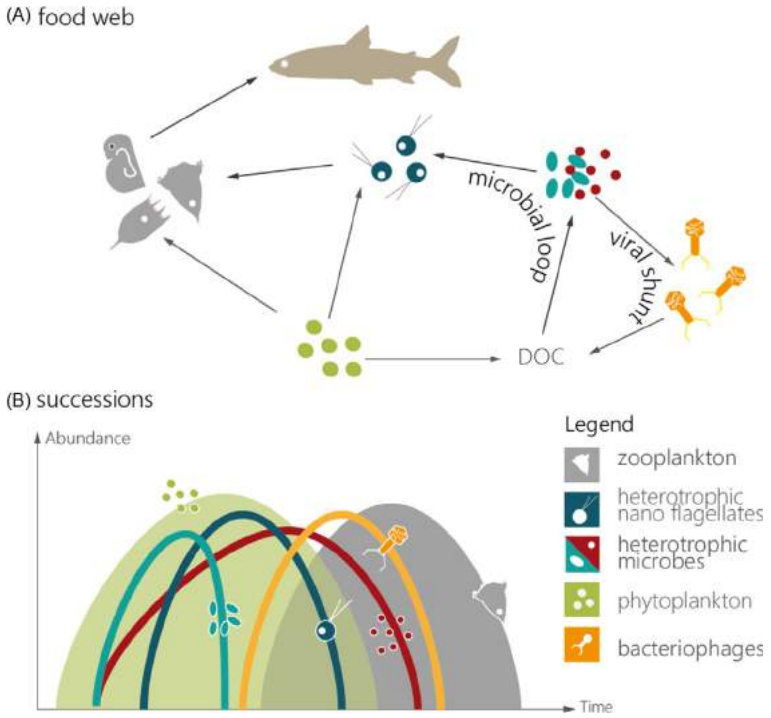


Fig. 4 (A) Simplified scheme of an aquatic food web including the microbial loop and the viral shunt; (B) examples of potential successions of microorganisms observed in aquatic habitats.

simplified food chain, it has been termed "the microbial loop" (Fig. 4A). This concept was established for marine bacterioplankton in the early 1980s and easily applies to freshwater ecosystems too. Given the privileged role that macroscopic life has been granted for most of the history of science, the microbial loop is generally considered an attachment to the classical food-chain. Nevertheless, it has been argued that, given the impressive biomass and energy turnover in the microbial loop, its diffusion, and its importance to the biosphere and its evolution, the classical food chain should be seen as an appendage to microbial trophic interactions. The two main pathways of the microbial loop (DOC—Dissolved Organic Carbon uptake, and predation) highlight the two most important factors defining the niche space of aquatic microbes: availability of organic and inorganic nutrients (bottom-up control) and agents of mortality such as viruses and HNF (top-down).

Consequently aquatic bacteria are often classified in two main guilds: fast-growing and grazing-resistant bacteria. Some of the fast- or opportunistically-growing bacteria are free-living, while others attach to particles, where organic carbon concentrations and turnover are typically higher. These bacteria are known for their efficient substrate uptake machinery, as well as their high vulnerability to protistan grazing. Typical members of this guild are *Alteromonas* and *Flavobacteria*, and *Limnohabitans*-affiliated bacteria, in marine and freshwaters respectively. A different survival-strategy is the investment in protection against protistan grazing, which is characteristic of the so-called grazing-resistant bacteria. Such protection can be due to morphological and structural features, as very small cells (*ultramicrobacteria*) with rigid cell walls, and long filamentous cells, are less commonly ingested by predators. Ultramicrobacteria are highly abundant in aquatic systems and include the ac1 group of *Actinobacteria* in freshwaters and *Pelagibacter ubique* (SAR11) in the oceans.

Dissolved organic carbon found in natural waters is composed by a variety of substrates including sugars, amino acids, humic acids and many others. On the one hand, different phytoplankton species release different carbon compounds, that is, *Chryptophytes* release more sugars than many *Cyanobacteria*. On the other hand, there are also numerous other sources of organic carbon, such as bacterial growth and mortality, sloppy-feeding by zooplankton or substrates deriving from terrestrial input. As a consequence, co-occuring heterotrophs have been seen to have very different substrate uptake preferences: whereas some genotypes seem to be specialized in the uptake of amino acids, other prefer glucose, for example. Such differences in meal-preference help explain why so many different bacterial genotypes can co-exist.

As we already mentioned, the two main causes of mortality in microbial communities are flagellate predation and viral lysis. The effect of flagellated predators on microbial communities is often seen as a shift in the composition and morphology. Some bacteria are able to change their growth strategy in the presence of predators by, for example, aggregating with other cells or elongating into filamentous shapes, which allow them to avoiding predation due to increased size. But such grazing defense strategies are often evolutionarily costly. This becomes clear when considering an experiment involving a bacterium (*Sphingomonas* sp.) kept under grazing pressure for a long time: after the experiment, the bacterium aggregates also in absence of the predator, but lost part of its genome related to carbon degradation. In fact, many highly abundant marine and freshwater ultramicrobacteria are characterized by so called *streamlined* genomes, genomes that contain only the minimum genes needed for survival. Simultaneously, though,

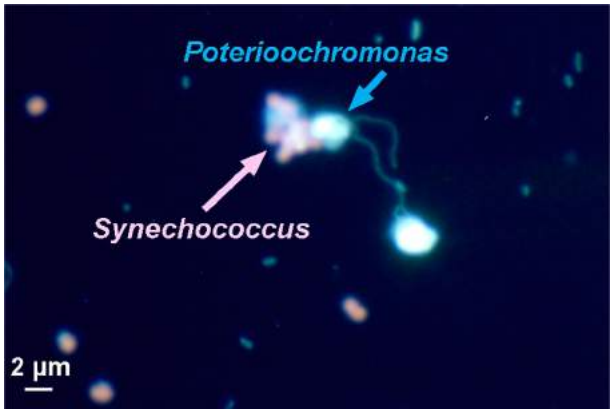


Fig. 5 *Synechococcus* cells (prey) and *Poterioochromonas* sp. (mixotrophic nano-flagellate predator) interaction in a culture, giving rise to the presence of microcolonies (photo by C. Callieri).

flagellate grazing is also known to stimulate bacterial growth, probably also due to the egestion of incompletely digested prey, which can serve as a substrate to other bacteria and regenerate nutrients. In the case of picocyanobacteria, the effect of nanoflagellate predation was to affect the shift from single cell morphotypes to microcolonial aggregates (Fig. 5), thus contrasting the effect of this intrinsic driver.

Viruses are potentially able to infect all of the components of aquatic microbial communities, however they primarily infect the most abundant bacteria and eukaryotic microbes (killing-the-winner hypothesis). Due to their ability to infect microbial hosts, in order to survive and replicate they directly affect, at the community level, the structure and composition of the microbial life and, at level of the single infected hosts, their physiological status. Therefore, viruses also play an important role within the food web. In fact, they lyse the infected bacteria, determining the release of their cellular content, which is converted in particulate and dissolved organic carbon usable by non-infected prokaryotes (viral shunt) (Fig. 4A). The actual consequence of this flow of matter through the food web is the increase of the respiration rate of the community and the decrease of the efficiency of carbon transfer toward the higher trophic layers. Together with carbon, also other nutrients are released, which can sometimes almost satisfy the particular nutrient need of certain organisms.

Based on these food web interactions, typical successions of aquatic organisms can be observed (Fig. 4B), initiated by the steep increase of phytoplankton biomass (e.g., early spring). These primary producers exude organic carbon (OC), which serves as a substrate for heterotrophic bacteria. In turn, the increasing numbers of bacteria promote the growth of heterotrophic nanoflagellates (HNF), which feed on the bacterial assemblage and provide a niche for alternative bacterial groups, not affected by grazing. Mortality among these bacteria is probably mainly caused by viral lysis. Generally, phytoplankton blooms are terminated by high abundances of zooplankton species.

While these classical food web interactions provide a useful handle to imagine such dynamics, many more interactions characterize microbial communities alongside them; they are termed non-trophic interactions. One example of these interactions can be found in microbes that grow as symbionts, epibionts or parasites. Some parasites can be highly abundant, such as members of the marine Alveolata MALV-IV group that attach to many different higher organisms. Some of these associations are only transient: bacteria may benefit from zooplankton as a refuge from threats such as grazing and abiotic stressors, or use zooplankton as a mean of transport from lower to higher water layers. The association of the cholera-causing agent *Vibrio cholerae* with planktonic copepods enables the survival, proliferation and transmission of the disease. Attached bacteria and eukaryotes can be metabolically highly active. Since they are preyed upon together with their host, they might form a shortcut through the trophic chain, directly transferring organic carbon from the dissolved fraction to the top predators. Another example is co-aggregation or attachment, through which bacteria can attach, for example, to phytoplankton cells and reduce substrate competition with other bacteria due to their vicinity to the primary producer.

Conclusions and Perspectives

In this article, we focused on the broader processes that characterize microbial communities in aquatic environments. This scope provides a better grasp of the main questions and concerns informing the study of these organisms, and allows to foreground their impact on a global scale. This is also because much research on these communities concentrated on such more readily apparent dynamics. These planetary biogeochemical transformations are the results of processes working at the microscale and resulting in a fragmented and diverse micro-heterogeneity of aquatic habitats (Fig. 6). The prevalence of the microscale gradients in aquatic systems forces us to seriously reconsider the importance of studying aquatic microbes at single-cell level, and at a scale that is more commensurate with their own dynamics and lives. The new opportunities offered by genomics, single-cell technologies like nanoSIMS (nano secondary ion mass spectrometry), microfluidics, and atomic-force microscopy opened new frontiers for

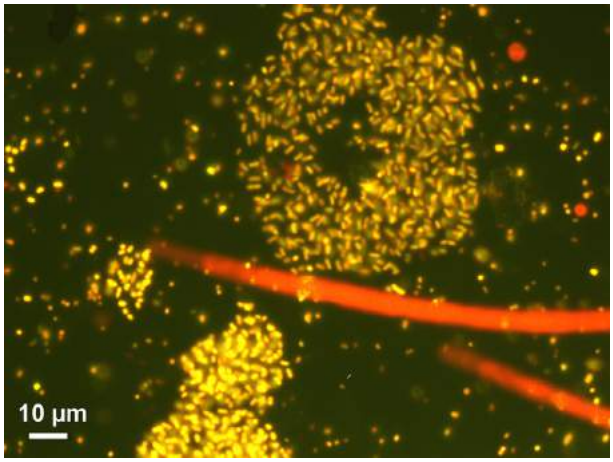


Fig. 6 Micro-heterogeneity in a drop of lake water (photo by C.·Callieri).

understanding the functioning of microorganisms and their interactions with the environment, paving the way to a micro-scale focus on microbial communities.

Despite these novel approaches, and the vast range of molecular and computational techniques that revolutionized our understanding of microbial life since the 1980s, our understanding of these life forms and the vital processes they shape still faces many challenges, and constantly stumbles upon surprising findings that redefine how we understand life on Earth and beyond. For this reason, the study of the microbiota, its micro-scale dynamics, and its interactions with the macrobiota require us to think outside the box, to work collaboratively across disciplinary boundaries, and to remain open to the awe-inspiring workings of our planet's ecosystems. This is the hard and yet rewarding task inherited by the younger generations of scientists, thinkers and explorers to whom this article is dedicated.

Further Reading

Amann R and Rosselló-Móra R (2016) After all, only millions? *MBio* 7: e00999-16.
Azam F, Fenchel T, Field JG, Gray JS, Meyer-Reil LA, and Thingstad F (1983) The ecological role of water-column microbes in the sea. *Marine Ecology Progress Series* 10: 257–263.
Doblin MA and van Sebille E (2016) Drift in ocean currents impacts intergenerational microbial exposure to temperature. *Proceedings of the National Academy of Sciences of the United States of America* 113: 5700–5705.
Giovannoni SJ and Stingl U (2005) Molecular diversity and ecology of microbial plankton. *Nature Reviews Microbiology* 437: 343–348.
Jürgens K (1994) Impact of Daphnia on planktonic microbial food web. A review. *Marine Microbial Food Webs* 8: 295–324.
Kent AD, Yannarel AC, Rusak JA, Triplett EW, and McMahon KD (2007) Synchrony in aquatic microbial community dynamics. *The ISME Journal* 1: 38–47.
Locey KJ and Lennon JT (2016) Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences* 113: 5970–5975.
Newton RJ, Jones SE, Eiler A, McMahon KD, and Bertilsson S (2011) A guide to the natural history of freshwater lake bacteria. *Microbiology and Molecular Biology Reviews* 75: 14–49.
Pedros-Alio C (2012) The rare bacterial biosphere. *Annual Review of Marine Science* 4: 449–466.
Pernthaler J (2005) Predation on prokaryotes in the water column and its ecological implications. *Nature Reviews Microbiology* 3: 537–546.
Ruiz-González C, Niño-García JP, and del Giorgio PA (2015) Terrestrial origin of bacterial communities in complex boreal freshwater networks. *Ecology Letters* 18: 1198–1206.
Seymour JR, Amin SA, Raina JB, and Stocker R (2017) Zooming in on the phycosphere: The ecological interface for phytoplankton-bacteria relationships. *Nature Microbiology* 2: 17065.
Stocker R (2012) Marine microbes see a sea of gradients. *Science* 338: 628.
Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, and Alberti A (2015) Structure and function of the global ocean microbiome. *Science* 348: 1261359.
Suttle CA (2005) Viruses in the sea. *Nature Reviews Microbiology* 437: 356–361.

Age Structure and Population Dynamics☆

Louis C Bender, New Mexico State University, Las Cruces, NM, United States

Age structure in populations can take a variety of forms and have a range of effects on population dynamics. Species can be comprised of multiple age classes in a single-stage life cycle (hereafter, age-class structuring), such as in most birds, fish, and mammals. Species can also show multi-stage life cycles, where different stages in the life cycle are analogous to age classes (hereafter, stage structuring). Such life cycles are common in invertebrates, amphibians, and plants. Species can also show both types of age structuring; typically this involves a stage-structured life cycle where the longest lived stage (usually the adult or reproductive stage) also shows ≥ 2 distinct age classes. For example, trees have both multiple life stages (seed, plant) and age classes within the plant stage (i.e., seedling, sapling, mature tree).

Both age-class structuring and stage structuring in populations can affect population dynamics, albeit typically at different scales. Stage-structuring is commonly an adaptation to habitats that show extreme temporal or spatial heterogeneity in resource availability. As such, stage-structured populations generally have ≥ 1 life stages which can persist through long periods of unfavorable environmental conditions, or are capable of rapid long distance dispersal. When favorable conditions return, these species can "irrupt" or rapidly increase in numbers and distribution because of their high reproductive potential. This allows these species to quickly colonize or occupy all available ephemeral habitats. Such species are often characterized by very high fecundity and display extreme *r*-oriented population dynamics in ≥ 1 life stage. Examples of stage-structured populations include many irruptive insect "pests" such as the eastern spruce budworm (*Choristoneura fumiferana*).

In contrast, age-class structured populations are comprised of ≥ 2 age classes of individuals born at the same time, or cohorts, in a single-stage life cycle. Often age-class structuring can be complex, such as in long-lived vertebrates which can have many age classes, including newborns and multiple subadult and adult age classes. In contrast to irruptive population dynamics, age structure in age-class structured species generally affects population dynamics through more subtle changes in annual population growth rate in response to annual variation in habitat quality and population density. Because of differing sensitivity to habitat conditions, this variation can affect survival and fecundity rates of different age classes unequally, resulting in often subtle effects on population rates of increase and compensatory effects among age classes. However, the effects of age-class structuring on population dynamics are extremely varied, and go well beyond simply differences in age-specific survival or fecundity. Such species are typically more *K*-selected and show more typical *K*-selected population dynamics, such as a general trend towards density dependent regulation, although this is not always the case. Extreme examples include megaherbivores such as African elephants (*Loxodonta africana*) and black rhinoceros (*Diceros bicornis*).

Some species show both age-class structuring and stage structuring, and their population dynamics include aspects of each. Usually, survival and fecundity is dependent upon age and phenotypic quality as in age-class structured species, but individual reproductive output can be enormous and ≥ 1 life stages can persist through long periods of dormancy to allow rapid colonization or recolonization of habitats. Examples include fire-adapted tree species with serotinous seed cones, such as lodgepole pine (*Pinus contorta*).

Age-structuring is extremely important in management of populations and species. For example, harvest management of game species and commercial fisheries manipulates age structure to achieve optimal sustainable yields. Integrated pest management programs preferentially target certain life stages of pests to limit damages associated with agricultural or other pests. Thus, understanding age structuring in its varied forms can facilitate biodiversity, sustainable yield management, and limit conflicts between humans and other components of ecosystems.

☆*Change History:* October 2017. L Bender updated the text and further readings to the entire article, including adding additional examples and new Figures 1 and 3.

Age Structure Effects in Age-Class Structured Species

In age-class structured populations, the effects of age structure on population dynamics arise primarily because of varying age-specific mortality and fecundity rates and differences in the viability of individual cohorts. Age structure can also influence population dynamics through effects on timing and length of breeding seasons, population-level productivity, total population-level mortality, and sensitivity to population regulating mechanisms such as density dependence. In turn, age structure can be influenced by actions such as harvesting and environmental variations including severe weather events, age-specific predation or parasitism, and presence of disease.

Age class effects are most definitively seen in populations that show ≥ 3 distinct age classes, including juveniles, prime-aged adults, and senesced adults. Although effects of population age structure can be seen in populations with ≤ 2 distinct age classes, they are generally less pronounced. Moreover, individuals in each cohort may be independently influenced by environmental conditions experienced by their mothers (maternal effects) or experienced by all individuals in the cohort as juveniles (cohort effects), and this variation can persist throughout life and lead to differences in fitness among cohorts that affect many aspects of population dynamics (see *Cohort effects* below). Because age structure has such a strong effect on the productivity and mortality patterns of populations and can vary markedly between years within a population, even for long-lived *K*-selected species, the annual rate-of-increase of populations can also vary markedly simply because of annual changes in the age structure of a population.

Age Structure in Populations

The age structure of populations is driven primarily, but not exclusively, by age-specific mortality rates (Fig. 1). In general, juveniles have relatively low but highly variable survival; prime-aged adults have high and relatively constant survival; and aged, senesced adults have lower and variable survival. Age-specific mortality rates determine both how long individuals in a population can potentially live and the proportions of individuals surviving into each successive age class.

For example, if the natural mortality rate of all age classes in a population is 0.10, then on average 90% of individuals in a given age class will survive the year and be recruited into the following age class. The result of incremental mortality through years is that populations tend to show a "pyramid-shaped" age structure, with fewer individuals in each successive age class (Fig. 1). Numbers of individuals in each successive age class is dependent upon (1) numbers in the previous age class the previous year, (2) age-specific survival, and (3) possibly emigration and immigration. However, if additional mortality such as additive hunting harvest is placed on adults in a population, mortality rates for each age class would increase and the number of individuals in each successive age class would be lower (Fig. 1). Similarly, mortality factors that preferentially act on juveniles, such as predation, can result in cases

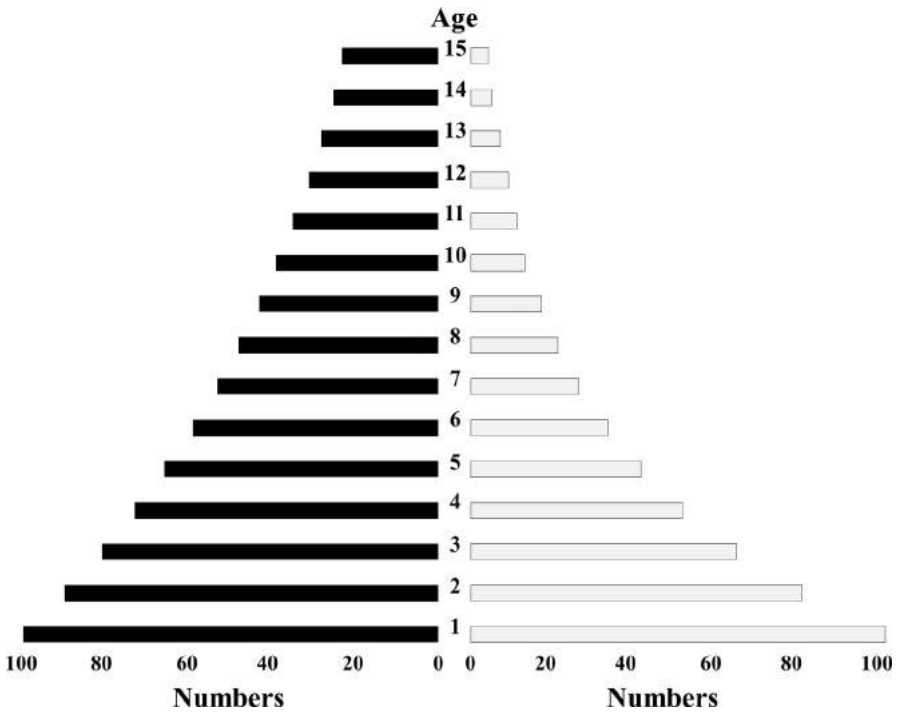


Fig. 1 Age structure of a North American elk population showing "pyramid-shaped" age structure and differential survival into older age classes based on an adult survival rate (S_{adult}) of 0.90 (▪) and 0.80 (□). Note that more individuals survive into older age classes as survival rates increase. Mean age of the population is 6.1 and 4.4 years-old at $S_{adult} = 0.90$ and 0.80, respectively, showing that population age structure declines (becomes younger) as population mortality rates increase. In this simulated population, no individuals survive beyond age 15.

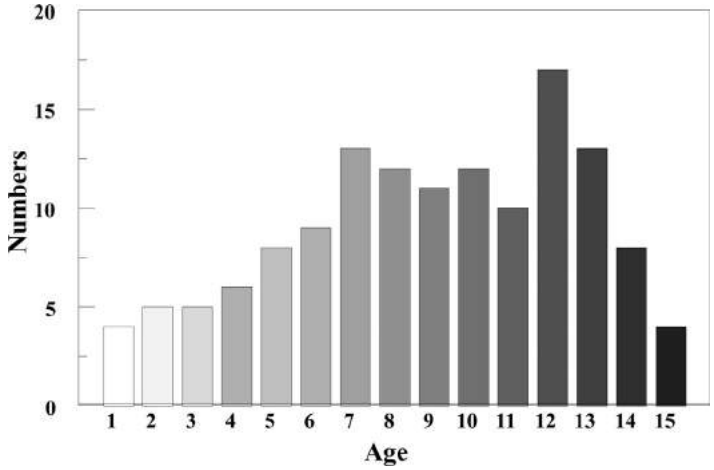


Fig. 2 Age structure of a North American elk population experiencing very high levels of predation on newborn calves. Note that successive recent years of poor recruitment into the population result in a population age structure that is shifted to the right or dominated by older age class individuals. Such an age structure would further decrease potential recruitment because older individual elk show reproductive senescence and produce fewer juveniles than do individuals in their reproductive prime (ages 2–8).

where a population is primarily comprised of older adults (Fig. 2). Therefore, under most conditions the age structure of a population, and the effects of that age structure on population dynamics, are driven by sex and age-specific mortality rates. Because changes in age structure affect the numbers of individuals that will die or be born each year, the age structure of a population can have a marked effect on year to year changes in population size.

Management of populations in most cases involves managing population productivity, or production and survival of juveniles, because productivity typically accounts for the majority of variation in annual rate of population increase. The more productive a population, the greater the potential return, regardless of whether that return is a sustainable harvest or a high rate of population increase for recovery of a rare species. The effects of age structure on population dynamics are most obviously manifest in population productivity through effects on individual fertility, fecundity, and probability of raising a juvenile to weaning or recruitment (see *Effects of female age structure*, below). However, age structure can also influence many other components of population dynamics including sensitivity to density dependent effects (see *Effects on population regulation*, below) and consequently population-habitat relationships, and numbers of individuals that die annually due to differing age-specific mortality rates (see *Effects of age-specific mortality*, below).

Effects of Female Age Structure

Age of females is an important contributor to different levels of fertility, fecundity, and survival of offspring. Generally, juveniles have much lower fecundity than do adults, and prime-aged adults have much higher fecundity than do older or senesced adults (Fig. 3). The larger and longer-lived a species, the greater these age-related differences. For example, desert mule deer (*Odocoileus hemionus eremicus*), fecundity of which is illustrated in Fig. 3, can live to be 20 years-old. Thus, there are many more age classes in the older, senesced category ($\geq$ 7 years-old = 14 age-classes) than in the juvenile categories (fawns and yearlings). Consequently, a population with an older adult age structure (i.e., one that has a greater mean age, because of more individuals in older age classes) tends to be less productive than a population of the same species with a younger age structure because more individuals are present in reproductively senesced age-classes. Actions aimed at maximizing productivity thus often try to increase mortality rates on the population above those attributable to natural mortality alone. The goal of these strategies is to decrease the age structure of the adult population in order to have a greater proportion of the population in prime reproductive categories (Figs. 1, 3, and 4), thus maximizing per capita productivity.

Maternal age can also influence the likelihood of successfully raising a juvenile to recruitment, or age of reproduction. This can be due to experience of mothers; prime-aged mothers are more likely to be larger and have a higher social rank in long-lived species, which leads to better territories and greater capture of resources, both of which result in greater production and survival of juveniles and hence greater lifetime reproductive success for these individuals. Moreover, birth attributes of neonates such as size or birth mass tend to be lower for older, senesced mothers and younger, inexperienced mothers, and birth attributes are strongly related to juvenile survival in many species. However, this can also be due to phenotypic quality of eggs or juveniles. For example, older (but non-senesced) fish of many species (e.g., Atlantic cod [*Gadus morhua*], black rockfish [*Sebastes melanops*]) produce higher quality eggs with consequent greater larval survival. The higher quality offspring show faster growth rates and greater resistance to starvation. While female size has long been recognized as affecting production and survival of offspring in many fish species, female age may be a much better predictor of larval performance in some species.

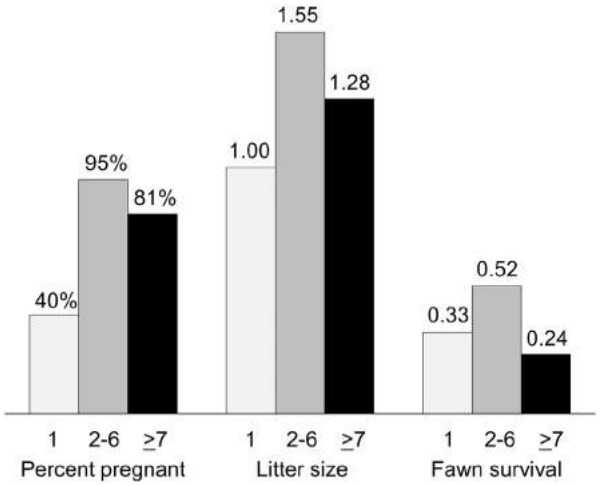


Fig. 3 Age effects on probability of pregnancy, numbers of juveniles produced, and fawn survival of desert mule deer. Because deer can live to be 20 years-old, there are potentially many more older, reproductively senesced age classes in the population. Thus, populations that show older mean ages tend to be less productive than populations with younger age structure. In this example, prime-aged adult females recruit 0.76 (90% CI = 0.60–0.94) fawns/year, compared to 0.11 (0.00–0.44) and 0.24 (0.10–0.42) for yearling and senesced females, respectively.

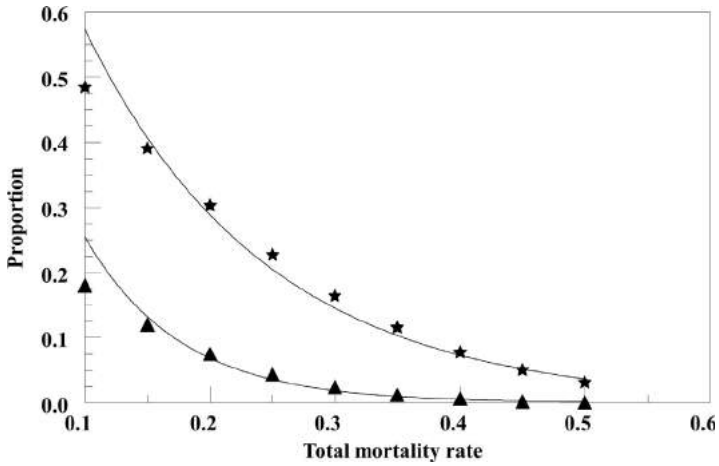


Fig. 4 Effects of increasing adult mortality rate on the proportions of oryx (*Oryx gazella gazella*) > 5 (★) and >10 (▲) years-old on White Sands Missile Range, south-central New Mexico, USA. Note that as total mortality rate increases, a smaller proportion of individuals are able to survive into older age classes. Observed total mortality rate in this population (approximately 0.26) results in a decrease of approximately 52% and 78% from the proportions of >5 and >10-year-old oryx in the population given an approximate "natural" annual mortality rate of 0.10.

Maternal behavior also varies according to age of females and number of pregnancy experiences, which can affect survival of juveniles. Prime-aged mothers are more successful in rearing juveniles than are younger females, particularly when threatened by predation. In large mammals, age of the mother has been associated with losses of neonates in the first week of life, when most newborns die; far more juveniles from primaparous females are lost compared to multiparous females. Experienced mothers are less likely to orphan juveniles due to a breakdown in the imprinting process. For many species, prime-aged females protect juveniles better than younger females. For example, when subjected to human disturbance or simulated predator threats, prime-aged female white-tailed deer (*Odocoileus virginianus*) move their fawns to more secure bedding sites, whereas young mothers often do not. Experienced females also commonly show complex distraction behaviors that can lead a predator away from juveniles, and may also actively defend juveniles by attacking predators.

Effects of age are often correlated with body mass or size, as older females tend to be larger. Juveniles born to larger mothers often show greater survival, likely because juveniles born to heavier mothers are larger and size at birth is strongly related to survival to recruitment (e.g., mammals) or because juveniles show faster growth rates and greater resistance to starvation (e.g., fish). Similar effects may be seen as litter size increases. For example, roe deer (*Capreolus capreolus*) fawns born to relatively light mothers or in twin or triplet litters have higher mortality rates than those born to heavy mothers or in smaller litters, because single-born fawns usually weigh significantly more than individual twins or triplets. For many species maternal mass in late pregnancy is correlated with

juvenile birth mass, and probability of survival is lower for juveniles born to females with lower than average body mass. Consequently, maximum juvenile-rearing success occurs when physically mature, multiparous females comprise the bulk of the breeding population, rather than juvenile or senesced adults. Senescence may not be solely a product of age, but may also be influenced by reproductive performance early in life. For example, birds with large clutches early in life tend to senesce more quickly, as do red deer (*Cervus elaphus*) that rear more calves early in life.

Age of the mother may also affect the sex ratio of juveniles, although climatic and other environmental conditions can also affect sex ratio of juveniles through effects on condition of mothers during gestation. In longer lived species, prime-aged females in good condition often produce more male offspring than females in poor condition, and both senesced and younger females are frequently in poorer shape and thus may produce more female young. The age structure of a population can consequently affect both juvenile and adult sex ratios, and thus the potential rate of increase of populations, simply by influencing the proportion of females born into the population. Having more adult females in prime reproductive classes increases population rates of increase because of their higher fecundity and greater likelihood of successfully raising a juvenile (Fig. 3).

Effects of Male Age Structure

Age structure of males can influence breeding dynamics of age-structured populations. In polygynous species, populations with greater numbers of older, prime-aged males have earlier, shorter, and less socially-disruptive breeding periods. Conversely, where fewer older males are present, breeding periods tend to be longer and females are frequently bred later in the season. Because later breeding may lead to later birth dates, and later birth dates lower juvenile survival, the age structure of the male population can potentially influence both pregnancy rates and survival of juveniles, thus affecting population rate of increase. However, this cascade of effects has not been conclusively demonstrated in free-ranging populations. Other factors such as female nutritional condition may overwhelm any effect of male age structure and breeding date by allowing females to shorten the length of gestation. Further, ages of males tending harems may not be dominated by prime-aged males until male/female ratios are very high even in polygynous species; thus, younger males may breed a significant proportion of females regardless of male age structure. Consequently, observed recruitment of juveniles has been shown to be independent of adult sex ratios and male age structure in several polygynous ungulate species. High harvest rates of males have also driven male age structure and male/female ratios well below thresholds theorized to affect population-level productivity, without any significant decreases in population productivity being documented in free-ranging populations.

Effects of Age-Specific Mortality

Survival rates of age-structured species can vary among age classes. Generally, juveniles have low but extremely variable survival whereas prime-aged adults have high and relatively constant survival. As individuals become older and move into the age classes that show reproductive senescence, survival rates typically decline. Gaillard et al. (2000) reviewed survival studies of cervids and found preweaning juvenile, postweaning juvenile, prime-aged adult female, and senescent adult female survival (CV) of 0.62 (0.25), 0.71 (0.28), 0.86 (0.09), and 0.79 (0.14), respectively, with similar patterns seen in bovids. Although fairly invariant with respect to large mammals, these patterns may be less pronounced in other taxa, such as rodents. Importantly, not only is survival of juveniles and older, senesced females lower than prime-aged females, but it is also from 2 to 3 times more variable.

Increased mortality rates in senesced adults can be due to a variety of factors, including decreased body mass, hormonal imbalances, lower individual nutritional condition, wear of teeth, and decreased immuno-competence. For juveniles, smaller body size, lower energy reserves, greater energy requirements per unit body size, and inexperience all increase vulnerability relative to adults. Because of their sensitivity to environmental variation, juveniles are particularly affected by temporal variation in environmental conditions and hence the annual variation associated with survival rates (as well as fecundity) is much higher than in adults. Lowered survival rates in senesced adults results in their lowered fecundity having less impact on overall population-rate of increase than if survival remained comparable to prime-aged adults.

The sensitivity of juveniles to annual variation in resource availability can strongly affect population-level production and mortality. If density or resource stress is low, juveniles can attain sexual maturity and breed earlier in life, and productivity of young females may equal that of prime-aged adults. These effects are important when one considers proportional contributions to population rate of increase. Adult female survival has the greatest potential effect on rate of increase (i.e., the highest elasticity) but, as noted above, typically varies little annually. Conversely, juvenile fecundity and juvenile survival have much lower elasticity, but vary greatly annually. Consequently, the majority of the variation in annual changes in population size is a result of changes in juvenile survival and fecundity. Population age structure can thus exert a strong regulating effect on population rate of increase by dampening potential population declines when mortality of adults is higher as well as increasing population rates of increase during periods of high resource availability through density dependent changes in survival and fecundity of juveniles. Despite their lower fecundity and survival as compared to prime-aged adults, juveniles can be critical contributors to population-level productivity.

Compensatory Mortality in Age-Class Structured Populations

The age structure of a population can help buffer populations from changes in the mortality patterns of specific age classes. An important but poorly understood aspect of population-level mortality patterns is the concept of compensatory mortality, which

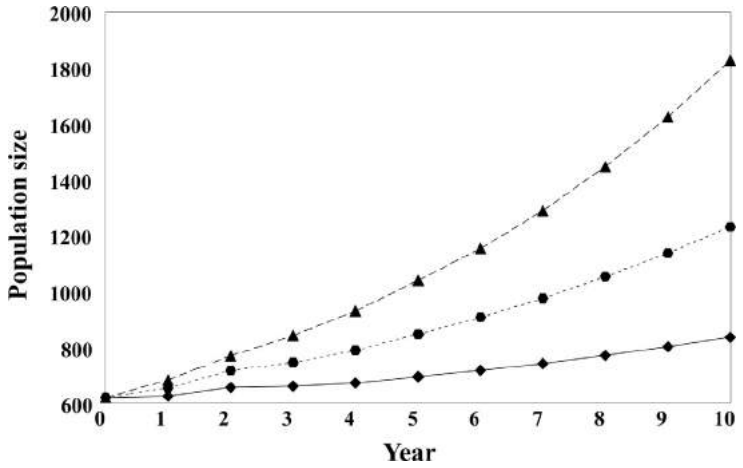


Fig. 5 Dynamics of females in a North American elk population showing effects of partial compensation in juvenile survival rate for changes in adult survival (compensatory mortality). Curves show population growth when $S_{adult} = 0.85$ and $S_{juvenile} = 0.50$ (▲); growth when S_{adult} is decreased to 0.75 by some additional mortality factor such as harvesting and $S_{juvenile}$ remains unchanged (♦); and partially compensatory growth when S_{adult} is decreased to 0.75 but $S_{juvenile}$ increases to 0.60 in compensation (●).

postulates that an increase in any mortality factor(s) can result in declines in other mortality factors, such that the total mortality rate either does not change or increases less than if the additional mortality was completely additive to existing mortality (Fig. 5). Because adult survival in most age-structured populations is high and shows little variation under conditions less than significant resource stress, less opportunity exists for mortality to be compensatory among adults in a population simply because adult mortality rates are typically at or near the chronic minimum, i.e. there is little excess mortality to "trade off" among mortality sources. However, age structuring of a population presents a mechanism whereby mortality at the population-level can be compensatory even if additive within certain or most age classes (i.e., adults) through compensatory responses in juvenile survival (Fig. 5).

Because juveniles are smaller and have a higher metabolic rate, they require greater resources per unit of body mass, are more susceptible to density stress and other resource limitations, and thus are more vulnerable to most causes of mortality as compared to adults. Consequently, any reduction in population size disproportionately benefits juveniles because they are more resource stressed than adults in any given set of environmental circumstances. Therefore, even if mortality is primarily additive on adults, the reduction in numbers of adults may result in a density dependent increase in juvenile survival. Hence, age structuring within a population provides a mechanism whereby compensatory effects can act to dampen changes in population size due to mortality factors such as harvesting. This is another means by which age structuring helps to regulate populations, buffering populations from large annual changes in population size potentially brought on by age-selective mortality factors such as harvesting (Fig. 5). The need for replacement of adults is also one reason why most *r*-selected and even some *K*-selected species significantly overproduce young. This "doomed surplus" provides ample recruits to replace adults despite very high juvenile mortality.

Cohort Effects

A "cohort effect" is a phenomenon where cohorts of a population differ from each other in some key attribute(s), such as body mass. Fetal sex ratio, birth mass, birth dates, rate-of-growth of juveniles, survival over the first winter, age of first reproduction, and adult survival rates are often related to the influence of the year of birth, which itself is a product of the environmental conditions cohorts face at birth as well as the nutritional condition of their mothers (itself a reflection of environmental conditions faced at or prior to birth). Environmental conditions that can influence birth and early growth attributes include droughts, abnormally high rainfall, late snowfall, and high population density. These and other factors influence birth attributes and early development of juveniles, and these early effects may persist and affect phenotypic quality throughout the lifetime of the cohort. Because many cohort effects such as lower body mass are tied to survival and reproductive fitness, cohort effects can influence population dynamics above and beyond the effects associated with typical age-specific reproductive potential or survival patterns.

Cohort effects are most often expressed as lowered reproductive output in the population as a whole as well as lowered lifetime reproductive success of individuals in that cohort. Differences among cohorts in lifetime reproductive performance have been demonstrated in a diverse array of taxa including large mammals, marine mammals, birds, fish, insects, and plants. Although populations need not have complex age-class structure to show cohort effects (e.g., annual plants), cohort effects most commonly occur because of environmental or density stress, which most commonly affect relatively *K*-selected species that also tend to have a complex age-class structure.

Cohort effects can be either short-term or long-term. Short-term cohort effects influence age structure and population dynamics by affecting the numbers of individuals in the cohort that live to be recruited. Long-term cohort effects affect overall reproductive

success of cohorts over time through their effects on phenotypic quality. Cohort effects can also vary by sex; for example, cohort may affect growth of males but not females due to the greater reproductive activities of males in polygynous species.

Age-Class Structure and Population Regulation

Age structuring in a population can help regulate numbers through a variety of processes noted above, usually associated with density-dependent changes in juvenile survival and fecundity rates in response to resource stress. These mechanisms allow age-class structured populations to exhibit much higher levels of population productivity during periods of resource abundance because of earlier sexual maturity and greater fecundity of juveniles. Further, increases in juvenile survival rates can compensate for decreases in adult survival to some degree, making the population more resilient to additional mortality such as associated with harvesting (Fig. 5). The mechanism behind this response is also density dependence. Thus, age structuring in a population, through variation in age-specific survival and fecundity rates, contributes significantly to density dependence as a regulating mechanism of populations. Typically, the more age-structured a population, the greater the number of potential density dependent responses available (i.e., increases in juvenile survival or fecundity; increases in young adult [yearling] survival or fecundity; increases in fecundity of older, senesced females, etc.) and hence the more sensitive density dependence becomes as a means of population regulation.

However, extreme age structuring in a species can limit the sensitivity of density dependence as a population regulating mechanism. Megaherbivores such as the African elephant can potentially have a dramatic effect on the vegetation of their habitat through foraging. The long life ($>$50 years) and large number of juvenile (pre-reproductive) age classes ($\geq$12) contributes significantly to the potential of these species to impact vegetation communities through herbivory. Even if reproduction is shut down due to density effects, there are still $\leq$12 cohorts of juvenile elephants that will grow and increase in individual mass, adding to the existing adult biomass. Larger elephants consume greater amounts of plant biomass, and it is the total biomass of elephants that determine herbivory levels more so than total numbers. Because adult biomass declines only very slowly due to extremely high survival rates (which allow the long life span) and juveniles continue to grow, total elephant biomass may actually increase for up to 12 years after density dependence has shut down reproduction in the population due to resource (food) stress. Thus, plant communities can be impacted by feeding above their sustainable levels; the increase in total elephant biomass continues well after the capacity of plant communities to support elephants is exhausted. Consequently, elephants can potentially overutilize their habitats because the extreme age structuring in the population makes density dependence a relatively insensitive means of balancing population biomass with available resources.

Age Structure Effects in Stage-Structured Species

Multi-stage life cycles generally represent an adaptation to environments that show extreme temporal or spatial heterogeneity in availability of key habitats. The *r*-selected dynamics, storage effect, and high dispersal potential seen in some life stages allows for very rapid population responses when ephemeral habitats become available. Many of the effects of age-class structuring on population dynamics detailed above are seen to some degree in at least 1 life stage in stage-structured species as well. However, population dynamics of stage-structured populations are also influenced by each life stage potentially having a unique life history. For example, different life stages of many invertebrate and amphibian species may have completely different physical forms and ecological requirements, and behave as distinct life forms with unique niches rather than as different age classes of the same life form. Because effects of age structure on age classes in the different life stages of multi-stage species can be similar to those detailed above for age-class structured single-stage species, this section will emphasize unique aspects of population dynamics driven by the distinct stage-structuring of these species.

Dynamics of Stage-Structured Populations

High fecundity and ecological complexity resulting from the multi-stage life cycle result in stage-structured populations showing a greater range of population dynamics than age-class structured populations. Although generally much more density independent than age-class structured populations, population dynamics of any life stage can range from density independent to intensely density dependent (i.e., amphibian larvae in ephemeral pools). At the population level, dynamics can range from density independent to populations that show multiple stable states or equilibrium points. At the extreme, high fecundity and extremely differing life stages can result in chaotic behavior in population dynamics.

For example, populations of spruce budworms and other insect "pests" may show multiple stable equilibrium points. Usually, the lower is associated with habitat conditions of low quality (such as immature balsam fir [*Abies balsamea*] and white spruce [*Picea glauca*] forests in the case of spruce budworms) in combination with predation and parasitism on the 6 larval instars (caterpillar) and adult (moth) stages. As habitat quality increases (forests mature) budworm populations can achieve a "critical mass" that allows them to irrupt to population levels orders of magnitude greater than the low density equilibrium. The irruption is facilitated by the highly mobile adult (moth) stage, which can disperse hundreds of kilometers in a single generation. Because of high mortality of host trees associated with large-scale outbreaks, conditions necessary for outbreaks may occur approximately only every 30 years. The time span between irruptions allows affected forests to regenerate and adjacent forests to attain a suitable mature forest structure, highlighting both the temporal and spatial variability in suitable habitats for this stage-structured species.

Stage Structuring and Colonization Ability

Multi-stage life cycles allow full use of habitats that are ephemeral in space or time. Generally, these more *r*-selected species are able to colonize new areas much faster than *K*-selected species. Thus, they are often early pioneers in succession. Some of this mobility is due to age structuring. For example, seeds can lie dormant for years until ecological conditions initiate germination and development into adult plants. Similar dormant phases are seen in many invertebrates and this phenomenon is called the "storage effect." Age structuring in such species allows the species to persevere in an area despite adverse environmental conditions, and thus allows the population to rapidly increase when conditions again become favorable. Such an age structure effect on population dynamics is a significant advantage over species which maintain a similar life form in all age classes; the latter may be extirpated during adverse periods if adverse conditions last longer than a generation, and would be dependent upon dispersal from favorable areas to recolonize. An age structure comprised of distinct life forms, especially if one or more is capable of extended dormancy, allows rapid population responses during favorable periods and is thus a significant advantage for early successional or pioneer species, or for species existing on the fringe of habitable conditions. Highly mobile life stages allow rapid colonization of habitats even if no "storage effect" is present.

Habitat and Population Dynamics of Different Life Stages

A diverse array of species, including plants, insects, and amphibians, have life stages with different life forms, and thus show unique population dynamics because of this life history. Ecological factors that limit one life stage (such as insect larvae) may have no effect on other age classes (such as the adult insect). Consequently, population dynamics of these species are less influenced by subtle changes in environmental conditions than are *K*-selected species. This, in combination with high fecundity, results in population dynamics being far more density independent. For example, outbreaks of many insect pests depend upon an environment that has (1) habitats which provide for growth and development of eggs and larvae and (2) habitats that allow adults to survive, reproduce, and disperse. Habitat limitations for any life stage can result in populations being limited at relatively low levels regardless of how suitable environmental conditions are for the other life stages. Even under suitable conditions, a minimum population size may be necessary to achieve a "critical mass" and allow the population to rapidly increase numbers and extent of distribution. Such a dynamic is called a manifold or break-point in the species growth curve; if ecological conditions allow the species to grow beyond this breakpoint, the species can irrupt.

The environmental conditions necessary to trigger rapid population establishment and growth may be extremely different among life stages. At one extreme, the environmental effect that triggers the irruption of juveniles may be the destruction of the adult life stage. For example, fire-adapted shade-intolerant trees such as lodgepole pine require full sunlight, mineral soil, and heat (to open serotinous cones and free seeds) to germinate. These environmental conditions usually only occur following a stand-replacing fire, which kills the existing population of overstory lodgepole pine trees. This in turn creates the conditions necessary for rapid germination of seeds and widespread establishment of seedlings. Without loss of the adult overstory, germination and establishment of the new cohort of lodgepole pine would not occur.

Conversely, habitat conditions necessary for rapid population growth can be similar for all life stages. Outbreaks of spruce budworms require extensive areas of mature spruce/fir forest to occur, as this forest structure provides necessary habitats for larvae to feed, molt, and pupate, as well as for adult moths to lay eggs. Thus, large scale outbreaks may only be ended by extensive mortality of preferred host trees due to defoliation, which decreases available feeding habitat for larval instars and egg laying habitat for adult moths. Consequently, numbers decrease to levels where population size can be limited by natural predators, limited resources (i.e., lack of mature spruce/fir forests), and weather, despite the ability of the adult moths to disperse widely and find most suitable trees.

Implications of Different Life Stages

The irruptive dynamics of many stage-structured species can result in significant conflicts with humans, such as seen with agricultural or forestry pests. In such cases, the differing habitat requirements and unique vulnerabilities of life stages are frequently exploited in control strategies. Integrated pest management specifically targets the most vulnerable life stage(s) of economic pests; once the lifecycle is understood, optimal intervention points in the lifecycle can be identified. For example, noxious annual weeds can be controlled with mulches and pre-emergent herbicides to prevent germination of seeds more easily than the adult plant stage can be controlled. Species such as mosquitoes (Culicidae) are frequently targeted with larvaecides or by limiting the availability of larval habitat, i.e., stagnant pools of water.

The greater the number of life stages, the greater the opportunity to find and exploit a vulnerable life stage. For example, the diamondback moth (*Plutella xylostella*) is a severe economic pest of cruciferous vegetables including cabbage, kale, and broccoli. Its life stages include egg, 4 larval instars, pupae, and adult, and the egg, larval, and pupae stages are each vulnerable to a different species of parasitoid wasp. Thus, rapid control of outbreaks can be attained by concurrent application of all wasp species, whereas long-term control may by attained by limiting management to use of only 2 species such as the larval and pupal parasitoids. Such life stage-based biological control allows management of the pest without eradication or the use of pesticides (both of which may have other undesirable consequences in the ecosystem) by exploiting vulnerabilities in stage (age) structure to interrupt irruptive population dynamics.

Further Reading

Beldade R, Holbrook SJ, Schmitt RJ, Planes S, Malone D, and Bernardi G (2012) Larger female fish contribute disproportionately more to self-replenishment. *Proceedings of the Royal Society B* 279: 2116–2121.
Clutton-Brock TH, Guinness FE, and Albon SD (1982) *Red deer: Behavior and ecology of two sexes*. Chicago, IL: University of Chicago Press.
Gaillard J-M, Festa-Bianchet M, Yoccoz NG, Loison A, and Toïgo C (2000) Temporal variation in fitness components and population dynamics of large herbivores. *Annual Review of Ecology and Systematics* 31: 367–393.
Gaillard J-M, Loison A, Toïgo C, DeLorme D, and Van Laere G (2003) Cohort effects and deer population dynamics. *Ecoscience* 10: 412–420.
Jewell PA, Holt S, and Hart D (eds.) (1981) *Problems in management of locally abundant wild mammals*. New York, NY: Academic Press.
Lindstrom J and Kokko H (2002) Cohort effects and population dynamics. *Ecology Letters* 5: 338–344.
Link WA, Royle JA, and Hatfield JS (2003) Demographic analysis from summaries of an age-structured population. *Biometrics* 59: 778–785.
Metcalf RL and Luckmann WH (1994) *Introduction to insect pest management*, New York, NY: John Wiley and Sons.
Radcliffe EB, Hutchison WD, and Cancelado RE (2009) *Integrated pest management: Concepts, tactics, strategies and case studies*. New York, NY: Cambridge University Press.
Rhodes OE Jr., Chesser RK, and Smith MH (1996) *Population dynamics in ecological space and time*. Chicago, IL: University of Chicago Press.
Shelton AO, Munch SB, Keith D, and Mangel M (2012) Maternal age, fecundity, egg quality, and recruitment: Linking stock structure to recruitment using an age-structured Ricker model. *Canadian Journal of Fisheries and Aquatic Sciences* 69: 1631–1641.
Silvertown JW and Charlesworth D (2001) *Introduction to plant population biology*. Oxford, UK: Blackwell Scientific Publishing.
Skalski JR, Ryding KE, and Millspaugh JJ (2005) *Wildlife demography: Analysis of sex, age, and count data*. Burlington, MA: Elsevier Academic Press.
Turchin P and Taylor AD (1992) Complex dynamics in ecological time series. *Ecology* 73: 289–305.
Williams BK, Nichols JD, and Conroy MJ (2002) *Analysis and management of animal populations*. San Diego, CA: Academic Press.

Altruism

KR Foster, Harvard University, Cambridge, MA, USA

Introduction

> This brings us to the central theoretical problem of sociobiology: how can altruism, which by definition reduces personal fitness, possibly evolve by natural selection? (Wilson, 1975, p. 3)

Altruistic behaviors, which reduce the personal reproduction of an actor and benefit another individual (**Fig. 1**), are found in a diverse set of organisms, ranging from microbes, through social insects, to higher vertebrates and humans (**Fig. 2**). Altruism presents a conundrum for evolutionary thinking because Darwin's theory of natural selection appears to suggest that selfish and competitive strategies are favored over evolutionary time. Why would natural selection select for a behavior that reduces personal reproduction?

As we will see, altruism can evolve when the actor and recipient carry the same genes, at one or more loci – the actor can then increase copies of their genes through the recipient's reproduction. This explanation, which comes from what is called inclusive fitness (or kin selection) thinking, remains the key solution for the problem of altruism, as originally defined in the evolutionary literature. However, more than one usage of altruism has developed in behavioral ecology and with alternative definitions came other explanations, which will be discussed.

Care with definitions becomes even more important when one looks outside of biology. In common parlance, altruism is often taken to indicate an actor's psychological `intention' to act selflessly. The biologist's focus on `outcome' and evolutionary fitness (**Fig. 1**), therefore, can contradict the mainstream meaning of altruism in at least two ways. First, it allows the possibility of altruism in simple organisms, like microbes, that lack conscious intention. In addition, a gene for altruism will only be selected when the action increases its carrier's fitness – genes cannot be selected to produce behaviors that decrease their frequency. Evolutionary discussions of altruism, therefore, typically involve hidden genetic benefits, which can be troublesome for those that require altruism to be truly selfless.

The Birth of the Idea

It may be no coincidence that the concepts of altruism and natural selection were developed simultaneously in the mid-nineteenth century. Social philosophy was being much discussed and contrasting opinions abounded. On the one hand, Auguste Comte was popularizing altruism as part of his secular positivist religion, which argued for selfless acts that aid humanity and founded the new science of sociology. On the other hand, Herbert Spencer's individualism was fueling the fires of British industry. It was into this environment that Darwin proposed his individual-centered theory of evolution – natural selection.

With altruism based upon selflessness, and natural selection on selfishness, their conceptual collision would appear inevitable. However, this collision was barely evident at first. While Darwin did not use the term, his writings sowed the seeds for all modern explanations for altruism: the *Origin of Species* confidently proposes a mix of family relations, colony-level benefits, and parental manipulation to explain social insect workers (**Fig. 2**); and the *Descent of Man* appeals to both group-level thinking and reciprocation to explain what he called human sympathy. Furthermore, Herbert Spencer explicitly discussed altruism in biology and explained it through both family life and competition among tribes. It is also noteworthy that Spencer often took an outcome-based definition, showing that there have long been parallel traditions of thinking about altruism, one based on intention and the other on behavior (see the introduction). This said, Spencer's views differed significantly from modern definitions by taking reproduction itself to be altruistic.

In the hundred years following the *Origin*, evolutionary discussions of cooperation and altruism are spotty, and often less clear than Darwin's original writings. This includes Kropotkin's extensive discussion of cooperation, which appeals to both group selection and a, sometimes flawed, species-level argument. By the mid-twentieth century, however, it is clear that many authors understood how cooperative acts like worker sterility and human sociality could evolve through kinship, group selection, and reciprocal benefits. These include H. G. Wells (with Julian Huxley and G. P. Wells), R. A. Fisher, A. H. Sturtevant, A. E. Emerson, J. L. Lush, and Sewell Wright. However, these authors rarely used the term `altruism' – the notable exception being J. B. S. Haldane who colorfully compared his reader altruistically rescuing some drowning relatives to sterility in insect workers – and the concept anyway was given little space or attention. No one seemed to think that altruism was all that important:

> There will also, no doubt, be indirect effects in cases in which an animal favours or impedes the survival or reproduction of its relatives… Nevertheless such indirect effects will in very many cases be unimportant… (Fisher, 1930, p. 27)

 doi:10.1016/B978-0-444-63768-0.00003-2

Effect on actor \ Effect on recipient	+	–
+	Mutualism	Selfishness
– or 0	Altruism	Spite
	Cooperation	Competition

Fig. 1 The four types of social action based on their effect on the direct fitness (lifetime personal reproduction) of the actor and recipient. Altruism and spite can either have no or a negative fitness effect on the actor.

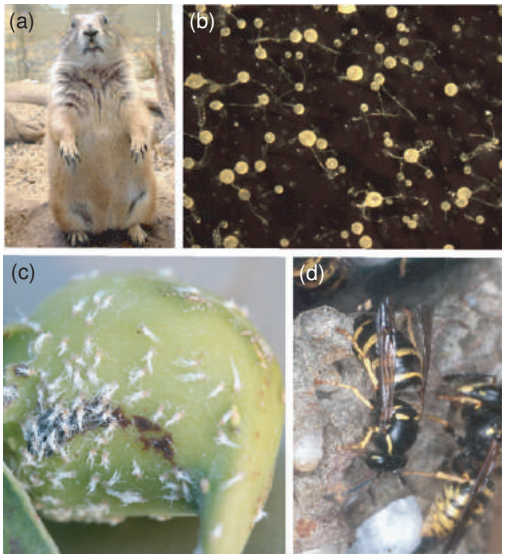


Fig. 2 Species that display altruistic behaviors. (a) Prairie dogs live in family groups in communal burrows or `towns'. When danger approaches, guard individuals will bark and warn others, at apparent cost to themselves. They also display cooperative brood care. (b) Fruiting bodies of the slime mold *Dictyostelium discoideum*. Thousands of cells aggregate together in these groups and many die altruistically to form a stalk that holds the others aloft as dispersal spores. (c) The gall-dwelling aphid *Pemphigus obesinymphae*. When disturbed, soldier aphids emerge and attack intruders. (d) The yellow-jacket eusocial wasp, *Dolichovespula saxonica*. Workers both altruistically work and lay eggs (shown) in this species. The level of worker reproduction, however, is kept low by both genetic relatedness and policing behaviors (Fig. 4). (c) Photo used with kind permission of Patrick Abbot.

Altruism via Inclusive Fitness (Kin Selection)

This all changed in the hands of a lonely London student, called Bill Hamilton, who dedicated himself to the first formal evolutionary analysis of altruism. His results are summarized with the following simple rule: altruistic behaviors will be favored by natural selection when

$$rb > c \qquad [1]$$

where b is the reproductive benefit to the recipient, c is the cost in terms of lifetime reproduction for the actor, and r is the genetic relatedness between actor and recipient (Fig. 3a). For example, selection can favor helping a sister ($r=0.5$) to raise her offspring when one can raise more than twice as many of her offspring (indirect fitness), than one's own (direct fitness), because this will increase the overall propagation of copies of the actor's genes. The sum of fitness effects through indirect effects and direct effects is the `inclusive fitness effect' of a behavior.

Semantics

Hamilton's definition of altruism requires the action to carry a cost to lifetime reproduction; a position solidified by E. O. Wilson who used this altruism as a center piece for his highly influential book *Sociobiology*. Hamilton's work also emphasizes the clarity that can come with gene-level thinking, which was later popularized by Dawkin's *The Selfish Gene*.

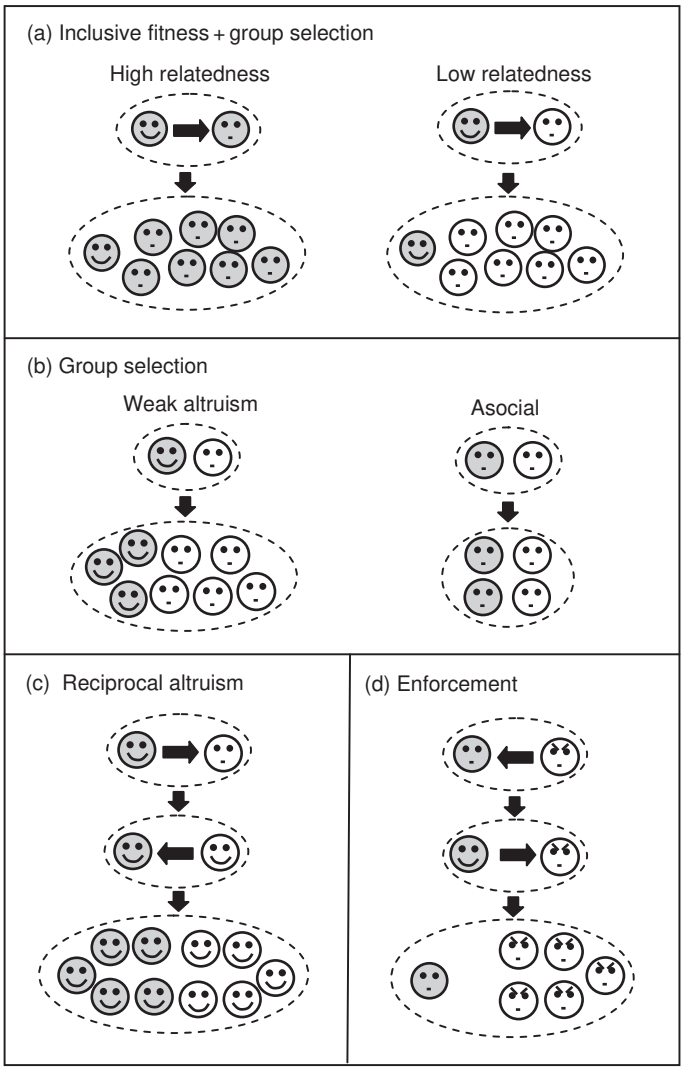


Fig. 3 Four nonmutually exclusive processes that generate altruism or altruism-like behaviors. Altruists are smiling and same-color individuals are genetically related. (a) Strong altruism can be selected when individuals are genetically related (left-hand side) but not when they are unrelated (right-hand side). (b) Weak altruism (gray, left-hand side) can be selected when helping the group feeds back on the actor, even though this increases the fitness of other group members more (white, left-hand side), because it increases reproduction relative to the population as a whole (right-hand side). (c) Reciprocal altruism can increase personal reproduction. (d) Enforcement: one individual forces altruism-like behavior from another individual that may or may not obtain a fitness benefit from their action. Note that the behaviors in (b) and (c) increase the personal reproduction of the actor, and are therefore not altruism in the original strict sense of Hamilton, which required a decrease in the personal reproductive fitness of the actor. Also, actions that arise purely through enforcement (d) are better viewed as adaptations of the enforcer, rather than altruistic adaptations of the helping individual.

Examples

The social insects are among the best and most discussed examples of Hamilton's altruism in behavioral ecology (Fig. 2d). Not only are they social, they are eusocial, with their division of work and reproduction among colony members. Comparable altruism occurs in other insects including some gall-forming aphids and thrips, which have a defensive soldier caste (Fig. 2c). In social vertebrates, sibling care is common that is no doubt often formally altruistic (Fig. 2a). However, individuals can usually reproduce later on, making it difficult to distinguish between true altruism, and behaviors with a delayed reproductive benefit. An interesting potential exception, however, is human menopause, which appears to reduce personal reproduction in order to help raise grand-offspring and under some definitions would constitute altruism. Altruism is also found in microbes (Fig. 2b). For example, individual cells often pay a growth cost to release a shared product, like digestive enzymes, which benefit other cells. There are good data to support the idea that relatedness drives altruism in the social insects (Fig. 4) and vertebrates, and the altruistic release of shared products in microbes has been shown to require genetic relatedness among cells.

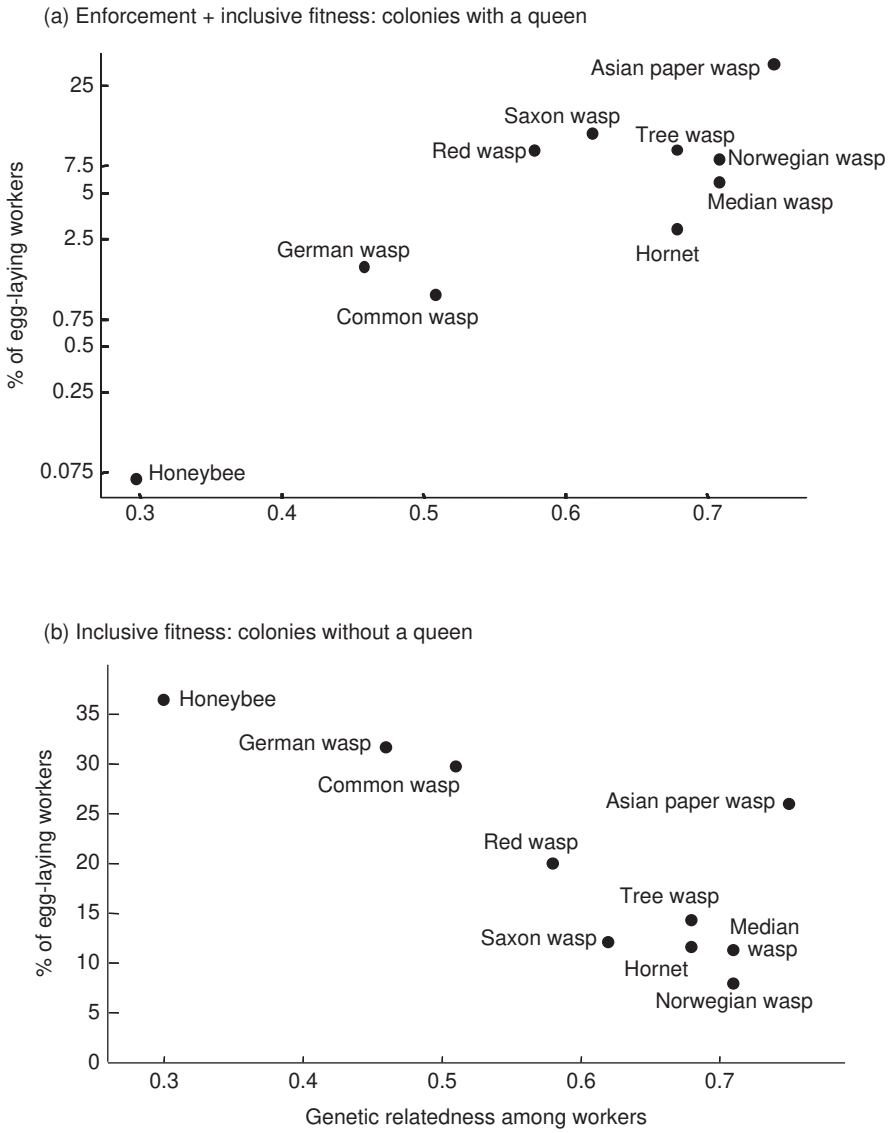


Fig. 4 Worker altruism is driven by a combination of inclusive-fitness effects and enforcement in social insect colonies. (a) Altruistic self-restraint due to enforcement. In colonies where the mother queen is alive, the workers can raise either the queen's or other workers' eggs. In species where relatedness among workers is high, they tend to raise the workers' eggs because they are highly related to them, but in species where relatedness among workers is low, like the honeybee, workers `police' each others' eggs and remove them. This reduces the benefits to worker reproduction which, alongside indirect fitness benefits, promotes reproductive self-restraint. (b) Altruistic self-restraint due to inclusive-fitness effects. If the queen dies the workers compete to lay eggs. However, when relatedness is high, many show altruistic self-restraint and do not attempt to reproduce. Reproduced from Wenseleers T and Ratnieks FL (2006) Enforced altruism in insect societies. *Nature* 444: 50.

Altruism via Group Selection

Another way to phrase the above explanation for altruism is in terms of group selection: when groups contain genetically related individuals (there is between-group genetic variance), selection can favor altruistic actions that invest in the group and increase its productivity. Importantly, and despite occasional misguided claims to the contrary, this logic is fully compatible with and complementary to inclusive-fitness theory: one can explain worker sterility by focusing on benefits to relatives (inclusive fitness), or the benefits at the colony level (group selection), but in the end both genetic relatedness and benefits are required for Hamilton's altruism (**Fig. 3a**). Like inclusive fitness, group-selection thinking can be traced back to Darwin (and also Spencer), and there were brief but explicit mathematical models by Haldane and Wright in the mid-twentieth century. However, it then got a bad name when Wynne-Edwards applied it uncritically to groups of unrelated individuals, such as large vertebrate populations, where individual-level selection will dominate and suppress altruism. It was correctly reformulated in the 1970s with the work of George Price, D. S. Wilson, and, once more, Hamilton. Price's work, specifically the Price equation, has since been central to the development of many branches of social evolution theory. This includes the development of cultural models of cooperation,

where imitation within groups increases between-group variance and promotes the spread of cooperative traits through `cultural group selection'. But there remains a point of departure between group selection and inclusive fitness when it comes to definitions.

Semantics

In the group-selection framework, altruism has been defined as cooperative acts that lower reproductive share in the group. However, this can include actions that increase personal reproduction (Fig. 3b), which is not altruism by Hamilton's definition. Consider, for example, a prairie dog (Fig. 2a) that contributes to the tunnels in its town and suffers a 10% decrease in its reproduction relative to another group member. This can evolve through selfish benefits alone if the tunnels allow all town members to double their reproduction. This is illustrated by a simple extension of Hamilton's rule:

$$\overbrace{\frac{b}{n}(n-1)r}^{\text{Indirect/kin benefit}}+\underbrace{\left(\frac{b}{n}\right)}_{\text{Direct/individual benefit}}.1>c \qquad [2]$$

where n is group size, b is the group benefit of which each individual gets a share b/n, and c is the individual cost. The individual-benefit term contains relatedness of the actor to itself, $r_{self}=1$, and even with no relatives in the group ($r=0$), tunneling can still evolve if there are feedback benefits to the actor. This type of behavior has been termed `weak' altruism (Fig. 3b) because it carries a personal (direct fitness) benefit, which distinguishes it from Hamilton's (strong) altruism, like that of sterile insect workers (Fig. 3a).

Examples

Because of the conceptual overlap, group-selected altruism includes all of the inclusive fitness examples above. Furthermore, feedback benefits of the sort that generate weak altruism must be common in many societies but are difficult to distinguish from inclusive fitness benefits. One example of weak altruism, however, is cooperative next founding by unrelated social insect queens. Here, co-investing in the colony can provide feedback fitness benefits when queens are later able to contribute to sexual offspring.

Altruism via Direct Fitness

In addition to weak altruism, several other processes that increase the personal reproduction of the actor (direct fitness) have been proposed to explain altruism-like behaviors. In the 1970s, Robert Trivers showed that helping can be selected when it increases the chance of return help, which he termed reciprocal altruism (tit for tat; Fig. 3c). A closely related idea is that of indirect reciprocity, whereby helping others improves reputation, which then increases the chance of being helped. More generally, feedback benefits to personal reproduction (direct fitness) are central to all manner of cooperative behaviors, including cooperation among genes and species, for example, plants provide nectar and insects pollinate in return:

> individual flowers which had the largest glands or nectaries, and which excreted most nectar, would oftenest be visited by insects, and would be oftenest crossed; and so in the long-run would gain the upper hand. (Darwin, 1859)

Semantics

A focus on direct fitness has led to a third general approach to modeling social evolution, called direct fitness or neighbor-modulated fitness theory, which again complements the inclusive-fitness and group-selection approaches. However, an action that evolves purely through direct-fitness feedbacks means increased personal reproduction and departs from Hamilton's altruism. Curiously, however, Hamilton started his original papers with a neighbor-modulated model (the fitness effect of others on the focal individual), before making a switch to inclusive fitness (the fitness effect of the focal individual on others) on which he based his rule.

Examples

Reciprocal altruism and indirect reciprocity are extremely important in human cooperation, but the requirement for recognition and memory of others means that they occur in relatively few other species. Potential examples include other primates and vampire bat blood-sharing, but inclusive fitness and group benefits also occur in these systems. More generally, however, cooperation that is selected due to direct-fitness feedback benefits is fundamental to social evolution, including between-species cooperation.

Altruism via Enforcement

Most recently, explanations for altruistic-like behaviors have focused upon a somewhat sinister mechanism: enforcement. This idea can be traced not only to the 1970s and Richard Alexander who proposed parental manipulation to explain insect workers (Fig. 3d), but also to Darwin, whose writings suggest something similar. While policing and punishment can explain apparent acts of altruism, however, one still needs an explanation for how policing, which carries a personal cost, can evolve the so-called `second-order problem'. For this, one must appeal again to some or all of the above theories: inclusive fitness, group selection, and direct benefits.

Semantics

If a helping behavior has arisen completely through enforcement, the primary evolutionary adaptation is in the enforcer, rather than the helping individual. The helping behavior, therefore, should probably not be considered an altruistic adaptation. This objection can be overturned, however, when an altruistic action evolves through a combination of enforcement and inclusive-fitness effects, as occurs in the social insects (below).

Examples

Enforcement, punishment, and policing are central to cementing the altruism in many social groups. This includes queen and worker policing in many species of social insects, whereby the queen and workers suppress the reproduction of other workers. The suppression means that natural selection favors workers that invest more in the indirect fitness from helping than direct fitness from their own reproduction, which increases altruistic self-restraint (Fig. 4). In addition, dominant males in macaque societies police and punish noncooperative individuals, and dominance hierarchies help to resolve breeding conflicts in many insect and vertebrate groups.

A Synthetic View of Altruism and Cooperation

Altruistic behaviors are a central component of many social systems. Any judgment on the extent of altruism in the natural world, however, will always depend upon definition. A requirement for conscious intention restricts altruism to creatures with sophisticated cognition, such as humans. However, the fitness-based definition of behavioral ecology reveals a wealth of additional examples, which typically arise through a combination of mechanisms. Centrally though, actions that decrease lifetime reproduction can readily evolve when there are indirect benefits that increase overall inclusive fitness. This is nowhere more obvious than in the social insects, where workers spend their entire life building, guarding, and foraging to raise a myriad of their relatives' offspring.

See also: Behavioral Ecology: Kin Selection. General Ecology: Cooperation

Further Reading

Abbot, P., Withgott, J.H., Moran, N.A., 2001. Genetic conflict and conditional altruism in social aphid colonies. Proceedings of the National Academy of Sciences of the United States of America 98, 12068–12071.
Bourke, A.F.G., Franks, N.R., 1995. Social Evolution in Ants. Princeton, NJ: Princeton University Press.
Darwin, C., 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. Edinburgh: John Murray.
Dawkins, R., 1976. The Selfish Gene. Oxford: Oxford University Press.
Dixon, T., 2008. The Invention of Altruism: Making Moral Meanings in Victorian Britain. Oxford: Oxford University Press for the British Academy.
Dugatkin, L.A., 2006. The Altruism Equation: Seven Scientists Search for the Origins of Goodness. Princeton, NJ: Princeton University Press.
Fisher, R.A., 1930. The Genetical Theory of Natural Selection. Oxford: Oxford University Press.
Foster, K.R., Ratnieks, F.L.W., 2005. A new eusocial vertebrate? Trends in Ecology and Evolution 20, 363–364.
Foster, K.R., Wenseleers, T., Ratnieks, F.L., 2006. Kin selection is the key to altruism. Trends in Ecology and Evolution 21, 57–60.
Gardner, A., Foster, K.R., 2008. The evolution and ecology of cooperation: History and concepts. In: Korb, J., Heinze, J. (Eds.), Ecology of Social Evolution. Berlin, Heidelberg: Springer.
Griffin, A.S., West, S.A., 2003. Kin discrimination and the benefit of helping in cooperatively breeding vertebrates. Science 302, 634–636.
Hamilton, W.D., 1996. Narrow Roads of Gene Land: The Collected Papers of W. D. Hamilton. Oxford: W.H. Freeman/Spektrum.
Lehmann, L., Keller, L., 2006. The evolution of cooperation and altruism – A general framework and a classification of models. Journal of Evolutionary Biology 19, 1365–1376.
Trivers, R., 1985. Social Evolution. Boston: Benjamin/Cummings.
Wenseleers, T., Ratnieks, F.L., 2006. Enforced altruism in insect societies. Nature 444, 50.
West, S.A., Griffin, A.S., Gardner, A., 2007. Social semantics: Altruism, cooperation, mutualism, strong reciprocity and group selection. Journal of Evolutionary Biology 30, 415–432.
Wilson, D.S., 1990. Weak altruism, strong group selection. Oikos 59, 135–140.
Wilson, E.O., 1975. Sociobiology: The New Synthesis. Cambridge, MA: Belknap Press of Harvard University Press.

Animal Home Ranges☆

Paul R Moorcroft, Harvard University, Cambridge, MA, United States

Glossary

Contact structure A description of the probability of contacts between individuals within a population. Relevant for the spread of infectious diseases.

Cost-benefit models A framework for assessing the performance of an individual exhibiting a particular type of behavior (such as territorial defense of a home range) compared to an alternate pattern of behavior (such as no defense of a home range). The metric of performance varies, but often is measured in terms of energy or resource acquisition.

Negative density dependence A reduction in the per capita rate of population growth caused by increasing population size or density.

Triangulation The process of determining the location of a point by calculating the angle between the point and two or more known locations.

Trophic level The position of an organism within a food chain or food web (e.g., primary producer, primary consumer (herbivore) or secondary consumer (predator)).

Introduction

Fig. 1 shows the spatial extent of relocations of two carnivores, a wolf and a coyote, as a function of time from the first measurement of an individual's location. Initially, their space-use increases rapidly, but as sampling continues, the spatial extent of the relocations saturates. This phenomenon is widespread among mobile animals and reflects the fact that they typically do not move randomly through their environment, but instead restrict their movements to particular areas. In some species, such as many carnivores and birds, this localizing tendency arises from the need to provision offspring located in a den or nest site resulting in these locations acting as a focal points for the movements of adults during periods of breeding—so-called "central place foraging." In other species, such as primates and deer, the existence of a localizing tendency in the movement of individuals is linked to the exploitation of particular resources such as foraging areas or watering holes. Observations such as these underlie the concept of an animal's home range, "the area in which an animal normally lives, exclusive of migration, emigration, or other large infrequent excursions" (Burt, 1943).

Ecological Correlates of Home Range Size

The spatial scales at which individuals exhibit a localizing tendency in their movement behavior can vary widely, even among closely related species. Not surprisingly, in many animal groups, home range size is correlated with body size. The energetic

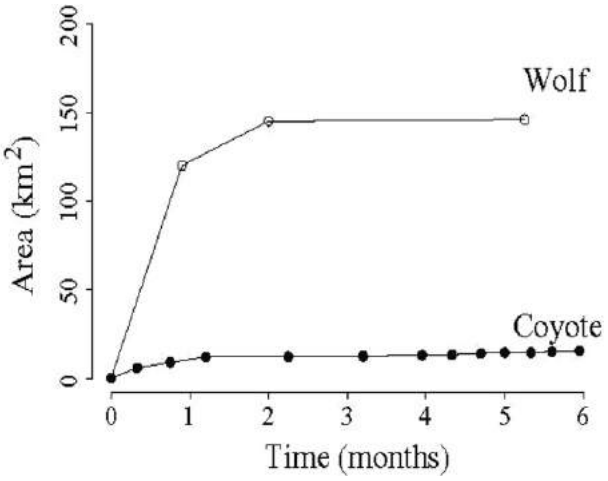


Fig. 1 Increase in the areal extent of wolf and coyote relocations as a function of the length of radio-tracking. Data re-plotted from Messier, F. (1994). Ungulate population models with predation: A case study with the North American moose. *Ecology* **75**(2), 478–488.

☆*Change History:* February 2018. Paul R Moorcroft updated the text and further readings to the entire article.

This is an update of P.R. Moorcroft, Animal Home Ranges, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 174–180.

 doi:10.1016/B978-0-12-409548-9.11095-4

requirements of endotherms generally scale in proportion to $M^{0.75}$, where M is the body mass of a species, and therefore it might be expected that home range size would scale in a similar manner. However, recent analyses have shown that home range size scales as $M^{1.0}$, indicating that home range size increases more rapidly with body size than would be expected from simple metabolic considerations (Fig. 2, solid line). One explanation is that as a result of increased difficulty of defending large home ranges, home range overlap increases with increasing body size, scaling as $M^{-0.25}$. When combined with the scaling of metabolic rate with body size ($M^{-0.75}$), this results in the observed scaling of home range size as $M^{1.0}$. Evidence in support of this explanation comes from the fact that while home range size scales as $M^{1.0}$, per individual area (i.e. the inverse of population density) scales as $M^{0.75}$ (Fig. 2, dashed line). This difference between these two relationships is consistent with the notion of increasing home range overlap in larger-sized animals.

Another important factor influencing animal home range size is its diet. At higher trophic levels, the losses of energy associated with the capture, digestion and utilization of resources markedly reduces the availability of food resources per unit area, and

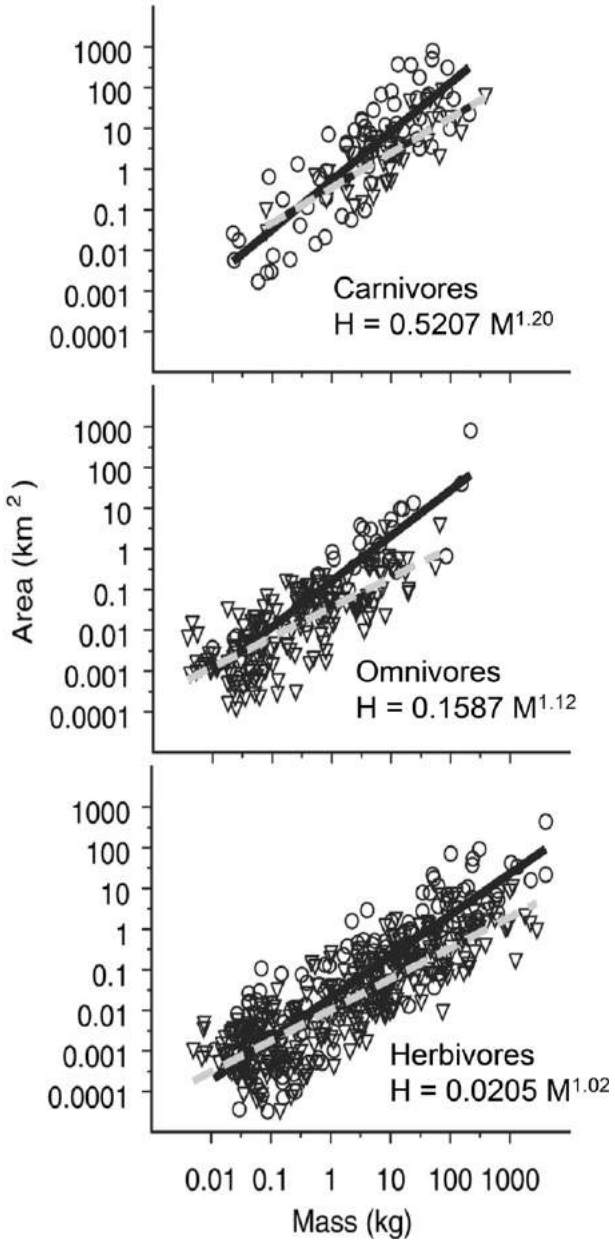


Fig. 2 The scaling of home range size with body mass across carnivores, omnivores and herbivores. In each panel, the *circles* and *solid line* indicate, respectively, the observed and fitted relationship between body size (M in kg) and home range size in (H, km). The coefficients of the relationship are also shown. The relationship between 1/population density (i.e., area per capita) and body mass is also plotted (triangles, and dashed line). In all three cases, the slope of this line is shallower than the slope of the relationship between home range size and body size.

consequently home range is strongly affected by trophic level. As the relationships plotted in Fig. 2 illustrate, the home range sizes of omnivores are approximately 10–15 times higher than those of equivalently-sized herbivores, and the home range sizes of predators are 25–60 times higher than that of equivalently-sized herbivores. What is also apparent in Fig. 2 is that while body size and trophic level account for a significant amounts of variation in animal home range size, there is a ~1000-fold level of variation around the relationships plotted in Fig. 2, emphasizing the fact that the characteristic home range size of a particular species is also significantly influenced by a variety of behavioral and ecological factors other than its body size and trophic level.

Measurement and Analysis of Animal Home Range Patterns

The measurements of home range size shown in Figs. 1 and 2 come from animals fitted with radio telemetry collars, a technology developed in the late 1950s that revolutionized the study of animal movement, enabling routine, systematic measurement of animal locations by triangulating their position on a landscape. Radio telemetry has been successfully used to study the movement behavior of mammals, birds, reptiles, amphibians, fish, and even insects. However, the subsequent advent of global positioning system (GPS)-based telemetry in the 1990s has begun to yielding important new insights into the fine-scale movements of animals, allowing researchers to track animals—in some cases in near real-time—regardless of weather conditions, distance moved, and terrain covered.

The spatial distributions of animal relocations recorded in telemetry studies are translated into estimates of home range size using statistical home range models. A widely used approach is the minimum convex polygon (MCP) method, which characterizes the animal's home range as the smallest sized polygon encompassing the observed relocations (Usually 5%–10% of the outermost relocations are excluded as outliers) (Fig. 3A). A number of density estimation methods have also been developed, in which the animal's home range is characterized using two-dimensional statistical probability density distribution fitted to the observed distribution of relocations (Fig. 3B).

Statistical home range models such as those shown in Fig. 3 provide a useful way to summarize observed spatial patterns of space-use; however, the models are purely descriptive, and thus yield little insight into the underlying causes for an animal's pattern of space-use. Another approach, resource selection analysis, has become a widely-used method for identifying underlying environmental correlates of animal space-use patterns. In contrast to the spatially-explicit nature of statistical home range models, resource selection analysis uses a spatially-implicit approach to identify habitats that are used disproportionately in relation to their occurrence through the examination ratios of habitat utilization relative to a measure of habitat availability. For example, Table 1 shows a resource selection analysis of elk home range relocations in western United States. The measurements in the table indicate that the elk preferentially utilize habitats that have intermediate levels of forest canopy cover. A plausible explanation for this is that the elk utilize habitats that balance their competing needs of having access to open areas that contain forage, and forest cover that provides a degree of protection from predators.

More recently, a new framework for analyzing patterns of animal home ranges has emerged in the form of mechanistic home range models. This involves formulating a mathematical model for the process of individual movement in which the fine-scale movement behavior of individuals is characterized as an underlying stochastic movement process that specifies the probability of an animal situated at a given location moving to a subsequent location in the time between relocations (Fig. 4A). Relevant behavioral and ecological factors influencing the movements of individuals can be incorporated into this description of the fine-scale stochastic movement process. In contrast to resource selection analysis, mechanistic home range models yield a spatially-explicit prediction for patterns of animal space-use that results from the process of individual movement. For example, a recent analysis of coyote home ranges in Yellowstone used a "prey availability plus conspecific avoidance" (PA + CA) mechanistic home range model to account for the observed patterns of coyotes home ranges within the park. In the PA + CA model, individuals exhibit: (i) an avoidance response to encounters with foreign scent marks, (ii) an over-marking response to encounters with foreign scent marks, and (iii) a foraging response to prey availability, in which individuals decreased their mean step length in response to small mammal abundance.

As Fig. 4B shows, the patterns of space-use predicted by the PA + CA mechanistic home range model correctly capture the observed spatial distribution of relocations of five adjacent coyote packs in the study region, implying that the combined influence of resource availability and avoidance responses to neighboring groups is responsible for the observed pattern of coyote space use across the region. A nice feature of the mechanistic approach of "modeling the movement process" is that mechanistic home range models can be used to predict patterns of space-use following perturbation. For example, analysis showed that the PA + CA model shown in Fig. 4B correctly predicted the shifts in patterns of coyote space-use that occurred following the loss of one of the packs in the study area. Mechanistic home range models, similar to one described above, have been used to characterize patterns of space-use in other species including wolves, badgers, and meerkats.

The emergence of a characteristic home range from mechanistic model of individual movement behavior requires that individuals have a centralizing tendency in their fine-scale movements. In the model for coyotes described above, the centralizing tendency arose from the scent-mark mediated avoidance interactions between individuals in neighboring groups, which biased the movements of individuals towards their respective den-sites. This model is appropriate for species with a well-defined focal point for their movement behavior, such as den-sites in many carnivores, burrows in small mammals, and nest sites in birds. However, as noted in the "Introduction" section, other animals, such as deer and many primate species, occupy characteristic home ranges in the absence of a well-defined center of attraction. This implies that other forms of underlying movement behavior must be responsible for the formation of home ranges in these species. This has led to interest

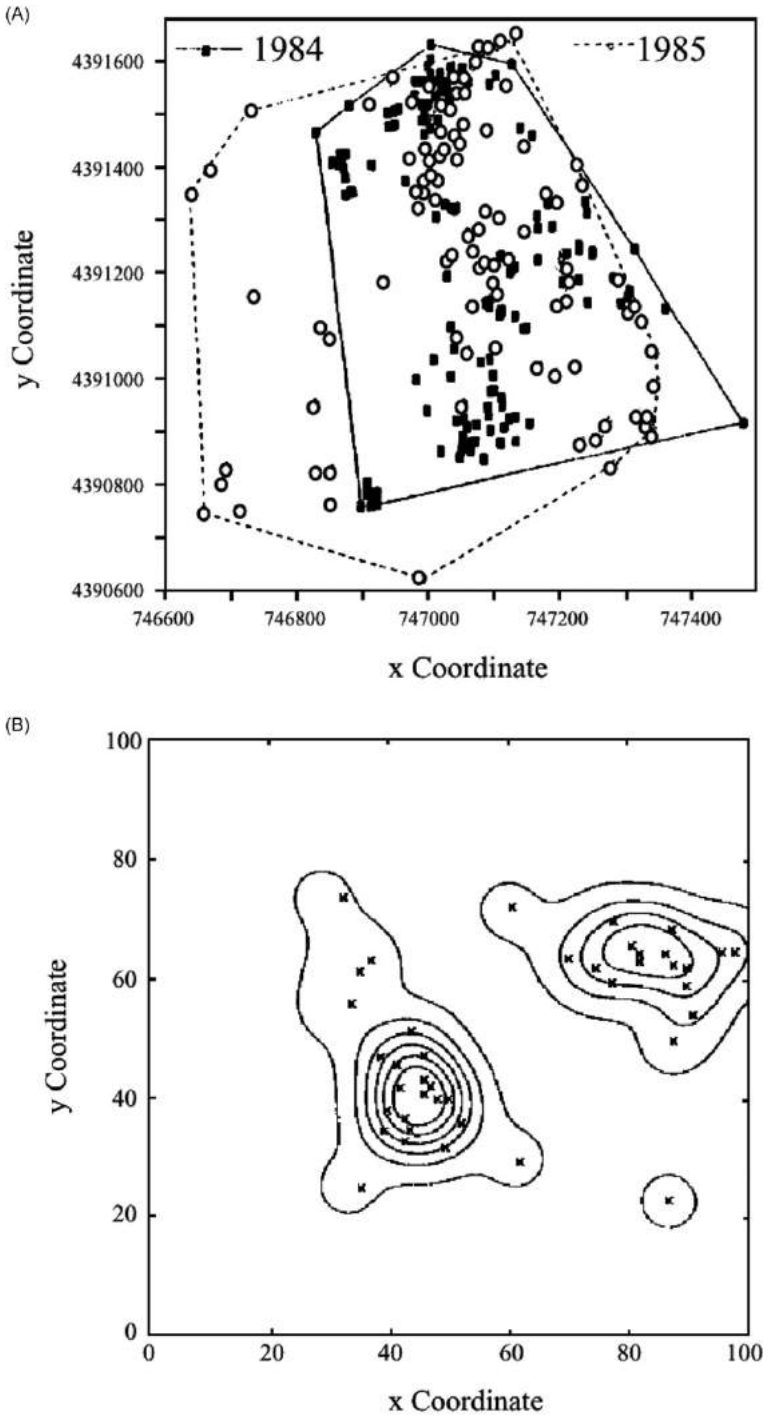


Fig. 3 Examples of statistical home range models. (A) Minimum convex polygon method, and (B) kernel method. Redrawn from White, P. J. and Garrott, R. A. (1997). Factors regulating kit fox populations. *Canadian Journal of Zoology* **75**(12), 1982–1988 (panel A) and Worton, B. J. (1989). Kernel methods for estimating the utilization distribution in home-range studies. *Ecology* **70** (1), 164–168 (panel B).

in the role that memory plays in the formation and maintenance of animal home ranges. The impacts of memory on movement can be described mathematically in terms of so-called "self-attracting random walks"—mechanistic movement models in which an individual's movements are biased towards previously visited locations. Recent work has shown that depending on the details of the underlying formulation, models of this kind can give rise to stable home ranges, or quasi-stable home ranges in which an animal's movements are spatially-localized into a home range, but on a slower timescale the center of the animals' home range drifts around the landscape.

Table 1 Estimated habitat preferences of elk for habitats with different levels of forest canopy cover

Forest canopy cover class	*# of relocations* (U_i)	*Landscape availability* (A_i)	*Expected utilization* $E_i = A_i{*}SU_i$	*Selection ratio* $w_i = U_i/E_i$	*Standardized selection index* $B_i = w_i/Sw_i$
0%	3	0.075	24.4	0.12	0.04
1%–25%	90	0.305	99.1	0.91	0.29
26%–75%	181	0.420	136.5	1.32	0.42
>75%	51	0.200	65.0	0.79	0.25
Total	325	1.000	325	3.14	1.0

Values of the selection ratio >1 indicate habitats that are utilized at a higher frequency than their availability. As the numbers indicate, elk preferentially use habitat with intermediate (26%–75%) canopy cover rather than habitats with either more open or more closed canopy cover.

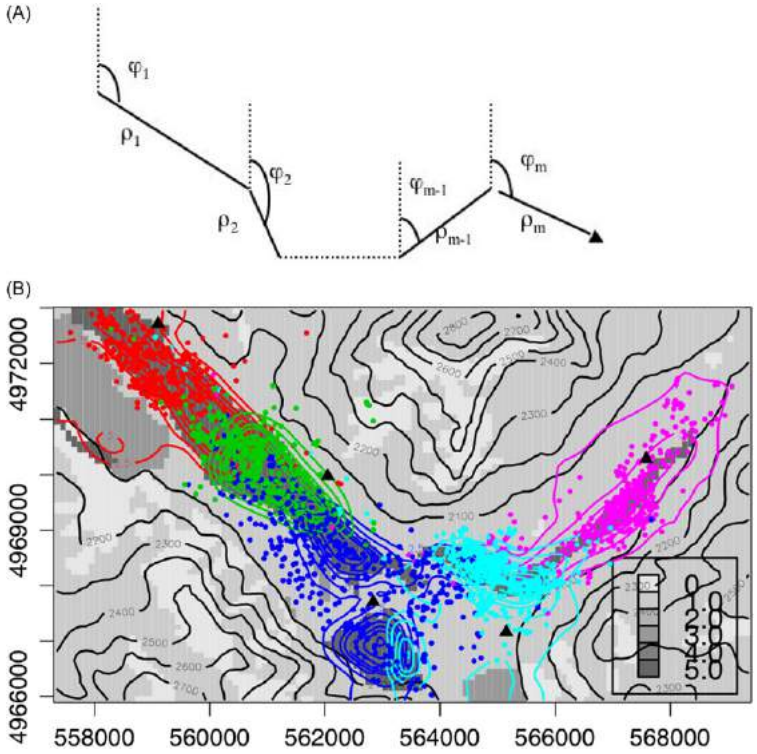


Fig. 4 (A) Schematic illustrating the underlying model of individual movement behavior that underpins a mechanistic home range model. The movement of trajectory of individuals is characterized as a stochastic movement process, defined in terms of sequences of movements between successive relocations ($i = 1...m$) of distance ρ_i and directions φ_i drawn from statistical distributions of these quantities that are influenced by relevant factors affecting the movement behavior of individuals. (B) Colored contour lines showing fit of a mechanistic home range model to relocations (filled *circles*) obtained from five adjacent coyote packs in Lamar Valley Yellowstone National Park. As described in the text, the PA + CA mechanistic home range model used in this study incorporates a foraging response to small mammal prey availability plus a conspecific avoidance response to the scent-marks of individuals in neighboring packs. The home range centers for each of the packs are also shown (▲), and the grayscale background indicates small mammal prey density (kg ha^{-1}) in the different habitat types.

The Functional Significance of Animal Home Range Patterns

Since an animal's home range is intimately linked to its utilization of resources (food, shelter, mates, etc.), and some of these will be in limited supply, competition often favors a spacing-out of animal home ranges across landscapes. This can arise from overt territorial aggression between individuals, such as the contests between male dragonflies that occur over ponds during their breeding season, and the contests that occur between red squirrels occupying neighboring home ranges. In other cases, however, spacing out arises from individuals passively avoiding other individuals, either via direct detection and avoidance of individuals, or, more commonly, in response to auditory and olfactory signals that indicate occupation, such as singing in birds and scent marking in mammals.

A central question in the study of animal home ranges has been to ask under what conditions should animals engage in territorial defense of their home range? A logical argument is that evolution should favor territorial defense of home ranges whenever relevant limiting resources for individuals are "economically defendable." In other words, animals should engage in

territorial defense whenever this results in increased fitness relative to alternate strategy, such as passive avoidance, or ignoring the presence of other individuals.

Key factors influencing the defendability of a given resource are: its distribution in space, its predictability, and the number of competitors for the resource. Sparsely distributed, low quality resources, such as the spatial distribution of forbs and grasses grazed by elk in Yellowstone National Park, are difficult to defend energetically, and thus do not tend to favor territorial defense. In contrast, resources that are clumped into patches in which resource levels are high tends to favor territorial defense. Examples of this phenomenon include: the clumped distribution of tree fruits within a tropical forest that favors territorial defense by groups of primates; and the clustered distribution of female southern elephant seals, which arises when the females haul up onto beaches to breed, that favors territorial defense of the females by male elephant seals. Resource predictability is also important. For example, pied wagtails defend territories along river banks where winter food resources are predictable, but in areas where food resources occur in transient, unpredictable patches, they abandon territorial defense and move around in flocks. The effects of competitor density on resource defendability are more difficult to characterize since intruder pressure is often correlated with food availability. One of the best examples of an effect of competitor density comes from a study that used a statistical approach to separate out the effects of food availability and intruder density on home range defense of sanderlings foraging along the California coast. The analysis showed that in areas with low resource levels, where the density of competitors was low, individuals defended home ranges, while in resource rich areas where competitor density was high, individuals abandoned territorial defense.

Simple cost-benefit models have been widely used to predict how changes in the environment should affect the size of home range that an animal defends (**Fig. 5**). In general, we might expect that the fitness payoff arising from occupying and defending home range will initially increase rapidly as home range size increases, but then saturate as its size increases further (curve B). The costs of defending the home range (curve C) will also tend to increase with increasing size. In the example shown, defense costs increase in an accelerating manner, as would occur for example, if the difficulty of detecting and evicting intruders increases markedly as a home range gets larger.

The optimal home range size is one that maximizes the difference between the two curves. An increase in resource richness causes the benefit curve (B) to steepen, shifting its position from B to B′, which causes the optimal territory size to decrease as resource richness increases (**Fig. 5**A). In a similar manner, increased costs of defense cause the cost curve (C) to steepen, shifting its position from C to C′, which, also causes a reduction in the predicted optimal territory size (**Fig. 5**B). Evidence in support of these kinds of responses has come from a number of studies. For example, blue tits in England have been shown to decrease their territory size in response to increased food resources supplied by artificial provisioning, and the breeding territories of male great tits have been shown to be inversely related to their population density during the previous year.

Note however, that the apparent simplicity of cost-benefit analyses such as the one depicted in **Fig. 5** belies a great degree of complexity in predicting how different ecological factors affect the size of an animal's home range. Theoretical analyses have shown that the predictions of cost-benefit models can vary depending on the proximal measure of fitness used, the precise shape of the benefit and cost curves considered, and the constraints on performance incorporated into the analysis. These differences can lead to entirely opposing predictions for the consequences of changes in resource availability and number of competitors on home range size. Additional complications arise because in reality changes in the cost and benefit curves depicted in **Fig. 5** are often not independent. For example, as noted above in the case mentioned earlier of sanderlings foraging along the Californian coast, increases in resource density also strongly affect home range defense costs due to increases intruder density in resource rich areas. Thus, like other aspects of animal home range behavior, analyzing the functional significance of differences in territory size is inherently challenging because the fitness of an individual exhibiting a particular behavior depends strongly on the behavior of other competing individuals present on the landscape (i.e., it is an *n*-person game), leading to complex responses to changes in ecological conditions.

Home Range Size and Population Social Structure

When faced with shifts in resource levels or intruder pressure, individuals that are occupying a home range can respond in ways other than adjusting their home range size. One alternative is to accept one or more additional individuals into their home range, and then cooperatively defend it against competing individuals. This kind of response is seen in the pied wagtails, which, as noted earlier, defend territories along river banks in England during the wintertime. At moderate resource levels, individuals occupy and defend individual home ranges. However, as resource levels increase, rather than decreasing their range size as might be expected from a simple cost-benefit model of territory size such as the one depicted in **Fig. 5**, owners begin to share their territory with an additional individual who shares the home range, but forages independently of the owner.

A simple model for evaluating the economic profitability of sharing a home range considers the costs individuals incur as result of reduced food availability versus the benefits that arise from reduced per capita defense expenditures when sharing a home range with one or more additional individuals. A key factor influencing the costs of sharing on the food intake of individuals is the dynamics of resource renewal (i.e., how quickly resource levels increase after having been depleted). A resource that is shared between individuals will be exploited at a higher rate than one exploited by a single individual. If the resource renews slowly, then the increase in the exploitation rate will have a significant impact on the foraging efficiency of the individuals. In contrast, if the

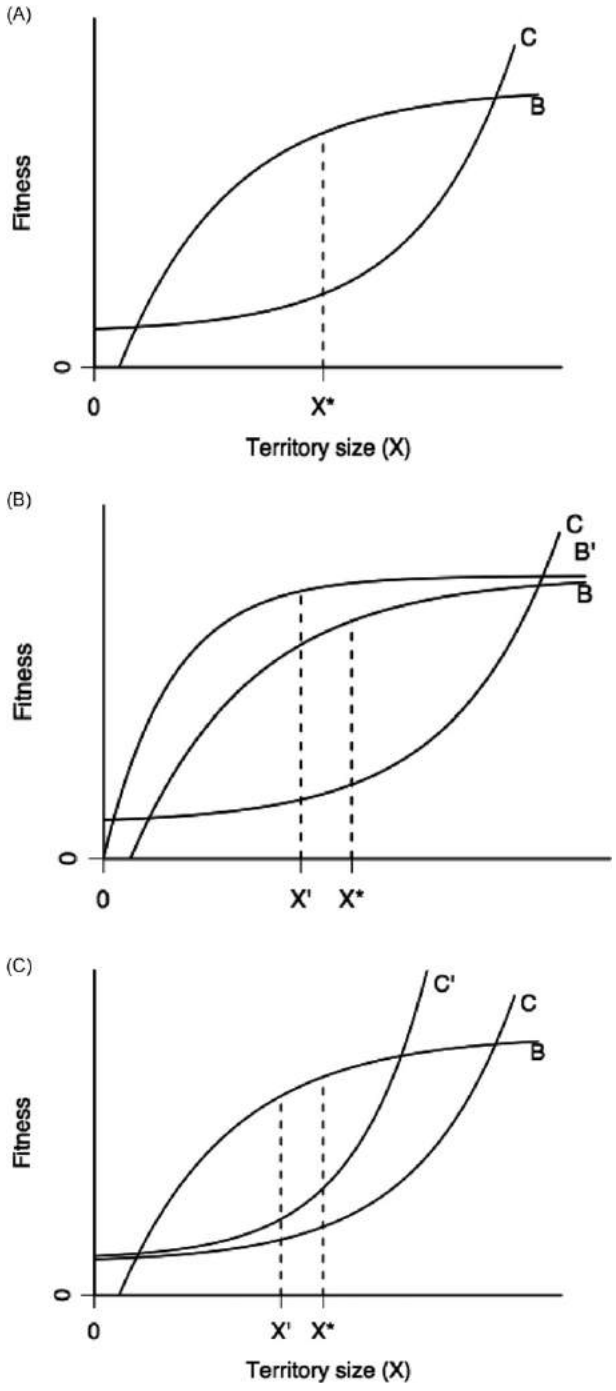


Fig. 5 A simple graphical cost-benefit model of territory size. (A) The optimal size occurs at X* where the difference between the fitness benefit (B) and the fitness cost (C) is maximized. (B) Increased resource availability shifts the benefit curve from B to B′, causing a decrease in the optimum territory size. (C) Increasing defense costs shifts the cost curve C to C′, also causing a decrease in the optimum territory size.

resource renews quickly, then the increased exploitation rate may have a relatively small effect on the foraging efficiency and resulting per capita food intake of individuals.

A second key factor affecting the costs of sharing a home range is the spatial distribution of resources. The "resource dispersion hypothesis" proposes that in situations where a species is foraging on a patchily-distributed, ephemeral resource in which each patch, when available, can supply enough resources to meet the needs of more than one individual, an economically-defendable home range that is large enough to sustain an individual or a mated pair of individuals is likely to support the foraging needs of one or more additional individuals. Two animals whose home range patterns appear to fit this reasoning are red foxes and badgers in England. Due to the patchy, ephemeral nature of earthworm abundance – a key prey resource for both species – mated-pairs of

badgers or mated pairs of foxes defending a home range large enough to satisfy their own energetic requirements incur relatively little cost from resource consumption of additional individuals and benefit from the cooperative defense of a shared home range. Similar arguments have been made for other carnivores, and for species in other animal groups, including primates, ungulates and birds.

A distinguishing feature of the resource dispersion hypothesis is that it argues that it is the spatial and temporal dynamics of the underlying resources, rather than the benefits of cooperative hunting or cooperative defense against predators, that favor the occupation and defense of shared home-ranges by groups of individuals. Additional benefits of group living, such as improved hunting success, improved predator detection and defense, and the inclusive fitness benefits when sharing a home range with relatives, will, however, increase the benefits and reduce the costs of group living. Distinguishing predictions of the hypothesis are that: (i) group size should more strongly correlate with patch richness and heterogeneity rather than territory size; and (ii), that territory size is primarily determined by the patterns of resource dispersion. These predictions have been borne out in several studies of fox and badger populations, as well in studies of other species, including spotted hyenas, mara, and magpie jays. Thus, the spatio-temporal distribution of resources, in conjunction with other benefits of group living, such as shared defense costs, improved foraging success, or predator defense, can favor the occupation and defense of shared home ranges by groups of individuals. An important implication of this is that the size, shape and degree of exclusivity of home ranges is a key determinant of the different social organization and mating systems found within species.

Effects on Population Demography

Since the spatial extent and degree of exclusivity of an animal's home range influence its access to important resources – such as food, shelter, and mates – patterns of animal space use can also exert a powerful influence on the demography of animal populations. This is particularly the case in species where individuals are territorial and actively defend home ranges against other individuals for all or part of the year as occurs in many birds, such as great tits and red grouse; in many vole and other rodent species; and in carnivores, such as coyotes and wolves. In such populations, the active defense of home ranges results in significant numbers of individuals being forced into either dispersing and setting up home ranges in marginal habitats, or existing as non-resident "floaters" within the population. Usually juveniles or lower-ranking adults, these non-resident individuals tend to have diminished rates of survival, and have either reduced fecundity or do not breed at all.

It has been argued that in species with cyclical population dynamics, such as red grouse and a number of vole and other rodent species, home range defense acts as a destabilizing factor on population abundance, due to delayed negative density dependence between the response of individual home range sizes and levels of defense to changes in population abundance. More commonly however, the existence of a reservoir of "surplus" non-resident individuals arising from home range defense is considered to act as a stabilizing factor on population size, reducing the propensity for fluctuations and increasing a population's resilience to perturbation. This has implications for efforts for management of populations in which it occurs: for example, research has shown that in coyotes and badgers non-resident individuals rapidly replace breeding individuals killed in control efforts, severely hampering efforts to reduce their population sizes through culling.

Secondary Ecological Interactions

Evidence from field studies indicates that the spatial pattern of home ranges within a species can affect the spatial distribution of prey and competitors. For example, in northeastern Minnesota, white-tailed deer are found primarily in the "buffer zones"—areas of low space use that occur between adjacent wolf pack home ranges. This negative correlation between the spatial distribution of wolves and deer arises as a result of differential predation rates between the interior of wolf home ranges and the buffer zones that separate them.

For the same reasons that within-species (intraspecific) competition for resources often favors defense of a home range area against utilization by individuals within a population, between-species (interspecific) competition can favor individuals defending their home range against utilization by individuals of competitor species. This phenomenon occurs in the carnivore community of Yellowstone National Park. Prior to wolf re-introduction, packs of coyotes, usually 4–6 adults, occupied and defended contiguous home ranges across the landscape. However, following the wolf re-introduction, coyotes have radically altered their patterns of space-use: the packs have broken up and individuals now move around individually or in pairs, restricting their movements in space and in time to areas where wolves are not present. These kinds of interactions between the spatial distributions of home ranges in co-occurring competing species can have important consequences for animal conservation. For example, in Africa, efforts to conserve the endangered African wild dog have been complicated by their competitive interactions with lions and hyenas, which have prevented them from setting up home ranges in favorable habitats, reducing their survival and breeding success.

The spatial distribution of animal home ranges also has implications for the incidence and spread of infectious diseases within populations. This because the rate at which a disease spreads through a population is strongly influenced by the probability that different individuals encounter one another, a property known as the contact structure of the population. For example, a study of Channel Island foxes living on islands off the coast of California found that both the number of contacts per day and the duration of contacts between individuals were positively correlated with their degree of home range overlap. Such effects of animal home

ranges on the contact structure within animal populations can have important implications for attempts to control disease outbreaks. For example, studies examining the impacts of culling badger populations upon the transmission of Tuberculosis (TB) from badgers to cattle in the United Kingdom found that, while culling reduced the incidence of TB in cattle in the areas where badger population sizes were reduced, it increased its incidence in neighboring areas due to shifts in badger home range patterns caused by the culling that elevated the rate of contacts between badgers and cattle.

See also: Behavioral Ecology: Habitat Selection and Habitat Suitability Preferences. General Ecology: Tolerance Range; Demography

Further Reading

Krebs, J.R., Davies, N.B., 1993. An introduction to Behavioural ecology, 3rd edn. Oxford: Blackwell Publishing.
Manly, B., McDonald, L., Thomas, D., 1993. Resource selection by animals: Statistical design and analysis for field studies. New York: Chapman and Hall.
Millspaugh, J.J., Marzluff, J.M., 2001. Radio tracking and animal populations. San Diego: Academic Press.
Moorcroft, P.R., Lewis, M.A., 2006. Mechanistic home range analysis. Princeton University Press: Princeton.
Potts, J.R., Lewis, M.A., 2014. How do animal territories form and change? Lessons from 20 years of mechanistic modelling. Proceedings of the Royal Society B 281.20140231. https://doi.org/10.1098/rspb.2014.0231.
Woodroffe, R., Donnelly, C.A., Cox, D.R., Bourne, F.J., Cheeseman, C.L., Delahay, R.J., Gettinby, G., McInerney, J.P., Morrison, W.I., 2006. Effects of culling on badger *Meles meles* spatial organization: Implications for the control of bovine tuberculosis. Journal of Applied Ecology 43, 1–10.

Anti-Predation Behavior

Lee A Dugatkin, University of Louisville, Louisville, KY, United States

Introduction

The biology behind antipredator behavior is, in a sense, unique, in that mistakes with respect to predators can lead to an animal having a future fitness of zero. As such, natural selection should operate very strongly on antipredator behavior. And it does: the array of antipredator adaptations in nature is dazzling. Consider, for example, antipredator behaviors in schooling species of fish. In addition to the potential hydrodynamic and foraging benefits accrued by living in groups, fish in schools display a wide array of antipredator tactics. When a predator is sighted, schooling fish school more tightly, allowing for the following antipredator tactics:

1. Fountain effect—schools maximize their speed, split around a stalking predator, and then reassemble behind the putative danger.
2. Trafalgar effect—when animals school tightly, information about a predator spreads from individual to individual more quickly than in loose schools. Because this resembles the quick transfer of information of battle signals in Lord Nelson's fleet at the battle of Trafalgar, it has been dubbed the Trafalgar Effect.
3. Flash Explosion—schools of fish "explode," with individuals swimming off in all directions, confusing predators and allowing individuals to escape. Schooling fish often add to this effect by moving around in erratic, unpredictable patterns.
4. Predator Inspection—a small number of individuals break away from a school and approach a predator to gain various types of information (e.g., is the predator hunting?), and then return to the school, where this information may spread across individuals.

Group Size and Antipredator Behaviors

One of the most fundamental antipredator strategies is to live in large groups, as group life confers a suite of potential antipredator benefits. The most basic benefit of living in large groups results from a simple statistical property. If a predator is to going to strike at one member of a group of size N, then the odds that any particular individual will be its victim is 1/N, which increases with group size. The larger the group size, the safer is each individual in the group (the situation is more complicated if predators prefer to attack larger groups). Individuals obtain benefits with respect to antipredator behaviors because as group size increases, there are more and more individuals vigilant for predators, making all group members safer: the "many eyes" benefit of group life. One benefit of "many eyes" is that each group member can spend less time being vigilant, and more time doing other activities, since vigilance can be parceled across more individuals as group size increases.

Large groups also provide positional benefits because individuals in the center of the group are often safer than those on the periphery (though again, the situation is complicated, as this will often lead to a scramble among prey to get positioned at the center of their group).

Living in groups provides other benefits with respect to safety. Groups can often respond to potential predation threats in coordinated ways that solitary individuals cannot, and the activity levels in large groups may overload a predator's sensory input mechanisms, making a successful strike less likely.

Behavioral Tradeoffs Associated with Predation

There are opportunity costs associated with antipredator behavior. Instead of avoiding predators, individuals may be foraging, mating, resting, and so forth. To examine such trade-offs, consider the case of antipredator behavior and foraging. Predation pressure affects virtually every aspect of foraging, from when a forager begins feeding to when it resumes feeding after an interruption, to where it feeds, what it eats, and how it handles its prey.

Work on predation and foraging in the gray squirrel (*Sciurus carolinensis*) has demonstrated that squirrels alter their foraging choices as a result of predation pressure from redtailed hawks (*Buteo jamaicensis*). Squirrels who could either eat their food items where they found such items or carry the food to cover were more likely to carry items to an area of safe cover when predation threat was significant—the closer the refuge from predation, the more likely they would use such a shelter when foraging.

Alarm Calls

Alarm calls are one of the more dramatic forms of antipredator behavior. Belding's ground squirrels give alarm calls when a terrestrial predator is sighted, with females being much more likely to emit such calls than males. The reason for this sex difference

in alarm calling is tied to the demography and genetics of Belding's ground squirrels. Male Belding's ground squirrels emigrate to new populations when they mature, but female squirrels spend their entire lives in their population of birth. This difference in dispersal creates an asymmetry in the way that adult males and females are related to others living in their populations. By remaining in the populations in which they were born, females—both young and old—are surrounded by genetic relatives. Mature males, who emigrate to new populations, however, find themselves interacting with fewer genetic relatives than do adult females. By giving alarm calls, females get indirect benefits, in that they help protect their blood kin—males receive no such benefits, and call at much reduced rates.

Alarm calls need not be vocal. In ungulates, individuals are known to "flag" their tails after a predator has been sighted. Such flagging occurs as part of a sequence of antipredator behaviors, and often involves an individual lifting its tail and "flashing" a conspicuous white rump patch. Flagging often, but not always, occurs when a predator is at a relatively safe distance from its potential prey. Flagging may: (1) warn conspecifics (kin and nonkin) of potential dangers, (2) "close ranks" and tighten group cohesion, (3) announce to the predator that it has been sighted and should therefore abandon any attack, (4) entice the predator to attack from a distance that is likely to result in an aborted attempt, and (5) cause other group members to flee, thereby confusing the predator, and making the flagger itself less likely to be the victim of an attack.

Alarm calls can be very complex. Vervet monkeys, for example, make very distinct and different calls in response to leopard, snake, and eagle predators. Vervets run to the trees when they hear a leopard alarm call, but they hide in bushes when a fellow vervet utters an eagle alarm call, suggesting that these monkeys are using alarm calls to indicate the mode of predator attack.

Interpopulational Differences in Antipredator Behavior

One way to examine how natural selection has operated on antipredator behavior is to compare the behavior of individuals that live in populations under different predation pressures. If antipredator tactics differ across these populations, it suggests that natural selection has been operating on these behaviors.

The population comparison method for studying selection and antipredator behavior has been employed in many animal systems including ground squirrels, guppies, sticklebacks and minnows. For example, researchers have examined antipredator behavior in two different populations of minnows (*Phoxinus phoxinus*). Minnows from the Dorset area of southern England and the Gwynedd area of northern Wales were chosen, because the Dorset minnow population is under strong predation pressure from pike predators, while pike are absent from the Gwynedd population of minnows. While individual fish from southern England and northern Wales both look like minnows, and are generally the same size, their antipredator repertoires are different.

In the laboratory, before exposure to a predator, Dorset minnows (those from high pike predation areas) swim around in larger groups than Gwynedd minnows. Dorset minnows also seemed to have more stable groups, with less movement of individuals from group to group than the Gwynedd fish. Once a predator was experimentally introduced to these laboratory populations, individuals in both minnow populations dramatically increased their group size but it took the Gwynedd minnows significantly longer to adjust their group size back to normal: not only did the high-predation Dorset minnows have generally stronger antipredator responses, but they were also quicker to respond to the removal of danger by resuming normal, nonpredator-based activities. In addition, when a predator was present, Gwynedd minnows ceased eating, while Dorset minnows, who are accustomed to foraging in the face of danger, curtailed their foraging activity, but not nearly to the extent of Gwynedd fish.

Interpopulation differences in antipredator behavior have also been studied in guppies native to the Northern Mountains of Trinidad and Tobago. In many of these streams, guppies can be found both upstream and downstream of a series of waterfalls. These waterfalls, however, act as a barrier to many of the guppy's predators. Upstream of such waterfalls, guppies are typically under only slight predation pressure from larger species of fish; while downstream populations of guppies are often under strong predation pressure from numerous piscine predators.

Guppies from high-predation sites mature faster, produce more broods of (smaller) offspring, and tend to channel their resources to reproduction when compared to guppies from low predation cites. These all appear to be antipredator adaptations. At high predation sites, guppy predators tend to be large and can eat a guppy no matter how large it gets. At these sites, producing many smaller fish should be favored by natural selection, as this is akin to buying lots of lottery tickets and hoping that one is a winner.

At low predation sites, only a single small fish predator (*Rivulus hartii*) of guppies exists. If guppies can get past a certain size threshold, they are safe from *R. hartii*. As such, natural selection favors females producing fewer, but larger offspring, who can quickly grow large enough to be out of the zone of the danger associated with *R. hartii*, and this is what researchers have found. Indeed, transplant experiments demonstrate that when low predation fish are transferred to high predation sites, natural selection quickly acts, and after only a handful of generations, descendants of the transplanted fish have converged on the traits associated with living in high predation. Reciprocal experiments have found the same result when high predation fish are transferred to low predation sites.

The interpopulation, experimental, approach to studying antiprdator behavior can be also be used in the wild. A nice example of this comes from work in the Bahamas archipelago, which has thousands of small rock islands, often just a few hundred square meters in size. Lizards abound on many of these rock islands. One of the prey items of the curly-tailed lizard (*Leiocephalus carinatus)* is the smaller brown anole lizard (*Anolis sagrei)*. Although many of the rock islands of similar size and vegetation are home to *A. sagrei*, only a subset of those also have curly-tailed lizards. This sets the stage for experimentally introducing predators

to a series of rock islands where they are otherwise absent, and examining both the short-term and long-term affects of the introduction on prey behavior and prey morphology.

Researchers added Curly-tailed lizard predators to five rock islands where they were absent, but brown anoles were present. As controls, they observed brown anoles on islands of similar size and vegetation, but which had no curly-tailed lizard predators added. Brown anoles responded behaviorally to the addition of predators in a quick and dramatic fashion. On the control islands, brown anoles typically were found in vegetation about 10 cm above the ground. On islands to which predators had been added, brown anoles moved to higher, thinner branches, presumably to avoid the new danger present. This behavioral response occurred within a year of the experimental manipulation.

These results allowed the evolutionary ecologists to examine a long-standing question regarding evolution and behavior: if a behavioral change occurs as a result of predatory pressures, will natural selection be weaker or absent on any additional morphological traits that might be important with respect to predators? In the case of the rock islands, does the behavioral shift to higher perches mean that selection will be weak or absent on morphology in prey, since the prey have moved to safer areas? What researchers found was strong evidence for natural selection acting on morphology, even after behavioral changes. Selection on islands with experimentally introduced predators produced males with longer limbs who were faster at escaping from predators, and larger females, who were both faster at escaping predators and harder for predators to consume should they be caught.

Social Learning and Antipredator Behavior

Motmots, a group of tropical bird species, *instinctively* fear poisonous coral snakes. The coral snakes that are dangerous to motmots have a specific color pattern—red and yellow bands. When baby motmot chicks are presented with a wooden dowel with red and yellow bands painted on it, the chicks instantly fear it. However, if green and blue bands or even red and yellow stripes—neither of which resemble snakes dangerous to the motmot—are painted on a dowel, motmot young no longer treat it like a danger.

The motmot solution to knowing who the enemy is works well under certain conditions, namely when the predatory species involved are few and constant through evolutionary time. If, however, there are lots of predators to handle and/or if the kinds of predators are constantly changing, innate fears may be an inadequate or inappropriate solution to the "know your enemy" problem. Under such conditions, selection may favor learning who the enemy is by observing how others respond to potential threats. This type of antipredator learning has been documented in blackbirds.

Once a flock of blackbirds spot a predator, some of them join together, fly toward the danger, and aggressively attempt to chase it away. Such mobbing behavior often works well enough to force predators to leave the blackbirds' area. Another function of this mobbing behavior may be to help predator-naive blackbirds identify what constitutes a predator. Experiments indicate that when young blackbirds see a particular species being mobbed, they learn that this species is in fact a predator.

Predation and Hatching Time

Antipredator behavior has even been documented in embryos. Work on red-eyed treefrogs (*Agalchnis callidryas*) and their predators demonstrates how natural selection can produce behaviors on the part of embryos that reduce their risk of predation. Red-eyed treefrogs attach their eggs to the various types of vegetation that hang over water, and once tadpoles hatch, they immediately drop down into their aquatic habitat. Both the terrestrial habitat of the egg and the aquatic habitat of the tadpole have dangerous, but different, predators that feed on treefrogs. If terrestrial predation from snakes and wasps is weak, embryos hatch late in the season. Such late hatching allows the frogs to grow to a size that lowers the levels of fish predation, once they hatch and fall into the water.

Both snakes and wasps are terrestrial predators on treefrog eggs, with the latter taking one egg at a time, but the former capable of much more damage per attack. When predation from snakes and wasps is high, it often pays to mature early and drop into the water, away from heavy terrestrial predation. Eggs in clutches that are not disturbed by predators often hatched at about 6 days. When comparing eggs from these undisturbed clutches to clutches that have already suffered some predation by wasps, hatching rates are significantly different. Eggs hatched at a much quicker rate when their clutch had been the victim of some wasp predation, with most eggs from attacked clutches hatching at 4 or 5 days (as opposed to six).

See also: Ecological Processes: Predation and Its Effects on Individuals: From Individual to Species. General Ecology: Communication; Dominance

Further Reading

Caro, T.M., 1994. Ungulate antipredator behavior: preliminary and comparative data from African bovids. Behavior 128, 189–228.

Chase, J.M., Abrams, P.A., Grover, J.P., Diehl, S., Chesson, P., Holt, R.D., Richards, S.A., Nisbet, R.M., Case, T.J., 2002. The interaction between predation and competition: a review and synthesis. Ecology Letters 5, 302–315.

Elgar, M.A., 1989. Predator vigilance and group size in mammals and birds: a critical review of the empirical evidence. Biological Reviews of the Cambridge Philosophical Society 64, 13–33.
Groenewoud, F., Frommen, J.G., Josi, D., Tanaka, H., Jungwirth, A., Taborsky, M., 2016. Predation risk drives social complexity in cooperative breeders. Proceedings of the National Academy of Sciences of the United States of America 113, 4104–4109.
Kruuk, H., 1972. The spotted hyena: a study of predation and social behavior. Chicago: University of Chicago Press.
Lima, S.L., 1998. Stress and decision making under the risk of predation: recent developments from behavioral, reproductive, and ecological perspectives. Advances in the Study of Behavior 27, 215–290.
Losos, J.B., Schoener, T.W., Spiller, D.A., 2004. Predator-induced behavior shifts and natural selection in field-experimental lizard populations. Nature 432, 505–508.
Magurran, A.E., Seghers, B.H., Shaw, P.W., Carvalho, G.R., 1995. The behavioral diversity and evolution of guppy, *Poecilia reticulata*, populations in Trinidad. Advances in the Study of Behavior 24, 155–202.
Pulliam, R., 1973. On the advantages of flocking. Journal of Theoretical Biology 38, 419–422.
Reznick, D.A., Bryga, H., Endler, J.A., 1990. Experimentally induced life-history evolution in a natural population. Nature 346, 357–359.
Sherman, P.W., 1977. Nepotism and the evolution of alarm calls. Science 197, 1246–1253.

Biological Rhythms

R Refinetti, University of South Carolina, Walterboro, SC, USA

Introduction

Rhythmicity is found in a multitude of biotic and abiotic factors. The examples of environmental cycles shown in **Table 1** document the wide range of frequencies of cycling abiotic factors. Some environmental cycles have well-known effects on organisms – such as the rhythmic change in foraging behavior of intertidal organisms caused by the ebb and flow of the tides. Other cycles, such as the change in the path of the Earth's revolution around the Sun (the orbital eccentricity rhythm), can potentially affect organisms by causing slow weather changes over thousands of years. A few well-known biological rhythms – some of which are and some of which are not controlled by environmental cycles – are listed in **Table 2**. Biological rhythms involve repetitive processes ranging in frequency from more than once per second, such as the spontaneous firing rate of neurons in the mammalian central nervous system, to less than once every decade, such as the oscillation in the wild population of the Canadian lynx (*Lynx canadensis*).

Most research on biological rhythms conducted during the past 50 years has dealt with daily/circadian and annual/circannual rhythms and has concentrated on basic behavioral, physiological, neural, and molecular processes studied in a small number of species in the laboratory. Few studies have examined interspecies differences in basic processes, and fewer yet have addressed ecological issues in natural environments. Thus, this article will start with a brief review of what has been learned in the laboratory about the basic processes of biological rhythmicity and will then examine what little we know about the behavioral ecology of biological rhythms.

Basic Processes

Environment, Endogenesis, Entrainment

The matching of the cycle duration of some environmental cycles with that of some biological rhythms might suggest that biological rhythms are mere responses of organisms to the rhythmicity of their abiotic environment. Extensive research has clearly established, however, that the relationship between environmental cycles and biological rhythms is much more complex. In some cases, biological rhythms are indeed mere responses to environmental cycles. An example is the tidal rhythm of burrowing in some (but not all) species of crabs. In other cases, biological rhythmicity is endogenously generated and does not respond to environmental cycles. An example is heart beating: the rhythmic activity is endogenously generated by the cardiac pacemaker, which can be modulated by sympathetic and parasympathetic stimulation but is not synchronized by natural environmental cycles. Finally, in some cases, biological rhythmicity is endogenously generated and this rhythmicity is modulated (entrained) by environmental cycles. The best-known example is the entrainment of circadian rhythms by the regular alternation of day and night (i.e., the alternation of light and darkness).

In strict sense, only rhythms that are endogenously generated and that can be entrained should receive the circa designation. Laboratory research has documented only four classes of circa rhythms: circatidal, circadian, circalunar, and circannual rhythms. If a given variable does not exhibit circa rhythmicity, or has not been proved to exhibit it, it should be named without the circa designation. Tidal, lunar, and annual are commonly used descriptors. Dian is never used; instead, daily is recommended – although diel and nycthemeral are also used.

It should be pointed out that observation of rhythmic behavior in a natural environment is necessarily insufficient to characterize the nature of the rhythmicity. Whenever a new rhythmic pattern is observed, controlled laboratory investigation is necessary to attribute the rhythmicity to environment, endogenesis, or entrainment. Accurate identification of the modality of rhythmicity is essential for the understanding of ecological and evolutionary significance of the observed rhythmic process.

Rhythmic Variables

Different types of biological rhythmicity may affect the regulation of one or more physiological or behavioral variables. Estrous rhythmicity in rodents, for example, has been shown to affect at least hormonal secretion, behavioral sexual receptivity, the pattern of vaginal discharges, and the amount and temporal organization of locomotor activity. Circadian rhythmicity has been shown to affect locomotor activity, eating and drinking, excretion, learning capability, heart rate, blood pressure, body temperature, hormone secretion, sexual activity, parturition, suicide, susceptibility to heart attack, and many other variables. It is still unclear which of these multiple rhythmic variables are controlled directly by the circadian pacemaker and which are simply caused (masked) by the rhythmicity of variables controlled by the pacemaker. It has been demonstrated that the circadian rhythm of body temperature

Table 1 Some environmental cycles on the Earth

Duration of cycle	Phenomenon
2×10^{-15} s	Oscillation of electromagnetic waves in visible light
2×10^{-2} s	Voltage oscillation in alternated current (home electricity)
12.4 h	Tides (attractive forces of the Sun and the Moon)
24 h	Days (Earth's rotation)
30 days	Months (Moon's revolution around the Earth)
365 days	Years (Earth's revolution around the Sun)
10 years	Cycle of sunspots
22 000 years	Precession of the equinoxes
41 000 years	Variation in Earth's obliquity (axial tilt)
96 000 years	Variation in Earth's orbital eccentricity

Table 2 Some biological rhythms

Duration of cycle	Phenomenon
10^{-1} s	Spontaneous firing rate of cortical neuron
1 s	Human heart rate
1 h	Pulsatile secretion of hormones
12.4 h	Tidal and circatidal rhythms
24 h	Daily and circadian rhythms
4 days	Estrous cycle of rat
7 days	Human work-rest week
30 days	Lunar and circalunar rhythms
110 days	Estrous cycle of elephant
12 months	Annual and circannual rhythms
10 years	Oscillation in the wild population of Canadian lynx

is not caused by the rhythm of locomotor activity, whereas the rhythm of urea secretion is simply a consequence of the rhythm of food ingestion, but generally very little is known about the inter-relationships among the various rhythms.

Entraining Agents

The circa rhythms can, by definition, be entrained by environmental cycles. Circatidal rhythms are often entrained by the cycle of inundation, whereas circannual rhythms are often entrained by the seasonal variation in photoperiod (i.e., the fraction of daylight in a day). Circadian rhythms are strongly entrained by the light–dark cycle and less strongly by daily variations in ambient temperature, food availability, and physical exercise. Many environmental cycles that may not entrain a rhythm can nevertheless mask it.

Substrates

Circadian rhythms have been demonstrated in almost ever species ever tested, from bacteria to humans. Although transcriptional/translational loops seem to underlie the intracellular process of generation of circadian rhythmicity in all organisms, the specific genes involved are not conserved across domains, kingdoms, or phyla. At the systems level, likewise, the pacemaking structures and the sensory receptors necessary for entrainment vary with the complexity of the organism. In mammals, a major circadian pacemaker is located in the suprachiasmatic nucleus (SCN), a small nucleus in the ventral hypothalamus composed of several thousand neurons. Each neuron in the SCN is an autonomous pacemaker, and the various cells are synchronized mainly through synaptic communication.

The mammalian circadian system relies exclusively on the eyes to acquire photic information necessary for entrainment, although other vertebrates and invertebrates possess a variety of additional photosensitive structures. Both classic photoreceptors (rods and cones) and photoresponsive ganglion cells in the retina of the eyes provide photic information to the mammalian circadian system. Very little is known about how the circadian pacemaker acquires the information about temperature and nutritional state that is needed for nonphotic entrainment. Temperature signals are available from cold- and warm-sensitive cells on the skin and in the body core. Hunger and satiety signals are available from the blood concentration of nutrients, taste and smell of the food being ingested, gastric distension, gastric contents, and blood levels of various hormones secreted by the stomach, by the intestines, and by fat cells.

The efferent pathways responsible for communication of circadian rhythmicity to the various organs are not well known but seem to involve neural as well as humoral mechanisms. One mammalian efferent pathway has been described in detail: the control of rhythmic melatonin secretion by the pineal gland is achieved through a tortuous pathway from the SCN to the

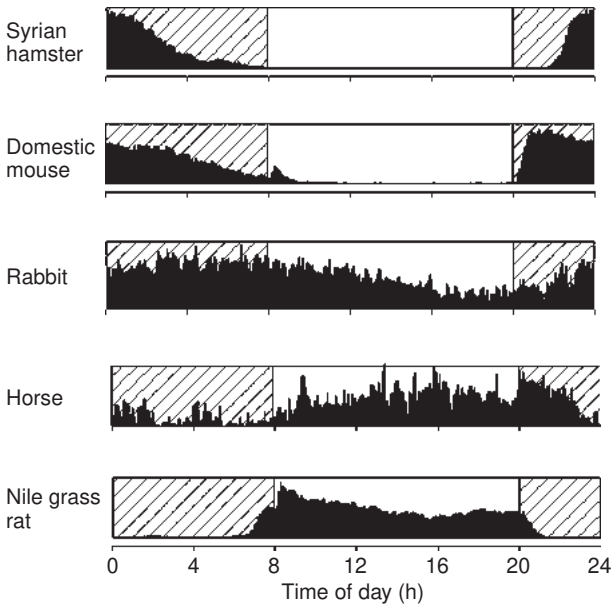


Fig. 1 Representative activity records of five mammalian species. In all cases, data were collected in a controlled laboratory environment where the light–dark cycle (indicated by the hatched and clear boxes) was the only prominent 24-h environmental cycle. The ordinates in each graph are set in arbitrary units to facilitate comparison between different species. Notice the differences in diurnality/nocturnality among the five species. Original figure from data collected in the laboratories of the author and his research collaborators.

paraventricular nucleus of the hypothalamus, to the intermediolateral column of the thoracic spinal cord, to the superior cervical sympathetic ganglion, and finally to the pineal gland.

Ecological Aspects

Evolutionary Advantage

Because extant bacteria exhibit circadian rhythmicity, it is usually assumed that endogenous rhythmicity was present already in the earliest life forms and was retained in all divergent branches along the evolutionary tree. In the absence of fossil evidence, however, it is equally possible that circadian rhythmicity evolved *de novo* multiple times in various taxonomic groups. Early life forms exposed to sunlight had to deal with the conflict between obtaining life-sustaining energy from solar radiation and being damaged by the Sun's strong ultraviolet emissions. Resolution of this conflict – in the form of daily vertical migration in the ocean – may have been the driving force for the evolution of circadian rhythmicity.

In general terms, it is often assumed that endogenous rhythmicity evolved as a mechanism that allowed organisms to prepare for predictable daily changes in the environment. For instance, photosynthetic plants could wait for sunlight each day, but those with an innate mechanism capable of anticipating sunrise would get an early start by initiating preparatory adjustments during the last part of the night. Similarly, nocturnal rodents could wait for the darkness of the night before getting ready to leave their burrows, but those with an innate mechanism capable of anticipating sunset would prepare in advance for the rigors of foraging. On a limited scale, experimental research has demonstrated enhanced reproductive fitness or survival in normal organisms as compared to organisms with deficient circadian systems.

Diurnality and Nocturnality

Phenomenology

Perhaps the most fundamental ecological issue in circadian physiology is an organism's adoption of a nocturnal niche or a diurnal niche. Evolutionarily, it is not certain whether the choice of a temporal niche was relevant to early life forms. If the first organisms were photoautotrophic and relied on energy from the Sun, then the choice of a diurnal niche would certainly have been important. However, if the first organisms were chemoautotrophic and relied on geothermal energy from deep-ocean vents, then the alternation of day and night on Earth's surface would have been of very little importance. Millions of years later, when living beings – particularly heterotrophic ones, such as animals – abandoned the ocean and colonized terrestrial environments, the choice of a nocturnal niche was probably necessary as a means of preventing desiccation. Thus, invasion of the diurnal niche likely became possible only after the evolution of integuments capable of preventing water loss.

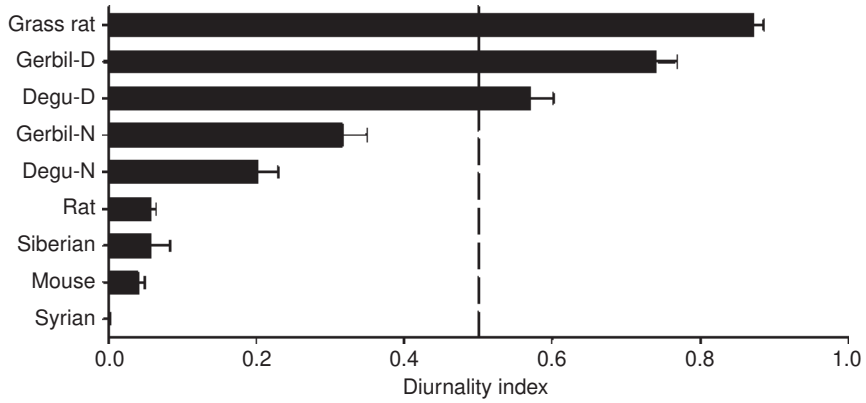


Fig. 2 Mean diurnality scores of nine groups of small rodents. The diurnality score was computed as the ratio of the number of activity-wheel revolutions during the light portion of the light–dark cycle and the number of wheel revolutions during the whole day, so that larger scores reflect greater diurnality. The dashed line indicates the theoretical separation between nocturnal and diurnal animals (i.e., equal amounts of activity during the light and dark portions of the light–dark cycle). Notice the gradient from predominant nocturnality to predominant diurnality. Adapted from Refinetti R (2006) Variability of diurnality in laboratory rodents. *Journal of Comparative Physiology A* 192: 701–714.

Although many organisms today can be classified as either nocturnal (night-active) or diurnal (day-active), many others defy classification. Representative activity records for five mammalian species are shown in **Fig. 1**. Under a light–dark cycle with 12 h of light and 12 h of darkness per day, Syrian hamsters (*Mesocricetus auratus*) are exclusively nocturnal. Domestic mice (*Mus musculus*) are predominantly nocturnal, but their active phase is quite long and extends slightly into the light portion of the light–dark cycle. Rabbits (*Oryctolagus cuniculus*) are not clearly nocturnal or diurnal. Horses (*Equus caballus*) are predominantly diurnal but have a long active phase that extends into the dark portion of the light–dark cycle. Finally, Nile grass rats (*Arvicanthis niloticus*) are almost exclusively diurnal.

A laboratory study involving seven species of small rodents revealed a gradient of temporal niches running from predominantly diurnal species to predominantly nocturnal species with many chronotypes in between, including species exhibiting wide intraspecies gradients of temporal niche (**Fig. 2**). Domestic mice (*Mus musculus*), laboratory rats (*Rattus norvegicus*), Syrian hamsters (*Mesocricetus auratus*), and Siberian hamsters (*Phodopus sungorus*) were found to be predominantly nocturnal, with small intra- and interspecies variability. Nile grass rats (*Arvicanthis niloticus*) were found to be predominantly diurnal, again with small intraspecies variability. Curiously, degus (*Octodon degus*) and Mongolian gerbils (*Meriones unguiculatus*) were found to be naturally distributed into two distinct groups – one predominantly diurnal and one predominantly nocturnal – so that a downward gradient of diurnality was observed from Mongolian gerbils classified as diurnal, degus classified as diurnal, gerbils classified as nocturnal, and degus classified as nocturnal.

Great intraspecies variability in diurnality, with some individuals showing predominantly diurnal activity patterns and others showing predominantly nocturnal activity patterns, has been described in other species as well. In goldfish (*Carassius auratus*), about 80% of individuals tested in the laboratory were found to be diurnal, whereas 10% were nocturnal, and 10% displayed very weak rhythmicity. In carpenter ants (*Camponotus compressus*), approximately 70% of individually tested animals were found to be nocturnal, whereas 30% were diurnal. Likewise, in subterranean mole-rats of various species, some members of the species were found to be diurnal and some were found to be nocturnal. Even some instances of intraindividual variability (i.e., the same individual being diurnal under some circumstances and nocturnal under other circumstances) have been reported. For instance, wolves (*Canis lupus*) are normally nocturnal; however, when traveling over long distances, they travel during the day. Conversely, migratory birds are normally diurnal, but they do most of their migratory flight at night.

Some authors have used the term cathemerality to refer to activity patterns that are not clearly diurnal or nocturnal. It has been suggested that cathemerality (lack of circadian rhythmicity) may be an adaptive feature that allows animals to optimally exploit the available resources without the temporal restrictions imposed by circadian rhythms. Of course, this reasoning is the very opposite of that used to explain the existence of circadian rhythmicity, but it is not absurd to assume that circadian rhythmicity provided selective advantage to some species and not to others.

Causality

Very little is known about the causes of temporal niche selection beyond the obvious fact that some species inherit a diurnal preference while others inherit a nocturnal preference or no preference at all. In animals, eyes specialized for day vision (i.e., eyes possessing retinal cones in addition to retinal rods) evidently facilitate adaptation to a diurnal niche, but image-forming photoreception is not essential for circadian entrainment because the photosensitive ganglion cells can provide sufficient photic input to the SCN. Researchers who have tried to identify the mechanisms responsible for diurnality or nocturnality have generally found that there is no clear difference between diurnal and nocturnal organisms except for the obvious difference in the phase angles of entrainment – that is, diurnal animals are diurnal because they are active during the day, and nocturnal animals are nocturnal because they are active at night.

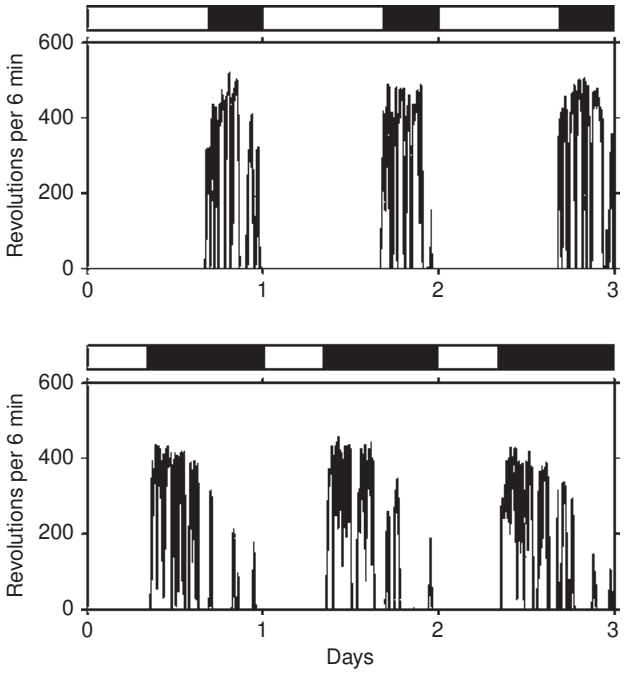


Fig. 3 Three-day segments of the records of running-wheel activity of a domestic mouse housed in the laboratory under a light–dark cycle with 8 h of darkness per day (top) or 16 h of darkness per day (bottom). The white and black horizontal bars denote the light and dark portions of the light–dark cycle, respectively. Notice that the active phase of the activity rhythm is longer when the nights are longer. Original figure from data collected in the author's laboratory.

Although we do not know why diurnal organisms differ from nocturnal ones, we do know that temporal niche selection depends on the interplay of two basic mechanisms: entrainment and masking. Entrainment results from the resetting of the pacemaker by photic stimulation at the appropriate time of the circadian cycle, whereas masking refers to the inhibition (negative masking) or disinhibition (positive masking) of behavioral activity without a direct effect on the pacemaker. Resetting of the pacemaker follows species-specific phase-response curves that do not differ between diurnal and nocturnal organisms except for the fact that diurnal organisms are responsive to light during their inactive phase, whereas nocturnal organisms are responsive to light during their active phase (i.e., both diurnal and nocturnal organisms are responsive to light at night). Similarly, the masking effects of light are equivalent in diurnal and nocturnal organisms except that light generally causes positive masking in diurnal organisms and negative masking in nocturnal organisms.

Naturally, entrainment and masking need not be restricted to photic stimuli. Outside the controlled conditions of the laboratory, organisms are subject not only to a light–dark cycle but also to rhythmic and nonrhythmic variations in food availability, ambient temperature, and intra- and interspecies competition. One case of interspecies competition that has been relatively well studied is that involving mice of the genus *Acomys*. In natural settings in rocky deserts of the Middle East, common spiny mice (*A. cahirinus*) share a foraging microhabitat with golden spiny mice (*A. russatus*). Normally, common spiny mice are nocturnal, whereas golden spiny mice are diurnal. However, if the common spiny mice are removed from the area, the golden spiny mice become nocturnal. This suggests that the golden spiny mice are normally forced into the diurnal niche by the competition for resources. Indeed, when golden spiny mice are trapped in the field and immediately tested individually in the laboratory, they exhibit a nocturnal pattern of activity. Thus, the phase reversal in spiny mice is quite interesting from an ecological point of view. It shows how masking mechanisms may supplant entrainment mechanisms in the determination of the temporal niche of species in the wild.

Seasonal Adjustments

Much research has dealt with the interaction between annual rhythms and circadian rhythms. Except at the equator, nights are longer in the winter and shorter in the summer, and it is well known that this seasonal variation in photoperiod causes a temporal compression or expansion of circadian rhythms. The phenomenon has been observed in natural settings (where the change in photoperiod is accompanied by changes in temperature and food availability) as well as in the laboratory (where only the photoperiod is changed). An example of expansion of the active phase (α) of a mouse under long nights in the laboratory is shown in Fig. 3. Notice that the expansion of α is accompanied by a reduction in exertion at each time point, so that the overall amount of activity (number of wheel revolutions, in this case) is conserved. As it would be expected, α is expanded under long nights in nocturnal organisms but under long days in diurnal organisms. As a rule, wintertime is associated with rhythm compression in

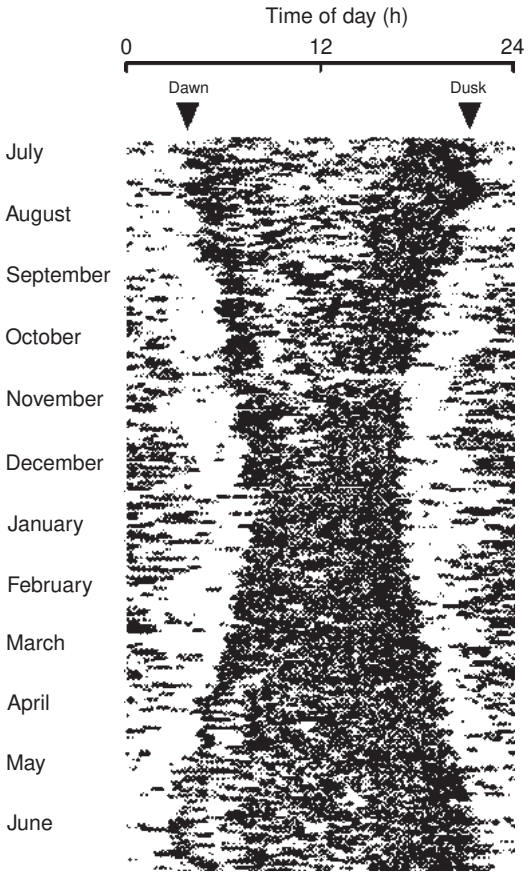


Fig. 4 The daily feeding rhythm of a mouflon sheep (*Ovis musimon*) maintained outdoors in Germany for a full year. Notice the gradual contraction – and later expansion – of the feeding rhythm as the days become shorter in the winter – and longer again in the summer. Adapted from Berger A, Scheibe KM, Michaelis S, and Streich WJ (2003) Evaluation of living conditions of free-ranging animals by automated chronobiological analysis of behavior. *Behavior Research Methods, Instruments, and Computers* 35: 458–466.

diurnal animals and rhythm expansion in nocturnal animals, whereas summertime is associated with rhythm expansion in diurnal animals and rhythm compression in nocturnal animals. A full-year record of feeding activity of a mouflon sheep (*Ovis musimon*) housed outdoors in Germany is shown in **Fig. 4**. This diurnal animal spent many more hours grazing during the summer than during the winter.

Seasonal variations have also been documented in other parameters of circadian rhythms, such as phase, amplitude, and period. An interesting seasonal modulation of rhythm amplitude is observed in beavers (*Castor canadensis*). During the winter, in Canada and northern United States, beavers remain essentially sequestered in their lodges or underneath the ice cover, so that their daily rhythm of activity is almost flat, whereas robust rhythmicity is present in the summer.

The interaction between annual and circadian rhythms occurs also in the opposite direction, as circadian rhythmicity can affect annual rhythms. Perhaps the best example of this interaction is the circadian modulation of entry into and arousal from hibernation. Studies conducted on several species of squirrels and hamsters have generally shown that entry into torpor is restricted to a narrow segment of the day (and, therefore, is modulated by the circadian system), although there is disagreement about the circadian modulation of arousal. **Fig. 5** shows the results of a study conducted on European hamsters (*Cricetus cricetus*). The animals were kept in the laboratory under simulated winter conditions of short photoperiod (8 h of light per day) and low ambient temperature (8 °C). Each dot in the figure corresponds to an episode of entry into or arousal from a deep hibernation bout. Although the temporal distribution of entries into torpor is not very tight, almost all entries occurred between 18.00 and 06.00. Arousals from torpor were scattered all over the day in this study. Some investigators have found that arousal from hibernation is restricted to a narrow segment of the day (and, therefore, is modulated by the circadian system), whereas others have not. Because the conflicting findings have been obtained in different species, they may be explained by species differences. The central question is whether the circadian system remains functional during hibernation, as a functional clock is required for the timing of arousal. Some researchers have observed circadian rhythmicity of body temperature (with very small amplitude) during hibernation, whereas others have not. A study of metabolic activity of various brain areas identified high activity in the site of the master circadian pacemaker (SCN) during hibernation, which constitutes evidence that the circadian system remains functional during the maintenance stage of hibernation.

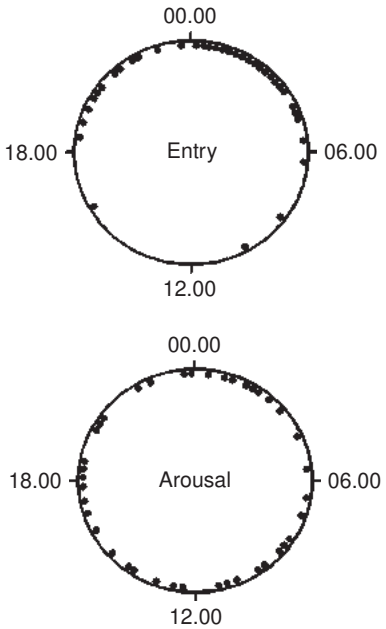


Fig. 5 Distributions of the times of entry into and arousal from deep hibernation bouts of eight European hamsters (*Cricetus cricetus*) maintained in the laboratory under simulated winter conditions (8 °C, 8 h of light per day). Notice that the distribution for entry into torpor is clustered around the late evening and early morning. Adapted from Waßmer T and Wollnik F (1997) Timing of torpor bouts during hibernation in European hamsters. *Journal of Comparative Physiology B* 167: 270–279.

Future Directions

Laboratory research on biological rhythms has concentrated on basic processes studied in a small number of species, mostly rodents. Few studies have examined interspecies differences in basic processes, and fewer yet have addressed ecological issues in natural environments. Field studies, on the other hand, have lacked the methodological sophistication necessary for the attribution of rhythmicity to environment, endogenesis, or entrainment, which is an essential step in the understanding of the ecological and evolutionary significance of the observed rhythmic process. The future hopefully will bring numerous studies that combine the rigor of laboratory experimentation with the realistic complexity of natural environments into groundbreaking investigations of the interplay of environmental cycles that modulate biological rhythmicity.

See also: Ecosystems: Tropical Seasonal Forest. General Ecology: Seasonality; Temperature Regulation

Further Reading

Berger, A., Scheibe, K.M., Michaelis, S., Streich, W.J., 2003. Evaluation of living conditions of free-ranging animals by automated chronobiological analysis of behavior. Behavior Research Methods, Instruments, and Computers 35, 458–466.
Dunlap, J.C., Loros, J.J., DeCoursey, P.J. (Eds.), 2004. Chronobiology: Biological Timekeeping. Sunderland, MA: Sinauer.
Foster, R.G., Kreitzman, L., 2004. Rhythms of Life: The Biological Clocks that Control the Daily Lives of Every Living Thing. New Haven, CT: Yale University Press.
Koukkari, W.L., Sothern, R.B., 2006. Introducing Biological Rhythms. New York: Springer.
Refinetti, R., 2006. Circadian Physiology, 2nd edn. Boca Raton, FL: CRC Press.
Refinetti, R., 2006. Variability of diurnality in laboratory rodents. Journal of Comparative Physiology A 192, 701–714.
Takahashi, J.S., Turek, F.W., Moore, R.Y. (Eds.), 2001. Handbook of Behavioral Neurobiology, Vol. 12: Circadian Clocks. New York: Kluwer/Plenum.
Waßmer, T., Wollnik, F., 1997. Timing of torpor bouts during hibernation in European hamsters. Journal of Comparative Physiology B 167, 270–279.
Young, M.W. (Ed.), 2005. Methods in Enzymology, vol. 393: Circadian Rhythms. San Diego, CA: Academic Press.

Competition

John R Wallace, Millersville University, Millersville, PA, United States
Mark Eric Benbow, Michigan State University, East Lansing, MI, United States

Glossary

Allelopathic chemistry The activity where one organism or species produces biochemicals that have negative or positive effects on the germination, growth, development, survival, and reproduction of other species.
Fecundity The ability to produce offspring.
Natural selection A mechanism of evolution, where there is differential survival of individuals based on differing phenotypes that lead to changes in heritable traits of a population.
Parasite-mediated competition An ecological relationship where a parasite disproportionally affects one of two species that compete for common resources and influences the outcome of the competitive interaction.
Quorum sensing A system of stimuli and responses associated with the biochemical activity of microbial populations that changes with population density.
Sexual cannibalism An activity where one sex of a species consumes the individual of the same species prior to, during or after copulation; this is often in the form of the female consuming the male.
Sexual dimorphism A phenotype where two sexes of the same species have different morphological or behavior differences that are different from gonads or sexual organs (body size, color, song).
Sexual selection A form of natural selection where organisms make preferential selection of mates based on differing traits or characteristics.

An Introduction Using a Narrative Case Study: Competition of Microbes and Insects

The existence of microbes on earth has been documented to extend nearly 3.5 billion years ago during the Precambrian period (Crippen *et al.*, 2015; Waggoner, 1996). Most bacteria form community assemblages of cells attached to a substrate creating a biofilm (Lang *et al.*, 2016). Such biofilms are ubiquitous, and include being found on and in plants, animals, substrates and decomposing animal carcasses, or carrion; and while these microbial assemblages are critical the decomposition of carrion (Parmenter and MacMahon, 2009; Crippen *et al.*, 2015; Nemergut *et al.*, 2011), interactions with microbes and other organisms (e.g., insects) may be quite intense and rather complex creating a competitive struggle for the resource. These interactions can be in the form of competition among the microbial communities, invertebrates, and vertebrate scavengers (Rozen *et al.*, 2008; Benbow *et al.*, 2015b). Carrion has often been characterized as a rich yet ephemeral resource patch, or a decomposition island (Carter *et al.*, 2007), and whether it takes the form of a deer lying alongside a roadway or a whale at the bottom of ocean, it can be the source for intense competition among organisms representing different kingdoms and domains of life. Because of the patchy and ephemeral nature of carrion (a temporary, but highly rich resource) (Benbow *et al.*, 2015a), it stands as an excellent example of how competition for any resource affects individual organisms and structures populations, communities, and ecosystems. We introduce this article on competition with a brief example of how questions related to carrion decomposition demonstrate several aspects of competition. The questions and hypotheses posed here can be applicable to other resources that are under competition within and among species, such as space and mates. An initial question is: What are the evolutionary and ecological factors that have driven contest between eukaryotic and prokaryotic organisms to secure such an ephemeral resource?

To answer this question when pondering decomposition ecology in the tropics, entomologist Janzen (1977) hypothesized of the evolutionary reason, if any, of why fruits rot, seeds mold, and meat spoils. Janzen (1977) reasoned that the stakes are high in this contest for the microbial communities associated with these resources, since arthropods often quickly find and consume the decomposing organic matter, including the consumption of the microbes competing for the newly available nutrients and energy. For the microbes, the cost of defeat for this resource is death, additionally, the arthropods may be exposed to similar risks by feeding on microbes in a form of chemical warfare (Janzen, 1977). In other words, toxins produced by microbes may be detrimental to both other microbes in the form of antibiotic-type toxins as well as making it objectionable and toxic to higher organisms (e.g., the consequences of eating spoiled meat in humans is foodborne illness from bacterial toxins). Such selective pressures might favor arthropods evolving ways to ignore such objectionable flavors or odors or select for behaviors that allow for detoxification; these interactions result in an evolutionary "arms race" between microbes and arthropods for competing for the limited resource.

How could the microbes win out in this battle? As Janzen (1977) later proposed, bacteria could render this food resource unpalatable via a number of evolved strategies. For example, he hypothesized that microbes could alter the nutrient composition, thus rendering it objectionable by the consumer; or they might select an alternative strategy by producing toxins that might repel the consumer. In turn, the animals that fed on fruit (or carrion) should evolve counterstrategies to compensate for the negative effects of competition with microbes. The early work by Janzen (1977) not only provided the foundation for a more thorough

 doi:10.1016/B978-0-12-409548-9.10597-4

Fig. 1 A larval mass of *Nicrophorus* sp. that was feeding on a salmon carcass. Photo credit: Bob Armstrong.

examination of the adaptive significance of animal–microbe competition but also unknowingly suggested at the time how this example may have been more complex than he realized. More than four decades later, Rozen *et al.* (2008) realized that there were little experimental data that tested Janzen's hypotheses related to decomposing fruit and other organic matter like carrion. Data were limited with respect to antimicrobial strategies that might allow animals to utilize decomposing organic matter be they fruits or meats. As Rozen *et al.* (2008) suggested, such work would need to be done to understand the evolutionary and ecological ramifications of competition between animals and microbes.

As Janzen (1977) discussed the cost of defeat for a carrion resource for the microbes was death, Rozen *et al.* (2008) reasoned that there were likely differences between facultative and obligate vertebrate scavengers, where both would suffer fitness effects during competition with microbes, obligate scavengers could also suffer with death if they lose the resource. This would be facultative scavengers can move and consume different food sources if competition with other consumers (or microbes) is intense, but the obligate carrion scavengers are restricted to the patchy and limited nature of such resources: if they are outcompeted, they suffer severe fitness costs or death. To test this scenario, researchers investigated how an obligate carrion consumer, the burying beetle (*Nicrophorus vespilloides*) utilized parental care and carcass choice as antimicrobial strategies (**Fig. 1**). From a fitness perspective, Rozen *et al.* (2008) understood that larval growth and body size were important fitness components that would influence adult survivorship and overall success in competition for breeding by adult beetles. To examine this relationship, they tested the effects of competition of microbes with beetles and evaluated their reproductive success.

In a suite of lab experiments, Rozen *et al.* (2008) used small containers with either fresh or old mouse carcasses, introduced mated female beetles and then manipulated beetle brood size. This manipulation allowed observation of parental behavior in terms of choice to breed and larval behavior associated with food begging on the two resources. The fresh carcasses had not decomposed and represented a resource with less intense microbial competition and presumed less microbial metabolic activity producing toxins and noxious odors. The old carcasses represented increased microbial activity, and thus increased competition for the resource.

The researchers found that larval size was significantly reduced on old mouse carcasses compared to those reared on fresh carcasses, indicating that the increased competition negatively affected the offspring of the beetles. They further explained that adult female and larval behaviors may be altered depending on the degree of microbial competition they encounter. For example, they also demonstrated that with old, microbe-laden carcasses, larvae begged significantly more indicating that they were stressed from a nutritionally poor diet associated with old carcasses, suggesting that the microbes had lowered the quality of the resource for the larvae. Female beetle behavioral results were mixed, where females preferred old carcasses to feed upon but then switched to fresh carcasses upon which to rear their broods (**Fig. 2**). This finding implies the level of parental care may influence any negative effects of larval competition with microbes.

The work by Rozen *et al.* (2008) confirmed and expanded upon Janzen's (1977) hypotheses explaining how higher carrion feeding organisms compete with microbial communities associated with decomposition through a variety of behavioral strategies. In reality, the interaction may be more complex than either group of scientists fathomed. Scientists have documented that the toxins and secondary metabolites produced by microbes do serve multiple purposes from repelling higher organism consumers to having antibacterial capacities that may reduce competition with other microbes (Lam *et al.*, 2009).

Recently, it has been shown that many microbes in such biofilms on carrion may be sending chemical signals using quorum sensing (bacterial cell-to-cell signals) mechanisms that communicate not only among other microbes but also to invertebrate carrion feeders (Ma *et al.*, 2012). Scientists know that these microbial communities change during the process of decomposition and affect how insect scavengers colonize the resource (Pechal *et al.*, 2014). The process by which chemical signals sent from microbes to insects affects competitive outcomes is an alternative to the toxin hypothesis proposed by Rozen *et al.* (2008); however, such signals may also change competition outcomes among microbes over the decomposition process as insects come

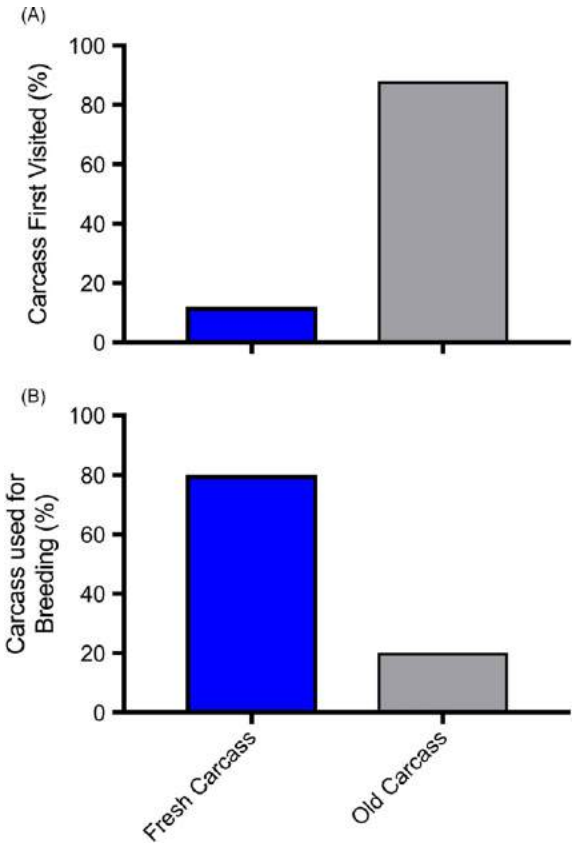


Fig. 2 The effect of carcass age (fresh versus aged) on the choice of *Nicrophorus vespilloides* for (A) female first visit to carcasses and (B) breeding on carcasses. Graphs recreated from data in Fig. 4 of Rozen, D. E., Engelmoer, D. J. P., & Smiseth, P. T. (2008). Antimicrobial strategies in burying beetles breeding on carrion. *Proceedings of the National Academy of Sciences*, ***105***, 17890–17895.

and go from a carcass. Such interactions imply that the complexity of competition is driven both evolutionarily and ecologically and that interactions such as in the above case study among microbes, insects, and vertebrates can determine that limited resources such as carrion can control the abundance of competing populations. In this article, we will expand on this case study and explore how communities are structured and how they change through a variety of species interactions, how competition is defined and identify the factors that affect competitive outcomes.

The Niche Concept

A common contention in evolutionary biology is that competition is a major driver of natural selection whether it is among conspecifics within a given population or different species with overlapping niches. Ecologists recognize that competition has been characterized in numerous fashions ranging from interactions among organisms of a single species to more than two populations of different species exerting negative effects on each other by for example, affecting growth, survivorship, and reproduction via a shared use of resources such as food, water, space, mates, or shelter (Ricklefs and Relyea, 2014; Bowman *et al.*, 2017). Shared resources are important, as they limit the degree of population growth even without constraining abiotic pressures such as temperature or humidity or biotic factors such as competition over those resources by other species. An important aspect of population biology is the *carrying capacity* of the environment, or the maximum number of organisms that can be sustained in a given habitat. Competition affects how two or more species utilize the resources in the environment and the carrying capacity can have important consequences on the outcome of competitive interactions: if the carrying capacity is high, there are more resources to be shared or competed for, while when it is low, there are fewer resources and the assumed competition for the limited resources becomes more intense. But just how species are able to occupy a place in natural ecosystem also depends on the different abiotic and biotic constraints on that species. When considering competition for a given resource, population limits are set based on the abiotic and biotic requirements that differentiate between the fundamental and a realized niche for each species (**Fig. 3** and see the following text).

It is important to distinguish the difference between these two niche types. The fundamental niche constitutes the range of abiotic conditions under which a particular species can exist, such as temperature, humidity, precipitation (Ricklefs and Relyea,

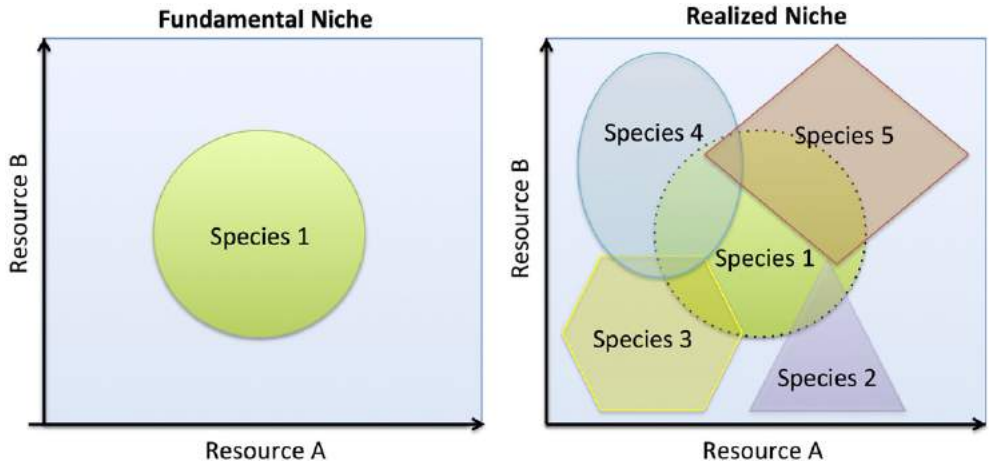


Fig. 3 This conceptualized schematic illustrates how in the left panel (A) a single species may be able to exploit two different resources in a fundamental niche within the *green circle*; however, resource use in reality is regulated by the interactions of two or more species as shown in the right panel (B) thus establishing the boundaries for a realized niche for species 1 but also the other species competing for resources 1 and 2. Modified from Fig. 14.3 in Bowman, W. D., Hacker, S. D., & Cain, M. L. (2017). *Ecology*. Sunderland, MA: Sinauer Publishers.

2014); or as the niche that would be occupied in the absence of competition from other species. For instance, certain plants are usually found in salt marsh environments and are rarely reported from freshwater habitats, and are often assumed to be "salt restricted." However, these water plant species can and do grow in freshwater but only in the absence of several other, but better plant species competitors. When the competitors are removed or are absent the "salt restricted" plants will expand their niche to include freshwaters. While many favorable geographic locations may be included in the fundamental niche range of species, the presence of competitors, predators, and parasites may exert biotic pressures that constrain the fundamental niche to a realized niche. Thus, the realized niche includes both abiotic and biotic conditions under which a species persists (Ricklefs and Relyea, 2014). The use of resources is shared by multiple species and their interactions affect the niche size of the other species. Thus, species interactions constitute a more restricted set of conditions that defines a realized niche for different species (Bowman *et al.*, 2017). In the example above, the "salt restricted" plant species have a realized niche of salt marshes, but have a fundamental niche of both salt and freshwater habitats: when their competitors are present they occupy their realized niche, but when the competitors are absent, they expand to their fundamental niche. Understanding the differences between fundamental and realized niches allows for a conceptual foundation important for determining how species interactions, and therefore competition, influence how species are distributed in space and change with time.

Species Interactions

Traditionally, competitive interactions among species have been characterized as pairwise comparisons that designate a negative, positive or zero effect type on the different individuals or species (Pianka, 1994). These interactions take various forms and are summarized in **Table 1**. A predatory interaction dictates that one organism, such as a cheetah, kills, and consumes another like the Thompson's gazelle (Lang and Benbow, 2013). In general, most predation occurs as interspecific interactions; however, cannibalism is common among both terrestrial and aquatic organisms (Greenwood *et al.*, 2010; Huss *et al.*, 2010). For example older mantids and spiders can be cannibalistic on younger individuals and sometimes even on mates (Elgar and Nash, 1988). For instance sexual cannibalism and sexual dimorphism have been well described for Golden Orb Weaving spider species (**Fig. 4**). In this species, the female is substantially larger ($>2\times$) than the males, demonstrating sexual dimorphism in body size; and as a result of this size difference, will often consume the much smaller male after they have mated (Yip *et al.*, 2016).

Similar to predation, herbivory is when a primary consumer such as grazing mayfly nymph grazes an algal covered rock in streams or how a three-toed sloth feeds on *Cecropia* leaves in the tropical forest canopy. Predation and herbivory differ in that herbivory does not always mean the death of the plant, whereas with predation one organism is killed (Lang and Benbow, 2013).

Symbioses such as mutualism, commensalism, and parasitism can either be obligate or facultative where two or more different species live in a close relationship, often one colonizing a host (Lang and Benbow, 2013). A mutualistic relationship dictates that the interaction is beneficial for both individuals such as the ant/acacia relationship where the acacia tree provides thorns and fruiting bodies to protect and feed the ants, and the ants provide protection against herbivores from grazing the acacia and against other competing plant species. A commensal relationship occurs when one individual benefits but the other is not affected positively or negatively. An example of commensalism is where phoretic aquatic mites colonize Carabidae beetles and the mites use the mobile insect to hitch a ride and disperse in the ecosystem without any measurable effect on the beetles (**Fig. 5**). Parasitism involves two organisms, the parasite and a host. In such interactions, the parasite benefits at the expense of the host by feeding on

Table 1 Summary of the types of pairwise biotic interactions between and among species

Interaction type	*Pairwise designation*		*Nature of interaction*
	Species A	*Species B*	
Mutualism, müllerian mimicry	+	+	Favorable interaction for both species
Commensalism	+	0	Species A benefits, but Species B is unaffected
Predation, herbivory, parasitism, batesian mimicry	+	−	Species A exploits Species B
Competition	−	−	Each species inhibits the other
Amensalism	−	0	Species A inhibited, Species B unaffected
Neutralism	0	0	Species A and Species B coexist without interactions

Designations are positive (+), negative (−), or no effect (0).

Fig. 4 A classic example of sexual dimorphism where the female is much larger than the male Golden Orb Weaving Spider (*Nephila* sp.). In this photo, a large female (~3 cm) is preparing a fly in silk as a food resource. The male spider (~3 mm) is under the female and above the fly. Photo Credit: M. Eric Benbow.

Fig. 5 Many insects acquire phoretic mites that hitch a ride for dispersal and do no harm to the insect, like this Carabidae beetle (*Pterostichus* sp.). (Photo credit: Bob Armstrong).

tissues or fluids of the host. A parasite also can influence competition between two species when one species is more susceptible to the parasite than a second (Lang and Benbow, 2013). For example, two species of *Anolis* lizards found islands are known to coexist under certain ecological conditions in Caribbean. Under average conditions, one species (*A. gingivinus*) is a better competitor and excludes the second species (*A. wattsi*) from certain parts of the islands; however, this better competitor is more susceptible to lizard malarial parasites, thus lowering its competitive advantage when the parasite is frequent in the populations (Schall, 1992).

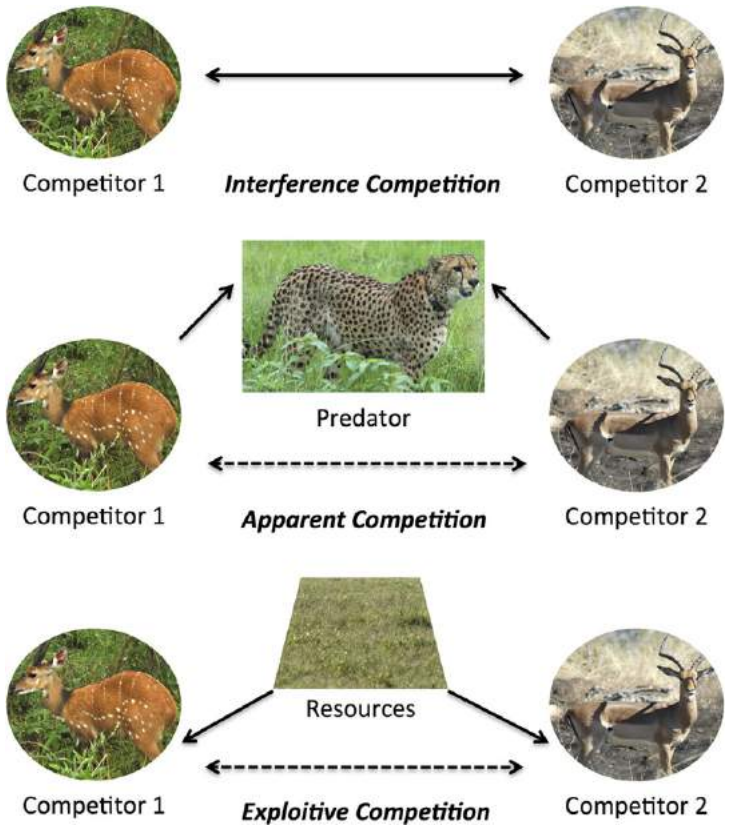


Fig. 6 A conceptual diagram showing the three major forms of competition: interference, apparent, and exploitive.

The ultimate outcome is that the two *Anolis* species only coexist when the malarial parasite is present (Schall, 1992), and this interaction is an example of parasite-mediated competition. Summarizing the breadth of ecological interactions that occur in nature, the next section will focus on the details and characteristics of competitive interactions.

Characteristics of Competition

Competitive interactions in ecological communities are often described within three main types: exploitive, interference, apparent and competition (**Fig. 6**). These forms of competition can either be an intraspecific interaction, where individuals of the same species compete for one or more resources; or through interspecific competition where individuals of different species compete for available resources (**Fig. 7**). Within these parameters, competition can occur indirectly between two or more species (or intraspecifically among individuals of a single species) that use a common and limiting resource in which such competition causes a depletion of the resource(s) and one species acquires the resource faster or more efficiently than the other species (Pianka, 1994). This type of competition has been characterized as resource or exploitative competition. For example, some plant species can acquire water and nutrients from the soil faster than other, nearby plant species, thereby reducing the resources for the less competitive species. Another example is the interspecific territoriality among birds for nesting sites, where some birds are better at occupying and defending common nesting locations that are limiting to another species (Stilling, 2002; Pianka, 1994). The bird species that is the better competitor for the nesting locations produces more offspring, thereby increasing their fitness and the population numbers in the next generation. The less competitive bird species produces fewer offspring with a resulting smaller population in the next generation. When this competitive advantage of one species over another continues for multiple generations without interruption it can lead to competitive exclusion, where the less competitive species is eliminated from that habitat. An example of competitive exclusion is the *Anolis* lizard species interactions of Caribbean islands that occur when the malarial parasite is absent in the populations, as described above.

Competition can also occur directly between two or more species (or individuals within a species) where one organism prevents access of another organism to a resource. This can take place in many forms; for example, via the production of toxins, such as with microbial populations described earlier in the case study of decomposing fruit and carrion; or by causing some other type of harm to the competing organisms that can range from behavioral threats to physical force (Stilling, 2002; Pianka, 1994). This type of competition is referred to as interference competition.

Fig. 7 (A) A romp of otters competing for a fish meal illustrates competition within a single species, or intraspecific competition (Photo Credit: John R. Wallace). (B) Hyenas compete with lions in the Serengeti for zebra prey but vultures also compete with hyenas for the scraps of lion, thus illustrating interspecific competition for food resources. Photo credit: John R. Wallace.

Another form of interaction that is not an obvious form of competition is apparent competition, where the negative affect of one species on another is mediated by the actions of shared natural enemies such as predators or parasites (Wootton, 1994; Morris *et al.*, 2004). The complexity of indirect interactions involving many species in changing environments makes apparent competition difficult to identify in nature. However, when two species that normally do not compete directly for a resource but can negatively affect each other through a mutual predator, apparent competition can exist. For example, Tompkins et al. (2002) determined that parasite susceptibility between ring-necked pheasants and gray partridges in the United Kingdom negatively influenced growth rate, fecundity, and survivorship in gray partridges simply because they were more susceptible to parasitic infections than pheasants. In some cases, however, apparent competition can occur in instances other than a shared enemy such as a predator or parasite, such as when one species facilitates the enemy of another (Ricklefs and Relyea, 2014). For example, when scientists were studying several species of shrubs near the California coast, they observed bare patches of soil around these shrubs (Bartholomew, 1970), an environmental characteristic that can be found in other habitats (**Fig. 8**). Hypothesizing that the patches were caused from allelopathic chemistry produced by the shrubs, they later determined that was not the case. What they found was that the purple sage (*Salvia leucophylla*), through its growth form, provided shade and refuge for mice herbivores to selectively consume seeds of the other, competitor shrub species (Ricklefs and Relyea, 2014). What appeared to be an example of interference competition was actually a case of apparent competition. The type and degree of competition often is dependent on the quality and type of shared resource, its availability and other environmental factors that mediate its acquisition (e.g., seasons in temperate forests).

Plants and animals compete for different resources. For example, plants often compete for sunlight, water, and nutrients (e.g., nitrogen and phosphorous) from the soil. Competition for these resources plays the primary role in how plants are distributed in the environment such as how desert grasses are distributed in central Idaho deserts (**Fig. 8**). In some ecosystems the distribution of certain plant species also has linked effects on the spatial distribution and feeding activity of many animal, that interact among themselves for the plant resources (**Fig. 9**).

Animals may compete for food, water, space, and mates. The latter two reasons for animal competition, space and mate procurement will be discussed in this section, as the other resources have been described above. A classic example for competition

Fig. 8 Nutrient competition for soil resources has contributed to this evenly spaced distribution of desert grasses in Crater-of-the Moons State Park, Idaho. (Photo Credit: John R. Wallace).

Fig. 9 The distribution of plant species in the African savannah may also have indirect effects on intra- and interspecific competition among grazing herbivores. In this photo, zebra, wildebeest, and gazelles or antelope compete within and among species for plant resources. Photo Credit: Richard W. Merritt.

for space has been well documented in Joseph Connell's foundational research where two genera of barnacles, *Balanus* and *Chthalmalus*, were described to compete for space along the high tide line within intertidal pools (Connell, 1961). In these habitats, the larger barnacle, *Balanus* would outcompete *Chthalmalus* for attachment sites on rocks; however, *Balanus* is unable to be survive exposure during low tides. Because *Balanus* requires total submersion in these tidal zones, *Chthalmalus* was able to outcompete *Balanus* for attachment sites above the low tide lines in these zones (Connell, 1961) (**Fig. 10**). Competition for attachment space based on desiccation tolerance differentiates the fundamental niche from the realized niche for *Balanus* barnacles (see the preceding text).

Competition for mates is best understood within the concept of sexual selection, and has been the primary driver in terms of the evolutionary mechanism for sex-specific traits for reproductive purposes: this can be in the form of males battling with other males (and sometimes females vs. other females) for access to female mates; or females choosing males based on characteristics such as size, plumage colors, song or behavioral characteristics. In other words, sexual selection is a result of natural selection in animals for sex-specific traits (Ricklefs and Relyea, 2014). For example, sexual dimorphism, defined as phenotypic differences between males and females of the same species (e.g., Golden Orb Weaving spider described above) includes traits of color,

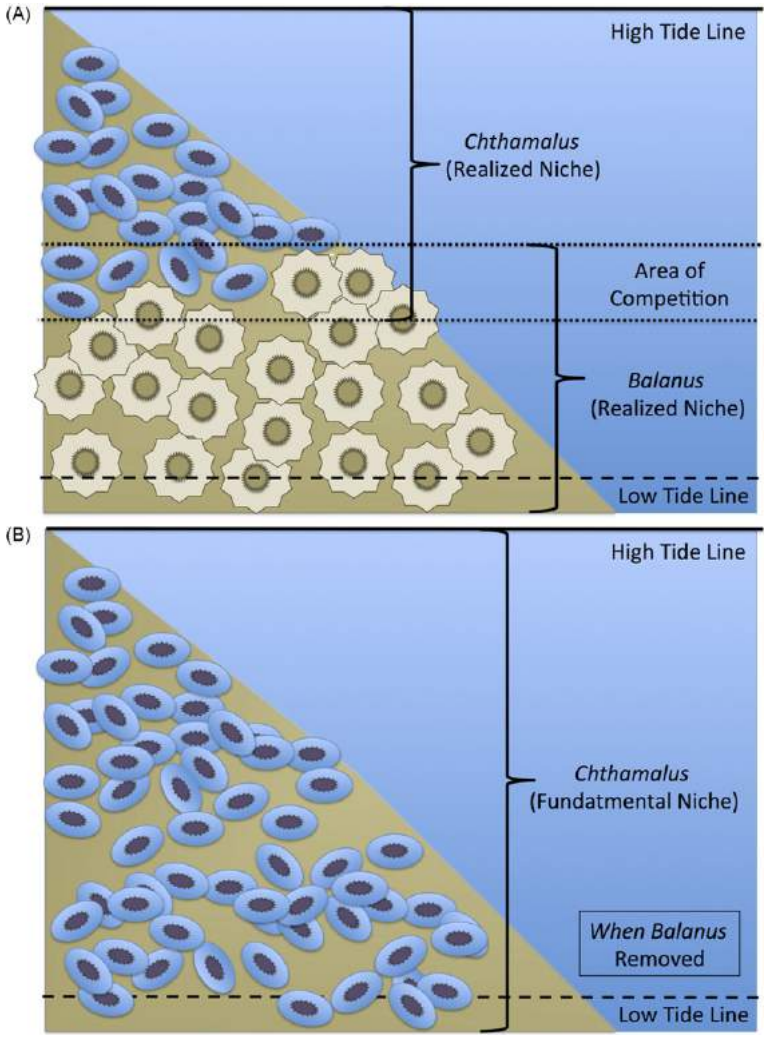


Fig. 10 Competition between two genera of barnacles, *Balanus* and *Chthalmalus*, for attachment space on intertidal rocks with (A) showing the outcome of competition when both species are allowed to colonize while (B) shows how *Chthamalus* spreads when *Balanus* is experimentally removed from the substrates. Figures redrawn from Connell, J. H. (1961). The influence of interspecific competition and other factors on the distribution of the barnacle Chthamalus stellatus. *Ecology*, ***42***, 710.

ornaments such as antlers, body size, gonads, song or behavior (Ricklefs and Relyea, 2014). Male size is often given as an example of sexual selection, where larger males often win out in competitive interactions with other males and either acquire the female, or based on the outcome the female chooses the winning male. This has been shown in many animal species, including classic examples of primates such as gorillas and other mammals (e.g., bears), fish, reptiles, and birds. When males compete with other males, sexual dimorphism can favor enhanced weaponry among males for physically competing to procure female mates, such as in some sheep (**Fig. 11**). Temporary weapons such as antlers of deer, elk, and moose as well as more permanent weapons such as horns of some sheep and mountain goats.

The evolution of female choice among competing males is tied to those features or behaviors that improve her fitness. Preferences by females generally are grouped within two types: (1) those that provide material benefits, such as food resources and; (2) those do not provide material benefits, such as shared parental care (Ricklefs and Relyea, 2014). Female choosiness may be connected to physical differences among males such as the covert feather length among male Resplendent Quetzals (*Pharomachrus mocinno mocinno*) or Wire-tailed manakins (*Pipra filicauda*), or behavioral differences such as courtship displays and dances among a variety of birds. In both cases, females choose males with longer outer covert feathers. The evolution of courtship behaviors especially among birds has driven male-to-male competition in terms of display dances or longer courtship songs especially among birds of paradise, bower birds and manakins in which the female chooses the male with the best dance, song or bower. There are two explanations as to why females might choose a male with a longer tail feather or a bower with better ornamentation, these include the good gene hypothesis and the good health hypothesis. The former refers to the female choosing a mate that might possess a superior genotype and the latter refers to female choice based on choosing the healthiest male (Ricklefs and Relyea, 2014): both traits are hypothesized to be beneficial to the females ultimate offspring.

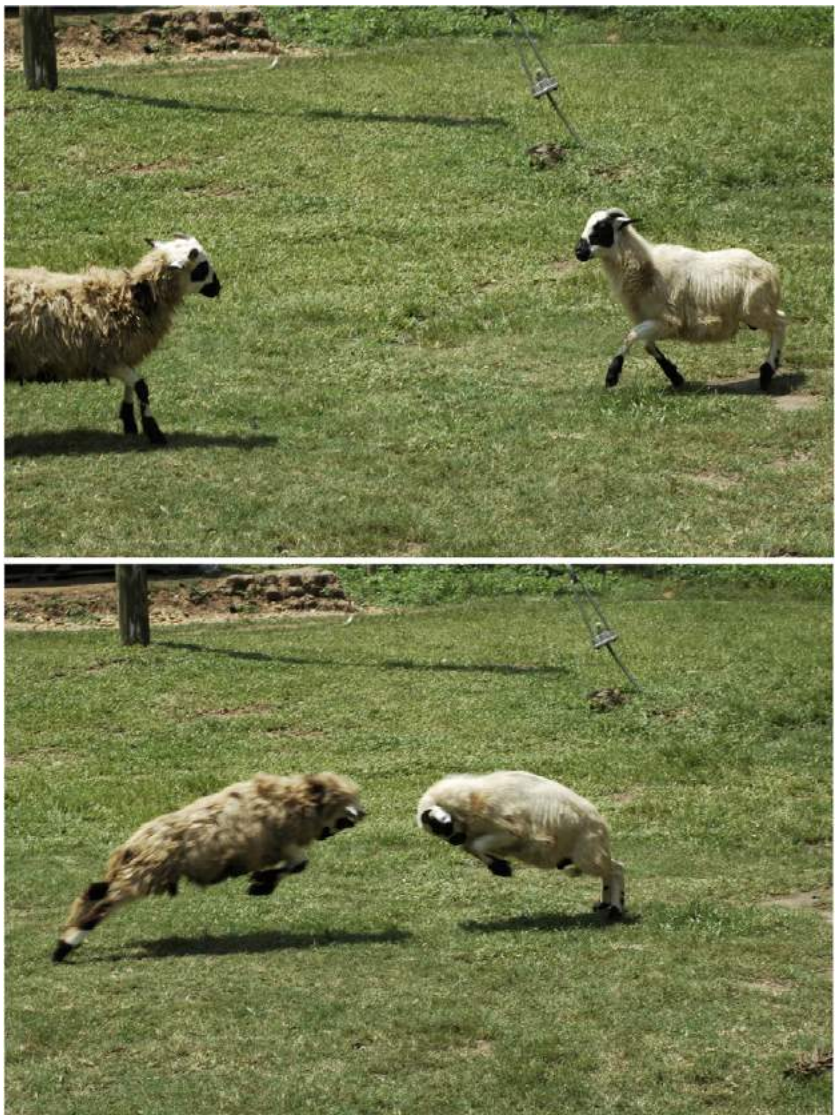


Fig. 11 Mate competition by two male sheep for access to a female. *Upper panel* showing males backing up from each other to get a running start. *Bottom panel*, males begin running toward one in other for a head-to-head collision. This competition can last hours until one is too hurt or tired to continue, losing access to the female. Photo Credit: M. Eric Benbow.

Limited Resources and Population Effects

Resources can be viewed as either renewable, or constantly regenerated, of finite and limited. An example of a renewable resource is the annual regeneration of seeds that are competed for by herbivores such as rodents and ants. A nonrenewable resource is one that is not regularly regenerated, such as colonizable substrata in a rocky intertidal zone (e.g., barnacle competition discussed above), or shelters as in the case of hibernating bears that require protection from freezing temperatures during winter months of very limited food (Ricklefs and Relyea, 2014).

Other types of resources that are the objective of competition are known to limit survivorship, growth, development, and reproduction of organisms. For example, silica in freshwater systems can decline in concentrations that limit growth and survivorship of some species of diatoms in aquatic ecosystems. While the silica is not a food resource, it is necessary for the structure of diatom cells (see below). A similar limiting resource is calcium carbonate in aquatic habitats that provides the structure to protective shells of clams and snails: if calcium carbonate is low in certain habitats, the protective shells are weaker, making the snails or clams more susceptible to predation. If there are multiple species in these aquatic habitats, the better competitor may use more of the silica or calcium carbonate, thereby having a negative effect on the other species that require and compete for these resources. Thus, competing organisms can reduce the availability and abundance of both renewable and nonrenewable resources having significant effects on the population growth and fitness of competing organisms. This phenomenon where a species population growth and survivorship are impacted by a limiting resources is known as Libieg's Law of the minimum.

By understanding what the minimum amount of a particular resource is for species, scientists are able to predict test for the outcome of competitive interactions by using both field and laboratory experiments. For instance, laboratory experiments by Tilman *et al.* (1981) illustrated how diatoms, *Asterionella* and *Synedra* competed for silica. Both freshwater diatom species require silica to construct their cell walls and in a series of elegant laboratory experiments Tilman *et al.* (1981) demonstrated that when either *Synedra* or *Asterionella* colonies were grown alone, both species of algae were able to reach stable populations levels even when the concentration of silica was experimentally reduced (see figures in Tilman *et al.*, 1981). However, when both species were grown together in the same aquarium, *Synedra* outcompeted *Asterionella* by reducing silica levels to a point at which *Asterionella* became extinct. The outcome between these two species of algae competing for silica is termed asymmetrical: while both species suffer negative effects, the harm is greater to *Asterionella* being the inferior competitor for the resource. As Darwin suggested and has been demonstrated through many studies, competition between species can influence the structure of ecological communities as well as the processes that create them. This result, where the superior species becomes dominant and thrives and the inferior species becomes locally extinct illustrates the principle of competitive exclusion (Bowman *et al.*, 2017). This principle states that when two species are limited by the same resource, they cannot coexist indefinitely if this resource is used the same way and one differentially acquires it over time (Ricklefs and Relyea, 2014).

In natural communities many species are able to coexist in some fashion despite sharing similar limiting resources. This situation is known as competitive coexistence (Bowman *et al.*, 2017). In order for competitive coexistence to occur, the competing species must be able to partition the limiting resource—sometimes referred to resource or niche partitioning. For example, in headwater streams of the northeastern United States, two species of stream dwelling caddisflies, *Pycnopsyche luculenta* and *Pycnopsyche gentilis* compete for leaves, wood and stones to construct larval cases used for to maintain balance in flowing systems. Both species are shredder caddisflies that feed on coarse organic material such as leaves. The life histories of these species overlap exactly, emerging as adults in the fall during leaf abscission, laying eggs from September to November, and growing as larvae from November to June the following year. Both species utilize leaves that they cut into disks and either arrange them so that they are made into flat cases or triangular in cross section during the winter or until larvae molt into the fifth larval stage (usually in March). By late spring, leaves are extremely limited as food or for case construction. To accommodate this reduction in leaf resource, *P. luculenta* will switch from leaves to using sticks for case construction and moves to the slow water along stream margins until pupation. In another strategy, *P. gentilis* switches its case from leaves to small stones from winter into spring and moves to the fast water of streams until pupation (Wallace *et al.*, 1992) (**Fig. 12**). Thus, competitive coexistence results from each species switching the limiting resource for case construction to alternative sources and by also moving to occupy different habitats where the alternative resources are more common.

In a terrestrial example, four species of Jamaican *Anolis* lizards were used as models to study resource partitioning (Schoener, 1974). Living together in the same trees and eating similar food, these four species separated themselves along the height and thickness of the perches of the trees. Similar to the caddisfly example above, the lizards spatially partitioned tree habitat and their food resources in order to compete less for these resources and coexist in the same tree. In addition to these examples of how competition can lead to coexistence, there are also other factors that affect the dynamics and outcomes of competition in ecosystems.

Anthropogenic Factors Affecting Competition

Competitive interactions among taxa can be fueled by either natural or anthropogenic activities. So far in this article only natural activities; however, anthropogenic (human-associated) activities can also affect competitive interactions and their outcomes. Fire suppression in parts of the southeastern United States in the early part of the 20th century is one example of how human modification to ecological processes affected the outcomes of competition. In this example, longleaf pine trees (*Pinus palustris*) dominated many forests of the southeast when human-induced fire suppression allowed this woody tree species to outcompete the grassy, herbaceous understory (Ricklefs and Relyea, 2014). The competitive interactions among herbaceous and woody species in this region were altered when fire suppression was stopped and fires were allowed to sweep through the area. Grasses such as the wiregrass were able to become established and reproduce because the fire frequency reduced the establishment of the longleaf pine trees, thus reducing its competitive ability. Fire managed habitats that were allowed fires at an intermediate frequency provided conditions where both species could coexist and also allowed for a greater overall plant diversity in these areas.

Another example is the anthropogenic reintroduction of once locally extinct wolves back into the Yellowstone National Park ecosystems. In this case, the disturbance was the hunting and removal of wolves from these areas by humans. When these apex predators were reintroduced into these systems the interactions of herbivores such as deer and elk, were found to shift the outcomes of competition among plant species: the grazing pressure of large ungulates such as deer and elk allowed for different plant species to thrive and flourish, allowing for competitive coexistence, and having additional effects on the numbers and diversity of smaller herbivores (such as mice and rabbits) that compete for herbaceous plant species in the ecosystems (Ripple and Beschta, 2004). This example demonstrates how complex ecosystems can be and that changes in certain biotic interactions (such as predation) have widespread ecological effects that are often mediated through shits in competitive outcomes for resources.

The introduction of nonnative species into new environments has had profound impacts on not only competition among native taxa but also reducing biodiversity in these systems. Native taxa have succeeded and evolved because of competitive success in these systems among native competitors. However, introducing an invasive species from another region may introduce a new

Fig. 12 Comparison of *Pycnopsyche gentilis* larval cases made from winter, transitional, and late spring materials. The larvae change their case construction to reflect resources in the habitat to avoid competition with other case building caddisflies (e.g., *P. luculenta*). Photo Credit: Bob Henricks.

competitor into the native ecosystem. Many examples have demonstrated that these invasive species may be better competitors than the native species, thus, becoming established and successful in spreading. For example, introductions of nonnative taxa into aquatic systems are a growing problem worldwide. The introduction of freshwater mussels whether they are zebra mussels (*Dreissena polymorpha*) into the Great Lakes or quagga mussels (*Dreissena rostriformis*) into the Colorado River in the southwest United States and other aquatic systems, is changing competition among other filter feeding organisms in these massive aquatic systems, as well as other organisms competing for food and space on substrate attachment sites (Strayer, 1999).

Another example is the introduction of the Nile perch and Nile tilapia into Lake Victoria in the late 1950s that not only caused a devastating decline in native haplochromine cichlid diversity but also caused shifts in cichlid mouthpart morphology and habitat preference. These biotic shifts changed the competitive outcomes among cichlids for different food types after the introduction nonnative Nile perch (Bwanika *et al.*, 2006). The Nile perch introduction led to the disappearance of several trophic groups of cichlids that in turn released macroinvertebrate species from predation. The released predation of macroinvertebrates, a major food resource of the insectivorous haplochromine cichlid (*Haplochromis pyrrhocephalus*), caused a resurgence of the haplochromine ciclid. The resurgence of this species was due to the increase in macroinvertebrate prey that was the shard resource by the other cichlid competitors; thus, the introduction of a nonnative species had foodweb wide effects that shifted competitive interactions and outcomes of several cichlid species (Bwanika *et al.*, 2006).

In addition to aquatic examples, terrestrial nonnative introductions have also had significant impact on ecosystems through changing competitive interactions. One classic example was the introduction of rabbits onto the continent of Australia. This introduction had numerous negative impacts, where rabbits were so abundant that they outcompeted native grazers such as

kangaroos for food. As burrowers, the high populations of rabbits also had significant impact on competing native burrows such as the only burrowing kangaroo, the boodie, but also with burrowing bandicoots (Stilling, 2002). Most recently, the introduction of Burmese pythons and African rock pythons into the Everglades National Park has created food resource competition with the American alligator. The alligator was the apex reptilian predator in the region but it has been replaced with these large nonnative pythons. Burmese and African rock pythons have caused a decline in mammalian diversity within the park but they have also outcompeted alligators for this food base such that alligators have now entered the food chain for pythons (Dorcas *et al.*, 2012). The outcomes of both native and anthropogenic species competition continue to be studied at both mechanistic and descriptive levels of inquiry with much research into developing and testing predictive models.

Modeling Competition and Predicting Outcomes

Population ecologists have long been interested in the factors that lead to population growth and decline, and one such factor is the role competition plays between populations of different species that share common resources. Such interactions can be modeled in several ways, but one classic and relatively simply approach has been the Lotka–Volterra competition model when such a resource is limiting between two different species. In general terms, the exponential growth of a population can be described by the following logistic growth equation:

$$\frac{dN}{dt} = r_1 N_1 \left(1 - \frac{N_1}{K}\right)$$

In this equation, $N =$ population size and the term on the left represents the rate of growth over a period of time. On the right, r is the intrinsic rate of increase, or maximum possible growth rate, for population N. The part of the equation on the far right factors in the carrying capacity of N_1, or the number at which the population tends to stop increasing: this term of the equation represents intraspecific competition for resources. This equation is quite useful if one wishes to examine the effect of competition between two different species on their population growth rate. To account for species interactions, one must consider competition coefficients (α and β). As defined by Ricklefs and Relyea (2014), these competition coefficients account for the interactions between the number of individuals of species 1 and the number of individuals of species 2. Now, using two equations that denote species 1 and 2 with subscripts, we are able to consider the effects of either species on the growth rate of the other; thus, allowing one to examine interspecific competition for a single resource:

$$\frac{dN_1}{dt} = r_1 N_1 \left(1 - \frac{N_1 - \alpha N_2}{K_1}\right)$$
$$\frac{dN_2}{dt} = r_2 N_2 \left(1 - \frac{N_2 - \beta N_1}{K_2}\right)$$

Both sets of equations examine how population size fluctuates relative to the carrying capacity of the environment. Moreover, these equations can be used to explain the dynamics between populations of species 1 and 2; that is, how their respective abundances change together. These relationships can be viewed graphically (**Fig. 13**). In this figure, when population N_2 is absent,

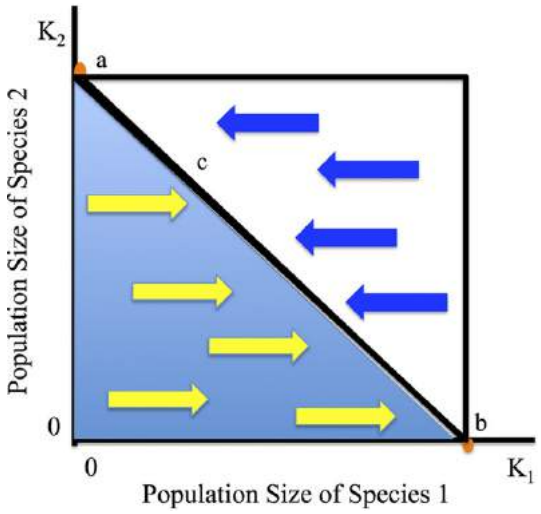


Fig. 13 Schematic represents changes in the population size of species 1 when competing with species 2. Populations in the *blue shaded* area will increase and reach an equilibrium along the diagonal line. Populations in the *unshaded area* will decrease as they approach the diagonal line. The letter a represents all resources being taken up by species 2 and none left for species 1; letter b represents no individuals of species 2 present and; letter c represents the zero growth isocline. From Stilling, P. (2002). *Ecology: Theories and applications.* Upper Saddle River, NJ: Prentice Hall Publishers—with permission.

population N_1 will increase until it reaches the carrying capacity. However, if this habitat is filled with population N_2, then N_1 cannot increase because N_2 has reached K_2. There are many combinations of N_2 and N_1 and these points fall along the diagonal line called the zero-growth isocline as described by Stilling (2002). When these figures are overlayed and vectors are added, there are four possible outcomes between these interacting populations: (1) species 1 becomes extinct; (2) species 2 becomes extinct; (3) depending on initial densities of both species, either one goes extinct and; (4) the two species coexist (Stilling, 2002). While quite simple, these modeling and visualization approaches are useful for testing how competition over resources affects the population growth of species. By understanding how and why populations change in nature and in response to biotic interactions, scientists can better predict how certain species will change over time and in response to a changing environment. This information also is important to understanding why species are distributed in certain environments and why some species are able to disperse and establish in many habitats, while other species are restricted to certain regions, ecosystems or local habitats. Many models are more complex and incorporate population responses to both abiotic and biotic conditions in efforts to predict the outcomes of changing environments on the distribution and abundance of species. These predictions are important for both understanding natural dynamics of ecosystems and for guiding natural resource management locally, regionally, and globally.

Conclusions

Ecosystems are community networks of organisms and populations of many species that interact within habitats that are defined by abiotic conditions that establish the initial restrictions to growth, development, and reproduction of a population. These conditions define the background within which organisms of each species interact in ways that change how they acquire necessary resources such as food, water, space, shelter, and mates that allows for continuous feeding, dispersal and reproduction; all factors that allow for each species to be sustained in the environment. While there are many ways that species interact for access to these resources, competition both among individuals of the same species and among individuals of multiple species often have strong and impactful effects on the coexistence or extirpation of species from habitats and ecosystems. Sometimes competition comes in the form of directly interacting with the other species, inhibiting them from acquiring the resources, or more indirectly through consuming the resources, and thus reducing them for the other species. It can also be much less obvious and more difficult to observe or quantify, as in the case of apparent competition. Nevertheless, the competitive interactions of species often define how scientists understand why and how species are distributed in different habitats; act as an area of research for understanding why certain human activities may affect the distribution and existence of species; and can be quantified to make predictions of species population biology dynamics in a changing world. Understanding and having the capacity to model such interactions provides scientists tools for defining factors that define ecological systems and the natural world, and for making predictions of how organisms, populations and communities respond to human-induced environmental changes.

See also: Evolutionary Ecology: Gauses Competitive Exclusion Principle. General Ecology: Dominance

References

Bartholomew, B., 1970. Bare zone between California shrub and grassland communities: The role of animals. Science 170, 1210–1212.
Benbow, M.E., Pechal, J.L., Mohr, R.M., 2015a. Community and landscape ecology of carrion. In: Benbow, M.E., Tomberlin, J.K., Tarone, A.M. (Eds.), Carrion ecology, evolution and their applications. Boca Raton, FL: CRC Press.
Benbow, M.E., Tomberlin, J.K., Tarone, A.M., 2015b. Carrion ecology, evolution and their applications. Boca Raton, FL: CRC Press.
Bowman, W.D., Hacker, S.D., Cain, M.L., 2017. Ecology. Sunderland, MA: Sinauer Publishers.
Bwanika, G., Chapman, L., Kizito, Y., Balirwa, J., 2006. Cascading effects of introduced Nile perch (*Lates niloticus*) on the foraging ecology of Nile tilapia (*Oreochromis niloticus*). Ecology of Freshwater Fish 15, 470–481.
Carter, D.O., Yellowlees, D., Tibbett, M., 2007. Cadaver decomposition in terrestrial ecosystems. Naturwissenschaften 94, 12–24.
Connell, J.H., 1961. The influence of interspecific competition and other factors on the distribution of the barnacle *Chthamalus stellatus*. Ecology 42, 710.
Crippen, T.L., Benbow, M.E., Pechal, J.L., 2015. Microbial interactions during carrion decomposition. In: Benbow, M.E., Tomberlin, J.K., Tarone, A.M. (Eds.), Carrion ecology, evolution and their applications. Boca Raton, FL: CRC Press.
Dorcas, M.E., Willson, J.D., Reed, R.N., Snow, R.W., Rochford, M.R., Miller, M.A., Meshaka, W.E., Andreadis, P.T., Mazzotti, F.J., Romagosa, C.M., 2012. Severe mammal declines coincide with proliferation of invasive Burmese pythons in Everglades National Park. Proceedings of the National Academy of Sciences 109, 2418–2422.
Elgar, M.A., Nash, D.R., 1988. Sexual cannibalism in the garden spider *Araneus diadematus*. Animal Behaviour 36, 1511–1517.
Greenwood, M.J., Mcintosh, A.R., Harding, J.S., 2010. Disturbance across an ecosystem boundary drives cannibalism propensity in a riparian consumer. Behavioral Ecology 21, 1227–1235.
Huss, M., van Kooten, T., Persson, L., 2010. Intra-cohort cannibalism and size bimodality: A balance between hatching synchrony and resource feedbacks. Oikos 119, 2000–2011.
Janzen, D.H., 1977. Why fruits rot, seeds mold, and meat spoils. American Naturalist 111, 691–713.
Lam, K., Geisreiter, C., Gries, G., 2009. Ovipositing female house flies provision offspring larvae with bacterial food. Entomologia Experimentalis et Applicata 133, 292–295.
Lang, J.M., Benbow, M.E., 2013. Species interactions and competition. Nature Education Knowledge 4, 8.
Lang, J., Erb, R., Pechal, J., Wallace, J., Mcewan, R., Benbow, M., 2016. Microbial biofilm community variation in flowing habitats: Potential utility as bioindicators of postmortem submersion intervals. Microorganisms 4 (1),
Ma, Q., Fonseca, A., Liu, W., Fields, A.T., Pimsler, M.L., Spindola, A.F., Tarone, A.M., Crippen, T.L., Tomberlin, J.K., Wood, T.K., 2012. *Proteus mirabilis* interkingdom swarming signals attract blow flies. The ISME Journal 6, 1356–1366.

Morris, R.J., Lewis, O.T., Godfray, H.C.J., 2004. Experimental evidence for apparent competition in a tropical forest food web. Nature 428, 310–313.
Nemergut, D.R., Costello, E.K., Hamady, M., Lozupone, C., Jiang, L., Schmidt, S.K., Fierer, N., Townsend, A.R., Cleveland, C.C., Stanish, L., Knight, R., 2011. Global patterns in the biogeography of bacterial taxa. Environmental Microbiology 13, 135–144.
Parmenter, R., Macmahon, J., 2009. Carrion decomposition and nutrient cycling in a semiarid shrub-steppe ecosystem. Ecological Monographs 79, 637–661.
Pechal, J., Benbow, M., Crippen, T., Tarone, A., Tomberlin, J., 2014. Delayed insect access alters carrion decomposition and necrophagous insect community assembly. Ecosphere 5.art45
Pianka, E.R., 1994. Evolutionary ecology. New York, NY: HarperCollins.
Ricklefs, R.E., Relyea, R., 2014. The economy of nature. New York, NY: W.H. Freeman and Company Publishing.
Ripple, W.J., Beschta, R.L., 2004. Wolves, elk, willows, and trophic cascades in the upper Gallatin range of southwestern Montana, USA. Forest Ecology and Management 200, 161–181.
Rozen, D.E., Engelmoer, D.J.P., Smiseth, P.T., 2008. Antimicrobial strategies in burying beetles breeding on carrion. Proceedings of the National Academy of Sciences 105, 17890–17895.
Schall, J.J., 1992. Parasite-mediated competition in Anolis lizards. Oecologia 92, 58–64.
Schoener, T.W., 1974. Resource partitioning in ecological communities. Science 185, 27–39.
Stilling, P., 2002. Ecology: Theories and applications. Upper Saddle River, NJ: Prentice Hall Publishers.
Strayer, D.L., 1999. Effects of alien species on freshwater mollusks in North America. Journal of the North American Benthological Society 18, 74–98.
Tilman, D., Mattson, M., Langer, S., 1981. Competition and nutrient kinetics along a temperature gradient: An experimental test of a mechanistic approach to niche theory. Limnology and Oceanography 26, 1020–1033.
Tompkins, D.M., Parish, D.M., Hudson, P.J., 2002. Parasite-mediated competition among red-legged partridges and other lowland gamebirds. The Journal of Wildlife Management. 445–450.
Waggoner, B.M., 1996. Bacteria and protists from Middle Cretaceous amber of Ellsworth County, Kansas. PaleoBios 17, 20–26.
Wallace, J.R., Howard, F.O., Hays, H.E., Cummins, K.W., 1992. The growth and natural history of the Caddisfly *Pyncopsyche luculenta* (Betten)(Trichoptera: Limnephilidae). Journal of Freshwater Ecology 7, 399–405.
Wootton, J.T., 1994. The nature and consequences of indirect effects in ecological communities. Annual Review of Ecology and Systematics 25, 443–466.
Yip, E.C., Berner-Aharon, N.A., Smith, D.R., Lubin, Y., 2016. Coy males and seductive females in the sexually cannibalistic colonial spider, *Cyrtophora citricola*. PLoS One 11. e0155433

Further Reading

Odum, E.P., Barrett, G.W., 2004. Fundamentals of ecology. Boston, MA: Cengage Learning Publishing.

Dispersal–Migration

AP Ramakrishnan, Portland State University, Portland, OR, USA

Introduction

Dispersal, or the movement and subsequent breeding of individuals from one area to another, strongly influences the population dynamics of a species. Dispersal can help regulate population size and density; many animals, such as aphids and female root voles, have increased dispersal rates under high density situations. Sometimes low density instead of high density is associated with greater dispersal rates. For example, during range expansions, peripheral populations of some grasshoppers may experience higher dispersal rates though they are of lower density than central populations, probably because of fitness costs associated with morphologies specialized for dispersal.

Such dispersal events can have large effects on neighboring populations. Marginal populations that are subject to high rates of immigration may experience a rescue effect, where despite poor genetic or ecological conditions, populations are able to persist. On the other hand, high dispersal rates can inhibit adaptation to novel environments due to constant influx of nonadapted individuals. Small populations that experience high rates of emigration may have a higher probability of extinction under such situations.

Natural populations in highly fragmented areas, such as agricultural or urbanized settings, may not experience sufficient levels of dispersal. Lack of dispersal can lead to high rates of inbreeding, which can lead to decreased fitness in many species. Because dispersal can have such strong effects on populations, dispersal patterns and processes are important when considering the potential spread of a biocontrol agent, pathogen, or invasive species into a new range. Dispersal also has implications for species redistributions due to climate change, as the dispersal rates and distances of a species will affect its potential to shift its range in response to climate change.

Two types of dispersal are commonly distinguished: natal dispersal, which is movement and subsequent breeding away from the birth territory or area, and breeding dispersal, which is movement from one area to another after the first breeding season. Dispersal of spores, or haploid life stages (such as pollen), strongly affects patterns of gene flow in a species, but the process is not generally considered to be directly associated with population dynamics. Dispersal in plants is generally limited to natal dispersal, as little to no secondary movement is possible, while many animals disperse multiple times.

All species disperse to some extent, in part because resources become limited locally as populations grow. Seedlings of plants must grow at some distance from the parent plant in order to obtain enough water, nutrients, and light to survive. Similarly, animals must disperse to avoid competing for resources such as mates, food, and territory. Depending on intraspecific patterns of resource limitation, dispersal is often sex-biased. In mammals, females tend to disperse more often than males; the trend is reversed in birds.

In areas with high temporal environmental variation, or in areas prone to frequent disturbances, species with greater dispersal abilities are expected to have a greater likelihood of survival. When one population's habitat is rendered untenable, if the species has a high dispersal rate, many individuals in that population will be able to move to a more suitable area. In the case of nonmotile organisms such as plants, high dispersal rates increase the likelihood that another population may be established even as the original population is rendered extinct. When studying populations that specialize in habitats with high temporal environmental variation, it is sometimes appropriate to distinguish between spatial and temporal dispersal. For example, many animals and plants that live in deserts with unpredictable rainfall will produce desiccation-resistant embryos that delay maturity until favorable environmental conditions cue further development. Instead of traveling long distances to reach suitable habitat, the individuals produce offspring that are able to lie dormant until the habitat is once again suitable for survival and reproduction. Because dispersal can enable escape from low-quality environments and access to higher-quality resources, many species that specialize in colonizing disturbed areas tend to have greater dispersal abilities than species that live in relatively stable habitats.

In some cases, dispersal can have a high cost associated with it, especially if individuals that disperse experience a higher mortality rate than those that do not disperse, or that disperse only a short distance. Because individuals are moving to an area that may not be as productive, and because they may have to travel through unsuitable habitats, mortality rates during the dispersal process may be high. The number of individuals that successfully establish in a new area may be far fewer than the number of individuals engaging in the dispersal process. In plants and other organisms with no choice involved in the dispersal process (passive dispersal), many propagules may never establish simply because they land in an unsuitable habitat. In animals where some choice may be involved in the final dispersal location (active dispersal), survival of dispersing individuals may be higher than individuals of species with passive dispersal, but there are still risks associated with dispersal, such as locating an appropriate territory, finding a mate, and successfully breeding in the new area. However, the benefits of dispersal can overcome the costs if mates and/or resources are limiting in the home range.

The process of dispersal is not necessarily as simple as suggested above, as it involves both emigration (leaving the original patch) and immigration (entering a new patch). The entire process of dispersal can be divided into approximately four different

stages: (1) emigration, (2) exploring or traveling through the surrounding habitat, (3) immigrating to a different patch, and (4) successfully breeding in the new patch. Each of these stages has a cost involved. Leaving the original patch involves leaving an area where resources are known to exist, but may have become limiting. The exploratory phase of dispersal can involve a high risk of mortality, as the individual may have to travel through territories with inadequate resources. In many plants and other passive dispersers, the exploratory phase entails a high rate of mortality, as seeds often land in areas unsuitable for growth. Even when a propagule successfully disperses to a hospitable environment, it may not be able to establish there, due to mortality rates associated with establishment. The risks involved with emigration, exploratory movement, and settling in a new patch can be outweighed by the potential benefits of dispersal if successful dispersal significantly increases the fitness of the individual.

There are varying degrees of active and passive dispersal, with many species exhibiting intermediate levels of participation in the dispersal process. In many animals, dispersal is active, involving a high level of choice during the dispersal process. In passive dispersal, there is little or no choice involved in selection of the final location. In many insects, many marine animals, and all plants, dispersal is largely passive, depending on air currents, water currents, or on the actions of vectors transporting the propagule. Larvae of many marine animals are often dispersed solely at the whims of the currents or in ship ballast. Insects are often at the mercy of the wind when entering a dispersal phase, especially if they cannot generate enough speed to overcome wind velocities. However, even dispersal of small insects need not be completely passive. Small insects, even if they are not large enough to overcome wind velocity, can have some level of choice as to where they land. They can begin exiting a wind stream when they decide to settle, then make short, self-powered trips to explore the surrounding area and find a suitable habitat.

Though considered passive dispersers, plants can regulate dispersal to some extent. Seed size, shape, and seed coat construction vary among species. Seed morphologies that aid dispersal include barbs (for attaching to animals), eliasomes (for attracting ants as dispersal vectors), or pappus scales (to assist in wind transport). However, because the seed itself is not actively involved in the decision process, it is still a passive process.

A species with little innate dispersal ability may be able to move greater distances and have higher survival than expected if it has the ability to be spread by a vector, such as ants, birds, or other animals. Plants commonly use vector-assisted dispersal, and there are many instances of adaptations by plants to use animals as dispersal agents. For example, mistletoe seeds are eaten by birds which then fly to another tree. The seeds are adapted to survive the digestive tract, and are subsequently deposited on the tree where the bird lands, which is usually a suitable tree for growth. Such assisted dispersal can lead to dispersal distances that would be impossible to achieve otherwise.

Most vector-associated dispersal regimes have evolved over hundreds of generations. Recently, however, many species of both plants and animals have serendipitously become associated with novel and extremely efficient dispersal vectors. Species associated with humans have always been dispersed in concert with human movements. However, the last few generations of humans have seen an exponential increase in the rates of movement around the globe. Many terrestrial and marine species have been spread at unprecedented rates through ship ballast and packing materials. In addition, ornamental plants and agriculturally associated species are deliberately transported from one location to another by humans, at distances and rates that would be impossible for each species to accomplish under its own power. Hundreds of species involved in these accidental experiments in dispersal and evolution have benefited tremendously, becoming the world's invasive species. Species such as cheatgrass in North America, *Caulerpa taxifolia* (an alga) in the Mediterranean, and the Nile Perch in Africa have successfully outcompeted hundreds of native species, often driving them to extinction.

Dispersal Patterns

There are several terms associated with dispersal patterns. Dispersal distributions, or dispersal curves, are frequency distributions of the proportion of individuals moving different distances; dispersal kernels are probability density functions used in modeling dispersal. Because a dispersal kernel describes the probability of a seed landing within a particular region, dispersal curves can be calculated from dispersal kernels. Dispersal modeled with a simple diffusion equation (a Gaussian kernel) generates a normal distribution curve, due to the random Brownian motion assumed in such simple diffusion models. Though accurate for some species, most species appear to have leptokurtic, or long-tailed dispersal curves.

The shape of a dispersal curve drastically affects estimated rates of population expansion, with normal curves having lower expansion rates than dispersal curves with longer tails. The fatter the tail of the distribution, the greater is the speed of the range expansion. Accurately estimating the shape of the curve is important, for example, in predicting spread rates of spread of the emerald ash borer. Rates and pattern of spread are often consistent with simple diffusion; however, some infestations in Michigan do not spread via simple diffusion, but have a higher frequency of long-distance dispersal events. If all control efforts and spread rate predictions are based on simple diffusion, management programs will be unprepared for long-distance dispersal events and the efficacy of management efforts will be greatly diminished.

If the tail is exponentially bounded, as in curves with moderately fat tails (i.e., a Laplace distribution) (**Fig. 1**), then models predict that the rate of expansion of a population will remain constant. However, in curves with unbounded tails (very fat tails, i.e., a Weibull distribution), the rate of expansion of a population can actually increase over time. Because different dispersal curves can drastically affect estimates of population expansion, it is important to choose appropriate dispersal curves when attempting to predict species' population dynamics.

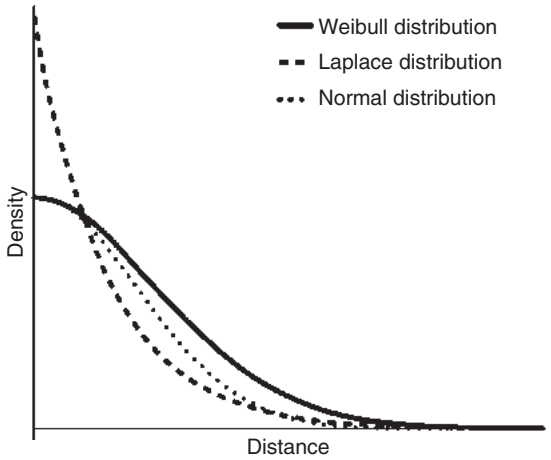


Fig. 1 Normal curves are drawn in each graph for comparison. The Laplace distribution is more peaked than the normal curve, and leptokurtic. The Weibull distribution shown here has a fatter tail than both the normal and the Laplace distribution. Graphs by Hyrum Paulsen.

Modeling Dispersal

A simple model of a species consists of one population, infinite in size, randomly mating, with no immigration or emigration. This idealized situation is never found in nature, though populations can approach this equilibrium. In real life, species are divided into populations that are subdivided to some extent, with dispersal occurring between the subdivisions at varying degrees. However, the idealized situation is often a good starting point for modeling species dispersal. Models investigating the effects of dispersal on population dynamics use several different approximations of natural situations. One of the simplest models of dispersal is a model proposed by Levins in 1969, one of the first metapopulation models. This model describes the colonization and extinction dynamics of sites under different rates of colonization (dispersal) and mortality. An individual produces offspring which then disperse to other sites. When a propagule reaches an empty site, it occupies it completely, and any propagules reaching that same site are subsequently eliminated. This model is termed a metapopulation model because it deals with colonization and extinction of sites. Though originally formulated as entire sites that are either colonized or not colonized, it can be viewed as a collection of sites where each site is the size of a single individual. In this way, a simple equation can be used to model the spread of a population through dispersal:

In this equation, the basis of the Levins model,

$$\frac{\partial s}{\partial t} = cs(1 - s) - ms$$

s represents the proportion of sites colonized, c is the rate of propagule production, and m is the mortality rate. Though not spatially explicit, spatial relationships are implied in this model by having a proportion of sites either available or not available. Some of the assumptions this model makes are that propagules are distributed randomly, and that each propagule can occupy any space in the habitat. A proportion of sites will be filled until the population reaches an equilibrium, determined by the relationship between the rates of mortality and propagule production. The population will grow at a rate that increases with lower mortality rates and increased propagule production. As long as mortality rates are less than propagule production, the population can persist. Though extremely simplified, this model has led to some interesting predictions about population dynamics. Because the final density of individuals is dependent on mortality rate in addition to propagule production, it will be impossible for a species to occupy every suitable patch in the population's habitat. When one individual dies, the point it occupied is empty until recolonized by another propagule. Thus, the population reaches an equilibrium where a certain percentage of sites are empty, rather than completely filling the entire habitat.

Though this model yields some interesting predictions about the way population dynamics depend on propagule production and mortality rates, one of the drawbacks of this simple approximation of dispersal is that propagules in real life cannot occupy any unoccupied site in a habitat. Real dispersal is restricted in distance, with a large spatial component that is only implied in the basic Levins model. The spatial component of the model is unrealistic, being essentially infinitely large and infinitely accessible.

To observe the effects of dispersal on population dynamics in more realistic situations, greater spatial detail is required. Incorporating spatial information into a model of dispersal is relatively simple by using a cellular automaton or lattice model. This type of model does not have an explicit mathematical solution but can be evaluated through simulations that can lead to predictions about population dynamics based on certain assumptions. In this type of model, individuals are placed on a grid of polygons. Each individual has a certain chance of mortality and of producing propagules. A propagule can travel a certain distance away from its current position, and its direction can be determined by either having the propagule move through a side or corner of the parental polygon. The propagule can be restricted to land within a certain area or not restricted at all. Again, any propagule that lands in an occupied site is lost. Parameters in these models are relatively simple to modify, and any combination of

requirements can be included. Different dispersal distribution curves, rates of long-distance dispersal, habitat extinction, propagule production rates, and mortality rates can all be included. However, more complex models will require more computational power, and they may not be as applicable to multiple species and/or situations as are more general models. One of the interesting predictions made by relatively simple cellular automaton models is that the distributions of individuals will be aggregated, even in a homogenous environment. The amount of aggregation of individuals depends on the average dispersal distance, the mortality rate, and fecundity. Some of these models have demonstrated that patchy spread is likely during population expansion simply through the stochastic nature of mortality and propagule dispersal distances. This is an important prediction, as patchiness need not be the direct result of unsuitable habitat, but instead can be merely the product of different dispersal patterns. Cellular automata models are often used in epidemiology, and have also been applied to species invasion dynamics.

Cellular automaton models are accurate if space, time, and population dynamics are best represented as discrete variables. In a species where many parameters are known, it is possible to reach a high degree of specificity in the simulations, which may be useful when trying to predict what may occur for a particular species under different scenarios. Less-specific models, though less applicable to a single species, can lead to more generalized predictions about the effects of dispersal on population dynamics.

Often, it is convenient to model a population or set of populations more generally, using mathematical equations. One such model involves a reaction-diffusion equation, describing the rate of change of the number of individuals in a population (N) over time (t):

$$\frac{\partial N}{\partial t} = rN\frac{(K-N)}{K} + D\frac{\partial^2 N}{\partial x^2}$$

This model consists of two parts: the reaction portion of the equation (which describes how the population acts in the absence of dispersal) and the diffusion portion (which is a partial differential equation that describes the movement of individuals in the population). The reaction portion can be as simple as the logistic equation for growth of a population, shown here immediately to the right of the equals sign, where r is the intrinsic rate of increase, and K is the carrying capacity. Other models of population growth can be used as well, such as adding an Allee effect, where propagule production rate declines at a lower than expected rate at low densities (e.g., when a mate is difficult to find). The diffusion portion of the equation is usually the mathematical representation of simple passive diffusion

$$\frac{\partial}{\partial t}N(x,t) = D\frac{\partial^2 N}{\partial x^2}$$

where D is the diffusion coefficient, t is time, N is the number of individuals, and x represents space. The diffusion portion of the equation can be more complex, incorporating directional movement, changes in velocity, or interactions among individuals. Increasing complexity is not always necessary to make meaningful predictions, as the simplest version of this equation matches observations seen in mark–release–recapture studies in several animals. In models of invasion dynamics, a reaction-diffusion model produces waves of invaders that advance at a rate dependent in part on the rate of increase of the population. This type of equation is most appropriate when the environment is homogenous, all individuals have similar dispersal patterns, and reproduction/dispersal occurs constantly. Such models generally produce a smooth traveling wave front with a linear rate of spread.

To incorporate the discrete reproduction events (and hence dispersal events) frequently seen in plants and animals, it is appropriate to use an integrodifference equation. Integrals allow each moment in time to be dependent on the previous moment in time, thus incorporating discrete time intervals into the model, as opposed to differential equations, which assume that time is continuous. Integrodifference models consist of two main parts: a dispersal kernel (a function that describes the dispersal patterns of a population) and a density function.

$$N_{t+1}(x) = \int_{-\infty}^{+\infty} k(x-y)f[N_t(y)]\mathrm{d}y$$

In this equation, $k(x-y)$ represents the dispersal kernel, or an equation that describes the movement of a propagule from the natal territory, x, to the final breeding place, y. The density function describes the density of individuals at the location y and time t. The various dispersal kernels used with these types of models are often derived from empirical data. There are many different dispersal kernels that can be incorporated into these models, including long-tailed (leptokurtic) kernels and other non-normal dispersal distributions. When production of propagules is not continuous, as is true in many animals and plants, this type of integrodifference model will better approximate reality than a reaction-diffusion model. In contrast to reaction-diffusion models, integrodifference models can show an accelerating rate of spread over time if long-distance dispersal is relatively common. Integrodifference models are more consistent with observed patterns of range expansion than reaction-diffusion models because of the accelerating rate of spread often observed. The waves of invasions seen in these models have smooth fronts, similar to the reaction-diffusion models. Space is still represented as continuous in these models, and spatial relationships are implicit, not explicit as in cellular automata/lattice models.

There are many other models that can be used to investigate the effects of dispersal on population dynamics. Stratified diffusion is a variant of the reaction-diffusion equation, where a proportion of individuals are assumed to travel long distances. Stratified diffusion models have accelerating rates of spread, similar to the integrodifference models. Metapopulation models are very applicable to modeling dispersal patterns and rates of spread, and are also applied to other problems in ecology such as persistence of groups of populations in specific spatial arrangements under various scenarios of habitat destruction. They are commonly associated with biogeographical studies. A variation on integrodifference equations that includes stochastic population dynamics (nonlinear integrodifference models) shows waves of invasion that are patchy, with variable spread.

Measuring Dispersal

Direct Measures

Generating an accurate picture of the entire process of dispersal in a species involves detailed demographic analyses in addition to tracking emigrating and immigrating individuals. In order to know what demographic parameters drive effective dispersal, it is important to know how many individuals leave, survive the exploratory process, and breed successfully in the new area. Ideally, all parameters of dispersal should be quantified. However, because the dynamics of a population are directly driven by effective dispersal, it may be unnecessary to conduct detailed studies of each stage of dispersal, depending on the particular goals of the researcher.

Mark–recapture methods and demographic analyses can assist in the estimation of many dispersal-related parameters, and though the route traveled by the individual captured in a new patch is often unknown, it is still possible to gain an estimate of immigration and emigration rates. Many studies focus on relatively local effects of dispersal, studying population dynamics in a few interconnected populations that are spatially tractable. These studies involve either mark–recapture methods, genetic methods, seed traps (for plants), or radio- or satellite-tracking methods. Animals and seeds can both be marked using tags, paint, or dyes. Tracking methods, such as by radio telemetry or satellite, show great promise for obtaining detailed information on dispersal patterns, especially on the tail of the dispersal curve. Genetic methods are becoming popular as well, because they can detect effects of very low rates of dispersal over long distances. Each method has advantages and disadvantages, and all these methods have assumptions and uncertainties associated with them, which must be taken into account when analyzing data and estimating dispersal curves.

Collecting data from the tail of the dispersal curve can be difficult, either hampered by the difficulty of maintaining sampling densities, or due simply to the rare and stochastic nature of long-distance dispersal events. The importance of long-distance dispersal in estimating the spread of populations was highlighted in scientific literature when the rate of post-Pleistocene expansion of treesin Europe estimated with models neglecting long-distance dispersal could not account for the rapid expansion rates observed as the glaciers receded. Incorporating long-distance dispersal by modeling spread with leptokurtic dispersal curves matched the estimated rates of spread more closely. Unfortunately long-distance dispersal events are extremely difficult to measure empirically, and hence estimating them has since received much attention.

For animals, one way of estimating dispersal patterns involves marking and releasing animals, then observing the animals when they are collected, usually during an annual harvest. In the case of mark–harvest methods, animals are only viewed twice, once during the marking process, and once when harvested. This type of data may be useful for estimating mortality rates associated with movement from one site to another if it is possible to assume that the animals in question always return to either the original marking site or to the final capture site. Prior knowledge of movement patterns is important; mortality cannot be estimated if a significant number of animals disperse outside of the sample area. If multiple mark–capture episodes are accomplished in one season, it is possible to estimate probabilities of survival and movement for a specific area. However, no models currently can estimate dispersal from one area to another using this type of data.

In a similar method, animals are marked, released, then re-sighted or recaptured and released again. Animals may be sighted multiple times with this method, and with a robust sampling design, immigration and temporary emigration rates can be estimated. If this type of method is employed on multiple sites, with site-specific markers, immigration and emigration probabilities in addition to transition rates can be estimated. If possible to employ, this type of design is quite useful, as it provides data necessary for estimating population dynamic parameters associated with dispersal. Long-distance dispersal events are difficult to detect with this method, due to the stochasticity associated with the occurrence and detection of such events.

Seed dispersal is often measured using seed traps to capture seeds at varying distances from the source. Seed traps usually involve pit traps or sticky traps placed in or near the ground. To identify seeds' origins, individual fruits can be marked, a chemical tagging method can be used, or a rare genetic variant can be used as a marker. The most common method is to measure the densities of seed deposited at various distances from a source. Because individual plants are not identified when only density of seeds can be recorded, likelihood methods are used to model dispersal curves. Seed traps work well for estimating dispersal curves near the source, but as distance from the source increases it becomes more and more difficult to detect dispersal events. If enough traps are used, long-distance dispersal events can be detected; however, such events will be rare, and their detection will be dependent on the resources available to the researcher.

Radio telemetry and satellite tracking provide excellent data, when practical. Such studies have documented that long-distance dispersal events are more common than estimates from mark–recapture methods suggest. Most studies involve large- to medium-sized animals, including marine mammals. Invaluable information about the long-distance travels of these animals has been collected, including information about movements of some seabirds. Ideally, a large proportion of a population could be followed individually, and detailed analyses made of their movements. In order to accomplish this, the radio or satellite transmission units should not inhibit movement or survival, and the batteries should be strong enough to allow signal detection at a distance for a long period of time. As technology advances, smaller tags can be used. For example, very small radar tags have lately been adapted for use on bumblebees, showing promise for generating detailed dispersal data for larger insects.

Indirect Methods

Genetic methods hold promise for estimating dispersal patterns, though it is important to remember that genetic methods only measure effective dispersal, and not dispersal of individuals that did not successfully breed in the new population. In addition, for

organisms with motile gametes, genetic patterns will likely reflect the movement of gametes among populations as well as the movement of diploid individuals. Most genetic methods involve collecting DNA from immature or mature individuals, then analyzing the DNA to try and identify the origin of a particular individual. In addition, if a dispersed individual has the same genetic signature as the individuals in the new population, the dispersal event will be undetected; this becomes less likely when highly variable markers are used. Another possible drawback to genetic data is the potential for unsampled source populations to contribute to apparent gene flow estimates between two sampled populations. Also, it is difficult to generate a detailed dispersal curve using solely genetic data, especially at local distances, due to the large amounts of data that would have to be collected and the heavy expense involved. Nevertheless, genetic data enable estimation of many parameters of interest, such as historical amounts of gene flow, effective dispersal rates among differentiated populations, and dispersal rates over long distances. Genetic data are also often (but not always) easier and faster to collect than detailed demographic data. These advantages can outweigh the potential difficulties with genetic data, depending on the parameters of interest.

The first genetic estimates of dispersal were derived from Sewell–Wright's equation

$$N_e m = \frac{1}{4}\left(\frac{1}{F_{ST}} - 1\right)$$

where N_e is effective population size (an estimate of the number of individuals in an idealized population that would show the same patterns of genetic diversity or levels of inbreeding), m is migration (dispersal) rate, and F_{ST} is a measure of population structure. This equation can be used to estimate dispersal among populations using DNA sequence data, DNA markers of variable length, or allozyme (protein) markers. However, more recently the use of this equation to estimate dispersal rates has come into question based both on unlikely assumptions made when calculating F_{ST}, and on the applicability of the above equation to natural populations as opposed to idealized populations. Some of the assumptions made include equal and constant population sizes, and equilibrium between gene flow and genetic drift (stochastic variation in genetic frequencies over time). In addition, F_{ST}-based methods do not distinguish between historical and contemporary gene flow. In some cases, however, as in well-established, large populations, F_{ST}-based methods may be sufficient for estimates of dispersal rates.

Parentage analyses and assignment methods are both techniques that use genetic marker data to estimate dispersal. They both assume that source populations are discrete and that there is no genetic linkage disequilibrium (nonrandom associations of genetic marker variants due to inbreeding or chromosomal proximity) among the different markers used. Parentage analyses are based on multilocus genotypes, and can generate data about local animal movement and both seed and pollen dispersal. However, parentage analyses require extensive sampling to ensure that all possible parents in the source area have been included in the study. If a parent present in source populations is not sampled, dispersal from an unsampled (possibly quite far away) population may be inferred, potentially altering dispersal estimates. Depending on the situation, parentage analysis can be expensive enough to outweigh the potential advantages of genetic analyses over demographic studies.

Assignment methods, on the other hand, use allele frequencies, or frequencies of different variants of genetic markers, to predict from which source a particular individual came. This means that exhaustive sampling of source populations is unnecessary. It is still desirable to have representatives from all possible sources. In order to distinguish sex-biased dispersal, sex-specific markers must be used. Assignment methods assume discrete source populations and no linkage disequilibrium, and often assume equilibrium of populations. However, methods that enable dispersal estimates when populations are not in equilibrium are being developed.

A method that has come into use recently for many questions in genetics involves Bayesian analyses. Bayesian methods can be used in nonequilibrium situations, such as during range expansions, with high levels of inbreeding, or with unequal population sizes. Information known about the populations in question is entered in the analysis in the form of prior probability distributions. This information, commonly known as a prior, is basically a guess about how the populations might act based on data already available, such as experimental data. If no information is already known about the populations, an uninformative prior can be used. Then the genetic data is used, in conjunction with the prior, to calculate posterior probabilities of the data using a maximum likelihood algorithm, given the parameters currently in place in the maximum likelihood model. Markov chain Monte Carlo resampling is then used to explore parameter space and find values that optimize the fit of the parameters in the model to the data. The accuracy of detecting dispersal events with these methods is still being explored. Factors affecting accuracy of dispersal detection include the level of genetic diversity in the populations, amount of dispersal occurring among populations, how many genetic markers are used, and the level of variability in the markers themselves. Depending on the population structure in the system in question, these methods may be equally viable both for detection of local dispersal and long-distance dispersal events.

As mentioned in the beginning of this article, dispersal in the context of population dynamics usually refers to the dispersal of diploid individuals. However, especially when genetic methods are used, movement of haploid gametes can affect dispersal estimates. Male and female gametic dispersal patterns often differ, and if not accounted for, can skew dispersal estimates from genetic data. Nuclear genetic markers come from both paternal and maternal sources, due to the combination of nuclear genetic material during fertilization. If gametic gene flow is significantly different than diploid dispersal patterns, such as with pollen in many plants, care must be taken not to confuse gametic patterns of gene flow with movements of diploid individuals. In plants, for example, the male pollen often travels farther than seeds, especially if the species in question is wind-pollinated. In order to distinguish the dispersal of diploid seeds from haploid pollen, sex-specific genetic markers must be used. In plants, the chloroplast genome is generally (but not always) maternally inherited, and comparing patterns of genetic differentiation between the

chloroplast and the nucleus can be used to estimate pollen versus seed dispersal patterns. If only seed dispersal is of interest, it may be sufficient to focus on chloroplast genetic markers, if the chloroplast is indeed maternally inherited in the species in question. Similarly, markers based on mitochondrial DNA will only show dispersal patterns of the female in animals. If dispersal patterns of both males and females are desired, both nuclear and mitochondrial markers must be used.

Because different analysis techniques have different strengths, weaknesses, and assumptions, it is important to consider what the goals of a given study are, and what methods are best suited to the questions at hand. Genetic methods hold promise for dispersal estimations, as data can be gathered relatively quickly and with less cost than demographic data. However, depending on the species and/or populations under consideration, estimates may require more genetic data than are currently available to optimize the parameters of a genetic analysis technique. If this is the case, then the cost of genetic analyses could equal or exceed the costs of demographic data collection. In addition, some data must be gathered through observation, such as detailed movement patterns among sites or populations and breeding success rates. Other data can best be estimated using genetic data, such as historical patterns of gene flow or long-distance dispersal events, depending on the organism in question. As techniques for measuring dispersal and its consequences improve, we will be better able to predict the survival, extinction, range expansion, and range contraction of populations and species. These predictions will improve estimates of the effects of habitat destruction and fragmentation on populations, a growing concern at multiple scales worldwide. In addition, accurate predictions of range expansion and contraction rates based on models of global warming will enable us to prepare for the possible effects of a rapidly changing climate on both marine and terrestrial species.

See also: Behavioral Ecology: Orientation, Navigation, and Search. Ecological Data Analysis and Modelling: Modeling Dispersal Processes for Ecological Systems. General Ecology: Migration and Movement

Further Reading

Bullock, J.M., Kenward, R.E., Hails, R.S. (Eds.), 2002. Dispersal Ecology. Oxford: Blackwell Science.
Cain, M.L., Milligan, B.G., Strand, A.E., 2000. Long-distance seed dispersal in plant populations. American Journal of Botany 87, 1217–1227.
Clark, J.S., Lewis, M., Horvath, L., 2001. Invasion by extremes: Population spread with variation in dispersal and reproduction. American Naturalist 157, 537–554.
Clobert, J., Danchin, E., Dhondt, A.A., Nichols, J.D. (Eds.), 2001. Dispersal. Oxford: Oxford University Press.
Estoup, A., Beaumont, M., Sennedot, F., Moritz, C., Cornuet, J.M., 2004. Genetic analysis of complex demographic scenarios: Spatially expanding populations of the cane toad. Bufo marinus. Evolution 58, 2021–2036.
Kot, M., Medlock, J., Reluga, T., Walton, D.B., 2004. Stochasticity, invasions, and branching random walks. Theoretical Population Biology 66, 175–184.
Levin, S.A., Muller-Landau, H.C., Nathan, R., Chave, J., 2003. The ecology and evolution of seed dispersal: A theoretical perspective. Annual Review of Ecology Evolution and Systematics 34, 575–604.
Nathan, R., Muller-Landau, H.C., 2000. Spatial patterns of seed dispersal, their determinants and consequences for recruitment. Trends in Ecology and Evolution 15, 278–285.
Nathan, R., Perry, G., Cronin, J.T., Strand, A.E., Cain, M.L., 2003. Methods for estimating long-distance dispersal. Oikos 103, 261–273.
Okubo, A., Levin, S.A. (Eds.), 2002. Diffusion and Ecological Problems: Modern Perspectives. New York: Springer.
Olin, E., Rhodes, J., Chesser, R.K., Smith, M.H. (Eds.), 1996. Population Dynamics in Ecological Space and Time. Chicago: The University of Chicago Press.
Paetkau, D., Slade, R., Burden, M., Estoup, A., 2004. Genetic assignment methods for the direct, real-time estimation of migration rate: A simulation-based exploration of accuracy and power. Molecular Ecology 13, 55–65.
Simmons, A.D., Thomas, C.D., 2004. Changes in dispersal during species' range expansions. American Naturalist 164, 378–395.
Skalski, G.T., Gilliam, J.F., 2003. A diffusion-based theory of organism dispersal in heterogeneous populations. American Naturalist 161, 441–458.
Tilman, D., Kareiva, P. (Eds.), 1997. Spatial Ecology, the Role of Space in Population Dynamics and Interspecific Interactions. Princeton, NJ: Princeton University Press.

Dominance Hierarchy

Hilde Vervaecke, Odisee University College, Sint-Niklaas, Belgium
Jeroen Stevens, Royal Zoological Society of Antwerp, Antwerpen, Belgium and University of Antwerp, Antwerpen, Belgium

Glossary

Agonistic dominance In an agonistic dominance relationship, the dominant animal has the power to limit the behavior of a subordinate to some extent by means of aggression or fighting abilities. Agonistic dominance is expressed in the outcome of agonistic interactions.

Agonistic interaction Agonism refers to the combination of aggression and submission. Indices of winning or losing conflicts are frequently used as operational measures of agonistic interactions such as attacks or threats and withdrawals after an attack, or the spatial yielding of one individual to another.

Cardinal dominance rank Cardinal dominance rank quantifies rank distances between individuals, based upon their relative probability of winning or losing agonistic interactions.

Competitive ability Competitive ability reflects the capacity of an individual to obtain access to limited resources (usually food access is measured).

Dominance hierarchy It is the network of dominance relations among all the dyads in a group. A dominance relationship indicates a consistent asymmetry in the probability of winning a contest of one individual over another.

Dominance position Dominance position is based on the outcome of the dominance interactions, animals occupy an ordinal position in the hierarchy, being high, mid or low ranking, or more specifically number one, two, three, etc.

Formal dominance This is characterized by ritualized communication signals and greeting rituals of which the direction among dyads is consistent and does not vary with social context.

Linear hierarchy The network of dominance relationships can result in a figurative ladder with the highest ranking animal on top and the lowest ranking animal on the lowest rung. In a linear hierarchy one individual dominates all the other individuals, followed by the second ranking animal that dominates all the others except for the highest ranking individual, etcetera. The relationships are transitive: A dominates B, B dominates C, and A dominates C.

Ordinal dominance rank The rank order of individuals based on their probability of winning or losing an agonistic interaction.

The Concept of Dominance

The concept of dominance was developed to help to describe, explain and predict social relationships. It can be defined as "an attribute of the pattern of repeated, agonistic interactions between two individuals, characterized by a consistent outcome in favor of the same dyad member and a default yielding response of its opponent rather than escalation. The status of the consistent winner is dominant and that of the loser subordinate" (Drews, 1993). Individuals rarely engage in costly fights and contests are generally resolved quickly by one retreating individual.

A *dominance relationship* indicates a consistent asymmetry in the probability of winning a contest of one individual over another. Within each dyad, the *dominance status* of each individual refers to its relative higher or lower dominance rank. The *dominance hierarchy* is the network of dominance relations among all the dyads in a group. The *rank order* refers to each animal's position in the hierarchy, that is, number one, two, three, etc. based on the outcome of the agonistic interactions. In a linear hierarchy one individual dominates all the other individuals, followed by the second ranking animal that dominates all the others except for the highest ranking individual, etcetera. The hierarchy is completely linear if for every dyad A and B, either A dominates B or B dominates A. For every triad of individuals A, B, and C, the following relationships will be transitive and not linear, meaning: if A dominates B and B dominates C, then A dominates C. This results in a figurative ladder with the highest ranking animal on top and the lowest ranking animal on the lowest rung. The index of linearity in such a system will be consistent. In a nonlinear hierarchy, intransitive dominance relationships can be found: A dominates B, B dominates C, but C dominates A, resulting in an inconsistent linearity.

The mechanisms that can generate the linearity of hierarchies are learning by experiencing dyadic wins and losses, by watching others fight (Chase and Seitz, 2011), and by assessing the fighting ability and behavioral dispositions. Hemelrijk (1999) studies the characteristics of the hierarchy by manipulating simple parameters in a computer model such as frequency or intensity of aggression in order to detect the emergent phenomena. Interactions on a lower level clarify the macrostructure on a higher level and shifts the focus from the internal complexity of individuals (learning, fighting abilities, etc.) to social interaction patterns.

The study of dominance hierarchies has received considerable interest in behavioral studies. Its value lies in its' *general predictive value* of the outcome of future, not just agonistic interactions, but other social interactions as well (Syme, 1974). Based on the agonistic hierarchy, you can generally more broadly predict the outcome of contests and the direction of other social interactions.

 doi:10.1016/B978-0-12-409548-9.10565-2

Different Types of Dominance Relationships

In general, the descriptive, predictive and explanatory value of the dominance concept has greatly improved by distinguishing between *different types of dominance relationships*. de Waal (1989) made a distinction between three types of dominance: agonistic dominance, formal dominance, and competitive ability.

In *an agonistic dominance relationship*, the dominant animal has the power to limit the behavior of a subordinate to some extent by means of aggression or fighting abilities (Nöe *et al.*, 1980). Agonistic dominance is expressed in the outcome of agonistic encounters, and indices of winning or losing conflicts are frequently used as operational measures. Since conflicts may also be resolved by deference of the loser without escalation (Drews, 1993), the spatial yielding of one individual to another may provide a good operational measure of submission or deference. When an aggression is ignored or not followed by submission of the target individual, the interaction may not express a mutually acknowledged dominance relationship. Submissive interactions are therefore usually considered as better indicators of a dominance relationship.

Formal dominance is characterized by ritualized communication signals and greeting rituals of which the direction does not vary with social context. When the agonistic dominance relationship is accepted by the subordinate, aggressive conflicts are few and the subordinate acknowledges the higher dominance status of the other by showing formalized submissive signals. This implies that the formal and agonistic dominance relationships coincide. Unidirectional signals of subordination such as the teeth baring in rhesus macaques, the bowing and pant-grunting in chimpanzees or unidirectional dominance signals such as mock-biting in stump-tailed macaques are reliable ritualized expressions of formal rank (de Waal and Luttrell, 1985; de Waal, 1989). These behaviors are unidirectional and multicontextual, are expressed by different individuals, and covary with other selected measures of agonistic rank. The *competitive ability* reflects the capacity of an individual to obtain access to limited resources (usually food access is measured). The derived rank order does not necessarily correspond to the agonistic dominance rank. The ability and motivation to compete may vary according to the resource that causes competition (Syme, 1974). Temporal variation in competitiveness implies that an individual not always shows the same tendency to use competitive ability. It can show a certain degree of respect for possessions of others or can allow others to share a resource.

Van Hooff and Wensing (1987) suggest several criteria that a behavioral variable should meet in order to justify the adoption of a dominance model: the behavior should allow for a linear ordering in the group; it should be expressed in most dyads as a predominantly unidirectional interaction; and, it should be expressed not just in a few but in most of the relationships in the group expressed in its' "coverage" or proportion of dyads in which the relevant behavior is expressed at least once.

Structural Aspects of the Dominance Hierarchy

We can describe the structural aspects of dominance hierarchies at dyadic, group and individual level (overview Langbein and Puppe, 2004; de Vries, 1998; de Vries *et al.*, 2006; Flack and de Waal, 2004; Van van Hooff and Wensing, 1987).

At dyadic level, dominance characteristics can be expressed as the number of unknown (or zero or blanco) dyads: that is, number of dyads in which the two members of the dyad have not been observed to perform any agonistic interaction towards each other; the number of one-way dyads, that is, the number of dyads in which the behavior is shown in one direction only irrespective of the frequency of interaction within the dyads; the number of two-way dyads, that is, dyadic dominance interactions occurred at least once in both directions (from A to B and from B to A); the number of tied dyads in which the two members of the dyad have performed an equal number of agonistic interactions towards each other; and the number of circular dyads in which an intransitive relationship exists among A, B, and C.

At group level, the ordinal rank order can be obtained by reorganizing data on the outcome of dyadic agonistic interactions in actor-receiver matrices, with, in order to put the individuals in a linear rank order based on the minimization of the number of inconsistencies in the matrix and on the minimization of the total strength of the inconsistencies in the matrix (I&SI-method). The cardinal dominance rank quantifies the rank distances between the individuals, based upon their relative probability of winning or losing dominance interactions. The Clutton-Brock's Index and the David's scores provide a measure of individual competitive success in absolute terms and in relation to the wins and losses of all other individuals in the group. The relative strengths of the other individuals are taken into account and defeating a high-ranking animal is weighted heavier than defeating a low-ranking one, minimizing the chance of a highly dominant individual skewing the index. Alternatively, a Bayesian method can be used to provide cardinal scores with confidence intervals. The Elo rating method calculates the hierarchy by taking the sequential order of interactions into account and updates the rank of individuals as interactions occur. This method takes into account that the outcome of an agonistic interaction can be influenced by prior interaction success or failure.

Characteristics of a dominance hierarchy can be described by referring to the strength of the linearity, steepness and directional consistency. Linearity in a set of binary dominance relationships depends on the number of established relationships, and on the degree to which these relationships are transitive. If none of the relationships is undecided and all relationships are transitive, a perfect linear order exists. Linearity is expressed in Landau index (h) based on the number of dominated animals or the Kendall index K based on the concept of circular triads or the improved Landau index (h') in case there are unknown relationships. A value of 1 indicates complete linearity and a value of 0 indicates that each individual dominates an equal number of other individuals. The steepness or strength of a hierarchy refers to the size of the absolute differences between adjacently ranked individuals in their overall success in winning dominance encounters (i.e., dominance success) and can be inferred from cardinal dominance ranks. When these differences are large the hierarchy is steep; when they are small the hierarchy is shallow.

The directional consistency index (DCI) gives the frequency with which the behavior occurred in its more frequent direction (H) minus the number of times the behavior occurred in its' least frequent direction (L), divided by the total number of times the behavior occurred. The index varies from 0 (complete bidirectionality) to 1 (complete unidirectionality). The stability of a hierarchy can be tested by calculating the rank correlation coefficient between hierarchies over time and by a matrix correlation test of two matrices of each time period with agonistic interactions upon which the hierarchy is based.

At the individual level, dominance can be described as the dominance index DI_{dom} with the ratio of animals which are dominated by an individual in relation to all animals with which it has interacted, the dominance index DI_{AI} being the sum of the number of wins minus defeats to the number of all interactions, the ordinal rank number placing each individual into an ordinal rank according to the wins and losses of dyadic dominance encounters, the cardinal rank number quantifying rank distances between individuals based upon their relative probability of winning or losing dominance interactions, the agonistic index AGI referring to the number of agonistic interactions per individual and time unit, the aggressive index ARI describing the number of initiated agonistic interactions per individual and time unit.

Linear Dominance Hierarchies

In several animal species like dolphins, chimpanzees, hyenas and baboons the hierarchies are not strictly linear (Chase and Seitz, 2011) and in some species males and females form separate hierarchies, such as in mountain goats (*Oreamnos americanus*), and bighorn sheep (*Ovis canadensis*). Finding a linear dominance hierarchy requires sufficient observation effort. de Vries *et al.* (2006) warn that one should be cautious in interpreting the results with respect to observational zeroes in which some dyads have had no dominance interactions. Absence of linearity can only be concluded in case of sufficient observation effort and unbiased sampling. In some cases, a subordinate animal keeps a safe distance from a dominant one because there is a clear dominance–subordination relationship that has priorly been aggressively established (Hemelrijk, 1999). In this case it may be advised to include dyadic "avoiding at a distance" interactions to reduce the number of dyads with missing values. Alternatively, two animals may stay at a relatively large distance from each other and have an unresolved dominance relationship. In case two animals are regularly seen in each other's neighborhood, without having any dominance interactions with each other and without avoiding each other at a distance, these animals show that a clear dominance–subordination relationship between them is absent. In certain contexts, a dominance hierarchy can remain hidden and be rarely expressed in agonistic interactions. At periods of resource scarcity and competition, the asymmetry in agonism and in resource access can suddenly become overtly expressed, for instance in female European badgers (*Meles meles*) in case of breeding competition over males (Hewitt *et al.*, 2009). Bottleneck periods of scarcity have shaped the potential for hierarchy formation of many group living species and are a part of their environment of evolutionary adaptiveness. Animals that show few asymmetric hierarchical interactions at times of low competition, may express a high degree of agonistic asymmetry at times of scarcity. This may explain ambivalent claims in popular literature on certain species (e.g., dogs, wolves and cats) as either lacking or possessing a clear hierarchy.

Dominance Styles

The term "dominance style" was initially used to describe patterns of agonistic behavior by de Waal and Luttrell (1989), and later was broadened to include the covariation of traits related to conflict or conflict management (Flack and de Waal, 2004). In macaques, a despotic dominance style was linked to asymmetrical patterns of agonistic behavior, low rates of reconciliation, and a strong kin bias in affiliation. Macaques with a tolerant dominance style do not show a strict inheritance of dominance rank, they show low rates of severe biting, tolerance around limited resources, little agonistic behavior in response to the approach of others, maternal tolerance for infant handling, and triadic male–infant interactions (Thierry, 2000) (**Fig. 1**).

Stricto sensu, the dominance style describes the asymmetry in competition in terms of the frequency of aggression of the dominant individuals and the submission of the subordinates (de Waal and Luttrell, 1989). In a group of individuals in which a perfect linear dominance hierarchy exists and for which the wins in every dyad are completely unidirectionally distributed, the steepness score of the hierarchy will be high, indicating a steep and despotic style. Where cardinal rank distances are large the hierarchy is steep and despotic; where they are small it is shallow and egalitarian. The lower the unidirectionality, the shallower and more egalitarian the dominance style (de Vries *et al.*, 2006). This dominance style depends on socioecological factors, such as the distribution of the resources, the type and intensity of inter- and intragroup competition and the adaptive value of group living, and will determine the rank-dependent inequality among group members with regard to costs and benefits of group living (de Waal, 1989; de Waal and Luttrell, 1989). Within a group, dominant and subordinate animals reach some equilibrium between exploitation and cooperation (Vehrencamp, 1983). The higher the costs for the lower-ranking animals to leave the group, the higher the liberty of the high ranking animals to despotically claim the access to limited resources. In a despotic dominance hierarchy especially the low ranking animals may evaluate the fitness gain obtained by leaving the group. The outcome of this trade-off will be determined by the options outside of the group. A high predation pressure or a high degree of between group competition, may increase the value of remaining in a group. Whenever the outside options are limited, the subordinates will be likely to appease dominants to safeguard their membership to the group. Whenever the presence of the lower ranking animal confers certain benefits, we expect to find suppressed competitiveness by dominants and more egalitarian dominance

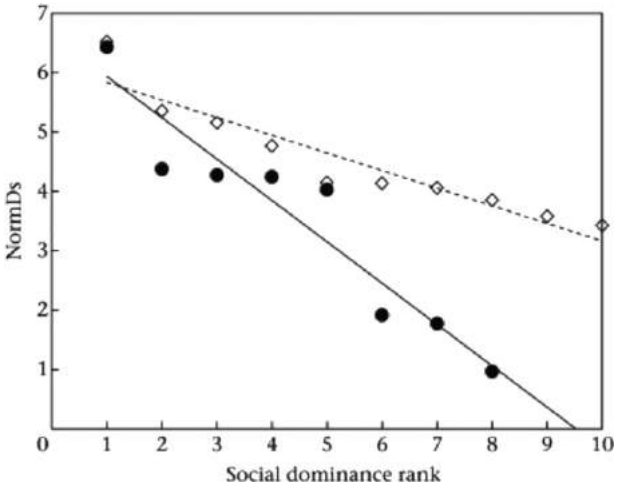


Fig. 1 Illustration of the quantification of dominance styles (Kaburu and Newton-Fisher, 2015). Dominance hierarchy steepness, indicated by the slope of the best-fit line of normalized David's score (NormDs) against dominance ranks among adult male chimpanzees of two study communities. Sonso chimpanzees (*solid line*, diamonds, $N = 8$) show a steep (despotic) hierarchy, while for M-group chimpanzees (*dashed line*, circles, $n = 10$), the hierarchy is shallow (egalitarian).

relationships. In the model proposed by Hemelrijk (1999) the differences in social behavior between despotic and egalitarian societies are the consequence of changing one simple parameter only, representing the intensity of aggression. A higher intensity of aggression increases the degree of variation in dominance values, resulting in a steeper hierarchy. A steeper gradient of the hierarchy in turn leads to several emergent phenomena, affecting cohesiveness, frequency of interaction, initial decline of aggression, rank relatedness of interaction, rank overlap between the genders and spatial centrality of dominants, etc. Between species, linear and steep hierarchies may be easier to detect among close-knit groups of social animals, but more difficult to find among individuals living in loose societies (de Vries *et al.*, 2006).

The concept of dominance style, sometimes referred to as "dominance gradient," has been used in several biological market models on behavior (Barrett *et al.*, 1999). The steepness of the hierarchy was used to predict the economic exchange and interchange of social interactions among dyads, for example, in Kaburu and Newton-Fisher (2015) who found that male chimpanzees trade grooming for agonistic support where hierarchies are steep (despotic) and consequent effective support is a rank-related commodity, but not where hierarchies are shallow (egalitarian).

Dominance Determining Factors

Several studies (review Chase and Seitz, 2011) showed that animals that are heavier and larger can dominate the lighter and smaller individuals. Some examples are weight in female American bison (*Bison bison)*; size in male semiterrestrial crab (*Neohelice granulate)*; size in cichlid fish; size in Japanese fluvial sculpin (*Cottus pollux*). Other physical attributes or signals that correlate with dominance include beak size in Greater sheathbills (*Chionis alba*), plumage characteristics in birds, urinary composition in the European lobster (*Homarus gammarus)*, facial patterns in paper wasps (*Polistes fuscatus*), comb size in roosters (*Gallus gallus*), ornament size in the swordtail fish (*Xiphophorus helleri*), UV reflectance in male wall lizards (*Podarcis muralis*).

Personality can predict the outcome of dominance interactions, such as boldness in trout (*Salmo trutta*), proactivity in rainbow trouts (*Oncorhynchus mykiss*), speed of exploration in mountain chikadees (*Poecile gambeli*), aggression, boldness and exploration in hermit crabs (*Pagurus bernhardus*), aggressiveness in sheepshead swordtail fish (*Xiphophorus birchmanni)*.

The genetic basis of dominance and aggression was examined in wild red deer (*Cervus elaphus)* and in graylag geese (*Anser anser*), as well as the tendency to be agonistically victimized in yellow-bellied marmot (*Marmota flaviventris*). A heritable component of aggressive behavior and dominance-related traits was found in many species, such as male bank voles (*Clethrionomys glareolus*), in dogs (*Canis familiaris*) and insects. The complex interaction between environmental and genetic determinants was studied in the cichlid (*Astatotilapia burtoni)*, where males can change in color and dominance several times throughout their lives depending on the available territories.

In many species, dominance will affect the *hormone levels* and vice versa. Watching fights will raise hormone levels in fish. Maternal aggression and *androgens* influence the rank-related behavior in the young. High ranking female spotted hyenas (*Crocuta crocuta*) have higher levels of androgens in the uterus resulting in more aggressive infants. Testosterone levels in the eggs are determined by maternal aggression, and affect the growth and dominance of young tree swallows (*Tachycineta bicolor*) in the nest. Kittiwake (*Rissa tridactyla*) nestlings with higher maternal androgens show increased sibling aggression and dominance. *Metabolic speed* was higher in high ranking rainbow trout (*Oncorhynchus mykiss*). Reproductive state affects dominance in wild female Hanuman langurs (*Semnopithecus entellus*).

Experience is known to play an important role in *dominance acquisition and maintenance*. An animal that has won an interaction in a previous conflict, has a higher probability to win the next conflict, known as the "winner effect." Winner effects are less common and less strong than loser effects and they last shorter. The loser effect implies that an animal that has lost a previous interaction, has a higher probability to loose the next conflict. It has been observed in several species and can last several days. The bystander effect implies that the animal that observed a conflict among two individuals, will alter its behavior in a subsequent encounter with one of the participants, in comparison to animals that did not observe the interaction.

Fitness Benefits

The costs and benefits of group living may be divided unequally among group members depending on their dominance rank. The probability of engaging in an agonistic interaction to attain or maintain dominance or of developing alternative strategies as subordinate individual can be calculated in a cost-benefit model. The success can be expressed in *fitness benefits* obtained directly or indirectly by rank-related access to resources (Smith, 1979). In groups with despotic dominance relationships, we expect to see a higher variation in reproductive success (Vehrencamp, 1983). Direct or indirect benefits of high rank may be multiple and varied including for instance food, peaceful cooperation by subordinates by the threat of punishment in the Goby (*Paragobiodon xanthosomus*), a coral-reef fish, increased growth and survival benefits in blue-footed boobies (*Sula nebouxii*) nestlings. Dominance hierarchies may have important consequences for the patterns of mating and reproduction. High ranking females can obtain a variety of sexually competitive benefits (review by Stockley and Bro-Jorgensen, 2011). In several species dominance relates to a higher female reproductive success, in primates (review Majolo *et al.*, 2012), in mice (*Mus musculus)* or rabbits (*Oryctolagus cuniculus*). In males, reproductive success has been related to dominance rank in many studies (such as in primates: Majolo *et al.*, 2012, American bison (*Bison bison*) or Japanese fluvial sculpin (*Cottus pollux)*.

See also: Behavioral Ecology: Social Behavior and Interactions. General Ecology: Communication; Dominance

References

Barrett, L., Henzi, S.P., Weingrill, T., Lycett, J.E., Hill, R.A., 1999. Market forces predict grooming reciprocity in female baboons. Proceedings of the Royal Society of London, Series B 266, 665–670.

Chase, I.D., Seitz, K., 2011. (2011). Self-structuring properties of dominance hierarchies: A new perspective. Advances in Genetics 75, 51–81.

de Vries, H., 1998. Finding a dominance order most consistent with a linear hierarchy: A new procedure and review. Animal Behaviour 55, 827–843.

de Vries, H., Stevens, J.M.G., Vervaecke, H., 2006. Measuring and testing the steepness of dominance hierarchies. Animal Behaviour 71, 585–592.

de Waal, F.B.M., 1989. Dominance "style" and primate social organization. In: Sanden, V.F. (Ed.), Comparative Socioecology. Oxford: Blackwell Scientific, pp. 243–264.

de Waal, F.B.M., Luttrell, L.M., 1985. The formal hierarchy of rhesus macaques: An investigation of the bared-teeth display. American Journal of Primatology 9 (2), 73–85. doi:10.1002/ajp.1350090202.

de Waal, F.B.M., Luttrell, L.M., 1989. Toward a comparative socioecology of the genus Macaca: Different dominance styles in rhesus and stumptailed macaques. American Journal of Primatology 19, 83–109.

Drews, C., 1993. The concept and definition of dominance in animal behaviour. Behaviour 125, 283–313.

Flack, J.C., de Waal, F.B.M., 2004. Dominance style, social power, and conflict management: A conceptual framework. In: Thierry, B., Singh, M., Kaumanns, W. (Eds.), Macaque societies: A model for the study of social organization. Cambridge: Cambridge University Press, pp. 157–181.

Hemelrijk, C.K., 1999. An individual-orientated model of the emergence of despotic and egalitarian societies. Proceedings of the Royal Society of London. Series B, Biological Sciences 266, 361–369. doi:10.1098/rspb.1999.0646.

Hewitt, S.E., Macdonald, D.W., Dugdale, H.L., 2009. Context-dependent linear dominance hierarchies in social groups of European badgers, Meles meles. Animal Behaviour 77, 161–169.

Kaburu, S.S.K., Newton-Fisher, N.E., 2015. Egalitarian despots: Hierarchy steepness, reciprocity and the grooming-trade model in wild chimpanzees, Pan troglodytes. Animal Behaviour 99, 61–71.

Langbein, J., Puppe, B., 2004. Analysing dominance relationships by sociometric methods, a plea for a more standardised and precise approach in farm animals. Applied Animal Behaviour Science 87, 293–315.

Smith, J.M., 1979. Game theory and the evolution of behaviour. Proceedings of the Royal Society of London B 205, 475–488.

Majolo, B., Lehmann, J., de Bortoli Vizioli, A., Schino, G., 2012. Fitness-related benefits of dominance in Primates. American Journal of Physical Anthropology 147, 652–660.

Nöe, R., de Waal, F.B., van Hooff, J.A., 1980. Types of dominance in a chimpanzee colony. Folia Primatologica 34, 90–110.

Stockley, P., Bro-Jorgensen, J., 2011. Female competition and its evolutionary consequences in mammals. Biological Reviews 86, 341–366. doi:10.1111/j.1469-185X.2010.00149.x.

Syme, G.J., 1974. Competitive orders as measures of social dominance. Animal Behaviour 22, 931–940.

Thierry, B., 2000. Covariation of conflict management patterns across macaque species. In: Aureli, F., de Waal, F.B. (Eds.), Natural conflict resolution. Berkeley: University of California Press, pp. 106–128.

van Hooff, J.A.R.A.M., Wensing, J.A.B., 1987. Dominance and its behavioral measures in a captive wolf pack. In: Frank, H. (Ed.), Man and wolf. Dordrecht: DRW Junk Publishers, pp. 219–252.

Vehrencamp, S.L., 1983. A model for the evolution of despotic versus egalitarian societies. Animal Behaviour 31, 667–682.

Environmental Stress and Evolutionary Change☆

B van Heerwaarden and VM Kellermann, Monash University, Clayton, VIC, Australia
AA Hoffmann, The University of Melbourne, Melbourne, VIC, Australia

Introduction

Changes in the normal types of environmental conditions to which species are exposed can lead to dramatic reductions in the fitness of organisms, and these conditions are often referred to as 'stressful' (Hoffmann and Parsons, 1991). They may be catastrophic such as droughts or mass flooding events, or subtle such as small increases in sea or air temperature. Both types of changes potentially have an enormous effect on population density and ultimately lead to changes in species distributions and community composition. However, while stressful conditions are often seen in a negative light because they decrease biodiversity, they may also have positive effects by being linked to rapid evolutionary responses; for instance, species radiations within the fossil record have often been linked to periods of environmental change, and geographic regions with a high level of animal and plant diversity are often exposed to marked fluctuations in environmental conditions that produce periods of local extinction and expansion.

One reason why stressful conditions might produce evolutionary changes is that they represent periods of intense selection. Numerous laboratory-based experiments in model organisms like *Drosophila melanogaster*, mice, and *Escherichia coli* bacteria have demonstrated that rapid evolution can occur following intense selection (Hoffmann and Parsons, 1991). There are also numerous examples of natural populations of organisms adapting under intense selection due to stressful conditions, such as the evolution of highly resistant pest populations following exposure to chemicals, and evolutionary changes in both morphological and physiological traits in birds exposed to food shortages resulting from climatic stressors. Apart from producing more intense selection, stressful conditions can also influence evolutionary rates in other ways explored further below.

The question of how stressful conditions influence the evolutionary potential of populations has become increasingly important because human activities are rapidly producing habitat fragmentation, thermal and precipitation shifts associated with global warming and increasing quantities of pollutants entering the environment. There are numerous examples of species adapting in the face of such adversity but similarly there are many examples highlighting that not all species successfully adapt (Bradshaw and Holzapfel, 2006; Merilä, 2012). The low plant diversity typical of metal-contaminated soils and the loss of endemics in areas that are becoming warmer highlight the limitations of evolutionary responses. Under many circumstances strong directional selection will lead to adaptation but why do we see an absence of a selection response in other cases, and how might evolutionary changes be promoted in such situations?

In this article we consider conditions under which stress might increase or decrease the likelihood of adaptation. We focus on current human-induced disturbances such as habitat fragmentation, the direct and indirect effects of global warming, and the effects of insecticides and pollutants. Within this framework we discuss potential genetic and ecological limits that may prevent future adaptation and what this means for species conservation. We finish by highlighting current evidence for adaptation to human-induced environmental stressors and the potential to use genes and genetically-based polymorphisms as markers of environmental change. A theme emerging from work undertaken to date is that populations and species markedly differ in their propensities to evolve. This makes predictions difficult in specific instances, but generalizations may emerge about specific groups of organisms as more data are collected.

☆*Change History:* April 2014. Van Heerwaarden et al. Updated Table 1 with new example and reference, updated text on gene flow limits with new examples, updated climate change coverage, rewrote Conclusion section, updated references, incorporated references into text citations.

Stress and the Expression of Phenotypic Variation

Although selection acts at the phenotypic level, adaptive responses to stressors ultimately will be dictated by the presence of genetic variation in traits in the direction which selection acts (Falconer and Mackay, 1995; Lynch and Walsh, 1998). Typically, narrow sense heritability for a single trait ($h^2 = V_A/V_P$), which is the slope of the relationship between the additive genetic variance (V_A) and total phenotypic variance (V_P), is used to predict a population's response to selection ($R = h^2S$). An alternative predictor is evolvability, an expression of the ratio between the additive genetic variance and mean (x) of a trait, such as V_A/x^2. The additive genetic variance (V_A) is the proportion of the genetic component that contributes to the resemblance between parents and offspring, while the phenotypic variance includes additive genetic variance, genetic interactions between alleles within or between loci and environmental variance. Changes in any one of these components may alter a population's response to selection.

The debate as to whether narrow-sense heritability (V_A/V_P) or evolvability will decrease or increase as a consequence of stressful environments is still to be resolved (Charmantier and Garant, 2005; Hoffmann and Parsons, 1991). Empirical studies have emphasized the absence of simple answers to this question by showing both increases and decreases in heritability depend on traits and species measured (**Table 1**). Heritability is predicted to decrease under stressful conditions due to an increase in environmental variation, increasing phenotypic variation without changing V_A. In addition, stressful conditions can directly decrease V_A, as seems to happen in some animal populations when food is scarce and the proportion of V_P due to genetic factors decreases relative to environmental factors. Heritability and selection responses can increase under stressful conditions due to the release of cryptic genetic variation as a by-product of canalization, or through an increase in mutation or recombination rate in

Table 1 Trends in heritability estimates in stressful environments

Species	*Trait*	*Environment*	*Heritability*
D. buzzati (insect)	Wing Length	High temperature	(+)[a]
D. melanogaster	Wing Length	High temperature	(−)[b] (−)[a] (+)[a]
		Low temperature	(−)[c] (−)[a] (−)[a]
	Bristle number	High temperature	(+)[b]
		Low temperature	(+)[c]
	Development time	High temperature	(−)[b]
		Low temperature	(=)[c]
Stator limbatus	Egg length	Environmental quality	(+)[d]
	Lifetime fecundity		(+)[d]
Pholcus phalangioides	Body size	Food abundance	(−)[d]
	Development time		(+)[d]
	Body mass		(+)[d]
Rana temporaria (amphibian)	Development time	Desiccation	(−)[d] (−)[d]
	Mass		(+)[d] (+)[d]
	Body length		(+)[d] (+)[d]
	Tail length		(+)[d] (−)[d]
Bufo calamita	Survival to 25 stage Gosner	Osmotic stress	(+)[d]
Parus caeruleus (bird)	Nestling tarsus length	Parasite load	(+)[d]
Pica Pica	Tarsus length	Food abundance and quality	(+)[d]
	Body mass		(+)[d]
	Immune response		(+)[d]
Crassostrea gigas (bivalve molluscs)	Survival	Food abundance	(+)[d]
	Growth		(−)[d]
	Reproductive effort		(+)[d]
Salmo salar (fish)	Juvenile length	Habitat quality	(+)[d]
Ovis Canadensis (mammal)	Adult body mass	Seasons	(+)[d]
Ovis aries	Male parasite resistance	Seasons	(−)[d]
	Female parasite resistance		(−)[d]
Tamiasciurus hudsonicus	Growth rate	Annual quality	(−)[d]
Rhizoctonia solani (fungus)	Growth rate	High temperature and fungicide	(+)[e]

'+' indicates an increase in heritability while '−' represents a decrease and '=' means no change. Multiple estimates represent more than one study, while brackets indicate statistical significance was not tested or results were only suggestive.

[a]Hoffmann AA and Merilä J (1999) Heritable variation and evolution under favorable and unfavorable conditions. *Trends in Ecology and Evolution* 14: 96–101.

[b]Bubily OA and Loeschcke V (2001) High stressful temperature and genetic variation of five quantitative traits in *Drosophila melanogaster*. *Genetics* 110: 79–85.

[c]Bubily OA and Loeschcke V (2002) Effect of low stressful temperature on genetic variation in five quantitative traits in *Drosophila melanogaster*. *Heredity* 89: 70–75.

[d]Charmantier A and Garant D (2005) Environmental quality and evolutionary potential: Lessons from wild populations. *Proceedings of the Royal Society of London Series* B 272: 1415–1425.

[e]Willi Y Frank A Heinzelmann R Kälin A Spalinger L and Ceresini PC (2011) The adaptive potential of a plant pathogenic fungus, *Rhizoctonia solani* AG-3, under heat and fungicide stress. *Genetica* 139: 903–908.

stressed organisms. The results of numerous laboratory experiments and those undertaken in natural populations have emphasized that the heritability or evolvability of a trait are not stable measures but depend on environmental conditions.

Canalization refers to the process that maintains phenotypic constancy despite fluctuations in both the genotype and environment. A classic example of a 'canalized' trait is scuttelar bristle number in *Drosophila* (which rarely varies from four). A highly canalized phenotype will vary little within a range of conditions and this is often referred to as the zone of canalization; outside this range the trait may show an increase in variation. Canalization is thought to be maintained through genetic mechanisms and result in an accumulation of cryptic genetic variation not expressed within a trait's zone of canalization. However, stressful conditions may disrupt the buffering mechanism and lead to an increase in variation (Hoffmann and Parsons, 1991). A heat shock protein (Hsp90) has been implicated as a candidate mechanism for canalization; this protein is present in plants, fungi, and animals, and plays a role in folding proteins, particularly those involved in signal transduction, cell cycle, and cell death. Experiments on *Drosophila*, *Arabidopsis*, and fungi, have demonstrated increases in variation following suppression/disruption of Hsp90 with chemical inhibitors or temperature stress. Experiments manipulating Hsp90 expression in fungi demonstrated its involvement in the evolution of drug resistance; the suppression of Hsp90 allowed new mutations to have immediate effects on the phenotype (Cowan and Lindquist, 2005). Furthermore, the drug-resistant phenotype became stable without the suppression of Hsp90 following intense selection. This highlights the important role cryptic genetic variation and thus canalization may have in rapid evolution following an environmental stress. The idea of cryptic genetic variation in itself is interesting from an adaptive perspective; is cryptic genetic variation a by-product of canalization or a pre-adapted mechanism to cope with stressors?

Mutation is the ultimate source of genetic variation, and under stressful conditions mutation rates may increase, introducing new variation available for selection when variation is potentially needed most. For example, some studies of the bacterium *E. coli* suggest that the stressful starvation conditions encountered during stationary phase incubation result in a temporary increase in mutation rate due to decreased fidelity of DNA replication and a reduction in DNA repair (Bjedov et al., 2003). There is also evidence that recombination frequencies may be associated with stress levels and produce an increase in combinations of different alleles in offspring. Several experiments on *Drosophila* have shown increases in recombination frequencies under extreme culturing temperatures and starvation, while overcrowding in mice has also been found to increase recombination (Hoffmann and Parsons, 1991). Abiotic stress factors like heat and increased salinity can stimulate somatic recombination in *Arabidopsis*.

Although an increase in mutation and recombination rates may promote novel variation during environmental stress, there are also likely to be deleterious effects in stressful environments. Most mutations are deleterious and selection acts against them; an increase in mutation rate may increase the mutational load in a population exposed to stressful conditions. Many new combinations of alleles generated by recombination may also be deleterious, generating costs. It is often not clear if an increase in recombination/mutation rate is adaptive or simply a consequence of exposure to stressful conditions (Hoffmann and Parsons, 1991).

In addition to influencing the heritability and evolvability of traits directly, stressful conditions can also exert indirect effects through changes in population size. A decrease in population size is expected to lead to lower levels of additive genetic variance and decrease the adaptive potential of populations as alleles are lost in small populations due to genetic drift; however, the decrease in size often needs to be quite substantial to have much impact on trait heritability (Willi et al., 2006). The adaptive potential of small populations can also be reduced through a loss of fitness due to inbreeding.

Theory suggests that the effects of population size on evolvability can be unpredictable when there are nonadditive genetic interactions among alleles within (dominance) or between (epistasis) loci (Goodnight, 1988). In these cases additive genetic variation may increase when population size declines rapidly. This additive genetic variance, previously hidden from selection by epistatic or dominance interactions, may now contribute to a selection response, although the levels of inbreeding required may be substantial (Turelli and Barton, 2006). There is some empirical evidence that additive genetic variance can increase after a population bottleneck for morphometric and behavioral traits in the house fly, and viability in the flour beetle and *D. melanogaster*; however, these increases may not have much evolutionary relevance. In the case of viability in *D. melanogaster*, although the response to selection was higher in the bottlenecked lines, it was not high enough to overcome the initial level of inbreeding depression caused by bottlenecking (van Heerwaarden et al., 2008). As such, although a population bottleneck increased additive genetic variance, in this case it is not likely to be evolutionary significant.

Stress and Limits to Selection in Single Traits

Animal breeding and laboratory-based selection experiments on model species have shown that many species have the ability to evolve large phenotypic changes over short periods of time. Despite this there are many examples of evolutionary stasis within populations. Clearly, a simple explanation for this is an absence of genetic variance in traits under selection; however, a lack of genetic variance is rarely implicated when selection fails to produce a response. Nonetheless, examples of stasis in the presence of genetic variation are scattered throughout the literature so other factors must also be involved. There are several reasons why an evolutionary response may be low in populations including an absence of genetic variation, selection, negative consequences of low population sizes, asymmetrical gene flow and finally, complex genetic, ecological, or environmental interactions (**Figure 1**). These limits to stress adaptation are discussed below.

The most obvious factor that may limit stress adaptation is low levels of genetic variance. Variation is thought to be maintained in most traits through a balance between mutation and selection, and there is almost no empirical evidence in the literature for a

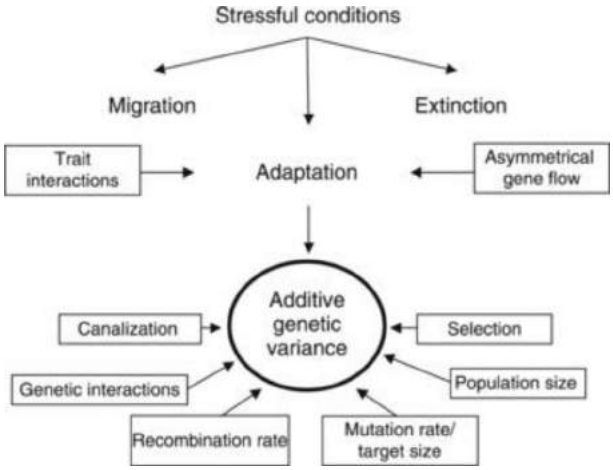


Figure 1 When faced with stressful conditions, organisms have three options: adapt, migrate, or face extinction. Migration will be influenced by ability to disperse and the availability of suitable habitat elsewhere. Adaptation is primarily influenced by the presence of genetic variance, which is influenced by population processes, including selection, population size, and genetic factors such as genetic interactions, mutation, recombination, and canalization. Trait interactions and asymmetrical gene flow may also limit adaptation to stress even when additive genetic variance is present.

lack of genetic variation in quantitative traits. However, the presence of additive genetic variance is required in specific traits for adaptation to occur. Although directional selection may increase genetic variation while rare beneficial alleles increase in frequency, if continued selection causes these alleles to increase in frequency and approach and reach fixation, genetic variance is predicted to be reduced. Mutation may be adequate to maintain variation in traits even under intense levels of selection if the mutational target is large (enough loci are involved). However, a small mutation target size (small number of loci) or a low mutation rate may limit the amount of new variation being generated, resulting in lower levels of additive genetic variance for key traits involved in adaptation.

Adaptive shifts may also be influenced by low population sizes (Willi et al., 2006). Low size can adversely affect populations due to the negative effects of inbreeding and the increased action of genetic drift decreasing additive genetic variance. Breeding between relatives increases homozygosity, unmasking unfavorable allele combinations which tend to lead to a reduction in fitness and limit evolutionary change. Furthermore, the effects of genetic drift are greater in small populations where favorable alleles may be lost by chance. Human-induced habitat fragmentation is increasingly reducing the size of many populations. Habitat fragmentation isolates populations, preventing gene flow and thus reducing effective population size. Restoring gene flow between habitats represents one effective way of reversing the effects of inbreeding and genetic drift.

Finally, interactions among alleles within (dominance) or between (epistasis) loci may enhance or constrain adaptation, depending on the intensity and directionality of the interactions. For example, positive epistatic interactions, where alleles systematically reinforce the effect of one another can increase a selection response, while negative epistatic interactions may constrain evolution by diminishing the effect of the interacting allele, 'hiding' this variation from selection. Chromosome inversion polymorphisms, where the arrangement of genes on particular chromosomes is reversed, can lock up particular combinations of alleles that are involved in an adaptive shift – because gametes become inviable when recombination occurs between inverted and noninverted chromosomes. Inversions and epistatic interactions are thought to evolve because certain combinations of alleles are favored in an adapted population. However, if environmental conditions shift, these interactions may no longer be favored and having alleles locked up could prevent an adaptive response.

While a lack of additive genetic variance can reduce the evolutionary potential of a population (Kellermann et al., 2009), many traits reach an adaptive limit with ample genetic variation for selection to act upon. There are several reasons why genetic variation does not always ensure a selection response.

One reason is that a trait may still fail to respond because of the way that selection influences a trait. Selection may act on a nonheritable component of the phenotype rather than the heritable component. For instance while there is heritable variation for antler size in red deer (*Cervus elaphus*) and directional selection for increased size, this trait has not changed for many years (Kruuk et al., 2002). It appears that only the environmental component of variation in antler size is under selection – this component reflects the nutritional state of the organisms and is unrelated to their antler genotype.

Another reason is that while low or infrequent migration can adversely influence the amount of genetic variation available for selection, high and/or asymmetrical gene flow can negatively influence a population's ability to respond to selection. Unidirectional gene flow from central/source populations may prevent local adaptation within marginal border/sink populations. As a consequence of asymmetrical gene flow, border populations will not be locally adapted to their surrounding environment, preventing further range expansion. This hypothesis may explain the presence of species borders in the absence of any geographical barriers. There is evidence for poorly adapted genotypes lacking diapause in high latitude populations of the cricket *Allonemobius socius* despite strong selection for diapause, in support of this hypothesis (Fedorka et al., 2012). However, there are also examples of

genetic differentiation in populations despite the presence of high gene flow. While asymmetrical gene flow has the potential to explain evolutionary stasis in the presence of stressful conditions, especially in border populations, empirical evidence for gene flow limiting adaptive responses exists in just a few cases.

Stress and Selection on Multiple Traits

Although we have so far focused on the effects of stressful conditions on single traits, there is a growing realization that the direction of evolution can be restricted by interactions among traits. Different traits can be genetically correlated, mostly due to the fact that the same genes can influence the expression of multiple traits (i.e., they exhibit pleiotropy). This means that variation in one trait will often be correlated with variation in different sets of traits. When selection is pushing a trait in one direction, the response to selection can be constrained because selection is at the same time pushing different traits into other directions (Lynch and Walsh, 1998).

The simplest form of pleiotropy that constrains selection on traits involves antagonistic pleiotropy. This process is commonly envisaged as occurring among two traits; for instance, increased cold resistance may be associated with a decrease in the early reproductive output of females. Selection under stressful conditions may act to increase cold resistance, but there will be an associated reduction in reproductive output, ultimately providing a limit on the extent to which cold resistance might increase. In reality, pleiotropic interactions involve multiple sets of traits, and several of these might be under selection. Identifying constraints to evolution depends on understanding whether the direction of selection can proceed when levels of genetic variation in a set of traits and genetic interactions among these traits are understood. The response to selection for a complex of traits is given by $\Delta z = G\beta$, where z is a vector of the response of individual traits, G is the genetic variance–covariance matrix, and β is the vector of linear selection gradients. The response of traits within the vector z depends critically on G, which is effectively a matrix that includes V_A in its leading diagonal (as a measure of genetic variance) and genetic interactions among traits in its off-diagonal components. Particular forms of G may mean that traits might not respond at all to selection or they may respond in a direction opposite to the one in which selection acts (Blows and Hoffmann, 2005).

A difficulty in making predictions based on G is that genetic interactions among traits, just like the genetic variance of traits, can change dramatically depending on environmental conditions particularly when these are stressful. For instance, there is a well-known interaction between development time and egg production (faster developing genotypes often have lower levels of egg production), but this interaction can change sign when organisms are exposed to stressful conditions such that rapidly developing individuals have a higher reproductive output.

Detecting Evolutionary Changes to Stressors

Anthropogenic disturbances such as pollution, habitat modification and destruction, climate change, and species exploitation are impacting on the distribution and abundance of many sensitive species. Monitoring and predicting the ability of populations to respond to different environmental stressors is vital to understand and minimize their impact on species. Individuals respond to stressors in various ways, depending on intensity and duration. Below we discuss how stress responses can be detected and provide examples of how stressors are already influencing natural populations.

Evolutionary Responses Detected Through Traits

Populations have the potential to evolve and thereby shift phenotypes over short periods of time. Impacts of stressful environmental conditions on populations can be detected through phenotypic shifts, such as increases in tolerance levels, changes in the timing of life history traits including flowering and reproduction, or changes in morphology. These phenotypic shifts can provide direct evidence of rapid responses to stressful environments and there are many examples in the literature. These include the many cases of rapid evolution of pesticide resistance in a variety of insects and mites to a range of chemical classes such as cyclodienes, carbamates, formamidines, organophosphates, and pyrethroids. Resistance is often first detected from the failure of a pesticide to control a pest species and this is usually validated by experimentally comparing tolerance levels of putative-resistant and sensitive populations. Resistance has also evolved in organisms not targeted by pesticides; for example, natural populations of the vinegar fly *D. melanogaster* are resistant to insecticides such as dichloro diphenyl trichloroethane (DDT) even though they have never specifically been targeted for control by these pesticides (Catania et al., 2004).

Evolutionary responses to pollutants have been less widely documented than responses to pesticides. Nevertheless, there are now well-studied cases of plant species evolving in response to waste from mining operations. Plant populations growing close to smelters commonly show a much higher level of resistance to heavy metals than other populations. The evolution of pollution tolerance has also been demonstrated in aquatic fauna, including crayfish, worms, and midges. However not all plants or animals successfully evolve in response to pollutants: levels of biodiversity are often much lower in polluted compared to unpolluted areas, suggesting that many species are unable to adapt.

Pollution stressors can indirectly influence the evolution of traits. There is the classic case of melanism in the peppered moth evolving due to increased pollution levels in England in the 1800 s. Although this species of moth is usually a light color, a darker

melanic form increased in frequency in some polluted areas during this time. The increase in this form was not a direct consequence of pollution but a response to predation by birds; pollution caused the trees which these moths reside on to darken and the melanic form was less likely to be predated upon than the previously camouflaged lighter form. A recent example is the decreasing frequency of grey body colour morphs (and concomitant increase in the brown morphs) in tawny owls as a consequence of decreasing snow cover that provides camouflage for the grey morphs (Karell et al., 2011). These examples highlight evolutionary responses arising as a consequence of changes in biotic interactions due to an environmental stress rather than its more direct effects.

Phenotypic evidence is accumulating on organism responses to recent climate change. Average global surface temperatures have risen 0.8 °C in the past century and more than 0.2 °C per decade over the past 30 years, and several studies that span decades are linking these changes to shifts in the timing of life history traits and geographical shifts in species ranges. A variety of birds, butterflies, and alpine herbs in the Northern Hemisphere have altered their geographic boundaries; a meta-analysis on 99 different species found significant range shifts, averaging 6.1 km per decade away from the equator. Additionally, many birds are now first laying eggs earlier and also becoming smaller in size.

To test whether these patterns reflect evolutionary shifts rather than just phenotypic variants of the same genotype (phenotypic plasticity), a genetic basis needs to be established through controlled breeding experiments or family studies (Hoffmann and Sgrò, 2011). Changes in the timing of diapause induction in the pitcher plant mosquito is an example of adaptive evolution in response to longer growing seasons induced by recent climate change (Bradshaw and Holzapfel, 2006). Parent–offspring comparisons have shown the timing of diapause induction in the pitcher plant mosquito has a heritable genetic basis and this trait has also been shown to cline with latitude along eastern North America. When southern and northern populations are exposed to mid-latitude lengths, they enter diapause either too late (southern) or too early (northern) and experience between a 74% and 88% decline in fitness, indicating that diapause timing is an adapted phenotypic response. Over the last 30 years, this latitudinal cline has shifted, with populations from the north shifting toward a more southern shorter day length form as growing seasons have become longer due to warmer temperatures. In contrast, shifts in the breeding time of birds often appear to represent a plastic response rather than an evolved genetic response.

The rigorous experimental approach needed to demonstrate a genetic basis for a phenotypic shift cannot be readily carried out for many organisms, particularly when they have long generation times. For this reason some of the recent attention on detecting evolutionary changes under stress has moved to the genetic and genomic levels.

Evolutionary Responses Detected Through Allele Changes

Underlying evolutionary changes are allele frequency changes – alleles that are at a selective advantage in particular environments are favored and increase in frequency. As evolutionary responses to stress directly involve changes in allele frequencies, detecting changes at this level is potentially the most direct way to measure and predict the biological consequences of stressors in populations. Initial studies used neutral markers to compare levels of genetic diversity between polluted and nonpolluted populations was one of the first ways DNA markers were used to assess the genetic impacts of stress. However, these markers do not reflect adaptive changes unless they are closely linked to a gene under selection. Genetic markers based on loci coding for proteins (allozymes markers) have been used in some studies, but these can only be applied to well-preserved specimens, which limits their application in longitudinal comparisons of specimens that are often poorly preserved.

One example where changes in allozyme markers were successfully linked to climate change is the alcohol dehydrogenase enzyme (Adh) in *D. melanogaster* (Umina et al., 2005). In the 1980s, the allele frequencies of the *Adh* gene (Adh^F and Adh^S) were found to cline with latitude across the continents of Asia, Australasia, and North America. This association with latitude has been implicated with thermal tolerance where flies with the Adh^S allele are more likely to survive heat shock than flies with the Adh^F allele. The rapid formation of these clinal patterns across continents highlights strong climatic selection on these genes. To determine whether recent changes in climate have influenced the clinal patterns in this marker since the 1980s, the Australian east coast cline was reinvestigated in 2000 and 2002. The slope of the latitudinal association had not changed over the 20 year study period; however, a shift equivalent to 4° in latitude was detected in the height of the graph, with the southern high latitude populations now having the genetic constitution of the more northerly populations. Frequencies of the chromosome inversion polymorphism (*In(3R)P*) in *D. melanogaster* have also changed drastically in the last 20 years (**Figure 2**). Furthermore, changes in chromosome arrangement frequencies of *D. robusta* from North America have been linked to shifts in minimum temperature since the 1970s. Similar parallel shifts in latitude association in numerous inversion polymorphism analyzed simultaneously have also been detected in populations of *D. subobscura* across three separate continents. These studies reflect the power of latitudinal markers to act as sensitive indicators of climatic change.

Ideally, shifts in allele frequencies in genes specific to particular environmental stressors could be used to track genetic responses to stress. Progress has been made in the area of pesticide resistance. Microarray technology has led to the discovery that a single chromosome P450 gene, Cyp6g1, confers resistance to DDT in *D. melanogaster*. Cyp6g1 was found to be over-expressed in *D. melanogaster* lines from natural populations that exhibit resistance due to the insertion of a transposable element in the promoter region that regulates expression of the gene (Catania et al., 2004). Examination of the patterns of molecular variation around this gene showed a sharp reduction in the level of molecular variation in this area of the genome indicative that it had undergone recent selection (selective sweep). Under a selective sweep, variation decreases in the genomic region around the candidate gene, as alleles at adjacent loci spread with the new allele favored by selection. Associations between insecticide resistance

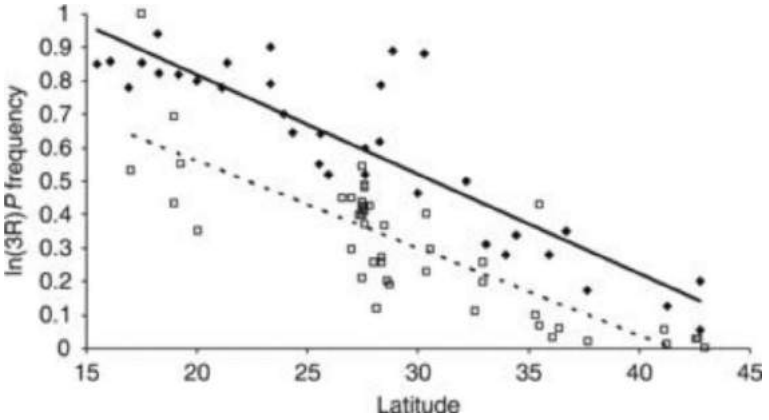


Figure 2 Change in latitudinal patterns, between 1979 and 1982 and between 2002 and 2004, of the inversions *In(3R)Payne*. Open symbols and dashed lines indicate 1979–82 pooled data solid symbols and solid lines indicate 2002–04 pooled data. From Umina PA, Weeks AR, Kearney MR, McKechnie SW, and Hoffmann AA (2005) A rapid shift in a classic clinal pattern in *Drosophila* reflecting climate change. *Science* 308: 691–693. Reprinted with permission from AAAS.

and specific gene changes have been identified now in many insects including moths and mosquitoes, and they are being increasingly used to track changes in resistance in natural populations and to inform management actions.

Concluding Remarks

Ongoing research on model and non-model organisms is leading to the identification of genes and genetic processes linked to evolutionary shifts in response to climatic and chemical stressors. As more genomes are being sequenced, it is becoming possible to extend this search to a wide range of organisms, particularly sensitive species that respond differently to environmental stressors. Furthermore, the ability to extract DNA data from museum specimens or specimens that can be resurrected such as in the small crustacean Daphnia from lake beds is providing opportunities to collect historic data. Both phenotypic and genotypic data on model organisms in natural and laboratory-based environments provide evidence for rapid evolutionary responses to stress. It remains to be seen if sensitive species and those with restricted distributions can also show rapid evolutionary adaptation. There is a huge potential for genetic markers emerging from genomic studies to reveal current and past effects of stress on natural populations and to indicate the potential of populations and species to adapt in the future.

References

Bjedov I, Tenaillon O, Gérard B, Souza V, Denamur E, Radman M, Taddei F, and Matic I (2003) Stress-induced mutagenesis in bacteria. *Science* 300: 1404–1409.
Blows MW and Hoffmann AA (2005) A reassessment of genetic limits to evolutionary change. *Ecology* 86: 1371–1384.
Bradshaw WE and Holzapfel CM (2006) Evolutionary response to rapid climate change. *Science* 312: 1477–1478.
Catania F, Kauer MO, Daborn PJ, Yen JL, Ffrench-Constant RH, and Schlotterer C (2004) World-wide survey of an *Accord* insertion and its association with DDT resistance in *Drosophila melanogaster*. *Molecular Ecology* 13: 2491–2504.
Charmantier A and Garant D (2005) Environmental quality and evolutionary potential: Lessons from wild populations. *Proceedings of the Royal Society of London Series B* 272: 1415–1425.
Cowan LE and Lindquist S (2005) Hsp90 potentiates the rapid evolution of new traits: Drug resistance in diverse fungi. *Science* 309: 2185–2189.
Falconer DS and Mackay TFC (1995) *Introduction to quantitative genetics*, 4th edn. Harlow: Longman.
Fedorka KM, Winterhalter WE, Shaw KL, Brogan WR, and Mousseau TA (2012) The role of gene flow asymmetry along an environmental gradient in constraining local adaptation and range expansion in the ground cricket *Allonemobius socius*. *Journal of Evolutionary Biology* 25: 1676–1685.
Goodnight CJ (1988) Epistasis and the effect of founder events on the additive genetic variance. *Evolution* 42: 441–454.
Hoffmann AA and Parsons PA (1991) *Evolutionary genetics and environmental stress*. Oxford: Oxford University Press.
Hoffmann AA and Sgrò CM (2011) Climate change and evolutionary adaptation. *Nature* 470: 479–485.
Karell P, Ahola K, Karstinen T, Valkama J, and Brommer JE (2011) Climate change drives microevolution in a wild bird. *Nature Communications* 2.
Kruuk LEB, Slate J, Pemberton JM, Brotherstone S, Guinness F, and Clutton-Brock T (2002) Antler size in red deer: Heritability and selection but no evolution. *Evolution* 56: 1683–1695.
Lynch M and Walsh B (1998) *Genetics and analysis of quantitative traits*. Sunderland: Sinauer Associates.
Merilä J (2012) Evolution in response to climate change: In pursuit of the missing evidence. *Bioessays* 34: 811–818.
Turelli M and Barton NH (2006) Will population bottlenecks and multilocus epistasis increase additive genetic variance? *Evolution* 60: 1763–1776.
Umina PA, Weeks AR, Kearney MR, McKechnie SW, and Hoffmann AA (2005) A rapid shift in a classic clinal pattern in Drosophila reflecting climate change. *Science* 308: 691–693.
van Heerwaarden B, Willi Y, Kristensen TN, and Hoffmann AA (2008) Population bottlenecks increase additive genetic variance but do not break a selection limit in rainforest Drosophila. *Genetics* 179: 2135–2146.
Kellermann V, van Heerwaarden B, Sgrò CM, and Hoffmann AA (2009) Fundamental evolutionary limits in ecological traits drive Drosophila species distributions. *Science* 325: 1244–1246.
Willi Y, Van Buskirk J, and Hoffmann AA (2006) Limits to the adaptive potential of small populations. *Annual Review of Ecology, Evolution, and Systematics* 37: 433–434.

Food Specialization☆

Richard Svanbäck, Uppsala University, Uppsala, Sweden
Daniel I Bolnick, University of Texas at Austin, Austin, TX, United States

Nomenclature

BIC The variance in resource use between individuals
TNW Total niche width of the population
WIC The variation in resource use within individuals

Glossary

Extrinsic factors In this article it refers to factors not attributed to the individual such as for example, intra- and inter-specific competition, predatoion, parasitism, and ecological opportunity.
Individual diet specialization Where the population niche consists of subsets of specialized individuals that compromise the population niche.
Intrinsic factors In this article it refers to factors attributed to the individual such as for example, sex, size, age, or morphology.
Ontogeny The development or course of development especially of an individual organism.
Optimal foraging theory A mathematical theory used to model foragers' decision-making in choosing among multiple available resources, assuming selection has favored a benefit-maximizing strategy.
Polymorphism Heritable variation among individuals within a population, for a focal trait or gene.
Sexual dimorphism The condition where the two sexes of the same species exhibit different characteristics beyond the differences in their sexual organs.

Introduction

Energy acquisition has been important in ecological and evolutionary theory to predict the survival and fitness of individuals and species. Energy acquisition is affected by prey density and behavior, as well as an individual's ability to detect, capture, and consume the prey (Fig. 1). These foraging abilities are shaped by both morphological and behavioral traits, which generally reflect genetic and environmental influences. To the extent that morphological or behavioral variation leads to different rates of energy income, these phenotypic traits will be subject to natural selection. Evolution of foraging traits will then influence the structure of ecological communities through changing predator–prey dynamics. Consequently, a major goal in evolution and ecology research has been to understand the mechanisms that determine patterns of resource use. In particular, researchers have tried to explain why species are generalists (using a diverse set of resources) or specialists, and why they use particular subsets of available resources. The resulting "niche theory" was a major focus of ecology from the 1960s through the 1980s, during which period ecologists tended to view the niche as a property of a species or a population as a whole. However, to understand niche evolution we must think of the niche instead as a property of individuals, based on heritable variation in foraging traits. Many animal populations with generalized diets actually consist of individuals specializing on a small subset of the population's diet. In some cases such diet variation among individuals may exceed differences between distinct species. Within-population diet variation can be related to phenotypic variation, including sexual dimorphism, ontogenetic niches, or related to morphological or behavioral variation in the population.

Defining Food Specialization

The term "specialization" is used in a variety of ways that require some initial clarification. A species or individual may be considered "specialized" (an adjective) if it uses a narrower diet than some reference group. For example, a specialized species might use fewer prey than are available in the environment, or than another species uses. An individual may be specialized if it uses fewer prey than its conspecifics (individually or as a group). Alternatively, the adjective has been used to indicate a species that uses a unique set of prey (not necessarily less diverse) compared to related species. Specialization can also be a verb, the evolutionary process of becoming more specialized. In this case we are comparing the focal group to its evolutionary progenitor. Finally, a specialization can also be a noun and refer to a phenotypic trait that confers a derived ability to perform some ecological task, for example, hypertrophied pharyngeal jaws are a specialization for molluscivory. In the context of individual food specialization, these three views correspond to asking whether or not an individual has a narrower niche than its population,

☆*Change History*: October 2017. Richard Svanbäck and Daniel I Bolnick updated the text and further reading to this entire article, especially a whole section on extrinsic factors have been added.

This is an update of R. Svanbäck and D.I. Bolnick, Food Specialization, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 1636–1642.

 doi:10.1016/B978-0-12-409548-9.10924-8

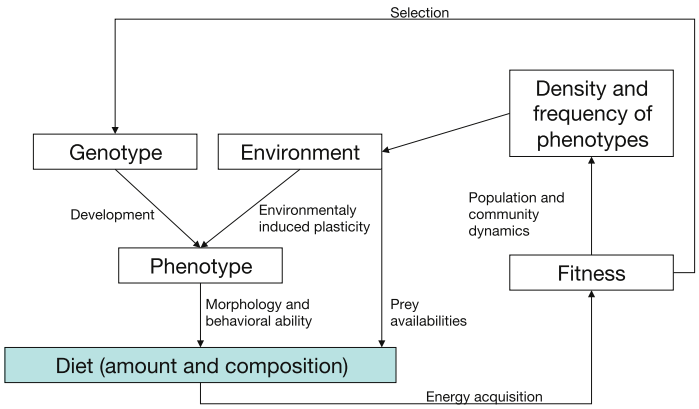


Fig. 1 Possible scenarios for how individual diets are determined and how it will feed back on ecological and evolutionary dynamics.

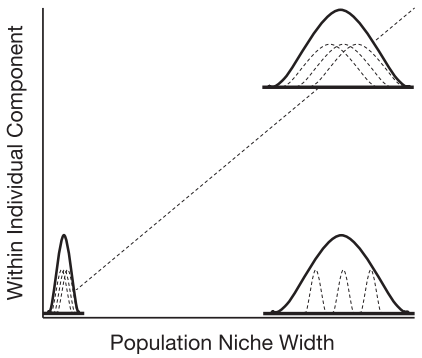


Fig. 2 Individual diet specialization is part of a continuum from where the within-individual component (WIC) equals that of the total population niche width (TNW) (on the hatched diagonal) to where WIC is a small part of TNW (close to the *x*-axis). The three schematic diagrams represent how individuals (*hatched curves*) can subdivide the population's niche (*thick curve*). The two diagrams on the hatched diagonal represent cases where individual diet specialization is low, that is, WIC is a large part of TNW whereas the diagram at the bottom right represents a case where there is a high degree of individual specialization.

describing the process of becoming more specialized and finally, the morphological or behavioral traits underlying food specialization among individuals.

When adopting the relativistic definition of specialization, it is possible to measure the degree of specialization (how much narrower a niche is than its reference). Consider a food resource distribution (niche distribution) that can be described by a single continuous variable such as prey size. The total niche width (TNW) of the population is simply the variance in size of all the prey captured. The total niche width can be partitioned into two components; the within-individual component (WIC) and the between-individual component (BIC) so that TNW = WIC + BIC. WIC is the average variance of prey sizes found within individuals' diets, whereas BIC is the variation between individuals. Individual specialization occurs when individuals niche widths (WIC) are much smaller than their population as a whole (TNW), or equivalently when BIC is a large proportion of TNW such that BIC/TNW is large. An increase in BIC could occur in a number of different ways. For example, BIC can be large if individuals of different age- or size-classes use different diets, or if males and females use different diets. Furthermore, an increase in BIC, could also be due to phenotypic variation among individuals (morphological or behavioral). Thus, individual food specialization is part of a continuum from where the within-individual component equals the total niche width (WIC/TNW = 1), that is, no individual specialization, to where WIC is only a small part of TNW, that is, great individual specialization (**Fig. 2**).

Causes of Diet Specialization

Why does an individual forager use a given set of resources? This is a problem often addressed with optimal foraging theory. Although optimal foraging theory has not always been successful in generating qualitative predictions of foraging behavior it can serve as a rough tool to understand what makes an individual choose its resources. According to optimal foraging theory, an individual is expected to choose its diet from among available resources to maximize its expected fitness. This might be achieved by maximizing energy income per unit time, or by some balance between energy income and another goal such as risk avoidance. The diet of an individual then depends on a variety of factors such as the energetic content of different prey, handling times, search

efficiencies, resource abundances and risk such as predation. All these factors are a combination of the resources traits (e.g., size, caloric content, abundance and defensive traits) and the forager's phenotype (e.g., search or handling behavior, morphology, physiology, and experience). Consequently, an individual's preferences for alternative resources will reflect a complex interaction between resource traits and the forager's phenotype.

Given a shared environment, why would different conspecific individuals specialize on different prey types? The most proximate answer is that individuals will use different resources if they have different resource use efficiencies. Different resource use efficiencies may reflect variable morphological, behavioral or physiological capacities to detect, capture, or handle alternative resources. This poses a second question: why does phenotypic variation result in efficiency variation? Without trade-offs constraining efficiencies on alternative resources, phenotypic variation would not produce efficiency variation. By trade-offs, we mean that an individual adopting one strategy that increases efficiency on one prey type will lose the ability to efficiently use alternative prey. In situations with trade-offs, a generalist forager may be unable to perform either strategy as well as the respective specialist and may therefore be selected against. Such trade-offs are known to occur in many aspects of foraging including search efficiency, handling time efficiency (prey capture and digestion), and prey recognition. Thus, trade-offs remain one of most plausible mechanisms for limiting an individual's niche breadth. Consequently, food specialization within a population can occur as a result of particular age- or sex-based characteristics or by individual level variations in phenotype. Phenotypic variation is often associated with discrete polymorphisms with distinct modes along a quantitative axis such as beak depth. However, this variation may also be simple unimodal quantitative variation, or may have a complex behavioral or biomechanical basis that is not effectively summarized by a single quantitative axis.

Not all diet variation need to be related to variation in efficiencies. Even individuals with equal efficiencies can nevertheless have different diets reflecting variation in social status, mating strategy or territories. For example competitive dominant individuals may defend and monopolize preferred resources, and suboptimal individuals will be forced to rely on lower quality resources. In this case, diet variation arises because individuals differ in their ability to achieve their optimal diet. Such interference competition is facilitated when the optimal resource is patchily distributed and can be defended by territories. Individual specialization in several mammal species is caused by territoriality in patchy environments. For example, individuals of both bears and pine martens whose territories abut streams have more fish in their diets compared to neighbors whose territories do not include streams. However, territorial foragers with similar resource use efficiencies may also end up with similar diets due to homogeneously distributed resources, so that each territory contain the full resource spectrum in the same proportions. Furthermore, coarse-grained environmental heterogeneity may sometimes also create differences among individual diets not related to foraging efficiencies. An example of that is the moose in the boreal forest region are severely limited in their movement in winter due to deep snow. Individuals will forage intensely in small patches of often homogeneous vegetation, while another will forage on very different vegetation. The next winter, a given individual may or may not find itself in a similar patch.

Although most of the examples of food specialization are probably due to maximization of energy intake, other food constituents can also be important in developing specialization. For example, the muskox in the Canadian arctic spends more time foraging for protein-rich arctic vetches on low productivity upland tundra than in the energy rich sedge meadows in the valleys, where in relatively little time they can gorge themselves on what is not a limiting resource in summer. Such food specialization is probably due to lack of essential nutrients in the most energy rich food source. Thus, organisms may forage to maximize metrics other than energy. Morever, it is plausible that individual foraging differences might arise from among-individual differences in optimization goals, but this possibility has not been widely tested.

However, most cases of individual diet specialization are not due to territoriality or lack of essential nutrients, but could instead be related to intrinsic (properties attributed to the individual, e.g., sexual differences, ontogenetic niche shifts, or phenotypic variation) or extrinsic (properties not attributed to the individual, intra- and inter-specific competition, predation, parasitism, and ecological oportunity) factors. We will here briefly describe causes and consequences of diet specialization related to intrinsic and extrinsic factors.

Intrinsic Factors

Sexual dimorphism

Sexual dimorphism in trophic structures is a common phenomenon in most animal phyla, and has attracted considerable interest from evolutionary biologists. In most animals, males and females differ in size, sometimes substantially. The usual cause cited for sexual dimorphism in animals is sexual selection acting through female choice and/or male-male competition.

Natural selection acting on the fitness advantages of reduced resource competition between the sexes, however, is also an important alternative evolutionary scenario that can produce sexual dimorphisms. This alternative idea is that sexual differences in body size or morphology may evolve due to ecological causes—that is, the ecological benefit of the sexes occupying different ecological niches, which reduces inter-sexual food competition and expand the species' overall feeding niche. Adaptation to different diets may also be due to sex specific nutritional requirements, as for example females may have additional dietary components compared to males that directly relates to nutrients needed for the production of eggs. However, sexual selection and natural selection for sexual dimorphisms are not incompatible, for example, it may well be that the direction of dimorphism in body size is determined by sexual selection, but that the degree to which the sexes diverge is either constrained or amplified by ecological factors.

Ontogeny

Intraspecific resource partitioning due to ontogenetic niche shifts occurs in many species. This is probably most obvious when age groups live in different habitats and use different types of food. Examples of such ontogenetic niche shifts include particle-feeding amphibian larvae that turn into carnivorous adult amphibians, immature stages of aquatic insects that turn into adult terrestrial insects, planktonic marine invertebrate larvae that settle down to be sessile adults. Other species live in the same general habitat but use distinct food types depending on size or age class, as for example, fish where the young are gape-limited planktivores or herbivores and the adults are predators. There might also be large ontogenetic changes in diet within the same general food type as for example for predators that start to feed on small prey species but as they grow they can and will include larger prey types in their diet.

Differences in diet over the ontogeny of an animal can have different reasons. Search and handling efficiencies can change as animals grow and gape-limitation eases. Other animals change their diet with growth because predation risk declines with size, allowing them to use previously risky habitats. Still other species exhibit a fundamental change in habitat use as a result of life history changes or dispersal (e.g., marine invertebrates with planktonic larvae that settle to become sessile adults).

Morphological and behavioral variation

Morphological variation within a species may reflect adaptive genetic variation or phenotypic plasticity, whereby individuals are able to modulate their morphology in response to foraging conditions. Food specialization in relation to morphology is thought to be a response to competitive pressures within a population favoring niche partitioning among conspecifics, and hence niche expansion. Diet–morphology relationships are mainly due to trade-offs in either search or handling time efficiencies where adaptations to maximize foraging intake have long been recognized to be dependent on specific habitats or prey types. Morphological specialization is thus presumed to lead to an increase in foraging efficiency in the habitat to which they are adapted but on the other hand, it is likely to constrain their efficiency in alternate habitats. For example, the beak size in birds can be related to feeding efficiencies on different sized seeds. Larger beak sizes are better adapted at large seed sizes but they loose their efficiency on smaller seed sizes. Smaller beak sizes on the other hand are more efficient on smaller seed sizes. In fish, species (or individuals) that are specialized for living in open water and searching for widely dispersed prey have fusiform bodies that minimize drag and allow for efficient cruising. In contrast, fish that are adapted for searching for prey in structurally complex habitats have a deep and laterally compressed body and extended fins and are well suited for slow and precise maneuverability.

Behavioral variation can also influence feeding performance and thus diet use of individuals. Behavioral variation can arise from morphological variation in the sense that morphology may determine an individual's ability to carry out a specific prey-capture behavior. However, behavioral variation may also arise from variation in cognitive ability, perception, risk aversion, which may all have a genetic and/or environmental basis. In particular, early experience may lead to learned behaviors that influence prey use. For instance, individuals that are initially exposed to different prey acquire divergent search images that influence later foraging abilities. For example, sea otters (*Enhydra lutris*) show extensive among-individual variation in prey preferences. These differences are maintained through cultural transmission, as mothers teach their offspring how to forage. In *Pieris* butterflies, individuals form a new search image for a single flower type at the start of each day, and specialize on that flower for the rest of the day. Individuals that feed on multiple flower types form less effective search images and thus forage less efficiently.

Extrinsic Factors

Intraspecific competition

Increased intraspecific competition can lead to increased or decreased among individual diet variation, depending on the context. It is generally thought that intraspecific competition will favor diversification. In fact, several correlational studies have shown a positive relationship between intraspecific competition and among individual diet variation, alternatively that reduced resource abundance lead to increased among individual diet variation. This positive effect of intraspecific competition on diet divergence has also been shown in experimental studies. Yet, other studies have found that among individual diet variation is lower at higher levels of intraspecific competition. In general, competition has been shown to be diversifying in species with small effects on their resources and non-diversifying in species with large effects on their resources. Increased density of consumers might besides increase resource competition as described above also increase interference competition. Although not as thoroughly investigated as resource competition, interference competition can lead to diet variation among individuals when individuals defend territories with different resources.

Interspecific competition

In 1965, Van Valen proposed the "niche variation hypothesis," which suggests that "populations with wider niches are more variable than populations with narrower niches." Van Valen suggested this hypothesis in an attempt to explain the observation that bird populations inhabiting oceanic islands tend to be more morphologically variable than their mainland counterparts. Van Valen suggested that island birds evolve to use a wider diversity of resources when they are released from selection imposed by interspecific competitors. Similarly, theory on character displacement and character release relies on the assumption that a population's niche width is limited by a competitor species, and when released from competion, the population can diversify by increased among individual diet variation. The only experimental study so far on how interspecific competition affects among

individual diet variation was conducted on threespined sticklebacks (*Gasterostues aculeatus*). This study showed that among individual diet variation in sticklebacks increased as predicted, when were they released from competition with juvenile cut-throat trout (*Oncorhynchus clarki*). On the other hand, sticklebacks showed reduced among individual diet variation as a consequence of expanded individual niche widths (and constant population niche width) when they were released from competition with prickly sculpin (*Cottus asper*). Some correlational studies have also been performed and found a negative effect of species richness on resource use divergence, as predicted by the niche variation hypothesis. Thus, studies on the effect of species richness on among individual diet variation have yielded conflicting results and more research has to be done to understand when interspecifc competition will increase and when it will decrease among individual diet variation.

Predation

Foraging in different habitats or on different prey types may come with differences in predation risk. As predation and predation risk can have strong influence on prey populations, it is expected to have large effects on among individual diet specialization. Among individual diet specialization can be affected by predation either via reduction in forager densitity or changes in forager behavior in reponse to predation risk. The reduction of the density of forager species by predation, which in turn will reduce the competition within the forager species can either increase of decrese among individual diet specialization. If among individual diet specialization increases with intraspecific competition, then predation will reduce diet specialization. The reduction of competitors by predation can also counter the effect of over-exploiting resources leading to an increase in diet specialization (see "Intraspecific competition" above). Refuge seeking is a common reponse in prey to predation risk. In this case, predation can close off niche space by driving forager individuals to a safer habitat, thus reducing among individual diet variation. However, forager individuals may differ in vulnerability or risk aversion and in this case predation risk might exaggerate among individual diet variation as less vulnerable or less risk-avers individuals can stay in the more profitable habitat whereas other individuals forage in the safer refuge habitat.

Parasites

It is well known that niche specialization may drive parasite infections, as variability in niche use can differentially expose individuals to prey species containing intermediate hosts. However, recent findings have shown that habitat or diet-specific parasites may also influence individual niche use. For example, manipulative parasites, parasites that alter aspects of phenotypic traits and behaviors of the host, can drive divergence in diet between paraitized and unparasitized host individuals. For example, cestode parasites modifies their behavior of intermediate fish hosts to use habitats that increase predation from the definitive host (fish eating birds). Parasites can also affect growth and competitive ability of infected hosts, potentially leading to divergence in niche use between parasitized and unparasitized host individuals. Lastly, parasites that are acquired by consuming a particular food type, may drive selection favoring foragers that avoid that prey type. Thus, parasitism could reduce individual specialization by driving the evolution of a narrower niche width. However, despite some evidence that parasites can drive among individual niche specialization, the potential role of parasites in trophic niche specialization remains unclear, and further research is needed.

Ecological opportunity

Both predation and interspecific competition can restrict populations from using specific resources or habitats. By ecological opportunity we mean other factors, that hinders niche divergence, such as patch size, and microhabitat/resource diversity. As larger patch sizes can hold more diverse prey populations habitat fragmentation may have large effects on among individual diet variation. For example, it has been shown that habitat fragmentation in eustarine tidal wetlands leads to lower resource diversity consequently decreasing among individual diet variation in the gray snapper (*Lutjanus griseus*), a predatory fish. Similarly, it has also been shown that among individual diet variation in the fruit bat, *Rousettus aegyptiacus*, was higher in spring when the number of plant species bearing fruits was also higher.

Measuring Diet Specialization

There are many different indices for calculating the degree of individual diet specialization. Most measures of niche breadth compare the frequency distribution of the individuals' resource use with that of the average population resource use. The total niche width of a population (TNW) can be broken down into the average variance of resources found within individuals' (within-individual component WIC), and the variance between individuals (between individual component) so that TNW = WIC + BIC. Assuming that variation in diet parameters can be expressed along a single continuous dimension x (e.g., prey size) and x_{ij} is the size if the jth prey item in individuals i's diet. Then,

$$\mathrm{TNW} = Var\left(x_{ij}\right)$$

$$\mathrm{WIC} = E\left[Var\left(x_j i\right)\right]$$

$$\mathrm{BIC} = Var\left[E\left(x_j i\right)\right]$$

The relative degree of individual specialization can be measured as the proportion of TNW that is explained by the within-individual variation, WIC/TNW, or its converse BIC/TNW. The latter is more intuitive, since it is larger in populations with greater diet variation, but the former has been used more frequently.

This index is, however, limited to continuous diet data. An alternative is to use the Shannon–Weaver index as a proxy for variance for discrete data such as the frequency with which individuals use a set of prey taxa. In this index n_{ij} represents the number (or mass, or volume) of diet items in individual i's diet that fall into category j. n_{ij} is then transformed into p_{ij}, which describes the proportion of the jth resource category in individual i's diet. Then

$$\mathrm{TNW_s} = -\sum_j q_j \ln\left(q_j\right)$$

$$\mathrm{WIC_s} = \sum_i p_i \left(-\sum_j p_{ij} \ln\left(p_{ij}\right)\right)$$

$$\mathrm{BIC_s} = \sum_i p_i \ln(p_i) - \left\{\sum_j q_j \left[-\sum_i \gamma_{ij} \ln\left(\gamma_{ij}\right)\right]\right\}$$

Subscript "s" in $\mathrm{WIC_s}$, $\mathrm{BIC_s}$, and $\mathrm{TNW_s}$ distinguishes this from the continuous index. The variable p_i, is the proportion of all resources used by the population that are used by individual i, q_i is the proportion of the jth resource category in the population's niche, and γ_{ij} is the proportion of the population's total use of resource j that was used by individual i, so that

$$p_{ij} = \frac{n_{ij}}{\sum_j n_{ij}}$$

$$p_i = \frac{\sum_j n_{ij}}{\sum_i \sum_j n_{ij}}$$

$$q_j = \frac{\sum_i n_{ij}}{\sum_i \sum_j n_{ij}}$$

$$\gamma_{ij} = \frac{n_{ij}}{\sum_i n_{ij}}$$

As before, $\mathrm{TNW_s} = \mathrm{BIC_s} + \mathrm{WIC_s}$ and the relative degree of individual specialization is also as before; $\mathrm{WIC_s/TNW_s}$. The Shannon–Weaver index has been widely used, but suffers serious disadvantages. First, the index increases both with the number of prey categories and/or with the evenness with which those categories are used, making its interpretation rather ambiguous. Second, the index can substantially over-estimate diet variation (low WIC/TNW) when the population as a whole relies heavily on a single prey taxon. In this case individuals whose diets largely match the population as a whole may have WIC = 0. The resulting WIC/TNW would thus be close to zero, falsely implying that there is pronounced diet variation.

The proportion similarity index (PS), is a distribution-overlap measures that provide discrete data alternatives to $\mathrm{WIC_s/TNW_s}$. This measures the mean pair wise overlap between each individual and the population where the diet overlap between an individual i and the population is

$$PS_i = 1 - 0.5\sum_j \left|p_{ij} - q_j\right| = \sum_j \min\left(p_{ij}, q_j\right)$$

where p_{ij} and q_j are the same as above. For individuals that consume resources in direct proportion to the population as a whole PS_i will equal 1. For an individual that specializes on a single diet item j, PS_i takes on the value q_j. The population-wide prevalence of individual specialization (IS) is then measured by the average of individuals' PS_i values.

These three indices all vary from values close to 1—indicating that all individuals have the same diet (may it be broad or narrow diet breadth)—to near 0, indicating strong individual specialization. All these indices yield very similar though not identical values when applied to the same data. But which index should one use? Although the indices are quantitatively very similar, each has particular advantages and disadvantages. Both WIC/TNW and $\mathrm{WIC_s/TNW_s}$ offers the attractive advantage of quantifying both relative specialization and population niche width. This can be used when testing hypotheses such as whether niche expansion during competitive release occurs by increased inter-individual variation (greater BIC), implying that higher TNW is associated with greater individual specialization (low WIC/TNW). Both WIC/TNW and $\mathrm{WIC_s/TNW_s}$ make assumptions about the resource distribution. The continuous version assumes that niches are normally distributed, whereas the Shannon–Weaver index assumes that resources are evenly distributed. The Shannon–Weaver-based index can also be biased to overestimate individual specialization due to its natural log of a proportion. Additional advantages of the two overlap indices are that they make no assumptions about the shapes of the resource distributions, and they yield estimates of specialization for each individual. The estimate of specialization for each individual further makes it possible to study the variation in specialization among individuals in a population so one can study the ecological or evolutionary (fitness) consequences of individual specialization.

It can also be important to characterize how the among individual diet variation is structured in the population, for example, the level of nestedness and clustering of the diets within the popoulation. Together with a measurement of individual

specialization, measuretments of nestedness and clustering can be useful to identify pattern of resource usage within the populations. A nested pattern of diet use indicates that the diet of selective individuals is a subset of less selective individuals. The degree of clustering in diet specialization can be estimated with network analysis. Network analysis of nestedness will then compare the overall density of connections in the network to the density of connections around individual nodes and can be used to analyze whether the network of diet use is totally random or if it is organized in clusters. Note that the clusters do not mean that individuals are clustered in space, but rather that individuals in the same cluster use the same subset of resources. Clustered diets would in this analysis represent discrete diet variation commonly refered to as discrete resource polymorphism, whereas totally randomly organized networks would represent more continuous diet variation. Indices of nestedness and clustering as well as the described indices of individual diet specialization can be calculated using the R-package RInSp (see "List of Relevant Website').

It is important that the niche axis or diet categories have been chosen appropriately when measuring diet specialization. For example, coarse-grained niche studies that pool functionally distinct resources may underestimate individual specialization. When resources that a forager distinguishes among are lumped together by an ecologist, individuals may appear more generalized than they really are. Conversely, high between-individual variation may not be biological significant if it is based on "snap-shot" sampling regimes. This risk can be minimized by several sampling schemes that allow one to establish the temporal consistency of diet variation. The most direct method is to follow individuals through time. Alternatively, a significant phenotype-diet correlation in a snapshot sample provides strong inferential support for consistent diet differentiation but does not guarantee that the quantitative measure is accurate. This is because it suggests that diet variation is due to functional morphology rather than random effects such as patchy prey distribution. Stable isotope ratios have been used to estimate the contribution of different prey to a forager's diet, as prey have characteristic isotope signatures. Isotope ratios in a forager's tissues turn over slowly, so isotope signatures thus represent a long-term average of prey use.

Why should we quantify individual specialization? The obvious answer to that question is that there are large-scale trends in diet variation that reveal more fundamental patterns about trade-offs, character release, and these effects of competition. These trends would not be detected if we simply tested whether diet variation was present or absent using a simple hypothesis testing approach. For example, one study showed experimentally that individual specialization in three-spine stickleback (*Gasterosteus aculeatus*) was stronger when competition was more intense. Furthermore, by quantifying individual specialization, we enable broad scale comparative studies where we can investigate similarities and differences between different taxonomic or geographic groups.

Implications of Diet Specialization

Why is it important to study individual diet specialization? Ecologically, studying individual level diet variation represents a more complete description of a biological system. Information on individual resource use is necessary if we want to make the transition from phenomenological models of population dynamics to mechanistic models in which the dynamics are predicted from the properties of its components (e.g., individual foraging decisions). Evolutionary, since variation is the raw material for evolution by natural selection, intraspecific niche variation thus represents an important target for natural selection. Furthermore, individual specialization, species specialization and similarly generalization are all affected by intra- and inter-specific competition, predation, and prey dynamics, so the topic of food specialization is at the interface of both ecological and evolutionary studies.

Intra- and inter-specific competition, types of social interactions, and the risk of predation or parasitism are factors generally used in describing a population's ecology, all of which can depend on an individual's resource use. Risk factors connected to diet are common because foraging individuals can be vulnerable to predators and parasites associated with a particular diet. Social and competitive interactions between individuals are strongest among individuals that use the same subset of resources especially if different resources are associated to different microhabitats. As a result, populations with large between-individual variation can be divided into subgroups that may compete within themselves but with low between-group competition. Consequently, censuses of total population size will not serve as a good proxy for the level of intraspecific competition. Instead, exploitative competition will be both density and frequency dependent.

Resource-specific ecological interactions mean that individuals within the same population can be subject to different selective pressures. Habitat and resource use affect an individual's energy income, mating options and exposure to risk. Niche variation among individuals is therefore likely to be a major source of variation in fitness and may play a central role in evolutionary diversification. In many models of evolutionary diversification, divergence is driven by disruptive selection because phenotypically average individuals (or generalists) experience disproportionately more intense competition than rare phenotypes (or specialists) with access to exclusive resources. When resource-specific fitness and individual specialization leads to frequency dependent interactions, individual specialization may lead to disruptive selection that will facilitate adaptive speciation. Whether such effects lead to trait evolution and speciation will depend on the heritability of the traits and the temporal consistency of the inter-individual variation.

Summary

Many generalized species are in fact composed of individual specialists that use a small subset of the population's resource use. The most proximate reason for why individuals will specialize on different resources is if they have different resource use efficiencies,

that is, there are efficiency trade-offs between different prey. Different resource use efficiencies may reflect variable morphological, behavioral or physiological capacities to handle alternative resources. Individual diet specialization can occur in a number of ways, they can be related to sexual dimorphism, ontogenetic niches, or related to morphological or behavioral variation in the population. However, not all diet variation needs to be related to variation in efficiencies, diet variation may also be related to variation in social status, mating strategy or territories. Studying diet variation is important both in ecology and evolution. Ecologically, diet variation is important when looking at predator–prey interactions, competitive interactions but also in population dynamical studies. In evolutionary studies, diets of an individual are equivalent to the energy acquisition of the individual and are thus strongly related to fitness. An individual with a relatively high rate of energy (high foraging rate) income will thus have a relatively higher fitness compared to the average individual. Thus, differential foraging rates on different prey types will have a big influence on both the ecological and evolutionary dynamics of populations.

See also: Behavioral Ecology: Optimal Foraging Theory. Conservation Ecology: System Omnivory Index; Trophic Index and Efficiency. Ecological Data Analysis and Modelling: Climate Change Models. Evolutionary Ecology: r-Strategists/K-Strategists. General Ecology: Hunting

Further Reading

Araújo, M.S., Bolnick, D.I., Layman, C.A., 2011. The ecological causes of individual specialization. Ecology Letters 14, 948–958.
Araújo, M.S., Guimarães, P.R., Svanbäck, R., *et al*., 2008. Network analysis reveals contrasting effects of intraspecific competition on individual vs. population diets. Ecology 89, 1981–1993.
Bolnick, D.I., Svanbäck, R., Fordyce, J.A., *et al*., 2003. The ecology of individuals: Incidence and implications of individual specialization. American Naturalist 161, 1–28.
Bolnick, D.I., Yang, L.H., Fordyce, J.A., *et al*., 2002. Measuring individual-level trophic specialization. Ecology 83, 2936–2941.
Britton, J.R., Andreou, D., 2016. Parasitism as a driver of trophic niche specialization. Trends in Parasitology 32, 437–445.
Jones, A.W., Post, D.M., 2016. Does intraspecific competition promote variation? A test via synthesis. Ecology and Evolution 6, 1646–1655.
Polis, G., 1984. Age structure component of niche width and intraspecific resource partitioning: Can age groups function as ecological species? American Naturalist 123, 541–564.
Roughgarden, J., 1972. Evolution of niche width. American Naturalist 106, 683–718.
Shine, R., 1989. Ecological causes for the evolution of sexual dimorphism: A review of the evidence. Quarterly Review of Biology 64, 419–461.
Shine, R., 1991. Intersexual dietary divergence and the evolution of sexual dimorphism in snakes. American Naturalist 138, 103–122.
Skúlason, S., Smith, T.B., 1995. Resource polymorphisms in vertebrates. Trends in Ecology and Evolution 10, 366–370.
Stephen, D.W., Krebs, J.R., 1986. Foraging theory. Princeton, NJ: Princeton University Press.
Sutherland, W.J., Ens, B.J., Goss-Custard, J.D., Hulscher, J.B., 1996. Specialization. In: Goss-Custard, J.D. (Ed.), The oystercatcher. New York: Oxford University Press, pp. 105–132.
Svanbäck, R., Bolnick, D.I., 2007. Intraspecific competition drives increased resource use diversity within a natural population. Proceedings of the Royal Society B 274, 839–844.
Taper, M.L., Case, T.J., 1985. Quantitative genetic models for the coevolution of character displacement. Ecology 66, 355–371.

Relevant Website

R-package, n.d., "R-package to calculate indices of population and individual niche widths"—https://cran.r-project.org/web/packages/RInSp/RInSp.pdf.

Habitat Mapping

Vincent Lecours, University of Florida, Gainesville, FL, United States

Glossary

Abiotic Not living, nonbiological, without any life.
Geomatics Discipline dedicated to the acquisition, analysis, and management of spatially referenced data; includes the technologies and theories of remote sensing, cartography, spatial analysis, geostatistics and geographic information systems.
Habitat Space characterized by distinct environmental characteristics and that is associated with the presence of a particular species or community, or where a particular species or community could be found.
Spatial Autocorrelation (spatial dependence) Quantification of the tendency of spatially closer objects to be more similar than spatially distant objects.
Spatial heterogeneity (spatial nonstationarity) The level of variation of a property across space; whether a variable varies locally or globally.
Surrogate (proxy) A measurable characteristic of the environment that can substitute another one that is more challenging to measure or map.

Introduction

The use of habitat maps has become relatively ubiquitous in certain fields of ecology. Because species have a range of environmental preferences and requirements, habitat mapping approaches focus on the structure and quantity of potential habitats, either instead of, or in addition to, the distribution of biological populations at the time of sampling. The structure and spatial arrangement of habitats can be important predictors of species distribution, abundance, and richness. Habitat maps enable the interpretation of the nature, distribution, and extent of distinct types of environments, and allow predictions of species or communities distribution based on their habitat requirements or associations with the environment. Maps representing the actual or predicted spatial distribution of species and habitats have become key for data integration and synthesis to assist in the study of distribution patterns and ecological dynamics. When properly implemented at the relevant scales, habitat mapping provides accurate, quantitative and spatially explicit information that can inform decision-making. Habitat maps are among the best available spatial decision-support tools and are critical in contexts like conservation, the implementation of scientific management, the monitoring of environmental change, and the assessment of anthropogenic impacts on ecosystems.

The field of habitat mapping has evolved rapidly since the early 1990s (Fig. 1), following technological and methodological developments in geomatics, more specifically in Geographic Information Systems (GIS), spatial analysis methods and remote sensing technologies. Within a GIS environment, spatial analytical techniques are combined with different data to facilitate the

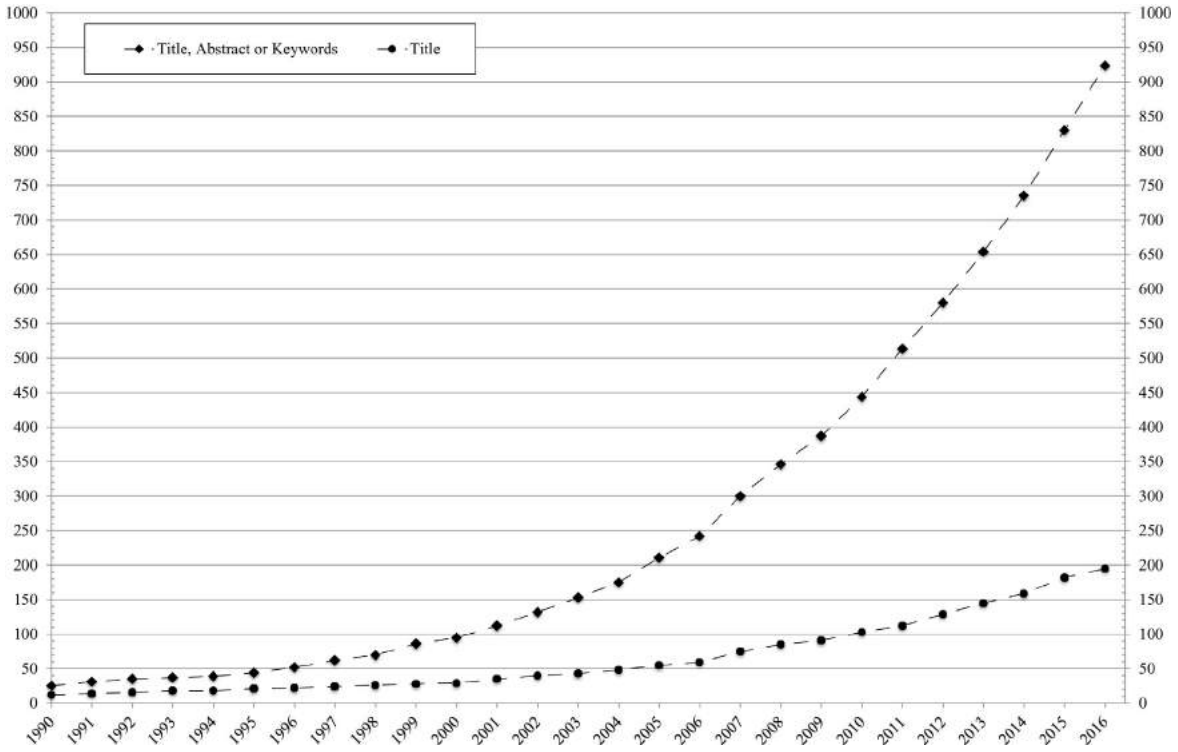


Fig. 1 Cumulative number of articles and reviews listed in the Scopus databased by the end of 2016 with the term "habitat map*" in their title ($n = 195$), or in their title, abstract or keywords ($n = 924$).

quantification and spatial representation of habitats, which provides a framework for mapping species/communities distribution and interpreting spatial patterns of biodiversity. The different approaches to habitat mapping thus generally involve the combination of various types of data (e.g., species occurrences, topographic and climatic variables) into a same geographic framework.

Definitions

Habitat

The concept of habitat has become increasingly used in ecology since the mid-1970s (Fig. 2). Many definitions of this concept have been proposed in the literature (Table 1), and many authors have called into question the ambiguous uses of the term. Some of the older definitions strictly define habitats as vegetation types, while others involve the condition that a species can survive and reproduce in the habitat, thus leaning towards the concept of ecological niche. A common element to many of those definitions is that of a multidimensional space associated with the presence of a species, the multiple dimensions representing characteristics of the physical environment. More recent definitions have generalized the scope of habitats by explicitly integrating the chemical environment—particularly relevant for aquatic and marine habitats—and by not requiring the presence of a species, thus delineating potential habitats that can be driven by different factors like vegetation types, substrate types, or altitude. The use of techniques from the spatial sciences and developments in spatial ecology have led to the recognition that habitats need to also be defined by their spatial and temporal components as opposed to only by environmental characteristics.

Mapping

Mapping, or cartography, is recognized to be as much an art as it is a science. Mapping as a science helps to define quantifiable spatial relationships (e.g., distance, relative position) among the different objects represented in maps. Maps display a selection of objects or phenomena that have been generalized, for instance by reduction, simplification or classification. The concern for the artistic component of mapping is now more about the efficiency of communicating information without jeopardizing simplicity than it is for purely artistic reasons—although maps need to be visually appealing to be effective means of communication. They are recognized to be a unique and powerful form of communication when properly drawn or produced. Good maps render the displayed information accessible to a wide variety of readers by putting the information into a common frame of reference, in this case, a geographic one, that can be read and understood by many. These characteristics of maps have made them very popular as

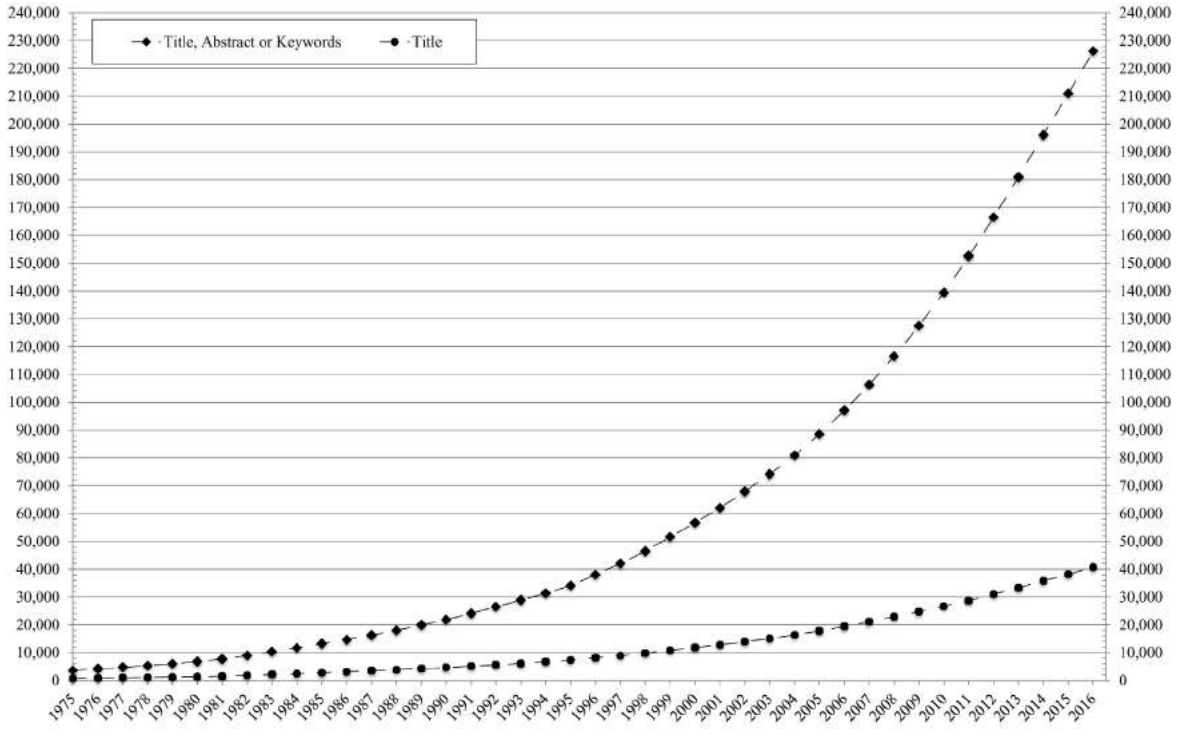


Fig. 2 Cumulative number of articles and reviews listed in the Scopus databased by the end of 2016 with the term "habitat" in their title ($n = 40{,}735$), or in their title, abstract or keywords ($n = 226{,}232$).

spatial decision-support tools in fields such as conservation and management as they facilitate communication between scientists and other stakeholders like land users or decision-makers.

The science of mapping has evolved dramatically in the last 50 years, particularly with the emergence of GIS, and digital and remotely sensed data. The increased ability to store more data in geodatabases and availability of user-friendly GIS have put mapping within the reach of scientists from a wide range of backgrounds and expertise. Developments in newer types of remote sensing (e.g., LiDAR, acoustic remote sensing) in recent decades has enabled the collection of spatial environmental data at an ever-increasing resolution, providing details on the environment in greater quantity and of better quality. These data have revolutionized habitat mapping and are invaluable for the scientific community to advance its understanding of various types of ecosystems.

Habitat Mapping

A simple etymological analysis of the term "habitat mapping" highlights the multidisciplinary and interdisciplinary nature of this field; habitat mapping integrates the ecological concept of habitat into a spatial framework for analysis and representation. Depending on contexts, habitat mapping can integrate theories and methods from the spatial sciences and ecology with those of fields such as geology, climatology, chemistry, oceanography, geomorphology, soil sciences, and many more. The adequate and simultaneous spatial representation of the different biological, physical, chemical and ecological elements of species and the environment make the accurate mapping of habitats challenging. Consequently, cartographic generalization techniques like classifications are commonly used to statistically regroup these characteristics into distinct classes organized such as the similarity inside each of them is maximized and the similarity between them is minimized.

Three broad categories of habitat maps are commonly recognized: single species habitat maps, community maps, and abiotic (or potential) habitat maps. The purpose of the map usually defines which of these types is produced, although habitat mapping is often data-driven and thus also depends on the available data. Single species habitat maps are regularly associated with the characterization of a species' fundamental niche. Community maps often link a particular combination of environmental variables with all the species forming a community. The interpretation of this type of habitat maps can be more intricate than for single species maps, as it requires a greater understanding of the biological and ecological interactions among the different species of the community. Multivariate statistics like multidimensional scaling or analysis of similarity (ANOSIM) can be used to regroup species into communities based on their similarity in spatial distribution and/or associations with the environment. Abiotic habitat maps, or maps of potential habitats, are based on the assumption that variations in biological communities follow changes in the abiotic environment. Such maps are often produced in areas with limited biological observations (e.g., the deep sea) to delineate broadly defined potential habitats.

Table 1 Sample of different definitions of the term "habitat" in the literature

Sources	*Definitions*	*Contexts*
Abercrombie et al. (1961) ISBN: 9780140510034	Place with a particular kind of environment inhabited by organism(s)	Biology (Dictionary)
Milne and Milne (1971) ISBN: 9780045740116	The place where an animal or plant is characteristically found, [. . .] recognizable from information about the nonliving environment	Ecology (Textbook)
Odum (1971) ASIN: B01MS2RT0N	The place where an organism lives, or the place where one would go to find it	Ecology (Textbook)
Villee (1972) ISBN: 9780030253836	The natural abode of an animal or plant species; the physical area in which it may be found	Biology (Textbook)
Whittaker et al. (1973) https://doi.org/10.1086/282837	The range of environments or communities over which a species occurs	General
Bates and Jackson (1984) ISBN: 0-385-18101-9	The environment in which the life needs of a plant or animal are supplied	Geology (Dictionary)
Hutto (1985) ISBN: 9780121780814	A spatially contiguous vegetation type that appears more or less homogeneous throughout and is physiognomically distinctive from other such types	Physiological ecology: land birds
European Environmental Agency (2014) ISSN: 1725-2237	Terrestrial or aquatic areas distinguished by geographic, abiotic and biotic features, whether entirely natural or semi-natural	Terrestrial habitat mapping
Hall et al. (1997) www.jstor.org/stable/3783301	The resources and conditions present in an area that produce occupancy—including survival and reproduction—by a given organism	General
Allaby (1998) ISBN: 9780199567669	The living place of an organism or community, characterized by its physical (for plants) or vegetative (for animals) properties	Ecology (Dictionary)
Krebs (2002) ISBN: 9780321507433	Any part of the biosphere where a particular species can live, either temporarily or permanently	Ecology (Textbook)
Mader (2004) ISBN: 9780070284814	Place where an organism lives and is able to survive and reproduce	Biology (Textbook)
Kearney (2006) https://doi.org/10.1111/j.2006.0030-1299.14908.x	A description of a physical place, at a particular scale of space and time, where an organism either actually or potentially lives	General
Campbell et al. (2009) ISBN: 9780321489845	A place where an organism lives; an environmental situation in which an organism lives	Biology (Textbook)
Knox et al. (2010) ISBN: 9780074716649	The environment of an organism; the place where it is usually found	Biology (Textbook)
Harris and Baker (2012b) ISBN: 9780123851406	Ecological or environmental area that is inhabited by a particular species of animal, plant, or other type of organism	General
Harris and Baker (2012a) ISBN: 9780123851406	Physically distinct areas [. . .] associated with suites of species (communities or assemblages) that consistently occur together	Benthic habitats
Hickman et al. (2012) ISBN: 9780073524252	The place where an organism normally lives or where individuals of a population live	Animal diversity (Textbook)
Lecours et al. (2015) https://doi.org/10.3354/meps11378	Areas [. . .] that are (geo)statistically significantly different from their surroundings in terms of physical, chemical and biological characteristics, when observed at particular spatial and temporal scales	Benthic habitats

Over broad areas, habitats are often delineated according to classification schemes, which can integrate different types of habitat maps. Classification schemes are often hierarchical, ranging from the delineation of the physical environment at a broad scale (e.g., physiographic provinces) to the delineation of finer-scale habitats. Abiotic habitat maps are thus used to define the broad levels of the schemes, while community maps and single community maps can be associated with lower levels of classification schemes (e.g., bogs, fens, seagrass meadows, soft substrate). An example of such classification scheme used for community habitat mapping is the European Nature Information System (EUNIS), which is based on six hierarchical levels that classify both the terrestrial and marine environments.

Surrogacy

Unlike broadly defined, well-known and easily mappable habitats (e.g., shallow-water coral reefs, croplands), many habitats are challenging to define. Due to the difficulties associated with sampling ecological data at sufficient spatial and temporal resolutions, habitat mapping practices often rely on the concept of surrogacy to understand species distribution and ecological processes. A surrogate, or proxy, is a measurable characteristic of the environment that can substitute another one that is more challenging to measure or map. For many habitat maps, the characteristic that is more difficult to measure or map is an element of biodiversity (e.g., presence, abundance). Surrogates can be sampled in situ at specific locations (e.g., soil content, water pH) or can be measured

continuously or near-continuously to provide broader coverage (e.g., vegetation indices, sea surface temperature). Several types of environmental variables, described below, were found useful for habitat mapping, with differing degrees of importance depending on species, locations, settings, and scales. While the next subsections introduce some of the potential surrogates of species and habitat distribution that have been used in habitat mapping, they do not make for a comprehensive review of all of them. Readers are invited to consult the list of recommended readings for more information.

Types of surrogates

Physical surrogates

The physical environment is arguably the most extensive and most used source of surrogates for habitat mapping. For example, temperature is often used for mapping habitats on land and in the oceans, as it has been associated with reproductive success, growth, and speciation rates, thus having the potential to drive or restrain species distribution. Geology, including the nature, size, and location of substrate, influences vegetation types and global biodiversity on land and is among the main drivers of benthic biodiversity in aquatic and marine environments. Topography and geomorphology, which respectively represent the arrangement of physical features and their shape, are also widely used in all types of environments. These two elements can be quantified using geomorphometry, or terrain analysis, which is the quantitative study of the terrain. Hundreds of algorithms from the geomorphometry literature enable the derivation of terrain attributes from digital terrain models representing elevation or depth. Many user-friendly GIS tools facilitate the derivation of these attributes without requiring prior knowledge of the algorithms behind them. Terrain attributes have been used extensively in ecology and habitat mapping. They have become particularly relevant for marine habitat mapping where bathymetric data (i.e., depth values) are often the only available environmental data. The derivation of terrain attributes thus provides a greater number of potential surrogates to be tested against biological data. Among the most popular terrain attributes, slope, rugosity and topographic position indices are often tested as surrogates as they together quantify terrain complexity. Terrains that are more complex are known to shelter a higher level of biodiversity than less complex ones. Aspect, which measures the orientation of the slope, can act as a surrogate of sun exposure in terrestrial habitat mapping or food supply in marine habitat mapping, and curvature can be used in relation to water flow on land and of sedimentation and current patterns underwater.

A few potential physical surrogates are only relevant for marine habitat mapping, and to a lesser extent aquatic habitat mapping. For instance, current patterns and other hydrodynamics conditions are excellent mediums of transport for elements essential for many marine species, including food and oxygen, and are therefore relevant to test for surrogacy when data is available. Hydrostatic pressure is another potential surrogate only applicable to marine studies as the physiology of many marine organisms is pressure-dependent. Pressure affects proteins, biological membranes and rates of enzymatic catalysis, thus having the potential to restrict habitat for particular species.

Chemical surrogates

Potential surrogates originating from the chemical environment are underused in habitat mapping, likely due to the difficulty to obtain continuous coverage of the relevant variables using remote sensing. Chemical surrogates are particularly relevant for habitat mapping of vegetation as plant growth and survival are dependent on particular pedological characteristics (e.g., soil pH, nutrients availability). In aquatic and marine environments, oxygen concentration is a potential driver of species distribution as some species have a low tolerance for hypoxic conditions. Many skeleton-forming organisms like corals also require specific chemical conditions to grow or survive, particularly in terms of phosphate, calcium carbonate, or salinity.

Biological and ecological surrogates

These types of surrogates are particularly relevant for species or communities associated with the presence of a habitat-forming species (e.g., vegetation, corals); ecological interactions like predation or commensalism can lead to the definition of surrogate for a species based on the nature of the relationship and the presence or absence of the other species. Chlorophyll α concentration at the surface of the ocean can also be used as a surrogate of primary production to determine the estimated nutritional resources available to different organisms and thus defining areas as suitable habitat. Knowledge of specific behaviors (e.g., feeding, growth, reproduction strategies) may help in the delineation of habitats when these behaviors are spatially varying, for instance when patterns of larvae dispersal can be assessed or nurseries identified.

Spatial and temporal surrogates

When expressed as ecogeographical data (i.e., spatial ecological data, or ecological data with a geographic component), potential surrogates have a spatial dimension defined by their latitude, longitude, and elevation (or depth). These spatial variables can themselves be used as surrogates as they sometimes are found to drive species and habitat distributions. For instance, latitude and altitude are known to drive patterns of vegetation distribution globally. A derivative of these components of location is geographical distance, which can be important in defining habitats. For instance, distance from the coast, distance from tree cover, and distance from a reef can be used to characterize transitional habitats or to quantify the influence of a specific habitat type on its surroundings (e.g., edge effect). The third dimension is also important. In some contexts, small changes in elevation or depth can better explain changes in habitat types than larger changes in latitude and longitude. The fourth dimension, time, is rarely used in habitat mapping but can be critical in some contexts. For instance, if there is a seasonal pattern in habitat selection for migratory species, an area may

be a suitable habitat for only a few months of the year. Time can be used as a surrogate of species distribution to capture the varying habitat suitability in temporally explicit habitat maps.

Direct and indirect surrogates

Once an environmental variable is identified as being a surrogate of a particular species or habitat distribution, one needs to be careful with the interpretation of that relationship as it is not necessarily a causal-effect link: it may be an indirect surrogate, as opposed to a direct surrogate. Topography is one of the most common indirect surrogates. On land, topography is often found to drive species and habitat distribution. However, topography influences gradients in moisture, energy, and nutrients across the landscape. These gradients are often the true drivers of species and habitat distribution (i.e., the direct surrogates) while topography is an indirect surrogate. The situation is similar underwater. Topography is often identified as a driver of the distribution of suspension-feeders, but is likely an indirect surrogate since topography influences circulation patterns and accelerates currents that bring food to these organisms and rid them of sediments. The hydrodynamic conditions are thus the direct surrogates of species and habitat distribution and topography is an indirect one.

Spatial variables such as latitude and longitude are often not direct drivers of habitat patterns as they reflect other driving gradients. For instance, the strongest environmental gradients in the oceans, such as temperature, pressure, sedimentation, food availability, and community structure (the direct surrogates), are all associated with changes in depth (the indirect surrogate).

Approaches

The very general approach to habitat mapping involves the spatial integration of different datasets representing potential surrogates and/or species or habitat distribution, most often within a geospatial environment. There are however three specific approaches to habitat mapping: the abiotic surrogate approach, the top-down (or unsupervised) approach, and the bottom-up (or supervised) approach. Each of them uses a different way to combine the different geospatial environmental datasets. In all cases, the available data are integrated into a habitat model. Different models are used with different approaches and types of maps. The choice of the methods is therefore adapted to the purpose and needs of the application.

Abiotic Habitat Mapping

Abiotic surrogate mapping does not consider biological data and produces maps of potential habitats. Maps created with abiotic surrogates allow the identification of broad environmental classes (e.g., geomorphic features) from which species' distribution can then be predicted; abiotic surrogate mapping is primarily based on the environmental attributes—or potential surrogates—deemed relevant for the distribution of the species or communities of interest. When there are no particular species or communities of interest, abiotic habitat mapping can be used in classification schemes. In general, this approach is used to map habitat features that can be resolved within the environmental data, thus not attempting to delineate attributes beyond what the habitat model is capable of resolving.

Habitat models used in the production of abiotic habitat maps are often unsupervised classification algorithms, more specifically statistical clustering techniques, from the remote sensing and digital image analysis literature (Table 2). The environmental data are used as input to the algorithm (Fig. 3), which usually performs a classification by regrouping individual pixels with others that share similar characteristics. If an object-based approach is preferred to a pixel-based approach, the algorithm first segments the dataset, meaning that the algorithm divides the data into distinct spatial units before classifying them. In some cases, segmentation can also be performed manually by expert interpretation, although this practice is dwindling because of the objectivity of algorithms as opposed to the subjectivity of expert interpretation. An expert validation and visual interpretation of the resulting units is however recommended. Like for pixel-based approaches, the different spatial units resulting from the segmentation are then regrouped with other units that share similar characteristics in order to be classified. The interpretation of these classes can be done based on existing knowledge of the area or by analyzing the characteristics of the environmental data within each class. Methods like multivariate ANOSIM, nonmetric multidimensional scaling, and percentage of similarity (SIMPER) can be used to investigate differences between classes and provide an estimation of the quality of the segmentation and classification.

Table 2 Examples of some specific techniques used in habitat mapping

Unsupervised classification algorithms	*Habitat suitability modeling techniques*	*Geostatistical techniques for multiscale analysis*
ISO cluster	Boosted regression trees	Bivariate scaling
K-means	Generalized additive model	Coarse-graining
Mahalanobis classifier	Generalized linear model	Lagging
Minimum distance	Maximum entropy	Scalewise variance
Nearest neighbor	Random forest	Spectral analysis

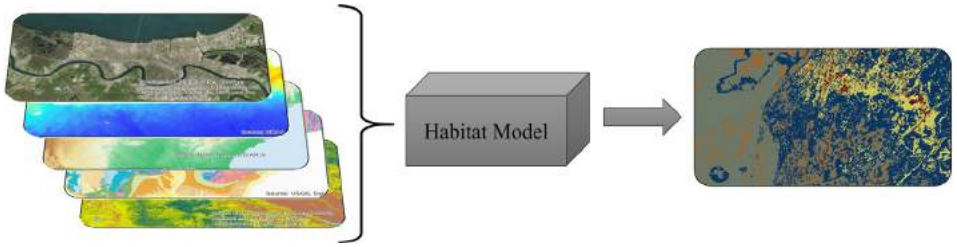


Fig. 3 In the abiotic habitat mapping approach, the environmental data (left) are used as inputs to the habitat model (center), which classifies the data to create a map of potential habitats (right).

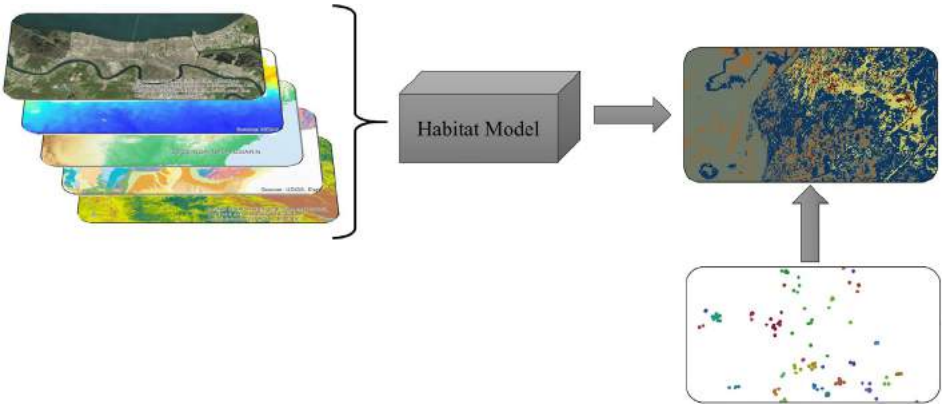


Fig. 4 In the top-down (unsupervised) approach to habitat mapping, the environmental data (left) are used as inputs to the habitat model (center), which classifies the data to create a map of potential habitats (right) that is then spatially compared with ground-truth data or biological observations to be interpreted and produce the final maps. Some of the classes resulting from the application of the habitat model may be further regrouped when compared with the sample data, thus creating a different but final habitat map.

Top-Down Approach

The top-down, or unsupervised approach is very similar to the abiotic surrogate approach: classification and, if needed, segmentation are performed, but the interpretation of the classes is informed by their spatial comparison with ground-truth data or biological observations (Fig. 4). When the ground-truth data are not of biological nature (e.g., sediment grain size), it then produces abiotic habitat maps (not to be confused with the abiotic surrogate approach to habitat mapping). If compared to biological observations, then single species habitat maps and community habitat maps are produced.

Maps produced using this approach can be validated and their accuracy measured with confusion matrices. Confusion matrices are built by spatially comparing the predicted habitat classes from the classification to the actual habitat classes provided by the ground-truth data or biological observations. A number of measures that quantify classification accuracy can be derived from a confusion matrix. The overall, producer and user accuracies, the kappa coefficient, the true/false positive/negative, and sensitivity and specificity are among the most commonly used measures of accuracy assessment.

Bottom-Up Approach

The bottom-up, or supervised approach, combines the ground-truth data or biological observations with the environmental data in the habitat model rather than after the model is applied, meaning that the segmentation of the environmental data into spatial units is informed and driven by the ground-truth data or biological observations (Fig. 5). Those data are usually divided between a training dataset and a validation dataset: model validation is critical in habitat mapping and is used to assess whether or not a model meets the performance requirements of its intended use. In its simplest form, the bottom-up approach to habitat mapping can be applied using simple statistics (e.g., correlations), multivariate statistics (e.g., linear discriminant function), and remote sensing algorithms for supervised classifications (e.g., maximum likelihood). With these techniques, the classification of the data is informed by the training samples, which are used to define the environmental characteristics, or signature, at their location. Then, the rest of the environmental data is classified according to this signature, and the validation samples are used to quantify the accuracy of the classification. The use of these techniques has however diminished in recent years because of the upwelling of habitat suitability models developed in ecology.

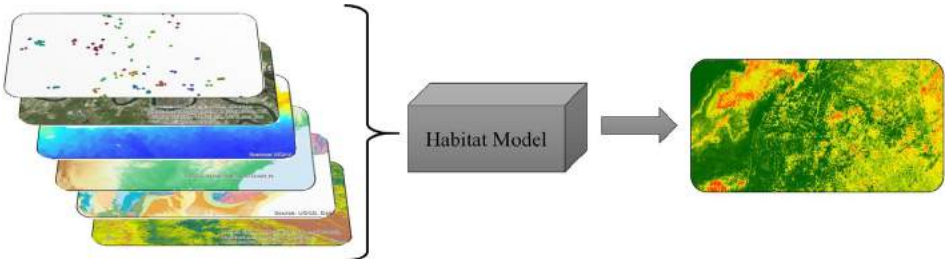


Fig. 5 In the bottom-up (supervised) approach to habitat mapping, the environmental data and the ground-truth or biological data (left) are used as inputs to the habitat model (center) to create the habitat map. In this example, the map represents habitat suitability.

Habitat suitability models, also called ecological niche models, species distribution models, resource selection functions or bioclimatic envelopes, have become the most common applications of the bottom-up approach to habitat mapping. These models can produce single species habitat maps or community maps by modeling the location of the species or community as a function of the environmental data. These models can then be used to predict the extent and location of the species or community distribution in areas where only environmental information is known, whether it is within the geographical area of the existing samples (i.e., interpolation), or outside the range of existing samples (i.e., extrapolation). While there are many types of habitat suitability modeling techniques operating on different frameworks (e.g., statistical, tree-based, machine learning), new models are frequently being introduced (Table 2). The wide availability of techniques and the lack of consensus on which one to use make the choice of modeling techniques challenging. The selection of the appropriate techniques depends on the availability and the spatial or temporal resolution of data in relation to the geographic extent of the study area. Models are often developed for a specific type of data, whether it was for use with presence-only data, presence/absence data, or abundance data. The criteria to consider in the selection of a model for a particular application are (1) species characteristics, (2) data availability, (3) spatial scale, (4) stability in time, and (5) the ecological underpinnings of the model.

Validation techniques specific to habitat suitability models have been developed. Models are expected to have some level of prediction error, and although ecological representativeness and model credibility should be considered, the validation of habitat suitability model usually focuses on the model's predictive performance. While the area under the curve (AUC) of a receiver operator characteristic (ROC) plot is the most commonly used performance measured, it has been highly criticized in the last decade.

Outcomes and Further Analyses

In general, two types of outcomes can result from the application of these approaches: habitat maps with classes represented as groups of pixels associated with a particular habitat type, or habitat maps in which each pixel represents a quantitative measurement such as the probability of occurrence, the relative likelihood of occurrence, or predicted abundance. If needed, the latter type can be categorized (e.g., 0%–25%, 25%–50%, 50%–75%, 75%–100%) to produce the former. The first type of outcome can also be simplified and transformed into vector data for esthetic purpose, although this may alter the accuracy of the displayed information.

Depending on their purpose, habitat maps are not always the final product of an analysis and can be used in further spatial analyses. For instance, they can be overlaid with other data to establish relationships between habitats and other spatial phenomena (e.g., human activities, the distribution of different species, municipal zoning). Habitat maps can also be used with other maps to assess temporal changes in habitat distribution, given that they were produced with the same methodology. They can also be used to analyze patterns in spatial and landscape ecology. Habitat-based approaches to estimate organism response to landscape heterogeneity have been used for decades in landscape ecology. The environmental heterogeneity represented in habitat maps is analyzed with spatial pattern metrics that quantify the spatial structure of the environment and delineate patch-based models of habitat type. The size of habitat patches can be an indicator of the spatial scale at which species use an environment when linked to species distribution and behavior.

Particular Considerations for Habitat Mapping

Scale

Species-environment relationships and ecological patterns and processes are scale-dependent, making habitats scale-dependent. A habitat occurring at a specific spatial or temporal scale may not occur at another one, and no single scale is suitable for the study of all habitats. The importance of scale became widely acknowledged in the 1980s and research on scale remains of utmost importance in many sciences. It is critical for the elements being mapped to be situated in context with the appropriate scales. Many types of scale are recognized in the literature. The types relevant for habitat mapping are spatial, temporal, thematic, management and social scales. Spatial and temporal scales commonly refer to the spatial and temporal characteristics of an object or process, including both

its resolution (i.e., the level of spatial detail or the length of the unit of time) and extent (area or range of time). These two types of scale can be further divided into three types: ecological (or intrinsic), observational, and analytical. The ecological scale is that at which a pattern of process occur, while the observational scale is that at which we measure this pattern or process. Consequently, the observational scale is often associated with the resolution and extent of environmental data. The analytical scale is the scale at which the data are analyzed, for instance in focal statistics used to measure terrain attributes. The analytical scale can also be related to the minimum mapping unit, which is characteristic of habitat maps used in a context of landscape ecology to determine the smallest area to be mapped as a discrete unit for the calculation of spatial pattern metrics. The thematic scale is linked to the level of organization at which objects of study are described (e.g., family vs. genus). Management and social scales are not directly related to the production of habitat maps but are important when the maps are used in some particular contexts. Management scale is related to spatial, temporal and thematic scales as management plans and actions can have a spatial, a temporal and a thematic component. Finally, the social scale is the scale at which people or industry use resources.

Scale has always been a challenge in both ecology and the spatial sciences, making it even more so in a discipline like habitat mapping that integrates them. Scale issues have been widely reviewed in the literature (see further readings). It is now widely recognized that species-habitat relationships are scale-dependent, that environmental characteristics may act as surrogates for species distribution as some scales but not at others, and that a spatial continuum-based approach should be used to identify the ranges of scales over which organisms associate with their habitat. The observed relationships between any two variables can vary across the different types of scales, and therefore understanding the effects of scale in habitat mapping is challenging but essential. Issues can arise when there is a mismatch between ecological, observational, and analytical scales as the appropriate detection of species-habitat relationships and ecological patterns and processes is dependent on the observational and analytical scales to encompass the ecological scales of the habitat being mapped. Producing habitat maps using unsuitable scales for the area or species being studied may affect map accuracy and model performance and reliability. Since habitat maps are used as spatial decision-support tools in some particular contexts (e.g., to inform management or conservation), the situation can be more complex as the appropriate matching of methodological (observational and analytical), ecological, social and management scales is needed to be effective. Much work remains to fully understand the underpinnings of attempting to match the many different types of scale to produce sound science, better inform decisions, and answer the needs of different stakeholders.

Habitat maps are often produced at one particular scale, namely the observational scale at which data were available, and thus regularly fail to capture multiscale environmental drivers of habitat and species distribution. However, recent years have seen the rise of methods for the adoption of a multiscale perspective in habitat mapping that would guide the choice of objective and nonarbitrary methods to select observational and analytical scales. The most relevant scales will vary depending on the study area, organisms considered, and environmental variables available. Some of the suggested approaches are truly "multiscale" and integrate data from multiple scales simultaneously in the habitat model, while others are "multiple scale approaches" that integrate data into a habitat model at multiple successive scales, giving the possibility to combine the multiple maps at different scales in further analyses. Geostatistics, or spatial statistics, put multiscale analysis on a sound mathematical basis (Table 2). Current limitations for a proper implementation of multiscale approaches are associated with data collection methods, which limit the resolution and extent of data, and the lack of availability of user-friendly tools to easily implement such analyses. Until such tools become available, scale information should always be explicitly stated as part of the metadata, and habitat mapping efforts should always be associated with a scale assessment.

Data and Map Quality

Another element that requires careful consideration in habitat mapping is data quality, which encompasses concepts of accuracy, precision, and uncertainty and can be of different types (e.g., spatial, thematic, and temporal). All data carry errors and uncertainty, and the parameters and structure of habitat modeling techniques themselves can introduce uncertainty in an analysis. Understanding data limitations is crucial when integrating them into a workflow as errors and uncertainty propagate throughout analyses and ultimately affect their outcomes. While research has been done on individual components of data quality (e.g., positional accuracy, uncertainty), there is still a general misunderstanding of how all the components integrate and add up together throughout the workflow and to influence habitat maps. While there are no readily available tools to assist in comprehensive data quality assessments, it is important that habitat map producers do not simply disregard it since it is an inherent part of any ecological analyses. Known limitations of data and methods should be acknowledged in the metadata associated with habitat mapping efforts, together with their potential impact on analyses. Such assessment should be made both at the data level before they are integrated into the analysis, and at the end of the workflow on the habitat maps. The assessment of the data before they are used in the habitat model may assist in deciding if the data should be at all integrated into the model. Measures of map accuracy (e.g., commission and omission errors) should always be reported. The need for maps of ignorance has been acknowledged in the literature; maps of ignorance spatially display errors and uncertainty that are associated with habitat maps. They can be provided as probabilistic outputs that account for model uncertainty, or by producing and combining multiple habitat maps with, when possible and appropriate, multiple habitat models. The comparison of the multiple maps may help identifying areas that are consistently classified as the same habitat type, thus indirectly providing a measure of confidence in the mapping outcome. Combining multiple maps can reduce uncertainty and help identify and understand data and ecological patterns where maps agree or disagree with each other. Ensemble techniques are promising methods for the implementation of these solutions.

Other Spatial Concepts

By its spatial and data-driven natures and the near ubiquitous use of GIS, remote sensing and spatial analysis in its workflow, habitat mapping and its practices are directly influenced by spatial concepts. Beyond the questions of scale, considering spatial properties of the data is vital in understanding ecological complexity in all types of habitats. The spatial attributes of measurements are thus essential to be considered when making habitat maps. Two of these attributes are spatial heterogeneity (spatial nonstationarity) and spatial autocorrelation (spatial dependence). They are properties of most ecogeographical data: spatial heterogeneity refers to the level of variation of a property across space, i.e., if an observed variable varies locally or globally, while spatial autocorrelation is the quantification of the tendency of spatially closer objects to be more similar than spatially distant objects. These two properties can affect the measurements of observed species-habitat relationships, particularly when regular statistical methods (e.g., those based on assumptions of independence between variables and identical distribution) are used. Spatially aware statistical methods (e.g., locally and geographically weighted methods) and techniques from geostatistics (e.g., factorial cokriging) should be adopted more consistently in habitat mapping to produce maps and measurements of species-habitat relationships that are sound and consider the spatial effects of the data.

Another concept from the geospatial literature from which habitat mapping practices would greatly benefit is the concept of fuzzy logic. Most habitat mapping efforts impose discrete boundaries to habitat types. However, it is well recognized that the natural environment rarely shows sharp transitions from a habitat type to another—except in the presence of anthropogenic disturbances. In order to improve spatial representation of habitats, there is a need to characterize these transition zones between habitat types. Theories from fuzzy logic demonstrate a lot of potential for application to habitat mapping. Instead of using a binary system in which a location either is a habitat type or is not, fuzzy logic quantifies the probability of that location to be any of the proposed habitat types, thus considering the possibility that the location could be of two or more types. An area having a relatively equal probability to be of two or more types can then be interpreted as being an area of transition between those particular habitat types. Some geostatistical methods have been developed to integrate the concept of fuzzy logic in analyses and should be used more often in habitat mapping.

Ecological Representativeness and Fitness for Use

Not all data are equally good at capturing the relevant information, and not all information is relevant for a particular species, habitat or area. Habitat maps are highly sensitive to the methods used to produce them and to the data selected, which includes their nature, quality and scale components. A critical evaluation of the data has to be performed before their inclusion into a habitat model. Rather than selecting data based on availability, the fitness for use of these data has to be assessed in terms of ecological, biological or environmental relevance (including scale), but also in terms of quality. An inappropriate selection of data can produce results that do not accurately represent the studied habitats. Requirements for the assessment of fitness for use is different based on the type of habitat map and approach used as it requires an intended purpose and specific application goals or questions. Abiotic habitat maps may not require data to be as much ecologically representative as data used for single species habitat maps, which should consider variables that are realistic for the focal species' life history and traits. The assessment of fitness for use may also guide the number of variables to be included in an analysis: too few variables can result in overly general habitat characterizations while too many variables can reduce the generalization and reproducibility of the approach. Fitness for use assessments also need to consider the scale of the data: it is as important to identify the relevant environmental factors as it is to identify the scales at which these factors drive species distributions. In addition, there is often a trade-off between spatial scale and data quality.

Conclusions

The application of geomatics-based habitat mapping to ecological questions is relatively recent and new developments and techniques are commonly introduced (e.g., object-based image analysis, ensemble mapping). In the near future, standards and protocols are likely to be updated to integrate new tools and methods that facilitate the implementation (1) of multiscale approaches that consider the spatial nature of data, (2) of error and uncertainty quantification in data and analyses, (3) of fuzzy logic techniques to study habitat edges and transitional habitats, and (4) of dynamic cartographic representations that provide a better multiscale, multi-temporal, and multidimensional representation of habitats. Assessments of fitness for use should always be performed, and complete metadata that explicitly state scale information, error and uncertainty quantification, and any other relevant information should always be associated with data and habitat maps. The availability of tools that streamline the workflow from data collection to analysis, in addition to data quality assessment, will be key in enabling scientists with a wide range of background and experience to produce habitat maps that are grounded on a sound statistical and spatially explicit basis. Finally, given their sensitivity to data and methods characteristics, habitat maps should be critically interpreted to try to understand not only the ecological patterns and processes that may have been captured by the habitat mapping, but also the cartographic patterns that are dependent on the procedures rather than the subject of the mapping.

References

Abercrombie M, Hickman CJ, and Johnson ML (1961) *The penguin dictionary of biology*, 284 p. London: Penguin Books.
Allaby M (1998) *Oxford dictionary of ecology*, 432 p. Oxford: Oxford University Press.
Bates RL and Jackson JA (1984) *Dictionary of geological terms*, 3rd edn., 576 p. New York: Anchor.
Campbell NA, Reece JB, Taylor MR, Simon EJ, and Dickey JL (2009) *Biology: concepts and connections*, 6th edn., 928 p. San Francisco: Benjamin Cummings.
European Environment Agency (2014) *Terrestrial habitat mapping in Europe: an overview*, 154 p. Copenhagen: European Environment Agency Publications Office.
Hall LS, Krausman PR, and Morrison ML (1997) The habitat concept and a plea for standard terminology. *Wildlife Society Bulletin* 25(1): 173–182.
Harris PT and Baker EK (2012a) *Seafloor geomorphology as benthic habitat: GeoHAB atlas of seafloor geomorphic features and benthic habitats*, 900 p. London: Elsevier.
Harris PT and Baker EK (2012b) Why map benthic habitats? In: Harris PT and Baker EK (eds.) *Seafloor geomorphology as benthic habitat: GeoHAB atlas of seafloor geomorphic features and benthic habitats*, pp. 3–22. London: Elsevier.
Hickman CP, Keen SL, and Roberts LS (2012) *Animal diversity*, 479 p. McGraw-Hill.
Hutto RL (1985) Habitat selection by nonbreeding, migratory land birds. In: Cody ML (ed.) *Habitat selection in birds (physiological ecology)*, pp. 455–476.
Kearney M (2006) Habitat, environment and niche: What are we modelling? *Oikos* 115(1): 186–191.
Knox B, Ladiges PY, Evans BK, and Saint RB (2010) *Biology: an Australian focus*, 1232 p. Sydney: McGraw-Hill.
Krebs CJ (2002) *Ecology: the experimental analysis of distribution and abundance*, 5th edn., 688 p. Harlow, Essex: Pearson.
Lecours V, Devillers R, Schneider DC, Lucieer VL, Brown CJ, and Edinger EN (2015) Spatial scale and geographic context in benthic habitat mapping: Review and future directions. *Marine Ecology Progress Series* 535: 259–284.
Mader SS (2004) *Biology*. Maidenhead: McGraw-Hill Education.
Milne LJ and Milne MJ (1971) *The arena of life: the dynamics of ecology*, 350 p. New York: Doubleday/Natural History Press.
Odum EP (1971) *Fundamentals of ecology*. Philadelphia: Saunders College Publishing/Harcourt Brace.
Villee CA (1972) *Biology*. New York: Harcourt Brace College Publishers.
Whittaker RH, Levin SA, and Boot RB (1973) Niche, habitat, and ecotope. *The American Naturalist* 107(955): 321–338.

Further Reading

Bamford MJ and Calver MC (2014) A precise definition of habitat is needed for effective conservation and communication. *Australian Zoologist* 37: 245–247.
Brown CJ, Smith SJ, Lawton P, and Anderson JT (2011) Benthic habitat mapping: A review of progress towards improved understanding of the spatial ecology of the seafloor using acoustic techniques. *Estuarine, Coastal and Shelf Science* 92: 502–520.
Carl G and Kühn I (2017) Spind: A package for computing spatially corrected accuracy measures. *Ecography* 40: 675–682.
Ferree CF and Anderson MG (2013) *A map of terrestrial habitats of the northeastern United States: Methods and approach*. Boston: The Nature Conservancy.
Greene GH, Bizarro JJ, Tilden JE, Lopez HL, and Erdey MD (2005) The benefits and pitfalls of geographic information systems in marine benthic habitat mapping. In: Wright DJ and Scholz AJ (eds.) *Place matters: Geospatial tools for marine science, conservation, and management in the Pacific northwest*, pp. 34–46. Corvallis: Oregon State University Press.
Harris PT and Baker EK (eds.) (2011) *Seafloor geomorphology as benthic habitat*. London: Elsevier.
Li X and Wang Y (2013) Applying various algorithms for species distribution modeling. *Integrative Zoology* 8: 124–135.
Monmonier M (1996) *How to lie with maps*, 2nd edn. Chicago: The University of Chicago Press.
Morris LR, Proffitt KM, and Blackburn JK (2016) Mapping resource selection functions in wildlife studies: Concerns and recommendations. *Applied Geography* 76: 173–183.
Pittman SJ and McAlpine CA (2001) Movements of marine fish and decapod crustaceans: Process, theory and application. *Advances in Marine Biology* 44: 205–294.
Scott JM, Heglund PJ, and Morrison ML, et al. (eds.) (2002) *Predicting species occurrences: Issues of accuracy and scale*. Washington: Island Press.
Todd, B. J. and Greene, H. G. (eds) *Mapping the seafloor for habitat characterization*. St. John's: Geological Association of Canada.

Relevant Websites

www.emodnet-seabedhabitats.eu—European Marine Observation and Data Network, seabed habitats.
http://eunis.eea.europa.eu—The European Nature Information System.
https://gapanalysis.usgs.gov—United States Geological Survy National Gap Analysis Program.
www.geohab.org—GeoHab (Marine Geological and Biological Habitat Mapping).
www.geomorphometry.org—International Society for Geomorphometry.

Habitat Selection and Habitat Suitability Preferences

B Doligez, CNRS – Université Claude Bernard Lyon 1, Villeurbanne, France
T Boulinier, CNRS – Centre d'Ecologie Fonctionnelle et Evolutive, Montpellier, France

Why Select or Prefer a Given Habitat?

Definition of Habitat and Habitat Patch

For most species, the environment is heterogeneous at various spatial and temporal scales (Fig. 1). A habitat can be defined by a given type of environment characterized by general physical features (e.g., type of vegetation, water, or soil structures). The habitat of a species can also be defined by the general characteristics of the areas used by individuals, which must be suitable enough for the species' activities. This definition results from the observed spatial distribution of individuals. Finally, at a finer scale, a habitat can also be defined by the portion of the environment devoted to a particular activity of individuals (e.g., breeding or foraging). Usually, the term `habitat' does not encompass conspecifics, that is, the social components of individuals' activities, contrary to the term `environment'. A habitat patch can be defined as a continuous and homogeneous portion of a habitat (Fig. 1).

Definition of Habitat or Patch Suitability and Quality

A habitat is suitable for a species when it contains all resources needed for a given activity in sufficient quantity (e.g., food when foraging, nest sites when breeding). Individuals can only live in suitable habitats; thus by definition, habitats where individuals are found to live or perform a given activity must be suitable. Habitat suitability can be difficult to define using other criteria than the observed repeated and long-term presence of individuals.

Suitable habitats and patches for a given species can differ by some intrinsic characteristics affecting individuals' fitness (e.g., available resource quantity and quality, level of competition for resources or predation). The quality of a habitat or patch is usually defined by the fitness (measured, e.g., by energy gains per time unit or reproductive output) that can be achieved by individuals in this habitat or patch: a habitat or patch in which individuals achieve high fitness is defined as a high-quality habitat or patch, relative to other habitats and patches. Within the framework of foraging and breeding habitat choice, the quality of habitats is usually evaluated in terms of energy intake rate and reproductive success, respectively.

Spatiotemporal Variability of Habitat Suitability and Quality Leads to Habitat Selection

In heterogeneous environments, natural selection will favor individuals capable of occupying the most favorable areas for activities linked to fitness, that is, survival and reproduction. Spatial and temporal environmental heterogeneity (Fig. 2) thus leads to selective pressures favoring behaviors that allow individuals to select high-quality habitats or patches, that is, the evolution of individual strategies of habitat choice, for any activity considered (e.g., foraging for food, searching for a sexual partner, finding shelter from predators, breeding).

Spatial heterogeneity and temporal predictability of habitat or patch quality are required conditions for habitat choice or preference to evolve (Fig. 2). In a homogeneous and equally exploited environment, there is no need for individuals to choose because the expected fitness will be equal in all habitats or patches. Moreover, if the environment is not predictable at the relevant timescale for the activity considered, a location cannot be chosen based on given characteristics since they could randomly change in the time between gathering information on those characteristics and individual decision, thus preventing individuals to achieve the expected fitness.

Habitat choice can have a major impact on fitness. Wrong decisions can lead to highly reduced survival, or complete breeding failure. Habitat or patch choice will be all the more critical as the temporal scale involved is long and individuals' movements are spatially constrained, for example, breeding compared to foraging habitat selection. The habitat used for breeding also determines the conditions to which breeders will be exposed during this period of life, sometimes representing most of an individual's lifetime. Selective pressures on habitat choice are thus strong.

Sources of Variation in Habitat Suitability and Quality

Spatial and temporal heterogeneity in habitat or patch suitability and quality can be due to many different factors linked to abiotic, biotic, and social characteristics. The physical characteristics of the environment that can affect fitness vary from climatic conditions (rain, wind, and temperature regimes), soil nature (e.g., for species that dig burrows), stability of the substratum (e.g., when breeding on a slope or in a tree), level of salinity (for marine species), etc., depending on the species considered. Biotic sources of environmental variation include the availability of biotic resources (e.g., food and nest-building materials), which may

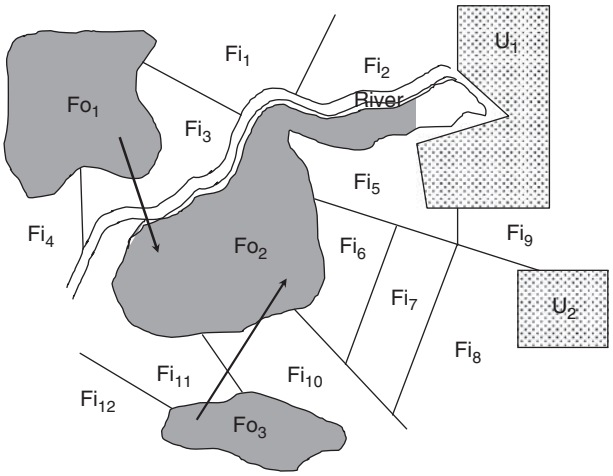


Fig. 1 An illustration of environmental spatial heterogeneity defining habitats and habitat patches. In this schematic farmland landscape, portions of the environment are constituted by a river, fields (delimited by straight lines), forests (gray zones), and urban areas (dashed zones), which have very different general physical characteristics and define four different habitats. Within each habitat type, several continuous and homogeneous subareas can be found and define habitat patches: Fo_1 to Fo_3 (forest), Fi_1 to Fi_{12} (fields), U_1 and U_2 (urban areas). Patches of the same habitat can vary in quality; for instance, different types of crops may generate different amounts of food or breeding sites availability in different fields.

be required in sufficient quantity as well as quality (e.g., required nutrients may only be available in specific food items); predators and parasites are often spatially heterogeneous biotic factors.

Finally, social components can play a major role. The density of conspecifics or heterospecifics exploiting the same resources can vary drastically in time and space, and competitors' presence may reduce the fitness of a given individual directly (e.g., when resources are limited) or indirectly (e.g., via the attraction of common predators). Because conspecifics also have to select and secure resources, their distribution among habitats and patches will affect the relative quality of potentially available resources. Conspecific decisions affect fitness gains expected by individuals choosing a particular habitat or patch (i.e., a frequency-dependent process). Conversely, conspecifics' or heterospecifics' presence can also have positive fitness effects, for example, when individuals interact with each other to capture preys, deter predators, build nests, etc., so that when conspecific density decreases below a certain value, individuals' fitness decreases (Allee effect). Conspecifics' presence can also be beneficial by providing information about habitat or patch quality. Both conspecifics' quantity and quality may vary and thus affect breeding habitat quality. In particular, the relatedness between individuals can affect local habitat quality, through kin competition or cooperation.

In many cases, individuals will require several different critical resources simultaneously. All the fundamental ecological requirements for a given activity thus have to be accounted for. For instance, a breeding patch may provide large amounts of food, but lack breeding sites, and thus will not be used. The different factors affecting fitness are also likely to interact with each other. Furthermore, spatiotemporal variations of important factors likely show different patterns at different scales, generating tradeoffs between factors, since the values of the different factors that maximize individual fitness may not occur in the same locations at the same time. These tradeoffs may themselves differ in time and space.

Scales of Variation in Habitat Suitability and Quality

The importance of a given factor for habitat choice varies with its spatiotemporal variability. Local habitat quality can vary at different scales, both spatially (e.g., between and within habitat patches) and temporally (**Fig. 2**). Detecting spatial heterogeneity and temporal predictability is tightly linked with the scale considered. The environment may be homogeneous at a given spatial or temporal scale (e.g., within patches or hours), but heterogeneous at another scale (e.g., among patches or years). Habitat selection can involve a cascade of nested scales: individuals can first choose among habitat types, then a general area within a habitat type, and then within this area, a patch that may comprise several sites. The relevant spatial and temporal scales at which habitat choice needs to be investigated thus have to be identified. These scales are constrained by habitat heterogeneity itself, but also by the ability of individuals to detect heterogeneity. Individuals may choose between habitats only if they are aware of habitat or patch quality variation. The scales at which an individual perceives spatial heterogeneity (depending, e.g., on its movement ability) constrain habitat selection and define the upper scales of possible habitat choice. Only environmental factors affecting fitness and varying in time and space at the scales individuals can explore are therefore relevant to habitat choice.

Definition of Selection/Choice/Preference

Because of environmental heterogeneity, individuals face alternatives with different fitness outcomes. When confronted with multiple alternative situations, individuals eventually select one of them, and are said to perform a choice or prefer one option

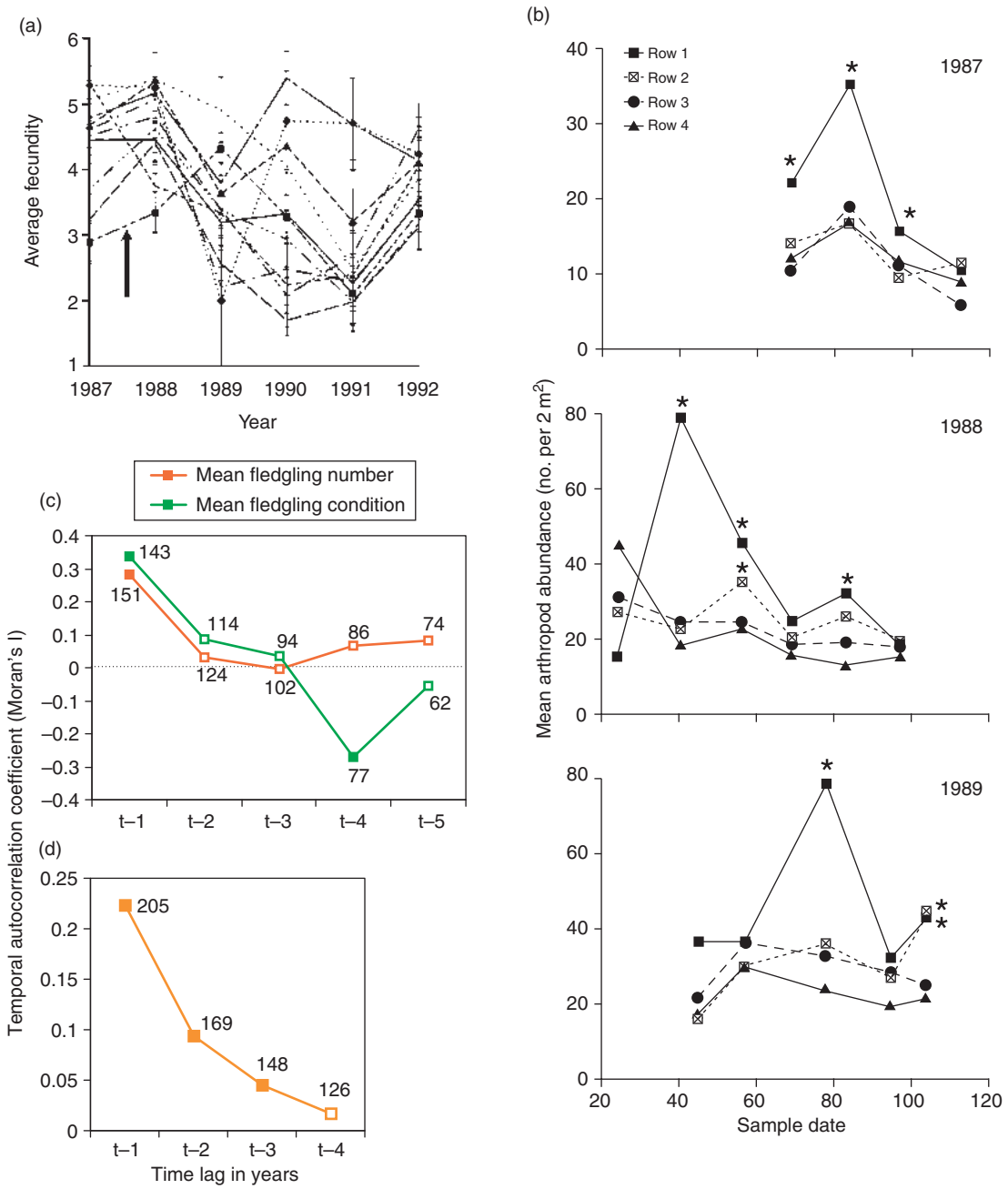


Fig. 2 Illustration of the spatiotemporal variation of habitat quality from several field studies. (a) Mean fecundity (number of young fledged per reproductive event ±1 SE) of collared flycatchers (*Ficedula albicollis*) in a series of 11 forest patches over 6 years. The relative quality of a patch differs between years: for example, the patch indicated by an arrow has the lowest mean fecundity in the first year, but the second highest fecundity 2 years later. (b) Mean prey abundance per 2 m^2 leaf surface in a flooded vegetation along shores that constitutes breeding habitat of prothonotary warblers (*Protonotaria citrea*) over 3 years, according to the localization of breeding site within patch. Sites are located in rows parallel to shoreline, with row 1 closest to, and row 4 farthest from, the shore (stars indicate significant differences). (c) Temporal autocorrelation of mean fledgling number and condition as measures of patch quality in collared flycatchers in a series of 20 forest patches over 20 years (closed symbols: significant coefficients). The autocorrelation fades after a time lag of 1 year. (d) Conversely, the temporal autocorrelation of patch reproductive success (mean number of fledged young per nest) for black-legged kittiwakes (*Rissa tridactyla*) stays significant for 3 years. (a) From Doncaster, C.P.D., Clobert, J., Doligez, B., Gustafsson, L., Danchin, E., 1997. Balanced dispersal between spatially varying local populations: An alternative to the source–sink model. American Naturalist 150, 425–445. (b) Reproduced from Petit, L.J., Petit, D.R., 1996. Factors governing habitat selection by prothonotary warblers: Field tests of the Fretwell–Lucas models. Ecological Monographs 66, 367–387 (c) Data from Doligez, B., Pärt, T., Danchin, E., Clobert, J., Gustafsson, L., 2004. Availability and use of public information and conspecific density for settlement decisions in the collared flycatcher. Journal of Animal Ecology 73, 75–87. (d) Data from Danchin, E., Boulinier, T., Massot, M., 1998. Conspecific reproductive success and breeding habitat selection: Implications for the study of coloniality. Ecology 79, 2415–2428.

when the probability to choose this option is significantly higher than expected by chance, and is affected by the variation in the expected fitness among potential alternatives. In the classic expression `habitat selection' (synonym of `habitat choice'), the term `selection' describes a process of individual choice using some decision rule (despite implying no conscious mechanism), which includes the processes of information acquisition about the quality of alternative habitats or patches, and information use to select the alternative expected to maximize fitness.

Because of competition, all individuals may however not be able to settle in the highest-quality, preferred habitat or patch, that is, the realized choice may not reflect individual's preference. This may be because individuals are prevented from choosing a habitat or patch by dominant competitors despite attempting to do so, or because individuals evaluate competition intensity beforehand and choose not to use these best habitats or patches. Furthermore, individuals may use different strategies than optimal habitat or patch choice, that is, choose a suboptimal patch but compensate for the decrease in fitness using other strategies. For instance, individuals may choose to exploit different food sources, or adopt strategies limiting risks, rather than changing patch.

Which Habitats and Which Constraints for Different Choices?

Foraging Habitat Selection

The issue of finding and exploiting food is crucial. In heterogeneous environments, foraging habitat patches differ in quality depending on the availability of food resources and their intrinsic quality (e.g., nutritive value), but also on the costs associated to exploit these resources, involving access to food items in terms of energy, time, and injury risk when competing, risk of being preyed upon, of getting scrounged by con- or heterospecific competitors, etc. Spatial and temporal scales involved in foraging habitat selection can vary (e.g., from a few seconds up to days in large predators), but in most species, foraging habitat choices are made by individuals a large number of times over their entire lives. Decisions linked with foraging habitat selection include the choice of patches where to start foraging and duration of patch exploitation following progressive resource depletion over time, thus decision to depart to another foraging patch (**Fig. 3**).

Theoretical models have been built and tested empirically to investigate which conditions affect these two types of decisions. Classical optimal foraging theory addresses the patch-time allocation that maximizes an individual's fitness by referring to the marginal value theorem (**Fig. 4**), which states that the optimal strategy is to leave a foraging patch when the instantaneous fitness gain rate from the current patch falls below the average gain rate that can be achieved in the environment. The model predicts that individuals will stay longer (1) in a more profitable patch, (2) as the distance between patches (and thus travel time) increases, and (3) when the environment as a whole is less profitable (**Fig. 4**). The marginal value theorem has been a useful tool but has however been criticized on the grounds that it makes simplified assumptions, in particular that foragers are optimal and have a complete knowledge of resource abundance and distribution in the environment, which is unrealistic. Linking the optimality predictions of the model with proximate mechanisms of patch departure decisions involved is necessary. Simple mechanistic rules for patch-leaving decisions have therefore been proposed and tested experimentally (**Fig. 5**):

(a) *Incremental rule*. The probability to stay in the current patch decreases with unsuccessful search time spent in the patch, but increases each time a resource is found; individuals will find more resources items, and thus stay longer, in rich patches.
(b) *Decremental rule*. The probability to stay decreases each time a resource is found; individuals will thus stay shorter in high-quality patches.
(c) *Giving-up-time rule*. The tendency to stay decreases with unsuccessful search time spent on the patch, but each time a resource item is found, the tendency to stay is reset to a maximum level; individuals leave after a fixed unsuccessful search time (giving-up time).
(d) *Fixed-number rule*. The individual leaves the patch after a given, fixed number of resource items have been found.
(e) *Fixed-time rule*. The individual forages for a fixed period of time in each patch and leaves the patch independently of the number of resource items found.

Which decision rule will be adaptive depends on (1) the spatial distribution of resource items in the environment, which conditions the information about patch quality that can be derived from finding a resource item, and (2) the individual's *a priori* knowledge about the environment (**Table 1**).

Breeding Habitat Selection

Individuals will also have to make a series of habitat choices for breeding. As for foraging, breeding requires availability of high-quality food resources necessary for offspring development, but also many other resources affecting breeding success, in particular the presence of mates and availability of safe breeding sites (**Fig. 6**). In some species, decisions made during the course of breeding are sequential in time and space, and made independently based on different criteria. In other species, all resources have to be secured simultaneously, which may generate tradeoffs between optimal choices for each resource. In many species, the number of breeding attempts is limited over an individual's lifetime, because breeding involves longer timescales than foraging (up to several years), and/or is a seasonal activity implying yearly timescales. Thus constraints associated with breeding habitat selection often

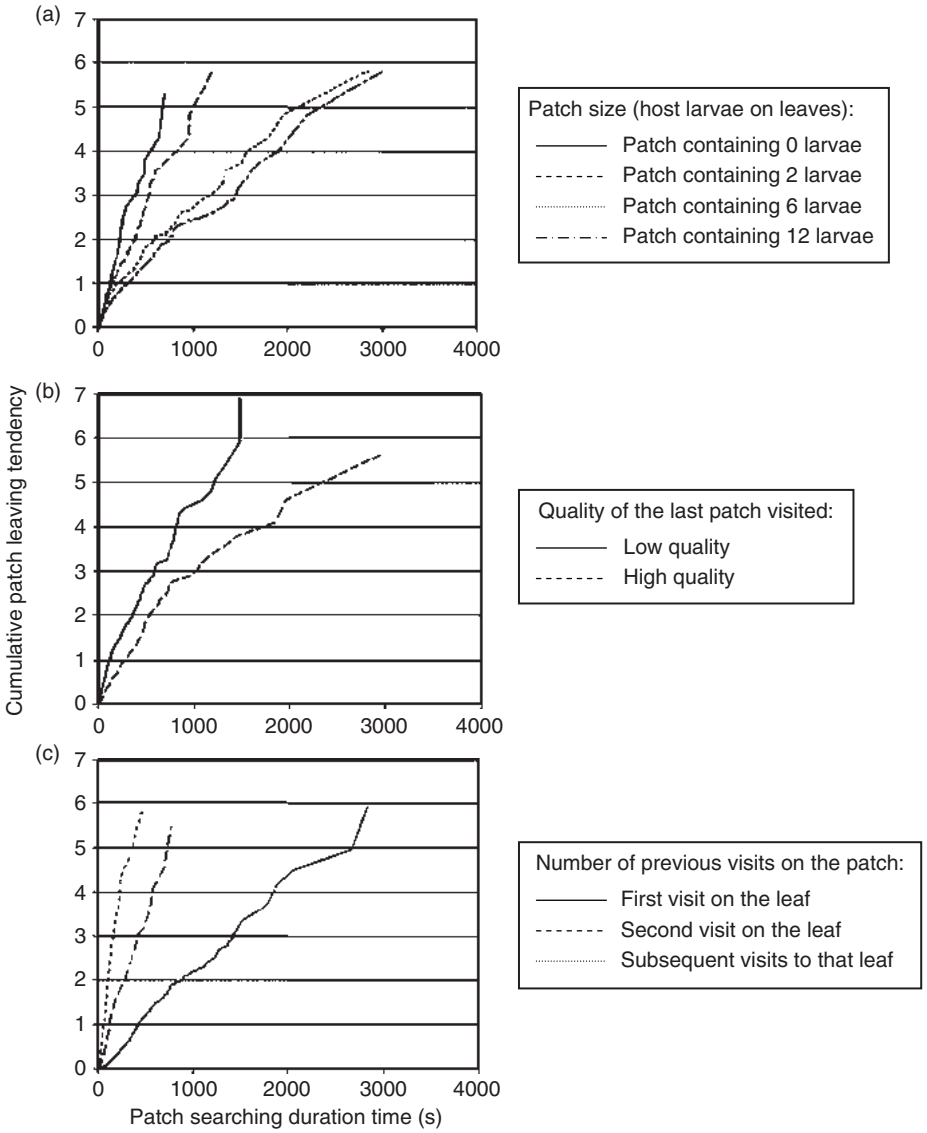


Fig. 3 Foraging decisions in an insect parasitoid, *Aphidius rhopalosiphi*. Females of this parasitic wasp lay eggs in grain aphids *Sitobion avenae*, which are spatially distributed in discrete patches. Parasitic wasps have to adjust their searching time within a given host patch and allocate their foraging time among the different patches available in the habitat, to maximize their fitness. These decisions have to be based upon information on patch quality, obtained through both host encounter rate in the patch and previous searching experience by the wasp on the same or other patches. Patch-leaving decisions in this species depend on (a) host patch size, (b) quality of the last patch visited, and (c) previous experience in the current patch. Females spent more time in a patch when it contains more resources, when they just visited a high-quality patch, and during their first visit in the patch. From Outreman, Y., Le Ralec, A., Wajnberg, E., Pierre, J.-S., 2005. Effects of within- and among-patch experiences on the patch-leaving decision rules in an insect parasitoid. Behavioral Ecology and Sociobiology 58, 208–217.

differ from foraging habitat selection. Habitat choice may vary depending on the type of breeding site and species breeding ecology, and can occur at variable spatial scales, both in absolute values (from millimeters for some parasites to hundred of kilometers for large vertebrates), and relative values, depending in particular on the spatial range used by a given species.

Other Habitat Selection Behaviors

Individuals have to choose a habitat or patch in many other situations. In species where mating occurs in a different place than the remaining of breeding activity (e.g., lekking species), individuals have to select an optimal displaying habitat or patch, and within the patch, an optimal site, for example, close to a dominant male, or within a light spot in low-light-intensity environments. For instance, if visual signals are used in mate choice, the environment chosen to display can affect mating success because signal appearance depends on the joint effect of ambient light and individual's reflectance spectra (Fig. 7). Similarly, individuals may

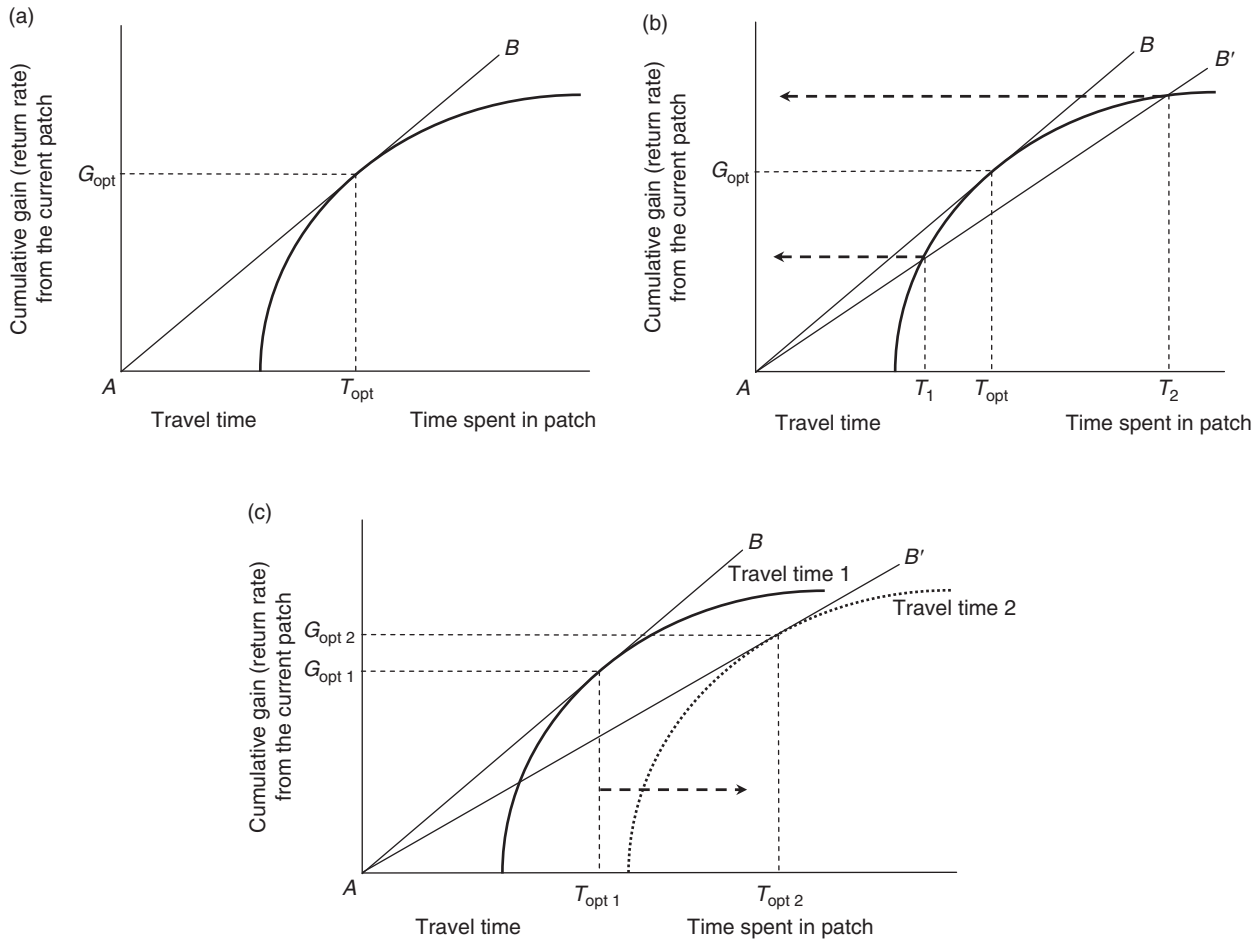


Fig. 4 The marginal value theorem yields the `giving-up time' when an individual should leave its current foraging patch. (a) As an individual forages, its cumulative fitness gain gradually slows down as food becomes scarcer in the patch and food items take longer to find. An individual should aim at maximizing the net rate of fitness (or energy) gain, including time during which it cannot feed because it travels between patches. The rate of fitness gain corresponds to line *A–B*. The steepest slope line (gain/time), which maximizes the rate of energy gain, corresponds to the tangent line to the gain curve. When time on the patch reaches the point of contact between line *A–B* and the gain curve (T_{opt}), the individual should leave the patch. (b) An individual that leaves too early (T_1) will gain less energy per unit of time relative to the maximum (line *A–B′*). Similarly, there is no benefit in staying too long (T_2) as food items are running out. (c) When an individual should stop exploiting its current patch and leave will depend on the travel time between patches, even though the gain curve once on the patch does not change. When travel time is long, individuals should leave the patch after spending more time (T_{opt2}). Adapted from Charnov, E.L., 1976. Optimal foraging: The marginal value theorem. Theoretical Population Biology 9, 129–136.

have to choose among alternative resting habitats or patches. In this case, the main resource is a safe site from predators or a site allowing individuals to optimize energy expenditure (e.g., against cold or rain).

Migration can be considered as an extreme form of habitat selection, when individuals change habitat because resource availability is seasonal while individuals' requirements remain unchanged. However, migration behavior is a fixed habitat selection process, individuals changing habitat similarly year after year. During migration, individuals will choose stopover areas, but this choice can be considered as classical foraging habitat choice, constraints of which include energy requirements and costs of long-distance flights.

Differences between Choices: Spatiotemporal Scales Involved, Tradeoffs

Habitat selection shows fundamental differences depending on the activity considered, in particular in spatiotemporal scales involved. Estimating a foraging patch quality may take only up the time to try to find a food item (**Fig. 5**). The equivalent rule for assessing a breeding patch quality implies attempting to breed to obtain information on expected breeding success in this patch, thus spending time and energy for one breeding attempt there, which may represent a significant portion of life span. In other words, foraging decisions are more dynamic than breeding habitat decisions. Breeding habitat choice may occur only once in life, when individuals decide whether to stay on or leave the natal site (natal site fidelity vs. dispersal). In mobile iteroparous species,

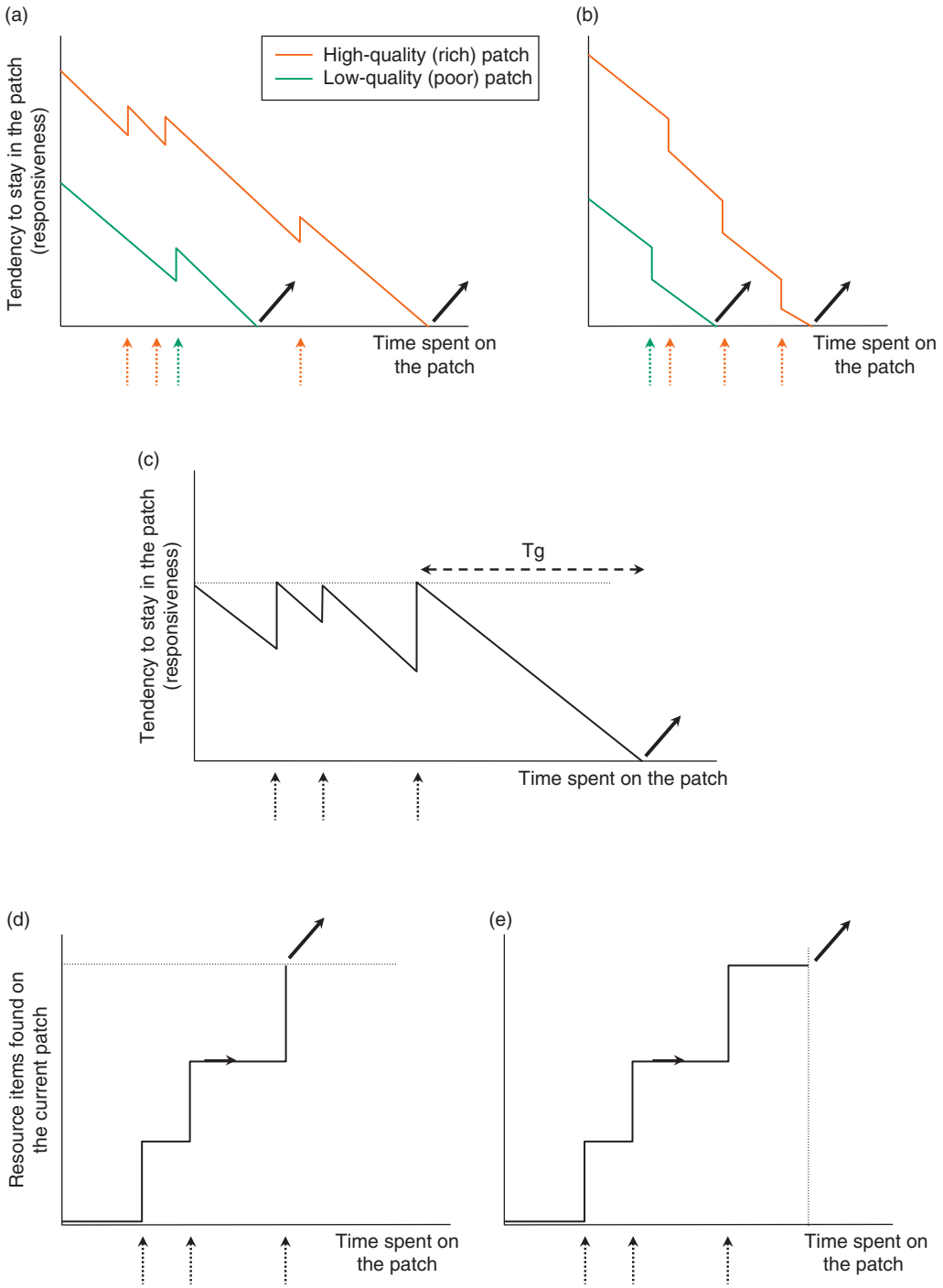


Fig. 5 Patch-departure decision rules in discrete foraging environments. When resources are not distributed evenly but are aggregated in patches whose size is not *a priori* known, individuals can adopt several simple rules for deciding when to stop exploiting their current patch and leave for another patch (black arrows). Dashed arrows indicate when resources items are found by individuals. The probability to stay in the current patch, also called responsiveness, is shown depending on time spent in the patch. When the probability to stay in the current patch drops below a critical threshold, the individual leaves the patch. (a) Incremental rule; (b) decremental rule; (c) giving-up-time rule (giving-up time: Tg); (d) fixed-number rule; (e) fixed-time rule. (a, b) Adapted from Waage, J.K., 1979. Foraging for patchily distributed hosts by the parasitoid. Nemeritis canescens. Journal of Animal Ecology 48, 353–371; and Van Alphen, J.J.M., Bernstein, C., Driessen, G., 2003. Information acquisition and time allocation in insects parasitoids. Trends in Ecology and Evolution 18, 81–87. (c) From Stephens, D.W., Krebs, J.R., 1987. Foraging Theory. Princeton, NJ: Princeton University Press. (d, e) From Iwasa, Y., Higashi, M., Yamamura, N., 1981. Prey distribution as a factor determining the choice of optimal foraging strategy. American Naturalist 117, 710–723.

Table 1 An overview of the different rules for patch-departure decisions in the context of foraging habitat selection, and the conditions in which each rule is likely to be selected for in terms of type of environmental variation in patch quality (i.e., spatial distribution of resource items) and individuals' knowledge on the environment. All proximate rules assume that individuals cannot *a priori* assess patch quality upon entering the patch

Patch-departure decision rule	*Conditions in which the rule will be selected for*
Ultimate mechanism	
Marginal value theorem	
Leave patch if when the instantaneous intake rate from the current patch falls below the mean intake rate in the environment	
Proximate mechanisms	
Incremental rule	
Initial probability to stay on patch depending on its quality (size); probability to stay decreases linearly with unsuccessful time on patch; each resource item found adds an increment to the current level of probability to stay; leave patch when threshold probability is met	High variability of patch quality (aggregated spatial distribution of resource items) Limited individual knowledge about patch size
Decremental rule	
Initial probability to stay on patch depending on its quality (size); probability to stay decreases linearly with unsuccessful time on patch; each resource item found subtracts a decrement to the current level of probability to stay; leave patch when threshold probability is met	Low variability of patch quality (evenly dispersed spatial distribution of resource items) Good individual knowledge about patch size
Giving-up time rule	
Leave patch if time since last resource item found exceeds a given threshold	High variability of patch quality (aggregated spatial distribution)
Fixed-number rule	
Leave patch when a fixed number of resource items has been found	Low variability of patch quality (constant number of resource items per patch)
Fixed-time rule	
Search patch for a fixed period of time and leave patch independent of the number of resources items found	Poisson distribution of number of resource items per patch

individuals can change breeding sites between breeding events (breeding dispersal), but in many species, once the breeding place has been selected, individuals remain on that place for the whole breeding season.

Selecting a habitat or patch often implies different habitat requirements for different activities, and may thus follow different decision rules. However, different types of habitat selection may strongly interact and therefore be traded off against one another, for instance, in species searching for food and reproductive sites simultaneously. Even when selecting a breeding and foraging location are distinct activities during the life cycle, selecting a breeding habitat constrains foraging during the whole breeding period, especially in species spatially constrained during reproduction, for example, sessile and territorially breeding species. Furthermore, breeding habitat choice is often strongly linked with mate choice, which implies that constraints linked to sexual selection also influence breeding habitat choice, while they are unlikely to influence foraging habitat choice.

How Select a Habitat?

The Use of Information

To choose between alternative habitats or patches, individuals have to assess the relative quality of (i.e., the expected fitness for) each alternative. Choosing a habitat or patch thus implies gathering and using *a priori* information on environmental variability, and this information may be critical. Due to the strong selective pressures on habitat selection, the use by individuals of any kind of information allowing them to improve their choice should be favored. Therefore, the existence of information-based habitat choice behaviors may be expected. This raises the questions of which type of information should be used, and how individuals sample the environment and acquire information. In theory, individuals should evaluate all characteristics affecting the success of the activity considered in each site. This could clearly become prohibitive in terms of time and energy when many factors influence this activity independently or at different moments. Therefore, individual strategies for choosing a habitat or patch based on cues integrating the effect of various factors on expected fitness, or that mix different information, may be especially favored.

What Defines Information, Its Value, Quality, and Reliability?

To be informative about habitat or patch quality, a cue should allow individuals to predict reliably their expected fitness in this habitat or patch and compare alternative patches. This depends on several factors:

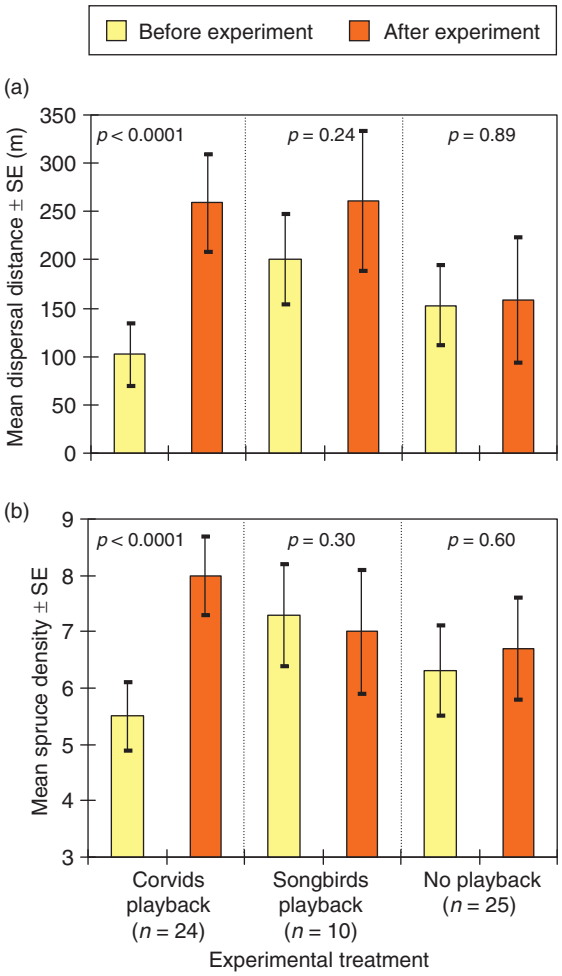


Fig. 6 An example of breeding habitat selection. Influence of the perceived nest predation risk on (a) dispersal distance and (b) nest site preferences in Siberian jays (*Perisoreus infaustus*). When exposed to nest predator playbacks (corvids) compared to control playbacks (songbirds) or no playback during breeding, individuals changed their breeding site choice in the following year. Data from Eggers, S., Griesser, M., Nystrand, M., Ekman, J., 2006. Predation risk induces changes in nest-site selection and clutch size in the Siberian jay. Proceedings of the Royal Society London Series B 273, 701–706.

1. Temporal predictability of habitat quality between the time of information gathering and use is one of these factors (**Fig. 8**); in seasonally breeding species, predicting environmental conditions from one year to the next may be easier than from the beginning to the end of the breeding season.
2. Degree of covariation between the cue and environmental variation is also important: an informative cue will in particular reflect environmental variation without time delay. Furthermore, the standard error of the cue measurement should be low compared to environmental variation; this will be the case when the cue is assessed on large samples.
3. Reliability of the cue as reflecting environmental variation is also one of the factors influencing information value. In particular, if the cue is linked to conspecifics' activity, they should not be able to affect it to provide false information; also, phenotype– or genotype–environment interactions should be limited.
4. Integrative power of the cue, which should be closely related to fitness, is another factor.
5. Easiness of cue assessment, that is, costs of acquiring or gathering information, which depend on (a) direct costs of sampling the environment (e.g., reduced survival due to predation or aggressive interactions) and (b) indirect costs (e.g., time lost in sampling and not used for other activities), is also factored in.

Subsequent costs may also be paid through competition between individuals using the same information and thus making the same choice. The costs (and thus the value of information) will depend on both species biology and spatiotemporal environmental variation of habitat quality, for example, individuals' mobility, length of the breeding period, and synchronization of breeding events among and within patches. In conclusion, the value of information will depend on the balance of information-gathering costs and benefits gained from its use.

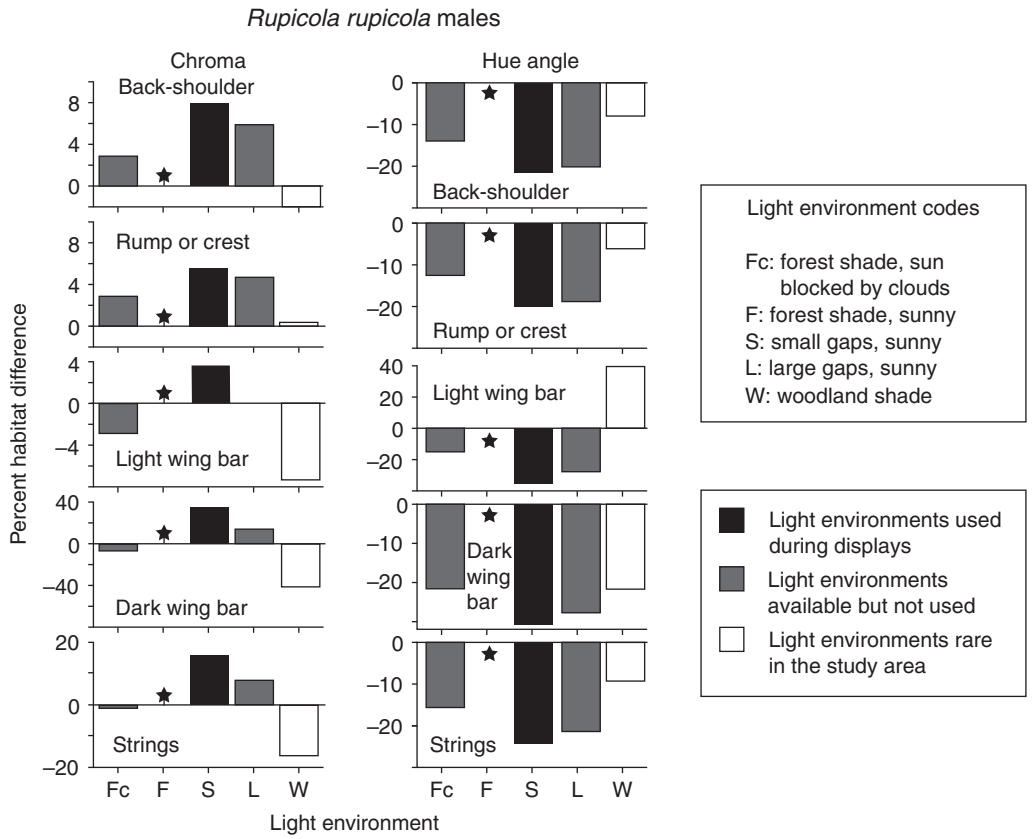


Fig. 7 Display site selection by a neotropical forest-dwelling, lekking bird species, the cock-of-the-rock *Rupicola rupicola*. Forests exhibit a mosaic of spectral environments arising from both vegetation cover and weather. Males display to females both in sunny forest shade (F) and small gaps (S), and these two spectral environments show the highest percentage difference (F taken as standard, indicated by a star) with respect to chroma and hue angle for each male color plumage element (shoulder, chest, wing, etc.). Cock-of-the-rock males therefore select for displaying the two light environments that illuminate different parts of their plumage and maximize the visual contrast during displays. From Endler, J.A., ThÅry, M., 1996. Interacting effects of lek placement, display behaviour, ambient light, and color patterns in three neotropical forest-dwelling birds. American Naturalist 148, 421–452.

Different Types of Information for Habitat Selection

Depending on species biology and activity considered, individuals can use many types of information, ranging from physical and biological habitat characteristics to conspecifics.

Nonsocial cues

Individuals may directly evaluate potential resources and constraints affecting success in a given activity (e.g., food availability, parasitism load, predator presence – **Fig. 6**). If success is mainly linked to one factor, then this strategy should prove efficient to assess habitat quality. However, when this factor is difficult to assess, when many factors affect success, or when information on some factors is not available at the time of information gathering, an alternative is to use indirect cues revealing the effect of important factors, for example, chemical compounds revealing the presence of predators. Individuals can use search images of suitable habitats acquired during development (imprinting) or later (learning).

Individuals may also use as information source their own experience and history, in particular their own performance in the activity considered in the habitat or patch, called personal information (and sometimes also private information, despite that other individuals can access it – **Fig. 9**). In the context of foraging, different strategies involve gathering information via direct environment sampling by individuals, using in particular trial-and-error tactics (stay after success, leave after failure). Differences in timescale between foraging and breeding decisions imply that trial-and-error strategies, which can be optimal in foraging, are unlikely to be selected for alone in breeding habitat choice, because they would imply settling at random to breed and using only the breeding success achieved to decide about future habitat choice. This might be very costly when the total number of breeding attempts is limited, and personal information in the case of breeding habitat selection will often be mixed with other sources of information. Philopatry, that is, fidelity to the natal site, can be considered as a form of personal information use: individuals choose a site whose quality has allowed their own growth and survival.

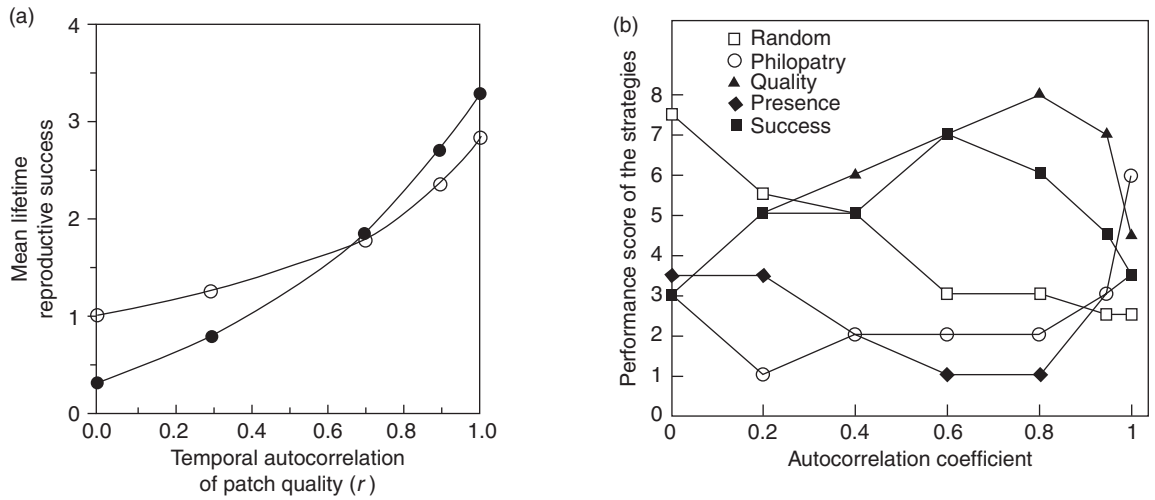


Fig. 8 The value of `public information', that is, the local reproductive success of conspecifics, for breeding habitat selection depending on environmental temporal predictability. In most seasonally breeding species, this cue will be available at the end of the breeding season, and can therefore only be used in the next year. It will thus be valuable only if the environment is autocorrelated from one year to the next. (a) In an optimality model, individuals choosing their breeding patch according to local reproductive success in the previous year (closed dots) achieve a higher lifetime reproductive success compared to individuals settling at random on a patch (open dots) only when the level of temporal autocorrelation of the environment is high. (b) Similarly, in a game theory model, strategies based on local reproductive success (`quality' and `success') are selected for only when the level of temporal autocorrelation of the environment is sufficiently high. (a) From Boulinier, T., Danchin, E., 1997. The use of conspecific reproductive success for breeding patch selection in territorial migratory species. Evolutionary Ecology 11, 505–517. (b) From Doligez, B., Cadet, C., Danchin, E., Boulinier, T., 2003. When to use public information for breeding habitat selection? The role of environmental predictability and density dependence. Animal Behaviour 66, 973–988.

Social cues

Conspecifics can also be used as a source of information about local habitat quality, that is, social information, because they share the same needs. Social information may be provided either intentionally through signals (communication), or inadvertently (inadvertent social information), when individuals monitor the behavior and performance of their conspecifics. When individuals use social information, they benefit in particular from environment sampling performed by others. Individuals can be expected to use social information more often for breeding than foraging habitat choice, and the importance of this information for breeding habitat choice has recently been emphasized (**Fig. 10**).

Conspecifics' presence on a habitat patch as an information source has received much attention (social attraction process). It can reveal good enough conditions for a local population to persist (**Fig. 10**). However, the mere presence of conspecifics may be misleading because the correlation between local density and habitat quality can prove weak in certain conditions. Conspecifics' activity and their success may better reveal habitat quality. Where conspecifics are the most successful can indicate where an individual is the most likely to be successful itself. Conspecific success integrates in a single parameter the effect of all components of environmental quality, including social interactions (**Fig. 10**). It can also be more precise than personal information when based on large samples (e.g., many conspecifics), and when phenotype–environment interactions are limited. The information derived from the performance of other individuals sharing ecological requirements has been called public information, in contrast to personal information.

The use of social information can be extended to heterospecifics, provided that they share the same needs (e.g., food or breeding sites), leading to interspecific social attraction or the use of heterospecific public information (**Fig. 10**). Because their characteristics will slightly differ from individuals of the focal species, they could even provide additional information compared to conspecifics.

Individuals are likely to combine and use several information sources, in order to refine their assessment of local habitat quality and adjust their future decisions, depending on which factors affect fitness and relative costs of gathering each information (**Fig. 11**).

Sampling the Environment and Gathering Information: Prospecting

Information gathering about relative habitat quality via prospecting behavior involves sequential visits of potential occupied or nonoccupied patches or sites by an individual that does not currently feed or breed there (called a prospector). Despite the major impact of prospecting on habitat choice, data on prospecting remain fragmentary. Prospectors on breeding patches are usually immature individuals before recruitment, and nonbreeding or unsuccessful adults, which are

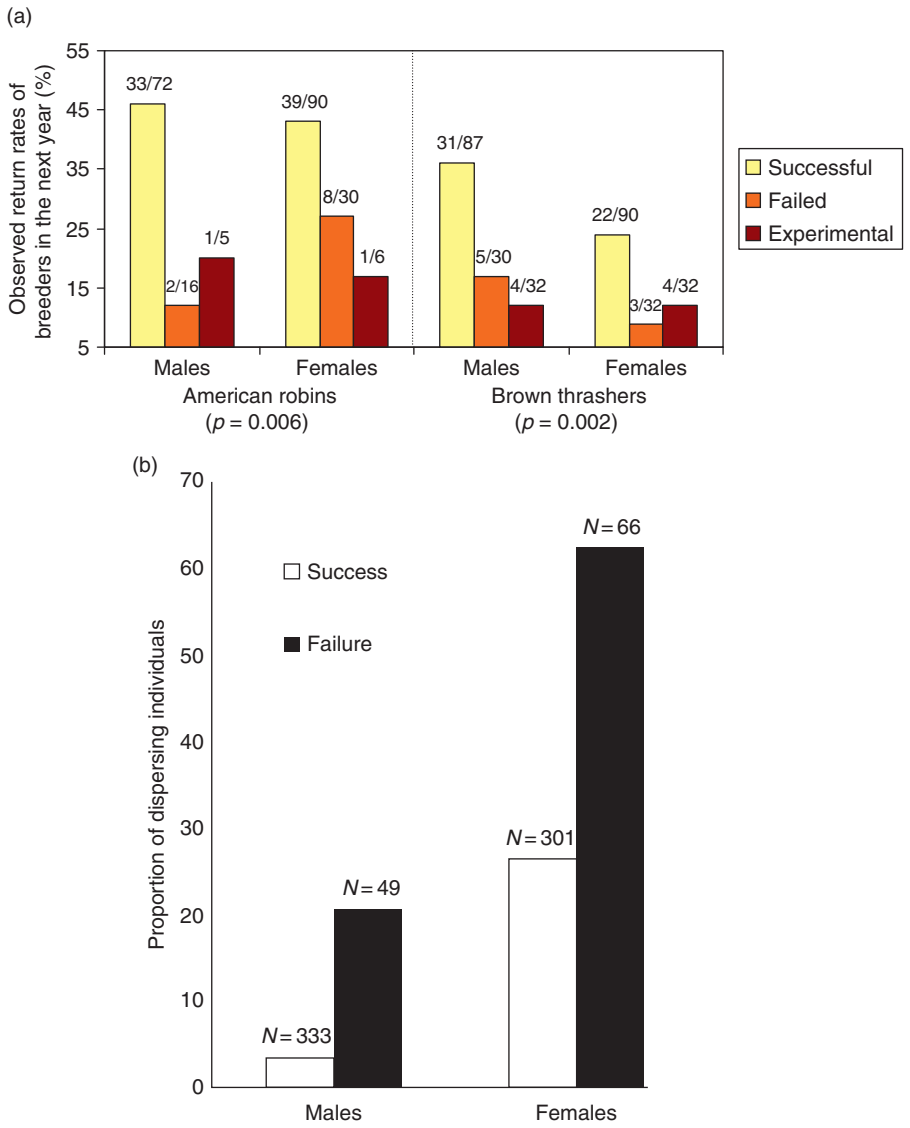


Fig. 9 Influence of personal information (individual's breeding success) in subsequent breeding habitat choice in (a) American robins (*Turdus migratorius*) and brown thrashers (*Toxostoma rufum*), and (b) collared flycatchers. Failed (naturally or experimentally) breeders are more likely to disperse to another patch in the following year compared to successful breeders. Individuals use their own reproductive performance as a cue to assess the current local breeding habitat quality and adjust their breeding patch choice in the next year. (a) Data from Haas, C.C., 1998. Effects of prior nesting success on site fidelity and breeding dispersal: An experimental approach. Auk 115, 929–936. (b) From Doligez, B., Danchin, E., Clobert, J., Gustafsson, L., 1999. The use of conspecific reproductive success for breeding habitat selection in a non-colonial, hole-nesting species, the collared flycatcher. Journal of Animal Ecology 68, 1193–1206.

likely to be looking for a breeding site for the following year. However, the links between prospecting, type of information gathered by prospectors, and subsequent habitat choice are still poorly investigated. Constraints acting on prospecting can however determine which types of information are available to individuals, and thus which habitat choice strategies can evolve. Prospecting may also shape the evolution of life-history traits such as age at first breeding when individuals have to prospect before settling for breeding.

Constraints on Habitat Selection

Habitat choice involves two important steps: (1) deciding whether to leave or stay on the current habitat or patch; (2) if individuals decide to leave, choosing where to settle next. These two sets of decisions may be based on different criteria, and be either independent or linked: individuals may decide to leave before having decided where to settle next; alternatively, they may

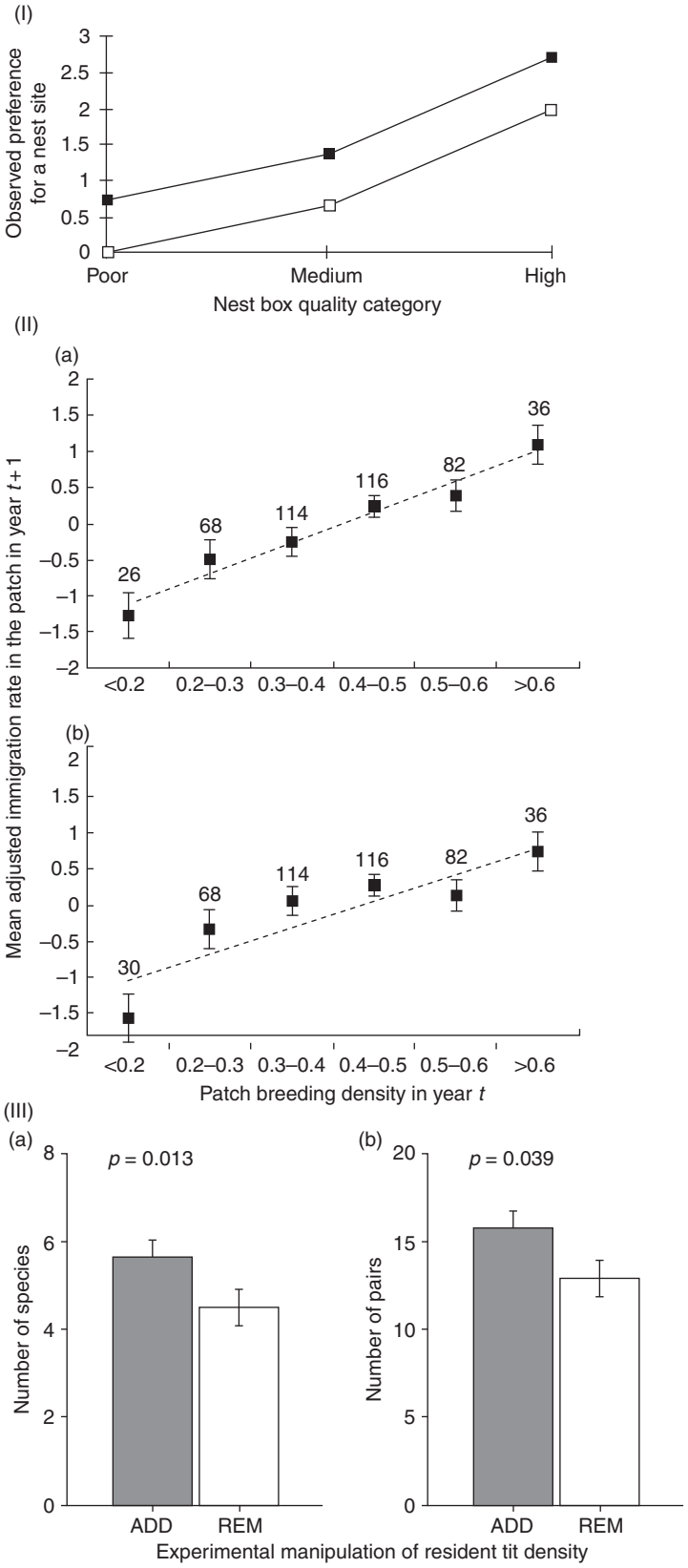
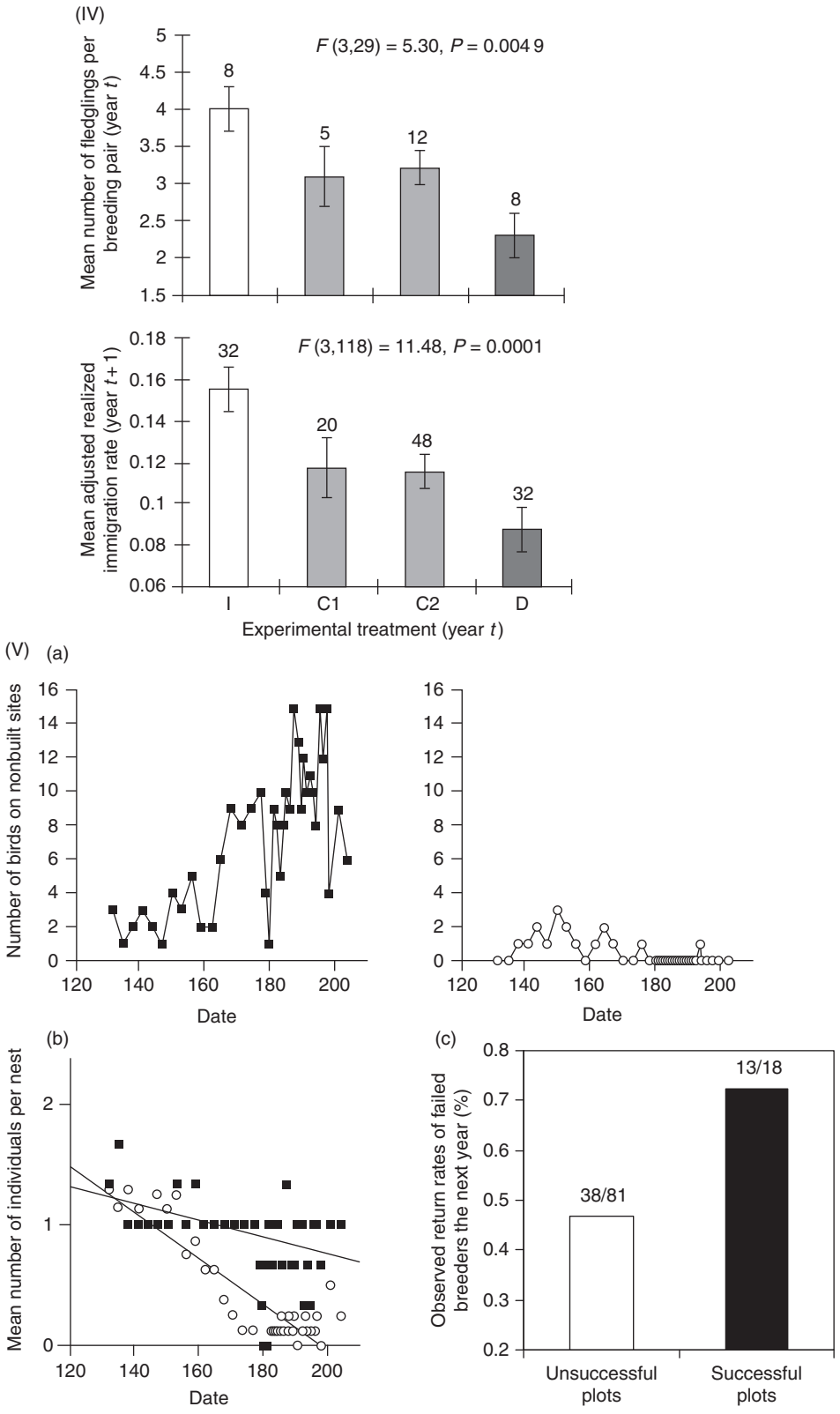


Fig. 10 (continued)


(IV)
Mean number of fledglings per breeding pair (year t)
F (3,29) = 5.30, P = 0.004 9
8
5
12
8
Mean adjusted realized immigration rate (year t + 1)
F (3,118) = 11.48, P = 0.0001
32
20
48
32
I
C1
C2
D
Experimental treatment (year t)
(V) (a)
Number of birds on nonbuilt sites
Date
(b)
Mean number of individuals per nest
Date
(c)
Observed return rates of failed breeders the next year (%)
38/81
13/18
Unsuccessful plots
Successful plots

decide to leave because they have already chosen their next patch. Choices can occur on repeated occasions, and thus be increasingly shaped by personal experience, except for breeding habitat selection in sessile species. Habitat choice is thus a complex process constrained by many parameters linked to: (1) species characteristics, in particular cognitive (spatial and temporal memory) and movement capacities (maximal movement rate or distance, especially when individuals have to travel across unsuitable habitat – arrows in Fig. 1); (2) species life-history strategy and tradeoffs involved (investment in different activities, in particular prospecting vs. breeding; tradeoffs in choosing multiusage sites, in particular year-round territories); (3) individuals' characteristics and interindividual differences in the ability to exploit the environment (phenotype– or genotype–environment interactions), in habitat preferences (through imprinting or habitat training due, for example, to acquired parasite resistance), or in selective pressures (e.g., sexual selection depending on individual's sex); and (4) environment variation (e.g., temporal predictability, spatial variation patterns).

Testing the Existence of Habitat Selection Processes

Processes, that is, mechanisms, of habitat choice have to be distinguished from patterns of space use, that is, distribution of individuals in the environment resulting from individual decisions. Patterns of individuals' spatial distribution or variation in fitness are often used to infer habitat choice processes by individuals, because determining whether habitat selection occurs can be difficult. Empirical studies often analyze habitat choice processes by comparing site characteristics and patterns of site use in different types of habitats. Occupied sites are expected to be of higher quality than unoccupied or randomly picked up sites if habitat choice occurs. However, the same patterns can result from different processes, and from processes other than active habitat selection by individuals.

A widely used concept is the ideal free distribution (IFD), defined as the distribution of individuals among habitat patches expected under the assumptions that individuals (1) distribute themselves so as to optimize their fitness, (2) are free to move among habitat patches, that is, without any cost or constraint, and (3) have a perfect (ideal) and instantaneous knowledge of the relative quality of habitat patches and local density dependence function. At equilibrium, (1) mean individual fitness is equal on all patches because individuals are distributed among patches proportionally to the relative quality (availability of resources) of each patch and (2) individuals cannot increase their fitness by changing patch (Fig. 12).

More realistic refinements of the IFD model including (1) different forms of density dependence (e.g., nonmonotonic density-dependent functions such as Allee effects) and (2) interindividual differences in competitive ability (allowing some individuals to monopolize resources and thus achieve higher fitness than others: ideal despotic or dominant distribution (IDD)) have been proposed, but the IFD represents a null model describing the spatial distribution of individuals in the environment that maximizes fitness at the population scale given the distribution of patches, and to which distributions generated via different habitat choice processes incorporating constraints on information accessible to individuals (the ideal assumption) and their movements (the free assumption) can interestingly be compared (Fig. 13).

Inferring habitat choice solely from patterns of habitat use by individuals can be misleading because high densities may be observed on low-quality patches, for instance, because of constraints for individuals in obtaining reliable information about potential habitat patches quality. A direct investigation of habitat choice, aiming at identifying information cues and decision rules used by individuals and determining the extent to which habitat choice strategies affect fitness, is often more appropriate than

Fig. 10 Examples of the use of different types of social information for breeding habitat selection: (I and II) presence of conspecifics, (III) presence of heterospecifics sharing the same needs, (IV and V) local reproductive success of conspecifics. (I) Naive house wren males (*Troglodytes aedon*) preferred to settle in nest boxes of higher quality (as measured by previous breeding success in the box), but also close to the nearest occupied box. Black squares: boxes located $\leq$70 m from the nearest occupied box (open squares for $>$70 m). (II) Immigration rate of new breeders in a patch strongly positively increased with patch breeding density in the previous year, for both (a) experienced ($\geq$2 years old) and (b) naive (yearling) collared flycatchers. (III) The number of migrant passerine bird species and densities of migrant individuals were lower in patches where the density of heterospecific resident tit (*Parus*) species was decreased (REM) by removing individuals than in patches where it was increased (ADD) by releasing them. (IV) Immigration rate of collared flycatcher breeders was higher in patches where the mean number of fledglings had been increased locally (by adding nestlings – patches I) in the previous year compared to control (C1 and C2) patches (unchanged mean fledgling number), and higher in control patches compared to patches where the mean fledgling number had been decreased locally (by removing nestlings – patches D). (V) In the black-legged kittiwake, (a) prospecting and (b) nest attendance by failed breeders were higher on patches where local success was unchanged (black dots) than where it had been experimentally decreased by removing eggs (open dots), and, in the following year, (c) failed breeders were more likely to return to breed on the same patch in patches where success was unchanged (black bar) compared to decreased patches (open bar). (I) From Muller, K.L., Stamps, J.A., Krishnan, V.V., Willits, N.H., 1997. The effects of conspecific attraction and habitat quality on habitat selection in territorial birds (Troglodytes aedon). American Naturalist 150, 650–661. (II) From Doligez, B., Pärt, T., Danchin, E., Clobert, J., Gustafsson, L., 2004. Availability and use of public information and conspecific density for settlement decisions in the collared flycatcher. Journal of Animal Ecology 73, 75–87. (III) From Forsman, J.T., Mönkkönen, M., Helle, P., Inkeröinen, J., 1998. Heterospecific attraction and food resources in migrants' breeding patch selection in northern boreal forest. Oecologia 115, 278–286. (IV) Reproduced from Doligez, B., Danchin, E., Clobert, J., 2002. Public information and breeding habitat selection in a wild bird population. Science 297, 1168–1170, with permission from AAAS. (V) From Boulinier, T., Yoccoz, N.G., McCoy, K.D., Erikstad, K.E., Tveraa, T., 2002. Testing the effect of conspecific reproductive success on dispersal and recruitment decisions in a colonial bird: Design issues. Journal of Applied Statistics 29, 509–520.

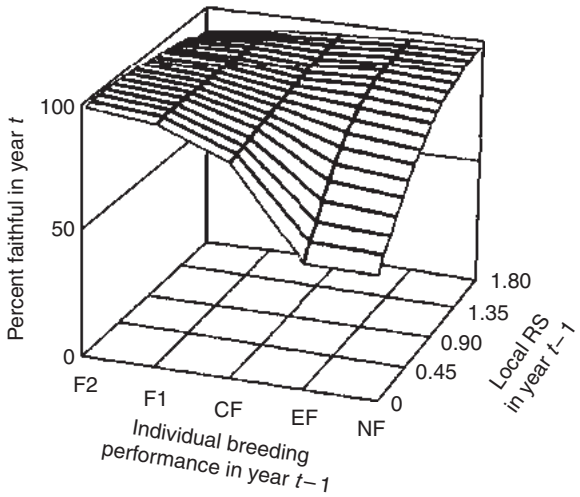


Fig. 11 Combining several sources of information for breeding habitat choice. In the black-legged kittiwake, the fidelity probability of breeders depended on both personal information (i.e., individual reproductive success) and public information (i.e., local reproductive success (RS) of conspecifics on the patch) in the previous year. Individual reproductive success decreases from F2 (high success) down to NF (nest failure: no eggs laid). Local reproductive success: average number of chicks fledged per nesting pair on the patch. Personal and public information significantly interacted. Successful individuals (F2 and F1) were always highly faithful to their breeding patch, while fidelity of early failed individuals (NF and EF) increased with local success. Thus, individuals used both personal and public information to decide whether to emigrate, but prioritize the different sources of information: public information was used only after breeding failure. From Danchin, E., Boulinier, T., Massot, M., 1998. Conspecific reproductive success and breeding habitat selection: Implications for the study of coloniality. Ecology 79, 2415–2428.

indirect inferences. Such an approach links proximate factors (elements of the environment used by individuals for choosing in a mechanistic way) and ultimate factors (evolutionary causes of individual choices, that is, linked to the relative fitness of individuals adopting different habitat choice strategies) involved in habitat choice.

Individual and Population Implications of Habitat Selection

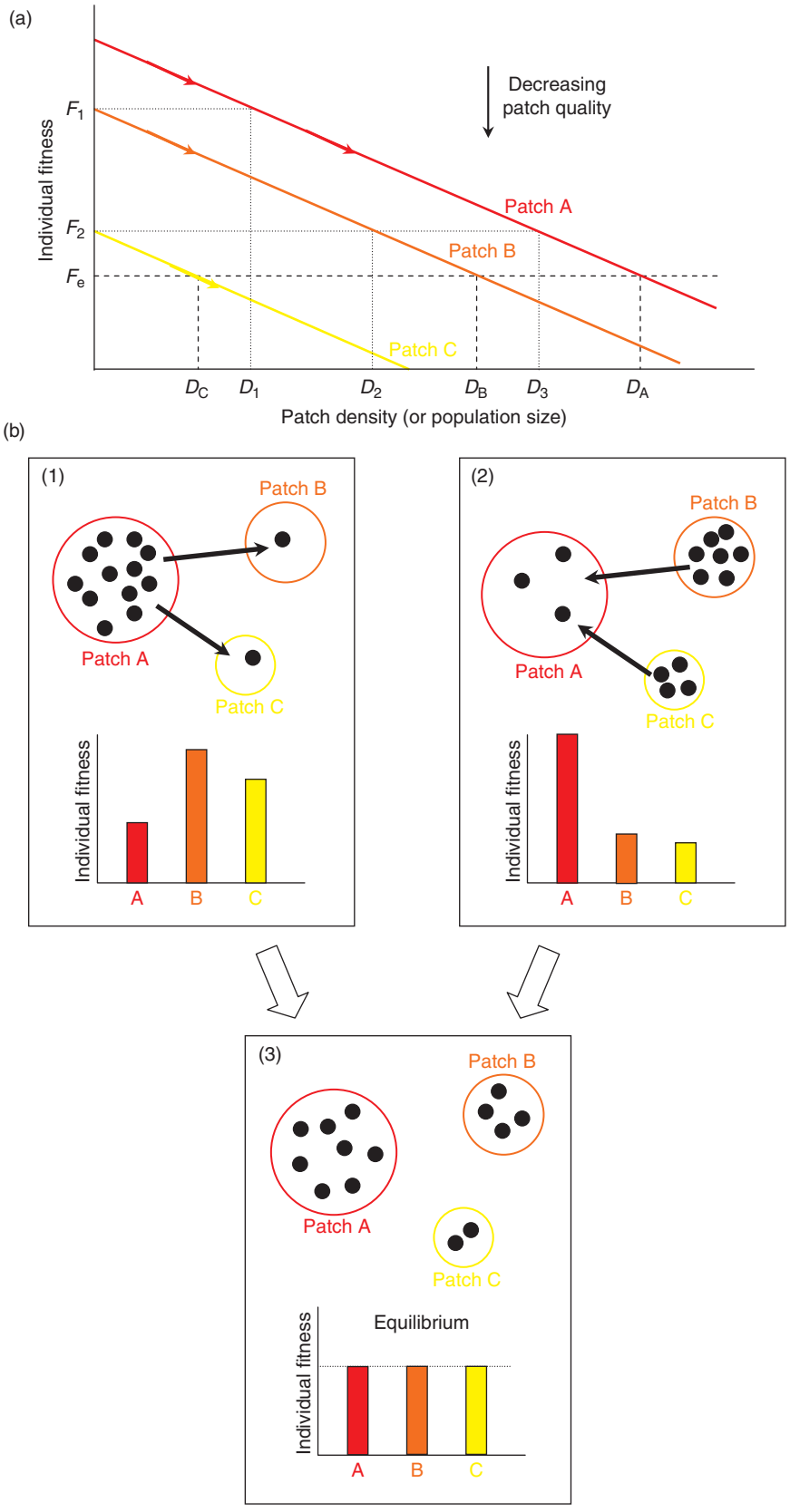
Spatial Distribution of Individuals and Evolutionary Consequences

Habitat choice directly affects the distribution of individuals in the environment and their use of habitat patches through movement and dispersal. Individual habitat choice is thus linked to spatial aggregation patterns at the population level, and the choice of each individual may have important consequences for the rest of the population (Fig. 14). Different habitat selection decisions, in particular foraging and breeding, will affect the dynamics of individuals' distribution in the environment, but at different timescales: foraging habitat selection is linked to short-term use of the environment, while breeding habitat choice will be directly linked to long-term persistence of local populations via reproduction and exchanges of individuals (thus genes) between populations, that is, dispersal.

Habitat choice strategies based on different information sources may generate different individuals' distributions among patches and temporal dynamics (Fig. 14). In particular, social attraction (i.e., use of conspecifics' presence) will progressively strongly aggregate individuals on the highest-density patch, and the whole population may end up in a single patch, independently from the relative qualities of the different patches. The presence of individuals prevented from breeding because of the lack of available sites on highly occupied patches while breeding sites are available on other suitable patches can be explained by the use of social attraction for breeding habitat selection. Breeders will indeed aggregate on a fraction of suitable patches rather than colonizing empty patches, and on these occupied patches, only a fraction of the individuals manage to secure a breeding site, the rest remaining nonbreeders because of patch saturation. Constraints in habitat selection may therefore lead to the evolution of floating and queuing strategies.

Effect of habitat choice on population regulation

Simple habitat choice decision rules can also participate in site-dependent regulation of populations, through the sequential occupation of sites of decreasing quality (Fig. 14). When the population increases due to high fitness on good-quality patches, an increasing proportion of individuals start settling on patches of lower quality. Thus average fitness at the population scale decreases, which reduces the overall population growth rate, and may lead the population to start decreasing. This regulatory effect is obtained simply through the variation in the mean quality of chosen and occupied sites, even in the absence of local crowding effects (i.e., negative density dependence at the individual level), that is, no decrease in fitness is observed for individuals occupying the best sites: the population growth rate varies simply because of the variation in the mean quality of occupied sites.


(a)
Decreasing
patch quality
Individual fitness
F_1
F_2
F_e
Patch A
Patch B
Patch C
D_C
D_1
D_2
D_B
D_3
D_A
Patch density (or population size)
(b)
(1)
Patch B
Patch A
Patch C
Individual fitness
A
B
C
(2)
Patch B
Patch A
Patch C
Individual fitness
A
B
C
(3)
Patch B
Patch A
Patch C
Equilibrium
Individual fitness
A
B
C

Habitat choice and local population viability

Individual choices are constrained by the accessibility of reliable information, and the use of suboptimal information sources may increase individuals' spatial aggregation. In a metapopulation (i.e., a set of populations connected by dispersal), individuals' aggregation on some patches leaving others empty may increase the overall extinction probability of the metapopulation by increasing the probability of simultaneous extinction of all subpopulations. In addition, if individual fitness negatively depends on local density, aggregated distributions further increase local extinction probabilities. Mixed strategies of habitat choice, using a combination of cues, or condition-dependent habitat selection strategies, may minimize extinction probability.

Evolution of sociality and coloniality

Individual strategies based on conspecific cues have also been suggested to lead to the evolution of group living. By affecting spatial distribution of individuals in the environment and thus population structure, habitat selection can lead to selective pressures favoring the evolution of group living behaviors. The use of social information for habitat choice may have led to different forms of group living, such as coloniality, since individuals using conspecifics' presence or performance for breeding site choice settle on already-occupied patches and thus aggregate breeding sites. Furthermore, feeding or breeding close to conspecifics may favor the gathering of social information on habitat quality, thus individuals may actively seek spatial aggregation to gather social information.

Conservation Biology

Through effects on the spatiotemporal distribution of individuals in the environment and population extinction probability, habitat selection behaviors have strong implications for conservation biology.

Small and fragmented populations

Threatened populations are usually small and subdivided within fragmented habitats. Habitat choice behaviors are critical for conservation issues because: (1) habitat choice strategies affect individual exploratory and prospecting movements between isolated habitat patches, which can lead to increased mortality risk depending on the degree of fragmentation; (2) the distribution and movements of individuals among habitat patches directly affect the dynamics and viability of the small subpopulations and thus the metapopulation; (3) individuals may end up settling on low-quality habitat because of constraints on mobility and information gathering, or on sites of decaying quality because of human activity. The study of breeding habitat choice is thus critical for the monitoring and management of threatened, reintroduced, or reinforced populations. Knowledge of factors affecting habitat choice has direct implications in such situations, and may greatly influence the design and monitoring of protected areas as well as the assessment of subdivided populations' viability.

Environments under human influence

Human activities can alter habitat structure and quality, and in particular break the natural correlations among habitat components. Thus, naturally selected habitat choice strategies may become maladaptive in environments modified by human activity: individuals may be lured to unsuitable patches because cues no longer reveal habitat quality when some habitat characteristics affecting fitness deteriorated but do not affect the cues used by individuals to assess site quality. Such a mismatch defines an ecological trap. Because small and subdivided populations have often greatly decreased in size in the recent past, a large proportion of potentially suitable patches may be unoccupied, but may nevertheless need to be preserved to allow individuals to move. Such situations require managing habitat in terms of metareserves aiming at protecting a habitat type independently from the current occupation by the species of interest.

Fig. 12 The ideal free distribution. (a) Individuals start occupying the highest quality patch (patch A). As population density on patch A increases, fitness return decreases as a result of a negative density-dependence function (e.g., due to competition). When density on patch A reaches level D_1, expected fitness on patch A is equal to expected fitness on patch B, which is of lower quality but still empty. The next individuals to arrive should thus start occupying patch B, as well as continuing to occupy patch A. The same applies for patch C (level D_2 on patch B and D_3 on patch A), etc. At equilibrium (dashed line), individuals are distributed among patches (densities D_A, D_B, and D_C) so that their fitness is equal on all patches (F_e). (b) This mechanism can also operate in a closed population when local densities of individuals change as a result of demographic or environmental stochasticity. Unbalanced local densities will generate different fitness gains on different patches, due to the overexploitation of rich patches (patch A, case b1) or poor patches (patches B and C, case b2). In this case, some individuals should move to a less exploited patch, so that individual fitness as equilibrium is equal on all patches again (b3). Adapted from Fretwell, S.D., Lucas Jr., H.L., 1970. On territorial behaviour and other factors influencing habitat distribution in birds. Acta Biotheoretica 19, 16–36.

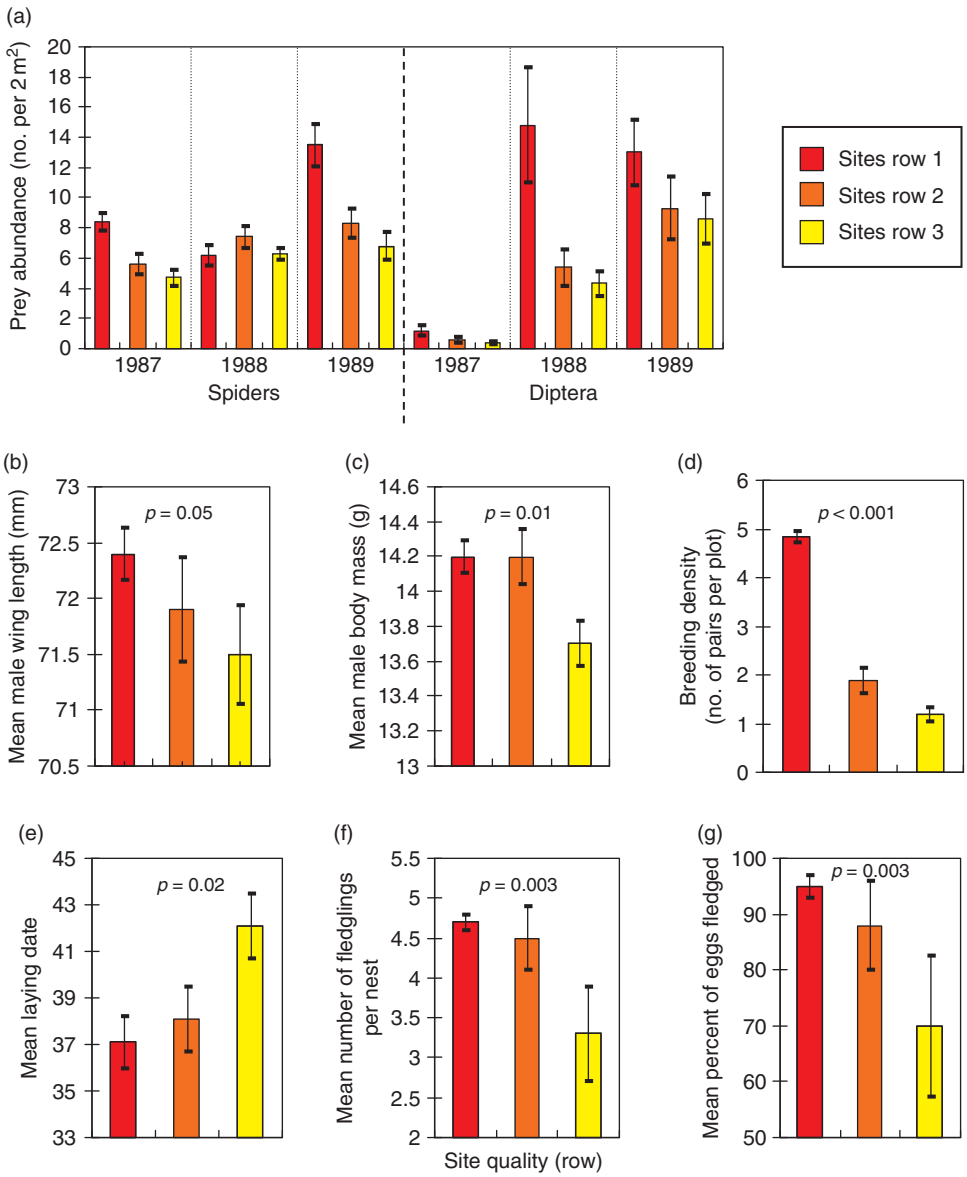


Fig. 13 Testing the predictions of the ideal free and ideal dominant distribution (IFD/IDD) models in a population of prothonotary warblers. Habitat and breeders characteristics were compared between sites of varying quality. Sites differed in their relative location with respect to shoreline, with increasing distance to the shore for increasing row level. (a) Prey abundance (thus intrinsic site quality) gradually decreased with increasing site row. Breeding male (b) wing length and (c) body mass (measuring male quality), and (d) breeding densities decreased with increasing row level, i.e., decreasing site quality (no differences in females). (e) Females initiated breeding earlier in high-quality (row 1) sites compared to low-quality (row 3) sites. Finally, breeding success measured by (f) mean fledgling number and (g) percentage of eggs that produced fledglings decreased with decreasing site quality. Thus the spatial distribution of prothonotary warblers in this population followed an IDD, with higher-quality males excluding lower-quality ones from the preferred, highest-quality areas, and thereby achieving higher reproductive success. Data from Petit, L.J., Petit, D.R., 1996. Factors governing habitat selection by prothonotary warblers: Field tests of the Fretwell–Lucas models. Ecological Monographs 66, 367–387.

Reintroduced populations

Finally, a thorough understanding of habitat choice behaviors is useful for increasing the efficiency of population reintroduction or reinforcement. The rearing conditions of individuals may affect their tendency to choose specific types of habitats, and habituation to a site can contribute to early individual settlement after release while a large mismatch between rearing and release habitats may result in individuals being unable to make optimal habitat choices. Social interactions can also be critical, such as attraction to active breeding conspecifics. Visual and/or sound decoys (e.g., mimicking successful conspecifics) can attract individuals to sites identified as suitable by managers. Decoys of predators can also be used to deter focal individuals from settling in areas identified as low quality. In other words, understanding the cues used by individuals for selecting a habitat patch allows manipulating these cues to alter individuals' choices.

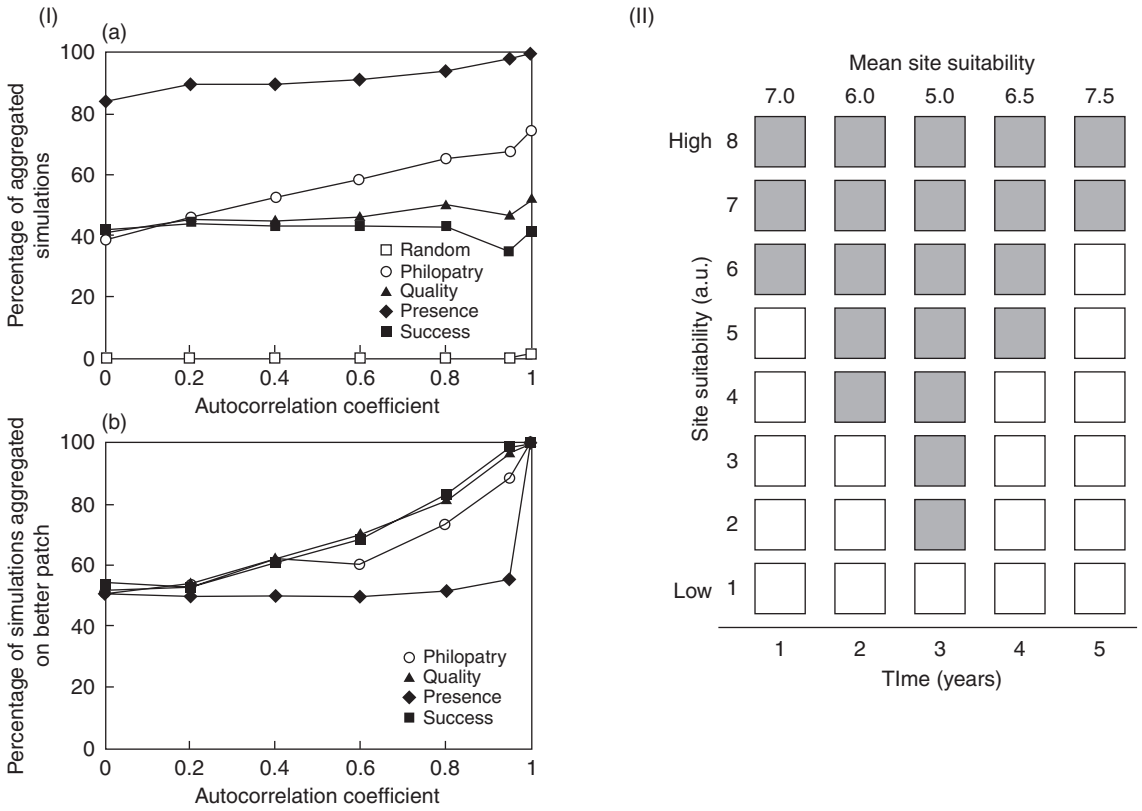


Fig. 14 An illustration of population consequences of individual habitat selection behavior. (I) Spatial aggregation of individuals. Strategies of breeding habitat selection based on different types of information lead to different levels of spatial aggregation of individuals among patches (a). In particular, the use of the presence of conspecifics generates spatial aggregation far above the IFD. (b) Individuals aggregate on the best patches as environmental predictability (and thus the value of information) increases. However, when breeding success is negatively density dependent, individuals using the presence of conspecifics pay a cost via decreased success, which limits the efficiency of this strategy. (II) Population regulation via site quality. A negative feedback can be created via individual habitat selection behavior. In small populations, individuals occupy the best patches, leading to a high growth rate (year 1). When population increases, individuals start settling on sites of decreasing quality, thus mean occupied site quality decreases (years 2 and 3). Consequently, population growth is slowed (year 4). As population declines again, mean quality of occupied sites, and thus population growth, increases again (year 5). (I) From Doligez, B., Cadet, C., Danchin, E., Boulinier, T., 2003. When to use public information for breeding habitat selection? The role of environmental predictability and density dependence. Animal Behaviour 66, 973–988. (II) From Rodenhouse, N.L., Sherry, T.W., Holmes, R.T., 1997. Site-dependent regulation of population size: A new synthesis. Ecology 78, 2025–2042.

See also: Behavioral Ecology: Habitat Mapping. Terrestrial and Landscape Ecology: Spatial Distribution

Further Reading

Battin, J., 2004. When good animals love bad habitats: Ecological traps and the conservation of animal populations. Conservation Biology 18, 1482–1491.
Boulinier, T., Danchin, E., 1997. The use of conspecific reproductive success for breeding patch selection in territorial migratory species. Evolutionary Ecology 11, 505–517.
Boulinier, T., Mariette, M., Doligez, B., Danchin, E., 2008. Choosing where to breed – Breeding habitat choice. In: Danchin, E., Giraldeau, L.-A., CÅzilly, F. (Eds.), Behavioural Ecology. Oxford: Oxford University Press.
Boulinier, T., Yoccoz, N.G., McCoy, K.D., Erikstad, K.E., Tveraa, T., 2002. Testing the effect of conspecific reproductive success on dispersal and recruitment decisions in a colonial bird: Design issues. Journal of Applied Statistics 29, 509–520.
Charnov, E.L., 1976. Optimal foraging: The marginal value theorem. Theoretical Population Biology 9, 129–136.
Clobert, J., Danchin, E., Dhondt, A., Nichols, J.D. (Eds.), 2001. Dispersal. Oxford: Oxford University Press.
Cody, M.L., 1985. Habitat Selection in Birds. San Diego, CA: Academy Press.
Dall, S.R.X., Giraldeau, L.-A., Olsson, O., McNamara, J.M., Stephens, D.W., 2005. Information and its use by animals in evolutionary ecology. Trends in Ecology and Evolution 20, 188–193.
Danchin, E., Boulinier, T., Massot, M., 1998. Conspecific reproductive success and breeding habitat selection: Implications for the study of coloniality. Ecology 79, 2415–2428.
Doligez, B., Cadet, C., Danchin, E., Boulinier, T., 2003. When to use public information for breeding habitat selection? The role of environmental predictability and density dependence. Animal Behaviour 66, 973–988.
Doligez, B., Danchin, E., Clobert, J., 2002. Public information and breeding habitat selection in a wild bird population. Science 297, 1168–1170.
Doligez, B., Danchin, E., Clobert, J., Gustafsson, L., 1999. The use of conspecific reproductive success for breeding habitat selection in a non-colonial, hole-nesting species, the collared flycatcher. Journal of Animal Ecology 68, 1193–1206.

Doligez, B., Pärt, T., Danchin, E., Clobert, J., Gustafsson, L., 2004. Availability and use of public information and conspecific density for settlement decisions in the collared flycatcher. Journal of Animal Ecology 73, 75–87.
Doncaster, C.P.D., Clobert, J., Doligez, B., Gustafsson, L., Danchin, E., 1997. Balanced dispersal between spatially varying local populations: An alternative to the source–sink model. American Naturalist 150, 425–445.
Eggers, S., Griesser, M., Nystrand, M., Ekman, J., 2006. Predation risk induces changes in nest-site selection and clutch size in the Siberian jay. Proceedings of the Royal Society London Series B 273, 701–706.
Endler, J.A., ThÅry, M., 1996. Interacting effects of lek placement, display behaviour, ambient light, and color patterns in three neotropical forest-dwelling birds. American Naturalist 148, 421–452.
Forsman, J.T., Mönkkönen, M., Helle, P., Inkeröinen, J., 1998. Heterospecific attraction and food resources in migrants' breeding patch selection in northern boreal forest. Oecologia 115, 278–286.
Fretwell, S.D., Lucas Jr., H.L., 1970. On territorial behaviour and other factors influencing habitat distribution in birds. Acta Biotheoretica 19, 16–36.
Giraldeau, L.-A., Caraco, T., 2000. Social Foraging Theory. Princeton, NJ: Princeton University Press.
Haas, C.C., 1998. Effects of prior nesting success on site fidelity and breeding dispersal: An experimental approach. Auk 115, 929–936.
Iwasa, Y., Higashi, M., Yamamura, N., 1981. Prey distribution as a factor determining the choice of optimal foraging strategy. American Naturalist 117, 710–723.
Kokko, H., Sutherland, W.J., 1998. Optimal floating and queuing strategies: Consequences for density-dependance and habitat loss. American Naturalist 152, 354–366.
Muller, K.L., Stamps, J.A., Krishnan, V.V., Willits, N.H., 1997. The effects of conspecific attraction and habitat quality on habitat selection in territorial birds (*Troglodytes aedon).* American Naturalist 150, 650–661.
Orians, G.H., Witenberger, J.F., 1991. Spatial and temporal scales in habitat selection. American Naturalist 137, S29–S49.
Outreman, Y., Le Ralec, A., Wajnberg, E., Pierre, J.-S., 2005. Effects of within- and among-patch experiences on the patch-leaving decision rules in an insect parasitoid. Behavioral Ecology and Sociobiology 58, 208–217.
Petit, L.J., Petit, D.R., 1996. Factors governing habitat selection by prothonotary warblers: Field tests of the Fretwell–Lucas models. Ecological Monographs 66, 367–387.
Reed, J.M., Boulinier, T., Danchin, E., Oring, L., 1999. Informed dispersal: Prospecting by birds for breeding sites. Current Ornithology 15, 189–259.
Reed, J.M., Dobson, A.P., 1993. Behavioural constraints and conservation biology: Conspecific attraction and recruitment. Trends in Ecology and Evolution 8, 253–256.
Rodenhouse, N.L., Sherry, T.W., Holmes, R.T., 1997. Site-dependent regulation of population size: A new synthesis. Ecology 78, 2025–2042.
Stephens, D.W., Krebs, J.R., 1987. Foraging Theory. Princeton, NJ: Princeton University Press.
Switzer, P.V., 1997. Past reproductive success affects future habitat selection. Behavioral Ecology and Sociobiology 40, 307–312.
Van Alphen, J.J.M., Bernstein, C., Driessen, G., 2003. Information acquisition and time allocation in insects parasitoids. Trends in Ecology and Evolution 18, 81–87.
Waage, J.K., 1979. Foraging for patchily distributed hosts by the parasitoid. Nemeritis canescens. Journal of Animal Ecology 48, 353–371.
Wiens, J.A., 1976. Population responses to patchy environment. Annual Review of Ecology and Systematics 7, 81–120.

Herbivore-Predator Cycles

AC McCall, Denison University, Granville, OH, USA

Introduction

One of the most long-enduring and fascinating phenomena of terrestrial ecology is the origin and explanation of synchronous population cycles in herbivores and some of their predators. Often, these cycles are synchronized over large areas, sometimes reaching over 1500 km in their effective range. Charles Elton, one of the founders of the modern science of ecology, was one of the first workers to realize the importance of extreme population fluctuations in snowshoe hares and Canadian lynx and lemmings, early in the twentieth century. Since Elton, researchers have found similar cycles or regular fluctuations in such herbivores as lepidopterans, voles, mice, grouse, and geese. The cycles of these fluctuations vary with the particular taxon under study and the mechanisms underlying the fluctuations are not uniform (see **Table 1**). So far, ecologists cannot conclusively prove the origin and maintenance of many cycles, but progress has been made in identifying the individual mechanisms responsible for both the increases and decreases in population densities.

How Common are Cycles in Animal Populations?

Although they seem unusual and dramatic, cyclic dynamics in animal populations are not particularly uncommon. Approximately 30% of all animals display significant cycles in population size, which is substantially greater than the 5% expected by chance alone. These cycles typically occur in disturbed temperate and boreal biomes, while cycles in tropical insects are much less common. Mammals and fish are the two major taxa that have the highest proportion of cyclic populations (0.33 and 0.43, respectively). It is likely that as more data are collected, examples of either separate populations or whole species that exhibit cyclic population dynamics may become more prevalent than we had realized. Latitudinal gradients in propensity to cycle are common, but are only statistically detectable in mammalian taxa.

Cycles in Relation to Lotka–Volterra Models

In the now-classic standard Lotka–Volterra model of population regulation by predators, the end dynamic is cyclical for both the consumer and the consumed. In particular, herbivores tend to increase in density until brought under control by a specialized predator whose population is dependent on herbivore density. This model gives us hints as to what forces may be responsible for the cycles we actually observe in nature. In particular, predation by specialists or other trophic interactions such as parasite infection could be important, especially if there is a significant lag between the prey density growth and predator density growth.

Factors Involved in the Regulation of Populations

Broadly speaking, herbivore populations can be regulated by density-dependent mechanisms or affected by density-independent mechanisms. These factors are key in understanding how herbivore populations cycle. Density-independent factors affect population densities of the animal in question, but are not affected by changes in densities themselves so that there is no feedback in the system. These factors may include climatic variables, sunspots, or some types of human landscape-alteration. Any particular factor may act as a density-independent or density dependent mechanism given a certain set of circumstances. For example, a disease may be density dependent if higher host population densities increase its transmission rate, but density independent if it is not contagious or otherwise is unaffected by the host population.

Density-independent factors are rarely the sole determinants in causing cycles, but they are important in providing the arena in which density-dependent mechanisms play out. For example, density-independent factors may influence the amplitude of the

Table 1 Examples of herbivores with strong cyclic population dynamics

Herbivore	*Period (years)*	*Synchronous area (km)*	*Dominant process order*	*Mechanisms*
Voles	3–5	500	2	Predation
Snowshoe hare	9–11	1500	2	Predation
LBM	9–10	1000	2	Induced plant susceptibility, parasitoids

Cycle periods, maximum spatial extent of cycle synchrony, important mortality factors, and dominant order of population dynamics are presented.

This work is greatly influenced by Turchin P (2003) *Complex Population Dynamics: A Theoretical/Empirical Synthesis*. Princeton, NJ: Princeton University Press.

 doi:10.1016/B978-0-444-63768-0.00647-8

cycles. Researchers at one time believed that density-independent factors were relatively unimportant in either predicting or influencing the increases and decreases of cycles. We now know that these factors can be vital in exaggerating, dampening, or stabilizing the effects of density-dependent factors.

One density-independent factor that was invoked to explain some aspects of the hare–lynx cycles is the activity of sunspots. Since sunspot activity affects the entire planet, the effects may be responsible for the synchronization of hare population cycles across vast areas of high-latitude boreal forest. The sun goes through a 10-year cycle of solar activity with distinct and longer periods of high-amplitude cycles every 80–90 years, or what is known as the Gleissberg period. Although the relationship between sunspot activity and hare density drifts out of synchrony over many hundreds of years, they are forced back into synchrony at the onset of a new Gleissberg period of high sunspot activity.

Thus, it appears that sunspot activity by itself is not sufficient to cause or maintain the cycles we observe in herbivores, but may be necessary to force the population synchrony over large geographic areas. Several hypotheses have been put forth to explain the mechanism behind this forcing, but the researchers believe that sunspot activity seems to affect large-scale climatic variables such as precipitation, which can affect plant growth. Increased solar radiation can also affect the defensive compound concentrations in certain plants. Since the biochemical pathways for protection against UV radiation and antiherbivore chemical defense often use the same precursor or intermediate molecules, increased resource allocation to UV-protection during periods of high sunspot activity may cause concomitant decreases in antiherbivore protection. This would presumably allow hares to eat more food, reproduce more, and thus increase population density.

Latitudinal Gradients

Climate may also play an important role in moderating the strength of herbivore cycles. Some herbivore cycles are more pronounced at higher latitudes than at lower ones, although the strength of the relationship depends on the particular taxa under scrutiny. Many extrinsic factors vary with latitude, most importantly seasonality, strength and body size (Bergmann's rule). The strongest patterns in cycling along latitudinal gradients are found in Fennoscandian (Finland, Norway, Sweden, and the Kola Peninsula in Russia) voles and lemmings. Population cycles are very strong at high latitudes and lessen in strength as latitude decreases. In these systems, it is hypothesized, and supported with small-scale manipulative experiments, that the type of predators associated with latitude may help explain the dampening of cycles at lower latitudes. In particular, specialized predators are more abundant at higher latitudes than at lower latitudes, and generalist predators are more abundant at lower latitudes. Specialized predators that depend on voles to survive will decrease in population density, following the dynamics of the vole increase and decrease. Generalist predators, on the other hand, have the ability to switch from voles as the herbivores begin their population decline, thus moderating the decline itself. In other words, specialists destabilize vole populations and generalists stabilize them. Thus, latitude itself does not cause the cycles, but influences the predators that are probably drivers of cycles.

Density-Dependent Mechanisms

When ecologists speak of population regulation, they are often referring to density-dependent mechanisms. These mechanisms may be a product of the population itself, or a reaction to the population density. For example, a population of voles may increase exponentially until competitive intraspecific interactions cause either the birth rate to decrease or the death rate to increase, leading to a net decline in reproductive rate and subsequent decrease in population density. Alternatively, as the population of voles increases, the population may become more apparent to predators, causing an increase in herbivore consumption and subsequent changes in predator densities.

First- and Second-Order Dynamics

Feedback loops that may be characteristic of density-dependent mechanisms can be broken down into having first- or second-order dynamics. First-order dynamics occur over on a short timescale, with short lags (1–2 years) between the response of the herbivore population and the feedback mechanism. Second-order dynamics have pronounced time lags (3–5 years) between the mechanism and the effects on the herbivore population. Second-order dynamics can cause cycles, but may also result in more complex phenomena, such as chaotic (very dependent on initial conditions) and bounded fluctuations in population density. In general, ecologists believe that population cycles are caused by second-order or higher dynamics, since an appreciable time lag between the factor involved and the herbivore density is needed to produce periodicity in herbivore populations.

Some properties of herbivore populations themselves may cause negative feedback loops and thus may contribute to the maintenance of population cycles by causing or exacerbating the decline portions of the cycles. For example, increasing population densities may cause more competition for food or more aggression among conspecifics, leading to increased rates of mortality or decreased birth rates. Increased competition for shelter or mating sites may result in many individuals succumbing to prey or failing to recruit offspring into the next generation. These mechanisms, however, may not be sufficient to cause cycling, since there may be only short or even nonexistent time lags between consumption of food or space and subsequent reductions in herbivore densities. One prominent exception is the population cycles of Soay sheep populations on the St. Kilda archipelago off the coast of Scotland. This population goes through abrupt and somewhat regular fluctuations in number, and has been well documented for

over 30 years. There are no predators or significant parasites on these islands, so starvation and regrowth of forage are thought to be the main driver of the cycles. Some recent analyses also conclude that growth of forage and concomitant starvation is also affected by climatic variables such as the North Atlantic Oscillation (NOA).

For herbivores, like the Soay sheep, an obvious limiting factor in the growth and maintenance of populations is their food sources – plants. Population cycles of herbivores could be maintained if the plants that they feed on also undergo large fluctuations in their population densities. For example, defoliation of trees often reduces the quality and increases toughness of vegetation in following seasons. This phenomenon, coupled with the intrinsic rate of population growth of the herbivores, may result in a feedback loop in the population size of the herbivore. This feedback can be immediate or delayed, and thus characteristic of first- or second-order dynamics, depending on the particular system involved.

If plants were the sole determining factor in the cycling process, then herbivore densities should mirror closely the densities of their host plants with some time lag, and would thus have low-order dynamics (zero or first-order) This explanation, however, has not been supported by data since plant densities rarely undergo such drastic cycles as herbivores in any system. For example, the larch budmoth (*Zeiraphera diniana*) (LBM), which cycles approximately every 9 years, plant mortality is only around 1% after attack, and plant biomass is only reduced by 50% during the outbreaks. Thus, it seems unlikely that the drastic crashes in herbivore densities are due to a lack of food resource in this system.

Similarly, in a large data set of Fennoscandian voles, oscillations in population density are probably not caused by concomitant oscillations in plant abundance alone. Oscillations in these populations have characteristics of second-order dynamics with relatively long lags in response time. Again, if determining factor was available plant biomass for consumption, then there will be little time lag between the action of the herbivores and the decline in herbivore population densities.

Methods of Determining Patterns and Mechanisms

One of the problems inherent in studying population cycles is that cycles often take place over many years. For example, the snowshoe hare and lynx cycles have a period of roughly 10 years. Ecologists were only able to appreciate this phenomenon because of the careful records taken by Hudson Bay Company fur trappers. For most organisms, we do not have the long-term population data sets needed to observe cycles, so ecologists will likely discover more cycling populations as long-term data sets are established and evaluated in the future.

Cycles in population may, at first glance, seem easy to detect. For example, we can look at the snowshoe hare and lynx data and observe clear up and down portions of the graph in a repeating pattern (Fig. 1). But how do researchers detect if there is a regular pattern to the fluctuations? There must be enough detailed data from several years of observation to even begin the analysis. If the cycle has a period of 10 years, then at least 20 and perhaps 30 years of data are needed. After these data are collected, ecologists generally use time-series analysis to detect whether a true repeating pattern occurs. In practice, this usually takes the form of calculating the autocorrelation function (ACF), which measures the correlation in population density between pairs of years in the data. The ACF can tell you the most probable period of the population cycle, that is, whether a cycle takes 5, 7, or 9 years to complete.

More complex methods can also detect if the fluctuations are characteristic of first-, second-, or higher-order functions in the time series. A cycle with a period of t years does not necessarily mean that the only effect of this year's population will be felt t years in the future. Population densities in the intervening years will also have an impact on densities t years in the future.

Which years have the greatest impact on current populations are related to the order of the system dynamics, as mentioned above. How do we determine these orders if they are not evident by looking at graphs? Ecologists use partial ACFs (PACFs) or their

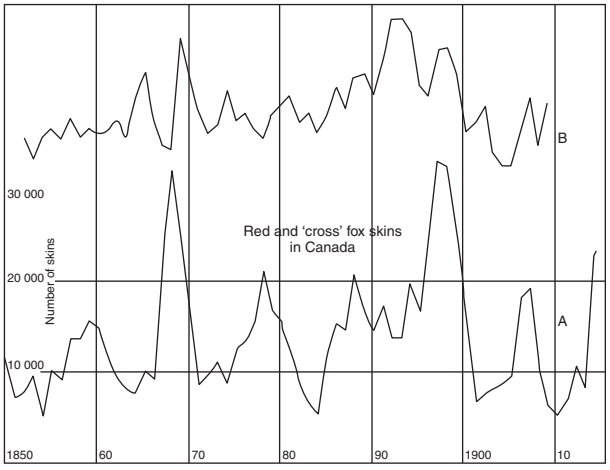


Fig. 1 Curve A shows the numbers of red fox and `cross' fox taken by the Hudson Bay Company in Canada each year. Curve B shows the same curve after a period of 10 years has been eliminated from it. The period remaining is one of about 3.5 years.

derivatives to determine the dominant process order of a density-dependent system. PACFs tell you where, in time, the greatest effects of population density occur. For example, systems with second-order dynamics usually have density-dependent lags of between 3 and 5 years, and systems with zero-order dynamics have almost instantaneous lags. It should be noted that although these analyses generally identify cycles of a certain period, there are many fluctuations around these periods in the actual data. Stochastic events are present in any ecological system and may perturb whatever periodic cycle we identify such that a population that has a general period of 3 years may have some periods of 4 years and others of 2.

These demanding forms of analysis have been used recently to help pinpoint actual mechanisms of population regulation in cycling populations. These methods are used to complement field manipulations, since field manipulations are rarely employed long-enough or at large-enough scales to replicate or change long-term dynamics.

Despite the difficulty of manipulations, such experiments have been executed, especially in the snowshoe hare–lynx systems in boreal Canada and the vole systems in Europe. Large-scale manipulations using exclosures or food additions have the twin advantages of being relatively realistic and having the ability to look in-depth at a single factor at a time over large areas.

Case Studies

Different ecosystems have different organisms, climates, primary productivity, and edaphic factors. Thus, it is not surprising that the strength of cycles, their periods, and the drivers behind the cycles vary with the particular plant–herbivore–predator system researchers examine. Since these idiosyncrasies are quite important, we will examine some of the major findings in the best-studied systems with the most-detailed records of herbivore abundance or density.

Snowshoe Hare

It is appropriate to begin with the well-known case of the snowshoe hare–lynx population cycles recorded in Canada and Alaska. These cycles were discovered by Elton while he was examining the hunting records from the Hudson Bay Company in Canada. He found that the hare and lynx pelt records as recorded by trappers exhibited strong periodic population fluctuations, and that the lynx populations seemed to track the hare densities, albeit with a significant time lag. The cycles are pronounced and occur over wide geographic areas; in 1997–99 hare population peaks were synchronous over large parts of Canada and Alaska. Since the cycles appear in the herbivore, the hares, and the predator, the lynx, the obvious causal mechanism seemed to be predation. So far, experimental data have supported this early hypothesis, with some caveats.

Although observing time-series data and making conclusions using models may give us insights into what causes the lynx–hare cycles, manipulative experiments are also necessary to determine what mechanisms are responsible for different parts of the cycles. Both the increases and decreases in herbivores may be caused by the same or different organisms. Although difficult to implement, some large-scale manipulations have helped ecologists examine individual portions of the hare–lynx cycles. The general strategy of these experiments is to try and change one major ecological factor in the hope that this manipulation will stop the cycles from occurring. These methods are very powerful, since a change in a single factor, such as food supply or predator abundance, may be shown to be responsible for the cycles.

Researchers have added food to hare populations in an effort to reverse the decline phases of a population cycle, with mixed results. When commercial-grade rabbit feed was added to large (1 km^2) experimental areas the amplitude of the cycles increased, but the cycles continued unabated. The increases in hare populations with food addition were not the result of increased fecundity by itself, but by immigration into the experimental areas from outside of the plots. Thus, food itself was not enough to `rescue' the populations from their decline phase. When more natural sources of food such as tree boughs were used to amend hare diets the same increase in population densities occurred during the increasing phase of the cycles, but this was again not enough to prevent the population crash portions of the cycles. Finally, even the addition of fertilizer to the extant vegetation cannot prevent the decline phases of cycles. On the other side of the cycles, the increases in population densities were also not significantly affected by food additions. Thus, it appears that food abundance is not solely responsible for either the increase or decrease portions of hare cycles.

In another large-scale manipulation with the hare–lynx system, researchers attempted to exclude large predators that were known to cause a great deal of mortality in hare populations. These included both lynx and coyotes, which caused very high proportions of all hare mortality in natural, unmanipulated populations. Hare populations in these exclosures did not show the marked decrease in survival rates seen in areas open to predators, but instead maintained relatively stable survival rates. Avian predators such as raptors did not seem to have a significant effect on survival rates, although it is important to realize that they still may be important in affecting certain parts of the cycle. Avian predators may, to some extent, be redundant with lynx and coyotes, replacing their numbers if they should decline.

Larch Budmoth

For lepidopterans subject to population cycles, two mechanisms have been proposed to explain the mechanism behind the pattern. Extrinsic factors, such as sunspots or weather, were popular explanations when these patterns were first discovered.

However there is little evidence that any such factors are as strictly periodic in their effects to produce the cyclical patterns we observe. Workers have come to accept the idea that negative feedback mechanisms involved with predators or parasitoids are probably responsible for most of the cyclical dynamics observed in lepidopterans. Although several lepidopteran systems exhibit relatively strong cyclical dynamics, the strongest exposition of the problem and the solution can be found in the larch budmoth.

The LBM system shows very regular population cycles in Europe. Since there are many long-term data sets for this organism, the population dynamics have received a great deal of scrutiny. Cycles generally have a period of 9 years, over large geographic areas. Based on time-series analysis, second-order or higher-order processes are likely to be causing the cycles, with declines in plant quality over many years of herbivory possibly being the mechanism. The decline in plant quality is hypothesized to occur as a result of LBM activity; this is sometimes termed induced susceptibility in plant tissues. Models employing declines in plant quality over several seasons as an exogenous factor and herbivore population growth dependent on plant material accurately mimic the observed dynamics in the system. Empirical evidence, however, suggests that other factors may be the main drivers of the cycles. During one LBM outbreak there was little decline in plant quality and the density of the larvae were quite low, relative to earlier outbreaks. However, the decline portion of the cycle still occurred, leading researchers to question the mechanism of induced susceptibility.

An alternative explanation for the production of cycles is the activity of specialized parasitoids. Many parasitoids prey upon LBM populations, although only two groups cause appreciable amounts of mortality. Simple models employing parasitoids and a time lag fit the observed dynamics significantly better than the plant quality models. There are, however, very few data regarding the exact proportion of budmoth larvae that are attacked successfully by parasitoids. A combined model that includes induced susceptibility in both plants and parasitoids actually explains the greatest amount of variation in the empirical data, suggesting that the mechanism may include the interaction between these two factors.

Owing to the lack of large-scale manipulations in this system, it is difficult to say conclusively what the true drivers of the LBM cycles are. It seems likely that many processes contribute to the effects that we see, with some redundancy in the factors. For example, population crashes still occur when plant quality stays relatively stable, so parasitoid activity may `take over' when plant quality does not decline.

Fennoscandian Voles

As mentioned above, Fennoscandian voles show dramatic population cycles over a wide geographic range with a pronounced decrease in cycle strength with decreasing latitude. Time-series analyses point to a significant second-order component in the system, again suggesting that delayed density-dependent time lags are in operation. In contrast to the previous systems, this example includes a suite of small mammals that cycle in partial synchrony: the field vole, *Microtus agrestis*, the sibling vole, *M. rossiaemeridionalis*, and the bank vole, *Clethrionomys glareolus*. The abrupt crashes in population densities that characterize the decline portion of the cycles occur despite seemingly adequate environmental conditions for survival and reproduction.

Time-series analysis again suggested that there is a significant lag dynamics, with an approximate order of 2. Thus, we are dealing with density-dependent lags that are of medium length (3–5 years). Model-fitting using specialist predators as the causal mechanism fit the dynamics of the actual populations quite well, and can mimic the declines in periodicity with increasing latitude given the correct model parameters.

In contrast to the LBM example, there are experimental data that support the time-series data and the mechanistic model results. Experimental removal of both avian predators and weasel predators eliminated the decline in population densities during two separate cycles. Elimination of weasels alone, however, did not preclude the population crashes, perhaps because weasels do not make up a majority of vole-eating predators in the study areas. Although these results are highly significant, they do not fully explain the cycles. They only are able to mechanistically account for the population crashes, but not the subsequent increases in population sizes as the next cycle begins. This may not be a large obstacle, as small herbivores have a high intrinsic rate of population growth, and the increases may be a case of standard exponential growth dynamics as described in Lotka–Volterra models.

General Conclusions

Cyclic population dynamics continue to fascinate ecologists, and as more long-term data are collected, more examples of population cycles are sure to be discovered. What broad conclusions can ecologists make regarding the forces responsible for the maintenance of these cycles, given the information we have at hand? First, it is apparent that cycles are usually not caused by a single factor at a single period of time. Indeed, this is the hallmark of many ecological phenomena, since science is largely the study of complex interactions among many factors. Many agents are probably responsible for the net effects ecologists have observed over time.

Given this caveat, if we examine the examples in **Table 1**, there seems to be a mechanistic bias toward trophic interactions, that is, predators or parasites of some kind seem to be involved in all of the examples listed. Secondly, the main drivers of the oscillations are second-order in nature, with significant time lags in density-dependent factors. This is not to say that first-order dynamics are unimportant; first-order dynamics are often needed to stabilize cycles, but they are not sufficient to cause cycles. In

general, population cycles in herbivores are rather common in nature and the cycles themselves tend to depend on significant time lags. These lags suggest that predation or other trophic processes may be operating, but only experimental manipulation over several population cycles can confirm the exact mechanism. Thus, as in most of ecological science, theories are rejected or supported based on many avenues of evidence. Since biological systems are so complex, we should expect that both the drivers of population regulation and the population trajectories themselves will be equally fascinating and multivaried.

See also: Behavioral Ecology: Optimal Foraging Theory. Conservation Ecology: Trophic Index and Efficiency. Ecological Data Analysis and Modelling: Climate Change Models. Evolutionary Ecology: r-Strategists/K-Strategists. General Ecology: Seasonality

Further Reading

Berryman, A.A., Stenseth, N.C., Isaev, A.S., 1987. Natural regulation of herbivorous forest insect populations. Oecologia 71, 174–184.
Elton, C.S., Nicholson, M., 1942. The ten-year cycle in numbers of the lynx in Canada. Journal of Animal Ecology 11, 215–244.
Kendall, B.E., Predergast, J., Bjørnstad, O., 1998. The macroecology of population dynamics: Taxonomic and biogeographic patterns in population cycles. Ecology Letters 1, 160–164.
Klemola, T., Tanhuanpää, M., Korpimäki, E., Ruohomäki, K., 2002. Specialist and generalist natural enemies as an explanation for geographical gradients in population cycles of northern herbivores. Oikos 99, 83–94.
Korpimäki, E., Brown, P.R., Jacob, J., Pech, R.P., 2004. The puzzles of population cycles and outbreaks of small mammals solved? Bioscience 54, 1071–1079.
Korpimäki, E., Nordahl, K., 1998. Experimental reduction of predators reverses the crash phase of small-rodent cycles. Ecology 79, 2448–2455.
Korpimäki, E., Nordahl, K., Huitu, O., Klemola, T., 2005. Predator-induced synchrony in population oscillations of coexisting small mammal species. Proceedings of the Royal Society of London B 272, 193–202.
Krebs, C.J., Boonstra, R., Boutin, S., Sinclair, A.R.E., 2001. What drives the 10-year cycle of snowshoe hares? Bioscience 51, 25–35.
Turchin, P., 2003. Complex Population Dynamics: A Theoretical/Empirical Synthesis. Princeton: Princeton University Press.
Turchin, P., Wood, S.N., Ellner, S.P., *et al.*, 2003. Dynamical effects of plant quality and parasitism on population cycles of larch budmoth. Ecology 84, 1207–1214.

Imprinting

T Slagsvold and BT Hansen, University of Oslo, Oslo, Norway

Introduction

The ability to learn is a naturally selected adaptation that enables an individual to adjust its behavior according to the current surroundings. Some of the most basic animal behaviors, such as recognizing conspecifics or knowing where to look for food, may be acquired through a learning process known as imprinting. A widely applicable definition of imprinting is that it is a learning process that restricts preferences to a specific class of objects. It implies some sensitive period when imprinting can occur. Usually, imprinting refers to the learning of social preferences that occurs relatively early in life, and that is stable once it is established in the individual. Under natural conditions, the social parents usually serve as stimulus objects for the developing young. A further characteristic of imprinting is that it occurs without any obvious reinforcement (this point has been widely debated). Although clearly a unique learning process, imprinting shares many characteristics with associative learning.

Two kinds of imprinting have been extensively studied. Filial imprinting concerns the development of a social preference of a young animal for its parent(s). Sexual imprinting is the process by which young animals learn the characteristics of future mates. Both kinds of imprinting may also function in individual and kin recognition. In a wider context, imprinting may determine the species recognition of many animals. Imprinting may even be of importance for establishing nonsocial ecological preferences, such as for food and habitat. The timing and duration of the learning process may differ between behaviors and species. Moreover, the kinds of behavior that may be affected by imprinting may vary with the life history and ecology of the species.

Filial Imprinting

Filial imprinting refers to the process where the social behavior of the young animal becomes limited to a particular object or class of objects, as a result of exposure to an object. In most cases studied, the stimulus object is the social mother, hence the young learn to recognize and attach to a parent through filial imprinting. The typical example is the young of ducks and geese, which instantly follow their mother. This is often called the following response. The chicks of these birds imprint on any individual present at hatching time, including a human, and will follow this individual as a mother figure. In fact, ducklings and goslings have also been shown to imprint on various inanimate objects in the laboratory. Generally, the attachment to the mother figure is very strong, and hence when a gosling is imprinted on a human, it will not change the attachment although it is given access to its true mother after only a day post hatch. Hence, a characteristic of this imprinting is that it is more or less irreversible. Cases of filial imprinting similar to those seen in birds are also found in mammals. For instance, a juvenile, single lamb will follow a human who provides milk from a bottle, also when the lamb is not hungry. The attachment may last well into adulthood.

At least two perceptual mechanisms are involved in the development of filial behavior, namely filial imprinting and predispositions. The latter refer to perceptual preferences that develop in young animals without any prior experience with the particular stimuli involved. The two mechanisms are neurally and behaviorally dissociable but are supposed to interact during the development of filial preferences, where the predisposition may bias the animal's responses. This bias may differ strongly among species, as observed in precocial birds. For instance, the bias may be relatively weak in a goose but strong in a wader.

Numerous experiments have shown that the young may imprint on a wide range of objects, and individuals that form an attachment to the wrong kind of object will certainly be disadvantaged under natural circumstances. However, in nature the adult present at this early stage of life will usually be the caring parent and so this imprinting method may work perfectly. In species with parental care, the benefits of filial imprinting are obvious because it helps an offspring to attach to its parent. If such a bonding does not take place, the parent may not start investing in that particular young and it will soon die. The following response of goslings and ducklings is adaptive because in these birds the nest is often placed at some distance from water and hence the mother has to lead the young to the water soon after hatching, often through various obstacles, like dense vegetation, where it is important that the young would follow closely. The strong attachment to the mother would also have other advantages, like avoiding dangerous sites and learning to recognize enemies, where the mother may give a specific call and the chicks would freeze and remain motionless until danger has passed. Attachment to the mother will, in many cases, also help the young to learn foraging behavior, to locate shelter, and to socialize with other members of their own species.

Typically, juveniles do not avoid any object initially but tend to approach and explore them. Imprinting on the caring parent would help them to avoid novel objects, and this acquired ability to discriminate may effectively bring the sensitive period for the filial imprinting to end. However, experiments have also shown that after having become familiar with an imprinting stimulus, the juvenile may begin to prefer stimuli slightly different from the initial one. This may have the effect of familiarizing the offspring with different aspects of the mother. It may be important to recognize a parent from many angles, which seems only possible if the juvenile builds up a composite picture of its parent's characteristics.

Apparently, the juvenile identifies as the caring parent the first object it meets that possesses some simple characteristics. Various tests have been made to study which imprinting stimuli are more important. In case of the geese studied by Konrad

 doi:10.1016/B978-0-444-63768-0.00017-2

Lorenz, most important was the movement of an object away from the chick. The effect was even stronger if the object produced some sound, although the sound did not need to be from a goose or bird; even the sound of a ticking clock could work. Furthermore, the model body did not need to be of an animal or covered by feathers but could be a simple box or a block of woods. However, more detailed studies have shown that young birds tend to show some innate preferences for certain features, such as color, shape, and size that may steer them toward the real mother figure rather than toward some arbitrary objects. Similarly, juvenile rhesus monkeys prefer a cloth surrogate mother to a wire surrogate mother. Such innate preferences are often referred to as predispositions. The variation among species in sensitivity to stimuli seems related to which stimuli are important in the wild. For instance, sound is very important in wood ducks. These birds nest in tree holes and the mother calls to induce the ducklings to leave the cavity. There also seems to be a change in relative importance of stimuli with time. For instance, in ducks, the following response is largely influenced by auditory cues from the mother soon after hatching whereas visual cues become important later on. Likewise, offspring of many species may show a preference for the more conspicuous signals, but if the signal is too startling, it may elicit fleeing rather than approach. The following response can also be enhanced with food rewards, which makes sense in species where the parent provides food or leads the young to a food resource.

Studies have also been made to identify the sensitive period for the filial imprinting, for instance, by quantifying the following response in ducks and geese to a proper mother model presented at various points in time. Some authors use the term `critical period' for the part of the sensitive period when the learning response is greatest. In precocial birds, like geese, where the chicks can run around and find food soon after hatching, the sensitive period for the filial imprinting is only the first one or two days post hatch. In fact, the young seem to learn the call of their mother already before hatching. However, the duration recoded for the sensitive period may be dependent on the method used, for example, whether the response is measured as the percentage of following responses after a single exposure, or by the percentage of birds following during the first exposure; the former measure may show a much more sharply defined period of sensitivity than the latter. One also has to take into account whether single individuals are tested, or a brood of young, because there may also be an effect of siblings (social effects, or imprinting on siblings). For instance, juveniles kept singly may remain responsive to moving objects much longer than juveniles kept in groups. Imprinting on sibs may help the juveniles to stay together.

The sensitive period for filial imprinting is much later in altricial birds than in precocial birds. Altricial young are blind and very helpless when hatching, which may limit their opportunity to imprint during their first days of life. Moreover, the necessity for very early imprinting on parents clearly differs between the two groups of birds. Precocial young move around and feed at least partly by themselves shortly after hatching, and may thus need immediate tuition in, say, feeding behavior. Moreover, this behavior may lead to many encounters with various adults apart from their mother. Altricial young, on the other hand, do not need to distinguish their parents from other birds until quite late in the nestling period since they receive food in their nest where there usually will not be any adults around except their parents.

Sexual Imprinting

Sexual imprinting is the process by which young animals learn a sexual preference for opposite sex conspecifics. The effects of this early learning process become manifest in adult mate choice. The social parents usually serve as templates for the young in the establishment of the mate preference. There is also some evidence that siblings may influence the development of mate recognition.

Most research on sexual imprinting has been performed with birds. Among birds, sexual imprinting has been documented in more than 100 species and seems to be the rule rather than the exception. Sexual imprinting has also been documented among mammals and fish. It is not limited to rapidly evolving species, or males only, as has been suggested. Most studies on sexual imprinting have been performed in a laboratory setting, particularly with zebra finches, mallards, and quail. However, the existing field studies show that sexual imprinting is also prevalent under natural conditions, and generally confirm the properties of the learning process as they have been documented in the laboratory.

Typical experiments demonstrating sexual imprinting involve some kind of manipulation of the parental phenotype. Interspecific cross-fostering is frequently used, that is, the experimenters let young of one species be reared by adults of another species. The result of this treatment is usually that cross-fostered individuals express a sexual preference for their foster species when tested in mate choice trials as adults **Fig. 1**. This is proof that mate choice is learned, since normally reared individuals do not express any preference for heterospecifics. Similarly, young raised by parents with an artificial ornament may learn a preference for this ornament.

Previously it was thought that sexual imprinting was confined to a short period early in life. However, a series of laboratory experiments on the zebra finch showed that this notion was incomplete, and redefined the understanding of the sensitive period and the irreversibility of sexual imprinting. They showed that birds raised by heterospecific foster parents develop a preference for their foster species if they also experience their first courtship with that foster species. However, if such cross-fostered birds are exposed to conspecifics during first courtship or breeding, they may shift their initial preference toward conspecifics. These findings have been corroborated by studies of the zebra finch forebrain where physical changes accompany the early experiences as well as the experiences during first courtship. It thus seems like sexual imprinting is accomplished in two separate stages.

Several studies have documented that interspecifically cross-fostered individuals in species that do show sexual imprinting may also have a sexual preference for conspecifics in spite of the experience of having been raised by heterospecifics. There are several

Fig. 1 A blue tit female paired to a great tit male, feeding young. Blue tits raised by great tits and great tits raised by blue tits imprint sexually on their foster species. Heterospecific couples may be formed between such interspecifically cross-fostered individuals. Photo copyright: Tore Slagsvold.

potential sources for such an own-species bias. First, the bias may be genetically encoded, that is, mate recognition may to some extent be inherited rather than learned. Second, an own-species bias may have arisen as a result of conspecifics initiating more courtship than do heterospecifics toward the interspecifically cross-fostered individuals. Hence, the behavior of the stimulus individuals in the experimental mate choice situation may influence and even constrain the mate choice of the cross-fostered individuals. Third, factors such as the amount of care received by the parents, or the number of siblings in the nest, may influence the degree of imprinting, and such factors may have differed between cross-fostered and control individuals. In sum, there are many potential sources of variation in the development of mate choice. These must be kept in mind when designing and drawing conclusions from experiments on imprinting.

Some studies have revealed that the recognition of same-sex individuals also may be influenced by imprinting. Interspecifically cross-fostered great tits and blue tits respond aggressively toward same-sex individuals of their heterospecific foster species during the breeding season, while normally reared controls of both species respond aggressively only toward conspecifics. This effect lasts for life (Fig. 2). It is not known whether the development of such rival recognition is different from sexual imprinting. However, the fact that the appearance of rivals also may be learned highlights the function of imprinting in the development of species recognition as a whole.

Species recognition is a prerequisite for adequate mate choice, and learning species-specific characteristics from social parents and/or siblings is reliable because social parents and offspring are usually of the same species. A notable exception is the case of interspecific brood parasites such as cuckoos and cowbirds, which leave their young to be reared in nests of heterospecifics. Initially, it was thought that recognition templates were genetically inherited in brood parasites. However, recent findings suggest that the early social environment affects choices of social partners and mates in some brood-parasitic species. The timing of learning may have been shifted to the postfledgling stage for species that do not associate with conspecifics earlier. It has been suggested that conspecific recognition in the obligately brood-parasitic brown-headed cowbird is initiated when a young brood parasite encounters an innate species-specific vocalization that triggers learning of additional aspects of the vocalizing individual's phenotype.

Is Sexual Imprinting Adaptive?

Individuals that fail in the very basic task of recognizing conspecifics are clearly disadvantaged – thus there is a strong selection pressure for optimal species recognition. Why may this ability be affected by a learning process? After all, the learned features seem categorical and quite straightforward (i.e., conspecific/heterospecific), so why is not this knowledge inherited rather than learned? Although the categories to be distinguished remain distinct, there is considerable variation within the categories. Species-typical properties might change over space and time, as might the environment. Inheriting species recognition is hence potentially disadvantageous, because a change in the gene frequencies of the inherited recognition will be much slower than learning the change in species appearance. A genetically inherited behavior thus gives less flexibility, and inflexible individuals may thus miss out on mating opportunities if they do not recognize a novel conspecific morph. Learning from parents or other tutors may enable the animal to track such changes when they appear, and hence to develop adequate preferences. An advantage of learning mate preferences through imprinting is thus that it offers some flexibility in the face of stochastic events. However, the flexibility offered by learning can also be costly. If an animal imprints on the wrong kind of stimulus object, it might end up courting heterospecifics. Erroneous species recognition may result in futile hybrid mating, or no mating at all. Also, forgetting and relearning of mate recognition in the course of an individual life span may be disadvantageous because repeated stimulus object choice could

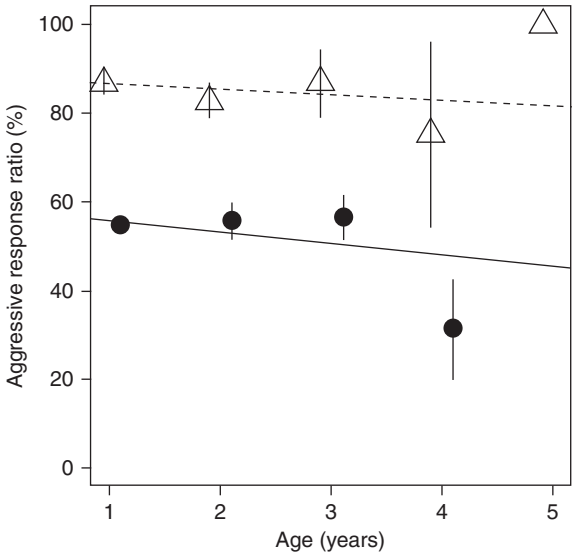


Fig. 2 Rival imprinting in blue tits persists with age. Aggressive response ratio (*y*-axis) is measured as the aggressive response toward blue tits divided by the sum of the aggressive responses toward blue tit and great tit intruders into the territory. Hence, a ratio of 50% means equal responses to great tits and blue tits, while a higher ratio means a stronger relative response toward blue tits than great tits. Control blue tits (triangles) respond mostly toward blue tit intruders into their territory, while blue tits cross-fostered to great tits (circles) respond aggressively toward both great tit and blue tit intruders, and this pattern persists throughout the life span of the respondents. Aggressive response is measured as the proportion of 5 min spent within 2 m of a caged intruder placed near the nest box of breeding respondents. All respondents were sequentially presented with individuals of both species. Symbols indicate arithmetic means ± SE. Reproduced from Hansen BT, Johannessen LE, and Slagsvold T (in press) Imprinted species recognition lasts for life in free-living great tits and blue tits. *Animal Behaviour* (doi:10.1016/j.anbehav.2007.07.023; available online 29 Oct. 2007), with permission from Elsevier.

introduce errors. Imprinting trades off the mixed blessings of flexibility. It limits the error potential of learning by optimizing the timing of the event relative to the presence of adequate stimulus objects. Further, imprinting minimizes the number of learning events and limits learning duration. Note, however, that there seem to be species differences in the degree of sexual imprintability, which indicates that the selection pressure favoring a somewhat flexible species recognition is not universal.

Sexual imprinting may also have a function in kin recognition. For instance, quails of both sexes develop a preference for the phenotype of distant relatives of the opposite sex. Kin recognition is advantageous for many reasons, not least optimization of mate choice with regard to inbreeding and outbreeding.

Evolutionary Consequences of Sexual Imprinting

Sexual imprinting promotes assortative mating, and may thereby restrict gene flow between populations. It may thus be instrumental in the establishment and maintenance of premating isolation between species or populations, which may result in speciation. If two populations diverge in some mate recognition trait, so will the populations' preferences for that trait. The tight link between the appearance of the trait and the preference for the trait may ensure premating isolation if the trait values of the two populations do not overlap.

It has also been suggested that sexual imprinting may be a driving force in the evolution of exaggerated traits by sexual selection. However, a preference in offspring for individuals that look exactly like their parents will not lead to directional selection for exaggeration of ornaments. Furthermore, frequency-dependent selection resulting from assortative mating will select against novel mates. How, then, may imprinted mate preferences drive sexual selection? To balance the costs and benefits along the inbreeding–outbreeding continuum, it may be advantageous to choose a mate who looks rather similar, but not identical, to one's own parents. In addition to facilitating inbreeding avoidance, this asymmetric mate preference may drive sexual selection, since deviations from the population mean are favored. Sexual imprinting may hence produce preferences for novel modifications of secondary sexual characters, and this process may in turn drive speciation.

Preferences for traits that deviate substantially from the population mean have been shown experimentally in several species. Moreover, theoretical models confirm that sexual imprinting and an asymmetric mate preference can lead to sexual selection. The exact mechanism for the development of an asymmetric mate preference has not been firmly established, but it could be a product of sexual imprinting itself, or sexual imprinting combined with a perceptual bias. Recent empirical evidence suggests that skewed mating preferences may be a result of a by-product of the learning process, called peak shift. If two stimuli are similar in appearance, and are differently reinforced in the responding individual, this process shifts the peak in response to increase the contrast between the stimuli. Zebra finches show natural sexual dimorphism in beak coloration, and it has been shown that males prefer females with a beak of a more extreme color than that of their social mothers. The preference is in the opposite direction of

the paternal beak color. It thus seems like sexual imprinting through peak shift can generate skewed sexual preferences for exaggerated phenotypes that have not been present at the time of learning.

Other Types of Imprinting

Several other behavior patterns seem to be acquired by an imprinting-like process. Most of these cases have been less studied than filial and sexual imprinting, particularly in a laboratory setting, and hence less is known about the details, for instance, of the timing and duration of the sensitive period, and the degree of stability of the learning in the individual.

The acquisition of some nonsocial ecological preferences seems to be influenced by early learning processes. Such ecological imprinting is illustrated by a recent study on great tits and blue tits. These tits inhabit the same woodland habitat, but differ in the ecological niche they utilize, the larger great tit foraging more on the ground and closer to the tree trunk than do blue tits. In an experiment where these tits were interspecifically cross-fostered, adults of these birds chose the ecological niche of their foster species rather than that of their genetic species, even though they possess species-typical adaptations for niche utilization (e.g., bill and feet morphology). There are indications from other species that food-type, habitat, and home area preferences also may be similarly affected. For instance, invertebrate cuttlefish have been shown to visually imprint on their preferred prey type. The development of homing behavior is another remarkable example. Salmons learn olfactory information as juveniles and use those odor memories as adults to home to their natal site for reproduction years later. This has been termed olfactory imprinting, and a similar mechanism for homing has been demonstrated in sea turtles and pigeons.

Parental recognition of the young may also be acquired by an imprinting-like process. This process is atypical in the sense that it occurs during adulthood rather than during the early stages of development. It is of course beneficial for parents to recognize their young so that they can direct care to their own kin. In birds, as in a colony of seabirds, vocalization may be particularly important in this regard. In goats, an olfactory imprinting mechanism may help mothers to recognize their own kin at a very early stage. Development of very early offspring recognition is necessary in herds of goats because kids may lose contact with their mother and may then try to approach other females. A few minutes of contact with a kid may be sufficient for the mother to imprint on it. During this period, the mother may learn the smell of the kid, and label it further by licking. Afterwards, she may only allow such labeled kids to suckle. The high arousal of a female giving birth may facilitate the imprinting on the young. In contrast, altricial birds associate the young hatching in their own nests as their own kin and hence do not need to recognize them individually until they leave the nest. However, there is a cost to the parents when following this simple rule of thumb, namely that the parental imprinting mechanism can be exploited by brood parasites, like cowbirds and cuckoos, dumping eggs into their nest. Parasitic eggs may be recognized and removed, but as soon as they hatch, the parasitic chicks are usually given full support by the foster parents.

Imprinting may also play a role in host choice in certain brood-parasitic species. Some brood-parasitic birds exploit hosts of a particular species, and the egg of the brood parasite has evolved to mimic the host egg. Hence, the brood parasite must recognize its optimal host species, and it has been suggested that they may have imprinted on the host only to such an extent that they prefer exploiting, but not mating with, the same kind of host that they grew up with.

Song learning and song preference learning in oscine birds, sometimes referred to as acoustical imprinting, have been studied in great detail, and possess some of the characteristics of imprinting. However, the sensitive period for song learning occurs later than in the classical cases of imprinting, and the time window for learning may be considerably extended. Also, tutor choice in song learning is often more diverse than is the case in filial and sexual imprinting. There may also be a greater degree of flexibility throughout life in song learning. Language acquisition in humans is an analogy to avian song learning.

See also: Behavioral Ecology: Kin Selection; Learning

Further Reading

Bateson, P., 1982. Preference for cousins in Japanese quail. Nature 295, 236–237.
Bischof, H.J., 2003. Neural mechanisms of sexual imprinting. Animal Biology 53, 89–112.
Bolhuis, J.J., 1996. Development of perceptual mechanisms in birds: Predispositions and imprinting. In: Moss, C.J., Shettleworth, S.J. (Eds.), Neuroethological Studies of Cognitive and Perceptual Processes. Boulder, CO: Westview Press, pp. 158–184.
Gottlieb, G., 1971. Development of Species Identification in Birds. Chicago: University of Chicago Press.
Hansen BT, Johannessen LE, and Slagsvold T (in press) Imprinted species recognition lasts for life in free-living great tits and blue tits. *Animal Behaviour* (doi:10.1016/j.anbehav.2007.07.023; available online 29 Oct. 2007).
Hansen, B.T., Slagsvold, T., 2003. Rival imprinting – Interspecifically cross-fostered tits defend their territories against heterospecific intruders. Animal Behaviour 65, 1117–1123.
Hess, E.H., 1973. Imprinting. Early Experience and the Developmental Psychobiology of Attachment. New York: Van Nostrand.
Immelmann, K., 1975. Ecological significance of imprinting and early learning. Annual Review of Ecology and Systematics 6, 15–37.
Kruijt, J.P., Meeuwissen, G.B., 1991. Sexual preferences of male zebra finches: Effects of early and adult experience. Animal Behaviour 42, 91–102.
Laland, K.N., 1994. On the evolutionary consequences of sexual imprinting. Evolution 48, 477–489.
Lorenz, K.Z., 1937. The companion in the bird's world. Auk 54, 245–273.

Slagsvold, T., Hansen, B.T., Johannessen, L.E., Lifjeld, J.T., 2002. Mate choice and imprinting in birds studied by cross-fostering in the wild. Proceedings of the Royal Society of London, Series B 269, 1449–1455.
Slagsvold, T., Wiebe, K.L., 2007. Learning the ecological niche. Proceedings of the Royal Society of London, Series B 274, 19–23.
ten Cate, C., Bateson, P., 1989. Sexual imprinting and a preference for `supernormal' partners in Japanese quail. Animal Behaviour 38, 356–357.
ten Cate, C., Verzijden, M.N., Etman, E., 2006. Sexual imprinting can induce sexual preferences for exaggerated parental traits. Current Biology 16, 1128–1132.
ten Cate, C., Vos, D.R., 1999. Sexual imprinting and evolutionary processes in birds: A reassessment. Advances in the Study of Behavior 28, 1–31.
ten Cate, C., Vos, D.R., Mann, N., 1993. Sexual imprinting and song learning: Two of one kind? Netherlands Journal of Zoology 43, 34–45.

Kin Selection

AS Griffin, University of Edinburgh, Edinburgh, UK

Natural selection is the process by which a trait can be favored because of its beneficial effects on fitness. Kin selection is an extension of natural selection theory that allows for the fact that a trait can be favored because of the beneficial effects on the fitness of relatives. The term itself was first coined by John Maynard Smith in 1964 although the idea that a gene could spread due to beneficial effects on relatives has been appreciated at least since the 1930s.

Kin Selection Optimizes Inclusive Fitness

Natural selection optimizes fitness. Understanding what kin selection optimizes was crucial to its emergence as one of the fundamental tenets of evolutionary theory. W. D. Hamilton was the first to explain this formally in terms of his theory of inclusive fitness in 1964. Hamilton pointed out that when we measure the fitness of a trait we must take into consideration the effect that trait has on the fitness of other individuals as well as the actor who performs the behavior. If a trait has a beneficial effect on carriers of the same gene, then that gene can spread by kin selection. Inclusive fitness therefore is made up of two components – direct fitness (from the production of offspring) and indirect fitness (from aiding the reproduction of relatives). Kin selection explains how a trait can spread by its effect on indirect fitness.

Kin Selection Explains Altruism

The gene-centered view of Hamilton's inclusive fitness theory solves Darwin's problem of how a gene that reduces the fitness of its carrier can evolve. One of the main aims of Hamilton's work was to explain altruistic behavior but in fact, the theory is much more general and applies to any social behavior (see **Table 1**). Examples of animals behaving altruistically toward one another are all around: animals feed one another, groom one another, build homes for one another, defend one another, babysit for one another, and even die for one another. If there is no direct fitness benefit to a helping behavior, then kin selection is the only explanation for the behavior. Crucially, the beneficiary of an altruistic act must have a higher probability of sharing genes in common with the altruist than a random member of the population. Helping relatives is simply the most common way in which this can be achieved.

The spread of a gene for altruism was formalized by Hamilton in what is known as Hamilton's rule. The rule states that a gene will be favored if the following condition is met:

$$rb > c$$

where r = relatedness between the actor of a behavior and the beneficiary, b = benefit to the recipient, and c = cost to the actor. The effect of the behavior to the actor's own lifetime reproductive success is $-c$ and this must be outweighed by the positive effect on the recipient rb.

Testing Kin Selection Theory

As well as being supported by a large body of mathematical theory, the parameters in Hamilton's rule are measurable and the theory is testable. However, it is also deceptively simple – in many cases this simple rule can hide a huge amount of complexity. For example, the term kin selection is used to refer to selection in two different situations: when the gene of interest is shared due to common ancestry alone or more broadly to any situation where the gene is shared. For example, in the latter case, the relatedness between two individuals who are known to carry the gene for an altruistic trait would be $r = 1$, regardless of kinship by co-ancestry.

Hamilton's rule clarifies the predictions of kin selection theory: traits will be more likely to spread if they maximize r and b and minimize c.

Table 1 Categories of social behaviors based on the effect on the fitness of the actor and the recipient

Effect on actor	*Effect on recipient*	
	+	−
+	Mutual benefit	Selfishness
−	Altruism	Spite

 doi:10.1016/B978-0-444-63768-0.00019-6

Maximizing Relatedness

There are two ways in which an appreciable relatedness between social partners can arise.

1. *Kin discrimination*. If, as is often the case, an animal is faced with a decision of who to help, potential beneficiaries may or may not be related. Kin discrimination refers to a process by which an altruist discriminates with respect to relatedness when deciding who to help. In long-tailed tits, in which case helpers have the choice of helping at several different nests in the territory, it has been shown that they preferentially provide help at the nests of relatives (see **Fig. 1**).
2. *Population viscosity*. This refers to a population structure where dispersal is limited from a natal patch. By chance, a potentially altruistic individual will be surrounded by relatives and so any altruistic act it performs will, by chance, benefit those who share the altruistic gene. This is the case in social insect colonies, which are typically founded by one or a few reproductive queens. Viscosity may account for the fact that there is no evidence for kin discrimination in eusocial insects. The costs of such a system are not worth the benefits as workers have not evolved in an environment where help is squandered on nonrelatives.

Although there is a wealth of evidence in support of kin selection theory, much of it is correlative. This is mainly because it is difficult to design experiments where relatedness and altruism can be manipulated. Recently, however, this has been made possible using microorganisms. In populations where relatedness between social interactants was higher, a higher level of cooperation was selected for (see **Fig. 2**).

Maximizing Benefit

One field where a great deal of work has been done on the ability of kin selection theory to explain altruistic behavior is in the study of cooperative breeding in birds and mammals. In such species, nonbreeding helpers remain in the natal territory to help raise offspring in subsequent breeding seasons rather than dispersing to reproduce. The question often asked is: do helpers at the nest preferentially give help to more related individuals? Cooperatively breeding species offer the opportunity to test this prediction of kin selection theory. In several species, such as the long-tailed tit and the Seychelles warbler, helpers were found to discriminate in favor of kin. However, in other species, such as the kookaburra and the meerkats, workers do not appear to discriminate in favor of kin. In such cases, other explanations for helping behavior are needed based on direct fitness benefits. However, it turns out that Hamilton's rule predicts this pattern. Further analyses have shown that the extent to which helpers discriminate depends on the amount of benefit provided by helping: if there is no benefit to help in terms of offspring raised then

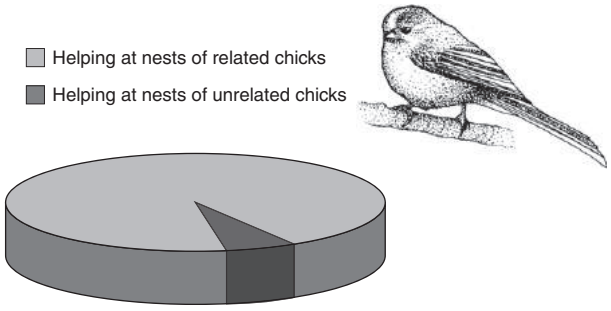


Fig. 1 The difference in the amount of help provided at the nests of relatives (94%) relative to nonrelatives (6%) in a cooperatively breeding bird, the long-tailed tit.

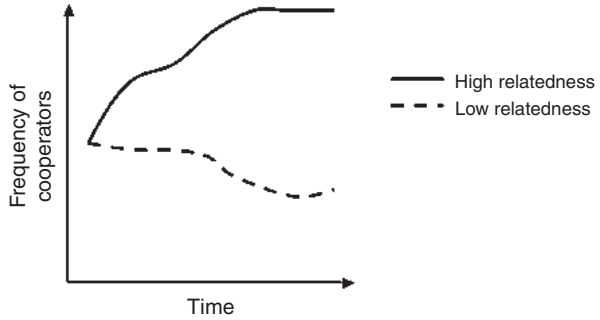


Fig. 2 Two populations of bacteria were maintained and the changes in frequency of an altruistic trait (production of a molecule involved in the scavenging of iron) was monitored across time. When there was high relatedness between cells the frequency of altruism increased, providing the first experimental support for the prediction that relatedness facilitates the evolution of altruism.

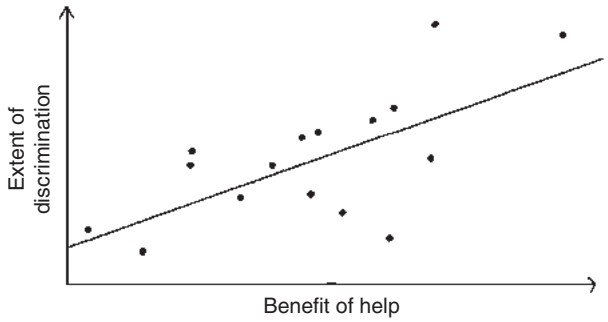


Fig. 3 Each circle represents a species of cooperatively breeding bird or mammal. For each species, the extent of kin discrimination was measured as the effect of kinship on the likelihood or the amount of help given by nonbreeding helpers. The extent of kin discrimination was correlated with the benefit of providing help which was measured as the proportion of offspring surviving to 1 year. As predicted by kin selection theory, when helpers were more helpful, they were also more choosy about which individuals they helped.

there is no incentive to discriminate (see **Fig. 3**). This provides an across-species test of Hamilton's rule – when b is higher, preferential helping of relatives is more likely to be favored.

Minimizing Cost

There is less empirical support for the prediction that cost, in terms of direct fitness, is minimized for the simple reason that it is a difficult parameter to measure. Whereas relatedness can be estimated using pedigree or genetics and benefit can be measured by counting offspring, the cost resulting from competition with relatives is more difficult to quantify. However, kin selection theory has been supported by a study which measured the cost of helping in terms of direct fitness in the hairy-faced hover wasp. In the hairy-faced hover wasp, there is a single dominant breeding female, and helpers that provide aid to the dominant, form a queue to reproduce. The queue is based on age: when the dominant female dies, the second oldest female takes over the breeding position. Cost in this example is easily measured in terms of queue length. Helpers help less when the queue to reproduce is shorter: a shorter queue means more to lose in terms of direct fitness and so selection does not favor investment in helping.

General Application of Kin Selection

The true power of kin selection theory is its generality: kin selection can help explain a huge range of social interactions and not just altruistic cooperation. The simplest cases are when interacting individuals are more closely related, they should be more likely to cooperate, show more selfish restraint, and show less aggression. A range of more subtle possibilities arise whenever there is the potential for cooperation or conflict between relatives. A few examples of these are:

- Individuals are expected to be more likely to give warning calls about the presence of predators, if they are in the presence of close relatives, as occurs in ground squirrels.
- In species where cannibalism occurs in response to food limitation, individuals should prefer to eat nonrelatives, as occurs in tiger salamanders and ladybirds.
- In social insects, such as wasps and bees, workers remove eggs laid by other workers, because they are more related to the queen's eggs, than the worker-laid eggs.
- In many insects, related males (brothers) compete with each other for mates (often their sisters), before these females disperse to lay eggs elsewhere. When this happens, mothers produce a female-biased offspring sex ratio, to reduce this competition between brothers.
- If the relatedness between the parasites infecting a host is high, they are expected to prudently exploit that host, causing less damage and mortality (virulence).

In other words, kin selection theory describes when individuals should behave altruistically and also when they should curtail their selfishness. Furthermore, kin selection theory also predicts the existence of spiteful behaviors, where an individual suffers a personal cost ($c>0$) in order to inflict harm upon a social partner ($b<0$) (**Table 1**). Such behaviors are favored when $rb>c$ is satisfied, which requires a negative relatedness ($r<0$) between spiteful actor and victim. Examples of spiteful behaviors include bacteria producing chemicals that kill nonrelatives, or wasp larvae preferentially attacking and killing individuals to whom they are less closely related.

See also: Behavioral Ecology: Learning. Global Change Ecology: Material and Metal Ecology

Further Reading

Bourke, A.F.G., Franks, N.R., 1995. Social Evolution in Ants. Princeton, NJ: Princeton University Press.
Clutton-Brock, T.H., 2002. Breeding together: Kin selection and mutualism in cooperative vertebrates. Science 296, 69–72.
Dawkins, R., 1989. The Selfish Gene, 2nd edn. Oxford: Oxford University Press.
Dugatkin, L.A., 1997. Cooperation among Animals. New York: Oxford University Press.
Grafen, A., 1991. Modelling in behavioural ecology. In: Krebs, J.R., Davies, N.B. (Eds.), Behavioural Ecology, 3rd edn. Oxford: Blackwell, pp. 5–31.
Griffin, A.S., West, S.A., 2003. Kin discrimination and the benefit of helping in cooperatively breeding vertebrates. Science 302, 634–636.
Hamilton, W.D., 1996. Narrow Roads of Gene Land: Evolution of Social Behaviour. New York: Freeman.
Maynard-Smith, J., Szathmary, E., 1995. The Major Transitions in Evolution. Oxford: Oxford University Press.
Queller, D.C., Strassman, J.E., 1998. Kin selection and social insects. Bioscience 48, 165–175.
West, S.A., Pen, I., Griffin, A.S., 2002. Cooperation and competition between relatives. Science 296, 72–75.

Learning

DR Papaj, EC Snell-Rood, and JM Davis, University of Arizona, Tucson, AZ, USA

Introduction

Learning is ubiquitous among animals and plays an important role in all manner of ecological processes including competition, predation, mutualism, species coexistence, and population regulation. Learning has important consequences for evolutionary change, including biological diversification. Learning is also important to consider when assessing the effect of human activity on animal populations. Here we provide an overview of learning from an ecological and evolutionary perspective.

What Learning Is

The definition of learning has been the subject of long, unresolved debate, and no definition is universally accepted. However, most students of animal learning agree that learning involves a repeatable change in behavior with experience that persists for some time after experience ends (i.e., there is evidence of memory). Furthermore, with a few important exceptions, learned behavior changes gradually with continued experience to some asymptote; wanes if not continually reinforced; can often be undone by a new type of experience; and is more suited to the environment in some way (i.e., associated with higher fitness) than before learning took place.

Learning is just one form of behavioral plasticity. Learning is sometimes referred to as a form of phenotypic plasticity in which the phenotype is behavior. This is not imprecise, but can be somewhat misleading because behavior is itself a form of phenotypic plasticity. Behavior, like any form of phenotypic plasticity, can be described by a `reaction norm' that relates an animal's phenotype, in this case a set of morphologies generated by motor outputs, to particular environmental states. Learning in turn can be described as a mechanism by which the reaction norm representing a particular behavior is modified by experience with the environment. Learning is thus a mechanism of plasticity in behavior, which is itself a type of plasticity.

While learning is an important mechanism of behavioral plasticity, behavioral plasticity also results from motivational and maturation processes. For example, females from a given butterfly population may respond to the same host plant cue in different ways (e.g., laying eggs or not), depending on the number of eggs that are currently matured. The more number of eggs that are matured, the more likely the female will lay eggs in response to the host plant cue. Thus, a response to an environmental cue (a behavior) is different as a result of reproductive state, rather than learning. It can be difficult to distinguish changes in behavior due to motivation or maturation from learning.

Learning involves but is not equivalent to information gain. Learning is more than just information gain. Animals acquire much, if not most, information without use of memory. Any behavioral response to a sensory stimulus is indicative of information gain. For example, a tap on our knee results in a knee jerk reflex. No learning is involved in this reflex, yet information about a mechanical stimulus impinging upon the knee has been acquired by our nervous system and responded to.

The changes in physiological state discussed above can also be used as source of information that can change behavior without requiring memory. For example, a bird's increasing hunger may indicate that food resources are locally scarce. If the bird chooses nest sites on the basis of relative food availability, its choice may involve learning but could also involve an estimate of availability derived from patterns in hunger level.

The equation of learning with information gain may reflect a human bias. For example, we say that we `learn' what time it is or that we `learn' that it is raining this morning. This language is not inappropriate, since humans store so-called declarative knowledge (or, `knowledge of what' such as names and images of things or places) in a type of memory termed declarative memory. However, there is only limited evidence of declarative memory in nonhuman species. This form of memory is difficult to identify in nonhuman animals and its prevalence in animals is unknown.

Learning is not equivalent to synaptic plasticity. There is growing evidence that changes in the structural connections between neurons (termed synapses) mediate learning and memory in animals as different as vertebrates and insects. Nevertheless, an ecological and evolutionary perspective on learning need not and probably should not require *a priori* that a common mechanism underlie all examples of learning and memory. In fact, there is evidence of phenomena that resemble learning in every way, but which do not involve synaptic plasticity. For instance, feeding experience affects diet choice in locusts in a manner that resembles discrimination learning (a form of learning in which an animal learns to discriminate one food type from another). The effect of experience involves at least in part a taste-feedback mechanism in which the sensitivity of sensory receptors to nutrients in the hemolymph is adjusted. Conversely, synaptic plasticity simply means that synaptic function is not fixed with respect to inputs, and there are many kinds of synaptic plasticity unrelated to learning.

 doi:10.1016/B978-0-444-63768-0.00020-2

Who Learns

Learning has been found in every major group of animals in which it has been examined. Although not all animal phyla have been examined with respect to learning and, although no formal phylum-level phylogenetic analysis of learning has been made to date, existing evidence suggests that learning may be ancestral to the animal kingdom. Evidence of learning phenomena in protists affirms this position.

Learning has been ascribed to organisms other than animals and protists, including bacteria, slime molds, and plants. For sure, elements of learning are found across almost all taxonomic groups of living organisms. In unicellular organisms such as phage, networks of cell proteins appear to perform analogously to neural networks, possessing properties that make them capable of a cell-level associative memory. Empirical results suggest that metabolic responses in bacteria are tuned by repeated exposure to particular conditions in a fashion that resembles how behavioral responses are tuned in animals. Evidence of learned behavior in microbes may not be long in coming. Plants, like bacteria, have processes that resemble learning at the level of protein networks within cells, and physiological responses of plant cells consequently change with experience. A kind of memory occurs in plant cells in which information about environmental stress is stored until receipt of a complementary signal required for a developmental change. Although these processes are physiological in scope, the behavioral repertoire of plants is quite rich and learned behavior in plants conceivable.

Why Learn

Benefits of Learning

The basic advantage of learning is that it permits an individual to behave adaptively even in the face of environmental change. The environment must be unpredictable for learning to be of benefit, but not too unpredictable; what happens in the present must relate in some way to what happens in the future for learning to be useful. Straightforward though this benefit of learning is, the benefit of learning in particular contexts may not be so apparent. For instance, despite decades of study of the ontogeny and neuroethology of song learning in birds, the issue of why birds learn song (and the related issue of why animals such as insects and frogs do not) is still not settled.

Costs of Learning

Learning also has costs that arise from the time, materials, and energy necessary to acquire and store information. `Behavioral costs' arise from sampling a range of behaviors early in the learning process. For instance, birds take large amounts of time learning foraging skills and suffer energy losses from their mistakes; species that use a wider range of resources have more pronounced exploratory behavior during this period. Exploration may be costly in terms of time and energy, or in terms of increased risk of attack by natural enemies. `Tissue costs' of learning arise from the neural tissue necessary to acquire and store information. Since neural tissue is particularly energetically expensive to maintain, tissue costs can be considerable.

Costs of learning translate into various tradeoffs with learning ability. In *Drosophila* fruit flies, artificial selection for faster learning in adults resulted in reduced competitive ability by larvae. In *Pieris* butterflies, natural variation in learning ability is causally linked with delays in reproduction.

Learning in an Ecological Context

What Animals Learn

Evidence of learning in animals has been obtained with respect to foraging, territoriality, predator avoidance, dispersal, migration, thermoregulation, communication, mating behavior, parental care, kin recognition, and other social interactions. In these contexts, animals learn relevant features of their environment. Animals can learn the caloric content, nutrient profile, and handling time associated with food types, and the genetic and direct benefits offered by different mates. They learn the quality distribution of essential resources such as food, breeding sites, and mates. In other words, information about virtually every aspect of an animal's ecology may be acquired through learning. Of course, not every animal of every species or population is expected to learn all possible behavior in all ecological contexts. What, if anything, is learned should depend on the net value of learning to the individual. Two observations support this statement. First, animals frequently show biases in what stimuli they are prepared to learn. Second, animals that differ in their ecological requirement for a particular kind of learning differ in how much they invest in such learning.

Learning biases. Animals do not learn all environmental stimuli in all contexts equally well. A classic example of learning bias has been described for food aversion learning. Food aversion learning, best studied in birds, rodents, and insects, involves avoidance of food stimuli associated with a digestive malaise. The sensory properties of food aversion learning in rats are intriguing. Rats readily avoid a flavor associated with an illness but not sound or light. This is not because sounds and light are

generally more difficult to learn than flavors: rats more easily learn to associate sound and light with an electric shock (a rapidly acting punishment) than they do a flavor. This pattern most likely reflects an evolutionary history in which foods and their flavors are sometimes associated with delayed illness, favoring a learned aversion toward them, but such illnesses are very rarely associated with sound or light. Similar learning biases are known for song learning in birds. In playback experiments, for example, sparrows prefer to copy their own species-typical song over that of other sparrow species.

Patterns of investment in learning. What information animals learn should depend on the relative costs and benefits of learning that should in turn depend on a species' or population's particular ecological demands. Spatial learning is critical to navigation (including homing) in many animals, as well as to formation of territories and home ranges. Such learning necessarily entails acquiring, processing, and storing large amounts of information. As such, spatial memory can be expected to be very costly and, in fact, large areas in metabolically expensive brains (e.g., the hippocampus in birds and mammals) are devoted to it. Not surprisingly, in view of its high costs, some of the strongest evidence of species differences in learning occurs in relation to spatial learning. In birds and mammals, for example, spatial learning and associated areas of the brain, like the hippocampus, are better developed in species, populations, or sexes in which spatial learning is relatively important. Home range size, for example, is larger in meadow voles than pine voles and thus the former species has correspondingly better spatial memory and larger hippocampi. In meadow voles, the male's home range is much larger than the female's, a difference associated with superior spatial learning and larger hippocampi in males. Food-storing behavior is correlated with similar species differences in corvids. Nutcrackers, which depend strongly on food caches to survive the winter, have better spatial memory ability than scrub jays or pinyon jays, each of which depend less on caches. Significantly, nutcrackers are no better at learning nonspatial tasks than these other jay species.

When Animals Learn

Ethologists have long been aware that animals are not uniformly sensitive to experience throughout their lives. `Sensitive periods' are periods during development when experience has a particularly strong effect on later behavior. For example, it has been shown in birds from several taxonomic families that mate preference is based largely on the phenotypes individuals experience as fledglings, and that those preferences are only slightly altered later in life. Sensitive periods have been described in the ontogeny of behavior related to habitat, host, food, and mate preferences as well as communication (e.g., bird song) and homing.

The occurrence and timing of sensitive periods are shaped by selection such that an individual's behavior is influenced most by those experiences that most improve reproductive success. For example, young salmon (*Onchorhyncus* sp.) become particularly sensitive to experience with stream odors as their morphological transition from stream-dwelling forms (termed parr) to ocean-dwelling forms (termed smolt) is initiated. In this manner, salmon learn olfactory landmarks as they migrate toward the ocean that they can use later when they move back up those same streams to spawn. Similarly, some bird species have a sensitive period during the fledgling stage in which mate preferences are learned. The phenotype most frequently experienced during this period is that of the parent. Fledglings thus develop a preference for a phenotype that, by evidence of their own existence, represents a reproductively successful conspecific.

Sensitive periods can have important population and evolutionary consequences. Animals whose behavior is shaped early in development may not be able to adjust to environmental changes that occur within their lifetimes, even when adjustments would be beneficial. In a world in which environmental change has been accelerated greatly by human activity, the occurrence of a sensitive period may result in a potentially tragic kind of behavioral obsolescence. For example, in homing species such as sea turtles and salmon, individuals choose their spawning grounds based entirely on early experience, and typically do not alter those choices. Animals may consequently be `trapped' into reproducing in suboptimal habitats if those habitats are degraded within their lifetime by human activity.

How Animals Learn

How animals learn has long been the domain of psychology, where learning processes have been described with the objective of defining underlying cognitive mechanisms. More recently, psychologists have joined with ecologists to ask why animals learn in the way that they do. This functional approach to how animals learn is readily introduced with respect to the basic categories of nonassociative and associative learning. Nonassociative learning is exemplified by habituation, defined as the waning of a response to a stimulus upon repeated presentation of that stimulus. Habituation permits an animal to suppress wasteful neural processing of and behavioral responses to stimuli that are irrelevant to fitness. Associative learning entails pairing a stimulus with another stimulus (S–S), or with a motor response (S–R), in space and time such that the response to the first stimulus is altered as a consequence of the pairing. Associative learning enables animals to describe the correlative structure of information in their environment and to adopt behavior that takes advantage of that description. Animals learn by association to orient toward stimuli predicting fitness gains (e.g., acquisition of food or mates) and away from stimuli predicting fitness losses (e.g., effects of heat, toxins, or natural enemies).

The use, or not, of social learning offers another example of a functional approach to learning. Social learning, long known for vertebrates but recently described in insects and other invertebrates, is a form of learning that is facilitated by interactions with other individuals, usually of the same species. Social learning is diverse in terms of mechanisms. One form, local enhancement, occurs when an individual is attracted to a locale where another individual is present, and consequently learns something in that

locale. Another form, observational conditioning, occurs when an observer learns a stimulus–reward association by watching a demonstrator experience that association. For example, nectar-foraging bees might learn to prefer a certain floral color by watching other bees visit and extract nectar from flowers of that color. Acquisition of a fear of snakes in monkeys who watch conspecific respond fearfully to snakes has been categorized as observational conditioning. Imitation involves learning in which a pattern of behavior engaged in by an individual is copied. Song learning in birds is viewed by some as a special case of imitation. True imitation that involves imitation of a novel behavior, is rare in animals and difficult to document, but noteworthy because it allows for the rapid spread in a population of truly novel behavior. Chimpanzees learn to use tools to obtain food by imitating other, usually older individuals.

Social learning has costs and benefits that overlap partly, but not entirely with those of individual (asocial) learning. An animal might therefore adjust its use of social or asocial learning as the costs and benefits of each change. This appears to be the case, at least with respect to food-foraging behavior in birds. When a task associated with finding food is made more difficult in terms of individual learning, birds are more likely to rely on social learning.

Memory processes are similarly open to functional explanations. Forgetting is intuitively considered to reflect a constraint on the durability of memory. However, forgetting might be adaptive if it permits animals to discard obsolete information in a way that facilitates the acquisition, storage, and use of new, more relevant information. Forgetting of obsolete information may also reduce `operating costs' associated with maintaining memory.

A variety of learning processes described mainly in the laboratory may have functional significance in nature. Generalization, a ubiquitous phenomenon in which animals trained to a stimulus S+ respond also to stimuli that resemble S+ along some perceptual dimension, was once assumed to reflect a constraint on learning capacity. From an ecological perspective in which variability in the environment is viewed as a given with which all organisms contend, generalization seems functional. A forager that learns to search only for an S+ associated with a food reward and ignores related stimuli may overlook perfectly suitable food items. Generalization is thus a mechanism for coping with stochastic variation in environmental stimuli.

An adaptive argument can also be made for a phenomenon in learning studies known as blocking. Blocking occurs when an animal that first learns to respond to a stimulus (A+), and is then reinforced on A and a novel stimulus, B, presented together ([AB]+), subsequently fails to show an enhanced response to B alone. Learning of stimulus B has been `blocked' by coupling it with the previously learned stimulus A. Blocking illustrates a kind of economy in learning: for a stimulus to be learned, it must convey new information. Given a cost to storing information, blocking is functional because it minimizes the storing of redundant information.

Functional explanations like those above have been put forward for a variety of other phenomena including attentional shifts, latent inhibition, peak shift, and effects of local context on learning.

Learning and Evolutionary Change

Effects on Adaptive Evolution

Learning can influence evolutionary change in a number of ways. First, if the environment changes, learning may allow a population of individuals adapted to the old environment to persist until adaptations to the new environment evolve. This assertion implies that animals can learn to cope with challenges different than those the population evolved to meet. In fact, learning does often enable individuals to solve novel problems (as evidenced by numerous examples of animals learning to coexist with humans and even to flourish in novel urban environments). Once a population has a `foot in the door' in a new environment, its members may evolve to learn the new behavior faster or to express the behavior congenitally. Simultaneously, selection may favor morphology, physiology, or biochemistry that is suited to the new environment. These new morphological/physiological/biochemical traits must add something beyond what is afforded by the learned behavior and existing traits. Otherwise, learning can actually retard their modification. For example, if learning enables an animal to utilize a new resource perfectly well with its existing morphology, there may be little or no selection to alter that morphology.

Learning can also facilitate evolutionary change in morphological, physiological, or biochemical traits by promoting adaptive linkages among such traits. For example, a genetic variant in bill morphology in a population of birds may learn to exploit food types to which the bill is best suited, food types which are different than those exploited by individuals with different jaw morphology. The new diet will select for gut physiology appropriate to the diet. In this way, gut physiology and jaw morphology may evolve as correlated traits. Such correlated traits can greatly accelerate adaptation to the new resource and rapidly drive genetic differentiation within and among populations.

Effects on Diversification

By allowing individuals to behave adaptively given their particular phenotype, learning can permit a broader range of phenotypes and underlying genotypes to coexist in a population. For example, in one parasitoid wasp species, individuals of different size learn to use host species that they can handle most efficiently, generating a correlation between wasp size and

host preference. Wasps of all sizes may be more or less equally successful, permitting a broader range of phenotypes to coexist in the population. In time, the different phenotypes may experience selection on whatever traits promote fitness in their different host niches. The end result could be greater genetic differentiation within the wasp population than would occur in the absence of learning.

Learning can also promote genetic differentiation by restricting gene flow between populations or subpopulations. In the apple maggot fly, conditioning has been suggested to facilitate the formation of host races in sympatry on two alternative host fruit species, apple and hawthorn. Conditioning may facilitate host race formation in at least two ways: first, by biasing allocation of eggs by a female to its natal host fruit species, and second, by suppressing dispersal by adults from the natal host tree. Both effects can be considered examples of `natal habitat preference induction' and both may reduce gene flow between the two host types. A similar process may occur in the indigobird, an avian brood parasite. In indigobirds, females exclusively lay eggs in the nests of the host bird species they were reared by as nestlings. In addition, they learn to prefer the host species' male call. Because male indigobirds mimic the songs of their host species, matings are most likely to occur between birds that were reared by, and thus adapted to, a particular host. Such assortative mating can promote sympatric differentiation and even lead to the formation of host races.

Analogous effects of learning on diversification may occur at higher taxonomic levels. For example, filial imprinting in birds has been proposed to promote speciation in birds. Avian learning has been linked to diversification: rates of diversification of species lineages are correlated positively with that lineages' rates of innovation, which is mediated in part by learning. Moreover, in birds, innovative behavior can spread particularly rapidly through social learning. Interspecific correlations between social learning and innovation frequency suggest that learning interacts synergistically with innovation to promote diversification.

Learning and Species Interactions

Learning is important to consider in all manner of species interactions, including mutualism, competition, predation, and parasitism. In all cases, learning by one species in the interaction has the potential to shape the evolution of traits in the other species. For example, learning by predator species is thought to have played a role in the evolution of warning coloration in potential prey species. Warning coloration is coloration that serves to advertise the noxiousness of potential prey, to the mutual benefit of both prey and predator. It is typically highly conspicuous. One advantage of conspicuous coloration is that it is easier for the predator to detect from a distance. However, conspicuous coloration has two additional advantages. It is learned faster by predators and remembered longer, effects which should favor the evolution of conspicuous coloration.

A prey's conspicuous coloration and a predator's propensity to learn conspicuous colors could evolve in tandem, a prediction that remains to be tested. The possibility that learning in one species may be shaped by interactions with another species has been suggested for imprinting in avian brood parasites. Female brood parasites imprint to their hosts before fledging, a process which causes them later to be attracted to the nests of host species. However, the degree of sexual imprinting by females on their hosts is dictated by a tradeoff. Imprinting must be strong enough to allow females to be attracted to the nests of the host species and thus continue the brood parasitic cycle, but not so strong as to prevent females from mating with males of their own species. One strategy circumventing the mate-recognition problem, which appears to be implemented in cowbirds, is self-phenotype-matching wherein the maturing cowbird learns what it looks like and interacts selectively with individuals that resemble it. In effect, the parasitic lifestyle has led to self-phenotype-matching filling the role of imprinting in cowbirds.

Learning and Environmental Change of Human Origin

Human activity and population growth is causing rapid, radical changes in the habitats of many, if not most, species. In contrast to many forms of cue-based phenotypic plasticity, such as seasonal polyphenisms, learning can provide a wide range of beneficial responses to novel environments. Trial-and-error learning may result in `innovative' behaviors, which may facilitate adaptation to environments quite different from those in an animal's recent evolutionary history. For instance, some research suggests that animals can learn to recognize novel, invasive predators in genera or families other than those to which native predators belong. Such innovative behaviors may help certain species invade novel habitats: birds with larger brains tend to establish themselves better in novel environments.

Global climate change represents one of the more striking consequences of human activity. As with other environmental change, learning will likely play an important role in the ability of species to cope with global climate change. Some birds and mammals have already begun to adjust their timing of reproduction to accommodate changes in temperature and the availability of their prey; it has been suggested that learning is involved in this accommodation. The extent to which learning facilitates species' abilities to cope with global change remains an open area of study with important implications for evolution and conservation. Although learning may allow species to adjust to global change in some behavioral contexts, climate change may conceivably select against learning. As noted above, learning is likely to evolve only under conditions of moderate levels of unpredictable environmental variation. If weather becomes highly unpredictable, as suggested by some global change models, learning may not be useful in tracking variability in weather. The links between learning and how species adjust to human activity remain to be elucidated.

See also: Behavioral Ecology: Kin Selection. Global Change Ecology: Material and Metal Ecology

Further Reading

Bond, B., Kamil, A.C., 2006. Spatial heterogeneity, predator cognition, and the evolution of color polymorphism in virtual prey. Proceedings of the National Academy of Sciences of the United States of America 103, 3214–3219.
Chittka, L., Thomson, J.D., Waser, N.M., 1999. Flower constancy, insect psychology, and plant evolution. Naturwissenschaften 86, 361–377.
Davis, J.M., Stamps, J.A., 2004. The effect of natal experience on habitat preferences. Trends in Ecology and Evolution 19, 411–416.
Dugatkin, L.A., 2004. Principles of Animal Behavior. New York: WW Norton.
Dukas, R. (Ed.), 1998. Cognitive Ecology: The Evolutionary Ecology of Information Processing and Decision Making. Chicago: University of Chicago Press.
Galef Jr., B.G., Laland, K.N., 2005. Social learning in animals: Empirical studies and theoretical models. Biosciences 55, 489–499.
Giraldeau, L.-A., Valone, T.J., Templeton, J.J., 2002. Potential disadvantages of using socially acquired information. Philosophical Transactions of the Royal Society of London 357, 1559–1566.
Irwin, D.E., Price, T., 1999. Sexual imprinting, learning and speciation. Heredity 82, 347–354.
Johnston, T.D., 1982. Selective costs and benefits in the evolution of learning. Advances in the Study of Behavior 12, 65–106.
Lynn, S.K., Cnaani, J., Papaj, D.R., 2005. Peak shift discrimination learning as a mechanism of signal evolution. Evolution 59, 1300–1305.
Mery, F., Kawecki, T.J., 2003. A fitness cost of learning ability in *Drosophila melanogaster*. Proceedings of the Royal Society of London Series B 270, 2465–2469.
Shettleworth, S.J., 1998. Cognition, Evolution, and Behavior. New York: Oxford University Press.
Skow, C., Jakob, E., 2006. Jumping spiders attend to context during learned avoidance of aposematic prey. Behavioral Ecology 17, 34–40.
Sol, D., Timmermans, S., Lefebvre, L., 2002. Behavioral flexibility and invasion success in birds. Animal Behaviour 63, 495–502.
Tumlinson, J.H., Lewis, W.J., Vet, L.E.M., 1993. Parasitic wasps, chemically guided intelligent foragers. Scientific American 268, 100–106.
West-Eberhard, M.J., 2003. Developmental Plasticity and Evolution. Oxford: Oxford University Press.

The Marginal Value Theorem in a Nutshell

Vincent Calcagno, Université Côte d'Azur, Sophia Antipolis, France

Glossary

Average rate of gain (E) Total gain acquired, per unit of time, over a large number of patch visits.
Gain function (F) Function giving the gains acquired by an individual after exploiting a patch for some time (t).
Giving-up density (GUD) Level of resource the individual lets in a patch upon leaving.
Functional response (h) Function giving the instantaneous rate of consumption as a function of the current resource level in a patch (n).
Marginal gain Additional gain obtained if staying a little longer in the current patch (time derivative of the gain function F).
Travel time (T) Average time needed to enter the next patch after leaving one.

History and Background

Optimal foraging theory has produced some very general and successful results, one of which is the marginal value theorem (MVT). Even today, 60 years after the seminal publication (Charnov, 1976; although one can trace it back to slightly earlier papers), the MVT remains one of the most cited theories in ecology. It has been repeatedly tested experimentally and confronted to data, in organisms are different as viruses, insects, plants, ungulates or humans. It is commonly considered as providing satisfactory predictions that resemble observed patterns, even though, of course, an abstract and general model cannot pretend to match many datasets exactly. Several reviews are available and these aspects will not be discussed in further detail here.

Sometimes referred to as the "patch model" in behavioral ecology textbooks (as opposed to the "prey model," another foundational result from optimal foraging theory), the theorem predicts how long an individual should keep exploiting a resource patch before leaving in search of another. Arguably, MVT is not the most informative name. First, marginal value is a general and earlier concept in optimization theory, largely employed in the economical sciences for instance. Second, while theorem puts emphasis on the calculations, the result is a rather straightforward consequence of the definition of an optimum. The defining feature of the MVT probably resides more in the modeling choices and hypotheses (the patch model), than in the use of marginal values or in the derivation itself. One can find strong analogies between the MVT and optimality models under variable-interval schedules in behavior and psychology (Staddon, 2001). However the two seemingly developed independently; the latter does not quite predate the original MVT papers, and cross references are almost inexistent. Also, albeit mathematically equivalent, assumptions regarding which quantities are fixed and which are flexible differ radically. Therefore the MVT is a genuine product of behavioral ecology in the early 1970s, at a time when the application of economical principles in ecology flourished. Let us have a closer look at the underlying biological model.

Model Assumptions

The model assumes that the environment of an individual consists in many patches: spatially discrete entities where the individual can perform a rewarding activity and thus acquire gains (**Fig. 1**). This has a clear virtue of conceptual clarity, and in practice patches range from an obvious property of biological systems (compartments such as cells, individuals, flowers or water ponds) to a useful simplification of more continuous landscapes. In that it follows the familiar strategy in ecology of binarizing habitat complexity, as does for instance metapopulation theory. An individual can be in one of two states: (i) in a patch performing the profitable activity (**Fig. 1**; *orange*) or (ii) outside patches, performing other activities (**Fig. 1**; *green*). The course of time is therefore characterized by an alternation of (i) phases ("exploitation," e.g., eating) and (ii) phases ("exploration," e.g., searching).

The MVT considers that individuals (through reasoning, learning, plasticity mechanisms, or natural selection screening many variants) can adjust the time of (i) periods, as they can decide to interrupt a visit and leave the current patch, thus starting a novel (ii) period. They therefore control how long to stay on any patch, and this is called the residence time. In contrast, they cannot act of the duration of (ii) periods. The latter are envisioned as unavoidable and incompressible; if anything, they should be as short as possible considering a variety of external constraints. An individual has to spend some time outside patches before finding a new patch, and this time constraint is classically called the travel time (T). It represents the average time it takes to enter a new patch after leaving one. The duration of these periods, though it needs not be constant and can vary randomly, is independent of the duration of (i) periods. For instance, there is no fixed duration for a (i)/(ii) cycle, and no correlation between the time spent in patches and the speed at which the next patch will be found. An individual can act on the number of patches visited per unit of time, but only by adjusting its residence times (by staying less on each patch, it visits more). A nice property of the MVT is that the

 doi:10.1016/B978-0-12-409548-9.10874-7

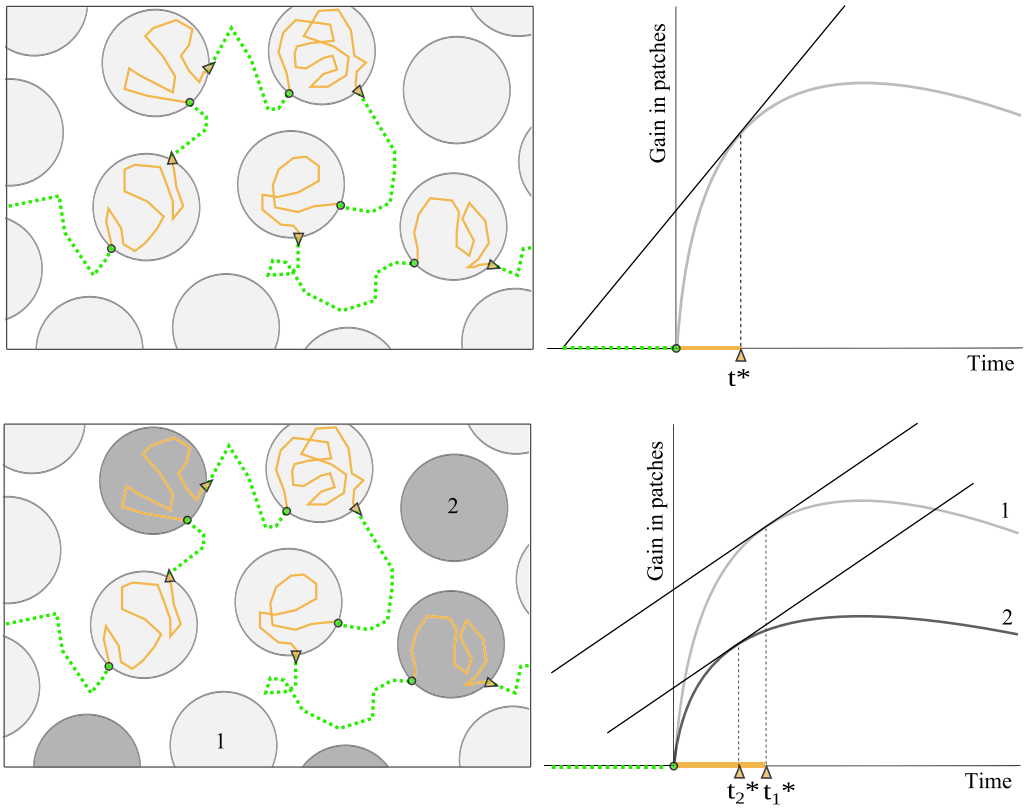


Fig. 1 Graphical interpretation of the MVT. *Top:* homogeneous habitat. *Left:* the habitat is composed of many patches (*gray disks*) among which the individual navigates (*dotted green lines*). When entering a patch (*green circles*), it spends some time exploiting it (*orange*) before deciding to leave in search of another (*triangles*). *Right:* while in patches the individual acquires gains as given by the gain function. If all patches have the same gain function, the individual maximizes its overall rate of gain by spending time t^* in each patch. This optimal residence time can be found by drawing the line tangent to the gain function starting from $(-T, 0)$, whose slope is E^*. *Bottom:* heterogeneous habitat. *Left:* two kinds of patches were assumed—good ones (1; *light gray*) and poor ones (2; *dark gray*). *Right:* having computed E^* from Eq. (1), optimal residence times are such that the gain functions have slope E^* in all patches. Here the individual should spend more time on good patches ($t_1^* > t_2^*$; that is, $\rho_{\text{INTRA}} > 0$). Modified from Charnov, E.L., 1976. Optimal foraging, the marginal value theorem. Theoretical Population Biology 9 (2), 129–136.

distribution of T does not matter, only the average does. The latter quantity is treated as a fixed parameter that depends both on the characteristics of the environment (e.g., the average distance between patches) and of individuals (e.g., their efficiency/speed at locating new patches and attaining them).

The process of gain acquisition during (i) sequences receives most attention. It is assumed that acquiring gains takes time, so that the total gains acquired during a patch visit of duration t are given by some gain function $F(t)$. Note that this gain function represents both the quality of a patch (Is it good or poor?) and of the individual (How efficient is it at extracting gains from the patch?). Quite universally, this function is considered to represent a law of diminishing returns, that is, the longer one stays in a patch, the less profitable it becomes to stay more. In mathematical terms, it means that F is an increasing function of time, but at a decreasing pace, that is, it should be concave at some point in time. It can otherwise have any shape. An important point is that F must represent net gains acquired, discounting potential associated costs. Therefore it may be negative for some time intervals (when the costs outweigh the gains acquired) and/or ultimately decreasing (if the rate of gain acquisition eventually drops below the costs of foraging in the patch).

Each patch can have its own gain function, and the environment of the individual is thus defined (besides the travel time) by the collection of gain functions encountered during consecutive patch visits. The simplest homogeneous case considers that all patches have one common gain function, but in general there will be a distribution such that a fraction p_i of patches will have gain function F_i. Importantly, this distribution is unaffected by what the individual does. In practical terms, it means either that all patches are visited only once (no revisits), or that revisits are sufficiently distant in time for patches to have restored their initial state by the next time they are visited. Recurrent patch revisits and less-than-perfect recovery of patch quality could be accommodated for in the MVT framework with limited changes, although this has not been much explored.

A second defining assumption is that individuals cumulate gains over time and seek to maximize the total gain accumulated (G) over the duration of an entire sequence (D). In practice, if the duration D is unaffected by what the individual does (for instance, if it is constant), this is equivalent to maximizing the average rate of gains over the period ($E = G/D$), and the MVT is classically presented as maximizing the latter. Maximizing this quantity implies that gains have the same value irrespective of when

they occur, and of what has already been accumulated during previous patch visits. For instance, it means that severe starvation does not occur in the short-term, so that an individual would never desperately need to find resource as soon as possible. Similarly, the total duration of the foraging sequence (D) should not be affected by patch leaving decisions, as would be the case if moving frequently between patches shortened average life expectancy (see also section "Going further and open questions").

In these conditions, the average rate of gain is the total of gains accumulated, divided by the total time elapsed ((i) and (ii) sequences). If there have been N patch visits (N large), the quantity an individual would seek to maximize is thus

$$E = G/D = \sum_{j=1}^{N} F_j(t_j) / \sum_{j=1}^{N} (T + t_j) = \frac{F_j(t_j)}{T + t_j}$$

where $< >$ represent averages over all patches visited.

The Classic MVT

Maximizing E implies that an individual should act in a way so that the residence time on patch i is t_i^*, as defined from the equation:

$$F_i'(t_i^*) = \frac{F_j\left(t_j^*\right)}{T + t_j^*} \tag{1}$$

These residence times are called optimal residence times, and the r.h.s. of the equation is therefore the optimal average rate of gain they permit to achieve (E^*). The equation defines a maximum as long as all gain functions have concave shape around the residence times t^*. Note that the maximum needs not be global: in particular, some of the patches (if there are too poor compared to the others) should be completely unexploited,[1] in which case their t^* must be set to 0 in Eq. (2). The equation holds even in the face of variability in travel times and the patch types, if the number of patch visits (N) is large enough. The exact meaning of large is not precise, but in practice it can be as low as 20–30. In other words, the total duration of the sequence (D) must be long enough for sufficiently many patches to be visited.

The use of "in a way so that the residence time on patch i is …" was purposely vague: the MVT, as many evolutionary predictions, does not pretend to specify how individuals should achieve the particular phenotype. The mechanisms brought up to achieve it will likely depend on the organisms studied, their evolutionary history and sensory and cognitive capacities. The ontogeny of t_i^* values, and the identification of rules-of-thumb that an individual could use in practice, are strictly speaking out of the scope of the MVT. For instance, an individual might compute its average rate of gain E and decide to leave a patch when the instantaneous rate of gain falls below E, or it might monitor the flow of time and leave when t^* has elapsed, or any other algorithm one can imagine. Any such algorithm may prove feasible or unrealistic depending on the study species, and clearly for many organisms the capacity to implement the optimal residence times may be a major constraint. We do not enter into these interesting considerations here, as they do not question the fact that t_i^* is an optimal strategy an organism would want to target.

In its simplest form, assuming that all patches identical, averages disappear and we get only one equation:

$$F'(t^*) = \frac{F(t^*)}{T + t^*}$$

This homogeneous MVT equation has the advantage of being easy to solve using an elegant graphical construct, which certainly contributed to the popularity of the MVT (Fig. 1). In a heterogeneous habitat, that is, when different patches visited do not all have the same gain function, the graphical construct cannot be used and one must solve Eq. (1) by other means.

So far the function $F(t)$ was not specified and can essentially have any shape. This is one strength of the MVT, allowing it to retain great generality and flexibility, as different situations will likely result in very different gain functions. We can obtain some general predictions at this level (Calcagno *et al.*, 2014a), but we may also want to be more specific and specify what F looks like, in order to obtain quantitative predictions or go further in the analyses. The simplest, and earliest, approach is to postulate some reasonable and convenient function, such as a Monod function or Holling disk equation. However, this has severe limitations in two ways: (i) in most cases we have no knowledge of the actual gain functions and similar functions are hard to tell apart from available data; and (ii), more importantly, some predictions thus obtained happen to be very dependent on the exact function chosen. To circumvent these limitations, a more recent approach was proposed that restricts the class of possible functions based on properties of the mechanisms at play, without going all the way to proposing specific functions. This is introduced briefly below.

The MVT for Resource Consumers

The gain function in a patch is usually assumed to be concave (saturating) because of resource depletion: as the individual consumes the resources initially present in the patch, patch profitability declines until virtually nothing is left to be consumed. This

[1]Quite simply, those patches are those for which $F(t^*)/t^*$ would be less than E^*.

emerges in a range of situations including predators feeding on prey populations or individuals, pollinators feeding on nectar, males fertilizing female eggs, or humans gathering fruit or catching fish. The particular case where gains represent the consumption of resources is of major interest in ecology. Processes of resource consumption are classically modeled by describing the dynamics of resource density through time, with the consumption rate at any time given by the so-called functional response of individuals (Fig. 2). By writing the initial density of resources in a patch as $n0$, the size (scale) of a patch as S, and given some functional response $h(n)$, the resource density in an exploited patch varies as $dn/dt = -h(n)/S$. We can thus deduce the gain function of an individual from the amount of resource consumed as:

$$F(t) = \gamma S(n0 - n(t)) = \gamma \int_0^t h(n(\tau))d\tau \quad (2)$$

with γ the energetic value of one resource unit (Calcagno *et al.*, 2014b).

Note that function h has not been specified. Just like in the classic MVT, $F(t)$ may thus have infinitely many different shapes, depending on h. However, some F functions one might consider in the classical MVT could not result from Eq. (2) and are thus not permitted, for they are incompatible with the proposed mechanism. As shown below, we get much more precise and easily interpretable predictions in the context of Eq. (2) than in the context of the general MVT, while retaining great generality.

What Questions Can Be Addressed?

The MVT can be used to predict several quantities of interest. First, it can tell what the average rate of gain (E^*) will be (or would be for an optimal individual) in a given habitat. This is important as the average rate of gain is very much a proxy of fitness, conditioning survival and/or reproduction. Second, it can tell us how the optimal residence times vary with environmental and patch characteristics. Optimal residence times are of special interest to behavioral ecologists and those interested in movement as they are closest to decision mechanisms and condition how frequently an individual moves over different patches in its habitat. Finally, the MVT can inform us on the level of patch exploitation, that is, for resource consumers, the amount of resource depletion: how much resource, if any, should be left in patches? The latter quantity, often called the giving-up density (GUD), can have intrinsic value (e.g., it may reflect the amount of pest suppression by a biological control agent) and is often more amenable to observation.

From these quantities, a broad range of questions can be addressed, which has sometimes generated some confusion in the literature. It is helpful to discriminate three intertwined, but slightly different, types of questions:

(1) What should be the relative exploitation of the different patches within a given habitat?
A typical question: should an individual stay longer on the best patches?
(2) What are the consequences of changing habitat characteristics, letting patches (i.e., gain functions) unchanged?
A typical question: should individuals stay longer on patches when those are longer to find?
(3) What are the consequences of changing the patches (i.e., the gain functions) in a habitat?
A typical question: should individuals stay longer on patches if patches contain more resources?

In the first case, we are comparing different patches in one given habitat: as a consequence E^* is a constant, which is a great simplification. We only need to understand the consequences of changing the gain function in one patch on the exploitation level

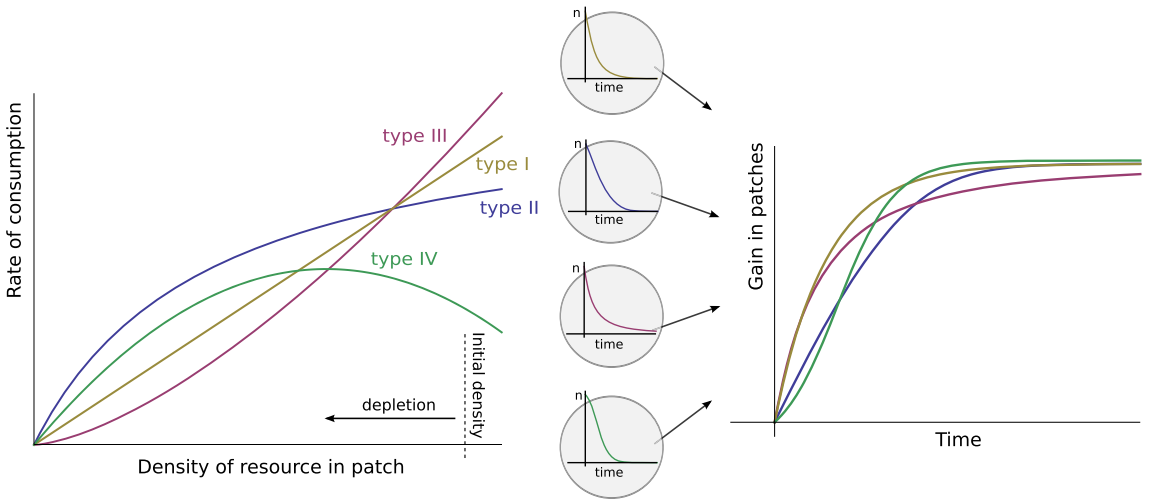


Fig. 2 From functional responses to the MVT. *Left*: the functional response (h) determines how the rate of consumption varies with the density of resource in a patch (n). *Middle*: As the individual exploits a patch, the resource level gradually drops from its initial level. *Right*: This in turn determines the gain function of the individual. Note that contrasted functional responses (*left*) can yield similar looking gain functions (*right*). Type IV functional responses (*green*) can yield sigmoid (initially accelerating) gain functions, as illustrated.

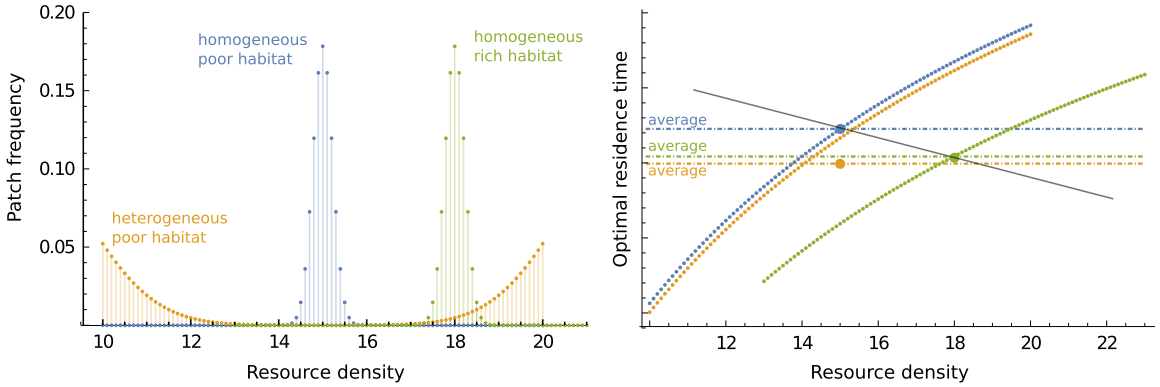


Fig. 3 An illustrative example of habitat modifications. *Top*: three habitats are considered that differ in resource distribution. *Blue* and *orange* habitats differ in the variance of patch resource density, with the same average (15). The *orange* habitat has patches with very contrasted resource densities, in the range 10–20, whereas the blue mostly has average patches. *Blue* and *green* habitats differ only in mean resource density (15 vs. 18). *Right*: An accelerating functional response was assumed and optimal residence times computed numerically. In all habitats an individual should spend more time on the best patches, that is, $\rho_{\text{INTRA}} > 0$ (**Table 2**). As predicted from **Table 2** for an accelerating functional response, increasing the mean resource density decreases the average residence time. The cross-habitat correlation of residence time and resource density is thus negative (*gray line*). The average residence time also decreases with increasing patch heterogeneity, keeping the mean constant (variance effect; see **Table 2**). Remark that in the heterogeneous (*orange*) habitat, contrary to the more homogeneous cases, the average residence time differs importantly from the residence time on the average patch.

of that patch, given some value of E^*. In the second case, the characteristics of the habitat, for instance the travel time (T), vary, and so E^* varies as well. However, since the gain functions stay unchanged, we only need to understand the consequences of changing E^* on the exploitation level in a patch, given its gain function. Clearly, in the third case, both effects must be taken into account simultaneously, as changing the gain functions in a habitat also impacts E^*. Two sorts of effects are particularly worth studying in this case: (i) the impact of the average quality of patches in the habitat (comparing good and poor habitats, that is, a 1st order effect); and (ii) the consequences of habitat heterogeneity (introducing variance in patch quality and comparing homogeneous and heterogeneous habitats; a 2nd order effect). **Fig. 3** provides an illustration of these two possibilities. Unsurprisingly, case (3) is not only the most interesting, but also the one for which predictions prove most challenging to obtain.

Below is an attempt to provide a comprehensive overview of MVT predictions, for the three types of questions, with emphasis on the third. Predictions from the classic MVT are first presented, before introducing the additional insights one can get in the case of resource consumers.

What Is Predicted by the MVT (and What Is Not)

Within a Given Habitat

As a direct consequence of Eq. **(1)**, patches in a given habitat should all be left at times such that the marginal rate of gain is the same (see **Fig. 1**): this "quitting rate" is invariant within a habitat.[2] It is a standard expectation from optimization and economics, with little specific to the MVT (it holds if the r.h.s. of Eq. **(1)** is replaced with any constant). In practice, it means that the marginal rate of gain of individuals when they leave patches can be used to infer the quality of the overall habitat. Since it is often challenging to measure the marginal rate of gain of individuals, for resource consumers people usually use the giving-up density as a surrogate.

As far as residence times are concerned, there is no unique prediction regarding their ordering with patch quality. Depending on what patch quality implies in terms of gain functions, the correlation between patch quality and residence time within a habitat (called ρ_{INTRA} henceforth) can have any sign (**Table 1**). Broadly speaking, when better gain functions get steeper, we expect a positive correlation, whereas when better gain functions get shallower, we expect a negative one. Distinguishing the two situations is sometimes tricky. The sign of ρ_{INTRA} also controls, if looking at a particular patch in the habitat, the sign of variation of the residence time on that patch if its quality were increased (keeping the rest of the habitat, and thus E^*, constant).

Changing the Habitat, but Not the Gain Functions

Provided the gain functions are left unchanged, MVT predictions are straightforward. Changes in the travel time (T) have unambiguous and intuitive impacts on the various quantities of interest, as summarized in **Table 1**. Remark that increasing T decreases the average rate of gain E^*, and it is obvious to see that any decrease in E^*, given a constant gain function for a patch,

[2]This of course only holds for patches effectively exploited ($t^* > 0$).

Table 1 Predictions of the classic MVT (from Eq. (1))

		E^*	t^*	GUD	ρ_{INTRA}
Travel time ↑		↓	↑	↓	NA
Patch quality	Mean ↑	↑	?	?	?
	Variance ↑	↑ or 0	?	?	

The first three *columns* show how each quantity of interest respond to changes in the environmental characteristics (*rows*) of a habitat. "?" indicates that no general conclusion can be reached (response depends on the specifics). For heterogeneous habitats, t^* and GUD should be intended as their spatial averages (over all patches). The last column shows how residence time correlates with patch characteristics within a habitat.

would similarly increase the optimal residence time on that patch. Therefore, a related prediction is that, looking only at one particular patch, any change in the rest of the habitat that increases E^* (e.g., an increase in the overall frequency of good patches) shortens the optimal residence time (and increase the GUD) on this particular patch. Those results are easily deduced from graphical arguments (play with the slope of the tangent lines in Fig. 1) and are probably the ones most often reported about the MVT, even though they only cover a fraction of the situations of interest.

Changing the Habitat Through the Gain Functions

Predictions were so far about changing one patch in a constant habitat (see section Within a given habitat), or changing the habitat for some constant patch (see section Changing the Habitat, but Not the Gain Functions). What happens in the general case, when an entire habitat is modified through the gain functions of its constituting patches? Predictions happen to be much scarcer (Table 1). Optimal residence times, in particular, can respond idiosyncratically to changes in patch quality depending on what exactly quality implies in terms of the gain functions (Calcagno *et al.*, 2014a). A couple of firm conclusions can nonetheless be established in all generality:

- Introducing variance in the gain functions (patch heterogeneity), keeping the average constant, generally increases the average rate of gain E^* (Table 1). In some special cases, fitness is unaffected by patch heterogeneity,[3] but this is not the general expectation. In biological terms, this prediction reflects the fact that an optimal forager is able to take advantage of patch heterogeneity, by disregarding the poorest patches and focusing on the best patches. Incidentally, it implies that knowledge of the average gain function over all patches is insufficient to predict neither the average rate of gain nor the optimal residence times. A forager that would not account for patch heterogeneity, devising its strategy from the average patch only, would achieve smaller fitness, and as a consequence over-exploit all patches, compared to a fully-informed optimal forager.
- If residence time correlates negatively with patch quality in a given habitat ($\rho_{\mathrm{INTRA}} < 0$), increasing the overall quality of patches in the habitat necessarily decreases all residence times (and thus average residence time). However, the reverse in not true (Calcagno *et al.*, 2014a). If residence time covaries positively with quality ($\rho_{\mathrm{INTRA}} > 0$), we cannot tell whether an increase in the quality of all patches would increase or decrease residence time. We can thus formulate this general expectation: it will be more common to observe a positive correlation between patch quality and residence time within habitats ($\rho_{\mathrm{INTRA}} > 0$), than a positive correlation between patch quality and residence time across different habitats.

From Table 1 it appears that frustratingly many answers are left unanswered in the classic MVT. Fortunately, much more precise predictions can be obtained in the context of resource consumers and Eq. (2), with patch quality intended as resource content (e.g., number of prey items).

Refining Predictions for Resource Consumers

Table 2 provides a synthetic overview of predictions regarding the effect of varying the resource distribution (mean and variance over patches) in a habitat, based on the mathematical analysis of Eqs. (1) and (2). Defining patch quality as resource content can mean two slightly different things. One may consider variation in the resource density ($n0$) of patches, or in their size (S). Both actions vary the total resource content, but since the rate of consumption is usually governed by resource density (rather than sheer quantity), they do not have the same implications. Predictions are thus given regarding the two possibilities. Of course, some organisms might respond to total resource amount irrespective of density and we would just treat n as the resource amount and forget about S. Also, some organisms respond to the fraction of resource items not-yet-attacked, for instance parasitoids that keep probing hosts already attacked, in which case one can take S to be the total amount of resource and n the fraction yet-to-be attacked. This framework thus encompasses all possible shapes of functional responses, and different types of consumption processes that have been studied separately.

It can be seen that predictions are clarified in several ways. Importantly, while some predictions remain uncertain if patch quality is intended as resource density (shaded cells), knowing, even qualitatively, the type of functional response suffices to lift all uncertainties.

[3]One such case is when patches differ in the energetic cost of reaching/entering them; Calcagno *et al.* (2014a).

Table 2 MVT predictions for resource consumers (Eq. (2)) regarding the effect of the distribution (mean and variance) of patch quality (resource density or size) of a habitat

		E^*	t^*	GUD	ρ_{INTRA}
Resource density (n)	Mean ↑	↑	↑ (II, IV) 0 (I) ↓ (III)	↑	+
	Variance ↑	↑	↓ (I, II, III) ↑ (IV)	↓ (I, II, III) ↑ (IV)	
Size (S)	Mean ↑	↑	↑	↑	+
	Variance ↑	0	0	0	

A major clarification is that there will invariably be a positive correlation between patch quality and residence time within habitats ($\rho_{\text{INTRA}} > 0$; Table 2). However this positive correlation within habitats does not imply that increasing patch quality in a habitat necessarily increases the average residence time. Indeed, an increase in the resource density of patches can have any effect on the residence time, and in particular it can decrease it. This illustrates the general expectation from the classic MVT on the links between within-habitat and cross-habitat correlations of residence time with patch quality (previous section). It turns out that the outcome of increasing the resource density of patches is entirely determined by the concavity of the functional response: with concave (type II or IV) functional responses, we would observe an increase of residence time, whereas with convex (type III) we would observe a decrease. Linear (type I) functional responses are the knife-edge case when patch quality has no impact on residence time (Table 2).

Variance in the resource density of patches can also have contrasted impacts of residence time and on the GUD (Table 2) but again, knowing the type of functional response removes uncertainty. In this case what matters is not the concavity of the response but its slope: increasing functional responses (type I, II or III) cause an increase of average residence time and average GUD with patch heterogeneity, whereas decreasing (type IV) responses cause a decrease of both. As the latter is probably not the most likely situation, we can safely state that greater variance in resource density would generally increase the average rate of movement in a habitat (Calcagno *et al.*, 2014b).

When patch quality is intended as patch size, predictions are much more straightforward: increasing patch quality invariably increases residence time and the GUD (Table 2). However, variance in patch size, unlike patch resource density, has no effect whatsoever on any quantity of interest: these are invariant to size heterogeneity. Patch size (at constant resource density) constitutes one of those special cases for which heterogeneity does not yield any fitness benefit for an optimal forager (see Table 1) and knowledge of the average patch size therefore suffices to determine the optimal strategy.

Overall, there is a fairly complete set of testable predictions regarding the consequences of altering the distribution of resources in a habitat, providing one knows the type of functional response. One often overlooked result is that individuals might decrease their residence time if patches are made better, even though they stay longer on the best patches within any habitat (see Fig. 3 for an illustrative example). This is predicted in the case of an accelerating functional response (Table 2), but not only (see next section about foraging costs).

It is also important to remark that in several cases, the mean and variance of the resource distribution have distinct effects. For instance, with type II functional responses, greater mean resource density increases the average residence time, while greater variance reduces it (Table 2). When the two effects operate simultaneously following habitat changes, their net outcome will therefore critically depend on the mean-variance scaling of the resource distribution. For instance, in Fig. 3, the green habitat is enriched (higher mean resource density) but as the variance stays constant, the variability (coefficient of variation) of the distribution is effectively reduced. From Table 2, decreasing the variability of resource density for a type-III functional response increases the average residence time. Therefore, in Fig. 3, a positive variance effect counteracted (but did not outweigh) the negative effect of average resource density. We can thus predict that the observed decrease in residence time would be even stronger if the habitats were strictly homogeneous (i.e., if all patches had exactly resource density 15 in the blue habitat, and resource density 18 in the green). In contrast, when shifting from the orange to the green habitat, the negative impact of increasing mean resource density was dominated by the positive variance effect (as the green habitat has much lower patch heterogeneity), and the net outcome was an increase in average residence time (Fig. 3). The bottom line is that predicting the consequences of patch enrichment in heterogeneous habitats requires paying attention to how enrichment is distributed over the different types of patches (Calcagno *et al.*, 2014b).

Going Further and Open Questions

For simplicity, foraging costs were not explicitly considered here, following a tradition of treating resource intake as gains. However, strictly speaking the MVT applies for net energy gains, discounting costs. We can easily assume the existence of some per-unit-of-time cost of moving between patches (μ) and of foraging in patches (ν). Obviously both cause a decrease in E^*, and the greater μ relative to ν, the longer residence times will be (as an individual will allocate more time in the compartments with lower cost). In fact, it is chiefly

the difference between μ and v that matters (Calcagno *et al.*, 2014b). All predictions reported above hold in the presence of costs, as long as μ is close to v. If this is not the case, there is only one change in **Table 2**: predictions regarding t^* and the GUD following an increase in mean patch quality must be shifted in some direction. When costs are greater between than within patches ($\mu > v$), predictions must be shifted in the direction of a decrease, whereas if costs are greater within patches, they must be shifted toward an increase. For instance, for a linear functional response we predict no effect of enrichment on the optimal residence time (0; **Table 2**). If $\mu > v$ we would predict a decrease ($\downarrow$), and if $v > \mu$, an increase ($\uparrow$). Quite unexpectedly, higher costs of traveling between patches thus expand the conditions under which decreased residence times are expected following patch enrichment.

Of course, foraging costs could be variable among patches, in which case further complications arise. For instance, the quitting rate would still be the same in all patches, but the GUD would not, as it would no longer be a reliable indicator of the rate of gain. For resource consumers, one can in principle still use Eq. **(2)** and the same mathematical techniques to obtain predictions regarding the consequences of patch variability in foraging costs, or in other aspects of the foraging process. One can for instance study the impact of resource quality (parameter γ) or of some parameters of the gain function,[4] such as attack rates and handling times. A comprehensive treatment of these aspects remains to be done.

More importantly, the MVT ignores the notion of risk. A foraging individual often faces the risk of being interrupted or even killed by a predator, and these risks typically vary over different parts of the habitat. This can strongly affect the foraging decisions and the optimal allocation of time. In addition, the total duration of the foraging session (D) would then depend on the residence times, violating one MVT assumption, and rendering the average rate of gain (E) an inadequate predictor of fitness. Marginal gains and optimization can still be used in the presence of such risks, and formalisms have been proposed in the optimal foraging theory to deal specifically with these aspects (Brown, 1988). However they remain quite distinct from the MVT and a better integration of the two frameworks would be valuable.

See also: Behavioral Ecology: Optimal Foraging Theory; Herbivore-Predator Cycles. Conservation Ecology: Trophic Index and Efficiency. Ecological Data Analysis and Modelling: Climate Change Models

References

Brown, J.S., 1988. Patch use as an indicator of habitat preference, predation risk, and competition. Behavioral Ecology and Sociobiology 22 (1), 37–47.
Calcagno, V., Mailleret, L., Wajnberg, É., Grognard, F., 2014a. How optimal foragers should respond to habitat changes: A reanalysis of the marginal value theorem. Journal of Mathematical Biology 69 (5), 1237–1265.
Calcagno, V., Grognard, F., Hamelin, F.M., Wajnberg, É., Mailleret, L., 2014b. The functional response predicts the effect of resource distribution on the optimal movement rate of consumers. Ecology Letters 17 (12), 1570–1579.
Charnov, E.L., 1976. Optimal foraging, the marginal value theorem. Theoretical Population Biology 9 (2), 129–136.
Staddon, J.E., 2001. Adaptive dynamics: The theoretical analysis of behavior. Cambridge, MA: MIT Press.

Further Reading

Stephens, D.W., Krebs, J.R., 1986. Foraging theory. Princeton, NJ: Princeton University Press.

[4]This is quite easy: varying γ is the same as varying resource density with a type-I functional response.

Mating Systems☆

Stephen M Shuster, Northern Arizona University, Flagstaff, AZ, United States

Introduction

Male–Female Differences

Males and females by definition are different. Male gametes are usually motile; female gametes are usually not. Yet most humans have the sense that males and females are distinct in ways other than mere gamete dimorphism. Sexual differences in human phenotypes, while moderate compared to some species, are undeniable, and consequently much has been made of battles between the sexes, of sexual dialectics, even of the possibility that men and women have different planets of origin. We are hardly unique in this regard. Any observant naturalist can list several species in addition to our own in which male-female differences are clear, if not extreme. Some of us can produce an even longer list of species in which external sex differences are inscrutable. Songbirds, for example, often lack external sexual differentiation. A large number of marine species, such as fucoid brown algae, sea urchins, polychaete worms and red snappers, all have separate sexes that are indistinct. Sexual differences in land plants too are often obscure. Botanists may assert that this is because both sexes coexist within each individual; but why should monoecy cause monomorphism? And if it does, why should sex in both cottonwoods and willows be apparent only upon close examination, whereas sex among marijuana plants, even for aficionados, is simple to diagnose? Invertebrate zoologists might now chime in with examples of physical uniformity among hermaphroditic barnacles, flatworms and freshwater snails, as well as among gonochoristic comb jellies, kinorhynchs and veneroid clams. What explanation can possibly exist for extreme sexual differentiation in some species, and for virtual monomorphism among others, especially those lacking gender? The answer to this question is mating system. That is, the circumstances in which reproduction occurs within individual species. It is here that sexual differences arise—or do not.

Historical Descriptions

Mating Systems in Plants, Fungi, and Protists

In current literature, mating systems are described in two distinct ways, and each description has different implications for how sexual differences may or may not appear. The first description of mating systems, one that is familiar to botanists and coevolutionary biologists, emphasizes the genetic relationships that exist between mating males and females. Selfing, partial selfing, random mating, positive assortative mating (inbreeding), and negative assortative mating (outbreeding), all are examples of mating systems described in terms of the genetic relationships that may arise among breeding pairs. This framework also serves to describe fungal and many protist mating systems in which, depending on the species, individuals may engage in selfing through homothallic fusion—the union of hyphae or gametes of like mating type; or outcrossing through heterothallic fusion—the union of hyphae or gametes of unlike mating type.

The diversity of floral traits in plants can be categorized, as Darwin noted, by the degree to which self-fertilization occurs or is prevented to various degrees by physical, temporal, or genetic separation of male and female elements (**Table 1**). Among species with "perfect" flowers, cleistogamous plants are entirely self-pollinating, whereas chasmogamous plants use pollen to fertilize ovules on different flowers, either within the same individual, as in geitonogamy, or on different plants, as facilitated by a wide range of mechanisms. For example, heterostyled or herkogamous plants show physical separation of the style and stamen, whereas dichogamous plants separate the development of male and female elements in time. Outcrossing may be further encouraged by genetic self-incompatibility systems that occur in either the sporophytic or gametophytic stages of the life cycle, or it is prevented entirely by separation of the sexes on different individuals with "imperfect" flowers, as in dioecy, or as in androdioecy, and gynodioecy, in which males or females may coexist with monoecious individuals, respectively. Because certain heritable traits tend to covary between the sexes within each breeding scheme, genetic correlations may arise that lead to, or prevent, the appearance of sex-specific phenotypes.

Mating Systems in Animals

A second description of mating systems, familiar to zoologists and behavioral ecologists, considers mating systems in terms of the number of mates acquired by males or by females. This scheme has historical precedence, but it can lead to inaccurate

☆*Change History*: March 2018. S M Shuster updated sections Introduction, Measuring the Sex Difference in Fitness Variance, The Mean and Variance in Mate Numbers; The Opportunity for Sexual Selection, and Genetic Correlations between the Sexes, Concluding Remarks, and Further Reading.

This is an update of S.M. Shuster, Mating Systems, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 2266–2273.

 doi:10.1016/B978-0-12-409548-9.11157-1

Table 1 A classification of plant mating systems based on fertilization mode and flower morphology

General category	*Fertilization mode*	*Definition*	*Associated terminology*
Perfect flowers (monoecious)	Self-pollination	Pollen fertilize ovules on the same flower	Autogamy, cleistogamy
	Cross pollination	Pollen fertilize ovules on different flowers	Chasmogamy
	Heterostyly	Physical separation of style and stamen	Distyly, tristyly
	Dichogamy	Temporal separation of anther and stigma maturation	Protandry, protogyny
	Self-incompatibility	Pollen cannot fertilize ovules on the same plant	Sporophytic, gametophytic self-incompatibility
Imperfect flowers	Outcrossing	Flowers are sex specific or exist as combinations of perfect and imperfect flowers	Dioecy; androdioecy, gynodioecy

characterizations, such as the misconception that on average, males can be more promiscuous than females. In fact, when adjusted for the sex ratio, the average number of mates, and of matings, if individuals mate more than once, for males and for females, must be equivalent. A terminology that accommodates this fact describes mating systems in terms of the *variance* in mate numbers among members of each sex. According to this scheme, monogyny and polygyny are descriptions of the variance in mate numbers among males, which for these terms equals 0, or exceeds 0, respectively. In contrast, monandry and polyandry are descriptions of the variance in mate numbers among females, which also for these terms, equals 0, or exceeds 0, respectively (**Table 2**). Additional descriptors address the variance in male and female mate numbers relative to that of the other sex (**Table 2**). For example, monogamy describes the condition in which no variance in mate numbers exists for either sex, whereas polygamy describes the condition in which non-zero, but approximately equivalent variance in mate numbers exists for both sexes. Polygynandry describes the condition in which non-zero variance in mate numbers exists for both sexes, but is greater among males than among females, whereas polyandrogyny describes the condition in which non-zero variance in mate numbers exists for both sexes, but is greater among females than among males (**Table 2**).

Each of these groupings describe convergent states that may arise from different ancestral conditions because a wide range of circumstances appear to favor these basic categories. Moreover, the spatial and temporal distributions of matings, the degree to which females tend to mate once or multiple times, whether females tend to aggregate or remain solitary, and whether male and female associations are brief or prolonged, all contribute to the "special circumstances," as Darwin called them, that make competition for mates more or less severe. When these conditions are considered, additional descriptors may modify the basic categories of animal mating systems.

For example, "mass mating" or "persistent pairs" may describe mating system variants in which breeding pairs form synchronously within a population, but are either aggregated or dispersed in space. In contrast, "attendance polygyny" or "feeding site polygyny" may describe variants of polygynous mating systems in which the variance in male mating success depends on whether males guard and mate with individual females in sequence, or guard the food resources that females require to reproduce. Again, most descriptions of animal mating systems focus on males, because males tend to be more active in courtship or mate monopolization than females. However, females in some species seek mates more actively than males, thus "cursorial polyandrogyny," or "dominance polyandrogyny," may describe variants of polyandrous mating systems in which females move among their several individual male consorts, or defend territories containing small aggregations of nesting males. In general, the basic categories of animal mating systems (**Table 2**) are fundamental to sex differences in fitness variance, that is, to the strength of sexual selection - or to its absence as an evolutionary force.

A Case for Combining Terminologies

Although usually considered separately, novel insights are possible when the schemes for describing plant and animal mating systems are combined. In particular, genetic covariances that arise between male and female mating phenotypes, because of the circumstances that cause mate number to vary among individuals, can be incorporated into discussions of mating system evolution. Such considerations provide simple explanations for observed interspecific variation in reproductive traits, including differences within and among species in promiscuity or mate guarding, as well as tendencies within and among species to release gametes synchronously or over prolonged durations. A combined description of mating systems identifies the nature, as well as the rates, of possible "run-away" processes. These evolutionary feedback loops arise when particular mating associations or preferences, cause genetic variation underlying male and female traits, as well as selection acting on those traits, to covary.

In this light, apparent examples of sexual conflict between males and females can be explained simply as the result of a genetic correlation between coercive mating behavior in one sex and resistance to such coercion in the other. Similarly, female tendencies to seek multiple mates may have less to do with females attempting to enhance the genetic quality of their offspring, and more to do with a genetic correlation that inevitably arises between the sexes when both sexes tend to mate more than once. Such trait associations may even occur between species, leading to interspecific genetic correlations, mostly mutualistic but occasionally not, involving plants and their pollinators. Such considerations explain more of the observed variation in sexual dimorphism, floral

Table 2 A classification of mating systems based on male and female mate numbers

Mating system	*Definition*	*Variance in mate number*	
		Females	*Males*
Monogamy	All individuals in each sex have a single reproductive partner (a mate).	0	0
Polygyny	All females have a single mate; male mate numbers vary.	0	+ + + +
Polygynandry	Both sexes have variable mate numbers; mate numbers vary more among males than among females.	+	+ + +
Polygamy	Both sexes have variable mate numbers; mate numbers are equally variable within each sex.	+ +	+ +
Polyandrogyny	Both sexes have variable mate numbers; mate numbers vary more among females than among males.	+ + +	+
Polyandry	All males have a single mate; female mate numbers vary.	+ + + +	0

morphology and patterns of gamete release and fusion, than classification schemes based on sex differences in mate numbers, or reductions in heterozygosity, can provide.

Recent Emphases in Mating System Research

Plant Mating Systems

Evolutionary research on plant mating systems began with Darwin, who combined the classification of species by their floral morphology, with documentation of whether these traits allowed or prevented self-fertilization (**Table 1**). Since Darwin, the population genetic consequences of these phenotypes, as well as the genetic and evolutionary mechanisms by which selfing, self-incompatibility, dioecy and gynodioecy may arise and persist, have all been given detailed theoretical expression. Frequency-dependent selection appears to maintain selfing and outcrossing as alternative stable states in some species, whereas mixtures of these traits in other species seem to be constrained by the deleterious effects of inbreeding. The stability of selfing and outcrossing within a given species also appears to be influenced by frequency- and density-dependent processes associated with pollen transmission. Here, patterns of floral morphology are expected to correspond to the spatial density of conspecifics, as well as to the relative amount of pollen available for export to other flowers.

Wind pollinated plants appear to fit these theoretical predictions more clearly than animal pollinated species, but this relationship may simply reveal that "animal pollination" is too broad a category to capture the distinct, or at times overlapping activities, of avian, mammalian and arthropod pollinators. Moreover, plant genetic diversity is affected by the relative distances pollinators can travel, further complicating classification. Most analyses of plant mating systems make use of the increasing sophistication of molecular methods available to document how population genetic structure is influenced by observed patterns of selfing and outcrossing. These tools have also proven useful for exploring how the evolution of sexual phenotypes is influenced by stochastic processes and by environmental influences on gender expression. Molecular phylogenetic analyses reveal with increasing precision that mating systems are among the most evolutionary labile of all plant traits.

Animal Mating Systems

Evolutionary research on animal mating systems also began with Darwin, who again combined classification schemes, this time according to differences in mate numbers for members of each sex, as well as according to the social contexts in which mate numbers become variable. Much of animal mating system research still focuses on whether sexual selection occurs within the contexts of male–male competition or female mate choice. As molecular methods have improved, animal researchers too have verified with increasing precision the number of mates males and females may obtain, allowing (but not always leading to) increasingly explicit classification of mating systems according the variance in mate numbers for males and females. As with studies of plants, molecular phylogenies have revealed considerable diversity in mating system expression within monophyletic animal taxa, suggesting that regardless of the kingdom considered, mating systems are responsible for considerable evolutionary innovation.

Analysis of animal mating systems for most of the last century have emphasized sex differences in parental investment as the source of sexual selection. According to this view, female reproduction is limited by the availability of resources required for energetic investment in ova and young. Given that resource abundance may vary in space and in time, male reproduction is presumed to be limited, in turn, by the spatial distribution of resources required by females, and by the temporal distribution of sexually receptive females themselves. This insight, combined with observed sex differences in gametic investment, suggests that the intensity of male–male competition reveals the intensity of sexual selection.

Two measures have served as the primary metrics for determining how female spatiotemporal distributions may influence sexual selection intensity. The first measure, the operational sex ratio (OSR), was conceived as the ratio of potentially mature males to potentially receptive females, although researchers often focus on those individuals who are receptive at a particular time and place. When OSR $>$ 1, females are rare, and male competition for mates and sexual selection can appear to be intense, although this depends on the consistency of mating success among males throughout the breeding season. When OSR $<$ 1, females are

abundant, and male competition for mates and sexual selection can appear to have relaxed; but again, depending on the cause of the surplus in females and how males respond to it, such conditions may still allow disproportionate success to certain males.

The second measure, the environmental potential for polygamy (EPP), identifies the degree to which social and ecological conditions allow males to monopolize females. Although useful for visualizing when male-male competition might arise, standardized methods for quantifying female distributions, or the scale on which EPP could be measured have remained undefined. As a result, while serving as conceptual surrogates for the intensity of sexual selection, the uncertain relationships between OSR, EPP and selection intensity itself, within, as well as among species, makes these measures unreliable for estimating sexual selection.

Researchers emphasizing sexual differences in parental investment as the source of mating system variation have encountered further difficulties in putting its assumptions to empirical tests. In particular, a sex difference in relative parental investment has proven extremely difficult to compare within and among species. Not only are the amounts of energy, cost and risk associated with relative parental investment difficult to quantify, but the correlation between sex differences in parental investment and sexual dimorphism itself is often poor, particularly in species with reversed sex roles. Measures of sexual selection intensity based on the predictions of parental investment theory, including comparisons of maximum potential reproductive rates among males and females, require laboratory conditions that are rarely encountered in nature. Other estimates, which emphasize how mating behavior may enhance or diminish the future fitness of individuals, require assumptions about who is breeding and who is not, that may underestimate the actual variance in mating success within the population; or they may be impossible to directly test at all.

Combining Methods for Mating System Analysis

Covariance Methods

The most obvious way to understand how selection acts within a mating system is to measure selection itself. Sexual selection gradients and parental tables isolate the statistical relationship between male and female trait values and fertilization success relative to other components of selection. Bateman gradients, named for this author's classic work with fruit flies, measure how the mating success for each sex correlates with offspring numbers. These gradients measure the selection that acts on every sexually-selected trait, thereby revealing the final common path between sexually-selected traits and fitness. Only when the gradient is zero, does it indicate that sexual selection is negligible and that selection in other contexts may prevail. Sex-specific gradients also provide a quantitative measure of "sexual conflict," measured as the sex difference in sign of this covariance. When the sign of this covariance is negative, additional matings will enhance the fitness of members of one sex, but diminish the fitness of members of the other sex. This approach provides a simple statistical explanation for apparent eagerness to mate in one sex and apparent choosiness before mating in the other, that is unrelated to sex differences in parental investment.

When a genetic basis for particular traits is associated with fitness, the covariance approach also provides a means for predicting evolutionary trajectories when genetic correlations become established by the cooccurrence of particular male–female relationships, such as when stigmas that are available for pollination appear synchronously with stamen dehiscence in wind-pollinated plants. However, direct estimates of selection also have their weaknesses. This approach assumes that the researcher has correctly identified and measured the trait that is under selection. Estimates of selection intensity can be erroneous if the actual target of selection is unmeasured, or if it is correlated in some unknown way with other traits that directly or indirectly influence fitness.

Parentage Data and the Zero Class

Both plant and animal mating systems increasingly use molecular parentage data. These assignments provide the easiest avenue for a combined approach to mating system analysis. Identification of genetic parentage reveals the degree to which progeny have arisen through inbreeding or outcrossing, and they establish whether males and females mate within or outside of their social pair bonds. To understand whether and to what degree the sexes may become distinct within a species, and to understand if a sex difference in selection intensity could be responsible for phenotypic divergence among related species, it is necessary to measure the variance in relative fitness for males and females within, as well as among taxa. This method illustrates when and why sexual selection can be strong enough to overwhelm the effects of natural selection and therefore, how it can produce the sex-specific phenotypes researchers find so compelling (Box 1). However, parentage data from plants and animals suffer from a common problem; they are both obtained from the fraction of their respective populations that are *successful* in producing progeny. The fraction of the population in plants and in animals that is *unsuccessful* in producing offspring typically goes unidentified in most analyses that assign parentage to experimental populations.

Why do such omissions matter? A focus on successfully reproducing individuals fails to capture the largest fraction of the variance in fitness that exists for a total population. To measure the total variance in fitness for a population, one must measure the average variance in fitness within the successfully reproducing individuals, as well as the variance in the average fitness between the successful reproducers and those individuals who fail to reproduce at all (Box 1). The latter component of the total variance in fitness is usually the larger of these two components. Thus, estimates of fitness that exclude nonbreeding individuals tend to *overestimate* the average fitness and *underestimate* the variance in fitness of the portion of the population that actually is measured. Moreover, the larger the fraction of the population that fails to reproduce and therefore goes unmeasured, the more severe this

Box 1 Measuring the sex difference in the variance in relative fitness

The Mean and Variance in Offspring Numbers

Consider a hypothetical population, say of albatrosses or penguins, consisting of 20 individuals and a sex ratio equal to 1. If, in a given breeding season, a single ovum from each of the 10 females in the population is fertilized by a different male, the total number of offspring, $N_{\mathbf{O}total}$, equals (1 ovum) $\times$ (10 females) $=$ 10. Because each mating pair produces 1 offspring, the total offspring produced by all females, $N_{\mathbf{O}♀}$, equals the total offspring produced by all males, $N_{\mathbf{O}♂} = 10$. Because there are 10 females and 10 males in our population, the average offspring per female, $\mathbf{O}_♀$, equals $N_{\mathbf{O}total}/N_♀ = 1$, which equals the average offspring per male, $\mathbf{O}_♂$, evaluated similarly as $N_{\mathbf{O}total}/N_♂ = 1$. Also, because each individual in the population produces the same number of offspring ($=1$), no variance in offspring numbers can exist for either sex. Thus, if $V_{\mathbf{O}♀}$ and $V_{\mathbf{O}♂}$ equal the variance in offspring numbers for females and for males, respectively, $V_{\mathbf{O}♀} = V_{\mathbf{O}♂} = 0$. This example shows that when the number of fertilized ova each female produces is 1, or even when it is larger, when each female is mated once by a different male, and the sex ratio equals 1, the mean and the variance in fitness for each sex are identical.

This example provides a simple explanation for why wind pollinated plants, marine species with external fertilization, and even hermaphroditic organisms, all tend to show little sexual dimorphism. In each of these cases, population sex ratios equal or are nearly equal to 1, pollen or sperm are either so widespread or so restricted in their distribution that individual males and individual females each have similar probabilities of reproduction, and members of each sex tend to contribute approximately equally to the next generation. In the case of hermaphrodites, both sexes are represented within every individual, thus reproduction by one sex usually allows reproduction by the other; and unless some hermaphrodites emphasize, or are forced to reproduce as only one sex (as occurs in certain flatworms and barnacles), the population-wide variance in fitness through male and female functions, are approximately equivalent. When such conditions apply, neither sex is likely to become distinct from the other, except as is specifically required for the production or ova or sperm.

Now consider a case in which 1 of the 10 males secures 2 mates instead of just 1. The total offspring produced by our population, $N_{\mathbf{O}total} = 10$, remains unchanged. Similarly, because $N_♂ = N_♀ = 10$, the average offspring per male, $\mathbf{O}_♂ = N_{\mathbf{O}total}/N_♂ = 1$, equals the average offspring per female, $\mathbf{O}_♀ = N_{\mathbf{O}total}/N_♀ = 1$. Because each female still secures 1 mate with whom she produces a single brood, the variance in offspring numbers for females, $V_{\mathbf{O}♀} = 0$. However, because 1 male has 2 mates, 1 male must be excluded from mating. When this happens, the variance in offspring numbers for males, $V_{\mathbf{O}♂}$, must increase.

If parentage data were available for our population, we could estimate the magnitude of the increase in $V_{\mathbf{O}♂}$ simply by calculating the statistical variance in offspring numbers for males. This quantity equals the average of the squared number of offspring produced by each male, minus the square of the average number of offspring produced by each male, or $(\Sigma o_{i♂}^2)/N_♂ - \mathbf{O}_♂^2$, where $o_{i♂}$ = the number of offspring produced by the i-th male, $N_♂$ = the total number of males, and $\mathbf{O}_♂$ = the average number of offspring per male as defined above. Note that the variance in the number of offspring per female, $V_{\mathbf{O}♀}$, can be calculated in the same way by substituting $o_{i♀}$, $N_♀$ and $\mathbf{O}_♀$ for females into the above expression, where $o_{i♀}$ = the number of offspring produced by the i-th female, $N_♀$ = the total number of females, and $\mathbf{O}_♀$ = the average number of offspring per female as defined above. When such data are available, this is indeed the simplest approach. However, as is more often the case, when parentage data are lacking, an equally accurate and in some ways more informative approach, involves partitioning the variance in offspring numbers, within and among the classes of mating and non-mating individuals. In the example above, only males were variable in their numbers of mates, but in many species, both sexes may vary in mate numbers, and in sex role reversed species, including certain sea spiders, giant water bugs and pipefish, females are consistently more variable in mate numbers than males. Although they are seldom used for this purpose, the data necessary to calculate the mean and variance in mate numbers for males, and the mean and variance in offspring numbers for females, are often available in standard life history analyses. This quantitative approach allows us to measure the fitness variance within each sex, which is proportional to the intensity of selection. The sex-difference in selection intensity, in turn, estimates the degree to which the sexes will diverge in phenotype.

The Mean and Variance in Mate Numbers

We begin by identifying the classes of mating males and their population frequencies. Here, we represent the proportion of the male population in each mating class as p_j, where j represents the number of females in the j-th mating class of males. There are three such classes: males who do not mate, p_0 [$=1/10$ males $= 0.1$], males who mate once, p_1 [$=8/10$ males $= 0.8$] and males who mate twice, p_2 [$= 1/10$ males $= 0.1$]. The sum of all male mating classes, $\Sigma p_j = (0.1 + 0.8 + 0.1) = 1$. Next, we use these values to identify the average offspring produced by males in each j-th mating class, $O_{♂j}$, as well as the average offspring produced by all males across all mating classes, $\mathbf{O}_♂$ [$= \Sigma p_j O_{♂j}$ where $O_{♂j} = (\Sigma o_{i♂})/N_♂$]. The average offspring that males in each mating class produce, $O_{♂j}$, equals the average offspring per female, $\mathbf{O}_♀$, as defined above, multiplied by the number of mates, j, that males in each j-th mating class obtains, or $O_{♂j} = j(\mathbf{O}_♀)$.

Clearly, the average number of offspring produced by males who do not mate, $O_{♂0}$, equals $(0)(1) = 0$. The average number of offspring for males who mate once, $O_{♂1}$, equals $(1)(1) = 1$, and for males who mate twice, $O_{♂2}$ equals $(2)(1) = 2$. The average number of offspring produced by all males, across all mating classes, $\mathbf{O}_♂$, is equal to the number of offspring produced by the average female, $\mathbf{O}_♀$, multiplied by the number of females mated by males in each j-th mating class [$= j(\mathbf{O}_♀)$], then, each quantity is multiplied by the fraction of the males belonging to that j-th mating class, p_j, and summed over all j mating classes, so that,

$$\mathbf{O}_♂ = \Sigma\, p_j\, j\, (\mathbf{O}_♀) \tag{2}$$

Using the values in our example above, we can see that $\Sigma p_j\ j = 1$ and that $\mathbf{O}_♂ = \mathbf{O}_♀ = 1$. Thus, although females are distributed unevenly among the 10 males, relative to females, as well as to the initial case in which all 10 males have equal mate numbers, the average number of offspring produced by all males in this example, again equals 1.

The distribution of females across all classes of mating males is equal to the population sex ratio, which we can call R. This value can be calculated as the number of females mated by males in each mating class, j, multiplied by the fraction of the males in each mating class, p_j, and summed over all classes of males, or, $R = \Sigma j p_j = 1$. Because the distribution of all females with all males equals the average number mates per male, R also equals $N_♀/N_♂ = 1$. That is, R is the reciprocal of OSR ($= N_♂/N_♀$); but whereas OSR measures the apparent intensity of male-male competition, R measures a slightly more useful quantity for estimating how selection works; R measures the population-wide average in the number of mates per male (in sex role reversed species, $1/R = R_O$ measures the analogous quantity). By substitution, we can see that the average offspring per male, $\mathbf{O}_♂$, equals the average mates per male, R, multiplied by the average offspring per female, $\mathbf{O}_♀$, or $\mathbf{O}_♂ = R\mathbf{O}_♀ = 1$. Furthermore, although the distribution of females is now uneven among males, the average mates per male, R, the average offspring per female, $\mathbf{O}_♀$, and the average offspring per male, $\mathbf{O}_♂$, all remain unchanged relative to our initial example.

We can now express the total variance in offspring numbers for males, $V_{\mathbf{O}♂}$, in terms of the average number of mates per male and the average number of offspring per female. As in an analysis of variance (ANOVA) problem, the total variance in male fitness can be partitioned into two components: (1) the average variance in offspring numbers for males within the classes of males who sire offspring, and, (2) the variance in the average number of offspring sired by all the classes of males.

The first component of variance in male offspring numbers is calculated in three steps. First, for each mating class of males, the variance in female offspring numbers, $V_{\mathbf{O}♀}$, is multiplied by the number of mates obtained by males in each j-th mating class ($= jV_{\mathbf{O}♀}$). Next, this product is multiplied by the proportion of males in the population, p_j, that belongs to each j-th mating class [$= p_j(jV_{\mathbf{O}♀})$]. Lastly, these products are summed over all j mating classes. Thus, the variance in offspring numbers within the classes of mating males equals,

$$V_{\mathbf{O}♂(\text{within})} = \Sigma p_j\left(jV_{\mathbf{O}♀}\right) \tag{3}$$

In this example, because all females produce exactly 1 offspring, there is no variance in offspring numbers for females, $V_{\mathbf{O}♀} = 0$, and consequently, the variance in offspring numbers within the classes of mating males is also zero ($V_{\mathbf{O}♂(\text{within})} = 0$). We will return to this point below.

The second component of variance in male offspring numbers equals the variance in the average number of offspring sired by males among these same categories. This quantity is calculated in four steps. First, for each j-th mating class of males, we calculate the difference between the average number of offspring per male, $\mathbf{O}_♂$, and the average number of offspring produced by that mating class, $O_{♂j}$ ($= [\mathbf{O}_♂ - O_{♂j}]$). Secondly, we square each difference [$= (\mathbf{O}_♂ - O_{♂j})^2$]. Third, we multiply each squared difference by the fraction of males belonging to each mating class, p_j [$= p_j(\mathbf{O}_♂ - O_{♂j})^2$], and fourth, we sum across all classes to obtain,

$$V_{\mathbf{O}♂(\text{among})} = \Sigma p_j\left(\mathbf{O}_♂ - O_{♂j}\right)^2 \tag{4}$$

Substituting in the values from above we have, $V_{\mathbf{O}♂(\text{among})} = 0.2$.

The total variance in offspring numbers among males is the sum of the within and among male components in offspring numbers, or,

$$V_{\mathbf{O}♂} = \Sigma p_j\left(jV_{\mathbf{O}♀}\right) + \Sigma p_j\left(\mathbf{O}_♂ - O_{♂j}\right)^2 \tag{5}$$

Because there is no variance in offspring numbers for females, $V_{\mathbf{O}♀} = 0$, the first term in Eq. (5) drops out. However, if $V_{\mathbf{O}♀}$ were nonzero, $V_{\mathbf{O}♂(\text{within})}$ would equal this quantity because the variance if fitness within the class of breeding males will always equals the variance in fitness within the class of breeding females. Thus, in this example, $V_{\mathbf{O}♂(\text{among})} = V_{\mathbf{O}♂}$, and we can easily see that the variance in fitness among males goes from 0 to 0.2 when a single male mates with 2 females instead of 1. Note too that the increase in fitness variance comes entirely from the among-male component of total fitness variance. Now, if 1 male mates with all 10 of the females, the mean and variance in offspring numbers for females remains unchanged ($\mathbf{O}_♀ = 1$; $V_{\mathbf{O}♀} = 0$), and there is no change in either the sex ratio, $R = 1$, or the average number of offspring per male, $\mathbf{O}_♂ = 1$. But, because 1 male mates 10 times, 9 males do not mate at all. Thus, $p_{♂0} = 9/10 = 0.9$, $p_{♂1}$ through $p_{♂9} = 0$ and $p_{♂10} = 1/10 = 0.1$. When these values are substituted into Eq. (5), we see that $V_{\mathbf{O}♂}$ now increases to 9, a 45-fold increase in the variance in male fitness!

This exercise shows three relationships. First, when the sex ratio equals 1, both sexes must have equal, average fitnesses. Secondly, when some individuals are excluded from mating, the variance in offspring numbers within that sex will increase. This is the source of sexual selection. Exclusion of some individuals from mating means that only the traits of the individuals who do mate will be represented in the next generation. Third, if the fraction of individuals excluded from mating is larger in one sex than it is in the other, a sex difference in the variance in offspring numbers appears. Because fitness variance is proportional to selection intensity, the magnitude of this sex difference in fitness variance determines the actual intensity of sexual selection. The larger the sex difference in fitness is, the more the sexes will diverge in phenotype.

The Opportunity for Sexual Selection

The above examples consider the absolute mean and variance in fitness for males and females, but selection is a relative process, and to account for this fact, certain adjustments are necessary. When the variance in absolute fitness, $V_{\mathbf{W}}$, is divided by the squared average fitness, $\mathbf{W}^2$, we obtain $V_{\mathbf{W}}/\mathbf{W}^2$, a quantity known as the variance in relative fitness, $V_{\boldsymbol{w}}$, or as $\boldsymbol{I}$, the opportunity for selection. The opportunity for selection provides a dimensionless, empirical estimate of selection's maximum strength that is comparable within and among species. When paternity can be assured, $\boldsymbol{I}$ can be calculated from estimates of mating success alone.

This approach is especially useful for understanding the strength of selection within each sex. Here, the value of $\boldsymbol{I}$ is expressed as the ratio of the variance in offspring numbers, $V_{\mathbf{O}}$, to the squared average in offspring numbers, $\mathbf{O}^2$, among the members of each

sex. Thus, $\boldsymbol{I}_{\sigma} = V_{\mathbf{O}\sigma}/\mathbf{O}_{\sigma}^2$ and $\boldsymbol{I}_{\varphi} = V_{\mathbf{O}\varphi}/\mathbf{O}_{\varphi}^2$. Because each offspring has a mother and a father, the opportunity for selection on males, $\boldsymbol{I}_{\sigma}$, and the opportunity for selection on females, $\boldsymbol{I}_{\varphi}$, are linked through the sex ratio and mean fitness, which must be equal for both sexes. However, when the sex ratio equals 1, the sex difference in the variance in relative fitness, $\Delta\boldsymbol{I} = (\boldsymbol{I}_{\sigma} - \boldsymbol{I}_{\varphi})$. This value may be greater for males, greater for females or zero. Its value determines whether and to what degree the sexes will diverge in character because fitness variance is proportional to selection intensity.

How can we express these relationships for a natural population? Rewriting Eq. (5), which expresses the total variance if male fitness, partitioned into within- and among-male components, we substitute values from Eqs. (3) and (4) and rearrange terms. We have,

$$V_{\mathbf{O}\sigma} = RV_{\mathbf{O}\varphi} + \mathbf{O}_{\varphi}{}^2 V_{\text{mates}} \tag{6}$$

When $R = 1$, Eq. (6) shows that the variance in fitness for males, $V_{\mathbf{O}\sigma}$, equals the variance in fitness for females, $V_{\mathbf{O}\varphi}$ (which $= 0$), plus the quantity, $\mathbf{O}_{\varphi}^2 V_{\text{mates}}$. This latter term equals the average female fitness squared, $\mathbf{O}_{\varphi}^2$, multiplied by the variance in mate numbers among males, V_{mates} $[=\Sigma p_j(R - j)^2]$. For the above example, $\mathbf{O}_{\varphi}^2 V_{\text{mates}} = 9$. This shows that the sex difference in fitness variance is due to the fitness effects of a sex difference in the variance in mate numbers, V_{mates}. In this case, variance in mate numbers exists among males, but not among females. Now recall that $\boldsymbol{I} = V_{\mathbf{W}}/\mathbf{W}^2$. We can obtain an analogous expression for the variance in relative fitness for males in terms of offspring numbers by dividing Eq. (6) by $[R\mathbf{O}_{\varphi}]^2$; that is, by the squared average offspring number for males.

When we do this, we obtain,

$$\boldsymbol{I}_{\sigma} = (1/R)(\boldsymbol{I}_{\varphi}) + \boldsymbol{I}_{\text{mates}} \tag{7}$$

Or, $\boldsymbol{I}_{\sigma} = (R_O)(\boldsymbol{I}_{\varphi}) + \boldsymbol{I}_{\text{mates}}$ because R equals 1/OSR $(= 1/R_{\text{O}})$. Thus, the opportunity for selection on males, $\boldsymbol{I}_{\sigma}$, equals the opportunity for selection on females, adjusted by the sex ratio, $(1/R)(\boldsymbol{I}_{\varphi})$, plus the opportunity for selection arising from differences in mate numbers among males, $\boldsymbol{I}_{\text{mates}}$. This expression, like Eq. (6), shows the relationship between male and female fitness. However, because this relationship is now standardized by the square of mean fitness, it provides estimates of relative fitness; that is, of selection opportunities for each sex. These expressions show that the sex ratio is only part of the total opportunity for selection. When the sex ratio equals 1 $(R = 1/R_{\text{O}} = 1)$, subtracting $\boldsymbol{I}_{\varphi}$ from both sides of Eq. (7) yields, $\boldsymbol{I}_{\sigma} - \boldsymbol{I}_{\varphi} = \boldsymbol{I}_{\text{mates}}$ $(= \Delta\boldsymbol{I})$. This relationship shows that the sex difference in the opportunity for selection, that is, the opportunity for sexual selection, is indeed due to differences in mate numbers between the sexes.

Inserting the values from our example into this latter equation, we see that when males and females have equal mate numbers, $\boldsymbol{I}_{\text{mates}} = 0$. When males vary in mate numbers, $\boldsymbol{I}_{\varphi}$ still equals 0, so all of the opportunity for selection on males is due to sexual selection or, $\boldsymbol{I}_{\sigma} = \boldsymbol{I}_{\text{mates}}$. If $V_{\mathbf{O}\varphi}$ becomes nonzero, either because females vary in their mate numbers, or because females vary in their offspring numbers, or for both reasons, $\boldsymbol{I}_{\varphi}$ will increase and $\boldsymbol{I}_{mates}$ will be eroded to a degree determined by the relative magnitudes of $\boldsymbol{I}_{\sigma}$ and $\boldsymbol{I}_{\varphi}$. If $\boldsymbol{I}_{mates} < 0$, the sex roles will reverse because sexual selection acts on females. However, erosion of $\boldsymbol{I}_{mates}$ may become negligible if the variance in mate numbers among individuals in one sex becomes large and in the other sex remains small. When this occurs, sexual selection on one sex can overwhelm the effects of natural selection acting on the other sex, leading to apparent cases of sexual exploitation. However, such situations are unlikely to last for long. Mating systems with strong sexual selection tend to be invaded by alternative mating strategies that reduce the variance in mate numbers within the sex in which it is large. Such invasions shut down strong sexual selection, and illustrate why sexual conflict is always self-limiting. The important empirical points are: (1) $\boldsymbol{I}_{mates}$ explains much about why sex differences exist, and (2) $\boldsymbol{I}_{mates}$ can be estimated for any population in which the mean and variance in offspring numbers among females, and the mean and variance in mate numbers among males are known.

error will be. This difficulty is part of the reason why estimates of OSR and potential reproductive rates (PRR) provide biased estimates of how selection operates within mating systems; they tend to focus only on the activities of successfully breeding individuals. To accurately measure their contribution to the total fitness variance, the identity and frequency of non-reproducing individuals must be clearly established.

All Individuals Have a Mother and a Father

How can experimenters account for the "zero class" of individuals in any given population, when they only obtain genotypes from the individuals they can actually collect? The answer is to recall three principles in reproductive biology that extend from R.A. Fisher's statement, "All individuals have a mother and a father." The first principle is that all offspring genotypes must be accurately assigned to both parents. If a sample of progeny and adults are genotyped from a population, it is certainly possible that every offspring collected and genotyped can be assigned explicitly to parents within the genotyped adults. However, it is more likely that some of the progeny can be assigned to only one parent, usually their mother, or they cannot be assigned to any of the genotyped adults at all. Although considerable work may have gone into identifying these individuals, and while they certainly can be used to answer other questions that parentage data can address, they must be excluded from the sample to account for the zero class. By focusing only on offspring whose parentage can be explicitly assigned to both parents, the experimenter is able to identify only the progeny produced within the population in question.

Next, in addition to the adults whose progeny can be identified, there are some genotyped adults within the population that can be shown to have failed to produce any offspring at all. These adults can be presumed to belong to the population in question because there were collected there. Moreover, if only the progeny that can be unambiguously assigned to both parents comprise the offspring produced within the population, and if the assignable adults comprise the breeding parents, the genotyped adults who have no assignable offspring, must comprise the fraction of the population that was unable to breed. They may have produced offspring in another nearby population, but the researcher is unable to account for these progeny. Therefore, the individuals without any assignable progeny within the focal population must constitute the zero class. The breeding and nonbreeding fractions of the adult population and the progeny they produced, within the focal population alone, are thus established.

A second principle possible from Fisher's statement provides a check on the accuracy of this parentage assignment. If all individuals have a mother and a father, the average number of progeny produced by males and by females must be equal. After parentage is assigned using the method described above, it is possible to calculate the mean and 95% confidence limits on the number of offspring produced by males and females. If parentage assignment has accurately captured both the breeding and the nonbreeding fractions of the population, there should be no significant difference in the average number of progeny assigned to each sex. If a significant difference exists, the accuracy of the parentage assignment is likely incorrect and should be reexamined.

A third outcome of Fisher's statement provides an additional means for examining total adult fitness in samples that appear to contain only the genotypes of individuals who successfully produce offspring. If all individuals have a mother and a father, and if equal fitnesses are found between males and females within the populations, then over time, the sex ratio of such populations is expected to equilibrate at unity. If the numbers of successfully and unsuccessfully reproduced individuals are added up within each sex, the resulting ratio of males to females, also should nearly equal 1. Again, a significant deviation of this ratio from unity could indicate that the parentage assignment is incorrect, or it could indicate that the population is in the process of reaching an equal sex ratio after some displacement. In this situation, examination of the population in the next generation is expected to yield a value closer to unity.

A Sex Difference in the Variance in Relative Fitness

Once the mean and variance in fitness can be identified for both sexes, it is possible to examine the sex difference in the variance in relative fitness ($=\Delta I$; Box 1) within the population. This value can be compared directly because the variance in fitness is proportional to the strength of selection operating within each sex. The statistical variance in fitness can be standardized by the squared average fitness to yield the opportunity for selection, I, for each sex ($=I_{\male}$ and $I_{\female}$; note that $I_{\male} - I_{\female} = \Delta I$). The value of I for each sex compares the fitness of breeding parents with that of all individuals of that sex within the population before breeding occurs, thus it measures the maximum intensity of selection as a result of an episode of selection (Box 1). While the average fitness for each sex must be equivalent, the variance in fitness can differ and the opportunity for selection will be stronger within the sex with the larger fitness variance. It is for this reason that the sexes are expected to diverge in character. When $I_{\male} > I_{\female}$, sexual selection will modify males; when $I_{\female} > I_{\male}$, sexual selection will modify females; when $I_{\male} = I_{\female}$, either sexual selection is nonexistent, or sexual selection operates with equal intensity in both sexes, causing both sexes to become modified relative to their ancestral state.

The Spatiotemporal Distributions of Matings and Fertilizations

The insight of Emlen and Oring has guided conceptual understanding of animal mating systems for nearly half a century. However, as mentioned, their proposed metrics for measuring the process of mating system evolution have proven unreliable for measuring selection on mating systems. Two related approaches for estimating the opportunity for selection within mating systems are available using data on the spatial and temporal distributions, of breeding adults and of fertilizations.

The Mean Crowding of Sexual Receptivity in Space and Time

The first method requires estimates of the spatial and temporal crowding of sexually receptive individuals. If males defend territories or resources where matings occur, the mean spatial crowding of receptive females is measured as the number of other receptive females the average female experiences when she becomes receptive, or $m^* = m + [(V_m/m) - 1]$, where, m equals the average number of receptive females per resource patch, and V_m is the variance around that average. When receptive females are maximally dispersed in space, m^* approaches zero. When receptive females tend to be aggregated on particular patches, the value of m^* increases.

If the breeding season can be divided into intervals, whose length equals the average duration of female receptivity, the mean temporal crowding of matings equals the number of other receptive females the average receptive female experiences with she becomes sexually receptive, or $t^* = t + [(V_t/t) - 1]$, where, t equals the average number of receptive females per interval and V_t equals the variance around that average. When female receptivity is maximally asynchronous within the breeding season, t^* approaches zero. When female receptivities overlap within fewer intervals, the value of t^* increases.

The opportunity for sexual selection, I_{mates} (also written as $\boldsymbol{I}_s$ or $\Delta\boldsymbol{I}$) measures the ability of particular males to secure more than one mate, at the expense of the mating success of other males (Box 1). Both m^* and t^* have been shown to estimate I_{mates}, but do so in different ways. The relationship between m^* and I_{mates} is proportional because when females are maximally crowded in space, one or a few males could defend and mate with all of the females in the population. In contrast, the relationship of between t^* and I_{mates} is reciprocal because when females are maximally crowded in time, the ability of one or a few males to mate with multiple females is reduced. The combined effects of m^* and t^* generate a surface that estimates the maximal selection intensity possible from empirical estimates of female spatial and temporal distributions.

The Total Variance in Fertilization Success in Space and in Time

The second method requires construction of a matrix, whose rows represent individual males, and whose columns represent intervals during the breeding season in which females may become receptive. This scheme works in species with separate sexes, as well as in monoecious, hermaphroditic and heterothallic species, as long as each individual and its gender or mating type can be identified. The cells of the matrix may contain zeros or larger numbers, identifying the number of mates or fertilizations that the individual within each row obtains within each interval of the breeding season. If multiple males inseminate or pollinate the same female, these numbers may also represent the fraction of the total fertilizations with a given female that a particular male obtains. This provision offers a specific means for estimating the strength of post-mating sexual selection within animals as well as plants, and provides a means for accommodating plant, animal, fungal, or other mating systems in which multiple fertilizations by different parents may occur within families.

The more precise parentage data are, the more detailed such analyses can be, and fractions of clutches instead of individual matings can be substituted into the matrix. This approach is especially powerful if parentage data are known, but even without it, the framework provides crucial information on whether or not particular individuals contribute disproportionately to fertilizations, and how such biases affect selection acting on sexual phenotypes. This approach captures the spatial and temporal elements of mating system variation in terms of total selection intensities, and avoids the statistical difficulties encountered when using estimates of EPP and OSR.

The matrix described above includes the total variance in fertilization success among the individuals the rows represent. Moreover, as in an ANOVA problem, the total variance in fertilizations can be partitioned into the within- and among-individual components, providing a means for evaluating individual fertilization success in space and time. Although the rows may represent males or females, depending on the structure of the breeding system, for simplicity, let us suppose that the rows represent males, and that these males obtain all of the fertilizations possible with each female they secure as a mate, making mating success and fertilization success equivalent. Then, as in ANOVA, we can identify two components of the variance in mating success; the component arising within each of the males representing the rows; that is, the variance in mating success that arises in space, and the component arising among all males, across all of the columns in which females become receptive; that is, the variance in male fitness that arises in time.

First note that within each interval (column) of the breeding season, the total number of receptive females present, divided by the total number of males in the population (all of the rows), equals the interval sex ratio; that is, the average number of females mated by each male at that time. Notice too that the average of the squared number of females mated by males within each interval, minus the squared interval sex ratio, equals the variance in mates per male, obtained for each interval. These two quantities provide the mean and variance in male mating success, estimated across the breeding season intervals in which females are receptive (see Box 1).

Remember that ANOVA problems partition population variance into within- and among-group components. In general, the within-group component equals the average of the variances estimated for each group, whereas the among-group component equals the variance of the averages estimated for each group. If the "groups" we are considering are the males in this population who vary in their mating success, and if the analysis we seek is to partition to total variance in mate number, in time, into within- and among-male components (again, mates or fertilizations can be distinguished as described above), then the within-male component of the total variance in mate numbers, in time, equals the average of the variances in mating success for the males, across all of the intervals in which females are present. Similarly, the among-male component of the total variance in mating success in time, equals the variance of the average male mating success, across all of the intervals in which females are present.

But this is only part of the problem we want to solve. We also seek to partition the total variance in fertilizations, in space, into within and among-male components. To do this, note that within each row of the matrix, the total number of receptive females who mate with that male, divided by the total number of intervals containing at least one female, equals the average number of mates for each male, estimated across all of the intervals containing receptive females; or, the average number of females mated by each male, over the entire breeding season. Notice too that the average of the squared number of females mated by each male, estimated across the intervals containing receptive females, minus the squared average number of mates for each male, estimated across the intervals containing receptive females, equals the variance in mate number, across the breeding season, for each male. These two quantities provide the mean and variance in male mating success, estimated within the male rows, across the breeding season intervals in which females are receptive.

Again, recall that ANOVA problems partition population variance into within- and among-group components. Because the "groups" we are considering are the males in this population, and now, the analysis we seek is to partition to total variance in mate

number, in space, into within- and among-male components, then the within-male component of the total variance in mate numbers, in space, equals the average of the variances in mating success, across all intervals in which females are present, for the average male. Similarly, the among-male component of the total variance in mating success, in space, equals the variance of the average mating success that the average male has, across all intervals in which females are present.

The separate estimates of the variance in mate numbers among males, in time and in space, can then be algebraically combined so that the total variance in male mating success is partitioned into three components. These variance components, when divided by the squared average mate numbers for males, yield estimates of the opportunity for sexual selection (Box 1) that arises among males in three contexts, due to the availability of mates in time and space. Specifically, the total opportunity for selection on males, I_{mates}, evaluated in this way equals,

$$I_{mates} = I_{sex\ ratio} + \left[{}^{*}I_{mates(t)} - {}^{*}I_{males(k)}\right] \quad (1)$$

in which, (a) $I_{sex\ ratio}$, is the opportunity for sexual selection caused by temporal variation in the sex ratio. This value is similar to the OSR, but unlike OSR does not overestimate the intensity of sexual selection within the intervals in which females are abundant; (b) ${}^{*}I_{mates(t)}$, is the weighted opportunity for sexual selection caused by temporal variation in the availability of females. This value varies in magnitude depending on whether female receptivity is synchronous or asynchronous, and is weighted by the number of females that appear in each interval, with larger numbers of females per interval contributing the largest effects. Lastly, (c) ${}^{*}I_{males(k)}$, is the weighted opportunity for sexual selection caused by variation in the availability of females for mating by the spatially distinct males. Note that ${}^{*}I_{males(k)}$ decreases the value of ${}^{*}I_{mates(t)}$ because ${}^{*}I_{males(k)}$ estimates the variance in mate number for individual males. Overall, increases in the variance in relative fitness within males tend to decrease the variance in relative fitness that occurs among males due to the temporal availability of females.

Concluding Remarks

As information on mating systems accumulates across all species, it is increasingly clear that typological approaches, while necessary for documenting the existing variation, cannot explain how mating system diversity arises, is modified, and persists or changes over evolutionary time. Increasing spatial mobility within species, either by individuals themselves or as mediated by pollinators or the environment, appears to increase the spatial and temporal scales at which matings and gamete transfers can occur, evidently favoring increased outcrossing and ultimate the separation of the sexes entirely. Expanding familiarity with the range of this variation in all species can lead to better metrics for documenting the sources and intensity of selection that shape mating systems evolution. The time has come to end taxonomic parochialism in mating system research and embrace its unifying principles.

See also: Behavioral Ecology: Kin Selection; Social Behavior and Interactions; Parental Care; Sexual Selection and Sexual Conflict. General Ecology: Communication

Further Reading

Barrett, S.C.H., 2002. The evolution of plant sexual diversity. Nature Reviews Genetics 3, 274–284.
Bateman, A.J., 1948. Intra-sexual selection in *Drosophila*. Heredity 2, 349–368.
Charnov, E.L., 1982. The theory of sex allocation. Princeton, NJ: Princeton University Press.
Clutton-Brock, T.H., 2017. Reproductive competition and sexual selection. Philosophical Transactions of the Royal Society B 372.20160310.
Crow, J.F., 1958. Some possibilities for measuring selection intensities in man. Human Biology 30, 1–13.
Darwin, C.R., 1874. The descent of Man and selection in relation to sex, 2nd edn. New York: Rand, McNally and Co.
Emlen, S.T., Oring, L.W., 1977. Ecology, sexual selection, and the evolution of mating systems. Science 197, 215–223.
Fisher, R.A., 1958. The genetical theory of natural selection, 2nd edn. New York: Dover Press.
Karron, J.D., Ivey, C.T., Mitchell, R.J., Whitehead, M.R., Peakall, R., Case, A.L., 2012. New perspectives on the evolution of plant mating systems. Annals of Botany 109, 493–503.
Krakauer, A.H., Webster, M.S., DuVal, E.H., Jones, A.G., Shuster, S.M., 2011. The opportunity for sexual selection: Not mismeasured, just misunderstood. Journal of Evolutionary Biology 24, 2064–2071.
Ni, M., Feretzaki, M., Sun, S., Wang, X., Heitman, J., 2011. Sex in fungi. Annual Review of Genetics 45, 405–430.
Shuster, S.M., 2009. Sexual selection and mating systems. Proceedings of the National Academy of Sciences 106, 10009–10016.
Shuster, S.M., Wade, M.J., 2003. Mating systems and strategies. Princeton, NJ: Princeton University Press.
Trivers, R.L., 1972. Parental investment and sexual selection. In: Campbell, B. (Ed.), Sexual selection and the descent of man. Chicago, IL: Aldine Press, pp. 136–179.
Wade, M.J., 1979. Sexual selection and variance in reproductive success. American Naturalist 114, 742–764.

Optimal Foraging Theory

DW Stephens, University of Minnesota, St. Paul, MN, USA

Foundations

Foraging is a fundamental aspect of animal behavior that has implications for predator–prey interactions, competition, and studies of animal cognitive abilities. Animal foraging may be as dramatic as a lion stalking a gazelle, or as mundane as a barnacle filtering plankton from seawater. Foraging theory seeks to bring order to this diversity, by recognizing and analyzing the common problems faced by foraging animals. Foraging theory, or optimal foraging theory as it was originally known, has its origins in seminal papers by Schoener, Charnov, Parker, MacArthur, and Pianka, and Pulliam and Emlen published in the 1960s and 1970s respectively. Taken together, these papers produced a remarkably cohesive body of theory based on two common foraging problems: patch exploitation and prey selection.

Basic Models

Both of these basic models assume that a forager encounters items (prey or patches) one at a time, according to some well-behaved process (often a Poisson process). For example, if a forager encounters and consumes prey items that take h time units to handle and provide e calories of energetic benefit when consumed, then Holling's disk equation gives the rate of energy intake:

$$\frac{e}{1/\lambda + b} = \frac{\lambda e}{1 + \lambda h} \quad [1]$$

where λ gives the encounter rate, so that $1/\lambda$ is the expected time between encounters with prey items. The form of the equation on the left hand side shows how Holling's disk equation is simply the expected energetic gains per encounter divided by the expected time (search time plus handling time) per encounter. One can generalize this basic structure to include multiple prey types or to situations where the forager encounters patches instead of prey items. Both models find the foraging behavior that maximizes the rate of energy intake as given by Holling's disk equation, but the models focus on different aspects of foraging behavior. The following sections outline the prey and patch models.

The prey model

As described above, this model assumes that we can associate an energy value e_i, a handling time h_i, and an encounter rate λ_i with each prey type. The model solves for the attack probability p_i for each prey type that maximizes the rate of energy intake. The model makes three predictions: (1) A forager should always take or always ignore a given prey type. Foraging models call this the zero–one rule because it follows from the mathematical observation that the rate-maximizing attack probabilities can only be zero or one. (2) Prey types should be ranked by their profitabilities, which we define to be quotients of the form e_i/h_i. That is, the prey type with the highest energy to handling time quotient is the `best' type (rank 1), and the next highest is rank 2 and so on. (3) One can determine the set of prey that maximizes intake rate by working through the possible `diets' in rank order. That is, first considering a diet by rank 1 prey only, then a diet of rank 1 and 2. Obviously, one can use Holling's disk equation to calculate the rate of energy intake for each of these diets. We can show mathematically that if, for example, a diet consisting of types 1–3 yields an intake that is smaller than the profitability of the fourth ranked prey type e_4/h_4, then adding this fourth ranked type will increase the intake rate. Therefore, to find the rate-maximizing diet, we simply add prey types to the diet in rank order until this is no longer true.

This result focuses our attention on the properties of the `best' or first ranked prey item. Obviously, a forager should always attack this best item upon encounter, and the properties of this type determine whether the forager should attack the second best item. Specifically, if the best type is abundant (has a high encounter), then a `best-type-only' diet may make sense. If, however, the best type is rare, then it typically makes sense to add the second best type to the diet. Notice that the abundance of the second best type is not important in this reasoning! Some investigators find this result counterintuitive, because they can imagine situations in which superabundant but mediocre prey items attract a forager's attention.

The patch model

The patch model assumes that a forager encounters patches rather than prey items. While we characterize a prey item by its available energy and handling time, we characterize a patch by its gain function. A gain function, $g(t)$, gives the relationship between the time spent exploiting a patch (t) and the energy gains extracted from a patch (g). We assume that a forager can extract energy from a patch at a fairly high rate initially, but this rate declines as the forager spends more time in the patch simply because the forager's exploitation depletes the patch. So, we typically draw gain functions as increasing curves that bend down. The patch

 doi:10.1016/B978-0-444-63768-0.00026-3

models solve for the patch residence time (t) that maximizes intake rate. For a situation with only one patch type, one can easily show that the rate-maximizing patch residence time, say t^*, satisfies

$$g'(t^*) = \frac{g(t^*)}{1/\lambda + t^*} \quad [2]$$

Where $g'(t^*)$ represents the derivative of the gain function with respect to patch residence time. Modelers call this condition the marginal-value theorem because marginal rate is a synonym for derivative. The condition tells us that at the rate-maximizing residence time, the instantaneous (or marginal) rate of intake equals the overall rate of intake. Algebraically, this marginal-value condition can be difficult to solve, but it has a very simple graphical solution (Fig. 1). The model predicts that in poor habitats (low overall rate of intake) foragers should spend more time in patches extracting more from each patch; while in rich habitats the model predicts that foragers should `skim the cream' from each patch – spend less time and extract less.

Opportunity costs

Notice that although these models seem to address quite different aspects of foraging behavior, they are logically similar. The marginal rate of intake in the patch model plays a role that is very similar to the profitability of lower ranked types in the diet model's inclusion algorithm. Both models maximize the long-term rate of energy intake, and the central properties of both models

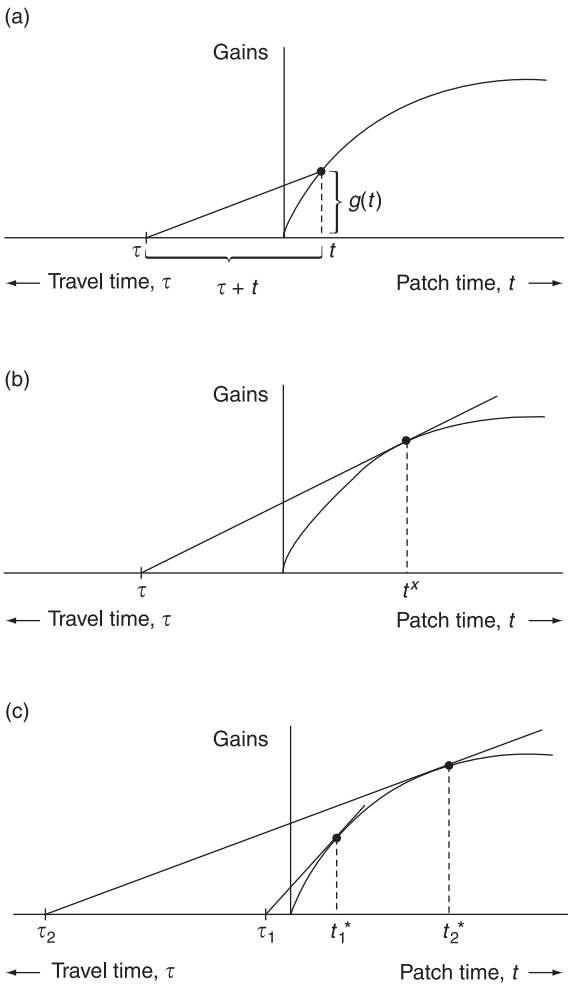


Fig. 1 The figure shows the graphical solution of the classical optimal patch exploitation model (the marginal-value theorem). The right hand side of all three panels shows a curve that gives the relationship between time spent exploiting a patch (patch time) and the resources extracted from the patch (gains). This curve is the function $g(t)$. The left hand side of each panel shows travel time (τ) increasing to the left. (a) shows a given travel time (τ) and an arbitrary chosen patch time (t). If we consider a slanting line from τ on the travel time axis to the point [t, $g(t)$] on the right side, we see that the slope of this line $g(t)/(\tau + t)$ is the rate of gain associated with patch time t. Clearly, the patch time shown in (a) is not the best, because we can increase this slope (the rate intake) by choosing a larger patch time. (b) shows that the optimal patch time corresponds to the case where the slanting `rate line' just touches (is tangent to) the gain function. (c) uses this solution to show the main prediction of the patch model, when travel times are long, say τ_2, then this tangent point just corresponds to a long patch time (t_2), but if travel times are short (e.g., τ_1) then we predict a short patch time (t_1).

stem from opportunity costs. For example, it can never be a mistake to attack the highest ranked item in the prey model, but it can be a mistake to attack the lower ranked items because in doing so, a forager might miss an opportunity to attack a higher ranked item. The role of opportunity costs is even clearer in the patch model. The model predicts that foragers should `skim the cream' from patches in rich habitats, because when a forager spends too much time in patches in a rich habitat, it loses opportunity to exploit fresh patches elsewhere.

Experimentalists and field workers have tested these models many times. Data broadly support many of their qualitative predictions. For example, virtually all patch-use studies support the prediction that foragers should spend more time exploiting patches in poor habitats. In addition, the claim that the abundance of high-quality prey types shapes animal selectivitiy is also well supported. Yet, the zero–one rule almost never holds. Each investigator in this area has a different interpretation of the empirical results. Some are encouraged by the pattern of qualitative agreement, while others emphasize the quantitative failures. In modern foraging theory, these models play a role like that of the Lotka–Volterra models in population ecology. They may not apply to any given situation, but modelers and students of foraging need to understand them so that they can use them as starting points for new models, and recognize their predictions within more complex situations. Literally hundreds of studies have used these models as building blocks. The work of Pirolli and Card on `information foraging' provides an especially novel example. These workers have used ideas from patch exploitation theory to analyze how computer users interact with websites and databases.

Most investigators recognize the patch and prey as the historical foundation of foraging theory, but as the theory has developed over the last 20 years, these same investigators have come to recognize two more ideas as basic building blocks of foraging theory: the ideal free distribution and dynamic optimization.

The ideal free distribution

Classical foraging theory, as represented by the patch and prey models, focused on the actions of solitary forager. The ideal free distribution provides a framework for extending foraging theory to cases in which animals forage in groups. The ideal free distribution considers how animals in a group should distribute themselves between two feeding sites. The foragers in this model are `ideal' in the sense they can immediately recognize the quality of sites and they share the resources at a site equally,that is, if $n$ individuals occupy a site, each obtains $1/n$th of the available food. The foragers are `free' because they can move freely between sites without cost. If site one produces food at rate r_1, then the n_1 individuals present there can expect to obtain food at rate r_1/n_1; similarly individuals at site two can expect to obtain r_2/n_2. The ideal free distribution holds that a stable distribution of individuals among sites can only occur when the intake rates in the two sites are approximately equal, because if intake rates differ in the two sites, then individuals at the lower-rate site can do better by moving to the higher-rate site. Data supports the ideal free model's predictions surprisingly well, even though the assumptions (e.g., equal sharing) are seldom satisfied. Although one can criticize the ideal free model at many levels, its focus on the economics of leaving and joining is fundamental to social foraging models.

Dynamic optimization

Dynamic optimization is not a model like the patch or prey model, instead it is a technique for modeling foraging and life-history problems. The patch model is a single-variable maximization problem like those studied by calculus students (e.g., find the height of tin can that maximizes volume). Dynamic optimization allows us to solve for an optimal trajectory or function, rather than a single variable. Consider the problem of foraging in the presence of predators. Typically, rich feeding sites are also the riskiest, while poor feeding sites are safer. As a rule, larger animals experience a lower risk of predation than small ones. So a small animal choosing the poorer but safer feeding site is also choosing a smaller future body size that could increase its risk of predation in the future. In situations like this, we cannot consider the best action at time 1 without also understanding the implications for time 2 and so on. Two pairs of workers (Marc Mangel and Colin Clark, and Alasdair Houston and John McNamara) have pioneered the application of dynamic models to foraging and behavioral ecology. Investigators have applied dynamic optimization to the study of predator avoidance and food storage (e.g., caching) with notable success, and dynamic models clearly represent a step forward in sophistication. The downside is that it can be difficult to solve dynamic optimization problems, and we must often resort to numerical techniques.

Where Is Foraging Theory Now?

Foraging theory has had a colorful intellectual history with vehement detractors (e.g., a paper with the title ``10 reasons why optimal foraging theory is a complete waste of time''), and zealous proponents. While the basic models and techniques outlined above represent basic building blocks of modern foraging studies, they are no longer the definitive core of `foraging theory.' Research that draws on this tradition is now so diverse that an outsider may find it difficult to recognize `foraging' as a single cohesive topic. This article recognizes three interconnected themes in contemporary foraging research. First, some investigators have focused on the neural, psychological, and physiological mechanisms underlying foraging behavior. Second, other investigators have generalized and extended traditional foraging models, developing a modern behavioral ecology of foraging that includes predator avoidance, social behavior, and food storage. Finally, a growing number of investigators considered the implications of foraging behavior for population and community level processes. The remainder of this article considers these three themes in turn.

Behavioral Mechanisms

Suppose we observe that foraging blue jays spend more time gleaning insects from clumps of foliage in poor habitats, as the patch model predicts. How do the jays achieve this? Do they learn the difference between rich and poor habitats? Does an internal clock trigger patch departures, or do jays use declining rates of prey capture to decide when to leave? Foraging theory, like the rest of behavioral ecology, emerged from a tradition of `mechanistic agnosticism', that ignores these types of questions. Natural selection, the argument goes, favors `outcomes' that might be achieved by any of several mechanisms, and one can study these behavioral outcomes and their fitness consequences without necessarily understanding the underlying mechanisms. The modern perspective holds that although one `can' study the fitness consequences of behavior without a complementary analysis of mechanisms, we achieve a richer and more complete understanding when we combine these approaches.

Many types and levels of mechanism influence foraging behavior. These include the physiological process of digestion and energy storage; the neural mechanisms of learning, memory, and sensation; and the psychological processes of learning, discrimination, and decision making. Each of these topics has something to add to the study of foraging. Readers should recognize that the interactions between the traditional models that follow from hypothesis about fitness consequences, and mechanistic analyses flow in two directions. Mechanistic approaches provide information about processes that can constrain foraging behavior, and we can use this information to refine traditional models. In the other direction, fitness-based models can inform mechanistic analyses by explaining the context and selective value of the behaviors controlled by a given mechanism. The following sections outline some key recent developments in this area.

Cognition and psychological phenomena

Many birds cache food items for later consumption. This raises basic questions about the fitness value of external food storage: how does food availability influence caching behavior (temporal patterns of abundance and dearth seem to be important); when is `external' storage superior to internal storage (e.g., as fat – predator avoidance may be important here). It also raises the question of how caching animals find and retrieve food items that they may have hidden weeks or even months ago. Studies of memory abilities of caching birds provide a textbook example of how to combine mechanistic and functional analysis. Evidence suggests that some caching birds have better spatial memories than related noncaching species. In addition to this, caching species tend to have larger hippocampuses in their brain (a structure implicated in memory), and some evidence suggests that the experience of caching food leads to an enlarged hippocampus. Finally, recent work by Clayton and Dickinson uses caching behavior to provide the first evidence of so-called episodic memory in a nonhuman species.

Neurobiology and neuroethology

Recent advances in neurobiology offer the promise of meaningful connections between neural mechanisms and naturally occurring foraging behavior. For example, worker honeybees undergo a well-defined progression from nurse to forager as they age. An impressive body of work now suggests that the worker's brain become modified for foraging as this sequence progresses. Particularly, the mushroom bodies (a structure crudely comparable to the vertebrate hippocampus) enlarge as the workers develop into foragers, and this structure appears to serve both the spatial abilities that foragers need to find and collect resources, and the abilities of foragers to form associations. The star-nosed mole provides an example of how foraging models can inform neurobiology. The `star' is a 22-lobed sensory structure covered with roughly 25 000 receptors and served by 100 000 nerve fibers, supported by a nearly literal `map' in the mole's brain. Why has this spectacular structure evolved and why don't other moles have such elaborate stars? The neurobiologist Kenneth Catania has turned to classical foraging models for insight. Behavioral studies suggest that the star and its associated neural machinery dramatically reduce the handling time associated with the small vertebrate prey that predominate in the mole's damp habitat.

Digestive physiology

Historically, students of foraging have paid little attention to the grimy fate of ingested food. Several recent studies have, however, shown that students of foraging must consider the `behavior' of the digestive system as part of an animal's overall foraging strategy. Internal physiological processes actively manage guts. Snakes provide a stunning example. A snake digesting a freshly caught meal dramatically upregulates its gut, increasing the weight of the intestinal mucosa, lengthening the microvilli, and increasing the overall metabolic rate by a factor of 40! All this happens over a very short timescale, and when the snake has finished digesting its meal, the process is reversed, downregulating the gut and associated metabolism. While the gut management of feast-or-famine foragers, like snakes, are obviously dramatic, studies with many types of animals document the flexible behavior of animal guts. Guts can be sensitive to the availability of food, as in the snake example, but they can also change in response to food types. For example, lengthening for a vegetarian diet and shortening for a carnivorous diet.

Modern Foraging Theory

While the basic foraging models provide insight into many foraging problems, they seldom provide a complete solution of any particular foraging problem. For example, they say nothing about tradeoffs with predator avoidance, or social interactions. In response, behavioral ecologists have extended and refined foraging theory in several directions. This section identifies three of these growing points in foraging theory.

Herbivory

Students of foraging have long recognized that the herbivores face different problems than carnivores. Recent work on herbivory shows two significant trends. First, students of herbivory have developed new intake rate models that are not based on Holling's disk equation. The best known of these is the Spalinger–Hobbs model. These `bite rate' models recognize that herbivores can often handle their food (i.e., chew it) while searching for the next `bite,' (a violation of the typical assumptions of the Holling approach). Whether constraints on chewing rate or constraints on food-discover limit, intake rate depends on the properties of the forager and the ecological situation. The second key trend in this area is the breakdown of barriers between applied students of herbivory (e.g., those who study domestic stock and insect pests) and behavioral ecologists interested in herbivory. These sort of barriers fall more slowly than one would hope, yet agricultural scientists offer many years of data and access to some impressive experimental systems and techniques, while behavioral ecologists offer comparisons to the natural setting and ideas about the evolutionary forces that have shaped the behavior of herbivores.

Foraging in the face of predation

Algae grow on top of rocks. Grazing snails would prefer to feed where the algae grow, but the exposed surfaces of rocks are also exposed to predation. Like grazing snails, many foragers face a fundamental tradeoff between obtaining food and avoiding predation. Indeed, the literature of `predator' effects is enormous and fast growing. Foragers may change where they feed, when they feed, or even the company they keep while feeding in the face of predation risk. This isn't news. Since the early days of foraging models, investigators have realized that a complete understanding of foraging would have to take predation into account. The problem is how to build quantitative models that account for foraging gains and mortality risk at the same time. The answer is a life-history approach. When we consider an animal's life history, we naturally think of current survival and future reproduction. Now the key problem is to specify how foraging behavior influences these two quantities. While there is, at present, no generally satisfactory way to specify universal relationships between foraging and life-history parameters, modelers recognize several important special cases. In one special case, for example, one can use the premise to simple maximization of survival probability to construct models that combine food acquisition and predator avoidance. Perhaps the most widely studied special case is Gilliam's 'μ over g' rule which minimizes the quotient of mortality rate (μ) divided by growth rate (g). Gilliam's rule applies to situation in which animals are not actively reproducing so that they allocate acquired resources to growth.

Social foraging

Many organisms forage and live in groups that range from loose aggregations to complex societies. The subtopic of `social foraging theory' pioneered by Giraldeau and Caraco takes a game theoretical approach to these problems. Group size models typically use a member–joiner paradigm that generalizes the approach taken in the ideal free distribution. In this approach, we plot the relationship between foraging gains (often, but not necessarily, measured as rate), and group size. The function typically increases to a maximum as new recruits contribute to the group's foraging efficiency and then declines as further recruits make excessive demands on the group's limited ability to detect and acquire new resources. Observations suggest that animal groups are typically larger than the rate-maximizing size. According to theory, this occurs because singletons can benefit from joining the group well beyond the rate-maximizing group size. The group size that is ultimately stable can depend on many factors such as the degree of relatedness between group members and the prospective joiners, that the degree to which current group member can limit joining. Models of within-group behavior have focused on the producer–scrounger framework. In this approach, some individuals (the producers) find food and others (the scroungers) parasitize their discoveries. Again, we need a game theoretical approach, because the scrounger tactic only makes sense if there are some producers. If group members can easily search for their own food and simultaneously `keep an eye out' for the discoveries of others, then we expect the group members should attempt to scrounge whenever the opportunity arises, since they lose nothing in doing so. If, however, watching for the discoveries of the other compromises your ability to find food on your own, this sets the stage for a more subtle balance between the producer and scrounger strategies.

Foraging Ecology

A key goal of early foraging models was to inform ecological theory, specifically population dynamics and community ecology. Early studies, however, tended to focus on the behavioral predictions of foraging models deferring studies of ecological implications for another day. This focus on behavior clearly frustrated some students of foraging ecology, but more recent work shows signs of fulfilling foraging theory's promise as an ecological tool. Within population dynamics, for example, the advent of individual-based models have given ecologists a tool with which they can study the population implications of foraging strategies, because one can specify the foraging strategies of the computational `individuals' within the model. Turner and colleagues have used this approach recently in a study of elk and bison populations within Yellowstone National Park making the prediction that the effects of fire on resource quality are keys to the winter survival of these two large grazers. Community ecologists have also begun to use ideas from foraging theory more frequently. For example, community ecologists have used an empirical technique derived from patch exploitation theory (Joel Brown's `giving-up density' procedure) to draw conclusions about species interactions and coexistence. Using this technique, Kotler and Blaustein found that one species of gerbil was more sensitive to the threat of

predation than the other, and this partially explains how the two species divide the available space; one uses more open habitat and the other uses habitat with dense cover. Perhaps the most exciting development in foraging ecology is the recognition of the indirect effects of predation. Ecologists have always recognized that predators (who are after all engaged in the business of foraging) affect populations by killing, but as foraging theory's models of predator avoidance tell us, predators influence prey behavior in a myriad of sublethal ways. These sublethal effects of predation can profoundly effect populations and communities. For example, changes in habitat use can influence plant communities, and this in turn can influence the other animals that depend on these habitats. While we are still in the early days of this research topic it is a very exciting patch for further exploitation.

The issues discussed in the article only represent a sampling of contemporary topics in foraging behavior. Other developments include models of energy storage (both external via caching, and internal via fat), studies of provisioning that recognize that animals must often balance the problems of feeding themselves and feeding their dependent offspring, and growing connections with conservation biology.

Conclusion

Over the last 20 years, the models of conventional foraging theory have provided a foundation for many new developments. These developments have produced an extremely diverse and vibrant body of inquiry. If there is a downside to this, it is that investigators have become so specialized that they may not recognize the commonalities between approaches. The challenge for the next 20 years is to keep the lines of communication open, so that we can recognize opportunities to understand better, the role foraging plays in behavior and ecology.

See also: Conservation Ecology: Trophic Index and Efficiency; Turnover Time; Trophic Classification for Lakes. Ecological Data Analysis and Modelling: Conceptual Diagrams and Flow Diagrams; Climate Change Models. General Ecology: Ecological Stoichiometry: Overview. Global Change Ecology: Nitrogen Cycle

Further Reading

Brown, J.S., 1988. Patch use as an indicator of habitat preference, predation risk, and competition. Behavioral Ecology and Sociobiology 22, 37–47.
Charnov, E.L., 1976. Optimal foraging and the marginal value theorem. Theoretical Population Biology 9, 129–136.
Charnov, E.L., 1976. Optimal foraging: Attack strategy of a mantid. American Naturalist 110, 141–151.
Clark, C.W., Mangel, M., 2000. Dynamic State Variable Models in Ecology: Methods and Applications. New York: Oxford University Press.
Clayton, N.S., Dickinson, A., 1998. Episodic-like memory during cache recovery by scrub jays. Nature 395, 272–274.
Emlen, J.M., 1966. The role of time and energy in food preference. American Naturalist 100, 611–617.
Fryxell, J.M., Lundberg, P., 1998. Individual Behavior and Community Dynamics. London: Chapman and Hall.
Giraldeau, L.A., Caraco, T., 2000. Social Foraging Theory. Princeton, NJ: Princeton University Press.
Houston, A.I., McNamara, J.M., 1999. Models of Adaptive Behaviour. Cambridge: Cambridge University Press.
MacArthur, R.H., Piankar, E., 1966. On optimal use of a patchy environment. American Naturalist 100, 603–609.
Pirolli, P., 2007. Information Foraging Theory: Adaptive Interaction with Information. New York: Oxford University Press.
Schoener, T.W., 1971. Theory of feeding strategies. Annual Review of Ecology and Systematics 2, 369–404.
Stephens, D.W., Brown, J.S., Ydenberg, R.C. (Eds.), 2007. Foraging: Behavior and Ecology. Chicago, IL: University of Chicago Press.
Stephens, D.W., Krebs, J.R., 1986. Foraging Theory. Princeton, NJ: Princeton University Press.
Vincent, T.L., Brown, J.S., 2004. Evolutionary Game Theory, Natural Selection and Darwinian Dynamics. Cambridge: Cambridge University Press.

Orientation, Navigation and Search☆

Jochen Zeil, The Australian National University, Canberra, ACT, Australia

Orientation in Space

Probably the most fundamental requirement for a defined orientation relative to the world is imposed by vision and by the need to appropriately direct motor commands. There are only two "absolute" and reliable reference cues animals can use for their alignment with the world: one is the direction of gravity and the other is the division of the world into a celestial and a terrestrial hemisphere, separated by a visual horizon line, with light coming from above. Animals have evolved a number of sensor systems to determine "where is up": they use collections of heavy materials bedded on shearing sensors (statocysts or otholiths) which signal the direction of gravitational pull. Some water insects trap air bubbles in sensory hair fields which serve the same role. In terrestrial animals, the distribution and orientation of body mass relative to gravity can be measured by proprioceptors, in vertebrates by muscle spindles and stretch receptors in the joints and around internal organs and in insects, by mechanoreceptors in the head-thorax and thorax-abdomen joints. The second absolute reference is the fact that light comes from above. Insects and fish are known to use this property of the world to align their vertical body axis and/or their eyes. The compound eyes of especially flying insects are aided in this task by an additional visual system, the ocelli, which are an assembly of three lens eyes with huge visual fields on the top of the head. In dragonflies and flies, ocelli have been shown to be exquisite horizon sensors, in addition to "dorsal light sensors," which elicit compensatory roll and pitch movements of the head, whenever it is not aligned horizontally.

There are a number of other animal sensors that cannot by themselves assure proper alignment with the vertical, but are essential in controlling and maintaining orientation. Vertebrates, crustaceans and flies possess vestibular systems that are designed to sense the rotational acceleration (in the case of semi-circular canals in vertebrates and crustaceans) and the velocity (in the case of fly gyroscopes) of the body and the head. The principle of operation in semi-circular canals is the fact that liquid in tubes embedded in the head does not immediately follow the acceleration of the head due to inertia. There is a brief moment then, in which the liquid moves relative to the canal walls and this movement is being sensed by hair cell mechanoreceptors, which are sensitive to shearing forces. The relative motion is caused by acceleration only: during prolonged rotation at a constant velocity, the liquid comes to rest due to friction. The main function of these vestibular systems is to stabilize gaze around the three rotational axes, the yaw, pitch and roll axes, by sending the appropriate commands to the eye muscles in vertebrates and crustaceans, and to the neck muscles in flies, which then move the eyes or the head into the opposite direction. The modified hindwings in flies, called halteres, are a special case of such inertial sensors: they consist of clubs on rigid stalks, which during flight rapidly oscillate up and down in anti-phase to the wings. Any rotation of the flying insect creates an inertial force which resists the change of orientation of the plane in which the halteres oscillate. This Coriolis force (which is also operating on the large gyral ocean and air current systems on earth) is then sensed by fields of strain sensors at the base of the halteres and their output drives compensatory head movements of the fly. Because these inertial systems provide feed forward information (their effect does not feedback directly onto the sensor itself) with little delay, the reflex movements they elicit belong to the fastest biological sensor-action loops. However, the information they provide is dynamical, lacks an absolute reference and is unreliable at low accelerations and low angular velocities. Inertial sensors by themselves therefore cannot prevent slow drift from the absolute reference orientation. They thus need to operate in concert with other sensory systems, most notably the visual system and its high sensitivity to image motion (Fig. 1).

All visual animals exhibit a strong reflex which counteracts large-field image motion on the retina. Such large-field motion is produced by rotational movements of the eyes, the head or the body, which cause the retinal image to shift in the opposite direction. Animals respond to such un-intended rotations by eyes, head or body movements in the opposite direction. This so-called optomotor response, or optokinetic reflex, is part of a negative feedback loop, with the speed of image motion as input and a motor command as output which—within a certain range of speeds—reduces image motion to zero. As a result, the retinal image is said to be stabilized, which is not quite true, because there will always be some residual image motion experienced by an animal when it moves through the world, that is caused by the animal's translational movement. However, the retinal image is cleaned from the effects of rotations by the optomotor response (around all three rotational axes) and thus helps to stabilize gaze and to keep the

☆*Change History*: March 2018. Jochen Zeil updated sections, reference added to Figs. 1 and 3 legend; Reference list culled and updated.

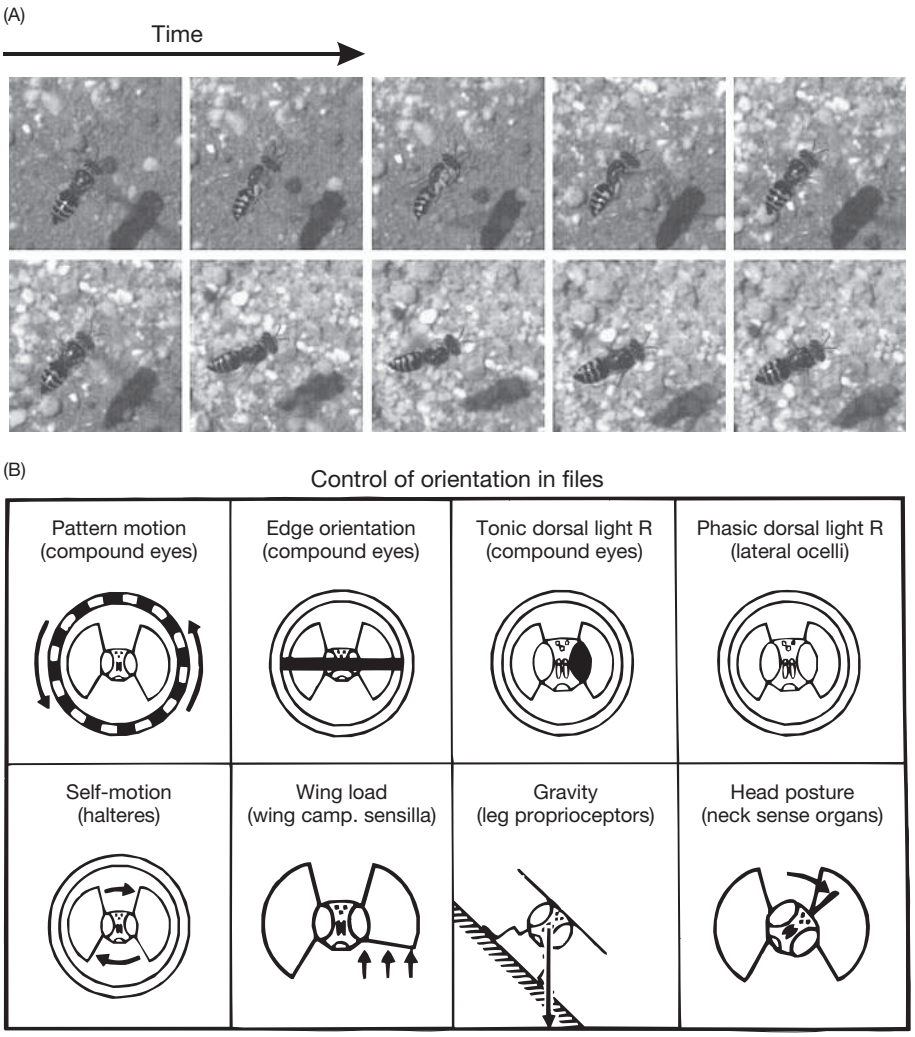


Fig. 1 Orientation in space. (A) Head stabilization in flight of a sand wasp (*Bembix* sp.). The sequence of high speed images shows the wasp as it executes a sideslip movement to the top left by rolling its body first to the left (top row) and then to the right as a breaking manoeuvre (bottom row). As the body rolls through nearly 180 degree, the head remains perfectly aligned with the horizontal. High speed video images were taken at 250 fps (from Zeil, J., Boeddeker, N. and Hemmi, J.M. (2008). Vision and the organization of behaviour. *Current Biology* **18**, R320–R323). (B) The different sensory inputs that contribute to head stabilization and orientation in flies (from Hengstenberg, R. (1993). Multisensory control in insect oculomotor systems. In: Miles, F.A. and Wallman, J. (eds.). Visual motion and its role in the stabilization of gaze, pp. 285–298. Amsterdam: Elsevier Science).

visual system aligned with the vertical. Although this sounds simple enough, there are some major computational problems animals have to solve, in order to achieve this optomotor stabilization: in flies, at least 60 large-field, motion sensitive neurons, many of them tuned to a particular rotational image motion component, are involved in the task of controlling the rotational movements of the fly's head!

Compass Direction and Distance Traveled

Even on their briefest journeys, like the 1 m or so excursions of fiddler crabs on their grazing trips away from their burrows, animals need to keep track of two kinds of information, if they are to eventually return home: their compass bearing and the distances they travel. This becomes even more important on longer forays, like the 600 m foraging excursions of desert ants or the many kilometers bees often cover when collecting nectar or water. The compass cues that are available on earth dependent on the scale at which animals move, a distant tree may suffice as a beacon for journeys of a few meters, a distant mountain can serve to provide compass direction for excursions of a few hundred meters, but movements beyond such "suburban" distances require compass information that is reliable and does not change over large distances of travel. The earth's magnetic field direction offers one such invariant cue, the celestial bodies sun, moon and stars, by virtue of their comparatively infinite distance, the other. The problem with celestial compass cues is that they change position as the earth rotates, and in order to use them for journeys that last longer than a few minutes, their movements need to be taken into account.

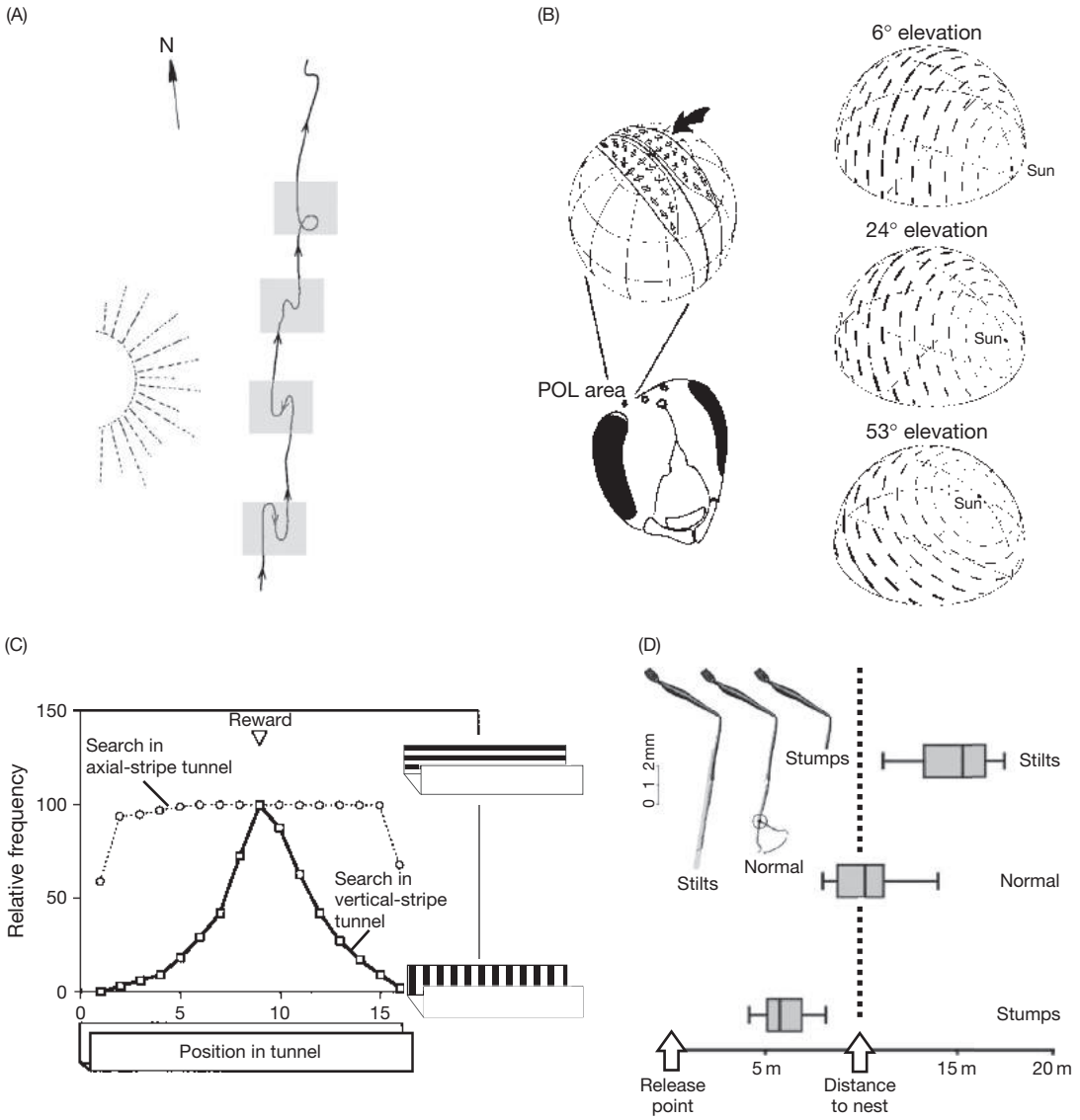


Fig. 2 Compass cues and odometry. (A) Insects are known to use the sun and the pattern of polarized skylight as a compass cue. *Left*: An experiment by Santschi (modified from Schöne, H. (1983). Orientierung im Raum. Stuttgart: Wissenschaftliche Verlagsbuchhandlung GmbH), in which he followed an ant on a straight path (line with *arrow* heads) and periodically screened off the sun on the left and reflected it back onto the animal with a mirror on the right during those parts of the ant's path that are marked with a gray patch. The ant always moves in such a way as to keep the sun to its left; making a u-turn whenever the apparent position of the sun has changed. (B) Many insects possess a specialized part of their eye, the dorsal rim, which carries photoreceptors that are particularly sensitive to the plane of polarization of light (*left*). Because the pattern of polarized skylight has an invariant structure with respect to the position of the sun, animals that are sensitive to the direction of polarization of light can infer the position of the sun from the direction of polarization of a few patches of blue sky: the sun must lie on the intersection of great circles perpendicular to the direction of polarization of these patches (modified from Wehner, R. (1982). Himmelsnavigation bei Insekten. Neurophysiologie und Verhalten. Neujahrsblatt der Naturforschenden Gesellschaft in Zürich **184**: 1–132). (C) Bees integrate optic flow to measure the distance they have to fly to a food source. Bees trained to find food about half-way into a small tunnel carrying pattern of vertical stripes at its walls, concentrate their search at a location where they had found the food source before. If the pattern in the tunnel is changed to horizontal stripes (which do not produce optic flow), the bees are unable to pinpoint a location in the tunnel and search throughout the tunnel (modified from Srinivasan, M.V., Zhang, S.W., Lehrer, M., Collett, T.S. (1996). Honeybee navigation en route to the goal: Visual flight control and odometry. *Journal of Experimental Biology* **199**, 237–244). (D) Desert ants (*Cataglyphis*) determine the distance they have walked by counting and measuring the size of their steps: when they walk on stilts on their home journey, they overestimate and when they walk on shortened legs, they underestimate home distance (modified from Wittlinger, M., Wehner, R. and Wolf, H. (2006). The ant odometer: Stepping on stilts and stumps. *Science* **312**, 1965–1967).

So, as far as compass cues are concerned, there are two kinds: cues that are static in relation to the world and those that are dynamic, but predictable. The static compass cues, in order of the scale over which they provide useful information, are landmark beacons, large oriented terrain features like edges of vegetation, mountain ranges, coastlines, and roads, and of course the polarity of the earth's magnetic field. The dynamic compass cues are all celestial: the sun, the moon, the stars, the pattern of polarized skylight and the distribution and spectral composition of skylight (Fig. 2A and B). Practically all mobile animals are known to use one or a

combination of these cues, when maintaining a straight course and those that derive compass information from the dynamic celestial cues do know about their movement across the sky. Bees, for instance, when communicating in the darkness of their hive to their nest mates the direction of a food source, can predict where the sun ought to be at that particular time of day and indicate the compass bearing of the food source relative to the sun.

When traveling, animals need in addition to a compass, also a way of measuring the distances they have moved. There is no direct way, in which this can be done, but animals could use a number of indirect measures of distance traveled: they could measure time, their energy consumption, they could count their steps, their fin- or wing-beats, or they could monitor the optic flow they experience, that is how much of the world has whizzed past their eyes. For insects, it is only in the last decade or so, that the sensory basis of odometry has been identified: bees integrate the image motion they experience during their foraging trips in order to judge (and communicate) the distance they have flown (Fig. 2C); while desert ants employ a step counter, which interestingly discards information on the three-dimensional topography of the ground they walk over, by only registering the horizontal component of the distance covered (Fig. 2D). Migrating birds, in turn, appear to come with instructions about how far to fly, that are based on the energy they would require to cover certain distances.

Animals that move over small and medium scales use these two navigational bits of information, compass direction and distance traveled, to determine a home vector, that is the direction and distance to the start of the journey (Fig. 3A). This process of path integration has been most thoroughly studied in desert ants, in bees, in crabs and in some small mammals, like hamsters and rats. Path integration thus seems to be a very old way of leaving an "Ariadne's Thread" behind throughout an excursion, that at the moment of return is being pulled taught and then defines a "beeline" home. The characteristic feature of path integration is that its result depends on the animal's own active movements. Passive drift or dislocations are not being registered. The navigational mechanism of path integration can therefore be identified by shifting an animal just before it heads home. If path integration is the only information an animal has about the location of its goal, like its nest, then after displacement it will move in the direction and the distance where it would find the goal, had it not been displaced (Fig. 3B). Path integration does not allow an animal to compensate for such passive displacements. Many details of the neural computations involved in this process are currently been unraveled, but one aspect of path integration has been well established: the process accumulates errors, because directions and distances cannot be measured with infinite accuracy, so that the information about the home position becomes less and less trustworthy, the longer the journey takes. There seem to be two solutions to this very basic problem. First, animals use visual, olfactory or even magnetic landmarks, which are the most robust way of pinpointing a goal. Landmarks in this context are any sensory cues that uniquely specify a location in space. Second, animals apply efficient search strategies to help them find the goal at the end of the home vector.

Position in Space

Probably one of the most fundamental abilities of animals is to know places. Many operate from nests and shelters, but also know and repeatedly visit other places of significance, especially places where they can find food and water. On a local scale, the most potent cue to the identity of a location in the world is its unique visual appearance. The fact that insects can utilize this property of the world in order to pinpoint their nest, was shown in the first half of last century in an elegant and simple series of experiments by Tinbergen and Kruyt (Fig. 3C): they displaced objects around the nest of a ground nesting wasp just a few tens of centimeters to the left or the right of the nest entrance, while the wasp was out foraging. The returning wasp searched not at its true location for her nest, but at a location defined by the displaced landmarks. These visual features, not the true geographical location, or olfactory cues, thus defined the goal location for this insect. Many experiments of a similar kind have been carried out since then, aimed at unraveling the detailed mechanisms underlying this ability of view-based homing in insects, but also in vertebrates. Homing insects behave as if they had stored snapshots of the world as viewed from the vicinity of the goal and upon returning then move to minimize the difference between these stored views and what they currently see. These stored views seem to contain the apparent sizes of salient objects, their shape and possibly also their color. However, insects also seem to use salient objects as beacons, not just as part of a panoramic scene and they can acquire information on the absolute distance between objects and the goal, not just on how large they ought to appear. View-based homing is so potent, because places in the natural world are uniquely defined by the view taken from them. The reason is that there are very few repetitive structures in nature that would make different places appear alike. On the other hand, the appearance of natural scenes change, because of changes in illumination and because objects are displaced by wind, water and by large animals. Ground-nesting insects, therefore, have to update their memory of the nest environment regularly, especially first thing every morning. To what extent our knowledge of view-based homing in insects can also explain the homing abilities of, for instance, birds which operate over distances of tens to hundreds of kilometers, is not entirely clear. It is only recently that the paths of birds at these scales can be accurately recorded. Like insects, birds visit the same food and nest sites over years—often after long migrations—and are attracted to and follow route landmarks. Both the pinpoint homing accuracy and the use of guiding landscape features, like roads, rivers or coastlines in the case of birds, are likely to be based on remembered views.

In barren landscapes, like featureless deserts, or the open and the deep ocean, view-based navigation fails. Yet, even there, animals are known to keep to idiosyncratic routes, to find tiny islands and to navigate on a global scale. One suggestion is that animals may have something akin to a global positioning system and the only cue they could access for that purpose would be the structure of the earth's magnetic field. By now, animals as diverse as bacteria, bees, lobsters, salamanders, fish, turtles and birds have

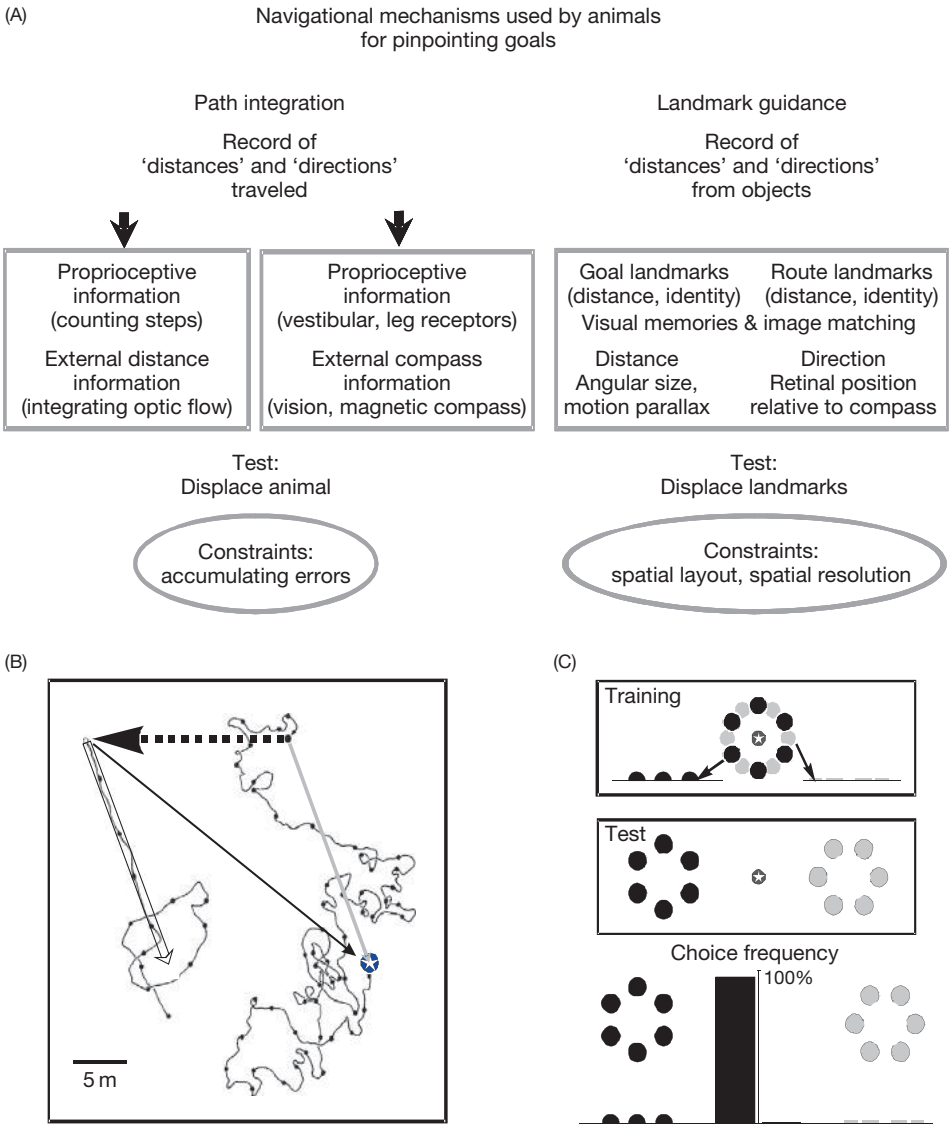


Fig. 3 Path integration and landmark guidance. (A) The two navigational mechanisms, the information required and examples of the cues animals are known to employ. (B) To test whether an animal employs path integration the animal is passively displaced before it heads home (*dashed arrow*). If the animal can compensate for the displacement (*thin black arrow*) it must have external (geocentric) information on its release site or on the nest location. An animal that relies exclusively on path integration information, like desert ants on featureless saltpans, runs into the direction and for the appropriate distance after displacement (*open arrow*) in which the nest would be, had it not been displaced (*gray arrow*). It then executes systematic search movements for the nest (see Fig. 8) (modified from Wehner, R. (1982). Himmelsnavigation bei Insekten. *Neurophysiologie und Verhalten. Neujahrsblatt der Naturforschenden Gesellschaft in Zürich* **184**: 1–132). (C) The demonstration that animals use the appearance of landmarks in order to pinpoint a location, like their nest, involves displacing the landmarks in the animal's absence and asking where it would search for the nest on its return. In many ground nesting insects, the nest location is defined by the surrounding landmark panorama, as this famous experiment by Tinbergen clearly demonstrates. The sand wasp *Philanthus* searches for her nest not at its true location (marked by a *star*), but in the centre of a ring of three-dimensional landmarks (*black*) rather than a ring of flat discs (*gray*) that both had surrounded the nest during training (modified from Collett, T.S., Zeil, J. (1997). Selection and use of landmarks by insects. In: Lehrer, M. (ed.) Orientation and communication in Arthropods, pp. 41–65. Basel: Birkhäuser Verlag).

been shown to be sensitive to at least some aspects of the magnetic field. Experiments with turtles in artificial magnetic fields that mimicked the situation in geographical locations that were hundreds of kilometers apart indicate that experienced animals actually recognized these locations by their magnetic field properties (although swimming on a tether in a backyard swimming pool) and moved in appropriate directions that would bring them back to their home grounds (Fig. 4). The magnetic field does not only vary along latitudes and longitudes across the earth, but also has local features, called magnetic anomalities, that could serve the same purpose as visual landmarks: something that uniquely specifies a place. Indeed, birds have been shown to change their behaviour in the vicinity of such "magnetic landmarks." Yet it is still unclear, whether any animal actually knows where it is at any point on earth

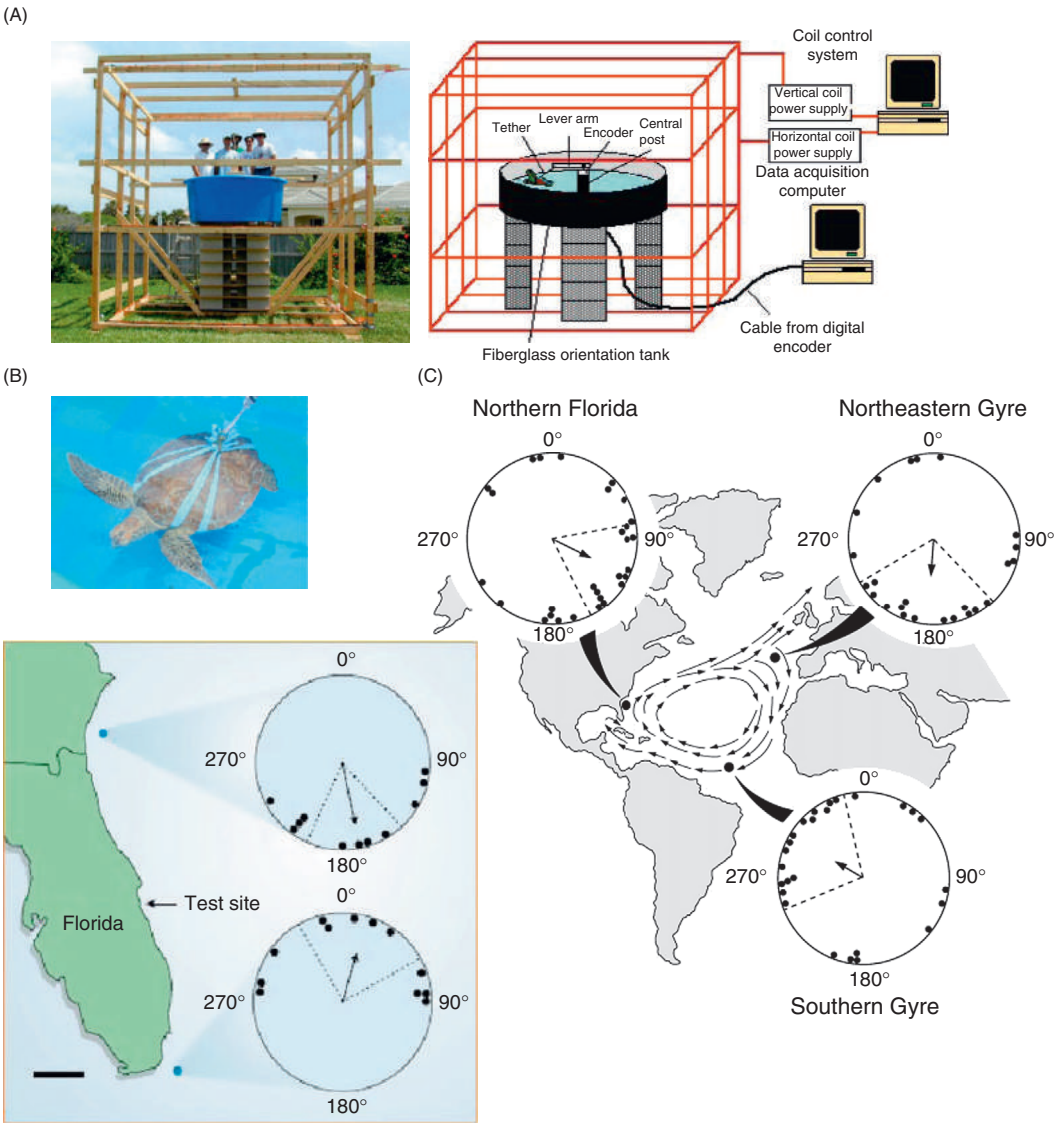


Fig. 4 Magnetic navigation. (A) The magnetic field properties of different locations on earth can be mimicked by sending appropriate currents through an arrangement of coils (from http:/www.unc.edu/depts/oceanweb/turtles, with kind permission of Kenneth Lohmann). (B) Adult turtles which are tethered in a large swimming pool and are subjected to a magnetic field characteristic of places several hundred kilometers north or south of their home ground (marked by stars), swim in directions that indicate that they perceive themselves to be north or south of their home grounds (from Lohmann, K.J., Lohmann, C.M.F., Ehrhart, L.M., Bagley, D.A., Swing, T. (2004). Geomagnetic map used in sea-turtle navigation. *Nature* **428**, 909–910). (C) Turtle hatchlings which are subjected in a similar way to magnetic fields that indicate different locations across the Atlantic ocean swim in directions that would keep them in the large Atlantic gyrus (from Lohmann, K.J., Cain, S.D., Dodge, S.A., Lohmann, C.M.F. (2001). Regional magnetic fields as navigational markers for sea turtles. *Science* **294**, 364–366).

by the particular property of the magnetic field that specifies that place, something that has been called the ability of "true navigation." The best evidence so far has been the behaviour of turtles and lobsters in artificially created magnetic fields that mimic the situation in specific places on earth (Fig. 4B and C). It is only very recently that the tools have become available that allow us to track animal movements over global scales. Knowing exactly where animals go, how they move and where they make navigational decisions is a prerequisite for understanding the knowledge base of their navigational abilities (Fig. 5).

Navigation: Compasses, Landmarks and Maps

Animals navigate across very different scales and the knowledge they require, the cues they have available and the mechanisms they employ differ depending on the task, the lifestyle, the habitat, the sensory limitations and the style of locomotion. Animals can also be expected to be very sensitive to the changing salience and reliability of navigational cues and adjust their reliance on these cues accordingly. Homing pigeons, for instance, follow roads and are attracted by large landscape features, like road intersections or

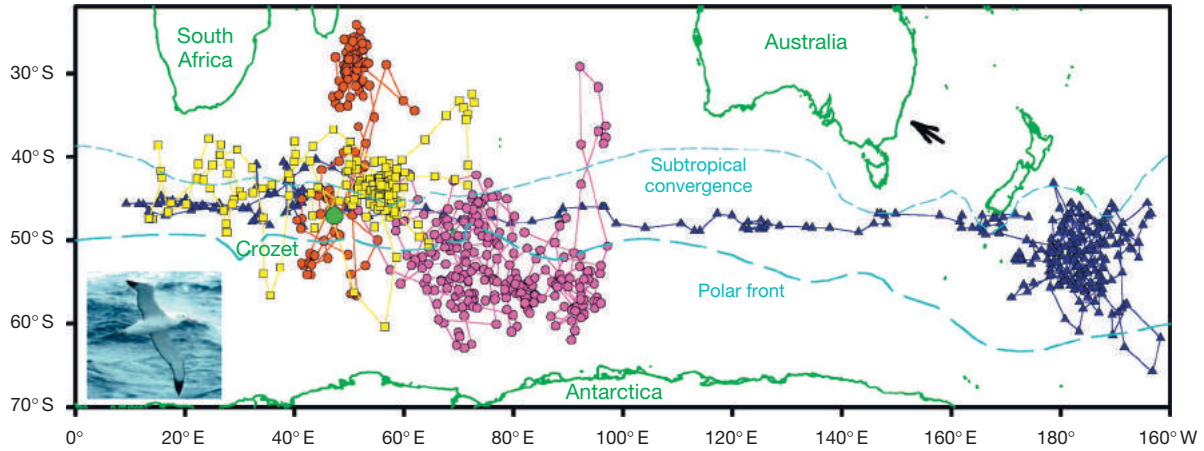


Fig. 5 Global navigation. Wandering albatrosses go on a year's sabbatical after breeding on Crozet Island. Two positions per 24 h cycle are shown for two male (*mauve dots* and *blue triangles*) and two female birds (*red dots* and *yellow squares*) as determined by light-intensity loggers. Each bird flies to its favorite wintering zones somewhere across the southern ocean (from Weimerskirch, H. and Wilson, R.P. (2000). Oceanic respite for wandering albatrosses. *Nature* **406**, 955–956).

bridges (Fig. 6A). Many insects use the sun as a compass on brief foraging trips, but switch to relying on the pattern of polarized sky light, or on landscape features, when the sun is not visible. Equally, some birds calibrate their magnetic compass using the twilight sky at the horizon; other birds calibrate their twilight and star compass with the aid of the magnetic field. Compass cue hierarchies thus differ in each specific case, possibly depending on the way experiments are being conducted, but also depending on the evolutionary history and the peculiarities of migratory behaviour.

The same flexible use of multiple navigational cues is also evident in the control of small- and medium-range movements by insects, in which information derived from path integration, from the visual scene and from olfaction interact. Honeybees, for instance, learn, remember and recall navigational instructions, when they encounter certain patterns, colors, or odors along a foraging route. Desert ants (*Cataglyphis*) can retrace their steps in complicated terrain, always detouring vegetation or artificial landmarks in the same direction on repeated trips (Fig. 6B). Wood ants (*Formica rufa*) are attracted by dominant landmarks as beacons and route guides, but can follow the same route even in the absence of such a landmark, indicating that they learned as well a sequence of other cues that defines their foraging path. The foragers of the Giant tropical ant (*Paraponera clavata*) ignore their own pheromone trails when they become more experienced with a route to a food source and rely themselves on route landmarks, while fresh recruits are guided by the trail.

These examples highlight several important aspects of animal navigation that are little or only partly understood: the organization of navigational memory, the saliency and reliability of navigational cues and the relative importance of genetically determined navigational instructions and of navigational learning. Large-scale migration routes in birds have been shown to be determined by genes that influence migratory direction, while the ability to pinpoint nest and food locations, has to involve learning and memory mechanisms. The close similarity of the complex behavioral organization of learning flights in phylogenetically distant species of nest-owning bees and wasps, however, suggests that the rules of how to learn about the visual scene around a goal have a genetic basis.

In recent years, a lively debate was concerned with the question, what kind of internal representation animals have of the different environments they navigate through. Is the knowledge they have similar to how we perceive the lay of the land, having been trained to read and interpret maps, which are basically bird eye views of the world? Or do animals know the world only as a set of paths and places they have traveled before, but without the ability to determine the shortest paths between these places? The debate revolves around the concept of a "cognitive map," the organization of the internal representation of space. In its hard definition, a cognitive map would mean that animals have a representation of the x–y–z coordinates of the landscape, much like a topographic map. In its softer definition, a cognitive map is any representation of the physical environment that allows an animal to navigate through it and to visit places of interest. One of the crucial experiments that were designed to test whether animals possess one or the other representation, involves training bees, for instance, to visit two locations that are some distance away from each other and then displacing them to a location half-way between the two training places, which is assumed to be unknown to them. The question then is whether the bees fly directly to one of the known places, which would indicate that they were able to judge, where the release location lies with respect to the training places. Such experiments and variants of it have so far not provided clear evidence, at least for bees, that they possess a topographic map-like representation of their environment. Similar experiments with dogs show that these animals do take short cuts between two places they know, possibly using path integration information. In contrast, insects seem to represent the environment they exploit as a set of instructions on how to reach and trap-line significant places from their nests and hives, but with no ability to take short-cuts between them. This question of how navigational knowledge in different animals is organized is still open and provides an inspiring and motivating challenge to scientists interested in the mechanisms of animal behaviour.

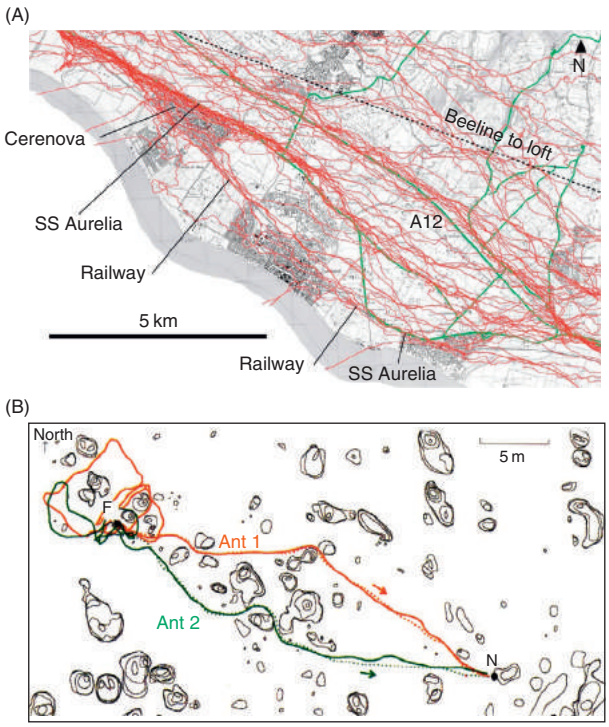


Fig. 6 Landmark guidance. (A) Homing pigeons are attracted to and are guided by large landmark features, like roads or large road crossings. Red lines show the GPS recorded paths of pigeons on their way back to the loft; *green lines* mark major highways (from Lipp, H.-P., Vyssotski, A.L., Wolfer, D.P., Renaudineau, S., Savini, M., Tröster, G. and Dell'Omo, G. (2004). Pigeon homing along highways and exits. *Current Biology* **14**, 1239–1249). (B) In terrain with vegetation, two desert ants (*Cataglyphis; green* and *red*) had returned from a feeder at (F) to their nest at (N) along the dotted paths. They were then picked up and returned to the feeder. After some search both ants practically retraced their steps (continuous *red* and *green lines*). Since *Cataglyphis* does not deposit pheromone trails, this experiment shows that they must remember route landmarks (from Wehner, R., Michel, B., Antonson, P. (1996). Visual navigation in insects: Coupling of egocentric and geocentric information. *Journal of Experimental Biology* **199**, 129–140).

Orientation by Olfaction: Trails, Flows and Plumes

Many animals are guided by olfaction. They are able to detect the gradient of molecule concentration to find the source that emits these molecules. However, homogeneous gradients are rare on earth, because molecules are transported by the media of air and water and therefore do not disperse in a homogeneous fashion. Rather, they are subject to diffusion and turbulence, processes that cause them to be very unequally and patchily distributed after a short distance of travel from the source. Animals have adapted to this pattern of molecule dispersal on earth and have acquired optimal strategies to detect the source of such turbulent plumes. The best way of doing so requires the following simple rules (Fig. 7): first, if a substance is detected, its most likely origin is from upwind or upstream. So first of all, move against the flow. Second, if you fail to detect it again, move across the flow direction in a zigzag fashion. Third, the moment the substance is detected again, move upstream or upwind. A recent theoretical study has shown that this is actually the only strategy that assures that an animal finds the source of a substance in turbulent media. Pinpointing the source of an odor is important for animals in a variety of tasks: to find food, like carcasses and flowers, to find mates, in which case very specific pheromones are being produced, and released into the medium, to detect predators and to find places that carry a specific olfactory signature, like a nest marked with pheromones or a river inlet. Some animals use odors to mark homes, territorial boundaries or in the case of ants, pheromone trails to a food source that serve to recruit and guide nest mates to such a significant place. It has even been suggested that locally varying blends of odors can be used like a map by animals, helping them to link places they visit by their distinct olfactory signatures.

Odors in the context of orientation and navigation have some interesting and unique properties: in the case of pheromone markers or trails, for instance, they carry information, not only about the sender, but also about time. Since molecules diffuse, depending on their volatility, their ambient concentration will decline with time and thus indicate the time elapsed after they were deposited. Olfactory trails and signposts are signals that do not require the sender to be present and they can be made very specific, compared to visual or sound signals. Chemical signals thus can help to avoid un-intended eavesdroppers. The problem with odors is that they are so affected by the movement characteristics of the medium. So, marking a nest entrance, for instance, will only help in homing, if returning animals are downwind from it. There is no signal to be detected from all other directions. For the same reason, it is also unclear at present—with the exception of rivers—how stable the distribution of natural odors are in the atmosphere

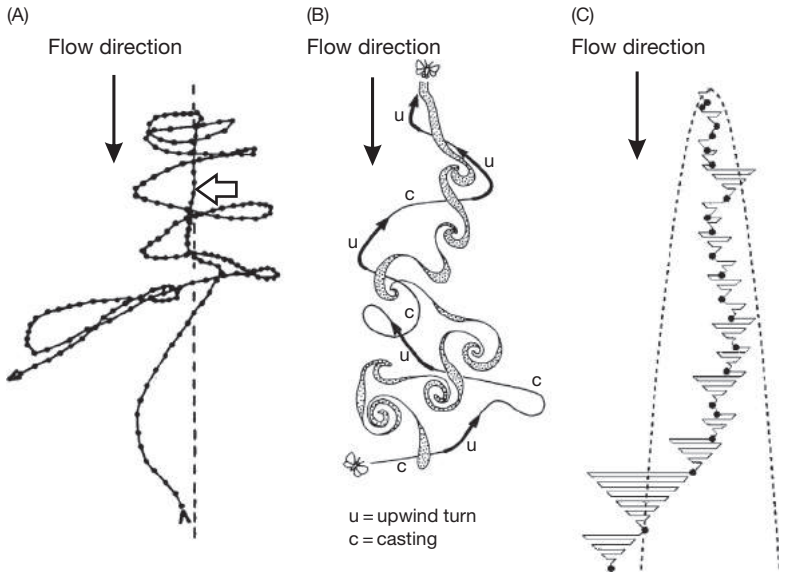


Fig. 7 Orientation in turbulent pheromone plumes. (A) A flight path of a moth silk moth male in a pheromone plume. The *open arrow* head marks the time when the pheromone release was switched off. Shortly afterwards, the insects starts casting, flying perpendicular to the wind direction. Moth positions are shown every 67 ms (modified from Baker, T.C. and Vogt, R.G. (1988). Measured behavioral latency in response to sex-pheromone loss in the large silk moth *Antheraea polyphemus*. *Journal of Experimental Biology* **137**, 29–38). (B) Principle of finding the source of a turbulent plume: if a molecule is encountered, move upwind or upstream (u); if no more molecules are encountered, move cross-wind or cross-stream (c) (modified from Kaissling, K.-E. and Kramer, E. (1990). Sensory basis of pheromone-mediated orientation in moths. *Verhandlungen der Deutschen Zoologischen Gesellschaft* **83**, 109–131). (C) The path generated by an optimal plume search algorithm. *Dashed line* encloses area of highest probability to encounter a patch of odor. *Dots mark* odor patch encounters (modified from Balkovsky, E., Shraiman, B.I. (2002). Olfactory search at high Reynolds number. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 12589–12,593).

and in water, which is crucial for their utility as navigational aids, as directional compass cues and as cues for localization. This gap in our knowledge is mainly a technical challenge at the moment, because monitoring biologically relevant chemical cues over long periods of time and with sufficient spatial and temporal accuracy is currently impossible to achieve. Being able to monitor, track, map and modify olfactory information would be truly exciting, because many experiments with homing pigeons seem to suggest that they attend to olfactory information in one way or other when being released tens of kilometers away from their loft. These observations lead to the suggestion that birds may possess an olfactory map, knowing places and their relative locations by their unique olfactory signatures.

Search

There are limits to the navigational abilities of animals because they face sensory and computational constraints to what they can know and remember. In many situations animals have to search for a goal or for food. Faced with these limitations and with the unpredictable aspects of the world, animals have evolved very efficient search strategies. Depending on the type, location and distribution of targets, these strategies fall broadly into two categories: systematic and random search. Systematic search movements are being employed in situations, in which animals have some information on the location of a target, for instance, when ants have been following path integration information on their return to nest (Fig. 8A). When they do not find the nest entrance at the end of their home vector, ants begin searching for it by running along paths that describe ever increasing loops centred on the location where the search began. This systematic search pattern is driven by the fact that the most likely location of the nest from the perspective of the returning ant is at the end of its home vector. Given that this home vector is associated with some uncertainty as to the direction and to the distance at which the nest is to be found, the probability of finding the nest can be assumed to be distributed like a three-dimensional Gaussian around the end-point of the home vector (Fig. 8B). The probability would be highest close to that point and fall off with distance from it. At the beginning, then, search should be concentrated at that location, a process that decreases the probability that the nest is there and consequently increases the probability that it is located further away from the start-location of the search. As search loops increase, the probability of finding the nest further out decreases and the probability at the start-location becomes relatively higher again. The ant should thus repeatedly revisit this central location. An alternative to this strategy would be to search along a path spiraling out from the estimated nest location. However, if the perceptual horizon of an ant is limited, she is in danger of missing the nest and continuing on a spiral path would not allow her to correct that mistake. Interestingly, in ants which run off only part of their home vector, the subsequent search is not centred on the beginning of search,

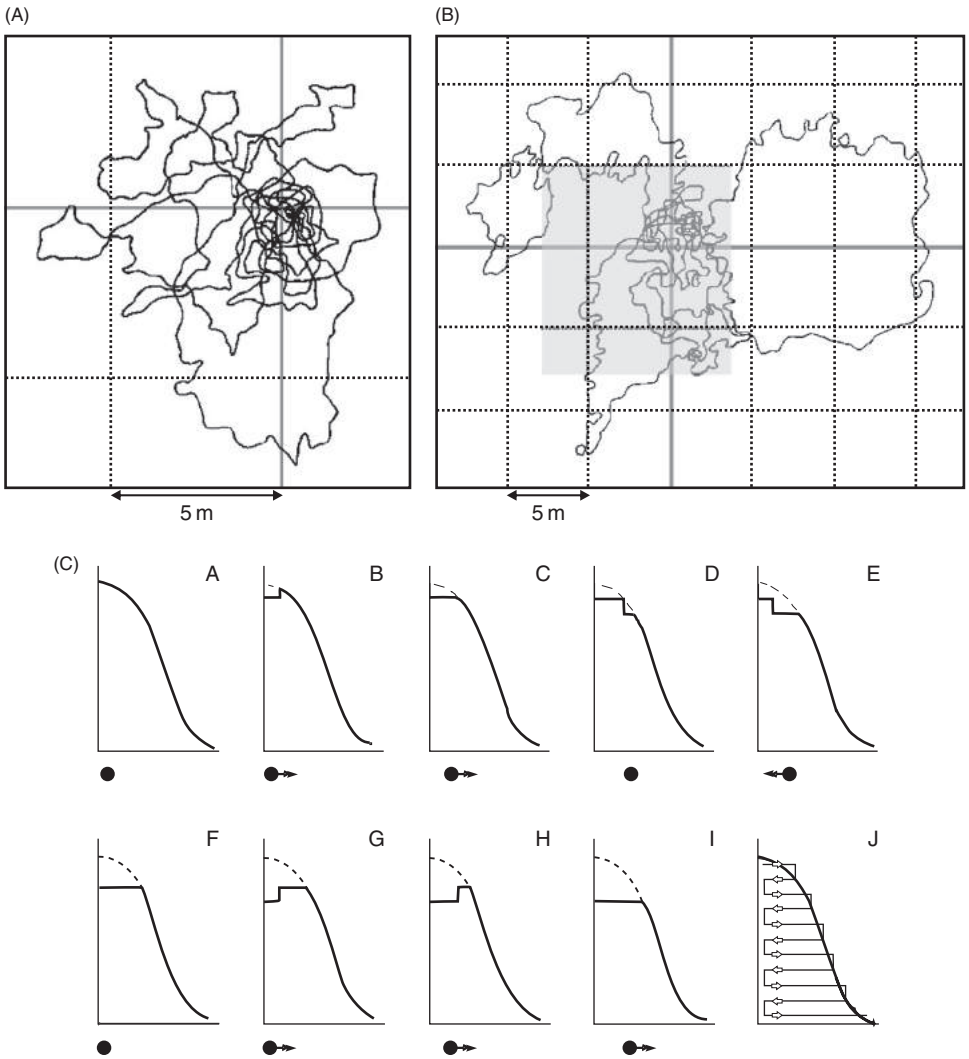


Fig. 8 Systematic search in desert ants. (A) and (B) The one hour long search path of a desert ant that had been displaced before running off its home vector (see Fig. 2B). The initial half hour path is shown in (A) and the subsequent path in (B). Note the different scales. The area shown in (A) is marked *gray* in (B). The ant searches in ever increasing loops, centred on the expected nest position at the intersection of the thick *gray* lines. (C) This systematic search pattern can be explained by considering that it is driven by the instantaneous probability distribution of finding the nest at the end of the home vector. The diagram shows the right part of a transect through a three-dimensional Gaussian distribution. Search is first concentrated at the centre of the distribution close to the end of the home vector (A) and consequently, the probability of finding the nest decreases there which causes more distant parts of the distribution to become relatively higher (B). If search loops lead the searching agent further away from the centre of the distribution, the probability of finding the nest decreases at these more peripheral parts of the distribution (D and E), leading to a relative increase of probability at the centre of the distribution, which attracts the searching agent back. Modified from Wehner, R. and Srinivasan, M.V. (1981). Searching behaviour of desert ants, genus *Cataglyphis* (Formicidae, Hymenoptera). *Journal of Comparative Physiology* **142**, 315–338.

but leads further and further away in the direction of the home vector. Such biased search movements are thus still informed by the direction of the home vector and reflect the probability of finding the nest in this direction, but with ever increasing uncertainty about both the distance and the direction at which it could be found relative to the beginning of search.

In contrast to situations—like the one described above—where animals have some information on the location of targets, many foraging situations are different: prey or food items are often randomly distributed and their location may be hard to predict. In such a case, random, rather than systematic, search strategies are being employed and have indeed been shown to be optimal for detecting randomly and sparsely distributed targets. The efficiency of random search movements is optimal, when the probability distribution of straight path lengths (between changes in movement direction) follows an inverse square power-law distribution. The foraging movements of so diverse animals as amoebas, bees, deer and wandering albatrosses, have all the characteristics of such optimized random search movements. They can also help animals to find favorable conditions—like places with preferred temperature or moisture—that are extended in space. In such cases, random search movements are modified by environmental cues, which for instance, trigger a decrease in the speed of movement and an increase in the size of turns whenever conditions become favorable. Both behavioral modifications tend to keep an animal within the range of favorable conditions.

The question of how animals orient in the world and what cues they have available to do so in an organized and systematic fashion can thus only be answered by understanding the specific ecology of information processing in each particular case. There are some universal constraints, like the ones imposed on vision, by its closed-feedback nature, and like the ones that are imposed by the physical properties of the world, like in the case of cues providing robust compass information, but how much a given animal is confronted with such constraints depends on the particular habitat of an animal, its style of locomotion, its active space and the tasks it has to solve. For most animals, we still know surprisingly little about these crucial aspects of their lives.

Further Reading

Alerstam T (2006) Conflicting evidence about long-distance animal navigation. *Science* 313: 791–794.

Borst A and Haag J (2007) Optic flow processing in the cockpit of the fly. In: North G and Greenspan RJ (eds.) *Invertebrate neurobiology*, Cold Spring Harbor, NY: CSHL-Press.

Cardé RT and Willis MA (2008) Navigational strategies used by insects to find distant, wind-borne sources of odor. *Journal of Chemical Ecology* 34: 854–866.

Chapuis N and Varlet C (1987) Short cuts by dogs in natural surroundings. *Quarterly Journal of Experimental Psychology* 39B: 49–64.

Collett M, Chittka L, and Collett TS (2013) Spatial memory in insect navigation. *Current Biology* 23: R789–R800.

Dacke M, Baird E, Byrne M, Scholtz CH, and Warrant EJ (2013) Dung beetles use the milky way for orientation. *Current Biology* 23: 298–300.

Gagliardo A (2013) Forty years of olfactory navigation in birds. *Journal of Experimental Biology* 216: 2165–2171.

Hardcastle BJ and Krapp HG (2016) Evolution of biological image stabilization. *Current Biology* 26: R1010–R1021.

Heinze S and Homberg U (2007) Maplike representation of celestial E-vector orientations in the brain of an insect. *Science* 315: 995–997.

Helbig AJ (1996) Genetic basis, mode of inheritance and evolutionary changes of migratory directions in Palearctic warblers (Aves: Sylviidae). *Journal of Experimental Biology* 199: 49–55.

Hengstenberg R (1993) Multisensory control in insect oculomotor systems. In: Miles FA and Wallman J (eds.) *Visual motion and its role in the stabilization of gaze*, pp. 285–298. Amsterdam: Elsevier Science.

Land MF (1999) Motion and vision: Why animals move their eyes. *Journal of Comparative Physiology* A185: 341–352.

Luschi P (2013) Long-distance animal migrations in the oceanic environment: Orientation and navigation correlates. *ISRN Zoology* 2013: 631839.

Menzel R and Greggers U (2015) The memory structure of navigation in honeybees. *Journal of Comparative Physiology* A201: 547–561.

Nalbach H-O (1990) Multisensory control of eyestalk orientation in decapod crustaceans: An ecological approach. *Journal of Crustacean Biology* 10: 382–399.

Pyke GH (2015) Understanding movements of organisms: It's time to abandon the Lévy foraging hypothesis. *Methods in Ecology and Evolution* 6: 1–16.

Vickerstaff R and Cheung A (2010) Which coordinate system for modelling path integration. *Journal of Theoretical Biology* 263: 242–261.

Webb B and Wystrach A (2016) Neural mechanisms of insect navigation. *Current Opinion in Insect Science* 15: 27–39.

Wehner R and Labhart T (2007) Polarization vision. In: Warrant E and Nilsson D-E (eds.) *Invertebrate Vision*, pp. 291–348. Cambridge: Cambridge University Press.

Zeil J (2012) Visual homing—An insect perspective. *Current Opinion in Neurobiology* 22: 285–293.

Parental Care

Per T Smiseth, University of Edinburgh, Edinburgh, United Kingdom

Glossary

Conflict resolution The outcome of familial conflict given how different family members influence the allocation of resources from parents to offspring.
Direct fitness The impact of an individual's actions on its own personal fitness. The benefits of parental care are assigned to the offspring's direct fitness whilst costs are assigned to the parent's direct fitness.
Forms of care Variation in how parents care for their offspring, ranging from the allocation of additional nutrients into the eggs to care for their offspring after they have become adults.
Indirect fitness The impact of an individual's actions on the fitness of a relative weighted by the coefficient of relatedness between them. Thus, the indirect benefit of care to the parent is found by multiplying the offspring's direct fitness by the coefficient of relatedness (usually 0.5), whilst the indirect cost of care to the offspring is found by multiplying the parent's direct fitness cost by the coefficient of relatedness.
Interference competition The negative effect that the density of competitors has on the amount of resources obtained by a focal individual.
Parental care Any parental trait that enhances the fitness of a parent's offspring and that seems likely to have originated and/or to be currently maintained for this function.
Patterns of care Variation in which parent provides care for the offspring, ranging from uniparental female and uniparental male care to biparental care.
Scramble competition Competition for access to shares of resources in a situation where the resource is accessible to all competitors.

Introduction

In virtually all birds and mammals, parents provide highly elaborate forms of care that enhance the offspring's fitness by neutralizing environmental hazards that otherwise are detrimental to offspring, including predation, cannibalistism by conspecifics, starvation, desiccation, extreme temperatures, and anoxia. For example, most songbirds build nests that conceal offspring from predators and protect them from extreme temperatures, place the nest in a safe habitat, incubate the eggs to supply a source of heat for the developing embryos, provision food to the nestlings after hatching, ensure nest sanitation to reduce the parasite burden, and ward off potential nest predators. Likewise, marsupial and eutherian mammals first nourish the developing embryo via a placenta, and then protect the young from predators and supply it with food in the form of milk after birth. In taxa outside the birds and mammals, parents often provide no care beyond supplying eggs with a small amount of nutrients. Yet, some reptiles, amphibians, fishes and invertebrates have evolved forms of post-fertilization care that match those of birds and mammals in terms of complexity. For example, females of the bromeliad crab *Metopaulias depressus* remove leaf litter and deposit snail shells into bromeliad pools to neutralize the low pH levels and boost the levels of calcium carbonate, defend the larvae from predators, and provision their larvae with food in the form of snails. Other species have evolved relatively simple forms of care, such as the treehopper *Publilia concava*, where females attend the eggs until hatching, thereby protecting them from ants and other egg predators.

The study of parental care and its evolution is a central topic in behavioral ecology (Clutton-Brock, 1991; Royle et al., 2012). As illustrated by the above examples, parental care is extremely diverse, varying both between and within species with respect to its form, magnitude and duration, as well as the extent to which it is provided by the female, the male, or both parents. Thus, an important aim in behavioral ecology is to understand the evolutionary origin and subsequent maintenance of this diversity in light of the benefits to the offspring and the costs to the parents. The costs and benefits of care may depend upon factors such as ecological conditions, life histories and mating systems (Clutton-Brock, 1991). Behavioral ecologists also take an interest in parental care because it often is associated with the emergence of family groups composed of one or two parents caring for one or several offspring. An important aim in behavioral ecology is to understand the mechanisms that mediate the resolution of conflict between family members over the allocation of parental care (Godfray, 1995; Mock and Parker, 1997).

Understanding the conditions promoting diversity in parental care is also of interest to behavioral ecologists because its evolution is closely linked with that of other important traits. For example, the emerging interest in parental care from the 1970s onwards was stimulated by Trivers's (1972) seminal paper in which he argued that the greater involvement in parental care by females makes them a limited resource to males, thereby leading to intense sexual selection in males. Although recent work has questioned this argument, the link between parental care and sexual selection remains an important topic in behavioral ecology (Kokko and Jennions, 2008). Furthermore, parental care plays an essential role in social evolution by providing a stepping stone from an ancestral solitary condition towards the advanced forms of eusociality seen in social insects, such as ants and bees. Thus, the study of parental care is fundamental to our understanding of the evolution of reproductive behaviors and complex sociality.

In this chapter, I start by providing a brief overview of the key terms used in the study of parental care and its evolution. I then describe diversity in how parents provide care for their offspring, and diversity in the extent to which female and male parents are involved in parental care for the offspring. Finally, I provide an overview of different forms of conflicts between family members over parental care and the mechanisms mediating the resolution of these conflicts.

What Is Parental Care?

Like any other field in behavioral ecology, the study of parental care and its evolution requires a clear terminology that facilitates comparisons between studies and translation between theoretical and empirical work. Unfortunately, the terminology in this field is described as 'diffuse and misleading' (Clutton-Brock, 1991), involving similar terms with distinct definitions in the literature. Furthermore, many terms have alternative definitions, and to make matters worse, the same definitions are sometimes applied to different terms (Smiseth et al., 2012). It is beyond the scope of this chapter to discuss the terminology in any great detail (for further details, see Clutton-Brock, 1991; Smiseth et al., 2012). Nevertheless, any student of parental care should be aware of the potential confusion caused by inconsistent use of terminology, and this section provides a brief overview of the six basic terms used in this field (Table 1).

The main distinction is between purely descriptive terms and terms used for the costs of care to parents. The use of three descriptive terms seems excessive but is a legacy of the history of the field. The terms *parental care* and *parental behavior* derive from behavioral ecology, and represent what Clutton-Brock (1991) referred to as 'broad-sense' and 'narrow-sense' definitions of parental care, respectively. The term *parental care* corresponds to the broad-sense definition and can be defined as any parental trait that enhances the fitness of a parent's offspring and that is likely to have evolved for this function (Table 1). Meanwhile, the term *parental behavior* corresponds to Clutton-Brock's narrow-sense definition that focuses specifically on behavioral traits. The term *parental effect* (or more specifically *maternal* or *paternal effect*) is defined as the causal effect of the parents phenotype on the offspring's phenotype over and above the effects of genes inherited from the parents. This term derives from quantitative genetics but is used increasingly by behavioral ecologists, often to describe prenatal influences of parents on their offspring. Adaptive parental effects can be divided

Table 1 Definition of different terms used in the study of parental care and its evolution (from Smiseth et al., 2012, with permission from Oxford University Press)

Term	*Definition*	*Currency*
Parental care	Any parental trait that enhances the fitness of a parent's offspring and that seems likely to have originated and/or to be currently maintained for this function	None (descriptive)
Parental behavior	Any parental behavior that seems likely to enhance the fitness of a parent's offspring	None (descriptive)
Parental effect (or maternal effect)	The causal effect of the parent's phenotype on the offspring's phenotype over and above direct effects due to genes inherited from the parent	None (descriptive)
Parental expenditure	Any expenditure of parental resources (including time and energy) on parental care of one or more offspring	Time and energy; benefits to offspring and costs to parents
Parental investment	Any investment by the parent in an individual offspring that increases the offspring's survival and reproductive success at the cost to the parent's ability to invest in other current or future offspring	Fitness; costs due to enhancing fitness of an individual offspring
Parent effort	The combined fitness costs that the parent incurs due to the production and care of all offspring in a given biologically relevant period, such as a breeding attempt	Fitness; combined costs due to production and care of offspring

between those that enhance the offspring's fitness and those that enhance the parent's fitness. Thus, the term parental effect is even more general than the term parental care as it includes both adaptive and non-adaptive parental effects and parental effects that may enhance or reduce offspring fitness.

The three remaining terms listed in Table 1 focus on the costs of care to parents and these differ with respect to the currency that is used to measure these costs. The term *parental expenditure* is used mainly by empiricists and quantifies costs in terms of the amount of time or energy spent on providing care. This term is also relevant for measuring the benefits of care to the offspring, which depend on the absolute amount of care provided by the parents. In contrast, *parental investment* and *parental effort* are important theoretical terms used in mathematical models. These two terms measure the fitness costs of care and differ with respect to whether these costs are measured in terms of enhancing the fitness of an individual offspring (parental investment) or producing and caring for all offspring in a given reproductive episode (parental effort). For example, models for how parents respond to their offspring's begging behavior are based on the term parental investment (Godfray, 1991). Unfortunately, it is often extremely difficult to measure fitness costs of parental care in the field because this requires information on levels of parental care and the future survival and reproduction of individual parents and their offspring.

There is an important distinction between energy and time costs and fitness costs. Given that it is challenging to estimate fitness costs in the field, it might be tempting for empiricists to use measures of parental expenditure as a proxy for parental investment or parental effort. Unfortunately, this practice is problematic because parental expenditure is measured in absolute terms (the amount of energy or time), while parental investment or parental effort should be measured in relative terms (the proportion of the total resource budget that is allocated to parental care). Given that the total resource budget is usually unknown for any species, it is not possible to estimate relative measures of costs from information of absolute measures of parental expenditure. This issue is particularly problematic when individuals differ with respect to parental quality or provide care under different ecological conditions. When this is the case, good quality parents, or parents providing care under benign conditions, may provide more care in absolute terms (i.e., expend more energy or time), yet pay a smaller cost in relative terms (i.e., allocate a smaller proportion of the total resource budget to care). In summary, readers are advised to be aware of the potential issues arising from the inconsistent use of terminology in this field.

Forms of Care

There is a great amount of diversity across animals with respect to how parents care for their offspring (Smiseth et al., 2012). The most basic, and presumably most widespread, form of care is for females to allocate additional nutrients into the eggs beyond what is required for successful fertilization. For example, females of the seed beetle *Stator limbatus* adjust egg size to match the environmental conditions as a means to enhance the larvae's survival after hatching. At the other extreme end, parents of some species provide care even after their offspring have reached adulthood. This form of care is extremely rare, but occurs in bonobos, where females support their adult male offspring during competitive interactions with rival males, thereby enhancing their son's social status and mating success. Between these two extremes, there are many other ways by which parents may enhance offspring fitness (Table 2). The precise ways by which parents enhance offspring fitness are often taxon-specific, and different schemes are often used to describe diversity in parental care in different taxa. Table 2 provides a general description of 11 basic forms of care observed across different animal taxa that are arranged in chronological order throughout the offspring's development (Smiseth et al., 2012).

Little is known about why different taxa or species have evolved specific forms of care. Potentially, this might reflect differences in the ecological conditions of a particular species and the specific hazards faced by the offspring. For example, selection might favor different forms of parental care when offspring are at a high of risk of predation versus a high risk of starvation. Furthermore, the evolution of different forms of care in different species may also reflect differences in life histories and the presence of pre-existing traits that selection can modify into parental care (Smiseth et al., 2012). For example, the risk of predation may only favor parental care if parents are capable of defending their offspring against predators and may originate from behaviors used by males or females to guard a territory or a partner. In some species, including most birds and mammals as well as some amphibians, fishes, arthropods and other invertebrates, parents provide elaborate and complex forms of care comprised of multiple components (see examples above). Relatively little is known about how such elaborate forms of care originate, but presumably they arise from relatively simple ancestral forms of care. For example, once egg attendance has evolved, it might evolve into offspring attendance by simply delaying the time of parental desertion, and additional components of care, such as food provisioning, might evolve once offspring attendance has evolved.

Patterns of Care

The extent to which females and males contribute towards parental care after laying or birth varies both between and within higher animal taxa (Maynard Smith, 1977; Kokko and Jennions, 2008). As a general rule, females tend to be more involved in care than males. For example, in the vast majority of mammals, females provide food in the form of milk and defend the young against predators after birth, while males make no contribution whatsoever. Likewise, in those arthropods that have some form of care for eggs or young, it is usually only the female that provides care. However, there are exceptions to this pattern. First, male-only care for

Table 2 Overview of different forms of parental care with examples (for further details, see Smiseth et al., 2012)

Form of care	*Description*	*Examples*
Provisioning of gametes	Deposition of energy and nutrients into eggs beyond the minimum required for successful fertilization. May include male nuptial gifts	Egg size in arthropods and birds enhances offspring growth and survival
Oviposition-site selection	Non-random choice of egg-laying site, including nest-site selection in nest-building animals. Excluding cases where oviposition-site selection increases female's own fitness	In the mosquito *Culiseta longiareolata*, females avoid ovipositing eggs in pool that contain larval predators
Nest building and burrowing	Concealment of eggs beneath substrate, building nest or burrowing in soil or excavating in wood. Nests may be built from materials found in environment (e.g., mud, plant material), processed plant material (e.g., paper) or materials produced parents (e.g., silk, mucus)	Many birds have a complex nest architecture where the outer layer conceals nest from predators and protects against wind and rain, while the inner layer insulates against extreme temperatures
Egg attendance	Remaining with the eggs at a fixed location, usually the oviposition site. Often associated with behaviors directed towards specific threats, such as egg guarding (directed at predators or oophagic conspecifics), egg fanning (to prevent hypoxia), egg cleaning (directed at fungal pathogens)	In the harvestman *Iporangaia pustulosa*, male egg attendance increases egg survival
Egg brooding	Carrying eggs after laying either externally (e.g., parent's back) or internally (e.g., specialized pouches or parent's mouth)	Male giant waterbugs carry eggs on their backs. In marsupial frogs of the genus *Gastrotheca*, females carry eggs within specialized brood pouches
Ovoviviparity and viviparity	Retention of eggs with the female reproductive tract, either (ovoviviparity) or until after hatching (viviparity). Viviparity may be associated with matrotrophy, where the embryo is provisioned with nutrients from sources other than yolk	Viviparity has evolved repeatedly in several invertebrate phyla as well as in fishes and squamate reptiles
Offspring attendance	Remaining with the young after hatching at a fixed location or escorting the offspring as they move around. Often associated with behaviors directed towards specific threats such as guarding (predators), and antimicrobial secretions (microbial pathogens)	The benefit of offspring attendance is demonstrated in a parental removal experiment in the lace bug *Gargaphia solani*
Offspring brooding	Carrying offspring after hatching or birth either externally (e.g., on the parent's back) or internally (e.g., in specialized pouches or parent's mouth)	Brooding of offspring in the female's stomach in gastric-brooding frogs
Food provisioning	Provisioning of food after hatching or birth. Mass provisioning occurs where food is provided before hatching, while progressive provisioning occurs where parents repeatedly provide food after hatching or birth. Progressive provisioning may be based on food found in the environment (e.g., arthropods) or specialized food produced by parents (e.g., milk, trophic eggs). Most extreme version is matriphagy, where hatched offspring consume their parent	In the caecilian *Boulengerula taitanus*, larvae feed on modified skin produced by female. In the freshwater leech *Helebdella stagnalis*, females provision food for their young
Care after nutritional independence	Assisting offspring after nutritional independence	In American red squirrels, offspring may inherit a cache of cones from their female parent
Care for mature offspring	Assisting offspring after sexual maturity	In bonobos, female parents assist mature sons during competitive interactions with rival males

eggs is relatively common among both fishes and amphibians. For example, in sticklebacks, males provide care by building nests and fanning the eggs to supply additional oxygen, and in African bullfrogs, males dig channels to adjacent pools to prevent the pool with their tadpoles from drying out. Second, both parents cooperate to provide care for their joint offspring in the vast majority of bird species. On top of these broad patterns, there is also variation in male and female involvement in care within each taxon. Thus, although patterns of care show a strong phylogenetic signal, they can undergo evolutionary change under the appropriate conditions.

Uniparental Female or Male Care

Why is it that in those species where only one parent provides care after laying or birth, this is usually the female? The fundamental difference between the two sexes is the difference in gametes size, termed anisogamy: females produce the larger gametes (eggs) while males produce the smaller ones (sperm). Thus, one potential explanation for the predominance of female-only care is that it is somehow a consequence of anisogamy. For example, Trivers (1972) argued that females should be under selection to provide additional care after birth to safeguard their higher initial investment into the zygote. However, this argument is logically flawed as it suggests that parents base decisions on current investment on their past investment; this sort of argument is known as the 'Concorde fallacy'. Instead, we should expect selection to favor parents that base their decisions on how much to invest in their offspring on the expected future returns on that investment. A formal theoretical model by Kokko and Jennions (2008) suggests that anisogamy alone is not sufficient to promote the evolution of sex role divergence and uniparental female care. The reason for this is that female-

only care leads to a male-biased operational sex ratio (OSR), which in turn promotes the evolution of male care because it becomes more difficult for males to find a mate. In this model, anisogamy alone is associated with the evolution of egalitarian sex roles and biparental care (see below). However, a more recent model by Fromhage and Jennions (2016) suggests that biparental care is evolutionarily stable only when there are synergistic effects due to both sexes being involved in care (see below), and that it otherwise will be replaced by uniparental female or male care due to genetic drift or small differences in the costs or benefits of care.

Why then is female-only care more common than male-only care? The theoretical model by Kokko and Jennions (2008) suggests that uniparental female care is favored only when there is strong sperm competition and/or strong sexual selection acting on males. The reason for this is that sperm competition reduces the coefficient of relatedness between male parents and the offspring in a given brood, thereby lowering the benefits of care to males. Meanwhile, strong sexual selection causes disparity in reproductive success among males, with successful males not providing care because they have more to gain from attracting additional mates and unsuccessful males not providing care because they have failed to sire any offspring. The more recent model by Fromhage and Jennions (2016) also predicts that strong sperm competition and/or strong sexual selection on males will favor female-only care. However, this model shows that anisogamy might lead to female-only care due to positive feedback between parental care and sexual selection. If the greater initial investment in the gamete by females is associated with greater investment among males in competition for mates, then the associated increase in mortality among males makes parental care more costly to males, thereby generating positive feedback between parental care and competition for mates.

Why does male-only care evolve in some taxa? As uniparental male care is much more rare than uniparental female care, the conditions under which male care is favored have attracted considerable attention from empiricists. As stated above, males may have lower benefits of care due to sperm competition and higher costs of care due to the trade-off between care and attraction of additional mates (Kokko and Jennions, 2008; Fromhage and Jennions, 2016). The evolution of male parental care may therefore be associated with the evolution of paternity assurance traits, such as mate guarding and frequent copulation, ensuring that the benefits of care may be almost as high for males as for females (Trivers, 1972). For example, in the ferocious water bug, *Abedus herberti*, males achieve a very high paternity for the eggs oviposited on their back by copulating frequently with the female. Furthermore, male care may evolve in species where caring males can continue to attract additional mates, thereby removing the trade-off between parental care and acquisition of additional mates (Kokko and Jennions, 2008; Fromhage and Jennions, 2016). In some species, caring males continue attracting additional females and there is even evidence that parental care may enhance the future mating opportunities of males due to their increased attractiveness to females. When this is the case, higher levels of male care might even be favored through sexual selection.

Biparental Care

Biparental care, which occurs when males and females cooperate to care for their joint offspring, is a rare pattern of care in most animal taxa (Royle et al., 2012). The exception is birds, where biparental care is found in about 90% of all species. Biparental care is rare because it is evolutionary stable under restricted conditions. First, for biparental care to evolve, two parents must improve offspring fitness beyond uniparental female or male care (Maynard Smith, 1977). Mate removal experiments on a wide range of species show two parents often can improve offspring fitness beyond uniparental care, suggesting that his condition is often met (Clutton-Brock, 1991; Harrison et al., 2009). Second, as suggested by a recent theoretical model, biparental care may only be evolutionarily stable if there are synergistic effects between male and female involvement in care (Fromhage and Jennions, 2016). Such synergy effects might occur when the two parents provide different forms of care that complement each other such that two parents are more than twice as efficient in providing care as a single parent. This condition is far more restrictive than the first one, and it may provide an explanation for why males and females often have distinct parental roles. Finally, biparental care is associated with sexual conflict between the two parents over the amount of care that each should provide, and the evolutionary stability of biparental care may depend critically on how this conflict is resolved (see Sexual Conflict below).

Conflict and Competition over Care

Parental care can be considered a form of altruism where the parent increases the fitness of its offspring at a cost to its own personal fitness. Given the overlapping interest in ensuring offspring recruitment into the population, we expect a high degree of cooperation among family members. Nevertheless, as with any other form of altruism and cooperation, parental care will be associated with conflict. Conflict in this context is defined as the divergence in the optimal levels of care between family members, and may not necessarily be expressed as behavioral squabbles or fights (Godfray, 1995; Mock and Parker, 1997). Familial conflict emerges as an inevitable consequence of two conditions: (1) a limited supply of resources, typically food provided by the parents, and (2) a relatedness asymmetry between family members. Relatedness asymmetries describe the probability that a focal individual shares a gene with another family member, which is usually determined by the coefficient of relatedness between them ($0 \leq r \leq 1$). For example, the probability that an individual shares a particular gene copy found in a focal offspring is obviously 1.0 for the focal offspring itself whilst it is 0.5 for its parents or its full siblings.

The nature of the conflicts that take place within a particular family group is contingent on its social structure, which in turn depends on whether it is comprised of one or two parents caring for one or multiple offspring. For example, when both parents care for multiple offspring, there may be three social dimensions of familial conflict (Mock and Parker, 1997; Fig. 1): (1) parent-

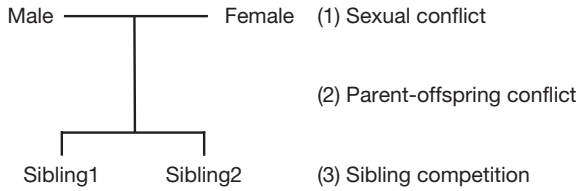


Fig. 1 Schematic diagram of the three different social dimensions of familial conflict: sexual conflict, parent-offspring conflict and sibling competition. All three social dimension of conflict are present in more complex family structures where both parents care for multiple offspring, whilst some social dimensions are absent from simpler family structures. For example, there is no sexual conflict if only one parent provides care and no sibling competition if there is a single offspring in the brood (redrawn from Mock, D. W. and Parker, G. A. (1997). *The evolution of sibling rivalry.* Oxford, UK: Oxford University Press).

offspring conflict; (2) sibling competition; and (3) sexual conflict. However, some of these social dimensions will be absent from simpler family structures. For example, there is no sexual conflict if only one parent provides care and no sibling competition if there is a single offspring in the brood.

Parent-Offspring Conflict

The first social dimension of familial conflict is parent-offspring conflict (Trivers, 1974). Trivers's theory of parent-offspring conflict introduced a major paradigm shift in how we think about the evolution of social interactions within families. Prior to Trivers, it was implicitly assumed that parent-offspring interactions were cooperative. Trivers showed that parents and offspring have overlapping but divergent interests with respect to the allocation of parental care. Fig. 2 illustrates parent-offspring conflict where the direct benefits of care are assigned to the offspring's fitness and the direct costs of care are assigned to the parent's fitness (Smiseth et al., 2012). Using kin selection theory, we can derive the indirect benefits to the parent and the indirect costs to the offspring by multiplying the direct benefits and costs by the relatedness coefficient ($r = 0.5$). Once we have the indirect benefit and indirect cost functions, we can identify the optimal levels of care to the parent (p_i^*) and its offspring (o_i^*). The conflict battleground is represented by the divergence in parental and offspring optima. The offspring's optimum exceeds the parent's optimum, suggesting that offspring are under selection to demand more resources than the parent is under selection to provide.

Trivers's theory of parent-offspring conflict met with initial criticism on the grounds that the theory was logically flawed (see Godfray, 1995). First, it was argued that any benefit gained as a more demanding offspring would be offset by a cost when producing demanding offspring as an adult. Second, it was argued that parents always win as they can control the allocation of resources. This criticism played an important role in the subsequent development of the field as it helped separate two key issues: (1) whether there is a conflict battleground and (2), if there is, how the conflict is resolved (Godfray, 1995). Theoretical models of the conflict battleground have shown that parent-offspring conflict can evolve when the perspective is shifted from that of an individual to that of a gene. Genes for conflicting behaviors in the offspring can spread in the population as long as they outcompete alternative non-conflicting alleles. Meanwhile, theoretical models for the resolution of parent-offspring conflict have shown that costly offspring begging behaviors can serve as a mechanism for resolving this conflict (Godfray, 1991). In the absence of any costs of begging, offspring are under selection to exaggerate their needs to acquire additional resources. However, such manipulative begging is not evolutionarily stable given that parents should ignore purely selfish offspring behaviors. In contrast, begging would be evolutionarily stable if offspring incur a cost from begging. Under such circumstances, the cost punishes offspring that misrepresent their true needs, and parents benefit from monitoring the offspring's begging behavior because they obtain honest information on their offspring's nutritional condition (Godfray, 1991).

The emergence of resolution models that made testable predictions stimulated empirical research on offspring begging in birds and other taxa. These studies find that, as predicted by theory (Godfray, 1991), begging reflects the offspring's nutritional state, and parents respond to offspring begging by adjusting their allocation of food (Kilner and Johnstone, 1997). There is also empirical evidence that offspring begging is costly in terms of attracting predators or slowing down growth or development (Kilner and Johnstone, 1997). Perhaps surprisingly, there is little empirical evidence for a conflict battleground, presumably reflecting that it is difficult to obtain empirical data on the shape of the benefit and cost functions of parental care that are needed to derive the parental and offspring optima (Fig. 2).

Sibling Competition

The second social dimension of conflict is sibling competition, defined as any offspring trait that promotes the fitness (survival or growth) of an individual offspring at the expense of the fitness of its siblings (Mock and Parker, 1997). Given that sibling competition occurs among close relatives, an important aim is to identify conditions that favor or constrain the evolution of selfish behavior towards close kin. Although relatedness constrains selfishness, sibling competition can evolve as long as there is a mismatch between the supply of resources from the parents and the total demand for resources by the offspring (Mock and Parker, 1997). Such a mismatch occurs because parents tend to overproduce offspring at the beginning of the breeding attempt (Mock and

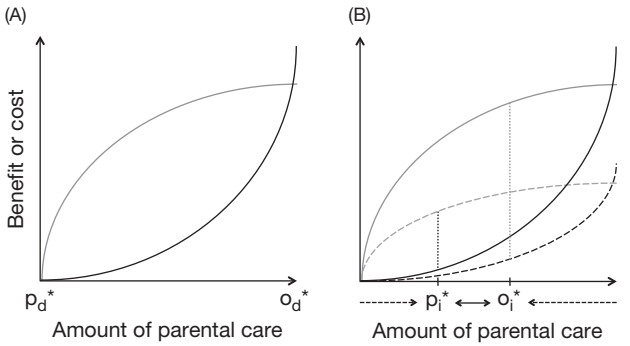


Fig. 2 Illustration of parent-offspring conflict, defined as the divergence between the optimal level of care for a parent and its offspring. The benefit of care to the offspring is assumed to increase at a decelerating rate, whilst the cost of care to the parent is assumed to increase at an accelerating rate. (A) The direct optima to the parent and the offspring when assigning the cost of care to the parent's fitness and the benefit of care to the offspring's fitness. The direct optima to the parent and the offspring, represented as P_d^* and O_d^*, are at the extreme ends (no care and maximum care, respectively). (B) The indirect optima to the parent and the offspring after assigning indirect benefits of care to the parent and the indirect costs of care to the offspring. The indirect benefits and costs can be found by multiplying the direct benefits and costs with the coefficient of relatedness between the parent and the offspring. The indirect optima to the parent and offspring are represented as P_i^* and O_i^*, respectively. The shift in the indirect optima illustrates that relatedness moderates conflict (dashed horizontal arrow), whilst the divergence in the indirect optima shows there nevertheless is a potential for conflict between close relatives (solid horizontal arrow).

Parker, 1997). Overproduction of offspring may be adaptive from the parents' point of view as it allows them to take advantage of favorable environmental conditions or to have an insurance against hatching failure of eggs and early mortality of hatchlings.

Sibling competition is common in species where parents provision their offspring with food, which may seem paradoxical given that parental food provisioning evolves to enhance the offspring's access to food and that there should be less sibling competition when there is better access to food. Parental care may nevertheless promote sibling competition because it is associated with interference among competing offspring. For example, in the burying beetle *Nicrophorus vespilloides*, larvae obtain food either by self-feeding directly from the carcass upon which they breed or by begging for food from the parent. In this species, there is interference when offspring compete by begging but not they compete by self-feeding. Sibling competition for parental resources can take many different forms, ranging from non-lethal scramble competition to lethal aggressive brood reduction (Mock and Parker, 1997). Diversity in the form of sibling competition may be driven by the extent to which an individual offspring can monopolize resources, which in turn is contingent on the size of resources, the existence of potential weaponry such as powerful beaks, and the presence of competitive asymmetries within the brood (Mock and Parker, 1997).

Theoretical models treat begging as a form of non-lethal scramble competition (Parker et al., 2002). However, unlike the honest signaling models of begging discussed above (Godfray, 1991), these models assume that it is the offspring (rather than the parents) that control the allocation of resources. Scramble competition models make the same predictions as honest signaling models; that is, begging reflects offspring nutritional state and parents respond to begging. Furthermore, scramble competition models also assume that begging is costly (Parker et al., 2002). Thus, distinguishing between honest signaling and scramble competition models of begging is a major challenge in this field. Nevertheless, there is empirical evidence that offspring adjust their begging behavior to the competitive environment in which they find themselves. For example, in many birds, individual nestlings adjust their begging behavior to the number of brood mates and/or to their competitive rank as determined by the hatching order. Thus, offspring begging is not simply a signal of the offspring's nutritional needs as suggested by honest signaling models.

A more extreme form of sibling competition is lethal aggressive brood reduction (Mock and Parker, 1997). This form of sibling competition has been studied in detail in great egrets. This species has a clutch size of 3–4 eggs, with a laying interval of 1–2 days between each egg. Females start incubation once the first egg is laid, leading to a pronounced age-based competitive hierarchy within the brood. Nestlings have powerful bills, reflecting the fish-catching lifestyle of the adults. The bills are used as weaponry during aggressive interactions between siblings. There are two types of feeding: indirect feeding where the parent deposits boluses of food on the nest floor and direct feeding where nestlings intercept boluses before they reach the nest floor. The nestlings do this by grabbing the parent's bill. The older chicks direct aggression towards their youngest chick, and the youngest chick get access to less food and have the lowest survival prospects within the brood (Fig. 3).

There is also scope for cooperation among siblings (Forbes, 2007). For example, in barn owls, siblings negotiate who is to get the next food item before the parents return with food as a means to reduce the amount of aggression within the brood. Meanwhile, in black-headed gulls, parents provide more food to the brood when nestlings synchronize their begging. Such cooperation does not preclude competition and a potential avenue for further work in this field is to investigate the balance between cooperation and competition among nestlings.

Sexual Conflict

The final social dimension of conflict is sexual conflict over parental care, which occurs because the benefits of care to the offspring depends on the combined amount of care by the two parents, while the costs of care to each parent depends on its personal

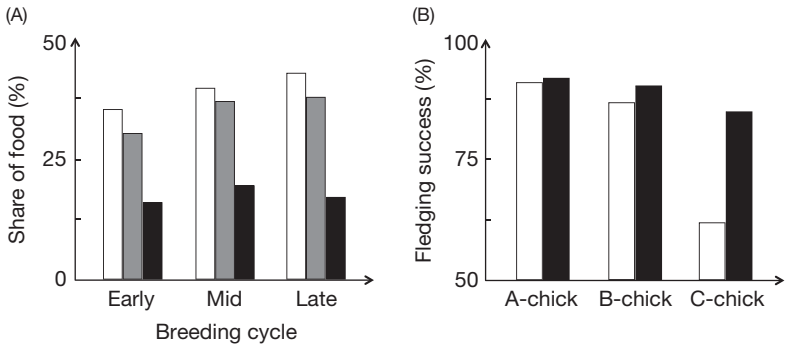


Fig. 3 Sibling competition by lethal brood reduction in the great egret. (A) The oldest A and B chicks (*white and gray bars*, respectively) receive the largest amount of food, whilst the youngest C chick (*black bars*) receives the smallest amount of food. (B) The A and B chicks have higher survival prospects, but the C chicks survival prospects improves if either the A or B chick dies first (*black bars*) (redrawn from Mock, D. W. and Parker, G. A. (1997). *The evolution of sibling rivalry*. Oxford, UK: Oxford University Press).

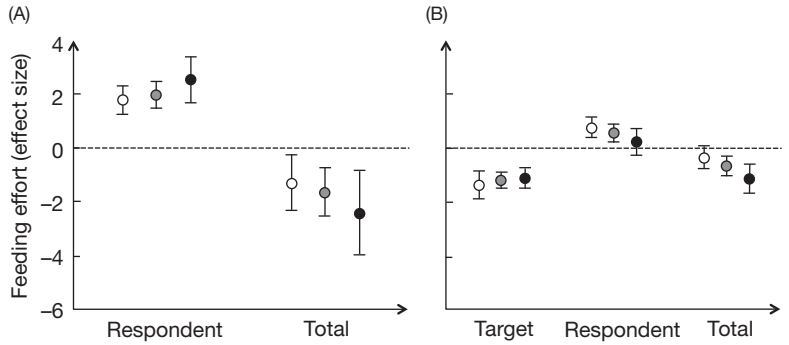


Fig. 4 A meta-analysis of published mate removal and handicapping experiments conducted on birds. (A) In removal experiments, the remaining parent (i.e., respondent) increases its level of care beyond the baseline (*dotted line*), but not such that it compensated completely for the removal of its partner (i.e., total). (B) In handicapping experiments, the handicapped parent (i.e., the target) reduces the amount of care it provides, its partner (i.e., respondent) increases the amount it provides but not such that it fully compensates for the reduction in the amount of care provided by the target (i.e., total). Females are represented by white circles, males by black circles, and both sexes by gray circles (redrawn from Harrison, F., Barta, Z., Cuthill, I. and Székely, T. (2009). How is sexual conflict over parental care resolved? A meta-analysis. *Journal of Evolutionary Biology* **22,** 1800–1812, with permission from John Wiley and Sons).

workload (Lessells, 2012). Owing to sexual conflict, each parent is under selection to reduce its own contribution and shift as much of the workload as possible over to its partner. Potentially, sexual conflict may undermine the evolutionary stability of biparental care and an important aim in behavioral ecology is therefore to identify mechanisms that allow biparental care to remain evolutionary stable despite sexual conflict.

Theoretical models suggest that sexual conflict over parental care can be resolved by three behavioral mechanisms: negotiation, matching and sealed-bid decisions (Lessells, 2012). Negotiation and matching occur when each parent adjusts its level of care in direct response to its partner's contribution. When there is negotiation, the focal parent responds to a reduction in amount of care provided by its partner by increasing its contribution, though only such that it compensates incompletely for the partner's reduction. In contrast, when there is matching, the focal parent responds by matching any increase or reduction in its partner's contribution. Finally, sealed-bid decisions occur when each parent makes an initial fixed decision about how much care to provide and that decision is independent of that of its partner. Experimental studies on birds and other taxa provide some support of all three mechanisms (Lessells, 2012). However, a meta-analysis of studies on birds found overall support for negotiation (Harrison et al., 2009; Fig. 4).

Less attention has been given to other mechanisms contributing towards the resolution of sexual conflict. For example, given that females control the production of eggs, females could potentially manipulate the behavior of caring males by adjusting clutch size, asynchronous hatching or maternal effects. Female manipulation is the outcome of any mechanism used by the female to influence the amount of care provided by her male partner in a way that increases the female's and/or offspring's fitness at the expense of the male's fitness (Paquet and Smiseth, 2016). Female birds deposit androgens such as testosterone in the eggs, thereby stimulating offspring begging, and this might provide females with a mechanism for increasing the male's contribution towards food provisioning. Currently, there is limited evidence that yolks androgens serve such a function, but further work is needed on the role of other mechanisms.

Familial Conflicts and their Resolution

An important aim in the study of familial conflict is to understand the mechanisms that mediate the resolution of such conflicts. This section highlights that each social dimension of conflict is associated with a specific mechanism of conflict resolution: (1) parent-offspring conflict with communication and honest signaling (Godfray, 1991, 1995); (2) sibling competition with scramble competition or aggression (Mock and Parker, 1997); and (3) sexual conflict with negotiation (Harrison et al., 2009; Lessells, 2012). Traditionally, the resolution of each conflict has been studied in isolation from each other. However, conflict resolution at one social dimension will impact on conflict resolution at other social dimensions. For example, the resolution of sexual conflict may determine the supply of food to the brood, which in turn may alter the level of sibling competition. Thus, an important future challenge is to understand interactions between conflict resolutions at different social dimension of conflict.

Acknowledgements

I thank past and present members of the burying beetle group at the University of Edinburgh and students attending the Evolution of Parental Care course at the University of Edinburgh for helping me develop my thoughts on parental care.

References

Clutton-Brock TH (1991) *The evolution of parental care*. Princeton, NJ: Princeton University Press.
Forbes S (2007) Sibling symbiosis in nestling birds. *Auk* 124(1): 10.
Fromhage L and Jennions MD (2016) Coevolution of parental investment and sexually selected traits drive sex role divergence. *Nature Communications* 7: 12517.
Godfray HCJ (1991) Signalling of need by offspring to their parents. *Nature* 352: 328–330.
Godfray HCJ (1995) Evolutionary theory of parent-offspring conflict. *Nature* 376: 133–138.
Harrison F, Barta Z, Cuthill I, and Székely T (2009) How is sexual conflict over parental care resolved? A meta-analysis. *Journal of Evolutionary Biology* 22: 1800–1812.
Kilner R and Johnstone RA (1997) Begging the question: are o spring solicita- tion behaviours signals of need? *Trends in Ecology and Evolution* 12: 11–15.
Kokko H and Jennions MD (2008) Parental investment, sexual selection and sex ratios. *Journal of Evolutionary Biology* 21: 919–948.
Lessells CM (2012) Sexual conflict. In: Royle NJ, Smiseth PT, and Kölliker M (eds.) *The evolution of parental care*, pp. 150–170. Oxford, UK: Oxford University Press.
Maynard Smith J (1977) Parental investment: a prospective analysis. *Animal Behaviour* 25: 1–9.
Mock DW and Parker GA (1997) *The evolution of sibling rivalry*. Oxford, UK: Oxford University Press.
Paquet M and Smiseth PT (2016) Maternal effects as a mechanism for manipulating male care and resolving of sexual conflict over care. *Behavioral Ecology* 27: 685–694.
Parker GA, Royle NJ, and Hartley IR (2002) Begging scrambles with unequal chicks: interactions between need and competitive ability. *Ecology Letters* 5: 206–215.
Royle NJ, Smiseth PT, and Kölliker M (2012) *The evolution of parental care*. Oxford, UK: Oxford University Press.
Smiseth PT, Kölliker M, and Royle NJ (2012) What is parental care? In: Royle NJ, Smiseth PT, and Kölliker M (eds.) *The evolution of parental care*, pp. 1–17. Oxford, UK: Oxford University Press.
Trivers RL (1972) Parental investment and sexual selection. In: Campbell B (ed.) *Sexual selection and the descent of man, 1871–1971*, pp. 136–179. Chicago, IL: Aldine.
Trivers RL (1974) Parent-offspring conflict. *American Zoologist* 14: 249–264.

Further Reading

Clutton-Brock TH (1991) *The evolution of parental care*. Princeton, NJ: Princeton University Press.
Godfray HCJ (1995) Evolutionary theory of parent-offspring conflict. *Nature* 376: 133–138.
Kokko H and Jennions MD (2008) Parental investment, sexual selection and sex ratios. *Journal of Evolutionary Biology* 21: 919–948.
Lessells CM (2012) Sexual conflict. In: Royle NJ, Smiseth PT, and Kölliker M (eds.) *The evolution of parental care*, pp. 150–170. Oxford, UK: Oxford University Press.
Mock DW and Parker GA (1997) *The evolution of sibling rivalry*. Oxford, UK: Oxford University Press.
Royle NJ, Smiseth PT, and Kölliker M (2012) *The evolution of parental care*. Oxford, UK: Oxford University Press.
Smiseth PT, Kölliker M, and Royle NJ (2012) What is parental care? In: Royle NJ, Smiseth PT, and Kölliker M (eds.) *The evolution of parental care*, pp. 1–17. Oxford, UK: Oxford University Press.

List of Relevant Websites

http://www.oxfordbibliographies.com/view/document/obo-9780199941728/obo-9780199941728-0014.xml.
http://www.oxfordbibliographies.com/view/document/obo-9780199830060/obo-9780199830060-0131.xml.

Sexual Selection and Sexual Conflict

Ulrika Candolin, University of Helsinki, Helsinki, Finland

Glossary

Effective population size, N_e The number of individuals in a population that contribute offspring to the next generation.

Mating system Describes the manner in which males and females associate during mating in an attempt to maximize their lifetime reproductive success. Animal mating systems are categorized according to the number of mates a given sex is able to monopolize (monogamy, polygamy, and promiscuity). Such categorizations can be based on social bonds (social mating system) or on genetic paternity and maternity of the resultant offspring (genetic mating system).

Operational sex ratio (OSR) The average ratio of fertilizable females to sexually active males at any given time.

Potential reproductive rate (PRR) The number of offspring that an individual can produce per unit time if unconstrained by mate availability. The rate is usually higher in males than females, as these produce many tiny sperm while females invest in larger ova and—more often than males—in time consuming parental care.

Sexually antagonistic co-evolution (SAC) A conflict between the evolutionary interests of the sexes that cause one sex to evolve traits that reduce the ability of the opposite sex to reproduce with other partners. The opposite sex can then evolve defense mechanisms against the manipulation, resulting in an "arms race." This can escalate and reduce female fecundity and, hence, population growth rate.

Transgenerational epigenetic effects Heritable alterations of gene expression without a change to the DNA sequence. Include DNA methylation, chromatin remodeling, histone modification, and non-coding RNA mechanisms.

Introduction

Many species possess conspicuous traits that appear to confer no advantage in the struggle for survival, such as colorful plumes and feathers, melodic songs, loud calls, and exaggerated structures like horns and antlers. The selection pressure behind these traits is usually sexual selection; in the competition for mates and fertilizations, organisms have evolved a variety of traits that increase either their attractiveness to the opposite sex, or their success in the competition among rivals for access to mates and fertilizations. How these traits evolve and why has intrigued scientists for decades and grown into a vivid field of research. In particular, the question of why females prefer males with cumbersome traits, which obviously reduce the probability of survival—the peacock's train being a classic example—has been difficult to solve and subjected to much research.

The foundation of the theory of sexual selection was laid by Charles Darwin. He realized that the competition for mates can result in traits that are of no advantage in the struggle for survival, but evolve because of the advantage they convey over rivals in the competition for mates. Since then, the field of sexual selection has grown rapidly and provided detailed knowledge of how and why sexual selection operates. The consequences of sexual selection for populations, on the other hand, are still poorly known and in need of more research.

In this article, I go through the mechanisms behind sexual selection—how and why does sexual selection operate—as well as the consequences of sexual selection for populations. I discuss both the benefits and costs of sexual selection, as well as the question of whether sexual selection enhances or depresses population viability. An inevitable part of sexual selection is a conflict of interest between the sexes, as the evolutionary interest of genetically different individuals always differ. This conflict can result in the evolution of traits that are beneficial to the individual but detrimental to the population. It can, hence, play a fundamental role in determining the effect of sexual selection on populations.

The Mechanisms of Sexual Selection

Intra- and Intersexual Selection

Sexual selection favors the evolution of traits that increase success in the competition for mates and fertilizations, i.e., of sexually selected traits. These traits can vary from flamboyant ornaments that attract the opposite sex, such as the peacock's train, to weapons used to combat rivals, such as the antlers of deer.

Sexual selection operates through two main pathways, intra- and intersexual selection. Intrasexual selection occurs when competition among members of the same sex for mates and fertilizations favors the evolution of traits that provide an advantage in this competition, like large antlers in male deer that are used in male–male competition for access to females. Intersexual selection occurs when members of one sex choose mates of the other sex based on some traits. When the preferences for these traits

 doi:10.1016/B978-0-12-409548-9.10886-3

are favored by selection, then the preferred traits may evolve and become exaggerated, like the train of the peacock (although strictly speaking intersexual selection is also competition within the sexes, but competition for being chosen and making the best choice). In most species, males are more competitive and females more choosy, but some degree of competitiveness and choosiness can occur in both sexes.

Pre- and Postcopulatory Sexual Selection

Sexual selection can be further divided into pre- and postcopulatory sexual selection. Precopulatory sexual selection takes place before mating and is the most obvious form of sexual selection. It was described already by Charles Darwin and was the focus of early sexual selection research. Postcopulatory sexual selection takes place after mating and received little attention until the publication of an influential paper by Geoff Parker in 1970 on sperm competition (Parker, 1970). Since then the field of postcopulatory sexual selection has grown rapidly and is today a highly active research field.

Precopulatory sexual selection can take many forms. It can be the competition for resources that are then used to attract mates —such as large territories—or the competition for direct access to mates—such as physical fights over mates, or ritualized displays of dominance. Similarly, mate choice can take many forms. It can be based on preferences for particular traits in the opposite sex—such as colorful ornaments—or take the form of resistance against mating attempts. The latter is common in species where the potential reproductive rate of males is much higher than that of females and males therefore attempt to copulate with unwilling females. A high potential reproductive rate reduces the cost of each mating attempt, which makes males less choosy.

Postcopulatory sexual selection includes sperm competition and cryptic female choice (Birkhead and Pizzari, 2002). Sperm competition arises when many males mate with a female, either because the female chooses to mate with multiple males or because multiple males overcome female resistance to mate (male harassment). It can favor the evolution of male traits that increase fertilization success, such as mate guarding, copulatory plugs that prevent female re-mating, or seminal proteins that influence female physiology. For instance, males of the fruit fly (*Drosophila melanogaster*) inseminate females with sperm that are embedded in a cocktail of seminal substances, which stimulate ovulation, deactivate sperm that are already stored in the reproductive tract of the female, and act as an anti-aphrodisiac, discouraging females from participating in future copulations (Chapman, 2001).

Cryptic female choice occurs when females have some control over which males' sperm fertilize the ova. It is "cryptic" in the sense that the choice takes place hidden within the reproductive tract of the female. Cryptic female choice is less investigated than sperm competition, and its prevalence is poorly known. An indication of the occurrence of cryptic female choice is complex female sperm storage organs that rapidly co-evolve with sperm and ejaculate traits. Cryptic choice has been documented in species such as fruit flies, stalk-eyed flies and feral fowls. Female feral fowls (*Gallus gallus domesticus*), for instance, have organs that can expel the ejaculates of lower-ranking males and, in so doing, increase the insemination success of higher ranking males (Pizzari and Birkhead, 2000; **Fig. 1**).

The Strength of Sexual Selection

The strength of sexual selection on traits used in mate competition and mate choice depends on the fitness costs and benefits of these traits. A common cost that constrains sexual selection is the expenditure of a limited pool of resources on competition and choice, as this reduces the availability of resources for other fitness enhancing traits. For instance, the allocation of time, energy and nutrients to mate competition and mate choice can reduce the amount available for feeding and growth. Other common costs are increased risk of predation and parasite infections. Mate searching, in particular, that increases activity can raise encounter rate with

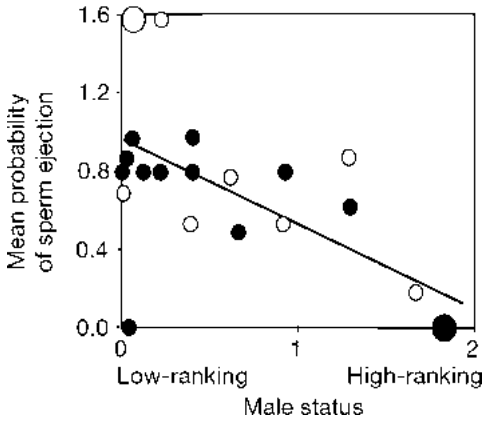


Fig. 1 An example of cryptic female choice; to increase the probability of ova being fertilized by sperm of high quality males, females feral fowls eject the sperm of low ranking males. Small data points represent one male, large data points, two males. Different colors refer to different years. Small data points represent one male, large datapoints, two males. Reproduced from Pizzari, T., & Birkhead, T. R. (2000). Female feral fowl eject sperm of subdominant males. *Nature 405*: 787–789, with permission from Elsevier.

predators or parasites, while ornaments and courtship behaviors can make individuals more conspicuous to these. An example is threespine stickleback males, who use a conspicuous red nuptial coloration to attract females and repel competing males, which increases their risk of predation from piscivorous fishes (Johnson and Candolin, 2017; **Fig. 2**).

The benefit of competition and mate choice, which favors the evolution of sexually selected traits, is improved fertilization success in relation to other individuals in the population. This benefit has to be larger than the costs, i.e., it has to increase lifetime fitness in terms of number of copies of the individual's genes that are transferred to subsequent generations. Relative mating success—not absolute mating success—is central, as relative success influences the representation of the genes of the individual in the next generation.

These costs and benefits of sexually selected traits depend in turn on a range of factors, both environmental and intrinsic to the individual. These are variation among individuals in the fitness benefits they can offer (if the variation is small, the benefit of mate choice is also small); mate encounter rate, which determines the possibility of choice; the operational sex ratio, which influences the strength of competition for mates; and life-history traits, such as number of lifetime reproductive opportunities, which in turn are determined by a range of factors, such as body condition, parental effort and age. For instance, the white plumage patches of collared flycatchers (*Ficedula albicollis*) show positive age-dependent expression, such that the size of the ornaments increases as individuals age (Evans *et al.*, 2011; **Fig. 3**). This increase in size could be a consequence of the cost of ornament expression decreasing when future reproductive opportunities decline towards the end of life, as this could allow individuals to increase their investment into the sexually selected trait as a terminal investment.

Sexual Conflict

A conflict of interest always arises when genetically different individuals attempt to maximize their fitness. This is because fitness is a relative measure, and the proportion of genes passed on to the next generation can only increase if that of other individuals decrease. In a mating context, a conflict of interest arises when the sexes have conflicting optimal fitness strategies (Arnqvist and Rowe, 2005). For instance, females may benefit if males increase their parenting effort, as this saves resources for investment into fecundity and future reproductive opportunities. Males, on the other hand, may benefit from leaving parental care to females, as this frees up time and energy for searching for new mates, to gain additional offspring. An example is males of the European starling (*Sturnus vulgaris*), who decrease their paternal effort when more nestboxes are available and the opportunity to attract additional mates is high (Smith, 1995).

Sexual conflict can be of two types: interlocus and intralocus. Interlocus sexual conflict occurs when one sex evolves traits that enhance its reproductive success at the expense of the fitness of its mating partners. The other sex may then evolve counter-adaptations, which are determined by other genes (loci), which leads to antagonistic coevolution between the sexes. For example, male fruit flies produce seminal proteins that reduce the probability that females will reproduce with other males. This has in turn induced females to evolve counteradaptations to resist their impact. Such sexually antagonistic coevolution can result in bizarre traits, such as the spines on the genitalia of male seed beetles, which harm females (Ronn *et al.*, 2007; **Fig. 4**). Other examples of traits that evolve through interlocus sexual conflict are mate guarding, physical harassment, and resistance against mating attempts.

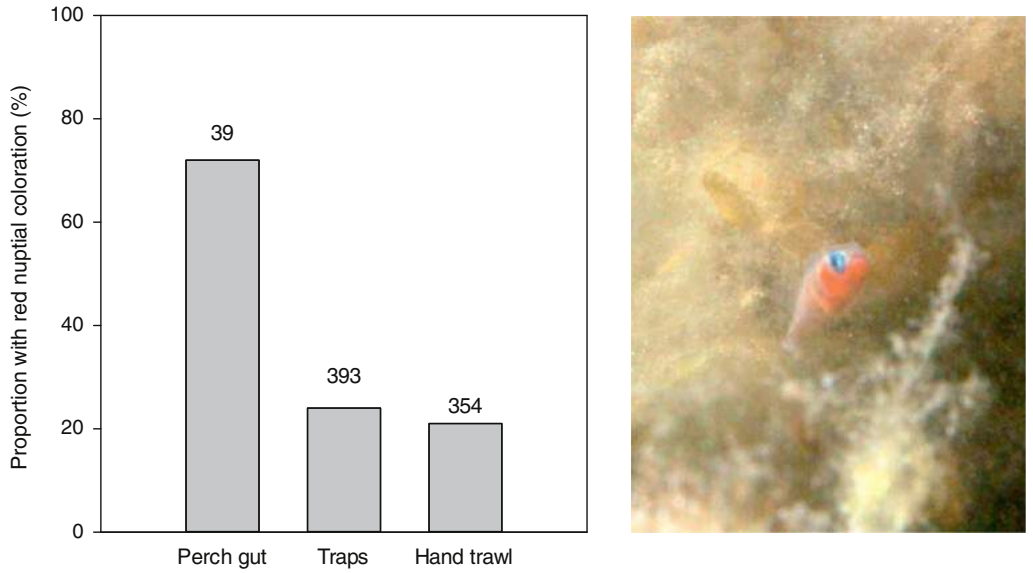


Fig. 2 The conspicuous *red* nuptial coloration of threespine stickleback males increases their predation risk. The proportion of *red* colored sticklebacks that are caught by a predator, the European perch (*Perca fluviatilis*), is higher than the proportion in the population. Reproduced from Johnson, S., & Candolin, U. (2017). Predation cost of a sexual signal in the threespine stickleback. *Behavioral Ecology*, *28*: 1160–1165, with permission from Elsevier.

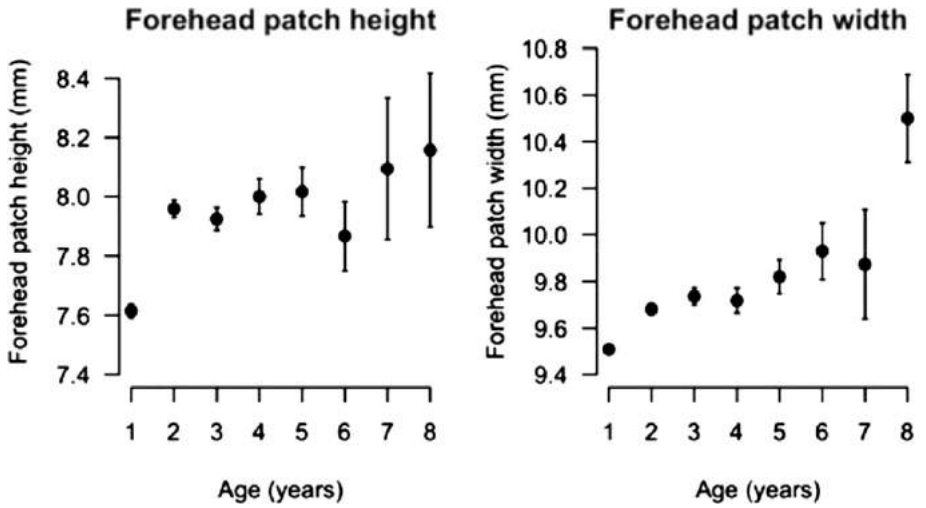


Fig. 3 The expression of a sexually selected trait, the *white* forehead patch size of the collared flycatcher, increases with age. This could be a consequence of reduced costs of expressing the trait in terms of loss of reproductive opportunities towards the end of the life, as reproductive opportunities then decrease. Reproduced from Evans, S. R., Gustafsson, L., & Sheldon, B. C. (2011). Divergent patterns of age-dependence in ornamental and reproductive traits in the collared flycatcher. *Evolution 65*: 1623–1636, with permission from Elsevier.

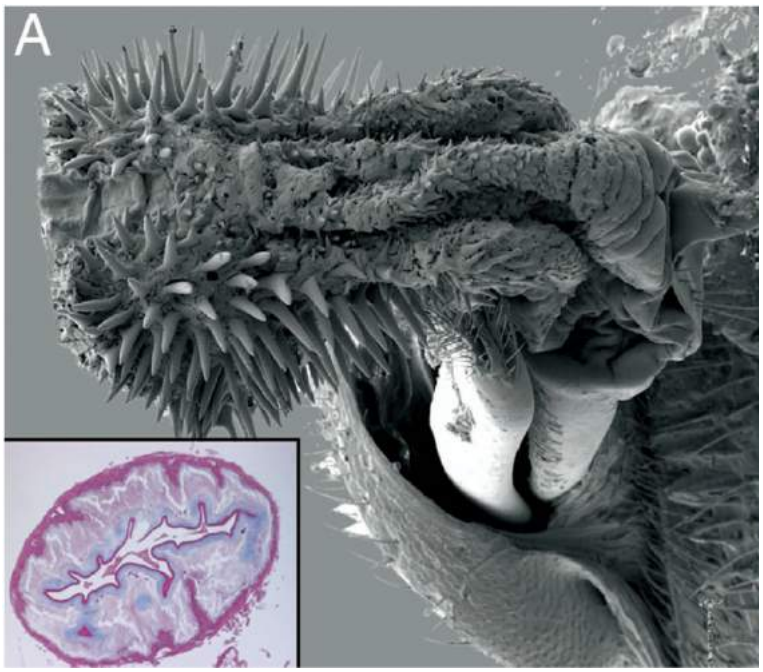


Fig. 4 Male seed beetles (*Coleoptera bruchidae*) have evolved spines on their genitalia that enhance stability during copulation, but cause harm to females. Females have in turn evolved resistance to these spines, through tougher copulatory ducts, which has led to sexually antagonistic coevolution (SAC) with males evolving more and more spines on their genitalia. Reproduced from Ronn, J., Katvala, M., & Arnqvist, G. (2007). Coevolution between harmful male genitalia and female resistance in seed beetles. *Proceedings of the National Academy of Sciences of the United States of America 104*: 10921–10925, with permission from NAS.

Intralocus sexual conflict occurs when the sexes have different optima for a trait that is expressed in both sexes and determined by the same genes, such as horn size or body coloration.

The conflict of interest is between individuals, between the male and the female attempting to reproduce. At the population level, the fitness of the sexes is equal when the sex ratio is equal, as each offspring belongs to both sexes. Thus, the variation in mating success is among individuals within each sex competing for mates. The variation is usually higher in males, as these typically have a higher potential reproductive rate, which, hence, are more competitive.

The Evolution of Sexually Selected Traits

Whether sexual selection results in the evolution of sexually selected traits depends on the strength of sexual selection, as well as on the heritability of the traits. The strength of sexual selection depends on the costs and benefits of sexually selected traits, as was discussed earlier in this article. The heritability of sexually selected traits depends in turn on the magnitude of variation among individuals in the traits, and the degree to which this variation is determined by genetic (i.e., allelic) and nongenetic mechanisms. If the variation among individuals in traits is genetically determined and in the direction of selection, then the traits may evolve through genetic changes. If, on the other hand, the variation is nongenetic and caused by phenotypic plasticity and environmental effects, then the traits might still evolve but through transgenerational effects, such as parental effects.

An increasing number of studies find nongenetic inheritance to contribute to evolution. An example is the loudness of an acoustic signal in a grasshopper (*Chorthippus biguttulus*), which is determined by the diet of the parents and, hence, inherited through parental effects (Franzke and Reinhold, 2013). The prevalence of nongenetic inheritance, and the persistence of the effects, are, however, poorly known. Transgenerational epigenetic effects (such as the methylation of DNA) might persists across many generations, as these alter gene expression, but how common epigenetic effects are in a sexual selection context, and for how many generations they can persist, are unknown.

Genetic and nongenetic inheritance may interact in influencing the evolution of sexually selected traits. This can complicate the prediction of evolutionary responses to sexual selection. Nongenetic inheritance may accelerate evolution by increasing the amount of heritable phenotypic variation that selection can act on, and by causing rapid changes in trait expression, as these may persist only from one generation to the next.

The heritability of sexually selected traits—whether genetic or nongenetic—is dependent on the magnitude of variation among individuals in the traits, as well as on correlations among the traits. Low variation and negative correlations can constrain evolution. For instance, in *Drosophila serrata*, female choice is based on male cuticular hydrocarbons (CHCs), but in populations with pleiotropic covariation among the traits, selection on several traits result in no evolutionary response (Hine *et al.*, 2014). This is probably because covariation among the traits reduces genetic variation in trait combinations in the direction of selection.

The Theory of Sexual Selection

The evolution of traits used in male–male competition for access to females is easily understood. A male that wins over other males in physical fights, or acquires a larger territory than other males, gains access to more females and have a higher mating success. The traits that ensures or correlates with this success, such as large body size or antlers, are then favored by sexual selection.

A more challenging question is the evolution of female mate preferences for traits that are harmful to survival, such as colorful ornaments that attract predators, or vigorous courtship displays that are energetically costly. This question occupied scientists during the early years of sexual selection research, and today we have a fairly good understanding of the factors that promote the evolution of female (or male) mate preferences. These mechanisms can be divided into two main groups; direct and indirect benefits of mate choice.

The direct benefits hypothesis states that females evolve preferences for traits that reflect some direct benefits of mating with the male, such as male parenting ability or fertilization success. By basing their mate choice on these traits, females improve their own reproductive success in terms of number of offspring produced. For example, females of the northern cardinal (*Cardinalis cardinalis*) use the brightness of the males' plumage to gain information on the proportion of the offspring feeding that the males would provide (Linville *et al.*, 1998; **Fig. 5**).

The indirect benefits hypothesis states that females evolve preferences for traits that reflect genetic benefits and, hence, improve offspring fitness, and thereby indirectly the fitness of the female. These models can be further divided into two groups: (1) good genes models that assume that offspring inherit genes that improve their survival, and (2) Fisherian runaway process that assumes that offspring inherit the attractiveness of their father and, hence, have a high mating success.

The Fisherian process includes self-reinforcing coevolution between the trait and the preference. Males evolve a trait that is preferred by females, and the trait and the preference then become genetically coupled. When the female preference increases in frequency in the population, exaggeration of the male trait yields an increasing advantage to the male and the trait evolves further, until the "runaway" evolution is stopped by opposing natural selection, i.e., by the costs of the trait. The degree to which Fisher's runaway process occurs in nature is poorly known, but it could commonly operate alongside other processes that drive the evolution of sexually selected traits.

Other mechanisms that can influence the evolution of mate preferences are the sensory drive mechanism and sexually antagonistic coevolution. The sensory drive model states that the presence of pre-existing sensory biases influence which traits females (or males) pay attention to and, hence, which traits will become exaggerated. For instance, females of a tropical fish—the swordtail characin (*Corynopoma riisei*)—feed on ants, which has favored the evolution of male ornaments with the shape of an ant (Kolm *et al.*, 2012). Although sensory drive can initiate the evolution of sexually selected traits and preferences, it is unlikely to drive the evolution alone. It needs to be coupled with direct or indirect benefits of choice for the sexually selected traits to persist in the population.

The sexually antagonistic coevolution model builds on the conflict of interest between the sexes, as discussed earlier. The sexes are caught in a perpetual sexual arms race where one sex attempts to overcome the resistance of the other sex, which then evolves

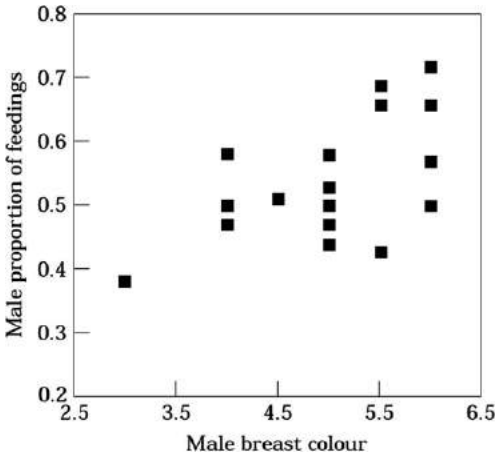


Fig. 5 The breast color of male northern cardinals reflects the proportion of the offspring feeding that the male will provide. Thus, the color can be used as an indicator of the direct benefit that the female would gain by mating with the male. Reproduced from Linville, S. U., Breitwisch, R., and Schilling, A. J. (1998). Plumage brightness as an indicator of parental care in northern cardinals. *Animal Behaviour 55*: 119–127, with permission from Elsevier.

stronger resistance, until the benefit of further escalation is outweighed by the cost. Antagonistic coevolution is assumed to operate alongside other processes and contribute to the evolution of sexually selected traits, but evidence for this is so far limited.

The Maintenance of Honesty

An intriguing question is how honest sexual signals of mate quality are ensured—i.e., what prevents individuals from deceptively increasing their signaling effort in order to attract more mates. A solution has been posed by the handicap principle; it postulates that sexually selected signals have to be costly so that poor quality individuals cannot express as costly signals as good quality individuals. This is analogous to the "big house and big cars" correlation, where individuals with much resources can invest in both, in the case of sexually selected signals, in both signals and other fitness enhancing traits. One mechanism that can ensure a positive correlation between viability and the expression of signals is condition-dependence of signals. For instance, the tail length of barn swallows (*Hirundo rustica*) indicates viability because long tails are costly in terms of energy and nutrient demands and, hence, only individuals in prime condition—who have a high viability—can sustain the cost (Møller, 1994).

Relation to Mating Systems

The strength of sexual selection is closely related to the mating system. In monogamous mating systems the number of mating opportunities is limited and the competition for mates is consequently less strong. Sexual selection is then weak and highly exaggerated traits are unlikely to evolve. In polygamous or promiscuous mating systems, on the other hand, competition for mates is stronger and exaggerated sexually selected traits are more likely to evolve. Exaggeration of male traits is particularly likely, as males usually have a higher potential mating rate than females. Male California sea lions (*Zalophus californianus*), for instance, can gather harems of 10–20 females on their territories, which results in fierce fighting among males for females. Such intense competition increases the strength of sexual selection on traits that improve success in male–male competition, such as large male size. Consequently, profound sexual size dimorphism has evolved, with males being three to four times heavier than females.

The mating strategy an animal adopts depends on the factors that determine the number of mates an individual can monopolize and fertilize, i.e., the mating skew. These factors can be demographic, social and ecological. For instance, the abundance and distributions of food can influence the number of females a male can provide for, and, hence, the intensity of competition for mates and fertilizations. Similarly, the need for male parental care can restrict the number of females a male can mate with and thereby the strength of sexual selection.

Alternative Mating Behaviors

When the competition for mates is intense, sexual selection may cause the evolution of alternative mating tactics. An individual with a low probability of gaining matings through the dominant mating tactic—such as fights or courtship—may adopt an alternative mating tactic, such as sneaking or forced copulations. For example, male guppies (*Poecilia reticulata*) that fail to gain matings through courtship may attempt forced copulations.

Such alternative tactics can be adopted throughout life, if the fertilizations success through them is higher than the success through the dominant tactic. For instance, males who are born at an unfavorable time of the year, and therefore suffer from small size throughout life, may remain as small sneaker males. On the other hand, if the males can grow in size with time, they may switch to the dominant tactic when reaching a critical size. If the lifetime fitness benefit of the two tactics is equal, then the tactics may become genetically fixed and persist throughout life, as both are then evolutionary stable.

Sexual selection can favor the adoption of an alternative tactic alongside the dominant tactic as a strategy to maximize fitness. For instance, stickleback males that have established a territory and successfully attracted females through courtship may, when given a chance, also capitalize on the courtship effort of neighboring males and sneak fertilize the eggs females lay in their nests (Candolin and Vlieger, 2013). Similarly, both males and females, of a range of species, engage in extra pair copulations in the pursuit of higher lifetime fitness. Sexual selection can then favor the evolution of traits that increases the success of such behaviors, such as a rapid switch to dull coloration when attempting to gain sneak fertilizations.

Population Consequences

The consequences that sexual selection has at the population level is debated. This is because sexual selection can both enhance and depress population viability, and whichever dominates depends on the characteristics of the species and the environmental conditions.

The good genes process predicts that populations benefit from sexual selection; individuals of high genetic quality—who are well adapted to prevailing conditions—are expected to have the highest mating success, which ensures that the population consists of individuals well suited to the environment. Similarly, the model of direct benefits of mate choice—according to which females enhance their lifetime reproductive success by being choosy—predicts that sexual selection can improve population viability as more offspring are born into the population.

The Fisherian runaway process, on the other hand, can result in reduced population viability. This is because the extravagant traits and expensive choice behaviors favored by the process can use up resources needed for population growth, or reduce survival by increasing predation or parasite infection risks. Similarly, the cost of sexual selection that ensures honest signaling of mate quality in both the good genes and the direct benefit models can reduce resources available for population growth, as well as increase mortality risk. An even more serious scenario arises when sexual conflict results in males evolving traits that harm females. These can drastically reduce female fecundity and, hence, population growth rate.

The degree to which the costs and benefits of sexually selected traits influence population viability depends on a range factors. In particular, any factor that alters female fecundity may alter population growth rate, as growth depends crucially on female fecundity. However, if only males suffer increased mortality because of the costs of sexually selected traits, and the surviving males can fertilize all females in the population, then the reproductive rate of the population need not decrease. On the contrary, increased male mortality could improve population growth rate if more resources are left to females to invest in fecundity. Similarly, if the use of resources in male–male competition has no influence on female fecundity, or the growth and survival of the offspring, then the cost of sexual selection may not reduce population growth rate. For instance, if males consume other food items, or inhabit different areas than the females and the offspring, then their use of resources may not influence population growth rate.

Antagonistic coevolution between the sexes, which results in males harming females and reducing their fecundity, can severely depress population viability. Little is known, however, about such effects in nature. Populations that suffer heavily from sexual conflict and male harm are not expected to be long-lived, as selection at the population level is expected to wipe out poorly performing populations. The frequency of harmful male traits could consequently be low in nature.

In general, the ultimate effect of sexual selection on population viability is poorly known. Sexual selection can be a strong force in driving the evolution of traits, but its influence on population viability, and the factors that determine its impact, need more investigations.

Sexual Selection and Speciation

While the influence of sexual selection on population viability and extinction risk is poorly known, its impact on hybridization and speciation is better understood. Sexual selection can be an important player in species divergence and convergence, given its power to drive changes in mate recognition traits. For example, the evolution of reproductive isolation between two frog species—*Pseudacris feriarum* and *P. nigrita*—is driven by sexual selection; by reinforcing differences in their acoustic signals and preferences—through reproductive character displacement—sexual selection acts against hybrids and promotes divergence (Lemmon and Lemmon, 2010).

Sexual selection may frequently interact with natural selection in driving species divergence or convergence. An example is sympatric species pairs of stickleback, which have diverged in parallel in both sexual and ecological trait in many lakes. Species that inhabit the benthic zone of lakes have evolved to base their mate choice mainly on body size, while species that inhabit the limnetic zone mainly use male nuptial coloration as a mate choice cue (Boughman *et al.*, 2005).

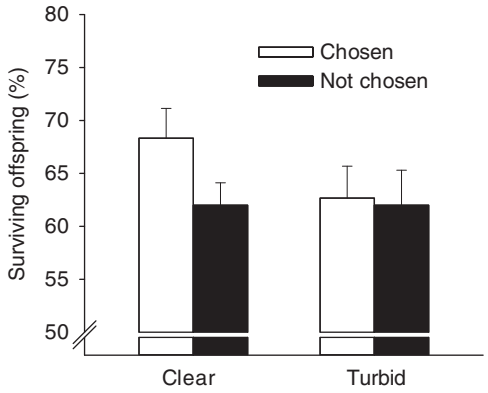


Fig. 6 Threespine stickleback females choosing between two courting males more often chose the male that sired offspring with low survival probability when the choice was made in turbid water rather than in clear water. Anthropogenic eutrophication can consequently cause maladaptive mate choices in the stickleback. Reproduced from Candolin, U., Tukiainen, I., & Bertell, E. (2016). Environmental change disrupts communication and sexual selection in a stickleback population. *Ecology 97*: 969–979, with permission from Elsevier.

Sexual Selection in Changing Environments

Environments are changing at an accelerating rate because of human activities. Whether species will be able to adjust to these changes can be influenced by sexual selection. Sexual selection can either facilitate or hinder adaptation to environmental change, depending on how it influences the number and quality of individuals reproducing in a population (Candolin and Heuschele, 2008).

If the change in the environment increases the costs of sexually selected traits, or, alternatively, decreases their benefits, then the survival or reproductive success of individuals may decrease. For instance, a shortage of food may increase the relative energetic cost of courtship activity, which can reduce survival, while a change in visibility may reduce the ability of females to evaluate ornamental traits and, hence, result in maladaptive matings. For example, female threespine stickleback that spawn in water that has become more turbid because of anthropogenic eutrophication are less able to evaluate the visual cues of males, such as their ornaments and courtship activity. This results in females increasingly spawning with males siring offspring with low survival probability (Candolin *et al.*, 2016; **Fig. 6**).

Correspondingly, if the costs of sexually selected traits decrease or their benefits increase in the altered environment, then sexual selection may improve the viability of the population. For instance, an increased variation among individuals in viability because of harsher conditions could increase the benefit of careful mate choice, which in turn could ensure that the individuals best adapted to the novel conditions have the highest reproductive success and, hence, accelerate adaptation. Moreover, if sexual selection had a negative effect on the population in the past environment, because of costs of sexually selected traits, then a relaxation of sexual selection in the altered environment could benefit the population. For instance, if antagonistic sexual selection reduced population growth rate in the past environment, a reduced ability of one of the sexes to manipulate the other sex in the altered environment could improve the growth rate of the population.

Sexual selection usually increases the mating skew, which decreases effective population size. This could increase the risk of extinction, as small populations have a higher probability of extinction than larger populations, because of higher demographic stochasticity and lower rate of adaptation. In support of this, populations introduced to island—which often involve only a few individuals—are less likely to establish themselves if sexual selection is strong rather than weak (Sorci *et al.*, 1998).

A factor that could limit the cost of sexually selected traits in altered environments, and prevent negative effects on the population, is condition-dependence of sexually selected traits. If individuals develop less exaggerated traits when conditions become harsher, then the cost could decrease and improve population viability. Currently, little is known about the impact that sexual selection has on the ability of species to adjust to rapid human-induced changes. It is clear, however, that sexual selection could play a major role given its influence on the reproductive success of individuals and their fitness.

See also: Behavioral Ecology: Kin Selection; Social Behavior and Interactions; Parental Care; Mating Systems. General Ecology: Communication

References

Arnqvist, G., Rowe, L., 2005. Sexual conflict. Princeton, NJ: Princeton University Press.
Birkhead, T.R., Pizzari, T., 2002. Postcopulatory sexual selection. Nature Reviews Genetics 3, 262–273.
Boughman, J.W., Rundle, H.D., Schluter, D., 2005. Parallel evolution of sexual isolation in sticklebacks. Evolution 59, 361–373.

Candolin, U., Heuschele, J., 2008. Is sexual selection beneficial during adaptation to environmental change? Trends in Ecology and Evolution 23, 446–452.
Candolin, U., Vlieger, L., 2013. Should attractive males sneak: The trade-off between current and future offspring. PLoS One 8 (3), e57992
Candolin, U., Tukiainen, I., Bertell, E., 2016. Environmental change disrupts communication and sexual selection in a stickleback population. Ecology 97, 969–979.
Chapman, T., 2001. Seminal fluid-mediated fitness traits in Drosophila. Heredity 87, 511–521.
Evans, S.R., Gustafsson, L., Sheldon, B.C., 2011. Divergent patterns of age-dependence in ornamental and reproductive traits in the collared flycatcher. Evolution 65, 1623–1636.
Franzke, A., Reinhold, K., 2013. Transgenerational effects of diet environment on life-history and acoustic signals of a grasshopper. Behavioral Ecology 24, 734–739.
Hine, E., McGuigan, K., Blows, M.W., 2014. Evolutionary constraints in high-dimensional trait sets. American Naturalist 184, 119–131.
Johnson, S., Candolin, U., 2017. Predation cost of a sexual signal in the threespine stickleback. Behavioral Ecology 28, 1160–1165.
Kolm, N., Amcoff, M., Mann, R.P., Arnqvist, G., 2012. Diversification of a food-mimicking male ornament via sensory drive. Current Biology 22, 1440–1443.
Lemmon, E.M., Lemmon, A.R., 2010. Reinforcement in chorus frogs: Lifetime fitness estimates including intrinsic natural selection and sexual selection against hybrids. Evolution 64, 1748–1761.
Linville, S.U., Breitwisch, R., Schilling, A.J., 1998. Plumage brightness as an indicator of parental care in northern cardinals. Animal Behaviour 55, 119–127.
Møller, A.P., 1994. Sexual selection and the barn swallow. Oxford: Oxford University Press.
Parker, G.A., 1970. Sperm competition and its evolutionary consequences in insects. Biological Reviews of the Cambridge Philosophical Society 45, 525–567.
Pizzari, T., Birkhead, T.R., 2000. Female feral fowl eject sperm of subdominant males. Nature 405, 787–789.
Ronn, J., Katvala, M., Arnqvist, G., 2007. Coevolution between harmful male genitalia and female resistance in seed beetles. Proceedings of the National Academy of Sciences of the United States of America 104, 10921–10925.
Smith, H.G., 1995. Experimental demonstration of a trade-off between mate attraction and paternal care. Proceedings of the Royal Society of London. Series B, Biological Sciences 260, 45–51.
Sorci, G., Møller, A.P., Clobert, J., 1998. Plumage dichromatism of birds predicts introduction success in New Zealand. Journal of Animal Ecology 67, 263–269.

Further Reading

Andersson, M., 1994. Sexual selection. Princeton: Princeton University Press.
Andersson, M., Simmons, L.W., 2006. Sexual selection and mate choice. Trends in Ecology and Evolution 21, 296–302.
Birkhead, T.R., Møller, A.P., 1998. Sperm competition and sexual selection. San Diego, CA: Academic Press.
Chenoweth, S.F., McGuigan, K., 2010. The genetic basis of sexually selected variation. Annual Review of Ecology, Evolution, and Systematics 41, 81–101.
Eberhard, W.G., 2009. Postcopulatory sexual selection: Darwin's omission and its consequences. Proceedings of the National Academy of Sciences of the United States of America 106, 10025–10032.
Emlen, D.J., 2008. The evolution of animal weapons. Annual Review of Ecology, Evolution, and Systematics 39, 387–413.
Hosken, D.J., House, C.M., 2011. Sexual selection. Current Biology 21, R62–R65.
Jones, A.G., Ratterman, N.L., 2009. Mate choice and sexual selection: What have we learned since Darwin? Proceedings of the National Academy of Sciences of the United States of America 106, 10001–10008.
Kokko, H., Jennions, M.D., Brooks, R., 2006. Unifying and testing models of sexual selection. Annual Review of Ecology, Evolution, and Systematics 37, 43–66.
Mousseau, T.A., Fox, C.W., 1998. The adaptive significance of maternal effects. Trends in Ecology and Evolution 13, 403–440.
Ritchie, M.G., 2007. Sexual selection and speciation. Annual Review of Ecology, Evolution, and Systematics 38, 79–102.
Servedio, M.R., Boughman, J.W., 2017. The role of sexual selection in local adaptation and speciation. Annual Review of Ecology, Evolution, and Systematics 48, 85–109.
Zahavi, A., 1975. Mate selection: A selection for a handicap. Journal of Theoretical Biology 53, 205–214.

Social Behavior and Interactions

Chelsea N Cook, Arizona State University, Tempe, AZ, United States
Noa Pinter-Wollman, University of California Los Angeles, Los Angeles, CA, United States

Glossary

Direct fitness Passing on genes to the next generation by having offspring.
Eusociality Extreme sociality defined by reproductive division of labor, collective care for offspring, and overlapping generations.
Inclusive fitness The sum of an individual's direct and indirect fitness.
Indirect fitness The fitness advantage that results from helping related individuals. Because closely related individuals likely share genes, helping kin may increase the chances of passing shared genes to the next generation.
Kin selection A hypothesized evolutionary process that results in groups of closely related individuals.
Selfish herd A hypothesis devised by W.D. Hamilton in 1971, proposing that animals live in groups to gain selfish advantages, such as not being eaten, and compete for safer spatial positions within the herd.
Trait group selection A hypothesized evolutionary process that acts at the level of the group, selecting for traits that increase the fitness of all group members.

Introduction

All animals interact with others at some point in their lives. Interactions can take the form of mating, parental care, communal living, and collective care for offspring. A wide range of species, from bacteria to elephants, live most of their lives in a social group. Living in a group provides many benefits but can also incur substantial costs. Below we expand on some of the benefits and cost of living in groups and the evolutionary mechanisms that might have led to group living.

Why Live in a Group?

Groups can benefit from social living in many ways. These benefits include avoiding predators or harsh environmental conditions, obtaining large food items, building shelter, and sharing information (Fig. 1).

When living in groups, the chances of any single individual being attacked and eaten tends to be lower than if the animal was alone. This is due to the dilution effect (Fig. 2). The dilution effect refers to the decreased chance of any individual to be attacked by a predator because there are other individuals surrounding it that might become the targets of attack. In 1971, W.D. Hamilton developed the selfish herd hypothesis which states that individuals in a social group will compete for the safest location within the group. This is observed in schooling fish and herding ungulates, where the individuals in the periphery of the group are often the subordinate individuals and are at the highest risk for being attacked by prey. The more centrally located individuals hold higher dominance positions and are less likely to be attacked. The confusion effect suggests that predators attacking a group can be confused by the prey animals scattering and moving in random directions.

Along similar lines, the detection, or many eyes effect suggests that groups can detect predators faster than a single individual because when a vigilant individual spots a predator, it will communicate this information to other group members. In some species such as prairie dogs and meerkats, individuals will "take turns" in being the sentinels who alert other individuals of approaching predators while the group is foraging. Birds learn alarm calls of other species inhabiting the same community, which allows them to flee when hearing alarms produced by another species.

In addition to avoiding predation, social living can help obtain and defend resources. Social groups can locate and process food sources more easily than individuals. In primates, the search for fruiting trees relies on the knowledge of experienced group members to find trees that a single individual might not be able to detect or consume on its own before all the fruits rot. Furthermore, Ammie Kalan and others show that chimpanzees modify their calls to others depending on what food they have discovered and how big the tree is. Song birds communicate to their group both intra- and interspecifically about new food sources and novel ways to access certain foods. A classic example are the blue tits in the United Kingdom who discovered how to open milk containers and eat the fat-rich cream layer at the top of the bottle. Opened milk bottles spread across the country. The birds were watching each other and socially learning how to get the high-calorie reward.

Once a food source is obtained, groups can be better at defending their resources than individuals. Meerkat groups defend their foraging territories from neighboring groups, small song birds can mob large raptors, and social insects defend their nests (Fig. 3).

Fig. 1 Collective food acquisition: (top) two African lions (*Panthera leo*) feeding while a hyena observes; (bottom left) two *Veromessor andrei* harvester ants carrying a large seed back to the nest; (bottom middle) a pack of wild dogs (*Lycaon pictus*) and (bottom right) a colony of the social spider (*Stegodyphus dumicola*) who both engage in elaborate collective prey attacks. Photos by Noa Pinter-Wollman.

Fig. 2 Grooming: (left) a *Veromessor andrei* harvester ant grooming its nest mate. (Right) A troop of baboons in which a female is grooming a juvenile. Photos by Noa Pinter-Wollman.

Furthermore, groups can obtain resources that individuals cannot. Wild dogs hunt in packs because a single dog cannot take down a large ungulate, but a pack can. Once a large prey has been subdued, there is enough food for the entire group who can then defend their food from other animals. Social spiders attack and capture prey that a single spider could not obtain on its own, and leaf cutting ants engage in elaborate agriculture that turns leaves into fungus gardens, which are the ants' food, a process that requires extensive effort of many individuals. Ants also cooperatively carry large food items back to the colony. Fig. 4 depicts some of these social food collecting behaviors.

Living in groups can increase an individual's lifetime reproductive success. Being part of a group can ease finding a mate (Fig. 5). Cooperative breeding, that is, when group members help each other care for young, is an advantage of living in a group. Such cooperative care can allow the reproductive pair to produce more offspring throughout their lives because of the higher survival of the offspring, and because they can reproduce more often. Andy Russell and others found that in the presence of helpers, fairy wrens lay small eggs which require a smaller energetic investment than large eggs, presumably because the helpers will assist in rearing the chicks after they hatch. Laying smaller eggs increases the reproductive female's chances of surviving into the next breeding season and producing more offspring throughout her life. Communal sea birds nest in large colonies that lower the risk of predation and increase the probability of offspring maturing to adulthood. Elephants, primates, ungulates, dolphins, birds, and social arthropods, all provide examples in which group members tend and raise the young of others in their group.

In social groups, juvenile animals often disperse from their natal habitat. Such dispersal lowers the chance of inbreeding depression which results from the propagation of recessive traits which can decrease offspring viability. However, the high costs of

Fig. 3 Group Thermoregulation: (left) Monarch butterflies (*Danaus plexippus*) aggregating presumably for thermoregulation (photo by Noa Pinter-Wollman). (Right) Honey bees (*Apis mellifera* L.) collectively fan to cool the colony (Photo by John Ternest).

Fig. 4 Dilution effect: herds of (top left) Plains zebra (*Equus quagga*), (top right) African buffalo (*Syncerus caffer*), and (bottom) wildebeest (*Connochaetes albojubatus*). Photos by Noa Pinter-Wollman.

Fig. 5 Territorial behavior: (left) *Forelius pruinosus* ants attacking a *Pogonomyrmex barbatus* harvester ant worker. (Right) Two hippopotamus engaging in a territorial fight. Photos by Noa Pinter-Wollman.

dispersal may keep young in their natal groups. Deborah Ciszek's work on the blind, hairless ground dwelling naked mole rat in Africa shows that dispersal is very costly. Within the nest a single reproductive female mates with one to three males, producing offspring that act as nonreproductive workers. These workers depend on the colony for thermoregulation and food, which are large

Fig. 6 Social interactions: (top) a pod of dolphins swimming together, (bottom left) a male and female African elephant (*Loxodonta africana*) interacting, (bottom right) *Pogonomyrmex barbatus* ants interacting near the entrance of their nest. Photos by Noa Pinter-Wollman.

tubers that animals reach by digging to them under ground. Dispersal is costly for these workers because they would have to excavate tunnels for protection and food searching on their own. Therefore, young tend to remain in their natal nest and help rear offspring. As a result, these mole rats are the only eusocial mammal on Earth.

Group living can be more efficient than living alone. For example, within a social insect colony, individuals are allocated to different tasks such as food gathering, taking care of offspring, and defending the nest. If individuals specialize on performing a certain task they might become efficient at performing it, similar to the division of labor in industrial factories. Dividing the labor allows individuals to become specialized on different jobs, and the group can accomplish more compared to a single individual that has to perform all the tasks. Which task is allocated to each individual can depend on their age, spatial position, needs of the colony, or morphological specialization.

When animals live in groups, certain tasks that would take a long time for a single individual are expedited. Collective shelter construction can range from a socially mating pair of birds digging out a cavity in a tree or constructing nests from twigs and grass, to hundreds or thousands of social insects constructing elaborate tunnel-chamber systems. A remarkable example of collective shelter building comes from social caterpillars, through research by Terrence Fitzgerald. Up to 200 madrone caterpillars will construct a single large nest from their own silk. This large, strong silk nest protects the caterpillars from predators and captures the warmth of the sun, providing protection from freezing temperatures. Termites are known for their enormous and complex nest structures which include a system of vents that maintain the temperature and gas composition within the nest. Dotting the African desert landscape, these structures jut out of the ground for miles. J. Scott Turner found that when the nest is damaged, termites know this immediately due to changes in carbon dioxide and humidity inside the nest. To repair the breach, workers use nest-building pheromones and vibrations to initiate nest building and repair, and immediately get to work carrying dirt in their mandibles. Without the work of many individuals, the construction of these elaborate and intricate structures would simply not be possible.

Group living allows animals to exploit potentially dangerous thermal landscapes. In large groups, animals can use collective thermoregulation to buffer brutal temperatures. Research from Caroline Gilbert and others explored huddling behavior in penguins. In the Antarctic, Emperor penguins huddle for up to 40% of their time to keep warm. The colder the temperatures became, the tighter the huddle is packed. When ambient temperatures dip to $-17^{\circ}C$, huddled penguins enjoyed temperatures as warm as $37.5^{\circ}C$. This protection from below freezing temperatures likely prevents body mass loss during blistering winters, as well as helping to successfully incubate eggs, which need an ideal temperature of $35^{\circ}C$. Honey bees are also excellent thermoregulators. Bernd Heinrich wrote extensively about his research on the honey bee's ability to maintain a constant internal hive temperature, even in subzero temperatures. During cold winter months, honey bees will form a cluster with the queen at the center, and shiver to keep warm. To ensure that everyone stays at a comfortable temperature, the workers will actually cycle through the cluster; as a bee becomes too warm in the middle, she moves toward the cooler surface, and when she is too cold, she begins to move inside again. During hot summer months, honey bees actively cool their colony by collecting water to spread on brood comb for evaporative cooling, and fanning to move hot air out of the colony, allowing cool air to move in. Chelsea Cook and Mike Breed found that honey bee fanning is an obligatory group-performed behavior; bees that are isolated will rarely fan, but they will begin to fan in small groups of just three bees. In Fig. 6, monarch butterflies cluster to keep warm during migration.

Why Not Live in a Group?

While there are many benefits to groups living, sociality also comes at a cost. When organisms live close together, they may compete over resources such as space, food, and mates. Further, a large group may help in avoiding predators, but predators may

Fig. 7 Access to mates: (top) Magellanic penguins (*Spheniscus magellanicus*) living in a colony and (bottom) Northern elephant seals (*Mirounga angustirostris*) in a rookery aggregate to gain access to mates. Photos by Noa Pinter-Wollman.

detect a group more easily than individuals, simply because of its size. If a predator can more easily detect the group, this may lead to more frequent attacks by more predators than any one individual experiences.

The increased competition for mates within a social group could decrease the chances that an individual reproduces. There could be lost breeding opportunities because of social hierarchies. Animals that fall on the lower end of the hierarchy may be beaten out of opportunities to breed. In extreme cases, individuals can become essentially sterile. In eusocial insects like ants, some wasps, and some bees, female workers rarely produce successful offspring. Even when eggs are laid, other workers actively police laying workers by eating their eggs.

Living in a group with many individuals can lead to the rapid spread of disease. This is especially true if the interactions between individuals are frequent and long lasting. Groups have developed collective methods to deal with the spread of disease. In mice, infected individuals decrease their interactions with other mice, which keeps a disease from spreading widely through the population. Bees exhibit a myriad of hygienic behaviors, including inducing a behavioral fever when sick or inspecting and discarding infected larvae. When bumble bees are infected with a conopid fly, or honeybees are infected with chalkbrood fungus or the *Nosema* microsporidian, they will regulate their temperature, essentially causing a fever for the hive that helps fight off the infection more effectively.

To decrease competition, individuals in social groups may kill other individuals. Craig Packer's work shows that female lions tend to aggregate and be more social than males who have only fleeing interactions with the females, which involve cofeeding and mating. When a new male joins the pride it may kill young cubs that were sired by another male. This infanticide causes the females to become reproductively active, as they are no longer nursing or caring for the cubs, and can then mate with the new male. If social groups remain stable and no new males join the pride, the rate of infanticide significantly drops. Competition over limiting food and water can become so intense that herbivorous insects resort to cannibalism. In migrating locusts, it was proposed by Sepidah Bazazi and others that the mass movements of hundreds of thousands of insects is initiated and perpetuated by biting behavior: when individuals get too close and touch, they begin biting each other, which forces them to move out of the way, but instead it leads to interacting more with other locusts, which leads to more biting. If movement does not happen quickly enough after being bitten, the animal may be eaten.

While there are many contraindications of being in groups, animals have adapted methods to mediate these issues. For example, to combat spreading disease, animals can groom each other (**Fig. 7**). Grooming not only solidify social bonds, it also plays a practical role of reducing potential pathogens from entering and spreading through the social group. We see grooming across animal taxa; from baboons to ants.

How Do Animal Groups Move?

One of the most striking behaviors animal groups engage in is collective movement. Moving groups include herding ungulates, schooling fish, and flocking birds. In these collectively moving groups, there is no leader, that is, no single animal is in charge of where the group is going. Marie Bourjade and other show that in herds of Przewalski horses certain individuals will prime the collective movement, but the actual decision to move is collectively agreed upon by a majority of individuals in the group. Building off of Karl von Frisch's pioneering work on the waggle dance, Tom Seeley has elucidated how honey bees choose new nest locations. In honey bee swarms looking for a new home, scouts will leave to find a suitable nesting place, then return to the swarm to signal the quality and location of the new found nests using an elaborate "waggle dance." More scouts will leave to examine the new potential nests, based on the advertised quality, and if they like it, they will return and add their "voice" to the other waggling scouts to recruit more bees to their preferred location. Once there are enough bees waggling to a particular site, that is, the swarm has reached a democratic consensus, it takes off. This is an example of quorum decision making, where a subset of individuals participates in making a decision about movement.

Individual animals within the group often base their movement decisions on very local information. This information includes what neighboring individuals are doing, for example, how far they are and in which direction they are moving, and the ecological context in which the behavior occurs. Both theoretical and empirical experiments from Iain Couzin and others show that using only local information, fish, for example, are repelled from each other only within very small distances. When the fish get farther away from each other, they attract. Speed and orientation within the group emerges from these simple rules which allow them to move away from predators without actually receiving direct information that a predator is there. While a fundamental assumption of these studies is that individuals tend to be behaviorally homogenous, it is becoming apparent that individuals vary in their behaviors and in how they respond to their neighbors. Such behavioral heterogeneity can lead to costs, for example, a group of both adults and juveniles will only be able to move at the speed of the youngest animals. In contrast, behavioral variation can expedite the response of a group to prey they capture, as seen in social spiders.

How Did Group Living Evolve?

Altruism and Cheaters

Group living requires cooperation. In some ways, however, cooperation goes against the idea of individuals working to maximize their own fitness. Altruism is defined as behaving in a way that increases the fitness of another individual at the expense of decreasing one's own fitness. Individuals who act selfishly maximize their own fitness, but groups of cooperators maximize the fitness of everyone in the group. However, in a group of cooperative individuals, selfish individuals can arise and take advantage of pro-social behavior of the altruistic individuals. The balance between altruistic individuals and cheaters is a strong force in the evolution of group living. Much theoretical work has attempted to explain why animals behave altruistically. One popular solution has been reciprocal altruism in which an individual will help another who will reciprocate and assist the helper at a different time. Interestingly there is no compelling empirical evidence for such reciprocal altruism. Other theories suggest that the cost of altruism to one's own fitness are not as great as they seem because the altruistic individual will benefit indirectly by being related to those it helps or because the group in which it lives will persist because of the altruistic behavior.

Kin Selection

Individuals can gain indirect fitness benefits by assisting with the production of closely related offspring. W.D. Hamilton hypothesized that individuals can give up their own direct fitness to enhance the fitness of another related individual. Why would animals ever forgo reproducing? When dispersal is very costly, closely related individuals will spatially cluster in groups. This means that group members share a high proportion of genes. For example, siblings share an average of 50% of their genes and offspring are 50% related to their parents. Depending on the relatedness of an individual to its group members, the costs and the benefits of assisting them, and environmental factors (such as difficulty of finding mates or a suitable territory), it may be advantageous to help parents rear one's siblings instead of dispersing and raising one's own offspring. If the genetic relatedness and the indirect benefits an individual gains when helping outweigh the costs of helping, sociality will emerge, this is known as "Hamilton's rule."

A critical component of kin selection is acting cooperatively with closely related individuals. How can one be certain that they are closely related to another individual? W.D. Hamilton proposed a hypothesis, popularized by Richard Dawkins as the "green beard" hypothesis, which postulates that individuals can perceive a shared allele through some phenotype (the hypothetical "green beard"). An empirical example of a "green beard" comes from Laurent Keller and Kenneth Ross on fire ants. In fire ant colonies, queens that are homozygous at a particular gene locus are killed by workers that are heterozygous in the same locus but heterozygous queens are allowed to live. The workers differentiate between the homozygous and heterozygous queens according to their smell, which is considered to be their "green beard."

Trait Group Selection

Trait group selection proposes that natural selection acts at the level of the group, in addition to the level of the individual. Thus, when one group does better than another, it will persist and the traits that allowed it to outperform other groups will be selected for. For example, groups that are more cooperative and collectively produce more offspring than others will outcompete groups that do not cooperate to increase the overall fitness gains. Work from Jonathan Pruitt and Charles Goodnight tests this hypothesis. In social spiders, the behavioral composition of a group is selected at the group level. The behavioral composition of a group determines its collective prey capture success and therefore the survival of the group. Certain behavioral compositions result in higher group survival in particular environments but not in others. Furthermore, groups' composition change over generations to fit the environments in which they are placed. In social animals where group living is critical for survival, group selection may have strong implications of which individuals persist through evolutionary time.

Overall, there are many levels upon which selection can act. While genetic and individual level selection are crucial, it is important to explore how group living may influence individual behavior, and how the survival of the individual depends on the survival of the group.

Applied Methods in Animal Behavior

Social behavior can be observed in both a natural setting, such as in the field, and in the lab where researchers can control for environmental influences. Scientists studying behavior have developed observational techniques that can be used in both a field and laboratory setting.

One of the foundational texts about these methods is Jeanne Altmann's 1974 paper "Observational Study of Behavior: Sampling Methods." In it, she outlines several, now common, methods by which to collect data that is suitable for both the field and the lab. One main question asked is what type of behavior is the researcher interested in studying. Importantly, care must be taken to avoid any influences the observer might have on the behavior of the animals and what Altmann calls "ad lib" sampling—where observations end up in field notes rather than structured observations that are taken at specific times, locations, situations, etc.

When designing a study of social behavior, one must establish whether the study will be based on experimental manipulations or observations.

Manipulative studies attempt to place the animal into a certain context to control the myriad of influences on the animal's behavior. These studies can take place both in the field and the lab. For example, providing primate groups with preferred food at particular locations to examine collective movement decisions; removing ant workers that engage in a certain task (such as nest maintenance or foraging) to examine task allocation processes; placing honey bees into a petri dish with a nonnestmate to evaluate aggression; and modifying the behavioral composition of social spider groups in the lab to determine how group composition influences collective behavior.

Social behavior can be studied by collecting data to characterize how the group as a whole behaves, or identifying a focal individual to keep track of during social interactions. Most sampling methods can be characterized as either a "state," or as an "event." States generally characterize the duration an animal is performing a behavior, while events are instances of behaviors. Group behaviors can be recorded as proportion of individuals performing a behavior in the group, for example, using scan sampling. Scan sampling is when the behavior of each group member is noted as an event at certain, predetermined, time intervals. Focal observations can be made when you have a single animal, while watching a group but focusing on one individual with that group, or while watching a group interact with another group. Characterizing focal individual (or individuals) behavior can be done in many ways, including counting all occurrences of one behavior, counting the order in which a behavior happens, or counting how long an animal spends performing a behavior. These simple experiments form the foundation of studying complex social behavior.

While Altmann and others provide an essential methodological framework to observing behavior, theoretical approaches have proven advantageous in developing hypotheses before and during experiments. Especially in social behavior, using mathematical models to predict how an animal will behave has given the field some of its most groundbreaking work. Here we discuss several important theoretical frameworks that contributed greatly to the study of social organization.

Theoretical Advances in Studying Social Behavior

Game theory has made a substantial impact on our understanding of social evolution. Game theory is essentially a mathematical puzzle, originally developed by economists, where during an interaction, one individual gains when another loses. The simplest model can be explored as a "zero sum" game, in which the sum of the exchange is zero (e.g., one animal benefits one food item and the other loses one food item). A popular application to animal behavior was led by George Price and John Maynard Smith who published "The Logic of Animal Conflict" in 1973, where game theory could be used to predict outcomes in animal interactions. The outcome is then translated into fitness and based on how individuals in a population will behave to enhance their fitness, the game theoretical model can help identify evolutionarily stable strategies (ESS). An ESS identifies a behavioral

strategy that will be selected by natural selection in a stable environment, Of course, environments are always changing, and therefore, ESSs are also shifting.

An example game theory test to explore cooperation is a game called the prisoner's dilemma. The thought experiment is as follows: two people are arrested for a crime. The two people cannot communicate. The police start to manipulate the criminals, stating that criminal #1 is going to give up information on criminal #2, and vice versa. The resulting potential outcomes could be:

(A) #1 gives up information on #2, but #2 keeps quiet. #2 goes to prison for 3 years, but #1 will be set free. The converse (#2 giving information but #1 staying silent, #2 goes free) is also true.

(B) If they both give up information on each other, they both go to prison for 2 years.

(C) If they both keep quiet, both will only go to prison for 1 year.

This experiment becomes especially interesting if the pair of actors in the situation know the outcome, and if they interact in similar ways over time. Examples of behaviors that can emerge from such repeated interactions include punishment from those who potentially suffered in the situation, or policing of individuals who do not cooperate, like when honey bee workers eat eggs laid by cheating sisters.

Payoff Matrix:

	Criminal #2 cooperate	*Criminal #2 defect*
Criminal #1 cooperate	Prison for 1 year	#1 Prison for 3 years #2 goes free
Criminal #1 defect	#1 goes free #2 prison for 3 years	Prison for 2 years

Optimal foraging theory is used to predict how animals will forage. The main predictors that go into optimal foraging are how much energy the food provides and how much energy the animal must use to collect the food. For example, Herring gulls will pick up mussels from the rocky coastline, fly into the air, and drop the mussels to crack the shells. Bigger mussels may yield more energy; however, the gull must then expend more energy to fly higher into the air to drop it a farther distance to crack it open. Living in a social group can make this relatively simple model more complicated. When a bumble bee is foraging, for example, is she cueing in on her own hunger, or the resources in her colony? Bumble bees use the amount of resources in the colony to make decisions about foraging versus their own satiation; if the colony needs food, a worker will collect more food, even if she has eaten recently. Social animals must collect different information and make decisions based on the balance of individual and group needs.

The organization of social groups has always fascinated and puzzled scientists. From herding antelope, schooling fish, and murmurating starlings, the ways in which animals organize themselves has been subject to many experiments. As discussed above, animal behaviorists have used the ideas of indirect and local information to explore how animals move as groups. In many complex systems, organization emerges from local behavioral rules and interactions. The study of these local rules has benefited from the use of Agent-Based-Modes (ABM). In these models, each individual is provided with a set of behavioral rules, such as which way to turn when it is a certain distance from another group member. Simulations are then run over many time steps and for various sets of parameters and situations. Comparing the simulation outputs for different parameters and with observed behaviors can guide experiments to test predictions generated by the models. For example, in an agent based model that was developed to examine local rules that lead to collective motion, Iain Cousin and colleagues in 2002 defined an attraction, repulsion, and orientation zones around a simulated agent. If another group member enters the repulsion zone of an agent, the agent will move away, if another group member enters the attraction zone, the agent will move toward the other individual, and if the other group member enters the orientation zone, the agent will orient in the same direction as its neighbor. By modifying the sizes of these zones, the researchers were able to reproduce movement patterns that can be seen in fish schools in different situations, producing predictions that can then be tested empirically. Manipulating the environment of fishes or using fish robots, whose behavior can be manipulated to test the response of school members, are some of the methods researchers can use to test the model assumptions and predictions.

In recent years, the use of social network theory has advanced substantially the way we study social behavior. A network framework allows the investigation of social interactions in the context of the group as a whole. Each individual is depicted as a node (shape) connected to other individuals with whom it interacts with an edge (line). Some network measures quantify the role of individuals, for example, how many others each individual interacts with (degree centrality), which individuals connect the most other individuals to one another (betweenness centrality), etc. Other network measures quantify medium scale structures, for example, identifying social cliques (modularity) and examining interactions with friends of friends (clustering coefficient and triads). Finally, at the group level, network measures can provide information, for example, on whether there are more connections in the group than one would expect at random (network density), or whether all group members tend to have similar interaction patterns or if some individuals tend to interact more than others (degree distribution). Social network analysis has allowed researchers to uncover what happens to the social interactions in a primate group when certain individuals are removed, revealing that older males take the role of policing (Flack *et al.*, 2006). Furthermore, a network approach has been used to identify individuals with high influence on social stability and whose removal would lead to social breakdown. Identifying such

individuals is important for preserving social structures of endangered species and when targeting individuals for vaccination, or when trying to prevent the spread of disease.

Overall, these applied theoretical methods allowed researchers to both develop and test hypotheses in social behavior.

Conclusion

Social behavior shapes individual, population, community, and ecosystem dynamics. The relationships between the benefits and costs associated with group living for each species and in any particular environment will determine whether sociality arises. There are still many open questions about why and how social behavior evolved across such a wide range of species, from bacteria to humans.

See also: Behavioral Ecology: Kin Selection; Parental Care; Sexual Selection and Sexual Conflict; Mating Systems. General Ecology: Communication

Reference

Flack, J.C., Girvan, M., De Waal, F.B.M., Krakauer, D.C., 2006. Policing stabilizes construction of social niches in primates. Nature 439 (7075), 426–429.

Further Reading

Alexander, R.D., 1974. The evolution of social behavior. Annual Review of Ecology and Systematics:. 325–383.
Couzin, I.D., Krause, J., 2003. Self-organization and collective behavior in vertebrates. Advances in the Study of Behavior 32, 1–75.
Dawkins, R., 1976. The selfish gene. Oxford: Oxford University Press.
Gardner, A., West, S.A., 2014. Inclusive fitness: 50 years on. Philosophical Transactions of the Royal Society B: Biological Sciences 369 (1642), 20130356.
Heyes, C.M., Galef Jr., B.G. (Eds.), 1996. Social learning in animals: The roots of culture. San Diego, CA: Elsevier.
Hoppitt, W., Laland, K.N., 2013. Social learning: An introduction to mechanisms, methods, and models. Princeton: Princeton University Press.
Seeley, T.D., 2010. Honeybee democracy. Princeton: Princeton University Press.
Strassmann, J.E., Queller, D.C., 2011. Evolution of cooperation and control of cheating in a social microbe. Proceedings of the National Academy of Sciences 108 (Suppl. 2), 10855–10862.
Wilson, D.S., 1975. A theory of group selection. Proceedings of the National Academy of Sciences 72 (1), 143–146.
Wilson, D.S., 1980. The natural selection of populations and communities. San Francisco, CA: Benjamin/Cummings.
Wilson, D.S., Wilson, E.O., 2008. Evolution "for the Good of the Group": The process known as group selection was once accepted unthinkingly, then was widely discredited; it's time for a more discriminating assessment. American Scientist 96 (5), 380–389.

Thermoregulation in Animals: Some Fundamentals of Thermal Biology☆

Udo Gansloß**er and Gianna Jann,** Jena University, Jena, Germany and Greifswald University, Greifswald, Germany

Glossary

Allen's rule A biogeographical rule stating that, the farther away from the equator a species/subspecies occurs the fainter or lighter colored it should be.

Bergmann's rule A biogeographical rule stating that the farther from the equator a species/subspecies occurs the bigger in size it should be.

Brown adipose tissue (BAT) A specific form of storage tissue. Within cells of BAT lipids are not stored in one, large, microscopically translucent vacuole but in many small ones, which in light microscopy causes a brown stain of the endoplasma. BAT is used in nonshivering thermogenesis to produce extra heat, for example, while awaking from hibernation.

Ectothermy A physiological condition in which an organism's body temperature is not significantly different from ambient temperature. No thermoregulatory processes are employed.

Endothermy A regulatory condition in which an organism is capable of regulating its core body temperature significantly independent of external temperature.

Gloger's rule A biogeographical rule stating that the closer to the equator a species/subspecies occurs the large and more slender its body appendages (ear pinnae, tails, beaks, limbs) should be.

Heat shock proteins A family of proteins whose genes are activated by a heat shock factor. The proteins then are produced and act as molecular controllers of correct folding of other protein molecules.

Homeothermy A special case of ➔ endothermy. Homeothermous organisms regulate their body temperatures, by internal mechanisms of heat production and cooling, within a very narrow range, often within one degree centigrade.

Nonshivering thermogenesis (NST) A set of cellular/organ-physiological processes that allow heat production without muscular contraction (i.e., shivering). NST includes, among others, the action of ➔ UCPs, or the activation of ➔ BAT.

Thermoneutral zone (TNZ) A range of ambient temperatures, species/population-typical, within which range of outer temperature the organism needs the least energy in basal metabolism. If ambient temperature is below or above TNZ the animal needs more energy for basal metabolism per unit of body weight, for cooling, heat production, sweating etc.

Torpor A state of physiological inactivity to reduce energy expenditure. Can be daily (during cold nights in continental areas, or during noon heat in arid countries) or annually, for example, hibernation. Torpor can be triggered by active lowering of metabolic rate and thus body temperature or, especially in ectotherms, as a passive consequence of dropping ambient temperature.

Uncoupling proteins (UCPs) A family of proteins that, in low temperature, act on the mitochondrial membrane. They cause proton leakage and thus separate cellular dissimilatory processes from ATP synthesis, allowing extra heat production directly.

Heat exchange between animals and their environment is basically restricted to their body surface. Direction and intensity of this exchange are governed by the temperature difference between body surface and environment (the latter temperature being called ambient temperature Ta), and can be of three different types: radiation, conductancy or evaporation. Radiation is characterized by the two bodies not necessarily being in direct contact, it is being achieved simply by electromagnetic waves. Intensity of radiation in heat exchange depends on the difference between temperatures, and increases in the fourth power, thus heat radiation very rapidly increases with any small increase in temperature difference. Contrary to commonly expressed expectations radiation heat loss in animals seems not to be dependent on the color of their fur or plumage. Body color may, however influence heat gain. Conductance, contrary to radiation needs a direct contact between both bodies, and is not only dependent on the difference in temperature but also by the specific thermal conductancy constant of the tissue(s) in question. This thermal conductivity constant for most tissues is bigger than for air, but far less than for water; water has a conductivity about 23 times that of air, fat has a conductivity about 8 times that of air, and fur between 0.9 and 2.4 times that of air. In absolute terms and values for substances are expressed in J cm^{-1} s^{-1}°C^{-1}, and some values are:

Air: 0.23×10^{-3}
Fat: $1.47–2.09 \times 10^{-3}$
Epidermis: $1.59–2.09 \times 10^{-3}$

☆*Change History*: March 2018. Gianna Jann B Ed contributed in literature search and compilation.

This is an update of U. Ganslosser, Thermoregulation in Animals, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 3550–3557.

 doi:10.1016/B978-0-12-409548-9.11170-4

Corium: $3.48–3.68 \times 10^{-3}$
Water: 6.28×10^{-4}

In terrestrial animals, conduction does not play a very important role for temperature exchange because it can only occur over the rather limited area of the feet or other body parts being in direct contact with the soil. This situation is totally different for aquatic animals. A specific form of conductance is called convection. This relates to temperature exchange between body surface and the immediately adjacent layer of air. The thicker this boundary layer, the less heat is transferred across it. The higher the density of this boundary layer, the larger the heat transfer. Should the boundary layer itself be in movement, then a turbulent movement leads to a higher convective coefficient than a laminar one. In general, however layers of still air (or fluid) have a much lower convective heat transfer. Thus erected plumages or furs, which lead to a still boundary layer, are almost twice as good for isolation as compressed ones.

The higher the wind speed, the thinner the boundary layer, thus the bigger the heat transfer. Heat transfer from a body surface to the environment is only possible as long as Ta is lower than the surface temperature. The only way to transfer heat from an animal into the environment under conditions of Ta being higher or similar to the temperature of the surface is by evaporation. A characteristic value for evaporative heat loss is called latent warmth. This is the amount of heat that is involved in transferring a substance from, for example, fluid to gaseous condition. Latent warmth is indirectly proportional to ambient temperature. Two examples from thermal biology: For evaporating 1 g of water at a Ta = 35 °C, you need 2413 J. An adult person normally loses about 1 L of water per 24 h via evaporation. For this, you lose about 2400 kJ of energy, which would be sufficient to freeze almost 5.8 L of boiling water! Evaporation thus is a very effective means of cooling, however water in many hot areas is a scarce resource, and behavioral adaptations to combine effective cooling and save water at the same time are much in need (see below). Most organisms are subject to changing ambient temperatures be they daily seasonal or unpredictable. Adaptation to these changes thus is one of the most important demands in physiological ecology. Basically there are two successful strategies as in other areas of physiological ecology: Conformers or regulators. Conformers in our case are organisms that have only a very limited capacity for internal heat production. They regulate their body temperature by staying in acceptable ambient temperatures. This rather limits their geographical and other ranges, but is energetically beneficial and can be kept with a reduced foraging activity. These so called ectotherms, sometimes also called poikilotherms, are typically invertebrates and so called lower vertebrates. Endotherms, on the contrary, are temperature regulators. They produce enough heat from their metabolism to keep their body temperature within a given range. This is however rather demanding on their nutritional needs, their food intake has to be 5–10-fold increased over that of ecotherms. Birds and mammals but also some reptiles, tuna fish, some insects (bees, moths) and even some flowers (Lotus, Philodendron) are endotherms. Among these, many species are indeed capable of regulating their temperature quite within a narrow range similar to humans, these are the homoiotherms. Many others, especially smaller species below 10 kg of body weight, are capable of reducing their body temperature daily, seasonally or spasmodically, in accordance with external conditions. These species are called heterotherms and shall be specifically discussed later.

In terms of temperature tolerance, eurythermal animals are those that can be active over a wide range of temperatures—many desert species are active at least between 8°C and almost 40°C.

Stenothermal species are active only over a small range of temperatures, and this also is the case for many ectothermal forms (e.g., polar fish can only exist between −2°C and +4°C, thus have a thermal tolerance of only about 6°C).

The range of temperatures at which we can find animals lies approximately between −5°C and slightly over +50°C (desert and hot springs insects reach about 48°C, hot-spring protists around 52°C). The hottest life forms may be hot springs bacteria at around 90°C.

Biochemical and Cell Biological Aspects

The so called Q10 value, the value that describes the change in reaction velocity at a temperature increase of +10°C, is normally regarded as being around 1.0, which means that an increase in 10°C would double the reaction velocity. This definitely in the case in most physical and inorganic chemical processes, however it is not necessarily true for organic chemistry or biochemistry. For biochemical (e.g., enzyme-modulated) and physiological processes Q10 often is between 2 and 3, in some cases (e.g., crustacean gill movements, thermal induction of insect diapause, or hemoglobin coagulation) for above 3.0. Another complication in this is the fact that Q10 is not independent of the temperature at which it is measured and an increase of Tb from 30 to 40°C might have a totally different value than an increase from 5 to 15°C. For an animal living over a temperature range of 30°C, this could mean differences in reaction velocities or decreases. Only very few obviously strongly selected processes in animals, such as the function of the biological clocks, seem to have Q10 around 1.0.

In species with rather narrow ranges of body temperature (endotherms and especially homeotherms), there is normally also a rather narrow "physiological comfort zone," called thermoneutral zone (TNZ). Within this range of ambient temperatures, oxygen consumption and basal metabolic rate (BMR) are lowest and curve function of metabolic parameters against ambient temperature forms a trough. Below TNZ, and above, the relationship between BMR/O_2 consumption and ambient temperature rises, mostly due to thermoregulatory processes.

Changes in biochemistry with temperature, and adaptive or regulatory processes, can be manyfold. Changes in effective concentration or activity of enzymes, for example, by phosphorylation, can lead to better activities at lower temperatures, changes

in compartmentalization of cells by mobilizing membrane—limited areas can change diffusion processes, production of cryoprotectants under hormonal control, changes of concentrations of ions altering pH, and many of these have also been found in animals such as cold-adapting insects.

These changes all can occur over a time span of already a few hours. Over a period of several days or weeks, different are found, such as an increase of mitochondrial protein content (i.e., increase of enzyme concentrations) in colder conditions or changes in isozymes. Thus in several cypriniform fish, the composition of myosin ATPase is changed by changing the light to heavy chain composition in the cold, which leads to a higher muscular activity. In trout, there are different forms of acetylcholinesterase at $+17^{\circ}$C versus $+2^{\circ}$C Tb, but at Tb of $+12^{\circ}$C both forms are present—giving these fish the possibility to migrate to both colder and warmer places. Over the long-term, evolutionary span of time, thermal stabilities of enzymes and other proteins can be altered genetically and it is often found that thermal stability is greater in eurytherms than in stenotherms. Enzymes from animals (species as well as populations) living at lower Tb seem to have a more open, less rigid, structure with fewer and weaker bonds, than these from a higher Tb!

Enzymes of animals from colder regions also have higher substrate turnover rates, and lower values for activation energy, etc. These fine-tuned biochemical changes even are found in sister species with almost identical DNA. Thus, in two species of gobiid fish (Genus *Gillichthys*) only four nucleotides may be different but isozymes can have different folding conformations.

Among the most temperature-sensitive tissues seems to be the nervous system. Neuronal functions, in terms of chemical and electrical conductivity, refractory periods etc. seem to depend strongly on temperature. Thus, action potentials, and the efficiency to create and forward them, greatly profit from increased temperature. But the optimum temperature here seems to be rather in a narrow range. Even slight increase of body temperature, for example, light fever can already cause severe impairment of neuronal functions.

Membranes in cells have a particularly high necessity to adapt to different temperatures because of their lipid structure. The proportion of short-chain and unsaturated fatty acids determines the viscosity of the biomembrane, and higher proportions of these can be found in cold-adapted animals. This even works the other way round: Lizards fed on a diet of highly polyunsaturated fatty acids afterward preferred temperatures of up to 5°C less than controls. Brain membranes of arctic fish have much higher proportions of polyunsaturated lipids and higher number unsaturated bonds per lipid chain. The activity of specific enzymes, cold desaturases, produces these changes in phenotypical adaption, and desaturase activity is increased up to 30-fold after transfer of the fish into the cold. To make matters more flexible for the animals and more confusing for the researchers, these changes are different from tissue to tissue. Biochemical changes can also help to produce body heat. In so called nonshivering thermogenesis (NST) biochemical pathways are operated that, for example, decouple the mitochondrial proton chain reaction from ATP synthesis thus leaving "only" heat as the product of catabolic metabolism. The best studied of these NST processes is in the brown body fat of mammals going into winter dormancy (see below).

An important biochemical process of temperature adaptation is mediated by "uncoupling proteins" (UCPs). Due to their action in mitochondrial synthesis, they cause proton leakage and generate heat directly. This process seems to occur in several plant species but also in birds and mammals (from bats through humans to cetaceans).

Leaving the adaptions to the cold there are also biochemical and cell biological processes to protect the animals from heat.

So called heat-shock proteins are being induced in production as soon as an animal comes into hot environment and these again are being produced at temperatures that depend on the normal range at which the population lives. In a commonly discussed model, this is started by a heat shock factor that starts to become active by trimerization and interacts with the promotor region of the heat-shock protein gene. Also, due to heat shock, ribosomes are detached from existing mRNAs, thus being available preferentially for translating the newly arriving heat-shock protein mRNA. Once heat-shock proteins are available, they seem to act as molecular chaperones controlling the correct folding of other proteins.

Inflammatory processes often are linked to increases in body temperature not only in endothermic species. Behavioral mechanisms of temperature regulation, by seeking warmer surroundings are often cooccurring with the development of inflammatory processes, allowing an increase in body temperature, hence fever. Experiments in rats do again demonstrate a high degree of variability in behavioral reactions with low doses of bacterial liposaccharides (LPS, inflammatory agents) causing warmth seeking, high doses however cold-seeking behavior. In snakes, an immunological challenge by injection of these LPS caused a change in shelter usage-corn snakes after injection of LPS spent more time in the open, instead of sheltering, compared to a control group of saline-injected conspecifics.

Physiological Adaptations and Regulatory Processes

Regulatory and Sensory Mechanisms

In order to regulate thermal conditions an animal first needs to become aware of possible thermal problems. The sensory and regulatory capabilities associated with this are best studied in mammals and reptiles. We find peripheral temperature receptors, as heat and cold sensors in the skin of an animal. These sensors are not evenly distributed. In mammals, the concentration of heat sensors is particularly high around the scrotal and inguinal area, whereas in and around the face there are mostly cold receptors. All peripheral temperature receptors studied so far are primary sensory cells. Color receptors often are branched into many small endings, with a high concentration of mitochondrial and glycogen particles. Thus we can infer a high metabolic activity.

Temperature sensors in mammals are mostly found at the boundary between corium and epidermis. Heat receptors in mammals seem mostly to be free nerve endings. Additionally to peripheral receptors there are sensors for temperature both in the central nervous system and in several other internal organs, such as the abdominal region and skeletal musculature.

Afferent nerve connections then connect sensory cells to the hypothalamus, however there are already some modifications and processing steps, at least with regard to heat receptors from the scrotal/inguinal area, within the spinal tract. In the hypothalamus we can find both thermosensitive and thermoresponsive neurons, and their input is then processed to influence both effector organs for heat dissipation or production, and to higher centers of thalamus and sensory cortex. Obviously the types of processing or reacting cells are different. Some seem to react proportionally, for example, increase firing rates proportionally to increase (or decrease) in temperatures others increase their firing rate when a certain set point is reached. Desert iguanas, for example, have set points of 36.4°C as the low set point for going into the sun again, and 41.7°C for going into shade. Whereas in normal cases, central thermoreceptors seem to take priority over the input from peripheral receptors.

In comparison to the important role of sensory and neural mechanisms the role of hormones in temperature regulation seem to be rather nonspecific, by increasing metabolic rate (e.g., thyroxine, adrenaline, or prostaglandins).

Excessive Heat, How to Cool and Conserve Energy

Heat production in endotherms is a function of basal metabolic rate (BMR) and as we know at least since the famous publications of M. Kleiber, this is dependent on a factor of body weight by the power of ¾ (or 0.75). If this were not so basal metabolic rate would directly increase proportional to the increase in body weight, then a ten ton elephant (at least some extinct species were as heavy as that) would need a skin temperature of >100°C to radiate his excessive metabolic heat, and a 1 kg rat would need 20 cm of fur length to insulate its body sufficiently.

Animals in hot areas would do well by reducing their BMR, because then metabolic heat contributes less to the amount of temperature they have to get rid of. Indeed, small desert rodents in many cases have BMR that reach barely 50%–75% of the expected values (as calculated from their body size). *Gerbillus nanus*, a small gerbil, only gets about 51%, the spiny mouse species, *Acomys russatus*, just 56% of their calculated values, for example.

Large species such as camels, are for better off in these respects. Not only are they capable of migration and other behavioral adaptations (see below), their large bodies also enable them to cope better from a physiological view point. Large bodies have relatively smaller surface areas, which means that they take up less heat from radiation. Large bodies also can store heat effectively. The larger an animal the more water (in its body fluids) does it contain, and these can store more heat. Camels, even under conditions of ad lib access to drinking water during the course of the day increase their body temperature from 36°C to about 39°C. Should they however not have enough water (thus must conserve water from evaporation), then their body temperature, over the course of the day, increases up to 41°C, and during the night they radiate heat again until their Tb is as low as 34°C. This span of 7°C helps a camel of 500 kg to conserve at least 5 L of water that would otherwise be necessary for evaporative cooling. The record in heat tolerance for large desert-living mammals is kept by Grant's gazelles, whose Tb can increase up to 46.5°C. However animals under these extreme conditions need some cooling system at least for their brains and other sensitive organs. In desert antelopes there is an effective cooling structure, a heat exchanger at the basis of their brain case. Here, the sinus cavernosus, a dilation of the nasal vene, is surrounded by an arteric rete mirabile in the Art. carotis. By means of this, "hot" blood from the heart is cooled down by cooler venous blood from the nose (where some evaporative cooling took place) before entering the temperature-sensitive brain.

Countercurrent heat exchanger are also found in ectotherms from hot environments. Thus, some desert reptiles (*Phrynosoma* spec, *Holbrookia* spec) operate heat exchanges between carotids and jugular veins while sun-basking. They only protrude their heads over the sand, and keep a temperature gradient between head and abdomen by a countercurrent system in their throat and neck areas.

Excessive secretion of saliva, and distributing it over the fur by grooming, is a widespread cooling mechanism not only in rodents but also in marsupials, for example, Red kangaroos. Nasal glands seem to contribute to evaporative cooling in Domestic dogs and probably other species of nonsweating mammals.

Heating, Isolation and Freeze Tolerance

In endothermal species the graphic curve describing relationships between heat production (or O_2 used for metabolism) and ambient temperature follows a characteristic trough form. Both with ambient temperatures below a so called lower critical temperature and above a so called higher critical temperature, O_2 consumption increases. The range of temperature between the lower and upper critical temperature values is called thermoneutral value for species or populations adapted to a certain environment.

As soon as an animal's ambient temperature falls below the lower critical point, heat has to be produced by physiological and biochemical means. One possibility we all know is shivering thermogenesis, caused by nonsynchronous contraction of skeletal muscles. The amount of heat produced by a given amount of muscular activity depends on body size, for example, 1 m V of muscular activity in a 390 g crow produces about 7.8 times as much heat than in a 14 g songbird. These differences are a consequence of the differing surface: volume ratio in larger versus smaller animals. At least in birds these phenomena are further complicated by different mechanisms of cold adaptation. On the one hand, electromyographic activity decreases with cold acclimation, on the other hand cold-adapted birds tend to have larger pectoral muscles in winter than in summer.

In mammals, there is also another mechanism of heat production called nonshivering thermogenesis (NST, see above). This biochemical decoupling of catabolic pathways from ATP production takes place under control of the sympathetic nervous system and the catechol-amines. It is particularly evident in muscles, liver and brown fat tissues. NST in small mammals often differs seasonally, and is higher in winter, especially in nonhibernating forms. Brown fat, which is characterized by having many small fat droplets per cell, and a high concentration of mitochondria, is particularly important for newborn mammals (even human beings have it perinatally) and small species. Capacity for NST is also depending on other ecological factors—desert living, and diurnal species have higher capacities than species from mesic environments or nocturnal habits.

Differences in thermal metabolism between active and resting animals are influenced by a complicated array of factors. The boundary layer is no longer undisturbed, increasing thermal conductance, wind passes over the moving body parts, more blood flows toward skin and muscles etc. Thus, there normally is no effective gain in heat production by just moving around—at least in small animals. Isolating morphological structures, such as thicker furs or plumages increasing the boundary layer, or thicker subcutaneous white fat tissue, also have to be considered, as well as regulatory changes in blood flow (triggered by hypothalamic activity, see above) toward skin and other peripheral organs. This is called regional heterothermy and it could be demonstrated in many species of temperate or cold-adapted ungulates (deer, Przewalski's horses) that this indeed, again depending on the season, significantly reduces heat loss.

Emperor penguins in Antarctic conditions also seem to be capable of changing surface blood flow, thermoimaging of penguins under clear sky demonstrated surface temperatures below ambient air temperature.

The perhaps most extreme and also most spectacular form of cold tolerance is found in ectotherms, mostly bony fish, living at or near freeze conditions in polar waters. In fact, living in very cold water can be achieved by different phenomena: we have to distinguish between animals that are freeze-tolerant (allowing the formation of ice crystals in their tissues), and those that are freeze-intolerant = freeze-avoiding—which means they can live in temperatures far below 0°C, but only if no ice is formed. Freeze-tolerant animals are mostly invertebrates (e.g., molluscs, insects), some specialized amphibians and reptiles also have this capacity. In invertebrates, this is achieved by several substances: ice-nucleating proteins control the formation of ice crystals in extracellular tissue fluid, by allowing the formation of very small icicles, whereas all the ions and soluble metabolites are concentrated intracellular. Cryoprotectant substances (sugars, polyolic alcohols) raise the osmotic concentration of body fluids, or bind to membranes to prevent ice crystals from forming there.

Freeze intolerance, on the other hand, if occurring under conditions of subzero temperatures has to be achieved by processes like supercooling. This, in animals again is achieved by adding substances, like glycerol, antifreeze proteins, and antifreeze glycoproteins, but also often by completely emptying one's gut in order to get rid of particles that could act as nucleators for ice formation. Cold-adapted invertebrates under these conditions can survive temperatures up to -60°C.

Size and Shape in Thermal Biology

As already discussed above, body surface to body volume ratios decrease with increasing body sizes. This leads to huge differences in mass-specific basal metabolic rates, 1 g tissue of a house mouse already having about a 10-fold higher BMR than 1 g tissue of a person. Also, small species consequently tend to have narrow TNZs, and increasing body size is, as Bergmann already formulated in his rule a common strategy for cold-adapted endotherms. In ectotherms, these correlations again are more complicated, because diffusional processes are much slower in the cold, thus insects, arachnids and other arthropods depending on O_2 diffusion into trachea or comparable structures can be larger in warm climate.

Another long-accepted rule of thermal biology has been formulated by Allen, also already in the 19th century: All body appendages of animals in the cold are smaller, thicker and rounder in shape than these of closely related taxa in warmer climates. Arid-adapted vertebrates often are small. This has to be discussed in connection with behavioral strategies see below. From a purely physiological point of view, a kangaroo rat (*Dipodomys* spec) would have to evaporate water of about 10% of its body weight per hour to keep its body temperature constant at a Ta of 43°C. Large and thin, sparsely furred ears (not the thickly furred ones of, e.g., a fennec), such as found in the Californian desert hares, *Lepus californicus* or *L. allerni*, which can reach up to 25% of their body surface areas, can help to radiate heat. Surprisingly on a typical day in the Californian desert sky radiation temperature is far below 20°C. As long as the hare is able to find a suitable micro-climate out of the direct sun, in the shade of a tree or other shelter, it is able by means of its "hot ears" (up to 38–40°C skin temperature have been found) to actually radiate heat back into the environment.

A comparative study of body size and degree of melanism in ant species from Africa, Australia and South America revealed similar correlations as predicted by these considerations: Species with larger body sizes were darker than smaller sized ones, color became lighter with increasing temperature but darkness again increased at the highest temperatures probably correlated to the high amount of UV-B.

Mammalian ear pinnae however, are not the only appendages allowing heat exchange. Horns, frills and antlers (be they of ungulates or dung beetles), large beaks of toucans and hornbills (*Ramphastidae*) also contribute significantly to heat flux. Hard bills, such as in ramphastids seem to have less possibility here than softer ones such as in toucans.

Torpor and Heterothermy

In temperature conformic ectotherms that are generally unable to increase their temperatures by internal means, torpor or hibernation in winter or aestivation in hot, dry summers (the latter being shown, e.g., by lungfish or desert snails), is superficially

similar to torpor in endotherms, however external energy sources are needed to end this state. In insects, a special form of arrested development called diapause, triggered by combination of hormonal, photoperiodic and nutritional factors, is a common strategy. Heterotherm animals are also characterized by changing their body temperatures and basal metabolism according to external conditions. A further distinction is often made between homeotherms that are capable of regulating their body temperatures with a very narrow range, and endotherms, species that are capable of regulating the body temperature at all, but not necessarily within a narrow range of only a few degrees centigrade.

Torpor is a condition which is characterized by a low body temperature and low BMR. In so far, it seems to be a variation of temperature conformity instead of regulation. However there is an important difference between heterothermy and ectothermy: heterotherms are capable of increasing body temperatures (mostly by increasing BMR) by their own, internal means whereas ectotherms need some external source of warmth or other energy for this waking up! Thermoregulation is never totally switched off during torpor, instead the setpoint for the onset of thermoregulatory activities is only temporarily lowered. Heterothermy has long been regarded as a primitive character of animals not yet ready for real homeothermy. It is a finely tuned adaptive strategy.

Torpor is further divided by the regularity and seasonality of its occurrence and triggering mechanisms. Long torpor, mostly in the form of hibernation, is often extended for several months and is characterized by a lowering of body temperatures under 10°C, and metabolic rate is about 5% of BMR during active phases. However even deep hibernating torpor in all species studied so far is interrupted by short periods of activity at normal body temperature, and these intervals are internally triggered.

Large mammals, such as bears, also go into torpor. However, this is only shallow torpor, with a reduction of their body temperatures by about 5°C, heart rates and metabolic rates are reduced by up to 30%. Nevertheless, hibernating bears can stay in their dens for several months, and their energy needs are covered by burning fat. Some other physiological adaptations, such as recycling urea into essential amino acids, and most probably also calcium storage and recycling, have been developed in these large carnivores as well. Large bears are not the only carnivores capable of larger torpor. Raccoons and Raccoon dogs at least in parts of their range also enter torpor for several weeks.

Short term torpor of several days or even daily torpor is much more widespread also among larger mammals: both American and European badgers have been shown to enter daily or short term torpors, with body temperatures of about 28°C. Daily or short term torpor in general reduces body temperatures to about 10–30°C, metabolic rates are reduced to values of about 30%. Most mammal species entering daily torpor are small and nocturnal such as small marsupials (dasyurids, petaurids and didelphids), mouse lemurs, hedgehogs, tenrecs, shrews, or bats. However, in almost all these taxa (except primates) we also find species exhibit deep torpor with body temperatures around 5°C and durations of 10 days to several months (marsupials: *Cercartetus nanus*, a burramyid, reaches values of 2% of its normal BMR for several weeks, European hedgehog: energy of about 4% normal rate, Tb around 5°C for at least 10 days, bats: Myotis -2°C– $+5$°C Tb, energy about 1% BMR etc.). Heterothermy among birds is different in several aspects: it mostly occurs during the night, Tb is lowered by about 5°C, it also occurs in rather large species such as Turkey vultures, but, and this is a phenomenon whose adaptive significance is still unclear, energetic demand is mostly higher than BMR! Only few species, such as some colibris, tend to reduce Tb to values below 18°C, some below 10°C, and only one species of bird, a night jar from North America, *Phalaenoptilus nuttallii*, goes into torpor for several days in consecution, and also reaches a Tb as low as 6°C. It is not yet totally clear, neither for birds nor mammals, which physiological mechanisms are responsible for reawakening. One hypothesis assumes that a combination of low blood pressure and accumulation of toxic metabolic products in the blood pressure to cleanse the blood from these waste products, another assumes a biological clock (maybe even the circadian on which is also being showed by lower body temperatures). In any case, the end of a torpor phase is achieved by active warming the velocity of which mostly depends on body size: small animals of about 10 g body weight can gain almost 1°C/min, species of about 1 kg only achieve 0.5°C/min and species over 10 kg are real slow walkers, with increases of about 0.1°C/min. This seems to be a constraint on the capability for deep torpor in large species.

Circadian and circannual changes in metabolic rates, circannual changes in body weight and weight of internal organs, changes in intestinal transport rates (especially peptide transport), and decreases of heart rate with decreasing subcutaneous (hence ambient) temperatures have been demonstrated in cold-adapted ungulates, for example, Red deer and chamois in the Alps.

Behavioral Adaptations

Social behavior can be an important component of temperature regulation. Many species hibernate or fall into torpor in larger groups, reducing the degree of heat loss and energetic demand for reheating. At least in the Alpine marmot it could be shown that, the closer related the members of such a hibernating group, the more synchronous their cyclic awakening rhythms.

Other behavioral strategies to cope with extreme temperatures depend on body size: Large species often migrate. At least in land-bound animals, such as mammals, energetic demands by locomotion are relatively lower in larger species, thus migrating mammals mostly are large ones. Swimming is the energetically easiest form of locomotion, thus seasonal migrations by aquatic vertebrates are quite common (chelonians, seals, sharks, cod…). Flying seems to be energetically the most demanding locomotion, however it is much easier to achieve high speed in flying than in walking. Gaining a certain distance for a small, flying bird thus is only about 1/10 as energetically costly than for a terrestrial mammals, thus even small birds (or bats) can migrate over long distances successfully. Smaller species of terrestrial endotherms, but also, for example, of reptiles, have different options. Many desert-living smaller mammals are totally nocturnal and it could be shown that even for rock-wallabies in northern Australia, rock clefts and crevices offer temperatures of at least xy°C lower and humidities at least yz% higher than outside. Another possibility, quite often found in

intermittent activity, retreating into burrows, rock crevices or other dark and cool places whenever one gets heated up too much. This strategy of intermittent activity bursts in an endotherm is well demonstrated by *Citellus leucurus*, a desert squirrel from the Southwestern United States. These animals use their bodies as heat-accumulators similar to camels, however, being small, they overheat much more quickly! As soon as their body temperature reaches dangerous values of around 43°C, they dash back into their burrows, or shady retreats, and closely press themselves on the ground or the walls. They rapidly lose heat by conductance thus and as soon as their body temperature return to safe values of about 38°C, within a few minutes they dash out again for the next foraging bout.

Intermittent activity bouts also are being used by desert reptiles. These animals additionally use changes in body orientation and body posture. Early in the mornings they start by directing their (mostly dark-colored) head and shoulder region toward the sun, and direct as much blood flow into these areas as possible. As soon as their blood has reached near critical temperature they turn away their body axes from the sun and later retreat into the shade. Again there are intermittent bursts of basking and seeking shade, and this helps them to keep their temperatures within acceptable ranges. Even nocturnal reptiles often have most of their activities during the first 3 h after sunset, thus also using a heat-storage strategy however in reverse to that of mammals.

Basking as a means of collecting heat from external sources is not limited to ectotherms. Hyraxes, a group of small Afrotherian mammals, living in rocky areas, also use this behavior to heat up in the mornings.

Withdrawing into a cooler, shady area beneath a tree, as being regularly seen by species in savannah areas, is the opposite, but similar type of behavioral regulatory adaptation. Australian koalas in hot conditions often hug cool trunks of large trees, hamadryas baboons sandbathe on hot days, Red kangaroos dig so called hip holes by scratching away the warm surface layer of earth, and lying down onto the then exposed cooler layers.

During sleeping periods, especially in REM sleep, thermoregulatory activities such as panting, shivering or sweating are discontinued, thus under conditions of severe heat or severe cold, most species do not fall into REM sleep at all whereas within the TNZ the amount of REM sleep significantly increases. Social mechanism such as huddling also correlate to ambient temperature. In many primate species there is an increase of social behavior, time spent socializing, and ambient temperature. Female vervet monkeys with more social partners had less hypothermia and reduced 24 h amplitudes of core body temperature.

Ontogeny of Thermal Reactions

The concept of ontogenetic niches also concerns different aspects of thermal biology. Not only the reduced body sizes of juvenile animals, but also different habitats, spending more time in shelters, or different diets can influence metabolic conditions and thus thermal biology.

Most endotherms need some time after completion of embryological development before really being able to thermoregulate. It is most obvious in species, both birds and mammals, that effective thermoregulation is achieved after some postnatal time span, particularly in species with altricial young. However, when comparing species of different degrees in precociality with respect to the onset of thermoregulation the postnatal period alone is not a good representation. Ideally we need to compare the time span between conception and onset of effective thermoregulation, because precocial species often have longer gestation periods than similar-sized altricial ones. Thus, postnatally a Virginia opossum needs about 85–95 days before achieving effective thermoregulation, a similar-sized porcupine about 3 days. However, taking into account the differences in gestation, the opossum needs 105 days from conception, the porcupine 115 days! Similar with small rodents: a gerbil takes 19 days, a spiny mouse only 7 days postnatally but taking into account gestation, the gerbil needs 39 days postconception, the spiny mouse 47 days.

It is also tempting to regard the nakedness of altricial young as a primitive character. But in reality this seems to be an effective and adaptive means to transfer heat from the parents to the offspring. And altriciality itself including the lack of effective thermoregulation, seems to be an adaptation to quick growth (channel all available resources into rapid growth!) instead of an imperfection.

Development of effective thermoregulation in precocial species also often seems to be ecologically influenced: cold-adapted birds, such as arctic ducks, or in general the young of diving-duck species, can attain effective thermoregulation very soon after hatching, species from temperate climates, or surface feeding ducks in general, achieve it at a later age.

In general, young animals have a much higher metabolic rate than would be expected simply from the Kleiber formula (see above). In fact, even in species who, as adults, do have a low metabolic rate, this value for growing young approaches the boundary curve for endothermy which means they operate at the upper limit of thermal possibilities in terms of metabolic rate. This obviously is an investment by the young, to allow high growth rates. It has been hypothesized that the switch to a high metablolic rate in juveniles, to achieve both growth and effective thermoregulation, is a taking-over of at least part of the investment into further growth from parents to offspring: In altricial young, this take-over is more gradual parents feed and huddle them much longer than in precocial species.

Ontogenetic influences, however can also be on an even longer time scale. A study on Australian *Amphibolurus* dragons revealed a significant correlation between the degree of parental basking behavior and personality (shy-bold syndrome, feeding and exploration) of offspring.

Evolutionary Considerations

This already leads to the question of phylogenetic development of thermoregulation. In order to understand this tradition, it may be helpful to look at some taxa that are somewhere in between ecto- and endothermy. Some insects, for example, large, nocturnal

moths (*Sphingidae*), bees, dragonflies or wasps, are able to regulate thoracic and in some cases also abdominal temperature. However, this endothermy is only achieved when they are active, they perform wing-movements, called shivering, decoupled from flight. Moths at least, due to their hairy scales, have values of thermal conductance similar to birds and mammals, and they can keep large differences between Ta and Tb (some North American moths can fly at a core body temperature of around 30°C at Ta = 0°C). However, these small animals cannot achieve continuous endothermy similar to same-sized vertebrates if they are not active day and night (which insects are not).

Larger species in the continuum between ecto- and endothermy are found among fish. Bluefin tuna (*Thunnus thymnus*) of 200–350 kg can uphold temperature differences of up to 20°C. In these fish, contrary to "cold- bodied" species, we find large amounts of red (=aerobic) skeletal muscles near the body core (along the vertebral column) instead of under the skin. Also a high BMR, and a countercurrent heat-exchanger in the circulatory system are further characteristics of these endothermic fish. Besides red skeletal muscles, endothermic fish also have local heat sources in stomach, gut and liver tissue. Also (again kept up by retia mirabilia = countercurrent heat exchangers) in the eyes and the brain of warm-blooded fish such as Mako sharks (*Isurus oxyrhynchus*) there is a temperature difference to the environment of >5°C. However, there are no heat generating tissues in the sharks, heads, instead warm blood from abdominal red muscles is transported directly to the eye and brain regions. In some bony fish (e.g., Swordfish, *Xiphias gladius*) contrary to sharks, eye muscles are working as local heat sources, the whole complex of heater muscles, brain and eyes is thickly isolated in fat, and temperature differences of up to 14°C can be upheld between brain and surrounding water.

There is also evidence for mechanisms of physiological and behavioral temperature control in these fish. Also some python snakes and Leatherback turtles (*Dermochelys coriacea*) are able to obtain a certain control over their body temperatures.

Thus, the question of when and why endothermy could evolve has to be approached very broadly. The adaptive value of real endothermy and effective thermoregulation could have been to allow a decrease in body size at a constant body temperature. This would not only have allowed an increase in activity, but also an increase in reproduction. However, endothermy also is costly, and thus certain preconditions had to be met before achieving it. On the biochemical level, changes in membrane permeability for ions, are discussed as necessary preconditions for increasing metabolic rates. On the organismic level, it seems plausible at least in the evolution of mammals to assume that the large (up to 250 kg) therosaurus reptiles, the ancestors of mammals, had, due to their large size, achieved a certain degree of thermal independence, and that a whole array of morphological and physiological changes (development of isolating fur, increasing efficiency of ventilation by developing a bony palate and diaphragm, etc.) then allowed the transition from large reptile (with so called inertial homeothermy, which means they simply were too large to lose enough heat for being poikilotherms) to small mammal with an active, regulatory endothermy.

Combining Strategies, Living in Extremes

Finally, some extreme habitats whose inhabitants combine several of the adaptations discussed above shall be briefly mentioned. Marine mammals, particularly those in arctic or deep-sea conditions, such as sperm whales, or elephant seals developed isolatory fat tissues of up to 50 cm in whales, at least 5–10 cm in seals, and a reduction in body surface area by shortening limbs and other appendages. Behaviorally in sperm whales it is obvious that only large, fully adult animals perform real deep dives. Mothers actively discourage their calves from following them, handing them over to a surface near babysitter instead, and in elephant seals the deepest dives are performed by the large males. Development of countercurrent heat exchangers in the fins and flukes of whales allows for a reduction of heat loss into the environment and a very high fat content in the milk (up to 70% in some arctic seals!) allows a very rapid build up of the isolating fat tissues in the young.

Living at high altitudes is another extreme condition with a lot of thermal challenges. Blood flow through the extremities in large species of high alpine (e.g., Andean) animals normally is increased to prevent tissue damage by freezing, skin temperature is higher at low Ta, metabolic rates and O_2 consumption are increased in high altitudes to uphold these regulatory mechanisms. Isolating structures (fur, fat) are well-developed, countercurrent heat exchangers exist in the limbs, and behavioral thermoregulation sets in: In Andean camelids, there are patches almost without any fur in the abdominal and inguinal regions. These areas can either be exposed to wind and air (by standing alone and with limbs stretched) or gradually covered more and more (by standing crouched, lying down with limbs folded under the belly) or finally huddling while lying down in a group. Thus, heat loss by radiance and conductance can be effectively altered according to different ambient temperatures.

Degree of ability to cope with previously unexperienced changes in climatic conditions following global warming may be one of the most important influential factors in adaptation to global warming in many species, not only from polar areas but also alpine and other montane regions.

See also: Ecological Complexity: Thermodynamics in Ecology. General Ecology: Tolerance Range. Terrestrial and Landscape Ecology: Thermodynamic Properties of Landscape Cover

Further Reading

Arnold W (2011) Überleben im Hochgebirge–Winteranpassungen des Gamswildes. Veterinärmedizinische Universität Wien, P 13–. In: Die Zukunft des Gamswildes in den Alpen. Schriftenreihe des Landesjagdverbandes Bayern.

Arnold, W., Ruf, T., Reimoser, S., Tataruch, F., Onderschenka, K., Schober, F., 2004. Nocturnal hypermetabolismas an overwintering strategy of red deer (*Cervus eiaphus*). American Journal of Physiology-Regulatory, Integrative and Comparative Physiology 286, R174–R181.

Bishop, T.R., Robertson, M.P., Gibb, H., Parr, C.L., 2016. Ant assemblages have darker and larger members in cold environments. In: Global ecology and biogeography. John Wiley, Sons Ltd.

Cloudsley-Thompson, J.L., 1999. The diversity of amphibians and reptiles. Berlin: Springer.

Ganslоßer, U. (Ed.), 1999. (Hrsg.): Spitzenleistungen. Fürth: Filander.

Harshaw, C., Blumberg, M.S., Alberts, J., 2016. Thermoregulation energetics, and behaviour. In: APA handbook of comparative psychology.

Mc Cafferty, D., Gilbert, C., Thierry, A.-M., Ancel, A., 2013. Emperor penguin body surfaces cool below air temperature. Biology Letters 9 (3), 20121192.

McNab, B.K., 2002. The physiological ecology of vertebrates. Ithaca: Cornell UP.

Ortega, Z., Mencia, A., Pérez-Mellado, V., 2016. Wind constraints on the thermoregulation of high mountain lizards. International Journal of Biometerology. doi:10.1007/s00484-016-1233-9.

Pierau, F.-K., 2001. Thermosensibilitat. P 315-332. In: Dudel, J., Menzel, R., Schmidt, R.F. (Eds.), (Hrsg.): Neurowissenschaft. Berlin: Springer.

Sauer, E.L., Sperry, J., Rohr, J.R., 2016. An efficient and inexpensive method for measuring long-term thermoregulatory behaviour. Journal of Thermal Biology 60, 231–236.

Schmidt-Nielsen, K., 1979. Desert animals. NY: Dover Publ.

Schnidt, K., 1993. Winters ecology of nonmigratory Alpine red deer. Oecologica 95 (2), 226–233.

Todd, G., Jodrey, A., Stahlschmidt, Z., 2016. Immune activation influences the trade-off between thermoregulation and shelter use. Animal Behaviour 118, 27–32.

Willmer, P., Stone, G., Johnston, I., 2000. Environmental physiology of animals. Oxford: Blackwell.

Biodiversity Indices

Peter Fedor and Martina Zvaríková, Comenius University, Bratislava, Slovakia

Glossary

Climax The final stage of biotic succession attainable by a community in an area under the environmental conditions present at a particular time.

Holistic Emphasizing the importance of the whole and the interdependence of its parts.

Niche A role (position) that a species plays in the environment, the status of an organism within its environment which affects its survival.

Resilience A property of ecosystems to respond to disturbance by resisting damage and recovering quickly.

Resistance A property of ecosystems to remain unchanged.

Introduction

Biological diversity represents the variety of life at all possible interpretation levels, from genes and populations to the most complex ecological systems (biomes), including morphological (genotypes and phenotypes), taxonomic (usually measured as species diversity), and ecological diversity (variety of ecological interactions). Even from a spatial point of view, global variety of life (gamma diversity) is a compound of local units (alpha diversity) to some degree of their mutual similarity (beta diversity) in a general concept:

$$\gamma = \beta \cdot \alpha$$

Biodiversity is not an isolated unit. On the contrary, it forms an open permeable structure in deep conjunction with abiotic conditions (habitat), behaving as a semiautonomous network, known as an ecological system (ecosystem), with all its general features and aspects. According to the Biological Diversity Convention, biodiversity has been officially defined as the variability among living organisms from all sources, including, "inter alia," terrestrial, marine, and other aquatic ecosystems and the ecological complexes of which they are part. This also includes species diversity, however, within the limits between the variety of elementary genetic subsystems on the one hand, and holistic ecological supersystems on the other. Unquestionably, species diversity indices have for a long time often been used as indicators of biodiversity.

This article presents an overview of the most well-known biodiversity indices in a concept of species, as well as ecological diversity from a single species richness to the proportional distribution of species in samples and communities (Fedor and Spellerberg, 2013). Apart from the mathematical expressions which are necessary for understanding specific features, disadvantages, and potential of numerous species diversity indices, with a particular emphasis on the Shannon–Wiener and Simpson index, there is strong postulation on its consequences in applied ecology, bioindication, monitoring of ecological change, and environmental assessment.

Introduction to Biodiversity Indices

When viewing a wide variety of ecological indices, it may seem very difficult and disputable in selecting the best model for measuring biodiversity. Without any doubt, there is absolutely no universal index suitable for all theoretical or real cases in ecological research, although some of them may appear more widely used in literature. Some indices are calculated in a very simple manner (e.g., Berger–Parker index, Margalef index, Menhinick index), while others are based on complex mathematics (e.g., Brillouin index). Before any ecological evaluation, a researcher should know and identify the real performance of biodiversity indices offered within a specific situation, and this predominantly includes:

- response to changes in ecological systems (from the smallest to the most significant);
- emphasis on species richness and evenness in ecological communities;
- dependence on a sample size (especially for rare species);
- discriminability as the ability to detect differences between sites and samples; and
- frequency in ecological research (how widely the index is used in literature, for comparing results more reliably).

In the most simple and elementary method, species diversity can be expressed as a "single species richness," indicating the number of species in a sample. However, for even the most basic ecological research, this value appears quite insufficient and hardly informative, raising more questions than answers. Therefore, ecologists prefer to express species richness in its more sophisticated form as species richness indices (e.g., Margalef index, Menhinick index). For many decades, the most advanced approach, widely used in environmental assessment and monitoring ecological change, has been established on a large cohort of species diversity indices with emphasis on the distribution of species within a sample or community, expressed as relative abundance or dominance. The performance and specific features of a range of the most well-known diversity measures were summarized by Magurran in 1988, and may be adapted as shown in **Table 1**.

Although calculation of some indices may appear more or less complex and difficult in current ecological research, manual calculations have been replaced by automatic and user-friendly processing within plenty of computer statistical applications, such as SPSS, SAS, and R software. Unfortunately, users often do not know any mathematical details of the indices in order to recognize their most elementary advantages and disadvantages.

From Single Species Richness to Its Indices

Species richness represents a measure of the variety of species based simply on a count of the number of species in a particular sample, although it can be expressed more usefully as species richness pre unit area, ranging from alpha (referring to a certain site) to gamma (for an entire study area) level.

The terms "biodiversity," "species diversity," and "species richness" are sometimes used in confusing ways. In some papers, the term "species richness" is used in the title and in the text it is assumed to mean the number of species, but this may not be made clear. By way of contrast, "diversity" is sometimes used in the title, but in the text the data seem to refer to the number of species only. However, in 2003 Spellerberg and Fedor suggested that "species richness" should be used to refer to the number of species (in a given area or in a given sample) and "species diversity" should be retained for use in this context, that is, as an expression of some relation between the number of species and number of individuals. Rather than using the terms "species richness" and "species diversity" interchangeably, it is helpful to distinguish between these two terms.

Apart from a single species richness (number of species), there are various simple species richness indices based on the total number of species and the total number of individuals in a sample or site. In fact, they represent the simple measures of species richness, taking into account only the number of species and the total abundance of all specimens in a sample. One of them was successfully developed by the American ecologist Edward F. Menhinick in 1964:

$$D = \frac{S}{\sqrt{N}}$$

with S being the species richness and N being the total number of all specimens in a sample.

The Margalef index, summarized by the Spanish ecologist Ramon Margalef López in 1958, several years before Menhinick, is calculated as follows:

$$D = \frac{S-1}{lnN}$$

with S being species richness and N being the total number of all specimens in a sample. This can also be a simple measure of mean population size and using the form $(S - 1)$ gives a zero value for just one species present in a sample. The maximum appears for $S = N$.

Ramon Margalef López (1919–2004): Professor of Ecology at the University of Barcelona, Spain, one of the most reputable Spanish ecologists. His scientific work is connected with applied information theory in ecology, including mathematical models in populations, presented particularly in his books *Natural Communities* (1962), *Perspectives In Ecological Theory* (1968), *Ecology* (1974), *The Biosphere* (1980), *Limnology* (1983), and *Theory of Ecological Systems* (1991).

Table 1 Basic biodiversity measures with their specific characteristics for ecological research (particularly based on Magurran, 1988)

Species diversity	*Discriminability*	*Sample size sensitivity*	*Emphasis on*	*Calculation*	*Use in literature*
Species richness	Good	High	Richness	Simple	Very frequent
Margalef index	Good	High	Richness	Simple	Frequent
Shannon index	Moderate	Moderate	Richness	Intermediate	Very frequent
Simpson index	Moderate	Low	Dominance	Intermediate	Very frequent
Brillouin index	Moderate	Moderate	Richness	Complex	Rare
Berger–Parker index	Poor	Low	Dominance	Simple	Frequent
McIntosh U index	Good	Moderate	Richness	Intermediate	Rare

Despite their simplicity, the species richness indices are undisputedly affected by sample size, and thus the sampling effort represents the investment in obtaining study material. The greater the sampling effort, the potentially higher index value; however, the different levels of sampling effort might be difficult and, in fact, incomparable. Furthermore, species richness indices could be misleading when they fail to take abundance patterns into account. As an example, in two theoretical communities A and B, both with the same species richness (S) of only 3 species and with the same total abundance (N) of 12 specimens, there is a significant difference in their proportion, as shown in **Fig. 1**.

Despite a difference in proportion of all three species X, Y, and Z, both species richness indices will be calculated with the same values, as shown in **Table 2**.

Nevertheless, from the species richness point of view, both communities appear perfectly alike; there should definitely be supportive statements coming from calculations of advanced diversity indices.

As a real example of applying species richness indices in environmental monitoring, Olawusi-Peters and Ajibare studied shellfish (Crustacea) communities in Nigeria to evaluate the degree of human-induced impact on marine ecosystems. Although the single species richness values did not appear significant across the four study sites, the Margalef index indicated a significant difference between two of them.

Very similar results were obtained by Jorgensen et al. when, during the analysis of the subtidal communities from the Mondego Estuary (Portuguese), they revealed that the Margalef index correlated significantly with phosphate concentration levels and was capable of detecting significant differences between polluted and unpolluted areas.

Species Diversity Indices

Species diversity indices express diversity in a more sophisticated way, correlating species richness with distribution of all elements in a sample or community within their relative abundance or dominance to avoid the same values for structurally different assemblages (**Fig. 1**). Undisputedly, there are more features describing the diversity of heterospecific groups (such as a community) apart from a single species richness. Except for the most well-known indices (e.g., Shannon and Simpson index) described in more detail in the following articles, the whole cohort of biodiversity measures include some that are less frequently used but should be

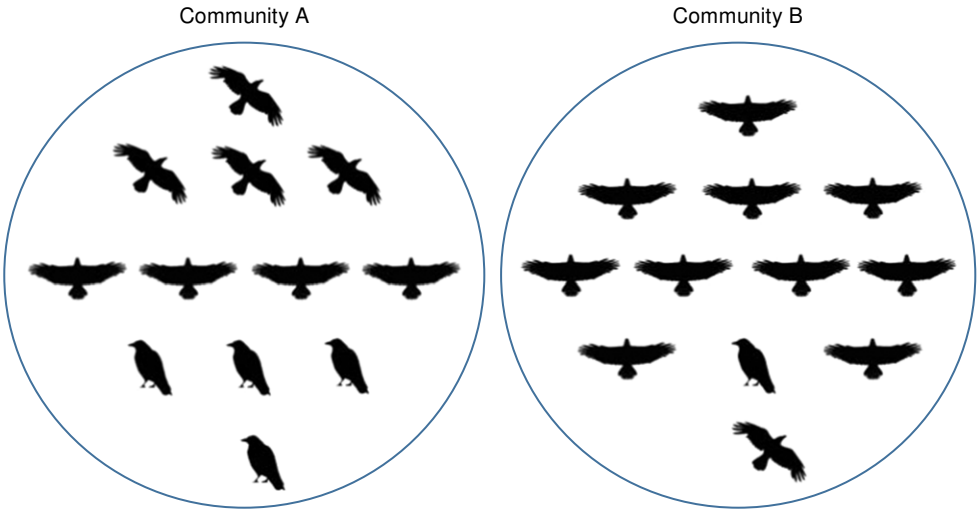


Fig. 1 Two different communities with the same species richness.

at least mentioned in the following survey.

Table 2 Structure of two different communities with the same species richness

	Community A	*Community B*
Species richness (S)	3	3
Total abundance (N)	12	12
Abundance of species X	4	10
Abundance of species Y	4	1
Abundance of species Z	4	1
Margalef's index	**0.80**	**0.80**
Menhinick's index	**0.87**	**0.87**

Fisher Alpha Index

A parametric index of diversity, with good discriminant ability and low sensitivity to the sample, assuming that the abundance of species follows the log series distribution.

$$S = \alpha \cdot ln\left(1 + \frac{N}{\alpha}\right)$$

with α being the index, S being the total number of species in the sample, and N being the total number of individuals.

Berger and Parker Dominance Index

A relatively simple index of species dominance was first introduced by Berger and Parker in their study on planktonic Foraminifera in sea sediments and later modified by Robert May as

$$d = \frac{N_{\max}}{N}$$

with $N_{\max}$ being maximum abundance among the species and N being total abundance in a sample. The lower the values of the index, the more equitable the community is.

Brillouin Index

$$H = \frac{1}{N} log_{10} \frac{N!}{N_1! \cdot N_2! \ldots N_s!}$$

with H being diversity index, N being total abundance, and N_1 being abundance of species 1…

This index has been discussed as inappropriate under certain conditions, for example, when random sampling is not guaranteed.

McIntosh *U* Index

$$D = \frac{N - U}{N - \sqrt{N}}$$

with N being total abundance and U given by the expression

$$U = \sqrt{\sum {n_i}^2}$$

where n_i is abundance of a species i.

This index has not been used in ecological research very frequently. The higher the value, the more homogenous the distribution species is in a sample.

Shannon (Shannon–Wiener) Index of Species Diversity

Claude Shannon was an expert in mathematics and electrical engineering. He undertook a great deal of research on theories of information communication. Following this research, he summarized his ideas in a technical journal. It was in this particular paper that there was a reference to his mathematical theory of communication. The main objective of this theory is to try to measure the amount of "order" (or "disorder") within a particular system. The concepts of order and disorder have long been a topic of discussion in natural history. In the 1950s and 1960s, ideas such as links between "information" and "diversity," were debated by ecologists. A commonly used index in ecology and ecological monitoring is the Shannon index (or Shannon–Wiener index). This has been derived from the Shannon function H and is expressed as follows:

$$H = -\sum_{i=1}^{n} p_i \ ln \ p_i$$

where H is the index of species diversity and p_i is the relative abundance of the ith species (n_i is the number of the ith species). For example, and with reference to **Table 2**, the diversity of the community A and B is calculated as follows:

Community A

$$H_A = -\left(\frac{4}{12} ln \frac{4}{12} + \frac{4}{12} ln \frac{4}{12} + \frac{4}{12} ln \frac{4}{12}\right) = 1.1$$

Community B

$$H_B = -\left(\frac{10}{12}ln\frac{10}{12}+\frac{1}{12}ln\frac{1}{12}+\frac{1}{12}ln\frac{1}{12}\right)=0.57$$

According to various ecological papers, any log-base is acceptable for calculations; however, *ln*-base is the most widely used and original. The Shannon function H has played a central role in information theory as a measure of information, choice, and uncertainty. This in turn led to its useful role as a measure of evenness or equitability.

There is some confusion in the literature in that the Shannon–Wiener index is sometimes mistakenly called the Shannon–Weaver index. This confusion has come about partly because Claude Shannon collaborated with the mathematician Warren Weaver on several occasions to publish papers and books, in particular the book The Mathematical Theory of Communication. Shannon first published an account of the entropy "H" in 1948. Weaver builds on this in 1949 in the second part of the above book. In his paper, Shannon acknowledges the fact that "communication theory is heavily indebted to the mathematician Norbert Wiener for much of its basic philosophy and theory," and cites several of Wiener's publications that refer to basic cybernetics. Shannon also refers to earlier work, including that of Boltzmann. In 2003, Spellerberg and Fedor suggested that the "mislabeling" of the Shannon Index "H" has come about partly because of the joint authorship of Shannon and Weaver's book, which has led to a belief that these two authors can be attributed to the index (the Shannon index is sometimes called the Shannon and Weaver index). However, the index should refer to "H" (the species diversity index) as the "Shannon index" or the "Shannon and Wiener index." Had Weaver's name been anything else and not similar to "Wiener," this confusion might not have arisen.

One of the advantages of the Shannon–Wiener index is that it is not greatly affected by sample size. One early study that examined the effects of sample size was by Wilhm and Dorris. They calculated species diversity indices from values that were pooled from successful samples. They found that sample size had a very small effect on the measures of species diversity.

Another advantage of using diversity indices such as the Shannon–Wiener index is that they capture a lot of information in one expression. This can be helpful when communicating large sets of data to a general audience. On the other hand, such expressions can appear to be very impressive only because they are derived from simple mathematics. It is essential, therefore, that anyone using such an index explains how it is calculated.

From many aspects, the Shannon index appears almost universal for general ecological use, as it behaves moderately in discriminability to detect different sites and samples with rather simple calculations and frequent use in literature. However, the measures of species diversity must be put into context. The context can be considered in two parts. One is the minimum and maximum that are theoretically possible, and the second is the range of values that could be expected in one particular ecological community. It is easy to measure the minimum and maximum for any species diversity index; the Shannon index, for instance, theoretically varies between 0 (at the lowest evenness) and lnS with S as a species richness. Similarly, it is relatively easy to assess the likely limits in any ecological community.

Claude Elwood Shannon (1916–2001): American mathematician and electrical engineer. In 1936, he worked in the Department of Electrical Engineering at the Massachusetts Institute of Technology and later spent a short time in advanced study at Princeton University. In 1941, he commenced work for the Bell Telephone laboratories in New Jersey where he spent 15 years among a well-respected scientific community. Following his research on information communication, he summarized his ideas on mathematical theory of communication in 1948 (*Bell System Technical Journal*) with the first form of the present Shannon expression as

$$H = -K\sum_{i=1}^{n} p_i \log p_i$$

where K is a positive constant. This expression has a central role in the information theory as a measure of information, choice, and uncertainty.

Norbert Wiener (1894–1964): American mathematician and philosopher, Professor of Mathematics at Massachusetts Institute of Technology. He is considered the inventor of cybernetics, inspiring future generations to develop sophisticated computer technologies.

Evenness or Equitability

The Shannon–Wiener index is particularly based on the concept of evenness or equitability. Simply put, the concept of evenness refers to the extent to which each species is represented among the sample. The extremes would range from one species being dominant and all other species being present in very low numbers (the lowest equitability close to 0), to all species being represented by equal numbers (the highest equitability of 1). For instance, in a sample of 10 species with total abundance of 100 specimens, the extreme A would be a sample with one species represented by 91 individuals and the other 9 being represented by 1 individual each. The other extreme (B) would be where each of the 10 species was each represented by 10 individuals. Equitability based on the Shannon index (E_H), assuming a value between 0 and 1, can be then calculated as

$$E_H = \frac{H}{lnS}$$

where H is the Shannon index of species diversity and S is species richness. Equitability is also known as the Pielou evenness index.

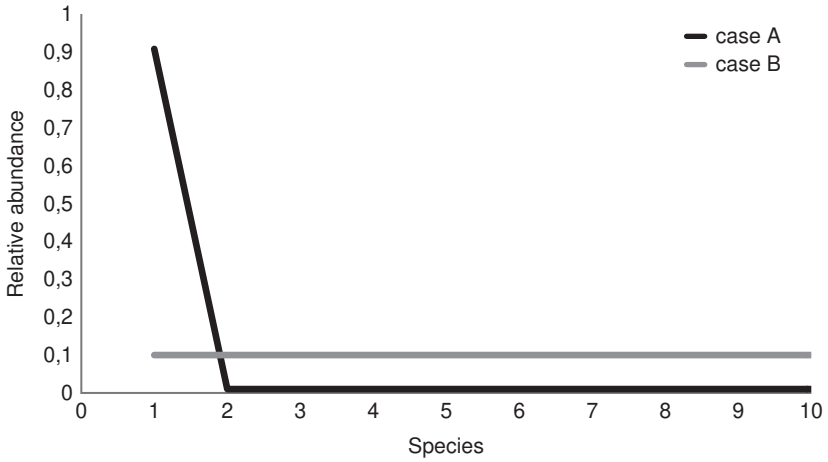


Fig. 2 Rank-abundance curves for two extreme cases A and B.

Equitability plays an important role in any ecological analysis using species diversity indices. For instance, for more equitable communities, the Shannon index becomes more sensitive to species richness.

Except for precise mathematical calculations based on the Shannon index, evenness in a studied assemblage can be expressed optically more suitable in a rank-abundance curve figuring distribution of relative species abundances (Pi) within the full array of Pi values by plotting Pi against rank. The rank-abundance diagram visually depicts species richness, as well as species evenness, with all species organized on the *X*-axis from the most to the least abundant with their relative abundances on the *Y*-axis. Species evenness is reflected within the slope of the final line from higher equitability with a shallow gradient to its lower values indicated by a steep gradient (Fig. 2).

Applied Cases

Species diversity indices are widely used in ecological monitoring and state-of-the-environment reporting. Although species diversity indices do summarize a lot of data, it is recommended that in ecological monitoring, diversity indices are used alongside other measures of the state of the environment. This is simply because different measures of the state of the environment are based on different parameters or variables.

Diversity in ecology has been the subject of much research. There have been debates about links or between diversity and resilience of ecosystems. In 1983, for example, Moore provided a brief summary of several research programs on ecological diversity and stress. In that summary, he referred to several authors including del Moral, who reported some research on subalpine meadows on the slopes of the Olympic Mountains in the western United States. In these grasslands, productivity is related to many biotic and abiotic factors. It was found that diversity (measured by the Shannon–Wiener index) had maximum values where there was moderate stress and where total productivity was suboptimal. Furthermore, it was reported that species transplanted from high-diversity sites to highly productive sites generally survived well if the surrounding plots were cleared of competitors. However, they did not do well if the plots were not cleared.

In another study (summarized by Moore), the researchers Hixon and Brostoff described complex links where predatory damsel fish in Hawaiian coral reefs control populations of herbivorous fish, which in turn influence the algae components of the reef ecosystem. They discovered that the greatest diversity (Shannon–Wiener index) was found inside the damselfish territories and the lowest outside.

In fact, the Shannon–Wiener index has been widely used to study the different effects of stress and disturbance on diversity of animal and plant communities. Because it provides more complex information than simple species richness and is based on the number of species as well as their equitability, the index serves as a valuable tool in monitoring ecological change.

Simpson Index of Species Diversity

The Simpson index was brought in by statistician Edward H. Simpson in 1949, but it had already been partially introduced in 1945 by the economist Albert O. Hirschman. Therefore, the same index is usually used as the Simpson index in ecology (D or λ), and as Herfindahl's index or the Herfindahl–Hirschman index (HHI) in economics.

The Simpson index belongs to dominance diversity indices because it counts more on common or dominant species. Thus, a few rare species that are only represented by a few individuals will not affect diversity markedly. The Simpson index measures the degree of concentration when individuals are classified into types, and takes into account the number of species present, including their relative abundance in a sample. The natural definition of the Simpson index can be interpreted as the probability that any

two individuals from the sample, chosen at random, will be found to belong to the same species (λ or D is always ≤ 1). It is expressed as follows:

$$\lambda = \sum_{i=1}^{s} p_i^2$$

or for diversity statistics it is taken as reciprocal:

$$D = \frac{1}{\sum_{i=1}^{s} p_i^2}$$

where p_i is the relative abundance of the i-th species (absolute abundance of each particular species (n_i) divided by the total number of individuals found (N)) and s is the number of species. The value of the index ranges between 0 and 1. The higher the score, the higher the probability that two individuals belong to the same species, along with a decline in diversity. Therefore, the index is sometimes expressed as

$$\lambda = 1 - \sum_{i=1}^{s} p_i^2$$

to reverse it. For the reciprocal it is $D > 1$, and the higher the score, the more diverse the community is considered to be. With reference to data in **Table 2**, the Simpson index will be calculated as follows:

Community A

$$D_A = \frac{1}{\left(\frac{4}{12}\right)^2 + \left(\frac{4}{12}\right)^2 + \left(\frac{4}{12}\right)^2} = 3$$

Community B

$$D_B = \frac{1}{\left(\frac{10}{12}\right)^2 + \left(\frac{1}{12}\right)^2 + \left(\frac{1}{12}\right)^2} = 1.41$$

Evenness or Equitability

The most stable communities usually have large numbers of species, almost evenly distributed. Seeing that dominance is the complement of evenness, an equitability index takes its largest value when all species have the same abundance. In this context, species diversity is a function of species richness and evenness; thus as species richness and evenness increase, diversity also increases. Analogous to "Shannon equitability," "Simpson equitability" is interpreted as follows:

$$E = \frac{D}{D_{max}}$$

where D stands for the Simpson index of species diversity and D_{max} stands for the maximal value of the index, which is the species richness of the sample (for the reciprocal form). The Simpson index is strongly influenced by the evenness of the species distribution, and it is weakly affected by species richness.

Edward Hugh Simpson (1922): British mathematician and statistician. In World War II, he worked as a cryptanalyst at Bletchley Park. In 1951, he described Simpson's paradox (also known as the Yule-Simpson effect) in which a certain trend appears in various groups of data but disappears or reverses after combination of the groups. He retired in 1982 as a British civil servant and lives in Oxfordshire.

There is a good example in soil-dwelling thrips (Thysanoptera) assemblages from Pannonian oak woods in Slovakia (**Table 3**) studied by Zvaríková and Fedor. Despite an increase in species richness during the vegetation season, the Simpson index remains stable with almost no response (**Fig. 3**). Due to the trend of high correlation between the Simpson index and equitability, there is no reason for a parallel increase of the index in accordance with species richness. The main emphasis refers to distribution of individual species.

Shannon-Wiener versus Simpson Index

Parallel use of species diversity indices in ecological studies is a general practice, and a typical case is the parallel application of the Shannon–Wiener and Simpson index. However, while the Shannon–Wiener index is strongly influenced by species richness and by rare species, the Simpson index gives more weight to evenness and common species. The effect of the sample size is generally negligible for both of them.

The use of both diversity indices improves the output information of the dataset, which is unique for each community or sample analyzed. Looking at the wider content—both emphasizing the richness, and the specimen distribution into the individual species—adds to the more complex information of the diversity in ecosystems.

In this sense, H has an advantage over D because it depends more on species richness and less abundant species, so it is very sensitive to even small diversity changes, and thus is widely used to assess the actual state of environment. On the other hand, D

Table 3 Soil-dwelling thrips (Thysanoptera) assemblages from Pannonian oak wood (SW Slovakia) (*S*—species richness, *N*—total number of individuals, *D*—reciprocal Simpson index, *E*—"Simpson evenness")

Species/date	*1 (June 22)*	*2 (July 3)*	*3 (July 22)*	*4 (Sep. 10)*
Thrips minutissimus	23	24	29	34
Haplothrips subtilissimus	28	34	23	25
Megathrips lativentris	24	33	41	0
Thrips major	0	21	0	0
Aptinothrips rufus	0	0	1	0
Limothrips denticornis	0	0	1	0
Liothrips pragensis	0	0	1	0
Mycterothrips albidicornis	0	0	0	30
Haplothrips aculeatus	0	0	0	1
Frankliniella intonsa	0	0	0	1
Hoplothrips ulmi	0	0	0	1
Phlaeothrips bispinosus	0	0	0	1
S	3	4	6	7
N	75	112	96	93
D	2.98	3.85	3.02	3.22
E	0.99	0.96	0.50	0.46

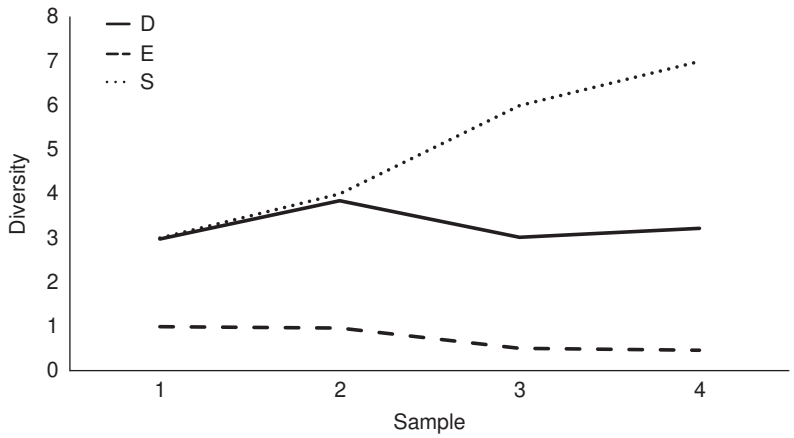


Fig. 3 Graph of species richness (*S*), reciprocal Simpson diversity index (*D*) and "Simpson evenness" (*E*) of thrips assemblages.

has the advantage over *H* in counting more on dominant species and is not affected by less abundant elements; therefore, it is used to show the trend of ecosystem diversity heading.

Applied Cases

There are many papers that assess diversity using species diversity indices. Without doubt, the Shannon–Wiener and Simpson indices are most often applied in ecological cases. Despite a recent effort to invent new measures that take species richness, as well as the number of individuals of each species into account, they mostly remain based on Shannon–Wiener and Simpson. For example, Hooker, in 2009, explains the importance of using both indices in the assessment of lake plankton diversity and community structure. The choice of using one index over the other strongly depends on what needs to be emphasized. If one is interested in assessing the recovery of a lake after a catastrophic event, the Shannon–Wiener index might be most useful because of its sensitivity to the return of initially rare species. On the other hand, evaluation of species that are relatively dominant, and therefore more important as a source of food for fish, might be of higher interest, so the Simpson index would be more suitable.

During the assessment of managed boreal forests in Finland, Pitkänen calculated several diversity indices (including Shannon and Simpson). Applying these indices to 166 forest sites, it was revealed that it is possible to classify forests in terms of biodiversity based on habitat characteristics. One could expect that species richness itself would be enough to distinguish between the forest sites, but in fact the mean values for species richness did not differ significantly. The indices had higher variability in their mean values, which may have been caused by the fact that they were influenced to varying extents by the evenness in the species abundances. The results indicate a correlation between stand structure and diversity of the ground vegetation in connection with forest classes, for which diversity measures can be analyzed. The Simpson index significantly correlated with every alpha diversity index and measurement of evenness ($P < 0.001$), but the Shannon diversity index mainly reflected species richness of the stands.

There is a small disadvantage of these indices when they are applied to large-scale monitoring of diversity intactness. Since information on species identity is lost, they cannot be used to monitor species turnover. At any rate, the Shannon–Wiener and Simpson diversity indices remain the most widely used in ecological research.

Bioindication and Biodiversity Indices

Species do not appear to be isolated. They form more or less complex systems (e.g., communities) with specific structural features, such as species richness, evenness or trophic interactions as a reflection of certain ecological conditions they live under and reflect. Many of them have been widely applied in bioindication within environmental assessment of monitoring ecological change. For example, some freshwater invertebrate communities have been the basis of some ecological monitoring programs. The community structure and the species present have also been bases for the classification of rivers and lakes. For instance, a software package aptly called RIVPACS (the river invertebrate prediction and classification system) is based on the assumption that the presence of certain taxonomic groups depends on certain physical and chemical variables. So there are two basic levels of bioindication potential: species dimension (presence, absence, morphological or physiological symptoms) and community dimension with its species diversity indices and ecological indices, expressed as numbers or scores that have been derived or transformed from quantitative data. For example, the Trent biotic index of water quality is derived from a mix of the presence or absence of certain indicator species and the number or diversity of taxa (or groups) or organisms present. One of the most well-researched indices is the Common Bird Index, which was established in the United Kingdom by the British Trust for Ornithology. This index came about because of a desire to monitor the state of bird species and because it was impractical to count every individual bird.

The climax ecosystem, as a final mature natural ecosystem along the complicated ecological succession, possesses the highest degree of stability based on the optimum network architecture of its components (e.g., species) and mutual interactions. All the niches (ecological positions) are unique and complementary. To provide the highest homeostasis according to May's criterion for stability, we use

$$\beta(SC)^{1/2} < 1$$

where β is the average strength of interaction between species, S stands for species richness, and C is the connectance of the web; (SC) should be constant. Thus, the tropical forests with high species richness are accompanied by lower connectance (linear and simple trophic chains); however, the temperate forests in Europe with lower richness work on forked chains. Both of them actually have the same ecological diversity and stability. Undoubtedly, diversity usually correlates with stability. The greater the diversity, most probably, the greater the stability. Therefore, measuring diversity plays a crucial role in the assessment of the ecosystem conditions within bioindication.

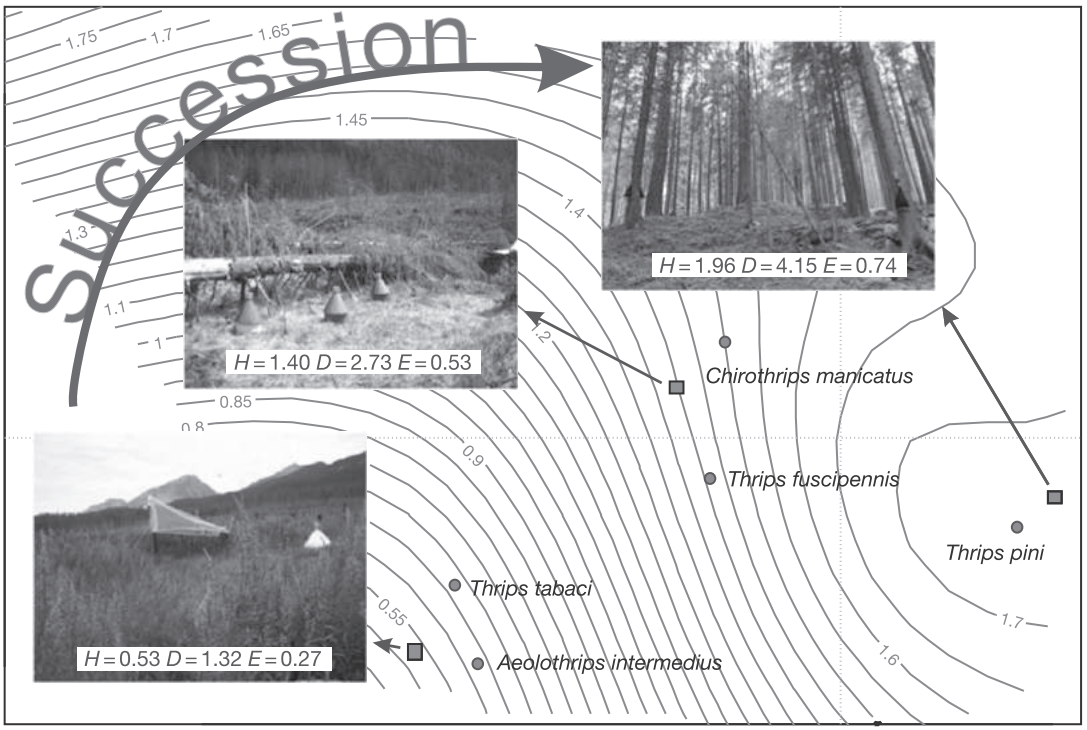


Fig. 4 A model case of complex bioindication.

As an example, in 2004, a massive windstorm devastated 13,000 ha (with 3 million m^3 of wood) of forest in the High Tatras, the oldest national park in Slovakia. The main damage appeared in artificial soft-wooded spruce monocultures with low resilience mechanisms. In climax ecosystems with no artificial bark beetle control, infestations reduce the dominant spruce stands, which are consequentially replaced by stronger tree species (e.g., silver firs), thereby building higher homeostasis. Simply put, the natural ecosystems with high biological diversity, including species richness and ecological interactions, are more resistant to disturbance and stress impacts. Ironically, the windstorm brought about a challenge for intensive ecological research. **Fig. 4**, adapted from NMDS analysis, describes the growth of three diversity measures (the Shannon index, its evenness, and the Simpson index) from the initial succession stage of the forest to 4 years after the windstorm, including massive fire devastation with all the wood destroyed or burnt, through to the stands affected by the windstorm's devastation with wood that remained, to the protected climax forest with the highest degree of ecological stability. Moreover, the figure includes some bioindicator species reflecting mature ecosystems (*Thrips pini*) or, on the other hand, open stands with low stability (*Thrips tabaci*) (**Fig. 4**).

Summary

The only pragmatic strategy to provide and maintain the huge potential of biodiversity in ecosystem services is associated with its sustainable use, meeting progressive human development within carrying capacity of environment and sustaining an essential network of ecological interactions. This approach, however, requires detailed knowledge on quantitative and qualitative parameters and bioindication potential of biodiversity, sophisticatedly expressed by its indices and predominantly represented by species diversity indices. On the other hand, without sufficient ecological knowledge, such analyses may bring more questions than answers.

For many years, the most advanced approach widely used in environmental assessment has been established on a large cohort of species diversity indices with emphasis on species richness, as well as distribution of species within a sample. Besides the most well-known indices (e.g., the Shannon and Simpson indices), the entire set of biodiversity measures includes some that are less frequently used, such as the Berger and Parker dominance index, the Fisher alpha index, and the Brillouin index. However, before any ecological evaluation, researchers should identify the real performance of biodiversity indices within a certain situation, including their specific response to changes in ecological systems, emphasis on species richness and evenness, dependence on a sample size, discriminability, as well as use frequency in literature to compare results more reliably.

See also: Conservation Ecology: Endangered Species. General Ecology: The Intermediate Disturbance Hypothesis. General Ecology: Biodiversity

Further Reading

Begon, M., Harper, J.L., Townsend, C.R., 1996. Ecology: Individuals, populations, and communities, 3rd edn. Cambridge, MA: Blackwell Science Ltd.
Duelli, P., Obrist, M.K., 2003. Biodiversity indicators: The choice of values and measures. Agriculture, Ecosystems and Environment 98, 87–98.
Fedor PJ and Spellerberg IF (2013) *Shannon-Wiener index, reference moduls in earth systems and environmental sciences*. Elsevier. Current as of 28 October 2016.
Heip, C.H.R., Herman, P.M.J., Soetaert, K., 1998. Indices of diversity and evenness. Oceanis 24 (4), 61–87.
Hill, M.O., 1973. Diversity and evenness: A unifying notation and its consequences. Ecology 54, 427–431.
Izsák, J., Papp, L., 2000. A link between ecological diversity indeces and measures of biodiversity. Ecological Modeling 130, 151–156.
Jorgensen, S.E., Constanza, R., Xu, F.-L. (Eds.), 2005. Handbook of ecological indicators for assessment of ecosystem health. Boca Raton: CRC Press.
Lamb, E.G., Bayne, E., Holloway, G., Schieck, J., Boutin, S., Herbes, J., Haughland, D.L., 2009. Indeces for monitoring biodiversity change: Are some more effctive than others? Ecological Indicators 9, 432–444.
Magurran, A.E., 1998. Ecological diversity and its measurement. Princeton: Princeton University Press.
May, R.M., 1975. Patterns of species abundance and diversity. In: Cody, M.L., Diamond, J.M. (Eds.), Ecology and evolution of communities. Cambridge, MA: Harvard University Press.
Morris, K., Caruso, T., Buscot, F., Fischer, M., Hancock, *et al*., 2014. Choosing and using diversity indices: Insights for ecological applications from the German biodiversity Exploratories. Ecology and Evolution 4 (18), 3514–3524.
Pielou, E.C., 1975. Ecological diversity. London: Wiley.
Rosenzweig, M.L., 1995. Species diversity in space and time. Cambridge: Cambridge University Press.
Shannon, C.E., 1948. A mathematical theory of communication. Bell System Technical Journal 27, 379–423.
Spellerberg, I.F., 2005. Monitoring ecological change, 2nd edn. Cambridge: Cambridge University Press.
Spellerberg, I.F., Fedor, P.J., 2003. A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the "Shannon–wiener" index. Global Ecology and Biogeography 12, 177–179.

Biological Integrity☆

Robert J Miltner, Ohio Environmental Protection Agency, Groveport, OH, United States

Glossary

Appreciative inquiry An approach to problem solving that focuses on the desired state and solutions to reach that state, rather than individual problems inhibiting that state.
Biocondition gradient The response of a biological system or assemblage to a range of disturbance or induced stress.
Biocriteria Standards for biological integrity relative to the biocondition gradient.
Biotic ligand model A water quality model that determines the bioavailability of metals to bind to membrane surfaces.
Cumulative distribution function A function that graphically illustrates the position of a given value in relation to all values in a continuous series.
Serotinous Delayed; in ecological contexts, usually referring to seed dispersal.

Integrity and Biological Systems

Integrity is defined as the state of being whole, undivided or uncorrupted. Applied to biological systems, the concept of integrity is hierarchical and dendritic: hierarchical in that biological systems can be categorized from the level of the biome (e.g., estuaries) to ecosystems (e.g., *Spartina* sp. marsh) to communities (e.g., fish and invertebrates living in the marsh) to an assemblage (e.g., benthic meiofauna), to populations of a given species within an assemblage (e.g., *Nereis diversicolor*). Biological systems are also organized by a trophic hierarchy from primary producers to consumers to predators, but brought full circle by decomposers. The trophic web is obviously one aspect of the dendritic nature of biological systems, but biological systems clearly depend on an interconnection of both biological and physical components. Littoral transport of sand nourishes the coastal barrier islands that protect the *Spartina* marsh from wave erosion. Nutrients delivered by rivers and tides sustain the *Spartina* that, in turn, interacts with the tides to provide a physically complex environment for juvenile fish and shellfish.

Apart from the collection of physical and biological parts that define a given biological system, a biological system cannot be considered whole unless it has the capacity to perpetuate itself, barring some epochal disturbance. So another aspect of biological systems is resilience to disturbance. In fact, many systems are dependent on periodic disturbance, with familiar examples being the boreal pine forest and fire, or rivers and seasonal flooding. For a boreal forest to be resilient to fire, it must contain species that are fire-adapted; hence, the serotinous cones of the jack pine. Although most jack pine cones only open from the heat of fire, some open from the heat of the day, demonstrating a range of traits, and implying a degree of genetic variation. Genetic variation is central to a biological system being able to persist through time. Perhaps nowhere has this been more apparent than with populations of lake trout in the Great Lakes of North America where restoration efforts following the collapse of the fishery due to overfishing and the invasion of the sea lamprey have been largely unsuccessful because, among other reasons, the hatchery reared fish used to supplant the native stocks are not genetically adapted to reproducing in specific open lake environments.

Taken collectively, the biological integrity of a system can be measured by the collection of its biological parts (from assemblages to species to genes) and the physical, chemical and biological processes that link those parts together and sustain the system. Biological integrity can be further measured by the resiliency of the system. The aspect of resilience is particularly important because it begins to frame how biological integrity can be measured against a gradient of human disturbance or corruption. Going back to the lake trout example, stocks of lake trout were resilient to fishing pressure (i.e., a human disturbance) until the introduction of the monofilament gill net, and the systemic corruption induced by invasive species (e.g., sea lamprey, alewives, and smelt) and industrial pollution. Stated another way, the system could be seen as having a high degree of biological integrity when it was capable of producing a desirable renewable resource. And that, in fact, has become the reference condition that animates the restoration efforts of the Great Lakes Fisheries Commission.

It is obvious from this example that measuring biological integrity hinges on a frame of reference, and that frame of reference is dependent on a necessarily subjective construct. Taking the reference condition as the state of the system in the absence or near absence of human disturbance is one way to attempt an objective framework. Similarly, but more realistically for developed landscapes and more commonly in practice, reference conditions are given by the least disturbed examples of a particular system available. More recently, the concept of the reference condition has been decoupled from a natural or least disturbed condition to one that can be identified through human-centered, appreciative inquiry. The lake trout example in the previous paragraph echoes

☆*Change History:* November 2017. RJ Miltner updated the text, replaced figures with three new figures, replaced Table 1 with a new table, and updated further reading of this article.

This is an update of J.R. Karr, Biological Integrity, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 408–412.

the appreciative inquiry approach by essentially asking the question of when was the system performing well, and then identifying that state as the desired condition. Similarly, the biocondition gradient approach to deriving expectations is informed by expert judgment to partition a given resource along a human disturbance gradient. For example, because the environmental tolerances and requirements of trout are generally well-understood, experts can partition the resource into tiers based on which species of trout a particular stream is capable of supporting, and set appropriate expectations for those tiers. Brook trout may only survive in relatively undisturbed headwaters, but brown trout are capable of surviving in modestly disturbed agricultural landscapes. This approach concedes loss (e.g., the erstwhile brook trout stream is now inhabited by brown trout), but recognizes the reality of the modern landscape.

Applying the Concept of Biological Integrity in Practice

The US Clean Water Act of 1972 called for the restoration and maintenance of the biological, physical and chemical integrity of the nation's waters, and more recently the European Union Water Framework Directive (EU WFD) called for surface waters in member states to meet "good ecological status" as a general requirement of ecological protection. These directives have led many countries in the EU and many states in the US to put in place biological monitoring, and develop methods to translate results from biological monitoring into information that can track attainment of water quality goals. Similarly, Australia and New Zealand employ robust biological monitoring programs and biological endpoints to track the condition status of their waters. More broadly, many countries or regional governing bodies have developed biological indicators to measure and track the status of their waters, either as a precursor to, or to help facilitate state adoption of biological requirements.

As is readily apparent, the focus thus far of biological monitoring and attendant condition indicators centers on surface waters, particularly rivers and streams. This is by dint of the fact that rivers have suffered and continue to suffer so heavily from pollution because they were or continue to be treated as simple waste conveyances rather than systems capable of producing multiple goods and services. Thus, measures of biological integrity for rivers and streams are the most widely developed and employed. Other environments where measures of biological integrity have been developed include estuaries, riparian areas, grasslands, coral reefs, lakes, wetlands, and upland forests.

There are two basic approaches widely employed to measure biological integrity: multimetric indexes, and observed-expected (O/E) models. Although the ensuing examples will center around rivers and streams, the concepts are generally applicable across systems. A multimetric index, as the name implies, includes various measures of the system parts previously described to form a single, composite index to gauge biological integrity against a reference condition. Individual candidate metrics are first proposed based on knowledge of the system, and natural attributes of the assemblage of interest. For example, to develop a fish-based index of biological integrity (IBI) for rivers and streams, one could divide the fish community into any number attributes, for example, trophic position, and then propose metrics based on the number or percent composition as herbivore, benthic insectivore, or top carnivore. Knowledge of the system in question is used to exclude inappropriate metrics; one would probably not consider the number of headwater-obligate taxa in a large river index. Other attributes such as tolerance or sensitivity to particular pollutants, temperature, dissolved oxygen, hydrologic stability, etc., can be included as candidate metrics. In practice, metrics used in fish biotic indexes typically fall into several categories: metrics reflecting habitat condition (e.g., species richness metrics), metrics reflecting trophic status (e.g., the percent composition of top carnivores), metrics reflecting pollution stress (e.g., number or percentage of pollution sensitive species), and general condition metrics like relative abundance and total species richness. **Table 1** maps attributes from three different indicator assemblages to a series of generic biotic metrics. Once candidate metrics are selected, samples are collected from reference sites and known stressed sites to ascertain which individual metrics are responsive to the disturbance gradient, and the direction of the response. Finally, of the metrics proving responsive, redundant metrics (i.e., those with very similar responses to disturbance) are identified and dropped from the final summary index. The samples collected from the reference sites are used to set scoring expectations for the individual metrics. Typically, either percentile ranges from the distribution of metric values are used to demarcate scoring bands, or the cumulative distribution function is used for continuous scoring.

O/E models also depend on an identified collection of reference sites to set expectations. Unlike multimetric indexes, however, the expectations in traditional O/E models are not assemblage attributes, rather the expectations are the probability of capture of a given species or taxon for the given environmental setting. The environmental setting is defined by classifying reference

Table 1 Broadly applicable multimetric index components and example metrics for three different indicator groups

Generic metric	*Fish*	*Macroinvertebrates*	*Diatoms*
Trophic guild	% Insectivore	% Scraper	% Eutrophic
Pollution tolerance	Sensitive richness	Presence of *Cricotopus* sp.	Bahls' tolerance
Habitat condition	% Gravel nesters	% Piercers	% Siltation tolerant
Hydrologic disturbance	% Pioneering	% Noninsects	Biovolume
General condition	Native abundance	EPT richness	Diatom richness

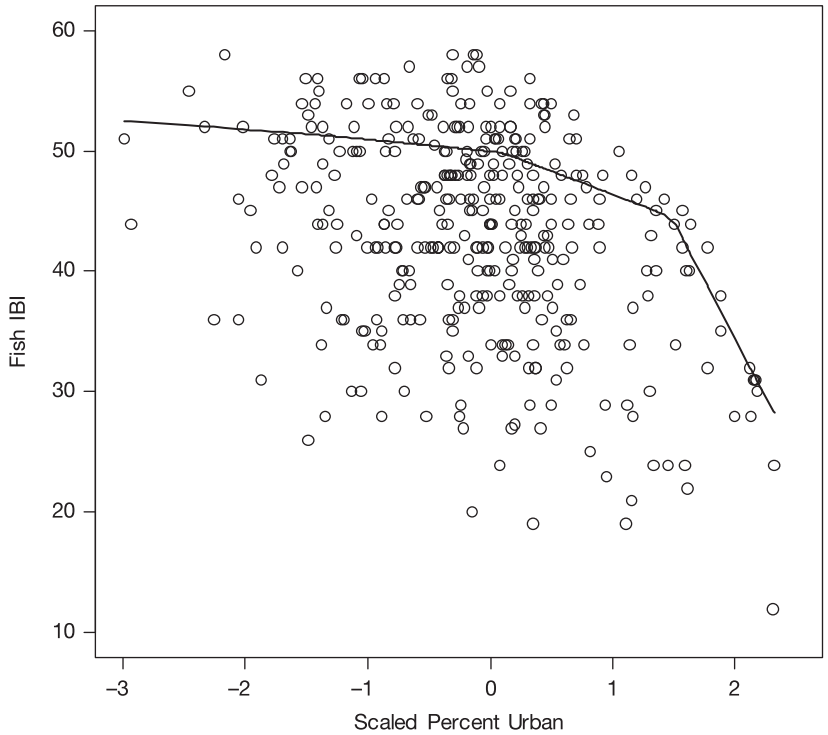


Fig. 1 Fish index of biotic integrity (IBI) scores plotted by scaled (i.e., *z*-scores) percent impervious cover. Zero on the *x*-axis corresponds to ~2% impervious cover; 1 corresponds to ~10%; 2 to ~50%. The line in the plot follows the local 75th percentile ($\lambda = 1$).

sites according to several or more geographic (e.g., latitude and longitude), physical (e.g., stream size, gradient), chemical (e.g., alkalinity, conductivity), physiochemical (e.g., temperature, dissolved oxygen), and hydrologic (e.g., base flow index) properties. For a novel site being sampled, the site is classed into its environmental setting, and the probabilities of capture for each taxon found at that site-type in the reference population is summed to give the expected number of taxa under the reference condition. The actual number of expected taxa found in the observed sample is then compared to the expected number under reference condition, and expressed as the ratio of observed over expected. The up-front effort to develop expectations is not trivial, and may be seen as a disadvantage compared to a multimetric index, but the advantage is that once done, scores from different stream classes can be compared on equal footing.

Although multimetric indexes and O/E models each have advantages and disadvantages, there are also advantages and disadvantages in the choice of assemblage used as a biological indicator. In larger rivers and streams, fish are relative easy to identify, but take great skill to sample effectively. Macroinvertebrates and diatoms are relatively easy to sample, but require great skill to identify. Fish are long lived, and can reflect episodic or intermittent disturbances that shorter lived macroinvertebrates or diatoms might recover from before being sampled. Conversely, macroinvertebrates and diatoms respond rapidly to disturbances.

Properties of a Multimetric Biological Index

Several properties of multimetric biological indexes make them particularly useful for evaluating ecosystem condition:

1. focus on biological endpoints to define condition;
2. use of reference condition (no disturbance or minimal disturbance) as a benchmark;
3. organization of sites into classes (e.g., large streams, small streams, wetlands), each with a select set of environmental characteristics;
4. assessment of change caused by human activities;
5. standardized sampling, laboratory, and analytical methods;
6. numerical scoring of sites to reflect position along a disturbance gradient; and
7. numerical and verbal expressions of biological condition that can be easily understood byscientists, citizens, and policy makers.

Unlike single-attribute chemical measures of water quality, analytical tools such as multimetric indexes or O/E models enhance practitioners' ability to measure condition in a manner that communicates the severity and extent of biological impairment. When combined with surrogate or direct measures of human activities such as water chemistry, pollutant loadings from wastewater sources, or land use, they function as response variables against those measured gradients to assess the severity of degradation. And

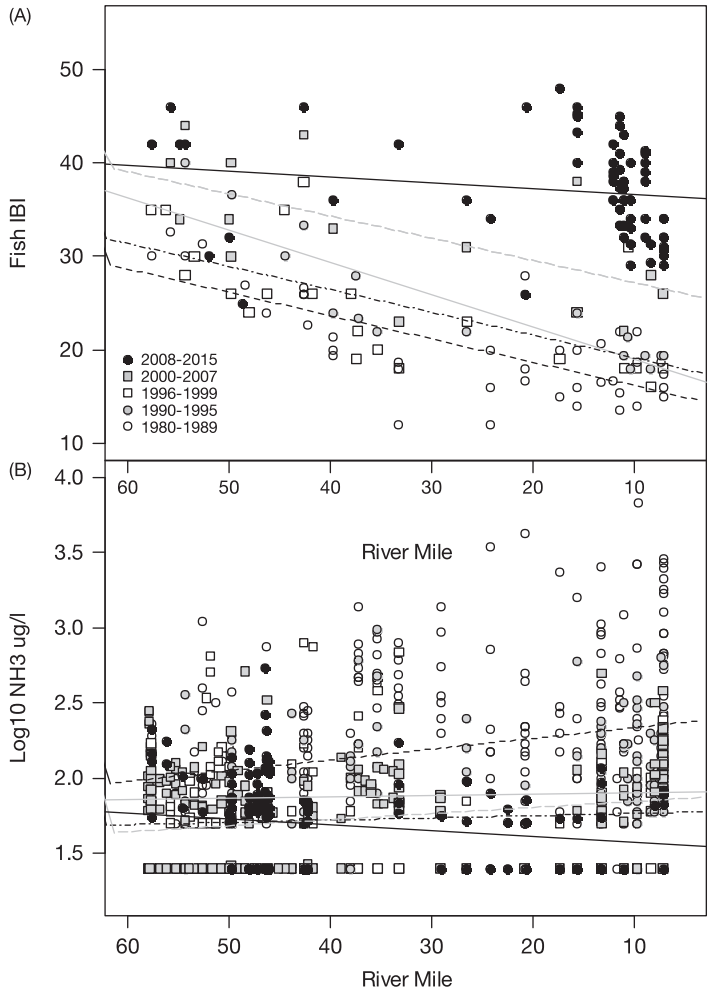


Fig. 2 (A) Fish index of biotic integrity scores obtained from the Cuyahoga River (Ohio, USA) grouped by survey years. (B) Ammonia nitrogen concentrations from the Cuyahoga River similarly grouped by survey year. Lines in the plots are from ordinary least squares regression.

when combined with multivariate ordination techniques like canonical correspondence analysis, help facilitate the identification of stressors when measured environmental variables are associated with diffuse sources of pollution. In addition to being scientifically rigorous, multimetric biological indexes or O/E models are also policy relevant. They are, for example, sensitive enough to provide reliable assessments of both existing and emerging problems and to evaluate the effectiveness of environmental policies and programs. Integrative approaches to biological monitoring directly support efforts to attain the integrity called for in national and international policy initiatives.

Applying Biological Integrity in Resource Management

Site-specific scores generated from either a multimetric index or an O/E model show how closely the biological condition at a given site matches the reference condition. When scores from a collection of sites are mapped along a gradient of disturbance, these scores quickly convey the effect of that disturbance on the biological integrity of the resource (Fig. 1). Impervious surfaces are a convenient measure of disturbance, and serve as a proxy for the combined effects of toxicity and hydrologic alteration associated with urban stormwater. In numerous case examples from disparate urban areas around the globe, biological index scores plotted against percent impervious cover essentially follow the same pattern as in Fig. 1. Because biological index scores were so effective at conveying the effect of stormwater on streams, enough public attention was drawn to the problem that management policies and strategies to address the problem became politically tenable. For example, these results informed modification of the general municipal stormwater permit issued to medium and large urban areas in the US. Additionally, many urbanizing areas have adopted construction performance standards that include setbacks, and provisions to treat runoff and maintain pre-construction hydrology.

Similar to the impervious cover example, biological index scores arrayed along specific stressor gradients have identified weaknesses in chemical water quality criteria. For example, water quality standards for ammonia'nitrogen were determined to be

under-protecting aquatic communities in general, and sensitive components (e.g., unionid mussels) in particular, leading to more stringent criteria. Conversely, water quality standards for some metals were determined to be over-protective, leading to more refined criteria like the biotic ligand model for copper.

A common thread running through these last examples is the benefit of biological monitoring on informing management decisions. In terms of managing pollution in surface waters, this thread can clearly be seen running through various provision of the Clean Water Act (CWA), conceptually if not in practice. Although the CWA calls for restoration and maintenance of biological integrity, few states have adopted numeric standards for biological criteria. In contrast, the EU WFD sets a biological goal, and defines operational methods for assessing that goal. For US states with biological criteria, information from biological monitoring is programmatically more integrated. For example, in Ohio, biological index scores obtained from monitoring near permitted wastewater discharges are consulted prior to making permit modifications. In Maine, biological criteria are a component of a numeric water quality standard for nutrients. In Florida, biological criteria and index scores are used in a stressor identification framework to identify impaired waters and associated causal pollutants, and set appropriate limits for the identified pollutants. Fully integrated, biological information and criteria can inform whether an increase in the discharge of specific pollutant has a reasonable potential to cause harm, even if that increase does not result in concentrations exceeding a numeric chemical criterion. Similarly, biological criteria can be invoked to prevent degradation of existing conditions if an activity, such as construction, threatens the status of a waterbody. Conversely, where degradation is allowed to occur, biological criteria can set mitigation standards.

Arguably the most important aspect of fully integrated biological information is that because biological communities are persistent through time, measures from those communities allow management actions to be translated into tangible and quantifiable results, and tracked through time. The success of the Construction Grants Program to modernize wastewater infrastructure is well-documented in Ohio (Fig. 2; IBI or Invertebrate Community Index [ICI]; ammonia nitrogen [NH_3] through time for Cuyahoga). The Cuyahoga River, once the poster child for the CWA, now meets the basic CWA goal throughout much of its length. Considering the starting point, that is no small accomplishment, and the degree of impairment and ensuing improvement would not be so apparent in the absence of biological index scores. More interestingly, however, is that biological monitoring has also been able to demonstrate the net effectiveness of a suite of agricultural conservation practices applied over a broad geographic region. Specifically, the net effect of conservation practices applied in Ohio's Eastern Corn Belt Plain Ecoregion between 2003 and 2012 was a mean increase of 4 IBI points (about 8% of the possible range) in headwaters.

Biological criteria provide a concrete goal for protection and restoration. However, because scores or expectations are calibrated to a reference population, caution should be applied when interpreting scores from sites that fall on the edge or outside the calibration range, be that geographically or physically (e.g., stream size). Similarly, because a defined percentile (e.g., the 25th) of index scores from the reference population is often used to set the biocriterion for a multimetric index, caution is needed when interpreting scores from sampled sites that fall close to either side of the criterion. This is especially important when examining scores collected from probability or systematic surveys of watersheds, where the distribution of survey scores may match (or even have a higher mean) than the distribution from reference sites (Fig. 3—box and whisker and longitudinal reach). Scores obtained from longitudinal reach surveys conducted around point sources are easier to interpret against the biocriterion because the natural variation within a reach is expected to be low (i.e., scores within a short reach should be autocorrelated).

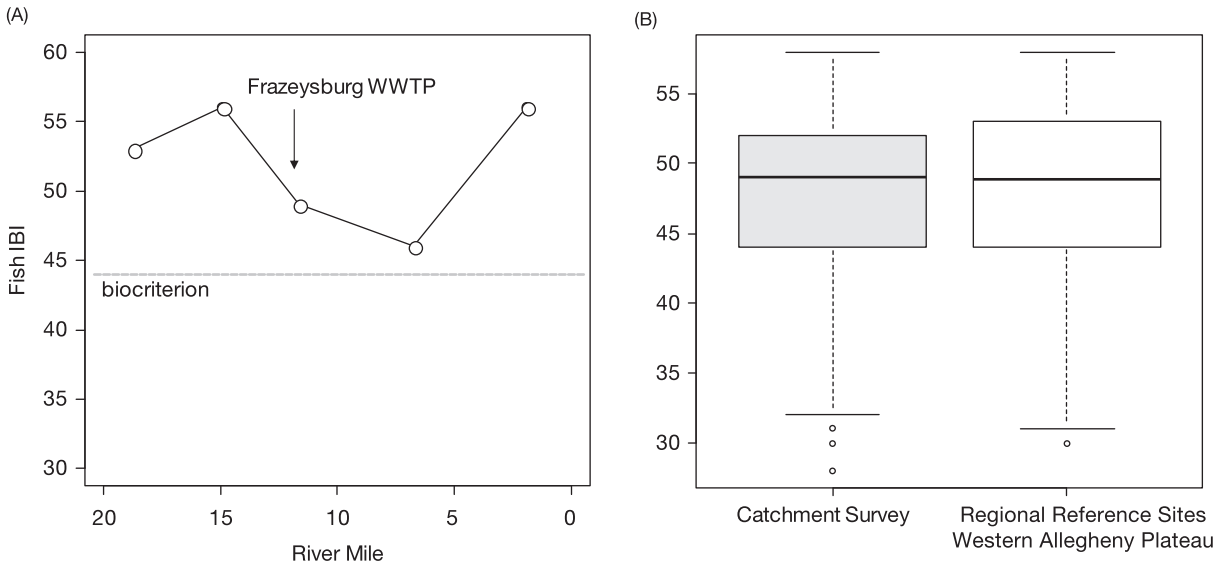


Fig. 3 (A) Fish index of biotic integrity scores from a longitudinal reach survey of a small stream receiving treated wastewater from a wastewater treatment plant (WWTP). The direction of stream flow in orientation to the *x*-axis is from left to right. The numeric biocriterion (IBI = 44) is shown. (B) Distributions of fish IBI scores from a localized catchment survey conducted in the Western Allegheny Plateau Ecoregion of Ohio (gray-shaded box plot), and from regional reference sites throughout the Western Allegheny Plateau of Ohio. The distributions are statistically identical.

Summary

Biological integrity depends on the sum of many parts, and systems with a high degree of biological integrity demonstrate that those parts are largely working and intact. Measures of biological integrity provide a framework for judging where a system is positioned on the human disturbance gradient, how susceptible that system may be to further disturbance, and for highly disturbed systems, determining feasible levels of restoration. Interpreting measures of biological condition and the desired state of that condition depend on reference expectations. The reference condition has traditionally been defined as "natural" or least disturbed, but also can be set to an agreed upon desired state. Integrated into resource monitoring and protection programs, measures of biological integrity can be used to assess and demonstrate the effectiveness of management.

See also: Human Ecology and Sustainability: Ecological Footprint

Further Reading

Identifying Reference Conditions

Angradi, T.R., Bolgrien, D.W., Jicha, T.M., Pearson, M.S., Hill, B.H., Taylor, D.L., Schweiger, E.W., Shepard, L., Batterman, A.R., Moffett, M.F., Elonen, C.M., 2009. A bioassessment approach for mid-continent great rivers: The upper Mississippi, Missouri, and Ohio (USA). Environmental Monitoring and Assessment 152 (1–4), 425–442.
Bouchard Jr, R.W., Niemela, S., Genet, J.A., Yoder, C.O., Sandberg, J., Chirhart, J.W., Feist, M., Lundeen, B., Helwig, D., 2016. A novel approach for the development of tiered use biological criteria for rivers and streams in an ecologically diverse landscape. Environmental Monitoring and Assessment 188 (3), 1–26.
Dufour, S., Piégay, H., 2009. From the myth of a lost paradise to targeted river restoration: Forget natural references and focus on human benefits. River Research and Applications 25 (5), 568–581.
Stoddard, J.L., Larsen, D.P., Hawkins, C.P., Johnson, R.K., Norris, R.H., 2006. Setting expectations for the ecological condition of streams: The concept of reference condition. Ecological Applications 16 (4), 1267–1276.

Developing Biotic Indexes

Clarke, R.T., Wright, J.F., Furse, M.T., 2003. RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. Ecological Modeling 160 (3), 219–233.
Davis, W.S., Simon, T.P.e., 1995. Biological assessment and criteria: Tools for water resource planning and decision making. CRC Press.
Garey, A.L., Smock, L.A., 2015. Principles for the development of contemporary bioassessment indices for freshwater ecosystems. In: Advances in watershed science and assessment. Springer International Publishing, pp. 233–266.
Karr, J.R., 1981. Assessment of biotic integrity using fish communities. Fisheries 6 (6), 21–27.
Van Sickle, J., Hawkins, C.P., Larsen, D.P., Herlihy, A.T., 2005. A null model for the expected macroinvertebrate assemblage in streams. Journal of the North American Benthological Society 24 (1), 178–191.

Introduction to Appreciative Inquiry

Watkins, J.M., Cooperrider, D., 2000. Appreciative inquiry: A transformative paradigm. OD Practitioner 32 (1), 6–12.

Biotic Inexes and Environmental Gradients

Riva-Murray, K., Bode, R.W., Phillips, P.J., Wall, G.L., 2002. Impact source determination with biomonitoring data in New York state: Concordance with environmental data. Northeastern Naturalist 9 (2), 127–162.

Relevant Websites

CEH—https://www.ceh.ac.uk/services/rivpacs-reference-database.
EPA—https://www.epa.gov/national-aquatic-resource-surveys.
EPA—https://www.epa.gov/sites/production/files/2015-10/documents/tech_memo_4_oct_15.pdf.

Biomagnification

KG Drouillard, University of Windsor, Windsor, ON, Canada

Introduction

This article provides an overview of biomagnification as it applies to bioaccumulation of hydrophobic organic contaminants in ecosystems. While the term biomagnification has been applied to other pollutant classes including certain metals, the origins of the term, major case studies, and advancements in the mechanistic interpretation of biomagnification have largely been focused on hydrophobic organohalogen compounds. The first section of this article defines the term biomagnification, related terminology, and traces the origins of the term. The second section provides a brief overview of major empirical evidence documenting biomagnification in various food webs, and the last sections provide a summary of alternative mechanisms that have been used to explain the biomagnification phenomena.

Definitions and Terminology Related to Biomagnification

The term biomagnification has classically been defined as the condition where the contaminant concentration in an organism exceeds the contaminant concentration in its diet when the major chemical exposure route to the organism is from food. By extension the term food web biomagnification has been defined as the increase in contaminant concentration with increasing trophic status of organisms sampled from the same food web.

Biomagnification and food web biomagnification were originally coined from observations of chlorinated pesticide bioaccumulation in aquatic food webs. However, the term biomagnification has been applied to other contaminants including mercury, heavy metals, and certain compounds of biogenic origin. The first demonstration of biomagnification was described for dichlorodiphenyldichloroethane (DDD), closely related to the pesticide dichlorodiphenyltrichloroethane (DDT), in Clear Lake California. Rachel Carson subsequently used the term `biological magnifiers' in her book, *Silent Spring*, to describe how earthworms concentrate DDT residues from soil in their bodies and transfer these residues to robins who consume them which in turn achieve even greater concentrations of the pesticides than worms. The term `biological magnification' was later used by Woodwell to describe the `systematic increase in DDT residues with trophic level' in his description of DDT trophodynamics in a salt marsh near Long Island, New York. Biological magnification subsequently became truncated to the commonly used term `biomagnification' in later years.

The mechanism of biomagnification as applied to organic chemicals, particularly compounds demonstrating physical properties of low water solubility and high hydrophobicity, was intensely studied and vigorously debated in the 1980s and 1990s. During this period, biomagnification was conceptually distinguished from the process of bioconcentration which refers to chemical bioaccumulation due to exposure of contaminant across respiratory exchange surfaces (i.e., gills and lungs). Research conducted in the 1960s and 1970s demonstrated how physical–chemical properties controlled environmental partitioning and diffusive flux of hydrophobic substances. Hydrophobic organic contaminants tend to distribute preferentially to organic phases which includes organic carbon of soils and sediments and lipid phases of organisms. Equilibrium partitioning theory was subsequently used to equate bioconcentration in animals to the equilibrium lipid/water distribution coefficient. Although equilibrium partitioning theory described laboratory bioconcentration data well and predicted laboratory bioconcentration factors (BCFs; defined as the lipid-normalized chemical concentration in the animal divided by the chemical concentration in water), it failed to fully account for elevated contaminant concentrations accumulated by upper-trophic-level animals in the field.

The failure to validate equilibrium partitioning as a theory explaining biomagnification prompted redefinition of the term, as applied to hydrophobic organic susbstances, to describe the thermodynamic context of biomagnification. Under this new definition, biomagnification refers to the condition where the chemical potential achieved in an animal's tissues exceeds the chemical potential in its food and its surrounding environment. Similarly, food web biomagnification was redefined as the increase in chemical potential of organisms with increasing organism trophic status. In practice, chemical potentials are not directly measured, but rather are compared relatively across different samples by normalizing the chemical concentration in a sample by the sample partitioning capacity for the chemical/sample matrix of interest. Since hydrophobic organic contaminants distribute primarily to neutral lipids within organisms, expression of lipid-normalized chemical concentrations have been used as surrogate measures of chemical potentials when comparing biomagnification between biological samples. Alternatively, chemical fugacity is used as a proxy for chemical potentials when comparing equilibration of contaminants between interacting abiotic and biotic samples. These data analysis methods apply to hydrophobic organic chemicals but do not necessarily apply to mercury or other contaminant classes which undergo biomagnification by the classic definition, but do not exhibit preferential internal distribution to lipids within animals.

The biomagnification factor (BMF) for organic contaminants is defined as the ratio of the lipid-normalized chemical concentration in the animal to the lipid-normalized chemical concentration in its diet. A BMF value greater than 1 indicates that the animal has achieved a greater chemical potential than its diet. Since organisms may include multiple food items in their diet, the

BMF can also be expressed according to the weighted average lipid-normalized chemical concentrations in its various food items. Similarly, when the BCF value is shown to exceed the *n*-octanol/water partition coefficient (K_{OW}; a standard laboratory-measured partition coefficient used as a surrogate measure of the equilibrium lipid/water partition coefficient) this indicates that the chemical potential achieved in the animal exceeds that of water. Similar expressions can be derived for air-breathing animals by comparing the lipid-normalized concentration in the animal/air concentration ratio with the octanol/air partition coefficient.

Determination of food web biomagnification requires establishment of the trophic level of different organisms included in the sampling program. Traditionally this has been carried out using diet analysis and establishing discrete trophic steps (see **Fig. 1**). More recently, emphasis has been placed on use of stable isotopes of carbon and nitrogen to define continuous trophic positions for different organisms in a sampled food web. The food web magnification factor (FWMF) has been defined as the slope generated from a regression of the logarithm of lipid-normalized chemical concentrations in biota expressed against trophic level on the independent axis.

Empirical Field Data Supporting Biomagnification

Hunt and Bischoff provided the first data demonstrating progressive bioaccumulation and increases in concentrations of the chlorinated insecticide DDD through an aquatic food web. DDD was applied to Clear Lake, California during three administration events in 1949, 1951, and 1957. Administrations were designed to achieve a nominal concentration of DDD in water of 50 mg l^{-1}, although reportedly water residues never achieved such levels. Mortalities of fish-eating birds were observed within months after the second and third applications, with the population of western grebes decreasing from 1000 pairs prior to DDD administration to less than 30 pairs in 1960. Food web sampling and residue analysis indicated phytoplankton achieved concentrations of approximately 5 mg g^{-1}, pelagic fish contained between 50 and 300 mg g^{-1} and a brown bullhead contained 2500 mg g^{-1} of DDD. DDD concentrations in western grebes and California gulls were reported at more than 2000 mg g^{-1}. Soon after, other studies began documenting DDT bioaccumulation in different food webs. Woodwell *et al.* determined DDT concentrations in water, soil, plankton, invertebrates, mussels, fish, and fish-eating birds in a salt marsh south of Long Island, New York. DDT increased from 0.04 mg g^{-1} in plankton to 75 mg g^{-1} in ring-billed gulls. Plankton concentrations were 800-fold higher than residues measured in water. Invertebrates and fish exhibited intermediate concentrations of DDT compared to plankton and birds, consistent with their trophic status. Unfortunately, the above studies did not determine lipid concentrations of samples submitted for insecticide residues. As such, these data could not be used to test the thermodynamic criteria associated with equilibrium partitioning theory and biomagnification.

Advancements in analytical technology in the 1980s, particularly with the development of capillary gas chromatography columns, greatly increased the ability of environmental scientists to examine individual chemical concentrations in more complex field matrices. This led to a plethora of food web data sets documenting biomagnification of other organic contaminants including polychlorinated biphenyls (PCBs). Two major studies documenting food web biomagnification of individual PCB congeners were

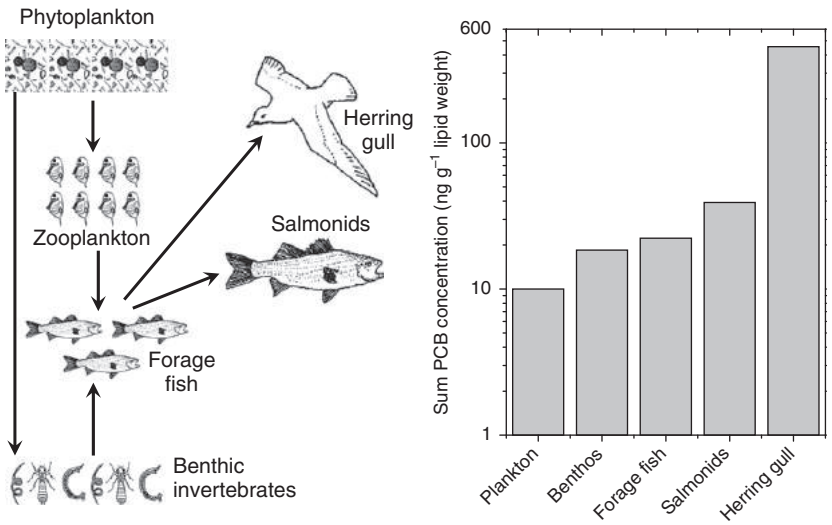


Fig. 1 Food web biomagnification of polychlorinated biphenyls (PCBs) in Lake Ontario. Data for aquatic organisms collected in Lake Ontario during 1984 from Oliver, B.G., Niimi, A.J., 1988. Trophodynamic analysis of polychlorinated biphenyl congeners and other chlorinated hydrocarbons in the Lake Ontario ecosystem. Environmental Science and Technology 22, 388–397. Data on Herring gulls collected in Lake Ontario during 1985 from Braune, B.M., Norstrom, R.J., 1989. Dynamics of organochlorine compounds in herring gulls: III. Tissue distribution and bioaccumulation in Lake Ontario Gulls. Environmental Toxicology and Chemistry 8, 957–968.

published in 1988. Oliver and Niimi measured individual PCB concentrations in water, sediments, amphipods, slimy sculpin, alewife, and lake trout from Lake Ontario. The authors also measured lipid contents in the biological samples allowing them to directly test predictions of the equilibrium partitioning theory. Their data demonstrated increases in lipid-normalized PCB concentrations with increasing trophic status (see Fig. 1). Salmonids were also shown to have fivefold higher lipid-normalized concentrations than predicted from equilibrium partitioning theory based on residues in water. Herring gulls from the same lake collected 1 year later in another study demonstrated lipid-normalized PCB concentrations that were tenfold higher than measured for salmonids by Oliver and Niimi. Conolly and Pederson also demonstrated that the fugacity ratio of rainbow trout/water exceeded a value of 1 for PCBs having a log K_{OW} value of 4 and greater in Lake Ontario. The authors demonstrated that the trout/water fugacity ratio for PCBs increased with increasing chemical K_{OW} up to values from 10 to 100 for PCBs having log K_{OW} values of 6 and higher. The same authors also demonstrated progressive increases in the animal/water fugacity ratio for PCBs with animal trophic status in the Lake Michigan food web. PCB animal/water fugacity ratios ranged from 3 to 5 for white fish and chub occupying a trophic level of 2 and up to a value of 14 for fish occupying a trophic level of 4. Similar case studies of food web biomagnification using lipid-normalized data sets have subsequently been demonstrated in several other aquatic systems including all five Great Lakes, Lake Baikal, and in agricultural and arctic terrestrial ecosystems.

Other data sets have shown the relationship between hydrophobic organic contaminant bioaccumulation and food chain length or number of trophic steps within the system. Using data generated from the Ontario sport fish contaminant surveillance program, Rasmussen *et al*. examined PCB bioaccumulation in lake trout from a large number of lakes in Ontario, Canada. The authors demonstrated that planktivorous lake trout from lakes lacking suitable forage fish exhibited lower PCB bioaccumulation compared to piscivorous lake trout from lakes containing forage fish. Finally, lake trout from lakes containing both forage fish and mysids achieved the highest contaminant residues. Lakes were also categorized and analyzed by location to remove confounding factors associated with different loading sources to individual systems. The authors attributed the lake to lake differences in PCB bioaccumulation by lake trout to reflect differences in food chain length. A more recent study documented enriched toxaphene bioaccumulation in fish from Lake Labarge, Yukon Territory, Canada as contrasted with other subarctic lakes which showed much lower bioaccumulation of the same contaminants in fish. Lake Labarge was isolated from known pollutant sources and thought to receive most of its inputs via atmospheric deposition. Water concentrations of toxaphene were also found to be similar in Lake Labarge relative to the other lakes which showed lower toxaphene bioaccumulation in fish. The major difference noted for Lake Laberge lake trout, burbot, and lake whitefish was that fish from this lake were feeding at higher trophic levels as revealed both by diet analysis and trophic enrichment of stable nitrogen isotopes. This study, similar to that of Rasmussen's work provided the empirical linkage between ecosystem structure, number of trophic links, and magnitude of biomagnification realized in top predator fish.

Mechanism of Biomagnification

Dietary Exposure

A number of mechanisms have been proposed to describe the biomagnification process as it applies to persistent, hydrophobic organic compounds. The first model published to describe biomagnification of the insecticide DDT described the lipid co-assimilation mechanism. In this model, both lipids and contaminants are efficiently assimilated from food; however, a smaller fraction of lipids are retained as a result of metabolism of these nutrients to satisfy energetic requirements. Recalcitrant contaminants are retained in tissues and over time, in conjunction with the number of feeding events, magnify in concentration over that of ingested food. Under this mechanism, the maximum biomagnification potential in nondeterminant growing animals is inversely related to growth-conversion efficiency (i.e., rate of tissue growth relative to food consumption) when contaminant elimination from the animal approaches a value of 0. For determinant growers, biomagnification will continue to increase with age as a function of number of feeding events. In practice, most environmental contaminants do exhibit elimination, which will attenuate biomagnification in proportion to the magnitude of the elimination rate coefficient. In this case, the steady-state biomagnification factor will be positively related to the feeding rate of the animal and chemical assimilation efficiency from the diet and inversely proportional to the elimination rate coefficient and growth rate.

The lipid co-assimilation mechanism was also used to explain food web biomagnification. In this case, biomagnification as achieved in top predator organisms is assumed to correspond to the inverse of ecological efficiency. Thus, the low energy transfer efficiency across trophic levels ($<10\%$) coupled with efficient chemical transfer efficiencies and high chemical retention among different organisms results in progressively increased residues through successive trophic steps. This model indicates that the number of trophic levels and trophic transfer efficiencies across each step specify the maximum contamination achieved for top predators whereas growth and elimination by individual organisms attenuates food web biomagnification.

Although lipid co-assimilation as a mechanism of biomagnification is consistent with concepts of bioenergetics and trophodynamics, environmental chemists were quick to point out that such a mechanism is not consistent with chemical bioaccumulation through passive partitioning mechanisms as is thought to occur for hydrophobic organic chemicals. Lipids, fat-soluble vitamins, and hydrophobic organic contaminants cross biological membranes by passive diffusion. In the case of the proposed lipid co-assimilation model of biomagnification, net chemical diffusive flux would have to proceed in a direction of low concentration (ingested food and digesta of the gastrointestinal (GI) tract) to high concentration (animal tissues). The GI

magnification model was subsequently developed as a competing model with lipid co-assimilation. This model is able to account for biomagnification while preserving chemical diffusion as the major mechanism of chemical flux between the organism and its gut contents.

Gastrointestinal magnification was first described in 1988. The model considers the GI tract and its contents as a separate compartment from animal tissues. The premise of the GI magnification model is that nutrient and lipid absorption occur independent of contaminants in the GI tract. The absorption of nutrients from digesta decreases both the volume and partitioning capacity of gut contents as digestion proceeds along the length of the intestines. The failure of mass balance within the GI tract compartment and loss of partitioning capacity raises the chemical potential of digesta in the GI tract above that of ingested food, providing the necessary gradient on which net diffusion can proceed from GI tract to animal even if the animal has a higher chemical potential than the food it has ingested. The animal thus equilibrates with the elevated chemical potential of its GI contents rather than its food.

Under the GI magnification model, the change in chemical partitioning capacity of feces relative to food and the volume reduction of feces produced relative to food consumed (i.e., diet absorption efficiency) provides the upper limit for the maximum biomagnification potential that can be achieved in an animal. These limits may vary according to the diet absorption efficiency for a given diet type and/or differences in digestive physiology between different species feeding on similar diets. The GI magnification model has been subject to experimental and field validation through studies that demonstrated increases in chemical potential of animal GI contents above that of ingested food. Careful laboratory measurements indicate maximum biomagnification potentials on the order of 3–12 in fish which appear to be consistent with field biomagnification factors experienced by fish. The GI magnification model has subsequently been adopted in numerous food web bioaccumulation models applied to hydrophobic organic contaminants. Food web bioaccumulation models allow consideration of relative exposures and biomagnification potentials in animals exposed to multiple diet items as defined by the diet matrix established for each species being modeled.

Recent studies have suggested amendments to the GI magnification model to account for additional physiological factors which may increase the biomagnification potential of an animal beyond diet assimilation and partitioning changes of feces relative to food. One of the simplifying assumptions applied in the original GI model solution was that the mass transport parameter describing chemical flux from gut contents to animal is equal to the mass transport parameter describing flux from animal to its gut contents. This asserts that uptake and elimination flux is diffusion-controlled and coupled throughout the length of the GI tract. However, evidence on hexachlorobenzene bioaccumulation in rats suggested that uptake occurs primarily in the upper GI tract whereas fecal elimination occurs predominately in the colon. Experimental studies on ring doves showed that PCB dietary assimilation efficiencies in ring doves during the uptake phase were higher than feces/animal exchange efficiency measured during the depuration phase. Similar observations were documented in humans.

The fat-flush model and micelle-mediated diffusion model have been suggested as potential submodels for use to augment the GI magnification model. Both the above models are used to explain the phenomena of decoupling of the site and timing of uptake and elimination processes in the GI tract. According to the fat-flush model, fatty acids assimilated by enterocytes of the intestinal mucosa of the small intestine become resynthesized into triglycerides and incorporated into growing chylomicron vesicles. This growth in lipid content of enterocytes increases the partitioning capacity of these cells and temporarily dilutes their chemical potential relative to blood and other body compartments. The lower chemical potential of enterocytes then favors chemical assimilation by diffusion. This process maximizes chemical absorption and minimizes losses of chemical from small intestine cells back to the lumen of the GI tract. When chylomicrons reach a critical size, they are released by active transport from the enterocyte into the circulatory system which again causes a temporary dilution of the blood compartment relative to other body tissues. Following the absorption of dietary lipids (and assimilated contaminant) from blood by other tissues, the chemical potential of blood once again re-equilibrates with other body compartments and becomes maximized. During the fasting state, when chemical potential in blood is highest, fecal elimination becomes more pronounced. At this time, the gut contents are found primarily in the large intestine where fecal elimination has been shown to take place. The fat-flush model therefore describes decoupling in both the site and timing of contaminant assimilation compared to fecal elimination.

The micelle-mediated diffusion model focuses on the physiological role of mixed micelles as vectors for lipid, hydrophobic vitamin and contaminant uptake in the GI tract. Mixed micelles are produced in the intestine as a result of the interaction of bile salts and fatty acids. These amphiphilic vesicles diffuse through the unstirred water layer (UWL) between the gut lumen and intestinal mucosa. Mixed micelles are capable of dissolving long-chain fatty acids, fat-soluble vitamins, as well as other hydrophobic compounds including contaminants in their interiors and transporting these compounds across the UWL. Recent physiological evidence indicates that mixed micelles are unidirectional in their movements between the lumen to enterocyte. A pH microgradient stimulates the breakup of mixed micelles at the interface of the intestinal mucosa of the small intestine. Thus, mixed micelles appear to be involved in the efficient assimilation of hydrophobic contaminants in the upper part of the digestive tract but do not facilitate elimination of chemical from enterocytes back to the lumen of the gut compartment.

Both the mixed-micelle and fat-flush models reflect extensions of the GI magnification model that bring about physiological realism to the digestive process. Current calibration of these models in birds and humans suggests that maximum biomagnification factors may be higher by a factor of 2–4 (i.e., total biomagnification factors ranging from 15 to 20 or higher) than predicted by the original GI magnification model. Calibration of the fat-flush model or mixed-micelle models in fish is yet to be completed. Further research to calibrate maximum bioaccumulation potentials in a wider variety of animal species as well as calibrated animal to gut and gut to animal transfer efficiency terms are required to substantiate these new model predictions and adopt them into food web bioaccumulation models as has been performed with the GI-magnification model.

Nondietary Mechanisms Explaining Biomagnification

The earliest critics of biomagnification identified equilibrium partitioning as the main mechanism of bioaccumulation and suggested that the phenomena of food web biomagnification could be explained primarily by differences in whole-body lipid content, and hence chemical partition capacities, of upper-trophic-level animals relative to lower-trophic-level organisms. As described above, food web bioaccumulation data sets generated in the late 1980s and 1990s provided lipid-normalized chemical concentration data and these studies were consistent with the thermodynamic definition of biomagnification. However, other alternative mechanisms of biomagnification, which do not involve special properties associated with dietary exposures, have been proposed.

Differences in spatially integrated exposures of sampled animals arising due to habitat size and/or differences in migration movements of organisms could potentially result in similar observations as biomagnification particularly in environments where the contaminant distribution in sediments and water is heterogeneous or subject to point sources. Smaller animals are likely to exhibit small spatial movements and be more reflective of contamination conditions at the local site of capture. Larger animals such as piscivorous fish may exhibit larger spatial movements and consequently integrate chemical exposures over broader spatial scales. Birds may carry residues over very long distances across their migration routes. Indeed the phenomena of biological vectors of pollution related to major spawning migrations of fish and seabirds flying to breeding sites have been recently described. Food web sampling programs for contaminants rarely consider the spatial scale of sample collections as it relates to the potential movements of organisms included within their collections. Spatial movements can confound interpretation of biomagnification factors when all animals are collected from the same location. If animals are collected at a highly contaminated site, food web biomagnification may appear attenuated as a result of high locally accumulated residues in benthic invertebrates and zooplanktons. Similarly, biomagnification trends may appear exaggerated when animals are sampled at relatively clean locations but are situated near enough hot spots that some of the larger animals are affected by the more distant contaminated areas.

Similar to the spatial scale described above, temporally explicit exposures may also confound biomagnification observations. Chemical elimination rate coefficients for negligibly biotransformed contaminants are inversely related to body size. Under conditions of pulses in environmental loadings, smaller, lower-trophic-level organisms are more likely to reflect equilibrium with water whereas larger organisms may exhibit lags in their ability to equilibrate with water during or after a pulse. For example, following reductions in water contamination after a seasonal pulse in inputs, as may be experienced during spring melt, phytoplankton and zooplankton may be capable of depurating their residues at a sufficient rate to maintain equilibration with the drop in water concentrations. However, larger fish will take longer to depurate their residues to water and will exhibit both higher concentrations and higher chemical potentials than their zooplankton/plankton counterparts. If the frequency of environmental pulse inputs is faster than the steady-state time of larger fish then this disequilibrium condition may be maintained. For some contaminants which do not achieve steady state in organisms, different animal ages also need to be considered when comparing residues among populations or between populations of species having different age structures.

Another confounding factor arises due to rapid changes in animal lipid contents, either through growth or weight loss. When an animal loses weight and lipids at a faster rate than it can lose contaminant, it concentrates its tissue residues and raises its chemical potential above its previous state even though net chemical flux proceeds in the direction of elimination. Rapid lipid depletion, and subsequent tissue concentration of contaminants, is likely to be common in animals that undergo seasonal cycles of weight gain and lipid loss or in animals that exhibit bioenergetic bottlenecks at critical times in their life history. Such observations were reported in depuration experiments involving birds and fish. In the case of birds, contaminant residues were found to become concentrated in blood following weight losses experienced by the animals during spring warming. The opposite was noted for warm-water fish, the yellow perch, where winter weight losses due to prolonged fasting caused an increase in chemical potential in animal tissues despite the fact that the study was measuring chemical elimination. Other examples documenting rapid weight and lipid loss during specific life-history points and subsequent tissue magnification of PCBs include metamorphosis in amphibians and pipping (hatching) of chicks.

The opposite condition of solvent depletion occurs during rapid growth. Extremely rapid turnover times demonstrated by algae during peaks in primary production result in growth rates that are faster than chemical-uptake coefficients. This causes phytoplankton to exhibit chemical potentials that are lower, and less than equilibrium, compared to water. Rapid growth dilution as experienced in juvenile animals will also reduce biomagnification factors due to high growth-conversion efficiencies experienced during these life stages.

Summary

Biomagnification is a well-documented phenomenon where persistent hydrophobic organic contaminant concentrations in an animal become elevated over and above its food. This increase in contaminant concentration propagates through successive trophic steps in a food web. Later definitions of biomagnification and food web biomagnification have imposed thermodynamic criteria specifying that biomagnification reflects a nonequilibrium process in which the chemical potential in an animal is elevated above that of its diet and environment. The GI magnification model and the new amendments to this framework explain how exposures through the diet can raise the chemical potential of the animal above that of its food and environment. The above mechanisms apply specifically to hydrophobic organic contaminants. However, other contaminants such as mercury are assimilated and distributed through tissues by different processes. Thus, while mercury conforms to the classic definition of

biomagnification and food web biomagnification it is difficult to evaluate the behavior of this compound in the context of the thermodynamic definition of biomagnification. Further understanding of uptake, assimilation, and elimination mechanisms and the free energy relationships associated with these processes are needed to provide a unifying theory of biomagnification across other contaminant types.

See also: Evolutionary Ecology: Endemism; Ecological Niche. General Ecology: Tolerance Range

Further Reading

Blais, J.M., Macdonald, R.W., Mackay, D., *et al*., 2007. Biologically mediated transport of contaminants to aquatic systems. Environmental Science and Technology 41, 1075–1084.

Braune, B.M., Norstrom, R.J., 1989. Dynamics of organochlorine compounds in herring gulls: III. Tissue distribution and bioaccumulation in Lake Ontario Gulls. Environmental Toxicology and Chemistry 8, 957–968.

Carson, R., 1962. Silent Spring. Greenwhich, CN: Fawcett Publications.

Connell, D.E., 1989. Biomagnification by aquatic organisms – A proposal. Chemosphere 19, 1573–1584.

Connolly, J.P., Pedersen, C.J., 1988. A thermodynamic based evaluation of organic chemical accumulation in aquatic organisms. Environmental Science and Technology 22, 99–103.

Gobas, F.A.P.C., Muir, D.C.G., Mackay, D., 1988. Dynamics of dietary bioaccumulation and faecal elimination of hydrophobic organic chemicals in fish. Chemosphere 17, 943–962.

Gobas, F.A.P.C., Wilcockson, J.B., Russell, R.W., Haffner, G.D., 1999. Mechanism of biomagnification in fish under laboratory and field conditions. Environmental Science and Technology 33, 133–141.

Hamelink, J.L., Waybrant, R.C., Ball, C., 1971. A proposal: Exchange equilibria control the degree chlorinated hydrocarbons are biologically magnified in lentic environments. Transactions of the American Fish Society 100, 207–214.

Harrison, H.L., Loucks, O.L., Mitchell, J.W., *et al*., 1970. Systems studies of DDT transport. Science 170, 503–508.

Hunt, E.G., Bischoff, A.I., 1960. Inimical effects on wildlife of periodic DDD applications to Clear Lake. California Fish and Game 46, 91–106.

Kidd, K.A., Schindler, D.W., Muir, D.C.G., Lockhart, W.L., Hesslein, R.H., 1995. High-concentrations of toxaphene in fishes from a sub-arctic lake. Science 269, 240–242.

Mackay, D., 1981. Calculating fugacity. Environmental Science and Technology 16, 274–278.

Oliver, B.G., Niimi, A.J., 1988. Trophodynamic analysis of polychlorinated biphenyl congeners and other chlorinated hydrocarbons in the Lake Ontario ecosystem. Environmental Science and Technology 22, 388–397.

Rasmussen, J.B., Rowan, D.J., Lean, D.R.S., Carey, J.H., 1990. Food-chain structure in Ontario lakes determines PCB levels in lake trout (*Salvelinus namaycush*) and other pelagic fish. Canadian Journal of Fisheries and Aquatic Sciences 47, 2030–2038.

Schlummer, M.G., Moser, G.A., McLachlan, M.S., 1998. Digestive tract absorption of PCDD/Fs, PCBs and HCB in humans: Mass balances and mechanistic considerations. Toxicology and Applied Pharmacology 152, 128–137.

Swackhamer, D.L., Skoglund, R.S., 1993. Bioaccumulation of PCBs by algae: Kinetics versus equilibrium. Environmental Toxicology and Chemistry 12, 831–838.

Woodwell, G.M., Wurster Jr., C.F., Isaacson, P.A., 1967. DDT residues in an east coast estuary: A case of biological concentration of a persistent insecticide. Science 156, 821–824.

Biotopes☆

Panayiotis G Dimitrakopoulos and Andreas Y Troumbis, University of the Aegean, Mytilene, Greece

Glossary

Community structure The range of species along with their relative abundances in a biological community.

Complementary resource use Resource partitioning in space and time among cooccurring species at a given community.

Conservation planning The identification of priority areas for conserving natural values.

Diversity-ecosystem function relationship The relationship between plant or animal diversity and ecosystem processes such as primary productivity, nutrient cycling, and decomposition.

Ecosystem functioning Energy flow and nutrient cycling through abiotic and biotic compartments of an ecosystem.

Habitat The living area of a species (plant, animal or microorganism).

Introduction

Biotope is a synthetic Greek work combining "bios" (meaning life) and "topos" (meaning place). The German scientist Friedrich Dahl introduced the term biotope in 1908 as the habitat in which a particular group of animal and plant species live. This term was a complement of the term biocoenosis (attributed to Karl Möbius in 1877) meaning the group of animals and plants living together in a specific habitat. In this sense, biotope is considered to be the physical (abiotic) conditions in which a biocoenosis exists. In 1935, biotope and biocoenosis were incorporated into the term of ecosystem coined by Arthur Tansley.

The term biotope is being used interchangeably with habitat, the latter more in the Aglo-Saxon ecological literature and the former more in other European one. The term habitat has a number of different uses, but is generally considered to represent the physical conditions that surround a species, a species population and an assemblage of species or community. However, biotope has to be assigned to the community concept and habitat to the species concept. In addition, the term biotope is not limited to encompassing only physical conditions that surround a community of organisms but it also includes the relative biota. Therefore, biotope is a division of the landscape (a topographic unit) characterized by similar environmental (physical) conditions and a specific assemblage of plant and animal species, that is, a set of adjacent places in a given geographic region having more or less similar biotic and abiotic features. Thus, a species has a certain habitat, but the group of species that share an ecosystem with that species in a geographic region, share a biotope.

As a structural component of the ecosystem, the spatial and temporal definition of a biotope comes up against the same methodological difficulties as its higher level organizational capstone concept, i.e., ecosystem. The definable boundaries of a biotope are more dependent on the aim of the study and on the criteria posed by the researcher than geographical, and, thus, are somewhat arbitrary. Biotopes are often described based on dominant species or suites of conspicuous species present in areas sharing similar physical conditions. When geographically defined discontinuities occur such as small islands, mountain peaks, caves, thermal vents, hot springs, vernal pools, nutritionally imbalanced substrates (e.g., serpentine, limestone) and isolated vegetation fragments in the landscape, distinct biotopes can be easily identified. Differentially, biotope identification, as well as ecosystem identification, is far more problematic because the natural environment can not be easily divided into a series of discrete and discontinuous units as it represents different parts of a high variable continuum.

The earliest static perceptions of biotope tended to link it to the location, extent and substratum of a given biocoenosis. As a "container" system, the biotope was viewed as qualitatively uniform, where ecological factors are invariable in space. The cognitive evolution towards the concept of ecosystem allowed a functionalist and dynamic approach since it has introduced the processes of transformation, circulation and accumulation of energy and matter between and within the components of the system. Within this ecosystemic framework and its primacy in modern ecological thinking, the biotope is connected directly and/or indirectly to values, functions, goods and services of natural systems and biodiversity in the perspective of global change. Biodiversity is being eroded across all levels of biological organization due to the changing pattern of land use and climate change. These changes are expected to have major impacts upon processes that maintain ecosystem functioning (e.g., productivity, decomposition, nutrient cycling and resistance to weed invasion) by altering species richness and composition, as well as the evenness of plant communities.

☆*Change History*: October 2017. Panayiotis G Dimitrakopoulos and Andreas Y Troumbis made changes in sections "Biotope, niche and ecosystem functioning," "Community succession and biotope changes" and "Biotope and nature conservation," added new references in "Further Reading" section and a list of "Relevant Websites," and added Table 1.

This is an update of P.G. Dimitrakopoulos and A.Y. Troumbis, Biotopes, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 471–475.

Thus, the concept of biotope appears instrumental in the investigation of theoretical issues such as the ecological niche, the functional relationship between biodiversity and ecosystem processes such as production and the applied problems of nature conservation such as designing nature reserve systems or networks.

Biotope, Niche and Ecosystem Processes

Assuming that the boundaries of a biotope are described in an adequate manner in terms of scale, structure or composition, or a specific combination of environmental factors, the explanation of the presence of a small or large number of species (the alpha diversity) is an open question that is intimately connected with the perception of the biotope. Among the major theories explaining variations of species diversity across environmental gradients, the so-called dynamic or energy theory well matches the view of a biotope as an entity of available resources such as water, nutrients and energy. These resources determine the total biomass and the population size of the biota, and then short-term ecological processes determine the species number while long-term evolutionary processes determine structural properties such as niche widths and species packing.

Biotope Space, Species Richness and Ecosystem Functioning

When diversity (species richness, taxonomic composition and abundance distribution) is considered as an 'ecosystem property' driven by the variation in the biotope's variables (abiotic and biotic)—such as productivity or soil nutrient content—then a unimodal or hump-back relationship between them is considered as widely valid at plot scale. The relationship between ecosystem productivity and diversity, produced from field survey data, varies among taxa, between spatial scales and within as against between biotopes. Recent studies conducted in grassland ecosystems found that productivity explains a small part of the variation in species richness, and that large-scale processes (e.g., the size of the local species pool) seem to govern local plant species richness.

Effects of plant species diversity on plant productivity can be analyzed in comparative studies in which a set of natural communities is selected to represent different levels of diversity, being as similar as possible in other characteristics except productivity, or in experiments where plant communities of different diversities were constructed. These experiments treat diversity as the independent variable driving variation in ecosystem processes, considered as the dependent or response variables. Field experiments (e.g., cross-European BIODEPTH experiment, Jena experiment or biodiversity experiments at Cedar Creek, Minnesota), as well as of theoretical models, suggest that the processes maintaining ecosystem functioning could be adversely affected by the loss of diversity. The increasing probability in presence of one or a few dominant species (selection effect), and more complete exploitation of available resources (complementarity effect, including facilitative interactions between species) in species-rich communities than less diverse ones have been proposed as underlying mechanisms to explain positive diversity effects on ecosystem functioning. In the long term, complementarity effects may become increasingly important in relation to selection effects.

In 1957, George Hutchinson discriminated niches as properties of species and biotope as the physical space in which all, part, or none of a species' niche may occur: "… niche may be regarded as a set of points in an abstract *n*-dimensional N space. If the ordinary physical space B of a given biotope be considered, it will be apparent that any point $p(N)$ in N can correspond to a number of points $p_i(B)$ in B, at each one of which the conditions specified by $p(N)$ are realized in B" (Fig. 1). Hutchinson presents a simple lake system as an example of a biotope that is a "segment of the biosphere with convenient upper and lower boundaries, which is horizontally homogeneously diverse."

Hutchinson's niche definition includes the combination of an organism's environmental tolerances and habitat requirements in determining the conditions and resources needed for its survival and reproduction. For example, species performance is limited by soil moisture but different species demonstrate different tolerances at a different range of values of this variable. This range determines one dimension of the species' niche. Various conditions (e.g., pH, temperature, salinity) or resource states (e.g., food size, water, nutrient or light levels) could serve as other dimensions of a species' niche. The combination of range of values of n dimensions fixes a subset of space—an abstract *n*-dimensional hypervolume—inside which the organism can perform its biological activities. The niche can become easily perceptible as the diversity and range of values of the biotope's resources that are exploited by a species (called niche width).

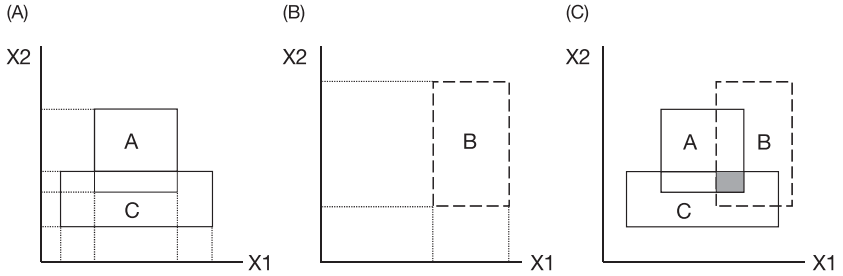


Fig. 1 The biotope and the niche. (A) A pair of variables (X1, X2) define two-dimensional areas that represent the niche of species A [N_A ($X1_A$, $X2_A$)] and of species C [N_C($X1_C$, $X2_C$)] respectively. (B) The ordinary physical space B(X1, X2) of a given biotope. (C) Part of species niches (N_A and N_C) represented within this biotope. Overlap of species niches in the biotope shows the amount of interspecific competition (*gray box*).

Consequently, niche overlap—the amount of competition—depends on the extent (or the degree) to which two or more species are able to utilize the same range of values of a resource (Fig. 1). According to niche theory, species can coexist in the same particular biotope as long as their niches do not fully overlap within this biotope. If species differ in their niches (i.e., more complete utilization of available resources) and the niches are considerably smaller than the entire biotope space (i.e., niche axes length) then communities richer in species may have higher productivity and other process intensities than species-poor mixtures or monocultures because of complementarity. These diverse communities would not become, as strongly dominated by individual species as expected under the selection effect. Fig. 2 illustrates mechanisms driving different biodiversity—ecosystem functioning relationships.

Does Reducing Biotope Space Weaken the Strength of the Diversity—Ecosystem Functioning Relationship?

In the example of soil volume, several species with different rooting architecture and depths may be combined, provided the biotope is sufficiently large and deep to fully accommodate these species niches. Therefore, complementarity and beneficial biodiversity effects should increase with biotope space. Experimental approaches, using constructed plant communities of various diversities growing on a gradient of increasing soil depths and volumes that offer increased rooting space to species and related to the niche dimension of nutrient acquisition by the roots of plants, showed that biodiversity effects due to species complementarity increased linearly with biotope space. Scaled up to agricultural systems, this means that benefits of intercropping may be greater in deep soils and that soil erosion may reduce intercropping benefits.

In the case of resource availability, biotope space can be defined in relation to a chemical, spatial or temporal aspect. For example, organic matter decomposition processes, derived by decomposer microorganisms and detritivores, regulate availability of different forms of nutrients (e.g., nitrogen, phosphorus) on a spatial and temporal scale, and, through enlarging biotope space for plant roots (e.g., utilization of deeper soil strata), increase the degree of complementarity among plant species. Along with resource availability, resource complexity has recently come into use in the definition of biotope space. Increasing resource complexity or heterogeneity in resource environments (i.e., large biotope space), can permit an increased number of species with complementary niches within communities, leading to positive diversity effects on community performance. In a smaller biotope

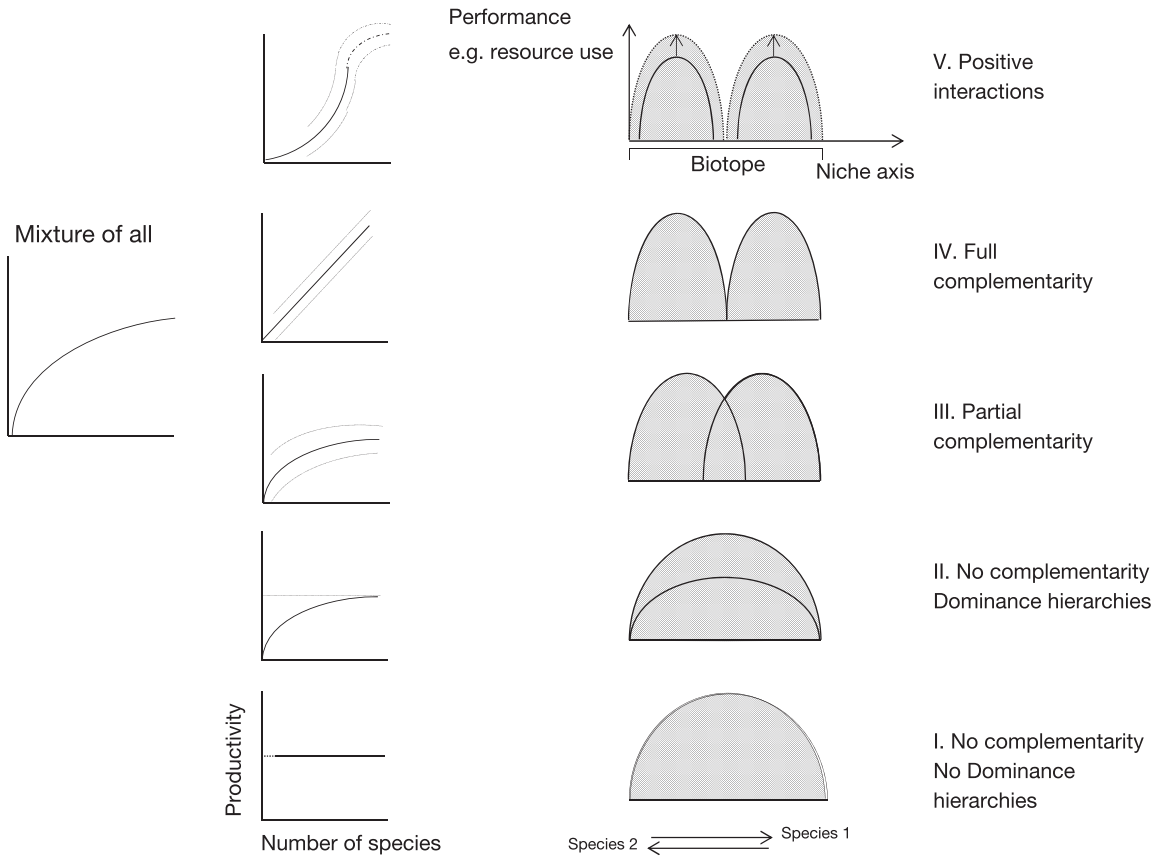


Fig. 2 Mechanisms explaining various responses of ecosystem processes in changing diversity. Two plant species performance curves along a niche axis represented in a particular biotope were used to depict potential niche separation and overlap. Species performances are hypothesized to be directly correlated with competitive ability. Each case (I–V) represents a different rule based on which species were assembled, shaping a different biodiversity—ecosystem functioning relationship. When a mixture of mechanisms operates together, the expected relationship is shown. From Kinzig, A.P., Pacala, S.W., Tilman, D. (Eds.), 2002. Functional consequences of biodiversity: Empirical progress and theoretical extensions. Princeton: Princeton University Press.

space (i.e., more homogeneous resource environments), competition will be more intense (i.e., species niches will overlap to a larger degree), reducing positive diversity effects or making them negative.

Comparatively, the strength of the positive relationship between species richness and productivity appeared to decrease in disturbed biotopes, as in grassland ecosystems where horse trampling and drought stresses were induced, indicating that reductions in biotope space through environmental stresses decrease the potential for positive effects of niche separation among species.

Community Succession and Biotope Changes

Biotopes are not constant in time; as they are intimately linked with a biocoenosis, they undergo gradual changes in a series of ecological factors generated by the biotic components of the community. In turn, these changes influence the nature of the biocoenosis modifying its species composition and therefore the structure of the entire ecosystem, and so on. In a given area these interactions determine a sequence of successive biocoenoses until equilibrium is reached. Various explanations of the causes of succession have been proposed, emphasizing key environmental (biotope) factors as determinants of the theoretical "end" stage of the succession. However, they converge in the hypothesis that when a biotope is physically modified at the limits of the capacities of the organisms composing the community, then compositional stability is reached.

There is a long tradition in ecology of studying the composition of plant communities. At the beginning of the 20th century, two contrasting views concerning plant community assembly were introduced by Fredrick Clements and Henry Gleason. Clements considered communities as well-organized associations of coevolved species, whereas Gleason viewed communities as more or less random aggregations of species in time and space. When species distributions are plotted along a biotope gradient, communities appear to have sharp boundaries based on Clements' model, while the Gleason model predicts that species are arranged independently of one another thus no distinct communities exist (Fig. 3). Compared with real communities, the two models can be considered as the extremes of a continuum. The Gleason vision seems to fit better communities of early secondary succession. These communities often show geometric dominance–diversity curves, indicating that the dominant species takes a certain proportion of the total resources offered by the biotope, the second most dominant species the same proportion of the remainder and so on. In contrast, dominance hierarchies are less strongly developed in late-successional communities, indicating less niche overlap and more niche separation, presumably due to there being more time for coevolutionary processes in shaping species interactions. The Clements model seems to better fit these late-successional communities. The species diversity and composition of a community is, in both models, treated as a function of succession.

Niche concept has been used in explaining species distribution patterns as well as in the understanding of community features (e.g., species richness) and of dynamic process patterns such as successional changes. For example, a community could support a greater number of species relative to another because biotope provides more available resources (longer length of the niche axes) and thus, more species niches can be accommodated. For a given range of resources, species richness can be increased because either mean species niche width within the community is smaller (i.e., species are more specialized in resource exploitation) or average overlap between adjacent species niches within the community is larger. In the end, when the available range of resources is not fully exploited, communities will contain a more limited richness of species than they potentially could.

Biotope and Nature Conservation

Given the alarming rates of biodiversity loss, conservation policies are being developed at various scales and within a range of institutional and legal frameworks. Conservation aims at maintaining biodiversity that contributes to the provision of key ecosystem services, i.e., the useful output for humans resulting from ecosystem functions. According to the Millennium Ecosystem Assessment classification, they are divided into four categories: (a) provisioning services—the variety of goods obtained from ecosystems, (b) regulating services—the benefits obtained from the regulation of ecosystem processes, (c) cultural services—

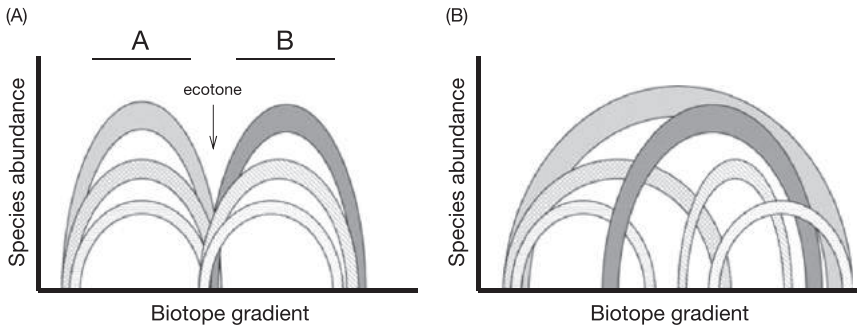


Fig. 3 Species distributions along a biotope gradient. (A) Communities (A, B) appear to have clear boundaries and are separated by intermediate zones (ecotones), (B) Species are arranged independently of one another, resulting in no distinct communities.

Table 1 Levels of hierarchical classification of the European marine habitats

No.	*Level*	*Description*	*Example*
1	Environment	Used to distinguish main types e.g., marine, terrestrial or freshwater	Marine
2	Broad habitats	Extremely broad habitat categorization	Littoral sediment
3	Main habitats	Very broad habitat divisions that mirror differentiation in their biological aspect	Littoral sand and muddy sand
4	Biotope complexes	Groups of biotopes that feature similar physical and biological attributes	Barren or Amphipod-dominated mobile sand shores
5	Biotopes	These are described based on dominant species or assemblages of conspicuous species present	Amphipods and *Scolelepis* spp.
6	Subbiotopes	These are characterized using less conspicuous species or on the basis of subtle physical habitat differences	*Eurydice* subbiotope

recreation and other nonmaterial benefits and (d) supporting services—the essential services for the production of all other ecosystem services. Ecosystem values measure the significance of ecosystem services to improve human welfare; they are dictated by cultural attitudes and society determines their relative benefits.

Evidently, conservation of biodiversity cannot be separated from maintenance of biotopes. From that perspective, a series of terms that correspond to conservation objects and refer to the biotope context have been developed and are mostly used in conservation planning strategies and language. Terms such as site, conservation area, ecotope, ecoregion and so on, often serve the conceptual needs of approaches to conservation strategies.

Selection and prioritization of areas i.e., biotopes is the most important stage for conservation planning, on a regional (or higher) scale. Conservation strategies can be species or ecosystem oriented. Species oriented strategies lay the emphasis on the analysis of population size and geographic distribution of individual species or the protection of biotopes important to particular taxonomic groups such as biotopes that are characterized by high levels of species richness or high concentrations of endemic, restricted-range or red-listed species. However, long-term maintenance of biodiversity within conservation areas can be achieved conserving both the biodiversity patterns and the natural processes that sustain those patterns. Ecosystem oriented strategies aim at protecting the range of ecological conditions found in a region of concern, using criteria such as species richness, endemism and uniqueness as well as elements of the abiotic environment, ecosystem processes, disturbance regimes and succession stages that assist in the definition of the ecosystems. Ecosystem approaches emphasize on protecting most species within conservation areas that are representative of natural or seminatural biotopes of a geographic region.

Conservation planning should define a hierarchy in conservation objectives that generates nested protection levels, necessary for the maintenance of the stability of the ecological landscape as a functional entity. The biotope or habitat-type approach that focuses on selection of conservation areas using criteria such as representativeness of vegetation types and high plant and animal species richness, is supposed to optimize ecological functions. The landscape-approach, where the landscape is understood to be a product of the combination of biotope attributes and human activities in a historical time frame, takes into consideration not only ecological but also socio–economic and aesthetic functions. Ideally, the integration of these approaches should incorporate dynamic aspects such as fluxes of organisms, energy and nutrients. Such a holistic conservation planning strategy will inevitably focus on the maintenance of the functional relation between sites through a spatial system of functionally interconnected elements, i.e., an ecological network. Certainly more difficult to design and achieve, it will probably require a creative integration of different policies (agricultural and agro-environmental policies, housing and tourism etc.), but will permit the collection of key areas for biodiversity conservation to become an ecological network.

Biotope Classification Schemes for Nature Conservation Purposes

The selection of sites to be included in an ecological network for conservation necessitates, apart from specific techniques, the clarification and classification of the correspondence between the conservation-driven terms (e.g., site, area, ecotope) and the meaning of fundamental concepts such as biotope, habitat, landscape, community etc. For example, sites are "geographically defined areas, whose extent is clearly delineated," within which conservation targets can persist. Based on this definition, sites are synonymous to landscape units or include a combination of both broad habitat divisions and landscape units. In the European Union, the CORINE-biotopes scheme uses the term site in order to avoid various etymological problems presented with the use of the term biotope in the various languages of the EU. The equivalence between the scientific term biotope and the practically used term site is provided by the definition of site as "…an area of land or a body of water which forms an ecological unit of Community significance for nature conservation, regardless of whether this area is formally protected by legislation." On a continental scale, ecoregions, considered as large-scale biodiversity units, are defined as "a relatively large area of land or water containing a characteristic set of natural communities that share a large majority of their species, ecological dynamics and environmental conditions." Recently, ecoregions representing the distribution of a wide range of flora and fauna species are

classified within major habitat types of the world (biogeographic biomes and realms). Comparisons between the biodiversity attributes of ecoregions (such as species richness, endemic species, unusual higher taxa, unusual ecological or evolutionary phenomena and the global rarity of habitats) in order to estimate their distinctiveness, have identified 238 important areas for global conservation goals—known as the Global 200. Biodiversity hotspots analysis, identifying the 36 most threatened terrestrial areas of the Earth richest in endemics, coincide, to a large extent, with the Global 200.

As to the classification problem of biotopes for conservation, various schemes have been proposed. In an analogy to species classification, which is based upon well-established rules of systematics and it structured in a hierarchy from genera to phyla, habitat classification schemes are designed for various ecosystem types. Habitats include assemblages of species coming from different taxonomic hierarchy that consistently occur together at a given spatial scale. Habitat classification can be arranged in a hierarchy from biotopes to broad habitats. For example, the marine habitat classification system of Britain and Ireland, very close to the corresponding European one, includes six levels of hierarchy (**Table 1**). In this hierarchy, broad habitat divisions (levels 2 and 3) are described based mainly on differences in their physical attributes reflecting differences in their relevant biota. At the lower end of the classification, biological features (e.g., characteristic species, species abundances, community structure) are combined with physical factors to define biotope types (levels 5 and 6), that can be grouped into higher levels with similar characteristics (level 4). Landscape units are comprised of sets of habitats consistently occurring together, coming from different levels of habitat classification hierarchy.

Criteria for the Identification of Important Biotopes for Biodiversity Conservation

Specific criteria or rules are used, tested or implemented at an international level. For instance, in the Ramsar Convention (or "Convention of Wetlands"), the criteria for the identification of internationally importance wetlands (Ramsar Sites) are organized into two groups based upon: (a) representativeness/uniqueness of wetland types included and (b) biodiversity supported (species and communities with specifications for water birds and fishes for which wetlands are important biotopes). A wetland classification system has been developed for the identification of the wetland habitat type within each Ramsar Site. 42 wetland types have been included and categorized into three broad groups: (a) coastal/marine, (b) inland, and (c) human-made. The European Union's CORINE-biotopes scheme selects priority sites based on species conservation attributes (e.g., degree of threat, richness for a taxonomic group of species) and on habitat characteristics (e.g., habitat sensitivity, richness of habitat type). A site should satisfy at least one from the following criteria in order to be evaluated as of community interest: (a) the site is one of 100 or fewer (or the 100 most important) in the Community or one of five or fewer (or the five most important) sites in a region for a threatened species or for a sensitive habitat and (b) the site supports at least 1% of the population of a threatened species at the Community level. This scheme proved an exceptionally useful tool in the creation of the Natura 2000 network of Special Areas of Conservation, known internationally as the "Habitat Directive," which is based on the idea of preservation of listed (i.e., endangered or sensitive) species and habitat types qualified as being of community interest. Sites are selected for inclusion into the network mostly on the basis of representativeness, quality of the habitat, surface area occupied, size and density of listed species and the degree of their isolation.

Such classification schemes are considered to support typical biotope/habitat-type strategies for nature conservation, since the sites are, in theory, selected on the criteria of the representativeness of phytosociological units and of the presence of threatened species. However, one can formulate the rather robust prediction that while conservation efforts focus on key areas for rare and endangered species and habitats, the remaining areas, large in scale, are exposed to an ever increasing intensification or to changes in management practices that devalue their long-term ecological and economic value which, consequently, can affect the environmental conditions within the protected biotopes.

See also: Ecosystems: Ecosystems

Further Reading

Begon, M., Townsend, C.R., Harper, J.L., 2006. Ecology: From individuals to ecosystems, 4th edn. Oxford: Blackwell Publishing.
Connor, D.W., Allen, J.H., Golding, N., Howell, K.L., Lieberknecht, L.M., Northen, K.O., Reker, J.B. (2004). *The marine habitat classification for Britain and Ireland*. Version 04.05. Peterborough: JNCC (Internet version).
Dimitrakopoulos, P.G., Schmid, B., 2004. Biodiversity effects increase linearly with biotope space. Ecology Letters 7, 574–583.
European Commission, , 2003. Interpretation manual of European Union habitats—EUR 25. . Brussels: DG—Environment.
Harper, J.L., 1977. Population biology of plants. London: Academic Press.
Hutchinson, G.E., 1978. An introduction to population ecology. New Haven: Yale University Press.
Jousset, A., Schmid, B., Scheu, S., Eisenhauer, N., 2011. Genotypic richness and dissimilarity opposingly affect ecosystem functioning. Ecology Letters 14, 537–545.
Kinzig, A.P., Pacala, S.W., Tilman, D. (Eds.), 2002. Functional consequences of biodiversity: Empirical progress and theoretical extensions. Princeton: Princeton University Press.
Naeem, S., Bunker, D.E., Hector, A., Loreau, M., Perrings, C. (Eds.), 2009. Biodiversity, ecosystem functioning, and human wellbeing: An ecological and economic perspective. Oxford: Oxford University Press.
Odum, E.P., 1953. Fundamentals of ecology. Philadelphia: Saunders.
Primack, R.B., 2012. A primer of conservation biology, 5th edn. Sunderland, MA: Sinauer Associates.
Tilman, D., Isbell, F., Cowles, J.M., 2014. Biodiversity and ecosystem functioning. Annual Review of Ecology Evolution and Systematics 45, 471–493.

Relevant Websites

http://www.cedarcreek.umn.edu/—Cedar Creek Ecosystem Science Reserve.
http://jncc.defra.gov.uk/Default.aspx—Joint Nature Conservation Committee (JNCC).
http://ec.europa.eu/environment/nature/natura2000/index_en.htm—Natura 2000: A European-wide network of protected areas.
http://www.the-jena-experiment.de/—The Jena experiment.

Connectivity and Ecological Networks

Robert HG Jongman, Wageningen University and Research, Wageningen, The Netherlands

Ecological Connectivity

Introduction

Isolation and connectivity are natural characteristics of ecosystems. Mountain tops and islands are typically isolated ecosystems while in an undisturbed situation natural forests and tundra's are extensive and connected systems. Connectivity is a key concept of landscape ecology as it relates to flows and movements of organisms trough landscapes (Burel and Baudry, 2005). Movement is a key process for survival of plants and animals. Faunal species move daily in search for food and to avoid predators.

Moving from one habitat to another can be done by passive transport (floating by wind, water or moving on or in other species) as well as actively by swimming, flying or walking. Passive transport is important for plant species. The easterly wind direction at the southern hemisphere and not the geographical position determines the flora relationships between the islands around Antarctica.

Migration is a prerequisite for many species from northern parts of the globe to survive the winter period. It is risky but essential for wetland birds (Boere and Stroud, 2006). Amphibians migrate from and to spawning habitats and to wintering habitats. For them distances are short and barriers hard to take. Migrating fish such as the Atlantic salmon (*Salmo salar*) migrate between spawning rivers and the northern Atlantic Ocean. The tropical fish species Pacu (*Piaractus mesopotamicus*) migrates from small rivers in the Brazilian savanna (cerrado) to the wetlands of the Pantanal and the Amazon. It feeds on plants, fruits and seeds and in this way it plays a role in the plant species migration.

Migration as a species or population's periodic movement is typically a long distance movement from one habitat area to another to avoid unfavorable seasons or conditions (Stenseth and Lidicker, 1992). The European stork (*Ciconia ciconia*), for instance, has adapted to the cultural landscapes of Europe, because they provide food and are accessible. Its wintering habitat is in Africa, 10,000 km away. Changes in area and habitat quality in both non-breeding and breeding sites have effect on species survival and the size of the populations. The biggest dangers that can affect populations of migrating species are the barriers and events during migration.

Corridors for dispersal and migration are species dependent and can vary from single wooded banks for voles, bats and badgers to small-scale landscapes for forest birds and from river shores for otters to whole rivers and coastlines for migratory fish species. Corridors may be continuous as linear connections or interrupted as stepping stones.

Ecological connectivity is, however, not a self-evident issue. The world has been shaped by humans nowadays and all species have to adapt or get extinct. Protecting the migration and dispersal process is one of the most difficult conservation challenges that we are facing as there is a growing conflict between nature and other land use. Habitat protection and international cooperation are essential to achieve this goal. This requires insight in the processes of species exchange through landscapes and designing and adapting the way we deal with land use with regard to the needs of other species.

Islands: Isolation and Metapopulations

The occurrence of wildlife species within a habitat is determined by its (1) habitat conditions, such as the availability of water, nutrients, energy, (2) habitat size, (3) connection to other habitat sites, and (4) by human disturbance. A habitat in a network depends on the landscape fluxes, such as air movements, water flows and species migration. Islands are in this respect special habitats. They are isolated from the mainland, are relatively small and mostly the variability in habitat conditions is restricted. They are on the one hand protected from outside influences, but on the other also restricted in species exchange. This isolation makes them also special as here new species or subspecies can develop such as on the Galapagos Islands, Madeira and Madagascar. Islands therefore have always attracted scientists to explore their wealth and specialized species, such as Linnaeus and Darwin.

In the second half of the 20th century Island biogeography was elaborated by several, mainly American biogeographers, among which May, Diamond, Ehrlich, McArthur and Wilson. They studied and discussed the role of islands, their isolation and habitat characteristics. In 1967 MacArthur and Wilson published the theory on island biogeography, bringing isolation, distance and habitat characteristics into a coherent mathematical model. The primary question to be answered was the explanation of the dissimilar number of species on different islands. They used as variables the relation between area and diversity, with isolation, extinction and colonization as key factors for establishing populations. Biodiversity of an island is by them expressed by the formula:

$$S = I + s - E$$

In which S is the number of species on the island, I is immigration, E is extinction and s is speciation (evolution). The number of species is determined by immigration of species from a continent or other islands and by extinction of species (**Fig. 1**). The variable s depicts evolution of new species, but this is a long term (evolutionary) factor.

 doi:10.1016/B978-0-12-409548-9.10619-0

Fig. 1 Human land use, urbanization in Shanghai, China, and large scale agriculture in Alentejo, Portugal.

In a set of islands with a more or less uniform climate and tiopography the number of individuals of a taxon on an island is according to MacArthur and Wilson (1967) linearly related with the area of the island:

$$J = \rho A$$

In which J is the number of individuals, A is the areas of the island and ρ the density of the individual organisms.

The relationship between the number of species and the area of an island can then be expressed as:

$$S = CA^z$$

In which S is the number of species, A the area, C is a constant, that varies among taxa and z is also a constant between 0.20 and 0.35 and consistent with lognormally distributed frequency curves of species.

Levins presented the mathematical basis of metapopulation theory in the 1979 in a lecture at the congress of the American Mathematical Society. This paper is not easily available. It describes the principle of a "population of populations," as a model for the basic properties of fragmented populations. Later Hanski and Gilpin (1991) defined metapopulations as ensembles of interacting individuals of which each ensemble has a finite lifetime. The metapopulation model focuses on the processes of population turnover by extinction and establishment of new populations, defining for each species the conditions under which these two processes are in balance. The assumption is that a population goes extinct when the area and quality of the required habitat goes below a critical level. *Re*-establishment of a population can take place through colonization by individuals from another population. The different populations form subpopulations within a metapopulation. Hanski (2001) formulated this in his spatially realistic metapopulation theory in which he provided a framework for integrating island theory and metapopulation theory.

The metapopulation concept laid a basis for understanding the dynamics of species populations not only in naturally fragmented situations, such as on islands or mountain tops, but also in fragmented landscapes as caused by human land use.

Changing Landscape: Homogenization, Fragmentation and Climate Change

To connect these ecological models with real life requires knowledge of landscape change processes. The dominant processes worldwide can be described as homogenization and fragmentation (Jongman, 2002). Homogenization means that landscapes are becoming more or less the same everywhere, without much variation. Some parts of the land are accessible for wildlife, but most is not. Fragmentation means that on the other hand that the land is dissected by barriers and increasingly difficult to go through. In the last 50 years new landscape types became dominant such as (sub) urban landscapes, motorway landscapes, recreation landscapes and industrial landscapes (**Fig. 1**). They replace forest landscapes, small scale agricultural landscapes with hedgerows, natural rivers and wetlands. Homogenization and fragmentation of the landscape contributes significantly to the decline and loss of wildlife populations and to the increasing endangerment of species.

In the entire world landscape fragmentation is increasing especially through the development of transport infrastructure, by building dams in rivers and by making agricultural land inaccessible by fencing and intensive land use. Landscape fragmentation caused by transportation infrastructure and built-up areas has four important ecological effects: direct habitat loss, traffic mortality, barrier for migration and dispersal and cutting populations into subpopulations (**Fig. 2**).

For Europe this process has been analyzed and made visible by the Swiss Environmental Agency together with the European Environmental Agency (FOEN-EEA, 2011). In general, landscapes are getting more connected for human use, but are at the same time ecologically disintegrating and fragmenting. (Sub-) urbanization is the driving force for ecological fragmentation. In Europe there are additional trends. National borders are getting less important and international issues are becoming more important, international transport and exchange is increasing (Jongman, 2002). The motorways and river regulation that are needed for it are the main ecological barriers in the landscape (**Fig. 3**).

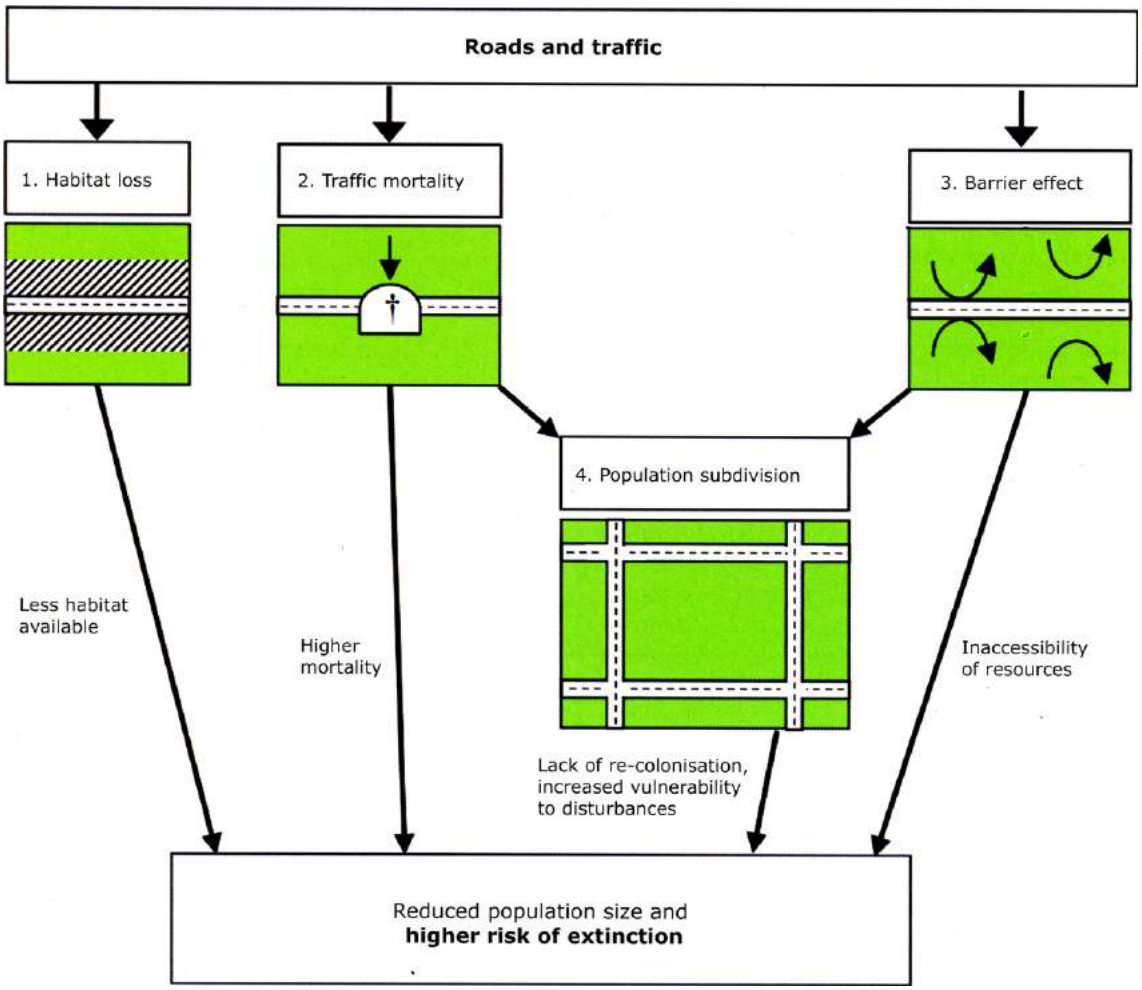


Fig. 2 Effects of transportation infrastructure on wildlife populations (Jaeger *et al.*, 2005).

Fig. 3 Human infrastructure as barriers: a motorway (A50, The Netherlands) and a river dam (Fukutomi dam, Honshu Island, Japan).

Climate change is complicating this development. Biodiversity depends on the protection and management of designated sites in an increasingly fragmented landscape. But habitats and species also have to adapt to a changing environment as a result of climate change. Over time, the conditions at habitats may change so much that species are moving to other sites. However, this is getting increasingly difficult because of homogenization and fragmentation of the landscape. Small, natural landscape elements within the agricultural landscape such as tree lines, hedgerows, road/waterway verges, ponds, and small woods can provide

suitable areas for dispersal and migration of species. However, often these natural landscape "routes" are too scattered and of poor quality from a biodiversity perspective. For improving connectivity between protected areas landscape and river management in a broader context is important.

Ecological Networks to Foster Ecological Connectivity

Introduction

From the beginning of the 19th century on, individuals as well as national and international conservation movements have elaborated strategies for species preservation and nature conservation through establishment of nature reserves and national parks. In the past these were sufficient to mitigate the human impacts on nature caused by animal use, urbanization and agricultural development. The restructuring of the landscape for urbanization and industrialized agriculture and the building of transport networks have isolated these protected areas, deteriorated ecosystems and made species extinct. Species survival is dependent on habitat quality, such as food availability, habitat size and for most species the ability to move through the landscape. This movement is needed for daily activities such as foraging, dispersal and reproduction migration and for avoiding unfavorable environments or seasons (Opdam, 1991).

Ecological networks can be defined as systems of sites of high biodiversity value and their interconnections that make a fragmented natural system coherent to support more biological diversity than in non-connected form. The definition of ecological network by Bennett (2004) is in line with this paradigm: "A coherent system of natural and/or semi-natural landscape elements that is configured and managed with the objective of maintaining or restoring ecological functions as a means to conserve biodiversity while also providing appropriate opportunities for the sustainable use of natural resources."

An ecological network is composed of core areas, usually protected by buffer zones and connected through ecological corridors (Jongman and Pungetti, 2004). The nature reserves and national parks, established in the past by traditional nature conservation policies form the majority of the core areas. The insight gained from island theory, metapopulation theory and landscape dynamics teaches us that these traditional conservation sites should be linked with each other through ecological corridors that have often multiple functions. Ecological corridors are the new elements in nature conservation policy and planning. However, they are also highly disputed because they are landscape elements where many functions coincide and that might lead to conflicts and discussion on, for instance, their use by migrating geese, wolves and invasive species.

At the heart of the ecological network approach is the awareness that without the full engagement of various sectors of economy and society in the management of ecosystems, there can be no effective biodiversity conservation. This "Ecosystem Approach" incorporates interaction of organisms, ecosystems, and the human component and it is the framework for holistic decision-making and action (Bennet, 2004). It is changing paradigms in protected area management moving from strictly nature oriented to a nature and society oriented approach. The goal of the network strategy is to conserve and restore ecological corridors and "stepping stones" (habitat islands), which function as habitat structures between core nature areas and facilitate ecological connectivity in the landscape. This cannot be done without finding compromises with other land use. It is however, the only strategy to cope with ecosystem fragmentation in the human dominated landscape and also with the changes to be expected through climate change. Changing climate will mean that land use will change, but also habitat quality and the suitability of the existing habitat islands; consequently there will also be a change in the role of connectivity. This means species will search for new, better adapted habitats and we have to facilitate this (European Commission, 2013a).

Ecological Corridors

Ecological links between habitats have always existed also in natural landscapes. Most obvious are migration routes for birds, ant-routes, badger routes and river corridors for fish migration like for the eel and the salmon. The term "corridor" has appeared very early in the literature to refer to long range dispersal. The current use of the term ecological corridor stems from the 1960s when it was used in spatial population dynamics for enhancing the chance of survival of smaller populations through exchange between patches. Metapopulation dynamics presupposes that some degree of connectivity exists between subpopulations within the fragmented landscape. Ecological corridors are based on either commuting, migration or dispersal movement functions. It can be defined as: A functional linkage between resource habitat of a species or a group of species, consisting of landscape structures that are different from the surrounding landscape resulting in a favorable effect on the genetic and species exchange (individuals, seeds, genes) as well as on species migration (Foppen *et al.*, 2000).

Ecological corridors therefore link isolated nature areas into one large "living" network (**Fig. 4**). An ecological corridor allows dispersal of plants and animals, and is often smaller and more linear than the natural areas it connects. These connections must, of course, be suitable for the species concerned, that is, they must contain the right habitat and be located in the right position in the landscape.

Due to differences in needs migration and dispersal routes can be manifold, from single wooded banks to small-scale landscapes and from river shores to whole rivers and coastlines. For fish it means that rivers do not only have a good water quality, but are also not blocked by dams. Most of the ecological corridors are primarily a mitigation of human disturbance regimes. Ecological corridors are therefore various landscape structures, varying in size and shape from wide to narrow and from

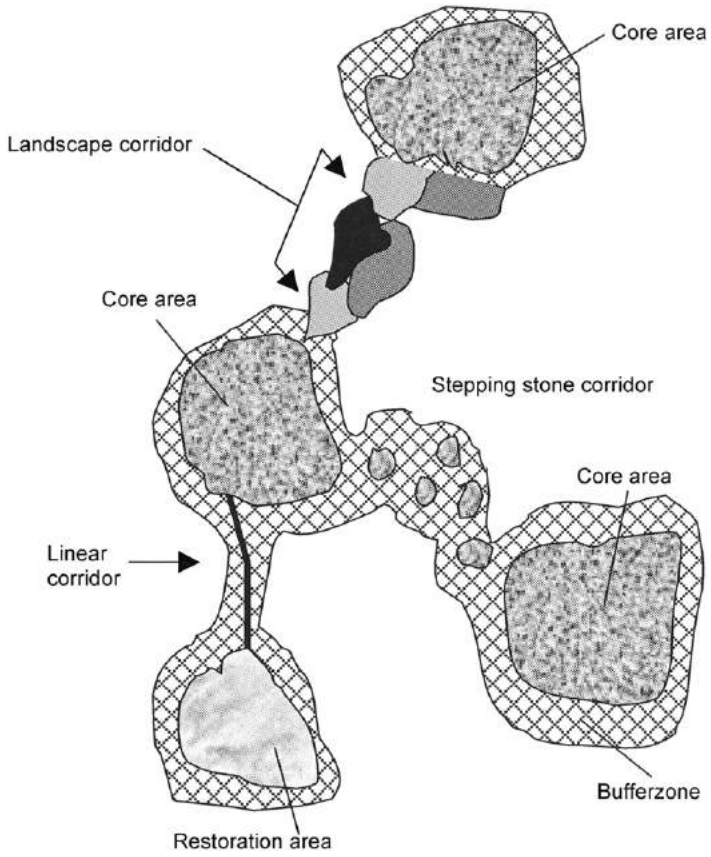


Fig. 4 Schematic example of an ecological network (Bouwma *et al.* 2002).

meandering to straight, providing ways to permeate the landscape, maintaining or re-establishing natural connectivity. Their density and spatial arrangement change according to the human land use. Nature needs different types of ecological corridors that have a complementary role to play in an interconnected habitat island system.

For mammals and amphibians it means that routes are available and that man-made barriers can be crossed. The need for these species to disperse can be between over distances from several meters to hundreds of kilometers. For small mammals ecological corridors can be hedgerows, brooks and all kind of other natural features that offer shelter. Migration is important for grazing animals like red deer (*Cervus elaphus*) and roe deer (*Capreolus capreolus*), but also for predators like the golden eagle (*Aquila chrysaetos*), the lynx (*Lynx lynx*), and the wolf (*Canis lupus*).

The nature of ecological corridors and their efficiency in interconnecting remnants and in permeating the landscape depend on their origin and the land use mosaic within which they are embedded and of which they consist. Forman (1995) makes a distinction in four main types of corridors:

1. line corridors, which are narrow strips of edge habitat, such as paths, hedgerows and roadsides,
2. strip corridors, with a width sufficient for the ready movement of species characteristic of path interiors (e.g., a wide power line corridor permitting movement of open country species through a forest),
3. stream corridors, which may function as one of the previous two, but which additionally control stream bank erosion, siltation and stream nutrient levels, and
4. network corridors, which are formed by the intersected corridors, usually resulting in the presence of loops, as well as subdividing the landscape matrix into many patches.

The Policy Context of Ecological Networks

Landscape changes are the reason that managing ecological connectivity is important and that nature conservation should take care of connectivity between reserves. Natural habitats in many parts of the world are now isolated islands. The smaller and more isolated these "habitat islands" are as a consequence of intensifying land use and road networks, the more likely it is that species decline. In many places in the world movements have been built to mitigate this development. In the 1970s already in several east European countries plans for nature networks have been developed, in the Netherlands in the 1980s the national Nature plan was designed (Jongman *et al.*, 2004), In Canada and the United States the Wildlands project and the Yellowstone-Yukon Conservation

initiative (https://y2y.net/) have been started also in the early 1990s. In 2010 connectivity has also been expressed in the 11th Aichi target of the Convention of Biological Diversity, to be reached by 2020.

Aichi target 11: By 2020, at least 17% of terrestrial and inland water, and 10% of coastal and marine areas, especially areas of particular importance for biodiversity and ecosystem services, are conserved through effectively and equitably managed, ecologically representative and well connected systems of protected areas and other effective area-based conservation measures, and integrated into the wider landscapes and seascapes (https://www.cbd.int/sp/targets/).

This target means that ecological corridors should link natural areas and especially protected areas into a coherent system: an ecological network. Accordingly, already years before the COP in 2010 in Nagoya various conservationists and conservation authorities have been changing their strategy from only of conservation of the existing, more and more isolated natural "islands" to the conservation and restoration of interconnected natural areas (Jongman and Pungetti, 2004).

Already in the end of the 19th century Parkways have been designed in both Europe and the United States but then mainly for other (recreational, hydrological) purposes. The most famous is the still existing park system of Frederick Law Olmstead in Boston, the Emerald Necklace as a connected systems of parks and greenways. This greenway approach (https://www.greenway.org/), is integrating local interests in hiking and biking with biodiversity conservation and building on the tradition of Parkway Planning (Jongman and Pungetti, 2004). Greenway systems have been created in many states such as in Florida (Florida Greenway Commission, 1994).

National authorities as well as non-governmental organizations around the world establish now ecological networks at different scales. If one looks at Europe, at present the Bern Convention and several EU directives support each other to realize the Pan European Ecological Network, the Natura 2000 network as spatial result of the Habitats- and Birds Directives and the Emerald Network (Bonnin *et al.*, 2007, Jongman *et al.* 2011).

Ecological networks and other habitat networks require planning also across borders. If decisions are not coordinated across borders then misunderstanding can lead to ineffective planning and waste of government budget. Bouwma *et al.* (in: Jongman and Pungetti, 2004) reviewed conventions and EU-legislation in order to assess, which European endangered vertebrates are considered to benefit from corridors with a European dimension. In total 420 vertebrate species were reviewed. It is estimated that of these species 104 species could benefit from European corridors. Of these 69 are bird species, 23 are mammals (mostly large herbivores and carnivores) and 12 are fish species. The highest number of potential species is in the Mediterranean and Continental biogeographic region. Realizing this means international cooperation in conservation policies.

The "Waterbirds around the world" conference in 2004 reviewed waterbird migration at many levels of detail, from the long distance migration routes to the relative short distance movements. In almost all cases the word "flyway" has been used to indicate the geographical region of migration. A general definition of a flyway is: "A flyway is the entire range of a migratory bird species or groups of related species or distinct populations of a single species, through which it moves on an annual basis from the breeding grounds to non-breeding areas, including intermediate resting and feeding places as well as the area within which the birds migrate" (Boere and Stroud, 2006). Flyways can be single species flyways or multi-species flyways. From knowledge of the various single migration systems it is possible to group the migration routes especially used by waterbirds into multi species flyways. This allows grouping the global migrations of waders into eight flyways: the East Atlantic Flyway, the Mediterranean/Black Sea Flyway, the West Asia/Africa flyway, the Central Asia/Indian sub-continent Flyway, the East Asia/Australasia Flyway, and three flyways in the Americas and the Neotropics.

Fish can migrate through oceans as well as rivers. Some species such as eel (*Anguilla anguilla*), Atlantic salmon (*Salmo salar)* and sturgeon (*Acipenser sturio*) combine both. Fish require river corridors like large rivers up to its headwaters dams and good water quality (Lafaille *et al.*, 2005). Fish species in larger wetlands or river complexes also migrate upstream for spawning as is the case in the Pantanal and the Mekong. Because of the great variety of feeding and reproductive niches for fish, these river wetlands harbor a high species diversity and abundance. Fish are an important resource, both ecologically, economically and socially. However, many of these river systems have been made inaccessible in the past due to damming and river regulation. Fish therefore often require adaptation by building fish ladders and removal of dams.

Green Infrastructure

The concept of ecological networks is developing further and is being accepted more widely by urban planners and river managers as principles for restoring nature and including it in their management plans. The name is changing towards "Green Infrastructure (GI)." Edward T. McMahon (The US Conservation Fund) says that Green infrastructure is "… an interconnected network of protected land and water that supports native species, maintains natural ecological processes, sustains air and water resources, and contributes to the health and quality of life for people." The European Commission (http://ec.europa.eu/environment/nature/ecosystems/) writes on its website that it: … is addressing the spatial structure of natural and semi-natural areas but also other environmental features which enable citizens to benefit from its multiple services.

Green infrastructure or blue-green infrastructure is a network providing the "ingredients" for solving rural, urban and climatic challenges by building with nature. Comparable approaches are being developed in the USA and in the EU. Habitat restoration and green infrastructure are key topics at present for the US Environmental Protection Agency (EPA, https://www.epa.gov/green-

infrastructure). This policy can be used to translate landscape ecological principles into practice concerning mitigating ecological homogenization and fragmentation.

Regional Implementation and Technical Solutions

The aim of establishing ecological networks is improved protection of nature and biodiversity. Its development is stimulated by science and nature management practice. The initiative taken in 1994 by the Council of Europe to formulate the Pan European Biological and Landscape Diversity Strategy (PEBLDS) and its action point on the Pan European Ecological Network (PEEN) was one of the drivers of a real European and global approach and recognition of corridors as inevitable part of nature conservation in a fragmenting world. The Pan-European Ecological Network (PEEN, Jongman *et al.*, 2011) aims to ensure that:

- A full range of good quality ecosystems, habitats, species and landscapes of European importance are conserved;
- Habitats are large enough to guarantee a favorable conservation status for key species;
- There are sufficient opportunities for dispersal and migration of species;
- Damaged parts of key environmental systems are being restored;
- The key environmental systems are buffered from threats.

In the meantime many countries have developed their own regional or national ecological networks (Jongman *et al.*, 2004) and have started to implement them in their own way (**Fig. 5**). In the meeting of the Standing Committee of the Bern Convention in December 2015 it was reported that at present ecological networks or conservation corridors do exist in 35 countries that signed the Bern Convention. Their status varies from pilots to implementation. The EU Green Infrastructure initiative is an important stimulus in this now.

Ecological networks are a way to make the landscape aesthetic again and to give a landscape back its regional characteristics. It should not only fulfill a role as linking element in the landscape but it should be possible to be used for education and recreation and fulfill ecosystem services such as flood surge and water storage.

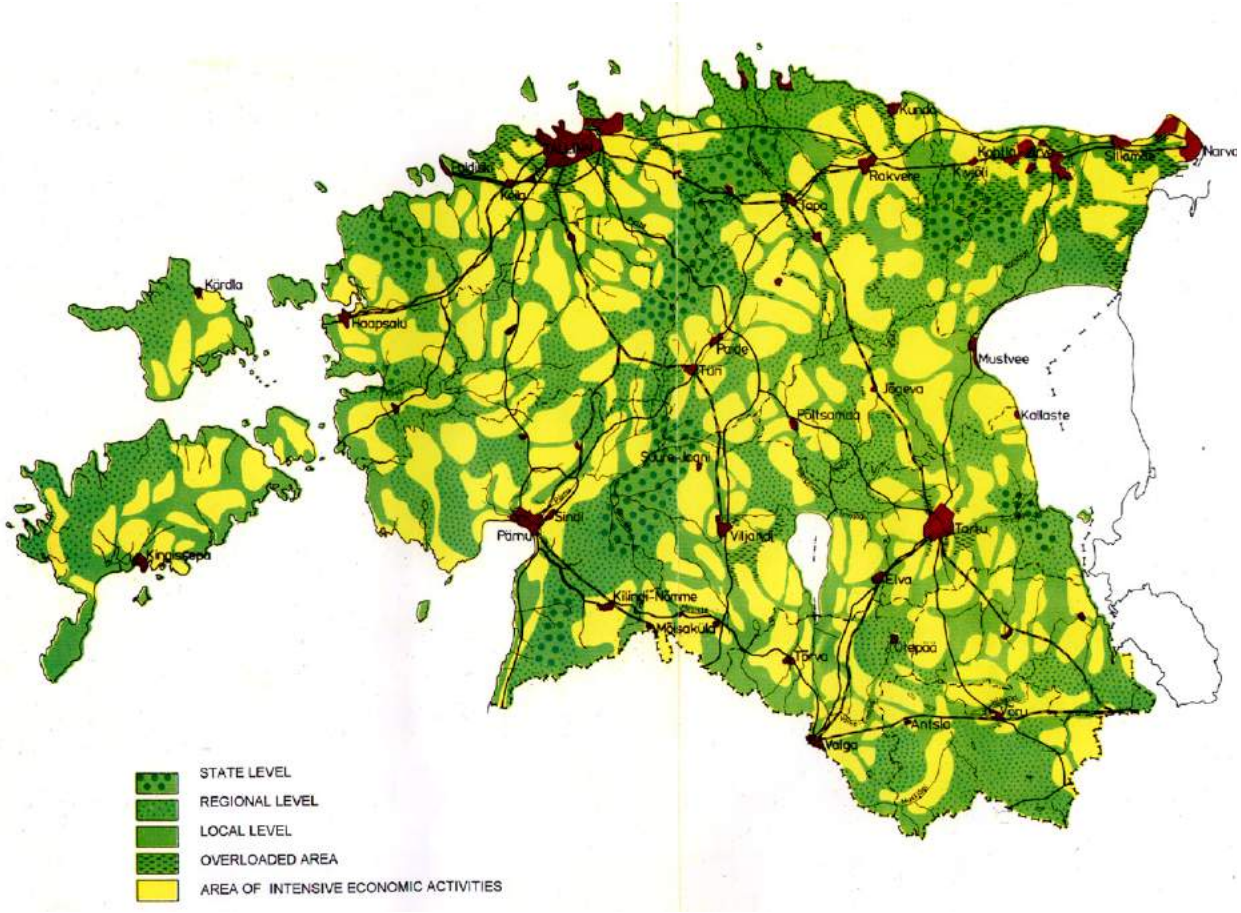


Fig. 5 Different approaches for development of ecological network or greenway system: in Estonia (Sepp *et al.*, 2013) and Florida (Florida Greenway Commission, 1994).

Motorway crossings for species are essential in a human dominated landscape. It has become an important issue in all countries in the world. In Europe about 50% of the country reports to the Bern Convention and to the CBD describe initiatives or the need to develop this kind of initiatives (http://www.coe.int/en/web/bern-convention/-/35th-standing-committee-meeting). The EU Directorate General for Environment was important in providing incentives for developing Green Infrastructure (GI) and large-scale infrastructure crossings for fauna. Such initiatives act as flagship actions and serve as examples at national, regional and local levels to boost the further implementation of trans-European GI in policy, planning and financing decisions. Member States and regions have in this way been stimulated to develop GI in a cross-border/transnational context (European Commission, 2013b).

Conflicts with human infrastructure can be overcome by spatial and technical solutions. The consequence is also, that the development of ecological networks and green infrastructure takes time for planning, finding funding and implementation. As it is a societal process, all stakeholders should be involved but they do not necessarily have to agree on all issues. Preparation takes time, the design has to be agreed based on landscape ecological knowledge, societal needs and spatial context. With a good landscape design, finances have to be found and afterwards maintenance and monitoring is needed as well. Technical solutions can be ecoducts for larger fauna (**Fig. 6**), culverts under roads for smaller fauna (**Fig. 7**) and fish ladders for fish (**Fig. 9**).

Fig. 6 Ecoduct Groene woud in motorway A2, the Netherlands (photo Rijkswaterstaat).

Fig. 7 Badger tunnel or underpass under Motorway A73, the Netherlands (photo Rob Jongman).

Fig. 8 Adapted culvert as wildlife underpass for small mammals (photo Rob Jongman).

Fig. 9 Fish ladder in one of the headwaters of the Tweed (Scotland) for migration of Salmon (*Salmo salar*) to its spawning waters. (photo Rob Jongman).

Implementation of fauna bridges and tunnels depend on willingness of sectoral (transport) policy makers and planners at national and regional level to consider biodiversity as an issue of the same importance and value as other societal interests. In some countries this change in perception and willingness to invest has been reached and solutions are being elaborated jointly by road planners, spatial planners and conservation planners. The Karpaten-Alpen corridor between Austria and Slovakia (www.alpenkarpatenkorridor.at/) has been elaborated and realized as an important European wildlife corridor system. It ensures the migration and genetic exchange between wild animal populations between Alps and the Carpathians by securing an ecological corridor through an urbanized area and to strengthen awareness of the importance of green connected areas and eco-friendly land use. In Croatia the highway from Zagreb to Rijeka stretches 68.5 km through a wildlife core area in Gorski Kotar. It has now a 100 m wide wildlife bridge at Dedin. Here on average 15.8 wildlife crossings (among others brown bears) per day have been counted (Kusak *et al.*, 2009). In The Dutch national Ministry for traffic and water management has set up a multi-annual defragmentation program in which 215 obstacles have been identified in national roads, rail and water infrastructure. Efforts to eliminate them involve measures like wildlife bridges (**Fig. 6**), wildlife underpasses (**Fig. 7**), and eco-culverts (**Fig. 8**). For supporting the design of fauna crossings and to integrate them in road construction there is a special guidance document (Kruidering *et al.*, 2005). The target of the program is to eliminate these obstacles by 2018 at the latest. At present it is expected that by then 94% is realized.

Defragmentation of rivers is an important target for the conservation of aquatic species, in particular fish. The US Forest Service is actually removing non-functional dams in rivers. In 1996, Benelux countries announced their intention of achieving free fish migration in all water catchments by 2010. In Belgium and the Netherlands this has not yet been realized, while Luxemburg does mention that in the 1990s multiple projects have been carried out leading to a more continuous water system. The Czech Republic mentions the existence of 6000 transverse barriers across rivers in the country. Hungary reports the disappearance of six fish species in its rivers due to damming of the Danube. Moldova announces agreements with Romania and Ukraine for the restoration of

common river systems. Also in the UK river connectivity is being restored by projects such as the revitalization of the Tweed catchment for the salmon by the Tweed water authorities and the Tweed Foundation boosting its Salmon population as well as regional tourism for salmon fishing (**Fig. 9**) and other projects by the Wildlife Trusts (http://www.wildlifetrusts.org/restoringrivers).

International corridors, such as in large rivers are difficult to manage as they require international agreements. The International Commission for the Protection of the Rhine (ICPR) has started in 1999 to restore the ecological river corridor of the Rhine and it has taken the salmon as its symbol and target species to be back as a healthy population in 2020 (Schulte-Wülwer Leidig, 2004). This project for accessibility of the river Rhine requires both national actions in seven countries and international political agreements on making the river accessible through among others fish ladders and other passages, long term projects for adaptation of locks and weirs including as final step the improvement of exchange between river system and the North Sea. This project is, of course, not only important for the salmon, but also for other migratory fish such as eel, sturgeon and shad.

See also: Ecological Complexity: Complex Ecological Networks

References

Bennet, G., 2004. Integrating biodiversity conservation and sustainable use, lessons learnt from ecological networks. Gland, Switzerland: IUCN.

Boere, G.C., Stroud, D.A., 2006. The flyway concept: What it is and what it isn't. In: Boere, G.C., Galbraith, C.A., Stroud, D.A. (Eds.), Waterbirds around the world. Edinburgh: The Stationery Office, pp. 40–47.

Bonnin, M., Bruszik, A., Delbaere, B., Lethier, H., Richard, D., Rientjes, S., Van Uden, G., Terry, A., 2007. The Pan European Ecological Network: Taking Stock. Strassbourg: Council of Europe, Nature and Environment series 146.

Bouwma, I.M., Jongman, R.H.G., Butovsky, R.O., 2002. Indicative map of the Pan-European ecological network for central and Eastern Europe. Technical background document. In: Technical Report Series., Tilburg/Budapest: ECNC.

Burel, F., Baudry, J., 2005. Habitat quality and connectivity in agricultural landscapes: The role of land use systems at various scales in time. Ecological Indicators 5, 305–313.

European Commission, , 2013a. Guidelines on Climate Change and Natura 2000. Dealing with the impact of climate change on the management of the Natura. 2000 Network of areas of high biodiversity value. Luxembourg: Publications Office of the European Union, doi:10.2779/29715.

European Commission, , 2013b. Building a green infrastructure for Europe. Luxembourg: Publications Office of the European Union, doi:10.2779/54125.

Florida Greenway Commission, , 1994. Creating a statewide greenway system, for people, for wildlife, for Florida. Tallahassee: Florida Department of Environmental Protection.

FOEN-EEA, , 2011. Landscape fragmentation in Europe. Joint EEA-FOEN report. EEA report 2/2011. Luxembourg: Publications Office of the European Union.

Foppen, R.P.B., Bouwma, I.M., Kalkhoven, J.T.R., Dirksen, J., Van Opstal, S., 2000. Corridors of the Pan-European ecological network: Concepts and examples for terrestrial and freshwater vertebrates. Tilburg: ECNC Technical Report Series.

Forman, R.T.T., 1995. Land mosaics, the ecology of landscapes and regions. Cambridge: Cambridge University Press.

Hanski, I., 2001. Spatially realistic theory of metapopulation ecology. Naturwissenschaften 88, 372–381. doi:10.1007/s001140100246.

Hanski, I., Gilpin, M., 1991. Metapopulation dynamics: Brief history and conceptual domain. Biological Journal of the Linnean Society 42, 3–16.

Jaeger, J.A.G., Bowman, J., Brennan, J., Fahrig, L., Bert, D., Bouchard, J., Charbonneau, N., Frank, K., Gruber, B., Tluk von Toschanowitz, K., 2005. Predicting when animal populations are at risk from roads: An interactive model of road avoidance behavior. Ecological Modelling 185, 329–348.

Jongman, R.H.G., 2002. Homogenisation and fragmentation of the European landscape: Ecological consequences and solutions. Landscape and Urban Planning 58, 211–221.

Jongman, R.H.G., Pungetti, G.P. (Eds.), 2004. Ecological networks and greenways, concepts methods and implementation. Cambridge: Cambridge University Press.

Jongman, R.H.G., Külvik, M., Kristiansen, I., 2004. European ecological networks and greenways. Landscape and Urban Planning 68, 305–319.

Jongman, R.H.G., Bouwma, I.M., Griffioen, A., Jones-Walters, L., Van Doorn, A.M.M., 2011. The Pan European ecological network—PEEN. Landscape Ecology 26 (3), 311–326. doi:10.1007/s10980-010-9567-x.

Kruidering, A.M., Veenbaas, G., Kleijberg, R., Koot, G., Rosloot, Y. and Van Jaarsveld, E. 2005. Leidraad faunavoorzieningen bij wegen. (Guidance for fauna adaptations in roads, illustrated, in Dutch) Delft: Rijkswaterstaat, Dienst Weg en Waterbouw.

Kusak, J., Huber, D., Gomerčić, T., Schwaderer, G., Gužvica, G., 2009. The permeability of highway in Gorski kotar (Croatia) for large mammals. European Journal of Wildlife Research 55, 7–21. doi:10.1007/s10344-008-0208-5.

Lafaille, P., Acou, A., Gillouët, J., Legault, J., 2005. Temporal changes in European eel, *Anguilla anguilla*, stocks in a small catchment after installation of fish passes. Fisheries Management and Ecology 12, 123–129.

MacArthur, R.H., Wilson, E.O., 1967. The theory of island biogeography. In: Monographs in population biology, 1. Princeton: Princeton University Press.

Opdam, P., 1991. Metapopulation theory and habitat fragmentation: A review of holarctic breeding bird studies. Landscape Ecology 5 (2), 93–106.

Schulte-Wülwer Leidig, A., 2004. Rhine & Salmon 2020, a Programme for Migratory Fish in the Rhine System. Koblenz: International Commission for the Protection of the Rhine (ICPR).

Sepp, K., Veersalu, T., Külvik, M., 2013. In: Green network applications in Estonia. IUCN, p. 33.

Stenseth, N.C., Lidicker, W., 1992. Animal dispersal, small mammals as a model. London: Chapman and Hall.

Further Reading

Levins, R., 1970. Extinctions. In: Some mathematical questions in biology: Lectures on mathematics in the life sciences. Providence Rhode Island: American Mathematical Society.

Ministry of Agriculture Nature Management and Fischeries, , 1990. Nature policy plan of the Netherlands. The Hague: Ministry of Agriculture Nature Management and Fischeries.

Relevant Websites

ecnc, n.d.—https://www.ecnc.org/programmes/green-infrastructure/.
ec.europa, n.d.—http://ec.europa.eu/environment/nature/index_en.htm.
epa, n.d.—https://www.epa.gov/green-infrastructure.
fs.usda, n.d.—https://www.fs.usda.gov/detailfull/srnf/home/?cid=STELPRDB5445190&width=full.
nature, n.d.—http://www.nature.com/news/dam-removals-rivers-on-the-run-1.15636?WT.ec_id=NATURE-20140731.
y2y, n.d.—https://y2y.net/.
wildlandsnetwork, n.d.—https://wildlandsnetwork.org/.
greenway, n.d.—https://www.greenway.org/.
alpenkarpatenkorridor, n.d.—http://www.alpenkarpatenkorridor.at/.
wildlifetrusts, n.d.—http://www.wildlifetrusts.org/restoringrivers.
tweedfoundation, n.d.—http://www.tweedfoundation.org.uk/.
iksr, n.d.—http://www.iksr.org/fileadmin/user_upload/Dokumente_en/rz_engl_lachs2020_net.pdf.
coe.int, n.d., http://www.coe.int/en/web/bern-convention/-/35th-standing-committee-meeting.

Conservation Biological Control and Biopesticides in Agricultural[☆]

ED Fountain and SD Wratten, Lincoln University, Canterbury, New Zealand

Prior to 1995, almost no research was conducted on biopesticides but recently a high number of new biopesticides are being researched, tested and introduced to the market. At the end of 2008, there were approximately 245 registered biopesticides in the USA. The Environmental Protection Agency (EPA) recognizes three major classes of biopesticides: microbial, biochemical, and plant-incorporated-protectants (PIPs). Although the use of biopesticides has increased, there is still resistance to their use. One of the major contributors to the negative perception of biopesticides is that growers do not know about biopesticides and the cost/efficacy. In an industry survey conducted in the USA, growers indicated that biopesticide companies need to place heavy emphasis education about the products and how to use them. Another barrier to biopesticide use is the highly competitive, capital intensive marketplace which is often dominated by multibillion-dollar agrichemical companies. However, these companies more recently have begun to recognize the benefit of biopesticides and have begun to produce new products. In 2008, Dow AgroSciences LLC, a major agrichemical company, won the Designing Greener Chemicals Award from the EPA for their product, Spinetoram.

Conservation biological control is the modification of the environment or of agronomic practices to protect and enhance the efficacy of natural enemies of pest organisms. Biopesticides are natural enemies of pest organisms released *en masse* to control the pest, where the agent does not usually persist in high numbers in the crop environment. These two methods are the two most recent forms of biological control and are relatively early in their developmental stages. However, these techniques are proving to have potential as effective tools in pest management and do not carry the sometimes high risks of classical biological control. This article will not concentrate on conservation biological control or biopesticides in particular crops, pest species, or diseases. Rather, it will assess the theories involved, progress to date, and prospects for further adoption of these practices and products.

Modern Agriculture

During the last 50–60 years, worldwide agricultural practices have been characterized by high inputs, high yields, unsustainable practices, and ecosystem damage. This production method, often termed 'substitution agriculture', relies on inputs of fertilizers and agrochemicals to maintain and enhance soil fertility and manage weeds, pests, and diseases. These chemicals damage the environment, reduce biodiversity and ecosystem function, and are increasingly becoming ineffective due to pest resistance and other factors mentioned above. Subsequently this form of agriculture has become dependent on these chemical inputs to maintain production. Public concerns about the unsustainability of these practices are being raised, along with concerns of the impacts of these practices on environmental and human health. Ecosystem services (ES) to humanity have been valued in excess of US$ 33×10^{12} pa worldwide. They include food production and quality, climate, soil and hydrological regulation, nutrient cycling, maintenance of genetic diversity/resources, and recreation. Recognition of the value of these services and a change in agricultural practices to restore ecosystem function are necessary to create sustainable growth, especially as the world population is expected to increase 50% in the next 50 years to 9 billion. In response to these social, economic, and marketing pressures, agriculture is being forced to discontinue the use of some agrochemicals (e.g., organochlorine insecticides) or they are to be phased out over time (e.g., methyl bromide). This has raised attention in the research community to find sustainable alternatives to 'substitution agriculture'. One area of focus has been biological control, especially conservation biological control and biopesticides, both as part of an agro-ecological approach. The early stages of these disciplines were rudimentary but current research is now showing great potential to play an important part in sustainable pest management in sustainable agriculture. However, there are several restrictions that impede a faster expansion of the discipline and its transfer to the agricultural community. These include the availability of selective biodegradable chemical pesticides, the lack of a strong information-transfer infrastructure, highly regulated restrictions on the

[☆]*Change History:* June 2013. ED Fountain and SD Wratten updated the text and further readings to this entire article.

 http://dx.doi.org/10.1016/B978-0-12-409548-9.00539-X

introduction and deployment of exotic agents, and poor grower perception and knowledge of biological control and how to use and enhance it. Also, there are not enough market-based incentives (or instruments) to deploy biological control technologies. This is in spite of the fact that ample evidence exists that the (ecosystem) service-providing unit (SPU) is well established and that the economic value of such provision/avoided costs, maintenance of, and increases in markets has been well demonstrated.

Conservation Biological Control

Insects

Conservation biological control has been the least-studied area of all biological control techniques and has been dominated by arthropod pest systems. This technique usually adds plant biodiversity to agricultural systems through the provision of shelter and nonprey food, especially in high-value crops such as wine grapes. One of the most successful conservation biological control techniques has been 'beetle banks' which were developed by the Game and Wildlife Trust and the University of Southampton in the United Kingdom. They comprise overwintering refugia for predatory arthropods in arable land and have the added benefit of providing nesting sites for rare farmland birds and mammals. Although floral resources for pests' natural enemies were planted in the 1980s and 1990s, their selection and deployment were not obviously informed by ecological science. There later developed a growing awareness of the possible impact of the plants on the pest itself and on any parasitoids of natural enemy species. Bioassays to measure the effects of flowers on fecundity and longevity in these trophic levels have recently been carried out so that the deployment of selective floral diversity in agricultural systems is now possible. Recently, floral nectar has been analyzed for its sugar ratio, then the same plants were assessed with pests' natural enemies. Flower structure and size also have an impact on the system. Long, narrow flowers may provide nectar only to larger insects with long mouthparts, whereas large, shallow flowers provide nectar to all species. Further knowledge of such aspects of floral resources will allow plant selection to benefit the third trophic level more than the second and the fourth, which will allow agriculturalists to maximize the effects of the technology.

An area of recent development in conservation biological control is the awareness of herbivore-induced plant volatiles (HIPVs). These (mainly terpenoids and indoles) are systemically released by the plant during insect herbivore feeding and attract predatory and parasitic insects to the plant. In tomato, this response led to twice as many parasitoid larvae (*Hyposoter exiguae*) developing on tomato fruitworm (*Spodoptera exigua*) than in the control. There is a large range of volatiles released by different plant species; some attract a narrow range of natural enemies, others a broad range. If natural enemies can be attracted into a crop in this way ('attract') and their 'fitness' enhanced by the provision of appropriate selective nectar resources ('reward'), then the prospects for an 'attract and reward' pest management program will have been developed.

Attention is now greatly focused on not only the importance of the floral resource and the species on which it impacts, but also the importance of each species with regard to ecosystem function. Ecologists generally agree that higher biodiversity *per se* does not necessarily increase ecological function, so the role of particular species or groups of species is being investigated. This will further refine research efforts to effectively maximize ecosystem services, including biological control. Also, adding resources to enhance a particular ES is likely to enhance others, too.

Plant Pathogens

Conservation biological control of plant pathogens is in the early stages of development but the technology shows great potential. The most recent work has been done on the grapevine/*Botrytis cinerea* system where the pathogen's life cycle was disturbed and levels of primary inoculum were reduced through the use of organic mulches or of cover crops, mulched *in situ*. Levels of primary inoculum from vine debris were reduced under mulch, through an increase in the activity of soil biota, both through competition with the pathogen for resources and through increasing rates of vine debris degradation. The changes in soil biota were linked to soil moisture and possibly soil nutrient levels. The vines under the organic mulches used in this work sustained half the rates of botrytis bunch rot at harvest, compared with nonmulch controls, averaged over 2 years (**Figure 1**), and brought the disease below the economic threshold of the region. Progressive grape growers are now using organic mulches for this purpose in their vineyards. This adoption sets a precedent for other growers and large-scale adoption of the technique is possible. Other advantages of the mulches that may help adoption are that they are easy to apply/manage and the materials they use are either cheap or waste-stream products from the vineyard itself. This biological system is potentially applicable to other plant pathogen systems (e.g., downy mildew (*Plasmopara viticola*)) which overwinter on plant debris. The technology could also be integrated into other understorey manipulation techniques, such as the provision of flowering plants (discussed above), where plants could be mulched after flowering.

Biopesticides

Insects and Pathogens

The technique does not include toxins or secondary metabolites that are produced from living organisms as these are nonliving. In 2000, biopesticides sales worldwide had a value of US $160 million of which about 90% contained *Bacillus thuringiensis* (Bt), a Gram-positive bacterium found in soil or on plant surfaces. The sales of biopesticides have been increasing rapidly since the 1980s.

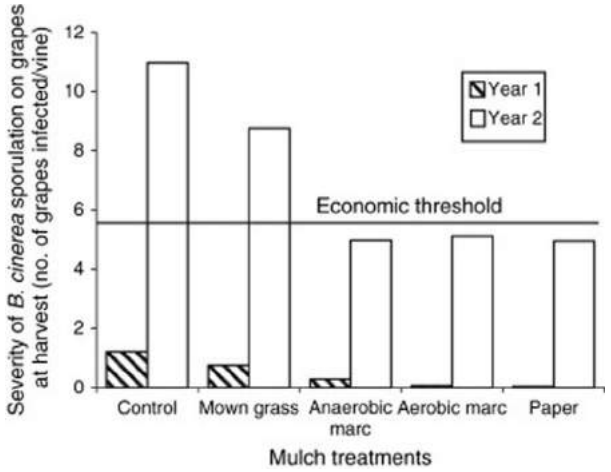


Figure 1 Severity of botrytis bunch rot in grape vines at harvest, under different mulch treatments over two consecutive seasons in Seresin Estate Vineyard, Marlborough, New Zealand. Three of the mulch treatments bring botrytis bunch rot below the economic threshold in the second year.

This market is mainly driven by consumer requirements for environmental and human health and the increasing problem of pest resistance. Currently, key markets for these products are organic and integrated pest management (IPM) systems and those with high pest resistance and high-value speciality crops.

The first example of a commercially available biopesticide was in the 1920s, where *Encarsia formosa* was sold to control greenhouse whitefly. This biopesticide is currently widely available commercially and has been sold in increasing amounts since initiation. This is one of the several examples where an invertebrate species has been used as a biopesticide agent and is largely successful because of the controlled environment of a greenhouse. Generally, invertebrates are not considered to be appropriate for the technique as they are too difficult to apply and are generally too mobile for field conditions. The first biopesticide containing a microorganism as the biologically active ingredient was Bt and was used in France in the 1930s; it is a product which has dominated the biopesticide market ever since. From this time, development in the area has been focused on creating a product that is easy to culture, store, and apply, has a fast impact, and is environmentally robust. Initially potential agents were usually tested and found to be effective under the ideal environmental conditions of the laboratory, then applied under field conditions where they often failed. Failures were commonly a result of poor spray coverage, low spore viability, or unfavorable environmental conditions. The importance of the environment for agent virility was then fully acknowledged and agents were then screened and tested for effectiveness under different climatic conditions. This approach not only assessed compatibility between the agent and the host, but also propagule hardiness and viability under adverse conditions. Efforts were also spent on formulating a nutrient-rich, stable medium in which the propagule could be stored and applied, in an attempt to maximize the agent's efficacy.

Currently, research in the area comprises many disciplines including pathology, physiology, genetics, mass culturing, formulation technology, and propagule stabilization and application; yet there is little collaboration between the disciplines and this hinders the rate of development of the technology. Interest in biopesticide products from agriculture is also generally low due to the fragmented nature of the industry, the lack of interest from agrochemical industries, the availably of 'soft' orthodox pesticides, low grower awareness, and their education about and perception of biopesticides. For the technology to develop further, these things need to change and growers need to change their view of biopesticides from purely a substitute for agrochemicals to a valuable biological organism with environmental requirements.

Biopesticides – Areas of Use

Biopesticides are currently used to manage invertebrate pests, plant pathogens, and weeds. The microorganisms used include bacteria, yeasts, fungi, and viruses. Most research has been conducted on fungal species; as many are easily cultured, their spores are environmentally stable and can be easily managed, amended, and manipulated – they can hence be applied with convention spray systems. To date, much more research effort has been spent on the development, registration, and commercialization of products for plant pathogens and pest species than of weed species, so only the former two will be covered here.

Plant Pathogens

Biofungicides attack plant pathogens through several mechanisms, including antibiosis, hyperparasitism, and competition, with the latter usually being the predominant and most important mechanism. There are many products available for a wide range of crop/pathogen systems which are effective on both foliar and root diseases. Many agents are highly competitive for space and nutrients but usually have a low capability of displacing developed microorganisms. These characteristics make biofungicides preventative treatments but some have limited curative potential. Biofungicides share the characteristics of the best biological

control agents of being highly reproductive, environmentally hardy, and highly antagonistic. Many of the current products contain a series of species from the same genus that operate through different mechanisms to obtain higher levels of control. To date, most products for foliar and soil plant pathogens are *Trichoderma* species and can help to manage economically important plant pathogens such as botrytis rot, powdery mildew, and *Sclerotinia* species. Products have also been developed as seed dressings from the genus *Bacillus* and *Pseudomonas* to control several soil-borne plant diseases. Species have also been investigated as potential biofungicides when they are suspected to have a role in suppressive soils, a concept where plants grown in soils high in particular soil flora are more resistant to disease. Fluorescent pseudomonades are one of these species and can inhibit germination of fungi through competition for ferrous minerals.

Microorganism biopesticides for the management of invertebrate pests

More than 1500 species of known fungi, viruses, bacteria, and protozoa can infect arthropod species. The first attempt to use them as bioinsecticides was in 1884 when a Russian scientist, Elie Metchnikoff, used *Metarhizium anisopliae* to infect two invertebrate pests, the cereal cockchafer, *Anisoplia austriaca*, and the sugar-beet weevil, *Bothynoderes punctiventris*. The first commercially available bioinsecticide, Bt, was produced in the 1930s and many bioinsecticides have been produced since, the majority of which are entomopathogenic hyphomycete fungi. Bioinsecticides have not always provided consistent control of insect pests, as the relationship between the agent, the pest, the timing, and the environment is inherently complex. Of these variables, the environment is the most important and the one most likely to lead to a breakdown in the system. Consequently, bioinsecticides have been most successful in the warm, humid conditions of the tropics. When environmental conditions change, both the pest and the biological control agent are affected. Both species have an optimum range of temperatures and humidities. In temperate climates, the agents' optimum temperature and humidity conditions are usually higher than those of the pest species, so the agents' performance can decline and the pests' fitness can increase with a decline in these environmental variables. Pest health also has impacts on resistance to fungal attack, especially pest nutrition, age, genetics, and physical damage through mechanical, chemical, or biological means. Agriculturalists can influence pest health and, as a consequence, increase pest susceptibility to bioinsecticides, by using resistant crop cultivars or manipulating the habitat to alter pest nutrition. Other variables that influence the success of the technique are the virulence and hardiness of the agent, its compatibility with the host, and the viability of the propagule after storage. To increase the effectiveness of the products, agriculturalists are urged to monitor environmental conditions to assist in deciding when bioinsecticides should be applied and when to apply them to optimize environmental conditions. Spray machinery should also be calibrated to ensure good spray coverage at the correct rate, as this will increase the likelihood of exceeding the pest propagule threshold and lead to effective management of the pest as a consequence.

Conclusion

Conservation biological control and biopesticides have great potential in organic agriculture or in IPM programs. These systems are complementary and elements of them can be integrated into conventional pest management systems. The two techniques do have their limitations, however, and most agriculturalists need to adjust their perception of the technologies, especially biopesticides. Currently, the latter are viewed as substitutes for agrochemicals, and the products are expected to operate at the same speed and under the same wide range of environmental conditions as conventional products. Agriculturalists require education about the biology of the biopesticide and the target pest. They need to set up pest monitoring systems, develop a network of contacts to assist in decisions concerning when and how to apply biopesticides, and realize the limitations of the technology. Limitations include being effective only pre-infection, of needing different biopesticides for particular pests, of environmental conditions, etc.; these are important influences on the efficacy of control. Acceptance of the limitations and the gaining of appropriate knowledge should allow biopesticides and conservation biological control techniques to be integrated into agroecosystems to produce a more sustainably produced, higher-value product. This approach can enhance the contribution of ecosystem services to pest, weed, and disease control, satisfying increasing environmental, energy, human health, and marketing demands. However, as long as conventional pesticides remain relatively inexpensive, because the economics of their negative external costs (human health and environment) are no included in the price, biopesticides may continue to have a small market share. Nonetheless, the factors mentioned under 'Modern Agriculture' are likely to change this system.

Further Reading

de Schutter O (2010) *Agroecology and the right to food. Report for United Nations*.

Fiedler AK, Landis DA, and Wratten SD (2008) Maximizing ecosystem services from conservation biological control: The role of habitat management. *Biological Control* 45: 254–271, http://dx.doi.org/10.1016/j.biocontrol.2007.12.009.

Gurr GM, Wratten SD, and Snyder WE (eds.) (2012) *Biodiversity and insect pests: Key issues for sustainable management*. Oxford, UK: Wiley-Blackwell.

Gurr GM, Wratten SD, and Altieri MA (2004) *Ecological engineering for pest management, advances in habitat manipulation for arthropods*. Collingwood, Australia: CSIRO Publishing.

Jacometti MA, Wratten SD, and Walter M (2007) Understorey management increases grape quality, yield and resistance to *Botrytis cinerea*. *Agriculture, Ecosystems and Environment* 122: 349–356.

Landis DA, Wratten SD, and Gurr GM (2000) Habitat management to conserve natural enemies of arthropod pests in agriculture. *Annual Review of Entomology* 45: 175–201.

Luck GW, Daily GC, and Ehrlich PR (2003) Population diversity and ecosystem services. *Trends in Ecology and Evolution* 18: 331–336, http://dx.doi.org/10.1016/S0169-5347(03)00100-9.

Simpson MG, Gurr M, Simmons AT, et al. (2011) Field evaluation of the 'attract and reward' biological control approach in vineyards. *Annals of Applied Biology* 159: 69–78.
Wratten SD, Sandhu H, Cullen R, and Costanza R (2013) *Ecosystem services in agricultural and urban landscape*. Oxford, UK: Wiley-Blackwell.
Wratten SD, Gillespie M, Decourtye A, et al. (2012) Pollinator habitat enhancement: Benefits to other ecosystem services. *Agriculture, Ecosystems and Environment* 159: 112–122, http://dx.doi.org/10.1016/j.agee.2012.06.020.
Zehnder G, Gurr GM, Kühne S, et al. (2007) Arthropod pest management in organic crops. *Annual Review of Entomology* 52: 57–80.

Relevant Websites

waiparawine.co.nz http://www.waiparawine.co.nz; – Research in the Waipara Valley, Waipara Valley Winegrowers Inc.
http://www.pure-ipm.eu/ - Innovative crop protection for sustainable agriculture in the EU.
http://www.sdupdate.org/home - Updates on sustainable developments.

Conservation Genetics

Richard Frankham, Macquarie University, Sydney, NSW, Australia and Australian Museum, Sydney, NSW, Australia

Glossary

Effective population size (N_e) The number of individuals that would result in the same loss of genetic diversity, inbreeding, genetic drift, or coalescence if they behaved in the manner of an idealized population.

Evolutionary potential The ability of a population to evolve, especially to cope with environmental changes. Often equated with genetic diversity, especially for quantitative characters such as fitness.

Genetic diversity The extent of genetic variation in a population, or species, or across a group of species e.g., heterozygosity, or allelic diversity, or heritability.

Genetic drift Changes in the genetic composition of a population due to random sampling in finite populations. Also referred to as random genetic drift.

Genetic load The load of harmful alleles in a population, some due to mutation-selection balance (mutation load), and others to heterozygote advantage and other forms of balancing selection.

Genetic rescue Improvement in reproductive fitness and increase in genetic diversity due to crossing to another distinct population, of a population previously suffering from inbreeding and low genetic diversity.

Inbreeding depression Reduction in mean for a quantitative trait due to inbreeding, especially manifest in reproductive fitness traits.

Kinship The probability that two alleles, one from each of two individuals, are identical by descent (also termed coancestry). The inbreeding coefficient of any real or imagined progeny of the two individuals.

Mutation-selection balance Equilibrium between the occurrence of harmful mutations and natural selection removing them, resulting in low frequencies (typically $<1\%$) of harmful alleles at individual loci in populations. This occurs at loci throughout the genome.

Outbreeding depression A reduction in reproductive fitness compared to either parent observed in the progeny of F_1, F_2 or later generation crosses between two genetically divergent populations (or sub-species, or species) caused, for instance by some combination of fixed chromosomal differences and/or disruption of local adaptation and long isolation.

What Is Conservation Genetics?

Conservation genetics is the application of genetics to understand and reduce the risk of population and species extinctions. It deals with genetic factors causing rarity, endangerment and extinction (inbreeding and loss of genetic diversity), genetic management to minimize these impacts, and the use of genetic markers to aid in resolving taxonomic uncertainties in threatened species, to understand their biology, and in wildlife forensics. It is an applied discipline drawing on evolutionary and molecular genetics and genomics.

The need to conserve species arises because the biological diversity of the planet is rapidly being depleted as a direct or indirect consequence of human actions. An unknown but large number of species are already extinct, while many others have reduced population sizes that put them at risk. Many species now require human intervention to ensure their survival. The scale of the problem is enormous; 26% of mammals, 13% of birds, 42% of amphibians and 40% of gymnosperms are categorized as threatened by the International Union for the Conservation of Nature (IUCN), with similar problems in other groups, but insufficient species evaluated to provide reliable statistics for them.

Four justifications for maintaining biodiversity have been advanced; the economic value of bioresources, ecosystem services, aesthetics, and rights of living organisms to exist. IUCN recognizes the need to conserve biodiversity at three levels; genetic diversity, species diversity, and ecosystem diversity. Genetics is involved in all three of these. In what follows, the emphasis is on the first two.

When Did Conservation Genetics Begin and How Has It Developed?

Sir Otto Frankel, an Austrian born Australian, was largely responsible for the recognition of genetic factors in conservation biology, beginning only in the early 1970s. Frankel collaborated with, and strongly influenced Michael Soulé of the United States, the founding father of modern conservation biology, and they wrote the first book on conservation biology that considered genetic factor (published in 1981). There was initially controversy regarding whether inbreeding had harmful effects on reproductive fitness (reproduction and survival) in wild species. In captivity, a series of pioneering studies by Katherine Ralls, Jonathan Ballou and collaborators in the late 1970s and early 1980s showed that inbreeding had harmful effects on juvenile survival: they found that inbred offspring had lower juvenile survival than outbred offspring in 41 of 44 mammal populations. The issue then turned to

 doi:10.1016/B978-0-12-409548-9.10559-7

the relationship between inbreeding and extinction in wild species in natural habitats, and whether nongenetic factors typically drive species to extinction before genetic factors could impact them. Empirical evidence refuted this hypothesis for most species: it showed that inbreeding caused harmful fitness effects, including elevated extinction risk, and the impacts were more severe than in captivity. The first textbook in the field was only published in 2002. More recently, there has been growing emphasis on the genetic impact of population fragmentation.

How Does Conservation Genetics Advance?

Conceptual advances in conservation genetics originate from mathematical modeling, computer simulations, laboratory experiments, field experiments, meta-analyses, and the feedback between them. Given that conservation genetics is a complex system with many variables, especially in the wild, single experiments and approaches rarely resolve issues relating to wild populations. Consequently, several independent approaches are often required to resolve issues, as was the case for the role of genetic factors in extinction (where theory, computer simulations, lab experiments, field experiments and meta-analyses all contributed).

Such approaches are complemented by technical advances, as with the improvements in statistical power and precision obtained with the progression from morphological and fitness traits, to protein electrophoresis, to DNA markers, and to genomics.

Why Is Conservation Genetics Important?

Conservation genetics is important because it deals with significant factors contributing to species and population extinctions (inbreeding, loss of genetic diversity and mutational accumulation), and by providing convenient techniques to obtain important information on species biology that contributes to their conservation, as discussed below.

Inbreeding

Inbreeding is the production of offspring from individuals that are related by (recent) descent from a common ancestor of their parents. It is measured using Wright's inbreeding coefficient (F), the probability that two alleles at a locus in an individual are identical by descent. This coefficient ranges from 0 to 1, and is 0.5 for the progeny of self-fertilization, 0.25 for the progeny of brother-sister mating, and less for progeny of matings between cousins. It is an unavoidable occurrence in small closed population (ones with no gene flow), accumulating slowly over generations in random mating populations: in hermaphrodites, F increases by the proportion $1/(2N_e)$ per generation, where N_e is the genetically effective population size.

Effective Population Size

All of the genetic impacts of small population size (inbreeding, loss of genetic diversity and genetic diversification among replicate populations) depend upon the genetically effective population size (N_e), rather than the census size (N). The genetically effective size is the number of individuals that would result in the same loss of genetic diversity, inbreeding, genetic drift, etc., if they behaved in the manner of an idealized population. An idealized population consists of hermaphrodites with equal number in each generation, random union of gametes, Poisson variation in family sizes, nonoverlapping generations, and no migration, mutation or selection.

How Are Effective and Census Sizes Related?

As real populations violate the assumptions of the idealized population in having fluctuations in population sizes over generations, greater than Poisson variation in family sizes, unequal sex-ratios, and overlapping generations, effective population sizes are generally much less than census sizes. Since we rarely have N_e values for threatened species, we often wish to infer it from census sizes and N_e/N ratios from related species where the ratio is known. On average the N_e/N ratios are ~ 0.13 based on three meta-analyses, but the ratio varies with life-history attributes, especially age at maturity and adult lifespan. In particular, species with high fecundity have lower ratios on the order of 10^{-3}–10^{-6}, based on data from fish, oysters, shrimp and seaweed. Low ratios mean that harmful genetic effects occur sooner and at a greater rate than assumed from census sizes.

Inbreeding Depression

Inbreeding results in reductions in fitness (inbreeding depression) in all naturally outbreeding species, and to a lesser extent in species with inbreeding mating systems. It has been known since as early as Darwin's work in the 1870s, but its full impact has only recently begun to be documented. All fitness traits are susceptible to inbreeding depression, but traits peripheral to fitness exhibit little or no inbreeding depression. For example, inbreeding depression is observed for fecundity and survival in *Drosophila* fruit flies, but not for number of bristles on abdominal segments.

How Does Inbreeding Depression Arise?

Inbreeding depression occurs because species contain a load of rare harmful partially recessive alleles due to mutation-selection balance, and because some loci exhibit heterozygote advantage. Inbreeding increases homozygosity at these loci exposing harmful recessive alleles in homozygotes. There must be directional dominance (beneficial alleles dominant, and harmful ones partially recessive) across loci for there to be inbreeding depression. Such directional dominance is found for fitness traits as they are typically subject to directional selection, whilst traits peripheral to fitness typically experience stabilizing selection (favoring phenotypic intermediates), or selection that varies over space and time. Inbreeding depression is found for all fitness components, including mating and fertilization ability, sperm production and quality in males, fecundity, survival, disease resistance, predator avoidance, mothering ability, etc., in animals, and equivalent traits in plants. Inbreeding of mothers and of zygotes both result in harmful effects on offspring fitness.

Inbreeding depression is found in all animal and plant species that are diploid or have higher ploidies. Haploids do not exhibit inbreeding depression, while haplodiploid species (e.g., Hymenoptera) may experience it in diploid females, but not in haploid males.

How Large Are the Effects of Inbreeding Depression?

Offspring resulting from brother–sister matings for mammals in captivity have on average 33% lower juvenile survival than noninbred offspring, with there being no clear differences among mammalian orders. However, the effects are greater in the wild than in captivity. Further, the effects accumulate across the life cycle and are devastating for total fitness in the wild (**Table 1**). For example, offspring of a full brother–sister mating in the New Zealand takahe bird experience a total reduction in fitness of 88%. Inbreeding depression is proportional to the amount of inbreeding, and is greater in more stressful environments, for outbreeding than inbreeding species, and for populations that have not previously been inbred versus ones subject to prior inbreeding.

Mutational Accumulation and Meltdown

Mutational meltdown describes the accumulation of new harmful mutations in small populations and their subsequent adverse effects on fitness and population persistence. It is a phenomenon closely related to inbreeding depression in that both involve exposing harmful recessive alleles. They differ in that inbreeding depression predominantly increases homozygosity for preexisting harmful mutations, while mutational meltdown focusses on accumulation and homozygosity of new harmful mutations.

While harmful alleles are kept at low frequencies in large populations due to the balance between mutation and natural selection, in small populations genetic drift becomes the primary determinant of whether alleles become more (or less) common each generation. Thus, mildly harmful alleles behave as if they are subject only to genetic drift (become effectively neutral), and some increase in frequency and reduce reproductive fitness. Over many generations, sufficient harmful alleles may accumulate to cause negative population growth and a decline to extinction. The contribution of mutational accumulation to extinction risk in small populations of outbreeding sexually reproducing species is controversial, but it appears to be minor compared to other genetic threats. However, it is expected to be somewhat more important in asexual and selfing species.

Loss of Genetic Diversity

Genetic diversity ultimately arises by mutation and is lost by selection favoring particular alleles over others, and by chance sampling effects (genetic drift). Genetics diversity is lost in small closed populations at rates per generation that are inversely

Table 1 Inbreeding depression (ID) for total fitness in wild species of animals and plants due to a brother–sister mating expressed as percentage reduction in mean of inbred progeny compared to outbred progeny in the same wild environment (Frankham *et al.*, 2017)

Common name	*Genus and species*	*ID (%)*
Red deer	*Cervus elaphus*	99
Collared flycatcher	*Ficedula albicollis*	94[a]
Great tit	*Parus major*	55
Song sparrow	*Melospiza melodia*	79
Takahe	*Porphyrio hochstetteri*	88
Deerhorn clarkia	*Clarkia pulchella*	100[a]
Rose pink plant	*Sabatia angularis*	38[a]
Wild radish	*Raphanus sativus*	56[a]

[a]Maternal inbreeding contribution not included, so it is an underestimate.

proportional to the genetically effective size of populations (N_e), and diversity progressively erodes across generations (t), as described by the following equation for loci not subject to selection:

$$H_t/H_0 = (1-1/[2N_e])^t$$

where H_t is the heterozygosity at generation t and H_0 that initially. For example, a population with an N_e of 25 is expected to lose 40% of its heterozygosity over 25 generations. Loss of genetic diversity for microsatellite DNA markers (variable number short tandem repeats of bases, that and are subject to little or no selection) in experimental populations occurs at a rate that is close to the predictions of this equation. Genetic diversity for loci subject to selection is also lost in small closed population. From theory we expect that genetic diversity in closed population will be related to population size, and this is observed within and across species. Further, as threatened species have, by definition smaller and/or declining population sizes than nonthreatened species, they typically have lower genetic diversity than taxonomically related nonthreatened ones.

Reduced Evolutionary Potential

Populations and species experience environmental change and in the face of persistent directional environmental change must evolve or they will go extinct. The ability to evolve depends on genetic diversity for fitness, selective forces, and genetically effective population size of populations. Consequently, threatened species are expected to have poorer ability to evolve than taxonomically related nonthreatened ones.

Do Genetic Factors Increase Extinction Risk?

Since inbreeding and loss of genetic diversity reduce reproduction and survival and the ability to evolve, they might be expected to increase extinction risk. However, it was hypothesized by Russell Lande in the late 1980s that demographic factors (human factors, plus demographic and environmental stochasticity and catastrophes) would often drive species to extinction before genetic factors could impact them, and this view was widely promoted. However, empirical evidence refuted this hypothesis for most populations. Further, genetic factors have been shown in increase extinction risk in wild habitats for outbreeding diploid species, based on direct observations and computer simulations parameterized with data from real threatened populations. For example, inbreeding explained 26% of the variation in extinction risk in wild populations of the Glanville fritillary butterfly in Finland. Further, populations of deerhorn clarkia plant (*Clarkia pulchella*) with an inbreeding coefficient of 4% had a 25% extinction rate, while those with an 8% inbreeding coefficient had a 69% extinction rate over 3 years in the wild i.e., a small increase in inbreeding resulted in a large increase in extinction rate. Computer simulations indicate that inbreeding increases extinction risk over a wide range of scenarios in outbreeding species. For example, across 30 vertebrate species, inclusion versus exclusion of inbreeding depression at realistic levels resulted in a 37% reduction in median times to extinction in wild populations over a range of population carrying capacities. Loss of genetic diversity has also been shown to increase extinction risk under environmental change.

It is important to recognize that human impacts, genetic impacts, catastrophes and demographic and environmental stochasticity (natural fluctuations) all contribute to extinction risk, and that they typically interact in a downward feedback loop termed the extinction vortex.

Extinction Vortex

Inbreeding and loss of genetic diversity typically exert their impacts on population persistence in concert with other catastrophic, demographic and environmental and human factors in a downward spiral towards extinction (termed the extinction vortex). Typically human impacts reduce population size, increasing susceptibility to demographic, environmental and genetic threats which further reduce the population size, and so on in a downward spiral towards extinction. Such extinction vortices have been observed in a number of populations. Various software packages exist to simulate such interactions of genetic and nongenetic factors, especially VORTEX (Lacy and Pollak 2014).

An important implication of this is that if genetic factors are ignored, conservation programs for threatened species may fail, even if the original cause of decline was nongenetic. This happened in the case of the greater prairie chicken population in Illinois. Restoration of habitat failed to alleviate the bird's decline, and it only recovered after its genetic deterioration was reversed by gene flow from isolated populations in other states.

Genetic Management of Captive Populations

Many species are incapable of persisting in the harsh conditions in the wild, often as a results of persistent human impacts, and so must be captive bred to save them from extinction. The three major genetic issues in management of captive populations of threatened species are inbreeding depression, loss of genetic diversity, and genetic adaptations to captivity that are harmful when species are returned to the wild. The first two issues were addressed by seminal work from Jonathan Ballou and Robert Lacy

published in 1995: they proposed that management by mean kinship was the optimal means for managing pedigreed populations. The kinship of two individuals is the inbreeding coefficient of their potential offspring. This procedure has been validated in a *Drosophila* fruit fly experiment, and applied widely in captive management of threatened species.

The success of reintroductions of threatened species into the wild depends on their fitness in their natural environment. Unfortunately, adaptations to captive environments are overwhelmingly harmful when populations are returned to the wild. For example, salmonid fish lose approximately 37.5% of their wild fitness for each generation of captive breeding. Similar processes occur in other species, but the rates of decline in wild fitness per captive bred generation are expected to vary widely according to the life-history of the particular species. Genetic adaptation to captivity can be reduced by minimizing selection, extending the generation interval, and especially by deliberate use of population fragmentation, but there has been little deliberate captive management to address this issue.

Genetic Management of Wild Populations

Despite the need to conserve genetic diversity being specified in the United Nations Convention on Biodiversity and legislation and regulations in many countries, genetic management of wild population has been slow to take off. For example, only ~50% of recovery plans (66% in United States, 55% in Australia and 33% in Europe) mention genetics, only 7% mentioned inbreeding, and even fewer inbreeding depression, and fitness-related parameters were generally overlooked in all plans (Pierson *et al.* 2016).

Much of the concern with genetic management of wild populations centers on genetic management of fragmented animal and plant populations, an issue considered as one of the most important, largely unaddressed issues in conservation biology. Most species have fragmented distributions, often with genetically isolated fragments. Adverse genetic impacts in these occur at a rate that is much more rapid than in the species as a whole, as inbreeding and loss of genetic diversity occur at rates dependent on the effective population size of the fragment, rather than that of the species. Consequently, these small isolated fragments have much elevated extinction risk due to genetic and combined genetic, demographic, environmental and human-associated factors. Further, population extinction is a step on the path to species extinction. Many of these small isolated population fragments could be rescued by gene flow from other population segments (genetic rescue). However, there are only ~30 cases worldwide where genetic rescue has been attempted for conservation purposes, yet there are ~1.4 million population fragments of threatened species that would benefit from it (Frankham *et al.* 2017).

Frankham *et al.* (2017) proposed a paradigm shift whereby evidence of inbreeding and loss of genetic diversity in isolated population fragments is followed not by inaction or separate management, but by asking:

- Does the population fragment needs genetic rescue?
- If so, is there any other population of the same species that can rescue it?
- If so, will augmentation of gene flow be beneficial or harmful?
- If beneficial, will the predicted benefits justify the cost and effort?

In this way they hope to stem the erosion in number of population fragments of species, and total species' population sizes.

Why Is Genetic Rescue Attempted So Infrequently?

There are nine main reasons why genetic rescues have probably been attempted so rarely:

- outbreeding depression/diluting local adaptation
- lack of quantitative data on its effectiveness
- causal relationship between genetic divergence among populations and genetic diversity within them
- overly stringent guidelines
- cost
- risks of disease, pest and parasite spread
- disrupting social systems in some animals
- moving biological material across political jurisdictions
- regulatory barriers

The first four are primarily genetic issues, and are discussed below. The remaining issues require consideration, but are rarely insuperable barriers to gene flow.

Crossing of populations can yield either beneficial or harmful effects. Harmful effects are sometimes found in the F_1, F_2, F_3 or later generation progeny of population crosses (outbreeding depression), and this is widely considered to be an impediment to genetic rescue attempts. However, such harmful effects are predictable, as they are dependent mainly on fixed chromosomal differences (ploidy, translocations, centric fusions and inversions) between the parental populations, or to the parental populations being adapted to different environments. A method for predicting the risk of outbreeding has been devised and shown to have a high degree of accuracy.

A second impediment to genetic rescue attempts has been a lack of quantitative date on the frequency and magnitude of beneficial effects of genetic rescue. In a recent meta-analysis, I found that gene flow into small inbred populations was beneficial in 93% of cases with a low risk of outbreeding depression. The median benefit on composite fitness (fecundity plus survival) in the wild was 148%, and was similar in vertebrates, invertebrates and plants. Benefits were larger in outbreeding than inbreeding species, in wild/stressful than captive/benign environments, for outbred than inbred immigrants, and increased with the magnitude of reduction in maternal and zygotic inbreeding coefficients as a result of crossing. Further, the fitness benefits persisted across generations: they were at least as large in the F_2 and F_3 (and later) generations as in the F_1.

A third impediment is that there is a causal links between genetic differentiation among populations and low genetic diversity within them, a fact that is underappreciated in the field. Thus, the most diverged small populations are likely to be assessed as different (even distinct taxa) and needing separate management, but these are the very ones with the lowest genetic diversity, and in most in need of augmented gene flow.

A fourth impediment to genetic rescue attempts has been overly stringent guidelines that do not adequately account for the urgency of action, or the recent evidence on the magnitude of genetic rescue effects on fitness, or of the ability to predict the risk of outbreeding depression. Frankham *et al.* (2017) contains guidelines that address these shortcomings.

Genetic Management for Global Climate Change

The impact of global climate change on species persistence has recently come to prominence as an issue, and increases the need for genetic management. Average global temperatures are increasing, their variability is growing, and oceans are acidifying as a result of anthropogenic carbon pollution of the atmosphere. Many species are expected to be unable to either evolve to keep up with environmental change, or to move to new locations, or both. Augmentation of gene flow into populations with limited genetic diversity may genetically rescue some of them so that they can evolve fast enough to keep up with climate change.

Genetic augmentation in the face of the predicted rapid environmental change may require sourcing new genetic diversity from other subspecies, or even closely related species. Use of genetic engineering to transfer specific adaptive genetic diversity has much to offer, and is less problematical than crossing genetically diverged species. Insertion of a wheat allele into the American chestnut confers resistance to a fungal disease that almost exterminated the tree species. Further, if species cannot evolve fast enough to keep up with climate change, genetic rescue may assist them to cope with new locations, whether they are translocated to within their historical range, or outside it (assisted colonization).

Aiding in Taxonomic Delineation

A crucial first step in conservation management is to decide which populations require separate, versus combined management. Taxa of conservation concern should be delineated such that species persistence is maximized. If populations and species are not appropriately delineated, genetically differentiated populations (including distinct species) may be inappropriately crossed leading to outbreeding depression. Alternatively, populations that are inbred and depleted of genetic diversity may be inappropriately classified as distinct taxa, causing genetic rescue of small populations to be blocked by regulatory and legal hurdles. Unfortunately, species delineations suffer from a multitude of problems, including:

- no standardized sampling regimes, list of characters to use, or analyses to perform
- many disparate definition of species that often lead to different delineations
- widespread use of diverse methods
- instability of delineations to technological change, especially with some species concepts
- poor repeatability of delineations, especially those done using different approaches
- over-lumping is common (especially in older delineations)
- over-splitting is common currently, and worsening

Frankham *et al.* (2017) have suggested means to overcome these and other problems. For example, they have shown that use of species concepts based on partial reproductive isolation between species, but not within species (such as the biological and the differential fitness species concepts) minimize each of the above problems. Conversely, the use of phylogenetic or general lineage species concepts is not recommended in conservation contexts, as they exacerbate several of the problems.

The most fundamental evidence required for species delineations for conservation purposes comes from crosses between populations carried out in the wild and taken to the F_3 generation. Populations with sympatric (overlapping) or parapatric (abutting) distributions carry out such experiments naturally. However, such evidence is rarely available or feasible to obtain for allopatric (geographically isolated) populations. When crossing data cannot be obtained for populations with allopatric distributions, they recommend the use of integrative taxonomy, the use of multiple lines of evidence in species delineations, especially evidence that predicts reproductive isolation (adaptive differentiation and fixed chromosomal differences). It does not make sense to ignore relevant information when it exists or is practical to obtain. The multiple lines of evidence they envisage being used in integrative taxonomy include:

- genotypes for multiple independent genetic loci
- morphological measurements for multiple characters
- chromosomal assays (essential), preferentially banded karyotypes to detect all relevant types of chromosomal change
- life history and behavioral information
- ecological (habitat) characteristics

The genetic loci used should include both near neutral and adaptive ones. Sole reliance on mitochondrial DNA (mtDNA) data is unacceptable.

Use of Molecular Genetics in Wildlife Forensics

Many species are threatened by poaching, including bears, large cats, elephants, parrots, whales and some plants. Molecular genetic approaches are now used to detect poaching and trade in threatened species and their parts, and to supply evidence for use in court cases against the perpetrators. Species can be identified from DNA in hair, horns, ivory, meat, eggs, turtle shells and plant material. As of 2016, there were 52 laboratories around the world that are members of the Society for Wildlife Forensic Science, including the US Fish and Wildlife Service forensics laboratory in Oregon, United States and the Australian Museum one in Sydney, Australia. Wildlife forensics came to prominence following an evaluation of the source of whale meat on sale in Japan and Korea. Studies using mtDNA revealed that some of the meat was not from minke whales for which Japan undertakes "scientific" whaling, but from protected blue, humpback, fin and Bryde's whales. Further, some "whale meat" was from dolphin, porpoise, sheep and horse. Not only was illegal whaling suspected, but consumers were being deceived. Forensic DNA analyses have also identified poaching in a variety of species including Eurasian badgers, Arabian oryx, and sika deer. Most species level identification in animals is based on mtDNA sequences for cytochrome B (*cyt b*) and cytochrome oxidase 1 (*COI*), while that for plants uses two chloroplast DNA loci, especially the large subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase (*rbcL*) and maturase K (*matK*) (Johnson *et al.* 2014). However, only limited molecular forensic work has been carried out on plants for conservation purposes.

Use of Molecular Genetics to Aid in Understanding Species Biology

For many threatened species, important aspects of their life-histories are unknown and difficult, time consuming and expensive to determine. Molecular genetic tools and chromosomal assays can aid conservation in a wide array of species. The molecular methods of choice have progressed from allozyme electrophoresis that detects heritable differences in proteins beginning in the later 1960s, to DNA methods such as microsatellites in the 1990s, and single nucleotide polymorphisms (SNPs) that rose to prominence post-2000. Extremely powerful genomic methods are now rising to prominence and can involve hundreds of thousands of SNP loci. Activities include the following:

- estimating population sizes using noninvasive sampling
- estimating effective population sizes
- determining demographic history
- detecting and dating population size bottlenecks
- measuring migration and gene flow
- delineating population structure
- identifying source populations to recover endangered species
- detecting introgression and genetic swamping
- detecting secondary contact
- identifying sites for reintroduction
- identifying populations for reintroduction
- determining mating systems
- determining mode of inheritance
- revealing paternity
- identifying founder relationships
- identifying sources of new founders for endangered populations
- sexing of animals
- detecting disease
- analyzing diet

These are detailed by Frankham *et al.* (2010) and Freeland *et al.* (2011), so only a selection of these are addressed below.

Censusing Populations

Many species are difficult and expensive to census, especially nocturnal, fossorial ones. Increasingly they are being censused using molecular methods. For example, critically endangered northern hairy-nosed wombats live in burrows in Epping Forest National

Park in inland central Queensland, Australia. They are nocturnal and difficult to census by trapping (which stresses the animals). Using DNA from hair collected on adhesive tape attached to frames at the opening of their burrows, individuals can be identified from their microsatellite genotypes, sexed using X and Y chromosome markers, and the population size estimated (~196 individuals in the National Park in 2013). Further, microsatellite analyses of museum skins revealed that animals from an extinct population at Deniliquin about 2000 km to the south belonged to the same species, thereby identifying a potential reintroduction site.

Migration and Gene Flow

Migration can be detected following multilocus genotyping (for example with microsatellites) and clustering to identify differentiated populations. Immigrants are individuals that belong to a genetic cluster other than the one they are found in. However, immigrants do not necessarily breed and result in gene flow. Consequently, more advanced analyses on similar data are used to measure gene flow. F_1 and F_2 individuals of crosses between populations will have constitutions lying approximately mid-way between their two parental populations, etc.

Detecting Introgression and Genetic Swamping

Introgression refers to gene flow from another taxon. While low levels of gene flow between species are not uncommon, it is a serious issue when an introduced invasive species hybridizes with a native species and overwhelms its genetic constitution (genetic swamping). This is a major threat in canids, ducks, fish and plants. For example, exotic rainbow trout introduced into the habitats of native cutthroat trout in the Western United States led to hybrid swarms that eventually overwhelmed the local species. Introgression is readily detected by molecular genetic and genomic methods.

Mating Systems

It is important to determine mating systems, as species with different systems have altered characteristics that often require modified conservation management. Mating system refers to the proportion of mating events resulting from self-fertilization e.g., selfing, mixed mating, outcrossing, and self-incompatible. These can be distinguished by genotyping mothers and offspring for genetic markers, such as microsatellites. In selfing species offspring will carry only alleles present in the mother, while outcrossing results in offspring with alleles not present in the mother.

Mode of Inheritance

This refers to deviations from transmission of haploid gametes from parents, such as polyploidy (greater than two sets of each chromosome), haploidy, haplodiploidy (typically diploid females and haploid males, as in Hymenoptera), and asexual reproduction. These are important to distinguish, as they result in different evolutionary genetic characteristics, and may require modified conservation management. For example, haploid species do not exhibit inbreeding depression, while haplodiploid species experience it in females and not males, and the extent is less than in diploids. Many plant species are polyploids, and they are less susceptible to loss of genetic diversity and (probably) inbreeding depression than similar size diploid populations.

If a species is exhibiting asexual reproduction, genotypes for genetic markers (such as microsatellites) in the mother and offspring will be identical, apart from rare mutations. Ploidy differences can be determined by chromosomes counts in dividing cells, or comparing DNA content per cell. For example, the endangered button wrinklewort daisy in Australia has forms that are diploid (22 chromosomes), tetraploid (44), and hexaploid (66).

Paternity

Paternity has been determined in a wide array of species, mainly by comparing microsatellite genotypes of offspring and potential fathers. This has overturned a number of presumptions about mating habits. For examples, most bird species were presumed to be monogamous. However, genotyping of attending "parents" and offspring has revealed that nearly 90% of socially monogamous passerine species have at least occasional extra-pair paternities.

See also: Evolutionary Ecology: Genetic Drift; Units of Selection; Metagenomics; Hardy–Weinberg Equilibrium

References

Frankham, R., Ballou, J.D., Briscoe, D.A., 2010. Introduction to conservation genetics, 2nd edn. Cambridge, U.K.: Cambridge University Press.
Frankham, R., Ballou, J.D., Ralls, K., *et al.*, 2017. Genetic management of fragmented animal and plant populations. Oxford, UK: Oxford University Press.
Freeland, J.R., Kirk, H., Petersen, S., 2011. Molecular ecology, 2nd edn. Hoboken, NJ: Wiley-Blackwell.

Johnson, R.N., Wilson-Wilde, L., Linacre, A., 2014. Current and future directions of DNA in wildlife forensic science. Forensic Science International: Genetics 10, 1–11.
Lacy, R.C., Pollak, J.P., 2014. *VORTEX: A stochastic simulation of the extinction process*, (version 10). Brookfield, IL, USA: Chicago Zoological Society.
Pierson, J.C., Coates, D.J., Oostermeijer, J.G.B., *et al*., 2016. Consideration of genetic factors in threatened species recovery plans on three continents. *Frontiers in* Ecology and Environment 14, 433–440.

Further Reading

Allendorf, F.W., Luikart, G., Aitken, S.N., 2013. Conservation and the genetics of populations, 2nd edn. Oxford, UK: Wiley-Blackwell.
Frankham, R., Ballou, J.D., Briscoe, D.A., 2004. A primer of conservation genetics. Cambridge, U.K.: Cambridge University Press.

Relevant Website

IUCN, 2016. http://www.redlist.org/—IUCN Red List of Threatened Species.

Ecological Health Indicators☆

Paul L Angermeier[1], U.S. Geological Survey, Virginia Cooperative Fish and Wildlife Research Unit, Virginia Tech, Blacksburg, VA, United States
James R Karr, University of Washington, Seattle, WA, United States

Glossary

Adaptive management An iterative approach to managing ecosystems and what humans do to alter ecosystems, whereby management actions are conducted like scientific experiments, and the outcomes are monitored to facilitate learning and improve management.
Biological integrity Wholeness of a living system, including the capacity to sustain the full range of organisms and processes having evolved in a region.
Biota All living organisms, including microbes, fungi, plants, and animals.
Biotic impoverishment Systematic reduction in Earth's capacity to support life.
Ecological health Condition of an ecosystem that is sustainable, that conserves all life-forms, and that society deems acceptable and beneficial.
Ecosystem A community of living things and the biophysical environment in which they interact, including genetic, physiological, morphological, population, energy-flow, and nutrient-cycling characteristics of living things.
Ecosystem service An ecosystem product, process, or condition that benefits or is valued by people.
Environment Surroundings; the complex of physical, chemical, and biotic factors acting upon a living system and influencing its form, function, and survival; the biophysical realities that govern everything on Earth.
Human well-being The state of wellness, happiness, security, and prosperity of a person or community; comprises personal health and safety, access to healthy living conditions, social relations, and personal freedoms.
Multimetric index A numeric measure reflecting combined scores (e.g., the sum) of selected metrics; such indexes are often used to represent ecological, economic, human health, and social well-being.

Introduction

Human societies have a vested interest in tracking the status and trends of their environment. Every human environment comprises physical, chemical, and biological elements, as well as social, cultural, and political elements. A healthy environment—one that nurtures and is sustainable—promotes a healthy society—one that can prosper, is equitable, and conserves other forms of life. Moreover, recognizing and monitoring instructive indicators of a healthy environment enable society to manage itself and its environment in beneficial ways. In this essay, we focus on physical, chemical, and biological environments. We examine key concepts related to indicators of ecological health and why those indicators are important to society. Simply stated, ecological health indicators are signs and symptoms reflecting the health of ecosystems in which human societies live and on which they depend. We contend that thoughtful measurement and analysis of these indicators can inform air, land, and water management in ways that decrease the impoverishment of Earth's living systems and increase human well-being.

What Are Ecological Health Indicators?

Defining Health

People have an intuitive understanding of what it means to be healthy (or unhealthy) and why being healthy is desirable. That said, our understanding of a "healthy" condition is strongly shaped by cultural context and values. For example, a person's body weight may be deemed healthy in one culture but unhealthy in another. A person's physical condition may range from the extreme fitness of an Olympic athlete to near death due to illness. Medical practitioners use established protocols and standards, based on indicators such as cholesterol level and blood pressure, to assess where each person lies along the healthy-to-unhealthy continuum. Other indicators—of diet and lifestyle, for example—may be used to assess a person's likelihood of becoming unhealthy. Further, appropriate standards vary

☆*Change History*: October 2017. Paul L. Angermeier and James R. Karr updated the entire chapter, including Abstract, Keywords, and Glossary; updated the sections "Introduction," "What Are Ecological Health Indicators?" "What Are We Trying to Indicate?" "What Makes a Good Ecological Indicator?" "How Are Ecological Health Indicators Used?" "Conclusions," and "Bibliography"; updated Tables and Figures; and added "Relevant Websites."

This is an update of J.R. Karr, Ecological Health Indicators, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 1037–1041.

[1]The Virginia Cooperative Fish and Wildlife Research Unit is jointly sponsored by the U.S. Geological Survey, Virginia Tech, Virginia Department of Game and Inland Fisheries, and Wildlife Management Institute.

with a person's age or home region. For instance, residents of the high Andes of South America have naturally higher tolerance of low atmospheric oxygen than residents of coastal Florida in the United States.

The health concept is often applied to systems other than individual people. For example, we may characterize the health of an ecosystem, an economy, a farm community, or an international collaboration. In each case, what is viewed as "healthy" reflects the values of those involved, benefits stakeholders, and implies the system is viable and beneficial over the long term. As with human health, assessing the health of such systems requires development of appropriate protocols, indicators, and standards. Below, we adapt this generic concept of health to assessing the condition of ecosystems and their many dimensions. We use "ecosystem" in the conventional sense, referring to a community of living things (including people) and the physicochemical environment in which they interact.

The notion that an ecosystem can range from being healthy to unhealthy is intuitively appealing, although its application has presented challenges because people often disagree on what makes up a "healthy" ecosystem. Experts concerned about undesirable trends in environmental quality have developed well-reasoned rationales for distinguishing healthy versus unhealthy ecosystems. These rationales are founded on the relative importance of the various values—such as intrinsic or utilitarian—through which we judge our relationships with other people and other life-forms. The health of both human and nonhuman life is typically central to how we view ecosystem health. Given some consensus on what constitutes a healthy ecosystem, ecologists can help identify and calibrate useful indicators.

A well-established frame of reference for the condition of an ecosystem is its natural state, which is shaped by natural evolutionary processes that align the adaptations of a biota with prevailing environmental conditions. Many people view natural ecosystems as inherently healthy, realizing, for example, that the flora and fauna of a remote desert are healthy, as well as distinct from those of a healthy rainforest. If people converted a rainforest to a desert (or vice-versa), however, the result typically would be judged ecologically unhealthy.

To evaluate a place's ecological condition, we must first identify and define natural attributes useful as measures of system condition. Physicochemical and biological attributes are chosen to collectively represent that place's ecological integrity (i.e., state of being whole or unimpaired)—a benchmark for measuring condition and judging health. The word *integrity* refers to the complex systems that evolved at a place, with proven capacity to persist in the associated physical, chemical, and biological environment. The evolutionary ties between natural assemblages and their home places validate integrity as a benchmark for assessing the ecological condition of places altered by human activities.

Naturalness also underpins biological integrity, a concept that is widely used to assess ecological condition. If a place has biological integrity, a multiscale, integrated, adaptive biological system is present, including the full range of evolved parts (e.g., genes, species, and assemblages) and processes (e.g., dispersal, biotic interactions, and energy flow) expected in the absence of substantial recent human influence. This definition recognizes several key complexities. First, ecosystems encompass spatio-temporal scales from individuals to landscapes, any of which can inform an assessment of ecological condition. Second, living systems include items we can count (the parts) plus the processes that generate and maintain them; any of these can be useful indicators. Third, living systems are embedded in dynamic evolutionary and biogeographic contexts that influence and are influenced by their physical, chemical, and biological environments. An important upshot of this definition is that the natural conditions embodied in integrity vary through time and space within some normal range. Understanding this range is essential to accurately assessing ecological condition.

Maintaining *ecological health* is the goal for sites managed intensively for human benefit—as for crop cultivation, tree harvest, urbanization, recreation, and the like (Karr, 1996). Protection of integrity in an evolutionary sense is unrealistic as a goal for such places; instead three criteria for their management are important: Ecological damage from intensive use should not threaten the long-term beneficial use of the area. Intensively managed areas should not degrade or damage other areas beyond their boundaries, such as downstream or downwind. Finally, what is viewed as "healthy" for a given site should reflect the values of those involved—the stakeholders at relevant spatial and temporal scales.

Identifying Indicators

Ecological indicators can be calibrated to measure deviations from integrity, including deviations that are acceptable (healthy) or unacceptable (unhealthy). The acceptability of any given deviation is negotiated and determined by society via its environmental policies and regulations. Compared with conditions embodied in integrity, conditions deemed unacceptable are more fluid, reflecting prevailing value systems. Further, what is unacceptable in one place may be acceptable in another, depending on societal goals. For example, the health of a wilderness forest and an industrial timberland might be assessed via different ecological criteria or indicators, even though they are both assemblages of trees with shared natural conditions. Even for intensively used ecosystems, however, we can set limits for the acceptability of ecological deviations. Consider a farm. If practices there damage the land for future farming or harm nearby waterways or people downstream, those practices and the resulting ecological conditions may be considered unhealthy. Regardless of which conditions are societally acceptable, ecological indicators can be calibrated objectively to distinguish the healthy from the unhealthy.

Numerous ecological indicators are in use, typically chosen to match the specific ecosystem and issue of interest (Feld *et al.*, 2010; O'Brien *et al.*, 2016; Yu *et al.*, 2017). Some indicators reflect conditions at a given point in time (e.g., pH, number of species), while others reflect processes over a given time frame (e.g., annual soil erosion, population growth rate). Although many potential physical, chemical, and biological indicators can be measured, the biological assemblages that persist in a place provide the most integrative and

instructive indicators of prevailing ecological conditions. Monitoring such assemblages is crucial to understanding the full array of ecological consequences produced by human actions or natural events. Without regular biological report cards, humanity is ill-equipped to protect ecologically intact places, restore degraded places, or make informed decisions about how to manage natural resources. Thus, development of reliable, instructive ecological indicators is vital to society.

Accurately assessing ecological condition requires attention to the key factors and processes that drive ecosystem dynamics. This knowledge may be synthesized into a conceptual model of how factors and processes produce ecological outcomes (Lindenmayer and Likens, 2010), including biological responses such as changes in the behavior of an organism, changes in the abundance of a population, or shifts in the interactions among species. Such responses are important links to how people value ecosystems and view ecological health. Knowledge of species' life histories and habitat use is crucial to such models. Because ecosystems are dynamic, ecological assessments involve teasing the signals of interest from background environmental noise. Accurately interpreting such signals requires care in designing a monitoring protocol and selecting indicators to monitor. Translating monitoring results into management action further requires setting up thresholds for action that relate selected indicators to societal goals. For example, suppose monitoring shows that the ecological condition of a stream is unhealthy because of excessive sediment and bacteria attributed to livestock; that is, the stream's condition is unacceptable to local or downstream users who would otherwise benefit from the stream's flow. Managers might implement practices meant to improve the stream's condition and restore beneficial uses to stakeholders, such as restricting livestock access and planting riparian vegetation.

What Are We Trying to Indicate?

The short answer to this question focuses on ecological conditions of interest to people. Fully answering the question, however, is complicated by the fact that indicators can serve many purposes, which differ from case to case. In each case, the purpose of monitoring ecological indicators reflects a subset of their many links to human well-being, plus the kinds of knowledge needed to make wise decisions about human actions with environmental consequences.

Ecological Indicators Have Long Been (and Still Are) Relevant

Living organisms, including people, experience and assess their environment through a cacophony of sensory information. Only some of that information is useful to stimulate physiological and behavioral responses that may enhance one's livelihood. For example, accurately characterizing important threats (e.g., predators) and opportunities (e.g., food) is crucial. Organisms selectively use available information as indicators of the condition of their world. If they respond appropriately to relevant indicators, ecological and evolutionary success typically follows.

The indicators deemed important to people have changed dramatically over history. Early humans, knowingly dependent on nature for their well-being, no doubt tracked indicators such as weather, plant fruiting, and animal migration. Technological advances—such as using tools, controlling fire, and tending crops—gave humans more control of their environment. Increased proficiency in procuring food facilitated rapid growth in human populations and technology. Communication technologies such as written language (and, eventually, electronic media) greatly expanded the scope of people's sensory capability, as well as the kinds of knowledge that were valued. Although some 21st-century human communities still depend directly on natural systems, most people are far less aware of any such dependence, now focusing on other kinds of indicators (e.g., economic) to assess and plan for their well-being. In addition, when environmental problems arise, we are more inclined to try to redesign an ecosystem (via engineering) than to adapt to prevailing ecological conditions.

Because modern people are far less connected to nature than their hunter-gatherer ancestors, some believe that tracking the status and trends of ecological indicators is far less important to their well-being than tracking economic indicators. Others, however, believe that such a view is shortsighted and risky, because ecosystems—even unnatural ones—are still the foundation of life support on Earth and contribute greatly to human quality of life. Only by carefully monitoring the condition of nonhuman living systems can we assemble the knowledge needed to manage the collective consequences of our actions for the continued benefit of people and Earth's biota. This management challenge is far more complicated than our ancestors' challenge to find food and mates and to escape predators.

Since the mid-20th century, concern about environmental trends—particularly, widespread chemical pollution—sparked the development of a variety of indicators to assess how environmental regulations were being enforced. Initially, such indicators described regulatory activity and pollution loading rather than the broader condition of ecosystems, thereby reflecting the dominance of toxicological approaches to solving environmental problems (e.g., see Karr, 1991). These indicators fall into five classes:

- Administrative indicators, such as numbers of pollution-discharge permits and enforcement actions.
- Technological indicators, such as numbers of implemented "best management practices" (e.g., storm-water detention ponds, conservation tillage).
- Stressor indicators, such as measurements of pollutant loads.
- Exposure indicators, such as measured or modeled concentrations of pollutants in water, soil, or air.
- Response indicators, such as measures of biological condition, including taxa richness, population abundance, production, and multimetric indexes that integrate multiple biological attributes.

Exposure indicators and response indicators gained use as scientists, environmental managers, and regulators recognized that the other indicators neither assessed ecological conditions nor ensured that legally mandated goals (e.g., protecting the integrity of living systems) were attained. Only exposure and response indicators merit use as ecological indicators.

Causes and Consequences of Ecological Change

Ecological change is continual, but much of it goes largely unnoticed by people. Of course we notice infrequent but dramatic changes, such as those caused by volcanoes, tornadoes, epidemics, or wars, but we may pay little attention to gradual changes, such as the daily shifts accompanying transitions between seasons. We can observe change at many spatial scales (e.g., local, regional, global) and temporal scales (e.g., daily, annual, millennial). Change at any of these scales can be induced by natural forces, human actions, or some combination. Regardless of cause, it seems prudent to keep track of ecological changes that may influence human well-being and to measure indicators that enable us to manage the outcomes. Over the long term, documenting conditions before important ecological changes occur can reveal factors that make ecosystems more or less sensitive to change. History repeatedly confirms the tragic consequences of ecological ignorance: over the past 4500 years, dramatic societal collapses have occurred in Mesoamerica, Mesopotamia, and South Asia when ecological indicators were ignored (Diamond, 2005).

Human action is now the principal driver of environmental change on Earth, rivaling astronomical forces in its impact and prompting many environmental experts to call the present epoch the Anthropocene (Waters *et al.*, 2016; Johnson *et al.*, 2017). People profoundly shape ecosystems—at all spatial and temporal scales—and ecosystems' capacity to support life. Whether a change is purposeful and obvious (e.g., a prairie converted to a cornfield or a wetland drained to build houses) or nearly invisible (e.g., altered biotic communities after a dam is built), when such changes happen widely across regions, the consequences add up. The upshot is that human uses of Earth's resources are causing worldwide changes in living systems, creating an unprecedented level of biotic impoverishment (Woodwell, 1990). These changes are manifest from genetic to ecosystem levels because of diverse human actions that degrade or destroy habitat; spread diseases, invasive species, or pollutants; and alter climate (**Table 1**; Scheffers *et al.*, 2016; Chu and Karr, 2017).

Loss of biodiversity is a common consequence—as well as indicator—of rapid or extensive ecological change. Although extinctions of species are inevitable, even under natural conditions, current human-induced extinction rates of mammals, birds, and fishes are at least 100 to 200 times greater than in prehistoric times (Burkhead, 2012; Johnson *et al.*, 2017). Anthropogenic changes to ecosystems may also depress the production of species with commercial or recreational value, as often happens in overharvested forests and fisheries. Moreover, all forms of biodiversity loss represent significant costs to society (Cardinale *et al.*, 2012). Monitoring trends in biodiversity can help minimize such societal costs.

Some consequences of ecological change are more apparent as shifts in ecosystem processes than as shifts in particular species. Myriad organisms—from megafauna to microbes—form the building blocks of ecosystems. Their interactions, which play out at multiple spatial and temporal scales, control the dynamic mixes and flows of air, water, nutrients, and energy that are essential to life on Earth. Although we may not readily observe such processes, we do see some of the outcomes that are enormously beneficial to human society. For example, biota play crucial roles in purifying air and water, forming soils, pollinating crops, and regulating local climate and floods. Collectively, these processes make civilization possible. Conversely, widespread human actions that alter ecological conditions have the potential to disrupt our life-support systems. Keeping track of well-chosen ecological indicators can help ensure that we recognize a dangerous aberration before an ecosystem is irreparably damaged.

Links between Ecological Health and Human Well-Being

In addition to revealing the condition of ecosystems, ecological health indicators also reveal information pertinent to human health and well-being (Chivian and Bernstein, 2008; Gottdenker *et al.*, 2014). Human well-being—physical and mental—is closely tied to ecological conditions, often in subtle or complex ways. Biotic impoverishment often precedes or accompanies diminished human health. Perhaps best known are the thousands of harmful chemical by-products of various technologies, including heavy metals, pesticides, and endocrine disruptors. Most of these find their way into our air, water, or food, with both predictable and unpredictable consequences for personal health. The plants and animals that live in the same ecosystems we do are often sensitive to the same contaminants and pathogens we are, which makes them ready-made indicators of conditions that pose risks to human health. Moreover, research increasingly shows how much the naturalness of our daily environment—such as visible water, trees, and sky—enhances our mental health and capacity to handle stress (Sandifer *et al.*, 2015). Medical and ecological experts alike conclude that healthy ecosystems promote healthy people.

Beyond the effects of pollutants on personal health, human-induced biotic impoverishment may also threaten human well-being more generally. For example, deforestation can accelerate soil erosion and undermine the capacity of an ecosystem to produce crops (Breure *et al.*, 2005); overharvesting fish and wildlife populations can destabilize biotic communities and deplete food resources for people; and introducing new weeds and pests can reorganize ecosystems and impair production of farmed or wild foods. Although the long-term outcomes of such changes for specific people are difficult to predict, it is certain that shifts in ecosystem conditions broadly affect the well-being of many.

Countering industrial societies' overemphasis on economic indicators, many current conceptions of sustainability recognize that human well-being and progress depend on three cooperative systems: ecological, social, and economic (Kates *et al.*, 2005; Wu,

Table 1 The many faces of biotic impoverishment

Indirect depletion of living systems through alterations in physical and chemical environments

1. Degradation of water (redirected flows, depletion of surface and groundwater, wetland drainage, organic enrichment, destruction and alteration of aquatic biota)
2. Soil depletion (destruction of soil structure, erosion, salinization, desertification, acidification, nutrient leaching, destruction and alteration of soil biota)
3. Chemical contamination (land, air, and water pollution from pesticides, herbicides, heavy metals, and toxic synthetic chemicals; atmospheric ozone depletion; ocean acidification; fish kills; extinctions; biotic homogenization and biodiversity loss; bioaccumulation; hormone disruption; immunological deficiencies; reproductive and developmental anomalies; respiratory diseases; intergenerational effects)
4. Altered biogeochemical cycles (alteration of the water cycle; nutrient enrichment; acid rain; fossil fuel combustion; particulate pollution; degradation of land and water biota; outbreaks of pests, pathogens, and red tides)
5. Global climate change (rising greenhouse gas concentrations, altered precipitation and airflow patterns, rising temperatures, effects on individual and community health, shifts among and within global ecosystems)

Direct depletion of nonhuman life

1. Overharvest of renewable resources such as fish and timber (depleted populations, extinctions, altered food webs)
2. Habitat fragmentation and destruction (extinctions, biotic homogenization, emerging and reemerging pests and pathogens, loss of landscape mosaics and connectivity)
3. Biotic homogenization (extinctions and invasions, lost biodiversity among food crops and livestock)
4. Genetic engineering (homogenization of crops, antibiotic resistance, potential extinctions and invasions if genes escape, other unknown ecological effects)

Direct degradation of human life

1. Emerging and reemerging diseases (occupational hazards, asthma and other respiratory ills, pandemics, Ebola, AIDS, hantavirus, tuberculosis, Lyme disease, West Nile fever, chikungunya, Zika virus disease, antibiotic resistance, diseases of overconsumption and stress, altered human microbiomes)
2. Loss of cultural diversity (religious wars and genocide, loss of cultural and linguistic diversity, loss of knowledge)
3. Reduced quality of life (malnutrition and starvation, failure to thrive, poverty)
4. Environmental injustice (environmental discrimination and racism; economic exploitation; growing gaps between rich and poor individuals, segments of society, and nations; environmental refugees; gender inequities; trampling of the environmental and economic rights of future generations)
5. Political instability (civil violence, especially under intransigent regimes; resource wars; international terrorism; increased number of refugees)
6. Cumulative effects (environmental surprises, increased frequency of catastrophic natural events, boom-and-bust cycles, interactions between disease and biodiversity, collapse of civilizations because of environmental degradation)

Modified from Karr, J.R., Chu, E.W., 1995. Ecological integrity: Reclaiming lost connections. In: Westra, L., Lemons, J. (Eds.), Perspectives in Ecological Integrity, pp. 34–48. Dordrecht: Kluwer Academic.

2013). This thinking presumes that human well-being is best served by promoting a "triple bottom line," which can be tracked by monitoring indicators of ecological, social, and economic well-being. Clean water and green space are common landscape features that contribute to a person's ecological well-being. As is the case for ecological health, human well-being can be measured at a range of spatial scales, such as a community or a nation.

Scientific understanding of nature's contributions to human well-being has expanded rapidly over the past three decades, with contributions commonly grouped into four classes of "ecosystem services": provisioning, regulating, cultural, and supporting. Provisioning services—the most readily recognized—include tangible goods such as water, food, and lumber. In contrast, regulating services derive from ecological processes that are often less apparent, including the purification of air and water and the regulation of climate and floods. Cultural services, which include aesthetic, recreational, educational, scientific, and spiritual benefits, are also largely intangible. Finally, supporting services include ecological processes such as primary production and soil formation, not all of which benefit people directly but which do underpin the maintenance of tangible ecological goods and other services. Many ecological health indicators inform us about whether ecological changes are likely to affect the capacity of ecosystems to deliver valued services.

Perhaps the most comprehensive assessment of ecosystem services—the Millennium Ecosystem Assessment—was completed in 2005 (**Fig. 1**). It synthesized input from hundreds of scientists and catalyzed discussion among policymakers regarding how to manage ecosystems sustainably. A key strength of this approach is that it provides an integrative framework to account for the many benefits provided by healthy ecosystems, well beyond those considered in conventional economic analyses.

In economic terms, human well-being can be viewed as forms of wealth stemming from various forms of capital—including but not limited to financial capital (Hancock, 1997). How we value different sources of wealth strongly influences how we value healthy ecosystems and their indicators. For example, ecological economists presume the human economy is a subsystem of the global ecosystem and that much of our wealth ultimately depends on the persistence of the natural capital that underpins healthy ecosystems. In contrast, neoclassical economists (who prevail in most policy arenas) presume that ecosystems are subsets of the

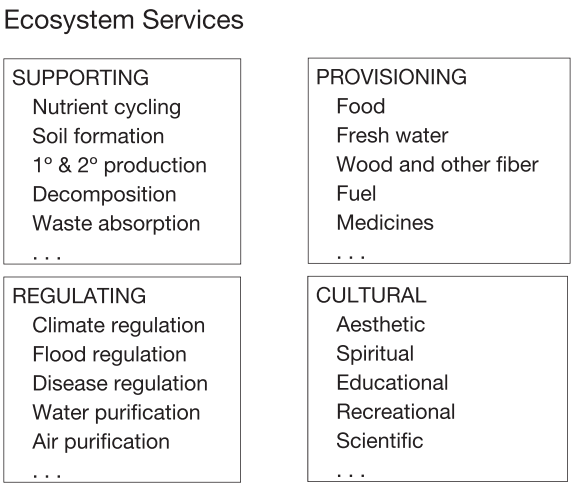


Fig. 1 Four major categories of ecosystem services with examples from each category. Modified from Millennium Ecosystem Assessment (2005). Ecosystems and Human Well-Being: Biodiversity Synthesis. Washington, DC: World Resources Institute, with permission from United Nations Environment Programme.

human economy and that natural capital can be traded indefinitely for other forms of capital, such as financial capital, without dire long-term consequences. Subscribing to one of these disparate worldviews can profoundly affect how an economist sees the relevance of ecological health indicators to socioeconomic policy. Because no one can be certain of the current or future value of natural capital, directly monitoring ecological indicators linked to human well-being helps ensure that ecosystems continue to provide crucial goods and services for people.

What Makes a Good Ecological Indicator?

The many efforts to develop and apply ecological health indicators over the past 40 years have taught us much about which indicators are or are not useful (Table 2). Most important, the knowledge gained from monitoring ecological indicators adds to understanding ecosystems and guides management actions. For example, data on a suite of indicators can help a manager interpret where an ecosystem lies along a gradient of biological condition; the manager can then choose specific actions, such as protection or restoration, aligning with societal goals. Overall, ecological indicators are simplified lenses through which we can better understand complex ecosystems and use that understanding to improve our management of people to protect those ecosystems. Although the most useful indicators tend to be case specific, several features apply generally.

Ecological indicators are most useful when they are readily measurable, integrative, ecologically and societally relevant, interpretable, cost-effective, anticipatory, measured at appropriate spatial and temporal scales, and able to detect status or trends. Further, indicators are most informative when they provide data that are quantitative, statistically robust, reliable, and comparable and when they can be used to diagnose problems. Data collection for ecological monitoring is typically designed to do no long-term harm or substantially alter the ecosystem being assessed. Ideally, indicators are sensitive to a broad range of known stressors and likely to be sensitive to unknown or as yet unidentified stressors. Most of these features require that indicators make sense in the context of empirical and conceptual ecological frameworks reflecting present scientific understanding of real ecosystems. Combining indicators that represent individual, population, assemblage, and landscape levels of ecological organization enriches our ability to understand the many dimensions of ecosystems. Depending on the ecological impact(s) of particular concern, indicators of genetic, physiological, morphological, population, community, energy-flow, nutrient dynamics, and landscape conditions may be especially informative. Multiple complementary indicators are often combined into composite ecological indexes, analogous to the widely familiar economic indexes based on multiple economic indicators (such as the consumer price index).

How Are Ecological Health Indicators Used?

Protocols for Measuring Ecological Health

The utility of the integrity concept is reflected in its widespread appearance in policy and planning documents. The phrase "biological integrity" was first used in 1972 to establish the goal of the US Clean Water Act for the nation's waters. Thanks to the conceptual and management success of Clean Water Act programs, maintaining biological or ecological integrity became a primary goal in other environmental policies, including the Canada National Parks Act, the US National Wildlife Refuge System

Table 2 Sample indicators for assessing ecological health. Indicators may target specific media (such as air, land, or water); habitat types (such as forests, grasslands, or streams); or taxa (such as plants, fishes, or insects). Indicators can be used individually and to construct multimetric indexes. Although not comprehensive, this table illustrates the breadth of ecological indicators in use or potentially usable

Indicators for nonhuman biological systems
Individual and population levels: body burden of toxic contaminants; condition index (ratio of a fish's weight to its length); population size and density, especially for threatened and endangered species; population sex and age structure; recruitment and population growth rate; shifts in phenology; migration rate; harvest rate by sport, commercial, and subsistence interests; rates of fish stocking, including supplementation; genetic diversity within a population; rates of spread by nonnative species; primary and secondary production; standing stock biomass; presence and condition of iconic species, such as bald eagles or giant pandas
Community level: species or higher-level taxonomic richness; proportional abundance of functional groups, such as predators, grazers, or carrion feeders; shifts in community composition and proportional abundances (biotic homogenization, tropicalization, desertification); biomass, primary production, and secondary production; standing stock of timber; forest regrowth rates; total basal area of a timber stand; persistence of coevolved species groups, such as plants and their pollinators and seed dispersers; persistence of predator-prey relationships; rates of pest and disease outbreaks; shifts in presence or abundance of tolerant or intolerant species; toxin levels in plant and animal tissues across taxa
Ecosystem level: altered rates of biogeochemical processes; altered rates of nutrient cycling; altered rates or pathways of energy flow; embodied energy (i.e., energy), solar energy required to produce goods and services; concentrations of key chemicals in ecological interactions; trophic state of lakes; chlorophyll concentrations; shifts in social or economic outcomes stemming from ecological relationships
Landscape level: rates of habitat fragmentation and destruction; patch size and patch distribution; width of riparian vegetation along streams or lake margins; measures of fragmentation, aggregation, and connectivity; shifts in extent, structure, or stability of habitat mosaics; abundance and type of soil organic matter; extent and stability of major communities, such as coral reefs, forests, or wetlands
Indicators of human well-being
Production systems (e.g., agriculture, forestry, marine and freshwater fisheries and aquaculture): proportions of organic versus conventional agriculture; proportions of grass-fed versus grain-fed livestock; spatial extent of bottom-trawling versus nontrawling fisheries; extent and intensity of aquaculture and its human health consequences; ratio of harvest of fish or other resources to annual production; production of genetically modified organisms and associated risks; social and economic stability of production
Human health: human body burdens of toxins; rates and extent of disease transmission in people (and wildlife); rates of malnutrition, overnutrition, and related diseases; rates of disease stemming from polluted air, land, or water; rates of mental illness; rates and extent of spread by pathogens resistant to antibiotics and other drugs; rates of emergence and reemergence by diseases associated with habitat alteration
Cultural ecology: rates of cultural diversity loss; rates of loss of specific cultures or traditional ecological knowledge, including languages; degree of social or economic isolation; extent and rates of environmental injustice, including toxic exposures and access to potable water; measures of ecosystems' ability to supply recreational opportunities outdoors; availability or consumption of wild foods
Abiotic indicators of ecological condition
Chemical: concentrations of contaminants or toxins in air, water, and soil; nitrogen and phosphorus concentrations in water bodies; rates of soil salinization from irrigation; rates of greenhouse gas emissions; application rates of fertilizers or pesticides in urban and agricultural areas
Physical: physical structure of an environment as determined by substrate type and topography, geology, hydrology, or biota (e.g., coral reefs, marine and terrestrial forests, grasslands, or deserts); rates of stream or river flow; frequency, duration, and magnitude of floods; measures of stream-channel morphology or substrate composition; degree of seasonal variation in surface and subsurface water flows; temporal shifts in stream channel morphology; rates of sediment loading into water bodies; soil mineral and organic content, structure, and tilth; soil generation and erosion rates; atmospheric structure and stability; frequency, duration, and magnitude of severe storms; diurnal and seasonal variation in water temperature; changes in sea levels; glacial melt rates; permafrost extent

Improvement Act, the Canada–US Great Lakes Water Quality Agreement, the European Union's Water Framework Directive, and the Earth Charter.

The integrity concept has catalyzed development of many programs that use ecological indicators to monitor and assess the condition of living systems. These programs recognize that wild living things are ideal indicators of ecological health and sensitive sentinels for human health. For example, healthy fish and insect assemblages are good indicators of healthy water bodies, which pose fewer health risks to people than do degraded water bodies. This relationship underpins the programs that use biotic assemblages to monitor the quality of waters in the United States and around the world. Similarly, the US Forest Service uses the ecological integrity concept to guide land-use planning and conservation in national forests, and the US National Park Service incorporates the concept into inventory and assessment programs (Wurtzebach and Schultz, 2016).

Ecological indicators are widely used to summarize technical information about the condition of ecosystems in a form nonexperts can readily understand. For example, the Heinz Center has reported (in 2002 and 2008) on the status of and trends in ecosystems of the United States. The center synthesized input from more than 300 scientists, who compiled data on 108 indicators for six ecosystem types (coasts and oceans, forests, farmlands, fresh waters, grasslands and shrublands, and urban and suburban areas) at local, regional, and national scales. The Heinz indicators are organized into four groups—system dimensions, chemical and physical condition, status of biological components, and ecosystem services (**Fig. 2**). Each group comprises more than 20 indicators. The resulting assessments are widely used by policymakers to understand the condition and use of ecosystems.

Composite ecosystem "report cards" are commonly used to broadly summarize ecological conditions for cities, states, estuaries, river basins, and islands (Harwell *et al.*, 1999; see also http://ian.umces.edu/ecocheck/report-cards/). Report cards simplify the complex relations among environmental goals, ecosystem stressors, and ecological endpoints. Report cards provide timely assessments based on selected indicators, usually presented as maps and tables derived from scientific studies. Such assessments are accompanied by numeric or letter grades, allowing for quick and straightforward understanding by nonexperts.

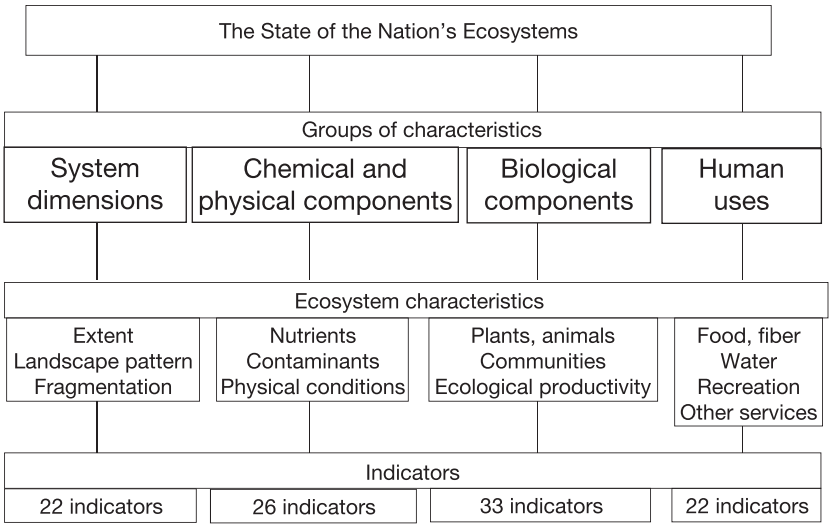


Fig. 2 Organization and examples of ecosystem characteristics, with number of indicators examined, in the *State of the Nation's Ecosystems* report. Modified from Heinz Center (The H. John Heinz III Center for Science, Economics and the Environment) (2002 and 2008). The State of the Nation's Ecosystems: Measuring the Lands, Waters, and Living Resources of the United States. New York: Cambridge University Press. Reprinted from Heinz Center 2002.

Ecosystem scientists and managers need standardized methods for sampling biota, analyzing data, and assessing ecological conditions. Further, integrative information from individual, population, assemblage, and landscape levels is needed to provide an accurate assessment of how far a system's condition diverges from integrity. A multimetric measurement approach—the index of biological integrity (IBI; Karr, 1991, 2006)—has been widely used since the 1980s to facilitate such assessments. IBIs have been applied in contexts such as resource management, engineering, and public policy, in the United States and at least 85 other countries (J.R. Karr, unpublished). Initially, IBIs were developed to assess streams and rivers, but the conceptual framework has since been applied to other environments (wetlands, lakes, reservoirs, coral reefs, estuaries, riparian corridors, forests, grasslands, sagebrush steppes, and caves) and to diverse taxa (vascular plants, algae, diatoms, bacteria, benthic invertebrates, fishes, zooplankton, crayfishes, freshwater molluscs, amphibians, coastal amphipods, nematodes, and birds; J.R. Karr, unpublished). Many states incorporate IBI values, based on fish or macroinvertebrate communities, into water quality standards (O'Brien *et al.*, 2016). Such applications are key components of implementing the US Clean Water Act and the European Union's Water Framework Directive.

IBI metrics are carefully chosen to collectively characterize human impacts on ecosystems. Each metric represents a verifiable biotic response, analogous to an organism's response to varying dosages of a toxin. Because their values reflect cumulative impacts of all human activities in an ecosystem, IBIs represent ecological dose-response relations (**Fig. 3**). Before inclusion in an IBI, individual metrics are validated empirically to ensure they (1) are ecologically meaningful, (2) increase or decrease as human influence increases, (3) are sensitive to a range of stresses, (4) distinguish stress-induced variation from natural and sampling variation, (5) are relevant to societal concerns, and (6) are scientifically robust, yet easy to measure and interpret. The ecological breadth and depth of metrics make the resulting IBIs more informative about ecosystem condition than are simpler chemical measures.

Specific IBIs vary in the number of metrics included and how numeric scores are assigned, but all IBIs share certain features. Scores for metrics are summed to yield a single index value. Metric and index scores enable comparisons between a locale's present condition and what integrity would look like there. Low scores indicate altered conditions, whereas high scores indicate natural conditions. The original IBI (based on stream fishes) comprises 12 metrics, each of which is assigned a score of 5, 3, or 1. Thus, this index ranges from 60 to 12. Similarly, the benthic invertebrate IBI (B-IBI) for streams comprises 10 metrics, so this index ranges from 50 to 10. In Washington State, monitoring results based on B-IBI are instrumental in enforcing the US Clean Water Act and in managing Pacific salmon (**Table 3**). Here, a stream is considered impaired when its B-IBI falls below 35, indicating it cannot support a salmon population.

Using Health Indicators in Ecosystem Management

Using health indicators to inform management of ecosystems has parallels in how people manage personal health. The ecological analog of a human patient might be a wild population, a local assemblage of many species, or the global biosphere. Indicators could be developed to assess ecological health at any of these scales. The ecological analog of a medical practitioner is a natural resource manager, such as someone working at a regional, state, or federal agency or nongovernmental organization. Another parallel is that regular "wellness exams" on the condition of an ecosystem can alert us to emerging problems and help us choose

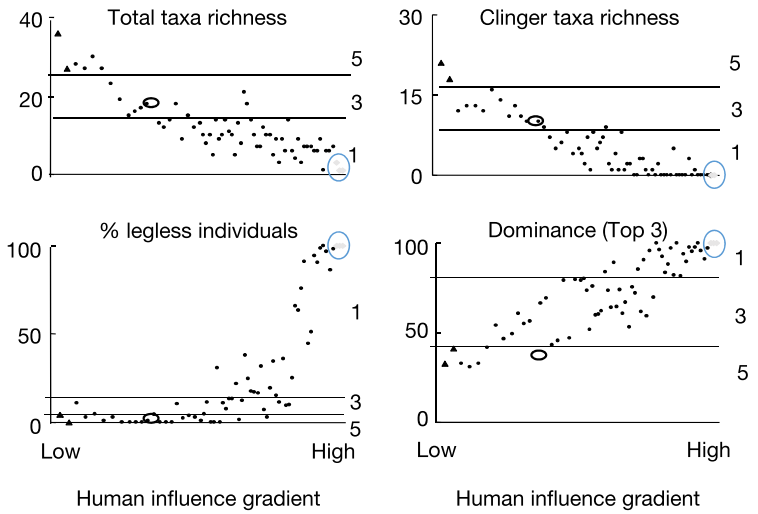


Fig. 3 Examples of dose-response relations for a hypothetical four-metric IBI based on data from Japanese streams. The four metrics are benthic invertebrate taxa richness (top left); taxa richness of clinger taxa (top right); percentage of individuals in sample that are legless, such as snails and worms (bottom left); and dominance, measured using the three most abundant taxa (bottom right). Black triangles represent values from reference streams considered to be minimally influenced by humans. Gray triangles (enclosed in blue ovals) represent values from highly disturbed streams. Black ovals represent a sampling site that has an IBI value of 16 (3 + 3 + 5 + 5) out of a possible 20. In each graph, high-quality sites appear on the left side and low-quality sites appear on the right. Modified from Karr J.R. and Chu E.W. (1999). Restoring life in running waters: Better biological monitoring. Washington, DC: Island Press. Reproduced by permission.

Table 3 Benthic invertebrate index of biological integrity (B-IBI) as applied in US Pacific Northwest streams. Index scores, which can range from 50 to 10, are partitioned into five levels designed to associate regulatory language with specific biological conditions. (Developed in consultation with Ed Chadd, Streamkeepers of Clallam County, Washington, USA.)

Score	*Regulatory language*	*Biological condition*
50–46	Healthy	Ecologically intact, supporting the most sensitive life-forms
44–36	Compromised	Showing signs of degradation: impacts expected to one or more salmon life stages; loss of some intolerant, long-lived, or other taxa (e.g., stoneflies)
34–28	Impaired	Ecosystem parts and processes demonstrably impaired; cannot support self-sustaining salmon populations
26–18	Highly impaired	Highly inhospitable to many native fishes and invertebrates
16–10	Critically impaired	Cannot support a large proportion of native life-forms; only the most tolerant taxa present

corrective actions. Effective treatment of any health problem requires basic understanding of relations between causes and symptoms (i.e., indicators). Two other lessons are important: (1) Preventively treating the causes of poor ecological health is more cost-effective than reactively treating the symptoms, and (2) evaluating the effectiveness of treatments to promote health goals is crucial to avoid repeating the same mistakes in the future.

Choosing how to manage the effects of human actions on ecological conditions requires analyzing trade-offs between benefits and costs and how these are distributed among stakeholders. Some trade-offs involve local stakeholders in the near term, whereas others involve stakeholders living in other regions or who will live in the future. For example, converting a forest to a cornfield diminishes the capacity of the converted land to purify water, regulate floods, and maintain biodiversity but enhances its capacity to produce food. The costs of increased water pollution and flooding are mostly borne by people living literally downstream, whereas the benefits of increased food production are enjoyed by local farmers, as well as distant consumers. Trade-offs can be difficult to sort out—and agree upon—if all ecosystem services and stakeholders are to be considered equitably. Additionally, management choices vary among places and times. In national parks, for example, we choose to emphasize aesthetic and spiritual benefits along with widespread ecological services, while in commercial forests we choose to emphasize timber production. As societal values shift, choices that once made sense may be altered to align with current notions of human needs or well-being. In any case, analyses of changes in ecological condition (divergence from integrity) and the delivery of ecosystem services—both informed by appropriate ecological indicators—can summarize trade-offs and inform management choices.

Successfully managing ecosystems is complicated by the fact that we often know little about how ecosystems work, and the problems we try to solve may be too "wicked" to have straightforward solutions (DeFries and Nagendra, 2017). Managers can

become more effective by embracing structured, adaptive approaches, which use the knowledge gained from monitoring to guide future management actions (Conroy and Peterson, 2013). Adaptive management emphasizes "learning by doing," where learning is not achieved by simple trial and error but instead by careful, hypothesis-driven monitoring and by treating management actions as experiments (Murphy and Weiland, 2014). Properly chosen ecological indicators are crucial to adaptive management's four phases: plan, do, check, and adapt. Further, an adaptive approach allows an indicator to be chosen on the basis of current knowledge, then revisited as more is learned. Sometimes adaptive management is combined with a process called "ecosystem-based management," in which nonscientist stakeholders help develop indicators. Although this approach commonly promotes ecological understanding among stakeholders, outcomes for improving ecological health have been mixed (Layzer, 2008). Unfortunately, although monitoring is essential to adaptive management, insufficient monitoring—especially of biological indicators—too often limits the ability of managers to improve ecological health.

Conclusions

Whether the issue is endangered species, threatened water supplies, or emerging human diseases, understanding patterns, trends, and causes of biotic impoverishment in Earth's dynamic ecosystems is essential to informed, effective policy. Monitoring indicators of ecological health helps improve our understanding of ecosystems and diagnose the causes of ecological degradation. Ecological indicators inform scientists, the public, and policymakers about the environmental, social, human-health, and economic consequences of ever-changing ecological conditions. Sound environmental monitoring based on scientifically well-chosen ecological indicators is the best means of supplying concrete evidence for the importance of ecological health—no matter who may try to deny or discount it. When put to use, knowledge of ecological health has enormous potential to enhance human well-being.

Some of the most serious and complicated challenges now confronting humanity come from the impoverishment and declining condition of Earth's living systems. No single set of indicators—economic or ecological—will be enough to resolve the resulting debates and social dislocations. Ecological indicators can and must play an irreplaceable role in informing those debates and turning decisions toward protecting Earth's ecosystems and the human society that depends on them. Achieving a more sustainable human society requires greater ecological awareness and stewardship, as well as greater commitment to conserving all of Earth's biota well beyond the 21st century. We need ecological health indicators—widely measured, analyzed, and discussed—to reach these goals.

See also: Aquatic Ecology: Ecosystem Health Indicators—Freshwater Environments

Bibliography

Breure, A.M., Mulder, C., Rombke, J., Ruf, A., 2005. Ecological classification and assessment concepts in soil protection. Ecotoxicology and Environmental Safety 62, 211–229.
Burkhead, N.M., 2012. Extinction rates in North American freshwater fishes, 1900–2010. Bioscience 62, 798–808.
Cardinale, B.J., Duffy, E., Gonzalez, A., *et al.*, 2012. Biodiversity loss and its impact on humanity. Nature 486, 59–67.
Chivian, E., Bernstein, A., 2008. Sustaining life: How human health depends on biodiversity. Oxford: Oxford University Press.
Chu, E.W., Karr, J.R., 2017. Environmental impact: concept, consequences, measurement. Elsevier Reference Module in Life Sciences. doi:10.1016/B978-0-12-809633-8.02380-3.
Conroy, M.J., Peterson, J.T., 2013. Decision making in natural resource management: A structured, adaptive approach. Hoboken, NJ: John Wiley.
DeFries, R., Nagendra, H., 2017. Ecosystem management as a wicked problem. Science 356, 265–270.
Diamond, J., 2005. Collapse: How societies choose to fail or succeed. New York: Viking Press.
Feld, C.K., Sousa, J.P., Martins da Silva, P., Dawson, T.P., 2010. Indicators for biodiversity and ecosystem services: Towards an improved framework for ecosystems assessment. Biodiversity and Conservation 19, 2895–2919.
Gottdenker, N.L., Streicker, D.G., Faust, C.L., Carroll, C.R., 2014. Anthropogenic land use change and infectious diseases: A review of the evidence. Ecological Health 11, 619–632.
Hancock, T., 1997. Ecosystem health, ecological iatrogenesis, and sustainable human development. Ecosystem Health 3, 229–234.
Harwell, M.A., Myers, V., Young, T., *et al.*, 1999. A framework for an ecosystem integrity report card. BioScience 49, 543–556.
Heinz Center (H. John Heinz III Center for Science, Economics and the Environment), , Economics and the Environment), 2008. The state of the nation's ecosystems: Measuring the lands, waters, and living resources of the United States. New York: Cambridge University Press.
Johnson, C.N., Balmford, A., Brook, B.W., *et al.*, 2017. Biodiversity losses and conservation responses in the Anthropocene. Science 356, 270–275.
Karr, J.R., 1991. Biological integrity: A long-neglected aspect of water resource management. Ecological Applications 1, 66–84.
Karr, J.R., 1996. Ecological integrity and ecological health are not the same. In: Schulze, P.C. (Ed.), Engineering within ecological constraints. Washington, DC: National Academy Press, pp. 97–109.
Karr, J.R., 2006. Seven foundations of biological monitoring and assessment. Biologia Ambientale 20 (2), 7–18.
Kates, R., Parris, T., Leiserowitz, A.H., 2005. What is sustainable development? Goals, indicators, values, and practice. Environment 47, 8–21.
Layzer, J.A., 2008. Natural experiments: Ecosystem-based management and the environment. Cambridge, MA: MIT Press.
Lindenmayer, D., Likens, G.E., 2010. Effective ecological monitoring. Collingwood, Victoria: CSIRO Publishing.
Millennium Ecosystem Assessment, , 2005. Ecosystems and human well-being: Biodiversity synthesis. Washington, DC: World Resources Institute.
Murphy, D.D., Weiland, P.S., 2014. Science and structured decision making: Fulfilling the promise of adaptive management for imperiled species. Journal of Environmental Studies and Sciences 4, 200–207.
O'Brien, A., Townsend, K., Hale, R., Sharply, D., Pettigrove, V., 2016. How is ecosystem health defined and measured? A critical review of freshwater and estuarine studies. Ecological Indicators 69, 722–729.

Sandifer, P.A., Sutton-Grier, A.E., Ward, B.P., 2015. Exploring connections among nature, biodiversity, ecosystem services, and human health and well-being: Opportunities to enhance health and biodiversity conservation. Ecosystem Services 12, 1–15.
Scheffers, B.R., De Meester, L., Bridge, T.C.L., *et al*., 2016. The broad footprint of climate change from genes to biomes to people. Science 354, 719–730.
Waters, C.N., Zalasiewicz, J., Summerhayes, C., *et al*., 2016. The Anthropocene is functionally and stratigraphically distinct from the Holocene. Science 351 (6269), doi:10.1126/science.aad2622.
Woodwell, G.M., 1990. The earth in transition: Patterns and processes of biotic impoverishment. Cambridge, MA: Cambridge University Press.
Wu, J., 2013. Landscape sustainability science: Ecosystem services and human well-being in changing landscapes. Landscape Ecology 28, 999–1023.
Wurtzebach, Z., Schultz, C., 2016. Measuring ecological integrity: History, practical applications, and research opportunities. Bioscience 66, 446–457.
Yu, D., Lu, N., Fu, B., 2017. Establishment of a comprehensive indicator system for the assessment of biodiversity and ecosystem services. Landscape Ecology 32, 1563–1579.

Further Reading

Bernhardt, E.S., Palmer, M.A., Allan, J.D., 2005. The National River Restoration Science Synthesis Working Group. Restoration of U.S. Rivers: A national synthesis. Science 308, 636–637.
Davidson, E.A., 2000. You can't eat GNP: Economics as if ecology mattered. Cambridge, MA: Perseus Publishing.
Haney, A., Power, R.L., 1996. Adaptive management for sound ecosystem management. Environmental Management 20, 879–886.
Jacobs, J., 2000. The nature of economies. New York: Modern Library, Random House.
Jungwirth, M., Muhar, S., Schmutz, S. (Eds.), 2000. Developments in hydrobiology: Assessing the ecological integrity of running waters. Dordrecht: Kluwer Academic.
Karr, J.R., Chu, E.W., 1999. Restoring life in running waters: Better biological monitoring. Washington, DC: Island Press.
Last, J.M., 1997. Public health and human health, 2nd ed. Stamford, CT: Appleton and Lange.
Naish, K.A., Taylor III, J.E., Levin, P.S., *et al*., 2008. An evaluation of the effects of conservation and fishery enhancement hatcheries on wild populations of salmon. In: Sims, D.W. (Ed.), Advances in marine biology. London and New York: Academic Press, pp. 61–194.
Ruddiman, W.F., Ellis, E.C., Kaplan, J.O., Fuller, D.Q., 2015. Defining the epoch we live in. Science 348 (6230), 38–39.
Summers, J.K., Smith, L.M., Harwell, L.C., *et al*., 2014. An index of human well-being for the U.S.: A TRIO approach. Sustainability. 3915–3935. doi:10.3390/su6063915.
World Bank, , 1995. Monitoring environmental progress: a report on work in progress. In: ESSD environmentally & socially sustainable development work in progress. Washington, DC: World Bank.
World Wildlife Fund, , 2016. Living planet report 2016; risk and resilience in a new era. Gland, Switzerland: WWF International.

Relevant Websites

https://www.epa.gov/waterdata/national-aquatic-resource-surveys-nars — US Environmental Protection Agency, National Aquatic Resource Surveys.
https://www.epa.gov/smartgrowth/creating-equitable-healthy-and-sustainable-communities — US Environmental Protection Agency, Equitable, Healthy, and Sustainable Communities.
http://www.heinzctr.org/ — The H. John Heinz III Center for Science, Economics and the Environment.
http://millenniumassessment.org/en/index.html — Guide to Millennium Ecosystem Reports.
http://www.teebweb.org/ — The Economics of Ecosystems and Biodiversity.
http://livingplanetindex.org/home/index — 2016 Living Planet Report: Risk and resilience in a new era.
earthcharter.org — Earth Charter Initiative.
www.globalecointegrity.net — Global Ecological Integrity Group.

Ecological Risk Assessment☆

Glenn W Suter II and Susan B Norton, National Center for Environmental Assessment, US Environmental Protection Agency, Washington, DC, United States

Glossary

Assessment endpoint An explicit expression of the environmental value to be protected. An assessment endpoint must include an entity and specific attribute of that entity.

Ecological risk assessment A process that evaluates the likelihood that adverse ecological effects may occur or are occurring as a result of exposure to one or more agents.

Effects characterization A phase in an ecological risk assessment in which the relationship between exposure to contaminants and effects on endpoint entities and properties and associated uncertainties are estimated.

Exposure characterization A phase in an ecological risk assessment in which the spatial and temporal distribution of the intensity of the contact of endpoint entities with contaminants and associated uncertainties are estimated.

Exposure–response The functional relationship between the degree of exposure to an agent and the nature or magnitude of response of organisms, populations, or ecosystems.

Measure of effect A measurable or estimable ecological characteristic that is related to the valued characteristic chosen as the assessment endpoint (equivalent to the earlier term "measurement endpoint").

Measure of exposure A measurable or estimable characteristic of a contaminant or other agent that is used to quantify exposure.

Problem formulation The phase in an ecological risk assessment in which the goals of the assessment are defined and the methods for achieving those goals are specified.

Risk characterization A phase of ecological risk assessment that integrates the exposure and the exposure–response profiles to evaluate the likelihood of adverse ecological effects associated with exposure to the contaminants.

Risk management The processes of deciding whether to accept a risk or to take actions to reduce the risk, justifying the decision, and implementing the decision.

Uncertainty Lack of knowledge concerning an event, state, or relationship. It may include bias or imprecision in an estimate due to the estimation process, natural variability, or a fundamental lack of knowledge.

Introduction

Managers and decision makers are challenged to solve complex environmental problems associated with the increasing pressures placed on vital natural resources by human activities. These challenges are made difficult by the sheer number and diversity of human disturbances and exacerbated by the complexity of imperfectly understood natural ecological systems. The process of ecological risk assessment (ERA) addresses ecological complexity and incorporates uncertainty in characterizing the impacts of natural and man-made disturbances on ecological resources (Fig. 1).

ERA integrates ecology, environmental chemistry, environmental toxicology, geochemistry, hydrology, and other fundamental sciences in estimating the probabilities of undesired ecological effects. In practice, ERAs derive from specific needs to assess human-induced effects on the environment. Many ERAs conducted in the United States are motivated by legislation, including the National Environmental Policy Act (NEPA), the Toxic Substances Control Act (TSCA), and the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA or Superfund). ERAs are also undertaken by private industry to determine future risks and liabilities associated with the development, use, and disposal (i.e., life cycle) of new or existing products (e.g., herbicides, pesticides, and industrial chemicals).

Several different approaches for performing an ERA have been developed internationally. However, the approach developed by the United States Environmental Protection Agency (US EPA) guides many ERAs performed in the United States and has served as a model for ERAs in other nations. It consists of three basic steps: problem formulation, analysis and characterization. The following discussion emphasizes this methodology.

☆*Change History*: March 2018. Glenn Suter and Susan Norton edited the original entry authored by Steve Bartell. They added information on the problem formulation phase, exposure–response modeling for ecological communities, and the use of weight of evidence. They also updated the Further Reading list and added a figure, a glossary of ecological risk assessment terms and a list of selected relevant websites.

The views expressed in this encyclopedia entry are the authors' and are not the views or policies of the US Environmental Protection Agency.This article is a revision of the previous edition by S.M. Bartell, Ecological Risk Assessment, In *Encyclopedia of Ecology*, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 1097–1101.

This is an update of S.M. Bartell, Ecological Risk Assessment, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 1097–1101.

 doi:10.1016/B978-0-12-409548-9.11137-6

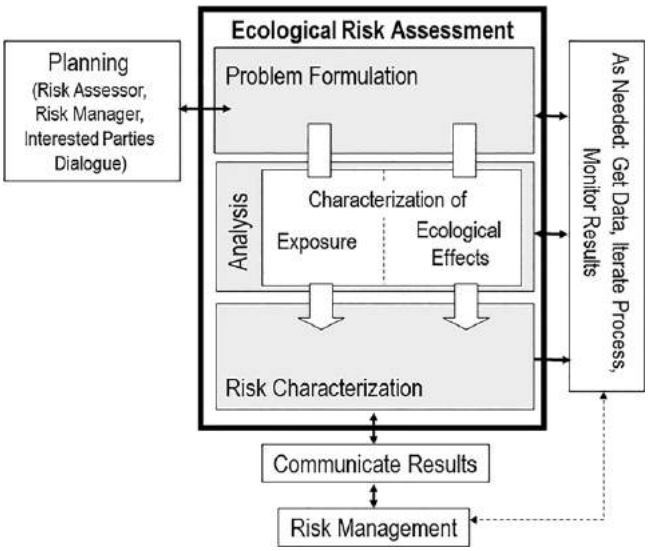


Fig. 1 US EPA's ecological risk assessment process Modified from US Environmental Protection Agency (1998). Guidelines for ecological risk assessment, EPA/630/002F, Washington, DC.

Definition of Ecological Risk

Risk is defined as the probability that an undesired event will occur. Correspondingly, ecological risk refers to the probability of the occurrence of an undesired ecological event. Alternative definitions of risk include an evaluation of the consequences of the undesired event along with estimation of its occurrence. For the most part, risk pertains to the probability of occurrence and this definition will serve this presentation.

ERA originally focused on the undesired ecological effects of toxic chemicals. As ERA evolved, the set of stressors has expanded to include physical, geological, hydrological, and biological stressors. Examples of these kinds of stressors include physical habitat degradation, erosion of soils or sediments, drought, floods, and introductions of exotic species.

Problem Formulation

This initial and perhaps most important part of the assessment defines the nature and scope of the ERA. It determines and describes the stressors that pose the risk and their sources, the processes by which the stressors may induce effects, and the undesired ecological effects (i.e., endpoints). This information is summarized in a conceptual model of the system to be assessed. This model should be viewed as dynamic and subject to change throughout the ERA in relation to modifications to the objectives and the development of new data and information.

Performing the problem formulation requires collaboration among risk managers and risk assessors to define the assessment objectives and scope and develop the corresponding conceptual model. The interactions might also involve other organizations and concerned members of the public (stakeholders). Initial discussions can help ensure that all important aspects of the assessment are identified, included as part of the problem formulation, and represented in the conceptual model. Such interactions can also ensure that the kinds of results produced by the ERA can be used effectively in the processes of decision making and risk management.

Following construction of the conceptual model, problem formulation continues by developing a plan to estimate the risks from each stressor to each endpoint. The resulting analysis plan characterizes the stressors, identifies specific ecological effects of concern, and identifies applicable data, as well as measures or models that can be used to quantitatively relate the stressors to the expected ecological effects.

Upon completion of risk estimation, risk managers, risk assessors, and stakeholders may reconvene to discuss the nature and interpretation (e.g., conclusions, assumptions, caveats) of the results in the context of the overall assessment objectives. Possible outcomes of these interactions include revisions to the conceptual model, collection of new data, and subsequent iterations of risk estimation. Once the requirements of risk managers and decision makers are fulfilled and documented, the process ends.

Following the problem formulation, the ERA continues with analyses that characterize exposure to the stressor(s) and the ecological effects of concern. The subsequent derivation of functional relationships between exposure and effects sets the stage for risk estimation.

Analysis: Exposure Characterization

Exposure is characterized by identifying the processes and mechanisms that bring organisms into contact with the stressor(s) of concern and quantifying the frequency, magnitude, and duration of such contact. The nature of the stressor(s) and the kinds of ecological effects of concern will strongly influence the exposure analysis. Each identified stressor will suggest a relevant spatial and temporal scale for analysis. The scales might be local and relatively short term, as for accidental spills of toxic, yet readily degraded or volatilized chemicals that result, for example, from hazardous waste management. Conversely, some stressors (e.g., fire, climate change) can exert ecological impacts over large expanses and for durations that greatly exceed the generation time of most organisms. Stressors are also evident at intermediate scales, for example, major oil spills and certain exotic species (e.g., in North America: gypsy moth, zebra mussel, Asian long-horned beetle).

The nature of specific stressor(s) can provide information concerning the processes or mechanisms of exposure that should be evaluated in an ERA. Chemical contaminants introduced into the environment are naturally transported by the movements of wind and water. Certain chemicals can accumulate in organisms and be transmitted throughout food webs. Some organic chemicals are comparatively insoluble in water and are rapidly adsorbed to soils and sediments, while other chemicals remain in solution and are transported by water. In contrast, movements of biological stressors such as invasive species might be augmented by private and commercial transportation. Corresponding characterization of exposure might emphasize delineation of transportation networks in place of physical–chemical processes.

The kinds of ecological effects included in the conceptual model can also provide insights into exposure analysis. Organisms occupy certain dimensions in space and time. Habitats have measurable spatial extent; ecological processes exhibit characteristic rates. Such observations can guide the analysis of exposure. For example, knowledge of the timing and duration of a sensitive life stage (e.g., eggs, larvae) can focus the corresponding measurement of stressors of concern and provide more meaningful quantification of exposure than longer-term averages or monitoring that might completely miss the critical time period for exposure. Similarly, seasonal changes in light, temperature, precipitation, and other physical factors can result in spatial–temporal variability in exposure. The important point is that variability in both the processes that influence the stressor and the characteristics of the ecological entities of concern should be addressed in performing a meaningful analysis of exposure.

Alternative approaches can be used to assess exposure. Worse-case scenarios can be developed that assume maximum values of the stressor. For example, end-of-pipe concentrations of toxic chemicals can be used without accounting for physical dilution, chemical alterations, or biological degradation that would otherwise reduce the concentrations experienced by the organisms of concern. This approach is biased toward overestimating exposure and risk, but it ensures that hazardous agents are not overlooked. If acceptable risks result from these extreme exposures, the assessment process might reasonably stop. As an alternative to worse-case scenarios, exposures might be measured. Actual measures of exposure are undoubtedly the most easily defended scientifically (presuming competent sampling and analysis) and the most realistic inputs to an ERA. Finally, exposures might be estimated using physical (e.g., microcosms, mesocosms) or mathematical models.

Exposure characterization generates an exposure profile. For chemicals, the profile includes the nature of the source; pathways of exposure; identification of environmental media of concern (e.g., soils, water, sediments, contaminated biota); estimates or measures of exposure concentrations (magnitude, timing, duration, recurrence); and uncertainties associated with these concentrations. Analogous exposure profiles are developed for nonchemical stressors addressed by an ERA.

Analysis: Effects Characterization

The large number and different kinds of ecological effects that are of potential concern distinguish, in part, ERA from more traditional human health risk assessment. The diversity of effects reflects the comprehensive nature of ecology and the environmental sciences. Ecologists have recognized several levels of organization as being useful in describing the natural world including organisms, populations, communities, ecosystems and landscapes. ERAs commonly identify more than one kind of ecological effect of concern in problem formulation. The ecological effects of concern identified during problem formulation should be ecologically important, sensitive to the stressor(s), and relevant to risk management.

Endpoints in ERA can include several effects at different levels of organization. An ERA might address alterations in basic physiological processes (e.g., photosynthesis, respiration) as long as they had implications for relevant endpoints the organism level (e.g., reduced survival or growth). In rare instances, effects to individual organisms might be selected as endpoints. In these cases, the individuals are likely to represent small populations of endangered species. Such instances would be equivalent to assessing risks to a maximally exposed human.

ERAs routinely assess effects on populations, expressed in terms of changes in survival, growth or reproduction in assessment populations. However, in cases in which the production or abundance of a particular population is important, assessments may be based on population dynamics. It is not uncommon for species officially designated as threatened or endangered or for a commercially or recreationally important species to be the focal points for ERA. Population models are well developed and these models are being used increasingly to estimate ecological risks.

Ecologists recognize that individual populations do not persist in an ecological vacuum. The number of species, their absolute and relative abundances, and their correlation of occurrence in space and time define community structure. The most common expression of effects on communities is the proportion of species affected. The most common exposure–response model is a

species sensitivity distribution, which is usually derived using toxicity test data for individual species but may also be derived from field data. Field tests such as mesocosm and field observational studies may also be used to derive exposure–response models for other community endpoints. These may include reductions in species diversity, indices of biotic integrity and measures of community similarity.

Ecosystem ecology addresses important feedback mechanisms between biotic and abiotic processes that determine ecosystem structure and function. The effects of stressors on fundamental ecosystem processes (e.g., primary production, system respiration, decomposition, nutrient cycling) can be important endpoints in ERA. The ecosystem concept also emphasizes the scale dependence and asymmetry of ecological interactions. Even in highly complex systems, not all components and processes are of equal importance. Delineation of critical scales and feedbacks can help define the relevant spatial–temporal scales in designing ERAs.

Assessments of stressors that operate at larger scales (e.g., acid precipitation, climate change) has led to the consideration of landscape-level effects in ERA. The most common landscape endpoints in risk assessment are alterations in the spatial distribution and extent of different habitat types within landscapes. Changes in the size, shape, and proximity of similar habitat areas (patches) can be measured or modeled.

Exposure–Response Relationships

The core of ERA is a causal relationship between exposure to the stressor and effects on the ecological endpoint, expressed as an exposure–response functions. For a given stressor, these functions estimate the severity of the expected ecological response in relation to the magnitude, frequency, and duration of the exposure. The derivation of exposure–response functions depends on the quantity and quality of available data.

Sources of data that might be used in the construction of exposure–response functions include: the results of toxicity tests performed under controlled laboratory conditions, direct measures of exposure and response in controlled field experiments, or uncontrolled field observations. Field observations are the most realistic and complete, but laboratory toxicity tests are less variable and less confounded by extraneous variables. In the absence of directly relevant data, the development of exposure–response functions may require extrapolations among similar stressors or ecological effects for which data are available. For example, effects might have to be extrapolated from an available test species to an untested species of concern. Similarly, toxicity data might be available only for a chemical similar to the specific chemical stressor of concern. An extrapolation from the known chemical to the unknown would be required to perform the assessment.

The most generally useful exposure–response relationships are statistical functions fit to paired exposure and effects levels. They generally increase monotonically and are nonlinear. Exposure levels corresponding to particular centiles of effects—may be used as benchmark values. The most common such value is the concentration (or dose) that produces effects in half of the organisms or 50% of the maximum response. For example, the concentration that results in 50% mortality during a prescribed period of exposure (e.g., 48, 96 h) defines the LC_{50} (lethal concentration that produces 50% mortality). An EC_{50} defines an exposure that results in a 50% decrease in an endpoint other than mortality, for example, growth. Dichotomous functions based on statistical significance such as the lowest observed effects concentration or LOEC have been used, but should be avoided. Significance tests do not measure the size of an effect and they do not correspond to biologically meaningful thresholds.

Risk Characterization

Risk characterization combines the exposure profiles with the exposure–response relationships to estimate ecological risks in ERA. A variety of methods and tools are available for risk estimation. For assessing risks posed by toxic chemicals, one simple method simply divides the exposure concentrations by a relevant toxicity benchmark value. A test endpoint such as an A LC_{50} or EC_{50} or a modeled threshold value such as a 10% loss of biodiversity may be used as a benchmark. Quotients equal to or greater than 1.0 are typically interpreted as implying risk; quotients less than 1.0 suggest minimal or no risk. Such quotients can prove useful in initial screening-level assessments to reduce the number of stressors that should be analyzed in greater detail. The screening assessments may be particularly effective if exposure estimates used in risk characterization are biased toward overestimating risk. However, 50% mortality or 50% reduction in growth or fecundity clearly are not thresholds for effects. Therefore, benchmark values, particularly when used for screening assessments often require safety factors.

More information is obtained by using the entire exposure–response relationship to characterize risk. That is, the model is solved for an estimated level of exposure that generates a predicted level of effect.

Depending on the availability of data, distributions of exposure and toxicity can be constructed and compared. Risk can be estimated by statistically comparing the degree of overlap between these distributions: the greater the overlap, the higher the risk. Using comparisons of distributions in assessments incorporates much more information than the quotient approach, including uncertainty, in the risk estimate. However, it must be remembered that they still are subject to the inherent uncertainties of using laboratory tests, standard species, and chemical transport and fate models.

Mathematical and computer simulation models can be used to estimate ecological risks. Following decades of model construction in support of basic ecological research and development, it stands to reason that some of these models might prove useful in estimating ecological risks posed by various stressors on organisms, populations, communities, and ecosystems. To be

useful in characterizing risk, the selected ecological model must include some representation of one or more of the assessment endpoints as a dependent variable. The model must also represent the stressor as an independent variable. The remaining critical aspect in selecting or adapting models for assessing risk is the ability to derive exposure–response relationships for the stressor(s) and ecological effects of interest.

Commonly in ERA, risks may be estimated by more than one method. Exposure–response relationships may be based on laboratory tests, microcosms, mesocosms, field tests, and field observations. Exposure may be estimated by modeling or field measurements. Weight-of-evidence methods can be used to evaluate those alternative types of evidence and to either combine them or to choose be one that gives the best estimate.

Uncertainty

Risk implies uncertainty. ERA was designed expressly to include uncertainty as an integral component of the assessment process. Sources of uncertainty, as used here, include natural variability in ecological and environmental phenomena, as well as lack of knowledge or bias and imprecision in estimates of exposure or exposure–response functions. This latter source of uncertainty can be exacerbated if extrapolations were involved in the derivation of the functions (e.g., laboratory to field, species to species). Incomplete and imperfect understanding of baseline ecological phenomena also adds uncertainty to ERA.

Uncertainties inherent to the risk assessment process can be quantitatively described using, for example, statistical distributions, fuzzy numbers, or intervals. Corresponding methods for propagating these kinds of uncertainties through the process of risk estimation include Monte Carlo simulation, fuzzy arithmetic, and interval analysis. Computationally intensive methods (e.g., the bootstrap) that work directly from the data to characterize and propagate uncertainties can also be applied in ERA. Implementation of these methods for incorporating uncertainty can lead to risk estimates that are consistent with a probabilistic definition of risk.

Methods of numerical sensitivity and uncertainty analysis can be used to examine uncertainty and identify the key sources of bias and imprecision in quantitative estimates of risk. Once identified, limited resources (e.g., time, funding) can be efficiently allocated to obtain new information and data for those major sources of uncertainty and reduce uncertainty. These analyses can be repeated until uncertainties associated with the risk estimates are of an acceptable degree or until uncertainties cannot be further reduced. Importantly, application of these methods for analyzing uncertainty will identify whether unacceptably high estimates of risk derive from the inherent severity of the stressor or from high uncertainty.

See also: Human Ecology and Sustainability: Precaution and Ecological Risk

Further Reading

Kapustka, L.A., Landis, W.G. (Eds.), 2010. Environmental risk assessment and management from a landscape perspective. New York: Wiley.
Norton, S.B., Cormier, S.M., Suter II, G.W., 2014. Ecological causal assessment. Boca Raton, FL: CRC Press.
Pastorok, R.A., Bartell, S.M., Ferson, S., Ginzburg, L.R., 2002. Ecological modeling in risk assessment. Boca Raton, FL: Lewis Publishers.
Suter II, G.W., 2007. Ecological risk assessment, 2nd edn. Boca Raton, FL: CRC Press.
Suter II, G.W., Efroymson, R.A., Sample, B.E., Jones, D.S., 2000. Ecological risk assessment for contaminated sites. Boca Raton, FL: Lewis Publishers.
US Environmental Protection Agency. 1988. Guidelines for ecological risk assessment, EPA/630/002F, US EPA, Washington, DC.

Relevant Websites

http://www.environment.gov.au/science/supervising-scientist/research/ecological-risk—Australian Department of Environment and Energy.
http://www.federalcontaminatedsites.gc.ca/B15E990A-C0A8-4780-9124-07650F3A68EA/ERA%20Guidance%2030%20March%202012_FINAL_En.pdf—Environment Canada.
https://echa.europa.eu/documents/10162/23036412/bpr_guidance_ra_vol_iv_part_b-c_en.pdf/e2622aea-0b93-493f-85a3-f9cb42be16ae—European Chemicals Agency.
http://www.oecd.org/publications/guidance-on-selecting-a-strategy-for-assessing-the-ecological-risk-of-organometallic-and-organic-metal-salt-substances-based-9789264274785-en.htm—OECD.
https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=156509—US Environmental Protection Agency.
https://cfpub.epa.gov/ncea/risk/era/recordisplay.cfm?deid=233809—US Environmental Protection Agency.
https://cfpub.epa.gov/ncea/risk/era/recordisplay.cfm?deid=253500—US Environmental Protection Agency.
https://www.epa.gov/caddis—US Environmental Protection Agency.
https://www.epa.gov/risk/ecological-risk-assessment—US Environmental Protection Agency.
https://www.regulations.gov/document?D=EPA-HQ-OPP-2002-0262-0162—US Environmental Protection Agency.

Ecosystem Health Indicators☆

Marion Kruse, Kiel University, Kiel, Germany

Authors of first version on which this updated is based on: B. Burkhard, Leibniz Universitat Hannover, Hannover, Germany; F. Müller and A. Lill, Kiel University, Kiel, Germany.

Introduction

The fundamentals of the ecosystem health concept have been already set in the 18th century with the ideas of the Scottish geologist James Hutton, who started to describe the Earth as an integrated system. Later, the writings about land sickness of the pioneering ecologist Aldo Leopold in 1941 created the first roots of the ecosystem health concept. Since then, the definition of ecosystem health has been constantly evolving toward an integration of more and more human and societal contexts, in order to understand what is considered a healthy ecosystem. Ecosystem health can be understood as a management concept that contributes to solve environmental problems. In the United States and Canada the concept of ecosystem health has been adopted in legislation and it is part of international political programs, for example the so-called Rio Convention on sustainable development. Therefore, the assessment of ecosystem health does not only take into account ecological components but also requires a comprehension of social, economic, and cultural dimensions. Hence, its definition is more comprehensive than the definition of human health.

☆*Change History:* April 2018. M. Kruse extended the abstract, introduction, and conclusion and introduced small edits to the other parts of this article. The further reading list was updated.

These facts have to be reflected while defining appropriate sets of indicators that can be applied within respective assessments of ecosystem health. Indicators must fulfill diverse scientific and practical requirements (e.g., a distinct relationship between the indicator and the subject to be indicated (indicandum), measurability and data availability). The selection and application of indicators determine the results of the quantification of the ecosystem health and have a major impact on the management targets and their success.

Suitable indicator sets have to consider both the structure and the function of ecosystems in different spatial and temporal scales. As pointed out in various studies, ecosystems on our planet already are, and will continue to be, degraded under the pressure of increasing human activities. Anthropogenic impacts, such as intensive and unsustainable land use causing erosion and eutrophication, pollution, urbanization and loss of biodiversity have become more and more evident during the last decades. There are no more ecosystems on Earth which are not impacted by any human activity. Hence, preventative and restorative strategies are needed in order to achieve the health of regional and global ecosystems on a long-term perspective. Due to the complexity of ecosystems and their interactions in human–environmental systems, there is a demand for holistic management concepts that can tackle these problems. With regard to different well-known and established indicators, for example species diversity or water quality, many of the ecosystems on our planet can be considered unhealthy. Several of their functions, especially those needed to sustain life of the human community, have become impaired. Consequently, ecosystem health has been changed significantly, that is, because it refers to systems that are manipulated to satisfy human needs. This makes the difference to other related concepts, for example, some notions of ecological integrity, which refers to the functioning of ecosystems based on nature-near, self-organized processes.

Generally speaking, an ecosystem is often called healthy if it is stable (respectively resilient) and sustainable in the provision of goods and services used by human societies (ecosystem services). This implies that the ecosystem has the ability to maintain its structure (organization) and function (vigor) over time under external stress (resilience). Further ecological principles and theories like homeostasis, diversity, complexity, emergent properties, or hierarchy principles are closely related or included in the concept of ecosystem health as it refers to complex adaptive systems. These systems are characterized by certain dynamics including sudden reconfigurations from one state of organization to another. Some changes can be inherently unpredictable. Both positive and negative feedback processes and cause and effect chains, operating over a range of spatial and temporal scales, predominate these dynamics.

Besides these theoretical considerations on ecosystem health, a list of further aspects, which are also, included in currently available definitions have been compiled in de Kruijf and van Vuuren:

- Healthy ecosystems are free from ecosystem distress syndrome, a set of signs present in most heavily damaged ecosystems;
- Healthy ecosystems are resilient—they recover from natural perturbations and disturbances;
- Healthy ecosystems are self-sustaining and can be perpetuated without subsidies or drawing down natural capital;
- Healthy ecosystems do not impair adjacent ecosystems—that is, healthy ecosystems do not result in stress in neighboring systems;
- Healthy ecosystems are free from risk factors;
- Healthy ecosystems are economically viable;
- Healthy ecosystems sustain healthy human communities.

Hence, ecosystem health represents the sustainability of an ecosystem as a whole that needs a minimal external support by management measures. It comprises biophysical, socioeconomic, cultural, and human dimensions of the environment. This notion of ecosystem health is closely interrelated to the ability of ecosystems to provide and sustain ecosystem services.

In the beginning of the article, a short review of the historic development of ecosystem health assessment and indicators is given. Afterwards, different categories and examples of corresponding ecosystem health indicators are presented. This part is supplemented by several examples of commonly applied ecosystem health indicators from aquatic as well as from terrestrial ecosystems. At the end, some examples of applied methods of ecosystem health assessment are given to show the diversity of ecosystem health indicators and application contexts.

Ecosystem Health Assessment and Indicators

The idea of ecosystem health assessment (EHA) has started in the late 1980s with an increasing amount of publications on this topic. Since then, there have been wide discussions about adequate concepts and methodologies. As the individual ecosystems are never identical with reference to basic features (e.g., forest ecosystem and a wetland ecosystem), it is challenging to find general and reproducible approaches instead of case specific techniques. Furthermore, appropriate assessments have to bridge the gaps between natural, socioeconomical, and health sciences, and must integrate human norms and values with the aim to support sustainable management of natural resources. Defining management goals and their indicators is a further challenge in the ecosystem health assessment.

As ecosystem health cannot be measured or observed directly, surrogate measures (indicators) have to be applied to assess it. These indicators should be backed by ecological principles and systems theory. They should be suitable for applications on varying temporal and spatial scales. Parameters needed for their quantification have to be obtainable and reproducible within ecological assessments. Moreover, indicators have to give adequate information about ecosystem quality trends, useful for managers and

scientists that take into account the complex cause and effect chains in human–environmental systems. Different values and processes have to be integrated in order to assess the health of an ecosystem in relation to economic and social aspects, regional and global environmental changes, and alterations in the provision of ecosystem services and their consequences for human well-being.

Categories of Ecosystem Health Indicators

As ecosystems are extremely complex, any single indicator cannot be completely representative with reference to all possible demands, features, and conditions. Nevertheless, there is a wide spectrum of indicators that can be used for the assessment of ecosystem health. They can be classified into different categories that include varying levels of integration, ranging from rather reductionistic to holistic indicators, integrating a broad range of environmental information:

- Indicators based on the abundance of selected species;
- Indicators based on the concentration of selected elements;
- Indicators based on ratios between different classes of organisms of elements;
- Indicators based on ecological strategies or processes;
- Indicators based on ecosystem composition and structure;
- Systems theoretical holistic indicators.

The identification of appropriate indicator-indicandum (the phenomenon that has to be assessed) relations has to be carried out thoroughly, especially while dealing with holistic and highly aggregated indicators.

Commonly Applied Ecosystem Health Indicators

In the following, some examples out of the huge amounts of specific indicators and sets of indicators are given.

Indicators Based on the Abundance of Selected Species

These biological indicators cover the presence or absence of selected species. Indicator species have to be selected in order to be representative for a certain phenomenon or to be sensitive to distinct changes. Hence, their appearance and dominance is associated with a certain environmental situation, the usage of this kind of indices implies on the one hand a certain level of uncertainty, on the other, the results can directly be linked to biological phenomena which is useful for proper management of resources. In the following, some examples will be outlined.

Saprobic classification

The saprobe system is a collection of organisms that give information about the degree of water pollution. The different pollution intensities (saprogenic stages) are related to certain indicator organisms (e.g., bacteria, fungi, algae, amoeba, mussels, worms, insect larvae, and fishes) and ranges from polysaprobic (very highly polluted), α-mesosaprobic (highly polluted), β-mesosaprobic (medium polluted) to oligosaprobic (rather clean and clear water).

Bellan's pollution index

Bellan considers aquatic species like *Platynereis dumerilii, Theosthema oerstedi, Cirratulus cirratus,* and *Dodecaria concharum* as water pollution indicators. Clear water is indicated by species like *Syllis gracilis* or *Typosyllis prolifera*. The equation of the Bellan's pollution index can be formulated as follows:

$$\mathrm{IP} = \Sigma\ \text{dominance of pollution indicator species} / \text{dominance of pollution (clear) water indicators}$$

If the index value is bigger than 1, a pollution disturbance in the community is indicated.

AZTI marine biotic index

For this indicator the soft bottom macrofauna is distinguished into five groups in accordance to their sensitivity to increasing stress:

I. species that are very sensitive to organic enrichment and eutrophication and that are only present under unpolluted conditions;
II. species that are indifferent to organic enrichment, occur always in low densities and show no significant variations over time;
III. species tolerant to excess organic matter enrichment (these species can also be found under normal conditions but usually their populations are supported by organic enrichment);
IV. second-order opportunist species, very often small polychaetes; and
V. first-order opportunist species (deposit feeders).

$$\text{AMBI (bioticcoefficient)} = \{(0\ \text{x\%I}) + (1.5\ \text{x\%II}) + (3\ \text{x\%III}) + (4.5\ \text{x\%IV}) + (6\ \text{x\%V})\}/100$$

The AZTI marine biotic index (AMBI) results can be classified as: normal (AMBI coefficient between 0.0 and 1.2), slightly polluted (1.2 and 3.2), moderately polluted (3.2 and 5.0), highly polluted (5.0 and 6.0), or very highly polluted (6.0 and 7.0). The AMBI has been considered useful in terms of the application of the European Water Framework Directive to coastal ecosystems and estuaries.

Bentix

The Bentix is based on the AMBI but has only three groups in order to avoid errors in the grouping of species and to make the calculation of the index easier:

- *Group I*: Species generally sensitive to disturbances.
- *Group II*: Species that are tolerant to stress or disturbance. Populations may respond to organic enrichment or other source of pollution.
- *Group III*: First-order opportunistic species (pioneer, colonizers or species which are tolerant to hypoxia).

$$\text{Bentix} = \{(6 \times \%\text{I}) + 2(\%\text{II} + \%\text{III})\}/100$$

The Bentix results can be classified as: normal (4.5–6.0), slightly polluted (3.5–4.5), moderately polluted (2.5–3.5), highly polluted (2.0–2.5), or very highly polluted (Bentix = 0).

Macrofauna monitoring index

This index comprises 12 indicator species. Each of them is assigned a score, based primarily on the ratio of its abundance in control versus impacted samples. The index value is the average score of those indicator species which are present in the sample.

Benthic response index

The Benthic response index (BRI) is calculated as abundance weighted average pollution tolerance of species that occur in a sample. This is similar to the weighted average approach used in gradient analysis.

Indicators Based on Concentration of Selected Elements

Many assessments are based on measurements or monitoring of concentrations or densities of selected elements that can be linked with altering system states. In the context of environmental management, a link to anthropogenic activities is convenient: for example, the estimation of the level of eutrophication on the basis of total phosphorus concentrations. Another typical measurement is the pH value, referring to the activity of hydrogen ions, which, for instance, can be linked to effects of air pollution (acid rain).

Indicators Based on Rations Between Different Classes of Organisms or Elements

Increase or decrease of one species in relation to another can give useful information about changes in systems. Depending on the problem and the system that has to be investigated, different indices are applicable.

Nygaard's algal index

This index is used to evaluate the nutritional status of lake ecosystems and their fertility. The index can be calculated as a compound quotient out of:

$$\text{Myxophycean} + \text{Chlorococcales} + \text{Centric} + \text{Euglenophyceae}/\text{desmids}$$

Oligotrophic systems have a quotient between 0.01 and 1.0, eutrophic systems between 1.2 and 2.5. Further Nygaard's indices are based on ratios between Myxophyceae and desmids, chlorococcales and desmids; centric and pinnate diatoms or Euglenophyta and Myxophyceae + chlorococcales.

Diatoms/non diatoms ratio

In the context of altering nutrient charges in aquatic systems, the ratio of the major phytoplankton groups, diatoms versus flagellates (diatoms–non diatoms ratio), can be used as an indicator. Regarding for example eutrophication, nutrient reductions can be seen in a decrease in flagellate abundance.

Indicators Based on Ecological Strategies or Processes

Regarding ecosystems as entities of complex processes and interactions, varying strategies and processes are related to altering system's conditions caused by human activities or different stages of natural development.

Process and rate indicators

One functional way to assess the health of ecosystems is to measure or model indicators that represent important processes in the ecosystem from which conclusions for the whole system can be drawn. Primary production or growth rates are among the most commonly used indicators. The cycling of matter, water, or energy between the different components of a system, and the transformation of matter, water, or energy into different forms, are main processes in natural systems. Examples include cycling of nitrogen or phosphorus, the fixation of energy by plant photosynthesis, or the water cycle based on precipitation, runoff, and evapotranspiration.

Indicators based on ecological strategies

Further indices consider the distinct behavior of different taxonomic groups under environmental stress situations, for example, the nematodes/copepods index. Some authors have criticized these indices because of their dependence on parameters such as water depth and sediment particle size, and also because of their unpredictable pattern of variation depending on the type of pollution. More recently, indices, for example, the polychaetes/amphipods ratio or the index of r/K strategists, which consider all the benthic taxa, were developed.

Index of r/K strategists

In a rather stable system with infrequent disturbances, the competitive dominants in most communities are k-selected, or conservative, species with the attributes of large body size and long life span. They are usually dominant in terms of total biomass, but not dominant in number. R-selected, or opportunistic, species with shorter life spans are usually numerically dominant but do not represent a large proportion of the total biomass of the community. After a more significant disturbance, conservative species are usually less favored, and the opportunistic species can become dominant as well in biomass as in number. Thus, the analysis of r- and K-strategists' distributions can be used for the indication of ecosystem health.

Often, different feeding strategies of organisms are used to describe ecosystem conditions and developments.

Infaunal index

For the assessment of the trophic infaunal index, the macrobenthos species are divided into four groups:

1. suspension feeders,
2. interface feeders,
3. surface deposit feeders, and
4. subsurface deposit feeders.

The infaunal index is calculated with the following equation:

$$\mathrm{ITI} = 100 - (100/3) \times (0n1 + 1n2 + 2n3 + 3n4)/(n1 + n2 + n3 + n4)$$

$n1$, $n2$, $n3$, and $n4$ are the number of individuals that are sampled in each of the four species groups. If the ITI value approaches 100, suspension feeders are dominant which indicates an environmental disturbance.

Indicators Based on Ecosystem Composition and Structure

Ecosystems are dynamic systems that show varying compositions and structures during their different stages of development or due to perturbations. Different indicators to describe systems' composition and structure exist.

Shannon–Wiener index

Indices based on the diversity values are very common. One of the most often used is the Shannon–Wiener index, developed by C.E. Shannon and W. Wiener. This index originates in information theory and assumes that individuals are sampled at random out of a community which is indefinitely large, and that all the species are represented in the sample.

The Shannon–Wiener index is calculated by the following equation:

$$\mathrm{H}' = -\Sigma p_i \log_2 p_i$$

In this equation, p_i is the proportion of individuals found in species i. The values of this index can vary between 0 and 5. H' reaches a maximum value if the individuals of all species occur with the same density. Other indices based on the diversity value are, for example, the Pielou evenness index, the Brillouin index, the Margalef index, the Berger–Parker index, the Simpson index, and K-dominance curves.

Food webs

Food webs describe the connections of plants and animals which depend upon each other referring to the flow of energy. Organisms are assigned to different trophic levels that classify their position in the food chain which is determined by the number of energy transfer steps to that level. In general, food webs become more complex during the development of a system. Hence, their structure

and composition can be used to assess the condition and stage of development of a system. Network theory plays an important role for the interpretation of food web structures.

Ascendency

Ascendency and related indices are abstract concepts for the quantification of the size and organization of flows in systems using information-theoretic terms. Ascendency values indicate the overall status of dynamic systems in a quantitative fashion and show the limits of system growth and development. The response of a system to perturbations can be measured which is useful in the context of ecosystem health.

Systems Theoretical Holistic Indicators

Indicators based on systems theory have high potentials to represent complex issues in a holistic manner, but they tend to be rather abstract and difficult to communicate.

Vigor, organization, and resilience (V–O–R model)

Measures of vigor, organization (or performance), and resilience are often used to assess ecosystem health. However, they are more easily described in theory than quantified in practice. Vigor is usually represented by activity, metabolism, or primary productivity. A study of the Great Lakes Basin (North America) showed the decline in the abundance of fish and infertility of agricultural soils within the basin as an example of reduced vigor. Organization represents the diversity and number of interactions between system components. An example also from the Great Lakes, is the reduced morphological and functional diversity of fish associations that occur under multiple stresses. Resilience is normally understood as a system's capacity to maintain structure and function in the presence of (external) stress. When resilience is exceeded, the system can shift to an alternate state. A prime example is the shift from benthic to pelagic dominated fish associations in the Laurentian Lower Great Lakes Basin. In this approach ecosystem health is closely related to the concepts of stress ecology, where vigor, systems organization, resilience and the absence of signs of ecosystem distress are the main factors for the health of a system.

Exergy indices

Further holistic indicators are the exergy index and the specific exergy index. Exergy is derived from thermodynamics and measures the energy fraction that can be transformed into mechanical work. In ecosystems, the captured exergy is used to build up biomass and structures during succession. Hence, exergy can be used as a measure of biomass, structure, energy, and information stored in the biomass. Therefore, more complex organisms and systems also have more built in exergy that simpler ones. Specific exergy is defines as exergy per biomass. Both exergy and specific exergy can be used as indicators for ecosystem health. Relations between the exergy values and other ecosystem health characteristics like diversity, structure, or resilience can be found. For example, specific exergy expresses the dominance of the higher organisms as they per unit of biomass carry more information (they have higher β-values). A very eutrophic ecosystem has a very high exergy due to the high concentration of biomass, but the specific exergy is low as the biomass is dominated by algae with low β-values. The combination of exergy index and the specific exergy index usually gives a more satisfactory description of ecosystem health than the exergy index alone, because it considers the diversity and the life conditions for higher organisms.

The Holistic Ecosystem Health Indicator

The holistic ecosystem health indicator (HEHI) was developed in 1999 in Costa Rica as an integrative indicator which might be an appropriate tool for assessing and evaluating health of managed ecosystems. The HEHI follows a hierarchical structure starting with three main branches: ecological, social, and interactive. Measures about the condition and trend of the ecosystems are organized within the ecological branch. Socioeconomic measures concerning the community dependent on the ecosystem or affected by management decisions are organized within the social branch. The interactive branch includes measures relating to land-use and management decisions that characterize the interactions between the human communities and the ecosystem. Furthermore, each branch is subdivided into categories or criteria.

The indicators belonging to the categories serve as measures for the performance of each category. If we, for example, take soil quality, this is a category within the ecological branch and it can be measured using indicators such as microbial biomass, water infiltration, compaction, etc. Each category is given a target or a benchmark, which is based on references available in scientific literature, policies, etc. For example, a water-quality indicator can have a target defined by legal limits specified by the administrative authority in charge, while a target for a productivity indicator may be defined by a combination of the capacity of the system and objectives set by stakeholders.

NRCS indicator selection model

The National Resource Conservation Service (NRCS) of the US Department of Agriculture assigned an NRCS indicators action team in 1994, which developed an indicator selection model for the use of indicators in evaluations of ecosystem conditions.

The team identified framing questions for four different ecosystem aspects, namely: ecosystem processes, recovery processes, landscape and community structure, and abiotic features. The framing questions represent a minimum set of diagnostic questions,

which need to be answered when doing comprehensive evaluations of ecosystem conditions or health. The questions are asked at all scales of ecosystem evaluation:

System Processes: The questions asked at "system processes" are as follows:

1. Are precipitation and groundwater resources captured, stored, used, and released in a safe and stable manner?
2. Are kinds and flows of chemicals (minerals, nutrients, other) and energy in balance and optimized for plant and animal communities and biomass production requirements?
3. Are annual cash flows, technical assistance, and conservation incentives timely and adequate for desired community and land user income?

Recovery Processes: The questions asked at "recovery processes" are as follows:

1. Are soil, water, air, plant, and animal resources and biophysical processes in place and in a condition to allow timely and full recovery from stresses and disturbances and to meet management objectives?
2. Are social and economic systems available to allow land users and communities and the resources they manage to recover from environmental and socioeconomic stresses?
3. Are there human and animal resource health concerns associated with the management of present or planned enterprises?

Landscape and community structure: The questions asked at "Landscape and community structure" are as follows:

1. Do landscape features and patterns facilitate use, protection and optimization of ecosystem processes?
2. Do commodity markets, investment capital and public programs encourage land uses, enterprises and resource management that are compatible with ecosystem processes?
3. Are decision-making processes available to communities and individuals to resolve conflicts regarding current and desired uses, management and protection of natural resources?
4. Does the social infrastructure (healthcare, education, multicultural recognition, etc.) support and promote the desired quality of life for the communities and individuals?

Abiotic features: The question asked at this scale is: Are current and planned land uses and desired future conditions suited to the abiotic conditions (e.g., stream temperature, flow velocities, riffle/pool ratios, riparian shading, climate, topography, soils, and geology)?

Within the next step, for all ecosystem components (environmental, ecological, socioeconomic, cultural, or political factors), which are considered to be necessary elements of the respective system, appropriate indicators are listed. These indicators are the quantitative or qualitative tools in this model to assess the status, condition, or trend of a given ecosystem attribute or component. The underlying assumption for the use of such indicators is that relationships can be inferred between a relatively easily measured ecosystem attribute (i.e., litter distribution and amount) and the more difficult to measure ecosystem components or processes (i.e., energy flow and nutrient cycling).

Applied Methods of Ecosystem Health Assessment

The concept of ecosystem health has been criticized for being too fuzzy and not concrete enough for practical application. Nevertheless, it found entry in different management strategies and definitions of political targets. For example, in principle 7 of the Rio Declaration of 1992, it has been claimed that "States shall cooperate in a spirit of global partnership to conserve, protect, and restore the health and integrity of the Earth's ecosystem." More recently, the consideration of ecosystem health was integrated into the ecosystem approach of the Convention on Biological Diversity (CBD) and in connection to the precautionary principle which is part of the environmental and nature conservation strategies in different countries. Furthermore, the ecosystem health concept was implemented in the strategies of the OSPAR (Convention for the Protection of the Marine Environment of the North-East Atlantic) and the HELCOM (Baltic Marine Environment Protection Commission—Helsinki Commission) commissions for the protection of marine environments or in the European Union Water Framework Directive. The development of appropriate monitoring and indicator systems and methods to assess ecosystem health in practice are main targets of these initiatives.

Three examples for methods for possible ways of ecosystem health assessment applications are given in the following. The first one, the ecosystem health index method (EHIM) is based on a combination of different sub indicators which are synthesized into index values using individual weighting factors. Whereas in the ecological model method (EMM), modeling procedures are used to quantify indicator values, the direct measurement method (DMM) is based on values that are measured directly or calculated indirectly for the assessment of ecosystem health.

Ecosystem Health Index Method

For this method, a synthetic ecosystem health index (EHI) in a scale of 0–100 has been developed in order to quantitatively assess the state of ecosystem health. The worst health state exists when EHI is zero. To make description easier, the EHI was divided into

five categories: 0%–20% (worst health state), 20%–40% (bad health state), 40%–60% (middle health state), 60%–80% (good health state), and 80%–100% (best health state). Five steps are necessary to calculate the EHI:

1. selection of basic and additional ecosystem indicators;
2. calculation of sub-EHIs for all selected indicators;
3. determination of weighting factors for all selected indicators;
4. calculation of synthetic EHI using sub-EHIs and weighting factors for all selected indicators; and
5. assessment of ecosystem health based on synthetic EHI values.

Ecological Model Method

Five steps are applied when using the ecological model method (EMM):

1. determination of the model structure and complexity according to ecosystem structure;
2. establishment of an ecological model by creating a conceptual diagram, developing model equations as well as estimating model parameters;
3. calibration of the model in order to assess its suitability in application to ecosystem health assessment process;
4. calculation of the ecosystem health indicators; and
5. assessment of the ecosystem health using the values of the indicators.

Direct Measurement Method

The direct measurement method (DMM) is used by applying the following three steps:

1. identification of relevant indicators which are needed for the assessment process;
2. direct measurement or indirect calculation of the selected indicators; and
3. assessment of ecosystem health based on the resulting indicator values.

Conclusion

Ecosystem health is not simply defined because it is a holistic management concept that must deal with the complexity of human–environmental systems and today's diverse environmental problems, which are jeopardizing the supply of ecosystem services and human wellbeing. There are several definitions presented in the literature and the term "health" with its easily transferable metaphoric character can act on the one hand as a positive connotation in some regions, where it was on the other hand not successfully integrated in management and decision-making in other areas.

There is a broad range of ecosystem health indicators available to cope with this spectrum of ecosystems and human pressure, to which they are exposed. Depending on the questions or environmental problem to investigate, such indicator sets can be combined and used to supplement each other in order to carry out holistic ecosystem health assessments since environmental problems are often diversified. Therefore, indicators applicable for all different kinds of research and management questions have not been developed so far. The presented indicators are examples of the wide range of possible approaches developed so far, which also reflects the versatile definitions and viewpoints of ecosystem health. Another critical point is the availability of data for the quantification of respective indicators on suitable spatial and temporal scales. If ecosystem health indicators are linked to existing monitoring networks or easily transferable methods (e.g., by means of remote sensing data or report cards), appropriate data sets can be on-hand. Further data can be derived by new modeling techniques. Especially the linkage with social, economic, and land-use data can give valuable information with regard to the effects of human action on the state of ecosystems. This has been presented and discussed in several published case study applications in different socioecological context and regions.

Further Reading

Callicott JB (1995) A review of some problems with the concept of ecosystem health. *Ecosystem Health* 1(2): 101–112.
Costanza R and Mageau M (1999) What is a healthy ecosystem? *Aquatic Ecology* 33: 105–115.
Costanza R, Norton BG, and Haskell BD (eds.) (1992) *Ecosystem health: New goals for environmental management*. Washington, DC: Island Press.
de Kruijf HAM and van Vuuren DP (1998) Following sustainable development in relation to the north–south dialogue: Ecosystem health and sustainability indicators. *Ecotoxicology and Environmental Safety* 40: 4–14.
Flint N, Rolfe J, Jones CE, Sellens C, Johnston ND, and Ukkola L (2017) An ecosystem health index for a large and variable river basin: Methodology, challenges and continuous improvement in Queensland's Fitzroy Basin. *Ecological Indicators* 73: 626–636.
Jian P, Yanxu L, Tianya L, and Jiansheng W (2017) Regional ecosystem health response to rural land use change: A case study in Lijiang City, China. *Ecological Indicators* 72: 399–410.
Jørgensen SE, Xu F-L, and Costanza R (2010) *Handbook of ecological indicators for the assessment of ecosystem health*, 2nd edn. Boca Raton: CRC Press.
Marquez JC, Salas F, Patricio J, Teixera H, and Neto JM (2009) *Ecological indicators for coastal and estuarine environmental assessment. A user's guide*. Southampton: WIT Press.

Muñoz-Erickson, T.A. and Aguilar-Gonzalez, B.J. (2003). The use of ecosystem health indicators for evaluating ecological and social outcomes of the collaborative approach to management: The case study of the Diablo Trust. Proceedings of National Workshop Evaluating Methods and Environmental Outcomes of Community-Based Collaborative Processes Utah: Snowbird Center.

Patil GP, Brooks RP, Myers WL, Rapport DJ, and Taillie C (2001) Ecosystem health and its measurement at landscape scale: Toward the next generation of quantitative assessments. *Ecosystem Health* 7(4): 307–316.

Rapport DJ (1995) Ecosystem services and management options as blanket indicators of ecosystem health. *Journal of Aquatic Ecosystem Health* 4: 97–105.

Rapport DJ (ed.) (2003) *Managing for healthy ecosystems.* Boca Raton: Lewis.

Rapport DJ, Costanza R, and McMichael AJ (1998) Assessing ecosystem health. *TREE* 13: 397–402.

Robertson BP, Savage C, Gardner JPA, Robertson BM, and Stevens LM (2016) Optimising a widely-used coastal health index through quantitative ecological group classifications and associated thresholds. *Ecological Indicators* 69: 595–605.

Ecotoxicology: The History and Present Direction[☆]

Hailong Zhou, Nan Xiang, Jia Xie, and Xiaoping Diao, Hainan University, Haikou, China

Glossary

Adverse outcome pathway (AOP) AOP is structured representation of biological events leading to adverse effects and is considered relevant to risk assessment.

Dynamic energy budget (DEB) DEB offers a systematic way to relate processes at different organizational levels: molecules, cells, organisms, populations and ecosystems, and which focuses on metabolic processes that are common to large groups of organisms.

Ecological risk assessment (ERA) ERA is the process for evaluating how likely it is that the environment may be impacted as a result of exposure to one or more environmental stressors such as chemicals, land change, disease, invasive species and climate change.

Lowest observed effect concentration (LOEC) LOEC is the lowest concentration or amount of a substance found by experiment or observation that causes an adverse alteration of morphology, function, capacity, growth, development, or lifespan of a target organism distinguished from normal organisms of the same species under defined conditions of exposure.

Mass-balance ecosystem model (MBEM) MBEM is a series of linear equations to describe biomass interactions between trophic levels, and gives a static, mass-balanced representation of food web structure.

No observed adverse effect level (NOAEL) NOAEL is the highest tested dose of a medicine at which there is no increase in the frequency of any adverse effects (biological or statistically significant) when compared to its control.

No observed effect concentration (NOEC) NOEC is the highest experimental point that is without adverse effect, denotes the level of exposure of an organism, found by experiment or observation, at which there is no biologically or statistically significant (e.g., alteration of morphology, functional capacity, growth, development or life span) increase in the frequency or severity of any adverse effects in the exposed population when compared to its appropriate control.

Omics Omics informally refers to a field of study in biology ending in -omics, such as genomics, transcriptomics, proteomics or metabolomics. Omics aims at the collective characterization and quantification of pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms.

Persistent organic pollutants (POPs) POPs are compounds that resist photolytic, chemical, and biological degradation. They are low water solubility and high lipid solubility, resulting in bioaccumulation in fatty tissues of living organisms.

Introduction

Ecotoxicology is the study of the effects of toxic chemicals on different scales, ranging from the biomolecule, cell, organ, system, whole organism, population, community, ecosystem, and biosphere levels. Ecotoxicology examines the movements of harmful chemical pollutants in the environment, including migration, transformation and degradation. Ecotoxicology is a multidisciplinary field, which integrates toxicology and ecology. Furthermore, Ecotoxicology is the science of contaminants in the biosphere and their effects on constituents of the biosphere, including humans.

[☆]*Change History*: March 2018. We have updated the abstract, keywords, glossary, the text, current trends in ecotoxicology. Additionally, we extended introduction, the further readings to this entire article, and deleted the old Fig. 3 and added new Fig. 3. Regarding the text, we added the latest theory and approaches in this field, such as Adverse Outcome Pathway, Reverse transcription real-time quantitative PCR, Effect-Directed Analysis, Omics technologies, Dynamic Energy Budget (DEB) model, Mass-Balance Ecosystem Model, Ecopath with Ecosim (EwE) approach, etc. In a word, these updates and extension can help us better understand and the knowledge and methods, also contribute to chase the latest trends in the field of ecotoxicology.

In 1962, the publication of Rachel Carson's seminal volume, *Silent Spring* catalyzed the separation of environmental toxicology from classical toxicology. This separation continued into ecotoxicology, the direct descendant of environmental toxicology. The revolutionary element in Carson's work was her extrapolation from single-organism effects to effects at the whole ecosystem and the "balance of nature." Ecotoxicology is a relatively young field that was first defined by Rene' Truhaut in 1969 as "the branch of toxicology concerned with the study of toxic effects, caused by natural or synthetic pollutants, to the constituents of ecosystems, animal (including human), vegetable and microbial, in an integral context (Bard, 2008)." Ecotoxicology includes environmental science, environmental chemistry, biology, molecular biology, physiology, environmental law and so on. It has been regarded as a critical and cutting edge field. From the standpoint of environmental compartments, ecotoxicology can be divided into aquatic ecotoxicology, terrestrial ecotoxicology, soil ecotoxicology, and sediment ecotoxicology mainly.

The impetus for this new science was the need to understand and make decisions about environmental contaminants. Between World War II and the 1960s, several pollution events occurred with consequences universally acknowledged to be unacceptable. These watershed events included population crashes of raptor and piscivorous bird species due to DDT's effects on reproduction, widespread water pollution, and high-profile incidents of heavy metal poisoning like the Minamata disaster. Expertise for dealing with such issues became critical to society and several practical sciences coalesced into the nascent science of ecotoxicology.

Furthermore, ecotoxicology can be applied to access the enrichment capacity of lipophilic persistent organic pollutants through the food chain. Nowadays, the given reference limits RfD of majority of environmental compounds are determined by the no observed adverse effect level (NOAEL) or LOAEL identified by toxicity study of animal experiments. This is unreliable calculating the health risks combined with uncertainty factors. Due to the differences between human and animals, it may be unsuitable for researchers to use animal models to study human diseases. Pharmacokinetic studies at present are difficult to simulate the background exposure and chronic low-dose exposure. Epidemiologic studies, as a new type of toxicology, playing a very important role in the search of sensitive biomarkers for health risk assessment. The exposure assessment can estimate the degree of people exposure on a kind of chemical pollutants.

Ecotoxicology is a comprehensive science that combines causal explanations and information from many sciences, particularly molecular biology, Oimcs, biogeochemistry, ecology, and mammalian, aquatic, and wildlife toxicology. The integration of paradigms and data from these disciplines is presently incomplete. Chief among remaining challenges is establishment of congruency among theories and data emerging from different levels of biological organization. Because ecotoxicology is an applied science, ecotoxicologists take on different roles that are also not fully integrated. Some ecotoxicologists are concerned chiefly with scientific goals, that is, organizing facts around explanatory principles. Others focus on the technical goals, that is, developing and applying tools to generate high-quality information about ecotoxicological phenomena. Still others focus closely on resolving specific, practical problems such as assessing ecological risk due to a chemical exposure or the effectiveness of a proposed remediation action. Associated activities overlap but are presently integrated inconsistently in many instances. For example, the ecotoxicity tests applied today focus on effects at molecular, cell, organ, system and individual organism's levels, but predictions of consequences to populations and communities are a very high priority for ecotoxicologists. A major theme in ecotoxicology today is finding the best way of achieving scientific, technical, and practical goals while organizing a congruent body of knowledge around rigorously tested explanations.

Some general trends exist in ecotoxicology relative to different levels of the biological hierarchy (Fig. 1). Causes of lower-level phenomena such as biomolecular effects tend to be easiest to identify and relate to immediately adjacent levels such as to cells or tissues. Techniques for study of lower-level effects often have the advantage of documenting quicker responses than those occurring at higher levels. Unfortunately, the ecological relevance of change at the lowest levels is more ambiguous than that for higher-level changes. This creates a dilemma for ecotoxicologists attempting to develop better technologies, such as Omics. The ecotoxicologist tries to avoid measuring precisely the wrong effect or imprecisely the right effect. Another trend is that lower-level effects tend to be used proactively and those at higher levels are applied reactively by ecotoxicologists trying to solve specific problems. Lower-level effects tend to be more tractable than those at the higher levels.

Ecotoxicologists rely on conventional methods although these conventions vary among scientists focused at different levels of organization. Controlled experiments tend to be practiced more during studies of lower-level effects such as biochemical shifts; whereas, while higher-level effect studies rely more on observation and natural experiments such as accidental toxic releases. Regardless of the manner in which insight is obtained, the scientific intent is to organize facts around rigorously tested paradigms. Some ecotoxicologists work to produce more precise or detailed information around existing paradigms while others work to rigorously test existing paradigms or to propose novel ones. Both of these activities are essential to the growth of ecotoxicology as a science.

Activities to develop new or enhance existing technologies produce tools with which to understand the effects and fates of contaminants in the biosphere. Ecotoxicological technologists develop analytical instruments, procedures for studying regional impact, and specific tools for documenting exposure or effects. As examples, biomarkers are continually developed and improved for documenting effects from the molecular to individual level. Biomarkers consist of measurable molecular, cellular, tissue, body fluid, physiological, or biochemical changes in individuals. These biomarkers can quantify the effect of exposure even in the absence of any discernible adverse effect. At higher levels of organization, biomonitors, changes in organisms or groups of organisms, can be used to infer adverse impact of contaminant exposure. Qualities valued in ecotoxicological technologies are biomarker or biomonitor effectiveness (including low cost and ease of application), precision, accuracy, appropriate sensitivity, consistency, and capacity to generate clear results.

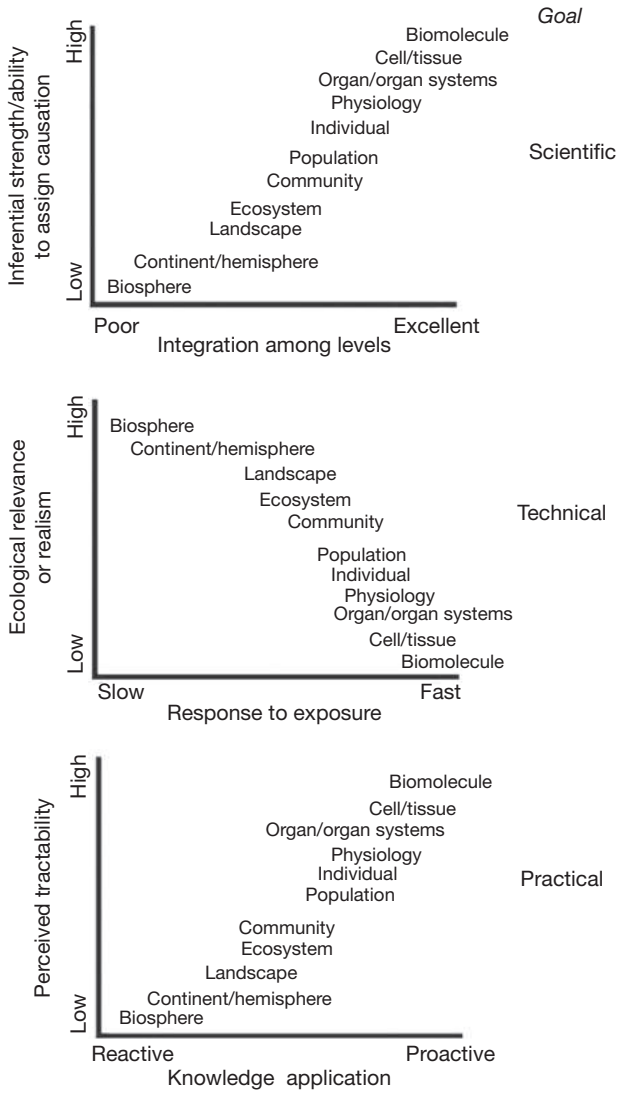


Fig. 1 Qualities of ecotoxicological knowledge based on the three goals to which it is applied and the biological levels for which it is generated.

The ultimate goal of ecotoxicology is to reveal and predict the effects of pollution within the context of all other environmental factors. Based on this knowledge, the most efficient and effective action to prevent or remediate any detrimental effect can be identified. In those ecosystems that are already impacted by pollution ecotoxicological studies can inform as to the best course of action to restore ecosystem services and functions effectively. Much of practical ecotoxicology is currently done within the Ecological Risk Assessment (ERA) framework (Fig. 2). ERAs can be retroactive, predictive, or comparative. A retroactive ERA estimates the risk from an existing situation such as a contaminated site, whereas a predictive ERA predicts the same for a future situation such as the proposed licensing of a new agrochemical. A comparative risk assessment might be done if the risks of two or more alternative actions are contrasted during environmental decision-making. Regardless of the type of ERA, the best available science is applied to formulate the plausible consequences of future action. The best available science and technologies are applied next for ecological effects and exposure characterizations.

Due to limited methods of toxicology experiments, the common application usually only focusses on the toxic effect in a particular field or a few molecular indicators. These limitations leave the methods open to questioning and challenging.

One method used in ecotoxicology is the adverse outcome pathway (AOP) (Fig. 3). Ankley et al. proposed the concept of the adverse outcome pathway framework and first and applied it to ecotoxicological research and chemical risk assessment. The AOP linearly links existing knowledge along a series of causally connected key events (KE). The AOP framework starts with a molecular initiating event (MIE) and leads to an adverse outcome (AO) that occur at a level of biological organization relevant to risk assessment. AOP related the structure of chemical pollutants, toxic effects, and the harmful outcomes of biological toxicity. This framework provided a new model to decipher the effects of chemical pollutants on populations of different species in the environment. AOPs offer a powerful approach to collect, organize and generalize toxicity-related information, allowing new

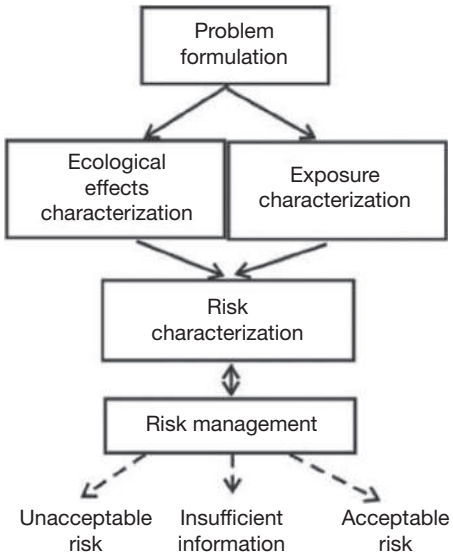


Fig. 2 The ecological risk assessment paradigm including problem formulation, exposure characterization, ecological effects characterization, and risk characterization. This paradigm was derived from that developed by the National Academy of Science.

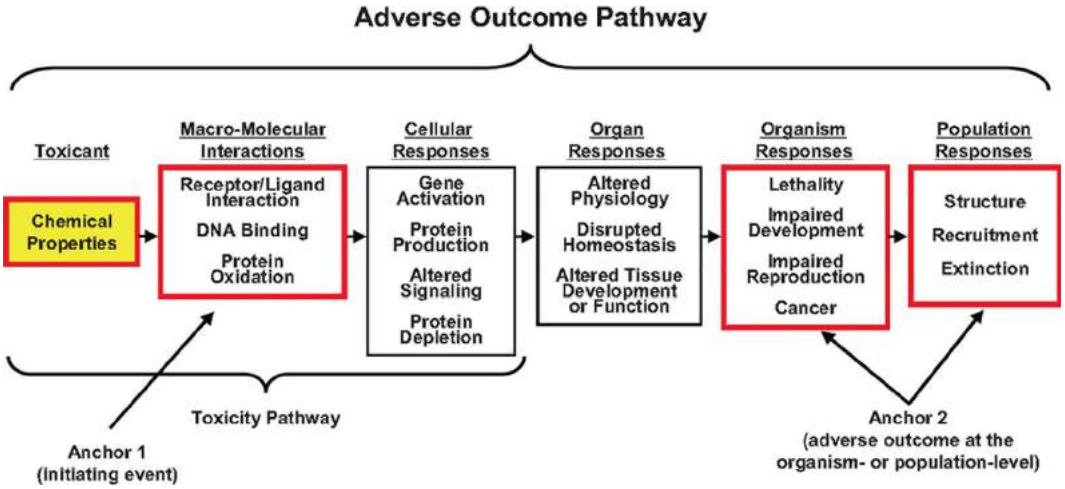


Fig. 3 Conceptual diagram of key features of an adverse outcome pathway (AOP). Each AOP begins with a molecular initiating event in which a chemical interacts with a biological target (anchor 1) leading to a sequential series of higher order effects to produce an adverse outcome with direct relevance to a given risk assessment context (e.g., survival, development, reproduction; anchor 2). The first three boxes are the parameters that define a toxicity pathway, as described by the National Research Council (cited from Ankley et al., 2010).

insights, guiding further research and targeted application of computational models, facilitating the development of alternative tests useful for practical risk assessment purposes and aiding prioritization of environmental exposures that warrant further toxicological assessment.

Additional methods of studying ecotoxicology, effect-directed analysis (EDA) and molecular toxicity identification evaluation (mTIE) are common in toxicology research. Chemical methods were first used to concentrate the pollutants in water. After this, indoor experiments are needed for biological research a concrete process consisting of four parts: Fractionation of toxicity test, Screening for toxic substances, Identification of induced toxic substances, and Confirmation of toxicity. The term mTIE refers to use acute exposure experiments to monitor and evaluate the toxic effects exposed directly on combined-pollution at gene levels. Recently, Altenburger et al. proposed a conceptual framework for combined toxicity studies making use of Omics approaches to support AOP development.

A computer model might predict movement of the contamination or the contaminant concentration might be measured in the relevant media. The best information is gathered to relate the contaminant concentrations to possible effects to valued ecological entities. All of this information is combined in the last stage of ERA to produce a risk statement. This information is shared with risk

managers who decide on the most appropriate action. This decision could mandate remediation, which would draw on ecotoxicological science and technologies.

Scales of Study

Ecotoxicological subject matter spans a wide range of biological levels. Levels from the molecular to the individual draw heavily on classical toxicology. Some population issues addressed by ecotoxicologists also benefit from the work of epidemiologists. Ecological issues at these levels of organization could be described as autecotoxicology, just as similar subjects in ecology are classified as autecology. Issues at higher levels would then be within the realm of synecotoxicology. Considerable biogeochemical and ecological knowledge are applied in synecotoxicological studies. A brief sampling of major autecotoxicological and synecotoxicological research approaches as follows.

Biomolecule

Information from the molecular level of organization is key to elucidating molecular mechanism of toxicity, differences in sensitivity among individuals, and adaptation of populations to contamination. Understanding the molecular mechanism of toxic action also helps to predict how contaminants might affect exposed individual. From a technological vantage, biochemical shifts are frequently used as evidence of toxic exposure or effect. Perhaps the best illustration would be the biomolecular shifts involved in phase I and II reactions of organic contaminants. The levels of associated biomolecules can quickly increase during exposure. Type I reactions are mediated by enzymes that catalyze contaminant hydrolysis, reduction, or oxidation, producing more reactive metabolites. The metabolites might be more readily eliminated from the cell or participate in phase II detoxification reactions. The best-studied phase I system is cytochrome P-450 monooxygenase which transforms contaminants such as polycyclic aromatic hydrocarbons (PAHs), chlorinated hydrocarbons, polychlorinated biphenyls (PCBs), hydrocarbons, dioxins, and dibenzofurans. Although phase I transformations are intended to facilitate detoxification, some transformed contaminants are more toxic than the parent compounds or might even be carcinogenetic. Phase II enzymes facilitate conjugation, that is, the addition of endogenous groups to contaminants or phase I metabolites which make the compounds more water soluble and readily eliminated. Other biomolecules provide mechanistic insight and a foundation for biomarker technologies. Elevated concentrations of metallothioneins, cysteine-rich proteins that bind and sequester metals, are often employed as evidence of heavy metal exposure. Also, commonly used as biomarkers are stress proteins, proteins induced by chemical stressors that function to reduce protein damage.

In the past, antoxidase enzyme activities are applied to ecotoxicology at the level of the cellular, and DNA damage is also applied at molecular level frequently. Biomarkers are biological macromolecules in organisms, whose changes can indicate a response to exogenous substances. For example, the Comet Assay can detect DNA strand breaks and, with the evaluation of other biomarkers, it can determine the effects of contaminants in exposed organisms. There were some positive correlations between the results given by the Comet Assay and other biomarkers such as enzyme activities. Gene expression analysis is another powerful tool to investigate the underlying mechanisms of toxicant exposure in nontarget organisms. Reverse transcription real-time quantitative PCR (RT-qPCR) has been established as a method of choice for the quantification of mRNA transcripts of a selected gene of interest (GOI) in biological samples. Although the defense mechanism of organisms respond to exogenous pollutants can be received, the big challenge is toxic effects of low dose relationship are hard to get.

The recent surge in genetic and Omics technologies provide another suite of biomarkers. Changes in DNA, RNA, protein products, and cellular metabolites are used separately or together to reveal mechanism of response or damage, and to document effects at the molecular level. Researchers increasingly apply omics technologies to carry out ecotoxicology studies to solve this boundedness, including the integrated study of transcriptomics, proteomics, metabolomics and other biological-omics subdisciplines. The major advantage of using omics methods in a systems approach is that one can assess hundreds to thousands of molecular responses simultaneously within an organism, facilitating a more holistic understanding of the organism's physiological status. Transcriptome research mainly probes the gene expression and the associated changes in biochemical pathway in the biological process, capturing the physiological response at the molecular level. This potentially reveals the mechanism of biological toxicity and population damage caused by environmental pollutants. Proteomics can reflect the function of the expressed genes effectively, including the interaction of protein to peptide and protein to protein. Proteomics integrated with bioinformatics can evaluate the response of the biochemical reactions of organisms exposed on environmental pollutants. Metabolites of abiological systems (e.g., cell, tissue or organism) exposure on environmental disturbance can be used to analyze the overall toxicity effect. Metabonomics can be applied to find a biomarker exposure on pollutants and analyze the metabolic pathways. Notably, High-throughput screening platforms have been demonstrated as a useful alternative to traditional in vivo experiments for drug development and agrochemistry.

Additionally, Molecular and ionic qualities of contaminants also influence the nature of exposure. At the same time, an organism's exposure to the same amount of a contaminant under different conditions can result in different realized doses and consequences. For example, the free ion form of a dissolved metal is considered the most bioactive. Metals dissolved in water form complexes with ligands such as dissolved inorganic anions and natural organic compounds. Depending on the ionic composition of the water, the same amount of dissolved metal will result in different concentrations of free ion and, consequently, concentrations of bioactive metal. Similarly, some organic contaminants are weak acids. If ingested with food, the capacity of such

contaminants to pass through the gut wall and cause harm is dependent on the amount of unionized compound present. Unionized compounds are generally more amenable to passage across the gut wall than ionized compounds. Under different pH conditions, different amounts of such a weak acid would be unionized as can be easily estimated with the Henderson–Hasselbalch relationship.

Cells and Tissues

Toxin-induced changes in cells and tissues are useful biomarkers. For practical and technical reasons, blood is the most common choice. However, tissues and cells like gills, sperm cells, early larval stages, coelomocytes, liver, and kidney have been also used for biomarker work. Due to the reduced use of experimental animals, low cost and rapid performance, the hemocyte of mollusks has become one of the most-used cell for biomarker work. Some changes used for biomakers reflect a cell's failure to remain viable in the presence of toxicants and others reflect partially successful attempts to maintain homeostasis. For example, histological examination of the liver from an exposed organism might reveal many necrotic cells. In the same tissues, inflammation might occur in an attempt to isolate, remove, and replace damaged cells. Both necrosis and inflammation are common biomarkers. Other changes such as the cellular accumulation of damaged biomolecules or cells modified to cope with toxicant damage are also useful histological biomarkers. Cancer is a cellular response to carcinogen exposure that is carefully studied by ecotoxicologists. Several ecological studies have demonstrated the role of environmental contaminants on cancer etiology. For example, Puget Sound English sole (*Parophrys vetulus*) taken from sites with elevated sediment contamination showed high prevalence of liver cancers. Another case of elevated cancer prevalence (i.e., 27% of dead adults) involved beluga whales (*Delpinapterus leucas*) inhabiting a contaminated region of the St. Lawrence estuary. Exposure studies at this level focus on the routes of contaminant movement into and out of cells, and differences in accumulation in various tissues. Generally, contaminant movement into and out of cells involves (1) simple diffusion across the membrane lipid bilayer or through an ion channel, (2) facilitated diffusion involving a carrier protein, (3) active transport, or (4) endocytosis. Some of these routes are designed for other purposes such as ATPase active transport of cations but also reveal the movement of contaminants such as cadmium. Other mechanisms are more specific. For example, the multixenobiotic resistance (MXR) mechanism specifically removes moderately hydrophobic, planar contaminants from the cell.

Organ and Organ Systems

Toxic effect on organs and organ systems is another major theme in classical toxicology that also has a role in ecotoxicology. Organs can be targets of toxicant effects as in the case of the liver cancer mentioned above or can be routes of toxicant entry into the body as in the case of the integument, breathing organs, and digestive tract. Contaminant effects on organs and organ systems are diverse. For example, pyrethroid pesticides modify essential ion exchange across amphibian skin. Alternatively, exposure to low pH or high metal concentrations impact fish gills in such a way that normal ion and gas exchange are altered. In other cases, some contaminants (i.e., teratogens) can cause abnormal organ development. In aquatic biota, a large number of studies have revealed a broad range of pesticides representing a variety of chemical classes to induce embryotoxicity and teratogenicity in nontarget fish, amphibia, and invertebrates, which result in organ malformations, delayed hatching, growth suppression, and embryonic mortality. For example, fish embryos develop cardiovascular abnormalities if exposed to high concentrations of PAH. Other contaminants compromise immunological competency, increasing susceptibility to infection. These examples represent only a few of the possible organ or organ system effects of contaminants on nonhuman species.

An issue attracting considerable attention at the moment is the ability of some environmental contaminants to modify endocrine functions such as those essential for sexual development and viability, or optimal metabolic activity. For example, the presence of the antifouling paint constituent, tributyltin, caused pervasive imposex (imposition of male features such as a penis or vas deferens on females) in whelk populations along the English and Northeast Pacific coasts. Contaminants that act as estrogen include DDT and its replacement, methoxychlor; nonylphenol from surfactant and detergent synthesis; and synthetic hormones from birth control pills that enter waterways from sewage treatment plants. Still other endocrine disruptors such as ammonium perchlorate from military munitions disrupt thyroid function. Exposure studies at this level of biological organization emphasize target organs.

Some organs or organ systems are more prone to toxic impacts due to their close contact with environmental media, location relative to blood circulation, or specific function. In the classic example, the gills of aquatic organisms are often target organs for dissolved contaminants because of their intimate contact with the surrounding water. The liver or analogous organs in invertebrates are often sites of harmful effects because of their prominent detoxification function, that is, the liver cancer noted above in English sole was caused by contaminant activation during phase I reactions in the liver.

Whole Organism

Effects to individuals are used to make inferences about contaminant impacts on individual fitness, and indirectly, on populations and communities. The most commonly measured qualities are mortality, development, growth, reproduction, behavior, physiology, and bioenergetics. Lethal effects are measured under different exposure scenarios. These effects might be measured during acute (4 days or shorter) or chronic (longer than 10% of an individual's lifespan) exposures. Especially, the critical effect concentrations

are the most common indicators to quantitatively assess risks for species exposed to contaminants. Three types of critical effect concentrations are classically used: lowest/no observed effect concentration (LOEC/NOEC), x-percent lethal concentration (LCx), and no effect concentration (NEC) lethal effects might also be measured for contaminant exposure via different media such as water, air, food, and sediment. Most are studied in the laboratory in such a manner that physical, chemical, and biological factors influencing response to exposure are controlled. Therefore, mortality predicted for a particular exposure concentration might not completely describe the mortality that would occur in the field where exposed individual must successfully forage, compete with individual of other species, avoid predators, and interact with individual of the same species in order to remain alive. Exposure studies that involve whole organism emphasizes bioaccumulation, the net accumulation of contaminant in an organism from water, air, or solid phases of its environment.

The thematic areas of its current employment in the evaluation of genetic toxicity are vast, either in vitro or in vivo, both in the laboratory and in the environment, terrestrial or aquatic. Genetic toxicity techniques have been applied to a wide range of experimental models: bacteria, fungi, cells culture, arthropods, fishes, amphibians, reptiles, mammals, and humans. Furthermore, there is a significant species difference in the spectrum of toxicity observed, for example, the LD_{50} for acute TCDD exposure varies from 1 $\mu g\ kg^{-1}$ in the guinea pig, 20–40 $\mu g\ kg^{-1}$ in the rat, 114 $\mu g\ kg^{-1}$ in the mouse and rabbit, and 5000 $\mu g\ kg^{-1}$ in the hamster. The diversity of AhR pathway genes and the species difference of the complicated regulation process of toxicology in different animals may throw light on the history of early molecular evolution. Fishes are clearly the most utilized group, reflecting their popularity as bioindicator models, as well as specific concerns over the aquatic environment health. Amphibians are among the most sensitive organisms to environmental changes, mainly due to an early aquatic-dependent development stage and a highly permeable skin. Moreover, in the terrestrial approach, earthworms, plants or mammalians are excellent organisms to be used as experimental models for genotoxic evaluation of pollutants.

Population

Most of the experimental testing in ecotoxicology takes place at the individual level, but the protection goals for environmental risk assessment are at the population level (or higher), population modeling can fill this gap. But only models on a mechanistic basis allow for extrapolation beyond the conditions in the experimental tests. The life-history traits of individual form the basis of population dynamics, and population modeling thus requires a detailed understanding of the individual's behavior. The dynamic energy budget (DEB) model offers a flexible platform for the development of model at the individual level. Applying DEB model to population model can provide a mechanistic basis for extrapolation. Some of these pesticide effects at the population level have at most been plausible linked to subindividual or individual effects by the application of Bradford-Hill's criteria of causation. For example, in birds, population effects of pesticides have been linked to neurotoxicity and endocrine disruption. It is well known that DDT and its metabolite DDE had a devastating effect on many bird species due to a reduction of eggshell thickness of up to 90% and consequently, cracking. Additionally, a growing number of ecotoxicologists study population-level effects directly. Such studies emphasize vital data such as birth, death, stage change, or migration rates. Demographic models based on vital rates improve our ability to project consequences such as a drop in the population growth rate or increase in local population extinction risk. Some population studies treat the population as one in which individual is uniformly distributed in the area of interest but others consider the population as a metapopulation composed of subpopulations inhabiting habitat patches of different qualities, including varying levels of contamination. The differences in vital rates, including exchange rates among patches, are used to project contaminant exposure consequences. With metapopulation models, effects at a distance from a contaminated patch can be explored. For example, a population member can be exposed in one patch yet the effects might manifest in an uncontaminated patch after migration.

Community

Community ecotoxicology explores the consequences of contaminant exposure of and movement of contaminants within ecological communities. The majority of such studies are field studies either addressing scientific questions or applying knowledge to assess risk or define remediation action for a contaminated system. Like the biomarkers applied at lower level of biological organization, bioindicators are applied by community ecotoxicologists. Bioindicators might be particularly sensitive species whose absence suggests an adverse impact. A community metric such as species richness, evenness, or diversity might also be used as an indicator of an adverse exposure consequence. Any study in which biological systems are applied to assess the structural and functional integrity of ecosystems is referred to as a biomonitoring study. Exposure within communities is often explored in the context of contaminant trophic transfer. Depending on its properties, a contaminant can increase (i.e., biomagnification), decrease (i.e., trophic dilution), or not change in concentration with progression through a food web. For example, the most detrimental effects of herbicides in aquatic systems stem from the reduction of the complexity and structure of the plankton and the submerged vegetation, including periphyton. These algae and plants all act as food sources and refuges for phytophagous species such as water birds and amphibian tadpoles. Furthermore, Species of higher trophic levels, such as salmon, are most likely to be affected in population growth and productivity by indirect pesticide effects.

Additionally, contaminants such as methylmercury or persistent organic pollutants (POPs) such as DDT biomagnify. Biomagnification can lead to adverse consequences to higher trophic level species such as the raptors and other piscivorous birds. Studies of

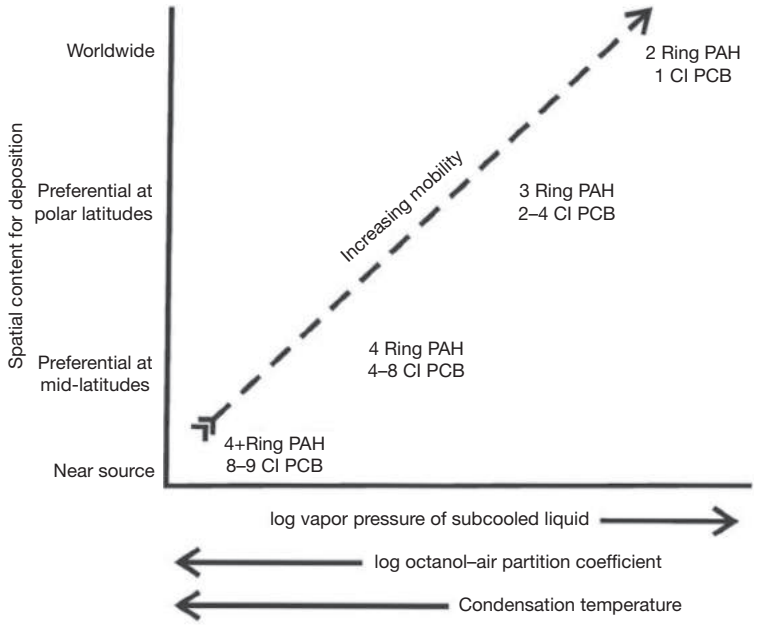


Fig. 4 Global movement of a POP is determined by several qualities including its (subcooled liquid) vapor pressure and condensation temperature which, together, determine its tendency to move into and remain in the atmosphere. A more volatile compound will be transported more by atmosphere movement than one that is less volatile. Its condensation temperature influences the latitudinal limits to which it might move. The octanol–air partition coefficient is also important because it reflects the POP's tendency to remain associated with the solid and liquid phases of the Earth versus the atmosphere. Here, polycyclic aromatic hydrocarbons (PAHs) with different numbers of aromatic rings and polychlorinated biphenyls (PCBs) with different numbers of chloride atoms are used to illustrate trends for global deposition of POP. This biospheric process is called global distillation.

food webs including omnivorous members require a measure of trophic status for species. A convenient measure of trophic status is provided by the nitrogen isotopic fractionation that occurs with each trophic exchange.

Ecosystems

Ecosystem-level studies vary widely in their spatial and temporal scales. Often ecosystem modeling techniques are applied to an easily definable ecosystem such as a contaminated lake or watershed. Fate and movement of contaminants are then modeled by computer or measured in extensive sampling programs. Larger-scale studies are required for contaminants prone to wide spatial dispersion via atmospheric transport. This category includes contaminants like mercury from coal power plants or contaminants used widely by society such as atrazine, an herbicide applied in the North American Corn Belt. For example, the results of mass-balance ecosystem model by Ecopath with Ecosim (EwE) approach suggested that salmon inputs specifically control PCB concentrations in stream-resident fish whereas Hg concentrations are more strongly influenced by diffuse background sources, and there exists species-specific differences in diet and growth, along with trophic pathways, can influence the magnitude of contaminant impacts by spawning salmon.

Often geographical information system (GIS) and remote sensing technologies are essential in these types of studies. In still other instances, a global perspective is required to adequately grasp the ecotoxicological consequences of contaminants. Current global issues are ozone depletion in the stratosphere due to chlorofluorocarbon (CFC) release, global warming due to greenhouse gas emissions, and global movement of persistent organic pollutants (Fig. 4). More and more frequently, large-scale issues are emerging as critical ones in ecotoxicology.

Current Trends in Ecotoxicology

A working knowledge of the movement and effects of environmental contaminants is recognized worldwide as essential to maintaining an acceptable quality to life. Ecotoxicology has emerged as the applied science that addresses the central issues of contaminants in the biosphere. Major challenges in this young science include the following: (1) the integration of causal explanations and knowledge arising at different levels of biological organization into a coherent whole; (2) integration of scientific, technical, and practical goals of ecotoxicologists; (3) consideration of ecotoxicological issues at increasingly wider spatial and longer temporal scales; (4) transformation from high-dose into low-dose effect; (5) the toxic effects of mixtures of contaminants; (6) the specify, effectivity and sensitivity, of biomarkers; (7) standardize experimental conditions in ecotoxicology; (8) "Omics" and bioinformatics in (eco)toxicology. Additionally, general improvements in technology are needed to develop the research of ecotoxicology.

Further Reading

Ankley GT, Bennett RS, Erickson RJ, Hoff DJ, Hornung MW, Johnson RD, Mount DR, Nichols JW, Russom CL, and Schmieder PK (2010) Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. *Environmental Toxicology and Chemistry* 29: 730–741.

Bard SM (2008) Ecotoxicology: The focal topics. In: *Encyclopedia of Ecology*, 1st edn., pp. 1194–1195. Amsterdam, The Netherlands: Elsevier.

De Lapuente J, Lourenco J, Mendo SA, Borras M, Martins MG, Costa PM, and Pacheco M (2015) The comet assay and its applications in the field of ecotoxicology: A mature tool that continues to expand its perspectives. *Frontiers in Genetics* 6: 1–20.

Hanson ML, Wolff BA, Green JW, Kivi M, Panter GH, Warne MSJ, Ågerstrand M, and Sumpter JP (2017) How we can make ecotoxicology more valuable to environmental protection. *Science of the Total Environment* 578: 228–235.

Henry TB (2015) Ecotoxicology of polychlorinated biphenyls in fish-a critical review. *Critical Reviews in Toxicology* 45: 643–661.

Jager T, Barsi A, Hamda NT, Martin BT, Zimmer EI, and Ducrot V (2014) Dynamic energy budgets in population ecotoxicology: Applications and outlook. *Ecological Modeling* 280: 140–147.

Koehler H-R and Triebskorn R (2013) Wildlife ecotoxicology of pesticides: Can we track effects to the population level and beyond? *Science* 341: 759–765.

Newman, M C. Ecotoxicology: The History and Present Directions. *Encyclopedia of Ecology*, 1st edn., 1195–1200. Amsterdam, The Netherlands: Elsevier.

Mikó Z, Ujszegi J, Gál Z, and Hettyey A (2017) Standardize or diversify experimental conditions in ecotoxicology? A case study on herbicide toxicity to larvae of two anuran amphibians. *Archives of Environmental Contamination and Toxicology* 73: 1–8.

Snape JR, Maund SJ, Pickford DB, and Hutchinson TH (2004) Ecotoxicogenomics: The challenge of integrating genomics into aquatic and terrestrial ecotoxicology. *Aquatic Toxicology* 67: 143–154.

Werner I, Aldrich A, Becker B, Becker D, Brinkmann M, Burkhardt M, Caspers N, Campiche S, Chèvre N, and Düring RA (2016) The 2015 annual meeting of SETAC German language branch in Zurich (7-10 September 2015): Ecotoxicology and environmental chemistry-from research to application. *Environmental Sciences Europe* 28: 1–12.

https://en.wikipedia.org/wiki/Ecotoxicology.

Endangered Species

P Kareiva and J Floberg, The Nature Conservancy, Seattle, WA, USA

Introduction

Biologists are concerned that the world is facing an extinction crisis, in which a large fraction of the Earth's species will disappear by 2100. The passenger pigeon, which numbered in the billions and was once the most abundant bird in North America gradually went extinct under our watch, with the last individual dying in the Cincinnati zoo in 1919. The Northern Right whale, once one of the North Atlantic's most important marine species, now numbers only 200–300 individuals, and the death of just one female is newsworthy. Although scientists agree that an unprecedented number of species are at risk of imminent extinction, and that the rate of biodiversity loss rivals the great extinction events of geological history, the science of extinction faces several big unknowns. Most notably, although scientists have named and described about 1.75 million species, they have still barely scratched the surface of Earth's total biological diversity. In fact, about 15 000 new species are described each year, and these discoveries are not limited to obscure or small-bodied groups – for example, a new species of large mammal is discovered approximately every three years. Clearly, it is hard to know how many species are going extinct when we do not even know how many species there are. It is not surprising then that estimates of extinction rates vary by an order of magnitude, ranging from lower than 10 species lost per day to over 100 species lost each day.

Even though quantifying global extinction rates is highly uncertain, we have a good understanding of what factors are likely to make a species endangered. Moreover, as nations have become alerted to the extinction crisis, international institutions and governments around the world have started to identify endangered species (species at risk of going extinct in the near future) and many countries have laws that afford special protection to these at-risk species. Our scientific understanding of species at-risk or endangered species comes from four main arenas: (1) our theoretical understanding of the population biology of small populations, (2) statistical studies relating the number of resident species to habitat availability and geography, (3) monitoring and tracking of endangered species in the United States, which was the first country to pass a law protecting endangered species and has since accumulated over 30 years of data on its endangered species, and (4) global lists and analyses of endangered species compiled by the International Union for the Conservation of Nature and Natural Resources (IUCN). If we can develop a predictive science of endangered species, then there is hope of avoiding the demise of the thousands of at-risk species currently on the brink of extinction throughout the world.

Predicting Future Extinction and Factors That Make Species Endangered

Dramatic statements regarding the loss of biodiversity have become commonplace in both the scientific literature and the popular media. However, if you look closely at these claims, you will see that, although they are consistently pessimistic, they do not seem to agree about exactly how bad the current extinction crisis is. Estimating extinction rates is complicated and involves a series of assumptions and educated guesses. Rather than attempting to count actual extinctions, most researchers use indirect methods to estimate extinction rates. Habitat destruction is widely considered to be the major cause of extinction and endangerment of species. Unlike other major causes of endangerment, such as pollution, exploitation, introduced predators, and disease, habitat destruction (typically deforestation for the terrestrial environment) is relatively easy to quantify over vast land areas using satellite imagery. Given an estimate for the rate of habitat loss, researchers might then ask `how many species are expected to become extinct with each million hectares of habitat destroyed?' To answer this question, one must know something about the relationship between habitat area and species diversity. Species–area curves show the relationship between the diversity of species and the physical size of an island, habitat patch, or sampling space. The general pattern is that species initially accumulate rapidly as area increases, but the rate at which new species are added gradually declines, until eventually all of the species present in a region are accounted for. This saturating relationship between diversity and area is one of the most robust patterns in all of ecology. Species–area curves have been successfully applied to many different taxa in a wide variety of settings and terrestrial habitats. The implication of the general shape of this relationship is that there will initially be few consequences of habitat destruction (i.e., a small percent of species will become extinct), but as the process continues and more and more habitat is destroyed, the impact on biodiversity becomes more extreme (**Fig. 1**). The relationship between species diversity, S, and area, A, is described by the equation

$$S = cA^z$$

where c is a constant reflecting the number of species in a habitat of area A, and z is the rate of species accumulation as A increases. Estimates of z are crucial if one wants to predict the impact of habitat destruction on diversity. Larger z values are typical of true oceanic islands and for species with relatively limited movement, while smaller values are common for virtual islands that really are areas embedded within a continuous landscape (such as parks surrounded by agriculture) and for highly mobile species.

 doi:10.1016/B978-0-444-63768-0.00487-X

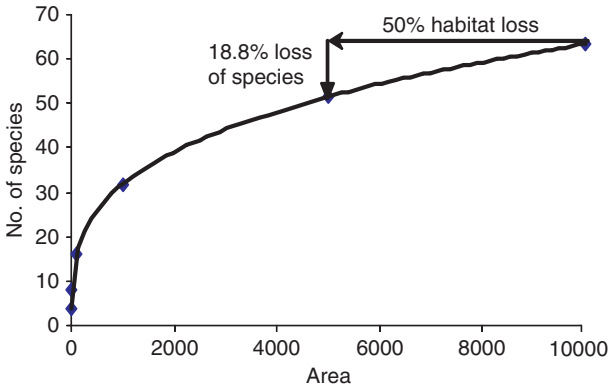


Fig. 1 A species–area curve ($S=cA^z$), where S is the number of species in habitat of area A, and c and z represent constants that depend on the type of species being considered and the type of habitat involved. In the curve shown $c=4$ and $z=0.3$. Using such a curve, one can estimate extinction of species associated with any fraction of habitat loss.

Although we may not be sure of the value of z in the species–area curve, the idea that habitat conversion or destruction causes loss of species is well-documented. This means that even without having data on particular species, ecologists can anticipate that species associated with particular habitats will be endangered if too much habitat is lost. For example, even though there are not counts of plants and animals declining where forests are logged, we can still be sure that some of the forest residents are endangered by habitat loss. The mechanism by which habitat loss endangers species is reduction in population size. If there is less habitat for a species, then that species must have fewer individuals. A small population size endangers a species for several reasons which we discuss below. But first it is worth recognizing that the tendency to use the phrases `habitat loss' and `habitat destruction' can be misleading. Habitat does not disappear – it is converted to a different type of landscape, such as farmland or logged forests with remnant patches of woodland left standing. Historical deforestation events have generally not produced the magnitude of extinction one would expect from a simple application of species–area curves. Between 90% and 99% of Puerto Rico was logged, yet only 10% of the resident terrestrial bird species in that forested habitat went extinct. Similarly at least 90% of the forests in eastern United States were logged, yet only three forest-dwelling birds went extinct. Although these percentages are qualified by the awareness that we have only recently been able to adequately track whole taxa such as birds for extinctions, it is still true that the risk of extinction can be tempered by the quality of converted habitat, which may not always be as uninhabitable as a species–area curve implies.

Population size is the best indicator of a species' endangerment. First, small populations are at greater risk of extinction because they are more susceptible to chance misfortune such as failing to mate successfully or to survive a winter in spite of there being an average likelihood that individuals in the population should succeed. For instance, if we flip a coin 100 times we will probably get roughly 50 `heads' and 50 `tails'. But if we flip a coin only ten times, there is a greater chance of getting a more lop-sided result, such as 10 `tails' – and if tails meant a chance death, then clearly a small population size is a great disadvantage. These effects are referred to as demographic stochasticity and usually apply to populations of reproductive females substantially less than 100. Of course, environments fluctuate due to droughts, heat waves, or unusually harsh winters, and species we normally think of as being much more numerous than 100 can be driven to unusually low numbers by temporarily poor environmental conditions, and then become at risk of extinction due to demographic stochasticity. Mathematical models of population extinction in fluctuating environments suggest that populations smaller than 1000 are likely to be at substantial risk because of a combination of fluctuating environments and demographic stochasticity. A second unfortunate attribute of small populations is that they usually occupy small areas. Any time a population is restricted to a small area it is at risk of being exterminated by one catastrophic event such as a wildfire or a large flood.

For some species there is a positive feedback whereby a decline in population size also causes a predictable (as opposed to chance) decline in survival or reproductive success of the remaining individuals. When reductions in population catalyze further population decline, it is called an Allee effect. Allee effects can occur in certain species that experience difficulty in finding a mate when at low density. Imagine, for example, the difficulties that a whale might have in finding a mate if there are only a few hundred whales left throughout an enormous ocean. One reason the Northern Right whale is not recovering, even though all hunting on it has been halted, is that mates cannot locate one another in the vast ocean. Allee effects have been well documented in certain small populations of plants that rely on insect pollinators such as bees. Bees often specialize on visiting abundant plant species because they can more efficiently gather pollen and nectar if they keep visiting the same common flower; as a result a plant species that becomes too scarce may be neglected by pollinators and hence fail to reproduce. Allee effects can also arise when social interactions become impaired at low population size. For example, animal species that rely on cooperative group hunting may be incapable of bringing down large prey if the group becomes too small.

Lastly, small population size can imperil a species because of genetic effects. Individuals in small populations may have no choice but to mate with related individuals. Mating with kin is called inbreeding, and inbreeding can severely reduce the survival or reproductive fitness of any offspring that are produced (a phenomenon called inbreeding depression). Inbreeding tends to

increase homozygosity because two related individuals are more likely to share the same version of a gene than are two unrelated individuals. Offspring that are homozygous for many genes may be worse off than individuals that are highly heterozygous for a couple of reasons. First, heterozygosity (having two different alleles for a particular gene) can be beneficial if the two alleles result in the production of two different versions of a protein. For certain genes, producing multiple versions of proteins can be correlated with greater physiological flexibility or improved immune system function. Homozygotes lose this advantage. Second, many harmful versions of genes are recessive or masked by the presence of a healthy version of the gene inherited from the other parent. But an individual that is homozygous for a harmful allele will express the genetic disorder. Examples of human genetic disorders caused by recessive alleles include cystic fibrosis, tay-sachs, and phenylketonuria. Both of these mechanisms can cause reduced fitness of inbred offspring, and reduced fitness of individuals will translate into reduced population growth and increased risk of extinction for the population as a whole.

Even in the absence of inbreeding, small populations tend to lose genetic diversity over time. Unique alleles (versions of genes) are lost due to genetic drift. Moreover, mutations, which are the source of new alleles, are rare and therefore fewer new alleles are generated in small populations. This loss of genetic diversity can be devastating to a population. Unlike natural selection, which increases the frequency of adaptive or beneficial alleles, genetic drift is completely random with respect to adaptiveness. In other words, either good or bad versions of genes may be lost, just by chance, and the odds of these random losses are higher when a population is small.

The rate at which genetic diversity is lost from a population does not depend on the total population size, but rather on what is called the effective population size. Effective populations are always less than the actual population because they refer to a concept called an ideal population. An ideal population is one with a 50:50 sex ratio, every individual being equally successful at mating and reproducing, and the same population size year-after-year. A population that departs from these ideal conditions has an effective population that is lower by an amount that is roughly proportional to the degree to which the ideal conditions are violated. Population genetics theory provides models for estimating effective population size given information on how a population departs from the ideal conditions. The reason the effective population size of any species is inevitably less than the numerical abundance of species is because sex ratios commonly depart from 50:50, the number of matings and offspring per parent is usually variable in wild populations, and populations in nature fluctuate. The important point is that species with small effective population size lose genetic diversity far more rapidly than species with large effective populations. Although small populations are always more at risk than larger populations, species that are represented by chronically small populations often do not show severe genetic inbreeding. These `naturally' and chronically small populations may have evolved mating systems and genetic mechanisms that have reduced inbreeding depression in comparison to large populations that are suddenly driven to low numbers by human activities. This distinction between chronically small populations and newly small populations is one reason that criteria for identifying which species are at risk typically factor in trends in abundance and recent declines in addition to population size.

When talking about endangered species it is important to note that anything that makes populations small imperils a species. Even if the cause of a species decline is halted because fishing or hunting has been halted, or the last remaining habitat is protected, a small population may still be doomed to extinction. Once populations shrink below a certain threshold level they enter an extinction vortex in which inbreeding depression, chance misfortune, disrupted social systems, and reduced genetic variability all conspire against their persistence and make the situation worse with each new generation. In those circumstances captive breeding may be the only hope. In short, small population size equals an endangered species, and declining populations can also equal an endangered species because prolonged declines will inevitably produce small populations. One key question is how small does a population have to be to be endangered? The answer depends on the mating system (do a few males get all the females? how uneven is reproductive success?) and rate of reproduction. Certainly any population below 1000 is at risk of extinction, and populations as high as 10 000 can be highly endangered if most of the individuals are males or too young or too old to reproduce.

Endangered Species in the United States

The USA was the first country to give endangered species legal status and in so doing helped define the science of endangered species. As early as 1964 the Department of Interior produced a list of 62 species at risk of extinction in the US. When the Endangered Species Act (ESA) was implemented in 1973 there were 392 endangered or threatened species on the list. The ESA itself defines endangered as a species "in danger of extinction throughout all or a significant portion of its range." Although the ESA does not give any population sizes or explicit scientific criteria, it is clear from the record of listings (when species are officially called endangered or threatened and afforded legal protection) that small population size is the dominant criteria. The median population size of species when listed is 1075 for vertebrates and 120 for plants. When one realizes that all of these individuals will not be capable of reproducing (e.g., too young, too old, cannot find mate), those population sizes are disturbingly low if one is to have hope of recovering the species so that they are no longer endangered. As of 2006, there are 1272 endangered and threatened species in the USA with a very uneven taxonomic distribution. The percentage of species in the US listed as endangered or threatened are by major taxa: 19% for mammals, 15% for fish, 12% for bird, 11% for amphibians, 8% for invertebrates, and 5% for plants.

When species are listed as endangered or threatened, the US government requires a description of the major threats that are placing the species at risk of extinction. By reviewing the listed threats for all the US species it is possible to arrive at an estimate of the major sources of endangerment. As expected, habitat destruction is by far the most widely identified threat, affecting

approximately 85% of all the species listed as threatened or endangered in the USA. Alien or introduced species is the second most common threat (affecting 49% of listed species), followed by pollution (24%) and overexploitation (18%). Notably underrepresented as endangered species in the US are marine organisms. Only 39 marine species (and many of these are fish such as salmon and steelhead that spawn in fresh water) are listed as threatened or endangered in the US. Only one of these marine endangered species is an invertebrate – the white abalone in California. The paucity of marine species that are listed under the ESA is not a reflection that marine organisms are not endangered. Rather marine organisms are more poorly known taxonomically, and population data are generally lacking. They are under water and hence their plight is less conspicuous than is, for example, a large terrestrial bird. Most biologists think that with further study we will find many marine species that are endangered – often because of habitat degradation and pollution.

Patterns of Endangerment at the Global Level

The most objective and authoritative information on global species rarity can be found in the IUCN Red List of Threatened Species, which was first conceived in 1963. The IUCN, also known as the World Conservation Union, is an international conservation network including over 10 000 scientists and 1000 organizations from around the world whose mission is to influence societies to conserve their natural diversity. Every 4 years, the IUCN hosts a World Conservation Congress, where the current status of global species rarity is scientifically evaluated in a Global Species Assessment that is based on the IUCN Red List of Threatened Species. The Red List, which currently lists 7181 endangered or critically endangered species, includes a regularly updated online database and represents an ongoing effort to track the status of all species. The Red List evaluation of species in terms of their risk of extinction falls into three steps. First a species is taken up for evaluation – to date only 38 047 species have been considered for evaluation. Of those species considered, some have to be dropped because the data are too scarce (roughly 10% of the species IUCN has attempted to evaluate have run into problems because of data scarcity). Obviously, the initial selection of which species to evaluate is also biased towards species that might be at risk – in no way is the selection of species random. If the data are adequate, then the third step is taken where each species is assigned to one of seven categories: extinct, extinct in the wild, critically endangered, endangered, vulnerable, near threatened, and least concern. Species at risk of extinction (defined by IUCN as threatened) include the three categories of critically endangered, endangered, and vulnerable. The criteria used to assign species to different risk categories are intended to be quantitative and objective, and to focus on data that could be realistically obtained if sufficient effort were made. Although the criteria in **Table 1** describe specific numerical thresholds between categories and appear to provide a data-driven way to assign rarity among species, caution is advised. One weakness of the approach is that the criteria do not reflect life-history variation or idiosyncratic natural history, and instead represent a `one size fits all' approach. In other words, the intent is that these criteria are to be used to evaluate all species, excluding microbes. However a population of 100 fruit flies has a much greater potential for recovery than a population of 100 sea turtles that do not reproduce until they are 20 years old. Also species differ greatly in how flexible or adaptable they are in their habitat – a small population size represents a much greater risk

Table 1 IUCN criteria to evaluate species for threatened status

	Critically endangered	*Endangered*	*Vulnerable*
A. Trend of loss in total numbers of mature individuals in a taxon[a]	$\geq$80%	$\geq$50%	$\geq$30%
B. Limited geographic range, based on either			
extent of occurrence[b]	$<$100 km^2	$<$5000 km^2	$<$20 000 km^2
area of occupancy[b]	$<$10 km^2	$<$500 km^2	$<$2000 km^2
C. Limited remaining individuals, based on either	$<$250 and 25% decline	$<$2500 and 20% decline	$<$10 000 and 10% decline
(i) numbers of mature individuals with decline	Expected in 3 years or one generation[c]	Expected in 5 years or two generations[c]	Expected in 10 years or three generations[c]
(ii) numbers of mature individuals	$\leq$50	$\leq$250	$\leq$1000[d]
D. Expected extinction rate in the wild based on quantitative study[e]	$\geq$50%	$\geq$20%	$\geq$10%

[a]Trend over a period of 10 years or three generations, whichever is longer (up to a maximum of 100 years). A greater percentage loss is required to be critically endangered, endangered, or vulnerable ($\geq$90%, $\geq$70%, and $\geq$50%, respectively) where threats to loss are clearly reversible and understood and have ceased.

[b]Also require at least two of the following: severe fragmentation/few locations, continuing decline, or extreme fluctuations in numbers.

[c]Criteria are still satisfied for C (i) with any continuing decline if each subpopulation has few individuals, if there is a high percentage of the total number of individuals in any one subpopulation, or if there is an extreme fluctuation in numbers of mature individuals.

[d]Vulnerable status also requires the area of occupancy to be less than 20 km^2 or number of locations $\leq$5.

[e]Trend over a period of 10 years or three generations for critically endangered status, 20 years or 5 generations for endangered status, and for all categories within 100 years maximum.

A species is critically endangered, endangered or vulnerable by meeting at least one of A–D in the appropriate column.

for an inflexible species than a generalist species with ample physiological and behavioral plasticity. The implications of life-history differences can become a focal point of debate about whether species should be listed or not. For instance, some marine experts have argued that certain decline criteria in **Table 1** do not apply to so-called resilient fish stocks which can be sustained at a tiny fraction of their precatch levels. Secondly, the numerical data likely embody huge uncertainties. Endangered species by their very nature are few in number and are often difficult to count and assess because they frequently occupy fragmented and/or impacted habitats and can be scattered over large areas. Moreover, analyses that are used to determine the expected population declines or probabilities of extinction for species are potentially misleading for species because data quality is often poor or population distributions are not well understood. There is a constant tension in developing the Red List between ensuring sufficient data are available to assign appropriate status and erring on the side of caution to list poorly understood globally endangered species so that they receive appropriate conservation attention (**Table 2**).

One of the biggest challenges in addressing endangerment is our limited ability to prioritize conservation actions because we have incomplete and skewed geographic and taxonomic knowledge. Assuming a conservative 10 million total species in the world, only a fraction have been described, a much smaller fraction described and evaluated for risk, and an even tinier fraction upon evaluation is judged to have enough data to make a risk determination (**Fig. 2**). In addition to not knowing much, what we do know is highly biased so that some taxa are much better studied than others; for instance, whereas nearly 40% of the known vertebrate species have been evaluated for rarity, less than 5% of the plants species have, and less than 1% of the invertebrate species have, and essentially no fungi have. Marine and freshwater species are also much less understood than terrestrial species. For example, though marine and freshwater areas cover more than 70% of the Earth's surface, only 6% of described fish species have been evaluated for rarity. Among the well-studied taxa where species are known and have been evaluated for risk of extinction, endangerment varies five to tenfold (**Fig. 3**). There is also a wide range of endangerment within taxa, with some classes being much more threatened than others. For example, amphibians stand out for having a rate of endangerment that is roughly twice that of mammals and four times that of birds. One reason for the vulnerability of amphibians might be their permeable skin which exposes them to pollutants and other environmental stressors, plus the fact that they typically require multiple habitats (aquatic and terrestrial) as opposed to just one habitat. If either their aquatic or terrestrial habitat has been lost or degraded they

Table 2 Definitions of threats used in IUCN Red List in order of importance[a]

Habitat loss/degradation	Human-induced land and/or water impacts associated with aqua/agriculture, logging, mining, land management, and infrastructure development
Intrinsic factors	Species specific factors that increase vulnerability such as limited dispersal, poor recruitment/reproduction/regeneration, high juvenile mortality, inbreeding, low densities, skewed sex ratios, population fluctuations, and restricted ranges
Harvest (hunting and gathering)	Reduction of species numbers for human use such as food, medicine, fuel, cultural, or scientific applications and materials for direct use or trade
Alien species	Invasive species directly affecting an endangered species through competition, predation, hybridization, and pathogens/parasites
Pollution	Land pollution, water pollution, and atmospheric pollution (including climate change). Pollution can be heat/cold, noise, light or chemical
Human disturbance	Impacts related to recreation/tourism, research, war/civil unrest, transport, and fire
Changes in native species	Changes in competitors, predators, prey/food base, hybridizers, pathogens/parasites, and mutualisms
Natural disasters	Includes storms/flooding, temperature extremes, wildfire, volcanoes, and avalanches/landslides
Other (accident, persecution)	Includes fisheries or terrestrial bycatch, collision with vehicles/buildings/pylons, and pest control

[a]Definitions are illustrative and may not include all conceivable examples for each category.

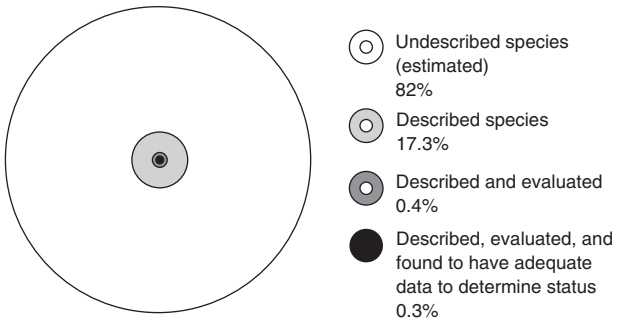


Fig. 2 The relationship between (hypothetical) total species and species that have been evaluated for globally threatened status. Only the bullseye area has been evaluated (0.3%) of a conservative total of 10 million species. In some rare cases (143), undescribed species were evaluated for threatened status where museum or herbarium collections were made and immediate listing would make some tangible benefit to those species' conservation.

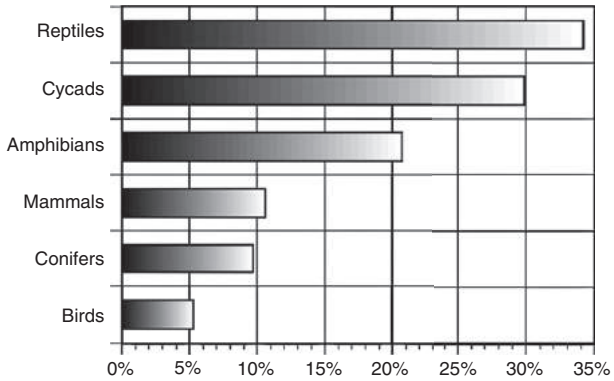


Fig. 3 The percentage of species that are classified as endangered or critically endangered by IUCN within taxa that have been comprehensively evaluated. Reptiles in this figure only include order Testudines, order Crocodylia, and family Iguanidae (~4% of all reptile species) and therefore may not represent the percentage of reptile endangerment as a whole.

will not be able to complete their life cycle. A handful of evolutionary lineages at the family and order level can be identified as especially endangered among mammals, amphibians, and birds. These higher-order evolutionary lineages with uniquely severe levels of endangerment include albatrosses, kiwis, carnivores, manatees, salamanders, and true toads.

Globally, the greatest source of endangerment is primarily habitat loss and degradation (Fig. 4); in fact, the next nearest risk factor after habitat loss is only one-fifth as frequent. The predominance of habitat loss as a source of endangerment provides the rationale for the widespread use of species–area curves in conjunction with deforestation rates to estimate global extinction in the terrestrial environment. Most terrestrial endangered species are in tropical rainforests, which parallels the fact most terrestrial species are in tropical rainforests. For well-studied taxa, one can ask how many endangered species are truly on the brink of disaster, which can be defined as those globally endangered species that are found at only one site. It turns out there are 795 such species spread among 595 sites in the world. If we were committed to zero extinction, these are the endangered species under imminent risk that would command our greatest attention. Whereas most extinction in the historical record has occurred on islands with their isolation and small population sizes, these contemporary imminent extinctions are increasingly in mainland areas where human activities have created isolated islands of habitat.

Ecology of Endangerment

Just as all individuals must die, all species eventually become extinct. However, the life spans of species are not random: some groups of species tend to have high rates of extinction and short life spans, whereas other groups have low rates of extinction. The risk of extinction is higher for some taxonomic groups than others. For example, within North America a high proportion of aquatic species are endangered. The factor most clearly related to endangerment is body size, where species with larger body size are more likely to become extinct. For example, the megafauna of North America, which included such species as the woolly mammoth, sabertooth cat, and the giant ground sloth, were lost approximately 8000 years ago. Analyses of extinction risk among families of birds showed that larger body size and lower fecundity were related to higher chance of extinction, and this remains the case today.

Another clear pattern is that species with small geographic ranges tend to be more prone to extinction. This is why many conservation efforts focus on endemic species – these are species whose distribution is limited to only one defined area (e.g., a country, ecoregion, or island). Depending on how small the geographic distribution is, an endemic species can be quite vulnerable to human perturbations because disruption over a limited area can lead to their extinction globally. Lastly, species with very specialized requirements are most likely to be endangered – largely because if something goes wrong with their diet, or habitat, they are unable to thrive in altered conditions. In sum, species with large body size, narrow geographic range, and specialized habitat or food requirements are more prone to extinction. The upshot of this is that as we accelerate the extinction rate, we will increasingly find ourselves in a world filled with small bodied, widely distributed, generalist species – science fiction projections of a world populated by rats, cockroaches, and gulls may not be far off the mark.

Discussion of extinction and endangered species often create an impression of hopelessness and despair. Fortunately, by paying attention to the science of endangered species, it is possible to take measures to recover species. Approximately 10% of the populations of threatened and endangered species in the USA are actually now increasing, and 30% are stable. Also globally, among species with a known trend, approximately 15% of the threatened bird species and 40% of threatened amphibian species have stable or increasing populations. In addition, several prominent species that were at one time endangered have in fact been recovered. The gray whale off the coast of California has now been restored to historical population levels thanks to a cessation of harvest. The American alligator may be an even more dramatic success story. This species was near extinction in the mid-twentieth century and now has a population of nearly a million in the US. Its source of endangerment was hunting for the skins and loss of

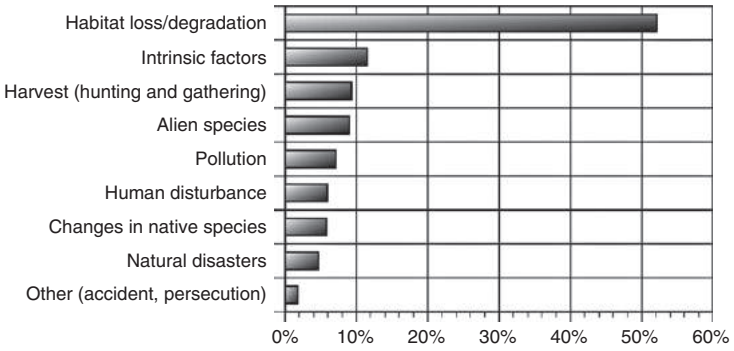


Fig. 4 The percentage of species that are classified as endangered or critically endangered for each IUCN threat category. Endangered species often suffer from multiple threats. Results are heavily biased toward amphibians, birds, and mammals, since threat data have only been comprehensively compiled for these groups. Definitions for threats are in **Table 2**.

habitat due to the draining of wetlands. With the abatement of these threats, the alligator population quickly recovered. In some cases Herculean captive breeding programs have rescued species from the brink of extinction. The Peregrine falcon had totally disappeared from eastern North America by 1964. A massive breeding program that drew on falcons from Europe, Australia, and South America yielded the release of 1300+ birds between 1974 and 2000. The reintroduction was so successful that these falcons are no longer considered endangered. If threats are abated and a species can be bred in captivity, then endangered species need not remain at risk forever.

There are also other reasons for optimism. Because habitat loss is understood to be the primary threat to species, many conservation-minded organizations and government agencies have shifted from a species to a habitat conservation focus. In the process they have secured larger areas where key ecological processes are more likely to remain intact, and larger populations of endangered species can potentially persist or develop. Recently, the rediscovery of the ivory-billed woodpecker in the Big Woods of Arkansas, USA was thought to be partially the result of previous efforts to conserve bottomland hardwood forests in the region. By preserving large habitat landscapes, we can overcome our lack of species knowledge, and thereby protect species we think we lost or do not even know exist. Beyond a habitat-based approach, our greatest scientific need when studying endangered species is likely to be more attention to evolutionary adaptability. Human impacts are so great and so rapid, that the best insurance against extinction will often be adaptability. It is unclear to what extent we can actually manage adaptability in endangered species, as opposed to simply having to work with the situation as it stands without any proactive efforts. Endangered species with ample genetic variability and behavioral plasticity may well be able to get through bottlenecks of low population size and recover from their at-risk status.

See also: General Ecology: Hunting. Terrestrial and Landscape Ecology: Island Biogeography

Further Reading

Baillie, J.E.M., Hilton-Taylor, C., Stuart, S.N. (Eds.), 2004. 2004 IUCN Red List of Threatened Species.A Global Species Assessment Gland, Switzerland and Cambridge, UK: IUCN, p. xxiv + 191.
Goble, D.D., Scott, J.M., Davis, F.W., 2006. The Endangered Species Act at Thirty. Washington: Island Press.
Lawton, J.H., May, R.M. (Eds.), 1995. Extinction Rates. Oxford: Oxford University Press.
McKinney, M.L., 1997. Extinction vulnerability and selectivity combining ecological and paleontological views. Annual Review of Ecology and Systematics 28, 495–516.
Ricketts, T.H., Dinerstein, E., Boucher, T., *et al.*, 2005. Pinpointing and preventing imminent extinctions. Proceedings of the National Academy of Sciences of the United States of America 51, 18497–18501.
Simberloff, D., 2003. Community and ecosystem impacts of single-species extinctions. In: Kareiva, P., Levin, S.A. (Eds.), The Importance of Species: Perspectives on Expendability and Triage. Princeton: Princeton University Press, pp. 221–233.
Stein, B.A., Kutner, L.S., Adams, J.S., 2000. Precious Heritage. London: University of Oxford Press.

Relevant Websites

http://www.zeroextinction.org–Alliance for Zero Extinction.
http://www.redlist.org–2006 IUCN Red List of Threatened Species.

Invasive Plant Species☆

Beth A Middleton, Wetland and Aquatic Research Center, Lafayette, LA, United States

Glossary

Adaptation Change in the genetic structure of a population due to natural selection, which leads to an improvement in function in relationship to the environment.

Biological control The control of invasive species via enemies such as predators, parasites, herbivores or competition.

Diversity-resistance hypothesis Communities with high levels of diversity are less invaded by exotic species.

Ecosystem function/service Major processes and their interplay within ecosystems including primary and secondary production, decomposition, mineral cycling, and energy transfer. In the social sciences, ecosystem service is a nearly equivalent term for ecosystem function, particularly as used to describe functions of ecosystems as they affect humans, for example, production as food support for wildlife or humans, biomass as a buffer from storm damage.

Enemy release hypothesis Invaders escape their traditional predators in a new geographical area, and are therefore at an advantage to native species.

Exotic species A species growing outside of its native range.

Introgressive hybridization The introduction of genes novel to the species (or type) by either by inter- or intra-specific hybridization. Introgressive hybridization of an invasive species with native or non-native species could create new genotypes of the invasive species.

Invasive species Alien species native to another continent.

Pangea Single "supercontinent" present during the Triassic Period before present day continents began to drift apart. In the context of invasive species, the "New Pangea" refers to the state produced by higher levels of species migration supported as modern day humans travel more freely between continents.

Resource-enrichment hypothesis/fluctuating resources hypothesis Communities are more invaded by exotics if amounts of unused resources within the community are higher.

☆*Change History*: March 2018. Beth A Middleton updated text, updated table 2 and some of the references.

Transgenic weeds Genetic modified (GM) types with genes from both wild and agricultural sources, which escape into either fields or natural habitats. Potentially modified traits include genes for drought tolerance, resistance to herbicides, resistances to herbivores.

The Invasive Species Problem

Global Perspective

Invasive species aggressively invade new continents so the world's flora is becoming more homogeneous. Benign components of their original habitats, invasive species can become dominant and aggressive in new geographical regions. These species are thought to be both drivers and symptoms of environmental change. Certain types of habitats have a propensity to be invaded by opportunistic exotics, for example, riverine wetlands, which have a relatively large amount of disturbance because of regular changes in water depth, flow, sediment, and nutrients.

Once established, invasive species cause many problems for humans in that they degrade natural communities. Invasive pests and diseases damage agricultural species. Exotic aquatic species can cut off local commerce by stopping boat traffic along rivers. These species also cause local electricity emergencies by clogging the operation of hydroelectric dams, particularly in countries that rely on hydroelectric power such as New Zealand.

Invasive species are those species that arrived on continents after the 16th century after human global travel, commerce and migration increased. While species have moved between continents for millennia, global travel by humans has greatly accelerated the rate of intercontinental movement of species. New invasions of species have paralleled the movements of humans worldwide, and the associated spread of invasive species has caused the decline of native species on their continents of origin. This relatively free movement of biota between continents has created a "New Pangea." Continents will have more species as new invasive species arrive, but the displacement and extinction of native species caused by invasive species ultimately will cause the worldwide number of species to decrease. Exotic species introductions are generally unintentional, although there are many documented cases of species being transported to other continents for horticultural or agricultural purposes.

From a philosophical perspective, invasive species are not just another for the species richness list because invasives cause environmental degradation. While some authors criticize the furor over invasive species as being akin to xenophobia, the perspective that invasive species are problematic in natural areas is fundamentally dissimilar to the idea that foreigners can cause harm to a society. The concern of ecologists over invasive species is due to the damage invasive species cause to natural plant communities.

In North America, there are 50,000 nonindigenous species, 3000 of which are invasive. Hawaii has more than its share of invasives with 860 invasive species. Each year, $137 billion per year are spent to eradicate invasive species in the United States (Figs. 1 and 2), and $26 billion are spent to combat crop weeds.

Types of Invasive Species

Alien species from other continents

Most commonly, invasive species are defined as aggressive alien species from other continents. Invasive plant species often share certain characteristics such as fast and high seed production, vegetative reproduction, rapid growth, and a tolerance for a wide range of environments.

Surprisingly, scientists do not always know on which continent an invasive species originated, but the source continent and population of the invader can be determined by genetic analysis. In the case of the North American populations of *Hydrilla verticillata*, the dioecious form of this aquatic invasive is similar to types found in Bangalore, India, while the monoecious form resembles plants from Seoul, Korea. *H. verticillata* in New Zealand may have been introduced from Australia, where the species invaded centuries ago. The determination of the geographic origin of an invasive species can be a useful first step toward determining factors that can control the invasive species. The place of geographical origin can be checked for parasites that control the spread of the invasive species there, and subsequently this knowledge could be used in the design of biological control methods.

Introduced disease species can also be devastating to native plant species. One example is the butternut (*Juglans cinerea*) canker, a disease which likely came to North America from Asia. The butternut canker has infected populations of butternut trees since the 1960s, a similar situation to that of Dutch elm disease and Chestnut blight, affecting American elm and chestnut, respectively. These three diseases have nearly eliminated important tree species and undoubtedly altered the function of the Eastern deciduous forests of North America.

Hybrid species with introgressed genes

Sometimes exotic species form hybrids with native species, so that subsequently the newly emerged hybrid becomes dominant in either the novel or native geographic location. For example, *Typha* x *glauca* is a hybrid of *Typha latifolia* and *Typha angustifolia*, which has displaced native *T. latifolia* in many locations in North America. Hybridization can create new genotypes either via inter- or intra-specific combinations of invasive species with either native or non-native types. Introgressive hybridization may make species more

Fig. 1 *Eichhornia crassipes* (water hyacinth) creating havoc in a boat channel in Jean LaFitte Historic Park and Preserve, south of New Orleans, Louisiana. Photo by Beth Middleton.

Fig. 2 *Eichhornia crassipes* (water hyacinth) is a floating aquatic species native to South America, but invasive in tropical and subtropical wetlands around the world. Photo by Beth Middleton.

invasive than they would be otherwise, and may be a contributing factor in the aggressiveness of *Lythrum salicaria* (Fig. 3) *Myriophyllum spicatum*, *Phalaris arundinacea*, *Phragmites australis* and *Typha* x *glauca*, especially in North America.

In restoration areas, genetic integrity may be compromised in situations where invasive species have either been intentionally planted or when invasive species have taken over sites. Hybrid cordgrass (*Spartina alterniflora x Spartina foliosa*) has taken over restoration sites in San Francisco estuaries. These cordgrasses differ in size from native cordgrasses, and are not used by the endangered clapper rail (*Rallus longirostris*), the species for which the coastal marsh restoration projects are designed. Invasive hybrid species threaten the success of restoration sites, both by comprising the genetic integrity of the vegetation, and altering the functional qualities of the site.

Transgenic Weeds

The potential exists for new types of invasive species to emerge with introgressed genetic material from transgenic crop species transferred to related native or invasive species. Transgenic sunflowers (hybrids of native and transgenic *Helianthus annuus*) are often found adjacent to GM (genetically modified) sunflower fields because of genetic introgression. Only a few studies have determined if transgenic weeds pose a threat similar to that of invasive species. In one study, transgenic oilseed rape, potato, maize and sugar beet were no more invasive or persistent in fields and natural areas than the traditional crop species. However, 42% of native sunflower populations adjacent to GM sunflower fields are transgenic hybrids, and these genes persist in these transgenic weed populations for at least five generations. Transgenic traits for disease, insects or herbicides could give a competitive advantage to wild species, thereby conferring a higher level of invasibility to these transgenic weeds. Transgenic traits could also be transferred to endangered species, in a situation paralleling the genetic erosion of cutthroat trout populations in the White Mountains of Arizona (see "Hybrid Species with Introgressed Genes"). The consequences of the presence of transgenic weeds in crop or natural plant communities will depend on their fitness, ability to spread, level of the genetic erosion of the populations of native species, and their overall ability to displace other species as invasives in natural communities.

Fig. 3 *Lythrum salicaria* (purple loosestrife) of short stature at the northern extreme of its invasive range near Amos, Quebec, Canada. Photo by Beth Middleton.

Invasion Hypotheses

Many hypotheses have been developed to describe the dynamics of invasion, and these explanations variously describe how a barrier to invasion is removed so that an invasive species is successfully transferred to a new location. The explanations are not mutually exclusive.

Hypotheses

Preemption hypothesis

Undoubtedly an important reason why invasive species do not always invade habitats, the preemption hypothesis suggests that invasives can be prevented from invading because of the presence of other species. In other words, the plant colonization window for an invasive species is closed, because other species are physically occupying the space. Though often overlooked, preemption is logically one of the most important forms of invasive species exclusion. If a species is already occupying a space, it is difficult for an invasive species to eject that species from the space.

Enemy release hypothesis

The Enemy release hypothesis is the idea that invasive species are less impacted by enemies (e.g., herbivores) than native species, because in the new geographical location, the invasives species are freed from the parasites that kept their growth in check in their native environment. Therefore, invasive species are thought to compete more successfully with native species in their invasive environment, because the native species have not been released from their traditional enemies. The enemy release hypothesis is the basis for biocontrol programs.

The introduction of the pests of invasive species can have a dramatic impact on the growth and spread of invasive species in biocontrol programs, which is evidence that the enemy release hypothesis is operating. Also, plant species are sometimes taller in their introduced ranges, and this is often attributed to the absence of their traditional parasites. Most invasive plant species have an average of 16 parasites in their native range, but only three in their introduced range.

Related to the idea that invasives may be larger in their introduced ranges is the idea that these species may be larger because these have introgressed genes from hybridization with native species. Following this idea, these invasive species may not be the same as their progenitors in their region of origin (Introgression Hypothesis; see section "Types of Invasive Species"). With novel genes introduced, the invasive species with introgressed genes may be larger and at an advantage to the original invasive type.

Resource hypothesis, resource-enrichment hypothesis, fluctuating resources hypothesis

Resource-related hypotheses assume that invasive species are limited by resource availability. According to the various resource hypotheses, invasion is related to the availability of resources including nutrients, light, and water. The resource-enrichment and fluctuating resources hypotheses assume that communities are more invasible if amounts of unused resources are higher.

Diversity-resistance hypothesis

The diversity-resistance hypothesis suggests that communities with high levels of diversity are less invasible by exotic species, as based on ideas of competitive exclusion and niche theory. If niches are narrower and already filled, competitive exclusion bars new species from entering very diverse communities. The opposite idea is also claimed that is, that communities with high levels of diversity are more invasible by exotic species. Actually, field studies support both claims. However, studies of the pattern of species invasion at large regional scales suggest that species rich landscapes are more invaded by invasive species than species poor landscapes.

Geographical range hypothesis

After an invasive species has initially established on a new continent, the best predictor of the invasive range is the climatic range of the species on its home continent. Thus, after the invasive species' successful establishment on a new continent, the species is likely to spread throughout the entire climatic range. This spread may take decades or centuries. The geographic range hypothesis is related to niche theory but carried to a landscape scale, because of the idea that species grow best in certain climates and environments. Modelers who attempt to project the eventual spread of an invasive species on a new continent use the ecological niche characteristics of the native range. An interesting wrinkle on this idea is that as world climates warm, the poleward migration of invasive species may occur, with their overall distributions enlarging or shrinking depending on their abilities to migrate, and occupy habitats within the new climate. In particularly, certain tropical or subtropical species are likely to move northward (e.g., *Eichhornia crassipes*).

Removing Isolation Mechanisms—Intercontinenal Migration

Invasive species transmission

Species are limited in their ability to move to new continents by geographical isolation, and this limits the majority of the world's species from becoming invasive species on another continent. Dispersal overcomes the isolating mechanisms that prevent the movement of species from one continent to another, and species have varying levels of abilities to disperse across the barrier and successfully open invasion windows. Species can move by natural dispersal vectors including wind, animal, and water transport. (See Middleton, 1999 for an extensive species list of seed dispersal mechanisms for world plant species.)

After plant species successfully disperse and establish on a new continent, many of these can spread locally by vegetative reproduction. One successful propagule can start a new population, and many plant species spread solely or mostly via asexual means (e.g., bulbils of *Dioscoria oppositifolia*, plant fragments of *E. crassipes* (Fig. 4), *Salvinia molesta*, *Elodea canadensis*, and *H. verticillata*). Wind-dispersed invasive species with prolific seed production can also spread rapidly after their initial introduction (e.g., *Phragmites* spp. and *Typha* spp.). Also, recently established northern invasive species may adapt quickly through selection for earlier flowering time, so that seed production may increase at the northern invasion front.

Transmission by travelers and commercial shipments

Historically, invasive species have moved between continents as human travel and migration increased. As climates and hardiness zones shift in North America, horticultural demands are likely to favor African and Middle Eastern species that are more successful in hotter and drier environments. Thus, a new wave of horticultural introductions are likely to accommodate those demands. Hitchhiking species cling to travelors' luggage, mud on shoes, and probably even the tires of airplanes. Recirculating air in airplanes transfers germs from one passenger to another on transcontinental flights, after which the travelers themselves deliver the novel diseases to new continents. Invasive species passively cling to transport vessels. The Antarctic now has at least 207 alien species, and these invasives mostly originate from ballast discharge.

With the current level of international travel, new introductions of invasive species are inevitable. However, international agreements attempt to address the problems associated with the international transport of organisms. The Global Ballast Water Convention is an international agreement passed in 2004, which will eventually require ships to comply with discharge limits, and establishes inspection and enforcement procedures.

Fig. 4 *Lythrum salicaria* (purple loosestrife) is an emergent perennial species native to Eurasia, and an invasive species of wetlands in northern North America. Photo by Beth Middleton.

Invasive Species in Natural Communities

Composition and Functional Change

Ecologists consider invasive species to be a serious threat to the biodiversity of natural areas. Invasive species may change the microenvironment for native species for example, by changing light, air and water environments in plant habitats. Most studies show that the numbers of native species decrease in habitats after exotic species invade, and the tall invading species are often the most serious problems. In some cases, invasive species may outcompete native species, for example, *Oenothera deltoides* declined in native communities with invasive grass species because of the thatch and associated shade created by the grasses. Subsequently, the germination of *O. deltoides* is reduced, and eventually the population numbers of the species decline. However, invasive species may not always eliminate native species. Recent studies of *L. salicaria* in North American wetlands suggest that the presence of this invasive species does not eliminate native species, but instead, reduces their size.

Emerging research suggests that the function of ecosystems dominated by invasive species differs from those dominated by equivalent native species (Table 1). One study of riparian floodplain function in the southwestern United States showed that after invasion by *Tamarix chinensis*, half of the functional traits of importance in riparian floodplain function differed from native *Populus fremontii* floodplain forests (Stromberg, 1998). The impacts of invasive species on ecosystem *L. salicaria* function may be subtle, and yet have far reaching effects for the function of the system. If dominant species that are important as food sources in the system are replaced by less useful invasive species, the secondary production of animals higher on the food chain may be affected. For example, *L. salicaria* did not support fish populations because the species decomposed at a different rate than a native sedge species (*Carex lyngbyei*) that it replaced in the Fraser River Estuary of the northwestern United States. Here, a shift of dominant species caused a loss of the productivity of the fisheries, because the invasive *L. salicaria* does not provide a seasonally equivalent food base for the invertebrate species that support the fish populations. Shifts in function due to invasive species may be profound, but we understand little about the overall threat of invasive species from this perspective.

Invasion and Disturbance

The majority of invasive species invade native communities after disturbance, so that these species are common in old field and restoration sites. However, some invasive species can invade natural areas without the help of human disturbance. In wilderness canyons of southern Illinois, *D. oppositifolia* (Chinese yam) can invade floodplains of streams via bulbils (asexually produced disseminules). *Lygodium microphyllum* (Old World climbing fern) invades tree islands in the Everglades of Florida. Also, *Impatiens glandulifera* is dispersed along rivers in Central Europe. These three examples are situations where flood pulsing along stream channels spreads the disseminules of invasive species. Invasives that are maintained by natural disturbances are particularly problematic from a management perspective, because the same natural disturbances, which are critical in maintaining plant community types, also maintain these invasive species.

Table 1 Potential shifts in ecosystem function in areas with invasive species

Function	*Specific shift in function*
Productivity	– Biomass of native plant species reduced – Size of animals feeding not supported as well by invasives, for example, bog turtles (*Clemmys muhlenbergii*) smaller in wetlands with *Lythrum salicaria*
Soil/water/salinity/geomorphological relationships	– Surface/ground water levels changed – Erosion/sedimentation levels changed – Soil content (silt/clay/sand proportion) changed – Organic matter dynamics in long-term succession changed – Organism function in stream channel, bank stabilization, or island building capacity changed – Water infiltration of ground is changed (repelled or absorbed more easily than native species)
Community attributes	– Species richness of native species lowered – Early succession dominated by invasives, so that maturation or establishment of later successional species is impeded – Displaces native species
Disturbance attributes	– Fire frequency changed – Differs from native species in its ability to tolerate disturbance/perturbations such as wind, flooding, herbivory/insectivory/parasitism, animal foraging or digging
Nutrient dynamics	– Nutrient availability lowered, for example, pH lowered by adding salinity or acidity so that nutrients become less available – Adds nutrients, for example, leguminous invader and associated bacteria fixes nitrogen
Decomposition dynamics	– Rate of decomposition shifts, so that litter or nutrient dynamics are altered
Secondary production	– Species at the top of food chain changed, for example, invasive species does not provide the same level of food support to specific higher order organisms
Habitat structure	– Habitat structure differs from that of native species in plant height, density, openness, or light characteristics so that animal foraging, nesting or burrowing is impacted

Invasive Species Control

Restoration of Natural Conditions Following Anthropogenic Disturbance

Often the best time to attempt to control an invasive species is immediately after the first individual establishes, and before it sets seed. Attentive natural areas managers and restorationists often have stories of killing that first invading purple loosestrife plant in its first year of growth in a newly restored wetland. Some countries (e.g., New Zealand) request that citizens report the first sightings of various invasive species, so that the invasive can be removed before it infests the area. However, after the first individuals have established, invasive species are difficult if not impossible to eliminate. To control invasive species, natural disturbances have been used such as fire, flooding, manual removal, shading, substrate removal, herbicide and biocontrol. However, the success of implementing natural disturbance to remove invasives is situation dependent. Fire may be a good means of controlling invasive shrubs in prairies in the midwestern United States; however, fire is not appropriate in Hawaii because indigenous species there did not evolve with fire. Also, invasive species may alter fuel loads so that fires may be more or less frequent in an ecosystem after these species invade.

Habitat Protection via Banning of Invasive Introductions

Many countries have laws to prevent the intentional introduction of invasive species. In the United States, the introduction of alien species is regulated by the USDA Animal and Plant Health Inspection Service, and the movement of species across international boundaries into the U.S. is highly regulated. Laws vary from country to country. A good source of information on the restrictions for the movement of organisms between countries is given on the website of The International Portal on Food Safety, Animal and Plant Health: http://www.ipfsaph.org/En/default.jsp.

Biological Control

Biocontrol of invasive species sometimes can be achieved through the introduction of a parasite from the region of the invasive species' origin (Table 2), but the danger exists of introducing yet another invasive species, which could create harm to native species. The negative effects of the introduction of organisms to control invasive species may not be immediately apparent, but may arise suddenly many years after introduction of the biocontrol agent. One biocontrol success story is in the control of *Opuntia vulgaris* in India, a species which was introduced from Brazil. A scale insect (*Dactylopius ceylonicus*) was released in 1795, and the insect completely controlled the cactus in India. However, in a similar attempt to control *Opuntia* on the island of Nevia in the Lesser Antilles, disaster occurred with wide reaching and unintended consequences. In 1957, the cactus moth (*Cactoblastis cactorum*) released to control *Opuntia* on Nevis, but the moth escaped to destroy populations of the rare *Opuntia spinosissima* in the Torch Wood Hammock Preserve in the Florida Keys. Having now reached the Gulf Coast and spread by hurricanes, the cactus moth has spread across the region to Louisiana, and could eventually threaten populations of cactus in the southwestern United States.

Insects pests of *L. salicaria* have been introduced from Eurasia to control this invasive in North America. However, these insect species may attack related species of *Lythrum* native in North America. Furthermore, some studies suggest that the insects do not eliminate *L. salicaria* but only reduce the biomass of the species as compared experimentally inside and outside of insect exclosures. Other studies suggest that *L. salicaria* does not necessarily pose as severe a risk to wetland communities, as the release of the insect species. *L. salicaria* does not reduce the species richness of other species, although the effect on the function of the wetland may be impacted.

Table 2 Mechanisms for the control of invasive plant species

Method	*Advantages*	*Disadvantages*
Hand or machine cutting	Labor intensive although less so than some other methods	Most species grow back from underground parts
Cover cut stems with object (e.g., dark plastic bag)	Plants usually don't grow back after cutting/covering	Labor intensive
Root pulling	Removes underground parts of plant that may regrow	Disrupts the soil and may encourage reinvasion of the exotic species from seeds or plant fragments
Herbicide application	Easy to accomplish relative to more labor intensive methods	May kill native species
Handcutting with herbicide application to cut shoots (Bradley Method)	Very effective removal mechanism, especially for woody species	Uses herbicides, although in very small amounts
Herbicide	Can be effective in the removal of certain species; less labor intensive than other mechanisms, so that it can be used to treat large tracts of land	Improper usage may harm the health of users and/or the environment
Biocontrol (release of insects or pathogens transferred from their continent of origin)	Can target the invading species only	Pests may be unpredictable; native species may be damaged

The introduction of parasites to control invasive species is not always very successful from the perspective of the survival of the introduced parasite. Sometimes the habitat requirements of the parasite are not met, so that these parasites become extinct shortly after introduction. Biocontrol can fail because the parasite can not be maintained in the invasive environment.

Many argue that biocontrol is worth the risk that the parasite might damage the environment because the cost of doing nothing may be higher than any potential cost incurred or harm caused by the biocontrol agent. While the benefits of releasing biocontrol agents to the ecosystem may be profound, the potential effects of harmful biocontrol agents are impossible to project. For example, who could have thought that a cactus moth released in the Lesser Antilles would be capable of making its own way across the Gulf Coast of North America? Others use the same reasoning regarding the lack of ability to project the effects of biocontrol agents, to argue that the usage of biocontrol agents is never warranted.

Global Warming

Species may increase in size with warming as many do from tidal marsh systems in experimental mesocosms. In some cases, the seeds are larger in exotic versus native populations, for example, seeds of the exotic shrub, *Cytisus scoparius* are larger in North America vs. Europe.

Physical Control

The physical control of invasive species by digging, removal, and harvesting (Table 2) may or may not be feasible depending on the size of the invasive population, and number of available people to help with the removal (Box 1). Sometimes, the invasive species has built a large reserve of seeds in the soil, so that seedlings from germinated seeds grow to replace the removed adults in the soil disturbed by the removal activities. Often, the invasion potential increases with the amount of disturbance created by the removal of invasive species, so care should be taken in physical removal programs not to disturb the sites. Especially in restoration projects, it is important to remove invasive species by hand, or the resulting vegetation will be unsatisfactory to meet the regulatory requirements

Box 1 Think globally, act locally: volunteer programs for invasive species control, and research data collection

Volunteers can help with the invasive species problem by volunteering their time to help in management and research projects. The control of invasive species can often be achieved by the removal of invasive species from natural communities, an effort greatly helped by teams of volunteers. The "weed pulling organizations" are regionally coordinated, and these organizations often have websites with information for volunteers. A short search on the Internet can help volunteers locate an organization in their region engaged in weed removal activities. A few agencies for volunteers to check include The Nature Conservancy, California Invasive Plants Council, Hawaiian Ecosystems at Risk, Invasive Plant Association of Wisconsin, Midwest Invasive Plant Council, Southeast EPCC, and Texas Invasives.

Volunteers also can donate their time to help researchers collect data, e.g., the U.S.G.S. purple loosestrife volunteer program. Volunteers in this program collect data on *Lythrum salicaria* around the world. See: http://www.nwrc.usgs.gov/special/purplel/.

of the restoration process. In such cases, the best approach is to remove invasive species immediately, before the invasives become widespread in the restoration area.

Invasive species can be snipped at the base by hand with a long handled clipper, while carefully avoiding native species. It is also possible to cut woody species, and then cover the stump with a heavy dark plastic bag, which will keep the stump from regrowing. After cutting invasives, the remaining native species are at an advantage since they are already established. After cutting invasives, the native species may regrow to shade and reduce the invasive species after the treatment. This selective clipping method is very time consuming, but can have very good results for reestablishing native plant communities that have become invaded with invasive species. Also, the removal of exotic species without attention to the restoration of native species can pave the way for further exotic species spread.

Mechanical mowing with a tractor or by hand with a scythe can be an ideal way to reduce the amount of an invasive species and encourage native species, particularly if the mowing is done in certain seasons. For example, mowing in the late spring and early summer reduced the amount of *Arrhenatherum elatius* and increased the native *Danthonia californicum* in a grassland. Harvesting by hand also can remove invasive species. Kudzu (*Pueraria lobata*) is sometimes harvested for cattle forage, and the invasive *Prosopus juliflora* can be used for fuel wood (the famed mesquite barbecue charcoal) (Fig. 5). The downside of putting invasive species to a "good" use after their harvest is that the transfer of biomass elsewhere can spread seeds or disseminules.

Herbicide

Herbicides are widely used to remove invasive species, but the success of their application depends on the situation and the species involved. Herbicide eliminated *E. crassipes* (Fig. 4) in Lake Hartbeespoort, South Africa, but only after four large scale aerial spraying episodes followed by spot spraying from boats. Sometimes herbicides work best in combination with other removal methods. For example, woody invasive species in grassland or forest situations can be controlled by a combination of hand cutting and application of an appropriate herbicide to the cut stump. Without herbicide or stump covering (see above), most woody species simply grow back from the stump after cutting. Direct application of herbicide to the stump is often an effective method of woody species removal, but at the same time, little herbicide is released to the surrounding environment.

Fig. 5 Stump digging of *Prosopus juliflora*, an exotic shrub in semi-arid grasslands in India. Photo be Beth Middleton.

References

Middleton BA (1999) *Wetland restoration, flood pulsing and disturbance dynamics*. New York: John Wiley and Sons.
Stromberg JC (1998) Functional equivalency of saltcedar (*Tamarix chinensis*) and Fremont cottonwood (*Populus fremontii*) along a free-flowing river. *Wetlands* 18: 675–686.

Further Reading

Baldwin AH, Jensen K, and Schönfeldt M (2014) Warming increases plant biomass and reduces diversity across continents, latitudes, and species migration scenarios in experimental wetland communities. *Global Change Biology* 20: 835–850.
Bartha S, Meiners SJ, Pickett STA, and Cadenasso ML (2003) Plant colonization windows in a Mesic old field succession. *Applied Vegetation Science* 6: 205–212.
Bauer JT (2012) Invasive species: "Back-seat drivers" of ecosystem change. *Biological Invasions* 14: 1295–1304.
Bhowmik PC (2014) Invasive weeds and climate change: Past, present and future. *Journal of Crop and Weed* 10: 345–349.
Blumenthal D (2005) Interrelated causes of plant invasion. *Science* 310: 243–244.
Bradley BA, Blumenthal DM, Early R, Grosholz ED, Lawler JJ, Miller LP, Sorte CJB, D'Antonio CM, Diez JM, Dukes JS, Ibenez I, and Olden JD (2011) Global change, global trade, and the next wave of plant invasions. *Frontiers in Ecology and the Environment* 10: 20–28.
Colautti RI and Barrett SCH (2013) Rapid adaptation to climate facilitates range expansion of an invasive plant. *Science* 342: 364–366.
Cox GW (2004) *Alien species and evolution*. Washington, D.C.: Island Press.
Crawley MJ, Brown SL, Hails RS, Kohn DD, and Rees M (2001) Transgenic crops in natural habitats. *Nature* 409: 682–683.
Hager HA and McCoy KD (1998) The implications of accepting untested hypotheses: A review of the effects of purple loosestrife (*Lythrum salicaria*) in North America. *Biodiversity and Conservation* 7: 1069–1079.
Hejda M, Pyšek P, and Jarošík V (2009) Impact of invasive plants on the species richness, diversity and composition of invaded communities. *Journal of Ecology* 97: 393–403.
Lee CE (2002) Evolutionary genetics of invasive species. *Trends in Ecology & Evolution* 17: 386–391.
Luken JO and Thieret JW (eds.) (1997) *Assessment and management of plant invasions*. NewYork: Springer.
Pimental D, Lach L, Zuniga R, and Morrison D (2000) Environmental and economic costs of nonindigenous species in the United States. *Bioscience* 50: 53–65.
Pyšek P and Prach K (1995) Invasion dynamics of *Impatiens glandulifera*—A century of spreading reconstructed. *Biological Conservation* 74: 41–48.
Pyšek P, Jarošík V, Hulme PE, Pergl J, Hejda M, Schaffner U, and Vilà M (2012) A global assessment of invasive plant impacts on resident species, communities and ecosystems: The interaction of impact measures, invading species' traits and environment. *Global Change Biology* 18: 1725–1737.
Seastedt TR (2015) Biological control of invasive plant species: A reassessment. *New Phytologist* 205: 490–502.
Stohlgren TJ, Barnett DT, and Kartesz JT (2003) The rich get richer: Patterns of plant invasions in the United States. *Frontiers in Ecology and the Environment* 1: 11–14.
Stohlgren TJ, Pyšek P, Kartesz J, Nishino M, Pauchard A, Winter M, Pino J, Richardson DM, Wilson JRU, Murray BR, Phillips ML, Li M-Y, Celesti-Grapow L, and Font X (2011) Widespread plant species: Natives versus aliens in our changing world. *Biological Invasions* 13: 1931–1944.
Zedler JB and Kercher S (2004) Causes and consequences of invasive plants in wetlands: Opportunies, opportunists, and outcomes. *Critical Reviews in Plant Sciences* 23: 431–452.

Relevant Websites

ipfsaph, n.d.—http://www.ipfsaph.org/En/default.jsp, The International Portal on Food Safety, Animal and Plant Health.
dnr.state.mn.us, n.d.—http://www.dnr.state.mn.us/invasives/terrestrialplants/woody/buckthorn/control.html.

k-Dominance Curves

RM Warwick, KR Clarke, and PJ Somerfield, Plymouth Marine Laboratory, Plymouth, UK

Introduction

Curvilinear plots or distributional representations extract information on patterns of relative species abundances in an assemblage without reducing that information to a single summary statistic, such as a diversity index. In fact, such plots provide the raw material that is extracted by the full range of possible measures that combine richness and evenness components of diversity. Unlike multivariate methods, these distributions extract universal features of community structure which are not a function of the specific taxa present, and which may be related to levels of biological stress. *k*-Dominance curves belong to this class of techniques, and have been quite widely used in the context of environmental assessment.

The Method

From a traditional ecological viewpoint diversity may be thought of as the number of groups (usually species) present in an assemblage, or how evenly those groups occur in the assemblage. These aspects of diversity are generally termed richness (species richness for example) and evenness. High richness equates to high diversity, and a highly dominated assemblage (i.e., one with low evenness) is considered to be less diverse than a more even one.

Diversity measures tend to be measures of richness (e.g., Margalef's richness index), evenness (e.g. Simpson's index), or are constructed in such a way as to combine the two components in one measure (e.g. Shannon–Wiener index). The Berger–Parker index) is the proportional abundance of the most abundant species in an assemblage and is a measure of dominance. Where no single species is overwhelmingly dominant, it is useful also to consider the dominance of the two most abundant species, the three most abundant, and so on. In an assemblage a family of indices may be defined: 1-dominance, 2-dominance, 3-dominance, and in general *k*-dominance, which is the combined dominance of the *k* most abundant species. Plotting values of *k*-dominance against species rank gives the *k*-dominance curve. Ranked species abundance (dominance) curves are based on the ranking of species (or higher taxa) in decreasing order of their importance in terms of abundance (or biomass). The ranked abundances, expressed as a percentage of the total abundance of all species, are plotted against the relevant species rank. The *k*-dominance curve, therefore, is simply a `cumulative' ranked species abundance curve in which cumulative proportional abundances are plotted against species rank, or often log species rank. The log scale compresses the information about low-ranked species so that the curves reflect a greater contribution from evenness than richness components of diversity. Obviously, the length of a curve along the *x*-axis is determined by the number of species in the assemblage under consideration, while dominance may be assessed by the height and shape of the curve. The higher the curve, the less diverse (and more dominated) is the assemblage it represents. To compare dominance separately from the number of species, the *x*-axis (species rank) may be rescaled from 0 to 100 (relative species rank), to produce Lorenz curves.

Fig. 1 shows *k*-dominance curves calculated from species abundances of infaunal invertebrates from a marine intertidal sand-flat. Animals were collected using different sieve meshes, and sample sizes were scaled with mesh size. Each curve is calculated from the averaged abundances from four replicates. Note that there is little or no overlap in species composition between samples collected on the largest mesh and on the two smaller meshes, but the method allows a simultaneous comparison of the dominance/diversity structure in each. The curve for invertebrates extracted using a 1 mm mesh is higher than the curves for invertebrates extracted using 250 or 63 μm meshes for all values of *k*. Thus, the assemblage in the 1 mm-mesh samples is unambiguously less diverse than the others.

The curves for invertebrates sieved on 250 and 63 μm meshes in Fig. 1 cross. It has been argued that diversity can only be unambiguously assessed when the curves to be compared do not overlap or cross, as different diversity indices biased toward species richness or evenness will rank these assemblages in opposite ways. However, intersecting plots are informative in that they illustrate differences in dominance relative to species richness in a way that a single univariate index does not.

Whether *k*-dominance curves are plotted from the species abundance distribution or from species biomass values, the *y*-axis is always scaled in the same range (0–100). This facilitates the abundance/biomass comparison (ABC) method of determining levels of disturbance.

Transformations of *k*-Dominance Curves

Very often *k*-dominance curves approach a cumulative frequency of 100% for a large part of their length, and in highly dominated communities this may be after the first two or three top-ranked species. Thus, it may be difficult to distinguish between the forms

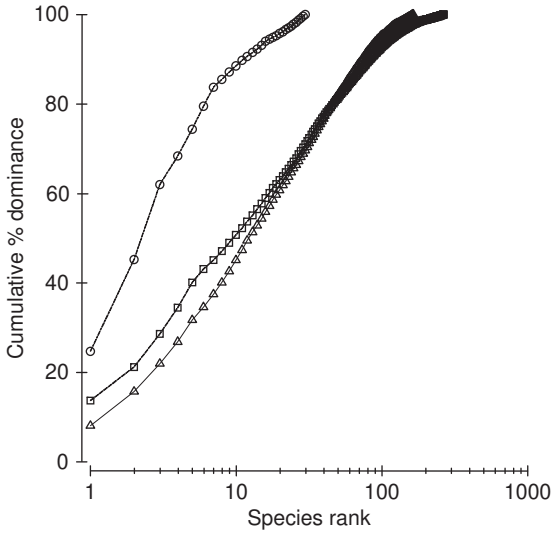


Fig. 1 *k*-Dominance curves for samples of invertebrates from a marine intertidal sand-flat. Each curve is based on average abundances in four replicate samples collected using a particular combination of sieve mesh and sample area: 0.1 m^2 sieved on a 1 mm mesh (circles); 0.006 25 m^2 sieved on a 250 μm mesh (squares); 0.000 39 m^2 sieved on a 63 μm mesh (triangles).

of these curves. The solution to this problem is to transform the y-axis so that the cumulative values are closer to linearity. The modified logistic transformation has been suggested for this:

$$y_i' = \log[(1 + y_i)/(101 - y_i)]$$

Partial Dominance Curves

A second problem with the cumulative nature of *k*-dominance curves is that the visual information presented may be over-dependent on the single most dominant species. The unpredictable presence of large numbers of a species, perhaps an influx of the juveniles of one species, may give a false impression of disturbance. With genuine disturbance, one might expect patterns of *k*-dominance curves to be unaffected by successive removal of the one or two most dominant species in terms of abundance, and the use of partial dominance curves has been recommended. These compute the dominance of the second-ranked species over the remainder (ignoring the first ranked species), the same with the third most dominant, etc. Thus, if a_i is the absolute (or percentage) abundance of the ith species, when ranked in decreasing abundance order, the partial dominance curve is a plot of p_i against log i ($i = 1, 2, \ldots, S - 1$), where $p_1 = 100a_1/\sum_{j=1}^{S} a_j$, $p_2 = 100a_2/\sum_{j=2}^{S} a_j, \ldots, p_{S-1} = 100a_{S-1}/(a_{S-1} + a_S)$.

Earlier values can therefore never affect later points on the curve. **Fig. 2** shows partial dominance plots from the sand-flat study. In this case it does not lead to a novel interpretation of the data, instead showing a similar pattern to that revealed by simple *k*-dominance plots but in a different form.

Hypothesis Testing

Plotting all of the curves from a fully replicated study can produce complex graphs in which the pattern may be difficult to discern. **Fig. 3** shows the replicate curves from the sand-flat study. It is clear that all replicates in the 0.5 mm samples are less diverse than all samples collected on smaller meshes, but is there evidence for differences in dominance/diversity between the 250 μm samples and the 63 μm samples? If *k*-dominance curves are calculated for replicates at a number of sites, times, or conditions, a measure of dissimilarity can be constructed between any pair of curves, for example, based on their absolute distance apart, summed across the species ranks. When computed for all pairs of samples in a study, this provides a (ranked) triangular dissimilarity matrix, essentially similar in structure to that from a multivariate analysis; thus, 1-way and 2-way ANOSIM tests that are used in multivariate analysis can be used in exactly the same way to test hypotheses about differences between *a priori* specified groups of samples. In this case the test shows unequivocally that there is a significant difference between the 250 μm samples and the 63 μm samples: curves within groups are more similar to each other than they are to curves in different groups. What the test does not reveal is the form of those significant differences. For this we need averaged plots (e.g., **Fig. 1**) or plots based on subsets of replicates.

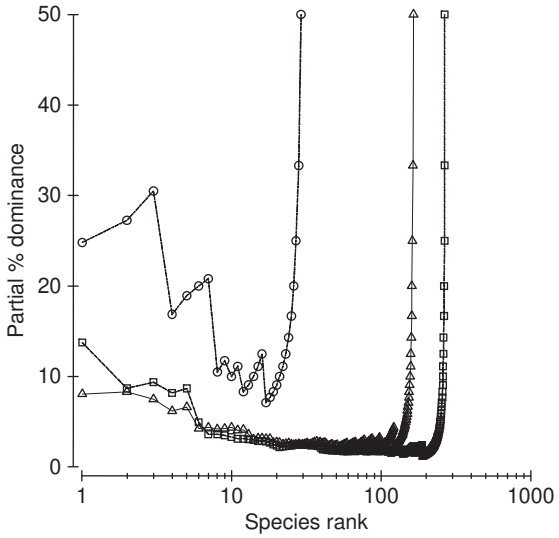


Fig. 2 Partial-dominance curves for samples of invertebrates from a marine intertidal sand-flat. Each curve is based on average abundances in four replicate samples as in **Fig. 1**.

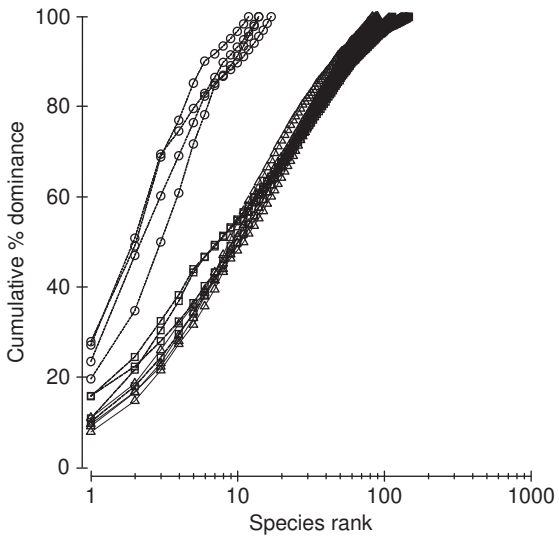


Fig. 3 *k*-Dominance curves for samples of invertebrates from a marine intertidal sand-flat. Each curve is based on abundances in a single sample. Four replicate samples were collected using three combinations of sieve mesh and sample areas as in **Fig. 1**.

See also: Aquatic Ecology: Abundance Biomass Comparison Method. Evolutionary Ecology: Dominance and Its Evolution. General Ecology: Succession

Further Reading

Clarke, K.R., 1990. Comparisons of dominance curves. Journal of Experimental Marine Biology and Ecology 138, 143–157.
Clarke, K.R., Warwick, R.M., 2001. Change in Marine Communities: An Approach to Statistical Analysis and Interpretation, 2nd edn. Plymouth, UK: Primer-e Ltd.
Lambshead, P.J.D., Platt, H.M., Shaw, K.M., 1983. The detection of differences among assemblages of marine benthic species based on an assessment of dominance and diversity. Journal of Natural History 17, 859–874.
Magurran, A.E., 2004. Measuring Biological Diversity. Oxford: Blackwell Science.
Warwick, R.M., Dashfield, S.L., Somerfield, P.J., 2006. The integral structure of a benthic infaunal assemblage. Journal of Experimental Marine Biology and Ecology 330, 12–18.

Polychaetes/Amphipode Index☆

C Chintiroglou and C Antoniadou, Aristotle University of Thessaloniki, Thessaloniki, Greece

Extended Synopsis

The urgent need to assess the state of marine coastal ecosystems and to preserve their biodiversity led to the invention of a variety of ecological indices. Polychaetes and amphipods are major taxa in coastal communities and several studies support their abilities to reflect the integrated conditions over a time period. The polychaetes/amphipode index, simply referring to the ratio of their abundances, is based on the different sensitivity of these two taxonomic groups to disturbances, with the former being more tolerant than the latter. The performance of this index has been tested, with varying degrees of success. Polychaetes/amphipode index is an easy, rapid, and cost-effective approach, and it appears sensitive to the assessment of various contaminants integrating the seasonal oscillations of the fauna in sorted soft sediments. However, its applicability is dubious in assorted sediments and on hard bottoms. The index, also, performs badly when the samples are impoverished, or when polychaetes are not the dominant taxon. Therefore, its applicability is limited and approaches ranking either the entire benthic fauna with respect to its sensibility, or a specific taxonomic group, gain importance in ecological quality assessment. Rather recently, to overcome the drawbacks in the applicability of the polychaetes/amphipode index, a modified version has been proposed by considering the ratio of opportunistic polychaetes to amphipods at first, and then, by considering the ratio of their frequencies after excluding the opportunistic *Jassa* amphipods. This latter index named as Benthic Opportunistic Polychaetes Amphipods index (BOPA), has been calibrated to meet the requirements of the Water Framework Directive. BOPA is a promising tool as it performs well in both coastal and transitional waters, and even in very poor communities.

Benthic Ecosystems: A Synopsis on Research Trends

The comprehensive study of marine benthic ecosystems started in the middle of the previous century (1950–60), focused at an early stage on the Mediterranean Sea. The pioneers of the research center of Marseille set the methodological groundwork of this scientific field, to be followed thereafter by a great worldwide endeavor for the exploration of benthic communities. The early, fundamental knowledge laid the ground for new enquiries about the development of a scientifically documented method to analyze and assess the state of marine coastal ecosystems. The creation and mathematical processing of adequate databases to simulate the dynamics of marine ecosystems is another objective imposed by international directives (WFD 2000/60/EC; MSFD 2008/56/EC) and the original convention for the biodiversity conservation of Rio (1992).

Ecological Quality and Ecological Indicators

The assessment of ecological quality is a current issue in marine research. The approaches to assess the ecological quality of a given water body, are based either on the physical, chemical (priority substances including), or biological elements of the system. The biological elements proposed for this topic have been manifold, ranging from molecular to ecosystem level. However, the most fundamental implement seems to be the benthic fauna as (1) it shows small natural variability compared to pelagic one, (2) it is well responding to organic pollution, and (3) it is rather localized, which means that most species are either sessile or discretely motile, thus not performing large-scale migrations. So they do reflect the integrated conditions over a time period. The concept of ecological indicators has been developed for this assessment; meaning any organism whose presence and/or dominance in a particular area can reveal the prevailed environmental conditions, characterizing the degree of community change or pollution effect. It is widely accepted that when more than a few species are taken into consideration, the indices' validity increases. Therefore, instead of working with a specific taxon, it is better to consider the ratio of taxonomic or functional units (i.e. species, genus, families, trophic guilds, reproductive mode, life strategies, ecological requirements); this approach has led to the invention of a wide number of ecological indicators which have been tested in several cases of pollution with varying degrees of success.

Polychaetes/Amphipode Index

Distribution of Polychaeta and Amphipoda in Coastal Communities

The taxonomic groups of Polychaeta and Amphipoda are among the prominent biota in coastal ecosystems, both in terms of species richness and abundance. Given their ecological significance, literature references confirm their status as sensitive

☆*Change History:* March 2013. C Chintiroglou and C Antoniadou updated Abstract, Extended synopsis, the Polychaete/Amphipods index the reference section, Figures 1, 3 and added Figures 2, 4 and 5, and added Table 1.

 doi:10.1016/B978-0-12-409548-9.00599-6

environmental indicators and support their utility as key taxa for biomonitoring. However, before these taxa can be used to specify disturbances, as demanded by international directives and conventions, it is necessary to provide a database on their distribution on natural benthic communities with comprehensive base line information on their presence at various geographic areas. The distribution of these two taxa shows significant differences with respect to the type of substratum, and therefore their ratio will be examined separately. In general, Polychaeta is the most dominant group in soft-substratum communities comprising over half of both the species richness and the numerical abundance; on hard substratum Polychaeta and Amphipoda co-occur, usually comprising over one third of both species richness and numerical abundance.

Definition of the Polychaeta/Amphipoda Index

The Polychaeta/Amphipoda (P/A) index simply refers to the ratio between the abundance of polychaete species to the one of amphipods. The basic thought behind this approach is the different sensitivity of these two taxonomic groups. Polychaetes, as a taxon, are in general resistant to pollutants, since many species are positively correlated with organic enrichment. As a result, after disturbances, several opportunistic polychaete species are favored and increase their abundance. Amphipods, as a taxon, are more sensitive and experience high mortality after severe pollution events. Still there are some species, such as *Elasmopus rapax*, *Jassa falcata*, *Podocerus variegatus*, and *Corophium* sp., which can proliferate by increased organic load. Several forms of this index have been tested (e.g., the opportunistic species instead of the total polychaete fauna, or some genera of polychaetes), since this taxon is the most important in disturbed areas because it contains a large number of both tolerant and sensitive species, whose gradient can reflect the degree of organic pollution. The use of log10 has been suggested for the calibration of the P/A index, with the addition of one monad to the denominator (i.e., log10 (P/A + 1)). So, the index has low values, below 1 (≤ 1), when the benthic communities are not disturbed and exceeds 1 (> 1) when they are affected from pollution events. Important attributes of such an index should be its reliability in ecological quality assessment, as well as the easy use and broad applicability to various ecosystems.

Modified Version of the Polychaeta/Amphipoda Index: The Opportunistic Polychaete/Amphipode Ratio (P_O/A Index) and the Benthic Opportunistic Polychaetes Amphipods Index (BOPA Index)

The Polychaeta/Amphipoda (P/A) index has been recently modified (Dauvin and Ruellet, 2007), again in accordance with the taxonomic sufficiency principle, that is, reduction of the taxonomic classification effort (Dauvin *et al.*, 2003; Ellis, 1985). Thus, only opportunistic polychaetes and amphipods were considered; the former animals have been exclusively included in the new formula of the index and the latter ones have been excluded from calculations. Latter on, relative frequencies were used instead of abundance data mostly to define the limits of BOPA in conformity with the requirements of the Water Framework Directive (WFD 2000/56 EU). According to the above BOPA can be written as follows: BOPA index = log $(f_P/(f_A + 1) + 1)$, where f_P is the opportunistic polychaete frequency (ratio of the abundance of the opportunistic polychaetes to the total abundance in the sample) and f_A is the amphipod frequency after excluding the opportunistic *Jassa* genus (ratio of the abundance of the amphipods except of *Jassa* to the total abundance in the sample). The addition of the two monad terms in the formula was done to allow the division when f_A is null $(f_A + 1)$ and to allow a logarithmic calculation when f_P is zero $(+1)$. The index ranges from 0, when there are no opportunistic polychaetes indicating an area with excellent environmental conditions, to log2, when there are no amphipods (except of *Jassa*) indicating an area with degraded environmental conditions. The index has been calibrated to define the five classes of environmental quality (**Table 1**) set by the WFD.

Function and Evaluation of the P/A Index and the BOPA Index

Soft-substratum communities

In soft bottoms, which constitute the largest ecosystem in terms of spatial coverage, granulometry is thought to be the major structuring factor. As the sediments become coarser, polychaetes diversity tends to increase, until larger clusters prevail (i.e., boulders, pebbles, and gravels) that will unavoidably cause a decline to the interstitial fauna. In soft sediments, few species of amphipods occur, mainly of the genus *Ampelisca*. The performance of P/A index has been comprehensively tested in soft-substratum communities of the English Channel and the NW Spain coasts after some severe oil spill events. At the most affected

Table 1 Lower limits of the BOPA index and correspondence to the five classes of ecological quality (EcoQ) set by the WFD Reproduced from Dauvin, J.C., Ruellet, T., 2007. Polychaete/amphipod ratio revisited. Marine Pollution Bulletin 55, 215–224.

BOPA index	*EcoQ*
≤ 0.04576	High
≤ 0.13966	Good
≤ 0.19382	Moderate
≤ 0.26761	Poor
≥ 0.26761	Bad

communities P/A index rapidly increased in values for about 3 years after the episode; afterwards, P/A gradually decreased. These results were attributed to the very low abundance of amphipods due to their high mortality after the oil spill, and to their subsequent recovery, as they started to re-colonize the sediment when the environmental conditions improved. An exception to the above pattern was observed at two more sites, where the ratio showed constantly high values during the study period, due to a very strong dominance of opportunistic polychaetes, such as *Chaetozone setosa*, *Spiochaetopteros costarum*, *Melinna palmata*, and *Diplocirrus glaucus*. So, the spill seemed to have little effect on these two sites, as the community structure is probably driven by some other ecological factors. The BOPA index has been also tested using the previous data sets reporting moderate to poor ecological quality, reflecting a decrease in amphipod abundance and a dominance of opportunistic polychaetes. After a decade BOPA values decreased again indicating the recovery of amphiopods that originally dominated the studied benthic community.

The performance of the same methodology in several communities of the eastern Mediterranean (Thermaikos Gulf), gave confounded results (**Fig. 1**).

Thermaikos Gulf has been characterized as moderately polluted by several methods (i.e., diversity indices, biotic indices, abundance/biomass comparison, multivariate analyses), with a northward gradient toward organic enrichment. Yet, P/A values were higher to the less-disturbed sites (four stations), as opposed to the organically polluted ones (two stations). This was rather unexpected, since all polychaetes collected from the last sites were typical opportunistic species. However, these two sites were characterized by significantly reduced species richness and abundance, linked to their proximity to a small outfall acting as a point source of various pollutants. Considering amphipods, a total of 14 species have been collected, among which *Elasmopus rapax* and *Ampelisca diadema* – two well-documented tolerant species (Antoniadou *et al.*, 2011; Simboura and Zenetos, 2002) – dominated. In that case we have two facts that violate the theoretical assumptions of P/A index: (1) some samples were very impoverished, creating large problems to the function of most ecological indicators and (2) although none *Jassa* amphipod has been recorded, most species are considered as tolerant according to updated information. Thus, the basic thought behind P/A (i.e. polychaeta=resistant to pollution; amphipoda=sensitive to pollution) seems inappropriate to these results. A probable explanation would consider the differences in the sediment type between these two examples; the sorted fine sands or muddy fine sands in Atlantic, as opposed to the unsorted sediments of Thermaikos with shell fragments of various sizes creating a type of hard substratum which attracts several amphipode species. Therefore, the wide-scale applicability of P/A seems to be restricted, as it is affected by the geography and the habitat heterogeneity that influence the biodiversity of an area. The application of the BOPA index at the same data set gave slightly different results (**Fig. 2**), although again overestimating the ecological quality at the most disturbed stations (5 and 6). Ecological quality ranged from excellent to poor, mostly classified as good or moderate, following a decreasing temporal trend. In all cases polychaetes dominated benthic communities. A clear problem in the applicability of the index derives from the fact that only *Jassa* amphipods are considered as opportunistic, whereas according to recent publications almost all amphipods collected from the study area (i.e. 10 of the 14 identified species) are assigned as tolerant.

Hard-substratum communities

Until recently, very few indices have been proposed for biomonitoring hard-substratum communities, despite the attractiveness of the subject, mostly due to the difficulties involved in the research of rocky bottoms. Recent efforts (biomonitoring of the photophilic, the sciaphilic, and port communities) conducted in the Mediterranean gave the opportunity for the establishment of such indices.

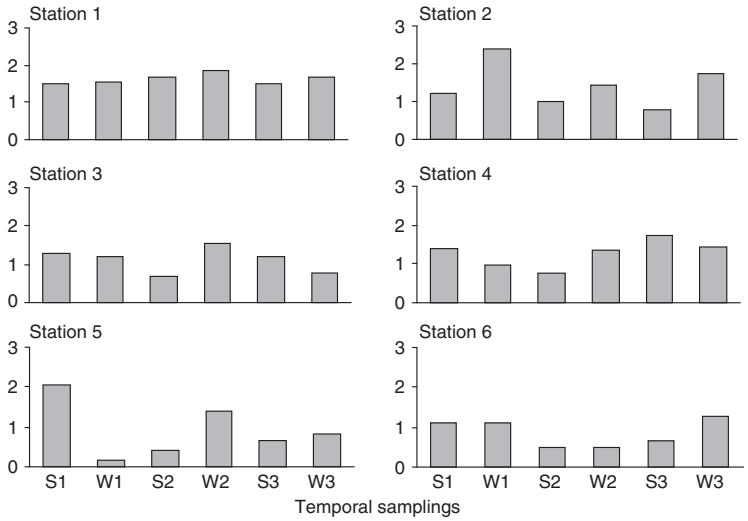


Fig. 1 The performance of P/A index in assorted soft sediments (Thermaikos Gulf), where S=summer, W=winter, and 1–3 the successive years of the study. Note that station 5 and 6 were the most affected sites.

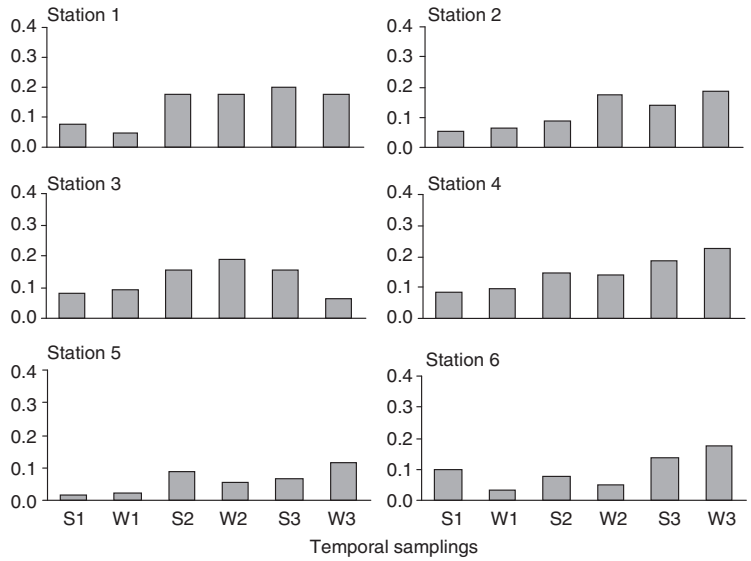


Fig. 2 The performance of BOPA index in assorted soft sediments (Thermaikos Gulf), where S = summer, W = winter, and 1–3 the successive years of the study. Note that station 5 and 6 were the most affected sites.

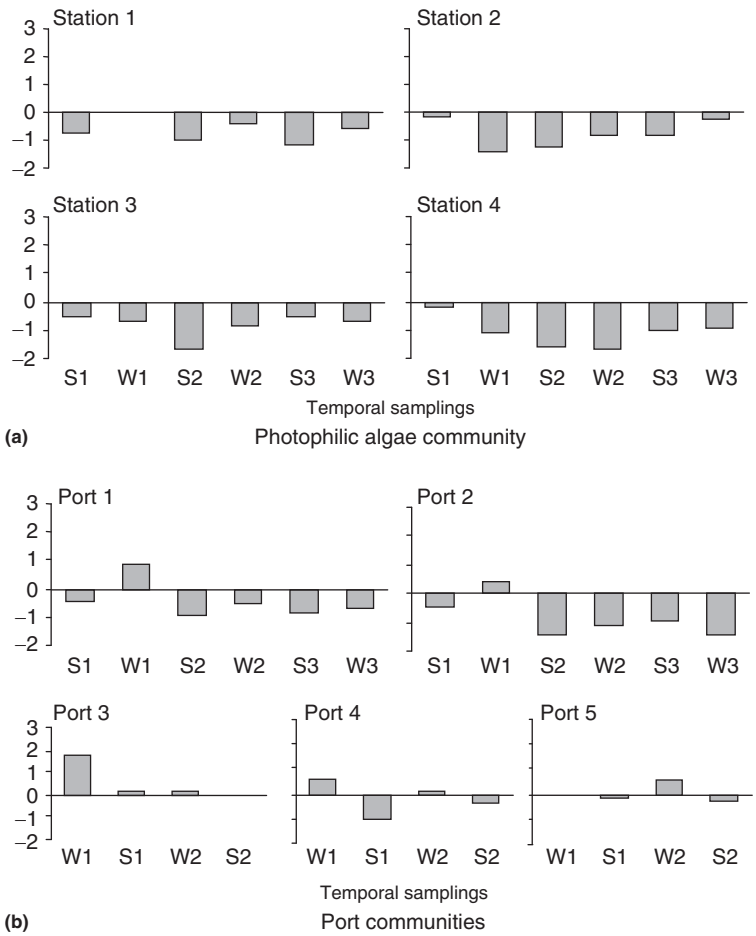


Fig. 3 The performance of P/A index on organically polluted hard-substratum communities (Thermaikos Gulf, North Aegean Sea), where S = summer, W = winter, and 1–3 the successive years of the study.

On rocky bottoms, zoobenthos is commonly related to algal structure and zonation. This does not necessarily mean dependence to the presence of particular macroalgal species, but rather to the morphological complexity and architecture of the algal cover and epiphytes. P/A index was applied on both a disturbed and undisturbed coastal area of the Aegean (Thermaikos Gulf and Chalkidiki peninsula, respectively). The results clearly show that the application of both indices is inappropriate in hard bottom (**Fig. 3–5**). P/A gains negative values in all the disturbed sites, whereas BOPA is below 0.1 in all cases classifying the ecological status as excellent or good considering both disturbed and undisturbed sites.

Thus, both indices clearly failed to detect any pollution effect. A possible reason of these results is the different response of the fauna in hard substratum with respect to pollution that contradicts P/A index assumptions. In hard bottoms, as opposed to soft sediments, amphipods constitute a major group with many species capable of proliferating by the increased organic load, building very dense populations. So, in the case of the photophilic algae community in Thermaikos, the species *Elasmopus rapax*, *Corophium sextonae*, *Caprella acanthifera*, and *Stenothoe monoculoides* increased their density, as did the species *Elasmopus rapax*, *Corophium acutum*, *C. sextonae*, and *Erichthonius brasiliensis* in ports. In contrast, in the undisturbed sciaphilic algae community, no amphipod

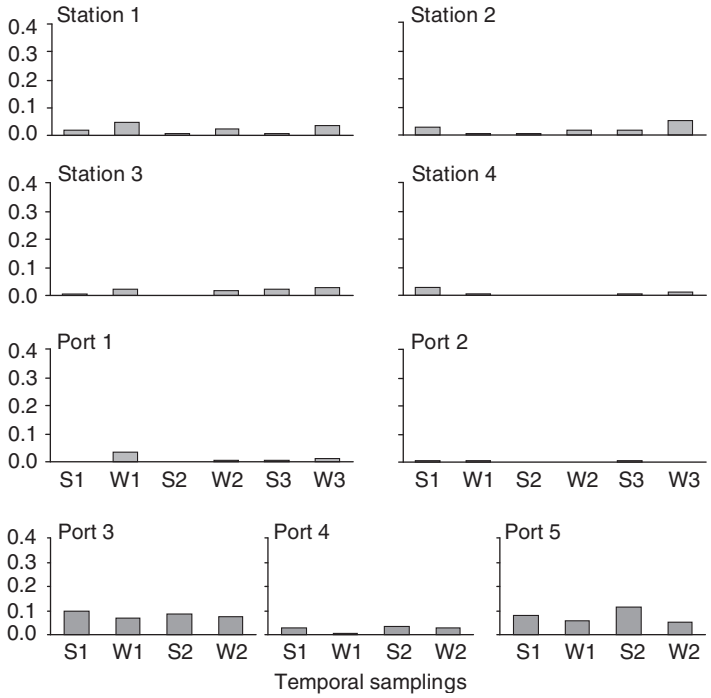


Fig. 4 The performance of BOPA index on organically polluted hard-substratum communities (Thermaikos Gulf, North Aegean Sea), where S = summer, W = winter, and 1–3 the successive years of the study.

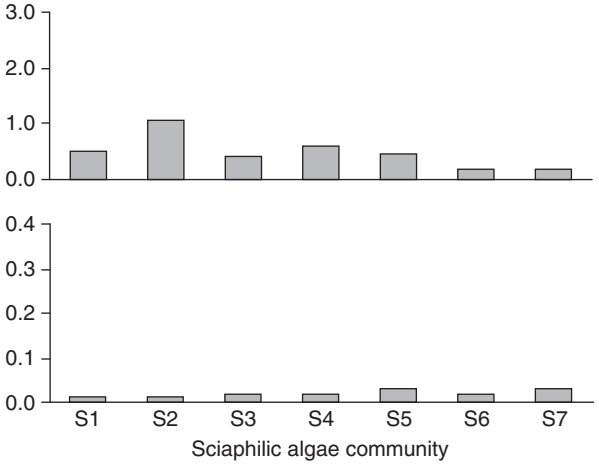


Fig. 5 The performance of P/A index (upper graph) and BOPA index (lower graph) on undisturbed hard-substratum communities (Chalkidiki peninsula, North Aegean Sea), where S1–S7 are the sampling sites.

species showed increased dominance. Consequently, the index performs unlikely to its predictions (P/A≤1=unpolluted; P/A1=polluted), since it is the denominator that increases with the pollution effect, as if the polychaetes were the sensitive group and amphipods the resistant. Moreover, as also stated previously, many amphipod species apart from the genus *Jassa* are tolerant to organic input violating the theoretical background of both indices and of BOPA in particular.

Perspectives and pitfalls

Any ecological index should hold some general attributes: it should (1) be robust and easy to handle, without the need of either laborious field or laboratory work, or complicated calculations to be used by non-experts; (2) be sensitive to any pollution event, as well as to detect gradients of disturbance; (3) assess different types of contamination; (4) incorporate seasonal variations; and (5) be applicable to various ecosystems and geographical areas.

P/A and BOPA indices follow some of the above desirable features. They are based on very simple calculation formulae and appear sensitive to the assessment of various contaminants, as they have been successfully implemented to detect alterations such as oil spill and waste material disposal. Both indices integrate seasonal oscillations of the fauna, whereas BOPA index can be applied in very poor communities, in which most indices fail to give valid results. BOPA seems to be more promising than P/A index for ecological quality assessment, mostly because it can be applied also in very poor communities, and because it is well fitted to the objectives of the WFD. However, BOPA index must be tested over geographic sectors and habitat types, and it should be, probably, modified by excluding all opportunistic amphipod species from its denominator to improve its reliability. A question arises on the discriminative power, robustness, and wide-scale applicability of both indices. Obviously the applicability of P/A and BOPA index is completely improper on hard bottoms, while it is dubious considering assorted sediments, in which the structure of polychaete and amphipode fauna is highly variable and driven mainly by a complex of ecological factors. Also, the performance of P/A index is bad when samples are impoverished or when polychaetes are not the dominant taxonomic group, while in the case of BOPA index the inclusion of only *Jassa* genus in the opportunistic amphipods can be strongly argued. Due to the above drawbacks P/A index has a rather limited applicability and approaches of ecological quality assessment should rank the entire benthic fauna with respect to either its sensibility/tolerance to pollution (e.g. biotic indices such as AMBI or BENTIX), or a specific taxonomic group (e.g. the polychaete index of Bellan, the amphipod index of Bellan-Santini, or the peracarid index of Chintiroglou). It is true, however, that the entirety of biotic indicators suffer from some of the above weaknesses (Blanchet *et al.*, 2008; Borja and Dauer, 2008; Borja *et al.*, 2008, 2009; Dauvin *et al.*, 2007; Diaz *et al.*, 2003), and it seems that no approach is a panacea offering a solution in all cases of environmental deterioration, demonstrating the necessity for intense scientific research on this topic under a multi-criteria approach.

General Remarks

Scientific research for the ecological quality assessment of coastal ecosystems has been intensified during the last decade. However, the general application of ecological indices is till date problematic and a complete proposal is still missing, mainly due to the high complexity of these systems and the general lack of adequate time-series data. Four basic concepts have been deployed (the diversity–stability hypothesis, the rivet hypothesis, the redundant species, and the idiosyncratic response hypothesis), among which the idiosyncratic response seems to better fit the marine environment. Each geographic sector contains inter-correlated populations. Several species thought as sensitive in one area, are more resistant to another. Most important is the national cooperation in broad geographic sectors to detect the environmental limits for each community, keeping in mind that each hypothesis should assess the divergence from a reference condition, in which the benthic community holds its normal state. To successfully meet this goal, the science of complexity, studying chaos and order, could be a relevant tool playing a leading role.

See also: Aquatic Ecology: Ecosystem Health Indicators—Freshwater Environments

References

Antoniadou, C., Chintiroglou, C., 2005. Response of polychaete populations to disturbance: An evaluation of methods in hard substratum. Fresenius Environmental Bulletin 14, 1066–1073.

Antoniadou, C., Sarantidis, S., Chintiroglou, C., 2011. Small-scale spatial variability of zoobenthic communities in a commercial Mediterranean port. Journal of the Marine Biological Association of the UK 91, 77–89.

Bellan, G., Desrosiers, G., Willsie, A., 1988. Use of an annelid pollution index for monitoring a moderately polluted littoral zone. Marine Pollution Bulletin 12, 662–665.

Bellan-Santini, D., 1981. Influence des pollutions sur le peuplement des amphipodes dans la biocénose des algues photophiles. Téthys 10, 185–194.

Bellan-Santini, D., Lacaze, J.-C., Poizat, C., 1994. Les biocénoses marines et littorales de Méditerranée, synthèse, menaces et perspectives. Paris: Collection Patrimoine Naturelle, Secrétariat de la faune et de la flore, Muséum National d'Histoire Naturelle.

Blanchet, H., Lavesque, N., Ruellet, T., *et al.*, 2008. Use of biotic indices in semi-enclosed coastal ecosystems and transitional waters habitats – Implications for the implementation of the European Water Framework Directive. Ecological Indicators 8, 360–372.

Borja, A., Dauer, D.M., 2008. Assessing the environmental quality status in estuarine and coastal systems: Comparing methodologies and indices. Ecological Indicators 8, 331–337.

Borja, A., Franco, J., Perez, V., 2000. A marine biotic index to establish the ecological quality of soft-bottom benthos within European estuarine and coastal environments. Marine Pollution Bulletin 40, 1100–1114.
Borja, A., Bricker, S.B., Dauer, D.M., *et al*., 2008. Overview of integrative tools and methods in assessing ecological integrity in estuarine and coastal systems worldwide. Marine Pollution Bulletin 56, 1519–1537.
Borja, A., Miles, A., Occhipinti-Ambrogi, A., Berg, T., 2009. Current status of macroinvertebrate methods used for assessing the quality of European marine waters: Implementing the Water Framework Directive. Hydrobiologia 633, 181–196.
Bustoz-Baez, S., Frid, C., 2003. Using indicator species to assess the state of macrobenthic communities. Hydrobiologia 496, 299–309.
Chintiroglou, C., Antoniadou, C., Baxevanis, A., *et al*., 2004. Peracarida populations of hard substrate assemblages in ports of the NW Aegean Sea (eastern Mediterranean). Helgoland Marine Research 58, 54–61.
Chintiroglou, C., Antoniadou, C., Krestenitis, Y., 2006. Can polychaetes be used as a surrogate group in assessing ecological quality in soft bottom communities (NE Thermaikos Gulf)? Fresenius Environmental Bulletin 15, 1199–1207.
Clarke, K.R., Warwick, R.M., 2001. Change in marine communities: An approach to statistical analysis and interpretation. Plymouth: Plymouth Marine Laboratory, Natural Environment Research Council.
Dauvin, J.C., 2000. The muddy fine sand *Abra alba–Melinna palmata* community of the Bay of Morlaix twenty years after the Amoco Cadiz oil spill. Marine Pollution Bulletin 40, 528–536.
Dauvin, J.C., Ruellet, T., 2007. Polychaete/amphipod ratio revisited. Marine Pollution Bulletin 55, 215–224.
Dauvin, J.C., Comez Gesteira, L., Salvande Fraga, M., 2003. Taxonomic sufficiency: An overview of its use in the monitoring of sublittoral benthic communities after oil spills. Marine Pollution Bulletin 46, 552–555.
Dauvin, J.C., Ruellet, T., Desroy, N., Janson, A.L., 2007. The ecological quality status of the Bay of Seine and the Seine estuary: Use of biotic indices. Marine Pollution Bulletin 55, 241–257.
Diaz, R.J., Solan, M., Valente, R.M., 2003. A review of approaches for classifying benthic habitats and evaluating habitat quality. Journal of Environmental Management 73, 165–181.
Ellis, D., 1985. Taxonomic sufficiency in pollution assessment. Marine Pollution Bulletin 16, 459.
Gaston, K., Spicer, J.I., 1998. Biodiversity: An introduction. Oxford: Blackwell.
Gómez Gesteira, J.L., Dauvin, J.-C., 2000a. Amphipods are good bioindicators of the impact of oil spills on soft-bottom macrobenthic communities. Marine Pollution Bulletin 40, 1017–1027.
Gómez Gesteira, J.L., Dauvin, J.-C., 2000b. Impact of the Aegean Sea oil spill on the subtidal fine sand macrobenthic community of the Ares-Betanzos Ria (Northwest Spain). Marine Environmental Research 60, 289–316.
Hardman-Mountford, N.J., Allen, J.I., Frost, M.T., *et al*., 2005. Diagnostic monitoring of a changing environment: An alternative UK perspective. Marine Pollution Bulletin 50, 1463–1471.
Rosenberg, R., Blomqvist, M., Nilsson, H.C., Cederwall, H., Dimming, A., 2004. Marine quality assessment by use of benthic species-abundance distributions: A proposed new protocol within the European Union Water Framework Directive. Marine Pollution Bulletin 49, 728–739.
Simboura, N., Zenetos, A., 2002. Benthic indicators to use in ecological quality classification of Mediterranean soft bottom marine ecosystems, including a new biotic index. Mediterranean Marine Science 3, 77–111.

Protected Area

Yan Xie, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

Introduction

Biodiversity is the key foundation of human survival and development. However, global biodiversity is declining rapidly and has become the most severe factor threating safety of humanity. From 1970 to 2012, the global wildlife populations fell by 58%, and we may witness a 2/3 decline during the period from 1970 to 2020. Protected areas (PA) have long been the cornerstone of biodiversity conservation tactics. PAs are critical for maintaining a healthy environment for people and nature. They are essential for biodiversity conservation and vital to the cultures and livelihoods of indigenous peoples and local communities. They also deliver clean air and water, bring benefits to millions of people through tourism, and provide protection from climate change and natural disasters.

IUCN defines protected areas as "A clearly defined geographical space, recognized, dedicated and managed, through legal or other effective means, to achieve the long-term conservation of nature with associated ecosystem services and cultural values." In view of the global current and future population, development trend and crisis pressure, the limited funds, scientific research, manpower, technology, time, etc. must be focused on the most important hot spots and key areas—PAs.

History and Current Status

PAs have a long history. For example, India set aside specific areas for the protection of natural resources over 2000 years, Europe protected some areas as hunting grounds nearly 1000 years ago. Punkaharju Esker Nature Reserve established in 1843 in Finland is one of the oldest PAs in the world. The first true national park came in 1872 with the dedication of Yellowstone by United States law "as a public park or pleasuring ground for the benefit and enjoyment of the people." 1866, the British Colony of New South Wales in Australia reserved 2000 ha of land for protection and tourism. Later additions created a park complex now known as the Blue Mountains National Park. In 1885, Canada gave protection to hot springs in the Bow Valley of the Rocky Mountains, an area later named Banff National Park. China's first PA could be the Nanhaizi—a royal hunting ground back to 1000 BC. Nanhaizi was the biggest wetland in South Beijing with 216.5 km^2 which was bigger than Beijing city in Qing Dynasty (1636–1912). It preserved the last population of David's deer which is endemic species of China. Before Nanhaizi was destroyed in 1900, some deer were sent to Europe and population expanded. After more than 80 years, 20 deer were reintroduced back to China and now the population increased to 3000. This demonstrated that Nanhaizi could be considered as a PA.

The modern PA movement has 19th century origins in the then "new" nations of Australia, Canada, New Zealand, South Africa, and the United States, but during the 20th century the idea spread around the world. The result was a remarkable expansion in the number of PAs. Nearly every country passed PA legislation and designated sites for protection. PAs continue to be established, and received a boost in 2004 when the Convention on Biological Diversity (CBD) agreed an ambitious Program of Work on PAs (Fig. 1).

The expansion history largely depending on political status of each country. The modern PAs in China were started to establish in 1956, with the first 20 years increased very slow due to cultural revolution, started to increase quickly from 1980s when the country opened to the world, and got its fastest development in numbers, areas and management quality since early 21st century. Now PAs have reached to over 10,000 occupying over 18% of the terrestrial land of China.

Realizing the importance of PAs to biodiversity conservation and therefore for survival of human being, the world's governments have committed to establish and manage PAs through Sustainable Development Goal Targets 14.5 (marine), 15.1 (terrestrial and freshwater), and 15.4 (mountains), Aichi Target of CBD 11 of the Strategic Plan for Biodiversity 2011–20, and numerous other international agreements such as the World Heritage Convention and the Ramsar Convention on Wetlands. By April 2016, under 15% of the world's terrestrial and inland waters, just over 10% of the coastal and marine areas within national jurisdiction, and approximately 4% of the global ocean are covered by PAs (Fig. 2). The Parties of CBD have agreed 20 "Aichi Targets" in which Aichi Target 11 addresses PAs, calling for at least 17% of terrestrial and inland water, and 10% of coastal and marine areas to be

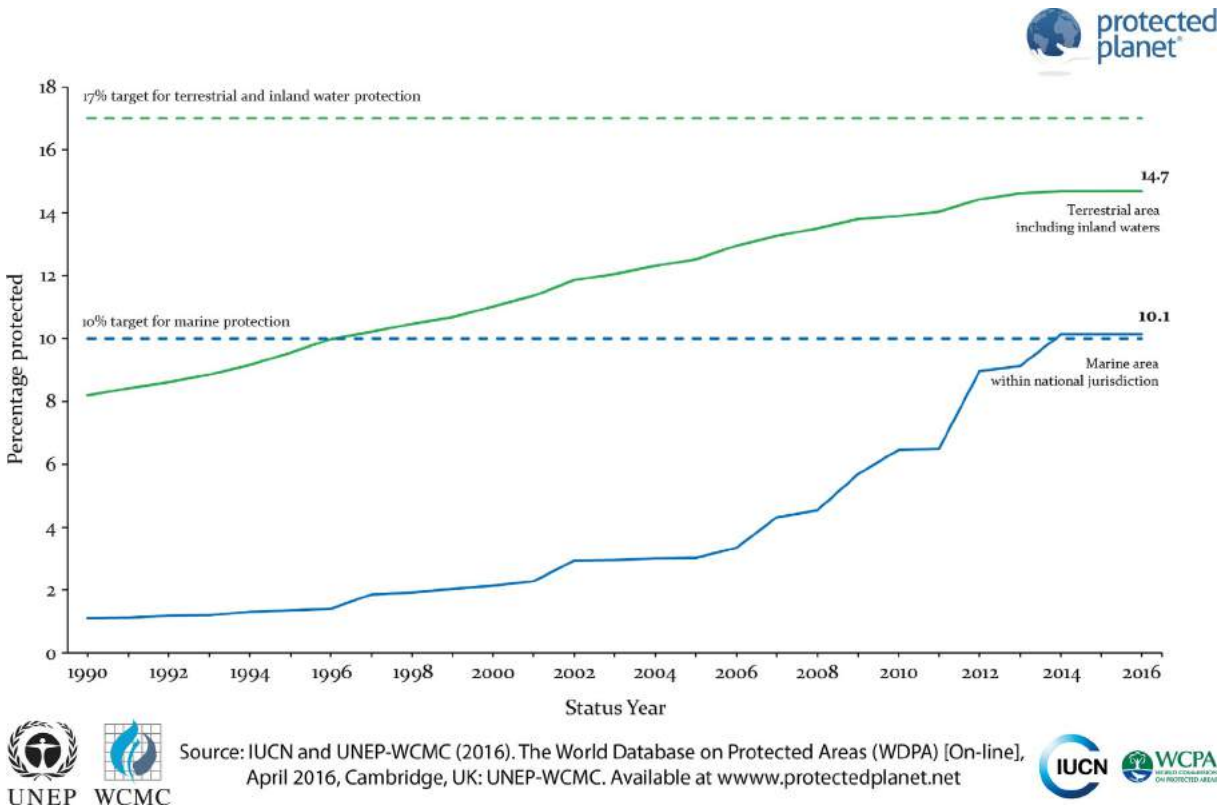


Fig. 1 Development of protected areas in the World.

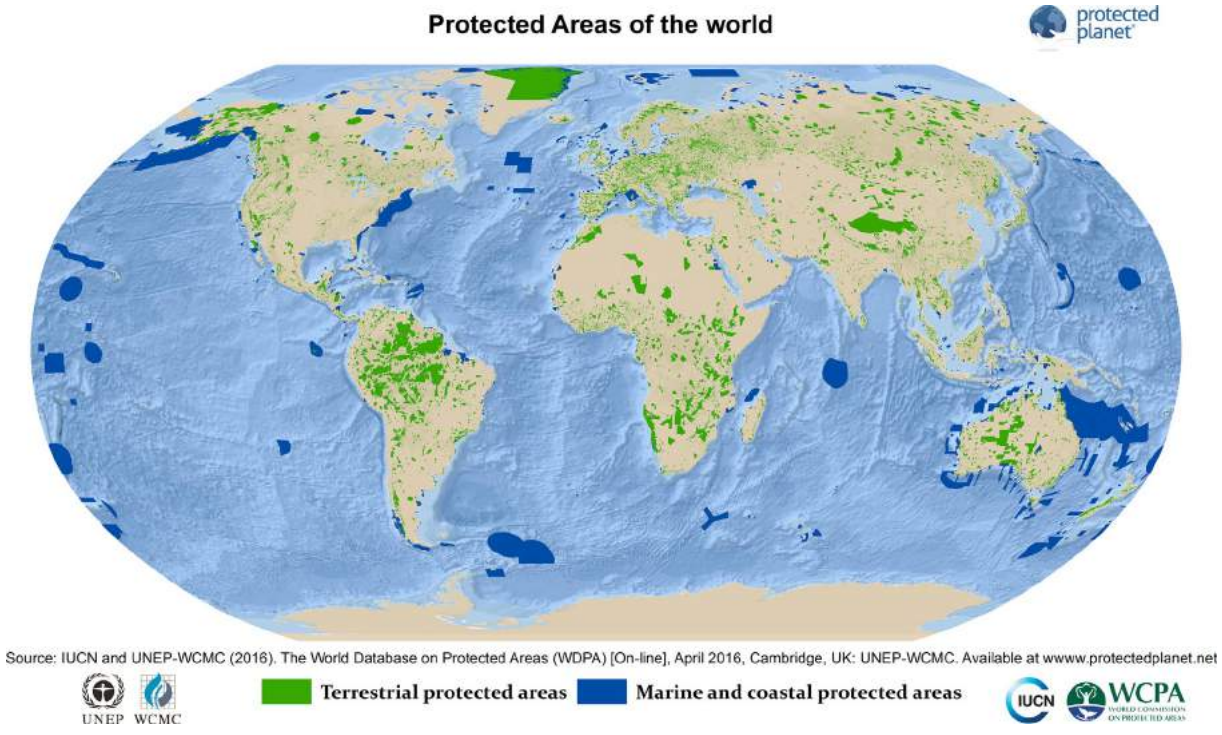


Fig. 2 Map of protected areas in the World in 2016.

established by 2020, including especially areas of particular importance for biodiversity and ecosystem services that are conserved through effectively and equitably managed, ecologically representative and well-connected systems of PAs and other effective area-based conservation measures, and are integrated into the wider landscapes and seascapes.

Role of PAs

Ensuring a more sustainable future for people and the planet will require greater recognition of the important role that PAs play in underpinning sustainable development. PAs are a key part of national and local responses to address harmful incentives to biodiversity, biological invasions, anthropogenic impacts and climate change challenges will help to halt biodiversity loss, improve food and water security, increase the resilience of vulnerable human communities to cope with natural disasters, and promote human health and well-being. PAs also play a key role in enhancing fish stocks and strengthening sustainable management of fisheries, and PAs in landscapes can promote sustainable production of natural resources in areas under agriculture, aquaculture and forestry.

However, the role of PAs largely depends on the management effectiveness. Conservation effectiveness of PAs varies in PAs and countries, with some showing positive on habitat or species conservation and some negative. However, selected studies that had counterfactual assessments of PAs, either of species populations or habitats, where outcomes could be attributed to PAs. Evaluation on change in the Red List Index for species for which PAs cover more than half of their important sites (key biodiversity areas) relative to those species for which less than half of their key biodiversity areas are protected, the result showed that the rate at which species are sliding toward extinction was halved for species with more than half of their range inside PAs. Globally, the Living Planet Index reports an average decline in vertebrate populations of 52% from 1970 to 2015. The same index of vertebrate populations for PAs shows a decline of only 18%. The synthesis in 2015 concluded that partial PAs significantly enhanced density and biomass of fish relative to open fishing areas, while fully no-take reserves in turn yielded a significantly higher biomass of fish relative to partially PAs.

A global analysis of the effectiveness of the full range of marine PAs found that, on average, coral cover within marine PAs remained constant, while coral cover on unprotected reefs declined, although the differences were not statistically significant. While the short-term differences between unprotected and protected reefs were not statistically significant, the trend in the differences could be significant over the long-term if the trends remained constant. Their results also suggested that older marine PAs were generally more effective in preventing coral loss.

Problems Existing in PA Management

These PAs are cornerstones of biodiversity conservation policy in the world and important in maintaining ecosystem functions and services. It is also the case, however, that obsolete, unreasonable, or unbalanced planning and legal protections for PAs make the overall management of these different PAs very poor. The many roles PAs could play in biodiversity conservation, therefore, far from being fulfilled.

Many PAs are not yet fully implemented or managed. Marine PAs are lagging far behind land and inland water PAs although there are now great efforts to rectify this situation. This is necessary because although the rate of growth has been impressive, many PAs have been set up in remote, unpopulated or only sparsely populated areas such as mountains, ice-fields and tundra and there are still notable gaps in PA systems in some forest and grassland ecosystems, in deserts and semideserts, in fresh waters and, particularly, in coastal and marine areas. Many of the world's wild plant and animal species do not have viable populations in PAs and a substantial proportion remain completely outside PAs.

The tension between community development and biodiversity conservation can be difficult to resolve. A large number of key ecological-function areas are located in impoverished and underdeveloped regions. Restrictions on resource use in these areas need to be accompanied by better compensation mechanisms and support policies.

Following are the key areas for improving effectiveness of PA management.

PA Management Categories

PAs are by no means uniform entities however; they have a wide range of management aims and are governed by many different stakeholders. At one extreme a few sites are so important and so fragile that no-one is allowed inside, whereas other PAs encompass traditional, inhabited landscapes and seascapes where human actions have shaped cultural landscapes with high biodiversity. Some sites are owned and managed by governments, others by private individuals, companies, communities and faith groups. PAs in China could be described using established national categories, such as nature reserves, scenic and historic areas, and forest parks, as well as international categories, such as natural and mixed heritage sites, wetlands of international importance, global geo-parks, and biosphere reserves, as well as new categories as national parks since 2013, and they are current now managed by a dozen governmental departments. There are also quick investment to establishing private PAs for commonweal with donation funding from private companies.

In order to speak a common language, The IUCN developed PA management categories as an important global standard for the planning, establishment and management of PAs, recognized by the CBD and adopted by many countries. Due to PAs exist in an astonishing variety—in size, location, management approaches and objectives, the IUCN PA management categories are not a straitjacket but a global framework to guide improved application of the categories.

IUCN first effort to develop the PA management categories was in 1934 and after many revisions, the IUCN General Assembly meeting approved the new system in 1994 and the first guidelines were published by IUCN and the World Conservation Monitoring Centre (WCMC) later that year. IUCN secured the endorsement of the system by the CBD, at the seventh Conference of the Parties in February 2004. The latest revision was done in 2008 with publication of Guidelines for Applying PA Management Categories. The current version of the IUCN PA management categories are summarized below.

Ia *Strict nature reserve*: Strictly protected for biodiversity and also possibly geological/geomorphological features, where human visitation, use and impacts are controlled and limited to ensure protection of the conservation values.

Ib *Wilderness area*: Usually large unmodified or slightly modified areas, retaining their natural character and influence, without permanent or significant human habitation, protected and managed to preserve their natural condition.

II *National park*: Large natural or near-natural areas protecting large-scale ecological processes with characteristic species and ecosystems, which also have environmentally and culturally compatible spiritual, scientific, educational, recreational, and visitor opportunities.

III *Natural monument or feature*: Areas set aside to protect a specific natural monument, which can be a landform, sea mount, marine cavern, geological feature such as a cave, or a living feature such as an ancient grove.

IV *Habitat/species management area*: Areas to protect particular species or habitats, where management reflects this priority. Many will need regular, active interventions to meet the needs of particular species or habitats, but this is not a requirement of the category.

V *Protected landscape or seascape*: Where the interaction of people and nature over time has produced a distinct character with significant ecological, biological, cultural and scenic value: and where safeguarding the integrity of this interaction is vital to protecting and sustaining the area and its associated nature conservation and other values.

VI *PAs with sustainable use of natural resources*: Areas which conserve ecosystems, together with associated cultural values and traditional natural resource management systems. Generally large, mainly in a natural condition, with a proportion under sustainable natural resource management and where low-level nonindustrial natural resource use compatible with nature conservation is seen as one of the main aims.

The category should be based around the primary management objective(s), which should apply to at least three-quarters of the PA—the 75% rule.

More and more countries have adopted the IUCN PA management category system, including Australia, South Korean, New Zealand, Canada, South Africa, Laos, and India.

PA Zonation

Many PA however promote multiple-uses (and are thus not only focused on nature conservation) or suffer from human disturbance inside their borders. One option to answer the challenges of combining the needs of economic and social development and the ecological requirements of the species and habitats that PA seek to protect is to design zoning schemes inside PA. Such zoning schemes can spatially and temporally refine the list of activities allowed and banned through the mapping of "functional zones." Different zoning models are used in different countries and within single countries. The UNESCO's Man and the Biosphere Programme (MAB)'s system of core, buffer and transition zones holds considerable influence and many countries adopted a similar system with three zones, given different names by different countries. Starting to become widespread in the 1980s, buffer zones are some of the most common zoning tool and are usually defined as zones where only a limited list of nondestructive activities are allowed, and located between strict conservation zones and zones outside the PA. However, the zonation system is too crude and failed guide PA management practices. More detailed zonation system, with clear zoning names referring to management strictness, clear allowed activities, level of impact to nature and permit management systems should be established and promoted around the world, is believed to be more effective to improve management level of PAs (Table 1).

System Planning

System planning is an organized way to carry out conservation planning for PAs at the macro level. It is recognized as a key management principle for increase the effectiveness of in situ biodiversity conservation. As explained by IUCN-WCPA, system planning in relation to PAs is about: (a) defining the priority of PAs as a worthwhile national concern; (b) defining the relationships between different units and categories of PAs, and between PAs and other relevant categories of land or sea; (c) taking a more strategic view of PAs; (d) defining roles of key players in relation to PAs and the relationships between these players; this may include building support and a constituency for PAs; (e) identifying gaps in PA coverage (including opportunities and needs for connectivity) and deficiencies in management; and (f) identifying current and potential impacts—both those affecting PAs from surrounding land or sea, and those emanating from PAs which affect surrounding land or sea.

Table 1 Proposed PA zonation system

Zone name	*Description*	*Activities allowed*	*Impact on biodiversity*	*Restrictions on activities*
1. Closed zone	Sanctuaries for species very sensitive to disturbances. Human disturbances strictly controlled	Nondestructive scientific monitoring Enforcement patrols and footpaths Fireproofing Very limited religious practice in sacred sites	Very low	Permit needed for scientific research. Religious practice limited to a very low number of individuals requiring permits
2. Active management zone	Areas where limited human intervention is necessary to manage and restore habitats in order to achieve the conservation goals	Habitat and species management (including through grazing) Ecological restoration Local species reintroduction	Positive to low	Permit needed for scientific research and expeditions
		Small-scale expeditions guided by PA staff Scientific research with little disturbance (e.g., specimen collection, interventionist conservation experiment) Trail for motorized vehicles (staff only)	Low	
3. Limited use zone				
3A. Limited visitor use zone	Areas for light visit, tourism and recreation	Nature tourism Access roads, weather shelters, rest pavilions, camp sites Traditional wilderness-based lifestyle and customs, (low densities and in ways compatible with the conservation objectives)	Medium	Tourism submitted to entry fees. No high impact accommodation or recreational activities unrelated to the nature area (golf, etc.). Control on access roads
3B. Limited resource use zone	Areas for sustainable natural resource utilisation	Hunting (if allowed by national laws) Fishing Collection of wild nontimber forest products (NTFP)	Medium	Resource use permit system ensuring the viability of populations
		Grazing	Medium[a]	
4. Reasonable use zone	Areas for PA management, heavier tourism and local residents	Tourist centre (accommodations, etc.) High intensity tourist areas Staff housing PA administration office areas Parking Residential areas Small-scale agriculture for auto-consumption in the PA	High	Ban on fertilizer and pesticide use wherever it can impact water quality or feeding habitats
5. External support zone (outside the PA)	Areas outside the PA but submitted to increased environmental scrutiny and collaborating closely with PA managers. The aim of such areas is to mitigate the impacts of surrounding human activities and PAs on each other	Large-scale agriculture Plantations Fish ponds	Very high[a]	Ban on fertilizer and pesticide use wherever it can impact water quality or feeding habitats. Extractive industries, etc. submitted to environmental impact assessment and a no net loss of biodiversity standard
		Extractive industries, pipelines, power lines, cell phone towers, etc.	Very high	

Unless stated otherwise, the impact described in column "impact on biodiversity" is negative.
[a]These activities can have a positive impact if they are part of species or habitat management actions (implemented in a two active management zone).

Over the years, IUCN-WCPA has developed guidelines on the main characteristics and considers that the main characteristics of a PAs system should include:

- *Representativeness, comprehensiveness and balance*: ability to represent or sample the full variety of biodiversity and other features such as landform types, and landscapes or seascapes of cultural value, so as to protect the highest quality examples, especially threatened and underprotected ecosystems, and species globally threatened with extinction.

- *Adequacy*: supporting the viability of ecosystem processes as well as species, populations and communities that make up the country's biodiversity.
- *Coherence and complementarity*: the extent to which each site makes a positive contribution to the system as a whole.
- *Consistency*: the application of management objectives, policies and classifications to individual sites under comparable conditions in standard ways.
- *Cost-effectiveness, efficiency and equity*: an appropriate balance between the costs of and benefits flowing from PAs, equity in their distribution, and efficiency in terms of the minimum number and size of PAs needed to achieve system objectives.
- *Persistence*: the ability to promote the long-term survival of biodiversity contained within a PA by maintaining natural processes and viable populations and by excluding or overcoming threats.
- *Resilience*: the ability to adapt and sustain primary conservation objectives of the site and the system overall in the face of climate change and other global change factors.

Legislation and Enforcement

The global conventions noted below as the main instruments concerning a nation's PAs are the CBD, UNESCO's World Heritage Convention, the Ramsar Convention and the Conservation of Migratory Species of Wild Animals (CMS). Some regional law instruments spell out important commitments and guidance for PA legal frameworks, such as African Convention on the Conservation of Nature and Natural Resources, Convention on the Conservation of European Wildlife and Natural Habitats, Convention on Nature Protection and Wild Life Preservation in the Western Hemisphere, UNESCO Man and the Biosphere Programme. Many of these commitments, as with global treaties, generate national obligations that require legislative action for implementation.

At the national level, over last 100 years, almost all countries have issued relevant legislation to secure PA supervision and management. However, legislation status varies largely among different countries and different regions. Some countries have issued laws or regulations for each PA types, such as Republic of Korean and Japan. Although diverse legislation make implementation more flexible, but duplication, gaps or contradiction among different legislation could bring problems when implementing laws. To address the issue, some countries have issued one unified law to manage all PAs. For example, Federal Nature Conservation Act 2002 of German, National System of Conservation Units 2000 of Brazil, PA Act 2003 of South Africa, PA Act (1949 revised 2015) of Kenya, National Integrated PAs System Act 1992 of Philippines, and PA System Act 1997 of Peru. Australian government issued Environment Protection and Biodiversity Conservation Law to replace several legislations issued before and unify management standards of all types of PAs. There are also many countries have include PAs managed by indigenous or local communities and private PAs into national PA management system. By 2011, Australia have 42 indigenous PAs, about 1/4 of national PA system, and India has several thousand community PAs around the whole country.

However, many countries PAs are governed not by legislation but by a patchwork quilt of administrative rules, regulations, and guidelines issued by various departments and agencies. This means, of course, that the existing legislative framework is insufficient to deal with some of the major threats to PAs, and this is a major reason why critics of PAs in these countries have referred to them as "paper parks," saying that they are characterized by the "threewithouts." They are without a management agency, without staff, and without recurrent funding.

Good legislation would largely improve PA management effectiveness. IUCN Guidelines to Legislation of PAs lists following important elements to be included into PA legislations based on good practices in the world: Policy and objectives, Institutional arrangements, Planning for PAs, Establishment of PAs, PAs management, Conservation agreements, Regulated activities, Compliance and enforcement, Environmental and social impact assessment, and Special financial tools.

Community Participation

Local people in and around PAs has long history of coexisting with nature, and their traditional culture and knowledge are part of the nature. Governments and PA managers should incorporate customary and traditional resource use, and control systems, as a means of enhancing biodiversity conservation. WWF and IUCN/WCPA have adopted principles and guidelines concerning indigenous rights and knowledge systems, consultation processes, agreements between conservation institutions, decentralization, local participation, transparency, accountability, sharing benefits and international responsibility. The five principles are as follows:

Principle 1: Indigenous and other traditional peoples have long associations with nature and a deep understanding of it. Often they have made significant contributions to the maintenance of many of the earth's most fragile ecosystems, through their traditional sustainable resource use practices and culture-based respect for nature. Therefore, there should be no inherent conflict between the objectives of PAs and the existence, within and around their borders, of indigenous and other traditional peoples. Moreover, they should be recognized as rightful, equal partners in the development and implementation of conservation strategies that affect their lands, territories, waters, coastal seas, and other resources, and in particular in the establishment and management of PAs.

Principle 2: Agreements drawn up between conservation institutions, including PA management agencies, and indigenous and other traditional peoples for the establishment and management of PAs affecting their lands, territories, waters, coastal seas, and

other resources should be based on full respect for the rights of indigenous and other traditional peoples to traditional, sustainable use of their lands, territories, waters, coastal seas, and other resources. At the same time, such agreements should be based on the recognition by indigenous and other traditional peoples of their responsibility to conserve biodiversity, ecological integrity and natural resources harbored in those PAs.

Principle 3: The principles of decentralization, participation, transparency and accountability should be taken into account in all matters pertaining to the mutual interests of PAs and indigenous and other traditional peoples.

Principle 4: Indigenous and other traditional peoples should be able to share fully and equitably in the benefits associated with PAs, with due recognition to the rights of other legitimate stakeholders.

Principle 5: The rights of indigenous and other traditional peoples in connection with PAs are often an international responsibility, since many of the lands, territories, waters, coastal seas and other resources which they own or otherwise occupy or use cross national boundaries, as indeed do many of the ecosystems in need of protection.

Further Reading

UNEP-WCMC and IUCN (2016). Protected planet report 2016. UNEP-WCMC and IUCN: Cambridge and Gland.

Eagles PF, McCool SF, Haynes CD, and Phillips A (2002) *Sustainable tourism in PAs: Guidelines for planning and management*. vol. 8. Gland: IUCN.

Butchart SH, Walpole M, Collen B, van Strien A, Scharlemann JP, Almond RE, Baillie JE, et al. (2010) Global biodiversity: Indicators of recent declines. *Science* 328(5982): 1164–1168.

Dudley N (ed.) (2008) *Guidelines for applying PA management categories*. Norwich: IUCN. https://books.google.com/books?hl=fr&lr=&id=pq4oEg58_08C&oi=fnd&pg=PR7&dq=protected+area+management+categories&ots=4CEXRSoYPx&sig=jZ7oGMSXXpwkIN64XEcePubzZv8.

Dudley N, Parrish JD, Redford KH, and Stolton S (2010) The revised IUCN PA management categories: The debate and ways forward. *Oryx* 44(4): 485–490.

Millennium Ecosystem Assessment (2005) *Ecosystems and human well-being: Synthesis*. Washington, DC: World Resources Institute/Island Press.

Watson JE, Dudley N, Segan DB, and Hockings M (2014) The performance and potential of PAs. *Nature* 515(7525): 67–73. https://doi.org/10.1038/nature13947.

Lausche B (2011) *Guidelines for PAs legislation*. Gland: IUCN. xxvi + 370.

Beltrán J (2000) *Indigenous and traditional peoples and PAs: Principles, guidelines and case studies*. Gland, Switzerland and Cambridge, UK: IUCN and WWF International, Gland, Switzerland. xi +133pp.

Reintroduction

Doug P Armstrong, Massey University, Palmerston North, New Zealand and Oceania Section Chair, IUCN Reintroduction Specialist Group
Philip J Seddon, University of Otago, Dunedin, New Zealand and Bird Section Chair, IUCN Reintroduction Specialist Group
Axel Moehrenschlager, Centre for Conservation Research, Calgary Zoological Society, Calgary, Canada and Chair, IUCN Reintroduction Specialist Group

Glossary

Assisted colonization The intentional movement and release of an organism outside its indigenous range to avoid extinction of populations of the focal species. Also called "assisted migration," especially when applied to plants.
Conservation introduction The intentional movement and release of an organism outside its indigenous range.
Conservation translocation The intentional movement and release of a living organism where the primary objective is a conservation benefit: this will usually comprise improving the conservation status of the focal species locally or globally, and/or restoring natural ecosystem functions or processes.
Ecological replacement The intentional movement and release of an organism outside its indigenous range to perform a specific ecological function.
Indigenous range All areas where a species is thought to have lived based on historic records and paleoecological evidence.
Population restoration Any conservation translocation conducted within the indigenous range.
Translocation The human-mediated movement of living organisms from one area, with release in another.
Reinforcement The intentional movement and release of an organism into an existing population of conspecifics.
Reintroduction The intentional movement and release of an organism inside its indigenous range from which it has disappeared.
Reintroduction biology Research designed to improve the outcomes of reintroductions and other translocations conducted for conservation purposes.

What Is Reintroduction?

Reintroductions are becoming increasingly frequent, and are one of many tools used worldwide to combat losses of biodiversity (Fig. 1). A reintroduction is an attempt to re-establish a wild population of a species in a location where it used to occur. It involves removing individuals from a source population, either captive or wild, and deliberately releasing them where they used to occur to restore a population. This deliberate movement with free-release is referred to as "translocation." Several types of translocations are conducted for conservation reasons. Like reintroduction, "reinforcement" is a form of "population restoration," but it differs from reintroduction in that it aims to bolster a population that still exists. In some cases it might be deemed necessary to establish a population of the species outside its indigenous range, in which case the translocation constitutes an "introduction" rather than a "reintroduction." Conservation introductions can constitute "assisted colonization" (or "assisted migration"), meaning species are introduced to new sites where they can escape from current or future threats. Alternatively, they can constitute "ecological replacement" where the species is introduced to perform the ecological function of an extinct species. The collective term "conservation translocation" is used to cover all of these types of translocation as long as they are conducted primarily for conservation benefit.

Conservation translocations are often part of species recovery programs, where the objective is to increase distribution and/or abundance of the species to ensure its long-term survival. The other main motivation is to restore ecosystems by reintroducing or trying to replace species that were components of those systems, a process often called "re-wilding" if large keystone species are involved (Donlan *et al.* 2005). Consequently, conservation translocation and ecological restoration are closely linked. Conservation translocation is also closely linked to the ecological concept of "habitat," meaning the suite of biotic and abiotic factors that determine whether a species can persist at a site. The most fundamental prerequisite for reintroduction is that there must be good reason to believe that the factors causing the local extirpation of the species have been addressed.

History of Conservation Translocations

People have moved other species from one place to another for thousands of years. Most of these movements have been accidental, for example the transmission of our parasites such as bacteria, viruses, and lice, and hitchhikers such as seeds and rodents. However, we have also deliberately tried to establish a wide range of species in new places. These species were predominantly food resources, that is, crops, livestock, and fish and game species, including many of the food resources used today. However, people also moved species for other reasons, such as erosion control, religious ceremonies, or simply because they

 doi:10.1016/B978-0-12-409548-9.10589-5

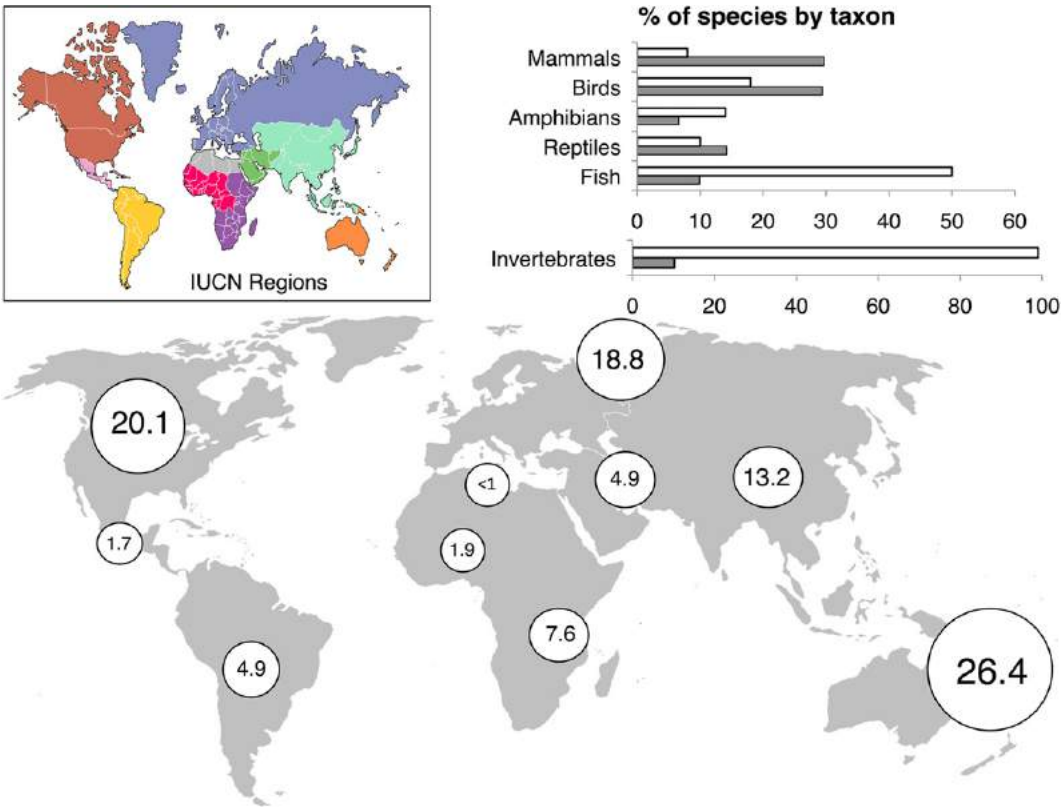


Fig. 1 Regional and taxonomic make-up of the 303 animal species featured in conservation translocations recorded in the IUCN database. The figures on the main map show the percentages of these species found in each IUCN region. The color inset map shows the 10 regions, which from west to east: North America and Caribbean (*brown*), Meso-America (*pink*), South America (*gold*), North Africa (*gray*), Central and West Africa (*crimson*), East and Southern Africa (*purple*), West Asia and the Middle East (*green*), Europe and the Mediterranean (*blue*), Asia (*pale green*), and Oceania (*orange*). The histograms compare the taxonomic makeup of the featured species (*gray bars*) to that found in nature (*white bars*). Modified from Seddon, P. J., Griffiths, C. J., Soorae, P. S., and Armstrong, D. P. (2014). Reversing defaunation: Restoring species in a changing world. *Science* **345**, 406–412.

enjoyed those species. The acclimatization societies of the 19th century tried to introduce a huge range of species to French and English colonies for a range of reasons, including nostalgia for European species and a desire to enhance the flora and fauna in the colonies. By the late 19th century it had become clear that many introduced species had become invasive, and were in fact degrading the flora and fauna through their impacts on native species.

Conservation translocations were motivated by a desire to reverse the impacts that humans have had on native species, not only through invasive species but also through our own hunting and other impacts to species' habitats. Two famous translocation programs occurred in the late 19th century on opposite sides of the world (**Table 1**). In New Zealand, it was recognized that invasive stoats (*Mustela erminea*) were driving native flightless birds to extinction. Consequently, Richard Henry was appointed caretaker of stoat-free Resolution Island which was designated a reserve. With the help of his dog, Henry caught hundreds of kākāpō (*Strigops habroptilus*; **Fig. 2**), a nocturnal parrot, and tokoeka (*Apteryx australis*), a type of kiwi, and rowed them out to an offshore island free of stoats. Unfortunately stoats swam to the island soon after and exterminated the introduced birds, but similar actions almost 100 years later saved the species from extinction (**Table 1**). In the late 19th century in the USA, it was recognized that snowy egrets (*Egretta thula*) were possibly being driven to extinction by people killing the birds for their showy plumes. The Tabasco sauce magnate Ned McIlhenny responded by reintroducing the species to a private sanctuary on Avery Island, Louisiana, using eggs sourced from remnant populations. This in turn served as a source population for other reintroductions, possibly saving the species from extinction. Ironically he later contributed to the deliberate introduction of the nutria (*Myocastor coypus*), an invasive rodent that has major negative impacts on Louisiana's wetland ecosystems to this day.

Translocations for conservation reasons have continued over the subsequent 120 years, with an apparent increase since about 1960 due to increased concerns about habitat loss, pollution, invasive species, overharvesting, and other forms of human persecution. Many of the translocations are considered great successes in terms of species recovery, ecosystem restoration, or both (**Table 1**). These successes have inspired further use of translocations, so that the scope of translocation has increased, both in terms of the conservation problems being addressed, but also the diversity of taxa and ecosystems involved (e.g, Germano and Bishop, 2009). However, it was realized by the 1980s that many conservation translocations were failing to establish populations. Although there being some failures is inevitable due to the uncertainties inherent in translocations, many translocations appear to

Table 1 Some notable conservation translocations

Date	*Translocation*
1894	First translocations of the kākāpō (*Strigops habroptilus*; **Fig. 2**) to Resolution Island in an attempt to save this flightless parrott that was rapidly declining on the New Zealand mainland due to exotic mammalian predators. This attempt failed due to stoats (*Mustela erminea*) colonizing the island. However, the kākāpō was saved from extinction by introducing it to Little Barrier and Codfish Islands in the 1980s, giving an example of a species being saved by assisted colonization
1895	Reintroduction of captive-bred snowy egrets (*Egretta thula*) to Ned McIlhenny's private wildfowl refuge, which provided a source population used to recover the species in the wild in the south-eastern USA
1907	Reintroduction of captive-bred American bison (*Bison bison*) to the Wichita Mountains Wildlife Preserve in Oklahoma, leading to the recovery of this once-abundant species that was decimated by hunting in the 19th century
1911	First reintroduction of captive-born Alpine ibexes (*Capra ibex ibex*) to the Swiss Alps, leading to the recovery of the population. Reintroductions were also conducted in Austria, Germany, France, Slovenia, and Italy
1961	Start of reintroduction program for the southern white rhinoceros (*Ceratotherium simum simium*) using stock from the remnant population in Hluhluwe–Imfolozi Park which had recovered following protection from poachers. The southern subspecies is now abundant in parks throughout South Africa and Swaziland despite increased poaching over the last decade, whereas the northern subspecies (*C. s. cottoni*) is extinct in the wild
1964	Translocation of 36 South Island saddlebacks (*Philesturnus carunculatus*) from the single surviving population on Big South Cape Island, off Stewart Island, New Zealand, after it was invaded by black rats (*Rattus rattus*). The birds were introduced to Big and Kaimohu Islands, resulting in the species being saved
1974	Start of peregrine falcon (*Falco peregrinus*) reintroductions and reinforcements to restore North American populations reduced by pesticides such as DDT. The species was removed from US list of endangered species in 1999 following the release of >6000 captive-bred falcons
1977	The black robin (*Petroica traversi*) endemic to New Zealand's Chatham Islands is saved from extinction by reintroduction. The last seven birds from Little Mangere Island, where the habitat had been degraded by seabirds, were translocated to Mangere Island following cat eradication and reafforestation
1982	Reintroduction of the Arabian oryx (*Oryx leucoryx*) to Oman using captive-bred animals from the San Diego Wild Animal Park. Although the population increased until about 1996, it subsequently declined to near extinction due to illegal hunting. However, reintroductions have also taken place in Saudi Arabia, Jordan, and the United Arab Emirates
1983	Initiation of the reintroduction program for the large blue (*Maculinea arion*; **Fig. 3**), a Lycaenid butterfly, in the UK. Reintroducing the species required understanding its complex habitat requirements, particularly its dependency on the ant *Myrmica sabuleti*
1984	Reintroduction of the greater Indian rhinoceros (*Rhinoceros unicornis*) to Dudhwa National Park, India, following decline of the species due to habitat loss and hunting
1984	Reintroduction of lake sturgeon (*Acipenser fulvescens*) to the Mississippi and Missouri Rivers begins
1987	Reintroduction of the red wolf (*Canis rufus*) to the Alligator River National Wildlife Refuge in North Carolina, the first reintroduction of a carnivore species that had been declared extinct in the wild
1989	First reintroduction in the Mallorcan midwife toad (*Alytes muletensis*) recovery program, which resulted in the species being downgraded from "Critically Endangered" to "Vulnerable"
1990	Start of translocation program for the Hawaiian monk seal (*Monachus schauinslandi*). This may be the first translocation program for a marine mammal, and has involved translocating pups to islands where they have higher survival due to reduced aggression from males and shark predation
1992	Start of reintroduction program for the takhi (Przewalski's horse, *Equus ferus przewalskii*) in the Gobi Desert in Mongolia
1994	Reintroduction of Atlantic salmon (*Salmo salar*) to tributaries of the Rhine River begins. Salmon were formerly abundant in the Rhine and an important food source to people, but were extirpated by the 1950s due to overfishing and pollution
1994	First reintroductions of native Australian mammals to a fenced sanctuary, Karakamia, where exotic mammalian predators had been eradicated, leading to a network of fenced sanctuaries where >60 reintroductions of at least 23 species have occurred. Native burrowing mammals such as the eastern bettong (*Bettongia gaimardi*; **Fig. 4**) act as "ecosystem engineers," hence their reintroduction has helped restore ecosystem processes. However, some species have probably become over-abundant due to lack of of regulation formerly performed by extinct native predators
1995	Reintroduction of the gray wolf (*Canis lupus*) to Yellowstone National Park, Montana, Wyoming and Idaho, resulting in substantial ecosystem changes through the reintroduction of this top predator
1996	Extensive reintroduction of the Mauna Kea silversword (ahinahina, *Argyroxiphium sandwicense* subsp. *sandwicense*) (>3000 individuals planted) following the control of the feral sheep and mouflon which had decimated this once common plant
2000	First reintroduction of a New Zealand bird, the Little spotted kiwi (*Apterx owenii*), to a fenced sanctuary, Zealandia, where exotic mammalian predators had been eradicated. This has led to a network of fenced sanctuaries where many native species have been reintroduced, including species that that were previously extinct on the mainland
2001	Translocation of two UK butterfly species (marbled white *Melanargia galathea* and small skipper *Thymelicus sylvestris*) into sites north of their range margins as a test case of assisted colonization in response to climate change
2004	First introductions of exotic Aldrabra giant tortoises (*Aldrabrachelys gigantea*; **Fig. 5**) to Mauritian offshore islands to restore the grazing and seed dispersal performed by the extinct Mauritian *Cylindraspis* species
2006	The Bolson tortoise (*Gopherus flavomarginatus*) reintroduced to the wild in New Mexico using captive-bred animals descended from the remnant wild population in Mexico. This is apparently the first wild population north of the Rio Grande for $>10{,}000$ years

(*Continued*)

Table 1 Continued

Date	*Translocation*
2008	Seedlings of the Florida torreya (*Torreya taxifolia*) planted >500 km north of its current range (**Fig. 6**). Although the translocation was originally conceived as a reintroduction to part of the species' prehistoric range, it was widely interpreted as an assisted colonization in response to climate change, leading to intensive debate on that topic
2011	Reintroduction of Atlantic salmon to tributaries of Lake Ontario begins more than 100 years since the species was extirpated from the lake
2012	Introduction of Tasmanian devils (*Sarcophilus harrisii*) to Maria Island, probably the first assisted colonization motivated by a disease threat (devil facial tumor disease)
2016	First translocation of juvenile western swamp tortoises (*Psuedemydura umbrina*) to sites south of the species' indigenous range, probably the first assisted colonization of a vertebrate motivated by climate change

Fig. 2 The flightless kākāpō (*Strigops habroptilus*) was saved from extinction by translocating the remaining birds to offshore islands where the species had probably never occurred, an example of assisted colonization being used in response to impacts of invasive predators. Photo credit: Neil Fitzgerald.

have been poorly conceived. In addition, most translocations at that time were poorly monitored, so little was learned from them. Consequently, in recent decades guidelines and regulations have been developed to improve the conception, planning, and monitoring of translocations. In addition, the science of reintroduction biology has been developed to improve translocation outcomes (Seddon *et al.*, 2007). These developments are covered in the sections below.

Scope of Conservation Translocations

Conservation translocations have traditionally been associated with reintroductions of birds and mammals. The books of case studies published by the IUCN Reintroduction Specialist Group since 2008 (available via the website below) give an idea of the taxonomic breakdown of recent translocations. Of the 303 animal species featured in the first four books, mammals, birds, and reptiles were overrepresented compared to their abundance in nature, while amphibians, fish, and especially invertebrates were vastly underrepresented (Seddon *et al.* 2014; **Fig. 1**). Nevertheless, there have been notable invertebrate translocations, such as the reintroduction of the large blue (*Maculinea arion*) to the United Kingdom (**Table 1**, **Fig. 3**). Recent conservation translocations have become more varied, the number of taxa have become more numerous, and their intended purpose has diversified.

Translocations of 279 animal species in North America from 1974–2013 mostly consisted of reintroductions and reinforcements, with these occurring at similar frequencies (Brichieri-Colombi and Moehrenschlager, 2016). However, assisted colonizations were also documented for species such as the fringed darter (*Etheostoma crossopterum*), white sands pupfish (*Cyprinodon tularosa*), wood frog (*Rana sylvatica*), ornate box turtle (*Terrapene ornate ornata*), gopher tortoise (*Gopherus polyphemus*) and wild turkey (*Meleagris gaallopavo*). Ecological replacements also occurred, as the yellow crowned night heron (*Nyctanassa violacea*) was released to replace the extinct Bermuda night heron (*Nyctanassa carcinocatactes*), and releases of greater prairie-chicken (*Tympanuchus cupido*) were being contemplated to replace the extinct heath hen (*Tympanuchus cupido cupido*).

Fig. 3 The large blue (*Maculinea arion*) was reintroduced to the United Kingdom after detailed research on its habitat requirements, which includes a parasitic relationship with the ant *Myrmica sabuleti* whose persistence is affected by grazing practices. Photo credit: David Simcox Habitat Designs Ltd.

While assisted colonization, such as the introduction of kākāpō to offshore islands (**Table 1**), has been used for over a century to protect species from inescapable threats such as invasive predators, it will be increasingly motivated by climate change. In the United States, the Torreya Guardians translocated seedlings of the Florida torreya (*Torreya taxifolia*) to areas beyond its current range to save the species from climate-induced threats. This translocation was originally conceived as a reintroduction to part of the species' prehistoric range. However, it was subsequently interpreted as an assisted colonization in response to climate change, starting intensive debate on that topic (McLachlan *et al.* 2007). Although strongly opposed by some people, assisted colonization appears to be an essential strategy for managing some long-lived tree species in the face of climate change, and is generally well accepted by foresters (Williams and Dumroese 2013). It may be equally essential for long-lived slowly-dispersing animals. The 2016 translocations of the western swamp tortoise (*Pseudemydura umbrina*) south of its indigenous range in Western Australia constituted the first vertebrate assisted colonizations that were clearly motivated by climate change.

Increases in the frequency of conservation translocations are particularly evident for plants. In China alone 154 species, of which 78% were nationally listed as threatened and included 87 endemics, have been translocated for conservation purposes. Surprisingly 26% of the 154 species, have been moved beyond their indigenous range, primarily as assisted colonizations to escape human-imposed threats such as large-scale habitat destruction.

While reintroductions specifically, and conservation translocations generally, are primarily associated with terrestrial ecosystems, recent evidence also shows that an increasing diversity of species are being translocated in marine environments for conservation purposes (Swan *et al.* 2016). Some iconic species translocated include the Hawaiian monk seal (*Monachus schauinslandi*) (**Table 1**) and the Antillean manatee (*Trichechus manatus manatus*) in Brazil, while the Atlantic walrus (*Odobenus rosmarus rosmarus*) was recently proposed for reintroductions in eastern Canada. Over the last 39 years, at least 242 marine species were translocated for reintroductions, reinforcements, assisted colonizations, or ecological replacements (Swan *et al.* 2016). Reinforcements were almost three times as frequent as reintroductions. Species consisted primarily of coastal invertebrates (44%) such as corals and sea fans as well as plants (30%) such as mangroves and sea grasses. Unlike most reintroductions of terrestrial fauna, 60% of conservation translocations in marine environments were aimed at restoring ecological functions within target ecosystems.

IUCN Reintroduction Specialist Group

Reintroductions are conceptually appealing for a wide variety of reasons, which can span potential benefits for individual species or for broader ecosystem function. Intertwined with such benefits are related motivations or outcomes that can span aesthetic, sociological, cultural, political, or economic aspects. These in turn may differ according to species, scale, and economic context. For example, considerations would differ for releases of corals on a small boulder in the ocean compared to those associated with reintroducing tigers into all Asian countries that might have habitat for them. Bringing species back to places from where they have disappeared is inherently appealing to many people, but good intentions alone do not mean that reintroductions will be effective. Perceived benefits for certain considerations, such as reducing the likelihood of global species extinction, could also entail risks on other levels if, for example, the release of an organism might harm other species in the receiving ecosystem, threaten the

Fig. 4 Jason Cummings releasing an eastern bettong (*Bettongia gaimardi*) at Mulligans Flat Woodland Sanctuary in Canberra, Australia. Bettongs are considered to be "ecosystem engineers," hence their reintroduction is critical to restoring the ecosystem processes. Photo credit: Woodlands and Wetlands Trust and Stephen Corey.

Fig. 5 Mauritian Wildlife Foundation staff filling water troughs for Albabra giant tortoises (*Aldrabrachelys gigantean*) on Round Island, Mauritius. The species was introduced to the island as an ecological replacement for the extinct tortoise that formerly inhabited it. Photo credit: Katie Macfarlane.

livelihoods of local people, or compromise industrial interests. Overeagerness may result in poorly planned reintroductions, but extreme risk aversion may result in useful translocations not being conducted. Scientifically based, evidence-driven guidance can help to curtail rash decisions on either end of the spectrum, while improving the efficacy of actions that are taken.

By the 1980s the frequency, taxonomic diversity, and geographic variation of reintroductions was beginning to grow due to surging interest spurned by actions for several high profile species (**Table 1**). At the same time, challenges and failures were starting to become more apparent, especially through assessments of bird and mammal translocations. In particular, captive-bred individuals and sensitive or threatened species had relatively low rates of success at establishing populations. A primary shortcoming of most programs was that extensive effort and resources were expended for release stages, but follow-up monitoring was generally limited. This lack of information compromised the ability of managers to respond to issues within reintroduction programs over time, and also hampered improvements to reintroduction practice in general.

To prevent potentially irresponsible activities including releases of invasive species, while harnessing potentially positive actions including reintroductions, the IUCN released "The IUCN Position Statement on Translocation of Living Organisms" in 1987. In the following year the IUCN Reintroduction Specialist Group (RSG) was founded to inform and encourage good practice in reintroductions, and to track such initiatives as well as possible. The practitioners that came to constitute the RSG were distributed globally, had diverse expertise, and spanned many sectors including academia, government agencies, and nongovernment organizations. The

Fig. 6 Connie Barlow planting a seedling of the Florida torreya (*Torreya taxifolia*) in North Carolina. This translocation was conceived as a reintroduction to part of the species' prehistoric range, but has been widely interpreted as assisted colonization in response to climate change. Photo credit: Torreya Guardians.

IUCN Guidelines for Reintroductions were drafted by 1995, and officially printed by the RSG in 1998. These were the first official guidelines that IUCN released on any topic, and have been among the most utilized conservation guidance documents worldwide. In addition to publications in the primary literature, the RSG has published five books of case studies.

Guidelines for Conservation Translocations

Over the decade after the original guidelines were released, increases in reintroductions worldwide, improvements in scientific knowledge and management tools, escalation of climate change as a biodiversity threat, and increasing occurrences of conservation introductions triggered a desire for them to be revised. Following 4 years of refinements, new guidelines were released in 2013 (IUCN/SSC 2013). These are now available in eight languages (English, French, Spanish, Portuguese, Korean, Russian, Arabic, Chinese), and further translations are currently under development. The 2013 Guidelines clearly distinguish conservation translocations from other translocations based on the primary motivations underlying such activities. The motivations for conservation introductions are also covered in greater depth, including the distinction between assisted colonization and ecological replacement.

Potential benefits for the translocated species and the receiving ecosystem are positioned against seven primary risks: risk to source population, ecological risk, disease risk, associated invasion risk, risk of gene escape, socio-economic risk, and financial risk. Compared to previous guidelines, there is particular emphasis on animal welfare and the reduction of adverse impacts on the cultural, sociological, or livelihood impacts of affected human communities. Community buy-in is also positioned as increasingly important since even iconic initiatives such as the release of Arabian oryx (*Oryx leucoryx*) in Oman became compromised by disenfranchised community groups. More than before, the 2013 guidelines present conservation translocations as a process of iterative decision making where alternatives to translocations are considered, feasibility and risk of action or inaction is continuously evaluated, and monitoring data inform subsequent management alternatives. Among such alternatives are decisions of what individuals and release areas could be selected and how their suitability can be improved through various management tools.

Sound planning within the guidelines is underpinned by tiered goals, objectives, and actions that are directly linked to the design of monitoring programs, and potential exit strategies when desirable results have been achieved or undesirable outcomes become overwhelming. Biological and social feasibility are assessed while considering the regulatory environment and availability of resources that would be required to conduct conservation translocations responsibly. The guidelines outline a process that is likely to improve translocation programs worldwide, and is potentially applicable to any species. The RSG has also worked with other IUCN specialist groups over time to further develop taxon-specific guidelines for species such as primates, bears, elephants, and galliforms.

The 2013 guidelines have been increasingly integrated into jurisdictional policy. Canadian federal and provincial governments are required to consider the IUCN guidelines for any conservation translocations in the country, Scotland developed a national code for conservation translocations, and the European Commission adopted the IUCN guidelines as their standard for managing conservation translocations of species within the context of changing climatic conditions. Governments worldwide develop regulations of relevance to conservation translocations on national or regional levels, and increasing opportunities should be pursued to formally integrate the guidelines into these regulatory policies. Such approval processes should also be based upon extensive consultation with affected stakeholders to improve the acceptance and efficacy of conservation translocations.

Recent developments in genomics and genetic engineering have raised the possibility of resurrecting species that have previously gone extinct. Indeed the birth in 2008 of a Pyrannean Ibex (*Capra pyrenaica pyrenaica*), a subspecies that had gone extinct at the beginning of the millennium, made these theoretical possibilities a reality. A rapid evolution of technological developments, and surging public or scientific interest, prompted the desire for IUCN to develop guidelines that would deal specifically with the potential "De-Extinction" of species. From a conservation perspective, many of the fundamental considerations for conservation translocations are also applicable to such scenarios, as proxies of previously extinct species may be created for release within or beyond the indigenous ranges they once occupied. A special task force was formed and in 2016 the IUCN published guidelines on creating proxies of extinct species for conservation benefit (IUCN/SSC 2016).

Reintroduction Biology

The term "reintroduction biology" refers to research designed to improve practice related to conservation translocations. This term appears to have been first used at the conference "Reintroduction biology of Australasian fauna" in April 1993, and first appeared in the scientific literature later that year in a paper from a presentation at the conference. Very few articles had been published on conservation translocations up to that time, but subsequent exponential growth in the publication rate has resulted in there now being >150 articles published each year. There have also been several recent books published on reintroduction biology (Hayward and Somers 2009; Ewen *et al.* 2012; Maschinski and Haskins 2012; Armstrong *et al.* 2015; Jachowski *et al.* 2016).

Like any applied science, reintroduction biology aims to provide information that will allow practitioners to make good decisions. Part of this simply involves monitoring translocations and making the information available so people can learn from previous experience. The early reintroduction literature was dominated by such descriptive accounts, and they continue to play an important role (see volumes of case studies available via the RSG website). However, learning from descriptive studies is a slow process, and often leads to unreliable inference. Consequently, over the last 15 years there have been efforts to advance the science of reintroduction biology by encouraging the development of a clear theoretical basis, and strengthening inference using experiments, quantitative modeling, and meta-analysis. To facilitate decision making, researchers need to be able to address questions such as:

- What is the probability that a species will persist at a proposed site?
- How is this probability affected by a proposed postrelease management regime?
- What is the expected postrelease survival under a particular release strategy?
- What is the expected population growth rate, and carrying capacity?
- How will this population trajectory affect retention of genetic diversity?
- What effects is the species expected to have on ecosystem processes at the site?
- What is the risk of a particular pathogen being introduced, and what impacts might occur?

Armstrong and Seddon (2008) suggested that such issues could be categorized into 10 key questions at three different levels of biological organization—the population, metapopulation, and ecosystem (**Fig. 7**).

The reintroduction biology literature has indeed become more question-focused over the last decade, and is using more sophisticated methods. However, the links to real management decisions is often unclear. Consequently, the RSG is now promoting increased use of structured decision making and adaptive management to ensure that improved information results in improved decisions. "Structured decision making" means choosing the optimal management option based on clearly defined objectives and predictions, and "adaptive management" means updating such decisions based on learning.

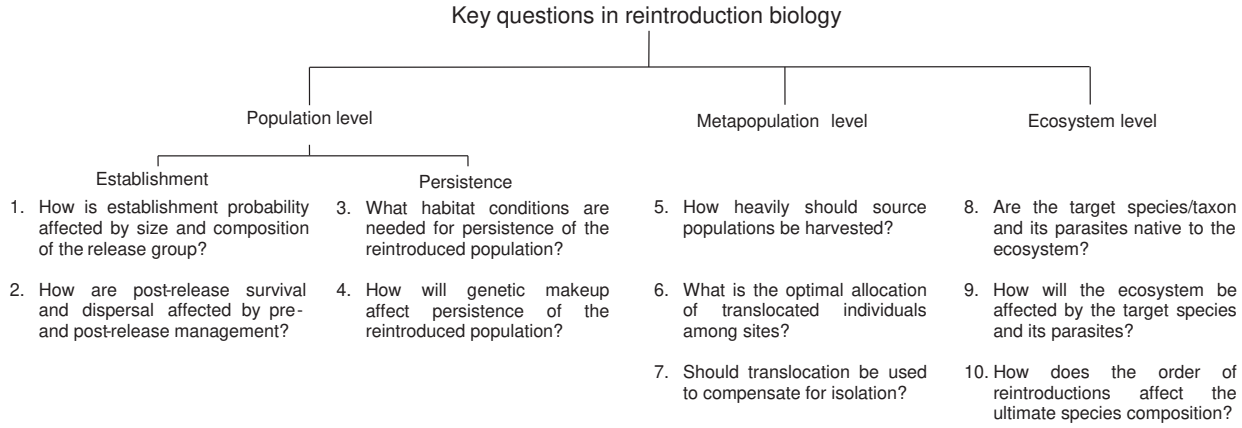


Fig. 7 Ten key questions in reintroduction biology, divided into questions at the population, metapopulation, and ecosystem level. From Armstrong, D. P. and Seddon, P. J. (2008). Directions in reintroduction biology. *Trends in Ecology and Evolution* **23**, 20–25.

A common misunderstanding is that reintroduction biology aims to improve the "success rate" of translocations, and therefore that this success rate should increase over time. This idea is misleading for three reasons. First, the concept of "success rate" is problematic because definition of success depends on the specific objectives of a given translocation—for example, is the main objective simply to establish a population at the site, or might it be to maintain a population with high genetic diversity, or to restore a particular ecosystem function? Second, maximizing success rate is an illogical goal because it means avoidance of risky translocations regardless of the potential benefits. Third, we do not necessarily expect success rates to rise, as the challenges involved in translocations do not remain constant over time. As discussed above, the scope of translocation as a conservation tool has broadened over time and will continue to change in the future. Reintroduction biology should aim to develop the best possible tools for making practical management decisions in the face of these challenges. Ultimately the iterative process of sound planning, evidence-based decision making, and integrated management will continue to increase the positive impact conservation translocations make in the restoration of species, reduction of extinctions, and maintenance of ecological processes.

See also: Conservation Ecology: Endangered Species; Invasive Plant Species. General Ecology: Migration and Movement

References

Armstrong, D.P., Seddon, P.J., 2008. Directions in reintroduction biology. Trends in Ecology and Evolution 23, 20–25.
Armstrong, D.P., Hayward, M.W., Moro, D., Seddon, P.J. (Eds.), 2015. Advances in reintroduction biology of Australian and New Zealand Fauna. Melbourne, Australia: CSIRO Press.
Brichieri-Colombi, T.A., Moehrenschlager, A., 2016. Alignment of threat, effort, and perceived success in North American conservation translocations. Conservation Biology 30, 1159–1172.
Donlan, J., Greene, H.W., Berger, J., *et al.*, 2005. *Re*-wilding North America. Nature 436, 913–914.
Ewen, J.G., Armstrong, D.P., Parker, K.A., Seddon, P.J. (Eds.), 2012. Reintroduction biology: Integrating science and management. Oxord, UK: Wiley-Blackwell.
Germano, J.M., Bishop, P.J., 2009. Suitability of amphibians and reptiles for translocation. Conservation Biology 23, 7–15.
Hayward, M.W., Somers, M. (Eds.), 2009. Reintroduction of top-order predators. Oxford, UK: Wiley-Blackwell.
IUCN/SSC, , 2013. Guidelines for reintroductions and other conservation translocations. Version 1.0. Gland, Switzerland: IUCN Species Survival Commission.
IUCN/SSC, , 2016. Guiding principles on creating proxies of extinct species for conservation benefit. Version 1.0. Gland, Switzerland: IUCN Species Survival Commission.
Jachowski, D., Millspaugh, J., Angermeier, P., Slotow, R. (Eds.), 2016. Reintroduction of fish and wildlife populations. Oakland, California: University of California Press.
Maschinski, J., Haskins, K.E. (Eds.), 2012. Plant reintroduction in a changing climate: Promises and perils. Washington, DC: Island Press.
McLachlan, J.S., Hellmann, J.J., Schwartz, M.W., 2007. A framework for debate of assisted migration in an era of climate change. Conservation Biology 21, 297–302.
Seddon, P.J., Armstrong, D.P., Maloney, R.F., 2007. Developing the science of reintroduction biology. Conservation Biology 21, 303–312.
Seddon, P.J., Griffiths, C.J., Soorae, P.S., Armstrong, D.P., 2014. Reversing defaunation: Restoring species in a changing world. Science 345, 406–412.
Swan, K.D., McPherson, J.M., Seddon, P.J., Moehrenschlager, A., 2016. Managing marine biodiversity: The rising diversity and prevalence of marine conservation translocations. Conservation Letters 9, 239–251.
Williams, M.I., Dumroese, R.K., 2013. Preparing for climate change: Forestry and assisted migration. Journal of Forestry 111, 287–297.

Relevant Websites

http://www.iucnsscrsg.org—IUCN/SSC Reintroduction Specialist Group.
http://www.massey.ac.nz/~darmstro/rsg.htm—Reintroduction Specialist Group, Oceania Section.
http://reviverestore.org/—Revive & Restore: Genetic Rescue for Endangered and Extinct Species.
http://www.snh.gov.uk/protecting-scotlands-nature/reintroducing-native-species/scct/—The Scottish Code for Conservation Translocations.
http://www.torreyaguardians.org/—Torreya Guardians.

Source–Sink Landscape

Wenwu Zhao, Lizhi Jia, and Stefani Daryanto, Beijing Normal University, Beijing, China
Liding Chen, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing, China
Yue Liu, Beijing Normal University, Beijing, China

Introduction

"Source" often refers to the origin of an object or matter, while "sink" often refers to the deposition of the same object or matter. These definitions originated from the research of atmospheric pollution in the context of global climate change in which "source" is the unit system where materials or bionts emigrate, while "sink" is the unit system where materials or emigrant bionts are deposited during the process of earth surface layer material (**Fig. 1**). In the study, "source" and "sink" were also used to analyze the cause and effect of atmospheric pollutants and the concept was then adopted by the United Nations Framework Convention on Climate Change which defines "source" as the process or activity which can emit greenhouse gases and aerosols, while "sink" as the process or activity in which the greenhouse gases and aerosols are absorbed (e.g., into the soil or organic materials during the carbon or nitrogen fixation) (United Nations, 1992). According to the above definitions, emissions from factories or residential areas and traffic are all considered as "source", while "sink" are the ecosystems such as forests that can absorb atmospheric pollutants.

With the application of the "source" and "sink" concept in different field of studies, the terminology of "source" and "sink" has also evolved accordingly and it often includes different meanings. In ecology, the concept of "source" and "sink" was originally applied to study the dynamic of heterogeneous population and to reflect the process of migration, diffusion and destruction of species (Boswell *et al.*, 1998; Finkenstadt and Grenfell, 1998; Fox and Fox, 2000; Foley *et al.*, 2005). In habitats with rich resources, higher birth rate than death rate is usually found, increasing the number of population. This population may then colonize other habitats and it is called the "source" population. In habitats with poor resources, on the other hand, population loss often outnumbers the new offspring and consequently, this "sink" population requires immigration of individuals from other habitats to maintain the number of population (**Fig. 2**). This definition is similar when plaques occur in which the "source" population is correspondingly referred to the plaque "source" while the population in which the plaque spreads is referred to the "sink" population (Dunning *et al.*, 1992; Green, 1994; Kindvall, 1995; Hill *et al.*, 1996; Holyoak and Lawler, 1996; Leibold *et al.*, 2004; Guo, 2014). For example, a small songbird blue tit (*Parus caeruleus*) population reproduces under two kinds of forest habitats: one

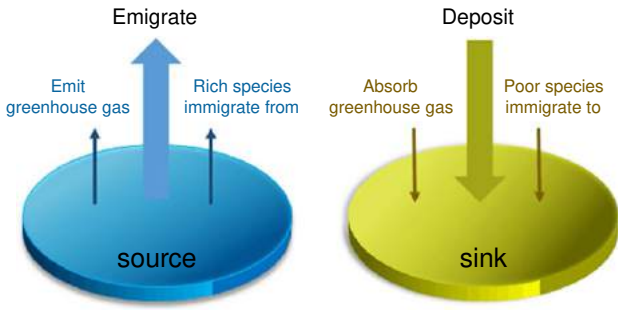


Fig. 1 Different applications of "source" and "sink" process.

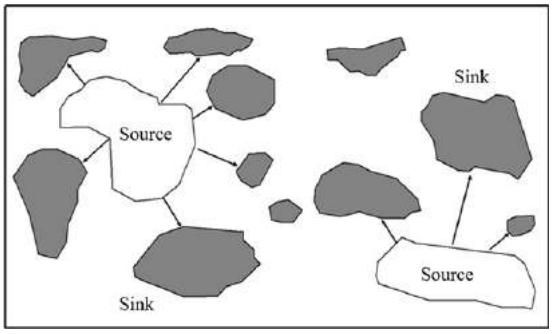


Fig. 2 Sources and sinks (Dunning *et al.*, 1992).

habitat is dominated by deciduous oak (*Quercus pubescens*) and the other is dominated by evergreen oak (*Quercus ilex*). Under normal condition and in the absence of migration, the annual blue tit population in deciduous oak dominated forest somehow increases while the population in the evergreen oak decreases. When the blue tit population perched in the deciduous oak dominated forest migrates to the evergreen oak forest, this population is considered as the "source" population, while that in the evergreen oak dominated forest is considered as the "sink" population.

The classification of a sink or a source, however, is not constant, but may change over time due to changes of climate and land management. The occurrence of climate change, for example, may reverse the above example on the blue tit population. If the blue tit population perched in deciduous oak dominated forest decreases with climate or vegetation change, the "source" population may be transformed into "sink" population following the migration of the blue tit population in the evergreen to deciduous forest. Similarly, the arctic tundra is likely to shift from carbon (C) sink to source as the temperature rises (Oechel *et al.*, 1993), which increases the oxidation of soil organic matter. In contrast, reforestation of degraded areas and the application of soil conservational practices will gradually change the cultivated soil in China and elsewhere from C source to C sink (Lal, 2004a, 2004b; Yu *et al.*, 2009).

Definition of the Source and Sink Landscape

Until the end of the last century, studies of landscape pattern and ecological process are stagnant due to complex ecosystem processes (Li and Reynolds, 1995). Chen *et al.* (2006) then introduced the concept of "source" and "sink" into the landscape pattern analysis and proposed the source–sink landscape theory which can be helpful to understand the coupling relationship between landscape pattern and ecological process. In the source–sink landscape theory, landscape types are classified as either "source" or "sink" landscape based on their function in an ecological process. "Source" landscape is a landscape type which contributes positively to the development of an ecological process, while a "sink" landscape is the one which is reversely contributes to the development of the same ecological process. From this point forward, "source" and "sink" always refer to the landscape-scale ecological processes (i.e., "source landscape" and "sink landscape"; **Fig. 3**).

The source–sink landscape theory which was originated from the source–sink theory of atmospheric pollution is now mainly used to study the effect of landscape patterns on ecological process. The theory has been used to the study of a wide range of fields, from biodiversity protection to urban heat island effect. For example, sloping croplands in a mountainous area, farmlands with high fertilizer application, or urban settlements play the role of "source" for nonpoint source pollution. In contrast, other landscape types such as grasslands, forests and wetlands located in the downstream direction of the "source" play the role of "sink". For soil and nutrient loss, "source" is the place where soil and nutrient are originated. Unless there are forests, grasslands or wetlands in the downstream position of the "source", this soil and nutrient will flow directly into the earth's surface or underground water, causing nonpoint source pollution. Similarly, some landscapes releasing CO_2 are considered as "sources" of greenhouse gas emission, while other landscapes absorbing CO_2 are considered as "sinks". A "source" in terms of biodiversity protection is an area which can provide habitat or resources for certain species, while "sink" is the area which has negative impact to the survival of the species.

The Characteristics of "Source" and "Sink" Landscapes

The Relative Categorization of "Source" and "Sink" Landscape

The categorization of a landscape as a "source" or "sink" is not absolute, but mainly depends on the role of that landscape type in an ecological process. Due to changes in numerous environmental variables over time, however, some sinks may act as sources and vice versa during the course of the same ecological processes. Although a landscape type can function both as a source and a sink, the landscape can be seen as a "source" when the source function is greater than the sink function. However, when the source function is less significant than the sink function, the landscape type can be viewed as a "sink" (**Fig. 3**). Due to the fact that source–sink landscape theory is based on specific ecological processes, it is important to define the ecological process studied when determining a landscape as a "source" or a "sink". During the process of soil loss, a grassland can intercept the sediment from the

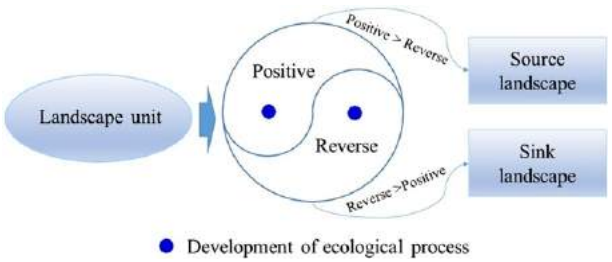


Fig. 3 The concept of source–sink landscape.

upslope position, and therefore functions as a sink. However, the same grassland can also produce sediments and therefore functions as a source. The balance between sediment yield and sediment interception thus determines the role of the grassland, whether it acts as a "source" or a "sink". A grassland is considered as a "source" when the amount of sediment yield is larger than the interception. In contrast, if the amount of sediment interception is larger than the yield, the grassland will then be considered as a "sink".

Linking the Categorization of a "Source" and "Sink" Landscape with the Process Studied

The fundamental difference between a "source" and a "sink" is that a "source" plays a positive role in promoting the ecological process studied, while a "sink" has a reversely influence on the process (Chen *et al.*, 2006). The classification of a landscape as a "source" or a "sink", however, may be different for different ecological processes. A landscape may be a "source" for a certain ecological process and is likely to be a "sink" for another ecological process. A farmland ecosystem is a "source" landscape for nonpoint source pollution due to the use of fertilizer and pesticide; however, it may become a "sink" for global C cycle because of its ability to fix CO_2. Therefore, one must always refer to an ecological process to determine the role of a landscape as a "source" or a "sink" in the process studied.

Different Magnitude of a "Source" or "Sink" Landscape in an Ecological Process

In an ecological process, different landscapes may have a different magnitude as a "source" or a "sink" by having different contributions to the corresponding ecological process. Despite farmlands, vegetable fields and orchards are all "sources" for nonpoint source pollution, one may contribute to a greater extent as a nonpoint source emitter compared to the others. Similarly, when forestland and grassland are both "sinks" for nonpoint source pollution, their contributions on nutrient absorption are different.

"Source" and "Sink" Landscape Evaluation Model

The source–sink landscape theory is supposed to regulate the ecological process from the point view of landscape pattern. The basic premises of source–sink landscape theory are to: (i) analyze the function of a landscape type (as a "source" or a "sink") in an ecological process, (ii) identify different "sources" or "sinks" in the landscape, and (iii) estimate the contribution of different landscape types in the ecological process. Because "source" and "sink" are defined within certain ecological process, the nature of the "source" and "sink" in a landscape will change when the ecological process changes. Chen *et al.* (2003b) proposed the location-weighted landscape contrast index. According to the theory and methods of Lorenz curve, this index includes three landscape properties: slope, distance and relative altitude to target points (e.g., monitoring stations and watershed outlet) (**Fig. 4**). The location-weighted landscape contrast index can be expressed as:

$$LCI = SODBC/SOFBC$$

where *LCI* is the location-weighted landscape contrast index relative to the watershed outlet monitoring point location (slope, distance and relative altitude to target points); *SODBC*, *SOFBC* are the irregular trilateral area composed of the cumulative curves of source and sink landscape area, respectively. Compared with *OFB* curve, the type of landscape displayed by *ODB* curve is located closer to the watershed outlet monitoring point.

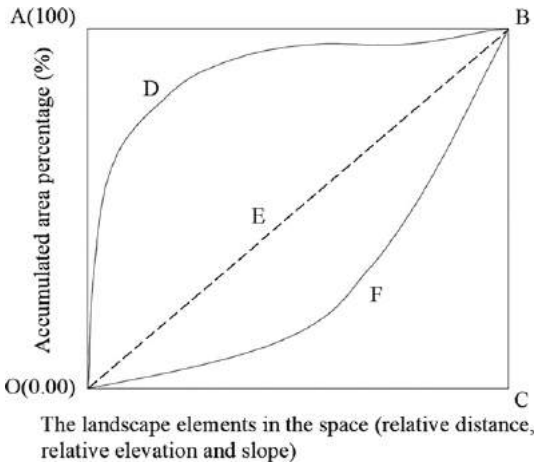


Fig. 4 Scheme figure of the location-weighted landscape contrast index.

The value of the index can well reflect the influence of landscape spatial pattern on ecological process (Cook, 2002). This model connects landscape pattern and ecological process and gives full consideration of the relationship between them. Furthermore, the model allows quantitative research and forecasts the relationship between landscape pattern and ecological process. For an ecological process in a landscape, the "source" and "sink" landscape evaluation model can be expressed as following:

$$LCI = \log\left\{\sum_{i=1}^{M} Si_{ODBC} \times Wi \times Pci / \sum_{j=1}^{N} Sj_{OFBC} \times Wj \times Pcj\right\}$$

where Si_{ODBC} and Sj_{OFBC} are the irregular triangular area composed of cumulative curve of the ith "source" landscape and the jth "sink" landscape in the Lorentz curve, respectively; M and N are the number of "source" and "sink" landscape, respectively; Pci and Pcj are the percentage of ith "source" landscape and the jth "sink" landscape in the catchment, respectively. Logarithm of the calculation results is mainly to control the range of LCI changes which will change around 0. When the value of LCI is 0, it means that the "source" landscape and "sink" landscapes are evenly distributed on the basin scale, and the contribution of this pattern to nonpoint source pollution is balanced on a watershed scale. When the LCI value is greater than 0, it indicates that the "source" landscape in the river basin contributes more to the monitoring point than the "sink" landscape, and the river basin may have more output of nonpoint source pollutants. When the value of LCI is less than 0, it indicates that "sink" landscape contributes more to the export monitoring point of the drainage basin than "source" landscape, and the pollutant output from this basin should be relatively small. In theory, the larger the value of LCI, the more nonpoint source pollutants should be output from the watershed, and vice versa.

O (0.00) is the watershed outlet, horizontal axis (OA) is the cumulative percentage of landscape type (range: 0 ~ 100); vertical axis (OC) is the relative distance between landscape type and watershed outlet (range: 0 ~ the maximum distance from the monitoring station) or slope gradient (range: 0 ~ the maximum slope gradient within the watershed); ODB and OFB is area cumulative curve of different landscape. OEB represents the absolute mean distribution curve, and if the "source" landscape and "sink" landscape is evenly distributed in the catchment, the OEB distribution curve will appear.

Application of Source–Sink Landscape Theory

The proposal of source–sink landscape theory is mainly based on the ecological balance theory in ecology. It explores the ways to regulate an ecological process by analyzing the spatial balance or distribution of "sources" and "sinks" in a landscape. The source–sink landscape theory, for example, can be applied in the following fields.

Landscape Pattern Design and Nonpoint Pollution Control

According to the source–sink landscape theory, some landscape types play the role of "source" and the other landscape types play the role of "sink" during the transport process of materials. The nonpoint source pollution, especially the eutrophication of aquatic ecosystems, is virtually induced by the imbalanced spatiotemporal distribution of nutrients. Identifying and managing the source of the contaminants (or nutrients), as well as minimizing the emission are therefore the most reliable methods to reduce the nonpoint source pollution. Generally speaking, there are two approaches to prevent excess nutrients from entering the aquatic system. One is to generate zero emission from each landscape unit so that there will be no contamination induced. The other one is to design the landscape pattern so that the contaminants are intercepted before entering the surface or subsurface water. Chen *et al.* (2003a,b) showed that there is a close relationship between the spatial distribution of "source" and "sink" in a landscape and the nonpoint source pollution. Thus, in order to reduce the risk of nonpoint source pollution, it is feasible to control nutrient loss from spatiotemporal scale by exploring the spatial combination of different landscapes in a watershed ecoplanning to control nutrient loss in a heterogeneous landscape (Turner and Gardner, 1991; Haines-Young and Chopping, 1996).

The source–sink landscape theory has been used in the research of nonpoint source pollution for many years. Chen *et al.* (2002) analyzed the relationship between the location-weighted landscape contrast index and nonpoint source pollution in the watershed scale by examining four typical small watersheds in the Yuqiao reservoir basin as the research object, and nonpoint source pollution as the target process. The result showed that the effect of different landscape patterns on the surface water quality can be evaluated by a source–sink landscape model at the watershed scale. Similarly, Chen *et al.* (2009) monitored the water quality of two adjacent watersheds for 3 years in the northeast black earth area, China. The results showed that the source–sink landscape index can well reflect the effect of landscape pattern on water quality and can be used to determine the risk of the occurrence of nonpoint source pollution. The source–sink landscape theory can be used to make the source–sink landscape risk identification and provide basis for decision-making in the treatment of nonpoint source pollution in Three Gorges Reservoir Area (Wang *et al.*, 2016a, b). According to the "source–sink" landscape theory, Fang *et al.* (2016) analyzed the concentration and amplitude difference of soil heavy metals (Pb, Cu, and Cd) in forest and construction land of Shanghai Pudong New District by using the sliding window analysis. The results showed that the source–sink landscape theory can be applied to the study of nonpoint source pollution.

The "source–sink" landscape index can reflect the spatial variation of total nitrogen (TN) and provides a reference for the evaluation of nonpoint source pollution in the Heihe river basin (Sun *et al.*, 2012; Zhang *et al.*, 2017). Similarly, the "source–sink"

landscape index can help researchers to clarify the role of infrastructure in reducing nonpoint source N pollution in a Mediterranean climate (Mcconaghie and Cadenasso, 2016).

In the study of the ecological process of soil erosion—a significant form of nonpoint source pollution—sloping croplands which increase soil erosion are considered as "source", while the forest and grassland which reduce soil erosion are considered as "sink". The contribution of different landscape to soil erosion can be characterized by the C factors (crop management factors) of the universal soil loss equation (USLE). By using USLE, researchers establish a multiscale soil erosion evaluation index in order to explore the relationship between landscape pattern and soil erosion. This index can effectively reflect the relationship between land use pattern and soil erosion process, and provide technical supports to optimize the design of regional land use pattern and comprehensive management of regional soil erosion (Fu *et al.*, 2006; Zhao *et al.*, 2012). The effect of watershed landscape pattern on soil erosion can be well-revealed by using the "source–sink" landscape theory (Xu and Zhou, 2008; Yang *et al.*, 2012). According to the "source–sink" theory of soil erosion, slope-HRUs landscape index (SHLI) can reflect the relationship between landscape pattern and soil erosion processes to a certain extent in the Yanhe watershed, China (Zhou and Li, 2015; Li and Zhou, 2015). By using the "source–sink" landscape theory, Chen *et al.* (2016a, b) also showed that there is strong relationship between soil erosion processes and land cover changes in the Loess Plateau in China observed in 1987, 1995, and 2007.

"Source" and "Sink" Landscape Pattern Design and Biodiversity Protection

The key to protecting biodiversity is to protect the habitats of endangered species (Tilman *et al.*, 1994; Lomolino and Perault, 2001; Marsh and Trenham, 2001). Using the source sink landscape theory, one can evaluate the spatial relationships between the dwelled patches and the surrounding patches to determine habitat suitability for species survival. Rich habitat patches and the surrounding resource patches act as the "source" for those species. Yet they may become unsuitable for the survival of the species if they are disturbed. When leaving them uncontrolled, disturbance can turn resource-rich habitat patches into "sinks". If more resource than disturbed patches are available for the target species, this landscape pattern is to be beneficial for the survival of the species. On the contrary, a landscape pattern would generate adverse effects for species survival and protection if it has more sinks than sources. Therefore, the suitability of the landscape pattern for a species can be evaluated by using "source" and "sink" landscape evaluation model, in which the role of different landscape types is identified.

For biodiversity protection evaluation, one should aim at actual "sources" and consider these source patches as points. Then, the landscape evaluation model is used to analyze the role of the surrounding landscape pattern on these "sources" (points), and to clarify how landscape pattern may assist the conservation of the target species. Li *et al.* (2014) classified the function of nature reserve and determined the development direction and the management framework based on the source–sink landscape theory. Liu *et al.* (2016) pointed that the "arthropod island"–shrub relationship, referred to as a source or sink in terms of the source–sink landscape theory during the ecological processes (Chen *et al.*, 2008; Wei *et al.*, 2012), could have important ecological implications on the succession processes of desertified ecosystems (Liu *et al.*, 2013; Zhao and Liu, 2013).

"Source" and "Sink" Landscape Pattern Design and Urban Heat Island Effects

An urban ecosystem is highly influenced by human activities. As the urban area develops, common issues related to urbanization (e.g., urban heat island effect and traffic jam) become increasingly serious, mostly due to inappropriate design of the urban landscape patterns. Using the source sink landscape approach, the occurrence of urban heat island effect and heavy traffic could be considered as the imbalance between "source" and "sink" in an urban area. Urban landscape consists of "gray landscape" (i.e., buildings and roads), "blue landscape" (i.e., rivers, lakes), and "green landscape" (i.e., gardens, turfs, vegetation belts). Different landscape types play distinct roles in the urban heat island effect. Since the over-centralized-distribution of gray landscape results in temperature increase, gray landscape can be considered as the source of the heat island effect. In contrast, the blue and green landscape may reduce temperature and alleviate the urban heat island effect. Since blue landscapes always have lower temperature than the gray landscape and provide a comfortable environment for the residents, adding the number of blue and green landscapes on heat island effect mitigation is always preferable to any city. However, in most cases, the proportion of the blue and green landscapes is limited. Consequently, it is necessary to spatially regulate the gray, blue and green landscapes based on their features as the "source" and "sink" landscape.

According to the source–sink landscape theory, the Lorenz curve and the Gini index were used to describe the distribution of "hot" towards "cool" land cover around a landscape center, and the first law of geography was echoed, which presents an interesting cross-disciplinary topic (Chen *et al.* 2016a, b). Ayanlade (2016) examined the seasonal and diurnal variations in land surface temperature (LST) between different landscapes in Nigeria. He found that differences in LST patterns between the urban areas and its surrounding rural areas might be caused by urban expansion resulting from the removal of vegetation and high heat capacity of construction materials in the central of urban area. Li *et al.* (2017) found that it is essential to design and transform the landscape in specific ways based on the heterogeneous effects of landscape pattern on the thermal environment to mitigate urban heat island.

Conclusions

The relationships between landscape pattern and ecological process are the core issues of landscape ecology. They are also the hot spots of landscape ecology research because it is not always easy to link landscape pattern and ecological process. Yet the source–sink landscape model, that originated from the source–sink theory of atmospheric pollution (Chen *et al.*, 2006), can be applied to the study of nonpoint source pollution control, biodiversity protection and urban heat island effect. Due to the variability of landscape pattern and ecological process, however, the same landscape type can be categorized either as a "source" or a "sink", depending on the ecological process studied. Similarly, different landscape types may have different contribution in the corresponding the ecological process. Determining the ecological process and identifying the "sources" and the "sinks" in the landscape including their extent of contribution are the key strategies to fully apply the source–sink landscape theory into practice.

Acknowledgments

This work was supported by National Key R&D Program of China (No. 2017YFA0604704), National Natural Science Foundation of China (No. 41771197), and State Key Laboratory of Earth Surface Processes and Resource Ecology (No. 2017-FX-01(2)).

See also: Aquatic Ecology: Estuarine Ecohydrology; Acidification in Aquatic Systems. General Ecology: Migration and Movement

References

Ayanlade, A., 2016. Seasonality in the daytime and night-time intensity of land surface temperature in a tropical city area. Science of the Total Environments 557–558, 415–424.
Boswell, G.P., Britton, N.F., Franks, N.R., 1998. Habitat fragmentation, percolation theory and the conservation of a keystone species. Proceedings of the Royal Society of London B 265, 1921–1925.
Chen, A., Zhao, X.F., Yao, L., Chen, L.D., 2016b. Application of a new integrated landscape index to predict potential urban heat islands. Ecological Indicators 69, 828–835.
Chen, L.D., Fu, B.J., Xu, J.Y., Gong, J., 2003a. Location-weighted landscape contrast index: A scale independent approach for landscape pattern evaluation based on "source–sink" ecological processes. Acta Eeologica Sinica 23 (11), 2406–2413.
Chen, L.D., Fu, B.J., Zhang, S.R., *et al.*, 2002. A comparative study on the dynamics of non-point source pollution in a heterogeneous landscape. Acta Ecologica Sinica 22 (6), 808–816.
Chen, L.D., Fu, B.J., Zhang, S.R., Qiu, J., Yang, F.L., 2003b. Seasonal change of solvable nitrogen in surface water of Yuqiao reservoir basin. China Environmental Science 23 (2), 210–214.
Chen, L.D., Fu, B.J., Zhao, W.W., 2006. Source-sink landscape theory and its ecological significance. Acta Ecologica Sinica 26 (5), 1444–1449.
Chen, L.D., Liu, Y., Lv, Y.H., *et al.*, 2008. Pattern analysis in landscape ecology: Progress, challenges and outlook. Acta Ecologica Sinica 28 (11), 5521–5531.
Chen, L.D., Tian, H.Y., Fu, B.J., 2009. Development of a new index for integrating landscape patterns with ecological processes at watershed scale. Chinese Geographical Science 19 (1), 37–45.
Chen, N., Ma, T.Y., Zhang, X.P., 2016a. Responses of soil erosion processes to land cover changes in the loess plateau of China: A case study on the Beiluo river basin. Catena 136, 118–127.
Cook, E.A., 2002. Landscape structure indices for assessing urban ecological networks. Landscape and Urban Planning 58, 269–280.
Dunning, J.B., Danielson, B.J., Pulliam, H.R., 1992. Ecological processes that affect populations in complex landscapes. Oikos 65 (1), 169–175.
Fang, S.B., Cui, Q., Pang, H.H., Yin, C.S., Luo, X.Z., 2016. Source–sink theory based distribution characters of soil heavy metals along an urban-rural gradient in Pudong New District. Chinese Journal of Ecology 35 (3), 772–780.
Finkenstadt, B., Grenfell, B., 1998. Empirical determinants of measles metapopulation dynamics in England and Wales. Proceedings of the Royal Society B-Biological Sciences 265, 211–220.
Foley, J.A., DeFries, R., Asner, G.P., *et al.*, 2005. Global consequences of land use. Science 309, 570–574.
Fox, B.J., Fox, M.D., 2000. Factors determining mammals species richness on habitat islands and isolates: Habitat diversity, disturbance, species interactions and guild assembly rules. Global Ecology and Biogeography 9, 19–38.
Fu, B.J., Zhao, W.W., Chen, L.D., Lü, Y.H., Wang, D., 2006. A multiscale soil loss evaluation index. Chinese Science Bulletin 51 (4), 448–456.
Green, D.G., 1994. Connectivity and complexity in landscapes and ecosystems. Pacific Conservation Biology 1, 194–200.
Guo, Q.F., 2014. Species invasions on islands: Searching for general patterns and principles. Landscape Ecology 29, 1123–1131.
Haines-Young, R., Chopping, M., 1996. Quantifying landscape structure: A review of landscape indices and their application to forested landscapes. Progress in Physical Geography 20 (4), 418–445.
Hill, J.K., Thomas, C.D., Lewis, O.T., 1996. Effects of habitat patch size and isolation on dispersal by Hesperia comma butterflies: Implications for metapopulation structure. Journal of Animal Ecology 65, 725–735.
Holyoak, M., Lawler, S.P., 1996. The role of dispersal in predator-prey metapopulation dynamics. Journal of Animal Ecology 65, 640–652.
Kindvall, O., 1995. The impact of extreme weather on habitat preference and survival in a metapopulation of the bush-cricket Metrioptera bicolor in Sweden. Biological Conservation 73, 51–58.
Lal, R., 2004a. Soil carbon sequestration impacts on global climate change and food security. Science 304 (5677), 1623–1627.
Lal, R., 2004b. Soil carbon sequestration to mitigate climate change. Geoderma 123, 1–**2)**, 1–22.
Leibold, M.A., Holyoak, M., Mouquet, N., *et al.*, 2004. The metacommunity concept: A framework for multi-scale community ecology. Ecology Letters 7, 601–613.
Li, D.J., Pang, Y., Qian, Z.D., Chen, H.P., 2014. Research of nature reserve zoning based on landscape ecology source–sink theory. Resources and Environment in the Yangtze Basin. 53–59.
Li, H., Reynolds, J.F., 1995. A new contagion index to quantify spatial patterns of landscapes. Landscape Ecology 8, 155–162.
Li, J., Zhou, Z.X., 2015. Coupled analysis on landscape pattern and hydrological processes in Yanhe watershed of China. Science of the Total Environment 505, 927–938.
Li, W.F., Cao, Q.W., Lang, K., Wu, J.S., 2017. Linking potential heat source and sink to urban heat island: Heterogeneous effects of landscape pattern on land surface temperature. Science of the Total Environment 586, 457–465.

Liu, R.T., Pen-Mouratov, S., Steinberger, Y., 2016. Shrub cover expressed as an 'arthropod island' in xeric environments. Arthropod-Plant Interactions 10, 393–402.
Liu, R.T., Zhu, F., Song, N.P., Yang, X.G., Chai, Y.Q., 2013. Seasonal distribution and diversity of ground arthropods in microhabitats following a shrub plantation age sequence in desertified steppe. PLoS One 8.e77962.
Lomolino, M.V., Perault, D.R., 2001. Island biogeography and landscape ecology of mammals inhabiting fragmented, temperate rain forests. Global Ecology and Biogeography 10, 113–132.
Marsh, D.M., Trenham, P.C., 2001. Metapopulation dynamics and amphibian conservation. Conservation Biology 15, 40–49.
Mcconaghie, J.B., Cadenasso, M.L., 2016. Linking nitrogen export to landscape heterogeneity: The role of infrastructure and storm flows in a Mediterranean urban system. Journal of the American Water Resources Association 52 (2), 456–472.
Oechel, W., Hastings, S., Vourlitis, G., 1993. Recent change of arctic tundra ecosystems from a net carbon-dioxide sink to a source. Nature 361 (6412), 520–523.
Sun, R.H., Chen, L.D., Wang, W., Wang, Z.M., 2012. Correlating landscape pattern with Total nitrogen concentration using a location-weighted sink-source landscape index in the Haihe River basin, China. Environmental Science 33 (6), 1784–1788.
Tilman, D., May, R.M., Lehman, C.L., Nowak, M.A., 1994. Habitat destruction and the extinction debt. Nature 371, 65–66.
Turner, M.G., Gardner, R.H., 1991. Quantitative methods in landscape ecology: An introduction. In: Turner, M.G. (Ed.), Quantitative methods in landscape ecology. New York: Springer, pp. 13–14.
United Nations, , 1992. United Nations framework convention on climate change. United Nations.
Wang, J.L., Shao, J.A., Wang, D., Ni, J.P., Xie, D.T., 2016a. Identification of the "source" and "sink" patterns influencing non-point source pollution in the three gorges reservoir area. Journal of Geographical Sciences 26 (10), 1431–1448.
Wang, J.L., Xie, D.T., Shao, J.A., Ni, J.P., Lei, P., 2016b. Identification of source–sink risk pattern of agricultural non-point source pollution in cultivated land in three gorge reservoir area based on accumulative minimum resistance model. Transactions of the Chinese Society of Agricultural Engineering 32 (16), 206–215.
Wei, W., Chen, L.D., Lei, Y., Fu, B.J., Sun, R.H., 2012. Spatial scale effects of water erosion dynamics: Complexities, variabilities, and uncertainties. Chinese Geographical Science 22 (2), 127–143.
Xu, S.L., Zhou, H., 2008. The landscape dynamics of 'source' and 'sink' and its quantification method. Research of Soil and Water Conservation 6, 64–67.
Yang, M., Li, X.Z., Hu, Y.M., He, X.Y., 2012. Assessing effects of landscape pattern on sediment yield using sediment delivery distributed model and a landscape indicator. Ecological Indicators 22 (17), 38–52.
Yu, Y.Y., Guo, Z.T., Wu, H.B., Kahmann, J.A., Oldfield, F., 2009. Spatial changes in soil organic carbon density and storage of cultivated soils in China from 1980 to 2000. Global Biogeochemical Cycles 23 (2), GB2021
Zhang, Y.J., Li, C.W., Hu, B.B., Xie, H.J., Song, A.Y., 2017. Impact of a "source–sink" landscape pattern in an urbanized watershed on nitrogen and phosphorus spatial variations in rivers: A case study of Yuqiao Reservoir watershed, Tianjin, China. Acta Ecologica Sinica 37 (7), 2437–2446.
Zhao, H.L., Liu, R.T., 2013. The "bug island" effect of shrubs and its formation mechanism in Horqin Sand Land, Inner Mongolia. Catena 105, 69–74.
Zhao, W.W., Fu, B.J., Chen, L.D., 2012. A comparison of the soil loss evaluation index and the C-factor of RUSLE: A case study in the Loess Plateau of China. Hydrology and Earth System Sciences 16, 2739–2748.
Zhou, Z.X., Li, J., 2015. The correlation analysis on the landscape pattern index and hydrological processes in the yanhe watershed, China. Journal of Hydrology 524 (5), 417–426.

Spatial Subsidy

DM Talley, University of California, Davis, Davis, CA, USA

Introduction and Definition

There is an increasing appreciation in ecology for the fact that even seemingly insular and discrete habitats are often deeply interconnected. This connectivity can arise from a number of mechanisms, including the provision of trophic resources from outside the focal habitat. This subsidization from across the habitat boundary is known as `spatial subsidy'. Synonyms and related terms include allochthonous inputs, trophic subsidy, and cross-boundary subsidy. Spatial subsidy arises when nutrients, food resources (detritus, primary producers, or prey items), or consumers move across boundaries, thus `subsidizing' local food resources in the `recipient' habitat with those from other `donor' habitats or ecosystems.

The conceptual origin for the integration of spatial subsidy into our understanding of ecosystems goes back at least to the late 1800s, when the strong trophic interactions between terrestrial and limnetic ecosystems were recognized. Spatial subsidy gradually became better integrated into ecology, in particular with respect to oceanic, riverine, and limnological ecosystems. More recently, we have seen an expansion of recognition of spatial subsidy due to the work of Gary Polis and colleagues, who integrated landscape, population, and food web ecology by focusing on the contributions of spatial subsidy to ecosystem dynamics.

Subsidies can occur between any two habitats, from those that are relatively similar (e.g., different water masses within the ocean) to those that are strikingly dissimilar (e.g., benthic and pelagic or terrestrial and aquatic habitats). Further, these subsidies can occur across a vast range of temporal and spatial scales, and can be driven by a range of biotic or abiotic processes, with ramifying effects throughout the ecosystem.

An understanding of spatial subsidies is critical to basic, theoretical, and conservation ecology. So despite the fact that ecologists often tend to focus tightly on a specific habitat or subdiscipline, external forcing due to spatial subsidies can often be the dominant factor driving the dynamics within a given habitat. Subsidies can have effects at the population level for individual species; at the community level, through changing species interactions; and at the ecosystem level, altering biodiversity and stability patterns.

This article presents examples of spatial subsidies that result from the movement of nutrients, food resources, and consumers. Along the way, key aspects of the mechanisms (e.g., unidirectional vs. bidirectional; biotic vs. abiotic) and scale (temporal and spatial) of subsidies are addressed, as well as the types of effects these subsidies have on the recipient ecosystem. Finally, the implications of understanding spatial subsidies for basic, theoretical, and conservation ecology are presented.

Movement of Nutrients

The movement of nutrients between habitats can greatly alter primary productivity in the recipient habitat, leading to trophic cascades and alterations of the community of consumers. Two specific cases are presented here: the Aeolian movement of nutrients from the African continent to the Amazon rainforest, and the input of marine-derived ornithogenic guano to island ecosystems in Baja California, Mexico.

The Amazon rainforest is a relatively nutrient-poor habitat, and the nutrient budget there is heavily subsidized through wind-blown deposition of sediments originating in Africa (**Fig. 1**). This nutrient subsidy from over 5000 km away and across a vast ocean deposits as much as 13 million tons of phosphorus per year in the Amazon Basin, and is responsible for much of the primary production in the recipient habitat. This increase in production, driven by abiotic mechanisms (wind) and occurring over vast spatial and temporal scales, has cascading effects throughout the rainforest ecosystem.

Another well-studied example of spatial subsidy of nutrients comes from arid islands in the archipelago of Bahia de los Angeles, in Baja California, Mexico (**Fig. 2**). As in many other systems, seabirds here have been shown to be effective transporters of marine-derived nitrogen and phosphorus from the aquatic to the terrestrial ecosystems, thus providing a surplus of nutrients for terrestrial plants. In most years, this nutrient subsidy has only minimal effects, since water is often the limiting factor for primary production in these desert environments. Yet during the occasional wet periods, this abundance of marine-derived nutrients allows those islands frequented by seabirds to have tremendous increases in plant growth relative to those islands less frequented by birds. This biotically mediated nutrient subsidy not only alters the composition and stability of the plant community, but the periodic pulse of production in turn alters the abundance, diversity, and composition of the island arthropod and in turn songbird communities, with effects that last well beyond the occasional wet year.

There are numerous other examples of nutrients being transported between ecosystems with dramatic effects for the recipient habitats. In the Galapagos, sea lions transport marine nutrients onshore, affecting primary productivity on beaches; in riverine systems, upstream habitats represent a source of nutrients, and downstream ones a sink; upwelling transports deep

 doi:10.1016/B978-0-444-63768-0.00829-5

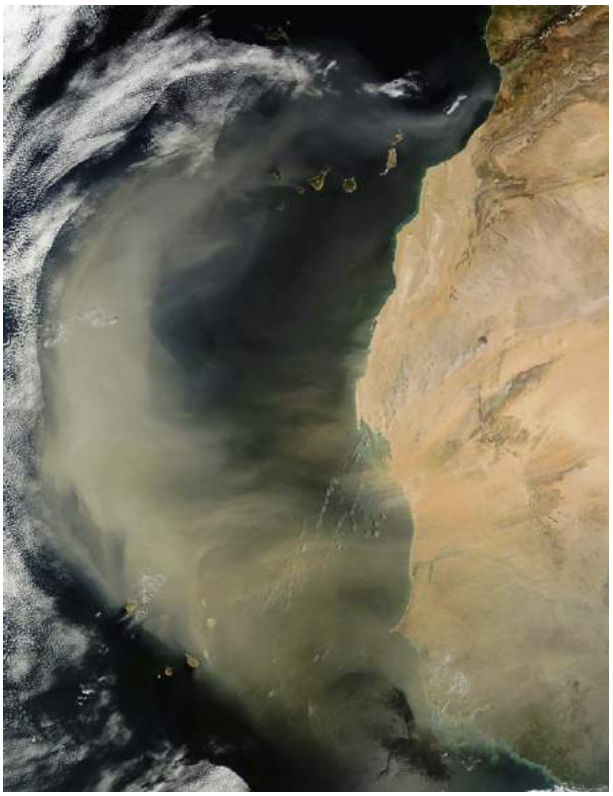


Fig. 1 Wind driving dust off the west coast of Africa. These dust storms can transport nutrients, sediment, and organic materials across thousands of kilometers of land and ocean, with dramatic effects for these distant subsidized habitats. Photo courtesy of NASA.

oceanic nutrients upward into the surface zone, where they can fuel plankton blooms; salmon transport marine nutrients to the rivers, where they die and are taken up by the local plants and algae, fuelling both the aquatic and terrestrial food webs. It is worth noting that humans mediate a number of these spatial subsidies – for example, anthropogenic nutrient inputs have dramatic effects on near-shore habitats in many areas, leading to shifts in plant and animal communities and occasionally eutrophication.

Movement of Food Resources

In many cases, food resources themselves (in the form of prey or detritus, the latter including carrion and dead plant material) are passively or actively moved between habitats. A thorough discussion of detrital inputs is available elsewhere in this volume, so here we will touch on that topic only briefly, and focus instead on the movement of prey.

Detritus represents a significant food resource for many systems, both terrestrial and aquatic, and spatial subsidies through detritus are common across many natural boundaries. Terrestrial coastal areas receive as much as 2000 kg m^{-1} of shoreline per year in the form of dead algae, seagrass, and carrion; leaf and litter fall represent massive inputs of detritus into stream ecosystems; benthic communities in aphotic zones are entirely dependent on detritus and food resources coming from the photic zone of the ocean, and new volcanic islands often have food webs based entirely on detrital input aerially deposited from other terrestrial ecosystems. Spatial subsidy through movement of detritus is a common phenomenon occurring across a range of spatial and temporal scales, and is generally driven by abiotic forcing (e.g., wind, waves, gravity, or currents).

The movement of prey items similarly impacts a vast number of habitats. A clear example is illustrated by the movement of emerging aquatic insects, where one study found as much as 97% of the biomass of emerging insects from a stream was transported to riparian terrestrial consumers including birds, herpetofauna, and arthropods. Other groundbreaking research on stream systems clearly showed a strong seasonal and reciprocal dynamic to spatial subsidy – during the summer, terrestrial invertebrates falling into the stream subsidized aquatic consumers, representing 44% of the total energy budget for fishes. Conversely, during the spring emergence of aquatic insects, allochthonous inputs from the aquatic environment provided the terrestrial ecosystem (specifically birds) with more than 25% of their total energy, and as much as 98% for a species of wren. This biotically mediated reciprocal flux between the forest and stream can alter community structure by causing top-down effects from subsidized predators. The dynamic and reciprocal nature of this spatial subsidy,

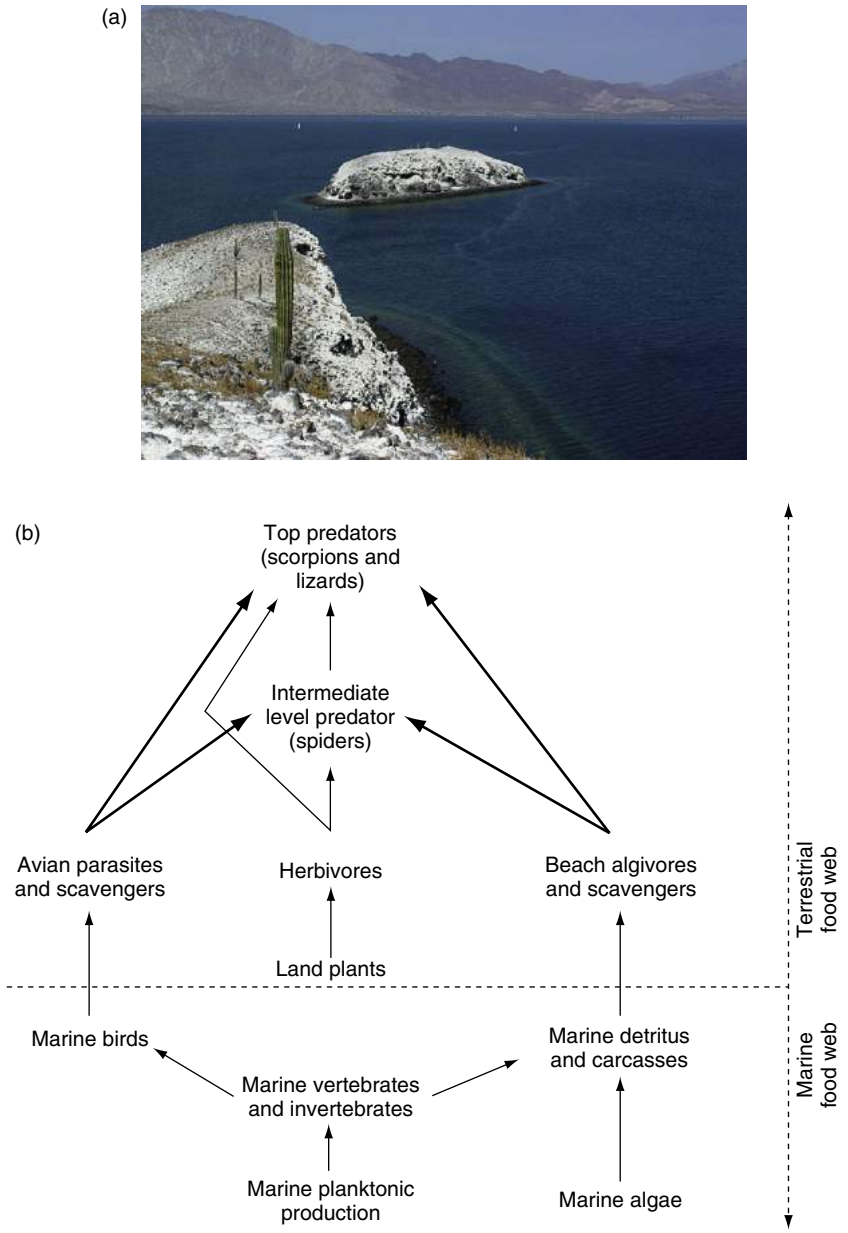


Fig. 2 (a) The Gemelitos islands in the archipelago of Bahía de los Ángeles, Baja California, Mexico. The high levels of marine-derived nutrients deposited on the islands by seabirds represent a subsidy to the terrestrial and intertidal systems, with dramatic effects on recipient communities. (b) A schematic food web for these small islands, showing the importance of marine subsidy. Redrawn from Polis, G.A., Hurd, S.D., 1995. Extraordinarily high spider densities on islands: Flow of energy from the marine to terrestrial food webs and the absence of predation. Proceedings of the National Academy of Sciences of the United States of America 92, 4382–4386.

while rarely experimentally tested, is likely common in nature, particularly wherever adjacent habitats undergo asynchronous pulses of productivity.

Spatial subsidies through the movement of prey items are, however, not confined to those that cross the aquatic–terrestrial boundary as in the examples above, but are commonplace in many environments and at many spatial scales. In particular, organisms which undergo `ontogenetic shifts' in foraging or habitat-use patterns often mediate these subsidies. Cicadas, termites, and other emerging insects provide a subsidy from belowground to aboveground ecosystems; ontogenetic shifts in foraging and movement patterns of California killifish (*Fundulus parvipinnis*, a small wetland resident fish) subsidize predators in subtidal ecosystems with energy these fish derived from foraging as juveniles in high intertidal pools; migrations of potential prey items, from Monarch butterflies to caribou, provide consumers in distant habitats with substantial trophic subsidies. Taken as a whole, subsidies resulting from the movement of prey (or detritus) can range in distance from a few centimeters to many thousands of kilometers, and over timescales from a matter of hours to decades.

Fig. 3 A coyote (*Canis latrans*) foraging on marine organisms in an intertidal salt marsh. Coyote densities in coastal regions can be more than an order of magnitude higher relative to areas where they receive no marine spatial subsidies. Photo Credit L. Goodwin, Aquatic Adventures Science Education Foundation.

Movement of Consumers

In a fashion similar to the movement of prey items, the movement of consumers can be a mechanism creating spatial subsidy. As in the other examples presented here, this can occur between two terrestrial habitats, two aquatic habitats, or between terrestrial and aquatic environments.

There are numerous examples of terrestrial mammals that make forays into aquatic habitats in order to forage. In fact, there are over 45 species of terrestrial mammals documented to feed at least in part on marine resources, consuming over 200 marine taxa. For some terrestrial mammals, such as coyotes (*Canis latrans*), this consumption allows predator population densities to persist several orders of magnitude over what it would be in areas without the ability to forage on marine foods (**Fig. 3**). This increase in population density of these omnivores in turn allows them to depress other prey resources such as small rodents, thus leading to `apparent trophic cascades', where subsidized consumers are able to exert a greater force on the local food web than would be possible based on *in situ* resources alone.

Similarly, migrating consumers can have profound effects in one habitat due to the trophic subsidy they receive from another. In the case of geese, the populations that breed in the Canadian arctic have been subsidized for several decades by agricultural production. This allochthonous agricultural resource provides a higher nutritional quality than do the historic feeding grounds, with profound results for both the goose populations and the habitats they inhabit. One is that the higher nutritional quality and abundant resources of the agricultural fields have caused a boom in population density for many species of geese. This, in turn, leads to an `apparent trophic cascade', with the abundant birds completely altering marshes in which they either did not feed historically or generally had lesser effects. The geese thus caused massive loss of vegetation and irreversible changes to soil conditions in many coastal marshes, with consequent loss of soil invertebrate species and impacts on those predators that use them (such as passerine birds and shorebirds).

As in the case of movement of prey items, consumer movements creating spatial subsidy occur across a range of habitat types and temporal and spatial scales. Krill move seasonally between feeding on algae beneath the Antarctic ice shelf to feeding on pelagic phytoplankton; a number of aquatic predators make foraging migrations over a broad range of spatial scales, in horizontal or vertical directions; a species of side-blotched lizards on the terrestrial-arthropod poor Islas Encantadas archipelago in the Gulf of California rely so heavily on intertidal food resources that they have evolved enlarged nasal salt glands to assist with removal of excess electrolytes; and humans forage across vast ranges of terrestrial and aquatic habitats, bringing back trophic subsidies from the interior of continents to the depths of the oceans.

The movement of consumers across boundaries is a key factor in the cross-system exchanges represented by spatial subsidies. Such exchanges are often driven by ontogenetic shifts in an organism's foraging or habitat-use pattern, but can also be realized whenever mobile consumers have the ability to cross a habitat boundary.

The Importance of Spatial Subsidy

Given the ubiquity of spatial subsidy and its potential to dramatically alter communities, populations, and environments, it is critical that these linkages be incorporated into ecological thinking. Understanding spatial subsidy illuminates both empirical and theoretical ecology, and informs conservation biology in theory and practice.

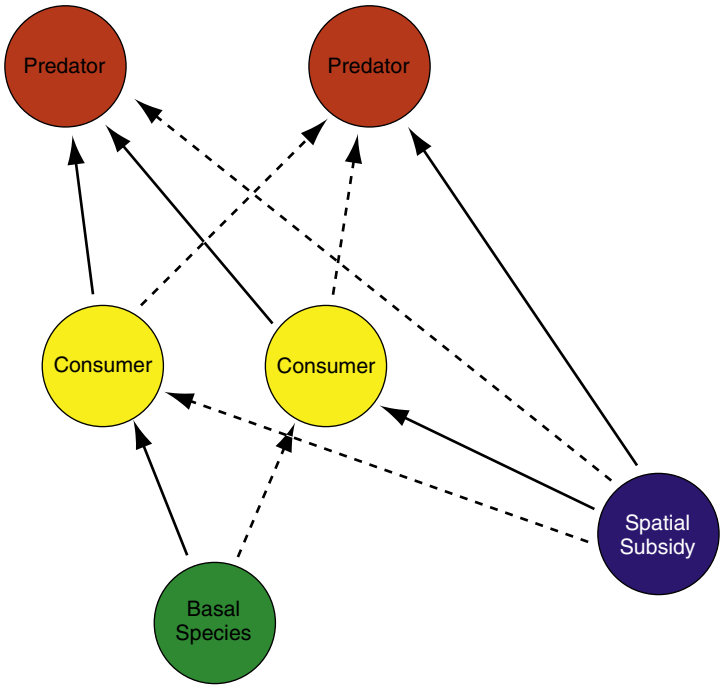


Fig. 4 Theoretical modeling suggests that the effect of spatial subsidy on food webs is dependent on a number of factors, including the abundance of subsidy and if the recipient species are specialists or generalists. In the tritrophic model illustrated here, a food web made up of generalist consumers (all arrows) would be stabilized by low to moderate levels of spatial subsidy. If the web were made up of specialists (dashed arrows removed), subsidies would be expected to destabilize the system. Adapted from Huxel *et al.* (2002).

As illustrated in many of the examples presented above, spatial subsidies are often the driving force structuring ecosystems, and as such are pivotal to developing a full understanding of the population of ecosystem dynamics. Subsidized consumers can push ecosystems to alternative stable states; allochthonous nutrient enrichment can completely alter the abundance and composition of primary producers, with profound effects throughout the ecosystem; and cross-system exchanges of resources can affect community stability and species diversity. In many cases, attempting to fully understand the dynamics of a focal habitat will be doomed to failure if external forcing through spatial subsidy is not considered. This has been appreciated for some time with regard to the effect of terrestrial systems on rivers, but only recently has the extent of spatial subsidy in so many ecosystems been acknowledged.

The importance to theoretical ecology of spatial subsidies has in recent years also seen a surge of interest, with exciting and illuminating results. Gary Polis made great strides in integrating the idea of spatial subsidy into food web dynamics and landscape ecology, and that work heralded the beginning of a number of studies examining the theoretical underpinnings of spatial subsidies (although it should be noted here that theoretical work on nutrients as a subsidy goes back somewhat farther). Modeling studies looking at the influence of allochthonous inputs across habitat boundaries, for example, predict strong effects on food web stability. Low to moderate amounts of allochthonous inputs relative to autochthonous productivity (that originating in the focal habitat) were shown to stabilize various kinds of food web dynamics. Other findings showed that food web structure influences the degree of impact of allochthonous inputs on the stability of food webs (**Fig. 4**). Specialization at the top trophic level (such as with scavengers) limited the indirect effects of allochthonous inputs. In contrast, generalists feeding on both local and allochthonous prey can exhibit increased densities due to the allochthonous resources, resulting in increased predation pressure on autochthonous prey and possible ecosystem-wide effects. These results match well with mensurative findings from nature, such as the coyote work referenced above.

Other models have suggested that differences among trophic levels in linkage among habitats significantly influence food web stability. If allochthonous resources are utilized by the second trophic level in a three trophic level system, then generalists (herbivores for example) will impact the first trophic level and provide increased resources for the third trophic level. Specialists at the second trophic level (perhaps detritivores) will not directly impact the first trophic level, but will provide increased resources for the third trophic level. These and other findings suggest that there is a wealth of information to be gained by incorporating spatial subsidies into theoretical models of communities and ecosystems.

Integrating an understanding of spatial subsidies is also critical to understanding ecological systems and theories, and this will provide various benefits to conservation, in both theory and practice. Several aspects of spatial subsidy relating to conservation

deserve particular attention: situations where subsidies have been altered by humans, and ways in which information about linkages can be used to improve the practice of conservation.

Anthropogenic Alterations

Human activity can alter the flow of spatial subsidy in a number of ways. Habitat fragmentation, for example, can increase subsidy (through small patches having greater `edge effects'), or can decrease trophic subsidy, by leaving some patches too distant from sources to allow natural flow. Further, human activity can directly block or increase subsidy relative to natural levels. For example, the movement of spatial subsidies between coastal consumers (utilizing aquatic resources) and inland habitats has been dissected in many regions by roads and highways, potentially restricting access to these trophic resources for wide-ranging species such as coyotes or small mammals. This lack of connectivity could have wide-ranging effects on terrestrial communities both by lowering population densities of consumers that are subsidized by aquatic resources and by restricting the flow of aquatic nutrients to inland.

There are numerous other examples: the creation of levees around rivers diminishes the aquatic subsidy to the surrounding terrestrial floodplain; urban development in coastal zones reduces the ability for terrestrial organisms to obtain spatial subsidy from the ocean; and destruction of foraging habitat along migratory bird flyways can greatly lessen the subsidy obtained from intermediate habitats. Over 90% of Pacific saltwater and Great Lake freshwater marshes have been lost, largely due to direct human habitat alteration, leading to limitation of aquatic resources available to terrestrial consumers. On southern California beaches in the United States, macrophyte wrack washing up in the intertidal provides a regular subsidy for macrofauna, which in turn provide food for shorebirds. The practice of `beach grooming' (using heavy equipment to remove macrophytes and debris from sandy beaches) in populated areas directly removes this subsidy and thus mitigates the associated bottom-up effects.

Anthropogenic processes and mechanisms of spatial subsidy can be even more direct, however. In any given year, millions of tons of crops are moved between continents, and some 75 million metric tons of marine biomass onto land worldwide, including 27 million tons of discarded nontarget animals (bycatch), as well as 126 500 t (dry weight) of algae, with ramifications for the entire oceanic ecosystem, as well as subsidizing human populations often quite distant from the ocean. Each of these movements of resources not only represents a spatial subsidy of its own, but also suggests that removing these quantities of organisms from the ecosystem may well be altering natural levels of subsidy occurring there.

Conservation Implications

Recognizing and understanding spatial subsidies is necessary for effective conservation. As most examples in this article illustrate, there are frequent, strong interactions between often seemingly disconnected systems, with one habitat's influence often dominating the dynamics of the other over vast temporal and spatial scales. Besides the conservation benefits inherent in an increased understanding of nature, there are some very specific areas in which recognizing spatial subsidies at the aquatic–terrestrial interface will benefit conservation science.

Siting decisions for areas of protection for threatened species or habitats will be greatly improved by integrating consideration of potential spatial subsidies into the decision-making process. For example, streams have traditionally been considered to be recipients, rather than sources, of trophic resources in terrestrial systems. Yet temperate forest herpetofauna have been shown to increase growth rates by as much as 700% through spatial subsidies arising from stream ecosystems.

A better understanding of these functional linkages also illuminates questions about buffer zones and spatial scale of protected areas. Studies of the effects of marine inputs on terrestrial ecosystems clearly demonstrate that it can be important to account for marine input to understand the dynamics of terrestrial systems. However, there is only inferential information regarding the distance offshore from which these resources originate. Fundamental questions are still unanswered about how far offshore one must protect a habitat to ensure adequate flow of allochthonous input. Nor is it clear how far inshore the effects of subsidies penetrate to impact terrestrial communities. An understanding of the spatial properties of these linkages will provide conservation biologists with better tools to protect threatened habitats and organisms. Efforts to establish effective marine protected areas will require a deep understanding of just how spatial subsidies between various marine habitats affect nearby (and distant) ecosystems.

Summary

Cross-system exchanges in the form of spatial subsidies are ubiquitous, even between seemingly distant and disconnected habitats. Subsidies can occur across a vast range of temporal and spatial scales, and can be driven by biotic or abiotic forcing (or some combination thereof). Spatial subsidies can have profound effects on entire ecosystems, with implications for species diversity, ecosystem stability, population dynamics, and persistence, often being the dominant factor driving ecosystem properties. Acknowledging and understanding spatial subsidies is critical to basic, theoretical, and conservation ecology.

See also: Ecological Data Analysis and Modelling: Spatial Models and Geographic Information Systems. Terrestrial and Landscape Ecology: Spatial Distribution

Further Reading

Carlton, J.T., Hodder, J., 2003. Maritime mammals: Terrestrial mammals as consumers in marine intertidal communities. Marine Ecology – Progress Series 256, 271–286.
Huxel, G.R., McCann, K., 1998. Food web stability: The influence of trophic flows across habitats. American Naturalist 152, 460–469.
Nakano, S., Murakami, M., 2001. Reciprocal subsidies: Dynamic interdependence between terrestrial and aquatic food webs. Proceedings of the National Academy of Sciences of the United States of America 98, 166–170.
Polis, G.A., Anderson, W.B., Holt, R.D., 1997. Toward an integration of landscape and food web ecology: The dynamics of spatially subsidized food webs. Annual Review of Ecology and Systematics 29, 289–316.
Polis, G.A., Hurd, S.D., 1995. Extraordinarily high spider densities on islands: Flow of energy from the marine to terrestrial food webs and the absence of predation. Proceedings of the National Academy of Sciences of the United States of America 92, 4382–4386.
Polis, G.A., Power, M.E., Huxel, G.R., 2004. Food Webs at the Landscape Level. Chicago: University of Chicago Press.
Power, M.E., 2001. Prey exchange between a stream and its forested watershed elevates predator densities in both habitats. Proceeding of the National Academy of Science of the United States of America 98, 14–15.
Summerhayes, V.S., Elton, C.S., 1923. Contributions to the ecology of Spitzbergen and Bear Island. Journal of Ecology 11, 214–286.
Talley, D.M., Huxel, G.R., Holyoak, M., 2006. Habitat connectivity at the land–water interface. In: Crooks, K., Sanjayan, M. (Eds.), Connectivity in Conservation. Cambridge: Cambridge University Press, pp. 97–129.
Willson, M.F., Gende, S.M., Marston, B.H., 1998. Fishes and the forest: Expanding perspectives on fish-wildlife interactions. Bioscience 48, 455–462.

System Omnivory Index☆

S Libralato, OGS (Istituto Nazionale di Oceanografia e di Geofisica Sperimentale), Sgonico (Trieste), Italy

Introduction

The system omnivory index (SOI) measures the distribution of feeding interactions among trophic levels of food webs, thus SOI allows for evaluating the complexity and connectivity of food webs.

Ecological networks are widely used in ecology to address theoretical concepts and to study properties of natural systems. In particular, networks of trophic interactions (food webs) are tools for describing ecosystems and for searching invariant properties in nature. Food webs represent elements (elsewhere also called nodes) connected through links (also termed edges): elements are group of species, a single species, or even the life-history stage of a population, while trophic links represent consumption (or predation) and are directed (predator eats prey). Complexity of food webs, i.e. the number and structure of trophic interactions, has been studied by ecologists since decades (Christensen, 1995; May, 1973; Neutel *et al.*, 2007; Pimm, 1980).

Complexity is an important feature of ecological food webs because it has been related to maturity and stability properties of ecosystems. Web-like and complex structured trophic networks imply the presence of omnivores, i.e., consumers feeding on more than one element of the food web. Thanks to prey-switching abilities, omnivores can possibly be only moderately affected by eventual fluctuations or decrease of one of the prey populations. Therefore, omnivory contributes in dampening perturbations affecting species of the ecosystem, allowing a short return time (resilience), and avoiding magnified propagation of disturbances (resistant). Moreover, omnivores' interactions increase stability through weak interactions that dampen the effects of alternative potential strong interactions (in an exemplified one consumer–two resources interactions). Web-like structured food webs are thus associated to ecosystems that are able to adsorb external shocks and to recover from perturbation. Understanding these capabilities of real food webs will enable to better manage ecosystems, promoting their persistence and avoiding anthropogenic disturbances that can lead to deep changes in ecosystem structure (Christensen, 1995; Fagan, 1997; Polis and Strong, 1996).

Measuring Complexity

The complexity of food webs is generally increasing with the number of elements (N), assuming a progressive increase of the number of links (L). However, for a given number of elements, different structures of the web (number of links) imply great differences in complexity. There are many measures applicable for quantifying the complexity of food webs (Banasek-Richter *et al.*, 2009; Neutel *et al.*, 2007), such as the direct connectance (C) represents the ratio between number of expressed links and number of theoretically possible links ($C=L\cdot N^{-2}$); the connectance index (CI) is excluding cannibalism and top predators effects and it is equal to $\mathrm{CI}=L\cdot(N-1)^{-2}$; the linkage density (LD) is the average number of links per element, that is, $\mathrm{LD}=L\cdot N^{-1}$. These measures of complexity have been developed for binary networks, in which for each couple of elements the interaction is set as absent (0) or existing (1), but there is no indication of its strength. Conversely, in weighted food webs, at each existing link is associated a strength of the interaction (weight), quantified by the consumption flow. Accounting for the interaction weights in the complexity measurement is relevant, given the importance of of weak interactions in determining stability properties of ecosystems. Moreover, topological indicators do not exploit the information on the direction of flows. However, direction of flows is a fundamental characteristic of predator–prey interaction, that is, the effects of the interaction are different for the prey and for the predator. Therefore, accounting of directionality of flows in trophic networks is important for capture real properties of ecological systems.

The SOI is intended to exploit the quantitative information carried by weighted and directed trophic networks for evaluating their complexity and thus for understanding the capabilities of ecosystems to respond to natural and anthropic perturbations. In particular, the SOI is a measure of the degree of omnivory in the trophic network, that is, a measure of the distribution of feeding interactions among trophic levels (Christensen and Pauly, 1993).

Definition

Given a trophic network of n elements, the SOI is calculated as the weighted average of the elements' omnivory, this latter calculated as the omnivory index (OI). The OI of the consumer element i with trophic level TL_i is quantified as the variance of the trophic levels of its preys (TL_j) (Williams and Martinez, 2004); thus

☆*Change History:* March 2013. S Libralato introduced small edits in the text of the article including citations, added the sections "Applications" and "References", and added **Fig. 6**.

$$\mathrm{OI}_i = \sum_{j=1}^{n} \left[\mathrm{TL}_j - (\mathrm{TL}_i - 1)\right]^2 \cdot \mathrm{DC}_{ij} \qquad [1]$$

where the contribution of each prey j to the variance of the consumer is proportional to the fraction of prey j in the diet of consumer i (DC_{ij}). The SOI of a given trophic network is quantified as the weighted average of the OI of all consumers of the network (Christensen and Pauly, 1993); thus

$$\mathrm{SOI} = \frac{\sum_{i=1}^{n} [\mathrm{OI}_i \cdot \log (Q_i)]}{\sum_{i=1}^{n} \log (Q_i)} \qquad [2]$$

where the weighting factors are taken as the logarithm of each consumer food intake (Q_i). This allows for accounting of the different strengths of consumer interactions and thus accounts for intra-element interaction strengths. The logarithm is used on the observation that consumptions are approximately log–normally distributed within the system (Christensen and Pauly, 1993).

Properties

The SOI is influenced not only by the number of links in the web but, other factors being equal, also by topological configuration of links and by their weights, these latter intended as the total flow to consumer. In the following, the effects of different topological configurations on SOI will be explored by comparing networks with equal number of elements but different structure.

Number of Links

From a chain-like web, the augmentation of the number of links imply an increase of the network's SOI. This is observed even if the total number of elements and total flows to each consumer are kept constant. The increase in SOI, therefore, is due to the inevitable increase of the number of omnivorous species in webs with increasing number of links, as shown in **Fig. 1**.

This is coherent with definition and main reasons for the develoment of the SOI as a measure of complexity of food webs (Christensen and Pauly, 1993). Increasing values of SOI measure the increase of omnivory elements that are found in nature in developed and mature ecosystems.

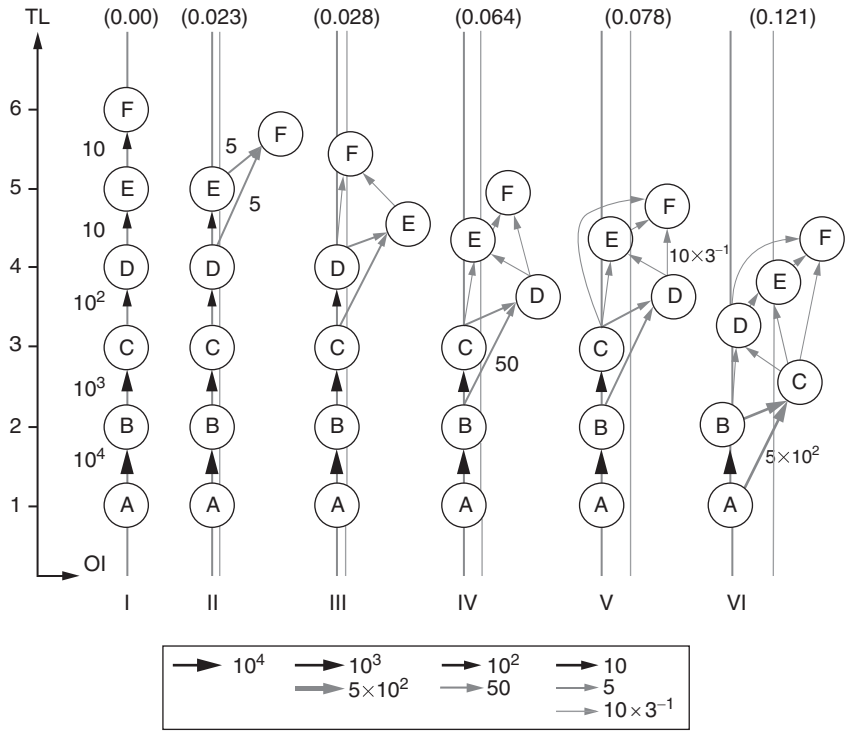


Fig. 1 Effect of network complexity on the value of the SOI. Elements of the networks are displaced on *y*-axis according to effective trophic level, and on *x*-axis according to OI values, with zero at the vertical continuous line for each network. Vertical dotted line represents the SOI of each web (values are reported above between parentheses). Keeping constant the number of compartments and total consumption for each consumer, the number of links in the trophic web (thus complexity) increases from case I to VI and so does SOI. Flows are expressed in hypothetical units (key legend at the bottom) but natural webs are mimed by assuming gross efficiency (ratio between consumption and production) for each element is equal to 50%.

However, keeping constant the number of links and thus of omnivorous elements, the SOI augments also when omnivores feed on nonadjacent trophic levels. As shown in Fig. 2, the increase in the number of levels between two trophic levels exploited by the same omnivore will produce a disproportionate increase in SOI, disregarding the number of links in the web (kept constant in the schematic example of Fig. 2).

Predators consuming preys at two very different trophic levels can produce high imbalances in natural systems. Productivity and ecological cost (total energy needed to produce biomass) are increasing for decreasing trophic levels and therefore equal exploitation by a predator has quite a different impact on the two populations: in particular, the switch between consumption on lower trophic level to the higher one can deplete this latter. In case of predation on nonadjacent trophic levels, therefore, high values of SOI are related to instability of the system (Neutel *et al.*, 2007).

This effect of high SOI due to extreme differences on trophic level of omnivores' preys is dampened down when omnivory is distributed over a range of trophic levels, even including extremely different ones (Fig. 3). Comparison with Fig. 1 implies that generalist species contribute to omnivory more than species characterized by low omnivory level (i.e., species that prey on lower number of trophic levels).

Weight of Links

SOI is also affected by the absolute amount of consumption flow rate. Keeping fixed the proportions of preys in the diet of predators but changing the absolute amount of flows, although not affecting OI of each element of the network, deeply influence the SOI of the network. Increasing flows to elements characterized by low omnivory definitely reduces the SOI of the system, while increasing consumption flows for omnivorous species will produce an increase of SOI of the web. An exemplification of such effects is reported in Fig. 4.

Therefore, generally, the SOI is not independent from the absolute values of consumption flows. This implies that particular attention should be paid when comparing food webs described with a different unit since their SOI estimates might not be comparable.

Biological Resolution

Building the food web of an ecosystem does not produce a univocal result in terms of structure. According to needs and focus of the analysis, the food web can have different biological resolution: that is to say, the same ecosystem can be described through a food web with a different number of elements. To be consistent with the same real ecosystem, however, there needs to be a correspondence between elements of the food webs and flows and biomasses should be coherent (thus total consumption of an element resulting from the lumping is equal to the sum of consumptions of aggregated elements). Equivalent food webs built by aggregating some elements illustrates the effect on SOI of the biological resolution used for building the food web (Fig. 5).

Topological indicators, such as connectance and linkage, are more influenced than SOI by aggregation as shown in Fig. 5. In particular, aggregation of primary producers has no effect on SOI, for it considers only consumers in food webs with directed interaction. Moreover, the accounting of consumption flows allows for balancing the aggregation of links in the SOI evaluation.

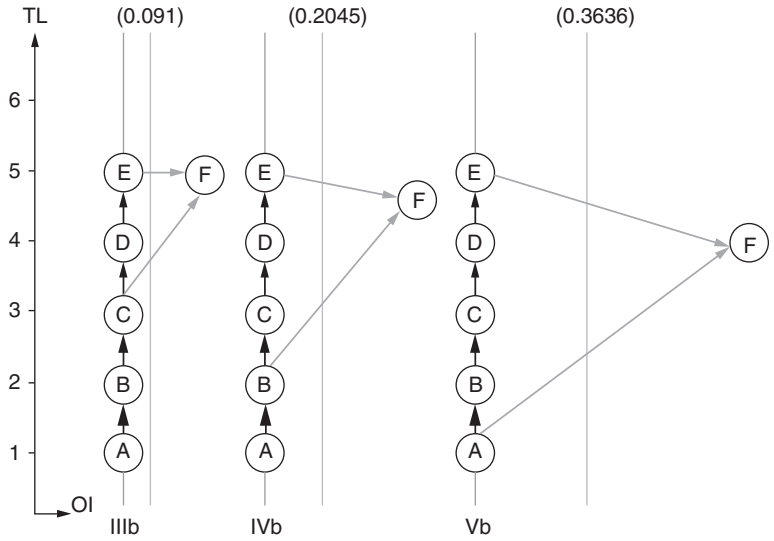


Fig. 2 Effect of feeding on nonadjacent trophic levels on the value of the SOI, other factors being constant. Number of compartments, total consumption for each consumer, and the number of links in the trophic web are kept constant. SOI is increasing as the omnivory of element F with regard to two groups of increasing distance in terms of trophic levels (network IIIb follows networks I and II of Fig. 1). Network representation is done according to Fig. 1.

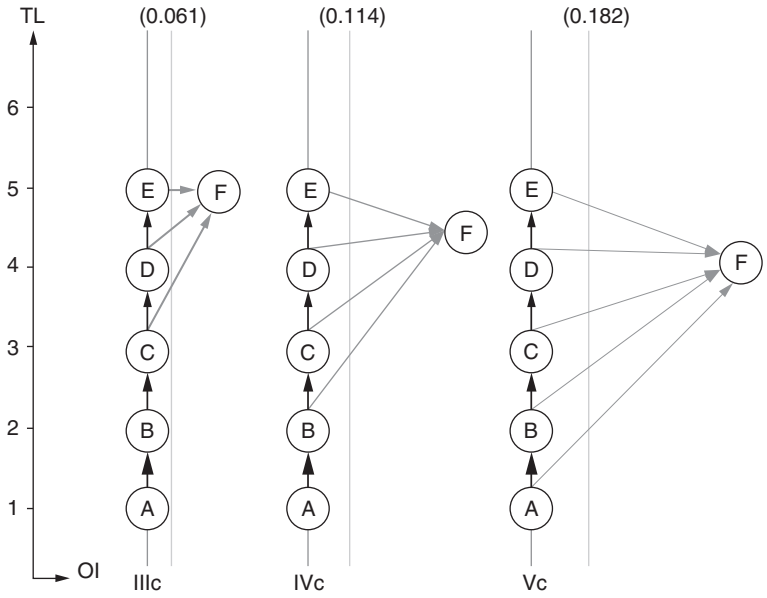


Fig. 3 Effect of feeding an increasing number of adjacent trophic levels on the value of the SOI. Number of compartments and total consumption for each consumer are kept constant. SOI increases as the number of links increases. Note that network IIIc follows networks I and II of **Fig. 1** and that the numbers of links of IIIc, IVc, and Vc are equal to the numbers of links of networks III, IV, and V of **Fig. 1**, respectively.

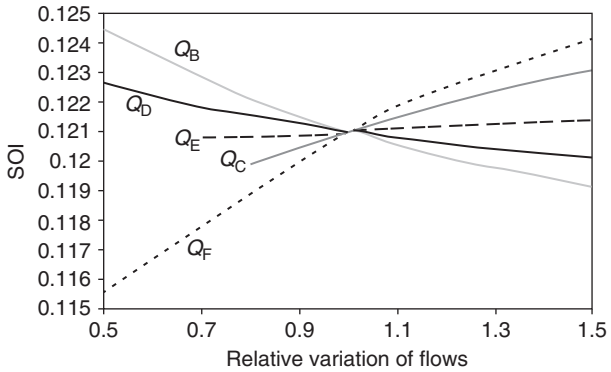


Fig. 4 Effects on SOI due to changes on consumption flows (Q) at different elements of network VI of **Fig. 1**. Elements with high omnivory with respect to average omnivory in the system (F, C, E) have positive relationship between SOI and induced changes on consumption flow. Increasing flows to elements with low omnivory (B, D) produces decreasing values of SOI.

Applications

SOI has been applied as an ecosystem indicator to evaluate complexity of real food webs and ecosystem health. Comparison of stability and complexity indexes (including SOI) for coastal marine food webs, highlighted positive correlation between SOI and magnitude of change and recovery time, thus suggesting that SOI is inversely related to stability at least in the marine ecosystems analysed (Perez-Espana and Arreguin-Sanchez, 1999).

Moreover, application of SOI and other ecological indicators on the basis of outputs of protected and fished marine food webs standardized by number of elements, showed that SOI is sensitive to fishing (Libralato *et al.*, 2010). This result suggests that SOI can be efficiently used as an ecological indicator for evaluate the effects of fishing on ecosystem food webs.

Recently, in a broad meta-analysis including 75 well documented marine food webs distributed globally (Heymans *et al.*, 2012), SOI and other ecological indices were calculated. Results showed that SOI differences were not significant when food webs are grouped by size, depth, latitude and longitude classes. Conversely, significant differences in SOI were found when the food webs were grouped by ecosystem types (Chi square; $P=0.04$; with SOI of reef > shelf > bay > lagoon > coastal > estuary) and by exploitation (Chi square; $P=0.03$; SOI exploited > SOI low or no fisheries) (Heymans *et al.*, 2012). Although SOI differences by ecosystem types do not have a clear explanation, results by exploitation further support SOI as a potential indicator for ecosystem stress induced by fishing, although further studies are needed.

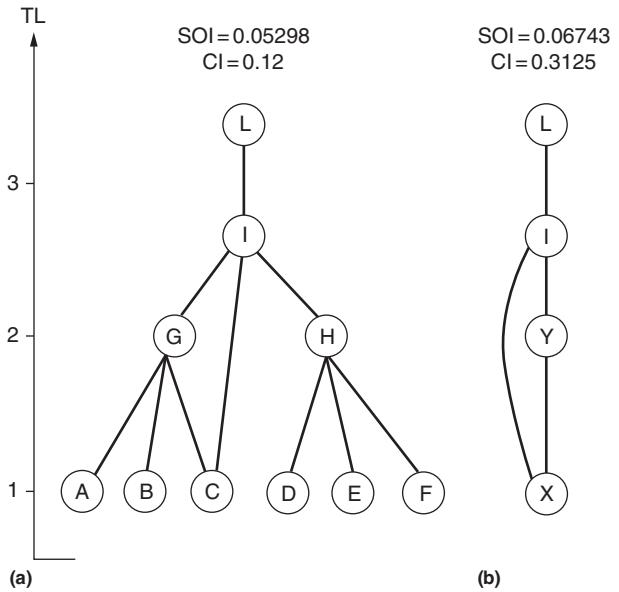


Fig. 5 Aggregation effect on SOI. Elements of network 'a' are aggregated in network 'b' according to trophic position and feeding elements. Element Y results from the aggregation of elements G and H; element X results from the aggregation of basal elements A, B, C, D, E, and F. Flows are aggregated accordingly. SOI and CI are reported above the respective food web. CI appears to be more sensitive to the number of links and elements than SOI; thus, this latter is less affected by lumping network elements.

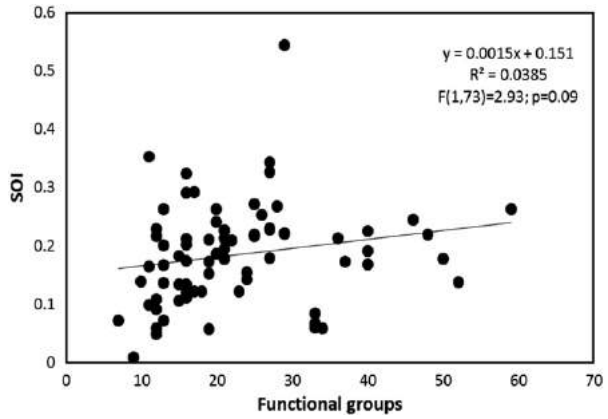


Fig. 6 Effect on SOI of the number of functional groups used in food webs to describe real marine ecosystems. Data regards the 75 models used in Heymans *et al.* (2012). Results show a linear trend non significant supporting theoretical invariance of SOI by the number of functional groups used to describe ecosystems.

Furthermore, using estimates from the 75 food web models collected in Heymans *et al.* (2012), **Fig. 6** shows no significant trend of SOI with number of functional groups thus confirming that SOI is an invariant measure independent from the number of components (nodes) of the model.

Capabilities

The SOI is therefore a good measure of the complexity of ecological networks of trophic interactions, and increasing values of SOI are, to a certain extent, representative of networks of increasing complexity that should be related to increase of maturity of ecosystems. However, extremely high values of SOI can be due to omnivores feeding on nonadjacent trophic levels that do not contribute to the complexity of ecosystems but can introduce instabilities. Therefore caution is needed when evaluating ecosystems characterized by the presence of those omnivores. In general, however, SOI appears to be a reliable measure of complexity and stability even in the presence of aggregation of elements of trophic networks and it is less affected by structural and topological changes introduced by lumping elements of networks. Application to real marine food webs,

moreover, highlighted the possibilities to use SOI as an ecosystem indicator also for stresses induced on ecosystems, such as effects of fishing on marine ecosystems.

See also: Ecological Data Analysis and Modelling: Climate Change Models

References

Banasek-Richter, C., Bersier, L.-F., Cattin, M.-F., *et al.*, 2009. Complexity in quantitative food webs. Ecology 90, 1470–1477.
Christensen, V., 1995. Ecosystem maturity – Towards quantification. Ecological Modelling 77, 3–32.
Christensen, V., Pauly, D., 1993. Trophic models of aquatic ecosystems. In: ICLARM conference proceedings, 26. Manila: ICLARM.
Fagan, W., 1997. Omnivory as a stabilizing feature of natural communities. American Naturalist 150, 554–567.
Heymans, J.J., Coll, M., Libralato, S., Christensen, V., 2012. Ecopath theory, modeling, and application to coastal ecosystems. In: Wolanski, E., McLusky, D., Baird, D., Ashish (Eds.), Treatise on Estuarine and Coastal Science. Waltham: Academic Press, pp. 93–113.
Libralato, S., Coll, M., Tempesta, M., *et al.*, 2010. Food-web traits of protected and exploited areas of the Adriatic Sea. Biological Conservation 143, 2182–2194.
May, R.M., 1973. Complexity and stability in model ecosystems. Princeton, NJ: Princeton University Press.
Neutel, A.-M., Heesterbeek, J.A.P., van de Koppel, J., *et al.*, 2007. Reconciling complexity with stability in naturally assembling food webs. Nature 449, 599–603.
Pascual, M., Dunne, J.A., 2005. Ecological networks: Linking structure to dynamics in food webs. Santa Fe, NM: Oxford University Press.
Perez-Espana, H., Arreguin-Sanchez, F., 1999. Complexity related to behavior of stability in modeled coastal zone ecosystems. Aquatic Ecosystem Health and Management 2, 129–135.
Pimm, S.L., 1980. Properties of food webs. Ecology 61, 219–225.
Polis, G.A., Strong, D.R., 1996. Food web complexity and community dynamics. American Naturalist 147, 813–846.
Williams, R.J., Martinez, N.D., 2004. Limits to trophic levels and omnivory in complex food webs: Theory and data. American Naturalist 163, 458–468.

Further Reading

McCann, K., Hastings, A., 1997. Re-evaluating the omnivory–stability relationship in food webs. Proceedings of the Royal Society of London B 264, 1249–1254.
Pauly, D., Soriano, M.L., Palomares, M.L., 1993. Improved construction, parametrization and interpretation of steady-state ecosystem models. In: Christensen, V., Pauly, D. (Eds.), ICLARM conference proceedings 26.Trophic Models of Aquatic EcosystemsManila: ICLARM, pp. 1–13.
Pimm, S.L., 1982. Food webs. London: Chapman & Hall.
Polis, G.A., Winemiller, K.O., 1996. Food webs: Integration of patterns and dynamics. New York: Chapman and Hall.
Ulanowicz, R.E., 1986. Growth and development – Ecosystems phenomenology. New York: Springer.
Wulff, F., Field, J.G., Mann, K.H., 1989. Growth and development – Ecosystems phenomenology. Berlin: Springer.

Trophic Classification for Lakes☆

Fu-Liu Xu and Yang Jiao, Peking University, Beijing, China

A Review of Ecological Indicators for Trophic Classification for Lakes

The classical system of distinguishing lake trophic states can be traced to August Thienemann. He classified lakes according to their trophic conditions into oligotrophic (low trophy), eutrophic (high trophy), and dystrophic (lakes of boggy character, with highly colored water due to the presence of organic matter from decaying vegetation), on the base of the composition of lake bottom sediments and the associated benthic fauna. Since the 1960s and the 1970s, a number of attempts have been made to quantitatively evaluate the trophic state of lakes. Ecological indicators so far used for trophic classifications of lakes may be divided into physical, chemical, biological, and system-level aspects. They have been used solely as single-variable trophic indices or synchronously as multiparametric approaches. The single-variable trophic indices can be divided into abiotic and biotic aspects. Among the abiotic parameters including physical and chemical indicators, plant nutrients (phosphate, nitrate), oxygen demanded (biochemical (BOD), chemical (COD)), and transparency were usually used to assess lake trophic levels. Also, biotic parameters were often employed to assess lake trophic conditions, given the sensitivity of aquatic organisms, especially algae and macroinvertebrates, to eutrophication processes. Phytoplankton, both in running waters and lakes, turned out to be a reliable environmental tool when estimating different levels of trophy. With the exception of chlorophyll *a* (Chl-a) concentrations, phytoplankton cell number, species number, and biomass, some form of index, such as Hurlbert's, Margalef's, Menhinick's, Shannon's, Simpson's, and McNaughton's, has always been used in assessing eutrophication conditions in aquatic environments. Like phytoplankton, zooplankton indices have also been adopted for lake trophic classifications. For instance, zooplankton community size, structure, abundance, and biomass of micro- and macrozooplankton, the shift of Rotifer communities, as well as zooplankton assemblages were applied to classify the lake trophic status.

However, these relatively simple single-variable trophic state criteria represent subjective judgments, and may be limited spatially. Further, the use of descriptive classifications for lake trophic states such as oligotrophic, mesotrophic, eutrophic, and so on, could create difficulties when attempting to describe continuous changes in a lake's trophic state or in studying quantitatively the eutrophication mechanism. The multidimensional nature of the eutrophication phenomenon means that no single variable is representative of the eutrophication status of a given water body. More robust trophic state criteria or indices using multivariate approaches have been proposed by a number of investigators. The contributions of Carlson, Walker, and Porcella offer a 0–100 scale providing continuous numerical classes of lake trophic states and a rigorous foundation for quantitative studies of the mechanisms behind eutrophication. This effectively eliminates the subjective labeling associated with the use of oligotrophic, mesotrophic, and eutrophic states as indicators. The trophic state index (TSI) based on several biological, chemical, and physical indicators, especially the Carlson-type TSI, offers the most suitable and acceptable method for trophic classifications of lakes. Mathematical methods play a very important role in lake trophic classifications in terms of parameters chosen, weighting factor calculation, and sample classification. Exploratory statistical regression analysis has been used to investigate relationships between the related parameters and eutrophication levels. Further, cluster analysis, fuzzy analysis, principal component analysis, and artificial neural networks have proven to be powerful tools in lake eutrophication assessment. Another important attempt at a multiparametric classification of trophic conditions was undertaken by Zurlini by combining the exact probabilities from Organization for Economic Cooperation and Development (OECD) of the log normal frequency distributions of chlorophyll, nitrogen, and phosphorus concentrations as well as Secchi depths. Additionally, several other researchers applied remote sensing technologies or geographic information system (GIS) technology to the trophic classification for lakes.

A Case Study: Trophic Classification for Lake Chaohu in China

Design and Calculation of Trophic Classification

A GIS-based method of lake eutrophication assessment was undertaken to study the spatial distribution of eutrophication conditions in lake environments. A trophic state index (TSI) consisting of six physical, chemical, and biological indicators including total phosphorus (TP), total nitrogen (TN), chemical oxygen demand (COD), Secchi disk depth (SD), chlorophyll-a concentration (Chl-a), and phytoplankton biomass (CA) was constructed to describe the eutrophication state of the lake environment. A 0–100 eutrophication scale was also developed to indicate seven different trophic levels within the lake environment: 0–30 representing oligotrophic, 30–40 lower-mesotrophic, 40–50 mesotrophic, 50–60 upper-mesotrophic, 60–70 eutrophic, 70–80 hypereutrophic, and 80–100 the extremely hypereutrophic. A representation of the spatial distribution of TSITP, TSITN, TSICOD, TSISD, TSIChl-a and TSICA was developed using the Inverse Distance Weighted (IDW) interpolation method. By

☆*Change History*: March 2018. Fu-Liu Xu updated Abstract, Keywords, Figure 1, Figure 3, and Figure 4.

This is an update of F.–L. Xu, Trophic Classification for Lakes, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 3594–3601.

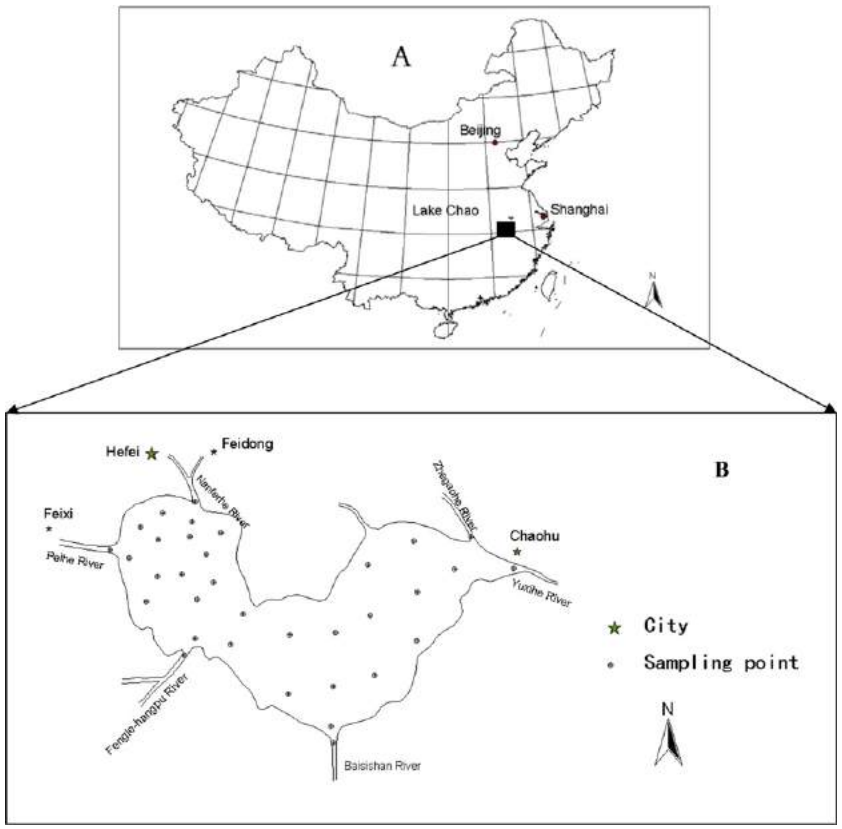


Fig. 1 Geographical location (A) and distribution of sampling points (B) of Lake Chao.

categorizing the interpolated values, a clear illustration of the different trophic levels was developed on six thematic maps. A Geographical Information System (GIS) overlay technique was applied to synthesize the information from the six thematic maps into a final map illustrating the spatial distribution of eutrophication conditions within the study area.

Study Area and Data Collection

Lake Chao is located in southeastern China (Fig. 1A). It has a mean surface area of 760 km^2, a mean depth of 3.06 m, and a mean retention time of 136 days. Lake Chao is one of the five largest fresh-water lakes in China, and the most eutrophic. Prior to the 1950s, the lake was well known for its scenic beauty and richness of aquatic life. Since that time, however, the lake has suffered from serious eutrophication. Increasing pressures from population growth and economic development in the drainage area are primarily responsible for the lake's current eutrophic state. The present conditions have had negative ecological, health, social, and economic effects on the lake and its utilization. Lake Chao was selected some time ago for inclusion in a nation-wide study of lake eutrophication conditions and processes. Accordingly, comprehensive research on Lake Chao has been carried out for more than 20 years (see Tu et al., 1990; Wang et al., 1995; Xu, 1994, 1996, 1997; Xu et al., 1999a,b,c).

Measurement and sampling of the lake's water for analytical purposes were performed monthly from April 1987 to March 1988. The water samples were collected using a van Dorhn sampler from 34 stations (Fig. 1B) at a depth of 0.5 m. The parameters chosen for measurement included both physic-chemical parameters (SD, pH, TN, TP, Si, COD, BOD, DO, etc.) and biological parameters (chlorophyll-a concentration (Chl-a), biomass concentration and the dry weight of both phytoplankton and zooplankton, and the number of phytoplankton cells). Thermal profiles were taken every 3 h at various points in the lake by means of a thermograph. A weather station located on the site provided 3-hourly measurements of atmospheric pressure, wind speed and direction, air temperature, humidity and solar radiation.

The flow chart of the GIS-based method for the lake eutrophication assessment is shown in Fig. 2.

TSI Calculation and Trophic Classification

Indicators selection

The TSIs was based on total phosphorus (TP, in mg/l of P), total nitrogen (TN, in mg/l of N), chemical oxygen demand (COD, in mg/l), Secchi disk depth (SD, in m), chlorophyll-a concentration (Chl-a, in mg/m^3), and phytoplankton biomass (CA, as C in mg/m^3) (see Table 1). The assessment standards for each indicator were based on those constructed for the Evaluation Standards

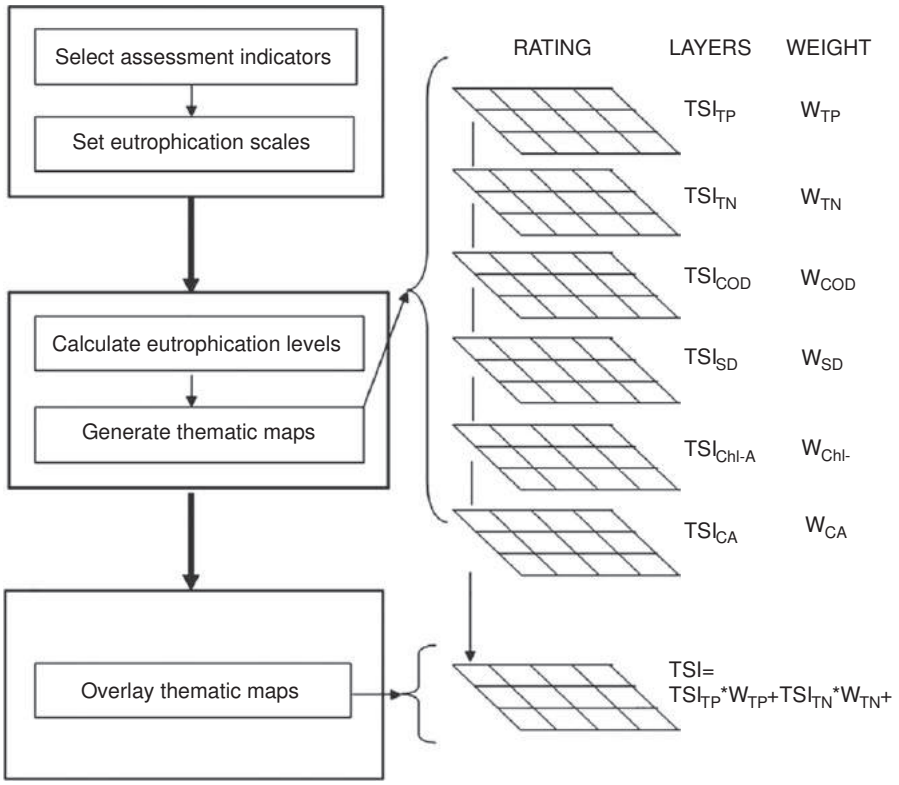


Fig. 2 The flow chart for the GIS-based method for lake eutrophication assessment.

Table 1 The scale of the trophic state index (TSI) and the evaluation standards for Lake Chao[a]

TSI	*Eutrophication level*	*TP (mg/L)*	*TN (mg/L)*	*COD (mg/L)*	*SD (m)*	*Chl-a (mg/m^3)*	*CA (mg/m^3)*
0	Oligotrophic	0.0004	0.010	0.06	48	0	/
10		0.0009	0.070	0.12	27	0.10	<50
20		0.0020	0.150	0.24	15	0.26	50
30		0.0046	0.300	0.48	8.0	0.66	100
40	Lower-mesotrophic	0.0100	0.600	0.96	4.4	1.60	150
50	Mesotrophic	0.0230	1.000	1.80	2.4	4.10	200
60	Upper-mesotrophic	0.0500	1.500	3.60	1.3	10.0	250
70	Eutrophic	0.1100	2.000	7.10	0.73	20.0	300
80	Hypereutrophic	0.2500	3.000	14.0	0.40	40.0	500
90	Extremely hypereutrophic	0.5500	4.600	27.0	0.22	100	800
100		1.2000	10.00	54.0	0.12	200	>800

[a]Consulting the trophic state index and evaluation standards for lake eutrophication of OECD (1982) and Japan National Environmental Institute (Aizaki et al., 1981), and for Lake Tai eutrophication in China (Jin et al., 1990).

for Lake Eutrophication designed for the OECD (1982) and the Japanese National Environmental Institute (Aizaki et al., 1981), and those used in assessing the eutrophication of Lake Tai in China (Jin et al., 1990) (see **Table 1**).

Calculation of eutrophication levels and generation of thematic maps

The following expression was used to calculate the lake eutrophication levels for each of the indicators:

$$\mathrm{TSI}_i = \left(\mathrm{TSI}_{k-1} + \left(\left(C_i - S_{i,k-1}\right)/\left(S_{i,k} - S_{i,k-1}\right)\right) \times \left(\mathrm{TSI}_k - \mathrm{TSI}_{k-1}\right)\right) \quad (1)$$

where C_i is the measured concentration of the i-th indicator (i = TP, TN, COD, SD, Chl-a and CA), TSI_k and TSI_{k-1} are the k-th and $(k-1)$-th *scales* of the i-th indicator, $S_{i,\,k}$ and $S_{i,\,k-1}$ are the evaluation standards of k-th and $(k-1)$-th scales of the i-th indicator (see **Table 1**).

The inverse distance weighted (IDW) interpolation method (Lam, 1983) with a spatial resolution of 500 × 500 m and ArcView Version 3.1 (ArcView 3.1, ESRI, Inc.) were used to generate the six thematic maps indicating the spatial distribution of eutrophication levels based on each indicator. The IDW interpolation method is based on the principle of assigning higher weights

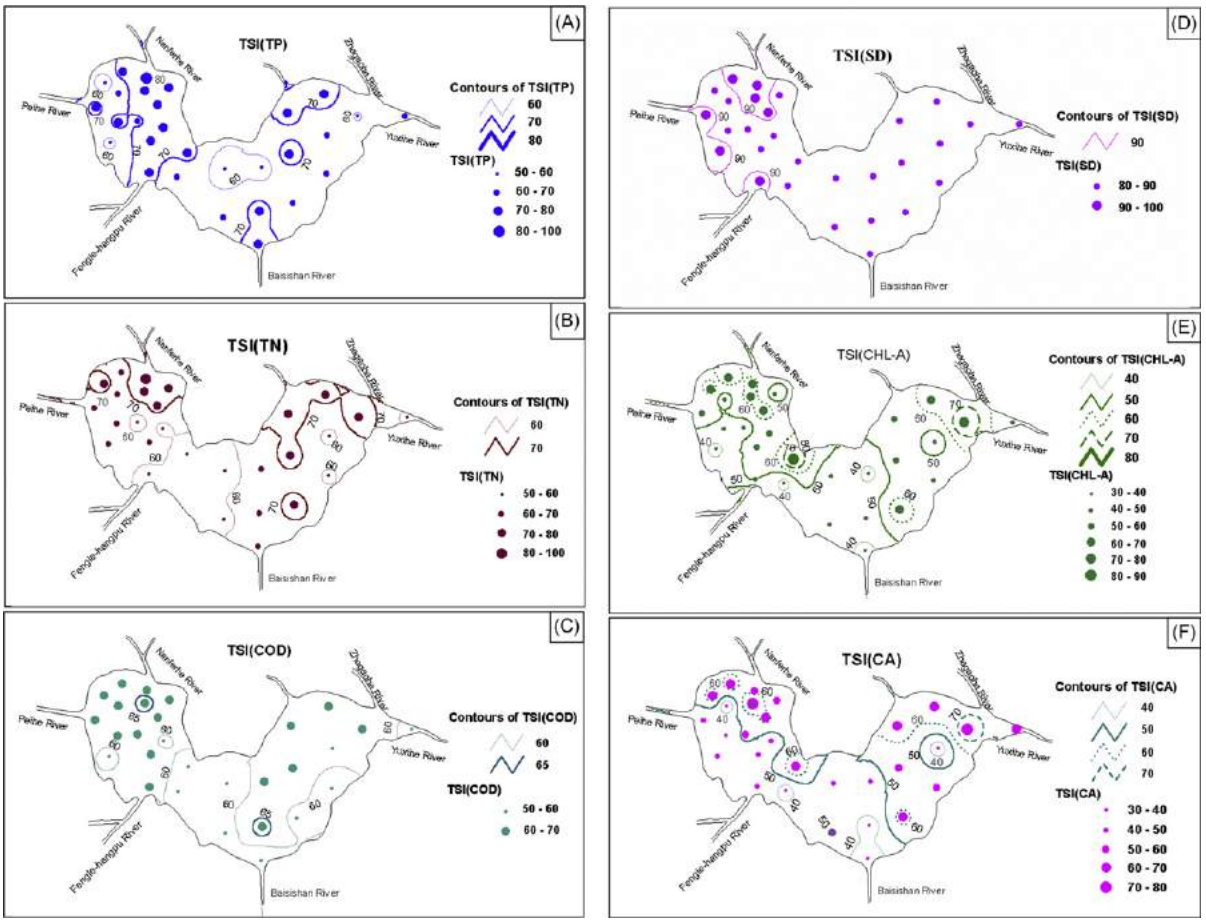


Fig. 3 Spatial distribution of the lake trophic state index based on each indicator: (A) TSI_{TP}; (B)TSI_{TN}; (C)TSI_{COD}; (D)TSI_{SD}; (E)$TSI_{Chl\text{-}a}$; (F) TSI_{CA}.

to data points closest to an unvisited point relative to those which are further away (Weber and Englund, 1992, 1994). In other words, the assigned weight is a function of inverse distance as represented in the following formula (Lam, 1983):

$$f(x,y) = \sum_{i=1}^{N} w(d_i) z_i \div \sum_{i=1}^{N} w(d_i) \tag{2}$$

where $f(x,y)$ is the interpolated value at point (x,y); $w(d_i)$ is the weighting function; z_i is the data value at point i; and d_i is the distance from point (x,y).

The interpolated values of any point within the dataset are bounded by $\min(z_i) < f(x,y) < \max(z_i)$, as long as $w(d_i) > 0$ (Lam, 1983). The IDW interpolation method has been widely used on many types of data because of its simplicity in principle, speed in calculation, easiness in programming, and credibility in interpolating surfaces (Lam, 1983).

The overlay of the thematic maps

The overlay technique, widely used in GIS applications (GIS by, 1994), was applied to synthesize the six thematic maps and develop the final eutrophication map. The following steps describe the synthesizing procedure used to analyze the six thematic maps:

(1) Development of an trophic state index (TSI) scale from 0 to 100 to label the different trophic levels (see **Table 1**).
(2) Application of this ordinal scale to all the pixels/cells on each thematic map; each pixel was assigned a value from 0 to 100 based on a comparison between its initial value and the eutrophication scales in **Table 1**.
(3) Analysis of the six thematic maps on a cell-by-cell basis. As a result, a final map illustrating the spatial distribution of eutrophication levels was produced. The following expression was used in the overlay operation to produce the TSI values:

$$TSI = (TSI_{TP}{}^{*}W_{TP} + TSI_{TN}{}^{*}W_{TN} + TSI_{COD}{}^{*}W_{COD} + TSI_{SD}{}^{*}W_{SD} + TSI_{Chl-A}{}^{*}W_{Chl-A} + TSI_{CA}{}^{*}W_{CA}) \tag{3}$$

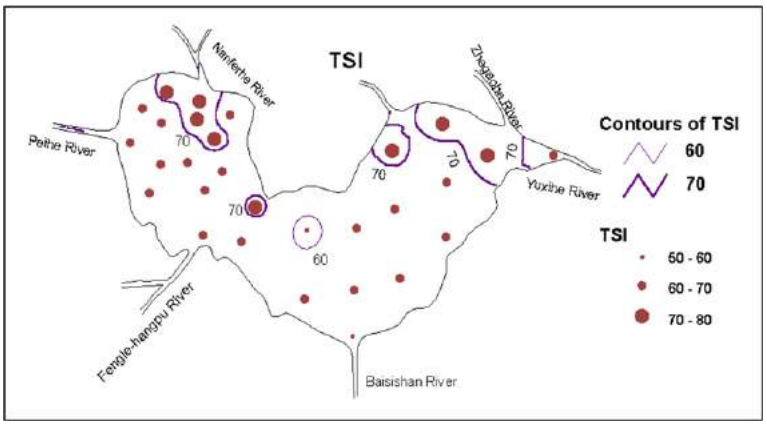


Fig. 4 Final map for the spatial distribution of the lake eutrophication based on overlay technique.

where TSI_{TP}, TSI_{TN}, TSI_{COD}, TSI_{SD}, $TSI_{Chl\text{-}a}$ and TSI_{CA} are the eutrophication levels for TP, TN, COD, SD, Chl-A and CA on the six thematic layers; W_{TP},W_{TN},W_{COD},W_{SD}, $W_{Chl\text{-}a}$ and W_{CA} are the weighting factors for each indicator (assumed as 1/6 for each indicator in the operation).

Trophic classification for Lake Chaohu

The thematic maps of TSI_{TP}, TSI_{TN}, TSI_{COD}, TSI_{SD}, $TSI_{CHL\text{-}A}$ and TSI_{CA} developed using the IDW interpolation method and the eutrophication scales are illustrated in Fig. 3A–F, respectively. The final TSI map developed as a result of the overlay technique is given in Fig. 4.

Fig. 3A shows the spatial distribution of TSI_{TP}. From it, one can see that the western part of Lake Chao is characterized mainly as severely eutrophic (TSI_{TP} 70–80), while the eutrophic field (TSI_{TP} 60–70) is distributed in the eastern part of the lake. The eutrophication levels near the river mouths are representative of eutrophic to extremely hypereutrophic (TSI_{TP} 60–90) conditions, while the upper-mesotrophic and eutrophic (TSI_{TP} 50–70) conditions can be observed in the open water areas.

Fig. 3B illustrates the spatial distribution of TSI_{TN}. The hypereutrophic field (TSI_{TN} 70–80) is mainly distributed in the northwestern and northeastern parts, near the river mouths. Most of the study area can be characterized as eutrophic (TSI_{TN} 60–70) centered in the southwestern and southeastern parts and upper-mesotrophic (TSI_{TN} 50–60) in the central part.

Fig. 3C shows the TSI_{COD} spatial distribution. Here, the upper-mesotrophic field (TSI_{COD} 50–60) is extended from the central to the southeastern part of the lake, with most of the remainder of the study area being characterized as eutrophic (TSI_{COD} 60–70).

Fig. 3D reveals the spatial distribution of TSI_{SD}. The extremely hypereutrophic field (TSI_{SD} 80–100) covers the entire lake. The eutrophication level near the river mouths in the western part of the lake is higher (TSI_{SD} 90–100) than that found elsewhere in the study area (TSI_{SD} 80–90).

Fig. 3E demonstrates the spatial distribution of $TSI_{CHL\text{-}A}$. The central and southwestern parts of the lake are characterized by mesotrophic and lower-mesotrophic fields ($TSI_{CHL\text{-}A}$ 30–50), with a strong mesotrophic field ($TSI_{CHL\text{-}A}$ 40–50) dominance. The northwestern and eastern parts of the study area are characterized as primarily upper-mesotrophic ($TSI_{CHL\text{-}A}$ 50–60), with the eutrophic field ($TSI_{CHL\text{-}A}$ 60–70) extending mainly near the river mouths in the same area. The severely eutrophic ($TSI_{CHL\text{-}A}$ 70–80) and the extremely hypereutrophic fields ($TSI_{CHL\text{-}A}$ 80–90) are limited to small areas in the central and eastern parts of the lake.

Fig. 3F indicates the spatial distribution of TSI_{CA}. The mesotrophic field (TSI_{CA} 40–50) extends from the southwestern to central parts of the lake, while the lower-mesotrophic field (TSI_{CA} 30–40) is limited to small areas in the same region. The upper-mesotrophic (TSI_{CA} 50–60) covers most of the northwestern and the eastern parts of the study area. The eutrophic range (TSI_{CA} 60–70) is found principally in the eastern part of the lake and in limited number of small areas in the northwestern part of the lake. The hypereutrophic field (TSI_{CA} 70–80) is only found near the mouth of the Zhegaohe River in the eastern part of the lake.

Finally, the overall TSI spatial distribution is illustrated in Fig. 4. Eutrophic conditions (TSI 60–70) cover most of the study area, while the hypereutrophic conditions (TSI 70–80) are distributed mainly near the river mouths in the northwestern and northeastern parts of the lake. A very limited area in the central part of the study area is characterized as upper-mesotrophic (TSI 50–60).

Discussions

The spatial distribution of Lake Chao's eutrophication levels derived from this study is closely correlated with the actual conditions of the lake. The northwestern and northeastern parts of the lake, especially the river mouths near the Nanfeihe and Zhegaohe Rivers, receive much more wastewater because of their nearness to Hefei City, the capital of Anhui Province, and Chaohu City, the second largest city in the lake's watershed (see Fig.1A). The nutrient contents which are primarily responsible for eutrophication (Rast and Holland, 1988; Ryding and Rast, 1989; Cooke et al., 1993) both in the lake's water and sediments, were far higher in

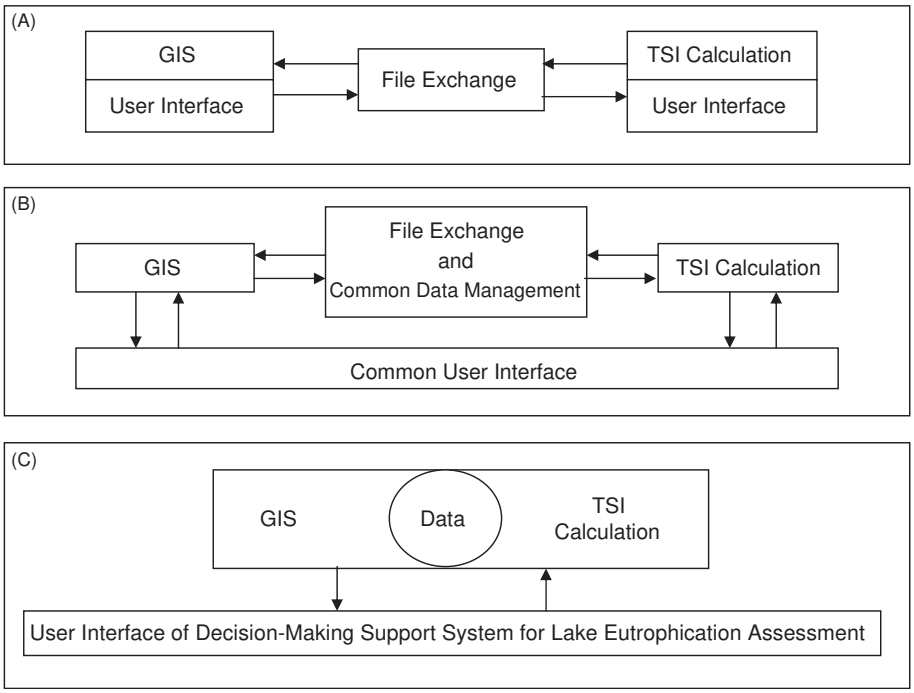


Fig. 5 The different periods or levels of GIS-based method for lake eutrophication assessment: (A) the first periods or levels; (B) the second periods or levels; (C) the third periods or levels.

these two regions than anywhere else in the lake (see **Fig. 3**A,B). The concentration of Chl-a and phytoplankton biomass, two of the more obvious symptoms of eutrophication (Rast and Holland, 1988; Ryding and Rast, 1989; Cooke et al., 1993), both followed similar trends in spatial distribution (see **Fig. 3**E–F).

Lake eutrophication, however, cannot be simply evaluated by a single physical, chemical, or biological parameter. It is a multidimensional feature (Shannon and Brezonik, 1972; Carlson, 1977; Cruzado, 1987). These single indicators cover different aspects of the lake eutrophication phenomenon. It is necessary, therefore, to apply several indicators simultaneously (including physical, chemical, and biological) to derive a more complete lake eutrophication assessment. Only through such a multi-dimensional approach can one capture all the features needed to yield a fully informative assessment of the eutrophic condition of a lake. Unfortunately, the sensitivity of a single-parameter and/or its weighting factor may be different in different lakes depending on the parameters used in the TSI (e.g., Therriault and Platt, 1978; Reckhow and Chapra, 1983; Powell et al., 1989; Boyle et al., 1990; Whitton and Kelly, 1995; Karydis and Tsirtsis, 1996; Danilov and Ekelund, 1999). The choice of suitable indicators and their weights is critically important in the eutrophication assessment of a specific lake. Hooper (1969) identified four important criteria useful in the development of eutrophication indices: (1) it should discriminate between changes associated with nutrient level and those associated with other categories of environmental change; (2) it should have considerable sensitivity to levels of enrichment; (3) it should have properties which are widespread geographically and short-lived geologically; and (4), it should be suitable for long-term surveillance and monitoring, that is, indices should both document past changes and serve as a predictive function.

Most of the current research has focused on the identification and selection of suitable indicators relative to assessing the eutrophic condition of a lake. The spatial distribution of eutrophication levels, however, is equally important in developing a complete picture of the trophic state of a lake. A GIS-based method to spatially assess lake eutrophication states was proposed in this paper. GIS technology is mainly adequate for distributed data which are within the resolution of the GIS grid, for example, land use, land cover (both can be deduced from sattelite inages), DEM, etc. The approach in this paper is not the same—the authors used rather limited number of observation points and a special technique (IDW) to calculate the GIS mesh values of variables. The methodology could have three different approaches depending on integration style and date exchange methodology between the GIS approach and the TSI calculations (see **Fig. 5**). The first is shown in **Fig. 5**A. Here, both the GIS and the TSI calculations belong to two separate systems, each having their own identifiable user interface. The GIS is used simply to display the results of the TSI calculations and their integration is limited to file exchange activity. The second approach is illustrated in **Fig. 5**B. Here, the GIS and the TSI calculations still belong to two separate systems. However, they possess a common user interface that is used to activate the GIS and TSI calculations as a single system, in addition to managing common data and file exchanges. This integrating style reduces errors associated with both data and file exchanges between separate systems. The third approach is presented in **Fig. 5**C. The GIS and the TSI calculations are now completely integrated as a Decision-Making Support System for Lake Eutrophication Assessment (DMSSLEA), having a uniform user interface and data sharing. In this approach, the TSI calculations are treated as an analysis function of the integrating system. The calculations are programmed using a specific GIS

programming language, for example, Avenue in ArcView, SML in Arc/Info PC version, MapBasic in MapInfo, or Genius in GanaMap. Here the GIS manages the spatial and attribute data, in addition to manipulating and displaying the results of TSI calculations.

The first approach was applied in the case study presented in this paper. The TSI calculations were developed using MS Excel 97, while the generation and synthesis of the six thematic maps was performed using the inverse distance weighted (IDW) interpolation method and overlay technique within the framework of a Geographic Information System (ArcView 3.1, ESRI, Inc.). The results of the TSI calculations were stored in Dbase format. The map indicating the boundary and sampling points in the lake served as the basis for visualizing the resulting spatial data (**Fig. 1**B).

Conclusions

A GIS-based method of lake eutrophication assessment was proposed with the purpose of studying the spatial distribution of eutrophication conditions in the lake environment. The method included the integration of GIS methodology into lake eutrophication assessment using a trophic state index. The inverse distance weighted (IDW) interpolation method was used for generating thematic maps indicating the spatial distribution of each of the Trophic State Indices. An overlay technique within the GIS framework was applied to analyze the information from the thematic maps in order to develop a final map illustrating the spatial distribution of eutrophication conditions in the lake. Results from the study indicate that the boundaries associated with different trophic levels (upper-mesotrophic, eutrophic and hypereutrophic) could be clearly defined in a final eutrophication map. This result comprises the principal advantage of the proposed methodology when compared with other attempts based on a multiparametric classification and assessment of lake trophic trends. The latter approach does not lead to a clear definition of the boundaries associated with differing trophic levels. As such, the proposed methodology could be of special interest to policy makers involved in lake management and planning, since policy makers need a more explicit view of trophic status and clear information on water quality.

See also: Ecological Data Analysis and Modelling: Climate Change Models. Ecosystems: Freshwater Lakes

References

Aizaki, M., Iwakuma, T., Takamura, N., 1981. Application of modified Carlson's trophic state index to Japanese lakes and its relationship to other parameters related to trophic state. Research Report. National Institute for Environmental Studies. Japan 23, 13–31.

Boyle, T.P., Smillie, G.M., Anderson, J.D., Beeson, D.R., 1990. A sensitivity analysis of nine diversity and seven similarity indices. Research Journal of Water Pollution Control Federation 62, 749–762.

Carlson, R.E., 1977. A trophic state index for lakes. Limnology and Oceanography 22 (2), 361–369.

Cooke, D.G., Welch, E.B., Peterson, S.A., Newroth, P.R., 1993. Restoration and Management of Lakes and Reservoirs, 2nd edn. Boca Raton, FL: Lewis Publisher, p. p. 548.

Cruzado, A., 1987. Eutrophication in the pelagic environment and its assessment. In eutrophication in the Mediterranean Sea: Receiving capacity and monitoring of long term effects. UNESCO Reports in Marine Science 49, 57–66.

Danilov, R., Ekelund, N.G.A., 1999. The efficiency of seven diversity and one similarity indices based on phytoplankton data for assessing the level of eutrophication in lakes in Central Sweden. Science of The science of the total environment 234 (1–3), 15–23.

GIS by ESRI, , 1994. Cell-based modeling with grid. California, USA: Environmental Systems Research Institute Inc.

Hooper, F.F., 1969. Eutrophication indices and their relation to other indices of ecosystem change. In: Eutrophication: causes, consequences, correctives. Washington, D.C.: USA National Academy of Sciences, pp. 225–235.

Jin, X., Liu, H., Tu, Q., Zhang, Z., Zhu, X., 1990. Eutrophication of lakes in China. Beijing: China Environmental Sciences Press, 614 pp.

Karydis, M., Tsirtsis, G., 1996. Ecological indices: A biometric approach for assessing eutrophication levels in the marine environment. The science of the total environment 186, 209–219.

Lam, N.S.N., 1983. Spatial interpolation methods: A review. The American Cartographer 10 (2), 129–149.

OECD, , 1982. Eutrophication of waters. Monitoring, assessment and control Paris: OECD.

Powell, T.M., Cloern, J.E., Huzzey, L.M., 1989. Spatial and temporal variability in South San Francisco Bay USA. I. Horizontal distributions of salinity suspended sediments and phytoplankton biomass and productivity. Estuarine, coastal and Shelf Science 28, 583–597.

Rast, W., Holland, M., 1988. Eutrophication of lakes and reservoirs: A framework for making management decisions. Ambio 17, 2–12.

Reckhow, K.H., Chapra, S.C., 1983. Engineering approaches for Lake management, volume 1: Dara analysis and empirical modeling. Boston\London\Sydney\Wellington\Durban \Toronto: Butterworth Publishers, p. p. 340.

Ryding, S.O., Rast, W., 1989. The control of eutrophication of lakes and reservoirs. In: Man and the biosphere series, 1. Paris: UNESCO, p. p. 256.

Shannon, E.E., Brezonik, P.L., 1972. Eutrophication analysis: A multivariate approach. Journal Sanitary Engineering Division ASCE 98 (1), 37–57.

Therriault, J.C., Platt, T., 1978. Spatial heterogeneity of phytoplankton biomass and related factors in the near-surface waters of an exposed coastal environment. Limnology and Oceanography 23, 888–899.

Tu, Q.Y., Gu, D.X., Yi, C.Q., Xu, Z.R., Han, G.Z., 1990. The researches on the Lake Chao eutrophication. Hefei, China: The publisher of University of Science and Technology of China, p. p. 225. (in Chinese).

Wang, S.Y., Jin, C.S., Meng, R.X., Xu, F.L., 1995. Environmental Research for Lake Chao in Anhui Province. In: Jin, X.C. (Ed.), Lakes in China (1). Beijing: China Ocean Press, p. 580.

Weber, D., Englund, E., 1992. Evaluation and comparison of spatial interpolators. Mathematical Geology 24 (4), 381–391.

Weber, D., Englund, E., 1994. Evaluation and comparison of spatial. Interpolators II. Mathematical Geology 26 (5), 589–603.

Whitton, B., Kelly, M., 1995. Use of algae and other plants for monitoring rivers. Australian Journal of Ecology 20, 45–56.

Xu, F.-L., 1994. Scientific decision-making system for environmental Management of the Lake Chao Watershed. Environmental Protection 21 (5), 36–39.

Xu, F.-L., 1996. Ecosystem health assessment of Lake Chao, a shallow eutrophic Chinese lake. Lakes & Reservoirs: Research and Management 2, 101–109.
Xu, F.-L., 1997. Exergy and structural exergy as ecological indicators for the development state of the Lake Chao ecosystem. Ecological Modelling 99, 41–49.
Xu, F.-L., Jorgensen, S.E., Tao, S., 1999a. Ecological indicators for assessing freshwater ecosystem health. Ecological Modelling 116, 77–106.
Xu, F.-L., Jorgensen, S.E., Tao, S., Li, P.G., 1999b. Modeling the effects of macrophyte restoration on water quality and ecosystem of Lake Chao. Ecological Modelling 117, 239–260.
Xu, F.-L., Tao, S., Xu, Z.R., 1999c. The restoration of wetlands and macrophytes in the Lake Chao: Possibility and effects. Hydrobiologica 405, 169–178.

Trophic Index and Efficiency☆

Timur Pavluk, Russian Research Institute for Integrated Water Management and Protection, Ekaterinburg, Russia
Abraham bij de Vaate, Waterfauna Hydrobiological Consultancy, Lelystad, The Netherlands

Introduction

This article describes the common scientific meanings of the widely used "trophic index" in classifications in the field of monitoring and assessment. The term "trophic" does not have a distinct application in ecology and is generally used in the description of mean processes like feeding, nourishment, production potential, and food web.

The word "trophy" originates from the Greek word *trophē* which means nourishment or pertaining to nutrition or connected with a source of nutrition. Many other ecological terms are derivatives of the initial word "trophy": for example, trophic level, trophic niche, trophic guild, trophic net, trophic structure, trophic status, and trophic index. In order to classify a plant or animal community or to assess its quality status, for example, as a result of anthropogenic disturbance, trophic indices were developed. They are mostly applied in aquatic communities since aquatic ecosystems are relatively stable in space and time.

Two groups of trophic indices can be distinguished. The first one, the group of trophic status indices, focuses on the primary production potential. In the aquatic environment four types of trophic statuses of the waterbody can be distinguished: the oligotrophic, mesotrophic, eutrophic, and dystrophic status. The second group includes those indices that reflect the complexity of trophic relations between organisms. These indices have been commonly used in bioassessment of the aquatic ecosystem health in general—especially, those systems that have been severely altered as result of anthropogenic activities (e.g., water pollution, physical disturbances). Results of assessments are commonly ranged into quality classes. In the forthcoming sections main types of trophic indices are discussed, mostly used in assessment procedures to evaluate the ecological status of a defined biotope.

Trophic Status Indices

TRIX Index

This index represents the linear combination of the logarithm of four state variables:

$$\mathrm{TRIX} = (\mathrm{Log}_{10}[\mathrm{ChA} + \mathrm{aD\%O} + \mathrm{minN} + \mathrm{TP}] + \mathrm{k})/\mathrm{m}$$

where ChA is the chlorophyll-*a* concentration ($\mu g\,L^{-1}$), aD%O is the dissolved oxygen concentration as absolute percentage deviation from saturation ($=100\%$), minN is the mineral nitrogen, dissolved inorganic nitrogen (sum of $N_{nitrate} + N_{nitrite} + N_{ammonia}$ in $\mu g\,L^{-1}$), and TP is total phosphorus ($\mu g\,L^{-1}$).

The coefficients $k=1.5$ and $m=1.2$ are scale coefficients, introduced to obtain results on a 0–10 scale. TRIX results are arranged in four classes according to **Table 1**. The index characterizes succinctly the trophic levels in coastal marine areas and was adopted for this purpose by the Italian national legislation.

Values exceeding 6 TRIX units are typical for highly productive coastal waters, where eutrophication effects determine frequency of anoxia episodes in the water layer above the bottom. Values lower than 4 TRIX units are associated with scarcely productive coastal waters, while values lower than 3 are usually found in the open sea. Because of the log transformation of the four variables used, annual distributions of TRIX data over homogeneous coastal zones follow or are nearly equal to a Gauss curve.

Carlson's Trophic State Index

Transparency of surface waters is often related to the amount of plant nutrients in the water. The more the nutrients, the more the phytoplankton, and as a result the less transparent the water is. Measuring transparency is a common but indirect way to estimate roughly the trophic condition of a waterbody. This condition, called the extent of eutrophication, is a natural aging process of lakes, which is unnaturally accelerated by too many nutrients. Trophic state determination is an important aspect of lake surveys. Trophic state is not the same as water quality, but is one aspect of it.

The concept of trophic status is based on the fact that changes in nutrient levels (measured as total phosphorus) cause changes in algal biomass (measured as chlorophyll *a*) which in turn cause changes in lake clarity (measured as Secchi disk transparency). The trophic state index (TSI) is a convenient way to quantify this relationship. TSI is calculated independently from Secchi disk depth, chlorophyll *a*, and total phosphorus concentration. It should be taken into account that TSI was

☆*Change History*: March 2017. T Pavluk updated Marine Trophic Index section, References and Relevant Websites.

Table 1 Categories of TRIX classes

TRIX value	*Trophic category*
<4	Low trophic level
4–5	Middle trophic level
5–6	High trophic level
6–10	Very high trophic level

Adapted from Moncheva, S., Doncheva, V., 2000. Eutrophication index ((E) TRIX)—an operational tool for the Black Sea coastal water ecological quality assessment and monitoring. International Symposium "The Black Sea Ecological Problems" Odessa: SCSEIO, pp. 178–185.

Table 2 Classes of TSI values and their ecological attributes.

TSI	*Chl* ($\mu g\ L^{-1}$)	*SD (m)*	*TP* ($\mu g\ L^{-1}$)	*Ecological attributes*
<30	<0.95	>8	<6	Oligotrophy: Clear water, oxygen throughout the year in the entire hypolimnion
30–40	0.95–2.6	8–4	6–12	Hypolimnia of shallower lakes may become anoxic
40–50	2.6–7.3	4–2	12–24	Mesotrophy: Water moderately clear; increasing probability of hypolimnetic anoxia during summer
50–60	7.3–20	2–1	24–48	Eutrophy: Anoxic hypolimnia, macrophyte problems possible
60–70	20–56	0.5–1	48–96	Blue-green algae dominate, algal scums and macrophyte problems
70–80	56–155	0.25–0.5	96–192	Hypereutrophy (light-limited productivity): Dense algae and macrophytes, algal blooms possible throughout summer
>80	>155	<0.25	192–384	Algal scums, few macrophytes

Adapted from Carlson, R.E., Simpson, J., 1996. A coordinator's guide to volunteer lake monitoring methods. USA: North American Lake Management Society Madison.

developed for use with lakes that have few rooted aquatic plants and little nonalgal turbidity. The formulas for calculating the TSI are presented below:

$$\mathrm{TSI\ (SD)} = 60 - 14.41\ \mathrm{Ln\ Secchi\ disk\ depth(meters)}$$
$$\mathrm{TSI\ (Chl)} = 9.81\ \mathrm{Ln\ chlorophyll}\ a(\mu \mathrm{g\,L^{-1}}) + 30.6$$
$$\mathrm{TSI\ (TP)} = 14.42\ \mathrm{Ln\ total\ phosphorus}\ (\mu \mathrm{g\,L^{-1}}) + 4.15$$

Each of these three variables can theoretically be used to classify a waterbody, because they are interrelated by linear regression. It is supposed that seasonal average values of variables are used for TSI calculation. If the three TSI values are not similar to each other, it is likely that algal growth may be light- or nitrogen-limited instead of P-limited, or that Secchi disk transparency is affected by erosional silt particles rather than by algae, or something else.

One considers that average TSI is a good indicator of water trophic status in general. Average $\mathrm{TSI} = (\mathrm{TSI(TP)} + \mathrm{TSI(Chl)} + \mathrm{TSI(SD)})/3$ and the values reflect ecological attributes that could be expected in temperate lakes (**Table 2**).

Trophic Diatom Indices

Diatoms are widely recognized and used as indicators of river and stream water quality, including trophic state conditions. The diatom trophic indices describe diatom distribution in relationship to either "dissolved" (~orthophosphate) or "total" phosphorus, that are mostly closely correlated with the nitrogen concentration in the water. Therefore, these indices are treated as broad indicators of the trophic status of water bodies.

Three trophic indices based on ecological properties of diatoms were developed in the 1990s:

1. The trophic diatom index (UK TDI), which is used in the United Kingdom. This index was developed by Kelly and Whitton (1995), later on revised by Kelly *et al.* in 2001.
2. The TDI developed by Coring *et al.* (1999), which is used in Germany (German TDI).
3. The TDI developed by Rott *et al.* (1999), which is used in Austria (Austrian TDI).

All these indices were designed for use in rivers, which means that they are not applicable in lakes and other stagnant water bodies. Calculation of the indices is based on the weighted average of the equation of Zelenka and Marvan for each diatom species. Two values are assigned to the species: a value reflecting the tolerance or affinity to a certain water quality (good or bad) and a value that indicates how strong (or weak) this relationship is. In addition, the index values are weighted by the relative abundance of the diatom in the sample (percentage of a particular diatom species in the sample):

$$\text{Index} = \frac{\sum_{j=1}^{n} a_j^* s_j^* v_j}{\sum_{j=1}^{n} a_j^* v_j}$$

where a_j is the abundance or proportion of species j in samples, s_j is the pollution sensitivity of species j (optimum), and v_j is the indicator value (tolerance).

Sensitivity value (s) is divided into five classes and varies from 1 (highly sensitive to phosphorous) to 5 (highly tolerant to phosphorous). Tolerance value (v) ranges from 1 (taxa with a broad distribution) to 3 (taxa that are restricted to a narrow range of nutrient conditions).

All TDIs mentioned are very similar in principal and differ in the number of species used and in the values of s and v which have been attributed to the species after compiling the data from literature and from ordinations. Also the number of the trophic classes distinguished may vary.

Benthic Trophic State Index

The ratio of gross production to respiration (P:R), measured quantitatively by oxygen flux rates, has long been used to characterize the trophic status of aquatic ecosystems. A ratio of 1.0 indicates a balance of photoautotrophic and heterotrophic processes, ratios <1.0 indicate heterotrophic dominance of processes, and ratios >1.0 indicate photoautotrophically dominated communities.

The benthic TSI is based on relative rates of sediment–water oxygen exchange in opaque (dark) and transparent (light) chambers incubated at or near ambient temperature.

The benthic TSI (BTSI) uses data from short-term metabolic measurements not as a quantitative measure of shoal community P:R, a continuous variable, but to assign a categorical value that broadly reflects the extent to which that environment supports the ecological processes associated with benthic autotrophy.

The BTSI, expressed in the values 0, 1, 2, or 3, is assigned to shallow sediments based on the relative magnitude of hourly rates of maximum net community production (NCP_{max}) and respiration (CR).

In calculating the BTSI, negative values are assigned to all oxygen fluxes into the sediment. When the BTSI increases, so does the degree of autotrophy, and thus the contribution to support grazing organisms, hypoxia reduction, nutrient sequestration, and biotic stability of the sediment. Ranges of gross production/respiration (P:R) associated with each BTSI value are shown in **Table 3** including examples of rates representative of each BTSI value (condition) and corresponding net community production at light saturation (NCP_{max}) and community respiration (CR).

Ranges of gross production/respiration (P:R) associated with each BTSI value are given including examples of rates representative of each BTSI value (condition) and corresponding net community production at light saturation (NCP_{max}) and community respiration (CR), both in mg O_2 m^{-2} h^{-1}. Negative values indicate fluxes from water to sediments. NA, not applicable.

The BTSI is a method to assess the functioning of shallow benthic ecosystems. It is methodologically relatively simple assessment method and reflects established ecological processes as the result of anthropogenic pressure. The BTSI relates potential benthic photoautotrophy to benthic respiration through discrete classification rather than the commonly used continuous variable of P:R ratio.

Oligochaete Trophic Index

The association of oligochaetes with organic enrichment of water was used to develop the "oligochaete trophic index." The index is based on the oligochaete community structure, where species were assigned to categories depending on their preference for, or tolerance of, oligotrophic, mesotrophic, or eutrophic conditions.

A number of modifications of the oligochaete trophic index have been developed since 1977. A modification index used in the Great Lakes (North America) is calculated as

$$c\frac{1/2 * \sum n0 + \sum n1 + 2\sum n2 + 3\sum n3}{\sum n0 + \sum n1 + \sum n2 + \sum n3}$$

where $n0$, $n1$, $n2$, and $n3$ are the total numbers of individuals belonging to each of the four ecological groups. Species characteristic

Table 3 Classification of shallow sediments by the BTSI.

BTSI	*P:R*	*Condition*	*NCP_{max}*	*CR*	*Sediment classification*
0	NA	$NCP_{max} \leq CR$	−25	−25	Totally heterotrophic
1	>0–0.5	$CR < NCP_{max} \leq 0$	−10	−25	Net heterotrophic
2	0.5–1.0	$0 < NCP_{max} \leq \lvert CR \rvert$	2	−25	Net autotrophic
3	>1.0	$\lvert CP \rvert < NCP_{max}$	50	−25	Highly autotrophic

Adapted from Rizzo, W.M., Berry, B.E., Wetzel, R.L., et al., 1996. A metabolism-based trophic index for comparing the ecological values of shallow-water sediment habitats. Estuaries 19, 247–256, with permission from estuarine Research Federation.

for oligotrophic waters are assigned to group 0, those for mesotrophic waters to group 1, those for eutrophic waters to group 2, and those for hypertrophic waters (*Limnodrillus hoffmeisteri* and *Tubifex tubifex*) comprise group 3.

The coefficient c depends upon the total oligochaete number per square meter as outlined as follows:

$c = 1,\ n > 3600;$
$c = 3/4,\ 1200 < n < 3600;$
$c = 1/2,\ 400 < n < 1200;$
$c = 1/4,\ 130 < n < 400;$
$c = 0,\ 0 < n < 130.$

Index values between 0.6 and 1.0 indicate mesotrophic conditions, while higher and lower values indicate eutrophic and oligotrophic conditions, respectively.

In general, when the index was applied to the Great Lakes, it appeared that the values give a reasonable evaluation of trophic conditions. Most sites in the upper lakes fall into the oligotrophic category, with areas of known higher productivity (near shore northern Lake Michigan; Saginaw Bay, Lake Huron) exhibit higher index values. Sites in Lake Erie generally fall into the mesotrophic range, while in Lake Ontario near shore sites were classified as mesotrophic, and offshore sites are oligotrophic.

Trophic Level Index

The trophic level index (TLI) is an indicator of lake water quality. Four parameters are combined to construct the TLI: concentrations of total nitrogen, total phosphorus, and chlorophyll *a*, and transparency. These parameters reflect the dynamics of the annual lake cycle. Nitrogen and phosphorus are essential plant nutrients. High levels of water-bound nitrogen and phosphorus most often come from agricultural runoff and urban wastewater, but can also come from geothermal inputs and deep springs that leach phosphorus from the rock geology.

Chlorophyll *a* is a good indicator of the total quantity of algae in a lake. Algae are a natural part of any lake system, but large amounts of algae decrease water clarity, make the water look green, can form surface scums, reduce dissolved oxygen levels, can alter pH levels, and can produce unpleasant tastes and smells. Transparency of the water is measured using a Secchi disk.

Calculation of the TLI:

- $TL_{nitrogen} = -3.61 + 3.01 \log (N_{total})$
- $TL_{phosphorus} = 0.218 + 2.92 \log (P_{total})$
- $TL_{transparency} = 5.10 + 2.27 \log (1/\text{Secchi disk depth} - 1/40)$
- $TL_{chlorophyl} = 2.22 + 2.54 \log (\text{Chl } a)$
- $TLI = (TL_{nitrogen} + TL_{phosphorus} + TL_{transparency} + TL_{chlorophyl})/4$

The higher the TLI, the worse the water quality. Trophic level ranges are grouped into trophic states for quantitative description as shown in **Table 4**.

Trophic states, as determined by the four key variables:

1. Microtrophic lakes are very clean, and often have snow or glacial sources.
2. Oligotrophic lakes are clear and blue, with low levels of nutrients and algae.
3. Mesotrophic lakes have moderate levels of nutrients and algae.
4. Eutrophic lakes are green and murky, with higher amounts of nutrients and algae.
5. Supertrophic lakes are fertile and saturated in phosphorus and nitrogen, and have very high algae growth and blooms during calm sunny periods.
6. Hypertrophic lakes are highly fertile and supersaturated in phosphorus and nitrogen. They are rarely suitable for recreation, and habitat for desirable aquatic species is limited.

Table 4 Trophic state and corresponding quantitative parameters of the trophic level index

Trophic state	*Nutrient enrichment category*	*TLI*	*Chl* a *(mg m^3)*	*Secchi disk depth (m)*	*$T_{phosphorus}$ (mg m^3)*	*$T_{nitrogen}$ (mg m^3)*
Ultramicrotrophic	Practically pure	0.0–1.0	<0.33	>25	<1.8	<34
Microtrophic	Very low	1.0–2.0	0.33–0.82	25–15	1.8–4.1	34–73
Oligotrophic	Low	2.0–3.0	0.82–2.0	15–7	4.1–9.0	73–157
Mesotrophic	Medium	3.0–4.0	2–5	7.0–2.8	9–20	157–337
Eutrophic	High	4.0–5.0	5–12	2.8–1.1	20–43	337–725
Surpertrophic	Very high	5.0–6.0	12–31	1.1–0.4	43–96	725–1558
Hypertrophic	Saturated	>6.0	>31	<0.4	>96	>1558

Adapted from Environment Bay of Plenty, Tropic Level Index, http://www.ebop.govt.nz/Water/Lakes/Trophic-Level-Index.asp, with permission.

Trophic Index of Macrophytes

The trophic index of macrophytes (TIM) is a tool for indicating the trophic state of running waters. Concentrations of soluble reactive phosphorus in both the water body and sediment pore water were assigned to macrophyte species and related to their phosphorus demand. The TIM is calculated with Zelenka and Marvan's formula for the determination of the saprobic index:

$$TIM = \frac{\sum_{i=1}^{n} IV_a{}^{*} W_a{}^{*} Q_a}{\sum_{i=1}^{n} W_a{}^{*} Q_a}$$

where IV_a is the indicator value for species a, depending on the assigned trophic category, W_a is the weighting factor for the tolerance of species a, and Q_a is the quantity of species a in the river section.

TIM values were classified into trophic categories to indicate the trophic status of the river sections examined (Table 5).

Trophic Relation Indices

Infaunal Trophic Index

The infaunal trophic index (ITI) was developed to identify changed and degraded environmental conditions as a result of organic pollution in coastal waters. The approach is based on the allocation of macroinvertebrate species to one of four groups based on the type of food consumed by the species and where the food was obtained from.

The ITI and its response to organic pollution is based on the principle that with increasing organic enrichment the dominant feeding type changes from those species which feed at the interface of the sediment and water, such as suspension feeders (which occur in areas of low organic enrichment), through to species which are predominantly deposit feeders (in areas of high organic enrichment).

After determining the abundance of taxa belonging to each of next four feeding groups, (1) detrital feeders, (2) interface detrital feeders, (3) deposit feeders, and (4) specialized feeders, the ITI is calculated by combining these groups in the following formula:

$$ITI = 100 - 33.33\frac{(0n_1 + 1n_2 + 2n_3 + 3n_4)}{(n_1 + n_2 + n_3 + n_4)}$$

where n_{1-4} are the numbers of individuals in feeding groups 1–4 distinguished, and 0–3 are factors in the numerator (scaling factors).

Values of the index range from 0 to 100 with low values indicating degraded conditions. With the ITI seabed areas are classified into either "normal" (values 100–60), "changed" (60–30), or "degraded" (30–0).

The ITI has been tested widely in fully marine conditions and showed most statistically powerful results. However, due to the natural prevalence of deposit feeders in estuaries and the generally lower number of taxa, the ITI can act in an aberrant way in transient areas and is therefore not a useful indicator for estuaries. The index has a direct link to eutrophication indicators.

Marine Trophic Index

Differences in the trophic level of selected groups of species provide a reliable indicator of the integrity of an ecosystem. The marine trophic index indicates changes in the mean trophic level of fish communities regionally and globally. Trophic level is

Table 5 Classification of TIM values into trophic categories

TIM value	*Trophic state*
1.00–1.44	Oligotrophic
1.45–1.86	Oligomesotrophic
1.87–2.24	Mesotrophic
2.25–2.62	Mesoeutrophic
2.63–3.04	Eutrophic
3.05–3.49	Eupolytrophic
3.50–4.00	Polytrophic

Adapted from Schneider, S., Melzer, A., 2003. The trophic index of macrophytes (TIM)—a new tool for indicating the trophic state of running waters. International Review of Hydrobiology 88, 49–67, with permission from Wiley-VCH, STM.

defined as the position of an organism in the food chain and ranges from a value of 1 for primary producers to 5 for marine mammals and humans. The method to determine the trophic level of a consumer is to add one level to the mean trophic level of its prey.

The equation corresponding to a species trophic level calculation is

$$TL_i = \sum_j TL_j{}^* DC_{ij}$$

where TL_j is the fractional trophic level of the prey j, and DC_{ij} is the fraction of j in the diet of species i.

Thus defined, the trophic level of most fishes and other aquatic consumers can have any value between 2.0 and 5.0. Trophic level changes through the life history of fish, with juveniles having lower trophic levels than adults.

Existent annual fishery database supplies sufficient information for marine trophic index computing. Therefore, mean trophic level for year k may be found as

$$\overline{TL}_k = \frac{\sum_i (TL_i)^* (Y_{ik})}{\sum_i Y_{ik}}$$

where Y_i is the landing (catch) of species (group) i, and TL_i is the trophic level of species (group) i.

Trophic level estimates for fish, based on their diet composition, can be found in FishBase, the global online database for fish, and for invertebrates in the Sea-Around-Us database.

The marine trophic index is a powerful indicator of marine ecosystem integrity and sustainability of fisheries at the global and regional levels.

Whether a fishery is balanced from ecological point of view we may define with a "Fisheries in Balance" index (FIB) (Pauly *et al.*, 2000). Since biomass transfer efficiencies between trophic levels are only about 10%, it follows that the rate of biological production is much greater at lower than it is at higher trophic levels. Fisheries catches, at least to begin with, will tend to increase as the trophic level declines. At this point the fisheries will target species lower in the food web.

The FIB index is defined, for any year i, by

$$FIB = \log\left(Yi * \left(\frac{1}{TE}\right)^{TLi}\right) - \log\left(Yo * \left(\frac{1}{TE}\right)^{TLo}\right)$$

where Y_i is the catch at year i, TL_i is the mean trophic level of the catch at year i, Y_o is the catch, TL_o the mean trophic level of the catch at the start (o refers to any year used as a baseline) of the series being analyzed, and TE is the transfer efficiency of biomass or energy between trophic levels.

The FIB index is stable (zero) over periods of time when changes in trophic levels are matched by appropriate changes in the catch in the opposite direction. The index increases if both catches and mean trophic level increase for any reason, for example higher fish biomass, or geographic expansion, suggesting that the fishery was expanding to stocks previously not, or lightly exploited. Decreases may be observed when TL shows stepwise decline by a corresponding increase in catches.

Index of Trophic Completeness

The index of trophic completeness (ITC) is based on communities of freshwater macroinvertebrates. Species were divided into 12 trophic groups on the basis of their trophic characteristics(**Table 6**). Each trophic group fulfills a function in the benthic community. The

Table 6 Characteristics of the macroinvertebrate groups distinguished in the ITC, including the relative number of taxa per group present in the database

Trophic group, no.	*Diet*	*Feeding behavior*	*Food size (mm)*	*Relative number (%)*
1	Carnivory	Active shredder/chewer	>1	9.8
2	Carnivory	Passive shredder/chewer	>1	3.6
3	Omnivory	Shredder/chewer/collector	>1	5.9
4	Herbivory	Shredder/chewer	>1	7.8
5	Herbivory	Shredder/chewer	<1	2.6
6	Herbivory	Scraper	<1	26.3
7	Herbivory	Collector	<1	22.7
8	Herbivory	Filtrator	<1	8.7
9	Carnivory	Sucker (incomplete food ingestion)	>1	6.6
10	Carnivory	Sucker (total food ingestion)	>1	2.4
11	Herbivory	Sucker	>1	1.9
12	Omnivory	Shredder/chewer	<1	1.7

Adapted from Pavluk, T. I., bij de Vaate, A. and Leslie, H. A. (2000). Development of an index of trophic completeness for benthic macroinvertebrate communities in flowing waters. *Hydrobiologia* 427, 135–141.

Table 7 Indicative value (weight factor) and respective score of the trophic groups.

Trophic group	C *(weight factor)*	*Ln* C (C_i)
1	10.2	2.3
2	27.6	3.3
3	16.9	2.8
4	12.8	2.6
5	39.2	3.7
6	3.8	1.3
7	4.4	1.5
8	11.5	2.4
9	15.2	2.7
10	41.4	3.7
11	53.2	4.0
12	57.3	4.1

Adapted from bij de Vaate, A., Pavluk, T. I. (2004). Practicability of the index of trophic completeness for running waters. *Hydrobiologia* 519, 49–60.

Table 8 Quality class score for the ITC with five classes.

Quality class	C_{tot}	*Quality description*
I	≥ 28	High
II	21–28	Good
III	14–21	Moderate
IV	7–14	Poor
V	0–7	Bad

Adapted from bij de Vaate, A., Pavluk, T.I. (2004). Practicability of the index of trophic completeness for running waters. *Hydrobiologia* 519, 49–60.

trophic characteristics of each group include the following criteria: plant/animal ratio in the diet, feeding mechanism, food size, food acquisition behavior, and energy and substance transfer ways. Any undisturbed benthic macroinvertebrate community should be represented by members of each of these 12 trophic groups, irrespective of the part of the river studied and its geographical region.

The ITC was developed to show the degree of benthic community functional completeness reflected via its trophic relations to other components of an aquatic ecosystem.

The modern modification of the ITC calculation presumes application of weight factor (**Table 7**), because the probability to meet species of different trophic groups in the aquatic community is unequal and quality of water and number of trophic groups present has no straight linear relationship.

The ITC is calculated using the formula

$$C_{\text{tot}} = \sum_{i=1}^{n} Ci$$

where C_{tot} is the total score for the index, n is the number of trophic groups present in the data set, and C_i is the Ln-transformed indicative value of trophic group i.

The relation between the ITC value and the quality classes is given in **Table 8** for an assessment system with five quality classes.

The index indicates the functionality of the community and is based on the assumption that in healthy environment all trophic groups will be present. Natural fluctuations (e.g., floods, drought, ice covering) may influence the community structure, but their long-term effects do not result in the extinction of trophic groups.

Recently a handy tool for the ITC calculation for macroinvertebrates was developed. It is called MaTroS (Macrozoobenthos Trophic Structure). The tool has a friendly interface with a simple algorithm for the index calculation (http://www.macro.nemi-ekb.ru/).

Summary

Applied research of aquatic ecosystems involves enormous amount of parameters to give a detailed description of ecological processes and to verify the degree of anthropogenic interference that takes place. To transform parameters studied into a clear and

integrated form indices are used of which the group of trophic indices is the most popular. Two groups of the indices are distinguished: trophic indices based on the primary production potential of ecosystems, and the indices based on the structure of energy and substance transferring between trophic levels of aquatic inhabitants. Trophic indices allow one to make comparative studies between very different aquatic ecosystems, even those that are located in different continents with completely different species compositions.

See also: Ecological Data Analysis and Modelling: Climate Change Models; Climate Change Models. Ecological Processes: Physical Transport Processes in Ecology: Advection, Diffusion, and Dispersion. General Ecology: Ecological Efficiency

References

Coring, E., Schneider, S., Hamm, A., Hofmann, G., 1999. Durchgehendes Trophiesystem auf der Grundlage der Trophieindikaation mit Kieselalgen. Germany: Deutscher Verband für Wasserwirtschaft und Kulturbau e.V Koblenz.
Kelly, M.G., Whitton, B.A., 1995. The trophic diatom index: a new index for monitoring eutrophication in rivers. Journal of Applied Phycology 7, 433–444.
Pauly, D., Christensen, V., Walters, C., 2000. Ecopath, Ecosim, and Ecospace as tools for evaluating ecosystem impact of fisheries. ICES Journal of Marine Science 57 (3), 697–706.
Rott, E., Pipp, E., Pfister, P., *et al*., 1999. Indikationslisten für aufwuchsalgen in österreichischen fliessgewässern. Teil 2: Trophieindikation. Wien, Austria: Bundesministerium für Land- und Forstwirtschaft.

Further Reading

bij de Vaate, A., Pavluk, T., 2004. Practicability of the index of trophic completeness for running waters. Hydrobiologia 519, 49–60.
Burns, N.M., Rutherford, J.C., Clayton, J.S., 1999. A monitoring and classification system for New Zealand lakes and reservoirs. Journal of Lakes Research and Management 15, 255–271.
Carlson, R.E., 1977. A trophic state index for lakes. Limnology and Oceanography 22, 361–369.
Carlson, R.E., Simpson, J., 1996. A coordinator's guide to volunteer lake monitoring methods. USA: North American Lake Management Society Madison.
Giovanardi, F., Vollenweider, R.A., 2004. Trophic conditions of marine coastal waters: experience in applying the trophic index TRIX to two areas of the Adriatic and Tyrrhenian seas. Journal of Limnology 63, 199–218.
Howmiller, R.P., Scott, M.A., 1977. An environmental index based on relative abundance of oligochaete species. Journal of the Water Pollution Control Federation 49, 809–815.
Milbrink, G., 1983. An improved environmental index based on the relative abundance of oligochaete species. Hydrobiologia 102, 89–97.
Moncheva, S., Doncheva, V., 2000. Eutrophication index ((E) TRIX)—an operational tool for the Black Sea coastal water ecological quality assessment and monitoring. International Symposium "The Black Sea Ecological Problems"Odessa: SCSEIO, pp. 178–185.
Pauly, D., Watson, R., 2005. Background and interpretation of the 'marine trophic index' as a measure of biodiversity. Philosophical Transactions of the Royal Society B 360, 415–423.
Pavluk, T.I., bij de Vaate, A., Leslie, H.A., 2000. Development of an index of trophic completeness for benthic macroinvertebrate communities in flowing waters. Hydrobiologia 427, 135–141.
Rizzo, W.M., Berry, B.E., Wetzel, R.L., *et al*., 1996. A metabolism-based trophic index for comparing the ecological values of shallow-water sediment habitats. Estuaries 19, 247–256.
Schneider, S., Melzer, A., 2003. The trophic index of macrophytes (TIM)—a new tool for indicating the trophic state of running waters. International Review of Hydrobiology 88, 49–67.
Word, J.Q. (1978). The infaunal trophic index Southern California Coastal Water Research Project, Annual Report. El Segundo, California. pp. 19–39.
Word, J.Q. (1980). Classification of benthic invertebrates into infaunal trophic index feeding groups Southern California Coastal Water Research Project, Biennial Report 1979–1980. Long Beach California, pp. 103–121.
Word, J.Q. (1990). The infaunal trophic index, a functional approach to benthic community analyses. University of Washington Seattle, Washington, PhD Thesis.

Relevant Websites

http://www.fishbase.net—FishBase.
http://www.NALMS.org—North American Lake Management Society (NALMS).
http://www.seaaroundus.org—Sea-Around-U.S. database; Web Products: Information by Species.
http://www.epa.gov—U.S. Environmental Protection Agency, External Links Disclaimer.
http://www.macro.nemi-ekb.ru—Macrozoobenthos Trophic Structure (MaTroS).

Turnover Time☆

EH Dettmann, US Environmental Protection Agency, Office of Research and Development, Narragansett, RI, USA

Introduction

Turnover time refers to the amount of time required for turnover or replacement by flow-through of the energy or a material contained in a system. Ecological units are generally open to flow-through of energy and materials. Carbon, nitrogen, phosphorus, and other materials, as well as energy, enter the system, may be incorporated into system components and later be released by respiration, excretion, decomposition, and other processes, and eventually leave the system. The material of interest may undergo a single cycle of incorporation and release within the system, or may be recycled many times before exiting. Some substances may be passively transported through the system. The concept of turnover is also applied to appearance and extinction of biological species in a local community, and can describe passage of individuals through a population through recruitment, emigration, and harvesting.

Turnover time provides a timescale for this replacement and is a tool for analyzing the significance of flow rates and the sizes of material and energy pools within a system or its components. The turnover time of a material is generally defined as the ratio of the quantity of that material in the system to its throughput or flow-through rate. This definition assumes that the rates of inflow and outflow are equal, that is, the system is in equilibrium (steady state). It is an average quantity, since the calculation provides no way to account for the transit time of an individual entity (e.g., an atom or molecule) through the system. Turnover time can be calculated at various spatial scales, for small ecosystems or large ones, such as the component subsystems of the global biogeochemical cycle of an element. Turnover times of subsystems will differ from that of the larger system of which they are a part.

Calculation of Turnover Time of Material and Energy

Turnover time (τ) for a flow of material is the ratio of the quantity of material in the system (M) to its flow rate (F) through the system, that is,

$$\tau = \frac{M}{F} \qquad [1]$$

Turnover time is the inverse of the fraction of material in the system that leaves per unit time. It may also be viewed as the amount of time required to put into the ecosystem an amount of the substance equal to that currently residing there, or for that amount to leave the system.

The units used to express the quantities in the numerator and denominator of eqn [1] must be consistent. For materials, the mass of the material in the system is M, and F is the flux (mass/time), with flux expressed using the same units for mass. If energy turnover is calculated, the ratio is of system energy content to energy flux, both expressed in consistent energy units.

Other timescales used in analyzing the dynamics of material flow through systems are age and transit time. The age of an entity is the amount of time that has elapsed since its entry into the system; its transit time is the total time between its entry into and exit from the system. Mean age is the mean age of all material in the system; mean transit time is the mean transit time of material leaving the system.

Generally, turnover time, mean age, and mean transit time may all change through time. Assuming constant flow rates, turnover time is constant only if the inflow rate equals the outflow rate. Otherwise, turnover time increases or decreases depending on whether inflow is larger or smaller than outflow. If the system is in equilibrium, the turnover time may be shown to be equal to the mean transit time, also called the residence time. The terms turnover time and residence time are often used interchangeably in the literature; this can result in confusion since these two quantities are generally not equal in systems that are not in equilibrium. Reentry of material that has exited the system can also complicate the analysis. If the system content and flow-through rate vary around central values, time-averaged values for M and F are often used to obtain an estimate of turnover time. If there is a rising or falling trend in the quantity of material in the ecosystem or its flow-through rate, the system is not in equilibrium, and one may need to calculate instantaneous turnover times at one point or a series of points in time.

Applications of Turnover Times in Ecosystem Analysis

Turnover time provides a useful tool for analysis of material fluxes, pool sizes, and the time required for pools to attain steady state after a change of inflow rate. The following examples illustrate a few of these applications for terrestrial and aquatic ecosystems.

☆*Change History:* August 2013. E Dettmann updated the section Species Turnover in Local Communities and added the article by Bolin and Rodhe (1973) to the Further Reading section.

These include applications to nutrient flows in the forest floor, changes in species composition in local communities, turnover of individuals in harvested populations, and pool sizes and dynamics for conservative materials and nutrients in aquatic systems. Similar techniques apply to calculation of pool sizes and transport properties of atmospheric ozone-depleting and greenhouse gases, but are beyond the scope of this section.

Turnover of Nutrients in Forests

Analysis of nutrient flows in forests must include consideration of turnover times of available nutrient pools. For instance, nutrients in litter and surface soils of temperate and wet tropical forests have very different turnover times. Storage of carbon and nutrients in litter and surface soils of wet tropical forests is relatively small, and any nutrients in dead plant or animal materials are rapidly released and made available to vegetation, resulting in short turnover times in this pool. In comparison, large quantities of carbon and nutrients are stored in the litter layer and surface soils of temperate forests, resulting in long turnover times for the nutrients stored in this compartment.

Species Turnover in Local Communities

The species composition of localized ecosystems may change, with new species immigrating and some species becoming locally extinct through mortality or emigration. Community ecology uses a number of definitions of turnover that estimate different aspects of change in community composition. The following describes some metrics of community change that are based on observations of species that are made at two successive times, t_1 and t_2 in a local ecological unit, such as an island or forest patch. N_{t1} and N_{t2} represent the total number of species in the community at times t_1 and t_2.

Change in species richness, one such metric, is defined as N_{t2}/N_{t1}, the ratio of the number of species present at the later time to that present earlier. This metric does not distinguish between numbers of species gained and lost during the time interval (t_1, t_2).

Two metrics of change in species composition of a community, turnover (colonization) rate and extinction rate, quantify local rates of appearance of new species (colonization) and extinction. These may be expressed as absolute numbers or as fractional gains or losses of species over the interval (t_1, t_2).

In the following, $C_{t1,t2}$ is the number of species common to observations at both t_1 and t_2. The number of new species that appear in the community during the time interval (t_1, t_2) is $N_{t2} - C_{t1,t2}$, and the number of species that become locally extinct in the same time interval is $N_{t1} - C_{t1,t2}$. The fractional turnover $(T_{t1,t2})$ of species in this time interval is defined as the ratio of the number of new species to the total number of species observed at time t_2:

$$T_{t1,t2} = \frac{N_{t2} - C_{t1,t2}}{N_{t2}} = 1 - \frac{C_{t1,t2}}{N_{t2}} \quad [2]$$

The fractional extinction $(E_{t1,t2})$ of species in this same time interval is defined as the ratio of the number of species that become locally extinct to the number of species at time t_1:

$$E_{t1,t2} = \frac{N_{t1} - C_{t1,t2}}{N_{t1}} = 1 - \frac{C_{t1,t2}}{N_{t1}} \quad [3]$$

Thus, $T_{t1,t2}$ and $E_{t1,t2}$ quantify the rates of local species appearance and disappearance within the time interval (t_1, t_2) as a fraction of the total number of species present; $N_{t2} - C_{t1,t2} = N_{t2}(T_{t1,t2})$ and $N_{t1} - C_{t1,t2} = N_{t1}(E_{t1,t2})$ express them as actual numbers of new species gained or old species lost. If these quantities are compared among studies over time intervals of differing length, they should be divided by the number of time units (e.g., years, decades) in the studies to which they refer so that they are all expressed in compatible time units. The ratios $(t_2 - t_1)/(T_{t1,t2})$ and $(t_2 - t_1)/(E_{t1,t2})$ are analogous to turnover times defined for mass or energy flow through a system.

Turnover and extinction rates for a community may or may not be equal. If the set of species in the community is the same at times t_1 and t_2, $N_{t1} = N_{t2} = C_{t1,t2}$, and both the turnover and extinction rates are zero. If the species composition changes in this time interval but the number of species stays the same, $N_{t1} = N_{t2} > C_{t1,t2}$, and the turnover and extinction rates are equal but nonzero. The total number of species in the community (species richness) may also change, in which case $T_{t1,t2} \neq E_{t1,t2}$.

Species turnover and extinction rates are used to investigate the influence of factors such as habitat fragmentation, patch size, and isolation from similar systems on species richness and the stability of species composition in local ecological units. Examples include the effects of island area and distance from other islands or the mainland on turnover and extinction rates of plant and animal species, and the dependence of local turnover and extinction rates of bird or other biotic communities in forest patches or other localized ecological units on system size.

Turnover concepts also apply to management of harvested populations such as marine fisheries, where increased adult mortality (from harvest) is theoretically compensated by increased recruitment of juveniles into the adult class due to reduced competition. Thus, entry and exit of individuals to and from the adult class create higher turnover within the population in response to harvest. The framework described here has also been applied to data for fossils to test hypotheses concerning appearance and extinction of taxa in the geological record and variation among phyla.

Lake Water Quality and Trophic State

Turnover time of water is a characteristic of a water body that is often used in analysis of its water quality and trophic state. Water acts as a carrier of substances such as nutrients and contaminants, and simple models employing the turnover time of water provide insight into the dynamics and pool sizes of these substances and ecological responses to their concentrations. The following examples for lakes illustrate this modeling approach.

The turnover time of water (τ_w) in a well-mixed lake is $\tau_w = V/Q$, where V is the lake volume and Q is the flow-through rate of water. Since water may enter lakes from numerous sources but often exits through a single localized outlet, flow through the outlet is often used to estimate flow-through rate. The terms water residence time and hydraulic detention time are often used in the limnological and engineering literature to denote water turnover time of a lake.

Turnover times vary widely among lakes; they are often on the order of months or years, but may be as long as a century or more for some very large lakes, for example, North America's Lake Superior and Lake Tahoe. Flows frequently vary on seasonal or shorter timescales that are less than lake turnover time, and long-term mean flow rates are often used to calculate mean turnover times.

The following examples illustrate use of turnover time of water in box models, sometimes called input–output models, to estimate concentrations in lake water of (1) conservative substances (those that are not removed from the water column, e.g., the chloride ion) and (2) nonconservative materials (ones that may be removed from the water column, for example, nutrients). The models are used to estimate steady-state concentrations in the water column for constant rates of input, and the dynamics of these concentrations after a change of input rate. Water flow rates are assumed throughout to be constant, so that the turnover time of water is constant. However, concentrations and turnover times of transported materials will vary if their loading rates change or their concentrations have not yet reached steady state. These models assume complete mixing, that is, they assume that concentrations of the materials of interest are uniform throughout the lake. Lack of complete mixing because of seasonal stratification or rapid flushing complicates the analysis, but does not change the underlying principles. The relationship between turnover times of water and transported materials will also be shown to depend on whether the latter are conservative.

Conservative substances

In the following, C denotes the concentration (mass/volume) in the water column of the material of interest, L its loading (input) rate (mass/time), Q the water flow-through rate (volume/time), C_{in} the mean concentration in inflows ($C_{in} = L/Q$), and t the time elapsed since the initial condition. The subscripts 'o' and 'ss' denote initial and steady-state conditions in the lake.

The rate of change of mass of a conservative substance in a lake is equal to the rate of input minus the loss rate:

$$V\frac{dC}{dt} = L - QC \qquad [4]$$

Therefore,

$$\frac{dC}{dt} = \frac{L}{V} - \frac{C}{\tau_w} \qquad [5]$$

For nonzero loading rates, this equation has the solution

$$C = \frac{\tau_W L}{V}\left[1 - \left(1 - \frac{V}{\tau_W L}C_o\right)e^{-t}/\tau_w\right] \qquad [6]$$

Since $\tau_w L/V = (V/Q)(L/V) = L/Q = C_{in}$, the mean concentration in the inflow, this may be written as

$$C = C_{in}\left[1 - \left(1 - \frac{C_o}{C_{in}}\right)e^{-t}/\tau_w\right] \qquad [7]$$

From its initial value of C_o at $t = 0$, the concentration asymptotically approaches C_{in}, the concentration in inflowing water. For $L = 0$ (no loading), the solution to eqn [5] is

$$C = C_o e^{-t}/\tau_w \qquad [8]$$

that is, the concentration decreases exponentially from the initial value, and asymptotically approaches zero. Thus, for both zero and nonzero loading rates, the steady-state concentration is

$$C_{ss} = \frac{\tau_w L}{V} = C_{in} \qquad [9]$$

The behavior of the concentration after onset of an increase in loading to the lake is shown in Fig. 1. The concentration increases (or decreases if $C_{in} < C_o$) from its initial value C_o and asymptotically approaches the steady-state value C_{ss}, with the rate of approach controlled by the water turnover time τ_w. At $t = 3\tau_w$, $C_{ss} - C$ will be 5% of the difference between C_{ss} and C_o. Since the asymptotic approach to steady state in eqns [6]–[8] is described by the factor e^{-t}/τ_w, τ_w is sometimes called the e-folding time for the concentration in this modeling framework. The turnover time of the conservative substance at steady state is the ratio of the mass in the water column at steady state to the throughput, or $(C_{ss}V/L) = \tau_w$, the same as that for water.

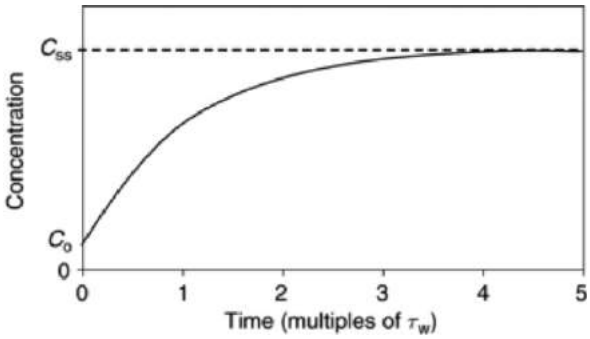


Fig. 1 Concentration versus time (in multiples of water turnover time τ_w) after an increase in loading rate. The initial and steady-state concentrations are C_o and C_{ss}, respectively.

Nonconservative substances

Many substances are nonconservative in aquatic systems. The macronutrients phosphorus and nitrogen, for instance, may be removed from the water column by biological uptake and subsequent incorporation into sediments. Nitrogen may also be removed by denitrification. Such permanent losses may be included in models as loss terms.

Extensive use has been made of input–output models that employ water turnover time as a parameter to calculate mean concentrations of total phosphorus in lakes. Since concentrations of total phosphorus and chlorophyll *a* in lakes are correlated, these models provide a useful means of analyzing the effects of phosphorus loading on the trophic states of lakes. Equation [10] assumes that the loss rate of phosphorus from the lake water column to the sediments is proportional to the water's phosphorus content, with first-order rate parameter σ. The term $-\sigma VC$ represents only long-term loss of phosphorus to sediments; it does not include the fraction that is recycled from sediments to the water column.

$$V\frac{dC}{dt} = L - QC - \sigma VC \qquad [10]$$

This equation is the same as eqn [4] except for the addition of the loss term that represents sedimentation. Equation [10] has the time-dependent solution,

$$C = \frac{\gamma L}{V}\left[1 - \left(1 - \frac{V}{\gamma L}C_o\right)e^{-t/\gamma}\right] \qquad [11]$$

and the steady-state solution,

$$C_{ss} = \frac{L\tau_w}{V}\left(\frac{1}{1+\sigma\tau_w}\right) = \frac{L\gamma}{V} \qquad [12]$$

where γ is defined as

$$\gamma \equiv \frac{1}{1}\Big/\tau_w + \sigma = \frac{\tau_w}{1+\sigma\tau_w}$$

to simplify presentation of the equations. Note that eqns [6] and [11] have the same form, with τ_w in eqn [6] replaced by γ in eqn [11]. Thus, the time dependence of phosphorus concentrations (eqn [11]) is also described in Fig. 1, with the asymptotic (steady-state) concentration given by eqn [12] and the abscissa representing time in multiples of $\gamma = \tau_w/(1+\sigma\tau_w)$. The turnover time of phosphorus in the water column at steady state is therefore $\tau_p = \tau_w/(1+\sigma\tau_w)$. This equivalence may also be seen by calculating the ratio of the mass of phosphorus in the water column at steady state to the input rate $(C_{ss}V)/L$ in eqns [9] and [12]. The turnover time of phosphorus is less than that of water because phosphorus is lost to both outflow and sedimentation, resulting in a smaller steady-state concentration than if the rate constant for phosphorus loss (σ) were zero (see eqn [12]). The value of the rate parameter σ may vary among lakes, requiring lake-specific measurements. Variations of this model use alternate methods of characterizing the phosphorus loss term and address issues such as seasonal changes and nonconstant loading rates.

This approach provides a basis for estimating phytoplankton abundance in the water column. A frequently used relationship between phytoplankton chlorophyll *a* and phosphorus concentrations in lakes is $[\text{Chl } a] = 0.0731[\text{P}]^{1.449}$, where [Chl *a*] is the summer concentration of chlorophyll *a* (mg m^{-3}), and [P] (mg m^{-3}) may be estimated from eqn [12].

Nitrogen in Estuaries

This example demonstrates the use of the turnover time of freshwater (τ_{fw}) in estuaries to analyze nitrogen throughput and concentrations. This measure (τ_{fw}), often called the freshwater residence time or flushing time in the estuarine literature, is particularly useful in estuaries when the watershed is the primary source of a material of interest.

Estuaries are semienclosed coastal water bodies having a connection with the sea, and in which freshwater mixes with seawater. Unlike lakes, estuaries are not simple flow-through systems since the seaward boundary is both a source of seawater and nutrients

and an outflow. Estuaries typically exhibit salinity gradients, with salinity lower in the inner estuary, near freshwater sources, than at the seaward boundary. Spatial distributions of nutrients and other materials also exhibit gradients that reflect their sources.

Freshwater entering the inner estuary from the watershed usually remains in the estuary longer than seawater, much of which may enter on a flood tide and exit on the following ebb tide. Therefore, there may be a number of measures of water turnover time in an estuary: including the turnover time of freshwater (the focus here), seawater, or that of all water in the estuary.

The freshwater turnover time of an estuary depends on both its content and the input rate of freshwater. The freshwater content of an estuary depends in turn on its inflow rate and physical factors such as tidal forcing that determine its distribution and removal rate from the system. The turnover time of an estuary is variable, changing with the rate of freshwater inflow and tidal range. Depending on the application, one may wish to calculate a turnover time for a specific set of conditions, or a long-term mean value for mean forcing conditions. Turnover times for most estuaries range from days for small systems to months for large ones.

The turnover time of freshwater in the estuary may be determined by measuring its freshwater content and inflow rate, and calculating the amount of time required to replace this freshwater, $\tau_{fw}=V_{fw}/Q$, where V_{fw} is the volume of freshwater in the estuary, and Q is the inflow rate of freshwater from all sources. For an estuary with total volume V, V_{fw} is determined by calculating the volume of freshwater that must be mixed with a volume of seawater ($V-V_{fw}$) that enters across the seaward boundary to yield the mean volume-weighted salinity (S_e) of water in the estuary, as determined by measurements. From mass balance considerations,

$$V_{fw}=\left(1-\frac{S_e}{S_b}\right)V \qquad [13]$$

where S_b is the salinity of seawater outside the seaward boundary. Another experimental method for measurement of turnover time, practical only for small estuaries, is by continuous introduction of a conservative dye into the freshwater source until concentrations of dye at individual points in the estuary reach equilibrium. Turnover time is determined by then terminating dye input and measuring the e-folding time for the spatially averaged dye concentration – the time required to decline to e^{-1} times its equilibrium value, where e is the base of natural logarithms.

The following example illustrates use of freshwater turnover time to analyze steady-state throughput and concentration of total biologically active nitrogen in estuaries. Biologically active nitrogen (N) is defined here as dissolved inorganic nitrogen (ammonium, nitrite, and nitrate) plus dissolved and particulate organic forms. Nitrogen gas (N_2) is not included.

The rate of change of the mass of N in the estuary may be represented as the difference between all source and loss rates:

$$\frac{dN}{dt}=L_l+L_s-E-\alpha N \qquad [14]$$

where L_l is the loading rate of nitrogen from the watershed (including direct discharges to the estuary) and atmosphere, L_s is the inflow rate across the seaward boundary, E is the export rate of nitrogen across the seaward boundary, and αN is the loss rate of nitrogen within the estuary to processes such as denitrification, permanent burial in sediment, and incorporation in fish biomass. The parameter α is a first-order loss coefficient analogous to that for phosphorus loss in lakes (σ). The term αN does not include temporary loss from the water column of nitrogen that is taken up by phytoplankton, subsequently sinks to the bottom, and is eventually remineralized and returned to the water column. The steady-state solution of this equation may be found by setting $dN/dt=0$. The rate of export of that nitrogen which has entered from the watershed and atmospheric deposition (E_n) is the difference between the total export and inflow rates of nitrogen across the seaward boundary: $E_n=E-L_s$. The fraction of nitrogen originating from the watershed and atmosphere that is exported is $F_{E(l)}=E_n/L_l$. If watershed and atmospheric inputs of nitrogen are substantially greater than those across the seaward boundary, an analysis similar to that for phosphorus in lakes shows that $F_{E(l)}$ may be approximated as

$$F_{E(l)}=\frac{1}{1+\alpha\tau_{fw}} \qquad [15]$$

Application of this equation to data for a group of North American and European estuaries has shown that α has a value of approximately 0.3 per month. This function is shown in **Fig. 2**. Note that the longer the turnover time of freshwater, the larger the consumptive losses of nitrogen from watershed and atmospheric sources within the estuary, and the smaller the fraction that is exported across the seaward boundary. This relationship applies specifically to nitrogen. For other substances, the existence and details of an analogous relationship would depend on details of loss mechanisms. For instance, all of a conservative substance ($\alpha=0$) entering an estuary is exported for all residence times, since none is lost in the estuary.

Solving eqn **[14]**, the steady-state concentration of total nitrogen from all sources, spatially averaged over the estuary, may be shown to be

$$[N]_{ss}=\left(\frac{L_l\tau_{fw}}{V}+[N_s]\right)\frac{1}{1+\alpha\tau_{fw}} \qquad [16]$$

where V is the volume of the estuary, the first term in the parentheses represents the contribution to the nitrogen concentration in the estuary by loading from watershed and atmospheric sources, and the second term ($[N_s]$) is the contribution from nitrogen input across the seaward boundary. The multiplicative term to the right of the parentheses corrects these contributions for losses within the estuary. The concentration $[N_s]$ may be estimated as the seawater content of the estuary (V_{sw}) times the concentration of

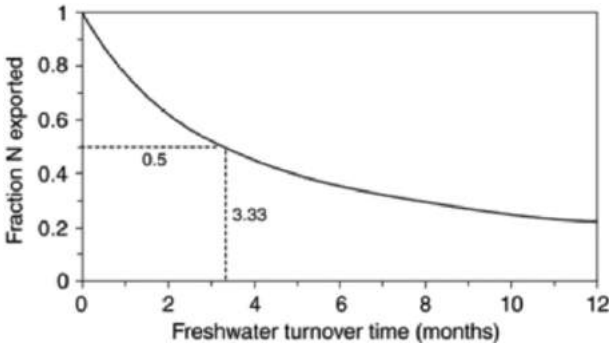


Fig. 2 The fraction of nitrogen (N) input from watershed and atmospheric sources that is exported across the seaward boundary of an estuary versus freshwater turnover time. The solid line shows the relationship $F_{E(l)}=1/(1+\alpha\tau_{fw})$ for $\alpha=0.3$ per month. For $\tau_{fw}=3.33$ months, half of the input from these sources is exported. For substances other than nitrogen, the existence and details of an analogous relationship would depend on loss mechanisms.

nitrogen at the seaward boundary ($[N_b]$), divided by V. Since $V_{sw}=V-V_{fw}$, where V_{fw} is given by eqn [13], $V_{sw}=(S_e/S_b)V$, and $[N_s]=[N_b](S_e/S_b)$. The spatially averaged concentration of nitrogen in the estuary from watershed and atmospheric sources alone,

$$\frac{L_1\tau_{fw}}{V(1+\alpha\tau_{fw})} \qquad [17]$$

increases as τ_{fw} increases and asymptotically approaches the limit $L_1/\alpha V$ for large values of τ_{fw}. Since estuaries generally exhibit spatial concentration gradients, the spatially averaged concentration $[N]$ is a measure of generalized estuary response to nitrogen loading, rather than a point measure.

Summary

Turnover time is defined as the ratio of the quantity of a material or energy in a system to its outflow rate. It may also be viewed as the inverse of the fraction of material or energy that leaves per unit time. For a system in equilibrium, the outflow rate is equivalent to the flow-through rate and the turnover time of a substance is equal to the residence time, that is, the mean transit time between entry into and exit from the system. Turnover time provides a timescale for the replacement of energy or materials in a system and is a useful tool for analysis of material and energy fluxes, pool sizes, and the time required for pools to attain steady state after a change in loading rate.

The concept of turnover is also applicable to analysis of species appearance and extinction rates in local systems, and to analysis of harvested populations such as in marine fisheries.

Acknowledgment

This is contribution number ORD-004917 of the US Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Atlantic Ecology Division.

See also: Ecological Data Analysis and Modelling: Biogeochemical Models. General Ecology: Generation Time

Further Reading

Bolin, B., Rodhe, H., 1973. A note on the concepts of age distribution and transit time in natural reservoirs. Tellus 25, 58–62.
Boulinier, T., Nichols, J.D., Hines, J.E., *et al.*, 2001. Forest fragmentation and bird community dynamics: Inference at regional scales. Ecology 82, 1159–1169.
Dettmann, E.H., 2001. Effect of water residence time on annual export and denitrification of nitrogen in estuaries: A model analysis. Estuaries 24, 481–490.
Dillon, P.J., Rigler, F.H., 1974. The phosphorus–chlorophyll relationship in lakes. Limnology and Oceanography 19, 767–773.
Nichols, J.D., Boulinier, T., Hines, J.E., Pollock, K.H., Sauer, J.R., 1998. Estimating rates of local species extinction, colonization, and turnover in animal communities. Ecological Applications 8, 1213–1225.
O'Neill, B.C., Gaffin, S.R., Tubiello, F.N., Oppenheimer, M., 1994. Reservoir timescales for anthropogenic CO_2 in the atmosphere. Tellus 46B, 378–389.
Pilson, M.E.Q., 1985. On the residence time of water in Narragansett Bay. Estuaries 8, 2–14.
Reckhow, K.H., Chapra, S.C., 1983. In: Engineering Approaches for Lake ManagementData Analysis and Empirical Modeling, Vol. 1. Boston: Butterworth.

Wildlife Ecology☆

Melissa Songer, Smithsonian Conservation Biology Institute, Front Royal, VA, United States

Introduction

Wildlife ecology is the scientific discipline of applying ecological principles to the study of wildlife species and their habitats. Research goals typically include developing applied strategies for managing natural systems to achieve and maintain balance among wildlife, habitats, and humans, and often result in baseline data that can be used to inform wildlife management plans. The term wildlife is not strictly defined and through the years its use has ranged from including only terrestrial vertebrates to encompassing all wild animal and plant life. Historically, wildlife management mainly targeted game species, particularly birds and mammals that were traditionally hunted, and often with the goal of increasing their populations to bolster sport and subsistence hunting opportunities. During the 1960s the focus of wildlife management and research began to expand to include nongame species and other taxa, such as reptiles, amphibians, and even invertebrates. In modern wildlife management, rare and endangered species are a top consideration and usually prioritized over game species in management plans. During the past decade many new quantitative methods have been advanced and are becoming essential to management, conservation, and the science of wildlife ecology. This article will cover underlying principles and concepts, applications, and current directions of wildlife ecology research.

Underlying Concepts

Wildlife ecology studies can be approached from various scales and organization levels, ranging from the study of individual animals and their relationships with abiotic and biotic aspects of their environment, relationships among individual animals within populations of a species (e.g., social behavior, demography, population changes), interactions between different species (e.g., predator–prey relationships, parasitism, pathogens, competition), community structure, and ecosystems. Ecological concepts foundational to wildlife study are discussed below.

Ecosystems and Species Relationships

Ecosystems are often the analysis framework for wildlife ecology studies. An ecosystem is the basic system that supports and sustains life, including all living organisms which interact with each other and with nonliving components of their environment within a given area. All wildlife species have roles within their ecosystem, called niches, comprised of how a species obtains food and water, and how it interacts with other living and nonliving components of the ecosystem. A species' niche is not always unique and can overlap in some ways with other species.

All individuals and species exist within a community composed of many different species and interact through various relationships, such as competition (with individuals of the same species and between different species), predation, and facilitation. Use and consumption of resources is the underlying driver for most of these interactions/relationships, whether it is between two different species or within the same species. By definition a resource is something essential to an animal (e.g., food, water, shelter) which, when used by an individual animal, is no longer available to another individual.

When two different species need the same resources and there is not enough available to support all it leads to competition. Competition can negatively affect both species, though typically one species will be a superior competitor, leaving one species more negatively impacted than the other. Wildlife ecology studies on competition can help inform strategies for reducing impacts on a species of concern, for example, when competition between wildlife species threatens a rare or endangered species, or when domestic animals are outcompeting wildlife for essential resources. In most ecosystems, humans also compete with wildlife, often using and consuming at levels much higher than any wildlife species.

Competition can reach the point of affecting the survival and reproduction rates of the species facing competition, creating a situation referred to as natural selection—an underlying concept of evolution and ecology. The concept asserts that individuals that are most fit (able to survive in their environment) are the most likely to reproduce and pass along their genes to the next generation, thereby increasing the chances of perpetuating the characteristics that made the individual more fit to survive. This process is called natural selection because the individuals which are the most fit are "selected" by nature to survive and reproduce. Species gradually change as a result of this process and over hundreds or thousands of generations it can lead to the evolution of a distinct species.

Predation is a relationship between a predator and a prey species in which the predator consumes part or all of the prey. There are four basic types of predation: (1) herbivory, in which an animal species eats parts or all of a plant; (2) parasitism, in which one

☆This is an update of J.L. Rachlow, Wildlife Ecology, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 3790–3794.

species (the parasite) feeds on another species (the host) in a way that may cause negative impacts but usually does not kill the host; (3) carnivory, in which the predator catches and consumes the prey; and (4) cannibalism, a rare occurrence when the predator species and the prey species are the same. Detritivory is the consumption of dead organic material by scavengers and does not affect the survival of the consumed species as in predation.

Though predator and prey often coexist in a stable way, various factors may cause the system to go out of balance. If predation levels are too high it could lead to a decrease in the prey population. Without sufficient prey, the predator species could also decrease. In human dominated systems, a more likely scenario impacting predator–prey relationships is the decline of predator populations, especially for larger mammals which require more habitat, are considered a threat to humans, and often are actively hunted. When predation levels are very low or even removed it can lead to a great increase a prey species and have an impact on other species as well. For example, if a prey species increases rapidly it could lead to insufficient food resources, creating a competitive situation that could also lead to the decline in populations of other species. Understanding predatory relationships are of critical importance to wildlife management.

Facilitation refers to a species relationship in which at least one of the species involved benefits from the relationship and neither species incurs a negative impact as a result. There are two basic forms; mutualism, in which the relationship is beneficial to both species, and commensalism, in which one species benefits and the other remains unharmed by the relationship. Mechanisms of facilitation include, a relationship that provides protection from environmental stress, reduces risk of predation, offsets competition from another species, improves access to resources, or aids in movement or migration. Facilitative relationships among species can influence species diversity and structure at the community level, not just the species that are involved in the relationship.

Altruism is a type of behavior that occurs within a species rather than between different species, when an individual acts in a way that increases the chance for survival of another individual while decreasing the chance of survival of the actor. This is usually seen between members of the same family, for example, parenting, in which altruism increases the survival rate for relatives that can pass on the same genes. It also occurs within species' social groups, for example, army ants forming body bridges to facilitate the movement of food to the group nests.

The Wolves of Yellowstone: A Case Study of Species Relationships

The history of wolves in Yellowstone National Park provides an example of the interdependencies among different species. Established in 1872, Yellowstone provided protection for wildlife, but excluded protection of predator species such as wolves, coyotes, and mountain lions. Wolves were generally considered a pest species and were hunted by park visitors and even rangers. They disappeared from the park by the 1920s and in 1967 became the first species listed as endangered in the United States. Over 30 wolves were released in Yellowstone during 1995–96 and have since established a relatively stable population (ranging from 9 to 16 packs). Wildlife ecologists have studied the impacts of the wolves on the ecosystem and found that their major prey species, the elk, decreased in number and were forced to forage more broadly into less preferred habitats, leading to lower consumption of willow trees—an essential resource for beavers. Reduced pressure on willows led to a substantial increase in the beaver population and more dams, and thus impacting the watershed by reducing erosion, replenishing the water table, and providing more habitat variety for fish, amphibians, and birds. The decrease in elk also led to higher availability of plants and berries eaten by grizzly bears, leading to an increase their population.

Competition and predation by wolves reduced coyote populations by half in only 2 years. In turn, the decrease in coyotes led to a rise in the number of foxes, and then followed an increase in coyote prey species (e.g., hares, deer, small rodents, and birds) and a change in the composition of local plants that are eaten by those prey species. Carcasses left by wolves provided food sources to scavengers such as vultures, ravens, eagles, bears, and martens. As a top predator, wolves fill a critical ecological role in the Yellowstone ecosystem and their return had far-reaching effects on a range of other wildlife species.

Population Growth and Regulation

Understanding factors that can influence the size of a wildlife population, such as species relationships and changing landscapes, is a primary focus of study in wildlife ecology. A population is a group of individuals living in the same area and sharing genetic information, while a species comprises all populations regardless of where they are located. A population can remain relatively stable or change drastically due to changes in their ecosystem caused by natural disaster, human activities, dynamics in species relationships, and other drivers. Population size changes are calculated based on the difference between the birth rate and the death rate of the group, along with the number of individuals leaving (emigration) or joining (immigration) the population. The difference between these numbers is the growth rate, which can be positive or negative. The maximum size that a population can grow and be supported within an environment without degrading the environment is called carrying capacity—a cornerstone concept of wildlife ecology.

Estimates of biological carrying capacity are based on availability of resources essential for a species and can be difficult to measure since resource availability is always changing. Survey methods usually do not account for other influences, such as competition with other species for the same resources (e.g., food, denning sites, etc.), wildlife diseases, or predation. In addition, some species have a cultural carrying capacity, which is the maximum number that will be tolerated by nearby human communities and can be much lower than the biological carrying capacity. A typical application in this area of wildlife ecology is

developing wildlife management strategies that can adjust for the factors that limit populations, and usually are designed to grow a particular population or increase its chance for survival. Another concern is the inflection point, which is the point at which individuals within a population can be removed and still maintain a maximum growth rate.

Population regulation refers to biological processes that counterbalance disruptive events (e.g., weather events, changing environmental conditions, disease outbreaks, etc.), causing wildlife populations to tend toward consistency in their numbers over time. Stability in a species' population is primarily maintained through changes in survival and reproduction rates. These rates can vary in response to changes in population density of a population and are, therefore, considered density dependent. Key feedback processes that help to maintain stability in populations are predator–prey relationships and competition for food or other resources. These provide negative feedbacks, that is, when a population size goes up, competition and predation also increase, causing a balancing decrease in that population.

Habitat and Habitat Selection

Habitat can be defined in relation to the land type (e.g., boreal, alpine, riparian, etc.), or related to its functional role for a species (e.g., nesting habitat, breeding habitat, shelter habitat, etc.). In the study of wildlife ecology, habitat is typically defined as the area that includes all the biotic and abiotic resources that an animal needs to survive and reproduce, including food and water, space, and shelter. Habitat assessment is a common focus in wildlife ecology studies and delineating a critical habitat for protection or improvement of habitat quality for a species, or community of multiple species, is a common application of these assessments. Assessing habitat quality can involve evaluating the availability of resources essential to the survival and reproduction of a species, a population, or a community of species, and understanding the influence of habitat quality.

The concept of species' habitat selection is based on the understanding that an individual animal will choose resources in a way that will increase their chances to survive and reproduce, and that those choices have evolved based on natural selection. Specific habitat selection behaviors may be learned from other individuals of the same species, or be may be instinctual or genetically inherent. Analyzing a species' selection of a habitat type within their ecosystem is typically done by observing the availability of all habitat types within that ecosystem and comparing it to the level of use of those available habitat types by a species. If a particular habitat type is used in higher proportion than its availability, then that habitat type is considered to be selected by the species. If a habitat type is being used at a level lower than its availability then it is considered to be not selected by the species. If the level of use of a habitat is similar to its availability then that is evidence that it is being used in a random way rather than being selected or not selected.

Studies comparing habitat use to availability help wildlife ecologists assess a species' habitat requirements, but are not as useful when trying to predict species distribution across landscapes. Resource selection functions provide methods for quantifying a species' habitat preferences and can include both categorical (e.g., vegetation, habitat type, etc.) and continuous variables (e.g., elevation, temperature, rainfall, etc.). Species presence can be detected via direct observations, such as capture, camera trapping, and radio or satellite telemetry, as well as indirect observations, such as through sign surveys (observing tracks, dung, and other evidence of species presence) and interviews with people who have observed the species. Species presence can then be assessed within a range of environmental variables to determine the key characteristics of areas where the species is present. Geographic Information Systems (GIS) has become an effective tool for extracting a range of abiotic and biotic variables relevant to resource selection and modeling habitat suitability.

Gurney's Pittas in Myanmar: A Case Study of Habitat Suitability Modeling

Gurney's Pitta was discovered in Myanmar and Thailand over 100 years ago, but little was known about this cryptic and rare species. Based on collection locations of 38 museum specimens, the Gurney's Pitta was known to be limited to semi-evergreen rainforest, located close to streams and gullies below 160 m in elevation, and using spiny palms at least 1 m tall for nesting. Conversion of lowland forest for crops, oil palm, and rubber deforested Gurney's Pitta habitat and caused major declines in their populations. By the early 2000s it was believed to be extirpated from Myanmar and on the verge of extinction, with only 11 pairs known to survive in a wildlife sanctuary in Thailand—making it one of the rarest birds in the world. Scientists from conservation nongovernmental organizations (NGOs) prepared to conduct ground surveys in Myanmar to determine whether there were any left in the country. Using land cover and deforestation data analyzed from satellite imagery along with vegetation maps, elevation, and slope data, wildlife ecologists were able to model potential suitable habitat based on known habitat requirements using GIS. Based on this map, field survey teams deployed to five sites identified as potential habitat areas where the birds had not been previously observed. The team found Gurney's Pitta at four of the five new sites and, combined with surveys at four historical sites, demonstrated that Gurney's Pitta survived in Myanmar, doubling the known global population. The findings provided justification for official protection of Lenya, an area proposed for national park status, and further spatial analysis assessed deforestation and mapped potential threats to aid in planning for conservation of the species.

Home Range

The concept of home range is foundational to understanding an individual's or species' use of space. An individual's home range is defined as the spatial area that encompasses its typical activities, including searching for food and shelter, predator avoidance, and

reproduction. All these activities are connected through an individual's movements. Though the concept is straightforward, there are complicating factors that need to be evaluated to obtain a home range that is biologically meaningful for a species, for example, assessing how to handle exploratory movements that are beyond the area of typical activities and determining how to exclude areas that are not being actively used but only crossed when an individual moves between areas of use. The home range can be assessed over 1-year or multiple years, or could target critical time periods such as a particular season or migration period.

Home ranges are typically based on location data obtained via radio telemetry or satellite/GPS tracking and there are various methods for delineating home range (e.g., simple convex polygons, adaptive or fixed kernel distributions, and local convex hulls are commonly used). However, more recent research shows that these methods are likely poor estimates that are biased by spatial autocorrelation, and estimates can be improved with an area-corrected kernel density estimation method. Recently, rapid advances in tracking (e.g., satellite/GPS collars, transmitters, tags, etc.) have made it possible for ecologists to obtain location data more frequently, over vast areas of land and sea, and for many years. In the past, tracking devices could only be used on medium or large mammals and birds, but miniaturization of devices, development of other remote detection devices (e.g., radar, fixed wing receivers), and improved battery life have allowed ecologists to study a broader range of species within a broader range of ecosystems. These technologies help increase the accuracy and reliability of home range and habitat selection analyses. Increases in availability and variety of satellite data along with improved tracking technology have elevated focus on the study of movement within wildlife ecology. Movement is fundamental to most ecological processes and relates to questions about impacts of habitat fragmentation, invasive species, wildlife disease, and climate change.

Traditionally, delineating home ranges is based on the assumption that they are clearly defined and established, though they may change seasonally. Many species are resident, and remain within a particular area throughout the year. Others species migrate in seasonally predictable patterns between two or more areas. However, there are some species which range widely over an annual period, not returning to any particular location and moving in a way that is seemingly not predictable through a movement system called nomadism. Understanding a species' movement system can be critical to developing effective wildlife management plans. especially for large ranging and rare species.

Monarch Butterflies: Challenges for Managing Migratory Species

A classic example of migratory behavior is the monarch butterfly, a species that, in the cold regions of North America, escapes the winter through massive movements south which create visual phenomena. Starting in October, thousands of monarchs in the north travel to Mexico and Southern California, sometimes over 5000 km, for the fall and winter months. Groups return to the same wintering grounds and sometimes the same tree for hibernation, even though they are not the same individual butterflies that came the year before. Wildlife ecologists continue to study how monarchs find their winter habitat; it appears to be a combination of the magnetic pull of the earth and the direction of the sun. Migration of species causes complexity in management planning for a rare species. Threats may be present in both summer and winter habitats as well as throughout migratory pathways, requiring research and conservation efforts in more than one country, sometimes separated by thousands of miles.

Territories are the subsets within a home range which are actively defended either by an individual, a pair, or a group of individuals. These areas support essential resources, space, mates, or offspring so that the efforts to maintain the territory are offset by the advantages of having control over the area. An individual, pair, or group dominates within its territory. Territories may be marked via scents or visual identifiers to warn potential invaders in an attempt to avoid conflict that uses energy and risks harm to individuals if a fight results. Not all species maintain territories. In some cases a species will form a social group that roams together and essentially functions as a moving territory, such as an ungulate species which form harems.

Dispersal and Distribution

Dispersal behavior refers to the movement of an individual animal from the location of its birth to its location of reproduction and differs from the cyclic patterns of migration behavior. Dispersal behavior, or the lack of dispersal, results in the distribution of a species or population. Distribution refers to the spatial arrangement or area occupied by a species or species' population. Dispersal and distribution help wildlife ecologists understand a fundamental question of ecology—why a species or populations of species are found where they are. Dispersal behavior can be instinctual or in reaction to something in the environment. There are three major causal drivers for dispersal—to improve opportunities to successfully compete for mates, to avoid breeding between closely related individuals, and to move away from areas where the population density and competition are high in favor of areas with more available resources. Whether an individual disperses or not can also be influenced by mating systems. For example, for species in which males are polygynous (having more than one mate), the males will often disperse to increase their chance to find new mates, while the females of the same species prioritize increased opportunities for obtaining resources. These females will usually not disperse because they have a better chance to find resources in areas which are familiar to them.

The distribution of a species is often limited by abiotic factors, such as temperature, precipitation, and climate. For example, reptiles and amphibians are especially affected by cold and are not found in arctic ecosystems. In general, species diversity decreases with increasing latitude (toward the poles) and increases with decreasing latitudes (toward the equator). Many species will move up and down in elevation in response to seasonal temperature changes. Some species are limited by temperatures that are too high, or areas that are too wet. Extremely dry environments, such as deserts, often do not have enough precipitation to support many species. Species only survive in these areas through behavioral or physical adaptations that make it possible to

survive, such as restricting foraging to nighttime, changing from eating grasses to succulents to increase moisture intake, and lighter colored coats to help reflect rather than absorb the sun's rays.

A species' distribution is limited by biotic and abiotic factors, and in general, species that have higher levels of abundance in a local population will have wider geographic distributions, while species that have lower levels of abundance and are more rare will have more narrow distributions. Studying a species' historic, current, and even predicted future range is a major focus of wildlife ecology. Understanding why a species distribution has changed or assessing how a species distribution might be expected to change can inform wildlife management decisions to help sustain a species in the face of increasing or changing threats.

Wildlife Genetics

Genetic diversity, the variability of genes within a species, represents the potential for of a species population to survive in the face of changing landscapes and increasing threats. When changes occur, a population with higher diversity will have a higher level of fitness, because there will be a better chance of having individuals with traits that can aid in their survival. If individuals are very similar (have low genetic diversity) the chances of the population having the traits needed to survive new conditions is lower, and can lead to reduction or extirpation (elimination) of the population from an area when conditions change.

When genetic diversity is reduced it will stay at the lower level or continue to be reduced, unless there is an addition of individuals from a population with different diversity, or over long time periods through chance mutations. When a population goes through a drastic reduction it is typically referred to as a population or genetic "bottleneck". Having fewer breeding individuals can lead to inbreeding (breeding between closely related individuals), which results in a higher likelihood of deleterious traits in the offspring and decreased fitness of the population overall. Small populations are also at higher risk of genetic drift, in which genetic diversity is reduced further due to random chance since few individuals are reproducing and passing on genes. The founder effect is an example of genetic drift that occurs when a small population separates from a larger population due to habitat fragmentation or other changes, and results in a population that is lower in genetic diversity and potentially very different from the original population. In some cases it can lead to speciation, the evolution of a distinct species.

Small populations, at higher risk due to their lower numbers and genetic diversity, must be carefully managed to maintain as much genetic diversity as possible. Population genetics research allows ecologists to compare genetics between populations that have become disjunct (genes no longer moving between them), assess the genetic structure and demographics of populations, and examine the progression of evolution and speciation. Habitat fragmentation has become one of the top threats to wildlife species, reducing available habitat and causing populations to become increasingly smaller and disjunct. In response, mapping land cover change and habitat fragmentation, assessing what causes barriers for a particular species or set of species (e.g., roads, unsuitable habitats, etc.), and developing strategies for maintaining connectivity among populations or restoring connectivity by developing corridors have become important areas of study in wildlife ecology.

Advances in genetic sampling techniques have greatly increased opportunities for applications in wildlife ecology. In the past, DNA was primarily extracted directly from captured wildlife through samples of blood or tissue. More recently, scientists have developed DNA extraction methods that can be accomplished noninvasively through hair, feces, skin, feathers, and other material that can be collected without direct contact with wildlife. DNA sampling allows wildlife ecologists to detect rare species, classify gender, and improve estimates of population size through mark-recapture studies.

Emerging Issues

We are currently facing the sixth mass extinction crisis in over 500 million years. Though some species go extinct due to natural processes, scientists estimate we are now losing species at 1000 to 10,000 times the typical rate, with an estimated 30,000 species going extinct each year. While the five previous extinction crises were brought about by major physical processes, the current crisis is attributed to human activities which have led to extensive loss and fragmentation of wildlife habitat, the spread of nonnative species within ecosystems, and climate change. Several wildlife ecology fields of study have emerged in response to the extinction crisis.

Effects of Climate Change on Wildlife

Increasing temperatures and changing precipitation patterns will have major impacts on ecosystems and the wildlife they support. As climate warms terrestrial habitat regimes are predicted to shift away from the equator toward the poles and to higher elevations, leading to changes in the distributions of wildlife species dependent on those habitats. We can already observe these impacts in some taxa, for example, in addition to shifting distribution patterns some bird species are changing migration routes, the timing of their migration, and locations of breeding grounds in response to the changing temperature patterns and availability of food resources. Disruption of natural ecological processes like migration can have cascading effects beyond the species initially impacted, because changes for one species can lead to changes in the dynamics of relationships among species. Terrestrial mammals cannot move as easily as birds, which can make it more challenging to adapt to the effects of climate change and changing landscapes. As natural areas are developed for transportation networks, expanding urban areas, and other human activities, this forms real barriers that could prevent animals from shifting into areas that have become suitable due to climate change.

Impacts of Climate Change on Lobsters

As oceans warm the latitudes and water depths also impacts marine and coastal wildlife species. Marine species are typically limited in the range of temperatures they can tolerate and many will need to move deeper and away from the equator to find suitably cold water. For example, changes in coastal water have pushed lobsters northward from southern New England, leading to a near collapse of the industry in Connecticut, Rhode Island, and New York, while Maine has seen more than double the number of registered lobster landings since 1994. As ocean temperatures continue to rise lobsters will continue their northward movement, eventually causing decline of the Maine industry as well. Less mobile marine species, such as coral, which cannot tolerate higher temps often do not survive, resulting in the bleaching of coral reefs. This phenomenon is already being observed around the world and even threatens the Great Barrier Reef.

In the face of imminent changes to climate, wildlife ecologists are working to monitor and predict these impacts in order to adapt management strategies to help mitigate negative effects on wildlife species and their ecosystems.

Invasive Species and Diseases

Globalization, along with increased and more rapid transportation, has led to increased travel and trade throughout the world. These interconnections have resulted in the introduction of nonnative, also called invasive, species to new ecosystems. In some cases these species survive and become superior competitors, leading to their rapid population growth which can degrade habitat, increase competition and predation to the detriment of native wildlife species, and set the system out of balance. It is estimated that invasive species have contributed to the extinction of 40% of species lost since the 17th century. In most countries invasive species have been detrimental not only to the environment, but costly in terms of impact to human health and the economy. Costs from pests introduced to the United States, United Kingdom, Australia, South Africa, India, and Brazil are estimated to be more than $100 billion annually.

Invasive House Sparrows and the Decline of Native Birds

Well-known examples of the impacts of invasive species include the intentional introductions of bird species from Europe to the United States. For example, beginning in the 1850s imported house sparrows were released in the northeast by various groups hoping to either reduce insect pests or bring birds that reminded immigrants of their homeland. Having evolved in different ecosystems, house sparrows did not have a natural predator and spread rapidly across the North American continent, except in far northern regions, growing to an estimated 150 million individuals. House sparrows do very well using human structures for nests, have a generalist diet, are productive breeders, and nest early in the year leaving fewer nesting sites available for native migrant species. They are aggressive birds that move other nesting birds out of nests by destroying eggs and nestlings. These characteristics make house sparrows superior competitors, leading to rapid increases in their numbers and the decline of native bird species such as American robin, chickadees, flycatchers, thrushes, tanagers, bluebirds, purple martins, and various sparrow species. Wildlife ecologists have developed strategies for supporting native species by building nest boxes designed to discourage nesting of house sparrows, for example, avoiding foods that attract house sparrows, placing boxes away from high traffic areas, keeping the box closed until native species returns from migration, and trapping out house sparrows if they become established. These strategies have helped reestablish native species, though they require continued maintenance. House sparrow populations have declined in recent years due to these types of interventions and other factors.

Globalization and invasive species have also increased threats to wildlife through facilitating the spread of diseases new to an area, for example, the invasion of the Asian tiger mosquito in the United States has been linked to the spread of the Zika virus. Wildlife trade and changes in land use bring humans and wildlife into contact, allowing diseases to spread among them. Out of all the diseases afflicting humans worldwide, an estimated 75% are zoonotic (originating from animals), for example, Ebola, SARS, influenza, and HIV/AIDS. With increasing interactions between humans and animals comes increased risk of disease and the need for research to understand and predict potential outbreak hotspots.

Wildlife species can also be devastated by disease outbreaks and it's estimated that they have had an even greater impact than invasive species. Humans are little prepared to intervene in these emergencies. For example, a disease called white-nose syndrome is a fungus that grows on the noses of bats in N. America while they are hibernating. It has spread rapidly and has killed over 5.7 million bats in just a decade. In some places it has wiped out over 90% of the bat fauna. The loss of bats could impact human food availability, it is estimated that in Wisconsin bats eating pests save $600 million to $1.5 billion by reducing the need for pesticides. Wildlife ecologists are working urgently to diagnose and surveil the diseases that impact wildlife and find management solutions for slowing the spread of the disease.

Human–Wildlife Conflict

Human populations continue to grow and encroach further into wildlife habitat, bringing increased opportunities for humans and wildlife to interface, which can result in human–wildlife conflict. As wildlife habitats are converted to agriculture and human residence the availability of food resources for wildlife is reduced, leading some wildlife species to seek forage from crops, gardens, and even garbage. This interface usually causes negative impacts on both sides, including injury or death of humans and wildlife, harm to human livelihood through livestock predation and damage to crops, destruction of homes, and potentially leading to the loss of populations of some species. Examples include baboons killing juvenile cattle in Namibia, Asian and African elephants

eating crops in countries across their ranges, orangutans eating palm oil plants in Indonesia, wolves and bear killing livestock in Europe, and deer foraging in gardens in the United States.

Historically, management techniques targeted removal or reduction of the wildlife populations involved in conflict through hunting, translocation, or other methods. More modern approaches seek to find coexistence among humans and wildlife, developing strategies for minimizing and mitigating conflict in ways that are not detrimental to wildlife species. Potential wildlife management solutions include fencing to keep wildlife away from crops and homes, careful land use planning, community management of natural resources, providing incentives for support of wildlife through ecotourism and wildlife friendly products, and human behavior modification such as avoiding dangerous times and places that are at high risk for wildlife encounters. Solutions must be designed for the community and situation as no one solution will work in all cases. For example, communities in Mozambique found that African elephants avoided chili pepper plants and began growing more as a means to deter them. Conversely, scientists in Sri Lanka report that Asian elephants there eat chili pepper and they are not deterred.

Restoring Species and Habitat

In the face of the Earth's sixth mass extinction crises, the field of wildlife ecology has expanded beyond efforts to maintain wildlife species and populations, broadening applications for restoring species and their habitats after they have been lost or greatly reduced. Restoration ecology research focuses on ecosystems that have been degraded or lost due to human activities, to understand impacts to ecological processes and species communities in order to inform strategies for restoring habitat for wildlife and ecosystem services (benefits to humans derived from healthy ecosystems). Examples of ecosystem restoration projects include planting of native trees and shrubs which have important roles in a damaged ecosystem; removal of invasive species through cutting, burning, or poisoning; and altering drainage patterns or soil content to encourage reestablishment of native species and ecological processes.

Beyond maintaining and restoring habitat for wildlife species at risk, wildlife ecologists also work to restore or return species to areas where the species' populations have become very small or extirpated from an area, or even have gone extinct in the wild. Strategies for bolstering or returning a species in an area include relocation, translocation, and reintroduction. When a population is small and faces an imminent threat due to invasive species, disease, habitat destruction, or other human activities that are underway or known to be coming, a population may be relocated (moved within its current home range) or translocated (moved to a new area outside its current home range) to another area where they have a better chance to survive. In extreme situations, such as in cases where there are very few individuals left in the world and they face a high likelihood of extinction or extirpation from an area, decision makers and managers may decide to remove a species from the wild. The objective is to capture the remaining individuals so they can be bred in captivity to build their numbers for eventual return to the wild.

Translocation can also be used as a strategy for bolstering very small populations, by moving individuals from a larger, stable population to a new area within its natural range. Reintroduction is a strategy that can be used to return a species to an area that was formerly part of the species' range, but where it has been extirpated from the area or gone extinct in the wild throughout the natural range. Individuals being released can be moved from another part of the range, but typically they are captive born, providing an opportunity for species from a successful breeding program to return to the wild. These individuals have no familiarity with the flora and fauna of their new environment and may struggle to find resources and survive. It can be especially challenging for a prey species that has not developed strong predator evasion skills. Often individuals are enclosed in the new area for an acclimation period and released, ideally with some form of tracking device so they can be monitoring and studied. When a species is reintroduced the threats which drove it to extinction are usually still present to some degree. Survival of these populations requires continued efforts to study and adapt management plans to help bolster the population, mitigate threats, and achieve coexistence with nearby human communities.

Black-Footed Ferrets: Back from the Brink of Extinction

Black-footed ferrets were thought to be extinct by 1979, due to humans targeting their prey (prairie dogs) and diseases such as canine distemper and sylvatic plague. When a population was discovered in Wyoming during 1981, the decision was made to capture any remaining ferrets to breed in captivity. Eighteen individuals became genetic founders of what has resulted in over 7000 kits born in six breeding centers. Based on captive breeding successes the United States Fish and Wildlife Service began reintroducing black-footed ferrets at eight sites in the western United States, one site in Mexico, and one in Canada between 1991 and 2009. Over time many sites were added by other organizations, with 24 reintroduction sites in total. There were many challenges, including lack of suitable habitat and prey, continued threat of disease, ferrets lack of experience catching prey, and opposition from land owners concerned about any efforts to increase prairie dog numbers or attempts to apply endangered species legislation to their land. The reintroduction effort has been slow and complicated, but through the work of governmental and nongovernmental land managers, scientists, and members of the public, black-footed ferrets are no longer extinct in the wild, with an estimated 500 individuals in release. Only four of the release sites are considered to have self-sustaining populations, and all wild populations will continue to be dependent for survival on conservation efforts to mitigate threats for the foreseeable future.

See also: Human Ecology and Sustainability: Biophilia; The Anthropocene. Terrestrial and Landscape Ecology: Forestry Management

Further Reading

Acevedo-Whitehouse, K., Duffus, A.L.J., 2009. Effects of environmental change on wildlife health. Philosophical Transactions of the Royal Society B: Biological Sciences 364, 3429–3438. doi:10.1098/rstb.2009.0128.

Bakun, A., Black, B.A., Bograd, S.J., García-Reyes, M., Miller, A.J., Rykaczewski, R.R., Sydeman, W.J., 2015. Anticipated effects of climate change on coastal upwelling ecosystems. Current Climate Change Reports 1, 85–93.

Dickman, A.J., 2010. Complexities of conflict: The importance of considering social factors for effectively resolving human–wildlife conflict. Animal Conservation 13, 458–466. doi:10.1111/j.1469-1795.2010.00368.x.

Eames, J.C., Hla, H., Leimgruber, P., Kelly, D., Aung, S.M., Moses, S., Tin, S.N., 2005. The rediscovery of Gurney's Pitta (*Pitta gurneyi*) in Myanmar. Bird Conservation International 15, 3–26.

Fleming, C.H., Fagan, W.F., Mueller, T., Olson, K.A., Leimgruber, P., Calabrese, J.M., 2015. Rigorous home range estimation with movement data: A new autocorrelated kernel density estimator. Ecology 96, 1182–1188. doi:10.1890/14-2010.1.

Fryxell, J.M., Sinclair, A.R.E., Caughley, G., 2014. Wildlife ecology, conservation, and management, 3rd ed. Hoboken, NJ: Wiley Blackwell.

Howard, E., Davis, A.K., 2009. The fall migration flyways of monarch butterflies in eastern North America revealed by citizen scientists. Journal of Insect Conservation 13, 279–286. doi:10.1007/s10841-008-9169-y.

Ichimura, T., Uemoto, T., Hara, A., Mackin, K.J., 2014. Emergence of altruism behavior in army ant-based social evolutionary system. SpringerPlus 3, 712. doi:10.1186/2193-1801-3-712.

Krausman, P.R., Leopold, B.D. (Eds.), 2013. Essential readings in wildlife management and conservation. Baltimore, MD: Johns Hopkins University Press.

Mueller, T., Fagan, W.F., 2008. Search and navigation in dynamic environments—From individual behaviors to population distributions. Oikos 117, 654–664. doi:10.1111/j.0030-1299.2008.16291.x.

Nathan, R., 2008. An emerging movement ecology paradigm. Proceedings of the National Academy of Sciences of the United States of America 105, 19050–19051. doi:10.1073/pnas.0808918105.

Ripple, W.J., Beschta, R.L., Fortin, J.K., Robbins, C.T., 2014. Trophic cascades from wolves to grizzly bears in Yellowstone. Journal of Animal Ecology 83, 223–233. doi:10.1111/1365-2656.12123.

Shaw, L.M., Chamberlain, D., Evans, M., 2008. The house sparrow (*Passer domesticus*) in urban areas: Reviewing a possible link between post-decline distribution and human socioeconomic status. Journal of Ornithology 149, 293–299.

Urban, M.C., 2015. Accelerating extinction risk from climate change. Science 348, 571–573.

Cellular Automata

AK Dewdney, University of Western Ontario, London, ON, Canada

Introduction

Somewhere between ecology and computer science a nascent science struggles to be born. Cellular automata provide a simple, yet flexible platform for simulating a large variety of phenomena, some of which resemble ecological processes, at least in a wider sense. One thinks of chemical oscillators, seashell patterns, and epidemics, among other things.

Whether one is discussing lively chemical solutions, seashell patterns, or epidemics, the question inevitably arises as to what degree cellular automata model ecological processes in a useful (i.e., predictive) manner. This crucial point is discussed in the final section of this article.

Cellular Automata in General

The term `cellular automaton' hints at the marriage of two concepts, automata and cellular space, the latter being essentially an infinite square lattice. Automata *per se* have been the subject of a vast amount of research into their computing powers, particularly the languages they produce or recognize. The theory of automata has been a core subject in computer science from the beginning. It must be stressed that finite automata have powers of computation that are severely limited in comparison with a fully programmable computer.

Finite Automata

A finite automaton consists of a finite set of states, an input tape and an output tape (which may be identical to the input tape, if desired). Time is divided into discrete units and, between ticks of the clock, the automaton remains in the same state. At each tick, the automaton enters a new state that depends on its current state and its input symbol. Normally we express the rules that govern this behavior in a table or a state-transitiuon diagram (Fig. 1). The figure shows a simple finite automaton represented in both ways.

The state-transition diagram consists of three nodes labeled with state names and transition arrows between them. The corresponding table consists of two columns (one for each possible input symbol) and three rows (one for each state). Below the diagram and table is a schematic drawing of the automaton, basically a simple visual context in which to imagine its operation (see Fig. 1).

The finite automaton pictured here has three states and always writes the same input symbol as it reads, making only one tape necessary. For example, in state zero, if the automaton happens to be reading a 1, it enters state zero, according to the state-transition diagram. (Simply follow the arc labeled 1.) According to the table, it does the same thing. (Simply examine the entry in column 1 and row 0.) Once the transition is complete, the clock advances by one tick and the input tape shifts past the read-head by one square.

The input symbols do not have to be fed to a finite automaton one cell at a time. For example, inputs might consist of triples, as in 010, 011, 111, etc., or even longer strings. In such an automaton, for example, there might be a transition from state 2 to state 5 if the input was 011, but the transition might well be to other states for the remaining seven possible inputs. Finally, although care is normally taken not to confuse tape symbols with state names, we are about to enter a context where they are identical.

Definition and Examples of Cellular Automata

Although many variations exist, the basic cellular automaton consists of a large rectangular grid of squares (called cells), a clock that ticks, and a finite automaton. We imagine that each cell has a copy of that finite automaton in it. Although the same finite automaton inhabits each cell, the automata so embedded do not have to be in the same state. The input for each automaton is not on a tape, but in the neighborhood of surrounding cells. Depending on how the `neighborhood' is defined, the inputs will be strings of four state labels or eight of them. An example of this particular kind of finite automaton is provided later.

In a cellular automaton all cells typically begin in state 0, except for a finite number that are in other states. The nonzero patterns that occur while a cellular automaton is running are called `configurations'. At each tick of the clock, many of the cells enter a new state and a new configuration develops. It is natural to refer to the sequence of configurations that develop as `generations'.

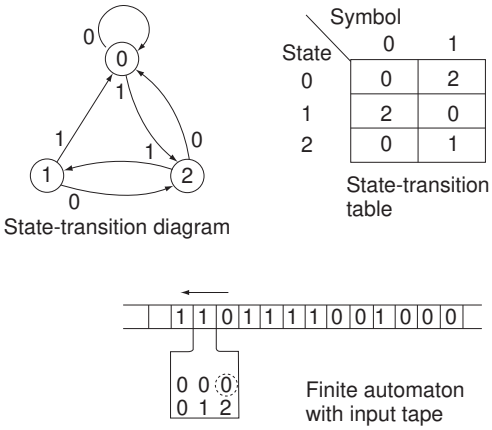


Fig. 1 Table and diagram for a finite automaton.

Perhaps the most famous cellular automaton to date is the one discovered by Cambridge mathematician John Horton Conway. He called it the Game of Life because it produced lifelike phenomena, as we shall presently see.

Conceptually, the space of Conway's Life is a two-way infinite grid of cells each of which can be in one of only two states, `alive' or `dead'. We will give these states the rather less exciting names of `0' and `1'. In this particular cellular automaton, each cell is considered to have eight neighbors, the four cells adjacent to the sides of the square plus the four cells adjacent at the corners. A cell in state 0 at a given tick of the clock will remain in state 0 at the next tick unless it has three living neighbors presently. In that case it will enter state 1 at the next tick, coming `alive', so to speak. A cell that is currently in state 1 will remain in state 1 at the next tick unless it has fewer than two neighbors in the state or more than three. Conway likens these rules to starvation (fewer than two living neighbors) or overcrowding (more than three).

These relatively simple rules astonished the mathematical and computing communities when first published in the journal *Scientific American* in 1970. No one would have predicted that such simple rules could lead to such complicated (and lifelike) behavior.

Fig. 2 shows a simple behavior of the Life cellular automaton in which an initial configuration of five cells in state 1 changes from one tick of the simulation clock to the next, turning into a somewhat larger configuration of nine cells in state 1 after five ticks of the clock. One can choose a particular cell in any of these stages and verify that its state in the next generation is the direct result of the appropriate rule as applied to the current generation.

An important feature in the Game of Life is the existence of `gliders', self-perpetuating patterns that `glide' across one's screen. **Fig. 3** shows the four consecutive configurations that constitute a glider. It will be noted that with every four generations, the glider has moved one cell diagonally in a direction that is determined by the orientation of the glider.

There is even a large configuration that produces gliders, requiring 30 generations for each glider to appear.

Even in the relatively simple milieu of a square-grid cellular automaton, a huge variety of rules is possible. The kind of rule used by Conway is called totalistic because state transitions are based on simple counts of neighboring cell states. However, totalistic rules account for only a tiny fraction of all possible rules, as applied to a particular space.

Totalistic rules can be easily implemented by a finite automaton in this context. For example, the totalistic rule for a cell in state 0 to enter state 1 requires that exactly three of the eight neighboring cells be in state 1. To manage this with a state-transition table is straightforward. If we number the neighboring cells in a fixed manner 1, 2, … , 8, the set of all strings in which exactly three 1s appear yield a list of transitions that one could begin as follows:

```
input       00000111 00001011 00001101
. . .
next state      1        1        1
. . .
```

Although cellular automata are generally conceived as inhabiting an infinite grid, the practicalities of finite computer memories impose a finite grid. To rid themselves of the boundary effects that this restriction inevitably imposes, many workers use a wraparound space in which opposite borders of the (finite) grid are identified, that is, counted as neighbors.

The underlying geometry of the cellular space need not be rectilinear. The cells may be hexagonal, for example, as in a honeycomb. (This style of cellular automaton has the advantage of having only one kind of neighbor, rather than two, as in the case of the square grid.) Nor is the underlying dimensionality of the underlying space restricted to just two dimensions. For example, there is a successful version of the Game of Life in three dimensions. In this version the cells are cubical.

The simplest possible cellular automata inhabit a linear (one-dimensional) space in which each cell has two neighbors, one to the right and one to the left. The rules for such automata are much simpler than for those having higher dimensions. In fact, for a two-state linear cellular automaton there are only four possible combinations for a neighborhood of two cells. Thus, the table would have at most four entries for each state. Yet even linear cellular automata may exhibit startling properties.

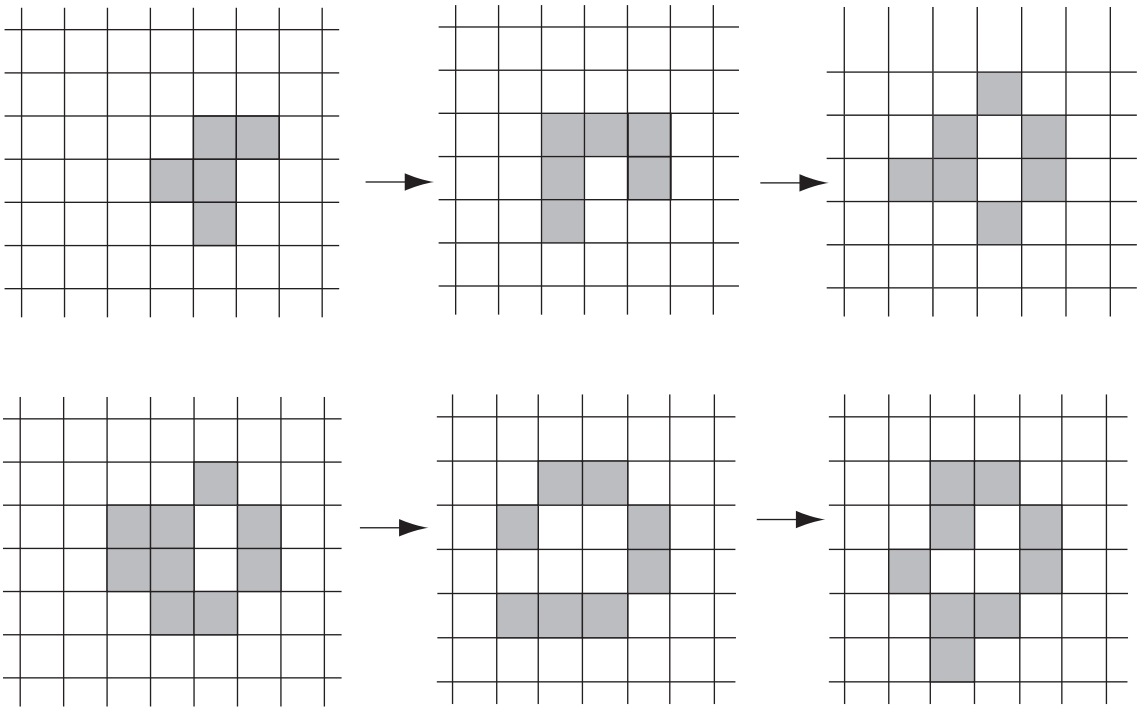


Fig. 2 Six generations.

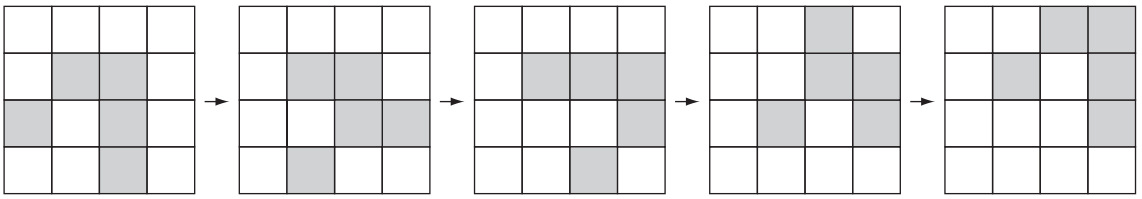


Fig. 3 Glider.

The most famous come from Stephen Wolfram, a physicist who became convinced that cellular automata amounted to a new paradigm for biological science and possibly for physics itself. Having devised a numbering scheme which encoded the many possible rule sets, Wolfram discovered Rule 30, as written according to Wolfram's scheme:

neighborhood	000	001	010	011	100	101	110	111
nextstate	0	1	1	1	1	0	0	0

The pattern of 1's and 0's, when written in reverse order, yields the binary representation for 30.

This cellular automaton, when started with an initial configuration consisting of a single cell in state 1 (the rest being in state 0), produces a succession of generations that appear to be essentially random, as in the accompanying illustration, which tracks a mere 12 generations (Fig. 4).

To this point deterministic rules have been assumed. The state of each cell depends on its own current state, as well as the states of its neighbors. The same configuration of states always gives rise to the same next state in a deterministic cellular automaton. In a probabilistic cellular automaton, the next state of a cell depends on probabilities associated with various configurations of its neighbors.

In general, probabilistic rules have the same form as deterministic ones except that the general statement would change from

Under the conditions X the next state is Y

to

Under the conditions X the next state is Y with probability p.

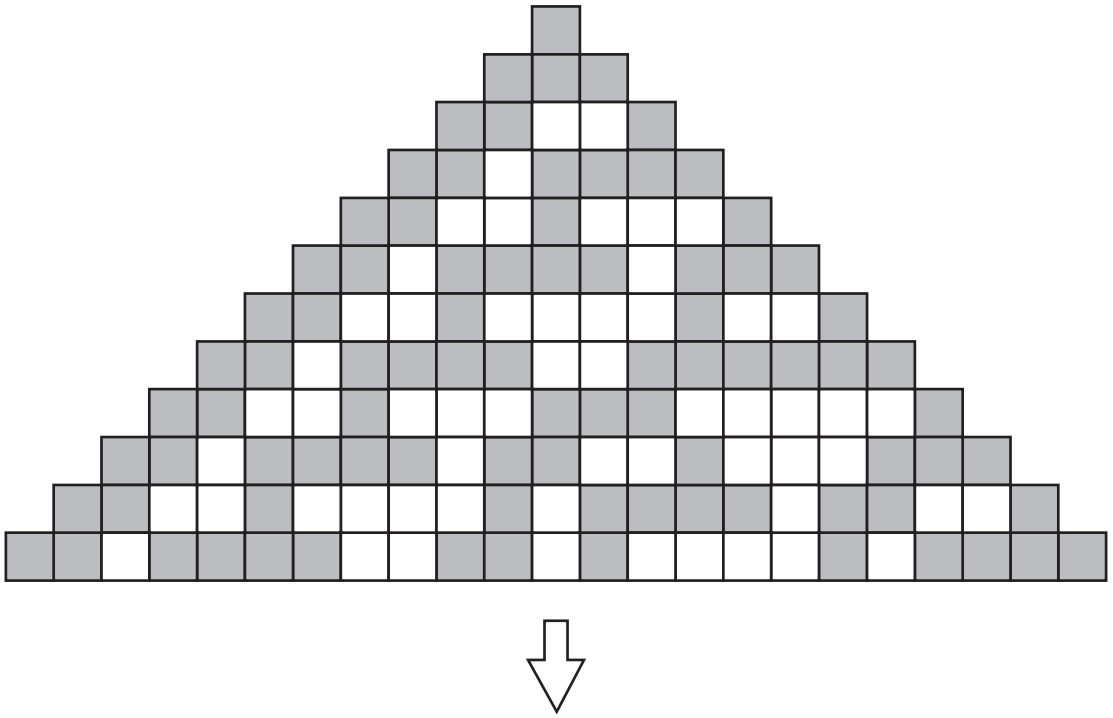


Fig. 4 Output of Rule 30.

Cellular Automata as Ecological Models

It is frequently useful, in building a cellular automaton model, to generalize the rules by making certain numbers available only at runtime. For example, the model might involve totalistic rules but with the totals treated as variables. In other words, the automaton might specify that when a cell is in a certain state, it will enter state 3 if n of its neighbors are in state 2. At run time, the user of the program may explore different behaviors by giving n a different value than he or she did last time. The probabilities used in probabilistic rules may also be made into variables, to be explored at runtime in much the same manner. Such variables, set at runtime, are called parameters of the model.

One can analyze some cellular automata readily enough to make certain predictions about their behavior without using a computer. But with a computer, one can see the rapid development of configurations through time, experiment endlessly with initial configurations, watch their fates, and adjust rules and parameters to explore certain possibilities.

History of Cellular Automata

A Strange Expectation

In the 1950s and 1960s, universities were creating computer science departments. One of the most important early subjects studied in these departments was automata theory, then being developed at the Los Alamos National Laboratory (US) and several universities, worldwide.

There has been a sense of expectation lurking in the halls of computer science from the very beginning. Leading thinkers like von Neumann and Alan Turing had pondered the lifelike properties of automata and the ability of the general purpose computer to simulate practically anything. If a soul did not lurk in the machine, then intelligence did. The field of artificial intelligence developed over the next decade into a persuasive metaphor, at least in the heady early days, with a hint of robots to come and, perhaps, new beings smarter than ourselves. This theme in computer science has had a long history, emerging two decades ago into a new field variously called artificial life or complex systems.

The Early Models

Before his death in 1957, the mathematician John von Neumann conceived of what he called the universal computer constructor, an abstract machine that was not only a fully programmable computer, but which made copies of itself, as well. The machine was designed in a cellular space of square cells that could be in any of 29 different states. von Neumann did not live to see his construction perform since he never completely specified the rules. His work, however, was taken up by the computer scientist E. F.

Codd, who reworked von Neumann's description, reducing the number of states from 29 to just 8 in the process. Codd's machine worked, at least conceptually, and von Neumann's dream was realized.

In truth, von Neumann could not be called the sole originator of cellular automata. Stanislaw Ulam, a fellow mathematician at Los Alomos National Laboratory in the United States, had devised a cellular grid to simulate the growth of crystals in a two-dimensional space. It was he who suggested to von Neumann that the dream of a universal computer constructor was best pursued in what we have now come to know as a cellular automaton.

Earlier, in 1951, Russian biophysicist Boris P. Belousov submitted a journal article that described a strange chemical reaction. When a solution of citric acid, acidified bromate, and a ceric salt was left to stand, the solution would first clarify, then turn yellow. Then it would clarify again, and so on. The paper was rejected because, according to the editor, the reaction was clearly `impossible', according to thermodynamical laws. The world of chemistry was not quite ready for lifelike properties emerging from a simple chemical soup.

Another Russian biophysicist, Anatol M. Zhabotinsky, refined Belousov's solution by replacing citric acid with malonic acid. He discovered that when the new solution was left undisturbed, it did not merely oscillate between two colors. Instead, amazing patterns consisting of concentric circles and spirals played across the surface of the mixture with no sign of ending. Zhabotinsky was ultimately able to publish his strange research, but it was not until 1989, when *Scientific American* carried an article about cellular automata that successfully mimicked the Belousov–Zhabotinsky reaction that the reaction was added to the steadily growing list of cellular automata incursions into the real world. The automaton was discovered more or less simultaneously by American David Griffeath, as well as a pair of German researchers, Martin Gerhardt and Heike Schuster.

These cellular automata are best described using an infection metaphor to express the interactions of neighboring cells. The Gerhardt–Schuster automaton has a great many states, typically 100. Cells with low state numbers are considered `healthy', while those with high numbers are considered `infected'. A weighted sum of infection in the neighborhood of a cell determines the degree of its infection at the next tick of the clock. A set of four parameters associated with the Gerhardt–Schuster cellular automaton suffices to specify the model completely. Depending on how the parameters are fixed at runtime, the cellular automaton will exhibit behavior (appropriate colors being assigned to its states) to produce a screen appearance that is virtually indistinguishable from the reaction itself. Refreshingly, Gerhardt and Schuster have issued a disclaimer that the simulation in any way `explains' the reaction. It is, they say, merely an interesting coincidence.

A Resurgence of Interest

In 1970, the journal *Scientific American* introduced Conway's Game of Life to the world in Martin Gardner's Mathematical Recreations column. Almost instantly, gliders began to flicker across the screens in graduate and faculty offices as scientists, some at least, came to appreciate the powers of cellular automata. The Game of Life was but one example. What other amazing machines awaited discovery?

Strangely, the appearance of Conway's Life came barely a year after the publication of *Calculating Space*, a book outlining the possibility that all space and physical phenomena in it were fundamentally discrete by nature. The appearance of Life gave added credibility to the book's author, Konrad Zuse. Twenty years later a kindred spirit, physicist Edward Fredkin, would maintain that the universe was one vast cellular automaton, a theme that was also being explored by physicist Stephen Wolfram.

In 1983 Wolfram began an investigation of one-dimensional cellular automata, the ones described in the previous section. He quickly discovered, while exploring the effects of different rules, that complexity would frequently arise in an unpredictable manner. In other words, it was never apparent, simply from an inspection of the rules themselves, which would produce complex and interesting generations and which would fizzle out or produce dully repeating patterns. Whatever might be the mathematical situation, Wolfram decided that the complexity he witnessed in the linear automata, produced by simple, interacting rules, had a parallel in the sunflower heads and seashells of nature. But would it be confined to occasional patterns connected with growth or would it, ultimately, hold a mirror up to nature itself? Wolfram, thinking more in the latter direction, published the boldly titled *A New Kind of Science* in 2002.

In the book, Wolfram explored elementary cellular automata, the totalistic linear automata described above. The book drew many significant parallels between patterns in nature and patterns produced by cellular automata. Perhaps the best known of these is the seashell pattern that appears on the cover of his book.

A significant question addressed by Wolfram is whether some elementary cellular automata are computation universal. Although not without significance for ecology, the question is typically asked in a computer science context. For an elementary cellular automaton to be computation universal means that the automaton is effectively capable of being programmed and is logically equivalent to a fully programmable computer. Did the marriage of humble finite automata and a cellular space boost the computing power to universality? Wolfram suspected that Rule 110, for example, was computation universal. Colleague Matthew Cook answered the question partially in 1994, proving that some of the elementary cellular automata had structures rich enough to support universality. Later, Wolfram provided the outline of a proof that Rule 110 is universal. Although the proof depends heavily on pictures and lacks mathematical rigor, Wolfram may well be correct. After all, one can prove that all sorts of simple systems are computation universal. One can, in fact, compute with ropes and pulleys, jets of water directed in a system of channels, systems of words that are assembled according to simple rules, pinball machines, and a host of other unlikely milieu.

One of the most interesting aspects of Wolfram's work with cellular automata involves a classification scheme for elementary cellular automata. When salted with an initial configuration of states, each of these automata exhibits one of four classes of behavior.

Class 1. After an initial flurry of activity, the automaton produces nothing but monochromatic generations (all cells in the same state).
Class 2. Once a class-two cellular automaton settles down, it produces strictly periodic generations. A finite set of configurations (often just one) repeats over and over again, always in the same order.
Class 3. These automata ultimately produce random-looking patterns that never repeat themselves. By the same token, they show little evidence of structure, not unlike snow on a television screen.
Class 4. Some elementary cellular automata produce a mixture of structure and randomness. These clearly interest Wolfram the most, being the most `lifelike'.

Wolfram has found interesting theoretical support for the reality of his classification system. If the state of a single cell is changed from one run to the next, the new pattern will of course be different. In class 1 automata, however, the change is rapidly `forgotten' on the way to uniform statehood. In class 2 automata, the change remains localized without affecting other areas of the pattern. Class 3 automata have a very different behavior. The effects of the change spread to the right and left at a near-linear rate, suggesting that any cell in the initial configuration can have an effect on the state of any other cell, no matter how distant, given enough time. Class 4 automata, on the other hand, show behavior that is intermediate between classes 2 and 3, with occasional bursts of long-distance communication mingled with local dieback.

Class 4 automata are poised midway between classes 2 and 3. At a metaphorical level, they seem to point to a digital future for biology and ecology. Somehow, everything alive will all end up in class 4. At a more practical level, this future can only appear if cellular automata, used appropriately, can make themselves indispensable vehicles for new knowledge that is expressed or expressible in traditional terms. Otherwise, there is a distinct danger, in this writer's opinion, that the cellular automaton will become a toy that generates the occasional `insight', ultimately to be discarded, as nobody knows what the results actually mean.

However things turn out, there are already a great many types of cellular automata employed in the pursuit of ecological insights. These have two- and even three-dimensional cellular spaces, sometimes distorted by landscapes, sometimes inhabited by finite automata but more frequently by miniature programs. The latter kind of model would be called simply and more accurately a cellular model, rather than a cellular automaton model.

Applications in Ecology

It is interesting to compare the present degree of penetration of cellular automaton models into the field of biology (including ecology) with that of a much older mathematical approach, the differential equation. A search on biosis using each term as a key phrase turned up 3993 instances of the latter, as compared with 357 instances of the former. Presumably a tenfold majority of authors still feel more at home with differential equations.

The two tools are markedly different. Differential equations describe continuous behavior, whereas cellular automata describe discrete behavior. Despite the differing conceptual frameworks, differential equations have been used in discrete situations (such as predator–prey relations) and, conversely, cellular automata have been applied in continuous situations (movement of swarms and flocks).

How Good Are the Models?

The essential question to be asked in the area of applications of cellular automata to ecology is: `How useful are cellular automata in generating insight into biological and ecological processes'? Among the most persuasive images in Wolfram's *A New Kind of Science* are those of seashells decorated with patterns that are eerily similar to those produced by some of his elementary cellular automata. Fig. 5 shows a fair rendition of a portion of one of these patterns as though it were produced by such a cellular automaton.

Of course, the surface of seashells are not marked off into tiny squares each of which is either one color or another (black and white in the case above), yet the patterns behave as if it were. On the other hand, if one tries to produce an elementary cellular automaton that actually produces patterns like the one shown here, the project will come to grief. The diagonal lines tell the story. A diagonal one-cell line amounts to a glider, that is, a structure that, in this context, moves either left or right. Although it is possible to have a right-moving one-celled glider – or a left-moving one – it is mathematically impossible to have both.

This problem illustrates the need to ensure that the `application' in question really does what most ecologists demand of their models, that they `generate insight' (a phrase which the author has encountered frequently).

An `insight', to be really useful, must go beyond merely noting a resemblance. It must lead to the discovery of a new structure or process. In the case of seashell patterns, Wolfram points out that the shell pigments are laid down by the edge of the gastropod mantle, essentially one line at a time. The cells, glands, or (as Wolfram calls them) `elements' of the mantle edge that deposit the pigments ``… have certain interactions with each other. And given this, the simplest hypothesis in a sense is that the new state of the element is determined from the previous states of its neighbours – just as in a one-dimensional cellular automaton." If an

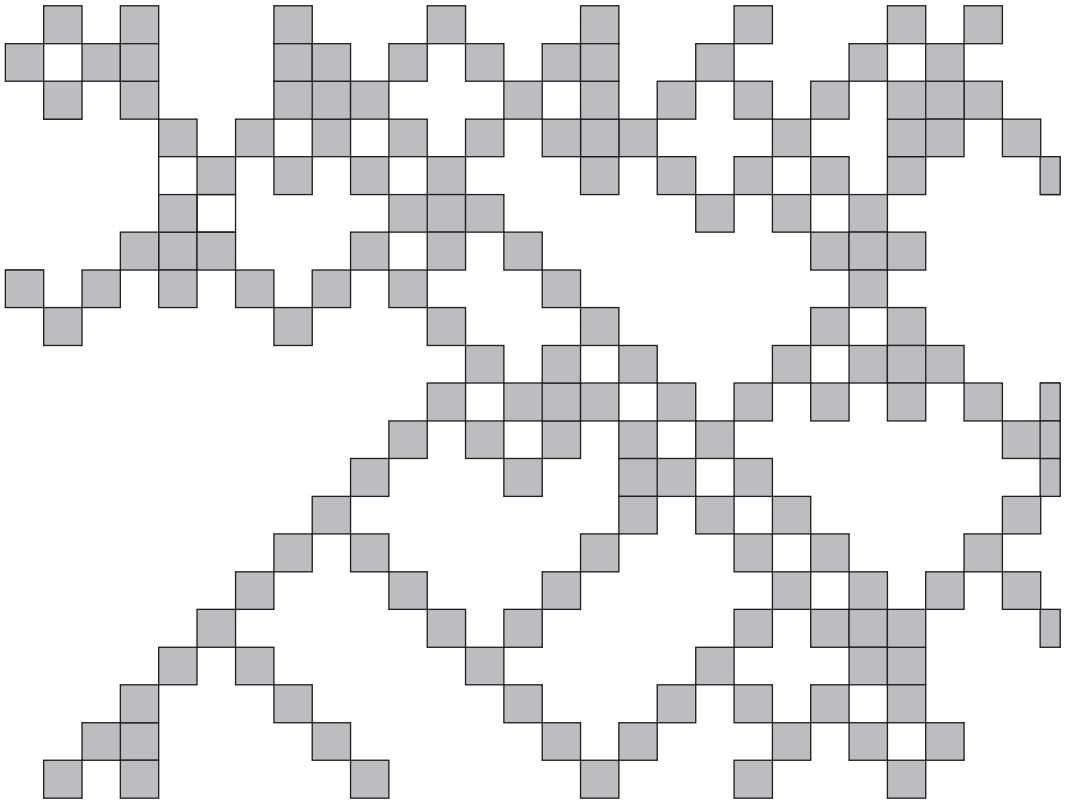


Fig. 5 Seashell pattern.

automatistic behavior is involved in such patterns, it is more complex than anything that elementary cellular automata are capable of.

A Typology of Applications

A review of the cellular automaton literature as it bears on ecology reveals a large range of applications, from abstract models with a somewhat fuzzy relationship to the real world to models that are focused on specific ecological problems with a view to eliciting new processes or structures in the underlying dynamic.

The former end of the spectrum could almost be described as solutions looking for problems. Early in the development of the subject, it became clear that certain cellular automata bore an uncanny resemblance to real-world objects or processes. Wolfram's molluskan patterns were but the tip of the cellular iceberg. A simple three-state cellular automaton captured the dynamic of the famed Belousov–Zhabotinsky (described earlier). Other cellular automata seemed to imitate the coat patterns of bi-colored animals such as zebras and leopards. Still others gave rise to `growths' that resembled root systems or rhizoids.

The most fascinating models to watch, however, were the early predator–prey cellular automata. Abstract predators occupied some of the cells, while prey occupied others. Rules could be formulated not only to allow the animals to pursue or to flee, but to reproduce. In many such models, the populations of both predator and prey oscillated in the manner predicted by the Lotka–Volterra equations. These were differential equations which operated not on discrete life forms, but continuous ones, a certain amount of predator in eternal pursuit of another quantity of prey, with both populations showing smooth, regular increases and decreases. Since the Lotka–Volterra equations were already suspected by many population biologists to capture the essence of simple predator–prey relations, the cellular model promised a more meaningful theater in which specific forms of predation could be examined.

As ecologists became intrigued with the possibilities of cellular automaton modeling, the subject took on a new and greatly expanded life. Of particular interest was the fact that thousands of interacting processes at the cellular (local) level could produce overall (global) changes or effects that might not have been predicted by simply examining the interaction rules. Such `bottom-up' or `emergent' effects include phenomena such as the complex swirling patterns of the Belousov–Zhabotisnsky reaction, the functional relationship between one variable and another, as in predator/prey models with their see-sawing populations, or long-term effects such as the disappearance of structure in Wolfram's first three classes of cellular automata.

Emergent or bottom-up properties of cellular automata not only have a fascination of their own, but motivate much of the work in this area. `What will happen if I change the density of plankton on which the prey fish subsist? Will they become extinct or

will it ultimately reduce the predator population'? It is nearly impossible to tell without adjusting the appropriate parameter and to run the model under the new conditions. Most cellular models in ecological and biological applications allow the alteration of fundamental parameters to explore emergent behavior under a great variety of conditions.

Although all cellular models share an emergent dynamism of one kind or another, the underlying spaces of the models vary in dimension. At present the great majority (90%) of applied cellular automata are two dimensional. The substrates vary from geographical areas such as grasslands to biofilms on submerged rocks, to the skin of animals, indeed any system that has a two-dimensional surface or zone connected with it. The remaining 10% involve a three-dimensional space in which animal assemblages (flocks of birds, schools of fish, myxobacterial associations, and so on), multicellular modeling (tissue growth, embryology, skeletal development), geological processes (water uptake and release, climatic effects, oil slicks), and several other areas with a paper or two each.

By far the largest number of journal articles based on two-dimensional substrates involve landscapes, with approximately equal numbers focused on urban growth/encroachment models and on vegetation dynamics in a great variety of settings. The remaining articles in this category focus on animal populations, from slugs to geese, as well as epidemiological models ranging from infections to alien plant invasions. Increasingly popular are the hybrid models, where programs replace automata, as in the example described next.

An Example Application

The black-legged tick study examined here involves an application of the cellular automaton concept that illustrates a (1) typical focus, (2) relatively successful outcome, and (3) variation in the concept that is increasingly common. An article on simulating the spread of the black-legged tick (*Ixodes scapularis*) in eastern North America by N. K. Madhav *et al*. may be placed in the epidemiology division (sensu lato) of the ecological literature.

The black-legged tick is a major vector of Lyme disease, a debilitating and ultimately fatal disease of humans. In recent decades there has been a noticeable expansion of this tick's range, offering new opportunities for human infection. Lyme disease involves a spirochete bacterium that lives in its tick hosts, then spreads to a variety of mammal and bird hosts when they are bitten by a tick. Principal vectors for the tick are the white-tailed deer, white-footed mouse, and American robin. It would be possible to predict more accurately where the tick range is most likely to expand if it were known which of these hosts played the most active role in its dispersal as a joint function of the number of such hosts and the distance they are likely to disperse the bacterium (mice: short, deer: medium, robins: long).

The study used a cellular grid to represent a large area (473 km $\times$ 473 km) within the tick's eastern range, each cell in the model representing 1 km^2. Each cell is inhabited by a process that is driven by both data and equations. Each process uses data from its own cell and neighboring cells to alter its own data. The process is also strictly deterministic. Notice that the `automaton' has departed entirely, along with its states.

The authors developed a number of hypotheses to test, including, (1) range expansion would be greater in areas dominated by hosts with larger ranges; (2) hosts with smaller home ranges would decrease the expansion of the tick range if they divert a sufficient proportion of ticks from hosts with larger ranges; (3) birds could expand the tick range if their densities and degree of infestation were high enough.

An important parameter in the model was the tick `burden' for each host in the model. Based on field data from the northeastern US, the authors calculated the mean number of ticks found on each of the host species over time. Other parameters included the average time that larvae (3 days), immatures (5 days), and adults (7 days) remained attached to their hosts, namely mice and birds for larvae, mice and birds for immatures, and deer for adults. Also important were the number of days spent by the tick at each stage of its life in seeking hosts.

Each run of the model began with a strip of cells down one side of the grid inhabited by ticks. Not to confuse the reader, each tick of the simulation clock represented one stage in the life cycle of the tick. Thus, three iterations would amount to one life cycle. The model was run for the equivalent of an 18-year period for each combination of parameter values under test.

At each iteration, this cellular model used host densities and their tick burdens in each cell, along with its four neighbors, to calculate the maximum number of ticks that can be supported in the neighborhood. This figure, in turn, yielded an estimate of the actual number of ticks feeding on a host species. The resulting engorged ticks were then distributed within the neighborhood to experience a degree of mortality, the survivors forming the basis of the next iteration.

The experiments produced some interesting results. First, the sensitivity analysis (a key component of all model building) revealed that the cellular model was highly sensitive to tick mortality. In other words, relatively small changes in the mortality parameter could produce large changes in how rapidly tick range expanded. The model was also sensitive to the home range of both mouse and deer hosts, as well as nymphal burdens. The results of the sensitivity analysis revealed the great importance of accurate estimates of certain field data, including on-host tick mortality, yearly tick burdens of host species, host behavior, and home range dynamics. In addition, accurate estimates of range expansion in the field (under appropriate conditions) would be necessary to validate the model.

The main experiment involved many runs of the model, in which each of the parameters was varied over three orders of magnitude. In this manner 125 host density combinations were tested. According to the model runs, deer density was significantly correlated with expansion of tick range, whereas with deer densities held constant at certain values, mouse densities were negatively correlated with tick range expansion. The authors sound a cautionary note that more work is needed, not only to further validate the model, but to make the cellular landscape more realistic by including geographical barriers, such as rivers.

Given that cell-to-cell interactions are represented by relatively realistic functions, the behavior of the model, though still requiring further validation, is correspondingly more realistic. Such a large proportion of work in this area involves nonautomaton modes of operation that the term cellular model would better categorize the research.

See also: Ecosystems: Alpine Ecosystems and the High-Elevation Treeline. Evolutionary Ecology: Fecundity

Further Reading

Codd, E.F., 1968. Cellular Automata. New York: Academic Press.
Dewdney, A.K., 1993. The New Turing Omnibus. New York: Computer Science Press.
Ermentrout, G.B., Edelstein-Keshet, L., 1993. Cellular automata approaches to biological modeling. Journal of Theoretical Biology 160, 97–133.
Fredkin, E., 1980. Digital mechanics. Physica D 45, 254–270.
Gerhardt, M., Schuster, H., 1991. A cellular automaton model of excitable media: IV. Untwisted scroll rings. Physica D 50, 189–206.
Hogeweg, P., 1988. Cellular automata as a paradigm for ecological modeling. Applied Mathematics and Computation 27, 81–100.
Madhav, N.K., Brownstein, J.S., Tsao, J.I., Fish, D., 2004. A dispersal model for the range expansion of black-legged tick (*Acari, Ixodidae*). Journal of Medical Entomology 41, 842–852.
Pollack, J. (Ed.), 2004. Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Artificial Life. Cambridge, MA: MIT Press.
von Neumann, J., 1966. Theory of automata: Construction, reproduction, homogeneity. In: Burks, A.W. (Ed.), The Theory of Self-Reproducing Automata part II. Urbana, IL: University of Illinois Press, pp. 91–381.
Wolfram, S., 2002. A New Kind of Science. Champaign IL: Wolfram Media.
Zhabotinsky, A.M., Buchholtz, F., Kiyatkin, A., Epstein, I.R., 1993. Oscillations and waves in metal-ion-catalyzed bromate oscillating reactions in highly oxidized states. Journal of Physical Chemistry 97, 7578–7584.

Relevant Websites

http://mathforum.org— Cellular Automata, The Math Forum @ Drexel.
http://www.hermetic.ch— Five Cellular Automata: The Belousov—Zhabotinsky Reaction, Hermetic Systems.

Chaos☆

Sven E Jørgensen, Royal Danish School of Pharmacy, Copenhagen, Denmark
Brian D Fath, Towson University, Towson, MD, United States

Appearance of Chaos in Models

Chaos can be explored through mathematical models or other techniques such as recurrence maps and phase diagrams. Model predictions are very sensitive to the initial conditions because the difference between predictions with slightly different initial conditions grows exponentially:

$$dN(t) = dN(0)e^{\lambda t}$$

where $d(t)$ is the difference between the two predictions at time t and $d(0)$ at time zero, and λ is the Lyaponuv exponent. Using mathematical models, the prediction uncertainty depends on the time scale of the system dynamics, referred to as the Lyapunov time. This causes long-term prediction problems because if the forecast time is doubled, then the proportional uncertainty is squared. A forecast horizon two or three times longer than the Lyapunov times is not meaningful because the system appears random. For example, for weather systems the Lyapunov time is thought to be around a few days.

An attractor is the set of numerical values to which the system evolves. These attractors can be fixed points, a finite number of points, a limit cycle, a limit torus, or "strange." Chaotic systems exhibit strange attractors, which have a fractal structure. An example of a strange attractor for the Lorenz equation is shown in Fig. 1. Note the obvious origin of the phrase, "butterfly effect."

Bifurcation and Population Dynamics

Chaos is also known in relation to bifurcation. One of the earliest models used to show this was the logistic model which describes the change in a population given its growth rate and carrying capacity. The following equation yields bifurcation:

$$N_{t+1} = N_t\left(1 + r\left(1 - {}^{N_t}/_{K}\right)\right)$$

where r is the growth rate per capita, t is the time, N the number of organisms, and K the carrying capacity. When r is above 2, the equation gives two stable fixed points; when above 2.6, four stable fixed points; when above 2.75, eight stable fixed points; and through successive bifurcations, an infinite hierarchy of stable cycles of the period $2n$ arises. Mathematically, it is clear that a high growth rate quickly pushes up against the carrying capacity causing this widely fluctuating pattern. Ecologically, chaos is rare in ecological systems because such high growth rates are not evident. Numerous researchers have demonstrated chaos in mathematical equations representing population dynamics. These studies have given support to theoretical questions of population dynamics. For example, the role of a top predator in a food chain can show chaotic behavior (Rai and Upadhyay, 2004), and the coexistence of a stable outcome for all three species in a two-prey and one-predator model occurs due to a strange attractor (Groll et al., 2017), among others. However, a few examples show that chaos may occur in insect populations and other species. A well-known field-based empirical example is the lemming population, where density (number per hectare) may fluctuate between two bifurcation values. Bifurcation pattern is also evident in seashell formation, as nicely demonstrated by Hans Meinhardt in his book *The Algorithmic Beauty of Sea Shells* (Fig. 2). Further work is needed to match the theoretical/mathematical work with empirical studies, and to explore the appearance and occurrence of chaotic behavior in population dynamics of real ecological systems.

Ecological Systems at the Edge of Chaos

A related concept that builds on chaos and complexity theory is the "edge of chaos," which is used to describe the transition zone between order and disorder. Systems operating in this region exhibit rich dynamical behavior and the concept has proven useful for understanding the evolution and adaptation of biological systems. For example, the abundance of other species determines which growth rate is optimal for any organism. If the growth rate is too high, then the resources (food) will be depleted and growth will cease. If the growth rate is too low, then the species does not utilize the resources (food) to the extent that is possible. If, in a well-

☆*Change History*: March 2018. Brian D. Fath prepared the update, revising the text in all sections and adding Figures 1 and 2.

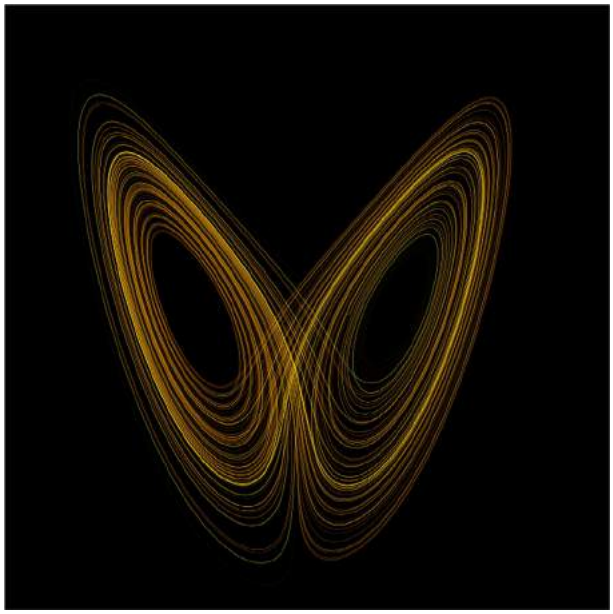


Fig. 1 A plot of Lorenz's strange attractor for values $\rho = 28$, $\sigma = 10$, $\beta = 8/3$. Figure in public domain by User:Wikimol, User:Dschwen—Own work based on images Image:Lorenz system r28 s10 b2–6666.png by User:Wikimol and Image:Lorenz attractor.svg by User:Dschwen, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=495592.

Fig. 2 A conus textile shell, similar to a model with chaotic behavior. Photo in public domain by Richard Ling—Own work; Location: Cod Hole, Great Barrier Reef, Australia, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=293495.

calibrated and validated eutrophication model (state variables include phytoplankton, nitrogen, phosphorus, zooplankton, fish, sediment nitrogen, and sediment phosphorus), we vary the zooplankton growth rate, then work energy capacity will show a maximum at a certain growth rate (which is frequently close to the value found by the calibration and approved by the validation). At both lower and higher growth rates, the "average" work energy capacity is lower because the available phytoplankton is either not utilized completely or is overexploited. When overexploitation occurs, the phytoplankton and zooplankton show violent fluctuations. When the resources are available, the growth rate is very high but the growth stops and the mortality increases as soon as the resources are depleted, which gives the resources a chance to recover and so on. Somewhere poised between these extremes is a value that gives the maximum work energy capacity. In this example, when a value of the growth rate is slightly higher than the value giving maximum work energy capacity, the model starts to show deterministic chaos: a minor difference in the initial value causes exponentially increasing changes as the time increases.

Stuart Kauffman has promoted and utilized this concept to show that the rate of biological evolution, as expressed in NK fitness landscapes, is maximized near the edge of chaos. The rationale is that biological systems tend to operate at the edge of chaos to be able to utilize the resources at the optimum. In response to constraints, systems move as far away from thermodynamic equilibrium as possible under the prevailing conditions, but this will imply that the system has a high probability to avoid chaos, although the

system is operating close to chaos. Considering the enormous complexity of natural ecosystems, and the many interacting processes, it is surprising that chaos is not frequently observed in nature, but it can be explained by an operation at "the edge" of chaos to ensure a high utilization of the resources—to move as far away from thermodynamic equilibrium as possible.

Further Reading

Gleik J (1998) *Chaos*. 352 pp. London: Vintage.
Groll F, Arndt H, and Altland A (2017) Chaotic attractor in two-prey one-predator system originates from interplay of limit cycles. *Theoretical Ecology* 10: 147–154.
Hastings A, Hom CL, Ellner S, Turchin P, and Charles H (1993) Chaos in ecology: Is mother nature a strange attractor? *Annual Review of Ecology and Systematics* 24: 1–33.
Kauffman SA (1993) *The origins of order*. Oxford: Oxford University Press.
Lorenz EN (1963) Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences* 20: 130–141.
Meinhardt H (2009) *The algorithmic beauty of sea shells*. Berlin Heidelberg: Springer-Verlag. 269 pp.
Rai V and Upadhyay RK (2004) Chaotic population dynamics and biology of the top-predator. *Chaos, Solitons & Fractals* 21(5): 1195–1204.
Turchin P (2003) *Complex population dynamics: A theoretical/empirical synthesis*. Princeton: Princeton University Press. 451 pp.

Citizen Science

Hiromi Kobori, Tokyo City University, Tokyo, Japan
Elizabeth R Ellwood, La Brea Tar Pits & Museum, Los Angeles, CA, United States
Abraham J Miller-Rushing, US National Park Service, Acadia National Park and Schoodic Education and Research Center, Bar Harbor, ME, United States
Ryo Sakurai, Ritsumeikan University, Osaka, Japan

Glossary

Amateur naturalist An individual with an interest in natural history who has not been formally trained in the subject.
Big data Large data sets that have been enabled by mobile devices and which, due to their immense size, often require specialized analysis techniques and devoted computational hardware and software.
Citizen science Scientific work undertaken by members of the general public, often in collaboration with or under the direction of professional scientists and scientific institutions.
Citizen scientist A volunteer participant in a citizen science project.
Do-it-yourself (DIY) science Scientific work undertaken by individuals, communities, and small organizations using the same or similar methods as traditional research institutions.
Open science The movement to make scientific research, data and dissemination accessible to all levels of an inquiring society, amateur or professional.
Typology Classification according to general type.

Introduction

The term "citizen science" has been around for just 40 years, but members of the public have done citizen science for most of recorded history—investigating scientific questions, often by noting observations of the world around them (see the glossary for the strict definition of citizen science, which only recently appeared in the dictionary). The field as grown rapidly in recent years, and encompasses an incredibly broad array of activities, common in that they include both science and the participation of volunteers—the essential components of citizen science.

Projects vary considerably, and can be classified in many ways. One of the useful ways to classify them is according to the level of participation by volunteers, the five "Cs"—contractual, contributory, collaborative, cocreated, and collegial projects. These projects vary from contractual projects in which volunteers ask questions and contract with professionals to do the research to answer them;

to contributory in which volunteers are asked by scientists to collect data; to collegial projects in which volunteers work independently of professional scientists to pursue a research project.

Recent advances in information and communications technology have facilitated rapid growth in all of these types of citizen science projects. Volunteers have unprecedented access to scientific resources and tools and can participate authentically in many research activities and at scales impossible in the past. As a result citizen science is contributing to ecology and conservation in new and exciting ways.

History of Citizen Science

Prior to the professionalization of science in the late 19th century, nearly all scientific research was conducted by amateurs—that is, by people who were not paid as scientists. These individuals were largely pursuing research because of an innate interest in particular topics or questions. Many amateurs were recognized experts in their field and conducted research indistinguishable from—and sometimes superior to—that done by most professional scientists of the time. For example, Isaac Newton, Thomas Jefferson, Charles Darwin, and many others (famous and anonymous) pursued much of their research outside the confines of academia or what we today think of as professional science. The terms "gentlemen scientists" and "naturalists" were often used to describe these individuals, at least those from upper parts of society; today they would be considered citizen scientists, following a collegial model of citizen science.

As early as the 17th century and probably earlier, some of these amateur experts as well as academic researchers had begun recruiting nonexperts to contribute natural history observations, engaging them in contributory citizen science projects. For example, in the mid-18th century, a Norwegian bishop created a network of clergymen and asked them to contribute observations and collections of natural objects throughout Norway to aid his research. These records were historical analogues to modern data collection systems, such as those used by eBird, iNaturalist, iSpot, and bioblitzes, in which volunteers record observations or collect photographs or specimens of plants and animals. Many early scientists, such as John Ray and Carl Linnaeus, used networks of volunteers to collect specimens and observations from across the known world. Such contributions have helped to build some of the most valuable collections of animals, plants, rocks, fossils, artifacts, and other specimens worldwide.

Others who have collected information and data about the natural world in the past include farmers, hunters, and amateur naturalists. For instance, wine-growers in France have been recording grape harvest days for more than 640 years, while court diarists in Kyoto, Japan, have been recording dates of the traditional cherry blossom festival for 1200 years. In China, both citizens and officials have been tracking outbreaks of locusts for at least 3500 years. Often, the people who made these observations were not intentionally participating in scientific projects, or were collecting data to address questions unrelated to their current use—but their observations and data are vital to many current scientific investigations, particularly those related to conservation and climate change.

The Professionalization of Science

Starting roughly 150 years ago, science started becoming professionalized. Many new professional scientists entered the workforce and the culture of science began changing toward that of the academic and professional laboratories and specialized scientists that we know today. Throughout this time, amateur scientists still pursued all aspects of scientific inquiry, much like gentlemen scientists and naturalists of the deeper past—amateur scientists, for example, still describe roughly 60% of all new species—but their volunteer efforts received little recognition in much of academic research. Rather, many amateur scientists were (and still are) supported by natural history clubs (e.g., New England Botanical Club, London Natural History Society, and Sapporo Natural History Society) that formed in the late 1800s or early 1900s with memberships that included mixes of professional and amateur scientists and naturalists.

During this time, advances in communications, transportation, and computing also made it easier for volunteers to contribute to large-scale data collection efforts—building on similar methods to those used by the Norwegian clergy, John Ray, and Carl Linnaeus—and for scientists and volunteers to manage and analyze the resulting data. These programs included the US National Weather Service Cooperative Observer Program (1890), Christmas Bird Count (Canada and the United States, 1900) and BirdLife Australia (initially Royal Australasian Ornithologist's Union (1901)). These surveys, and many that were developed later, yielded continental-scale datasets of biological and physical data that could not have been collected otherwise. For example, weather data gathered by volunteers for weather service agencies around the world have generated some of the most important long-term climate datasets—essential for agriculture, development planning, and assessment of recent climate change.

Recent Growth in Citizen Science

Most of the past approaches to citizen science continue today, but projects are more numerous and frequently bigger and more impactful than they were in the past. Expert amateur scientists still contribute, as they have for the entirety of the history of science; now they are supported by not only natural history and astronomy clubs, but also by do-it-yourself (DIY) laboratories and science communities (e.g., Public Labs, Genspace). Large-scale data collection projects and local and regional monitoring and research projects also still occur, many using similar methods to those used over the past 150 years or more. But mobile technologies and the

abilities to create apps and manage and analyze huge data sets have facilitated the proliferation of projects and the development of new and innovative methods. For example, projects like eBird build on methods that have been used by birders and ornithologists for decades or even centuries; but new technologies allow them to attract millions of observations and yield new insights into bird ecology and conservation.

Citizen science is also increasingly seen as a way to engage the public in science, improve scientific literacy and interest in science, and inform participants about particular topics, such as butterfly ecology, astronomy, molecular biology, or climate change. This is a major departure from most of the history of citizen science, when projects were set up mainly to achieve scientific objectives. Instead, many are now being organized primarily as means to improve participants' scientific literacy and understanding of the topics they are studying.

Characteristics and Classification of Citizen Science

In this section, we introduce some key characteristics of citizen science—namely, its unique feature that it can simultaneously contribute to research, learning, and problem solving. We also present an established citizen science project classification scheme.

Research

Citizen scientists contribute wide-ranging research that spans scales from observations of individual organisms (or even genes) to ecosystem-level assessments and analyses of images of landscapes taken by satellites. As mentioned earlier, citizen science has contributed to much of our understanding of the abundance and distribution of species, their evolution, and the functioning of ecosystems. Recent advances in technology, project design, and data standards have made the research potential of citizen science even greater. Project managers can more easily and accurately gather and curate data, and analytical tools and techniques enable easier cleaning and post-processing of huge volumes of citizen scientist-generated data. The Zooniverse suite of online citizen science projects, bird-focused projects like eBird and Project FeederWatch, and phenology projects such as those hosted by Canada's NatureWatch, are actively using big data from citizen science to contribute to solid ecological research.

Learning

A desire to learn is often a motivating factor for participants. Likewise, many citizen science project managers aim to increase the science literacy of participants. Learning through citizen science is unique because participants gain knowledge and skills, *and* contribute to authentic scientific research. This type of learning by doing is difficult to achieve in many traditional educational models. Participants of all ages can learn through citizen science activities. Youth may gain deeper understandings of the scientific method and careers in science, and their learning can be tailored to meet educational standards. Participation in citizen science may move adults to take action, alter their habits, and share what they have learned. Large institutions, such as natural history and science museums, can host citizen science activities to support learning opportunities for people of all ages and interests.

Problem Solving

Citizen science can directly contribute to solving problems—for example, by restoring habitat or removing invasive species—and can provide critical information or motivate changes in actions—for example, by informing air quality policies, increasing stewardship behavior, advocating for particular species or ecological communities, or inspiring volunteers to influence policy or vote for environmentally-friendly ballot measures. Governments are beginning to officially recognize the ability of citizen scientists to affect change and have started to provide incentives to support citizen science and encourage participation.

Classification of Citizen Science Projects

Shirk et al. (see Further Reading) established a system to categorize citizen science projects according to the level of involvement by volunteers. The scale ranges from minimal involvement in "*contractual* projects, where communities ask professional researchers to conduct a specific scientific investigation and report on the results" to a much greater degree of project management and leadership in "*collegial* contributions, where non-credentialed individuals conduct research independently with varying degrees of expected recognition by institutionalized science and/or professionals". The largest and most well-known citizen science projects exist in the second tier of this categorization, "*contributory* projects, which are generally designed by scientists and for which members of the public primarily contribute data." Many of the "grand challenges" in ecology, such as climate change and species and habitat conservation, are best supported by contributory projects, which can collect environmental data across large geographic scales. Many data collection activities require minimal training and equipment, and they can be deployed at a national or international scales year after year. Certain projects have goals that require greater volunteer involvement than occurs in contributory projects; this involvement can include participation in developing research questions, study design, data analysis and interpretation, and communication. These "collaborative" or "co-created" projects generally require more training or prior knowledge on the part of

the citizen scientists to be successful and are generally not designed to involve multitudes of people. With careful consideration, such projects can achieve deeper community engagement, improved resiliency and greater longevity.

Role of Information and Communications Technology

Information and communications technology facilitates citizen science across broad geographic scales, aids in the standardization of data collection, and broadens the appeal of citizen science to new audiences through the use of apps, games, and tools that run on mobile devices and the internet. Apps are ideal for citizen scientists monitoring light pollution (e.g., Germany's Verlust der Nacht, "Loss of the Night"), phenology (e.g., Chicago Botanic Garden's Project Budburst), or studying water quality for plankton health (e.g., the Secchi Disk Foundation's global Secchi Disk project). The ability to store and process large files also opens up opportunities for volunteers to help with the transcription of historical data sets and analysis of thousands of image files, creating valuable data for ecological research that would not be possible with professionals and automated technologies alone. Technology can also assist citizen scientists in contributing to research in a more substantial way. Web-enabled tools, such as the Collaborative Science online program developed for Virginia Master Naturalists, can help citizen scientists understand the complexities of research, and more deeply participate in data analysis, mapping, and other activities. And technologies, such as do-it-yourself kits or "hacks" for studying air quality or plant health (such as used by the Public Lab community) can allow the public to tackle research questions important to public health or other local problems that professional researchers are not interested in or cannot devote the time to. For example, community-based monitoring of water quality has grown substantially in recent years, supported by relatively easy-to-use monitoring and data management technologies, as well as the support of organizations, such as Canada's community-based environmental monitoring network.

Contributions to Ecology, Biodiversity and Conservation

Results from historical and newer citizen science have contributed tremendously to research in ecology, biodiversity, and conservation in ways often overlooked and at a variety of spatial and temporal scales (genes to global, and instantaneous to 1000 years or more). All three of these areas of research owe much of their most fundamental insights to the participation of amateurs and volunteers, whether to collect data and specimens, identify species new to science, or tackle key conservation issues.

Ecology

Citizen science is remarkably well-suited to studying the ecology of landscapes, macro-systems, urban areas, populations, communities, and ecosystems, as well as the particular areas of phenology, rare and invasive species, and disease. Landscape ecology and macroecology are relatively new subdisciplines of ecology that conduct research on large geographic scales. Large observational datasets composed of citizen science data have contributed to the development of these subdisciplines of ecology. For example, participants in the UK Breeding Bird Survey monitored the population changes of 224 common bird species in 3619 km^2 across the nation. And many of the specimens in museum collections and in large ecological databases, such as the Global Biodiversity Information Facility (GBIF), resulted from volunteer-collected data. These data sets are core to much of the work of macro-ecology.

For urban ecology, citizen science fills key roles, such as allowing sampling on private property, which is usually the majority of land in urban environments. The concentrations of people in urban areas also makes citizen science involvement in urban ecology studies—such as pollinator use of urban gardens, bird use of urban habitats, or searches for rare or underappreciated species—particularly appealing. For example, Garden Wildlife Watch is a nationwide, web-based citizen science project in Japan which monitors private gardens and the birds observed in them. Statistical analysis has revealed that larger gardens and gardens with clusters of trees and hedges support significantly more birds than other gardens, which allows researchers to make recommendations to gardeners who want to attract birds or provide important urban habitats for them.

Phenological data from citizen science projects have informed our understanding of how species are responding to climate change. In fact for much of history, volunteers led the study of phenology on their own, and now their long-term research and data sets provide much of the evidence we have describing one of the most sensitive biological responses to climate change (i.e., changes in phenology) on every continent and in the oceans around the world. For example, the UK Phenology Network revealed 250-year changes in first flowering dates for 405 plant species. Cherry blossom records from Japan show climate driven changes in flowering times, unprecedented in the flowering record (which extends back to the 9th century). More recent efforts, such as studies of the effects of climate change on winter birds in Yokohama, Japan (1986–2008) showed that birds decreased their average stay in Yokohama by about 1 month as annual temperatures warmed over the study period.

Citizen science can also be surprisingly important for studying the ecology of disease and invasive species. Data collected by FeederWatch, a project of the Cornell Lab of Ornithology and Bird Studies Canada, helped to reveal the impact of an emergent infectious disease, mycoplasmal conjunctivitis, that caused a wave of mortality in populations of house finches across the United States. And many areas rely on volunteers to help map species invasions through programs, such as Vital Signs and iMapInvasives; these mapping projects are key to managing the spread of invasive species, which can harm native biota.

Biodiversity Science

Citizen scientist contributions to biodiversity science come largely in the form of observations of the presence or absence of organisms. By and large, these projects require participants to go outside and either seek out a target organism or record the variety of life they see. Records of these observations are then shared with researchers, usually through an online portal. Biodiversity projects often require a vast quantity of observational data, making citizen science a good fit to help with data collection. In fact, most of the data submitted each year to GBIF, a primary global biodiversity database, comes from citizen science.

A distinguishing factor of observational projects is that they create *new* data. For example, a citizen scientist who records a salamander at their local park and uploads the observation to a database has, ideally, submitted a unique data point; the exact day, time, organism, or location have not been recorded before. En masse, these data provide information on distributions, range shifts, migration patterns, etc. However, these data are not always verifiable. Researchers compiling biodiversity data from citizen scientists usually employ tools to detect outliers and wrong information. Fortunately, machine learning, improved algorithms, and programmatic data analysis enables researchers to clean data efficiently and accurately.

Much biodiversity citizen science also occurs online. The two most common types involve volunteers deciphering images from camera traps and augmenting existing information with text transcription, georeferencing, or annotation. In these projects, citizen scientists are generally not creating new data per se, but are instead providing basic analysis of information, as is the case for projects where participants state the animals they see in a camera trap image. Or, they are making analog information digital by transcribing text (e.g., field notes or museum specimen records) or georeferencing textual locality information into approximations of latitude and longitude. Here, the data generated by citizen scientists is secondary to the primary data. This has implications for project design, methods, and downstream analysis as the original image, text, specimen, or other data source exists for data verification, as necessary.

Conservation

Citizen science also contributes greatly to conservation in three main ways: (1) by addressing key conservation-related science, like the science describe earlier in this article for ecology and biodiversity; (2) by involving volunteers in implementing key conservation actions, such as habitat restoration or the removal of invasive species; and (3) by encouraging citizen science participants to improve stewardship actions or get involved in decision making processes. The latter two of these really distinguish conservation citizen science from the ecology and biodiversity examples used earlier.

Many conservation projects would not be successful without the participation of volunteers—funding is frequently too limited. Thus, conservation organizations and government agencies frequently turn to citizen science to help with at least parts of their science-based restoration projects. For example, Earthwatch Institute has for a long time (since the 1970s) helped to recruit citizen science volunteers to participate in conservation-oriented research projects around the world, including the science-based restoration of mangrove forests in Kenya, clean-up of penguins and other species after an oil spill in South Africa, and restoration of sarus cranes to wetlands in Vietnam.

Many conservation organizations and government agencies also rely on citizen science specifically as a way to encourage volunteers to participate in decision making processes. The theory is that by participating in a project, volunteers will learn more and feel more invested in a particular issue and will want to get involved in decisions related to management and policy. The evidence that this type of causal relationship occurs—that is, that participating in citizen science increases participation in decision making—is still relatively weak, but there is anecdotal evidence and it is an area of active research.

Examples and Models of Citizen Science

Citizen science projects around the world have developed unique strategies to address specific themes and goals. Some projects engage thousands of participants to collect data across continents, while other projects enlist small groups of participants to focus on solving local issues.

The Cornell Lab of Ornithology, founded in 1915, works with partners to operate some of the most successful citizen science projects in the world. Most of their project focus on birds. Their largest projects, such as eBird, Great Backyard Bird Count, Nest Watch, Feeder Watch, YardMap, and Celebrate Urban Birds (many of which were created and are run in partnership with other organizations, such as Audubon and Bird Studies Canada), together have over 300,000 participants. They collect millions of observations each year and have led to key scientific insights (dozens or hundreds of scientific publications) and have informed management and policy. Most of the Cornell Lab of Ornithology's projects could be categorized as contributory projects and are designed to answer research questions initiated by scientists. Participants follow guidelines to provide high quality observational data, then scientists analyze the results, publish their findings, and provide policy suggestions. However, local organizations can use the methods and online data entry and data management infrastructure to design and implement their own projects, such as looking for spatial or temporal mismatches between species, monitoring for particular rare or invasive species, or identifying crucial conservation areas that are important to protect.

Other citizen science projects have found success through combined support from government agencies, NGOs, and scientific organizations. Monitoring Site 1000 is a citizen science project for monitoring biodiversity at about 1000 sites in Japan. It was

started in 2003 and is now a national project with about 2500 participants and with varied stakeholders including the Japanese Ministry of the Environment and more than 200 local NGOs. This project is a collaborative project with local NGOs and local citizens engaging in both data collection and dissemination of results. The United Kingdom and many countries of the European Union have similar government-supported citizen science monitoring schemes. And the United States government also recently issued policies explicitly encouraging the development of citizen science to support science, including key areas of ecology.

Zooniverse is one of the world's largest platforms of citizen-contributed research. Its online-only approach has proven to be an effective way of mobilizing data from a wide range of research-based topics including ecology, astronomy, history, and literature. Intriguingly, their approach frequently engages people in areas of science that they have not historically been interested in. People who participate in their online astronomy projects, for example, may not regularly participate in amateur astronomy. Similarly, their ecology-related projects, such as "Snapshot Serengeti" which asks participants to classify animals captured on camera trap footage, may attract new people to ecology and conservation.

Nonscience Outcomes

In addition to scientific outcomes, citizen science can benefits participant learning. By participating in research projects, volunteers may increase their awareness of local environmental issues, knowledge of scientific methods, and understanding of scientific topics. Participants can also develop scientific skills through learning data collection techniques, and may become interested in careers in science. Some of these benefits can be considered scientific literacy outcomes.

Some citizen science projects aim to influence the behaviors of participants. Evidence for these effects is limited, but it is an area of much research. Participants in some citizen science projects have gone on to improve wildlife habitat, contribute to publications, and apply for grant funding. Citizen science experiences can be transformative for some participants, but making this a consistent outcome is difficult to do and to measure.

Many citizen science projects also aim to influence policy, either through the science they produce or through the advocacy of their participants. The National Audubon Society's Christmas Bird Count, for example, engages thousands of participants each year, has contributed to more than 200 peer-reviewed papers, and has affected change in conservation policy, such as suggestions for reductions in hunting of American black ducks (*Anas rubripes*). Monitoring Site 1000 of Japan and citizen science projects in other countries have helped monitor progress toward achieving the Aichi Biodiversity Targets as a part of the Convention on Biodiversity.

Designing Citizen Science Projects

In the fields of ecology and environmental science, citizen science project managers consider four major axes when designing projects: (1) initiator of the project (professional scientists or the public); (2) scale (local or global) and duration (short or long term) of the project; (3) types of questions being asked, ranging from pattern detection to experimental hypothesis testing; and (4) goals, which include research, education, behavioral change, problem-solving, conservation and policy-making.

Information to implement projects is available through national and international citizen science organizations. New resources are being developed all the time—the field is moving quickly. Here we provide some of the major sources of information that will help lead to others.

The Citizen Science Association (CSA) was recently developed to facilitate communication and best practices. The CSA organizes a biennial conference, publishes a journal of citizen science research and ideas, *Citizen Science: Theory and Practice*, and hosts an email listserv where practitioners, researchers, and educators can exchange ideas and ask and answer questions related to citizen science. The European Citizen Science Association and Australian Citizen Science Association also promote citizen science and support the work of citizen science practitioners. CitizenScience.org, run by the CSA and Cornell Lab of Ornithology, provides a toolkit for project development and references for the professional network. CitSci.org supplies tools for creating, among other things, customized data-entry forms for volunteers. SciStarter provides tools, information about citizen science, and an extensive list of current citizen science projects. The Public Laboratory for Open Technology and Science offers citizen scientists a community atmosphere where citizens can collaborate with experts and access scientific resources.

Future of Citizen Science

As mentioned earlier, citizen science contributes to research, learning, and solving problems. It is also contributing to the "socialization" or "democratization" of science and to "open science" efforts. However, the full potential for citizen science has yet to be realized. New models, approaches, and collaborations will be helpful for researchers and society alike. Specifically, we highlight the following three areas for future work: new typologies, increased use of citizen science data, and social science.

New Typologies

Citizen science typologies to date have focused primarily on the integration of the public in scientific research. Typologies focusing on different aspects of citizen science could provide additional insights to evaluate and improve existing projects and forecast desirable future directions for the field of citizen science. One such typology focuses on project goals, such as action, conservation, investigation, virtual, and education. Projects with different goals require different approaches to design and management. Another typology focuses on mechanisms by which scientific and social aspects of the projects converge—for example, science-driven, policy-driven, and transition-driven models of citizen science. This typology could facilitate insights to increase the value of socio–ecological outcomes from citizen science projects. The transition-driven model, in particular, could be useful for tackling complex problems for which there are no ready-made solutions, where existing knowledge may be incomplete, or where perspectives are incompatible or conflicting.

Use of Data

Much of the data obtained from citizen science projects are not well analyzed. A recent review found that only 12% of biodiversity-related citizen science projects contribute data to peer-reviewed scientific articles, despite the fact that a third of these projects have verifiable, standardized data that are accessible online. It is not obvious why the data are not being analyzed—whether from lack of access or a lack of scientist time, interest, and trust. Newly developed machine learning methods may be able to help solve this problem. They require relatively little active scientist involvement relative to typical statistical data analyses, and may help researchers rapidly work through large quantities of citizen science data.

Social Science Approaches

Most ecology-related citizen science projects focus on natural science questions; social science approaches, questions, and insights are seldom included in the design and implementation of projects. This can be a serious shortcoming; insights from social science can help project organizers in many ways. For example, social science can help project managers understand and accommodate the concerns of volunteers or identify factors that influence volunteers' willingness to participate.

Further Reading

Bonney R, Cooper C, and Ballard H (2016) The theory and practice of citizen science: Launching a new journal. *Citizen Science: Theory and Practice* 1(1).
Bonney R, et al. (2014) Next steps for citizen science. *Science* 343: 1436–1437.
Dickinson JL and Bonney R (eds.) (2012) *Citizen science: Public participation in environmental research*. Ithaca, NY: Cornell University Press.
Danielsen F, et al. (2014) Linking public participation in scientific research to the indicators and needs of international environmental agreements. *Conservation Letters* 7: 12–24.
Dillon J, Stevenson RB, and Wals AE (2016) Introduction to the special section: Moving from citizen to civic science to address wicked conservation problems. *Conservation Biology* 30: 450–455.
Ellwood ER, Crimmins TM, and Miller-Rushing AJ (2017) Citizen science and conservation: Recommendations for a rapidly moving field. *Biological Conservation* 208: 1–4.
Henderson S (2012) Citizen science comes of age. *Frontiers in Ecology and the Environment* 10: 283. Introduction to a special issue on citizen science in ecology.
Kobori H, et al. (2016) Citizen science: A new approach to advance ecology, education, and conservation. *Ecological Research* 31: 1–19.
Kosmala M, Wiggins A, Swanson A, and Simmons B (2016) Assessing data quality in citizen science. *Frontiers in Ecology and the Environment* 14: 551–560.
McKinley DC, et al. (2015) Citizen science can improve conservation science, natural resource management, and environmental protection. *Issues in Ecology*. Report No. 19.
Miller-Rushing A, Primack RB, and Bonney R (2012) The history of public participation in ecological research. *Frontiers in Ecology and the Environment* 10: 285–290.
Phillips T, Furguson M, Minarchek M, Porticella N, and Bonney R (2014) *Users guide for evaluating learning outcomes from citizen science*. Ithaca, NY: Cornell Lab of Ornithology.
Shirk JL, et al. (2012) Public participation in scientific research: A framework for deliberate design. *Ecology and Society* 17: 29.
Theobald EJ, et al. (2015) Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biological Conservation* 181: 236–244.
Tweddle JC, Robinson LD, Pocock MJO, and Roy HE (2012) *Guide to citizen science: Developing, implementing and evaluating citizen science to study biodiversity and the environment in the UK*. Natural History Museum and NERC Centre for Ecology & Hydrology for UK-EOF.
Wiggins A, et al. (2013) *Data management guide for public participation in scientific research*. Albuquerque, NM: DataONE.

Relevant Websites

Australian Citizen Science Association—citizenscience.org.au.
CitSci.org.citsci.org.
Citizen Science Association—citizenscience.org.
European Citizen Science Association—ecsa.citizen-science.net.
Public Laboratory for Open Technology and Science—publiclab.org.
SciStarter—scistarter.org.

Complex Ecological Networks

Mathilde Besson*, **Eva Delmas***, **and Timothée Poisot,** Université de Montréal, Montréal, QC, Canada and Québec Centre for Biodiversity Sciences, McGill University, Montréal, QC, Canada
Dominique Gravel, Québec Centre for Biodiversity Sciences, McGill University, Montréal, QC, Canada and Université de Sherbrooke, Sherbrooke, QC, Canada

Glossary

Adjacency matrix Matrix representing species interactions. If two species *i*and *j* interact, the intersection of the matrix at *i*, *j* will be 1, and 0 if not.

Assembly rules Ecological processes leading to a specific species' composition of a community, for example, competition, predator–prey interactions, arrival history, etc.

Degree The degree of a node is its number of links (e.g., interactions per species). At higher level, the degree distribution represents the cumulative distribution of links per node within the network or a subnet of the network.

Ecological interactions Every type of contact between two species that alters the fitness of one or both species. Interactions can be directed or undirected, weighted or unweighted. They usually fall into one on these five main classes: competition, predation, parasitism, mutualism, and commensalism.

Ecosystem functioning Biotic and abiotic processes that sustain ecosystems, including flues of energy and nutrients between the components of ecological systems and the resulting stocks, for example, biogeochemical cycles.

Graph theory Mathematical framework used to represent the relationship between the objects of a network.

Network structure General shape of a network emerging from the organization of the interactions between its components. It is commonly described in ecology using connectance, link distribution, topological indices (such as nestedness, modularity, centrality), etc.

Nodes/links, vertices/edges Following graph theory, species are represented as nodes (or vertices), and interactions between them are represented by links (or edges).

Phylogenetic signal Tendency of phylogenetically close species to have similar traits (and as a consequence, similar interactions).

Unipartite/bipartite network The graphical representation of the entire adjacency matrix offers an unipartite network representation (see Fig. 1), where the hierarchy between nodes and their position into the network is not always visible. On contrary, a bipartite or k-partite network is a hierarchical representation of the network (Fig. 2), where nodes are separated depending on their position or function into the network (e.g., pollinator–plant as bipartite network).

Introduction

Interactions between the components of any ecological systems are organized nonrandomly. The species that form a community for example do not interact at random. The resulting organization of interactions between species drives some properties of the community such as stability, productivity, and the ability to resist extinctions, all of which eventually feedback on the organization of the system. The constant interplay between the organization of interactions and system dynamics constrains its structure. Studying the structure of ecological systems provides insights on the fundamental rules and processes that govern ecosystem formation, maintenance, and functioning.

The organization of interactions in a community is best represented as a network. *Graph theory* is a field of mathematics developed to analyze the structure of such systems. Every community can be abstracted by a *graph*, which is a representation of the system components and their arrangement (Fig. 1A). These components are called *nodes* and are linked together by *edges*. In an ecological system, nodes can be individuals, populations, communities or landscape patches and edges can represent trophic interactions, energetic flues and more generally every kind of interactions. Both nodes and edges can carry additional information such as weight (e.g., species abundance, intensity of the gene flow between two populations, etc.), location in space and time, and labels (e.g., species identity). Specific information can be attached to edges, modifying the characteristics of the graph, for example, the environmental dependence of an interaction. Graphs can be *directed* (i.e., interaction goes from A to B) or *undirected*, *weighted* (i.e., different strength of interaction among the network) or *unweighted* (Figs. 1 and 2). This information is summarized in the *adjacency matrix*, typically named *A* (Fig. 1B). The adjacency matrix *A* can be used to answer various ecological questions. Using it directly allows to follow direct interactions and the network structure, and using the inverse of *A*can be useful to obtain indirect interactions, and even more (Montoya *et al.*, 2009).

In this article, for simplicity, we will focus mostly on *Species Interaction Networks* (SIN). Ecological systems such as landscape, genetic or nutrient networks are not represented here, but they can be studying using the same framework as defined further.

*These authors contributed equally to the work.

 doi:10.1016/B978-0-12-409548-9.10564-0

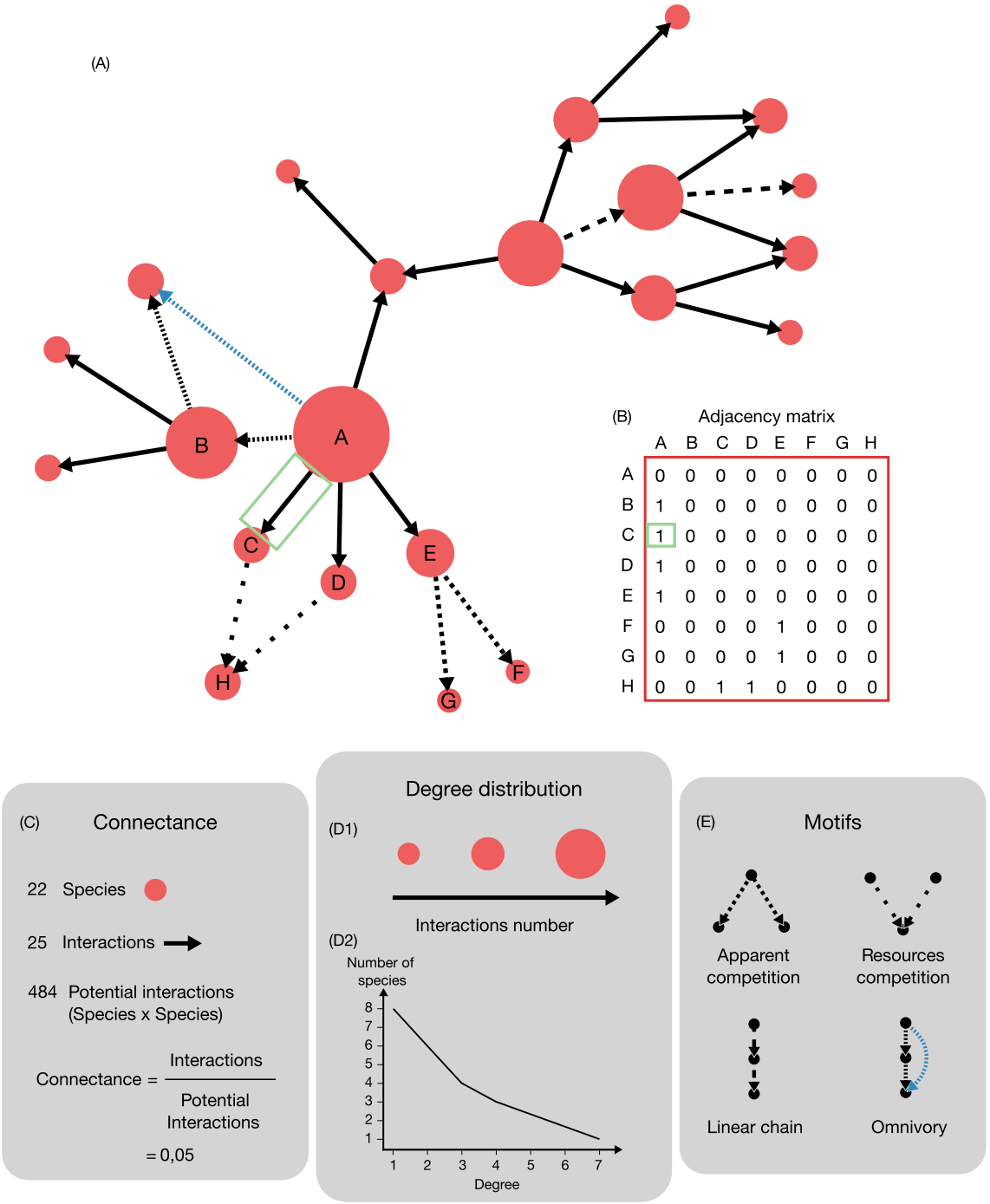


Fig. 1 Graphical representation of an ecological network (A), where species are represented by *circles* and their directed interactions by *arrows*. The representation is formalized in the adjacency matrix (B). In an unipartite representation as this one, each species is represented both as a column and a row. 1 indicates an interaction between two species (e.g., the *green square* in (B)), and 0 indicates the absence of interaction. This matrix facilitates computation of characteristics such as the connectance (C) and the degree distribution (D). (C) Represents the level of connection into the network and is calculated as showed in the figure. (D) Represents the distribution of interaction per species. The *circles* size is relative to the amount of interactions a species have (D1). This distribution is nonrandom and generally follows a power-law distribution (D2). The network can be split into subnets composed of 3 species, called motif (E). Among the 13 different possible motifs, we only represented the most commonly found in natural communities.

Describing and understanding the structure of SIN is an active, and growing, field of ecological research. We provide here an overview of some of the most prominent findings and areas of research from the last decade. Starting from a discussion of some invariant properties of the structure of species interaction networks, we will then discuss how this structure affects community

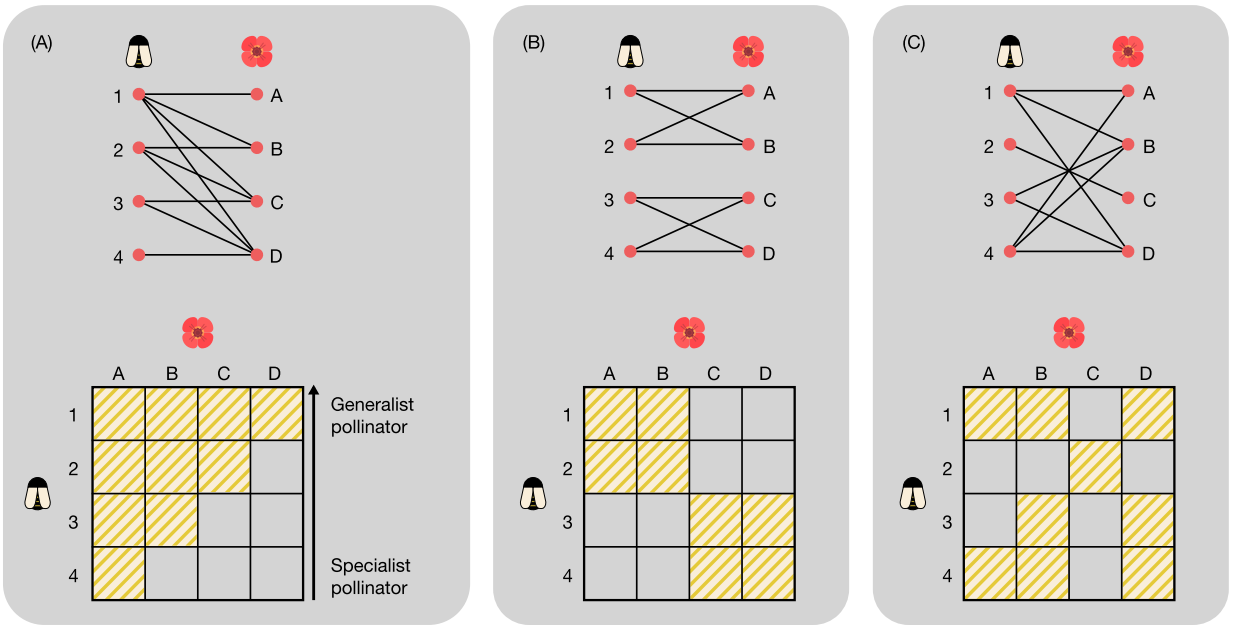


Fig. 2 Network topology, example of a fictional plant–pollinator network. (A) shows a perfectly nested network, where specialists pollinators are visiting plants embedded into the diet of more generalist pollinators. (B) Shows a perfectly modular network, where subgroups of species interact more strongly with each over than with the rest of the network. (C) Shows a random network. Two representations are possibles. Top: Bipartite representation using nodes and edges; Bottom: Ordered interaction matrix. Here, we used striped *yellow squares* instead of 1 for presence of interaction and empty squares in absence of interaction.

dynamics and properties. We will follow by a discussion of the ways ecological networks can be studied under familiar concepts from ecological theory, and finally how this approach scales up to larger temporal and spatial scales.

Invariants in Ecological Networks

One striking particularity of ecological networks is their consistency: even though they depict interactions between different organisms across all sorts of ecosystems, they all tend to look the same (Jordano *et al.*, 2003). Remarkably, even when interactions among species themselves vary, the overall network structure tends to remain unchanged (Kemp *et al.*, 2017). Most ecological networks have a very specific and similar *degree distribution* (Williams, 2011) (**Fig. 1**D), whereby most species have a small number of interactions, and a small proportions of species have a large number of interactions. In food webs, which represent interactions between prey and their predators, there is a well-described relationship between the number of species and the number of interactions. The number of interactions (L) increases proportionally to the number of species (S) raised to some exponent, or $L \propto S^k$. Martinez (1992) suggested that this exponent is approximately equal to 2, that is, the number of interactions is proportional to the squared number of species. Brose *et al.* (2004) showed that this relationship holds even across space; it is possible to estimate how many interactions a species will establish across its entire range. In other instances, networks may differ on some aspects of their structure, despite obeying to a shared underlying principle. For example, Fortuna *et al.* (2010) showed that in networks with a low connectance (**Fig. 1**C), nestedness (the degree to which the diet of specialists and generalists overlaps—**Fig. 2**) and modularity (the tendency of species to form densely aggregated clusters—**Fig. 2**) are positively correlated. In networks with higher connectance, this becomes the opposite: networks with a large number of interactions are either nested (and not modular) or modular (and not nested). In the recent years, it emerged that many aspects of network structure covary with connectance (Poisot and Gravel, 2014; Chagnon, 2015), suggesting that simply knowing how many species there are, and how many interactions they establish, is already very informative about the network structure.

Another remarkable generality of network structure is the distribution of particular interconnection between three-species subsets. Milo (2002) found that networks (not just ecological but other types of networks such as neuronal or electronical networks as well) can be characterized by the over or under representation of some of these three-species subsets, which they called motifs (**Fig. 1**E). Motifs can be more broadly defined as specific arrangements of interconnection between three (or more) nodes. The frequency at which they occur in a network can be computed and compared to randomized networks in order to reveal significant aspects of the structure. Three-species motifs represent the simplest building blocks of networks, and more importantly typical interaction modules found in communities. As such, they offer the possibility to integrate and test theories developed with simple modules in larger, more realistic networks (e.g., omnivory, McCann *et al.*, 1998, Holt, 1997). Food webs, for example, are characterized by an over representation of linear food chains and omnivory and an under representation of apparent and

exploitative competition (**Fig. 1**A and E) (Bascompte and Melián, 2005; Camacho *et al.*, 2007). Stouffer and Bascompte (2010) found that realistic motif distribution promotes stability in food webs, with over-represented motifs being more stable in isolation and correlated with higher stability in large realistic communities, and conversely. Motifs can also be used to characterize species role in networks. From the 13 different three-species motifs emerge 30 unique positions for species to occupy in these motifs, representing how the species is embedded in its community. The different positions a species will occupy, and the frequency with which it will occupy these different positions in networks are called species motif role (Stouffer *et al.*, 2012). These roles have been shown to be evolutionary conserved in food webs (Stouffer *et al.*, 2012) and to have less variability in time than expected in host–parasitoids bipartite networks (Baker *et al.*, 2015).

Another invariant network property relates to evolutionary history. Phylogeny is a key determinant of ecological network structure, being related to species position and interactions into the community. Phylogenetically close species tend to inherit traits from their common ancestors (e.g., body size, habitat, defensive strategy, metabolic type, phenology), increasing their propensity to interact with the same group of species or with similar species, a phenomenon called *phylogenetic signal*. This conservatism of interactions has been found to hold across different types of interactions such as antagonistic or mutualistic interactions (Fontaine and Thébault, 2015). However, depending in the species role (e.g., host or parasite, pollinator or plant) the link organization will be different, leading to an asymmetrical structure for pairwise interactions. For instance, closely related hosts tend to share parasites, while closely related parasites, because of competition for resources, tend to have different hosts (Krasnov *et al.*, 2012). The conservatism of interactions is consequently unequal all over the network. Following the logic that closely related species interact with the same group of species, Rezende *et al.* (2009) showed that phylogenetic structure of ecological networks explains almost entirely the formation and composition of modules and the connections between them. The species connecting modules together are indeed usually phylogenetically close. Cattin *et al.* (2004) also found, using a niche-hierarchic model, that diet is constrained by the phylogenetic origin of consumers. The nested structure of trophic networks is then influenced by the phylogenetic signal of interacting species and their traits compatibility. In contrast, the nested structure of mutualistic networks would be a consequence of trait complementary between species (Rezende *et al.*, 2007). For now, mechanisms underlying the nestedness–phylogeny relationship remain to be further investigated. Moreover, because of species plasticity, phylogeny alone does not fully explain the structure and evolution of ecological networks.

From Structure to Properties

The relationship between ecological network structure and stability is a long-lasting object of research in community ecology. MacArthur (1955) and Elton (1958) first proposed that diverse communities should have a more stable dynamic than simple ones because disturbances are more easily spread through highly connected nodes. May (1972) countered this hypothesis using a mathematical model based on random ecological networks and proposed there should be a limit to ecosystem complexity. This counter-intuitive proposition sparkled live debates still lasting today (McCann 2000; see Allesina and Tang, 2015). Two different approaches to the problem followed: one focused on dynamical stability and the other on the resistance of communities to species lost. Despite their dissimilarities, these approaches are not totally independent (Donohue *et al.*, 2013) and revealed that species diversity has no direct influence on community stability. However, the structure of ecological network such as the distribution of interaction strength and network topology seems to play a crucial role (Yodzis, 1981).

As mentioned above, the degree distribution of ecological networks often follows a power-law distribution (Montoya and Solé, 2002), indicating that few species are highly connected to the rest of the community and a large number of species are weakly connected to others. This organization combined with the myriad of weak interactions found across ecological networks buffers species variations and stabilizes the dynamics of the entire community (Bascompte *et al.*, 2005; Jacquet *et al.*, 2016). Other aspects of community structure, such as the predator–prey body–mass ratio (Emmerson and Raffaelli, 2004; Brose *et al.*, 2006a) and network architecture (Montoya *et al.*, 2006; Thébault and Fontaine, 2010), determine the distribution and strength of interactions and together drive the stability of ecological networks (Jacquet *et al.*, 2016).

Perturbations in ecological communities such as landscape fragmentation, habitat loss, or species invasion, are the primary drivers of species loss. Extinctions may happen directly, for instance if a particular habitat is eliminated, or indirectly following a first species loss (a phenomenon referred as secondary extinction or cascades). Such extinctions are used to measure the robustness of ecological communities. Simulation experiments revealed that the likelihood of secondary extinctions increases with community size (Lundberg *et al.*, 2008), decreases with network connectance (Dunne *et al.*, 2002) and primarily affects the most isolated species in the network. The loss of a highly connected species, also called a hub, induces a higher rate of secondary extinctions than the loss of a random and weakly connected species (Solé and Montoya, 2001). Similarly, species responsible for important energy-flow in the network (carbon, nitrogen or biomass) can trigger secondary extinctions (Allesina and Bodini, 2004).

The network architecture also affects the community response to perturbations. In agreement with MacArthur's intuition, it was found that species with low *degree* also more strongly propagate perturbations following permanent changes in the environment because of their tight connections (Montoya *et al.*, 2009). Alternatively, the most connected species diffuse such perturbations through the network and even though they affect a higher number of species, their average effect on other ones is much smaller. Overall network properties also affect the response to perturbation. Thanks to their structural properties (high nestedness and connectance, Jordano *et al.*, 2003), mutualistic networks persist longer than randomly structured networks (Memmott *et al.*, 2004;

Fortuna and Bascompte 2006). On the other hand, presence of modules in the community structure limits propagation of perturbations across the rest of the network and, as such, secondary extinctions (Stouffer and Bascompte, 2010).

Eluding the consequences of biodiversity lost for ecosystem functioning is also an important field where the network approach has been useful. The hypothesis that an increase in species diversity results in an increased productivity dates back to Darwin (1859) and a formal theory for what is now called the biodiversity-ecosystem functioning (BEF) relationship was proposed in the mid 1990s. In a trophic group (i.e., a group of species that all belong to the same trophic level, e.g., producers or herbivores), increasing diversity improves resource use efficiency and translates into larger productivity (Loreau, 2010) (e.g., nutrients for producers, or producers for herbivores). Yet, when the trophic group under focus is coupled to other(s), the action of diversity on functioning is more variable (Duffy *et al.*, 2007). This makes the BEF relationship unpredictable in real-world communities (Harvey *et al.*, 2013), composed of several trophic groups that are virtually never differentiable—as intraguild predation and omnivory blur the frontier between levels. The multiplicity of the factors influencing the BEF relationship calls for a more general framework that allows the integration of the theories developed for trophic groups and for simple modules or subsystems (Gravel *et al.*, 2016). By mapping transfer of biomass and energy and/or constraints on organism through the different compartments that compose a natural community, ecological networks—and food webs in particular—offer the possibility to perform this integration. Analyses performed on simulated food-webs with fixed species richness have shown that interactions, and more specifically their structure, have a significant influence on productivity (Thebault and Loreau, 2003; Thébault *et al.*, 2007; Poisot *et al.*, 2013). The structure of interactions is indeed a reflection of community properties, essential to ecosystem functioning. It seems then essential to integrate it in BEF studies.

Mechanisms Underlying Pairwise Interactions

Ecological interactions between species should be viewed as the result of low level processes involving pairs of individuals. A pollinator is able to effectively reach the nectar in a plant because their respective traits match, they have compatible phenologies, and they occur in the same environment. A virus can infect its host because it is able to attach to the cell surface, effectively penetrate it, and hijack the cellular machinery to its benefit. Interactions that are not allowed because trait values do not match have been called "forbidden links" (Olesen *et al.*, 2011). This prompted a search for "linkage rules" (Bartomeus, 2013) in ecological networks, that is, the relationships that must exist between traits of two organisms in order for an interaction between them to exist. These can be identified from existing data on traits and interactions (Bartomeus *et al.*, 2016), and then used to generate realistic ecological networks (Crea *et al.*, 2015). González-Varo and Traveset (2016) pointed out that interactions are happening between individuals, and as a consequence, it requires to consider not only how the traits are distributed at the individual scale, but also how different behaviors may allow organisms to overcome some of the forbidden interactions.

Although traits are an important part of what makes interactions happen, they are only relevant insofar as the organisms are able to encounter one another. The importance of neutral dynamics (i.e., how abundances of different species can determine the probability that they can interact, based on how often they would get in contact by chance) is, somewhat counter-intuitively, great. Canard *et al.* (2012) revealed that realistic food webs can be predicted with only knowledge of abundances. In a host–parasite system, local abundances has also been identified as a key predictor of species interactions (Canard *et al.*, 2014). More broadly, because interactions emerge from all of these ecological mechanisms, there is a need to develop a deeper understanding of their variability (Poisot *et al.*, 2015). Beyond the fundamental advance that this represents, this would allow to model interactions based on external information instead of documenting all of them (Morales-Castilla *et al.*, 2015).

The realization of an interaction between individuals has, by definition, an effect on population dynamics. But it is also archetypical of complex system dynamics, where low level processes propagate up to higher level of organization and impact emerging properties of the community. If we consider for instance a population A, its dynamic is not the same when it multiplies in isolation—where it can grow exponentially if resources are unlimited (Malthus, 1798) or logistically otherwise (Verhulst, 1938)—or when it is embedded in a real-world community, composed of several species interacting with one another through different processes. That population can lose individuals to predation, have parasitism increase its death rate and at the same time see its establishment eased through facilitation. It then becomes necessary to account for the entire set of interactions to understand population, community and ecosystem dynamics. But the effect of interactions on dynamics is not always straightforward to elude, both in terms of directionality and intensity, as there is different types of interactions and multiple factors influencing their occurrence and strength.

Ecological networks are also spatially and temporally variable (Trøjelsgaard and Olesen, 2016). There are two drivers to this variability: changes in species composition, and changes in the way these species interact (Poisot *et al.*, 2012). Changes in species alone are able to generate variation in network properties (Havens, 1992). Spatial variation in network structure can also reflect deep-time constraints; for example, Dalsgaard *et al.* (2013) revealed that historical climate change trends have a signature on the nestedness and modularity of pollination networks. Even when the same species are present, interactions between them can vary. Carstensen *et al.* (2014) and Trøjelsgaard *et al.* (2015) investigated this phenomenon in mutualistic networks. Interaction turnover results from variations in partner fidelity (some species pairs are extremely closely associated), but also from variations in the local environment in which the species interact. Interestingly, networks overwhelmingly tend to conserve their structure even when interactions within them change. Díaz-Castelazo *et al.* (2010) surveyed a pollination network over 10 years, and found important species turnover during this period. Nevertheless, the network retained its structure because species where replaced by their

functional equivalent; a generalist pollinator often succeeded to another generalist pollinator. Conversely, species tend to retain their role in different communities: Baker *et al.* (2015) showed that species keep occupying the same position in the network across space, regardless of the species they interact with at every location.

From the Regional Species Pool to Local Structured Communities

Describing the variation in ecological network structure at large spatial scales may represent an additional layer of information compared to simple species lists. As such, ecological networks are a powerful tool to shed new light on the processes underlying species distribution (Cazelles *et al.*, 2016) and variation in some ecosystem functions (e.g., trophic regulation). Until recently, the prevailing idea was that at large spatial scales, the role of biotic interactions on distribution is very small compared to that of abiotic conditions, and as such is important only locally (Pearson and Dawson, 2003; Boulangeat *et al.*, 2012). Empirical observations of species–environment relationship are used to approximate species physiological tolerance to environmental conditions and potentially predict their range under different scenarios of climate change (e.g., Araújo *et al.*, 2006). While these species distribution models provide a useful approximation of their potential range shift (Pearson *et al.*, 2002), there is mounting evidence that biotic interactions—both positive and negative—play a critical role in shaping communities not only at local scales (Boulangeat *et al.*, 2012), but also at macro-ecological scales (Davis *et al.*, 1998; Araújo and Luoto, 2007; Heikkinen *et al.*, 2007; Gotelli *et al.*, 2010; Araújo *et al.*, 2011).

It was proposed that the role of interactions in shaping species distribution could be approximated from knowledge of species cooccurrence (Araújo *et al.*, 2011). This very active field of research has been recently pushed by the development of joint species distribution models (JSDM), which account simultaneously for the effect of the environment and codistribution (Pollock *et al.*, 2014). But there are limitations to this approach. For instance, it does not allow to distinguish between cooccurrence caused by biotic interactions and correlated responses to unmeasured environmental variables (Pollock *et al.*, 2014). Conversely, the lack of association between species is no evidence of absence of interaction (Cazelles *et al.*, 2016). Further work is therefore needed to move from correlative species distribution models (SDM) toward more theoretically sound models. In particular, developing methods allowing to include prior information about the underlying ecological network when estimating (J)SDM could shed light on the fundamental processes underlying species distribution and thus making more accurate predictions (Cazelles *et al.*, 2016). Additionally, Poisot *et al.* (2017) recently showed that biotic interactions respond to environmental conditions on their own, independently of species.

Ecological networks also offer an ideal framework to study the conditions for the maintenance of biodiversity in communities. The competitive exclusion principle states that the number of coexisting species should be equal or smaller than the number of resources. This stands in contradiction with the existence of ecological communities containing species that overlap in some extent in their resources or consumers. Phytoplanktonic communities are often considered to illustrate this paradox (Hutchinson, 1961), as they exhibit a high biodiversity while species are competing for a limited number of shared resources (e.g., light, nitrate). Species coexistence mechanisms (Chesson, 2000) are based on species traits that either decrease fitness differences (equalizing mechanisms) and/or increase niche differentiation between species (stabilizing mechanisms).

The coexistence theory and the representation of ecological communities as networks of interactions has brought new perspective on species coexistence. Martinez *et al.* (2006) for instance showed that the global nonrandom structure of the food webs improve community persistence (i.e., species coexistence). The distribution of motifs in food webs (Stouffer and Bascompte, 2010, see section *Invariants in ecological networks*) as well as species' role within motifs (Stouffer *et al.*, 2012) are related to community persistence. In mutualistic networks for instance, the nested structure minimizes interspecific competition and increase the number of coexisting species (Bastolla *et al.*, 2009; Sugihara and Ye, 2009). Interactions structure also tend to impact species coexistence into communities, as highlighted by Bascompte *et al.* (2006), the fact that one species A depends strongly on another species B as resource for food or pollination, and the other species, B, only weakly depends on A, also called asymmetry of dependences, increases coexistence of species. As an other example, using food web structure Brose *et al.* (2006b) showed that the allometric scaling of metabolic rates of species improve community persistence. All these types of approach, whether they are based on motifs, species' role or allometric scaling, have highlighted the importance of network structure in species coexistence.

Ecologists have also questioned the way communities are formed and the hypothetical set of rules embedding their assembly. The network approach allows to explore in details the different processes influencing ecological communities assembly. Capitán *et al.* (2009), for instance, characterized the sequence of species arrival in a community with an assembly graph. It allows to follow step by step every possible path in community assembly from 0 to $x$species among several trophic levels, and to highlight underlying mechanisms. Verdú and Valiente-Banuet (2008), for instance, found that nested community provides generalists species which facilitate the presence of other species into the network. At the same time, Olesen *et al.* (2008) observed that newly arriving species tend to interact more easily with already well-connected or generalist species. Such results could let us think about the Drake's controversial idea that species arrival history would be an important factor driving community assembly (Drake, 1991). This proposition was supported by network analyses, such as in Campbell *et al.* (2011) for mutualistic networks, but still remains object of debate.

The addition of ecological networks into models of diversity dynamics fostered the development of theory of community assembly at both, fine and large spatial scales. Niche and neutral theories dominated most of community assembly research since the publication of Hubbell's book in 2001. A wide range of models have been used, most of them with very abstract and

phenomenological representations of the niche. But only recently, with the addition of trophic constraints (Gravel *et al.*, 2011) and other types of interactions (Cazelles *et al.*, 2016) to MacArthur and Wilson's (1967) model of island biogeography, that all types of interactions were considered in the process of community assembly. The model was first extended by assuming that predator could only colonize communities with prey already present, and go extinct with their last prey. This modification was sufficient to explain the observation of a sequential construction of food webs after the defaunation treatment of the famous experiment by Simberloff and Wilson (1969) and (Petchey *et al.*, 2008). The model was further use to illustrate a reciprocal feedback between colonization-extinction dynamics and local food web dynamics, where properties of the regional food web constrain the development of the local motif structure, and alternatively local dynamics influence the assembly process (Massol *et al.*, 2017). This modeling approach allows a general representation of the niche in studies of assembly dynamics (Jacquet *et al.*, 2017) and propose a unifying framework to explain the construction of local communities from a sample of the regional species pool.

Conclusion

Graph theory delivered important scientific discoveries, such as improved understanding of breakdown of electricity distribution systems or the propagation of infections in social networks. It is also a powerful tool to investigate key questions in ecology. Graph theory provides a remarkably simple way to characterize the complexity of ecological networks. Indices such as connectance, degree distribution or network topology serve as basic measurements to describe their structure. Such indices facilitate comparison between different systems and revealing commonalities and variations. Nowadays, the relatively important number of network studies leads to a myriads of ways to sample, analyze and interpret them (see Delmas *et al.*, 2017).

Studying ecological networks have however a larger purpose than just their description and classification. Basic measurements are correlated to several environmental conditions and network analysis appears to be helpful in different ecological fields. As we seen through this chapter, it can be used to study dynamics of ecological systems and their responses to changes, according to their stability over time or the BEF relationships in the system. It also highlights the understanding of mechanisms underlying ecological properties such as community assembly, coexistence and species distribution. Network studies were a key to reveal relationships between different properties of ecological network such as trait and structure.

See also: Ecosystems: Floodplains

References

Allesina, S., Bodini, A., 2004. Who dominates whom in the ecosystem? Energy flow bottlenecks and cascading extinctions. Journal of Theoretical Biology 230 (3), 351–358. doi:10.1016/j.jtbi.2004.05.009.

Allesina, S., Tang, S., 2015. The stability–complexity relationship at age 40: A random matrix perspective. Population Ecology 57 (1), 63–75. doi:10.1007/s10144-014-0471-0.

Araújo, M.B., Luoto, M., 2007. The importance of biotic interactions for modelling species distributions under climate change. Global Ecology and Biogeography 16 (6), 743–753. doi:10.1111/j.1466-8238.2007.00359.x.

Araújo, M.B., Thuiller, W., Pearson, R.G., 2006. Climate warming and the decline of amphibians and reptiles in Europe. Journal of Biogeography 33 (10), 1712–1728. doi:10.1111/j.1365-2699.2006.01482.x.

Araújo, M.B., Rozenfeld, A., Rahbek, C., Marquet, P.A., 2011. Using species co-occurrence networks to assess the impacts of climate change. Ecography 34 (6), 897–908. doi:10.1111/j.1600-0587.2011.06919.x.

Baker, N.J., Kaartinen, R., Roslin, T., Stouffer, D.B., 2015. Species' roles in food webs show fidelity across a highly variable oak forest. Ecography 38 (2), 130–139. doi:10.1111/ecog.00913.

Bartomeus, I., 2013. Understanding linkage rules in plant-pollinator networks by using hierarchical models that incorporate pollinator detectability and plant traits. PloS one 8 (7), e69200 Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0069200.

Bartomeus, I., Gravel, D., Tylianakis, J.M., *et al.*, 2016. A common framework for identifying linkage rules across different types of interactions. Functional Ecology 30 (12), 1894–1903. Available at: http://onlinelibrary.wiley.com/doi/10.1111/1365-2435.12666/full.

Bascompte, J., Melián, C.J., 2005. Simple trophic modules for complex food webs. Ecology 86 (11), 2868–2873. doi:10.1890/05-0101.

Bascompte, J., Melián, C.J., Sala, E., 2005. Interaction strength combinations and the overfishing of a marine food web. Proceedings of the National Academy of Sciences of the United States of America 102 (15), 5443–5447. doi:10.1073/pnas.0501562102.

Bascompte, J., Jordano, P., Olesen, J.M., 2006. Asymmetric Coevolutionary networks facilitate biodiversity maintenance. Science 312 (5772), 431–433. doi:10.1126/science.1123412.

Bastolla, U., Fortuna, M.A., Pascual-García, A., *et al.*, 2009. The architecture of mutualistic networks minimizes competition and increases biodiversity. Nature 458 (7241), 1018–1020. doi:10.1038/nature07950.

Boulangeat, I., Gravel, D., Thuiller, W., 2012. Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. Ecology Letters 15 (6), 584–593. doi:10.1111/j.1461-0248.2012.01772.x.

Brose, U., Ostling, A., Harrison, K., Martinez, N.D., 2004. Unified spatial scaling of species and their trophic interactions. Nature 428 (6979), 167–171. doi:10.1038/nature02297.

Brose, U., Jonsson, T., Berlow, E.L., *et al.*, 2006a. Consumer–resource body-size relationships in natural food webs. Ecology 87 (10), 2411–2417. doi:10.1890/0012-9658(2006)87[2411:CBRINF]2.0.CO;2.

Brose, U., Williams, R.J., Martinez, N.D., 2006b. Allometric scaling enhances stability in complex food webs. Ecology Letters 9 (11), 1228–1236. doi:10.1111/j.1461-0248.2006.00978.x.

Camacho, J., Stouffer, D., Amaral, L., 2007. Quantitative analysis of the local structure of food webs. Journal of Theoretical Biology 246 (2), 260–268. doi:10.1016/j.jtbi.2006.12.036.

Campbell, C., Yang, S., Albert, R., Shea, K., 2011. A network model for plant–pollinator community assembly. Proceedings of the National Academy of Sciences 108 (1), 197–202. doi:10.1073/pnas.1008204108.
Canard, E., Mouquet, N., Marescot, L., *et al.*, 2012. Emergence of structural patterns in neutral trophic networks. PLoS One 7 (8), e38295. doi:10.1371/journal.pone.0038295.
Canard, E.F., Mouquet, N., Mouillot, D., *et al.*, 2014. Empirical evaluation of neutral interactions in host-parasite networks. The American Naturalist 183 (4), 468–479. doi:10.1086/675363.
Capitán, J.A., Cuesta, J.A., Bascompte, J., 2009. Statistical mechanics of ecosystem assembly. Physical Review Letters 103 (16), 168101. doi:10.1103/PhysRevLett.103.168101.
Carstensen, D.W., Sabatino, M., Trøjelsgaard, K., Morellato, L.P.C., 2014. Beta diversity of plant-pollinator networks and the spatial turnover of pairwise interactions. PLoS One 9 (11), e112903. doi:10.1371/journal.pone.0112903.
Cattin, M.-F., Bersier, L.-F., Banašek-Richter, C., *et al.*, 2004. Phylogenetic constraints and adaptation explain food-web structure. Nature 427 (6977), 835–839. doi:10.1038/nature02327.
Cazelles, K., Araújo, M.B., Mouquet, N., Gravel, D., 2016. A theory for species co-occurrence in interaction networks. Theoretical Ecology 9 (1), 39–48. doi:10.1007/s12080-015-0281-9.
Chagnon, P.-L., 2015. Characterizing topology of ecological networks along gradients: The limits of metrics' standardization. Ecological Complexity 22, 36–39. Available at: http://www.sciencedirect.com/science/article/pii/S1476945X15000070.
Chesson, P., 2000. Mechanisms of maintenance of species diversity. Annual Review of Ecology and Systematics 31 (1), 343–366. doi:10.1146/annurev.ecolsys.31.1.343.
Crea, C., Ali, R.A., Rader, R., 2015. A new model for ecological networks using species-level traits. In: Methods in ecology and evolution. Available at: http://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12471/pdf.
Dalsgaard, B., Trøjelsgaard, K., Martín González, A.M., *et al.*, 2013. Historical climate-change influences modularity and nestedness of pollination networks. Ecography 36 (12), 1331–1340. doi:10.1111/j.1600-0587.2013.00201.x.
Darwin, C., 1859. On the origin of species by means of natural selection, or preservation of Favoured races in the struggle for life. London: John Murray.
Davis, A.J., Lawton, J.H., Shorrocks, B., Jenkinson, L.S., 1998. Individualistic species responses invalidate simple physiological models of community dynamics under global environmental change. Journal of Animal Ecology 67 (4), 600–612. doi:10.1046/j.1365-2656.1998.00223.x.
Delmas, E., Besson, M., Brice, M.-H., *et al.*, 2017. Analyzing ecological networks of species interactions. bioRxiv. 112540. doi:10.1101/112540.
Díaz-Castelazo, C., Guimarães, P.R., Jordano, P., *et al.*, 2010. Changes of a mutualistic network over time: Reanalysis over a 10-year period. Ecology 91 (3), 793–801. Available at: http://www.jstor.org/stable/25661111.
Donohue, I., Petchey, O.L., Montoya, J.M., *et al.*, 2013. On the dimensionality of ecological stability. Ecology Letters 16 (4), 421–429. doi:10.1111/ele.12086.
Drake, J.A., 1991. Community-assembly mechanics and the structure of an experimental species ensemble. The American Naturalist 137 (1), 1–26. doi:10.1086/285143.
Duffy, J.E., Cardinale, B.J., France, K.E., *et al.*, 2007. The functional role of biodiversity in ecosystems: Incorporating trophic complexity. Ecology Letters 10 (6), 522–538. doi:10.1111/j.1461-0248.2007.01037.x.
Dunne, J.A., Williams, R.J., Martinez, N.D., 2002. Network structure and biodiversity loss in food webs: Robustness increases with connectance. Ecology Letters 5 (4), 558–567. doi:10.1046/j.1461-0248.2002.00354.x.
Elton, C.C., 1958. The reasons for conservation. In: The ecology of invasions by animals and plants. Netherlands: Springer, pp. 143–153.doi:10.1007/978-94-009-5851-7_8.
Emmerson, M.C., Raffaelli, D., 2004. Predator–prey body size, interaction strength and the stability of a real food web. Journal of Animal Ecology 73 (3), 399–409. doi:10.1111/j.0021-8790.2004.00818.x.
Fontaine, C., Thébault, E., 2015. Comparing the conservatism of ecological interactions in plant–pollinator and plant–herbivore networks. Population Ecology 57 (1), 29–36. doi:10.1007/s10144-014-0473-y.
Fortuna, M.A., Bascompte, J., 2006. Habitat loss and the structure of plant–animal mutualistic networks: Mutualistic networks and habitat loss. Ecology Letters 9 (3), 281–286. doi:10.1111/j.1461-0248.2005.00868.x.
Fortuna, M.A., Stouffer, D.B., Olesen, J.M., *et al.*, 2010. Nestedness versus modularity in ecological networks: Two sides of the same coin? Journal of Animal Ecology 79 (4), 811–817. doi:10.1111/j.1365-2656.2010.01688.x.
González-Varo, J.P., Traveset, A., 2016. The labile limits of forbidden interactions. Trends in Ecology & Evolution 31 (9), 700–710. doi:10.1016/j.tree.2016.06.009.
Gotelli, N.J., Graves, G.R., Rahbek, C., 2010. Macroecological signals of species interactions in the Danish avifauna. Proceedings of the National Academy of Sciences 107 (11), 5030–5035. doi:10.1073/pnas.0914089107.
Gravel, D., Massol, F., Canard, E., *et al.*, 2011. Trophic theory of island biogeography: Trophic theory of island biogeography. Ecology Letters 14 (10), 1010–1016. doi:10.1111/j.1461-0248.2011.01667.x.
Gravel, D., Albouy, C., Thuiller, W., 2016. The meaning of functional trait composition of food webs for ecosystem functioning. Philosophical Transactions of the Royal Society B: Biological Sciences 371 (1694), 20150268. doi:10.1098/rstb.2015.0268.
Harvey, E., Séguin, A., Nozais, C., *et al.*, 2013. Identity effects dominate the impacts of multiple species extinctions on the functioning of complex food webs. Ecology 94 (1), 169–179. doi:10.1890/12-0414.1.
Havens, K., 1992. Scale and structure in natural food webs. Science 257 (5073), 1107–1109. doi:10.1126/science.257.5073.1107.
Heikkinen, R.K., Luoto, M., Virkkala, R., *et al.*, 2007. Biotic interactions improve prediction of boreal bird distributions at macro-scales. Global Ecology and Biogeography 16 (6), 754–763. doi:10.1111/j.1466-8238.2007.00345.x.
Holt, R.D., 1997. Community modules. In: Gange, A.C., Brown, V.K. (Eds.), Multitrophic interactions in terrestrial ecosystems.6th Symposium of the British Ecological SocietyBlackwell Science, pp. 333–350.
Hubbell, S.P., 2001. The unified neutral theory of biodiversity and biogeography (MPB-32). Princeton; Oxford: Princeton University Press, Available at: http://www.jstor.org/stable/j.ctt7rj8w.
Hutchinson, G.E., 1961. The paradox of the plankton. The American Naturalist 95 (882), 137–145. doi:10.1086/282171.
Jacquet, C., Moritz, C., Morissette, L., *et al.*, 2016. No complexity–stability relationship in empirical ecosystems. Nature Communications 7. 12573 doi:10.1038/ncomms12573.
Jacquet, C., Mouillot, D., Kulbicki, M., Gravel, D. (2017) 'Extensions of island biogeography theory predict the scaling of functional trait composition with habitat area and isolation', Ecology Letters. Edited by D. Storch, 20(2), pp. 135–146. https://doi.org/10.1111/ele.12716.
Jordano, P., Bascompte, J., Olesen, J.M., 2003. Invariant properties in coevolutionary networks of plant–animal interactions. Ecology Letters 6 (1), 69–81. doi:10.1046/j.1461-0248.2003.00403.x.
Kemp, J. E., Evans, D. M., Augustyn, W. J. and Ellis, A. G. (2017) 'Invariant antagonistic network structure despite high spatial and temporal turnover of interactions', Ecography, pp. n/a–n/a. https://doi.org/10.1111/ecog.02150.
Krasnov, B.R., Fortuna, M.A., Mouillot, D., *et al.*, 2012. Phylogenetic signal in module composition and species connectivity in compartmentalized host-parasite networks. The American Naturalist 179 (4), 501–511. doi:10.1086/664612.
Loreau, M., 2010. Linking biodiversity and ecosystems: Towards a unifying ecological theory. Philosophical Transactions of the Royal Society B: Biological Sciences 365 (1537), 49–60. doi:10.1098/rstb.2009.0155.
Lundberg, P., Ranta, E., Kaitala, V., 2008. Species loss leads to community closure. Ecology Letters 3 (6), 465–468. doi:10.1111/j.1461-0248.2000.00170.x.
MacArthur, R., 1955. Fluctuations of animal populations and a measure of community stability. Ecology 36 (3), 533. doi:10.2307/1929601.
MacArthur, R.H., Wilson, E.O., 1967. The theory of island biogeography. Princeton: Princeton University Press.

Malthus, T. R. (1798) 'An essay on the principle of population, as it affects the future improvement of society: With remarks on the speculations of Mr. Godwin, Mr. Condorcet, and other Writers.'
Martinez, N.D., 1992. Constant connectance in community food webs. The American Naturalist 139 (6), 1208–1218. Available at: http://www.jstor.org/stable/2462337.
Martinez, N.D., Williams, R.J., Dunne, J.A., 2006. Diversity, complexity, and persistence in large model ecosystems. In: Ecological Networks: Linking Structure to Dynamics in Food Webs., pp. 163–185.
Massol, F., Dubart, M., Calcagno, V., *et al.*, 2017. Island biogeography of food webs. In: Advances in Ecological Research. Elsevier, pp. 183–262. doi:10.1016/bs.aecr.2016.10.004.
May, R.M., 1972. Will a large complex system be stable? Nature 238 (5364), 413–414. doi:10.1038/238413a0.
McCann, K., Hastings, A., Huxel, G.R., 1998. Weak trophic interactions and the balance of nature. Nature 395 (6704), 794–798. doi:10.1038/27427.
McCann, K.S., 2000. The diversity–stability debate. Nature 405 (6783), 228–233. doi:10.1038/35012234.
Memmott, J., Waser, N.M., Price, M.V., 2004. Tolerance of pollination networks to species extinctions. Proceedings of the Royal Society B: Biological Sciences 271 (1557), 2605–2611. doi:10.1098/rspb.2004.2909.
Milo, R., 2002. Network motifs: Simple building blocks of complex networks. Science 298 (5594), 824–827. doi:10.1126/science.298.5594.824.
Montoya, J., Woodward, G., Emmerson, M.C., Solé, R.V., 2009. Press perturbations and indirect effects in real food webs. Ecology 90 (9), 2426–2433. doi:10.1890/08-0657.1.
Montoya, J.M., Solé, R.V., 2002. Small world patterns in food webs. Journal of Theoretical Biology 214 (3), 405–412. doi:10.1006/jtbi.2001.2460.
Montoya, J.M., Pimm, S.L., Solé, R.V., 2006. Ecological networks and their fragility. Nature 442 (7100), 259–264. doi:10.1038/nature04927.
Morales-Castilla, I., Matias, M.G., Gravel, D., Araújo, M.B., 2015. Inferring biotic interactions from proxies. Trends in Ecology & Evolution 30 (6), 347–356. doi:10.1016/j.tree.2015.03.014.
Olesen, J.M., Bascompte, J., Elberling, H., Jordano, P., 2008. Temporal dynamics in a pollination network. Ecology 89 (6), 1573–1582. doi:10.1890/07-0451.1.
Olesen, J.M., Bascompte, J., Dupont, Y.L., *et al.*, 2011. Missing and forbidden links in mutualistic networks. Proceedings of the Royal Society of London B: Biological Sciences 278 (1706), 725–732. doi:10.1098/rspb.2010.1371.
Pearson, R.G., Dawson, T.P., 2003. Predicting the impacts of climate change on the distribution of species: Are bioclimate envelope models useful? Global Ecology and Biogeography 12 (5), 361–371. doi:10.1046/j.1466-822X.2003.00042.x.
Pearson, R.G., Dawson, T.P., Berry, P.M., Harrison, P.A., 2002. SPECIES: A spatial evaluation of climate impact on the envelope of species. Ecological Modelling 154 (3), 289–300. doi:10.1016/S0304-3800(02)00056-X.
Petchey, O., Eklöf, A., Borrvall, C., Ebenman, B., 2008. Trophically unique species are vulnerable to cascading extinction. The American Naturalist 171 (5), 568–579. doi:10.1086/587068.
Poisot, T., Gravel, D., 2014. When is an ecological network complex? Connectance drives degree distribution and emerging network properties. PeerJ 2. e251 doi:10.7717/peerj.251.
Poisot, T., Canard, E., Mouillot, D., *et al.*, 2012. The dissimilarity of species interaction networks. Ecology Letters 15 (12), 1353–1361. doi:10.1111/ele.12002.
Poisot, T., Mouquet, N., Gravel, D., 2013. Trophic complementarity drives the biodiversity-ecosystem functioning relationship in food webs. Ecology Letters 16 (7), 853–861. doi:10.1111/ele.12118.
Poisot, T., Stouffer, D.B., Gravel, D., 2015. Beyond species: Why ecological interaction networks vary through space and time. Oikos 124 (3), 243–251. doi:10.1111/oik.01719.
Poisot, T., Guéveneux-Julien, C., Fortin, M.-J., *et al.*, 2017. Hosts, parasites and their interactions respond to different climatic variables. Global Ecology and Biogeography 26 (8), 942–951. doi:10.1111/geb.12602.
Pollock, L.J., Tingley, R., Morris, W.K., *et al.*, 2014. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). Methods in Ecology and Evolution 5 (5), 397–406. doi:10.1111/2041-210X.12180.
Rezende, E.L., Jordano, P., Bascompte, J., 2007. Effects of phenotypic complementarity and phylogeny on the nested structure of mutualistic networks. Oikos 116 (11), 1919–1929. doi:10.1111/j.0030-1299.2007.16029.x.
Rezende, E.L., Albert, E.M., Fortuna, M.A., Bascompte, J., 2009. Compartments in a marine food web associated with phylogeny, body mass, and habitat structure. Ecology Letters 12 (8), 779–788. doi:10.1111/j.1461-0248.2009.01327.x.
Simberloff, D.S., Wilson, E.O., 1969. Experimental zoogeography of islands: The colonization of empty islands. Ecology 50 (2), 278–296. doi:10.2307/1934856.
Solé, R.V., Montoya, M., 2001. Complexity and fragility in ecological networks. Proceedings of the Royal Society of London B: Biological Sciences 268 (1480), 2039–2045. doi:10.1098/rspb.2001.1767.
Stouffer, D.B., Bascompte, J., 2010. Understanding food-web persistence from local to global scales. Ecology Letters 13 (2), 154–161. doi:10.1111/j.1461-0248.2009.01407.x.
Stouffer, D.B., Sales-Pardo, M., Sirer, M.I., Bascompte, J., 2012. Evolutionary conservation of species' roles in food webs. Science 335 (6075), 1489–1492. doi:10.1126/science.1216556.
Sugihara, G., Ye, H., 2009. Complex systems: Cooperative network dynamics. Nature 458 (7241), 979–980. doi:10.1038/458979a.
Thébault, E., Fontaine, C., 2010. Stability of ecological communities and the architecture of mutualistic and trophic networks. Science 329 (5993), 853–856. doi:10.1126/science.1188321.
Thebault, E., Loreau, M., 2003. Food-web constraints on biodiversity-ecosystem functioning relationships. Proceedings of the National Academy of Sciences 100 (25), 14949–14954. doi:10.1073/pnas.2434847100.
Thébault, E., Huber, V., Loreau, M., 2007. Cascading extinctions and ecosystem functioning: Contrasting effects of diversity depending on food web structure. Oikos 116 (1), 163–173. doi:10.1111/j.2006.0030-1299.15007.x.
Trøjelsgaard, K., Olesen, J.M., 2016. Ecological networks in motion: Micro- and macroscopic variability across scales. Functional Ecology 30 (12), 1926–1935. doi:10.1111/1365-2435.12710.
Trøjelsgaard, K., Jordano, P., Carstensen, D.W., Olesen, J.M., 2015. Geographical variation in mutualistic networks: Similarity, turnover and partner fidelity. Proceedings of the Royal Society B: Biological Sciences 282 (1802), 20142925. doi:10.1098/rspb.2014.2925.
Verdú, M., Valiente-Banuet, A., 2008. The nested assembly of plant facilitation networks prevents species extinctions. The American Naturalist 172 (6), 751–760. doi:10.1086/593003.
Verhulst, P., 1938. Notice sur la loi que la population suit dans son accroissement. Correspondance mathématique et physique 10, 113–121.
Williams, R. J. (2011) 'Biology, methodology or chance? The degree distributions of bipartite ecological networks', PLoS One. Edited by J. Langowski, 6(3), p. e17645. https://doi.org/10.1371/journal.pone.0017645.
Yodzis, P., 1981. The stability of real ecosystems. Nature 289 (5799), 674–676. doi:10.1038/289674a0.

Further Reading

Bartomeus, I., Gravel, D., Tylianakis, J.M., *et al.*, 2016. A common framework for identifying linkage rules across different types of interactions. Functional Ecology 30 (12), 1894–1903. Available at: http://onlinelibrarywiley.com/doi/10.1111/1365-2435.12666/full.
Bascompte, J., 2007. Networks in ecology. Basic and Applied Ecology 8 (6), 485–490. doi:10.1016/j.baae.2007.06.003.
Cohen, J.E., 1989. Food webs and community structure. Perspectives in Ecological Theory. 181–202.

Dunne, J.A., Williams, R.J., Martinez, N.D., 2002. Food-web structure and network theory: The role of connectance and size. Proceedings of the National Academy of Sciences 99 (20), 12917–12922. doi:10.1073/pnas.192407699.
Evans, D. M., Kitson, J. J. N. and Lunt, D. H. et al. (2016) 'Merging DNA metabarcoding and ecological network analysis to understand and build resilient terrestrial ecosystems', Functional Ecology. Edited 30(12), pp. 1904–1916. https://doi.org/10.1111/1365-2435.12659.pdf.
Gravel, D., Canham, C.D., Beaudet, M., Messier, C., 2006. Reconciling niche and neutrality: The continuum hypothesis. Ecology Letters 9 (4), 399–409. doi:10.1111/j.1461-0248.2006.00884.x.
Gravel, D., Poisot, T., Albouy, C., Velez, L., Mouillot, D., 2013. Inferring food web structure from predator–prey body size relationships. Methods in Ecology and Evolution 4 (11), 1083–1090. doi:10.1111/2041-210X.12103.
Jordano, P. (2016) 'Sampling networks of ecological interactions. Functional Ecology Edited by D. Stouffer, 30(12), pp. 1883–1893. https://doi.org/10.1111/1365-2435.12763.
Jordano, P., Bascompte, J., Olesen, J.M., 2003. Invariant properties in coevolutionary networks of plant–animal interactions. Ecology Letters 6 (1), 69–81. doi:10.1046/j.1461-0248.2003.00403.x.
Kéfi, S., Berlow, E.L., Wieters, E.A., *et al*., 2012. More than a meal… integrating non-feeding interactions into food webs. Ecology Letters 15 (4), 291–300. doi:10.1111/j.1461-0248.2011.01732.x.
Kraft, N., Cornwell, W., Webb, C., Ackerly, D., 2007. Trait evolution, community assembly, and the phylogenetic structure of ecological communities. The American Naturalist 170 (2), 271–283. doi:10.1086/519400.
Leibold, M.A., Chase, J.M., Ernest, S.K.M., 2017. Community assembly and the functioning of ecosystems: How metacommunity processes alter ecosystems attributes. Ecology 98 (4), 909–919. doi:10.1002/ecy.1697.
Maherali, H., Klironomos, J.N., 2007. Influence of phylogeny on fungal community assembly and ecosystem functioning. Science 316 (5832), 1746–1748. doi:10.1126/science.1143082.
Martinez, N.D., 1992. Constant Connectance in community food webs. The American Naturalist 139 (6), 1208–1218. Available at: http://www.jstor.org/stable/2462337
Montoya, J.M., Solé, R.V., 2003. Topological properties of food webs: From real data to community assembly models. Oikos 102 (3), 614–622. doi:10.1034/j.1600-0706.2003.12031.x.
Montoya, J.M., Pimm, S.L., Solé, R.V., 2006. Ecological networks and their fragility. Nature 442 (7100), 259–264. doi:10.1038/nature04927.
Olesen, J.M., Bascompte, J., Dupont, Y.L., *et al*., 2011. Missing and forbidden links in mutualistic networks. Proceedings of the Royal Society of London B: Biological Sciences 278 (1706), 725–732. doi:10.1098/rspb.2010.1371.
Peralta, G., 2016. Merging evolutionary history into species interaction networks. Functional Ecology 30 (12), 1917–1925. doi:10.1111/1365-2435.12669.
Pilosof, S., Porter, M.A., Pascual, M., Kéfi, S., 2017. The multilayer nature of ecological networks. Nature Ecology & Evolution 1 (4), 0101. doi:10.1038/s41559-017-0101.
Souza, A.F., Bezerra, A.D., Longhi, S.J., 2016. Quasi-neutral community assembly: Evidence from niche overlap, phylogenetic, and trait distribution analyses of a subtropical forest in South America. Perspectives in Plant Ecology, Evolution and Systematics 23, 1–10. doi:10.1016/j.ppees.2016.09.006.
Tilman, D., 2004. Niche tradeoffs, neutrality, and community structure: A stochastic theory of resource competition, invasion, and community assembly. Proceedings of the National Academy of Sciences of the United States of America 101 (30), 10854–10861. doi:10.1073/pnas.0403458101.
Trøjelsgaard, K., Olesen, J.M., 2016. Ecological networks in motion: Micro and macroscopic variability across scales. Functional Ecology 30 (12), 1926–1935. doi:10.1111/1365-2435.12710.
Williams, R.J., Martinez, N.D., 2000. Simple rules yield complex food webs. Nature 404 (6774), 180–183. doi:10.1038/35004572.
Williams, R.J., Brose, U., Martinez, N.D., 2007. Homage to Yodzis and Innes 1992: Scaling up feeding-based population dynamics to complex ecological networks. In: From energetics to ecosystems: The dynamics and structure of ecological systems. Springer, pp. 37–51. Available at: http://link.springer.com/content/pdf/10.1007/978-1-4020-5337-5_2.pdf.

Complex Systems

Mikhail Prokopenko, The University of Sydney, Sydney, NSW, Australia

Glossary

Allelomimesis A range of activities in which the performance of a behavior increases the probability of that behavior being performed by other nearby animals.
Autocatalysis A single chemical reaction is autocatalytic, if one of the reaction products is also a reactant and therefore a catalyst in the same or a coupled reaction.
Homeostasis The property of a system to actively regulate a specific variable, keeping it practically constant.
Order parameter A macroscopic variable used to describe and measure the degree of order across the boundaries in a phase transition system (normally ranges between zero in one phase and nonzero in the other).
Phase transition A change in a feature of a physical system, often involving the absorption or emission of energy from the system and resulting in a transition of the system to another phase.
Propagule Any material that is used in propagating an organism to the next stage in their life cycle, for example, by dispersal of plant material (stem cuttings or seeds) for plant propagation in horticulture, or transmitting a disease by infectious components generated by pathogens.
Shannon entropy The quantity is defined in the context of a probabilistic model and measures the uncertainty, or the average (expected) surprise of an entity, given an ensemble in which it is embedded. Shannon information is a measure of the reduction of uncertainty.
Small-world network A graph in which the neighbors of any given vertex (node) are likely to be neighbors of each other (in other words, the average clustering is high), and most nodes can be reached from every other node by a small number of hops (formally, the average shortest path length is small). Small-world effect explains how strangers are linked by a short chain of acquaintances.
Stigmergy A mechanism of indirect coordination, through the environment, between agents or actions.
Tipping point (environmental) A threshold beyond which some (typically accumulated) small changes may force an (eco-) system to rapidly and often irreversibly change to a new state, with significant effects on biodiversity at a regional or global scale.

Introduction

A complex system typically contains a large number of individual components (often referred to as agents) that interact with each other and the surrounding environment. The number of agents is usually sufficiently large so that no specific feature controls or dominates the overall dynamics of the system, but not large enough to make individual features completely irrelevant. The agent interactions follow simple rules dependent on the agents' perceptions of their local conditions. For example, pheromone-depositing ants foraging for food indirectly interact with each other through their environment. The ants employ a form of autocatalytic behaviour called allelomimesis: the probability with which an ant chooses a trail, and therefore, deposits new pheromone there, increases with the number of ants that chose the same path in the past. As more ants, attracted by the pheromone, cross the same area and lay more pheromones, making the "chemical" path even more attractive. As pheromone slowly evaporates, shorter paths, which keep being reinforced by attracted ants, become more prominent than longer ones, and an optimal path emerges. The process is thus characterized by (i) positive feedback loops, making shorter paths stronger in terms of the deposited pheromone, and (ii) nonlinear dynamics of the pheromone concentration, which is increasing with the number of ants traversing the formed paths and decreasing due to the chemical's natural evaporation. Feedback loops and nonlinear dynamics are two other common features of complexity.

A variant of complex systems is complex adaptive systems (CAS), in which the agents adapt or learn during their interactions. Holland (2006) defined four distinguishing attributes of CAS: (i) parallelism: a large number of simultaneously interacting agents exchange signals concurrently (e.g., the reaction cascades and cycles within biological cells are continually coordinated in order to ensure that a cell maintains its functions); (ii) conditional action: the agents' actions depend on the received signals or other local perceptions, constructing flexible behaviors that involve intricate positive and negative feedbacks; (iii) modularity: an agent's behaviour often includes different groups of rules, acting as "subroutines" or building blocks, that can be recombined to respond to novel situations; the usefulness of such modules may be confirmed or disconfirmed based on their success in dealing with the challenging circumstances (e.g., the Krebs cycle—the citric acid cycle in biological cells, which incorporates eight proteins that interact to form a loop—is a basic series of chemical reactions used to release stored energy by all aerobic organisms, ranging from bacteria to elephants); (iv) adaptation and evolution: the agents adapt in order to improve their performance over time (the adaptations, that can be structural, behavioral and physiological, help organisms survive in their ecological niche or habitat, for instance, the chameleon camouflage, birds migrations and the snakes ability to make venom).

Adaptive behavior can take many forms. One important example is homeostasis: a dynamic process of self-regulation and adaptation by which an overall system or individual agents adapt their behavior over time in order to stay close to a certain state (or a set of states). Homeostatic behavior is related to Ashby's Law of Requisite Variety which states that in order to maintain a dynamic stability under perturbation (or input), an active controller requires as much variety (the number of states) as that of the controlled system. In other words, the greater the variety within the system, the more resilient it is.

Turning our attention to the interaction aspect of complexity, we point out that the agents may interact not only with each other but also with the surrounding environment. This may play an important role in sustaining or maintaining its ecosystem, for example, bees pollinate flowers, facilitating the production of fruits or seeds. Such agent–environment interactions may have both positive and negative feedback loops, and typically create nonlinear dynamics.

There are also examples of interacting sub-systems such as predator–prey interactions which shape an integrated ecosystem. Predator–prey models, for example, the well-known Lotka–Volterra model, include a pair of coupled differential equations that describe the dynamics of two populations, one of which (the predator) grows at the expense of the other one (the prey). Under some simplifying assumptions, these models are able to demonstrate the coupled dynamics which are characterized by (i) oscillations in the population sizes of both predator and prey, and (ii) a time lag between the prey and predator abundances: the peak of the predator's oscillation follows the peak of the prey's oscillation. However, as predators consume the prey, the prey's abundance greatly diminishes causing the predator population to reduce. This eventually removes immediate threats to the prey, swinging its population up and resetting the coupled cycles. From the evolutionary perspective, predators and prey can influence one another's evolution, and such co-evolution need to be modeled on longer (slower) time scales than the dynamics of population sizes.

The "mass-action" models, such as the Lotka–Volterra model, have a broad range. They can describe chemical reactions, as well as the compartmental models of disease spread in mathematical epidemiology, in which the population is split into different compartments, such as susceptible to the infection; infected; recovered, and so on, interacting with each other over time. The "mass-action" models conceptually differ from agent-based models which are more applicable to studies of complex systems. The agent-based models attempt to explicitly represent behaviors of individual agents and their local interactions, including their short-term adaptations and the long-term evolutionary changes, and so the resultant global dynamics are not specified directly, for example, by some differential equations, but are expected to emerge.

Emergence

In general, the global dynamics of a complex system which result from the agents' interactions are emergent: they cannot be predicted, or explained, as an aggregation of the agents' individual dynamics. In other words, the interactions among the constituent microscopic parts (at the bottom level) bring about macroscopic phenomena (at the top level) that cannot be understood by considering any single part alone. These macroscopic features often display synergy ("the whole is more than the sum of the parts"). Continuing with the ant colony example, we note that an optimal path, or a network of paths, connecting the colony nest with food sources, emerges as a result of the pheromone-depositing ants navigating around various obstacles present in the environment. The result of this distributed process cannot be reduced to the local actions. In other words, the optimal path is an emergent property of the distributed behaviour of the ants, constrained by the locations and shapes of the obstacles.

Another example of emergent behavior is given by swarms of animals. Animal groups in nature often exhibit spatial aggregation, for example, schools of fish, swarms of locusts, herds of wildebeest, and flocks of birds (Parrish and Edelstein-Keshet, 1999). Such aggregations may provide individuals with protection, mate choices, foraging, habitat assessment, migratory routes, etc. (Camazine et al., 2001). Being a typical case of complex systems, an aggregated flocking behavior results from the local behaviors of individual birds which try to stay in close proximity, while avoiding collisions and flying in the same general direction. Again we point out that complex large-scale patterns and structures emerge within swarms and flocks through individual decisions, based on perception of the individuals' local conditions. Interestingly, these local perceptions propagate through the collective as waves or cascades. Formation of signaling waves is a widespread phenomenon observed in animal groups, rapidly transferring information over long ranges (Potts, 1984). Information cascades in collective systems often create a positive feedback loop, due to a rapid autocatalytic

response to changing conditions. This heightened adaptive response allows the group to be sensitive even to weak or ambiguous external stimuli.

Emergent behaviour is related to tangled hierarchies exhibiting Strange Loops (Hofstadter, 1989):

> "an interaction between levels in which the top level reaches back down towards the bottom level and influences it, while at the same time being itself determined by the bottom level."

Emergence is often confused with self-organization and the primary difference between these two concepts lies in their relation to the levels of system. While feedback loops are possible both within and across the levels of the system, it is important to distinguish between (i) emergence, which assumes another level of description where the emergent properties can be used for better predictions by an external observer, and (ii) self-organization which occurs within a complex system, that is, within one level, affecting the functioning of the system itself (Prokopenko et al., 2009).

Self-Organization

Self-organization is typically defined as the evolution of a system into an organized form in the absence of external pressures explicitly guiding the local behavior of the sub-systems. For example, an eco-system may develop spatial structures in order to deal with diminishing resources. The prefix "self" should not be taken literally: it only indicates the absence of centralized ordering or explicitly defined external template prescribing the internal dynamics. In general, we deal with open systems, exchanging energy, matter and/or information with the surrounding environment, and assume that the systems contain components and constraints which are defined prior to the organization itself (Prokopenko et al., 2009). Camazine et al. (2001) offered the following definition in the context of pattern formation in biological systems:

> "Self-organization is a process in which pattern at the global level of a system emerges solely from numerous interactions among the lower-level components of the system. Moreover, the rules specifying interactions among the system's components are executed using only local information, without reference to the global pattern."

Another definition is offered by Haken (2006) from a more generic perspective:

> "A system is self-organizing if it acquires a spatial, temporal, or functional structure without specific interference from the outside. By "specific" we mean that the structure or functioning is not impressed on the system but that the system is acted upon from the outside in a non-specific fashion. For instance, the fluid which forms hexagons is heated from below in an entirely uniform fashion and it acquires its specific structure by self-organization."

This physics-based view is related to the approach pioneered by Kauffman (2000) who suggested that the underlying principle of self-organization is the generation of constraints in the release of energy. According to this standpoint, the constrained release allows for such energy to be controlled and channeled to perform some useful work. This work in turn can be used to build better and more efficient constraints for the release of further energy and so on. The ability to constrain and control the release of energy may allow a system to produce behaviors which, although possible, would be extremely unlikely in its non-organized state (Kauffman, 2000; Prokopenko et al., 2009).

An interesting form of self-organization is stigmergy. Grassé (1959) introduced the term "stigmergy" ("previous work directs and triggers new building actions") to describe a decentralized pathway of information flow in social insects. Stigmergy is a mechanism of indirect coordination among agents acting in the environment, where local traces left in the environment by their decentralized actions stimulate the performance of subsequent actions, by the same or a different agent (which is exactly the case of pheromone-depositing ants). Thus, stigmergy allows the environment to structure itself through the activities: the state of the environment, and the current distribution of agents within it, determine how the environment and the distribution of agents will change in the future (Bonabeau et al., 1998). Emphasizing the role of constraints in controlling the release of energy within a self-organizing system, we point out that an optimal path formation may happen by chance even with a random exploration of the landscape, but is much more likely when stigmergy is present.

One of the modern research themes in complex systems is Guided Self-Organization (GSO). The main objective of GSO is to leverage the strengths of self-organization (such as simplicity of the local behaviors, parallelization, adaptability of the agents, robustness, resilience and scalability of the overall system), while still being able to direct the outcomes of the self-organizing process. In other words, GSO approaches differentiate between the concepts of "control" and "constraint": rather than trying to precisely control a transition towards the desirable outcomes, one puts in place specific constraints on the system dynamics in order to mediate agent behaviors, interactions and adaptations. Thermodynamically, a spontaneous increase in order within a self-organizing complex system must be offset by a production and export of entropy to the external environment. Thus, GSO attempts to harness the order-inducing potential of self-organization by guiding the system towards a desired regime or state, while "exporting" the entropy to the system's exterior.

Critical Dynamics

Self-organization, and complexity in general, is strongly related to critical phenomena: the spatiotemporal behaviour of dynamical systems at an order/disorder phase transition: the critical regime separating the two phases is often referred to as the "edge of chaos."

A system is said to exhibit the property of chaos if a slight change in the initial conditions results in large-scale but bounded differences in the result (in other words, the system is sensitive to initial conditions; such outcomes are often referred to as the butterfly effect). Importantly, there are transitions separating ordered and chaotic regimes, and by varying control parameters (for instance, the system composition and the strength of internal interactions) it is possible to trigger these phase transitions. Following Ginzburg-Landau theory of phase transitions developed in physics, Haken (2006) introduced order parameters in explaining structures that spontaneously self-organize in nature. In an ordered phase, the system becomes low-dimensional as some dominant variables "enslave" others, making the whole system act in synchrony (e.g., a laser: a beam of coherent light created out of the chaotic movement of particles).

Phase transitions are often related to symmetry breaking. An example is a ferromagnetic system undergoing a second-order phase transition: (i) in the high-temperature phase the system has no net magnetization, is "disordered" and has a complete rotational symmetry (isotropy); (ii) at low temperature, the system becomes "ordered," and the net magnetization defines a preferred direction in space (anisotropy), breaking rotational symmetry. The low-temperature ordered phase is therefore less symmetrical and can be fully described by an order parameter—the magnetization vector. In this example, the temperature is a control parameter. But in a large complex ecological system, temperature can also be an order parameter, summarizing the effect of the sun, air pressure, and other atmospheric variables.

A well-studied model in the field of complex systems—random Boolean networks, used to simulate Gene Regulatory Networks—also exhibits a phase transition between ordered and chaotic dynamics, separated by a critical regime. The nodes of a random Boolean network may approximate interconnected genes with binary states, which switch on and off dependent on the input signals received along the network links from the connected neighbors. The "logic" of switching the nodes states on and off mimics the process of gene expression, and one of the main questions is whether the overall network, starting from a particular state or responding to a local perturbation, settles into a stable configuration, or keeps changing. A fundamental result is that, at relatively low connectivity (i.e., a low average number of network links) or dynamic activity (i.e., an extremely biased probability of state changes to either zero or one), a perturbed network remains in an ordered stable phase. This phase is characterized by high regularity of states and strong convergence of similar global states in state space, providing an analogue of high stability of the genotype. Alternatively, at relatively high connectivity and/or activity, the network never settles and ends up in a chaotic phase, characterized by low regularity of states and divergence of similar global states: an analogue of high adaptability of the genotype. In the critical regime (the edge of chaos), there is a balance between stability and adaptability.

A generic measure that has been found useful in studies of critical dynamics in complex systems is Fisher information. It quantifies the amount of information in an observable variable about a control parameter, and thus estimates sensitivity of the observed variable to changes in this parameter. Thermodynamically, Fisher information is proportional to the gradient of the corresponding order parameter, diverging when the system undergoes a phase transition at a critical point. Calculating this measure requires only appropriately defined probability densities and so the method can be applied to a wide range of systems, especially those in which computation of the order parameter is problematic.

Critical regimes are typical in proximity of various tipping points, occurring when a small change (or a number of previously accumulated changes) triggers a strong or even catastrophic response, amplified by positive feedbacks (Scheffer et al., 2009), for example, tipping points in climate and ecosystems such as the Amazon rainforest "die-back" and El Niño/Southern Oscillation effect. An example of a positive feedback intensifying the effects of the initial perturbation is a loop moving from deforestation reducing regional precipitation to increasing risk of fires to extending forest die-back to causing droughts.

Even robust and homeostatic systems may not withstand pressures accumulated in the lead up to a tipping point. A transition beyond a tipping point may be irreversible, but if the system which has undergone the transition is adaptive, then it may recover, and the speed of the recovery is dependent on the system's resilience. Resilience, thus, is the capacity of systems to survive and adapt despite acute shocks and chronic stresses they experience—it captures the ability to pull together and bounce back after a crisis. Resilience is directly related to self-organization, reflecting the degree to which a complex adaptive system is capable of self-organization while absorbing recurrent disturbances such as hurricanes or floods and retaining essential structures, processes, and feedbacks (Adger et al., 2005).

Entropy and Information in Adaptation and Evolution

Complex systems can also be viewed as distributed information-processing systems, with information being a crucial currency for animals and species from both a behavioral and evolutionary perspective. Behaviourally, many decisions benefit from a reduction in uncertainty: a school of fish reshapes to evade a predator, slime mold parallelizes its search for food, animals optimize their assortative mating choices, and so on.

Shannon entropy is a measure of potential knowledge, or if applied to a sequence, a measure of how much information a sequence could hold. For example, as argued by Adami (2002), Shannon entropy may quantify our uncertainty about the genetic

identity of a randomly selected individual. Shannon information is a nonlinear form of correlation, with respect to the system that the information is about. Without such reference, it is only potential information (i.e., entropy).

Following information-theoretic interpretation of Adami (2002), adaptation may then be seen to increase mutual (Shannon) information between a system and its environment. From this standpoint, the variation decreases the amount of information encoded in the system, whereas the selection acts to increase the information. The information loss due to variation must be less than the increase in mutual information due to selection. Assuming that the generations are coupled by inheritance with variation under selective pressure, the adaptive process reduces to evolution. In general, the evolution should increase the amount of information which a population (more precisely, its genome) contains about its niche, measured by "physical complexity" (Adami, 2002).

A similar idea about the relevant biological information is expressed by Haken (2006): the information "acquires its meaning only with respect to the surroundings and, in a way, with respect to its value for the survival of the whole species." Haken (2006) concludes that

> "by the interplay of mutation and selection new types of molecules and their corresponding phenotypes are then generated and in this way we observe the creation of new information. But whether this information is useful or not can be checked upon only by the interaction of the particular species with its environment."

In general, the increase in organization can be measured quantitatively as a decrease of Shannon entropy, exported by the self-organizing system into its surroundings (sometimes referred to as negentropy: the entropy that the system dissipates to keep its own entropy low). The systems which continuously export entropy in order to maintain their organization are called dissipative structures.

A recently developed framework of information dynamics systematically quantifies information processing in complex systems (Lizier et al., 2014), relating it to critical phenomena and phase transitions. This methodology suggests that discovering and quantifying information flows in complex systems could be a key to guiding the system dynamics towards desirable outcomes. Importantly, rather than trying to quantify the level of system's complexity with a single general-purpose measure, the information dynamics framework includes several dimensions, capturing different aspects of distributed processing within the system: memory, communications (interactions) and modifications. Memory refers to the storage of information by some agent(s) and can affect the future computation. Communication is understood as the transfer of information between one agent and another. Modification is the fusion of stored and/or transferred information into a new form. Within the space formed by these basic axes, the complexity of different systems may be measured and compared in terms of their specific components.

These three information-processing components originate in the field of distributed computation, but have been identified in many complex systems, ranging from natural to technological, where a system is seen as "computing" when it updates its states while moving along a global dynamic trajectory. For example, "ecological memory" can be seen in the biological legacies persisting after tipping points, various crises and disturbances, through mobile species and propagules that colonize and reorganize disturbed sites and refuges (Adger et al., 2005). The aspect of "ecological interactions" between species reflects diverse dependencies in ecological communities and food webs, exemplified by oppositional relationships such as predation and competition, as well as symbiotic relationships such as mutualism. Finally, "ecological modifications" are immediately evident in agricultural expansions and intensifications altering the quantity and quality of global water flows, deforestations or plant successions replacing forests with tundra, and ecological niche constructions, when organisms alter either their own environment or that of other species.

The concept of information modification is strongly related to the notion of synergy. Understanding and quantifying synergy, as well as modeling of critical phenomena, require a systematic consideration of system dynamics from a thermodynamic perspective which elucidates the analysis of phase transitions and synergistic interactions. Information thermodynamics is an emergent field of research attempting to treat dynamics of complex systems, while measuring and contrasting the entropic and energetic costs of manipulating information.

Complex Networks

Sometimes, a complex system is modeled as a network where the nodes represent the components and the links their interactions. Network topology is the specific type of an arrangement of the nodes and edges, for example, a star topology, in which each node (vertex) is connected to a central hub with a node-to-node link (edge), or a random graph topology, which is generated by a probability distribution, so that a possible link occurs independently with a given probability. Network representations create the possibility of studying phenomena at multiple scales under the same formalism, relating topological and dynamics properties.

In models of complex networks it is possible to study various propagation, diffusion or contagion processes, when a single event or disruption triggers a cascade of failures. For example, the emergence of a new pathogen in a remote village can give rise to a devastating global epidemic; the introduction of an exotic new species can eventually contribute to a chain of food-web disruptions and wide ecosystem collapses.

In complex networks, the ability to function effectively arises not from individual network nodes, but rather from the way they interact. This means that a complex network, like any complex system, cannot be completely understood by examining each of its parts in isolation. Specifically, for complex networks, the topology and function of such networks are tightly coupled (Newman,

2003): the function is constrained by the structure and the structure evolves due to function. Research into the structure, function, evolution, and design of complex networks has wide-ranging applications, from epidemiological modeling to understanding of food webs.

In a food-web network the nodes represent species in an ecosystem and a directed link indicates that one species preys on the other. This implicitly captures the flow of energy or carbon flow within the system, in the direction opposite to predation. As the structure of a food web typically affects the population dynamics, considering a network representation enhances the canonical predator–prey models. One complexity measure which has been found to be particularly useful in studies of food webs and their stability is "connectance": the fraction of all possible links that are realized in a network.

Complex networks is a vigorous field of research, comprising studies of network properties (such as the small-world effect, scale-free distributions, centrality); models of network growth (for instance, preferential attachment giving rise to scale-free topologies); mixing patterns such as assortativity (exemplified by the tendency where highly connected nodes are more likely to make links with other highly connected links representing assortative mating choices); community structure and modularity (for instance, communities in food webs might reflect subsystems within ecosystems); dynamic processes taking place on networks (e.g., epidemiological processes, spread of cascading failures or food-web disruptions, percolation, reaction-diffusion), network motifs (such as a diamond food-web motif with a predator consuming two prey species which in turn compete for a shared resource), and so on.

In random and small-world Boolean networks, the ordered phase of dynamics is typically dominated by information storage (representing "memory"), while the chaotic phase is dominated by information transfer (which includes both the "interaction" and "modification" components of distributed computation). Interestingly, the small-world topology allows the information dynamics to attain a balance, with both memory and information transfer approaching their maxima near the small-world transition between ordered and random graphs.

A very-well studied topology is scale-free networks, in which a relatively small number of nodes are connected to a tremendous number of neighbors, while the vast majority of other nodes have only a very few connections (formally, the distribution of connections follows a power law). Importantly, this feature is preserved at many levels of magnification, as zooming into any part of the network does not change the profile of the observed distribution of connections. Scale-free networks are robust to random node failures but become fragmented after coordinated attacks targeting the nodes with the highest number of neighbors (the hubs).

Food webs are not generally classified as small-world or scale-free networks. Nevertheless, food-web topology is consistent with some patterns found within those networks classes, and such patterns can be used to explore and predict functional responses of ecosystems to structural changes (Dunne et al., 2002).

Conclusion: Complex Versus Complicated

Complex systems should not be confused with complicated systems which may also contain a large number of components and conditional interactions. The two terms share a common Latin origin: complexus originates from complectⷭ ("to entwine or encircle"), derived in turn from com- ("together") and plectere ("to weave"), while complicātus is a form of complicāre ("to fold together') which augments com- ("together') with plecāre ("to fold'). So the etymological difference reflects the distinction between (flexibly) weaving and (rigidly) folding some parts together.

This difference becomes even quite apparent when one compares (complex) natural organisms which have evolved their adaptive and self-organizing responses with (complicated) engineered machines which conform to precise blueprints and operate under predefined protocols. The machines, and engineered structures in general, are designed and assembled to repeatedly perform within some well-defined cycles, by exploiting their design and operational constraints which are not generated anew but only reset, by a precise channeling of the external energy. On the contrary, the stability and resilience of biological and ecological systems are explained by their capacity to self-organize and homeostatically re-establish itself, while re-using the external energy and information in composing new internal constraints. These two approaches deal with the question of robustness in significantly different ways: (i) traditionally engineered systems are well-tested and validated for a wide range of known conditions, but may crumble at an encounter with a new situation or when an individual component fails, (ii) complex adaptive systems are able to re-combine the distributed information-processing behaviors of their constituent agents in innovating new solutions, often even beyond tipping points. Nevertheless, when faced with accumulated and chronic adverse pressures, even an adaptive system may undergo a shift to a post-critical phase incompatible with its primary purpose, highlighting the need for more accurate and predictive analysis and modeling of complex systems. Understanding and managing complexity at multiple scales improves our ability to maintain and restore ecosystems with timely and precise interventions.

References

Adami C (2002) What is complexity? *BioEssays* 24(12): 1085–1094.
Adger WN, Hughes TP, Folke C, Carpenter SR, and Rockström J (2005) Social-ecological resilience to coastal disasters. *Science* 309(5737): 1036–1039.
Bonabeau E, Theraulaz G, Fourcassié V, and Deneubourg J-L (1998) Phase-ordering kinetics of cemetery organization in ants. *Physical Review E* 57(4): 4568–4571.
Camazine S, Deneubourg J-L, Franks N, Sneyd J, Theraulaz G, and Bonabeau E (2001) *Self-organization in biological systems*. Princeton, NJ: Princeton University Press.
Dunne JA, Williams RJ, and Martinez ND (2002) Food-web structure and network theory: The role of connectance and size. *PNAS* 99: 12917–12922.

Grassé PP (1959) La reconstruction du nid et les coordinations interindividuelles chez Bellicositermes natalensis et Cubitermes sp. La theorie de la stigmergie: Essai d'interpretation des termites constructeurs. *Insectes Sociaux* 6: 41–83.

Haken H (2006) *Information and self-organization: A macroscopic approach to complex systems*. Berlin, Heidelberg: Springer.

Holland, J. H. (2006). Studying complex adaptive systems. Journal of Systems Science and Complexity, 19, 1–8.

Hofstadter DR (1989) *Gödel, Escher, Bach: An eternal golden braid*. New York: Vintage Books.

Kauffman SA (2000) *Investigations*. Oxford: Oxford University Press.

Lizier JT, Prokopenko M, and Zomaya AY (2014) A framework for the local information dynamics of distributed computation in complex systems. In: Prokopenko M (ed.) *Guided self-organisation: Inception*, pp. 115–158. London: Springer.

Newman MEJ (2003) The structure and function of complex networks. *SIAM Review* 45: 167–256.

Parrish JK and Edelstein-Keshet L (1999) Complexity, pattern, and evolutionary trade-offs in animal aggregation. *Science* 284(5411): 99–101.

Potts WK (1984) The chorus-line hypothesis of manoeuvre coordination in avian flocks. *Nature* 309: 344–345.

Prokopenko M, Boschetti F, and Ryan AJ (2009) An information-theoretic primer on complexity, self-organisation and emergence. *Complexity* 15(1): 11–28.

Scheffer M, Bascompte J, Brock WA, Brovkin V, Carpenter SR, Dakos V, Held H, van Nes EH, Rietkerk M, and Sugihara G (2009) Early-warning signals for critical transitions. *Nature* 461: 53–59.

Cybernetics

AM Makarieva, Petersburg Nuclear Physics Institute, St. Petersburg, Russia

Introduction

The word 'cybernetics' originates from the Greek word for 'steersman'. Cybernetics can be broadly defined as a field of knowledge seeking to offer a general mathematical approach to the study of complex systems irrespective of their nature (e.g., artificial or living). In this context, under complex systems, one understands the systems composed of elements exchanging energy, matter, and information with one another and with the environment, that is, systems where various control (feedback) processes operate (Fig. 1).

To give a simple example, stability of a system quantity x over time t can be described as $dx/dt = -kx$, where parameter k can, in general, depend on both x and t. If the value of k is positive, any disturbance $\Delta x = x - x_0$ of the initial value x_0 of this parameter will exponentially diminish with time, returning the system to the initial state, $\Delta x = \exp(-kt)$. At negative k, any slight perturbation will exponentially amplify, destabilizing the system. Such formal description does not depend on the nature of either parameter x or the positive or negative feedback processes in the system that are responsible, respectively, for the system's instability or stability (homeostasis).

Established in the 1940s and the 1950s with works of Norbert Wiener, William Ashby, and others, the notion of cybernetics became well-known in the 1970s and the 1980s, the time marked with intensive research and large expectations associated with the idea of creating artificial intelligence. Cybernetics studies in life sciences can be exemplified by the models of organic evolution like the Eigen's hypercycle; there is much modeling research in the fields of neurophysiology and sociobiology.

While generality of a scientific approach is obviously an important scientific merit, the overall success of the approach is determined by its concrete applications within particular fields. Ecology, the science of interactions among organisms and between the organisms and their environment features high complexity and has the problem of homeostasis versus change at its very heart. To apply the cybernetics approach in ecology, it is important to establish the correspondence between the major notions of cybernetics like information exchange, communication, control processes, etc., and measurable characteristics of the organism, ecosystem, and biosphere. This article describes the essential aspects of ecological cybernetics. Among the subsequent sections, the first section discusses the origin of information fluxes in the biosphere. In the second section, information stores and exchange fluxes within the biosphere, on the one hand, and within the modern civilization, on the other hand, are described and compared. The third section discusses control processes operating at the biota–environment interface, which allow life to persist on Earth for practically infinite time periods of the order of billion years, thus maintaining homeostasis of the living matter.

Solar Energy and Information

According to the second law of thermodynamics, closed systems ultimately reach the state of maximum entropy. The apparent high degree of orderliness of ecological systems and the persistence of this orderliness through time indicates that there is a continuous external input of order (information) into ecological systems. The source of this information is the solar energy, the primary source of energy for life on Earth. Both solar radiation and thermal radiation of Earth consist of particles – photons. Mean energy of one photon is proportional to absolute temperature measured in degrees kelvin. Absolute temperature of Sun is about $T_S = 6000$ K. Absolute global mean surface temperature of Earth is about $T_E = 288$ K (i.e., about 15 °C). Mean energy of one solar photon is about $T_S/T_E = 6000/288 \approx 20$ times larger than the mean energy of one thermal photon of Earth. According to the law of energy conservation, the cumulative energy of all solar photons coming to Earth is equal to the cumulative energy of thermal photons emitted by the Earth into space. It means that the number of thermal photons emitted by Earth into space is about 20 times larger than the number of solar photons reaching Earth's surface. Consequently, one solar photon decays on average into 20 thermal photons. Decay of solar photons gives rise to all ordered, information-rich processes on Earth, of which life is most powerful (Table 1).

Information capacity of a system is characterized by the available number N of memory cells. If a cell's memory can be characterized by only two possible values of a certain variable, the total number of possible combinations of these values in all memory cells is 2^N. The system possesses the maximum possible amount of information equal to N bits when the values of the measured variable are defined in all N memory cells. If states of N_1 cells remain unknown, the amount of information reduces to $N - N_1$. If the measured variable remains undefined in all memory cells, the information becomes zero while the entropy of the system reaches its maximum.

Solar photons interact with molecules of vegetation covering the Earth's surface. These molecules can be viewed as elementary memory cells of the ecosystem. Solar photons can excite molecules, that is, impart a certain amount of energy to

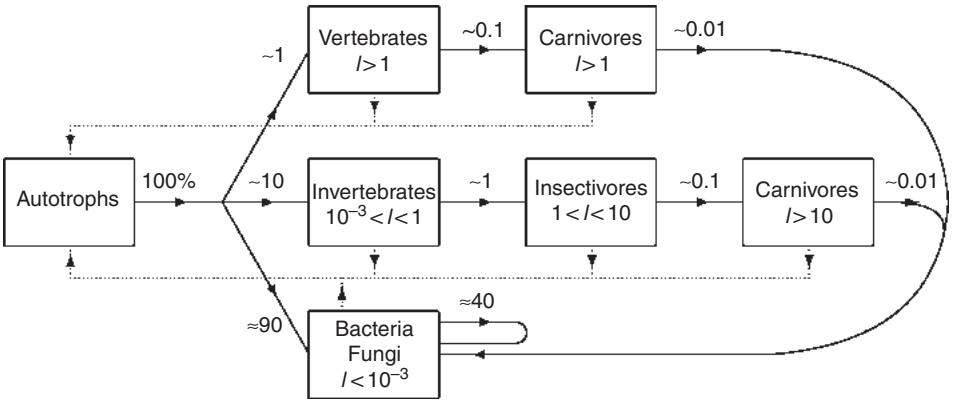


Fig. 1 Natural terrestrial ecosystem represented as a cybernetic system. The division into blocks is made on the basis of the trophic level and body size l (cm) of the organisms. Feedback loops between the system's blocks are exemplified by the fluxes of organic carbon (solid arrows) coming in and out of each block and fluxes of inorganic carbon (dotted arrows) coming out of the heterotrophic blocks into the autotrophic one. For heterotrophic blocks $dM/dt = P_{in} - P_{out} - R$, where M is the store of organic carbon in live biomass within the block, P_{in} is the incoming organic carbon, P_{out} is the organic carbon produced within the block and transferred to the next trophic level, and R is respiration (decomposition of organic carbon). 'Carnivores' indicate vertebrate-feeding heterotrophs, 'insectivores' indicate invertebrate-feeding heterotrophs. Numbers near solid arrows indicate the flux magnitude in terms of the percentage of ecosystem net primary productivity (100%). Numerical data of Gorshkov VG (1995) *Physical and Biological Bases of Life Stability*. Berlin: Springer; Makarieva AM, Gorshkov VG, and Li B-L (2004) Body size, energy consumption and allometric scaling: A new dimension in the diversity–stability debate. *Ecological Complexity* 1: 139–175.

Table 1 Solar power and some routes of its dissipation on Earth

Power source/sink	*Power*	
	10^{12} W	*Relative to the solar power*
Total solar power coming from Sun to Earth	1.7×10^5	1.0
Physical processes		
Wind power	2×10^3	10^{-2}
Oceanic waves	10^3	6×10^{-3}
Natural biota		
Transpiration	3×10^3	2×10^{-2}
Photosynthesis	10^2	6×10^{-4}
Modern civilization		
Energy consumption	10	6×10^{-5}
Consumption of the net primary production of the biosphere	9	6×10^{-5}

molecules and increase their energy above the average thermal level. A good approximation is to assume that molecular memory cells are characterized by two states – excited (a definite state) and nonexcited (indefinite state) compared to the average chaotic thermal level. During the process of decay, solar photons are able to excite molecules until their own energy becomes equal to the average energy of thermal photons of the Earth's surface. Each solar photon possesses an amount of energy equal to that of about 20 thermal photons of Earth. Consequently, one solar photon can excite about 20 molecules, that is,. impart information to about 20 molecular memory cells. Such consideration makes it possible to estimate the amount of information (in bit s^{-1}) coming from Sun to Earth per unit time. It is roughly equal to the number of thermal photons emitted from the Earth to space, because each thermal photon is emitted from an excited molecule, which represents a memory cell containing one bit of information. The number of Earth's thermal photons emitted to space in a unit of time is equal to the power Q ($Q \approx 2 \times 10^{17}$ W) of solar radiation reaching the Earth divided by the energy ε of one thermal photon, which is determined by the Earth's temperature T_E, $\varepsilon = k_B T_E$, where $T_E \approx 288$ K, k_B is Boltzmann constant, which is proportional to the reverse Avogadro number ($k_B = 1.4 \times 10^{-23}$ J K^{-1} molecule^{-1}). As far as one molecule represents a memory cell with two possible states, dimension molecule^{-1} in Boltzmann constant corresponds to bit^{-1}. The information flux F coming from Sun to Earth is $F = Q/(k_B T_E) \approx 10^{38}$ bit s^{-1}.

If one solar photon possesses energy equal to that of $T_S/T_E \approx 20$ thermal photons, the maximum number of molecules it can excite is $T_S/T_E - 1 \approx 19$, because after 19 acts of excitation its energy becomes equal to that of one thermal photon. After that it cannot impart any additional energy to molecules, and, therefore, cannot excite them. So, only $(T_S/T_E - 1)/(T_S/T_E) \times 100\% = (T_S - T_E)/T_S \times 100\% \approx 95\%$ of Earth's thermal photons come from excited molecules and characterize information

flux coming from Sun to Earth. The ratio $(T_S - T_E)/T_S$ describes the well-known Carnot efficiency of the solar radiation on Earth. If the Sun's temperature were equal to that of Earth, solar photons would have the same energy as thermal photons of Earth and could not excite any molecules on the Earth's surface. In such a case the information flux from Sun to Earth would be equal to zero.

Stores and Fluxes of Information in the Natural Biota and Civilization

Stores of Information

The maximum rate of information processing by the human brain in about 10 bit s^{-1}. Information is acquired most actively during the first 20 years of life of the individual, that is, during about 6×10^8 s. The amount of information acquired later in life does not change the order of magnitude of the total store. The amount of information stored in memory of an adult human can be estimated at about 6×10^9 bit.

An upper estimate of the total amount of cultural information of the modern civilization can be obtained multiplying the current population number of Earth ($\sim 6 \times 10^9$ people) by the average individual memory store of information ($\sim 6 \times 10^9$ bit), which gives a value of about 10^{19} bit. This is a gross overestimate of the real value, because most part of memory information is the same in all contemporary people. The unique nonoverlapping information of the civilization is stored in memories of specialists (professionals) – scientists, craftsmen, writers, musicians, artists. Working specialists constitute not more than about 10% of the whole population (multiplier 10^{-1}). Each field of knowledge can normally exist with no less than 100 specialists working in this field and sharing the same memory information (multiplier 10^{-2}). The real value of information store of the modern civilization can be obtained multiplying the upper estimate by 10^{-3}, which gives about 10^{16} bit.

Genetic information of most species of the biosphere is written in polymer double-strand molecules of DNA, which represent various sequences of the existing four different monomer units – nucleotide base pairs (bp). Human genome contains $G = 3 \times 10^9$ bp, that is, 3×10^9 memory cells each of which can be characterized by one of the four different values. The store of genetic information in the human genome is equal to $\log_2 4^G = 2G = 6 \times 10^9$ bit. The stores of genetic and nongenetic (memory) information in humans are of the same order of magnitude.

To quantify the store of the genetic information of the natural biota as a whole it is necessary to multiply the information content of an average genome by the total number of species in the biosphere, which is equal to about 10^7 species. The average genome size can be taken equal to 10^9 bp, which is the average genome size of insects that constitute the majority of species in the biosphere. The total amount of genetic information stored in the natural biota is of the order of 10^{16} bit and coincides in the order of magnitude with the information store of the modern civilization (Fig. 2a).

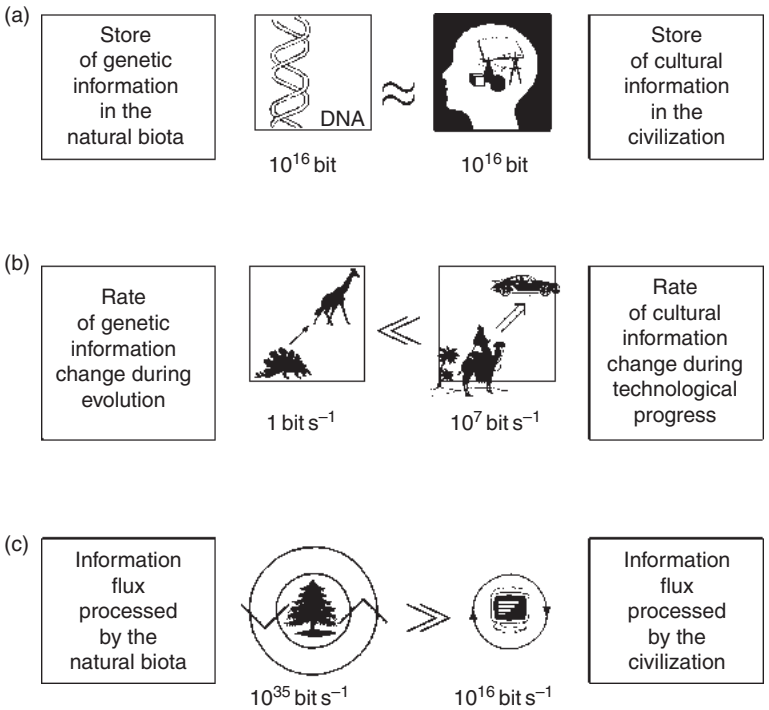


Fig. 2 Stores of information (a), rates of their change (b), and information exchange fluxes (c) in the natural biota and modern civilization. Reproduced from Gorshkov VG, Gorshkov VV, and Makarieva AM (2000) *Biotic Regulation of the Environment: Key Issue of Global Change*. London: Springer-Praxis, with kind permission from Springer Science and Business Media.

The amount of information that can be stored in a modern PC is of the order of 10^{11} bit, that is, it greatly exceeds the amounts of genetic and individual memory information of one person. The cumulative memory capacity of modern computer technologies is large enough to store both the cultural information of the modern civilization and the genetic information of the natural biota.

Fluxes of Information Exchange

Species composition of the natural biota changes with evolutionary transitions of old species to new related ones. The genetic information of the biosphere changes in the course of evolution. Closely related species differ from each other in about 1% of their genetic information. The average species life span is about 3 million years. Thus, 1% of the genetic information changes in a single act of speciation approximately every 3×10^6 years. A complete turnover of the genetic information of the natural biota, 10^{16} bit (Fig. 2a), takes about 3×10^8 years, that is, about 10^{16} s. The rate of evolution (i.e., the rate of change of genetic information of the natural biota in the course of evolution) is approximately equal to 10^{16} bit/10^{16} s = 1 bit s^{-1} (Fig. 2b). This extremely low rate of information change has been sufficient to ensure sustainable development of the biosphere, that is, to support evolution of the natural biota in such a manner that the latter has been able to compensate directional adverse environmental changes of cosmic and geophysical nature during the whole period of life existence, that is, during nearly 4 billion years.

The rate of information change during technological progress of the civilization is determined by the ability of people to generate and assimilate new cultural information. The present-day population of Earth can assimilate no more than $6 \times 10^9 \times 10$ bit $s^{-1} = 6 \times 10^{10}$ bit s^{-1}. The present-day rate of technological progress depends on the average time of renewal of modern technological systems, which is of the order of 10 years, that is, about 3×10^8 s. Given that the store of information of the modern civilization is of the order of 10^{16} bit, the modern rate of civilization progress is about 10^{16} bit/3×10^8 s $\approx 3 \times 10^7$ bit s^{-1}. (This estimate is obtained under the reasonable assumption that the most part of cultural information of the modern civilization is represented by information stored in memory of specialists dealing with modern technologies.) The ratio of the amount of newly generated information to the amount of assimilated information for modern people does not exceed $3 \times 10^7/6 \times 10^{10} \approx 10^{-3}$.

The information rate of the progress of the modern civilization, 3×10^7 bit s^{-1}, exceeds the information rate of evolution, 1 bit s^{-1}, by more than 7 orders of magnitude (Fig. 2b). This provides an explanation for the unprecedented (as compared to all other extant and extinct species) potential of *Homo sapiens* to destroy the natural environment.

An estimate of the total information flux going through all modern computers can be obtained multiplying the average information flux in one PC, $\sim 10^8$ bit s^{-1}, by the total number of people owning computers. Assuming that at present there is one PC for every hundred of people we obtain that the total flux of information in computers of the modern civilization is of the order of 10^{16} bit s^{-1}. This figure will hardly ever increase by more than 6 orders of magnitude (by providing computers for all people on the planet and ensuring a 4-orders-of-magnitude increase of information flux in each PC), that is, up to 10^{22} bit s^{-1}.

Even the present-day global computer information flux, 10^{16} bit s^{-1}, exceeds the assimilation capacity of the brain of modern people, 10^{10} bit s^{-1}, by a factor of million. Computers work on the basis of programs designed by people and speed up significantly the processing of information. But this only makes sense while people are still able to check and control the outgoing flux of information. All the information that is generated by computers and other mass-media devices (TV, cinema, video, theatre, music, etc.) above that threshold represents informational pollution of the environment. Informational pollution affects all the five organs of sense of people, and, among various types of pollution, presents the most dangerous threat to the mental health of humans.

In humans, metabolic power of existence in adults is equal to $q = 140$ W. The body temperature of humans is approximately equal to $T_b \approx 37\ ^\circ C = 310$ K. The average thermal energy of molecules in the human body is equal to $k_B T_b$, where k_B is the Boltzmann constant ($k_B = 1.4 \times 10^{23}$ J K^{-1}). Thus, $k_B T_b$ gives the order of the average amount of energy necessary to excite a molecule, that is, the additional energy committed to a molecule as compared to the average thermal level. Assuming that one molecule corresponds to one memory cell with two possible states (excited and nonexcited), that is, 1 molecule $\approx$ 1 bit, we obtain that the information flux going through all living cells of the human body is equal to $q/k_B T_b \approx 3 \times 10^{22}$ bit s^{-1}. This value exceeds the asymptotic information power of possible future computers and by more than 12 orders of magnitude exceeds the assimilation capacity of the modern humanity.

Human body contains about 10^{14} living cells. Thus, every living cell processes on average about $3 \times 10^{22}/10^{14} \approx 10^8$ bit s^{-1}, which is equal to information flux realized in a modern PC. The biosphere contains about 10^{28} living cells. Thus, the natural biota of Earth as a whole processes about $10^8 \times 10^{27} = 10^{35}$ bit s^{-1}, which is about 20 orders of magnitude more than the information flux of all computers of the modern civilization. Unlike in computers, molecular memory cells of living cells are directly coupled with the environment. Thus, the whole flux of information processed by the biota is used for correct interaction with the environment in control processes aimed at environmental and ecological homeostasis.

Energy consumption of the modern humankind is equal to 10^{13} W, which is only an order of magnitude less than the photosynthetic power of the natural biota (Table 1). But, due to the huge difference in the rates of information processing between the natural biota and civilization, any kind of anthropogenic energy use is inevitably characterized by a remarkably low efficiency, that is, low information content of most processes generated by the humankind with help of energy use.

The humankind spends a large portion of its energy on transport, that is, moving macroscopic objects – cars, trains, people, etc. Motion of a macroscopic object is totally determined by only four variables – its mass and three coordinates of the velocity vector. Motion of macroscopic objects can be fully described by a very small amount of information coded in corresponding macroscopic

memory cells. It is in principle possible to efficiently convert the kinetic energy of moving macroscopic objects to gravitational or electric energy that could be further used for generation of complex correlated molecular processes similar to those taking place in a living cell. But the kinetic energy of transported objects does not generate any ordered, information-rich processes. It is spent on friction, and finally dissipates converting to heat.

Macroscopic motion can be found in natural ecosystems as well (e.g., locomotive animals). However, in stable ecosystems, the amount of energy allocated by the natural biota to this less-efficient channel of energy use does not exceed 1% of the total biotic energy consumption. Meanwhile humans spend on transport more than one-third of the consumed energy. The remaining part of anthropogenically utilized energy is spent even more wastefully with respect to the information content of the generated processes (e.g., heating of buildings).

The total amount of energy consumed by the humankind (**Table 1**) does not characterize the amount of work that can be done by humans in order to stabilize the environment. Of critical importance is the total flux of information that can be processed by the modern civilization. And, as far as information fluxes are concerned, the modern civilization is inferior to the natural biota (**Fig. 2c**), which uses this flux to maintain ecological and environmental homeostasis.

Control Processes in Ecological Systems

Life is based on biochemical reactions that convert inorganic substances stored in the environment into organic ones and back. The existing power of the biochemical fluxes of synthesis and decomposition of organic substances is such that, were not these feedback processes rigidly correlated, the environment could change completely in time periods of several tens of years, reaching a state where life would be impossible.

For example, the global amount of atmospheric CO_2 is of the order of $M^- \sim 10^3$ Gt C (1 Gt$=10^9$ t). The mean global rates of biochemical synthesis P^+ and decomposition P^- are of the order of $P^+ \sim P^- \sim 10^2$Gt C yr^{-1}. If the rates of synthesis and decomposition were not correlated, that is, if they coincided by the order of magnitude only, their relative difference would be of the order of unity, $|P^+ - P^-|/P^\pm \sim 1$. In such a case, if synthesis exceeded decomposition, $P^+ > P^-$, the global biota would use up the entire store of atmospheric carbon on a timescale of $M^-/P^- \sim 10$ years. This would render further photosynthesis and existence of life impossible. The amount of organic carbon in the biosphere (living biomass, humus, and oceanic dissolved organic carbon) is of the same order of magnitude, $M^+ \sim 10^3$Gt C. If the rate of decomposition exceeded the rate of synthesis, the global biota would be able to destroy itself completely in equally short periods of time.

The fluxes of synthesis and decomposition cannot be correlated with each other directly. Synthesis and decomposition of organic matter represent independent biochemical processes that are generally performed by different species under different environmental conditions (temperature, humidity, etc.). While primary productivity is limited by the incoming solar radiation, there are no physical limitations on the rate of decomposition, since the latter is ultimately dictated by the population numbers of heterotrophic organisms. Characteristic ecosystem values of P^+ and P^- are determined by the individual design of every species, population abundance, and overall numbers of autotrophic and heterotrophic species inhabiting Earth. The values of P^+ and P^- cannot coincide with an infinite precision *a priori*.

For example, even if the mean global rates of synthesis and decomposition coincided, say, with a high accuracy of 1%, $\alpha \equiv |P^+ - P^-|/P^+ \sim 0.01$, such a biota would completely destroy its environment (or self-destroy) in $M^\pm/|P^+ - P^-| = M^\pm/(\alpha P^+) \sim 10^3$ years, that is, nearly instantaneously on a geological scale. The life span of the biota is short for any realistic accuracy of the coincidence of P^+ and P^-. To extend the biotic life span to the documented several billion years of life existence, $T \sim 10^9$ years, one has to demand that the living organisms and their ecological communities are designed such that the mean rates of synthesis and decomposition performed by them coincide to the accuracy of $M^\pm/(p^+T) \sim 10^{-8}$, which is improbable.

Correlation of the ecological fluxes of synthesis and decomposition of the organic matter is achieved indirectly, via continuous sensing of information about the current state of the environment that is performed by living organisms. The biota reacts to any environmental change as soon as its relative magnitude reaches some critical value, biotic sensitivity ε_b. As long as the magnitude of the environmental change remains lower than biotic sensitivity, synthesis and decomposition of organic matter by the biota may proceed in a noncorrelated manner at different rates. As soon as some environmental parameter changes by ε_b, the biota initiates compensating negative feedback processes and keeps them going until the disturbance is diminished to a level below ε_b, when the biota no longer notices it. The optimal state to which the ecosystem ultimately returns (the state of ecological homeostasis) is thus defined to an accuracy of ε_b. For example, if the amount of inorganic carbon in the atmosphere changes by $\varepsilon_b \sim 1\%$ (e.g., increases), the biota can enhance the rate of biochemical synthesis (that takes away CO_2 from the atmosphere) or reduce the rate of biochemical decomposition (that would further add to the atmospheric CO_2 amount) until the perturbed concentration relaxes to its optimal value. The same principle can be used to control temperature, humidity, and all other environmental parameters.

The huge information fluxes processed by the natural biota (**Fig. 2c**) are necessary for sensing the environment, reading the data about its state, and ensuring regulatory ecological processes aimed at compensation of possible environmental disturbances. This biotic regulation of the environment is equivalent to an operating system where the characteristic rate of information processing exceeds the maximum possible rate of automatic control provided by all computers of the modern civilization by 20 orders of magnitude. Biotic regulation is based on genetic programs of biological species of the biosphere. It can be viewed as an automatically controlled operating system where the program of automatic control has been tested for reliability in an experiment lasting for several billion of years (during the whole period of life existence).

The relative degree of unsteadiness in the work of a computer is defined as the ratio of the rate of human-induced changes in the computer program to the total flux of information processed by the computer. The relative unsteadiness of the regulatory program of the natural biota is fantastically low, 1 bit $s^{-1}/10^{35}$ bit $s^{-1} = 10^{-35}$. (Rate of program change corresponds to the rate of information change in the course of evolution, 1 bit s^{-1}. The total information flux processed by the natural biota is equal to 10^{35} bit s^{-1}). It means that each working regulatory program is maintained by the natural biota in a steady state for the maximum possible periods of time.

Genetic information of the natural biota changes completely every 3×10^8 years. Thus, during the whole period of life existence (3.8×10^9 years) there were no more than 12 completely different programs of biotic regulation of the environment. A working program of biotic regulation is presumably unique for each particular epoch. Evolution of the biotic regulatory program is possible due to acting geophysical and cosmic processes; that is, directional changes in parameters that cannot in principle be controlled by biota (e.g., solar activity) may lead to a situation when the old regulatory program is no longer the most effective one. As a result, there opens a possibility for a new more effective regulatory program of the biota to establish in the result of genetic modifications (i.e., appearance of new species) in the old program. New regulatory programs appearing in the course of evolution are exposed to a thorough experimental testing.

The humankind is unable to create a technological system equivalent to the natural biota, where each micron of the Earth's surface is controlled by dozens of independently functioning unicellular and multicellular organisms, each living cell processing an information flux similar to that of a modern PC. The genetic program of the natural biota cannot be substituted by any technological program of automatic control (even if this technological program is characterized by fluxes of energy and information similar to those in the natural biota), because search for appropriate technological decisions and their testing is performed by human beings and can take billions of years. Technological solutions of ecological problems can be only successful on a local scale. Globally, the only promising strategy for the modern humankind is therefore strategy of preservation of the remaining natural biota and gradual restoration of its global regulatory potential.

Further Reading

Ashby, W.R., 1956. Introduction to Cybernetics. London: Methuen.
Aulin, A.Y., 1982. Cybernetic Laws of Social Progress. Oxford: Pergamon.
Brillouin, L., 1956. Science and Information Theory. New York: Academic Press.
Eigen, M., Schuster, P., 1979. The Hypercycle. Heidelberg: Springer.
Glushkov, V.M., 1966. Introduction to Cybernetics. New York: Academic Press.
Gorshkov, V.G., 1995. Physical and Biological Bases of Life Stability. Berlin: Springer.
Gorshkov, V.G., Gorshkov, V.V., Makarieva, A.M., 2000. Biotic Regulation of the Environment: Key Issue of Global Change. London: Springer-Praxis.
Kauffman, S.A., 1991. Antichaos and adaptation. Scientific American 265, 78–84.
Makarieva, A.M., Gorshkov, V.G., Li, B.-L., 2004. Body size, energy consumption and allometric scaling: A new dimension in the diversity–stability debate. Ecological Complexity 1, 139–175.
Nicolis, J.S., 1986. Dynamics of Hierarchical Systems: An Evolutionary Approach. Berlin: Springer.
Ninio, J., 1998. Acquisition of shape information in working memory, as a function of viewing time and number of consecutive images: Evidence for a succession of discrete storage classes. Brain Research Cognitive Brain Research 7, 57–69.
Patten, B.C., Odum, E.P., 1981. The cybernetic nature of ecosystems. American Naturalist 118, 886–895.
Svirezhev, Yu M., Logofet, D.O., 1983. Stability of Biological Communities. Moscow: Mir.
Thribus, M., McIrvine, E.C., 1971. Energy and information. Scientific American 224, 179–188.
Wiener, N., 1948. Cybernetics or Control and Communication in the Animal and the Machine. New York: Wiley.

Ecological Data Archiving and Sharing

William K Michener, University of New Mexico, Albuquerque, NM, United States

Glossary

Archive A collection of data and information that is permanently stored in a data repository; also can be used as a verb to represent the process of preparing and storing data and information for the long-term in a repository.

Data curation The management, organization, quality checking and preservation of data.

Data life cycle The sequence of stages that data undergo in a research project from planning, through acquisition, quality assurance and quality control, documentation, analysis and visualization, and preservation and dissemination.

Data repository A location for long-term storage of data and metadata.

Data sharing The process of making data available for use by others.

Digital object identifier (DOI) A persistent identifier used to uniquely identify objects such as data sets or data packages (e.g., data and metadata).

Meta-analysis Combining the results of many research studies using statistical approaches; often used in synthesis efforts to reach broader generalizations.

Metadata Information about data that enables discovery, interpretation and use of the data (e.g., Who collected the data and how can one access and use the data?, Why were the data collected?, How, when and where were the data collected?).

Open science An emerging movement designed to make science (and data) more broadly accessible, transparent and reproducible.

Reproducibility The capability to independently reproduce one's own work or that of another researcher.

Workflow The sequence of tasks or operations followed in an analytical program including reading, processing, and visualizing data.

Introduction

Data sharing refers to the act of making data available for use by others. Data may be shared in many different ways. One mechanism is to physically copy and transfer data from the creator to others upon request. This transfer may be done via various media including paper copies, floppy disks, tapes, and hard drive(s)—an approach that remains necessary for transferring very large volumes of data. Another mechanism that is frequently used today is to email data files or upload data to a data or file-sharing service and provide a link so that others may download the data. Both mechanisms require the active and continual involvement of the data creator.

Increasingly, data creators voluntarily or are required by funders and journals to upload and share their data via a data repository. In this case, the data contributor initially prepares and submits the data and accompanying documentation (or metadata) to a data repository that may reformat, organize, preserve and advertise the existence of the data. Once the data are submitted by the contributor and accepted by the data repository, then the data repository, alleviating the need for further engagement by the contributing researcher, supports subsequent requests for data.

Data repositories are also often referred to as data centers or data archives and they may be associated with funding agencies, journals, professional societies, universities and other research, governmental and non-governmental organizations. Their role in preserving and sharing data is discussed further below.

Data Sharing: Perceived Impediments and Benefits

Researchers perceive that there are both impediments to and benefits derived from sharing data (Michener, 2015a; Tenopir *et al.*, 2011, 2015). Commonly expressed barriers to data sharing by data creators include concerns that potential new discoveries based on the data will be scooped by others and the belief that some individuals may misinterpret and misuse the data. Although such concerns are normally unfounded, there are real costs associated with organizing and preserving high quality data and metadata as well as making the data discoverable, interpretable and usable by others. Many researchers find data sharing to be challenging because of their lack of training in data management, absence of institutional support for data management, and paucity of user-friendly data management tools.

As data sharing has become more commonplace, attitudes and perceptions are changing and the majority of researchers—along with research sponsors and professional societies—recognize that there are many benefits that accrue from data that are preserved and shared. For example, the nature and pace of science expand and accelerate as researchers and others are able to discover and re-use relevant data, integrate data from different sources, and ask new questions. This benefit is particularly important as many of the

"grand challenges" addressed by ecologists require multi-, inter-, and trans-disciplinary perspectives as well as pertinent data that can support research efforts that bridge domain and disciplinary boundaries. Likewise, broad-scale studies, meta-analysis and major synthesis efforts normally depend upon the accessibility of large amounts of data from many different sources.

In addition to directly supporting science, data sharing can also benefit data creators and the public. Data creators (i.e., researchers) receive credit when their data are used and appropriately cited in new research studies. Further, studies have demonstrated that a researcher's publication citation rates increase when the underlying data are also made available. Public trust and the trust of research sponsors increase when findings can be reproduced and verified by others who access the data and run associated code and workflows.

Evolution of Data Sharing and Archiving Norms

Large national and multi-national efforts have played a major role in the evolution of data sharing and archiving practices (see Michener, 2015a). The International Biological Program (IBP) operated from 1964 through 1974 and was one of the first multi-national efforts to focus on ecosystem experiments, modeling and synthesis. IBP was viewed as a scientific success based on the significant publications and new theories that resulted. Although IBP drew attention to the need for cross-site data in synthesis and modeling studies, very little data from the program can be discovered today as there were no agreed upon standards for sharing, documenting and archiving data.

Attitudes towards data management and sharing changed with the emergence of the US Long-Term Ecological Research (LTER) Program in 1980 and the International LTER Program in 1993. Both Programs promoted data sharing and good data management practices and policies at the site and network levels. Consequently, the data underlying thousands of publications are preserved, well documented and available for further use.

More recently, ecological synthesis centers, ecological and ocean observatories, data aggregators such as the Global Biodiversity Information Facility, and community-based research networks and data federations have promulgated standardized data sharing and archiving policies and practices for the ecological and biodiversity sciences. Many research sponsors now require that researchers include a data management plan that describes how data will be managed, archived and shared as part of their research proposal. Furthermore, many leading professional societies and scientific journals associated with the ecological and evolutionary sciences established the Joint Data Archiving Policy (JDAP) in 2010 to ensure that data underpinning scientific publications are archived and shared (Box 1) (also see Whitlock, 2011). Most present-day ecologists and environmental scientists have an expectation that data will be preserved and shared including the data they have created (Tenopir *et al.*, 2015). Well-managed and shared data and code provide the foundation for open science where science is made more transparent and findings are reproducible (Hampton *et al.*, 2015).

Effective Practices for Sharing Data

Five practices can facilitate effective data sharing (Michener, 2015a). First, creating and following a data management plan from the inception of the research is recommended and, in many cases, required by research sponsors. A comprehensive data management plan covers all aspects of the data life cycle (**Fig. 1**) and fully describes how data will be treated and managed during the research as well as after the project has been completed (Michener, 2015b).

Second, researchers can promote the appropriate sharing, use, and attribution of their data by establishing or adopting and following reasonable data sharing and attribution policies. The Creative Commons Organization (CC) offers numerous free licenses to researchers that provide standardized ways for others to share and use their creative work based on conditions chosen by the researcher(s). CC conditions establish whether the data must be attributed to the originator, conditions under which the data can be shared with others, and whether the data can be modified or used for commercial purposes. Data archives also frequently have established guidelines for how data should be shared and cited. The Dryad Digital Repository, for example, provides recommendations for how to cite data products that are acquired from the repository (Box 2).

Third, data should be sufficiently described or documented so that others may discover, interpret and appropriately use the data. Such documentation is referred to as metadata and allows users to understand all facets of the data including who collected the data, as

Box 1 Joint Data Archiving Policy (created by a group of professional societies and publishers in 2010 and available through the Dryad Digital Repository)

"[Journal] requires, as a condition for publication, that data supporting the results in the paper should be archived in an appropriate public archive, such as [list of approved archives here]. Data are important products of the scientific enterprise, and they should be preserved and usable for decades in the future. Authors may elect to have the data publicly available at the time of publication, or, if the technology of the archive allows, may opt to embargo access to the data for a period up to a year after publication. Exceptions may be granted at the discretion of the editor, especially for sensitive information such as human subject data or the location of endangered species."

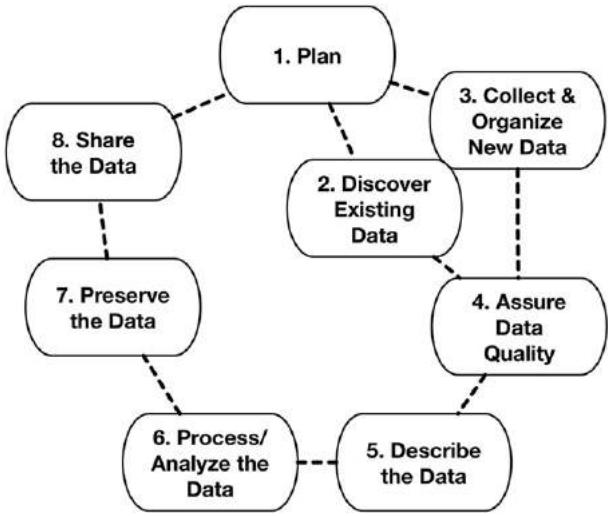


Fig. 1 A good data management plan encompasses the entire data life cycle: (1) starting with planning how the data will be managed, used and shared; (2) describing how existing data will be used and (3) how new data will be collected and organized; (4) summarizing the quality assurance and quality control procedures that will be employed to ensure data quality; (5) listing the metadata standards and supporting software tools that will be used to describe the data; (6) outlining anticipated data processing activities, algorithms and workflows that will be used to process and analyze the data; (7) describing how and where data will be preserved; and (8) providing plans for disseminating and sharing the data. Adapted from Michener, W. K. (2015b). Ten simple rules for creating a good data management plan. *PLoS Computational Biology* **11**, e1004525. https://doi.org/10.1371/journal.pcbi.1004525.

Box 2 Recommended guidelines for citing data that have been acquired from the Dryad Digital Repository

"How do I cite data from Dryad?

When citing data found in Dryad, **please cite both the original article as well as the Dryad data package**. It is recommended that the data package be cited **in the bibliography** of the original publication so that the link between the publication and data is indexed by third party services. Dryad provides a generic citation string that includes authors, year, title, repository name and the Digital Object Identifier (DOI) of the data package, for example

Westbrook JW, Kitajima K, Burleigh JG, Kress WJ, Erickson DL, Wright SJ (2011) Data from: What makes a leaf tough? Patterns of correlated evolution between leaf toughness traits and demographic rates among 197 shade-tolerant woody species in a neotropical forest. Dryad Digital Repository. https://doi.org/10.5061/dryad.8525

Dryad also assigns a DOI to each data file, which should only be used in contexts where the citation to the data package as a whole is already understood or would not be necessary (such as when referring to the specific file used as part of the methods section of an article).

If you are using a large number of data sources, it may be necessary to provide a list of the relevant data packages/files rather than citing each individually in the References. The list can then be submitted to Dryad so others who read your publication can locate all of the original data."

well as where, when, how and why it was collected (Michener *et al.*, 1997). Numerous metadata standards (Ecological Metadata Language (EML) for ecological and environmental data; Darwin Core for biodiversity data; Dublin Core for high-level descriptions of data and other resources; and ISO 19115 and others for geospatial data) and metadata management systems have been created to simplify the process of creating and managing standardized metadata (see Corti *et al.*, 2014; Goodman *et al.*, 2014; Hampton *et al.*, 2015; DataONE; Global Biodiversity Information Facility; and International Organization for Standardization for more information).

Fourth, effective sharing, interpretation and use of data typically require that the data and algorithms and workflows used to process the data be preserved and accessible to users. Hundreds of data repositories (also referred to as data archives and data centers) have been developed to serve different research communities. Repositories can be characterized by their hosting institution (e.g., university, funding agency, governmental or non-governmental organization, research network, etc.) and disciplinary focus (e.g., Arctic data, long-term ecological research data, or general purpose repository such as the Dryad Digital Repository). The Registry of Research Data Repositories provides a searchable catalog of hundreds of research data repositories worldwide ranging from general use to highly specialized disciplinary archives. Some repositories are free or inexpensive and offer a relatively small number of services, whereas others are fee-based and may offer numerous value-added services such as data curation, data quality

checks, peer-review, and advertising. Many repositories also provide access to metadata authoring tools and support the preservation of code, software tools, and associated resources in addition to the data and metadata.

Fifth, data products often are more broadly disseminated and cited when they are published as a data paper in a high-profile journal like *Scientific Data*. Several such data paper journals exist and, like traditional journal articles, the data papers normally undergo peer-review and are assigned a digital object identifier that allows them to be more easily cited and discovered. Another approach for promoting discovery and dissemination of data is to deposit the data and metadata in a trusted data repository that is associated with a data aggregation service or repository federation like DataONE, GBIF and VertNet. DataONE is one example of a federated repository network where researchers can discover ecological and other types of data associated with any one of a large number of member repositories by searching a single data portal. Likewise, VertNet and GBIF make it possible to discover vertebrate data and other species data that are associated with numerous museum and biological collections that are distributed worldwide.

Effective Practices for Creating Preservation-Ready Data

Data products or data packages (i.e., data and metadata) are ready to be submitted to an archive for long-term preservation once the data are logically and consistently organized, files and variables (i.e., file elements) are stable and appropriately named and defined, data quality is assured, and metadata is clear and complete. The following practices will help ensure that data are interpretable and usable by both the data creators and other researchers long after the data were collected and archived (from Olson and McCord, 2000; Cook *et al.*, 2001, 2017).

Logically and Consistently Organize the Data

It is good practice to organize similar types of measurements and observations in a single data set. For instance, a study of forest primary production may be most logically organized in three or more distinct data sets: (1) *tree data* including elements such as tree identification number, diameter (e.g., diameter at breast height), date and time the tree was measured, and other relevant data that are collected at some routine interval (e.g., monthly, seasonally, annually); (2) *site characteristics data* including location, forest or ecosystem type, geology, soils, and other relevant data that may only be collected once or at very infrequent intervals; and (3) *environmental data* including air and soil temperature, precipitation, soil moisture, and other relevant data that may be collected at relatively high frequency (e.g., minutes to hourly). In contrast, organizing all of the primary production data from this study in a single data set leads to many empty fields in the data as the three different types of data are collected and recorded at widely varying intervals with different instruments and, possibly, by different individuals. Likewise, parsing a data set like the *environmental data* set into daily or weekly data sets can lead to a large number of files—each of which must be individually named, quality assured, documented, managed, and preserved, thereby creating an undue management burden on the data creator and subsequent users.

Use Descriptive and Consistent File Names

Descriptive file names that follow consistent naming conventions enable data creators and users to more easily discover, organize and manage, analyze, and protect and preserve the data. For instance, organizing the *tree data* from the primary production example presented above into a series of files named "treedata_a.txt", "treedata_b.txt", …, "treedata_m.txt" provides little indication of what data are contained in the files other than the possible association with trees. Better file names uniquely and precisely relate to the file contents. For example, "Konza_trees_dbh_2014.csv" might reasonably be expected to include tree measurements (dbh—diameter at breast height) from the Konza site in 2014 that are stored in comma separated variable format. Assuming a consistent naming convention is followed, a file named "Konza_trees_dbh_2015.csv" would be expected to contain similar types of data from the following year (i.e., 2015).

Use Stable and Accessible File Formats

The long-term usability of data depends on storing the data in file formats that can be read long into the future. Companies fail; proprietary formats fall out of favor; and data, consequently, can become difficult or impossible to read and use. Appropriate file types include those that are non-proprietary and are neither encrypted or compressed. Several suitable file formats that are widely used by researchers are listed in Box 3.

Use Descriptive and Consistent Variable Names

Variable names such as "var_1", "var_2" through "var_n" may be understood initially by the data creator but convey no meaning to other potential users and typically become unusable even by the data creator as time passes and memory fails. It is good practice to use concise, but descriptive variable names such as "date_measured", "site", "tree_number", "scientific_name", "UTM_easting", "UTM_northing", "soil_moisture_10cm", and so on. If species names are variables, a common convention is to provide the first

Box 3 Stable file formats recommended by DataONE.

- ASCII formatted files are preferred for many types of data; tabular data should ideally be stored using ASCII comma-separated variables format (.csv).
- Geospatial data in raster format can be stored in a variety of formats including: GeoTIFF/TIFF; ASCII Grid; binary image files; NetCDF; and HDF or HDF-EOS.
- Image (vector) data can be stored as shapefiles, ENVI vector files (.evf), or ESRI Arc/Info export files (.e00).

Table 1 A data set or file in which each row represents a single observation

Location_name	*Date*	*Measurement*	*Value*	*Unit*
PawleysStation1	20170113	Max_air_temp	24	Deg_C
PawleysStation1	20170114	Max_air_temp	26	Deg_C
PawleysStation1	20170113	Precip_total	7	mm
PawleysStation1	20170114	Precip_total	0	mm

Table 2 A data set or file in which each row contains all measurements that were recorded at the same time

Location_name	*Date*	*Max_air_temp*	*Precip_total*
Units	YYYYMMDD	Deg_C	mm
PawleysStation1	20170113	24	7
PawleysStation1	20170114	26	0

letter of the genus, followed by an underscore and the complete species name (e.g., "C_florida" for *Cornus florida*); however, caution should be exercised since some abbreviated variable names could apply to two or more species.

Clearly Define Variables in the File and Their Associated Units

In addition to using descriptive variable names, it is good practice to also associate variables with their units directly in the file. This can easily be done in one of two ways. First, a data set or file can be constructed so that each row represents a single, well-defined observation such as in Table 1. In this case, each sensor measurement is clearly labeled and the values are associated with a specific measurement unit.

Another way to structure a data set is to include all measurements that are taken at the same time in the same row as in Table 2. This table contains the same information as Table 1, but is smaller in size since the information about measurement units is only listed once in the second row of header information. Such organization requires that the units of measurement be consistent throughout the file; thus, a new data file would need to be created if the measurement methods changed and observations were recorded in different units.

Assure the Quality of the Data

Errors of commission and omission are common in environmental and ecological data. For example, erroneous data can be introduced as a result of human transcription errors, sensor and instrument malfunction, power fluctuations, and environmental contamination (e.g., moisture, dust and temperature extremes that exceed operational guidelines). Such errors of commission can often be detected using graphical and visualization approaches (e.g., X-Y scatterplots, time series analysis, plotting sampling locations, and quality control charts), range checks (e.g., comparing new measurements to the range of values previously recorded and flagging those that exceed specific thresholds), and various parametric and non-parametric statistical methods. Of course, many errors of commission are best prevented before they occur through personnel training, establishing and following routine maintenance schedules, and following documented and standardized field and laboratory methods. Errors of omission may be difficult or impossible to detect and include failure to document unusual environmental conditions and other factors that may affect data quality. It is good practice to examine data frequently and routinely plot the data to look for both errors and anomalies that may indicate an interesting pattern or process. Data creators and users benefit when attention is paid to data quality throughout the project and when those efforts are well documented in the data files and metadata as appropriate.

Provide Complete and Understandable Metadata

A table of rows and columns of numbers (i.e., data) is meaningless unless there is sufficient metadata present (e.g., column headers and other information about the data) that allows one to understand how the data are organized and what the data represent. Some of this information may be included in a table header that defines the table's elements, units of measurement, format, and so on (see **Tables 1** and **2**). Much more information, however, may be required if the data are to be correctly interpreted, used and cited far into the future. Users will likely wish to know why the data were collected, who collected the data, what methods were used to collect and process the data, when and where the data were collected, how the data are organized and formatted, and how the data have been quality-assured. In addition, users may benefit from knowing whether the data have been analyzed and used in subsequent publications as well as how to best cite the data if they use it in a publication.

Protect Data and Metadata Throughout the Research Project

Accidents happen. All to often, valuable data are lost due to carelessness, theft (e.g., laptop computer), computer malfunctions, storage media degradation, and anthropogenic and natural disasters (e.g., fires, tornadoes, hurricanes, frozen and burst pipes). A good practice is to store data and metadata on three or more storage devices (e.g., desktop, ancillary mass storage device, cloud) in at least two geographically separated locations (e.g., office, home, commercial cloud-based mass storage facility).

The Role of Data Archives

The data archiving process generally consists of several steps that are illustrated in **Fig. 2**. First, the data creator or data contributor identifies and engages with a suitable repository (e.g., one identified by searching the Registry of Research Data Repositories; see Pampel *et al.*, 2013). Initial engagement with the repository may range from reading online policies and guidelines to meeting virtually (email, phone or videoconference) or in-person with a repository representative where fees (if any), requirements and best practices can be discussed. Following repository guidelines, the data creator prepares the data, metadata and supplemental files (e.g., software code, appendices, videos) and submits them to the data repository. Second, the repository initiates data curation activities whereby the files are quality-reviewed and errors or issues are noted. This stage is iterative and may require several exchanges between the curator and the data creator until all issues are resolved and the files are deemed ready for archival. Third, the repository archives the data, metadata and associated files, and normally associates a unique identifier (e.g., DOI) to the complete data product. Fourth, the repository links the data product with additional value-added services such as indexing that may enable rapid search and discovery by repository users, as well as analytical and visualization tools that may allow users to see where the data were collected or peruse statistics generated from the data (e.g., means, ranges). Some repositories may offer additional online or in-person support to users who wish to acquire and make use of the data. Fifth, a user discovers, reviews and, if deemed appropriate, may download and use the data.

The archiving process can vary considerably among data repositories. Many institutional repositories (e.g., universities, agencies) allow users to self-submit data and metadata and only require that the user be affiliated with the institution (e.g., faculty member, student, employee, fundee). Institutional repositories may provide online guidance and help desk support, and they may restrict data submissions to particular file types and sizes. Other repositories may provide higher levels of data curation. For example, the Dryad Digital Repository (Dryad) and the Oak Ridge National Laboratory's Distributed Active Archive Center for Biogeochemical Dynamics (ORNL DAAC, Cook *et al.*, 2017) verify that: (1) the correct files have been received by the repository; (2) the documentation adequately describes the files; (3)

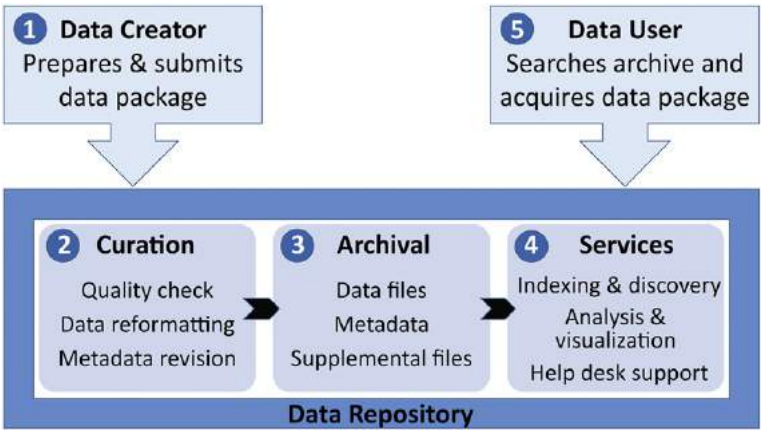


Fig. 2 The data archiving process starts with a data creator (1) who initially prepares and submits data and metadata; continues at the data repository where varying levels of data curation (2) may take place, the data package is archived (3), and ancillary services (4) add value to the data; and ends with data users (5) who may discover and potentially acquire the data for subsequent use.

files, parameters and units are appropriately named and defined; and (4) file content is consistent (e.g., variable formats are consistent throughout the file). To the extent possible, curators associated with Dryad and the ORNL DAAC will attempt to verify that values for parameters within the files are reasonable and may assist users with reformatting and reorganization if necessary.

Large government- or agency-supported repositories (e.g., ORNL DAAC) and publishers of data journals (e.g., *Scientific Data* and the publishers affiliated with Dryad) may also require that the data and metadata be peer-reviewed and undergo additional quality checks before the data products are accepted and disseminated. Dryad, for example, provides password-protected access so that peer-reviewers of journal articles may also simultaneously review the data and metadata; the data and metadata are only formally accepted and published in the archive when the journal article is accepted for publication (note that some journals affiliated with Dryad allow up to a one-year embargo on the data before they become publicly available).

Promoting Data Discovery, Use and Attribution

The data that can be most readily discovered and used are those that are well organized and documented, stored in a trusted and widely used data repository, and associated with a persistent unique identifier such as a DOI that facilitates indexing and discovery. DOIs provide a mechanism for permanently and unambiguously identifying a digital object, making it possible to easily access, cite, and use specific data products. Many data repositories assign DOIs as part of the archiving process. DOIs for data products are acquired from institutional members of organizations like DataCite and Crossref—global registration agencies for research data and related scholarly products.

Data discovery and use can also be promoted by archiving data in repositories that are associated with data aggregators and data repository federations that enable precise search and discovery of data. Data aggregators and federations (e.g., GBIF, DataONE) make it possible for users to discover and access particular data products that may be stored in one of many similar types of repositories without having to search each individual repository.

Appropriate citation of data is facilitated when the data repository and data contributor(s) provide specific recommendations for how the data should be cited; such guidance may be provided on the repository website and in the metadata associated with the data. Data creators can also promote discovery and citation by citing their own data products in the journal articles and other scholarly publications they produce that are based on the data. Of course, it also helps when the data creator can be uniquely identified since most personal names are not unique (e.g., J. Smith, B. Jones, W. Lee). Unique identifiers can easily be acquired from ORCiD—an organization that provides individual researchers with a free nonproprietary alphanumeric code that uniquely identifies them and can be associated with publications, data, proposals, and other scholarly contributions.

Conclusion

New scientific discoveries and knowledge emerge when the existing corpus of data, information and knowledge is archived, discoverable and interpretable. Like journal articles, data can be important products of the scientific enterprise and, thus, deserving of preservation, dissemination and re-use. Data sharing perceptions and practices have evolved over the last several decades. Large national, international and multi- and interdisciplinary research programs have been important in shifting cultural attitudes positively towards data sharing. Presently, many research sponsors, institutions and publishers require that data be well documented, archived in a data repository and available for further use. Five practices facilitate effective data sharing: (1) creating and following a data management plan that covers all aspects of the data life cycle during and after conclusion of the research project; (2) establishing and adopting a reasonable data sharing and attribution license or procedure; (3) comprehensively documenting project data following community standards and best practices; (4) protecting and making available data, metadata, and algorithms and workflows via a trusted community data repository; and (5) disseminating and advertising the existence of the data including, for example, publishing the data in a peer-reviewed data journal. Preparing a preservation-ready data product for submission to a repository requires that the data contributor logically, consistently and clearly name and describe the data package, including the variables, files and associated algorithms and workflows. Furthermore, the data should be quality assured, completely and comprehensively documented, and protected throughout the research. Data repositories often have specific data organization and submission guidelines and play an important role in preserving and disseminating data that serve as a valuable scientific resource for decades to come.

See also: Ecological Data Analysis and Modelling: Ecological Models: Model Development and Analysis; Ecosystems: Freshwater Marshes

Bibliography

Cook, R.B., Olson, R.J., Kanciruk, P., Hook, L.A., 2001. Best practices for preparing ecological data sets to share and archive. Bulletin of the Ecological Society of America 82, 138–141.

Cook, R.B., Wei, Y., Hook, L.A., Vannan, S.K.S., McNelis, J.J., 2017. Preserve: Protecting data for long-term use. In: Recknagel, F., Michener, W.K. (Eds.), Ecological informatics. Heidelberg: Springer.

Corti, L., Van den Eynden, V., Bishop, L., Woolard, M., 2014. Managing and sharing research data: A guide to good practice. London: Sage.
Goodman, A., Pepe, A., Blocker, A.W., *et al*., 2014. Ten simple rules for the care and feeding of scientific data. PLoS Computational Biology 10.e1003542
Hampton, S.E., Anderson, S.S., Bagby, S.C., *et al*., 2015. The Tao of open science for ecology. Ecosphere 6, 1–13.
Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B., Stafford, S.G., 1997. Nongeospatial metadata for ecological sciences. Ecological Applications 7, 330–342.
Michener, W.K., 2015a. Ecological data sharing. Ecological Informatics 29, 33–44.
Michener, W.K., 2015b. Ten simple rules for creating a good data management plan. PLoS Computational Biology 11. e1004525. doi:10.1371/journal.pcbi.1004525.
Olson, R.J., McCord, R.A., 2000. Archiving ecological data and information. In: Michener, W.K., Brunt, J.W. (Eds.), Ecological data: Design, management and processing. Oxford: Blackwell Science, pp. 117–130.
Pampel, H., Vierkant, P., Scholze, F., *et al*., 2013. Making research data repositories visible: The re3data.org registry. PLoS One 8.e78080
Tenopir, C., Allard, S., Douglass, K., *et al*., 2011. Data sharing by scientists: Practices and perceptions. PLoS One 6.e21101
Tenopir, C., Dalton, E.D., Allard, S., *et al*., 2015. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. PLoS One 10.e0134826
Whitlock, M.C., 2011. Data archiving in ecology and evolution: Best practices. Trends in Ecology & Evolution 26, 61–65.

Relevant Websites

www.creativecommons.org—Creative Commons Organization.
www.crossref.org—Crossref.
www.datacite.org—DataCite.
www.dataone.org—DataONE.
www.datadryad.org—Dryad Digital Repository.
www.gbif.org—Global Biodiversity Information Facility (GBIF).
www.iso.org—International Organization for Standardization (ISO).
www.daac.ornl.gov—Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics.
www.orcid.org—Open Researcher and Contributor ID (ORCiD).
www.re3data.org—Registry of Research Data Repositories.
www.nature.com/sdata/—Scientific Data.
www.vertnet.org—VertNet.

Ecological Indicators: Connectance and Connectivity☆

Marcel Holyoak, University of California, Davis, CA, United States

Glossary

Gene flow The movement of genes from a population (subpopulation, etc.) in one location to another, involving movement of individuals and their survival, successful reproduction, and survival of the offspring.

Genetic heterozygosity Variation in genetic material (DNA, RNA etc.) of organisms within a population (subpopulation, etc.) that indicates the presence of different genotypes and serves as future material for evolution and can enhance fitness under varied environmental conditions.

Habitat matrix or matrix The area between habitat patches that are suitable for an organism to survive in and complete at least a part of its life cycle.

Landscape The composition of an area of land including all of its elements and their spatially-explicit layout. An analogous term is a seascape for seas.

Metacommunity A population of communities in which individuals potentially move between local communities.

Metapopulation A population of populations. Individuals move between habitat patches, recolonizing them and replacing local populations that have gone extinct, which prevents regional (metapopulation-wide) extinction.

Introduction

Studies have reported that habitat loss and fragmentation are leading contemporary causes of imperilment for plants and animals. Very frequently habitat loss proceeds in ways that habitat fragments are left, reduced size and smaller more isolated areas of habitat that can be occupied by organisms. Humans also regularly erect barriers to movement in forms such as fences, roads, canals, dams, and urban areas. The extent to which the capacity of organisms to move through a landscape will be reduced is likely to depend on how an organism moves and its behavior and ecology as they relate to movement. For instance, how an organism reacts when it reaches an edge between two ecosystem types, the physical capacity of an organism to move, and whether it requires resources during movement. Anthropogenic changes in landscapes also often occur at the same time as increases in human activity or altered physical conditions, which potentially alter organismal movement. This entry reviews types of changes in habitat connectivity and their consequences. This is followed by a description of how we measure habitat connectivity, and a section on how habitat corridors are being used to attempt to reduce the effects of fragmentation.

Table 1 summarizes and defines some of the common terms used for connectivity. Habitat connectivity is defined as the degree to which a landscape facilitates or impedes movement of organisms among habitat or resource patches. The term is used synonymously with landscape connectivity. (This article does not review food web connectance, which refers to the density of trophic links in a food web—see food chains and food webs). Habitat connectivity includes elements of both the physical structure of habitat and the movement ability and behavior of the organisms in question. Structural connectivity denotes the physical structure of habitat, and is evaluated using measures of habitat extent, subdivision, and contagion, or some combination of these. By contrast, functional connectivity refers to the potential for movement of organisms among habitat or resource patches, and depends on the organisms' perception of, and reaction to, their environment, as well as on their ability to move and costs of movement. Therefore, functional connectivity may vary between species or between individuals of the same species. Functional connectivity is either a potential measure or an actual measurement of the number and frequency of individuals moving between locations. For some species, physical carriage (vectoring) of organisms by natural currents (e.g., stream flow or oceanic currents) or winds may influence functional connectivity. Both structural and functional connectivity may vary among natural environments and as a result of human activities that alter the composition and patterning of the environment and/or an organism's movement behavior. Some studies have defined connectivity on a purely structural basis, and others have used the presence of habitat corridors as an indicator of connectivity. A relevant question when reading papers about connectivity should be which kind of definition of connectivity are the authors referring to, and are parts of the movement of organisms among locations left out.

Habitat connectivity can also influence the flux of materials across ecosystem boundaries, which is considered under the topic ecosystem subsidy (Polis *et al.*, 1996). Here, the overlap between subsidy and habitat connectivity is considered. Ecosystem subsidy is defined as the transport of a donor controlled resource (prey, detritus or nutrients) from one habitat to another where it is utilized by a recipient plant or consumer. Transport here involves physical fluxes and flows, including river flows, winds, and

☆*Change History*: March 2018. M Holyoak updated the Introduction to reflect advances since 2007. Added Table 1 as an overview of terms used to describe connectivity. Updated section 2.4 on graph theoretic approaches to include more about network methods and metrics.

This is an update of M. Holyoak, Connectance and Connectivity, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 737–743.

Table 1 An overview of the various definitions of connectivity

Term	*Definition*
Connectivity	A generic overarching term that includes all of the types of connectivity below
Habitat or landscape connectivity	The degree to which a landscape facilitates or impedes movement of organisms among habitat or resource patches
Structural connectivity	The physical structure of habitat or resources across a landscape (seascape, etc.). Hydrological connectivity (below) is a specialized form of structural connectivity
Functional connectivity	Actual or potential movement of organisms between two or more locations
Genetic or population connectivity	Gene flow between populations (subpopulations etc.) resulting from movement and demography
Hydrologic connectivity	The connectivity of streams, rivers and similar water bodies as determined by physical factors like network topology, stream length, depth, width, any flows (etc.)

oceanic currents. A network of channels or rivers joining lakes or ponds fits naturally into ideas of structural connectivity. However, transport via winds and oceanic currents are not usually considered in landscape studies of structural connectivity, which usually involve the quantification of static (or occasionally dynamic) landscape patterns. Ecosystem subsidy is therefore dependent on connectivity, but the literature for patterns with fluxes and flows is largely separate from the landscape ecology literature about habitat connectivity. Studies of ecosystem subsidy typically investigate processes at ecosystem and community levels such as fluxes of nutrients and levels of biomass in different trophic levels. By contrast, landscape ecological studies of habitat connectivity are more frequently motivated by maintaining species diversity and population persistence. Hence connectivity in a broad sense influences processes at levels from individuals up to entire ecosystems, and at the ecosystem level it is termed ecosystem subsidy.

The prime importance of the movement of organisms to biodiversity is revealed by a variety of kinds of ecological and evolutionary studies. Below we briefly review the importance of island biogeography studies of species diversity, metapopulation studies of the population dynamics of one or a few species, and studies of inbreeding in relation to isolation.

MacArthur and Wilson's (1967) equilibrium theory of island biogeography theorizes that the level of species diversity on habitat islands arises as a balance between the processes of colonization and extinction of species. Colonization is related to distance from a large mainland area of habitat such that more isolated habitat islands should contain fewer species than islands that are closer to a large mainland area of habitat. Extinction of species is related to island size, such that smaller islands can hold fewer species than large islands. Hence, smaller more remote islands should contain fewer species than larger and more connected islands. Furthermore, there is a turnover of species and colonization is required to maintain species diversity. There is a broad range of empirical evidence supporting this theory in a diverse array of taxa. These ideas were brought to the attention of conservation biologists by Wilson and Willis, who in 1975 published a figure laying out the idea that islands closer together and those connected by habitat corridors support more species than islands that are less connected. The ideas both from the equilibrium theory of island biogeography and from Wilson and Willis's figure have been broadly applied to habitat islands of many kinds both aquatic and terrestrial, and have been influential in perpetuating the importance of habitat connectivity. The role of movement in colonization was also modified by James Brown and Astrid Kodric-Brown in 1975, who included the idea that movement may be from one island to another, both as a source of colonization and in a "rescue effect" where immigration might prevent populations of individual species from going extinct on individual and maintain species diversity on individual islands (and regionally).

Instead of exploring the effects of colonization and extinction on species diversity, metapopulation theory relates these processes to the regional population dynamics of one or a few species (reviewed by Hanski, 1998). Metapopulation studies show how reduced connectivity increases extinction rates of local populations and reduces rates of recolonization of vacant habitat patches. The net effects of such dynamics depend on whether species have their own independent dynamics, as represented in single species models, or whether species can be driven locally extinct by predators or competitors. As shown in **Fig. 1** the isolation of increasingly small habitat patches can only reduce the likelihood of regional persistence for single or non-interacting species, whereas interacting species can actually benefit from moderate reductions in connectivity. For example, in predator and prey metapopulation models a voracious predator might drive its prey species extinct from large well-connected areas of habitat. Reduced connectivity might reduce the ability of predators to reach areas containing prey, thereby weakening the net effect of predators on prey by providing prey with a refuge from predation. If connectivity is reduced too much, prey may be hindered from reaching habitat areas from which they have been driven extinct, so that local extinction rate exceeds the recolonization rate and eventually prey are driven extinct and predators would starve to death. In predator and prey metapopulation models the dynamics of prey at extremely low levels of connectivity are similar to those of single species. There is a moderate amount of evidence supporting such dynamics for predators and prey, but mostly from microcosms and other highly manipulable systems. Metacommunity models of many competing species show similar effects of reduced connectivity, with the added prediction that species that are the worst at dispersing should be least able to withstand reductions in connectivity even if they are strong competitors in local communities.

The isolation of a habitat patch refers to the rate of immigration into that habitat patch. Hence, isolation and connectivity are inversely related. Mortality during dispersal may add to isolation. In the absence of immigration populations are expected to suffer

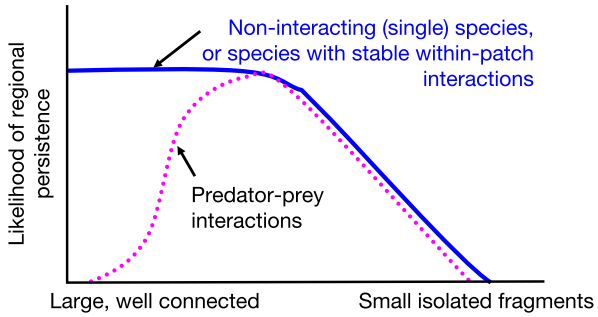


Fig. 1 The influence of habitat subdivision on single species and interacting species where one species can drive another extinct within a habitat patch (e.g., a predator and prey, or a dominant and subordinate competitor species). In highly connected habitat patches the predator is capable of driving the prey species regionally extinct whereas a single species can persist. At extremely low levels of connectivity in small habitat fragments all species go extinct regardless of their interactions. Overall the figure illustrates that the degree to which species interact may alter how they are influenced by changes in connectivity and habitat patch size.

from several effects. Small populations are likely to have reduced genetic variability as a result of the small number of individuals present. This may be compounded by the accumulation of deleterious mutations through inbreeding leading to mortality and reduced reproductive capacity in a population. Small populations are also more likely to go extinct, such that a metapopulation with lots of small populations in a region may experience a regional reduction in genetic diversity, which may exasperate the effects of small population size through reducing the potential for gene flow. Populations with low heterozygosity may have a reduced potential for future evolution. A metaanalysis (quantitative review) by Derek Spielman, Barry Brook, and Richard Frankham of over 170 threatened taxa from the IUCN-World Conservation Union Red List showed that on average threatened taxa had a 35% reduction in heterozygosity compared to non-threatened relatives. Several empirical studies indicate that the loss of heterozygosity is frequently associated with a reduction in reproductive rates through inbreeding. Furthermore, computer simulation models that involve reasonable assumptions suggest that this 35% level of loss of heterozygosity could cause 24%–78% reductions in the median time to population extinction because of reduced reproduction (estimates vary depending on assumptions about the extent of juvenile mortality caused by inbreeding). Habitat loss and fragmentation are the prime causes for imperilment of taxa (at least in the U.S.), and therefore reductions in connectivity are likely to be involved in these effects. It is important to realize that in severe cases reduced connectivity likely produces genetic changes that can impact populations before purely demographic processes cause these populations (or species) to go extinct. Hence, both demographic and genetic changes are important to understanding the effects of habitat fragmentation and reduced habitat connectivity.

Implicit in the idea of measuring connectedness is that there is a sharp demarcation between habitat and non-habitat. By contrast many organisms are not strict habitat specialists and may use different landscape elements to different extents. A step towards recognizing that species may not sharply demarcate habitat from non-habitat while still simplifying the landscape is to recognize the matrix that lies between habitat patches. Differentiating between species that use the matrix and those that do not may improve the predictability of analyses that look at species diversity in relation to connectivity because connectivity is only relevant to species that do not use the matrix as habitat. The same is true of metapopulation studies that investigate patch occupancy in relation to connectivity. A review by Jayme Prevedello and Marcus Vieira in 2010 examined whether the composition of the habitat matrix affected abundance or species richness in 104 empirical fragmentation studies. They found that although the composition of the matrix affected abundance or species richness in 95% of studies, effects were highly species specific, and the effect size was on average smaller than that of patch area and isolation (which is a habitat connectivity metric).

Measuring Connectivity

There are a variety of ways of quantifying the structural and functional connectivity of a landscape. Structural connectivity is usually quantified from aerial photographs, maps, or remote sensing data (GIS–geographic information systems data, satellite imagery). Many metrics require that an image is rasterized by overlaying a grid of cells, with each cell having a defined size or grain. The impression of structural connectivity may vary with the grain size that is chosen, with coarser grains making it more likely that small gaps between habitat areas will be overlooked because large cells average across the gaps. Similarly, organisms may have a particular spatial scale at which they sample or are affected by the environment, and move through it either in a conscious way or by the action of diffusion or a vector. The grain of our sampling a landscape to measure connectivity should be sufficiently fine that it is congruent with the scale selected by the study organism. However, this is often not known prior to commencing a connectivity analysis and therefore analyses at multiple scales are of value to identify the grains that have the highest ability to describe the movement or occupancy patterns of a species across a landscape.

There is a distinction between connectivity measured between pairs or landscape elements, or other portions of a landscape, and connectivity across an entire landscape. Often movement is measured between pairs of points, or along particular paths (e.g.,

for radio-tracked animals) and these data are then analyzed in relation to habitat type or the occurrence of particular landscape elements (rivers, roads, etc.). A typical approach is to start with pairwise measures of connectivity between two points and then repeat such measurements across a landscape. However, there are no general guidelines available to set the spatial scale at which connectivity should be measured.

It is useful to further consider functional connectivity, which can either be based on potential or actual measurements. Potential connectivity comes from combining structural connectivity with information about the movement behavior, distance and costs in the organism in question. Actual connectivity is that which is measured. Be it potential or actual, functional connectivity can be based on strict adjacency (touching) of habitat, a threshold maximum dispersal distance, or some other function of distance. Common forms include a decreasing function of distance, reflecting that movement frequency (or potential) is inversely related to distance, or a resistance-weighted distance function that incorporates the different costs of traversing various paths.

A useful framework for classifying connectivity metrics was presented by Justin Calabrese and Bill Fagan (2004), and this framework is used here to organize connectivity measures. Furthermore, the description of the advantages and disadvantages of different metrics, as well as the data requirements to estimate them draws heavily on their work. Six classes of metrics can be distinguished.

Nearest Neighbor Distance Metrics Using Patch Occupancy Data and Interpatch Distances

The crudest measures of connectivity are based on the distance from a given habitat patch to other occupied habitat patches, which might provide immigrants. This is an index of isolation, and is turned into an index of connectivity by taking its inverse. Indices of this type use only the distance to the nearest habitat patch, ignoring all other patches that might provide immigrants, and variation in the size of source populations in different patches. Metaanalyses and simulations both confirm that nearest neighbor metrics are crude estimates of connectivity and that they have lower predictive ability than the more sophisticated measures discussed below. Measures could be converted from structural measures to potential connectivity by converting links from distances to binary presences or absences, or probabilities using known movement distances and a dispersal kernel (describing the probability of dispersal as a function of distance), respectively. Nearest neighbor distances are the most widely used of the different kinds of connectivity metrics, despite their low predictive ability.

Spatial Pattern Indices Using Spatially Explicit Habitat Data

A variety of metrics can be calculated from a map portraying the spatially-explicit layout of habitat areas. These include things like the number of patches, patch sizes, patch perimeter to area ratios, fractal dimension, contagion and indices of patch shape. Such indices are readily calculated with software packages such as “Fragstats” and reflect the popularity of analyzing GIS data. However, simulated analyses of the ability of these indices to predict connectivity suggest that they are inconsistent when the characteristics of a landscape are varied. Empirical work is needed to establish relationship between actual connectivity and structural connectivity measured using spatial pattern indices.

Scale-Area Slope Indices Using Spatially Incomplete Data

Records of individuals at either random locations on a landscape or from a systematic survey of points on a landscape can be used to quantify the occurrence of a species across a landscape and how this varies with spatial scale. Specifically, a grid can be overlaid and the presence and absence of the species in each cell recorded. As grid cells are made larger (or adjacent grid cells summed) the number of records can be plotted as a function of cell size; specifically, the map area occupied by a species can be plotted against the size of the cells sampled and a power function can be fitted using regression. Such plots will have shallow slopes if species are uniformly distributed across space, which is taken as an indicator of highly connected landscapes. Conversely steep slopes are likely to result from aggregated distributions and these are presumed to arise because of limited movement. The slope is called the scale-area slope and is a measure of connectivity. Such approaches cannot distinguish whether steep slopes are because habitat is aggregated in its distribution across the landscape or whether habitat is uniformly distributed and the organisms are restricted in their movement for some reason other than habitat connectivity. The approach also assumes that proximity is the major determinant of connectivity, an assumption that had predictive power for the long-term dynamics of populations of fishes in desert springs and streams in the southwestern United States. The validity of the approach for other study systems requires evaluation.

Graph-Theoretic Measures Using Dispersal Data and Spatially Explicit Habitat Data

Graph theoretic measures of connectivity use spatially explicit habitat data and known biological information about dispersal to estimate potential connectivity. The approach consists of using a habitat graph that summarizes habitat patch arrangement (as “nodes”) and patch information in a concise way. One common way to convert a graph to measures of potential connectivity is to use either fixed dispersal distances to create links if the metric is less than the maximum dispersal distance or to do this probabilistically if a dispersal kernel is available. A dispersal kernel is a function describing the probability of dispersal as a function if

distance, and can be combined with structural landscape data by using random draws to decide whether a patch is connected or not depending on the appropriate probability of dispersal for the distance in question. Pairwise metrics such as the maximum distance able to be traveled in a random direction are then scaled up to the entire landscape. Three main types of approaches are in use, based on network analysis, percolation theory and the correlation length of spanning clusters, which are described below.

Networks and their application to connectivity are introduced in detail by Bronwyn Rayfield and colleagues in a 2011 review paper. A network consists of a weighted collection of nodes (locations), with weighting either being by population size at each node or the strength of links between nodes. The related subject of circuit theory considers the resistance of different links to flow, and emphasizes and analyzes the many ways in which circuits can be connected. Network analyses are more widely used in analyses of habitat connectivity, and building on available mathematical and statistical tools have gone from simple undirected equally weighted links to inclusion of factors like habitat suitability, different matrix permeability (see next paragraph), and movement behavior. Link weights can be assigned to reflect habitat characteristics (structural connectivity) or actual movement (functional connectivity). Analyses can consider either single (or least cost) pathways or multiple pathways through networks. In turn, metrics can be calculated for various properties at different scales from elements (nodes) within networks to whole networks or collections of "meta-networks". Rayfield and colleagues break down metrics into route-specific quantities, route redundancy, route vulnerability to breakage from loss of links or nodes, and area of connected habitat. There are a large number of such metrics and readers should refer to the literature on network analyses of connectivity for further details.

Percolation is most easily understood for a rasterized grid of habitat and non-habitat cells and is a measure of the probability that habitat cells are contiguous. Percolation theory comes from mathematics and was developed in physics before being applied to landscape ecology. The main concept is the existence of a percolation threshold, defined in the following way. Suppose p is a parameter that defines the average degree of connectivity between cells (in a grid) of a landscape classified into habitat and non-habitat. When $p = 0$, all patches are totally isolated from every other sub-unit. When $p = 1$, all habitat cells are connected to (touching) their neighbors. At this point, the landscape is connected from one side to the other, since there are paths that go completely across the system, through spanning clusters. Now suppose, starting at $p = 1$, habitat cells are randomly removed, so that p, the measure of average connectivity, decreases. The percolation threshold is that value of p, usually denoted p_c, at which there is no longer an unbroken path from one side of the system to the other. Measures of structural connectivity based on percolation can be changed into functional connectivity by allowing gaps between habitat cells of certain distances, corresponding to the organism's maximum movement distance across non-habitat.

The correlation length for a rasterized habitat patch map is based on the average extensiveness of connected cells. The correlation length is the average distance one might traverse across a landscape without leaving what is defined as habitat from a random starting point and moving in a random direction.

An advantage of graph-theoretic approaches is that they can be used to calculate how an individual patch contributes to landscape-scale connectivity. However, such approaches are data-intensive, requiring both movement data (maximum dispersal distance or a dispersal kernel) and spatially explicit landscape data representing habitat and non-habitat areas. Atte Moilanen in 2011 cautioned against using graph-theoretic measures in conservation and management because of their low ability to predict ecological patterns related to connectivity.

Buffer Radius and Incidence Function Metrics Using Spatially Explicit Patch Occupancy Data, Patch Area, and Dispersal Data

Building on the idea of nearest neighbor distances a buffer radius calculates the number of occupied habitat patches or total number of individuals of a species within a given distance. The metric uses more information than nearest neighbor distances, potentially improving its accuracy and can also be calculated at multiple spatial scales. The spatial scale of relevance will depend on typical movement distances and consequently movement data could greatly improve the utility of buffer radius connectivity metrics. A related idea is captured in Ilkka Hanski's (1998) incidence function model approach, which includes an explicit function of the number of individuals in habitat patches at different distance and these are weighted by distance in a power function. The approach can also be adapted to use information just on patch occupancy, or patch area as a surrogate measure, if the number of individuals per patch is not known. The approach has been tried and tested with a variety of taxa and appears to work well when organisms occur in discreet patches and are limited in their dispersal ability. As a minimum information on patch occupancy, or colonization are required and this needs to be complete spatially, without missing (uncensused) habitat patches in the study area. Such approaches are mainly limited by the high data requirements and limited areas that can be censused.

Observed Movement Rates Requiring Individual Movement Data

A broad range of metrics of movement are collected by ecologists and biologists. Common forms of measurement are tracking individuals using radio-telemetry, satellite tagging, radiotelemetry, or other methods, mark-release-recapture, mass mark-recapture, observations of colonization of habitat patches over some period of time, and indirect estimates from genetic data. Using genetic data requires making assumptions about effective population size and the equilibrium of dynamics, and the neutrality of genetic markers. Rolf Ims and Nigel Yoccoz reviewed the different measures of measuring movement and discuss the advantages and disadvantages of different metrics. The most accurate measures are those that are most labor-intensive to collect and these will necessarily be spatially limited.

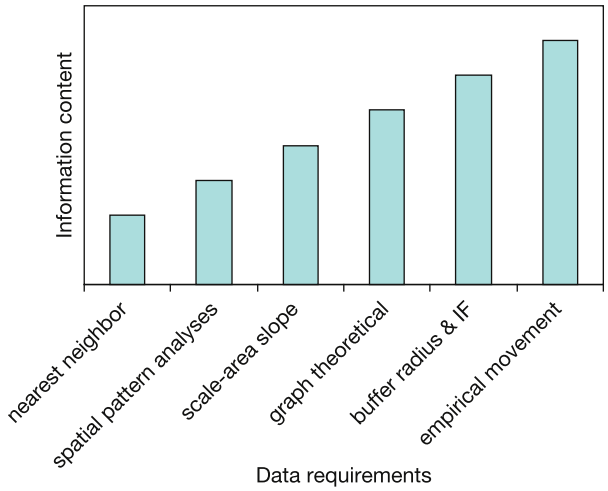


Fig. 2 A representation of the information contained in different metrics of connectivity as a function as the amount of data required for their measurement. IF = incidence function model approaches. Altered from Calabrese, J.M. and Fagan, W.F. (2004). A comparison-shopper's guide to connectivity metrics. *Frontiers in Ecology and the Environment*, **2**, 529–536.

Measuring Connectivity: Putting it All Together

The preceding discussion should have highlighted that connectivity metrics differ in their ease of measurement and the amount and type of data required to use them. There is a tradeoff between the ease of measurement of connectivity metrics and the amount of detail they capture. This is illustrated in Fig. 2. Hence, we face a choice between metrics in which we are uncertain about their utility or accuracy and investing the time and resources to measure metrics that are more information-rich. The choice of metric is likely to depend on the spatial scale of the problem which we are attempting to solve, be it a conservation strategy for a region, or the need for a migration route for deer across a highway, and so on. It should be noted that the metrics that have been discussed can often be improved by including additional information. Most commonly information about the number of individuals within patches is used, or patch size a surrogate for information about local population size. Alternatively immigration and emigration may scale differently with patch size, such that individuals are more likely to remain in or colonize smaller or larger habitat patches. Movement can also be adjusted for variation in the type of habitat being crossed or landscape features present in the area traversed. Such modifications to metrics have been shown to improve the accuracy in specific studies, but their general utility is unknown.

Corridors

Corridors are continuous strips of habitat connecting areas of habitat that would otherwise not be connected. They are one of the major ways in which we attempt to improve habitat connectivity. However, corridors serve many purposes that are related to connectivity, from facilitating the movement of individuals, enhancing social structure (e.g., for primates), population viability (e.g., for spotted owls in the Pacific Northwest of the U.S.), and community structure to improving ecosystem properties. Furthermore, corridors can be natural (e.g., riverine vegetation) or constructed (e.g., wildlife crossings on roads), deliberate or inadvertent (e.g., road verges), and vary in scale from highway crossings to ambitious international projects such as connecting the Greater Yellowstone area in Montana and Idaho to the Yukon and Alaska with a forest corridor to facilitate the movement of large mammals. Constructed corridors are often greenbelts or greenways in urban and suburban areas that serve recreational purposes, enhance ecosystem functioning (e.g., groundwater infiltration), or are low value areas, such as those subject to flooding.

When first conceived of, corridors were viewed as being non-selective pathways that facilitate movement. The reality is that certain species may benefit from corridors of a particular type, depending on width, length, habitat composition within them, the surrounding habitat matrix, and what they are connecting; all of these things are measured relative to the requirement of the species and individuals in question. Corridors may also be selective in terms of factors like the ages, social status or other traits of the individuals dispersing. Increasing connectivity through the use of corridors is not always good. For example, weedy plants are often spread along highway corridors, introducing nonnative invasive species that are selected for factors like rapid generation times, high movement capacity, being able to benefit from disturbance (e.g., open ground for germination) and high competitive ability. The spread of a disease, pathogen, predator, dominant competitor or nonnative invasive species may also occur and impact either particular species or whole communities and ecosystems. Nonetheless, the negative effects of isolation on population viability, genetic diversity, species diversity and community structure are generally severe so that the weight of evidence suggests that improving connectivity is usually desirable. It is clear that long term studies of the community-wide effects of altering habitat connectivity are especially desirable. Fig. 3 shows an aerial view of a long-term ecological experiment where Nick Haddad and colleagues manipulated the connectivity of pairs of habitat patches consisting of clearings within pine forest in South Carolina.

Fig. 3 False color aerial photograph showing a large-scale habitat fragmentation experiment conducted by Nick Haddad and colleagues. Three treatments are shown: pairs of patches connected by a habitat corridor (100 m by 100 m patches connected by a 25-m–by–150-corridor; an unconnected patch (of 137.5 m by 100 m) and winged patches that have the same total habitat area (100 m by 100 m, with two 25-m–by–75-m wings). The amount of edge habitat is also broadly similar in a connected patch and a winged patch.

The experiment controls for the amount of habitat edge and the increase in habitat area caused by having corridors by using winged patches that are of similar area to patches connected by corridors except the corridors do not connect to other patches. Nick Haddad and colleagues showed that connecting patches using corridors increased the interpatch movement rate of a diverse suite of taxa, including butterflies, small mammals, and bird-dispersed plants. In the same study system Ellen Damschen and colleagues demonstrated that by the end of a 5-year period after initiating the experiment the connected patches contained an average of 20 more plant species than unconnected patches.

Summary

Connectivity has important links to a broad range of population processes, genetic variability (and evolutionary potential), species diversity and community structure. Connectivity can be defined through structural and functional means, and functional connectivity can be further subdivided into potential and actual measurements. Connectivity can be measured between pairs of patches or across entire landscapes and many scales between. A complex diversity of metrics has been used to measure connectivity, with different metrics exhibiting different degrees of reliability, to the extent that this is known. Different metrics also differ in whether they require movement data or not, and spatially explicit landscape data or information on neighboring patches only. A key idea is that connectivity can be manipulated using habitat corridors, although such corridors are often constructed and their long-term consequences are less often explored.

References

Calabrese, J.M., Fagan, W.F., 2004. A comparison-shopper's guide to connectivity metrics. Frontiers in Ecology and the Environment 2, 529–536.
Hanski, I., 1998. Metapopulation ecology. Oxford: Oxford University Press.
MacArthur, R.H., Wilson, E.O., 1967. The theory of island biogeography. Princeton, NJ: Princeton University Press.
Polis, G.A., Holt, R.D., Menge, B.A., Winemiller, K.O., 1996. Time, space, and life history: Influences on food webs. In: Polis, G.A., Winemiller, K.O. (Eds.), Food webs: Integration of patterns and dynamics. New York, NY; London, England: Chapman and Hall, Inc., pp. 435–460.
Rayfield, B., Fortin, M.-J., Fall, A., 2011. Connectivity for conservation: A framework to classify network measures. Ecology 92, 847–858.

Further Reading

Damschen, E.I., Haddad, N.M., Orrock, J.L., Tewksbury, J.J., Levey, D.J., 2006. Corridors increase plant species richness at large scales. Science 313, 1284–1286.
Haddad, N.M., Bowne, D.R., Cunningham, A., Danielson, B.J., Levey, D.J., Sargent, S., Spira, T., 2003. Corridor use by diverse taxa. Ecology 84, 609–615.

Hilty, J.A., Lidicker Jr., W.Z., Merenlender, A.M., 2006. Corridor ecology: The science and practice of linking landscapes for biodiversity conservation. Washington, Covelo, London: Island Press.
Kool, J.T., Moilanen, A., Treml, E.A., 2013. Population connectivity: Recent advances and new perspectives. Landscape Ecology 28, 165–185.
Spielman, D., Brook, B.W., Frankham, R., 2004. Most species are not driven to extinction before genetic factors impact them. Proceedings of the National Academy of Sciences of the United States of America 101, 15261–15264.
Tischendorf, L., Fahrig, L., 2000. How should we measure landscape connectivity? Landscape Ecology 15, 633–641.
With, K.A., 1999. Is landscape connectivity necessary and sufficient for wildlife management? In: Rochelle, J.A., Lehmann, L.A., Wisniewski, J. (Eds.), Forest fragmentation: Wildlife and management implications. Leiden, The Netherlands: Koninklijke Brill NV, pp. 97–115.
With, K.A., Gardner, R.H., Turner, M.G., 1997. Landscape connectivity and population distributions in heterogeneous environments. Oikos 78, 151–169.

Relevant Websites

Conservation corridor, n.d., http://conservationcorridor.org/—"Conservation corridor" has a mission to bridge the science and practice of conservation corridors and contains a wide range of articles and tools about habitat corridors.
The Ecological Society of America's Issues in Ecology, n.d., http://www.esa.org/esablog/research/landscape-connectivity-corridors-and-more-in-issues-in-ecology-16/—"The Ecological Society of America's Issues in Ecology" series describes models and methods for reconnecting wildlife habitat in restoration and conservation planning and management.
Fragstats, n.d., http://www.umass.edu/landeco/research/fragstats/fragstats.html—"Fragstats" connectivity metrics calculations package.
Pathmatrix, n.d., http://cmpg.unibe.ch/software/pathmatrix/—"Pathmatrix" a GIS tool to compute distances among samples.
Conefor, n.d., http://www.conefor.org/—"Conefor" is a software package that allows quantifying the importance of habitat areas and links for landscape connectivity.
Marine Geospatial Ecology Tools, n.d., http://mgel.env.duke.edu/mget—"Marine Geospatial Ecology Tools (MGET)", also known as the GeoEco Python package, is an open source geoprocessing toolbox designed for coastal and marine researchers and GIS analysts.

Ecological Informatics: Overview

F Recknagel, University of Adelaide, Adelaide, SA, Australia

Introduction

Ecological informatics (ecoinformatics) is an interdisciplinary framework for the management, analysis, and synthesis of ecological data by advanced computational technology. Management of ecological data aims at facilitating data standardization, retrieval, and sharing by means of metadata and object-oriented programming. Analysis and synthesis of ecological data target elucidating principles of information processing, structuring, and functioning of ecosystems, and forecasting of ecosystem behaviors by means of bioinspired computation and hybrid models.

Ecological informatics currently undergoes the process of consolidation as a discipline. It corresponds and partially overlaps with the well-established disciplines of bioinformatics and ecological modeling but is taking its distinct shape and scope. In **Fig. 1**, a comparison is made between ecological informatics and bioinformatics. Even though both are based on the same computational technology, their focus is different. Bioinformatics focuses very much on determining gene function and interaction, protein structure and function, as well as phenotypes of organisms utilizing DNA microarray, genomic, physiological, and metabolic data (**Fig. 1a**). By contrast ecological informatics focuses to determine genotypes of populations by utilizing genomic, phenotypic, and environmental data as well as structure and functioning of ecosystems by utilizing community, environmental, and climate data (**Fig. 1b**).

A comparison is made between ecological modeling and ecological informatics in **Fig. 2**. Even though both rely on similar ecological data they adopt different approaches in utilizing the data. While ecological modeling processes ecological data top down by *ad hoc* designed statistical or mathematical modeling methods, ecological informatics infers ecological processes from ecological data patterns bottom up by computational techniques. The cross-sectional area between ecological modeling and ecological informatics reflects a new generation of hybrid models that enable to predict emergent ecosystem structures and behaviors, and ecosystem evolution. Typically hybrid models embody biologically inspired computation in deterministic ecological models.

Feature Areas

Current research in ecological informatics focuses at four major feature areas:

1. understanding information processing and evolution in ecosystems;
2. computational management of ecological data;
3. computational analysis and synthesis of ecological data; and
4. hybrid modeling of ecological data.

Great efforts are undertaken to address feature area (1) by studying both intraspecific population adaptations to changing climate and habitat conditions as well as interspecific population relationships controlled by infochemicals and allelopathy.

The feature area (2) aims at standardized archiving of highly complex and fragmented ecological data in order to allow ecological data sharing. The ecological metadata language EML (http://knb.ecoinformatics.org/software/eml/) is an example for developing computational tools based on metadata concepts that will facilitate ecological data warehousing at global scale.

The feature area (3) is being largely stimulated by both the availability of complex ecological data including genomic and phenotypic data, and the development of bioinspired computational techniques. The study of population genomics in their natural habitats without the need for isolation and lab cultivation of individual species has led to the new research area of ecogenomics (also called metagenomics) that promises to determine the impact of environmental and climate changes on biodiversity. Bioinspired computational techniques prove to be superior in unraveling highly complex ecological data, coping with distinct nonlinearities and inducing predictive models by learning from temporal and spatial patterns. The next section illustrates applications of artificial neural networks (ANNs) and evolutionary algorithms for ecological informatics.

Research on hybrid modeling in the feature area (4) promises ecosystem models with improved accuracy and generality. Cao and Recknagel provided a case study for multiobjective optimization of process and parameter representations in process-based ecosystem models by the embodiment of evolutionary algorithms in ordinary differential equations for food web dynamics and nutrient cycles in lakes.

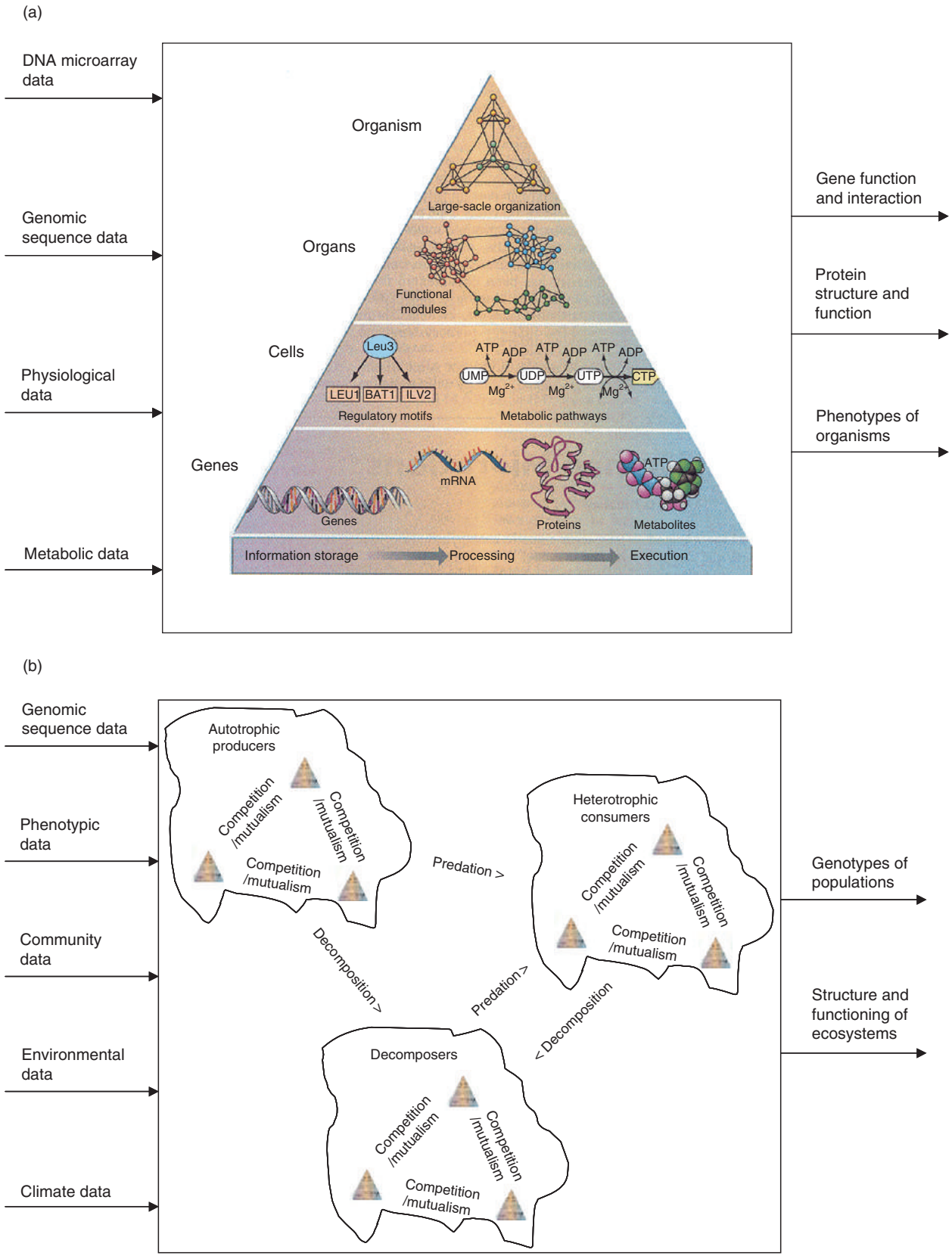


Fig. 1 Ecological informatics versus bioinformatics. (a) Scope of bioinformatics. (b) Scope of ecoinformatics. Modified from Oltvai ZN and Barabasi AL (2002) Life's complexity pyramid. *Science* 298: 763–764.

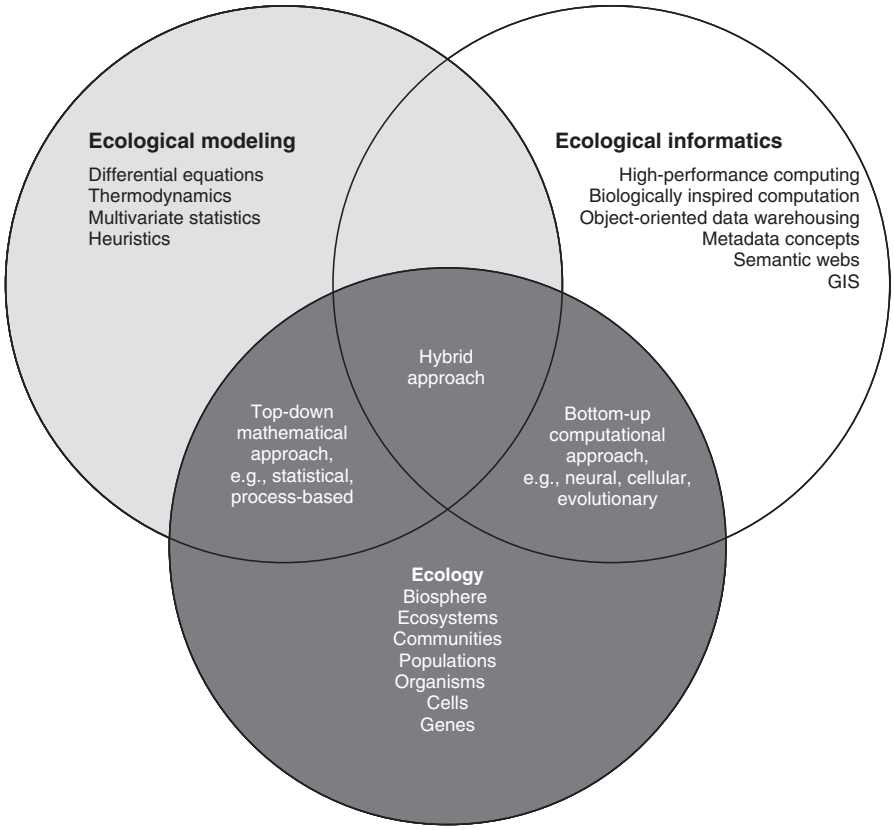


Fig. 2 Ecological informatics versus ecological modeling.

Ecological Informatics by Computational Analysis and Synthesis of Ecological Data

Artificial Neural Networks

ANNs are computer programs designed for inducing problem solutions (models, knowledge) from complex data by means of principles of information processing similar to biological neurons in the human brain. A biological neuron consists of three major components: the cell body, dendrites, and the axon (Fig. 3a). Connections between neurons are formed at synapses. Information is represented and transmitted by chemically generated electrical activity within the cell. Both excitatory and inhibitory inputs to the neuron enter through synaptic connections with other neurons. Input potentials are summed up within the cell body. If the total input potential is sufficient (e.g., meets a certain threshold value), the neuron acts. Ultimately an action potential is generated and propagated down the axon toward the synaptic junctions with other nerve cells.

The design of ANNs (Fig. 3b) has been inspired by the structure and functioning of biological neurons. The dendrites which are acting as input receptors were represented by input units. The cell body that acts as information accumulator was represented by activation units adjusting and summing up the weights of inputs, and the input–output transfer function. The axon that acts as the biological output channel was represented as the output.

ANNs gain adaptive capability by undergoing training similar to neural learning where two basic training modes are distinguished: supervised and nonsupervised. The supervised training aims at the optimal approximation of the calculated output Y_c to the observed (desired) output Y_o. An iterative adjustment of input weights takes place in order to minimize the error $(Y_o - Y_c)$. After training, the generalization of the supervised ANN is assessed by feeding it only with input values, not observed output values, and testing how close calculated outputs match observed outputs. The two most common methods for assessing generalization are the 'split-sample validation' and the 'cross-validation'. The 'split-sample validation' means that part of the data is reserved as a test set, which must not be used in any way during training. The test set must be representative for the problem to be modeled by the ANN. After training, the ANN is run on the test set, and the error on the test set provides an estimate of the generalization error usually expressed by the root mean square error (RMSE) or the correlation coefficient r^2. The disadvantage of split-sample validation is that it reduces the amount of data available for both training and validation. By contrast, 'cross-validation' allows use of all the data for training. In k-fold cross-validation, the data is divided into k equal-sized subsets. The net is trained k times, each time leaving out one of the subsets from training, but using only the omitted subset to compute the generalization error. If k equals the sample size, this is called 'leave-one-out' cross-validation. The disadvantage of cross-validation is that the ANNs need to be retrained many times.

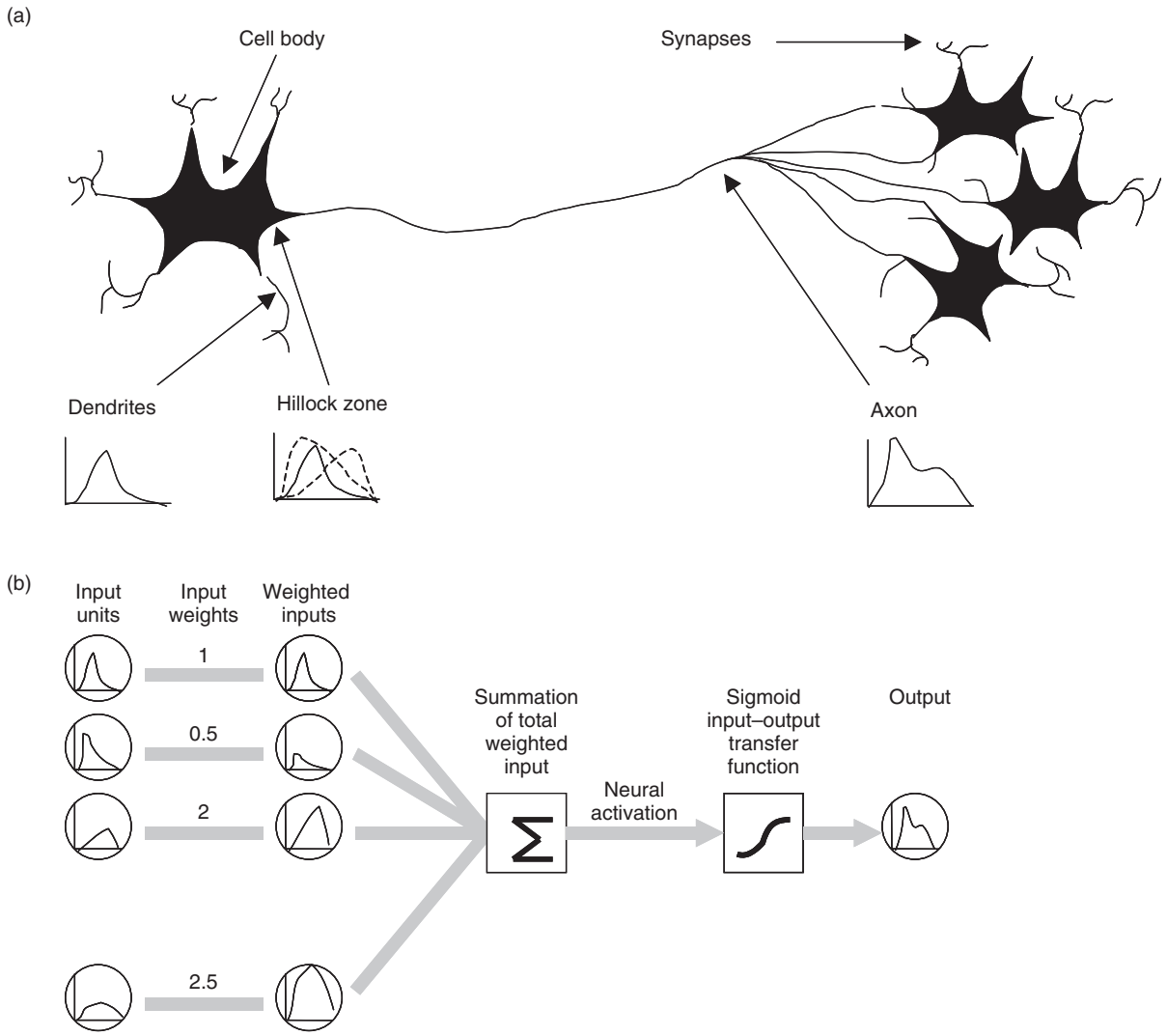


Fig. 3 Conceptual structures of biological and artificial neurons.

Depending on using external inputs only or feedback inputs as well, supervised ANNs are differentiated into feedforward or feedback ANNs (see **Figs. 4a** and **4b**). By contrast, nonsupervised ANNs process external inputs only without adjusting calculated outputs to known outputs (**Fig. 4c**).

Supervised feedforward ANN

The supervised feedforward ANN proves to be a universal approximator of multivariate nonlinear functions and is usually implemented as multilayer perceptron with back-propagation training. The multilayer perceptron represents input units as input layer, adjusted and accumulated input weights as hidden layer(s), and outputs as output layer. The back-propagation algorithm performs the iterative adjustment of input weights (activation units) in order to minimize the approximation error ($Y_o - Y_c$).

Supervised feedforward ANNs are widely applied in ecology either using cross-sectional data to predict discrete ecosystem states or using time-series data to predict continuous ecosystem behavior. Successful applications by means of cross-sectional data have been demonstrated for fish communities in streams, macroinvertebrate communities in streams, river salinity, primary productivity in esturaries, chlorophyll *a* concentrations in lakes, coastal vegetation, and bird populations.

Successful applications by means of time-series data have been demonstrated for marine fish and zooplankton communities, river hydrology, macroinvertebrate communities in streams, freshwater phyto- and zooplankton communities.

The majority of the supervised feedforward ANNs documented achieved forecasting results that were superior to conventional modeling techniques such as multiple linear regression. Even though supervised ANNs do not provide explicit mathematical representations of the underlying ANN model, most of the authors have conducted sensitivity analyses in order to identify inputs as key driving forces of the predictive ANN.

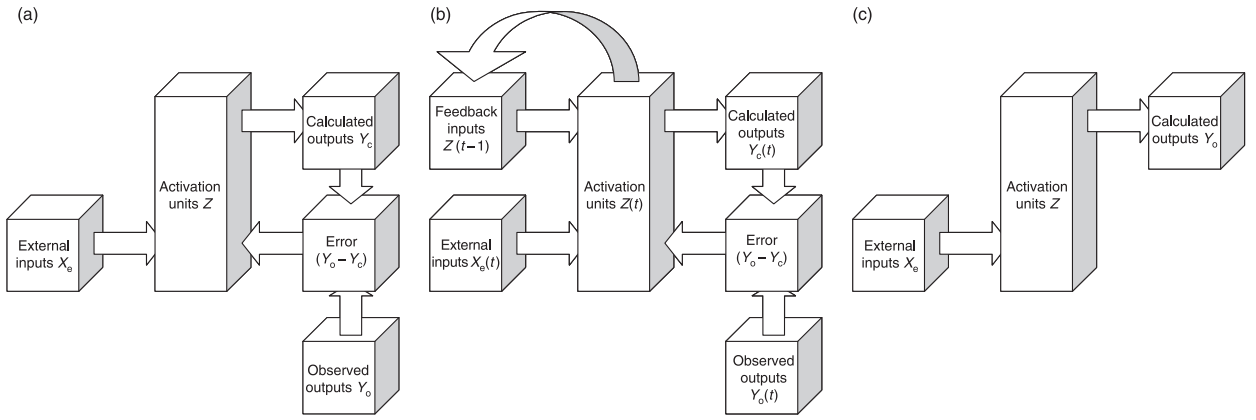


Fig. 4 Basic types of ANNs: (a) supervised feedforward ANN; (b) supervised feedback ANN; (c) nonsupervised ANN.

Supervised feedback ANN

Supervised feedback or recurrent ANNs are designed to use not only external inputs for training but also activation levels of the previous training iteration which are constantly fed back (see **Fig. 4b**). Their functioning can be compared with ordinary differential equations that calculate the current system state $Z(t)$ by taking into account current external inputs $X_e(t)$ and the system state $Z(t-1)$ of the time step before:

$$dZ(t)/dt = f(X_e(t), Z(t-1), P)$$

where P represents constant parameters.

Supervised feedback ANNs prove to be very powerful for modeling time-series data where the fed-back activation levels provide extra training information on the system state of the time step before.

Fig. 5 shows an example for a supervised feedback ANN that has successfully been trained and tested by split-sample validation for the forecasting of the algal populations *Microcystis* and *Stephanodiscus* in River Nakdong in South Korea. The weekly measured limnological data of the river study site were interpolated to daily values. The interpolated data from 1995 to 1998 were used as training set, and the interpolated data of 1994 were used as testing set. In order to achieve a 4-day-ahead forecasting, a 4 days time lag was imposed between the measured inputs and the measured outputs of the training data set. The design of the feedback ANN considered the following 18 external input variables: irradiance, precipitation, discharge, evaporation, water temperature, Secchi depth, turbidity, pH, DO, nitrate, ammonia, phosphate, silica, rotifera, caldocera, copepoda (see also for input sensitivity in **Fig. 5**), 21 hidden activation units, and the two output variables: *Microcystis aeruginosa* and *Stephanodiscus hantzschii*.

After 2100 training iterations, an RMSE of 0.001 7 was achieved and the generalization of the trained ANN was tested based on testing data of 1994. **Fig. 6** shows the visual comparison between the observed and the 4-day-ahead predicted data for *M. aeruginosa* ($r^2 = 0.68$) and *S. hantzschii* ($r^2 = 0.73$). The results indicate a high degree of accuracy in the forecasting regarding both the timing and the magnitudes of populations dynamics of the two algal species, which have their distinctive seasonal patterns.

This application has demonstrated that supervised feedback ANNs achieve a high generalization degree and forecasting accuracy after training by time-lagged time-series data. The typical rapid growth and blooming of the blue-green algae *Microcystis* under warm and calm conditions in mid- and late summer as observed in River Nakdong in 1994 was well reflected by the predicted data in **Fig. 6a**. By contrast diatoms tend to be abundant at moderate temperatures and turbulent conditions. Both observed and predicated data for *S. hantzschii* in River Nakdong correspond well by showing population densities in the highest range, in spring and autumn (**Fig. 6b**).

This case study has also convincingly demonstrated the benefits of sensitivity analyses in order to gain insights and test hypotheses regarding ecological relationships between input and output variables. Results in **Fig. 7** compare the input sensitivities of the two different algal populations that have been interpreted in great detail by Jeong *et al.* The most obvious differences between *M. aeruginosa* and *S. hantzschii* can be seen in their preferred water temperature, pH, and silica levels that comply with ecological theory.

Successful applications have also been demonstrated for time-series modeling of macroinvertebrate communities in streams and phytoplankton communities in freshwater lakes and rivers.

Nonsupervised ANN

Nonsupervised ANNs are designed to identify unknown input patterns based on similarities between inputs. So-called self-organizing maps (SOMs) developed by Kohonen are the most popular nonsupervised ANNs which can be applied to ordination, clustering, and mapping of complex nonlinear data.

The principal approach of nonsupervised ANNs according to Kohonen is represented in a simplified manner in **Fig. 8**. It shows that the neurons of the nonsupervised ANN learn to distinguish between similar and dissimilar features of the normalized input data, which are mapped as clustered inputs. The term nonsupervised in this context means that the learning algorithm is not

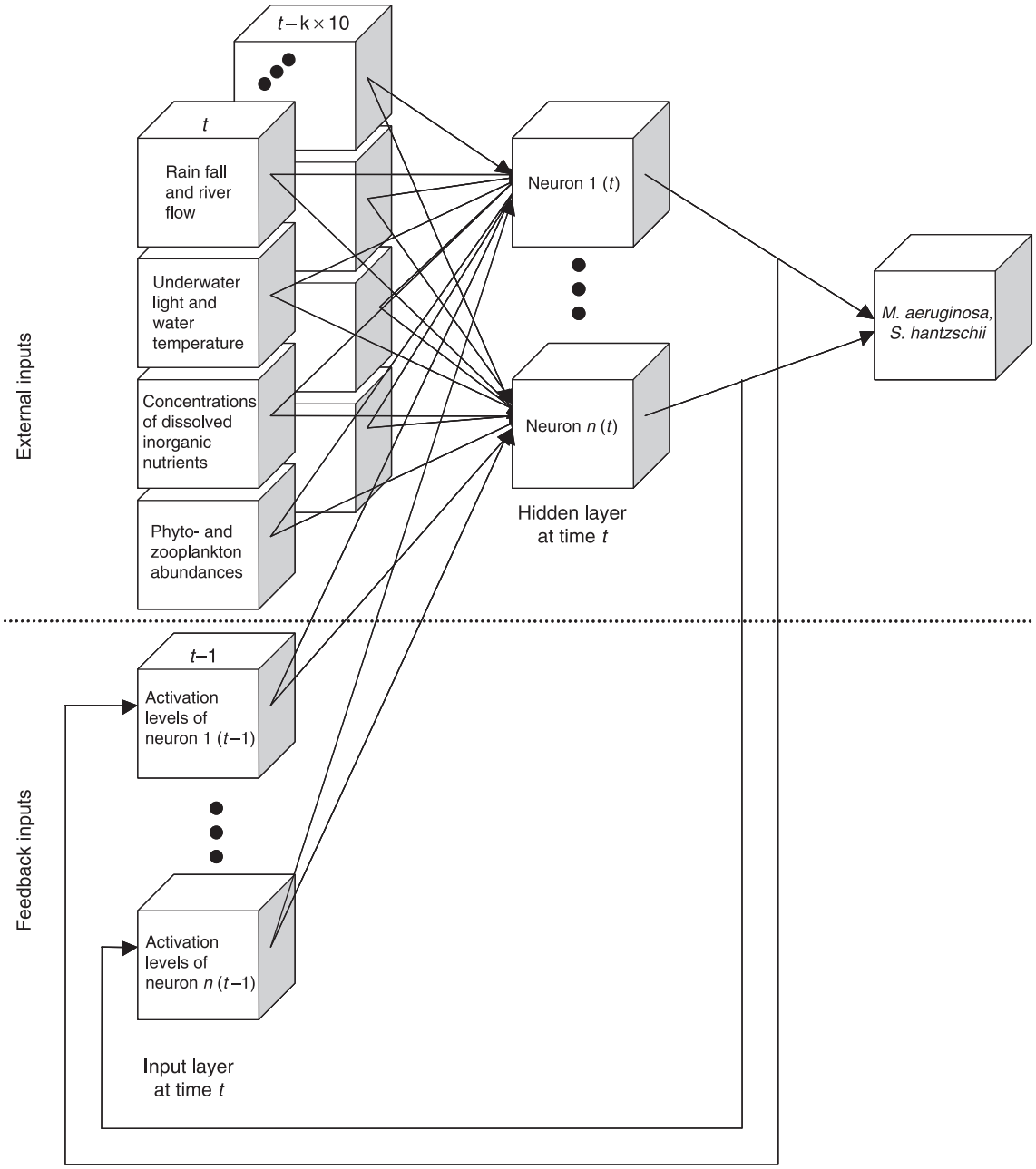


Fig. 5 Supervised feedback ANN for 4-day-ahead forecasting of population densities of *Microcystis aeruginosa* and *Stephanodiscus hantzschii* in River Nakdong (South Korea). Modified from Jeong K-S, Recknagel F, and Joo G-J (2006) Prediction and elucidation of population dynamics of the blue-green algae *Microcystis aeruginosa* and the diatom *Stephanodiscus hantzschii* in the Nakdong River–Reservoir System (South Korea) by a recurrent artificial neural network. In: Recknagel F (ed.) *Ecological Informatics: Scope, Techniques and Applications*, 2nd edn., pp. 255–273. Berlin: Springer.

guided by known output patterns but learns the patterns from features of the inputs. Those features can be expressed by Euclidean distances, which are calculated between the inputs and weights. Similarities between inputs in terms of Euclidean distances can be visualized and partitioned by the unified distance matrix (U-matrix) and the K-means map.

In order to illustrate opportunities of applications of nonsupervised ANN to ecological time-series data, **Figs. 9–12** show results of a case study carried out for limnological data of Lake Kasumigaura in Japan. **Fig. 9** represents seasonal clusters for Lake Kasumigaura as mapped by the U-matrix and K-means partitioning using the SOM Toolbox of MATLAB 5.3. The U-matrix map in **Fig. 9a** visualizes the relative distances between neighboring data of the input data space as shades of gray. The light areas in the U-matrix visualize neighboring data with distances in the shortest range belonging to a region or cluster. The black colors represent the distances in the longest range between neighboring data and denote borders between clusters. The K-means algorithm

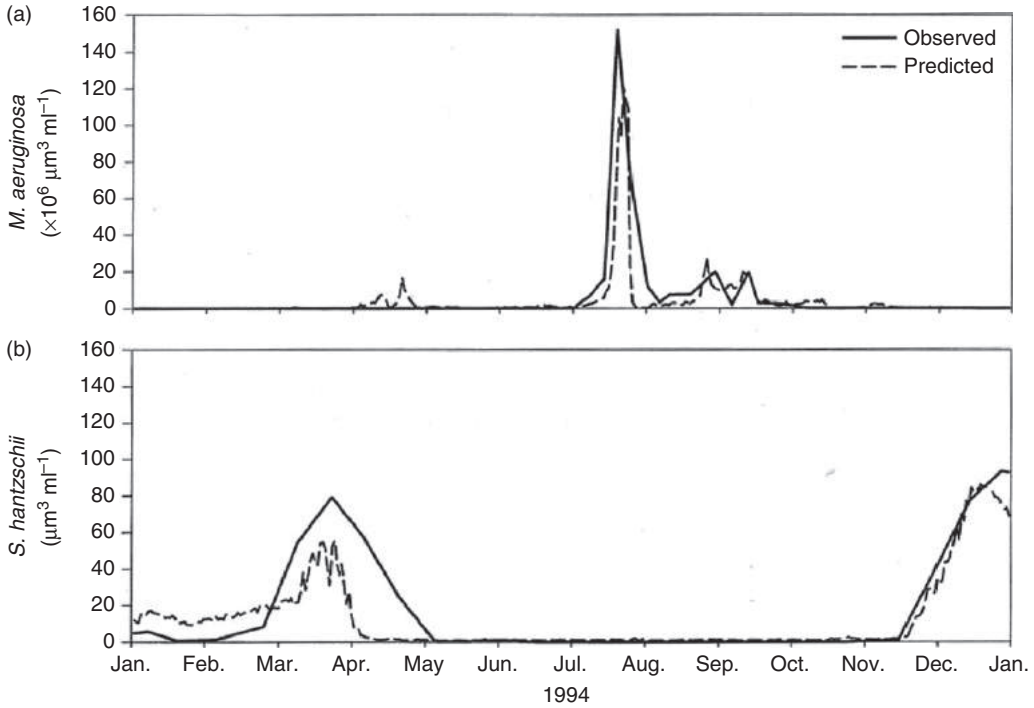


Fig. 6 Four-day-ahead forecasting of population densities of *M. aeruginosa* and *S. hantzschii* in River Nakdong (South Korea) by means of a supervised feedback ANN. Modified from Jeong K-S, Recknagel F, and Joo G-J (2006) Prediction and elucidation of population dynamics of the blue-green algae *Microcystis aeruginosa* and the diatom *Stephanodiscus hantzschii* in the Nakdong River–Reservoir System (South Korea) by a recurrent artificial neural network. In: Recknagel F (ed.) *Ecological Informatics: Scope, Techniques and Applications*, 2nd edn., pp. 255–273. Berlin: Springer.

partitions the input data space into a specified number of clusters based on the U-matrix. **Fig. 9b** represents the corresponding partitioned map for five seasons.

Fig. 10 visualizes seasonal distributions of abundances of the blue-green algae *Microcystis* and *Oscillatoria* in Lake Kasumigaura based on data of the years 1984–86 (left column) and 1987–89 (right column). **Fig. 11** represents the seasonal distributions of concentrations of NO_3-N and PO_4-P in Lake Kasumigaura in correspondence with the time periods differentiated in **Fig. 9**. **Fig. 10** highlights that while *Microcystis* declines in cell numbers by more than 50% between 1984–86 and 1987–89, *Oscillatoria* doubles in cell numbers. It also shows that seasonal dominance of two algal populations for the early and the late 1980s shifted for *Microcystis* from late summer to autumn, and for *Oscillatoria* from early summer to late summer. Takamura *et al*. pointed at changes of NO_3-N/PO_4-P ratios as possible explanations for the succession of the two blue-green algal populations during the 1980s in Lake Kasumigaura, that are indicated by the component planes in **Fig. 10**. From the early to the late 1980s, the NO_3-N concentrations increased by 50% while PO_4-P concentrations dropped to 50%, causing a significant change of the NO_3-N/PO_4-P ratios (from 8.5 to 32).

A combination of input sensitivity curves by supervised feedback ANN with component planes by nonsupervised ANN proves to be an informative approach (e.g., for hypothesis testing). While component planes allow to map nonlinear relationships of output variables with predefined input ranges in a qualitative manner (see **Fig. 12**, top), input sensitivity curves draw numerical relationships of output variables over the whole range of input variables as learnt from training data. Both the component planes and sensitivity curves in **Fig. 12** confirm theoretical assumptions that the diatoms *Cyclotella* have a preference of low to medium water temperatures typically occurring in spring and autumn, while the population growth of *Microcystis* reaches rates in the highest range at high water temperatures in mid- and late summer.

Successful applications of nonsupervised ANNs have been demonstrated for cross-sectional data of macroinvertebrate communities in streams and vegetation types.

Successful applications of nonsupervised ANNs have been demonstrated for time-series data of plankton communities in lakes and rivers.

Evolutionary Algorithms

Evolutionary algorithms (EAs) are adaptive methods for finding problem solutions (models, knowledge) based on principles of biological evolution by natural selection, genetic variation, and 'survival of the fittest' (see **Fig. 13**). Holland provided the theoretical framework for the development of genetic and evolutionary algorithms that are being widely used for pattern

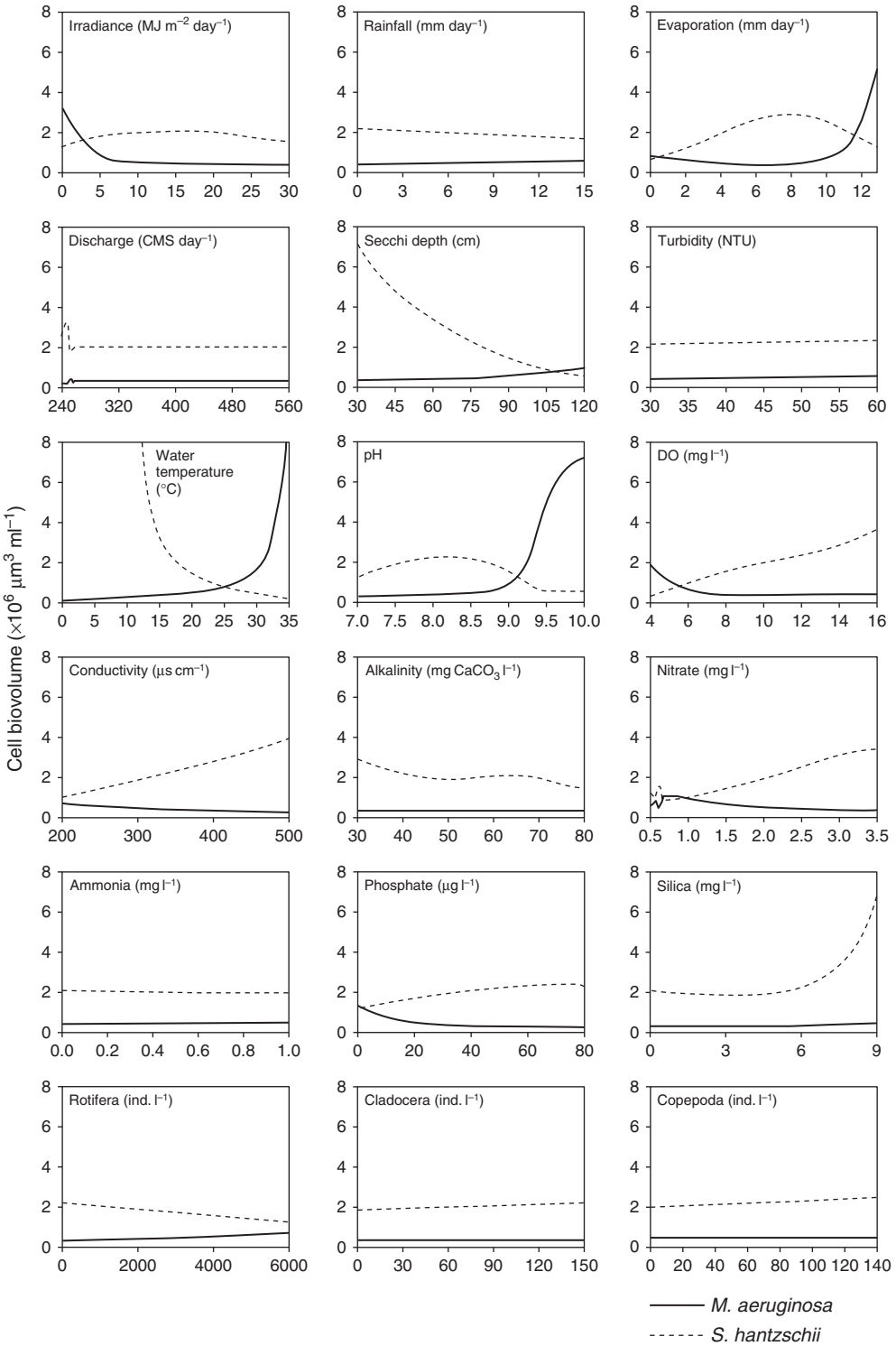


Fig. 7 Input sensitivity curves for the population densities of *M. aeruginosa* (solid lines) and *S. hantzschii* (dotted lines) in River Nakdong (South Korea) by means of a supervised feedback ANN. Modified from Jeong K-S, Recknagel F, and Joo G-J (2006) Prediction and elucidation of population dynamics of the blue-green algae *Microcystis aeruginosa* and the diatom *Stephanodiscus hantzschii* in the Nakdong River–Reservoir System (South Korea) by a recurrent artificial neural network. In: Recknagel F (ed.) *Ecological Informatics: Scope, Techniques and Applications*, 2nd edn., pp. 255–273. Berlin: Springer.

recognition, forecasting, knowledge discovery, optimum control, and parallel processing. Useful guides for history, current developments, and applications of genetic and evolutionary algorithms are provided by Goldberg, Mitchell, and Bäck *et al*.

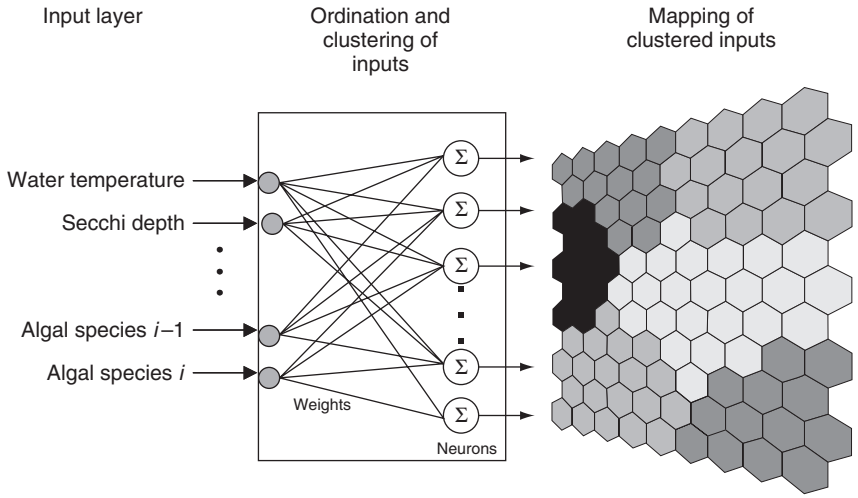


Fig. 8 Conceptual diagram of the structure and functioning of nonsupervised ANN.

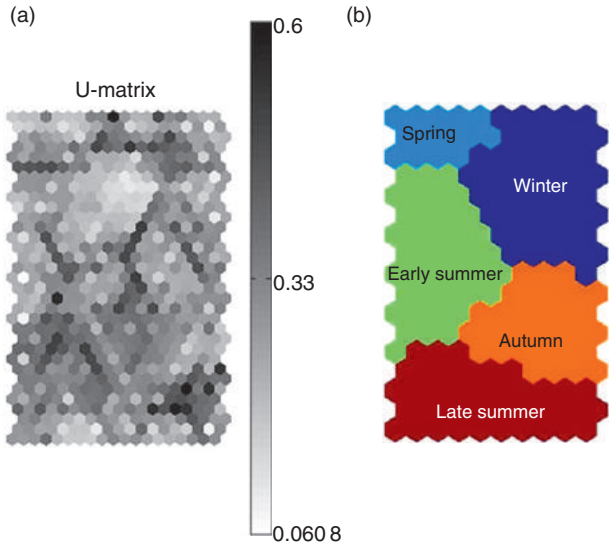


Fig. 9 Ordination and clustering of seasons of Lake Kasumigaura by means of nonsupervised ANN visualized as unified distance matrix map (U-matrix) (a), and as partitioned map (K-means) (b); the seasons were defined as follows: winter from 1 December, spring from 15 March, early summer from 1 June, late summer from 1 August, autumn from 1 October.

Successful implementations of EA as tools for solving complex economic and engineering problems have stimulated their application to solving ecological problems, which exhibit highest complexity. They allow to induce predictive models from ecological data sets similar to supervised ANN but rather than lacking an explicit model representation as typical for ANNs, EAs are distinctively designed for assembling the explicit model represented as multivariate functions or rule sets. Therefore EAs serve as powerful tools for knowledge discovery as well.

The hybrid evolutionary algorithms (HEAs) have been *ad hoc* designed as flexible tool for inducing predictive multivariate functions and rule sets from ecological time-series data. The conceptual framework of the application of HEA to rule discovery in limnological time-series data is represented in **Fig. 14**. It indicates that similar to supervised ANN, the training of HEA aims at the optimal approximation of the calculated output Y_c to the observed (desired) output Y_o. However, by contrast, HEA iteratively adjusts the rule structure and parameter values rather than input weights in order to minimize the error $(Y_o - Y_c)$.

The detailed algorithm for the rule discovery and parameter optimization by HEA is shown in **Fig. 15**. HEA uses genetic programming (GP) to generate and optimize the structure of rule sets and a genetic algorithm (GA) to optimize the parameters of a rule set. GP is an extension of GA in which the genetic population consists of computer programs of varying sizes and shapes. In standard GP, computer programs can be represented as parse trees, where a branch node represents an element from a function set (arithmetic operators, logic operators, elementary functions of at least one argument), and a leaf node represents an element from a terminal set (variables, constants, and functions of no arguments). These symbolic programs are subsequently evaluated by

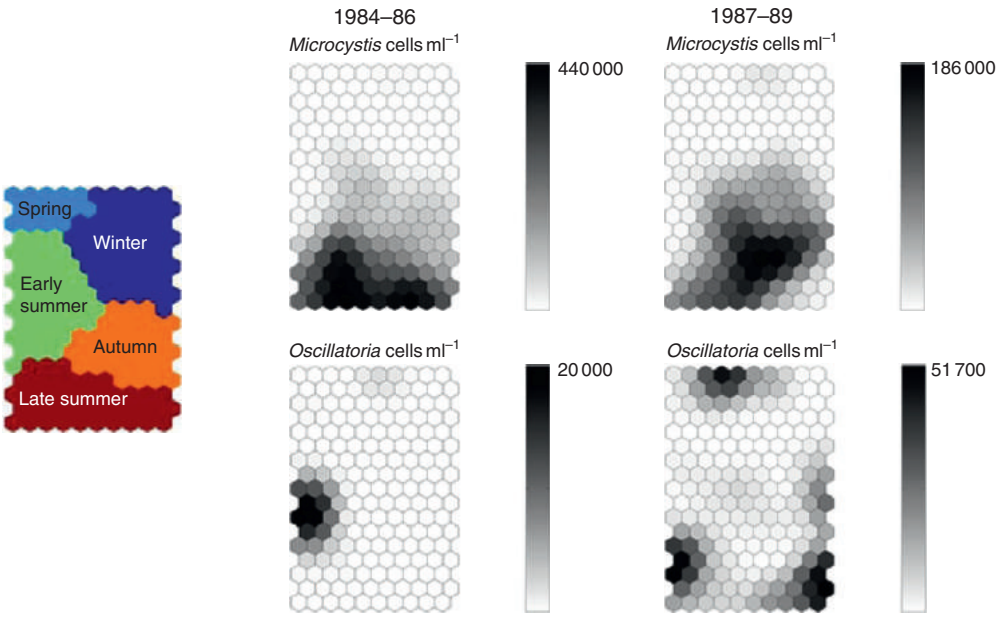


Fig. 10 Component planes for seasonal abundances of *Microcystis* and *Oscillatoria* populations in Lake Kasumigaura for the years 1984–86 (left column) and 1987–89 (right column).

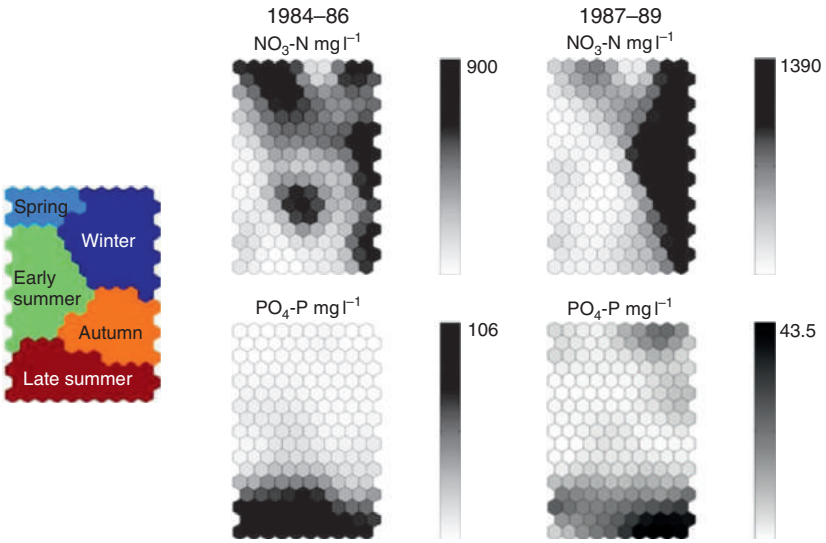


Fig. 11 Component planes for seasonal concentrations of PO_4-P and NO_3-N in Lake Kasumigaura for the years 1984–86 (left column) and 1987–89 (right column).

means of 'fitness cases'. Fitter programs are selected for recombination to create the next generation by using genetic operators, such as crossover and mutation. This step is iterated for consecutive generations until the termination criterion of the run has been satisfied. A general GA is used to optimize the random parameters in the rule set.

Figures 16 and **17** illustrate the structure, input sensitivity, and *k*-fold cross-validation of a rule-based agent for 7-day-ahead forecasting of *Microcystis* biomass developed by HEA.

The rule in **Fig. 16a** is the result of using 42 years of merged limnological data of the South African lakes Hartbeespoort, Roodeplaat, and Rietvlei for the training of HEA. The sensitivity analysis in **Fig. 16b** indicates that both water temperature and Secchi depth are key driving variables for low biovolumes of *Microcystis* of up to 14 $cm^3\ m^{-3}$ reflected by the THEN branch of the rule as well as for high biovolumes of up to 350 $cm^3\ m^{-3}$ reflected by the ELSE branch of the rule. As a result of *k*-fold cross-validation, the parameters p_1 and p_2 have been evolved to water temperature functions which provide the agent an extra mechanism for adaptation to lake-specific seasonal conditions.

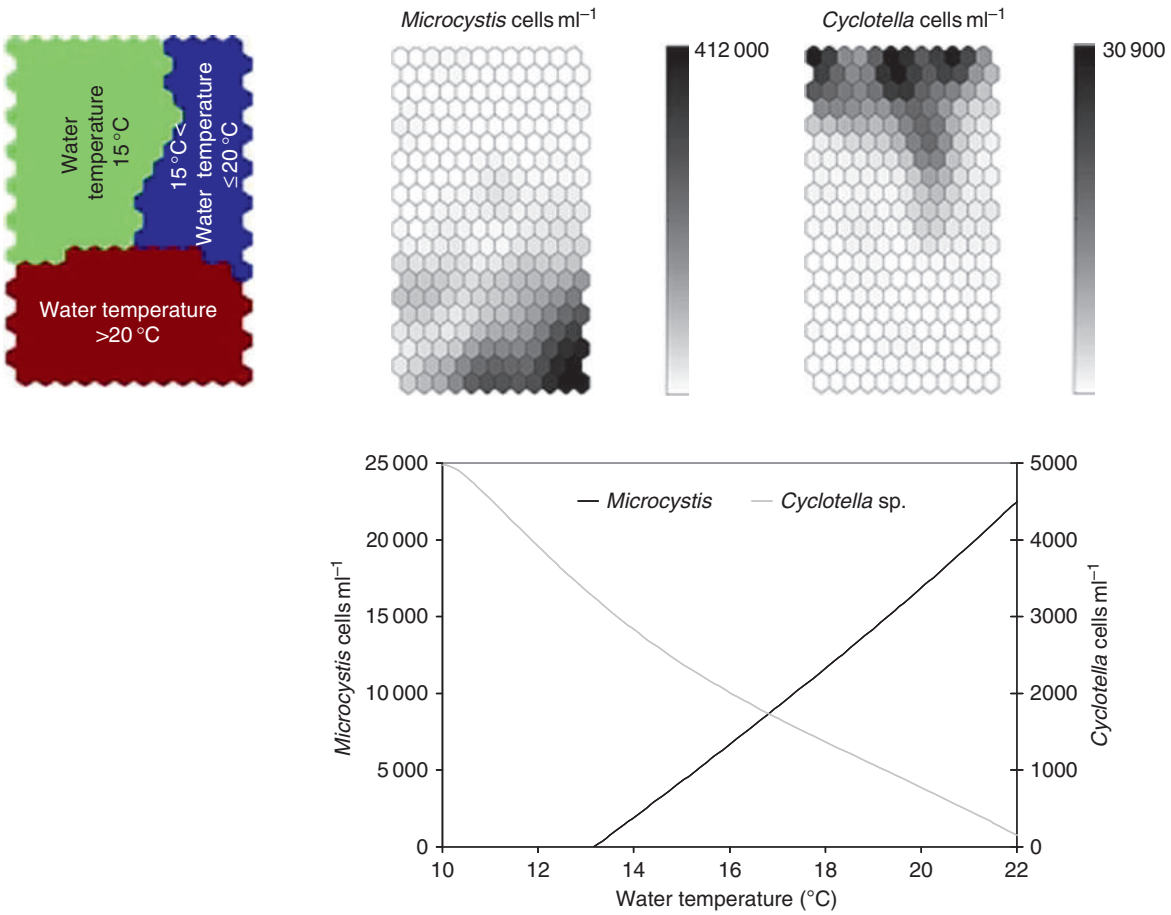


Fig. 12 Component planes for water temperature preferences of *Microcystis* and *Cyclotella* populations (top) and water temperature sensitivity curves for *Microcystis* and *Cyclotella* populations (bottom) in Lake Kasumigaura for the years 1984–93.

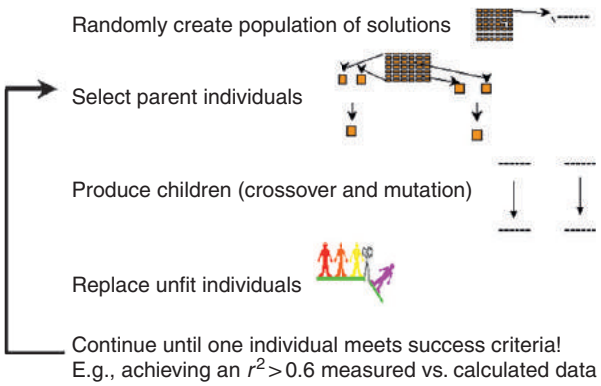


Fig. 13 Conceptual diagram for the design of evolutionary algorithms. Modified from Morrall D (2006) Ecological application of genetic algorithms. In: Recknagel F (ed.) *Ecological Informatics*, 2nd edn., pp. 69–83. New York: Springer.

The *k*-fold cross-validation of the rule-based agent for *Microcystis* achieved r^2 values of 0.31 for Lake Hartbeespoort, 0.34 for Lake Roodeplaat, and 0.75 for Lake Rietvlei (Fig. 17).

Successful applications of EA have been demonstrated for cross-sectional data of fish populations as well as macroinvertebrate communities in streams, and for time-series data of plankton communities in lakes and rivers, and biological wastewater treatment.

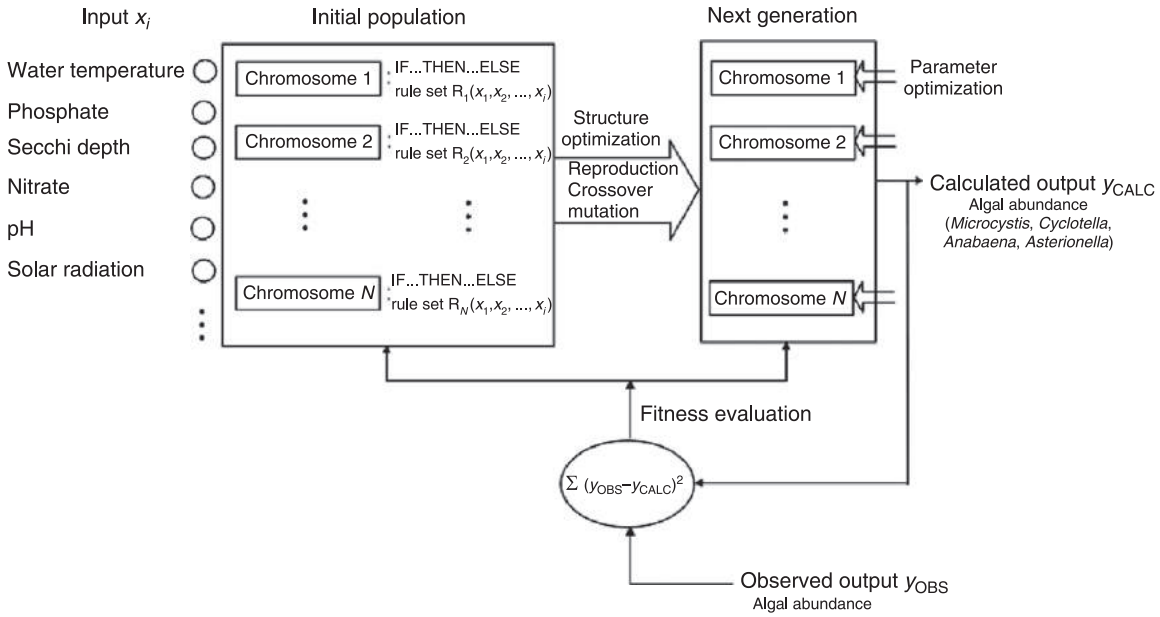


Fig. 14 Conceptual framework of the application of HEA for rule discovery in limnological time-series data.

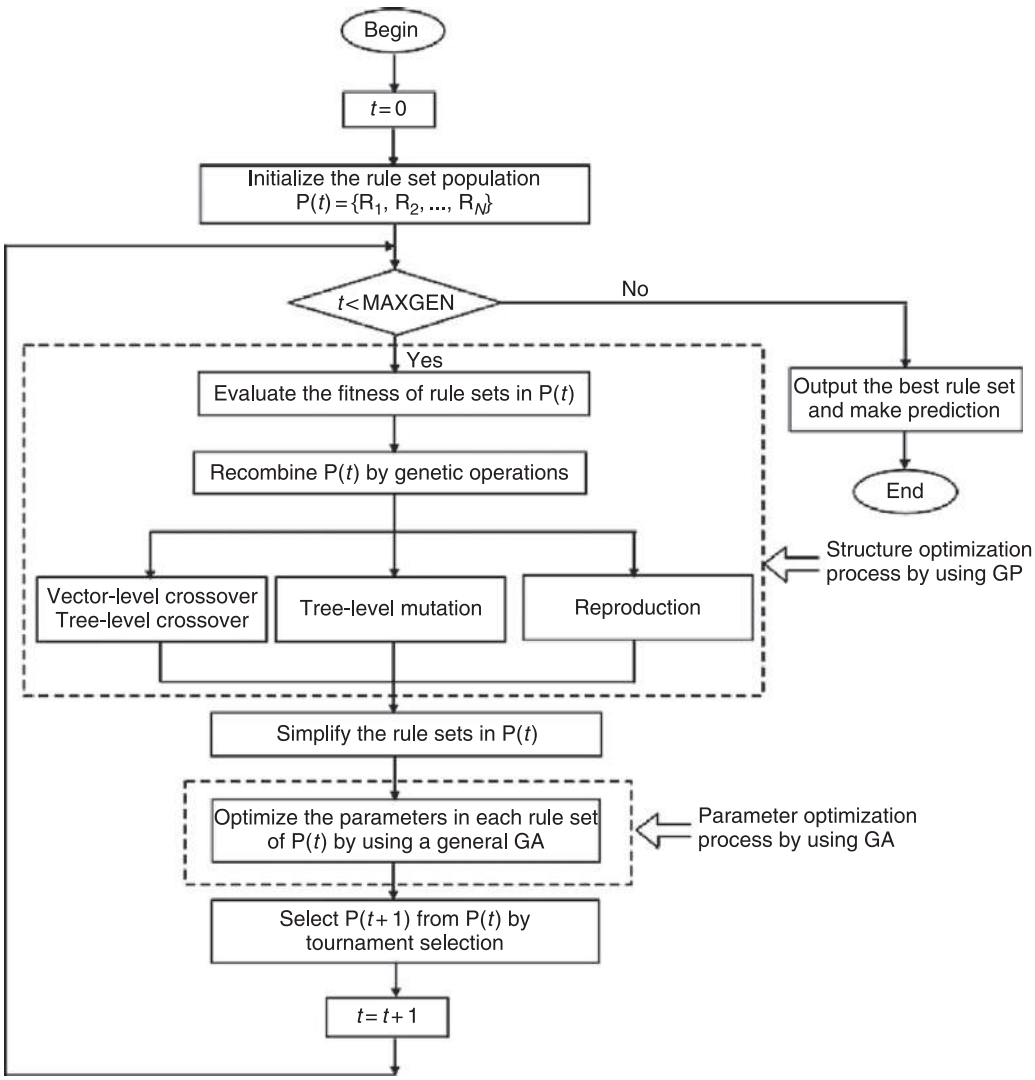


Fig. 15 Flowchart of HEA for rule discovery.

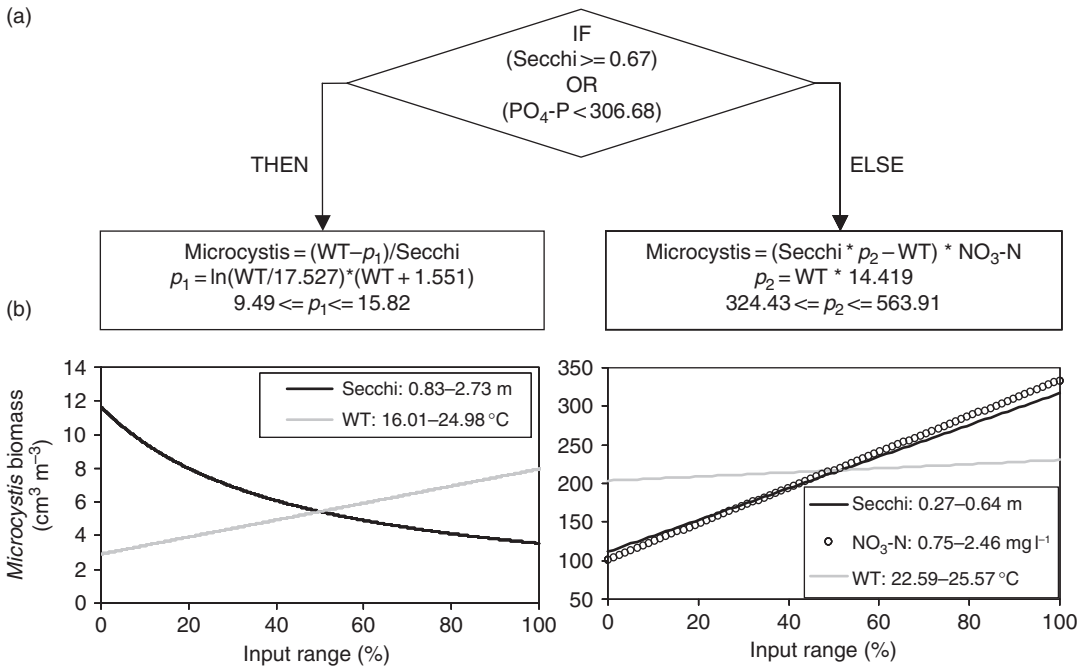


Fig. 16 Structure and input sensitivity analysis of a rule-based agent for 7-day-ahead forecasting of *Microcystis* biomass discovered in merged time-series data of the South African lakes Hartbeespoort, Roodeplaat, and Rietvlei by HEA.

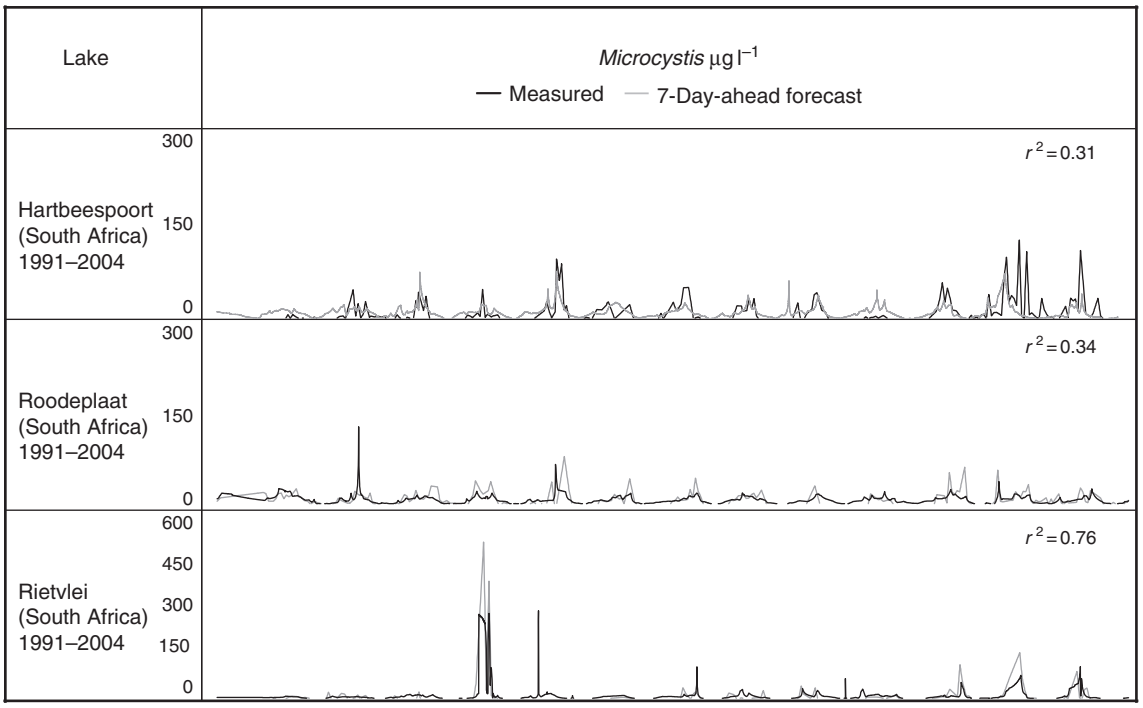


Fig. 17 *k*-Fold cross-validation of a rule-based agent for 7-day-ahead forecasting of *Microcystis* biomass by means of merged time-series data of the South African lakes Hartbeespoort, Roodeplaat, and Rietvlei.

Future Directions

Making informed decisions on preserving biodiversity and sustainable environments in spite of pollution, eutrophication, and climate change is of vital importance for the habitat Earth in the twenty-first century. Ecological informatics is challenged to

contribute ecological understanding and tools for integrating, analyzing, and synthesizing the wealth of ecological knowledge and data for making an informed decision at local, regional, and global scale.

It is anticipated that at the next stage ecological informatics will distinctively focus on: (1) integrated analysis of genomic, phenotypic, and ecological data in order to better understand biodiversity and ecosystem behavior in response to habitat and climate changes; (2) facilitating data sharing by www-based generic data warehouses tailored for ecosystem categories at global scale; and (3) implementing hybrid model libraries generic for ecosystem categories at global scale by object-oriented programming and interactive www-access.

See also: Behavioral Ecology: Learning. Ecosystems: Freshwater Marshes

Further Reading

Aoki, I., Komatsu, T., 1997. Analysis and prediction of the fluctuation of sardine abudance using a neural network. Oceanologica Acta 20 (1), 81–88.

Aoki, I., Komatsu, T., Hwang, K., 1999. Prediction of response of zooplankton biomass to climatic and oceanic changes. Ecological Modelling 120 (2–3), 261–270.

Bäck, T., Hammel, U., Schwefel, H.-P., 1997. Evolutionary computation: Comments on the history and current state. IEEE Transactions on Evolutionary Computation 1 (1), 5–16.

Banzhaf, W., Nordin, P., Keller, R.E., Francone, F.D., 1997. Genetic Programming: An Introduction on the Automatic Evolution of Computer Programs and Its Applications. San Francisco: Morgan Kaufmann.

Bobbin, J., Recknagel, F., 2003. Evolving rules for the prediction and explanation of blue-green algal succession in lakes by evolutionary computation. In: Recknagel, F. (Ed.), Ecological Informatics. Understanding Ecology by Biologically-Inspired Computation. New York: Springer, pp. 291–310.

Booth, G., 1997. Gecko: A continuous 2D world for ecological modeling. Artificial Life 3, 147–163.

Cao, H., Recknagel, F., Welk, A., Kim, B., Takamura, N., 2006. Hybrid evolutionary algorithm for rule set discovery in time-series data to forecast and explain algal population dynamics in two lakes different in morphometry and eutrophication. In: Recknagel, F. (Ed.), Ecological Informatics, 2nd edn. New York: Springer, pp. 330–342.

Cao, H., Recknagel, F., Joo, G.-J., Kim, D.-K., 2006. Rule set discovery for the prediction and explanation of chlorophyll-*a* dynamics in the Nakdong River (Korea) by means of a hybrid evolutionary algorithm. Ecological Informatics 1, 43–53.

Cao H and Recknagel F (in press) Hybridisation of process-based ecosystem models with evolutionary algorithms: Multi-objective optimisation of process and parameters representations of the lake simulation library SALMO-OO. In: Jorgensen SE, Recknagel F, and Chon TS (eds.) *Handbook of Ecological Modeling and Informatics*. Southampton, UK: WIT Press.

Capcarrere, M., Tettamanzi, A., Tomassini, M., Sipper, M., 1998. Studying parallel evolutionary algorithms: The cellular programming case. In: Eiben, A.E., Bäck, T., Schoenaver, M., Schwefel, H.P. (Eds.), Parallel Problem Solving from Nature – V. New York: Springer, pp. 573–582.

Chan, W.S., Recknagel, F., Cao, H., Park, H.D., 2007. Elucidation and short term forecasting of microcystin concentrations in Lake Suwa (Japan) by means of artificial neural networks and evolutionary algorithms. Water Research 41, 2247–2255.

Chon, T.-S., Park, Y.-S., 2006. Ecological informatics as an advanced interdisciplinary interpretation of ecosystems. Ecological Informatics 3, 213–218.

Chon, T.S., Park, Y.S., Cha, E.Y., 2000. Patterning of community changes in benthic macroinvertebrates collected from urbanized streams for the short time prediction by temporal artificial neural networks. In: Lek, S., Guegan, J.F. (Eds.), Artificial Neuronal Networks: Application to Ecology and Evolution. Berlin: Springer, pp. 99–114.

Chon, T.-S., Park, Y.S., Kwak, I.-S., Cha, E.Y., 2003. Non-linear approach to grouping, dynamics and organizational informatics of benthic macroinvertebrate communities in streams by artificial neural networks. In: Recknagel, F. (Ed.), Ecological Informatics. Scope, Techniques and Applications, 2nd edn. New York: Springer, pp. 187–238.

Chon, T.S., Park, Y.S., Moon, K.H., Cha, E.Y., 1996. Patternizing communities by using an artificial neural network. Ecological Modelling 90, 69–78.

D'Angelo, D.J., Howard, L.M., Meyer, J.L., Gregory, S.V., Ashkenas, L.R., 1995. Ecological uses of genetic algorithms: Predicting fish distributions in complex physical habitats. Canadian Journal of Fisheries and Aquatic Sciences 52, 1893–1908.

Dolk, D.R., 2000. Integrated model management in the data warehouse area. European Journal of Operational Research 122, 1999–2218.

Doney, S.C., Abbott, M.R., Cullen, J.J., Karl, D.M., Rothstein, L., 2004. From genes to ecosystems: The ocean's new frontier. Frontiers in Ecology and the Environment 2 (9), 457–466.

Downing, K., 1997. EUZONE: Simulating the evolution of aquatic ecosystems. Artificial Life 3, 307–333.

Eleveld, M.A., Schrimpf, W.B.H., Siegert, A.G., 2003. User requirements and information definition for the virtual coastal and marine data warehouse. Ocean & Coastal Management 46, 487–505.

Fielding, A., 1999. Machine Learning Methods for Ecological Applications. Amsterdam: Kluwer, 262pp.

Fischer, J.M., Klug, J.L., Ives, A.R., Frost, T.M., 2001. Ecological history affects zooplankton community responses to acidification. Ecology 82 (11), 2984–3000.

Foody, G., 2000. Soft mapping of coastal vegetation from remotely sensed imagery with a feed-forward neural network. In: Lek, S., Guegan, J.F. (Eds.), Artificial Neuronal Networks: Application to Ecology and Evolution. Berlin: Springer, pp. 45–56.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Reading, MA: Addison Wesley.

Goonatilake, S., Khebbal, S., 1995. Intelligent Hybrid Systems. New York: Wiley.

Grimm, V., Railsback, S.F., 2005. Individual-Based Modelling and Ecology. Princeton, NJ: Princeton University Press.

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., *et al.*, 2005. Pattern-oriented modelling of agent based complex systems: Lessons from ecology. Science 310, 987–991.

Gyllström, M., Hansson, L.-A., Jeppesen, E., *et al.*, 2005. The role of climate in shaping zooplankton communities of shallow lakes. Limnology and Oceanography 50, 2008–2021.

Hairston, N.G., Lampert, W., Caceres, C.E., *et al.*, 1999. Rapid evolution revealed by dormant egg. Nature 401, 446.

Handelsman, J., 2004. Metagenomics: Application of genomics to uncultured microorganisms. Microbiology and Molecular Biology Reviews 68 (4), 669–685.

Henikoff, S., Henikoff, J.G., Pietrovski, S., 1999. Blocks+: A non-redundant database of protein alignment blocks derived from multiple compilations. Bioinformatics 15, 471–479.

Holland, J.H., 1975. Adaptation in Natural and Artificial Systems. Ann Arbor, MI: University of Michigan Press.

Hong, Y.-S., Bhamidimarri, R., 2003. Evolutionary self-organising modeling of a municipal wastewater treatment plant. Water Research 37, 1199–1212.

Hongping, P., Jianyi, M., 2002. Study on the algal dynamic model for West Lake, Hangzhou. Ecological Modelling 148, 67–77.

Hornik, K., Stinchcombe, M., White, A., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2 (5), 359–366.

Horrigan, N., Bobbin, J., Recknagel, F., Metzling, L., 2005. Patterning, prediction and explanation of stream macroinvertebrate assemblages in Victoria (Australia) by means of artificial neural networks and genetic algorithms. In: Lek, S., Scardi, M., Verdonschot, P.F.M., Descy, J.-P., Park, Y.-S. (Eds.), Modelling Community Structures in Freshwater Ecosystems. Berlin: Springer, pp. 252–260.

Hraber, P., Milne, B.T., 1997. Community assembly in a model ecosystem. Ecological Modelling 103, 267–285.

Huang, W., Foo, S., 2002. Neural network modelling of salinity variation in Apalachicola River. Water Research 36, 356–362.
Huong, H., Recknagel, F., Marshall, J., Choy, S., 2001. Predictive modelling of macroinvertebrate assemblages for stream habitat assessments in Queensland (Australia). Ecological Modelling 146 (1–3), 195–206.
Huong, H., Recknagel, F., Marshall, J., Choy, S., 2003. Elucidation of hypothetical relationships between habitat conditions and macroinvertebrate assemblages in freshwater streams by artificial neural networks. In: Recknagel, F. (Ed.), Ecological Informatics: Understanding Ecology by Biologically-Inspired Computation. Berlin: Springer, pp. 179–190.
Huse, G., Strand, E., Giske, J., 1999. Implementing behaviour in individual based models using neural networks and genetic algorithms. Evolutionary Ecology 13, 469–483.
Jeong, K.-S., Joo, G.-J., Kim, H.-W., Ha, K., Recknagel, F., 2001. Prediction and elucidation of phytoplankton dynamics in the Nakdong River (Korea) by means of a recurrent artificial neural network. Ecological Modelling 146 (1), 115–130.
Jeong, K.-S., Recknagel, F., Joo, G.-J., 2006. Prediction and elucidation of population dynamics of the blue-green algae *Microcystis aeruginosa* and the diatom *Stephanodiscus hantzschii* in the Nakdong River–Reservoir System (South Korea) by a recurrent artificial neural network. In: Recknagel, F. (Ed.), Ecological Informatics: Scope, Techniques and Applications, 2nd edn. Berlin: Springer, pp. 255–273.
Jorgensen, S.E., 1995. Fundamentals of Ecological Modelling. Amsterdam: Elsevier, 628pp.
Karul, C., Soyupak, S., 2003. A comparison between neural network based and multiple regression models for chlorophyll-*a* estimation. In: Recknagel, F. (Ed.), Ecological Informatics: Understanding Ecology by Biologically-Inspired Computation. Heidelberg: Springer, pp. 249–264.
Kohonen, T., 1989. Self-Organization and Associative Memory. Berlin: Springer.
Lek, S., Delacosta, M., Baran, P., *et al.*, 1996. Application of neural networks to modeling nonlinear relationships in ecology. Ecological Modelling 90, 39–52.
Lek, S., Guegan, J.F. (Eds.), 2000. Artificial Neuronal Networks: Application to Ecology and Evolution. Berlin: Springer, p. 262.
Lek, S., Scardi, M., Verdonschot, P.F.M., Descy, J.-P., Park, Y.-S. (Eds.), 2005. Modelling Community Structure in Freshwater Ecosystems. New York: Springer.
Lockhardt, D., Winzeler, E., 2000. Genomics, gene expression and DNA arrays. Nature 405, 827–836.
Lupas, A., Van Dyke, M., Stock, J., 1991. Predicting coiled coils from protein sequences. Science 252, 1162–1164.
Lusk, J.J., Guthery, F.S., DeMaso, S.J., 2001. Northern bobwhite (*Colinus irginianus*) abundance in relation to yearly weather and long-term climate patterns. Ecological Modelling 146, 3–15.
Maier, H., Dandy, G., Burch, M., 1998. Use of artificial neural networks for modelling cyanobacteria *Anabaena* spp. in the River Murray, South Australia. Ecological Modelling 146 (1–3), 85–96.
Michener, W.K., 2006. Meta-information concepts for ecological data management. Ecological Informatics 1, 3–7.
Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B., Stanford, S.G., 1997. Nongeospatial metadata for the ecological sciences. Ecological Applications 7 (1), 330–342.
Minski, M.L., Pappert, S., 1969. Perceptrons. Cambridge: MIT Press.
Mitchell, M., 1996. An Introduction to Genetic Algorithms. Cambridge: MIT Press.
Morrall, D., 2006. Ecological applications of genetic algorithms. In: Recknagel, F. (Ed.), Ecological Informatics, 2nd edn. New York: Springer, pp. 69–83.
Mulderij, G., Smolders, A.J.P., van Donk, E., 2006. Allelopathic effect of the aquatic macrophyte, *Stratiotes aloides*, on natural phytoplankton. Freshwater Biology 51, 554–561.
Muttil, N., Lee, J.H.W., 2005. Genetic programming for analysis and real-time prediction of coastal algal blooms. Ecological Modelling 189 (3–4), 363–376.
Oltvai, Z.N., Barabasi, A.L., 2002. Life's complexity pyramid. Science 298, 763–764.
Overbeck, R., Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N., 1999. The use of gene clusters to infer functional coupling. Proceedings of the National Academy of Sciences of the United States of America 96, 2896–2901.
Park, Y.-S., Verdonschot, P.F.M., Chon, T.-s., Lek, S., 2003. Patterning and predicting aquatic macroinvertebrate diversities using artificial neural networks. Water Research 37, 1749–1758.
Pineda, F.J., 1987. Generalisation of back-propagation to recurrent neural networks. Physical Review Letters 59 (19), 2229–2232.
Poff, N.L., Tokar, S., Johnson, P., 1996. Stream hydrological and ecological response to climate change assessed with an artificial neural network. Limnology and Oceanography 41 (5), 857–863.
Recknagel, F., 1997. ANNA – Artificial neural network model predicting species abundance and succession of blue-green algae. Hydrobiologia 349, 47–57.
Recknagel, F. (Ed.), 2003. Ecological Informatics: Understanding Ecology by Biologically-Inspired Computation. Berlin: Springer.
Recknagel, F. (Ed.), 2006. Ecological Informatics: Scope, Techniques and Applications, 2nd edn. New York: Springer.
Recknagel, F., Bobbin, J., Whigham, P., Wilson, H., 2002. Comparative application of artificial neural networks and genetic algorithms for multivariate time-series modelling of algal blooms in freshwater lakes. Journal of Hydroinformatics 4 (2), 125–134.
Recknagel, F., French, M., Harkonen, P., Yabunaka, K., 1997. Artificial neural network approach for modelling and prediction of algal blooms. Ecological Modelling 96 (1–3), 11–28.
Recknagel, F., Kim, B., Takamura, N., Welk, A., 2006. Unravelling and forecasting algal population dynamics in two lakes different in morphometry and eutrophication by neural and evolutionary computation. Ecological Informatics 1 (2), 133–151.
Recknagel, F., Kim, B., Welk, A., 2006. Unravelling ecosystem behaviour of Lake Soyang (South Korea) in response to climate and management by means of artificial neural networks. Internationale Vereinigung für Theoretische und Angewandte Limnologie 29 (3), 1497–1502.
Recknagel, F., Talib, A., van der Molen, D., 2006. Phytoplankton community dynamics of two adjacent Dutch lakes in response to seasons and eutrophication control unravelled by non-supervised artificial neural networks. Ecological Informatics 1 (3), 277–286.
Recknagel, F., Wilson, H., 2000. Elucidation and prediction of aquatic ecosystems by artificial neural networks. In: Lek, S., Guegan, J.F. (Eds.), Artificial Neural Networks in Ecology and Evolution. New York: Springer, pp. 143–155.
Reick, C.H., Grünewald, A., Page, B., 2003. Multivariate time series prediction of marine zooplankton by artificial neural networks. In: Recknagel, F. (Ed.), Ecological Informatics: Scope, Techniques and Applications, 2nd edn. Heidelberg: Springer, pp. 369–383.
Reynolds, C.S., 1984. The Ecology of Freshwater Phytoplankton. New York: Cambridge University Press, 384pp.
Rummelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagation errors. Nature 323, 533–536.
Scardi, M., 1996. Artificial neural networks as empirical models for estimating phytoplankton production. Marine Ecology Progress Series 139, 289–299.
Schleiter, I.M., Borchartdt, D., Wagner, R., *et al.*, 1999. Modelling water quality, bioindication and population dynamics in lotic ecosystems using neural networks. Ecological Modelling 120, 271–286.
Sen, A., 2003. Metadata management: Past, present and future. Decision Support Systems 1043, 1–23.
Shapiro, J., 1984. Blue-green dominance in lakes: The role and management significance of pH and CO_2. Internationale Revue der Gesamten Hydrobiologie 69, 765–780.
Straskraba, M., Gnauck, A., 1985. Freshwater Ecosystems: Modelling and Simulation. Amsterdam: Elsevier, 302pp.
Takamura, N., Otsuki, A., Aizaki, M., Nojiri, Y., 1992. Phytoplankton species shift accompanied by transition from nitrogen dependence to phosphorus dependence of primary production in Lake Kasumigaura, Japan. Archive Hydrobiology 124, 129–148.
Tringe, S.G., von Mering, C., Kobayashi, A., *et al.*, 2005. Comparative metagenomics of microbial communities. Science 308, 554–557.
Van Donk, E., 2006. Food web interactions in lakes: What is the impact of chemical information conveyance? In: Dicke, M., Takken, W. (Eds.), Chemical Ecology: From Gene to Ecosystem. Berlin: Springer, pp. 145–160.
Van Donk, E., 2007. Chemical information transfer in freshwater plankton. Ecological Informatics 2, 112–120.
Van Ginkel, C.E., Silberbauer, M.J., Du Plessis, S., Carelsen, C.I.C., 2006. Monitoring microcystin taxin and chlorophyll in five South African impoundments. Internationale Vereinigung für Theoretische und Angewandte Limnologie 29, 1611–1616.

Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J., (2000) SOM Toolbox for MATLAB 5. Helsinki University of Technology, Finland.
Vijverberg, J., Doksæter, A., van Donk, E., 2006. Contrasting life history responses to fish released infochemicals of two co-occurring *Daphnia* species that show different migration behaviour. Archives of Hydrobiology 167, 89–100.
Voss, M., Vet, L.M., Wäckers, F.L., *et al*., 2006. Infochemicals structure marine, terrestrial and freshwater food webs: Implications for ecological informatics. Ecological Informatics 1, 23–32.
Walley, W.J., Fontana, V.N., 1998. Neural network predictors of average score per taxon and number of families at unpolluted river sites in Great Britain. Water Research 32 (3), 613–622.
Walter, M., Recknagel, F., Carpenter, C., Bormans, M., 2001. Predicting eutrophication effects in the Burrinjuck Reservoir (Australia) by means of the deterministic model SALMO and the recurrent neural network model ANNA. Ecological Modelling 146 (1–3), 97–114.
Wei, B., Sugiura, N., Maekawa, T., 2001. Use of artificial neural network in the prediction of algal blooms. Water Research 35 (8), 2022–2028.
Weiss, M., Kulikowski, C., 1990. Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning and Expert Systems. San Francisco: Morgan Kaufmann.
West, K., Cohen, A., Baron, M., 1991. Morphology and behaviour of crabs and gastropods from Lake Tanganyika, Africa: Implications for lacustrine predator–prey coevolution. Evolution 45 (3), 589–607.
Wolf, Y.I., Rogozin, I.B., Kondrashov, A.S., Koonin, E.V., 2001. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Research 11, 356–372.
Yabunaka, K.-I., Hosomi, M., Murakami, A., 1997. Novel application of a backpropagation artificial neural network model formulated to predict algal bloom. Water Science and Technology 36 (5), 89–97.
Yao, X., Liu, Y., Li, J., He, J., Frayn, C., 2006. Current developments and future directions of bio-inspired computation and implication for ecoinformatics. Ecological Informatics 1, 9–22.

Relevant Websites

http://knb.ecoinformatics.org
Ecological Metadata Language (EML), The Knowledge Network for Biocomplexity.

Emergent Properties

F Müller, University of Kiel, Kiel, Germany
SN Nielsen, Danmarks Farmaceutiske Universitet, Copenhagen, Denmark

Introduction

Many biologists will recognize the statement that "the whole is more than the sum of the parts" as a commonly used (but hardly understood) phrase. This formulation refers to the idea that there are systems which possess additional qualities or quantities, beyond easily measurable or predictable physical parameters. The resulting properties have been described on many levels of the biological hierarchy, from simple physical systems, like laser beams, to the organization of the whole biosphere within the Gaia concept of J. Lovelock. The emergent property at one level is in general finding its causality at the subsystem components and the interaction between them. For example, an organized form of cell functions, stemming from self-organized transformations of cellular compounds, known as hypercycling may be considered as an emergent entity; at the physiological levels we are, for example, dealing with the mating behaviors of organisms as results of hormone interactions. Similarly, motion, feelings, or intelligent behavior occur as a consequence of special couplings of neurons. In addition, the patterns in the development of ecosystems may not be predictable from knowledge of organisms alone. Therefore, emergent properties are not unusual phenomena, but simply consequences of hierarchical organizations.

The concept of emergence found its way into ecology through the proposal of E. P. Odum, who suggested that the study of emergence should lead to a 'new integrative discipline'. This idea was due to the fact that studies of complex systems had shown that the investigations of the details alone were not adequate in predicting ecosystem function and behavior. Neither were they sufficient to explain a more advanced pattern like behavior and performance of ecosystems, for example, during the succession from young systems toward more mature states.

The History of the Concept

The concept of emergent properties originates in the nineteenth century, finding the primary roots back in Kantian philosophy. The term was coined by G. H. Lewis as far back as 1875. A common definition from that time states that "emergence is the denomination of something new which could not be predicted from the elements constituting the preceding condition."

Throughout the last century, several scientists have addressed the concept from a more philosophical point of view, resulting in the appearance of different descriptions and explanations. The definitions have, in general, been referring to subjective arguments, such as surprise, unexpectancy, thus being clearly observer dependent. This has strongly influenced the present approaches, and this comprehension has often been connected with a flavor of mysticism. Thus, the seriousness of the concept has often been underestimated.

During the last decades, the use of the term emergent properties has found widespread use in biological sciences, especially, because it is clearly connected with the growing implementation of the system approach in ecology. The need for a holistic concept was due to the failure of the traditional reductionistic research strategies to explain the properties of ecosystems by the knowledge of the behavior and the properties of the ecosystem constituents alone. Ecosystems are highly complex middle-numbered systems dominated by nonlinear relationships between their constituents. In such systems, things are bound to happen that are not easy to predict from the basic knowledge of the system, no matter how extensive this knowledge is.

Highly relevant to biology and ecology is the question when an emergent property appears. This leads to the distinction of 'primary' and 'secondary' emergence, primary emergence being the first time an emergent property appears. To be conserved the property can be reproduced again and again but in this case it is nominated as a secondary emergence. Recent approaches to emergence have come up with three further notions of emergence: 'computational emergence', 'thermodynamic emergence', and 'emergence relative to a model'. The computational emergence deals with the patterns produced by different computer programs, for example, cellular automata systems developing complex distributions out of simple rules from game theory. Thermodynamic emergence covers the establishment of highly complex, self-organized structures and their relations to the nonlinear, far from equilibrium thermodynamics. Emergence relative to a model defines emergence as the deviation of the actual behavior of a physical system in comparison with an observer's model of it.

Summarizing these historical notions of emergence, the following features can be stated:

- Emergent properties are properties of a system which are not possessed by component subsystems alone.
- The properties emerge as a consequence of the interactions within the system.
- Two fundamental types of interactions are found that may be characterized as intra- and inter-connectedness, that is, connections within and between levels, including controls. This point does not consider the direction of the intra-level interactions. Emergence is based on both, upward and downward causation.

- The historically emerged properties are considered 'new' with reference to their primary appearance.
- These new properties appear at one level of a system and are not immediately deducible from observation of the levels or units of which the system consists.

Emergence and Hierarchy

Emergence has been described at many levels of the biological hierarchy. As argued above, the reason for emergence is to be found in the hierarchical organization of the system and the quantitative and qualitative characters of the 'linkages' within the structure. As biological structures are often complex, this makes it hard to determine the actual cause of emergence.

Hierarchy theory states that middle-numbered systems – such as ecosystems – can be comprehended if they are investigated on different levels of integration. Broad scale levels can be assigned to high spatial extents, and low typical frequencies, filtering the signals from lower levels. These scale levels are spatially smaller and their typical frequencies are higher. They are not able to filter constraints from the higher levels, but their potentials and interactions are building up the material basis and the coordination functions of the higher level. Emergent properties are created by both types of nonlinear interactions. Therefore, the properties of specific levels can be termed 'hierarchical emergent properties'. Of course, the interesting question is how these properties emerge. Some examples might be helpful to illuminate this question.

Prebiological Emergence

Several examples of emergent properties can be found in physical and chemical sciences. They form an important prerequisite for protobiology and evolutionary processes. Within the area of physics some examples are nearly classical: Water (e.g., its wetness), which is a simple molecule with a rather complex behavior, that is unpredictable from knowledge about oxygen and hydrogen alone, has often been used to demonstrate emergent properties. Similarly, the sense of colors by the eyes is not predictable by knowing a certain wavelength of light.

Two famous chemical examples related to self-organized behavior of systems may also be mentioned, the Bénard cells and the Bhelusov–Zabotinsky (BZ) reaction. In the case of Bénard cells, during specific conditions, hexagonal, convective cells (the emerging structures) form a fluid when a thermal gradient is imposed on the experimental setup containment. In the BZ-reaction a special ratio of chemicals causes a mixture to perform a pulsing pattern in colors with a period of about one minute. The structure of these physico-chemical processes, gradients resulting in convective cells, pulsing patterns, together with other observations like the occurrence of Turing structures in chemical fluids, spontaneous formation of lipid coacervates, might be crucial to our understanding of the emergence of life.

Protobiological Emergence

The appearance of the earliest life forms has often been referred to as a primary emergence. Although many of the properties occurring during this phase of evolution have been repeated, over and over again their appearance still qualifies them as emergent properties. As examples, the emergence of life, emergence of animals, or the emergence of bird feathers from reptile scales can be mentioned to characterize situations of primary emergence.

Many examples found in the literature deal with the formation of the earliest cells. Biochemical cycles, the organization and exchange of information by DNA or RNA and the compartmentalization of material within membranes are but a few examples. Molecular complementarity, defined as "nonrandom, reversible coupling of the components of a system," has been argued to be a widespread mechanism in biological systems and important for the understanding of the processes lying behind emergent properties. The seemingly (self)organization of molecules observed in prebiotic systems, such as Turing structures and auto-catalytic hypercycles, can be seen as emergent properties already at a very low level of organization.

Emergence in Biological Systems

Emergent properties really come into play when biological systems reach higher levels of complexity. This becomes evident already when cells or groups of cells communicate with each other as in the case of hormones and natural neural networks. Organs are composed of cells, their individual functions are important only to the organism as a whole. A heart, kidneys or lungs, are vital but their function is not existent when they are on their own. Organisms interacting as populations or societies provide properties which cannot be explained by properties of the individual organisms alone. They all go together in what we consider as ecosystems and thus are a part of the biosphere.

Cellular level. In regarding the outcome of interacting cells many studies have been concentrating on the organization of neuronal systems, which result in unexpected properties like the ability to move, to sense, to be intelligent, and to emote. The sensory systems, being connected to visual, auditory, or other communicative processes are all playing a major role in how successful living organisms are in performing specific life strategies. Reliable senses, and responding the right way to the received

stimuli are crucial to the existence of many life forms, in processes like finding food, knowing when and where to escape, or creating bonds with other members of the species, for example, during reproduction.

Neural networks, like in our brains, consisting of a huge number of interconnected neurons, are so complex that unforeseen patterns in responses are bound to occur and have also been reported to exist. During the evolution of the brain, emergent properties, together with new cell types, local and large circuits have added up to the increasing complexity of brain function. Motor control, the control and coordination of motor activity are taken care of by our brain passing on signals to the limbs or organs involved.

Organ level. Numerous cells, often during morphogenesis differentiated in certain, specialized directions, form organs, take up a particular task of the organism, like for instance liver cells secreting enzymes, or kidney cells filtering and cleaning the coelom. Although the formal 'layout' for this functionality is existent in the genetic material of all cells, the eventual determination occurs during the development of the organism and the actual function of the organs may be viewed as emergent. The brain as an organ may serve as an example of this emergence: Here differentiated cells, with highly specialized physiological properties, go together and create activity patterns that are far more complex than expected from knowing the physiology of neural cells alone. The whole becomes more than the sum of the parts.

Organism level. Complex behavior occurs among the individual organisms that cannot be determined exclusively by internal factors. The sending, reception and interpretation of signals from interagent organisms, the relationship(s) to the outside, and thus semiotics play an important role, creating patterns impossible to foresee if only the subsystems are known. For example, in trees, the formation of branches and leaf mosaics have been studied in a number of recent investigations with modeling approaches as well as the allocation of resources between above- and belowground biomass and the related physiological mechanisms. A modeling study of this problem indicates a 'complex integrated growth pattern' which may only be understood as an emergent property as it is claimed to have no direct or indirect mechanistic basis related to subcellular activities. In a similar manner it was shown that whole-plant behavior is an emergent property arising from a rule-based model of the system. Communications between individuals, that is, their social interactions within a population, are important to the function of the organism as a whole and are indistinguishable from the emergence of ethological features. Stressing the importance of communication, may lead to an interpretation of the communicative process as an emergent interpretation of signs, which is described within the discipline of semiotics.

Population level. Populations are composed of individual organisms, interacting in various ways, differing in quantity and quality, throughout the biological system. The interactions may vary in character according to the complexity. At the one end of the spectrum, we find the single cell organisms interacting mostly on a material basis (matter fluxes). At the other end, there are colonial organisms forming complex societies, where brains, senses, memory, and thus informational interaction become dominating. Emergent properties as a result of individual level behavior and interactions in populations of social insects have been argued in several studies. For instance, the distribution of food to larvae of the fire ant has been argued as emerging from interactions between individuals, workers, and larvae. Cellular automata models were used to study the short time oscillations in ant colonies. The nonlinear dependencies describing the relationships between, and the movement of, individuals explain this behavior. The resulting oscillations were found to be emergent properties of the colony.

Ecosystem level. Ecosystems are inherently complex as they are composed of an embedded hierarchy of all the previously mentioned subsystems in close interaction with abiotic factors. Emergence is to be expected, but surprisingly few reports exist at this level, before all analyzing microcosms, forest ecosystems, predator–prey relationships, food webs, and the organization of aquatic communities.

Ecosystem behavior is often analyzed through modeling studies. The relation to emergent properties becomes clear when looking at recent efforts of structural dynamic modeling, where the changes in ecosystem composition and structure over time are analyzed. Another example is the work of B. C. Patten on the propagation of matter–energy through the ecosystem network, leading to the discovery of the importance of 'indirect effects, quantitative and qualitative utilities' of the system, results that are highly surprising and unexpected, and as such are emergent properties. Both examples link to higher-level information expressions such as ascendency, different kinds of entropy or information derived descriptors like exergy.

The ability of the ecosystem to perform with systematic directional changes in some macroscopic characters, not predictable from knowledge about the single ecosystem members alone, has been discussed since 1967 on the basis of the 24 principles of ecosystem development during succession in the second edition of E. P. Odum's *Fundamentals of Ecology*. Many other factors, known as indicators, orientors, or goal functions have been presented since then (**Table 1**).

How Emergence Emerges

The concept of emergent properties refers very clearly to, and must be seen in tight connection with, at least two other concepts often occurring in literature on modern ecosystem theory, the concepts of hierarchy and self-organization. In connection with hierarchy, the emergent properties are seen as outcomes of ecosystem organization where supersystems are formed with subsystems as constituents and where the properties are observable at the supersystem level only. Here the emergent property is an outcome of a certain way of organization. To exemplify this point, we might look at the following hierarchical features:

1. Individual level: individual nutrition budgets – foraging strategies.

Table 1 Some ecosystem orientors

Immature state	*Mature state*
Properties of the dominating species	
Rapid growth	Slow growth
R-selection	K-selection
Quantitative growth	Qualitative development
Small size	Large size
Short life spans	Long life spans
Broad niches	Narrow niches
Properties of production	
Small biomass	Large biomass
High P/B ration	Low P/B ratio
Low respiration	High respiration
Small gross production	Medium gross production
Properties of nutrient flows and cycles	
Simple, rapid, and leaky	Complex, slow, and closed cycles
Small storage	Large storage
Extrabiotic	Intrabiotic nutrient distribution scheme
Small amounts of detritus	Large amounts of detritus
Rapid nutrient exchange	Slow nutrient exchange
Short residence times	Long residence times
Minor chemical heterogeneity	High chemical heterogeneity
Loose network articulation	High network articulation
Low diversity of flows	High diversity of flows
Undeveloped symbiosis	Developed symbiosis
Properties of the community	
Low diversity	High diversity
Poor feedback control	Developed feedback control
Poor spatial patterns	Developed spatial patterns
Thermodynamic and integrative system properties	
Poor hierarchical structure	Developed hierarchical structure
Close to equilibrium	Far from equilibrium
Low exergy storage	High exergy storage
Small total entropy production	High total entropy production
High specific entropy production	Small specific entropy production
Small level of information	High level of information
Small internal redundancy	High internal redundancy
Small path lengths	High path lengths
Low ascendency	High ascendency
Poor indirect effects	Developed indirect effects
Small respiration and evapo-transpiration	High respiration and evapo-transpiration
Small energy demand for maintenance	High energy demand for maintenance

The features, that are optimized throughout natural successions, provide several characteristics of emergent properties: They are only observable at the ecosystem level (which is the typical and the lowest logical level to describe, e.g., cycling phenomena), and they are based on self-organized processes. They can not be explained on the basis of knowledge of the parts alone, and the emergence-creating processual linkages between the sub systems are non-linear processes. From the hierarchy-based viewpoint also the additive features (e.g. size, biomass, life spans) can be categorized as emergent properties because their extensions are dependent on the scale of observation and because they also are based on internal system interrelations.

2. Population level: species nutrition efficiencies – intraspecific food competition.
3. Ecosystem level: nutrient cycling – food webs.
4. Landscape level: lateral nutrient transfers – food webs including large scale predators.

On the other hand, the ability of biological systems to arrange themselves in a special manner, for example, in a hierarchical way, is in itself a property which emerges as a consequence of the properties of its constituents, but the organization and the function for sure cannot always be foreseen. Thus, the capability of self-organization can be seen as an emergent property itself (**Fig. 1**).

The existence of emergent properties is based on the system's organization (built up by structures and functions) whereby the interrelations (energy, matter, water and information flows, communications) play an important role. Some conditions of the system's state add up to the increased chances that emergent properties will appear. For example, instabilities seem to be important

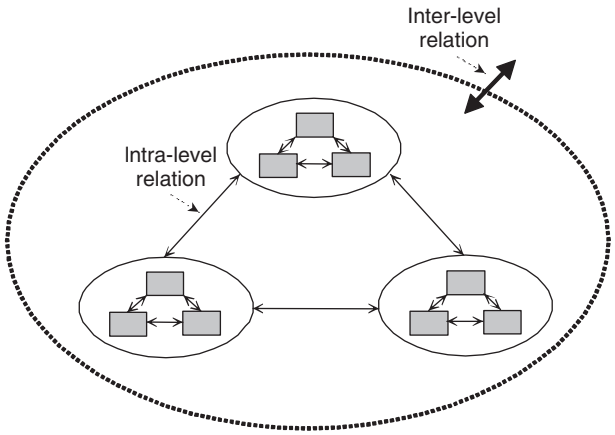


Fig. 1 Biological entities are often organized in a hierarchical manner, whereby the emergent properties of a certain level are based on the interrelations between the lower levels, while both are constrained from the highest level linkages.

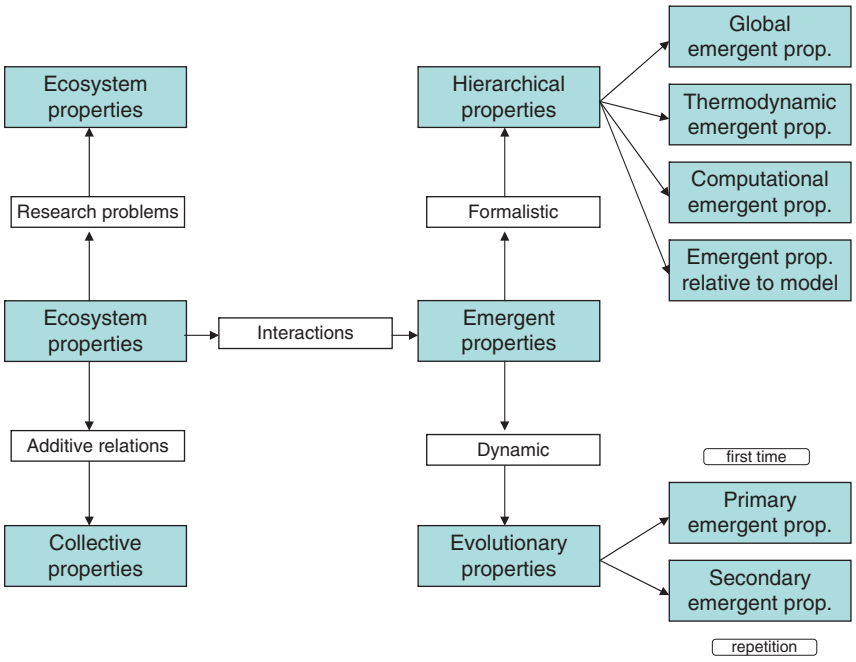


Fig. 2 An attempt to form a typology of emergent properties.

conditions that support emerging processes, especially referring to evolutionary emergence. Stable periods may lead to the emergence of new structures through bifurcations. As systems move toward the state of minimum dissipation they are, at the same time, moving toward bifurcation points with possibilities of further evolution to occur. Similarly broken symmetries, complementarity has been proposed as a global mechanism.

Classification of Emergent Properties

From the presentation of the concepts above it can be seen that emergence and emergent properties will not easily find a clear, consistent and unifying definition for covering all the cases described. The widespread and 'loose' use of the concepts over a vast range of areas at first glimpse simply shows confusion. However, it is possible to establish some typology of the areas where, and the ways in which, the concepts have been used, following Fig. 2.

First, emergent properties might appear through evolution of the systems, primary emergence, hereafter only being repeated. This characteristic may be called 'evolutionary emergence'. As structures are integrated, new organizational forms, as previously mentioned often hierarchical, occur ('hierarchical emergence').

Taking the view that emergent properties do exist and that the reductionistic approach to science will not (dis)solve the problem so it eventually disappears may allow us to establish a schematic relationship between the various categories of ecosystem properties.

One major line follows a direction of research problems, the search for the unexplained and not understood. This lies close to using emergent properties as research strategies, while the extreme leads to the reductionist approach. This is more or less the situation at the second line, where properties are 'collective' and additive, that is, that the properties are the sum of the whole, and may be explained at subsystem level, provided sufficient knowledge exists. At the other end, the attitude that only holistic studies will lead to increased understanding might be taken.

Along the third line, we find the core of emergence, and following the above points the respective features may be divided in an evolutionary line and in a hierarchal line. Here emergence is basically represented as a function of time and space. The evolutionary process was described above and deals with primary and secondary emergence. The organizational, hierarchical line includes four areas described in the previous sections: 'global emergent properties' as a function of local rules and local interactions, 'thermodynamic emergent properties' dealing with emergence as a consequence of mainly the second thermodynamic law, the emergence of (dissipative) structures as a result of thermodynamic gradients. 'Computational emergence' is also based on global patterns emerging form local rules. As mentioned above, emergent properties also appear as a result of models being used to analyze the problem, which is called 'emergent property relative to a model'.

Quantifying Emergence

Several authors argue that in any attempts to formalize or quantify the concepts, true emergent properties should be observer independent. This does not necessarily mean that emergent properties should be observation-independent. Observations undertaken by different methods result in differences in acquired knowledge. This means that emergent properties can be defined as the differences in knowledge gained by the observation of a system by two different methods. This is partly reflected by the computational emergence.

It is this observer dependency that leaves a way open for the quantification of emergent properties. Emergent properties could then be expressed in a semiquantitative way by the use of an 'index' derived of Kullback's measure of information (Fig. 3). This involves moving the normal reference frame in information theory assuming the *a priori* knowledge of the system to be zero, which is not necessarily the case.

Rather in ecology, we do possess some knowledge about the system and what we usually refer to are the deviations in what we observe in the systems or models of systems compared with our expectations built on previous knowledge. The way of quantifying emergence has to be built on the use of computers and models. If our knowledge gained hitherto is synthesized and treated in a computer model (from traditional ecological science) is p^{*}, and the outcome of an experiment or observations of a system differs by p^{**} the emergent properties can be calculated by the following:

$$\text{Emergence} = \sum p^{**} \ln \frac{p^{**}}{p^{*}}$$

which correlates emergence to the concept of exergy. Emergence now is a consequence of information gained between observations.

The question is if emergence in this manner will, at the end, dissolve itself and disappear as knowledge increases, which refers to the above debate of reductionism versus holism.

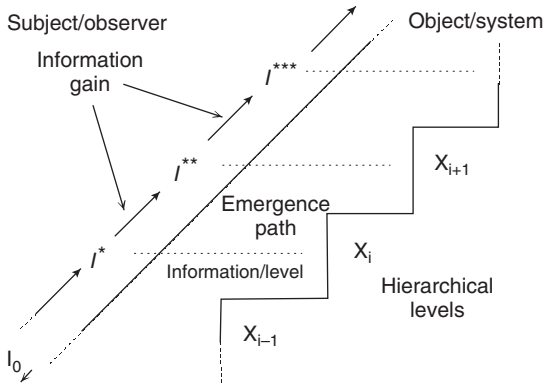


Fig. 3 Quantification of emergence, based on Kullback's measure of information, might be carried out from quantifying the difference between actual observed, *a posteriori*, behavior or composition of a system and what may be predicted from *a priori* knowledge about subsystems. The analysis may be carried out at various levels of hierarchy, differing in emergence value.

Many of the concepts used to characterize ecosystems are based on various numerical treatments of data observed in the ecosystem. Since the concepts are immediately deducible (calculable) from certain knowledge about the components of the ecosystem, for example, numbers, species, biomass etc. such concepts cannot be coined as emergent property but rather as a 'collective' property of the system. An interesting corresponding analog in this context are the macroscopic properties from thermodynamics such as entropy and parallels in formulation of formulas. Reductionism cannot win the debate since it will be impossible to achieve enough knowledge. If not for anything else, then for thermodynamic reasons, since the achievement of more and more detailed knowledge becomes more and more expensive in terms of not only energy but also dissipation.

Meanwhile, what strikes is that such a traditional, vertical organization of systems is not mandatory in order to produce emergent behavior. Vertical, here, refers to levels being either higher or lower in the hierarchy. Rather only parts are needed, of which none have actual regulatory functions and therefore should be evaluated or ranked higher than the other(s). Emergent properties can occur also in horizontally organized systems, emergence appearing alone as a consequence of interactions at the same level. The study of these intra-level relationships and their consequences to the higher levels in the hierarchy may be important to investigate in the future.

See also: Terrestrial and Landscape Ecology: Ecological Engineering: Design Principles

Further Reading

Bhalla, U.S., Iyengar, R., 1999. Emergent properties of networks of biological signalling pathways. Science 283, 381–387.
Breckling, B., Müller, F., Reuter, H., Hölker, F., Fränzle, O., 2005. Emergent properties in individual-based ecological models – introducing case studies in an ecosystem research context. Ecological Modelling 186, 376–388.
Cariani, P., 1992. Emergence and artificial life. In: Langton, G., Taylor, C., Farmer, J.D., Rasmussen, S. (Eds.), Artificial life II. Redwood City: Addison-Wesley, pp. 775–797.
Conrad, M., Rizki, M.M., 1989. The artificial worlds approach to emergent evolution. BioSystems 23, 247–260.
Emmeche, C., Køppe, S., Stjernfelt, F., 1993. Emergence and the Ontology of Levels. In In Search of the Unexplainable. Arbejdspapir. Afdeling for Litteraturvidenskab. Copenhagen: University of Copenhagen.
Morgan, C.L., 1923. Emergent Evolution. Williams and Norgate.
Nielsen, S.N., Müller, F., 2000. Emergent properties of ecosystems. In: Joergensen, S.E., Müller, F. (Eds.), Handbook of Ecosystem Theories and Management. Boca Raton, FL: Lewis Publishers, pp. 195–216.
Salt, G.W., 1979. A comment on the use of the term emergent properties. American Naturalist 113 (1), 145–148.
Wicken, J.S., 1986. Evolution and emergence. A structuralist perspective. Rivista di Biologia/Biology Forum 79 (1), 51–73.
Wieglieb, G., Bröring, U., 1996. The position of epistemological emergentism in ecology. Senckenbergiana maritima 27 (3/6), 179–193.

Goal Functions and Orientors

H Bossel, University of Kassel (retd.), Zierenberg, Germany

Introduction

The global ecosystem is made up of an ensemble of interacting local and regional ecosystems, each composed of biotic and abiotic subsystems. The evolution of these systems is constrained by physical and system laws and by the basic properties of their environment, including the constraints of exergy (energy that can be usefully transformed into work), material, and information flows. Sustainability (persistence) of a system in its environment therefore requires respecting these constraints. Conversely, the very fact of its persistence demonstrates that a system has successfully adapted to its operating conditions. Evolution has forced it to respect physical and system laws and the basic properties of its environment. To an observer, the system's behavior appears to be guided by a particular attractor state, or by attention to a number of orientors.

System Concepts

System Organization

'System' is anything that is composed of system elements connected in a characteristic system structure (Fig. 1). This configuration of system elements allows it to perform specific functions in its environment. These functions can be interpreted as serving a distinct system purpose. The system boundary is permeable for inputs from, and outputs to, the environment. It defines the system's identity and autonomy.

When we talk about a viable system, we mean that this system is able to survive, be healthy, and develop in its particular environment. In other words, system viability has something to do with both the system and its properties, and with the environment and its properties. And since a system usually adapts to its environment in a process of coevolution, we can expect that the properties of the system's environment will be reflected in the properties of the system; for example, the form of a fish and its mode of motion reflect the laws of fluid dynamics of its aquatic environment.

Systems are termed complex if they have an internal structure of many – qualitatively different – processes, subsystems, interconnections, and interactions. Besides assuring their own viability, the individual systems that are part of a complex total system specialize in certain functions that contribute to the viability of the total system. Viability of subsystems and the total system requires that subsystem functions and interactions are organized efficiently (or at least effectively). In the evolution of complex systems, two organizing principles in particular have established themselves: hierarchy and subsidiarity. They can be found in all successful complex systems: biological, ecological, social, political, technological.

Hierarchical organization means a nesting of subsystems and responsibilities within the total system. Each subsystem has a certain degree of autonomy for specific actions, and is responsible for performing certain tasks contributing to the viability of the total system. For example, body cells are relatively autonomous subsystems, but contribute specific functions to the operation of particular body organs, which in turn contribute to the viability of an organism.

Subsidiarity means that each subsystem is given the responsibility and the means for keeping its own house in order, within the range of its own abilities and potential. Only if conditions occur that cannot be handled by the subsystem would the suprasystem

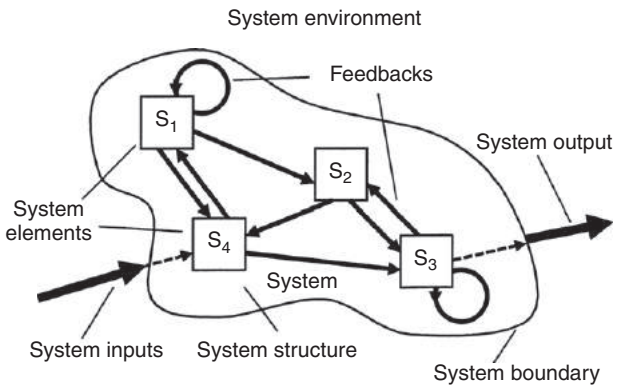


Fig. 1 System notation.

 doi:10.1016/B978-0-444-63768-0.00691-0

step in and help. The principles of hierarchical organization and subsidiarity require that each subsystem has a certain measure of autonomy. In its particular environment, each subsystem must be viable. The total system can only be viable if each of the subsystems supporting it is viable. Each subsystem reflects the properties of its individual environment; its behavior is informed (oriented) by that environment.

Note that this way of looking at complex systems is recursive. If necessary, we can apply the same system/subsystem dichotomy of viable systems again at other organizational levels. For example, a person is a subsystem of a family; a family is a subsystem of a community; a community is a subsystem of a state; a state is a subsystem of a nation, etc.

It is not enough to be concerned with the viability of individual systems. There are no isolated systems in the real world; all systems depend in one way or another on other systems. Hence their viability, and ultimately the viability of the total system are also preconditions for sustainable development. This means that a holistic system view must be adopted.

Evolution of Systems, and Emergence of Orientors and Goal Functions

The adaptation of a system to its environment is reflected in its structure, including its nonmaterial, cognitive structure. This system structure determines its behavior, and hence the adaptive response to its particular environment. System structures of material systems are dissipative; they require exergy and material flows for their construction, maintenance, renewal, and reproduction.

The dissipative structures of the global ecosystem are constructed and maintained by a finite rate of exergy input (mostly solar energy) and a finite stock of materials. The global ecosystem is therefore forced to recycle all of its essential material resources. The development of local ecosystems is constrained by the local rate of exergy flux (solar radiation input) and by the local rate of material recycling (weathering rate, absorption rate, decomposition rate, etc.) that it produces.

Evolution favors those species or (biotic) subsystems of the ecosystem that have learned to use available resources more efficiently and effectively than their competitors. This learning is embedded in their genetic code, and it is manifest in the dissipative structures they construct. Both will increase in complexity as a species evolves. At the ecosystem level, species evolution will cause increasingly better use of (exergy and material) resources. Species as well as ecosystems as a whole therefore tend to progress toward more complex dissipative structure producing more complex behavior.

Interacting species in a common ecosystem coevolve in the direction of increasing fitness of each individual species. Evolution of ecosystems therefore proceeds in the direction (arrow of time) of specialization, speciation, synergy, complexification, diversity, maximum throughflow of exergy, and more efficient use of material resources. This development becomes manifest in the corresponding emergent properties: exergy degradation, recycling, minimization of output, efficiency of internal flows, homeostasis and adaptation, diversity, heterogeneousness, hierarchy and selectivity, organization, minimization of maintenance costs, storage of available resources. These properties can be viewed as orientors, propensities, or attractors guiding system evolution and development. They are not limited to ecosystems; they are a general feature of living systems, including human organizations. When quantified and used in models, we refer to them as goal functions.

In particular, ecosystems will therefore build up in the course of their development as much dissipative structure as can be supported by the available exergy gradient. Available opportunities will eventually be found out by the processes of evolution, and will then be utilized. The ability to respond successfully to environmental challenges can be 'interpreted' as intelligent behavior, although it is strictly the result of nonteleological evolutionary development.

System Orientation in a Complex Environment

Basic concepts can be introduced by visualizing a simple animal with limited vision in a simple environment. The animal requires exergy for self-organization, motion, harvesting food, and maintenance. The environment provides food in certain locations, usually associated with obstacles that must be avoided since they have an exergy cost.

In a stable environment where sufficient (regenerating) food is distributed in a completely regular pattern, evolutionary adaptation would eventually lead to optimization of an animal's movements in a regular grazing pattern, with a single objective, optimum exergy uptake and use. The regular grazing pattern reflects the complete certainty of the next step, which the animal learns by accumulating and internalizing experience in a cognitive structure aiding its limited vision.

In more complex and diverse environments the animal, because of its limited vision, may not know for several steps which situation it will encounter next. It will therefore have to develop decision rules that have greater generality and are applicable to (and will be reinforced by) different motion sequences with different outcomes. In addition to the requirement of harvesting and using exergy resources effectively and efficiently, another objective is now implicitly added, to secure food under the constraint of incomplete information, that is, a security objective. Note that this is an emergent property that is not explicit in the reward system (which still rewards only food uptake). Failure to heed this implicit security objective will reduce food uptake and may endanger survival. On the other hand, the pressure to play it safe will occasionally mean giving up relatively certain reward. With other words, efficiency is traded for more security, and both are now prominent normative orientations (goals, values, interests) incorporated in the cognitive structure.

Orientation theory deals in a more general way with the emergence of behavioral objectives (orientors) in self-organizing systems in general environments. The proposition is that if a system is to survive in a given environment – characterized by a specific normal environmental state, sparse resources, variety, unreliability, change, and the presence of other systems – it must be able to physically exist in (be compatible with) this environment, effectively harvest necessary resources, freely respond to

environmental variety, protect itself from unpredictable threats, adapt to changes in the environment, and interact productively with other systems. These essential orientations emerge in the course of the system's evolution in its environment.

Properties of Environments

There is obviously an immense variety of system environments, just as there is an immense variety of systems. But all of these environments have some common general properties. These properties will be reflected in systems. These reflections, or basic orientors, orient not just structure and function of systems, but also their behavior in the environment. The term orientor is used to denote (explicit or implicit) normative concepts that direct behavior and development of systems in general. In the social context, values and norms, objectives and goals are important orientors. Ecosystems and organisms tend toward certain attractor states whose specific characteristics can be viewed as orientors. Orientors exist at different levels of concreteness within an orientor hierarchy. The most fundamental orientors, the basic orientors, are identical for all complex adaptive systems. Orientors are dimensions of concern; they are not specific goals. Their satisfaction can be determined by observation of corresponding indicators, which can also be used to define goal functions for model studies.

In addition to the physical constraints of exergy and material flows, ecosystem and species development is determined by the 'general properties of the environment':

1. *Normal environmental state*. The actual environmental state can vary around this state in a certain range.
2. *Scarce resources*. Resources (exergy, matter, information) required for a system's survival are not immediately available when and where needed.
3. *Variety*. Many qualitatively very different processes and patterns occur in the environment constantly or intermittently.
4. *Reliability*. The normal environmental state fluctuates in random ways, and the fluctuations may occasionally take it far from the normal state.
5. *Change*. In the course of time, the normal environmental state may gradually or abruptly change to a permanently different normal environmental state.
6. *Other systems*. The behavior of other systems changes the environment of a given system.

Basic Orientors

If evolution enforces fitness of (natural) systems, then persistent systems must reflect the properties of their environment in their structure. More generally, the basic properties of the environment require corresponding basic system features. Since the basic environmental properties are independent of each other, a similar set of independent system features must exist, and it must find expression in the concrete features of the system structure.

There is a one-to-one relationship between the properties of the environment and the 'basic orientors of systems' (Fig. 2):

1. *Existence*. Attention to existential conditions is necessary to insure the basic compatibility and immediate survival of the system in the normal environmental state.

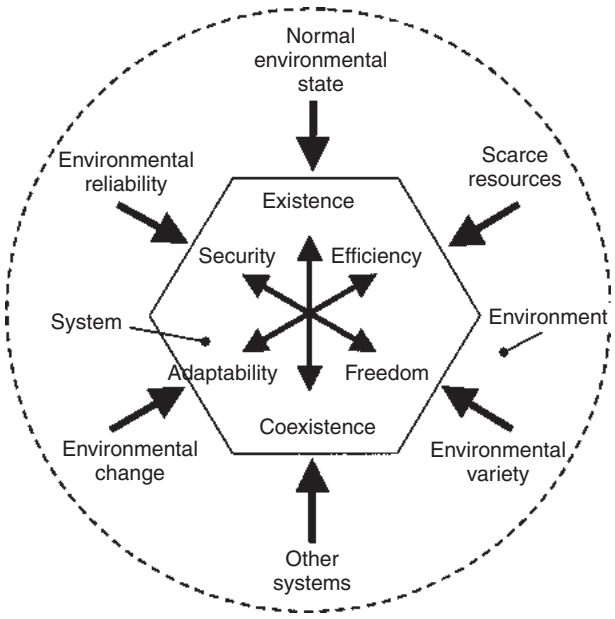


Fig. 2 A tentative typology of emergent properties.

2. *Effectiveness*. In its efforts to secure scarce resources (exergy, matter, information) from, and to exert influence on its environment, the system should on balance be effective.
3. *Freedom of action*. The system must have the ability to cope in various ways with the challenges posed by environmental variety.
4. *Security*. The system must have the ability to protect itself from the detrimental effects of variable, fluctuating, unpredictable, and unreliable environmental conditions.
5. *Adaptability*. The system should be able to change its parameters and/or structure in order to generate more appropriate responses to challenges posed by changing environmental conditions.
6. *Coexistence*. The system must modify its behavior to account for behavior and interests (orientors) of other systems.

Obviously, the system equipped to secure better overall orientor satisfaction will have better fitness, and will therefore have a better chance for long-term survival and sustainability. In persistent systems or species, these orientors will be found as emergent objectives (or system interests).

Properties of Orientors

Each of the basic orientors stands for a unique requirement. Attention (conscious or unconscious) must therefore be paid to each of them, and the compensation of deficits of one orientor by over-fulfillment of other orientors is not possible. Fitness forces a multicriteria response, and comprehensive (conscious or unconscious) assessments of system behavior and development must also be multicriteria assessments.

In the assessment and orientation of system behavior, we deal with a two-phase assessment process where each phase is different from the other.

Phase 1. First, a certain minimum satisfaction must be guaranteed separately for each of the basic orientors. A deficit in even one of the basic orientors threatens long-term survival. The system will have to focus its attention on this deficit.

Phase 2. Only if the required minimum satisfaction of all basic orientors is guaranteed is it permissible to try to raise system satisfaction by improving satisfaction of individual orientors further.

Adequate satisfaction of each of the basic orientors requires, on a lower level, system- and environment-specific satisfaction of thermodynamic, structural, functional, ecophysiological, and system orientors. Network analysis suggests complementarity of different formulations of extremal principles as orientors describing ecosystem development.

Characteristic differences in the behavior of otherwise very similar systems (animals, humans, political, or cultural groups) can often be explained by differences in the relative importance attached to different basic orientors (i.e., emphasis on freedom, or security, or effectiveness, or adaptability) in phase 2 (i.e., after minimum requirements for all basic orientors have been satisfied in phase 1).

The basic orientor proposition has three important implications:

1. If a system evolves in a normal environment, then that environment forces it to implicitly or explicitly ensure minimum and balanced satisfaction of each of the basic orientors (and of lower-level orientors contributing to this satisfaction).
2. If a system has successfully evolved in a normal environment, its behavior will exhibit balanced satisfaction of each of the basic orientors.
3. If a system is to be designed for a normal environment, proper and balanced attention must be paid to satisfaction of each of the basic orientors.

The third implication has particular relevance for the creation of programs, institutions, and organizations in the sociopolitical sphere, among other things. Note that for a specific system in a specific environment, each orientor will have a specific meaning. For example, security of a nation is a multifacetted objective set with very different content from the security of an individual particular organism. However, the systems theoretical background for satisfaction of the security orientor is the same in both cases.

Orientors as Implicit Attractors

Better orientor satisfaction (better fitness) for more participants in a system requires more dissipative structure, which requires more exergy throughput as well as exergy accumulation. Since the exergy flow of ecosystems is limited (capture of solar radiation by photoproduction), increasingly better utilization is to be expected in the course of system development. This saturates at maximum exergy flow utilization for the ecosystem as a whole. Ecosystems as a whole therefore move in the direction of using all available exergy gradients. For organisms in the ecosystem, this implies development tendencies (orientors, propensities, attractors) toward specialization (using previously unused gradients), more complex structure (greater use efficiency), larger individuals (less maintenance exergy required per biomass unit), mutualism, etc. For species development, this translates into a principle of maximum exergy use efficiency. On the basis of these principles, prediction of development trends in ecosystems is possible.

The selection for better fitness in evolutionary processes favors systems (organisms) with better coping ability. Aspects of the behavioral spectrum of a system that improve coping ability (basic orientors) can be understood as implicit goals or attractors: existence, security, effectiveness, freedom, adaptability, coexistence. In the developmental stage of ecosystems, emphasis is on the basic orientors: existence, effectiveness, and freedom; in the mature stage it shifts to security, adaptability, and coexistence (see **Table 1**, where orientor concepts have been linked to E. P. Odum's classical model of ecological succession).

Table 1 Orientor concepts in the context of ecological succession

	Developmental stage	*Mature stage*
	Basic orientor emphasis	
	Existence	Coexistence
	Freedom	Security
	Effectiveness	Adaptability
Ecosystem orientor	*Orientor emphasis (goal function)*	
Growth and change	High	Low
Life cycle	Short, simple	Long, complex
Biomass	Low	High
Energy conservation	Low	High
Nutrient conservation	Low	High
Nutrient recycling	Low	High
Specialization	Low	High
Diversity	Low	High
Organization	Low	High
Symbiosis	Low	High
Stability; feedback control	Low	High
Structure	Linear, simple	Network, complex
Information	Low	High
Entropy	High	Low

The existence of these implicit goals does not imply teleologic or teleonomic development toward a given goal (where the final state is specified). These attractors do not determine the exact future states of the system at all; they only pose constraints on choices (or evolutionary selection). The process and its rules are known, the product is unknown. The spectrum of (qualitatively different) possible future development paths and sustainable states remains enormous. The shape of the future, and of the systems that shape it, cannot be predicted this way. All one can say with certainty, however, is that (1) all possible futures must be continuous developments from the past, and (2) paths with better orientor satisfaction are more likely to succeed in the long run (if options to change paths have not been foreclosed).

In many systems, in particular ecosystems, specific attractors or functional orientors are often more immediately obvious than the basic orientors that cause the emergence of these orientors in the first place. These orientors can be viewed as appearing on a level below the basic orientors in the hierarchical orientation system (see **Table 1**). They translate the fundamental system needs expressed in the basic orientors into concrete attractor states linking system response to environmental properties. In models and ecosystem analyses, measures of ecosystem integrity can be based on corresponding ecosystem goal functions. Ecosystem attractor states emerge as general ecosystem properties in the coevolution of ecosystem and environment. They can be viewed as ecosystem-specific responses to the need to satisfy the basic orientors. Major ecosystem orientors are optimization of use of solar radiation, material, and energy flow intensities (networks); matter and energy cycling (cycling index); storage capacity (biomass accumulation); nutrient conservation, respiration, and transpiration; diversity (organization); hierarchy (signal filtering).

The emergence of basic orientors in response to the general properties of environments can be deduced from general systems theory, but supporting empirical evidence and related theoretical concepts can also be found in such fields as psychology, sociology, and the study of artificial life.

Orientor Guidance in System Development, Control, Adaptation, and Evolution

Environmental influences partially determine system behavior. The magnitude of their effect on behavior depends on the influence structure of the system.

Sometimes systems can be controlled by controlling the inputs from their environment. However, the feedbacks in the system itself are usually more important for system control and adaptation of behavior to environmental conditions. Feedback means that the system state influences itself. Behavior-changing internal feedbacks are possible on several hierarchical levels in complex systems with different typical response characteristics and time constants (typical response times). These possibilities are also shown in **Fig. 3**.

Response time	*Level*	*Response*
Immediate	Process	Cause–effect
Short	Feedback	Control
Medium	Adaptation	Parameter change
Long	Self-organization	Structural change
Very long	Evolution	Change of identity
Always	Basic orientors	Maintaining integrity

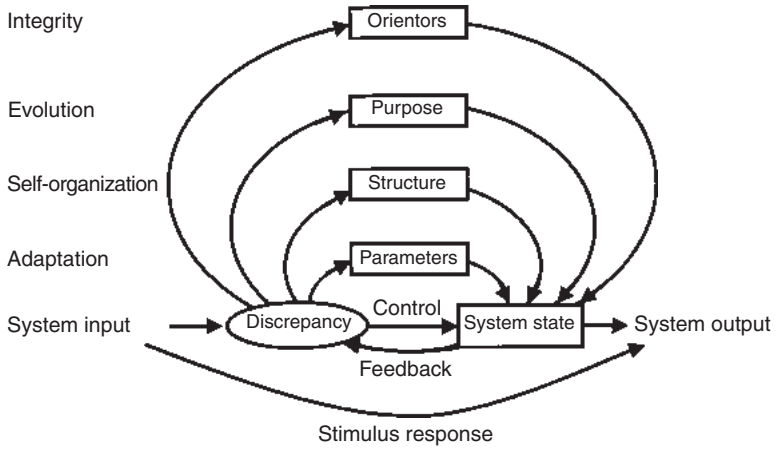


Fig. 3 System response can be caused by different processes with very different time constants: stimulus–response, feedback control, adaptation, self-organization, evolution, maintaining system integrity.

The simplest type of system response is the cause–effect relationship. It occurs at once as in for example, stimulus–response reflex. It is the only type of system behavior which can legitimately be described by relating the output directly to the input. Unfortunately, it is often assumed that the same simple relationship is also applicable to other types of system response (such as the following), and this erroneous assumption often leads to fundamental mistakes.

On the next higher level we find responses which are generated by feedback in the system, involving at least one state variable or delay – such as an empty stomach causing hunger and the search for food. Control processes belong to this category. The response time is short, and influence structure and system parameters remain invariant.

On the next higher level we find processes of adaptation. In this case the system maintains its basic influence structure, but parameters are adjusted to adapt to the situation, possibly changing the response characteristics in the process. For example, a tree may adapt to the gradual lowering of the groundwater level by growing its roots to greater depth. This constitutes a parameter change (root length and root surface). The fundamental system structure of a tree, in particular, the function of the roots, has not changed in this case.

On the next higher level we find processes of self-organization in response to environmental challenges. This means structural change in the system. Processes of this kind have a longer response time and can only be conducted by systems having the capability for self-organization. Adult organisms or technical systems rarely or never belong to this category; on the other hand, this characteristic is often found in the development of organisms, social systems, organizations, and ecosystems.

A system may also change its identity in the course of an evolutionary process. This means that its functional characteristics, and hence its system purpose, change with time. Adaptations of this kind take place as a result of reproduction and evolution of living organisms. It is characteristic of this process that the system change coincides with a possibly drastic shift in system identity (change of goal function and of system purpose). An evolutionary example is the development of flying animals (birds) from water-dwelling reptiles.

All of these system responses to challenges from the environment in essence constitute attempts to maintain system integrity (possibly over many generations and over a long time period) even if it means changing system identity, that is, system purpose. From this observation it can be deduced that a system must orient its development with respect to certain basic criteria (basic orientors) to assure its long-term existence and development in an often hostile environment. This orientation may be implicit (forced upon the system) or explicit (actively pursued by the system). It does not require conscious decision or even cognitive ability, although resulting action may appear to an observer as intelligent or even goal- or value-oriented behavior.

Simulation of the Evolution of System Orientation

Animats and Genetic Algorithms for Orientation

Orientation theory is not just a conceptual framework for understanding system evolution and behavior under the exergy availability constraint. It also allows quantitative and comparative analysis of system performance under different environmental conditions.

Genetic algorithms are models of biological adaptive processes that are being widely and successfully applied to a wide spectrum of adaptation and optimization problems. In particular, these algorithms have been used to simulate learning and adaptation of artificial animals (animats) in simulated environments containing food and obstacles. They can be used to demonstrate the emergence of basic orientors in self-organizing systems having to cope with complex environments.

The animat model incorporates essential features of a simple animal in a diverse environment. Being an open system, an animal depends on a flow of exergy from the environment. In the course of its (species) evolution, it has to learn to associate certain signals from the environment with reward or pain and to either seek or avoid their respective sources (exergy gain or exergy loss). This learning phase (of populations) will eventually lead to the establishment of cognitive structure and behavioral rules which are approximately optimal in the particular environment (with respect to maximization of reward, minimization of pain, and securing survival). These behavioral rules incorporate knowledge which enables intelligent behavior.

The animat is designed to simulate this process. It can pick up sensory signals from its environment (containing food and obstacles), and classify them with available rules to determine an appropriate action (direction of movement). After a successful move, the strength of rules leading up to it is increased by sharing in the reward (i.e., exergy gain). New rules are occasionally generated by either random creation, or by genetic operations (crossing-over and recombination). They are added to the existing rule set, and compete with the other rules for reward. Unsuccessful rules are not reinforced and lose strength and influence in the rule set.

The training process consists of placing the animat at a random empty location in an environment with specific environmental properties, and allowing it to move around searching for food. A collision with an obstacle causes a loss of exergy and throws the animat back to its previous position. Rules leading to success are rewarded. A genetic event of rule generation may occur with a prescribed probability. Random rules are created in unknown situations. The process is repeated for a large number of steps (typically 10 000). Eventually, a set of behavioral rules develops which allows optimal behavior under the given set of conditions.

Note that this optimal behavior has not been defined in terms of an objective function guiding the evolution of the set of behavioral rules. The rule set develops solely from the reinforcement of rules which lead to food or avoid collisions. An explicit exergy balance accounts for all exergy losses associated with movement, collisions with obstacles, and rule generation, and exergy gains due to uptake of food. The development of the rule set is then driven by the requirement to optimize exergy pickup in the given environment (with specific resource availability), while allowing for environmental variety, variability, and change specific for that environment. Neglect of these properties is penalized by lack of fitness, and threat to survival, and causes disappearance of deficient rules. Other criteria besides efficiency will therefore be reflected in the set of behavioral rules. Since these were not expressly introduced, we must recognize them as emergent value orientations or objective functions.

The animat experiment contains all components necessary for a study in the basic orientor framework. Animat fitness depends on the ability to maintain a positive exergy balance in the long term. This exergy balance is therefore at the core of the orientor satisfaction assessment. At each step, exergy uptake (by food consumption) and exergy losses (by collisions with obstacles, motion, and learning of rules) are recorded and used to compute the momentary exergy balance. Attention to all orientors is mandatory to ensure a positive exergy balance even under adverse environmental conditions.

Quantitative measures must be defined for characterizing the different properties of the environments used in the animat experiments. Animat performance in different environments is compared by using measures of orientor satisfaction. These have to be defined using relevant parameters of animat performance.

Emergence of Basic Value Orientations, Anticipation, and Individual Differences

Since the animat's training depends on a number of random factors, each animat develops a different cognitive system (classifier set and decision rules), even though final performance may be similar. In order to show general tendencies despite these individual differences, mean values over large populations were obtained. These dealt with (1) results of the training process in two (otherwise identical) environments having different variety and variability, and with (2) performance of animats after transfer from their training environment to environments challenging them with more variety, or variability, or change.

One remarkable result from these experiments is that individuals achieve comparable performance in a given training environment with very different cognitive systems, and in particular with different orientor emphasis. While this may not provide any particular advantage in the training environment, it may provide distinct fitness advantages if the animat is moved to a different environment. Three particular types of individuals stand out: generalists (type F) stressing freedom of action, specialists (type E) focusing on effectiveness, and cautious type (type S) emphasizing security. **Fig. 4** shows the different orientor stars for these three types.

The ability to develop a cognitive system reflecting its environment makes the animat a suitable vehicle for investigating goal function emergence and value orientation. Genetic algorithms are very effective processes that seem to capture the essentials of real processes found in the evolution of organisms and ecosystems. In the animat, they very effectively build up a cognitive model (or goal function) that enables anticipatory behavior; since rewards flow back to earlier rules leading to later pay-off, the activation of the initial rules in a pay-off chain means that the system suspects possible pay-off and anticipates the near future, that is, it has an internal model of the results of its actions under the given circumstances.

In the animat experiments (and similarly, in real life), implicit and (more or less) balanced multidimensional attention to the basic orientors emerges from the simple one-dimensional mechanism of rewarding success in the given environment. Thus, in the course of its evolutionary development in interaction with its environment, the system evolves a complex multidimensional behavioral objective function from the very unspecific requirement of fitness. Conversely, this also means that balanced attention to the emergent basic orientors is necessary for system viability and survival – they would not have emerged unless important for the viability of the system.

Balanced attention still leaves room for individual differences in the relative emphasis given to the different orientors. Individuals belonging to the populations used in the animat experiments evolve significant differences in value emphasis

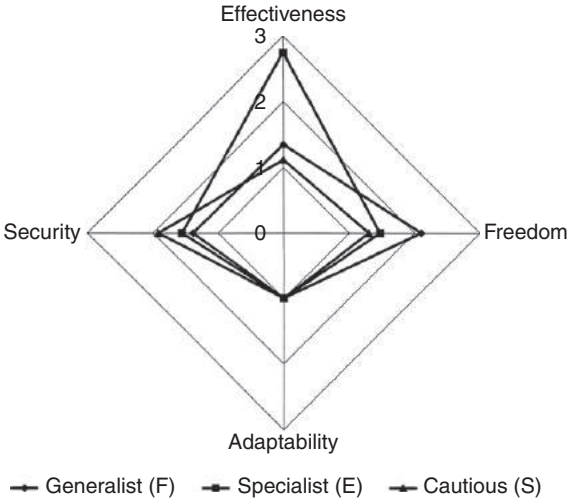


Fig. 4 In an identical training environment, different lifestyles may evolve. Generalists stress freedom of action, specialists focus on effectiveness, while cautious types emphasize security.

(e.g., specialist, generalist, cautious type). These individual variations, while not significantly reducing performance in the standard training environment, provide comparative advantage and enhanced fitness when resource availability, variety, or reliability of the environment change. They also result in distinctly different behavioral styles. However, pathological behavior will follow if orientor attention becomes unbalanced (e.g., dominant emphasis on one orientor).

Training of animats in different environments, the performance of animat individuals in environments that differ from their training environments, and the simulation of adaptive learning in a changing environment, lead to some general conclusions that are in full agreement with everyday observations and general systems knowledge:

- Generalists have a better survival chance than others if moved to an environment of greater variety.
- Cautious types have a better survival chance than others if moved to a less reliable environment.
- Training in more unreliable and/or more diverse environments increases satisfaction of the security and/or freedom of action orientors at the cost of the effectiveness orientor.
- Training in an uncertain environment teaches caution and improves fitness in a different environment.
- Learning caution (better satisfaction of the security orientor) takes time and decreases effectiveness, but increases overall fitness.
- Investment in learning (exergy cost of learning in the animat) pays off in better fitness; the learning investment is (usually) much smaller than the pay-off gain.

Animat individuals not only develop behavior that can be interpreted as intelligent, they also develop a complex goal function (balanced attention to basic orientors), or value orientation. Serious attention to basic values (basic orientors: existence, effectiveness, freedom, security, adaptability, coexistence) is therefore an objective requirement emerging in, and characterizing self-organizing systems. These basic values are not subjective human inventions; they are objective consequences of the process of self-organization in response to normal environmental properties.

Further Reading

Ashby, W.R., 1962. Principles of the self-organizing system. In: von Foerster, H., Zopf, G.W. (Eds.), Principles of Self-Organization. New York: Pergamon, pp. 255–278.
Bossel, H., 1977. Orientors of nonroutine behavior. In: Bossel, H. (Ed.), Concepts and Tools of Computer-Assisted Policy Analysis. Basel: Birkhäuser, pp. 227–265.
Bossel, H., 1999. Indicators for Sustainable Development: Theory, Method, Applications. Winnipeg: IISD International Institute for Sustainable Development.
Bossel, H., 2001. Exergy and the emergence of multidimensional system orientation. In: Jørgensen, S.E. (Ed.), Thermodynamics and Ecological Modelling. Boca Raton, FL: Lewis, pp. 193–209.
Fath, B.D., Patten, B.C., Choi, J.S., 2001. Complementarity of ecological goal functions. Journal of Theoretical Biology 208 (4), 493–506.
Holland, J.H., 1992. Adaptation in Natural and Artificial Systems. Cambridge, MA: MIT Press.
Jantsch, E., 1980. Self-Organizing Universe: Scientific and Human Implications of the Emerging Paradigm of Evolution. New York: Pergamon.
Jørgensen, S.E., 2001. A tentative fourth law of thermodynamics. In: Jorgensen, S.E. (Ed.), Thermodynamics and Ecological Modelling. Boca Raton, FL: Lewis, pp. 305–347.
Krebs, F., Bossel, H., 1997. Emergent value orientation in self-organization of an animat. Ecological Modelling 96, 143–164.
Mayr, E., 1974. Teleological and teleonomic: A new analysis. Boston Studies in the Philosophy of Science 14, 91–117.
Mayr, E., 2001. What Evolution Is. New York: Basic Books.
Miller, J.G., 1978. Living Systems. New York: McGraw-Hill.
Odum, E.P., 1969. The strategy of ecosystem development. Science 164, 262–270.
Müller, F., Leupelt, M. (Eds.), 1998. Eco Targets, Goal Functions, and Orientors. Berlin/Heidelberg/New York: Springer.
Wilson, S.W., 1985. Knowledge growth in an artificial animal. In: Grefenstette, J.J. (Ed.), Proceedings of the First International Conference on Genetic Algorithms and Their Applications. Pittsburgh PA and San Mateo: Lawrence Earlbaum and Morgan Kaufmann, pp. 16–23.

Hierarchy Theory in Ecology

TFH Allen, University of Wisconsin, Madison, WI, USA

Need for Hierarchy Theory

In ecology we need a body of theory to address relationships that are consequences of changing levels of analysis, which call for altered definitions. For instance, the contemporary fracas over definitions of plant competition could benefit from recognizing differences in the scope and type of the framework used by the respective partisans. With distinctions between levels of analysis made clear, each school may test their respective hypotheses in peace aware of which theories actually compete and when they merely address some other level of discourse. The contentious literature surrounding overcompensation of plants in response to losses to grazing was significantly a matter of pulse versus press consumption in relation to different timescales for assessing recovery. The Clements/Gleason debate over the proper definition of plant community might not have lasted the body of the twentieth century had hierarchy theory been available at the onset of hostilities when Nichols attacked Gleason's paper at the 1926 International Congress of Plant Sciences. Hierarchy theory's focus on level of analysis offers such clarification. It lays subtle distinctions bare, so that definitions work for the ecologist instead of ecologists working for their definitions.

If big, slow things were always on top, such that hierarchical levels were only a matter of scale, the problem would reduce to a straightforward technical scaling issue. Not to underestimate the challenges of scaling in engineering, but that technical setting does not need something as grand as a theory to deal with hierarchies. But in ecology, scaling is complicated by higher ecological levels giving lower levels meaning. In ecology, the move upscale to be more inclusive often changes significance more than it invokes a change of size, and so we do indeed need a special body of theory to deal with difference of quality, not just quantity. Differences in scale quickly become large enough to cause qualitative change in perception, which forces a change in the level of analysis. Thus scale is soon embroiled in values, judgment, and arbitrary choices, not just as an inconvenience, but as a necessity for proper understanding. While scaling as an engineering technicality actively ignores such messy issues, hierarchy theory explicitly includes value-based decisions of the observer in creating hierarchies.

To control for observer values, technical measurement and analysis in science keeps its criteria constant across the local discourse. But large discoveries precisely amount to the recognition of a change in value. New scientific ideas indicate a specific change in the preanalytic stage, before deciding what might be relevant data. In the terms of Russell and Whitehead (made accessible by Gregory Bateson), new scientific ideas amount to the definition of new logical types. Hierarchy theory's central activity is recognizing logical type. Logical types are tied to some new level of inclusivity, a new hierarchical level with its own meaning. Notice how left and right sides are possessed by organisms at their own level of existence. Meanwhile, the notions of up and down refer to a larger discourse that includes an environment, which is shared by many organisms. As a result, a mirror switches the image left and right, but with no switch in up versus down. The larger scope invoked by the idea of up introduces a new logical type, even though left and right may often be simply at right angles to up and down. If left and right contrasted with up and down can be problematic, ecosystems are a nightmare. While exquisitely holding criteria constant in formal scientific calibration will help, it is insufficient for large discoveries, which turn on recognizing when a new type is necessary to solve some puzzle.

Ecology in particular invites many logical types because its hierarchies are so rich. A new type invokes new aggregation criteria, which come explicitly from observer decisions. Consider the difference between a community conception of vegetation as opposed to the process-functional conception that prevails in ecosystem modeling. A forest can be considered as a collection of trees on a tract of land. Alternatively those same tree trunks may be aggregated as a separate class from the leaves (**Fig. 1**). If leaves in a forest are the production system independent of species, then the boles are part of the carbon storage function. This assignment has the peculiar effect of unifying the tree trunks with soil carbon in a single carbon storage compartment. A community focus aggregates trees set in an environment of soil and atmosphere. Meanwhile a flux-process conception splits the trees into at least two parts, one of which aggregates with the soil. But the soil was part of the environment in the community conception. Thus the same pieces of soil and plant biomass are aggregated into different higher units, depending on the type of system that is recognized as being in the foreground by the observer. Note how forests under either conception may be called forest ecosystems, suggesting that one use of hierarchy theory is to untangle alternative meanings in commonplace ecological terminology. The difference between a process-focused ecosystem and a community is a change in logical type.

Hierarchy and Hypothesis

Hierarchy theory is a body of thought that relates chosen levels of analysis to defined levels of organization, all in the context of scaling. It advises the scientist of subtle but crucial distinctions that follow from observing and making analytical decisions. A significant part of hierarchy theory is observation in relation to conceptions of order in complex systems that would otherwise

 doi:10.1016/B978-0-444-63768-0.00692-2

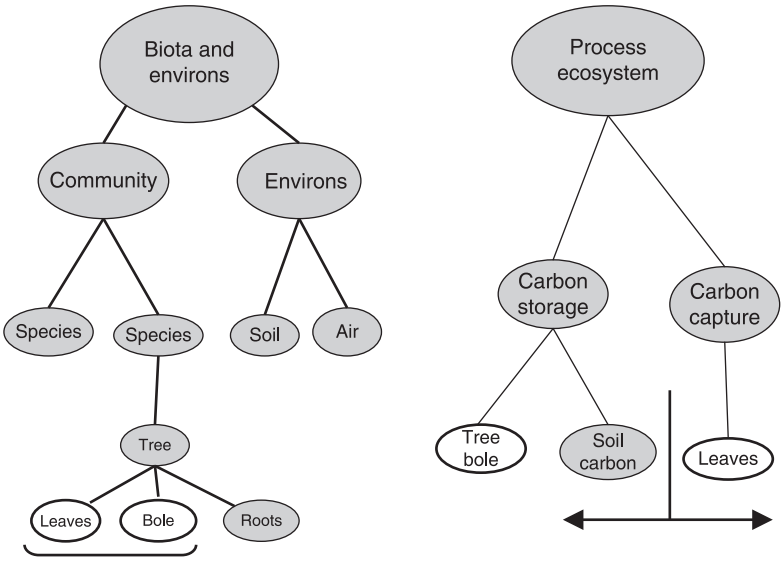


Fig. 1 A community conception leads to an expected situating of whole trees in an environment. But a process-functional ecosystem conception of that same forest can lead to a hierarchy where tree boles are separated from the leaves, and then united with soil elements in a carbon storage compartment. Each respective hierarchy takes its form from the purpose for which it is intended.

invite confusion. It is a metatheory that guides the generation, fine-tuning and testing of other bodies of thought, themselves more easily recognizable as theories in the conventional sense. Some theories in ecology are associated with answering questions taken as given. Such theory is validated in testing hypotheses. Other ecological theories may fine-tune questions, perhaps clarifying what is meant by competition, so that worthwhile hypotheses may be generated. By contrast, hierarchy theory applies in the pre-analytical stage, when the questions are being framed rather than when clarified or answered. In the preanalytical stage, the boundaries of things are established, and structures are assigned to types or classes. With the discourse laid out unambiguously by hierarchy theory, other theories may come into play, testing explicit hypotheses with measurement and models. Thus hierarchy theory does not have its own hypotheses *per se*, but rather opens the way for subsequent testing of specific hypotheses. Like multivariate description in ecology, hierarchy theory focuses on hypothesis generation and clarification. From a small number of first principles, it highlights what would be otherwise taken for granted and then forgotten in the muddle that ensues. Hierarchy theory is explicit, as it positions the tacit next to the focal. Its precision is in thought and choosing definitions, more than action in quantitative experimentation.

Hierarchical Levels

Entities in a hierarchy are recognized as belonging to levels. Levels are sets, but the sets become levels because of robust asymmetry between them in a hierarchy. Mathematically, the asymmetry between levels makes hierarchies partially ordered sets. Hierarchy theory is the set theory that may precede network theory. A level of analysis assigns entities to levels, and is often explicit about their relationship to other entities assigned to other levels. There is a distinction between levels of observation and levels of organization. Levels of observation are ordered relative to each other on matters of size and scale. Meanwhile relationships between levels of organization follow from definitions chosen by observers, sometimes as a prelude to actual observation. For instance, organisms are subsumed by populations only by a definition. Hidden in the definition is a requirement for equivalence between population members. Meanwhile, a host and its parasite, while both organisms, are generally not assigned to the same population, in part because of inequivalent size. Host versus parasite is the basis of a hierarchy employing levels subtly different from those in the population/organism distinction. Hierarchy theory places entities in levels, taking care to be explicit about the definitions that lead to those levels, and the criteria that create order and linkage.

Scale versus definition has potential for generating different sorts of hierarchies. Some hierarchies focus on size and containment, while others are control hierarchies where upper level entities simply control lower levels. A Watt governor may be placed at a higher level in a control hierarchy, while being smaller than the whole steam engine it controls. Whether it is a scalar or a control hierarchy depends on the use for which the hierarchical conception is intended, something for which the observer must take responsibility. Time against space plots are popular in landscape ecology. But such hierarchies can miss out on the interesting situations where large space maps onto short time spans, or long time spans map onto small places (**Fig. 2**). The globe is large enough to be the context of continental movement over hundreds of millions of years, but at the same time the rotation of the globe is also responsible for diurnal phenomena, at the fast end of ecological happenings. Surfaces arise when narrow space applies to large differences in time constants (strong temporal connection within, but weak connections across surfaces). Ecotones

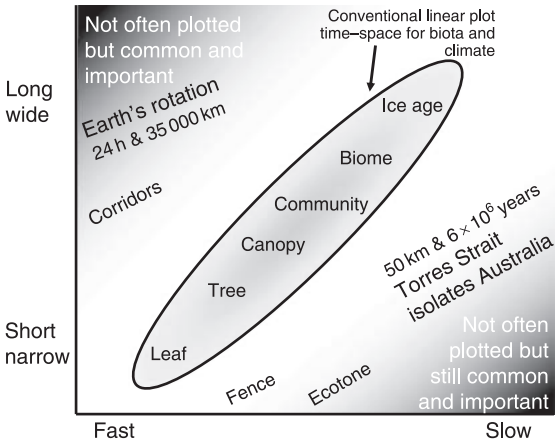


Fig. 2 A common graph appearing in landscape ecology plots increases in time against space, focusing on the quasi-linear pattern of larger things seen as behaving over longer time periods. But such plots ignore potential control systems and their hierarchies. Local intransigence can control large entities (Wallace's realms where Australia's fauna is isolated from Asia by the narrow Torres Strait, separating millions of years of evolution). Barriers and surfaces occur at the lower right, while communication channels and corridors appear upper left.

would be a case in point, because there is rapid exchange and fast process inside the abutting ecosystems or communities areas, while the exchanges across the narrow ecotone may be remarkably slow. Thus ecotones are spatially small, while representing slow exchanges that might cause the ecotone to move in a process of gradual encroachment. Conversely, in a communication channel, small differences in time constants apply along the long connection. Corridors would be an example here, where there is rapid movement along the extended length of the corridor. In ecology, these special places, such as ecotones and corridors, are at least as interesting as situations where time and space widen in concert. Complexity in hierarchies arises from the challenge of mapping between levels, as scale and definition entwine.

History of the Field

Hierarchy theory has its roots in economics and business administration of the 1960s, suggesting that the world appears nearly decomposable. We can decompose wholes into parts, but only to a degree, in that parts communicate and leak onto each other. Complete decomposability would deny upper-level structures' existence. Completely decomposed, the parts would not to be able to communicate with each other in making the larger whole. Parts have strong connections within, but weak connections between, and those weak connections may be precisely what links hierarchical levels (Fig. 3).

While some practitioners in subsequent studies have sought real hierarchies in an external world, much of the early literature of business administration hierarchies is agnostic about the ultimate reality of hierarchical structure. In this spirit, hierarchy theory in social organizations operates largely in the realm of epistemology, as a theory of observation and analysis. The discourse generally takes the position that hierarchies appear somewhere between the material world and human understanding. If there are complex

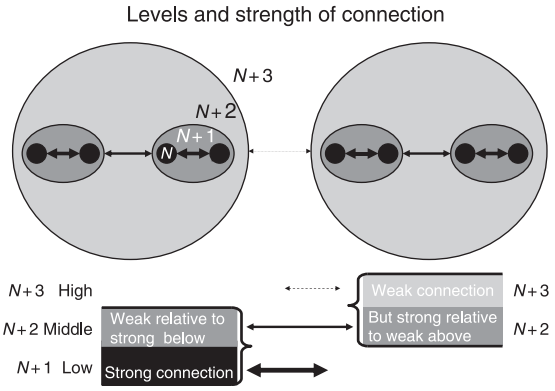


Fig. 3 In nested hierarchies, the bonds that unite members of the lowest level N are strong, as they make entities of level $N+1$. The bonds that create level $N+2$ are weaker. Nevertheless, these weaker bonds appear as the strong bonds making the entities at level $N+2$ when seen from level $N+3$. If the two largest units in the figure are molecules, then the entities at $N+1$ are atoms. Breaking their atomic bonds releases huge amounts of atomic energy as subatomic particles, N, are freed. Atomic bonds are stronger and release more energy when broken, compared to breaking the chemical bonds that make molecules, the $N+2$ entities.

material systems that are not hierarchic, we might expect to have great difficulty in observing or understanding them. There appear to be points of passage of information up hierarchies, where details are explicitly lost. A military command is a favorite hierarchic example of a human organization. There, details of how an individual soldier observed local enemy concentrations fall away as the intelligence passes up the command. Setting the detail aside allows the top brass to make sweeping decisions, without being encumbered by a blizzard of local happenings. Not only must the general in command let go of details of grains so as to get a handle on the wide extent, but so too must the observers of hierarchical structure. To understand what a general is doing, the observer of a command structure needs to integrate away the details inside the army. By the late 1960s, the notion of hierarchy had moved beyond administration systems, and was being taken up across a range of disciplines.

In the following decade, hierarchy theorists from physics addressed hierarchical complexity after Heisenberg, invoking dualities, uncertainty, and complementarity between dynamics versus structure. Important developments have turned on the tension across the dilemmas presented by dual structures, as in the holon, a generalized entity in a hierarchy. Holons have been equated with the concept of system, with the advantage that holon does not appear in common parlance. The holon can therefore escape the reification and slovenly usage in the vernacular where the model is mistaken for the materiality. Conceptual developments suggest that what is inside a given holon is chosen by an observer. This emphasizes that holons are abstractions more than material objects; a point forgotten when 'system' is used for 'holon'. In the holon, the tension is between system and subsystem. But the subsystem is a system in its own right, thus offering some sort of dual existence that invites contradiction.

The concept of the holon takes the whole to be a surface that integrates the parts to give a unified signal to the rest of the universe. At the same time, the holon is the surface that integrates the external environment for the parts to experience. In ecology, the environment falls away at the level of the holon when viewed from the perspective of the parts. A forest raises humidity and lowers temperature, thus allowing survival of some of its parts, tree seedlings that are its future. The parts are protected. Conversely, viewed from the context of the hot, dry environment surrounding the forest, the contributions of each tree to the water vapor inside the forest are lost in a more general flux of water from the canopy. Thus, loss of information occurs with movement both up and down the hierarchy. While the environment is too large to catch the details of the working of the parts, the parts themselves cannot span wide enough to see large, slow differences in the larger context. In all this, we see again the tension embodied in hierarchical discussions between scale, organization, and uncertainty in observation.

Earlier, five general principles for ordering ecological hierarchies were recognized:

(1) As to frequency of behavior, higher-level holons operate at a lower frequency, taking longer to exhibit returns in behavior than holons at lower levels.
(2) Higher levels in a hierarchy constrain lower levels by displaying intransigent constancy. Deans constrain faculty by not changing the budget, except once a year.
(3) Higher levels in a hierarchy will be contextual to lower levels. The environment would be seen as operating at a higher level.
(4) With regard to bond strength, higher-level holons are held together by weaker forces than those that integrate lower-level holons (e.g., chemical vs. nuclear bonds) (**Fig. 3**).
(5) As to containment, if higher-level holons consist of lower holons, which they contain, then the hierarchy is said to be nested. Not all the criteria apply to all hierarchies, but all five principles may apply simultaneously.

The distinction between nested and non-nested hierarchies matters (**Fig. 3**). In nested systems, upper-level entities contain and consist of lower-level entities. In non-nested hierarchies, containment is not a criterion, but principles (1)–(3) can still apply. In nested hierarchies, containment applies even if aggregation criteria between levels change type. Western medicine generally uses nested hierarchies for the human condition. Thus organelles may be aggregated into cells by biochemical interaction. Meanwhile, nesting of organs inside the whole body may invoke fluid mechanics as a principal on which parts make the whole person. When whole humans nest inside groups, relationships may be in epidemiological terms. In Western medicine, there are regular changes in aggregation criteria from biochemical, through fluid dynamic, to epidemiological. Despite inconstant criteria for linking levels, the nesting keeps such hierarchies straight. But in non-nested hierarchies, such as food chains or pecking orders, the top dog neither contains nor consists of the subordinate individuals. Because there is no nesting to maintain order, non-nested hierarchies embody only one specific rule for moving between levels. As a result, the criteria for moving up a food chain must be consistently 'is eaten by', or conversely going down it is 'eats'. In this way, the hierarchy is consistent top to bottom. Because of their robustness to changes in aggregation criteria, nested hierarchies are particularly useful for exploration before firm criteria connecting levels have been established. Concomitantly, non-nested hierarchies are for mature ideas, where focused sets of relationships are organized and abstracted in a control system.

In thermodynamic studies of ecological emergence, nested hierarchies are essential, because otherwise the bookkeeping of energy flow between the system and its environment would not sum. In complexity theory, self-organized emergence is a matter of thermodynamic gradients being applied to material systems that are pushed away from equilibrium. Thus, nested hierarchies apply when self-organization is invoked, when holons emerge at a new level without any plan. Planned systems often yield to a non-nested conception. A surprising and important new turn in applied ecology of human management systems links non-nested human socioeconomic hierarchies to nested thermodynamic hierarchies. The whole system is embodied in energy flow and control through the twinned social and biogeochemical hierarchies.

These thermodynamic approaches develop self-organizing holarchic open systems (SOHOs), using the term holarchy for nested hierarchies. The word holarchy appears in part to sidestep the political unacceptability of hegemonic hierarchical control.

Using the SOHO approach, the full power of hierarchy theory in solving real time problems has been developed by Waltner Toews and colleagues at NESH, a Canadian centered, complex systems group. They solved some critical problems in Peru, Kenya, and Nepal. For instance, a Kathmandu sewer had children playing around slaughterhouse waste. By linking the social hierarchy to the ecological process hierarchy, NESH identified that a street cleaner caste was being blamed for things out of their control. Blaming scapegoats had led to inaction and paralysis, but once the street cleaners were no longer held responsible, the SOHO thermodynamic methodology achieved significant rehabilitation as the social and ecological hierarchies began to function in concert.

The earliest explicit introduction of hierarchy theory into ecology in the 1970s spoke of decomposability as an issue in some of the biomes studied in the International Biological Program (IBP). At that time, terms, such as 'environ', 'creaon', and 'genon' were coined as extensions of the concept of holon. Environ addresses the environment acting as an integrated whole for its residents. The inward direction toward the holon pertains to the creaon, whereas the outward direction pertains to the genon generating new things and experiences for the environment and its residents. Holon remains the central concept. The first fully integrated treatments of hierarchies in ecology turned on epistemological implications of scale and dynamics. Following shortly, evolutionary ideas focused on the structural elements in hierarchies, in a more ontological spirit. The structural elements were cast as a triadic view of holons, where the level above and the level below, as well as the level of the holon in between, are all required for an adequate treatment. Recently, two more crucial levels were added: the level above the context keeps the context of the holon stable, while the level below the parts provides stability for the material of which the parts are made.

Scale and Type

Scale problems invited hierarchy theory into the discipline. Ecologists have long been aware of scale, investigating the properties of quadrats in obtaining estimates of vegetation in the 1950s. Then change in variance across quadrat size was used to measure aggregation of plants on the ground. Hierarchy theory remains associated with scale today. The observation protocol brings attention to a universe of a certain extent, while making a second distinction, the finest grain at which observation units are distinguished from one another. Grain and extent together characterize the scalar level in question in many ecological hierarchies. Grain and extent are connected. Wider extents require coarser grains, if the mass of data are to be remembered, analyzed, and understood. Modern computational power has widened the gap between grain and extent, where remotely sensed areas are captured in billions of pixels. Even so, explicitly linking items in the grain across the extent becomes difficult, and generally impossible as the extent widens by much.

In contrast to linking across scales, it is possible to unify ecology across types of ecological system that correspond to the main subdisciplines of ecology: organism; population; community; ecosystem; landscape; biome; and biosphere. These types for ecological subdisciplines are explicitly not scale based, and so are not required to be assigned to level in the order given in the previous sentence. When scale is parsed away from type, the various approaches to ecology achieve a sharper depth of focus, offering clear relief between types of investigation. The subdisciplines of ecology are not scalar levels. If they are levels at all, they are type-based levels of organization, with the different types related to one another by asymmetric relationships made explicit in the definitions. As a separate issue, a typed level of organization itself contains scale-based hierarchies, as in fractal landscapes. In that scaled universe, the ecosystem modeling strategy may apply across a range of sizes, where local processes are part of more global processes. Communities too may be variously inclusive of species across narrow or wider areas. Under the organism criterion, examples are found from redwood trees to mites. An ecological hierarchy may change the scale and type at the same time, but it is fraught with conceptual danger. Indeed, hierarchy theory is often invoked to clean up the mess in the aftermath of scale and type being mixed together. There is no prohibition changing both together, but only so long as the relationships at each new level are explicit. This matters because most descriptions of ecological material precisely do change type across widening scalar levels, although most of them do not follow the textbook ordering from organism to biosphere. For instance, in a forest community, a rotting tree trunk may be considered an ecosystem, whose upper surface is landscape, on which grows a community of bryophytes.

The copious variety of materials, entities, and sizes in ecology invites hierarchy theory into ecology. Indeed, it is in ecology that hierarchy theory has been used most often to significant effect, as in the NESH studies mentioned above. Hierarchy theory can capture a rich set of scaled examples across a mixture of types. Ecology is a multiple-scaled labyrinth of types. Hierarchy theory is the ball of string that we can trail behind, so that ecological scientists do not get lost.

Further Reading

Ahl, V., Allen, T.F.H., 1996. Hierarchy Theory, A Vision Vocabulary and Epistemology. New York: University of Columbia Press.
Allen, T.F.H., Hoekstra, T.W., 1992. Toward a Unified Ecology. New York: University of Columbia Press.
Allen. T.F.H., O'Neill. R.V., Hoekstra. T.W., (1984) Interlevel relations in ecological research and management: Some working principles from hierarchy theory. *General Technical Report R.M.110*. Fort Collins: USDA Forest Service (republished in 1987 in *Journal of Applied Systems Analysis* 14: 63–79).
Allen, T.F.H., Starr, T.B., 1982. Hierarchy: Perspectives for Ecological Complexity. Chicago: University of Chicago Press.
Kay, J., Regier, H., Boyle, M., Francis, G., 1999. An ecosystem approach for sustainability: Addressing the challenge of complexity. Futures 31, 721–742.
Koestler, A., 1967. The Ghost in the Machine. Chicago: Gateway.
O'Neill, R.V., DeAngelis, D., Waide, J., Allen, T.F.H., 1986. Monographs in Population Biology 23: A Hierarchical Concept of Ecosystems. Princeton: Princeton University Press.

Overton, W.S., 1975. Decomposability: A unifying concept? In: Levin, S. (Ed.), Proceedings of the SIAM–SIMS Conference on Ecosystems Analysis and Prediction. Philadelphia: Society for Industrial and Applied Mathematics, pp. 297–299.
Patten, B.C., 1978. Systems approach to the concept of environment. Ohio Journal of Science 78, 206–222.
Pattee, H.H. (Ed.), 1973. Hierarchy Theory: The Challenge of Complex Systems. New York: Braziller.
Salthe, S.N., 1985. Evolving Hierarchical Systems. New York: Columbia University Press.
Simon, H.A., 1962. The architecture of complexity. Proceedings of the American Philosophical Society 106, 467–482.
Waltner-Toews, D., Kay, J.J., Neudoerffer, C., Gitau, T., 2003. Perspective changes everything: Managing ecosystems from the inside out. Frontiers in Ecology and the Environment 1 (1), 23–30.
Webster, J.R., 1979. Hiearchical organization of ecosytems. In: Halfon, E. (Ed.), Theoretic Systems Ecology. New York: Academic Press, pp. 119–131.
Whyte, L.L., Wilson, A.G., Wilson, D., 1969. Hierarchical Structures. New York: Elsevier.

Relevant Websites

http://www.nesh.ca
James Kay Web Page, Network for Ecosystem Sustainability and Health (NESH).
http://www.nbi.ku.dk
Stanley N. Salthe Web Page, Center for the Philosophy of Nature and Science Studies (CPNSS), Niels Bohr Institute.

Panarchy

L Gunderson, Emory University, Atlanta, GA, USA

Panarchy is a conceptual model of how ecological systems change across scales of space and time. From the perspective of systems ecology, ecosystems are self-organized systems that cover a broad range of spatial and temporal scales. The interactions among biotic and abiotic components generate ecosystem dynamics that are both slow and fast as well as small and large. Panarchy theory attempts to explain qualitatively different types of change, by linking theories of hierarchy (how structures and processes are organized across scales) with a heuristic of adaptive cycles (how structure and process change within a given scale range).

The remainder of this article is structured into three parts. The first section defines terms of scales and dimensions in ecological systems. The second section contrasts two theoretical descriptions of hierarchy and panarchy. The third section describes cross-scale interactions.

Scale and Ecosystems

Scale has at least two meanings that relate to measurement of objects in ecological systems. One meaning of scale is the unit of measurement in any dimension. Much of ecosystem ecology is measured along dimensions of space and time. As such, meter and foot are different scales of measurement in space, and seconds and years are similar units in time. Ecosystem structures and processes are measured using multiples or fractions of these units. For example, the diameter of a tree is measured using scales of meters, centimeters, and millimeters. The other definition of scale has to do with ratios among measured units and is derived from the Latin word *scalaris* for ladder. For example, if the scale of a map is 1:250 000, then 1 cm on the map equals 250 000 cm on the ground. Both of these meanings suggest that relationships among units are set, so conversions can be made among units. The second meaning of scale, related to relationships among ecosystem components, is germane to panarchy, but requires defining two more terms.

Two other concepts are useful in understanding scales in ecological systems: grain and extent. Grain is defined as the unit of the smallest resolution of measure for a given system dimension. The resolution of a successional study may be a day or week in the time domain, at an area of a square meter in the spatial domain, depending upon what is being measured (species composition within a meter quadrat). The extent defines the bounds of measurement of a system under study. Continuing the successional study example, some long-term studies of forests may be multiple decades or even a century in extent. For two-dimensional spatial data, such as a map, the extent is also called the window of the map. In temporal data, the grain is usually defined as the minimal time unit, such as minute, day, or year, and the extent is the period of record used in analysis. Therefore, scale ranges can be determined by two components: the grain and extent.

One of the key features of ecological systems is that the components (structures and processes) cover large ranges of scales in both space and time. For example, the Atlantic Ocean ecosystem covers thousands of kilometers from the equator to almost the poles. Biological entities in the ecosystem range from the microbes whose volumes cover fractions of cubic micrometers to masses of *Sargassum* that cover thousands of cubic kilometers. In the time domain, the processes in the Atlantic Ocean vary from milliseconds to millenia.

One way of examining how ecological systems vary over scales of space is to map key structures and processes along dimensions of space and time. Ecosystem components can be mapped using the grain and extent to define ranges in these two dimensions. Across a range of scales in forested systems, leaves, branches, and trunks make up trees, trees make up stands, stands make up forests, forests make up landscapes, and landscapes comprise biomes (**Fig. 1a**). This can be identified as a vegetation hierarchy. Atmospheric hierarchy is comprised of similar structures (thunderstorms, frontal waves, El Niño Southern Oscillation) that cover ranges of scales.

A cross-scale examination of ecosystems leads to three observations. The first is that as the grain and extent of observation change, different objects (structures and processes) cover distinct scale ranges. For example, aerial photographs of forest stands cannot capture the detail of leaves (at smaller grains), nor spatial biome patterns (at larger extents). The second observation is that scalable processes and structures cover different extents in space and time. Some processes such as forest fires range from scales of a square meter to thousands of square kilometers. Other processes such as changes in carbon dioxide concentration in the atmosphere cover 20 or so orders of magnitude, from the cubic centimeters of cylinders in millions of internal combustion engines, to fossil-fuel-powered plants to regional-scale land clearing. The third observation is that ecological systems are comprised of self-organized processes that are not scale invariant; that is, they are not self-similar across scales (as measured by a constant fractal dimension). While many physical systems are self-similar or scale invariant, ecological ones are not because of the interaction between biotic and abiotic elements. Hence ecological systems are discontinuous in dimensions of space and time. These three observations lead to theories of cross-scale structures, described as hierarchy and panarchy theory.

 doi:10.1016/B978-0-444-63768-0.00695-8

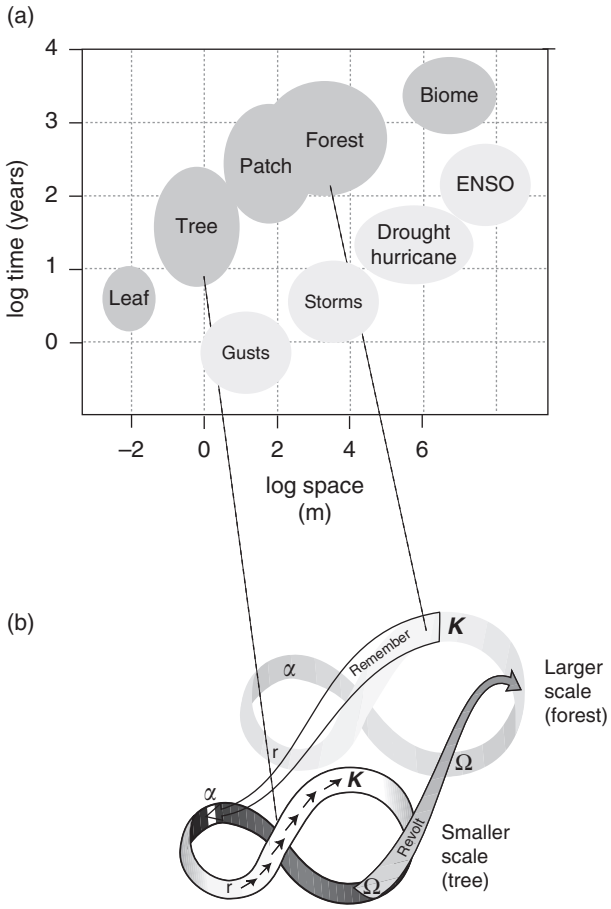


Fig. 1 (a) Cross-scale structures in a terrestrial ecosystem. Each object represents approximate grain and extent in spatial and temporal dimensions. Vegetation panarchy consists of leaves, trees, patches, forests, and biomes, while an atmospheric one is comprised of small, fast dynamics of wind gusts to global dynamics suggested by El Niño Southern Oscillation (ENSO) fluctuations. (b) Two levels of a panarchy, indicating dynamic systems within scales and cross-scale linkages or connections. Within a scale, systems undergo four phases of change. Initially, systems begin with a rapid phase (r phase) of growth. Growth slows as the system becomes connected and structure accumulates in the K phase. As the system is perturbed, rapid phases of creative destruction (Ω) and reorganization (α) occur. At key times, the release or creative destruction can spread to larger and broader scales. The up-scale connection or 'revolt' occurs when smaller-scale releases intersect with broader-scale vulnerability. Following release phases, often the system relies on larger- and longer-scale processes during periods of reorganization. The down-scale 'remember' connection facilitates renewal by drawing on the potential that has been accumulated and stored in a larger, slower cycles. National- or international-scale relief efforts following destructive storms or tsunamis are examples of remember connections.

Hierarchy and Panarchy

One major conceptual framework for understanding ecological cross-scale structure and dynamics is hierarchy theory. Ecologists in the 1970s and 1980s built on the seminal work of Herbert Simon to develop a theoretical base that emphasizes a pattern of aggregations (hierarchical levels, or 'holons') that are nearly separable across scales. Hierarchical levels can be identified by a stronger set of interactions within a hierarchical level than between levels. These hierarchical levels correlate to scales, and each level has characteristic spatial and temporal domains; that is, each level has a characteristic turnover time and spatial domain, as indicated in Fig. 1.

The interaction of hierarchical levels across scales has been the subject of much debate and is based on the interaction between slow (and broad) structures and processes and those that are fast and small. These interactions can be characterized in two ways: (1) hierarchical control and (2) panarchical relations. Hierarchical control is demonstrated when slow, broad features constrain and control the small, fast ones. This is an example of top-down control, when variables such as geology and soil types interact with climatic variables (temperature, photoperiod, rainfall) to determine the suite of plant and animal species that thrive. Much empirical evidence supports hierarchical or top-down controls. Panarchy theory was proposed to suggest that both top-down and bottom-up interactions occur; that is, while top-down control does exist, there are many bottom-up or cascading phenomena that occur. Many disturbance dynamics, such as forest fires or forest pest outbreaks are not the result of top-down control or control by slower variables, but examples where faster, smaller variables appear to control the system for periods of time. The interaction of processes and structures at different scales is one of the conceptual foundations of panarchy theory.

A critical feature of such hierarchies is the asymmetry of interactions among hierarchical levels. The larger, slower levels constrain the behavior of faster ones. In that sense, therefore, slower levels control faster ones. If constraint and control were all that mattered, then systems would have little, if any, opportunity for change and experimentation. For example, a forest stand moderates the climate within the stand to narrow the range of temperature variation that individual organisms experience. But missing in this representation is the dynamic of each level that is organized in the four-phase cycle of birth, growth and maturation, death, and renewal. That adaptive cycle is the engine that periodically generates the variability and novelty upon which experimentation depends. As a consequence of the periodic, but transient phases of creative destruction (Ω stage) and renewal (α stage), each level of a system's structure and processes can be reorganized (**Fig. 1b**). This reshuffling allows the possibility of new system configurations and opportunities from the incorporation of exotic and entirely novel entrants that had accumulated in earlier phases.

The adaptive cycle is a metaphor of temporal change in ecological systems. It suggests that systems at specific scale ranges exhibit four distinct and usually sequential phases of change in the structures and function of a system. As systems begin to form (such as primary or secondary succession in ecosystems), systems exhibit a growth phase. The growth phase is characterized by relatively rapid accumulation of structure (biomass and complexity). During this phase, competition is a scramble for resources, as winners are able to obtain and quickly convert raw materials to structure and organization. The first phase is also called the r phase, in reference to the specific growth rate parameter in an exponential growth model. Over time, structure accumulates and the system becomes more diverse and more connections appear among the system components. Gradually, system net growth slows, as more of the acquired resources and energy are allocated to system maintenance rather than growth of new structure. The accumulated structures and resources are conserved, hence the designation of the next phase is the conservation phase. This is also called the K phase in reference to the carrying capacity parameter in the logistic equation. During this phase, the system becomes increasingly connected, less flexible, and more vulnerable to external disturbances. These two phases, r to K, correspond to ecological succession and have been described as optimizing growth and energy throughput. These phases also represent system maturation and increasing vulnerability to external variations or disturbances.

Variation in inputs external to the system intersecting with increasing internal vulnerability leads to the next phase of the adaptive cycle. External disturbances generate a sudden release of accumulated capital or structure. This phase is called a period of creative destruction or the omega (end) phase. In pyric communities, for example, fuel slowly accumulates since the time of the last fire. As fuel accumulates, the system becomes more vulnerable to sources of ignition at small scales with an increasing probability of fire. Forest fires, grass fires, pest outbreaks are examples of the Ω phase, in which the structure is quickly destroyed. The collapse or release during the Ω phase is quickly followed by a reorganization (α) phase, where a new system emerges, leading to the growth phase of a new cycle. The new r phase may be very similar to the previous r phase, or it may be quite different. This pattern of rapid, then slowing growth, swift destruction and reformation, has been observed in many systems. These include ecological examples, such as pest outbreaks and fires in temperate forests; and social-ecological systems such as water management history of the Everglades, and aboriginal cultures in the eastern and western US.

In summary, many (but not all) systems exhibit a sequential series indicated in the adaptive cycle as it proceeds from an exploitation phase (r phase) slowly to conservation (K phase), very rapidly to release (Ω phase), rapidly to reorganization (α phase), and rapidly back to exploitation.

The dynamics conceptualized in the adaptive cycle are for systems at a particular scale range; that is, the dynamics of growth, conservation, destruction, and renewal can be observed for specific ranges of structures (**Fig. 1b**). Most plant leaves exhibit the phases on an annual cycle, growth senescence, fall, and the emergence of new buds. This cycle is driven in large part by external variation in seasonal cycles of insolation and temperature. Patches of forest go through these phases of succession on cycles of decades, as indicated by periodicities of fire or pest outbreaks. The blend of the adaptive cycle heuristic with older hierarchy theory forms the nucleus of panarchy theory.

Cross-Scale Interactions

A panarchy has three ingredients, (1) subsystems of adaptive cycles that represent system dynamics at a specific scale range, (2) dynamic systems that occur at different scale ranges (hierarchical levels), and (3) coupling of those systems across scales. All of these structures are posited to change in phases described by the adaptive cycle, but at a given scale. Panarchy links these structures across scales, and suggests that interactions can go from smaller, faster levels to broader, slower levels (or up-scale connections). Panarchy also suggests that the slow and broad processes and structures influence those that are faster and smaller (i.e., the subsystems and down-scale influences). But these connections are ephemeral, becoming dominant at certain times and dormant at others.

There are potentially multiple connections between phases at one level and phases at another level. The two most significant are the connections labeled as 'revolt' and 'remember'. **Fig. 1b** indicates two levels of a vegetation panarchy, each at distinct scale ranges or domains, and linked by cross-scale connections of revolt and remember.

Up-scale linkages or connections have been named 'revolt', suggesting that small events can cascade up to larger scales. Most of the time, the larger, broader variables control the smaller and faster processes. This is classic hierarchical control. However, when a level in the panarchy enters an Ω phase of creative destruction and experiences a collapse, that collapse can cascade up to the next larger and slower level by triggering a crisis, particularly if that level is at the K phase where resilience is low. One example is in the

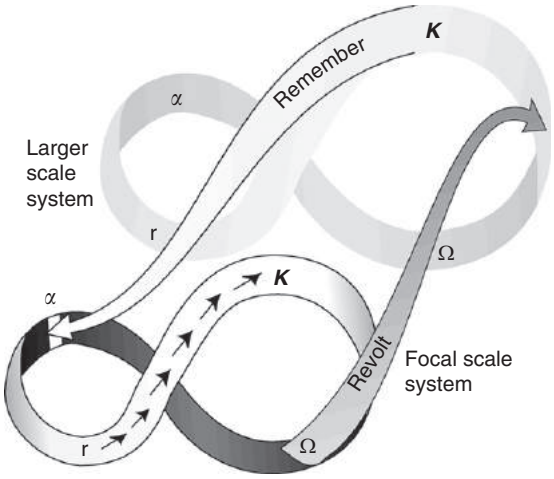


Fig. 2 Two key cross-scale connections of a panarchy are revolt and remember. The revolt pathway depicts how small-scale (local and fast) variables can interact to create an upscale cascade. Revolt dynamics have been described for forest pest, forest fires, and disease outbreaks. Resources in the form of capital, connections, and memory from larger systems often help in the recovery of focal scale collapse, as indicated by the remember pathway.

dynamics of fire-prone ecosystems. The lighting of a match in a forest, or a strike of lightning is a small, local phenomena. Under many conditions the local fire is either quickly extinguished or never begins a fire. However, under certain conditions (such as extreme droughts or low humidity), local ignitions can create a small ground fire that spreads to the crown of a tree, then to a patch in the forest, and then to a whole stand of trees. Each step in that cascade moves the transformation to a larger and slower level. A societal example occurs when local activist groups succeed in efforts to transform regional organizations and institutions, because they had become broadly vulnerable. Hence part of the connotation of revolt is used to describe how fast and small events overwhelm slow and large ones. And that effect could cascade to still higher slower levels if those levels had accumulated vulnerabilities and rigidities.

The down-scale interactions in panarchy are captured by the phrase 'remember'. As shown in **Fig. 2**, this type of cross-scale interaction is important at times of change and renewal. Once a catastrophe is triggered at a level, the opportunities and constraints for the renewal of the cycle are strongly organized by the K phase of the next slower and larger level. After a fire in an ecosystem, for example, processes and resources accumulated at a larger level slow the leakage of nutrients that have been mobilized and released into the soil. The options for renewal draw upon the seed bank, physical structures, and surviving species that form biotic legacies that have accumulated during the growth of the forest. It is as if this connection draws upon the accumulated wisdom and experiences of maturity – hence the choice of the word remember.

Panarchy theory describes at least three categories of change in complex ecological systems. These categories include gradual or incremental change, adaptive change, and transformative change. Incremental changes occur slowly and predictably, as systems mature. Adaptive change occurs after disturbances, when the system has the potential to change into a qualitatively different state. Alternative states in ecosystems occur when resilience of the system is exceeded. That alternative state occurs at one level of the panarchy. Transformational change occurs when multiple levels of the panarchy change as a result of cross-scale linkages and lead to fundamentally new types of structures, processes, and controls.

Summary

Panarchy is a term to describe cross-scale interactions in ecological systems. Other concepts such as hierarchy theory, bottom-up or top-down control are part of the systems ecology lexicon. Panarchy was proposed as a way of synthesizing these concepts. The derivation of the word hierarchy is from terms that suggest 'sacred rules'. As such, hierarchy suggests rules that are inviolate and top down. Much ecological research suggests the widespread presence of these relationships, where the broad and slow constrain and control the small and the fast. In contrast the derivation of panarchy was from Pan, the Greek god of nature. Panarchy in this sense is thus proposed as 'nature's rules'. It blends system dynamics at a particular scale domain that suggests that at least four phases of changes (growth, conservation, release, and reorganization) are observed in many systems. During the growth and conservation phases, internal controls dominate system dynamics. The release and reorganization phases are when cross-scale connections emerge, and the system becomes linked to dynamics and processes occurring at other scales. Panarchy also suggests that three types of changes occur in ecological systems. Incremental change occurs during growth to conservation phases of a system at a single scale. Adaptive change occurs as result of four-phase dynamics at a single scale, when the system can change structure and processes as a function of ecological resilience. Transformational change occurs when multiple levels of panarchy all undergo state changes.

Further Reading

Gunderson, L.H., Holling, C.S. (Eds.), 2002. Panarchy: Understanding Transformations in Human and Natural Systems. Washington, DC: Island Press.
Gunderson, L.H., Holling, C.S., Light, S.S., 1995. Barriers and Bridges to the Renewal of Ecosystems and Institutions. New York: Columbia University Press.

Population Dynamics: Stability☆

Peter A Henderson, Pisces Conservation Ltd., Lymington, United Kingdom and University of Oxford, Oxford, United Kingdom

Glossary

Density-dependence An effect in which the intensity changes with the increasing population size.

The abundance of all organisms varies through time and across space and many populations can occasionally undergo dramatic exponential change. No population can remain at a stable equilibrium in which the number of individuals is constant because deaths and births cannot be in perfect balance. However, many populations do not show the explosive increases in number that their reproductive potential would allow, nor the violent collapses that their pathogens or predators could potentially inflict. This constraint on abundance can be viewed as a form of stability and has been long recognized.

Before examining how populations may be stabilized, it is informative to briefly examine typical time series to see the way populations do change through time.

Common Patterns of Dynamic Behaviour

Populations can show a wide range of dynamic behaviors and a great variety of dynamics can occur contemporaneously within a single community. To illustrate this point all of the examples of different dynamic behaviour shown below are for species living together in Bridgwater Bay in the Bristol Channel, England between 1981 and 2018.

Seasonal Variation

The most common form of variability observed in the natural world is a seasonal change in numbers. Seasonal changes in abundance occur in both temperate and tropical latitudes. Whereas temperate changes relate to seasonal variation in sunlight, temperature, snow cover and wind, tropical seasonality is frequently related to rainfall or flooding. **Fig. 1** shows the clear seasonality in occurrence of the dab, *Limanda limanda*, a small flatfish which is only present in Bridgwater Bay during the autumn and early winter. The population is dominated by small fish only a few months old which are the size of a postage stamp. For the rest of the year dab lives in deeper, cooler waters.

Not all species show a single seasonal peak.

It is important to remember that many species are annual and for part of the year exist as resting stages such as eggs or pupae.

Random Variation Without Trend

Many populations show seemingly random between year fluctuations in abundance. For example, **Fig. 2** shows the abundance of whiting in Bridgwater Bay over a 37-year period. Note that whiting abundance jumps up and down between years but there is no long-term increase or decrease in the average abundance.

Common causes for this noise are changes in climatic conditions or food availability during the breeding season. However, for many species the cause may be far from apparent and quite likely multifactorial. A cursory examination of such time series might lead to the conclusion that these populations show no stability and the natural world is following a chaotic, wild, dance. However, there are clear indications that these are not the steps of a random sequence because the population does not progressively move to ever higher or lower levels of abundance with time. As shown in **Fig. 7**B, the level of temporal variability stabilizes as the time series lengthens, if it were a random walk it would continue to increase with time. Such populations may be considered stable because the fluctuations, however violent, are all around a long-term mean abundance that does not change. In fact Bulmer's test (Bulmer, 1975) when applied to the 37-year whiting abundance time series indicated significant levels of density-dependence ($R = 0.625$ which is smaller than $R0.5 = 1.53$). Essentially, after each perturbation the population tends to move back toward the same mean level.

☆*Change History:* March 2018. Peter A Henderson. The time series graphs were revised and increased in length (years of observation). A new section 'Density and Dependence and Stability Within a Single Community' was added.

This is an update of P. Henderson, Stability, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 3334–3341.

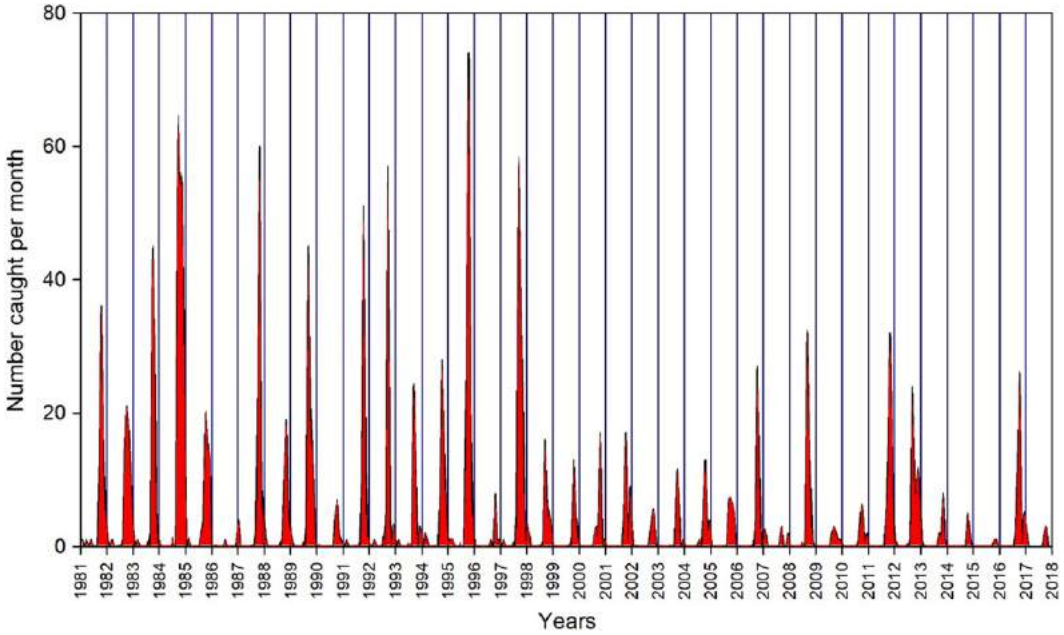


Fig. 1 The seasonal variation of the flatfish, *Limanda limanda*, in Bridgwater Bay, England. The time series was derived from monthly sampling and clearly shows an annual peak in abundance every autumn. The decline from 1999 is related to climate warming as this species avoids warmer waters.

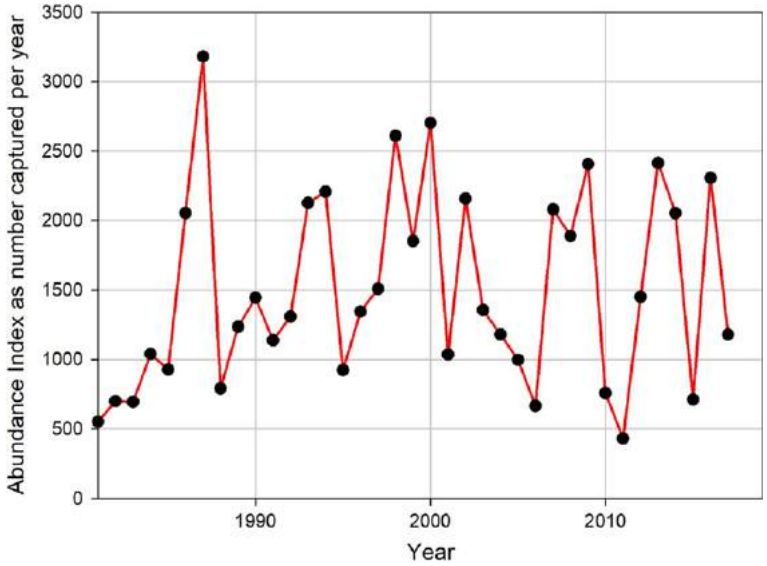


Fig. 2 The annual variation in abundance of whiting, *Merlangius merlangus*, a gadoid fish, in Bridgwater Bay, England. This species shows considerable between-year variation in number without any long-term trend (difference sign test, statistic = 0.56, $n = 37$, $P = 0.5741$).

Population Outbursts

Some populations maintain a generally low population size, but occasionally undergo abrupt bursts in number. Such behaviour is frequently observed with pest species, which in some years may be almost undetectable yet a few years later are present in huge numbers. Fig. 3 shows the between year abundance of the gurnard, *Eutrigla gurnardus*, a small bottom feeding fish. In 1992 there was a sudden peak in abundance, followed by a long period of low abundance before a second, much higher, peak in 2007 and subsequent lower outbursts in 2013 and 2016.

The reasons for the outbursts are often obscure, but, possibly relate to optimal conditions during the spring for larval survival and the arrival of large numbers of juvenile migrants into the study area. Again, there are indications of stability in the system as

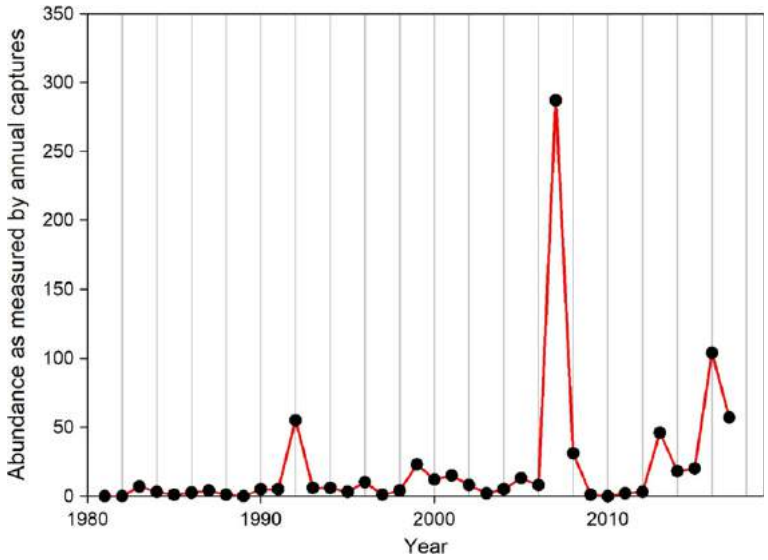


Fig. 3 The change in annual abundance of the gurnard, *Eutrigla gurnardus*, in Bridgwater Bay, England. While there is some evidence for a small upward trend in numbers, the most striking feature of this time series is the greatly increased abundances in 1992, 2007, 2013, and 2016.

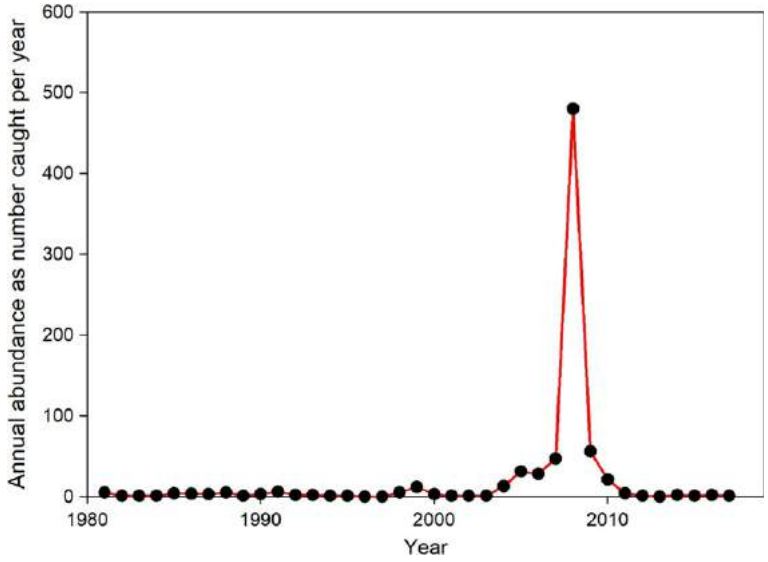


Fig. 4 The annual abundance of the snake pipefish, *Entelurus aequoreus*, in Bridgwater Bay, England showing the major outburst in abundance that occurred between 2003 and 2011.

the population explosions are contained and consumed so that the population equally quickly subsides to the basal level. Within a few years there is frequently no trace of the previous population explosion.

While the gurnard outburst described above comprised recently born individuals and can be thought of as reflecting years when reproduction was especially favorable some outbursts comprise adult age groups and do not reflect isolated incidences of reproductive success. An exceptionally clear example was the outburst of snake pipefish, *Entelurus aequoreus* in British waters between 2003 and 2011 (Fig. 4). The rise up to the peak abundance in 2008 is even faster than exponential growth and cannot be explained by reproduction alone. It appears that environmental conditions caused production from a wide region of sea to become concentrated in inshore waters. It is often the case that sudden bursts in population density are linked to migrations or environmentally forced concentrations into small regions of habitat (Fig. 4).

Exponential Increase and Decline

If the instantaneous rates of mortality and birth are consistently different for an extended period a population can show an exponential change in number. Fig. 5A and B shows that the large prawn, *Palaemon serratus*, underwent an approximately

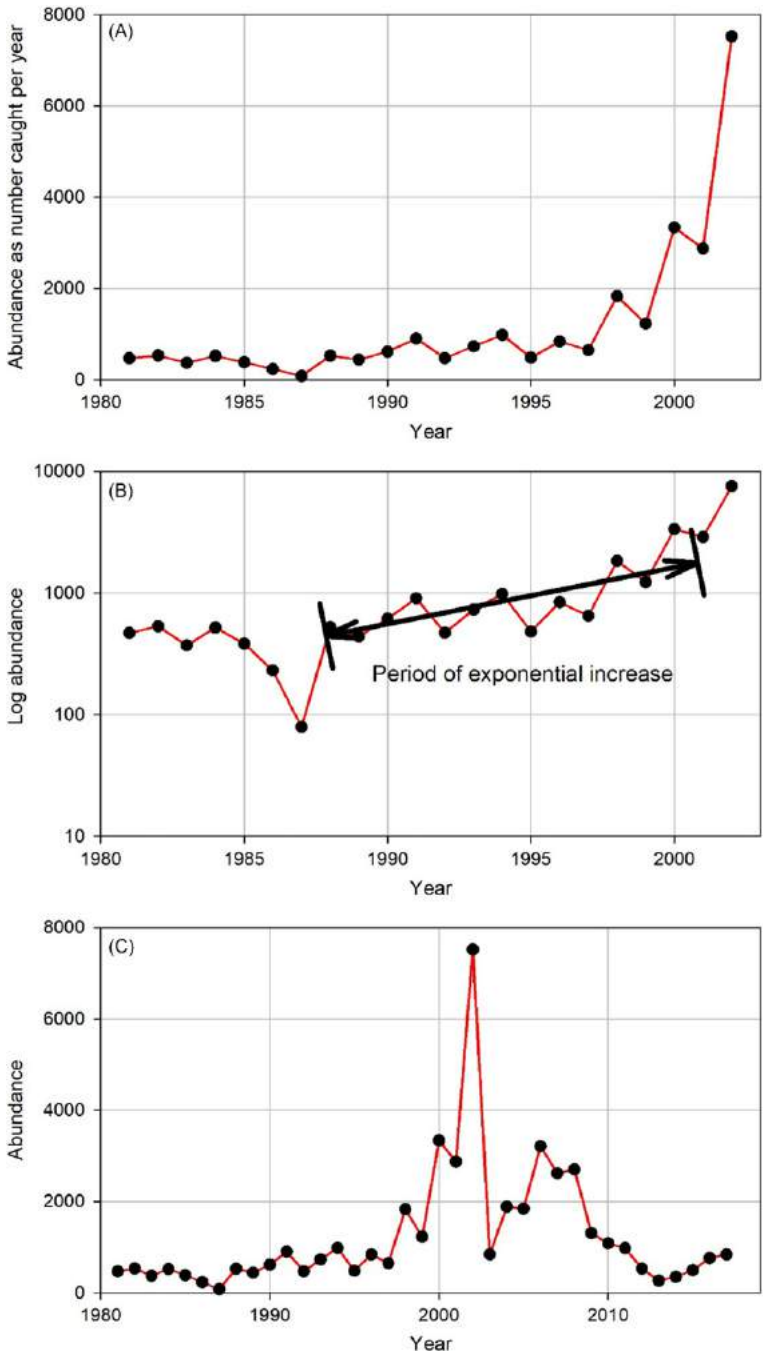


Fig. 5 The change in annual abundance of the prawn, *Palaemon serratus*, in Bridgwater Bay, England. (A) This species underwent an exponential increase in abundance between 1987 and 2002. (B) The exponential growth period is shown as a straight line in the log abundance plot. (C) The population trajectory for the longer period 1981–2017 shows that after 2002 the population then suddenly fell back in abundance.

exponential increase in abundance between 1987 and 2002 which is shown as a good fit to a straight line on the log plot (**Fig. 5**B). A major downward correction then occurred (see **Fig. 5**C), which was at some stage inevitable as any population growing exponentially must eventually experience resource limitations. The increased abundance of this prawn between 1980 and 2002 was probably linked to raised seawater temperatures allowing increased growth and a longer reproductive season. In recent years, average water temperatures in Bridgwater Bay have been declining, and the population has fallen to levels previously observed in the 1990s. It seems likely that the exponential growth phase was a short-term response as the population increased to take possession of a larger niche which has now been lost.

A clear example of exponential decline is shown by the common eel, *Anguilla anguilla* in **Fig. 6**. This dramatic decline toward possible local extinction of a fish that was once highly abundant is probably linked to a multifactorial increase in mortality rate.

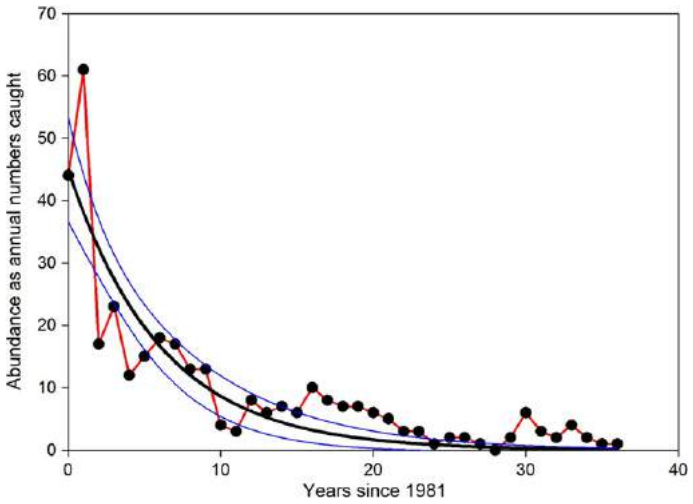


Fig. 6 The exponential decline in the annual abundance of eel, *Anguilla anguilla*, in Bridgwater Bay, England. This rate of decline has been observed at other localities across Europe. The smooth *black line* is an exponential curve fitted by regression and the two *blue lines* are the 95% confidence bands.

Man has been responsible for over-fishing, producing impediments to migration such as weirs and dams and finally introducing a deadly parasite. While such declines are becoming common because of human destruction, they can also occur naturally because of the outbreak of disease or climate change. This population decline must free resources allowing other species to increase as they use the resources previously taken by the eels.

The Balance of Nature

Our present understanding of the stability of natural populations shows a gradual, if uneven, increase in understanding that is still far from complete. That there is a balance of nature, so that all species can continue to coexist, is a common and ancient worldview and has been long held in western thought. Until the 20th century it was widely assumed that the sizes of populations were held in balance by divine power, occasional outbreaks of pests and a loss of balance, such as a plague of locusts, were seen as divine punishment. Belief in this benevolent balance gradually eroded, until C. S. Elton in 1930 could assert "The balance of nature' does not exist, and perhaps never has existed. The numbers of wild animals are constantly varying to a greater or less extent, and the variations are usually irregular in period and always irregular in amplitude." However, no 20th-century ecologist, including Elton, viewed populations as simply random variables changing without constraint for it was realized this would rapidly lead to extinction and the collapse of ecosystems. The natural world shows order that needs to be explained. The tightness of the constraints exerted by the environment, both biotic and abiotic, and the nature of the factors impacting population density are presumed to influence the degree of population stability. But, in addition, it came to be realized that the life history strategy followed by the organism also had an influence and was itself undergoing natural selection. Some species had evolved to produce large numbers of offspring and were capable of explosive growth when conditions allowed. In contrast, others produced a small number of offspring and their populations were considerably more stable. These two strategies were termed r (for reproductive rate) and k (for carrying capacity) selected. The populations of r-selected species were viewed as more variable than the k-selected species, which remain closer to their environmental carrying capacity. The distinction between r- and k-selected species is now less frequently applied in ecological arguments, in part because the distinction is too rigid. How population stability is determined and regulated remains a major area of research and is still capable of generating considerable controversy.

The Functional School and the Role of Density-Dependent Biotic Interactions in Population Regulation

The debate on how populations are regulated began in the mid-1930s when A. J. Nicholson proposed that controlling factors must act density-dependently. While a huge body of evidence has been collected demonstrating the existence of density-dependent mechanisms, the debate on their general role in determining population size and stability is still ongoing. Nicholson's key argument was that regulatory agents must increase their impact on a population as the size of the population increases. It is this change in impact with density that is termed density-dependence. Because predators, pathogens, and parasites can change their population size in response to changes in host abundance and thus change their level of impact with the host population size, they were viewed as key regulatory candidates. The core role of biotic factors in density-dependent regulation came to be a central belief of ecologists of the functional school. This view implies that populations have a long-term equilibrium size or density from which

they may be displaced by stochastic factors and toward which they are returned by density-dependent forces. However, the proponents of this view do not suggest that populations are necessarily held tightly to this equilibrium. As D. L. Lack stated in the 1954 "most wild animals fluctuate irregularly in numbers between limits that are extremely restricted compared with what their rates of increase would allow."

The Environmental School and the Role of Climate and Other Environmental Factors in Limiting Population Growth

In the 1950s a radically different view on the nature of regulation from that presented by Nicholson was proposed. In contrast to the views of the functional school, H. C. Andrewartha and L. C. Birch argued that density-independent agents such as climate and resource availability limited the growth of populations and there is only limited time when they had the opportunity to increase. Termed the environmental school, this viewpoint emphasizes constraint by predominantly physical conditions and by implication argues against the existence of long-term equilibrium population sizes. A strong argument for this viewpoint is given by T. C. R. White, "Surviving on this earth is, and always has been, especially for the very young, a struggle, a chancy business…. Nor is there an "optimum" or "equilibrium" density of a population in nature—only the maximum number that can survive each generation in a population that is pressing hard against the variable but limited supply of resources in its environment." From this viewpoint populations are constantly expanding and being pushed back by adverse events. Through time, they may occasionally expand exponentially, then, as conditions change, suddenly decline. This is a reasonable explanation for the dynamics of the Atlantic prawn graph shown in **Fig. 5**.

Combining Biotic and Physical Factors

Many ecologists take a hybrid position between the functional and environmental schools, noting the powerful influence of climate and other physical features while also acknowledging the existence of density-dependent regulation. Some have advocated a revision in terminology, arguing that, in part, the disagreement between ecologists on the nature of population regulation can be linked to poor or confusing terminology. A. A. Berryman suggests that ecology should use the vocabulary of dynamic systems theory and replace the term density-dependence by negative feedback. Negative feedback occurs whenever the rate of change of a population (dynamic variable) is inversely proportional to its current or past size (states). For a population to be regulated, a negative feedback must operate. It is clear that the individual scientist's viewpoint on population regulation is heavily influenced by field experience. Those working on large mammals or birds are far more likely to emphasize biological, density-dependent possesses than those working on single-celled organisms or small fast breeding insects. Entomologists frequently observe population explosions or collapses linked to a change in environmental conditions.

It is striking that much of the argument about population stability and the importance of density-dependence, population regulation, and species interactions such as competition has not been based on empirical observation and has been heavily influenced by simple mathematical models. This is notable, as generally ecologists put great emphasis on the importance of field observation. More recently, there has been a reduced emphasis on gaining insights from mathematical models as it came to be realized that quite simple models could produce an astonishing richness of dynamical behaviors. In effect, everyone could find a simple model to demonstrate that his or her view had a theoretical basis. For much of the 20th century the emphasis on theoretical considerations was inevitable as ecologists had too few long-term population dynamical time series to study. While all-natural populations vary in numbers in space and time, surprisingly little was and still is known about the statistics of this natural variability. The shortage of empirical data was, in part, because ecology is a relatively new science and population data was not systematically collected until the later part of the 19th century so even now we have few reliable data sets extending for longer than 100 years. When it is remembered that many larger species of animals can live for thirty or more years and large plants for hundreds of years, it is apparent that our time series cover too few generations to show the full dynamical range of long-lived organisms. In part, studies of fossils have supplied the long-term perspective required; however, such studies can rarely give reliable estimates of relative abundance. A continuous stream of new observations is now rapidly increasing the database upon which ecologists can test hypotheses and measure the degree of stability actually shown.

Measuring the Variability of Populations

Many measures of variability have been proposed, most of which are unsuitable for ecological purposes because they do not measure proportional change in density or population number. We generally require a measure of variability which gives the same result for a population which changes from 0.1 to 1, 10 to 100, or 1000 to 10,000 as in each case the population has increased 10-fold. Such a measure is then not dependent on the size of the sample. It is now recommended to use either the standard deviation of the natural logarithm of the abundances or the coefficient of variation of the abundances (CV(N)). As many time series of abundance have occasional zero values, CV(N) is the preferred measure as the logarithm of zero is undefined.

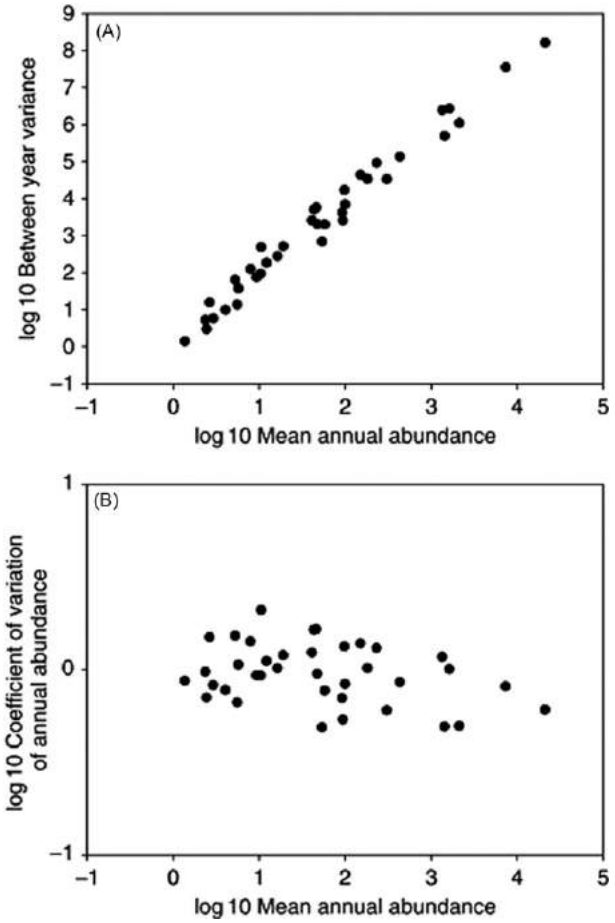


Fig. 7 (A and B) A comparison of the dependence of variance and coefficient of variation on the mean number caught per annum. The data are for the resident fish and crustaceans captured from monthly sampling over a 25-year period in Bridgwater Bay, England. These plots are based on the number caught per annum which may not be directly proportional to the actual number present. This is particularly the case for sampling methods such as trapping that can become saturated at high densities.

The merit of CV(N) over a nonproportional measure such as the variance of the time series is illustrated using as an example the 25-year variability of fish and crustaceans in Bridgwater Bay, England. Fig. 7A shows clearly that the variance is highly positively correlated with mean abundance and therefore is a useless measure of population variability as it will inevitably show that the most abundant species are the most variable. This linear relationship between log variance and log mean has long been recognized and is referred to as Taylor's power law. In comparison, the coefficient of variation of annual abundance does not always increase with the mean (Fig. 7B). In fact, the CV(N) for the Bridgwater Bay animals indicates that proportional variability tends to decrease with increasing abundance and hints at an increased stability with increasing population size (this is investigated and explained in the following section). This observation is important, as the choice of the CV(N) as the preferred measure of variation is not to remove all dependence on population size. Even if this were possible, it would be undesirable as we frequently wish to identify stabilizing density-dependent processes, which change the dynamics as mean abundance changes. The standard deviation of the logarithm of the abundances þ1 has been suggested as a measure of variability which can be used with time series that include zeros. It should be avoided as it is biased.

Having settled on the coefficient of variation CV(N) of the time series as an appropriate measure of variability, it is essential, when comparing population variability, to compare time series over the same time period. This is because the CV(N) tends to increase with the increasing length of the time series. This is illustrated in Fig. 8 using a 25-year time series of crustacean and fish data. While the common shrimp time series shows a long-term trend of increasing CV(N) (Fig. 8A), the whiting data (Fig. 8B) shows a time series for which the CV(N) initially rises with duration and then stabilizes. A tendency for the CV(N) to stabilize with time is indicative of population regulation and thus may be a general feature of time series monitored for sufficient time to display detectable density-dependent regulation. It is notable that the whiting time series showed no long-term trend in abundance and shows significant density-dependence using Bulmer's test (see Fig. 2). However, there is little empirical data on this subject and most time series have the potential to suddenly increase in variability. The jump in variability of the common shrimp population after 20 years (Fig. 7A) was linked to a sudden increase in water temperature from 1999.

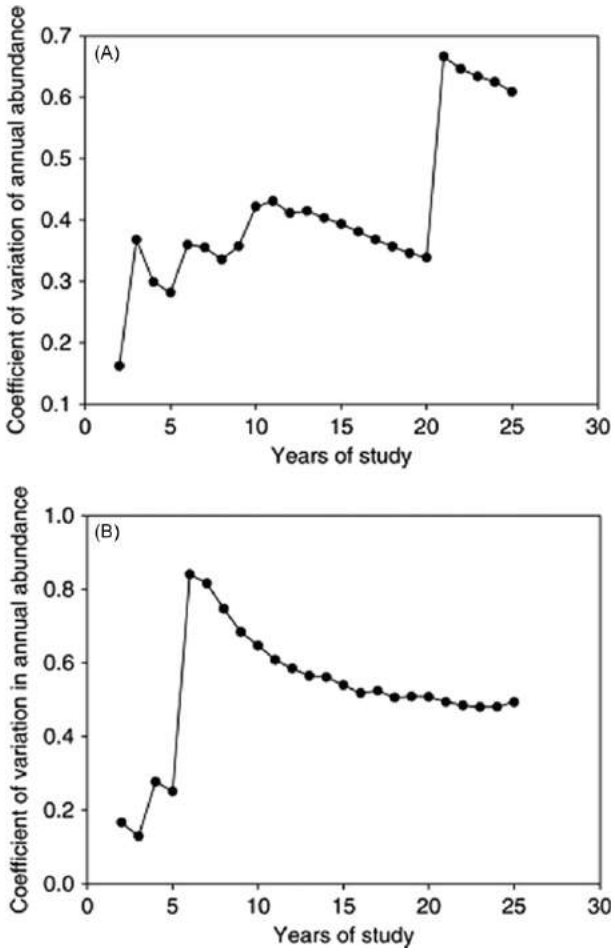


Fig. 8 (A and B) The change in the coefficient of variation with the length of the time series for (A), the brown shrimp, *Crangon crangon* and (B) whiting, *Merlangius merlangus*. The data were collected by monthly sampling over a 25-year period in Bridgwater Bay, England.

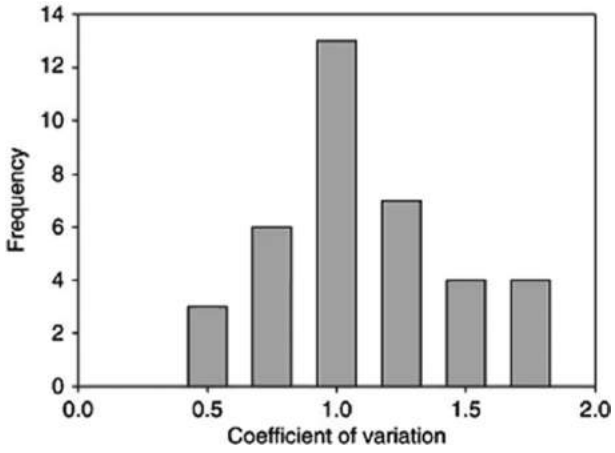


Fig. 9 The distribution of the coefficient of variation in annual abundance for the resident fish and crustacean species in Bridgwater Bay, England. The data were collected by monthly sampling over a 25-year period in Bridgwater Bay, England.

Once a measure of variability is defined and it is also noted that a time-series variability should only be compared between populations over the same time period, it is then possible to examine comparative population stability. Fig. 9 shows the frequency histogram of the CV(N) in annual abundance for all the species of fish and macro-crustaceans resident in Bridgwater Bay in the Bristol Channel, England. All the populations were monitored contemporaneously for a 25-year period commencing in 1981. It

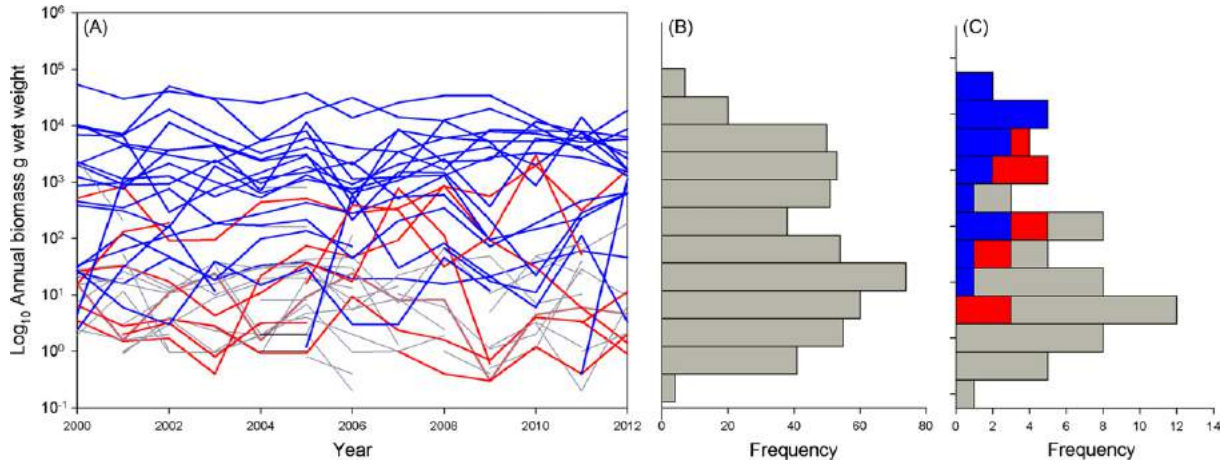


Fig. 10 Temporal variability within the fish community of Bridgwater Bay. (A) The variation in annual biomass. Core species showing density-dependence are shown in *blue*, core species with no evidence of density-dependence in *red*, and transient species in *gray*. (B) The frequency distribution of fish abundance over the 12 years combined. (C) The frequency distribution of average abundance over the 12-year period.

can be seen that CV(N) is approximately normally distributed around a mean of one with a slight skew toward higher CV(N) values. This skew would appear to be related to human impacts and climate change as high CV(N) populations are dominated by species that have been showing a trend in abundance over the entire 25-year period. Some species such as the herring, *Clupea harengus*, which has a CV(N) = 1.6, are slowly recovering from a population collapse caused by overfishing. Others, such as the gurnard, *Eutrigla gurnardus* (CV(N) = 1.4, see **Fig. 3**), have been increasing in abundance as seawater temperatures have increased. In summary, a wide range of vertebrate and invertebrate species, differing greatly in life history characteristics, display a notably small range of CV(N) between 0.5 and 1.5 over a 25-year period, suggesting that they all are tending to be constrained toward a similar level of population stability. Unfortunately, we do not have similar long-term data sets covering all the common vertebrates and invertebrates for a terrestrial ecosystem to test if this is a common feature of other communities.

It is possible to determine if the fish in Bridgwater Bay are showing a similar degree of variability to terrestrial vertebrates. B. Saether and colleagues in 2003 presented data for the CV(N) for 13 solitary birds over a 15-year time period. Values ranged from 0.08 for the sparrowhawk to 0.71 for the cactus finch with a mean of 0.31. In comparison, the 23 nonshoaling fish from Bridgwater Bay had CV(N)'s over a 15-year period of between 0.45 and 1.9 with a mean of 0.88. It is clear that North American birds have, on average, considerably more stable populations than estuarine fish. This difference is probably related to differences in reproductive behavior. Most of the fish have high fecundities and limited or negligible parental care and therefore their populations are able to respond rapidly to changes in climatic conditions. In general, the CV(N) in annual abundance varies greatly between habitats, taxonomic groups, and species because of differences in the reproductive and life history strategy and the degree of environmental variability. From an extensive study of bird population dynamics, B. Saether and colleagues in 2004 demonstrated that the stability of bird populations is clearly related to their life history strategy and concluded that "The demographic stochasticity decreased with adult survival rate, age at maturity, and generation time or the position of the species toward the slow end of the slow-fast life history gradient."

Density and Dependence and Stability Within a Single Community

An analysis of the Bridgwater Bay community shows how the views of the functional and environmental schools of population regulation are reconciled given sufficient data. Magurran and Henderson in 2003 noted that the Bridgwater Bay community comprised two groups of species they termed core and transient respectively. The core species are permanent members of the community and tend to have relatively high abundance population size in terms of biomass. In comparison, transient species are only occasionally present. These tourists tend to occur when the environment particularly favors their needs, but cannot maintain their populations in the long-term. Subsequently, in 2014, Henderson and Magurran showed that the core species, which have by far the greatest population size in terms of biomass, are under density-dependent control (**Fig. 10**). In comparison, the transient species tend to have highly variable populations often responding to physical conditions such as water temperature, ocean currents and wind. **Fig. 11** shows that for the Bridgwater Bay community population variability, as measured by the coefficient of variation, is far lower in core species under density-dependent control.

These results show that species make different contributions to the stability of a community. Species experiencing density-dependence exhibit relatively little temporal variation. These species also typically account for a large fraction of the overall abundance; in the case of the Bridgwater assemblage they represent >98% of the total biomass. Because of the stability in biomass of these core species it is inevitable that nutrient and energy flux will also be stabilized. The results demonstrated that density-

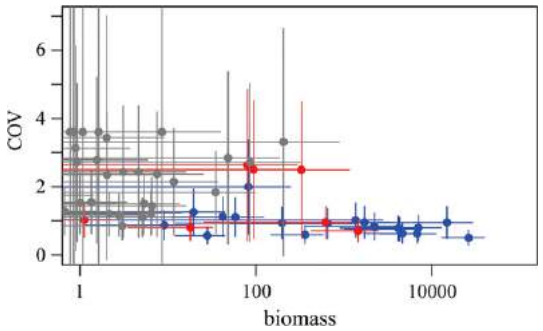


Fig. 11 Relationship between mean (±sd) coefficient of variation (COV) and mean (±sd) biomass. Core species showing density-dependence are shown in *blue*. Core species with no evidence of density-dependence in *red*. Transient species in *gray*. Note that the *blue*, density-dependent species have very low coefficients of variation in their population biomasses.

dependence is not uniform across species, operating in common species rather than rare ones. However, common species are defined as those always present at relatively high biomass. This need not imply high numerical abundance, for example the conger eel ranks 2nd in biomass but only 20th in terms of numerical abundance. When measured as the number of individuals per unit area of seabed the top predator conger eel would be considered rare, but is under density-dependent control.

This analysis leads to the hypothesis that animal communities (ecosystems) comprise a core group of resident species which have constrained populations because they are pushing against their respective carrying capacities and an often larger group of transients probing the system and ready to move in should conditions allow. The transient species abundances are influenced by environmental conditions and not generally constrained by resource availability and therefore do not exhibit density-dependent dynamics. The core species contribute the greater part of the biomass, but a far smaller proportion of total species richness. A healthy system needs both, the transients are a genuine part of a healthy system, they utilize resources which are only occasionally available and are preyed upon by the core assemblage. They must certainly be present if the system is to survive long-term as they allow adaptation to changed conditions.

Natural communities are hybrids of two dynamical behaviors, a core group, dominant in biomass terms, which display negative feedback dynamics and depressed variability (higher stability) and a larger group of species displaying quasi random abundance as they cannot establish permanent residency.

Species Richness, Stability and Regular Cycles

It has also been suggested on both theoretical and observational grounds that the species richness of a habitat may influence population stability. At present, the effect of species richness is unclear. The exceptional study of a plant community by D. Tilman and Downing has indicated that species richness actually increased stability by increasing resistance and recovery of primary productivity to drought. However, the opposite conclusion has been reached in other studies. Results from theoretical studies are equally contradictory. At present, the indications are that there is no universal relationship between the number of interacting species and individual species stability. However, it is difficult to imagine how a species rich system could maintain species number long-term if it held a dominant species that fluctuated greatly in abundance. Such a species would act like a bull in a china shop. The Bridgwater Bay fish data shows that the most abundant species are under the most powerful density-dependent constraints and have the most stable populations (Fig. 11). Within this system the dominants have remarkably orderly dynamics.

While the great majority of studied populations fluctuate with little obvious pattern and clearly respond to random climatic events, there has been great interest in those populations that show long-term oscillations. The most famous of these cyclic phenomena is the 4-year cycle of microtine rodents in boreal and artic regions first discussed by Elton in 1924. There has been considerable debate as to how these cycles form and are stabilized. One favored view is that they are caused by delayed density-dependence as time lags are frequently observed to produce cyclical or quasi-cyclical behaviour. It has been argued that in Fennoscandia, the cycle is imposed by time lags in the population response of mustelid predators and the time series of the rodents is actually chaotic. Studies in both Fennoscandia and Hokkaido, Japan, show that rodent cycles vary geographically and that cycles tend to be more apparent toward the north of the range where species richness is low. Studies on the gray-sided vole, *Clethrionomys rufocanus*, for example, show that the dynamics of this species is influenced by density-dependent, delayed density-dependent, and stochastic climatic changes. The appearance of quasi-regular cycles therefore seems to depend on the relative magnitude of the three types of influence: density-dependent, delayed density-dependent, and climatic. It is notable that as knowledge has deepened, the important role of both biotic and physical environmental factors has been recognized.

The geographical extent of species is an important factor in population stability. K. J. Gaston in a study of 263 species of British moth studied over a time period of up to 14 years concluded that there was a strong relationship between local population variability and species abundance and distribution—with the most abundant and widely distributed species showing the greatest local variability. These observations indicate that the spatial extent of a population may have a great influence on local stability.

This pattern may be related to the fact that widely distributed, well-dispersed, organisms can be opportunistic species, which take advantage of temporary niches to invade and then abandon when conditions worsen with low probability of global extinction. Spatial effects are also a key feature of pathogen population dynamics. For example, L. Ericson and colleagues report a 13-year study of the rust fungus, *Uromyces valerianae*, on the herb valerian, *Valeriana salina*, in an archipelago of islands in central Sweden. They found that the fungus populations frequently went extinct and re-colonized, producing very different and unstable population dynamics within different island populations. It is now well established that meta-population dynamics must be considered to understand the stability of many populations.

A meta-population comprises several spatially distinct populations. Stochastic variation acting on small populations almost inevitably leads to eventual extinction. However, the meta-population avoids collapse because immigrants from one population can re-colonize habitat which has been left open by the extinction. They may also reinforce a population and stop extinction. Such arguments rely on the ability of populations to exchange members. They are also favored by a lack of correlation between the abundance in different populations. This is not always the case. It has been frequently observed that insect pest outbreaks occur at about the same time over large geographical areas. The degree of spatial synchrony between forest insect populations has been used to test the differing hypotheses as to why populations fluctuate. While there is considerable difference in the degree of spatial synchronicity between insects, there is good evidence that synchrony is related to large-scale climatic variables. This in turn indicates that the notorious instability of insect forest pests is determined, at least in part, by their sensitivity to climatic conditions.

In conclusion, it is clear that populations can vary greatly in abundance through time and space. Huge though this variation can be, there is considerable evidence that populations are constrained within limits by biotic and physical factors and they do not show their full potential for explosive growth or sudden collapse. The amount of variability displayed differs between species, taxa, and habitat and is related, in part, to life history characteristics and, in part, to the variability inherent in the habitat. Populations are composed of individuals that are constrained by and respond to their physical and biotic environment. This response to the environment can have the characteristics of negative feedback control and results in sufficient stability to increase the likelihood of long-term population maintenance. However, many local populations repeatedly go extinct and re-colonization from successful populations is essential for long-term existence. Stability of existence for many species may only exist at the meta-population level.

See also: Ecological Data Analysis and Modelling: Metapopulation Models. Global Change Ecology: Material and Metal Ecology

Further Reading

Andrewartha, H.G., Birch, L.C., 1954. The distribution and abundance of animals. Chicago: University of Chicago Press.

Bulmer, M.G., 1975. The statistical analysis of density dependence. Biometrics. 901–911.

Cottingham, K.L., Brown, B.L., Lennon, J.T., 2001. Biodiversity may regulate the temporal variability of ecological systems. Ecology Letters 4, 72–85.

Cuddington, K., 2001. The "balance of nature" metaphor and equilibrium in population ecology. Biology and Philosophy 16, 463–479.

Egerton, F.N., 1973. Changing concepts of the balance of nature. The Quarterly Review of Biology 48, 322–350.

Elton, C.S., 1924. Periodic fluctuations in numbers of animals: Their causes and effects. British Journal of Experimental Biology 2, 119–163.

Elton, C.S., 1930. Animal ecology and evolution. New York: Oxford University Press.

Ericson, L., Burdon, J.J., Mulle, W.J., 1999. Spatial and temporal dynamics of epidemics of the rust fungus uromyces valeriancae on populations of its host valeriana salina. Journal of Ecology 87, 649–658.

Gaston, K.J., 1988. Patterns in the local and regional dynamics of moth populations. Oikos 53, 49–57.

Hanski, I., Turchin, P., Korpimäki, E., Henttonen, H., 1993. Population oscillations of boreal rodents: Regulation by mustelid predators leads to chaos. Nature 364, 232–234.

Henderson, P.A., Bird, D.J., 2010. Fish and macro-crustacean communities and their dynamics in the Severn Estuary. Marine Pollution Bulletin 61, 100–114.

Henderson, P.A., Seaby, R.M., Somes, J.R., 2006. A 25-year study of climatic and density-dependent population regulation of common shrimp, Crangon crangon, in the Bristol channel. Journal of the Marine Biological Association of the UK 86, 287–298.

Henderson, P.A., Seaby, R.M.H., Somes, J.R., 2011. Community level response to climate change: The long-term study of the fish and crustacean community of the Bristol channel. Journal of Experimental Marine Biology and Ecology 400, 78–89. doi:10.1016/j.jembe.2011.02.028.

Kland, B., Andrew, A.E., Liebhold, M., *et al.*, 2005. Are bark beetle outbreaks less synchronous than forest Lepidoptera outbreaks? Oecologia 146, 365–372.

Lack DL (1954) The natural regulation of animal numbers. Oxford: Clarendon Press.

Liebhold, A., Kamata, N., 2000. Are population cycles and spatial synchrony a universal characteristic of forest insect populations? Population Ecology 42, 205–209.

Magurran, A.E., Henderson, P.A., 2003. Explaining the excess of rare species in natural species abundance distributions. Nature 422, 714–716.

Nicholson, A.J., 1933. The balance of animal populations. Journal of Animal Ecology 2, 132–178.

Peltonen, M., Liebhold, A.M., Bjørnstad, O.N., Williams, D.W., 2002. Spatial synchrony in forest insect outbreaks: Roles of regional stochasticity and dispersal. Ecology 3, 3120–3129.

Saether, B.-E., Engen, S., Matthysen, E., 2002. Demographic characteristics and population dynamical patterns of solitary birds. Science 295, 2070–2073.

Saether, B.-E., Engen, S., Møller, A.P., *et al.*, 2004. Life-history variation predicts the effects of demographic stochasticity on avian population dynamics. American Naturalist 164, 793–802.

Saitoh, T., Stenseth, N.C., Viljugrein, H., Kittilsen, M.O., 2003. Mechanisms of density dependence in fluctuating vole populations: Deducing annual density dependence from seasonal processes. Population Ecology 45, 165–173.

Stenseth, N.C., Viljugrein, H., Saitoh, T., *et al.*, 2003. Seasonality, density dependence, and population cycles in Hokkaido voles. Ecology 100, 11478–11483.

Tamara, N., Romanuk, T.N., Kolasa, J., 2004. Population variability is lower in diverse rock pools when the obscuring effects of local processes are removed. Ecoscience 11, 455–462.

TCR, W., 1993. The inadequate environment: Nitrogen and the abundance of animals. Berlin: Springer.

Tilman, D., 1996. Biodiversity: Population versus ecosystem stability. Ecology 77, 350–363.

Self-Organization

DG Green, S Sadedin, and TG Leishman, Monash University, Clayton, VIC, Australia

Introduction

Self-organization is the appearance of order and pattern in a system by internal processes, rather than through external constraints or forces. Plant distributions provide examples of both constraints and self-organization. On a mountainside, for instance, cold acts as an external constraint on the ecosystem by limiting the altitude at which a plant species can grow. Simultaneously, competition for growing sites and resources leads to self-organization within the community by truncating the range of altitudes where plant species do grow. Self-organization can also be seen among individuals within a population (e.g., within an ant colony or a flock of birds) and within individuals (e.g., among cells during development) (Fig. 1).

A growing understanding of ways in which internal processes contribute to ecological organization has provided new perspectives on many phenomena familiar from traditional ecology. Self-organization usually involves interactions between components of a system, and is often closely identified with complexity. Also associated with self-organization is the idea of emergence: that is, features of the system emerge out of interactions, as captured by the popular saying, "the whole is greater than the sum of its parts." It is necessary to distinguish between emergent features and other global properties of a system. For instance, although biomass production in a forest is a global property, it is simply the sum total of production by all the organisms within the forest. A stampede, on the other hand, is behavior that emerges when panic spreads from one animal to another within a herd.

Semantic and philosophical issues sometimes lead to confusion about self-organization. Self-organizing systems are usually open systems, that is, they share information, energy, or materials with their surroundings. However this does not necessarily mean that the external environment controls or determines the way they organize. A growing plant, for instance, absorbs water, light, and nutrient from its environment, but its shape and form are determined largely by its genes.

Also, in considering self-organization, it is important to clearly identify the system concerned, and in particular, what is external and what is internal? This issue arises in the difference between a community and an ecosystem. For a community, which consists of the biota of an area, the effect of (say) soil is an external constraint. However, for the corresponding ecosystem, which would include soils, the interactions between plants, microorganisms, and soil formation are internal processes. Defining the physical limits of an ecosystem poses similar problems. A lake, for instance, is not a closed ecosystem. Among other things, water birds come and go, removing some organisms and introducing others.

Historical Comments

Self-organization as a widespread phenomenon first came to the attention of researchers during the mid-twentieth century. The interest in self-organization comes from many different fields of study. The biologist Ludwig von Bertalanffy drew attention to the role of internal interactions and processes in creating organization within biological systems. His 'general systems theory' drew heavily on analogies to highlight the existence of common processes in superficially different systems. Meanwhile, W. Ross Ashby and Norbert Wiener explored self-organization from the perspective of communications and feedback in the control of systems. Ashby introduced the term self-organizing in 1947. Wiener coined the term cybernetics to refer to the interplay of control systems and information. In the 1950s, systems ecologist H. T. Odum collaborated with engineer Richard Pinkerton to develop the principle of maximum power, which states that systems self-organize to maximize energy transformation.

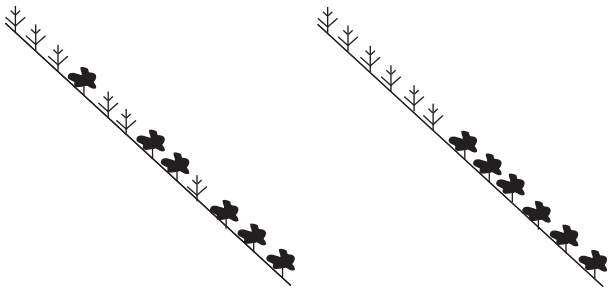


Fig. 1 Effect of competition on plant distributions on a gradient. The two plant species shown are adapted to different conditions, which are here found at either end of the slope. At left, there is no competition, so the distributions merge into one another. At right, competition truncates the distributions, leading to sharply defined altitudinal zones.

 doi:10.1016/B978-0-444-63768-0.00696-X

During the 1970s and 1980s, increasing computing power made it possible to use simulation to explore the consequences of complex networks of interactions. By the last two decades of the twentieth century, the nature and implications of biological self-organization were increasingly being explored as a part of the complexity theory. The new field of Artificial life (Alife), initiated by pioneers such as Chris Langton, Pauline Hogeweg, and Bruce Hesper, has produced a series of seminal models that demonstrate self-organization in a variety of ecological and evolutionary contexts. Around the same time, H. T. Odum introduced the systems concept of 'emergy' to represent the total energy used in developing a process.

By the 1990s, researchers were looking for broad-based theories of self-organization. John Holland stressed the role of adaptation in self-organization. He suggested that seven basic elements are involved in the emergence of order in complex adaptive systems. These include four properties – aggregation, nonlinearity, flows, and diversity – and three mechanisms – tagging, internal models, and building blocks. In contrast, Stuart Kauffman's work on autocatalytic sets within Boolean networks emphasizes ways in which self-organization may structure biological systems independent of selection. Likewise, embryologist Brian Goodwin suggested that to understand macroevolution, we require a theory of morphogenesis which takes account of physical, spatial, and temporal dynamics in addition to selection. The work of James Kay provided an interpretation of life from a thermodynamic perspective, arguing that self-organizing systems maximize the dissipation of gradients in nature. In particular, Kay argues that over time, ecosystems evolve to dissipate energy more efficiently by becoming increasingly complex and diverse.

Theories of Self-Organization

Thermodynamic Basis

In physical terms, the phenomenon of self-organization appears at first sight to be ruled out by the second law of thermodynamics, which states that in any closed system, entropy increases with time. In this sense, living systems seem to fly in the face of thermodynamics by accumulating order. However, self-organizing systems need not be closed. Open systems, including living things, share energy and information with the outside environment. In the late 1960s, Ilya Prigogine introduced the idea of dissipative systems to explain how this happens. He defined dissipative systems to be open systems that are far from equilibrium. Dissipative systems have no tendency to smooth out irregularities and to become homogeneous. Instead, they allow irregularities to grow and spread. Physical examples include crystal formation. Biological systems, including cells, organisms, and ecosystems, are all examples.

The Network Model

An important source of self-organization is provided by the interactions and relationships between the objects that comprise a complex system. Patterns of such relationships are captured by the network model of complexity.

Networks capture the essence of interactions and relationships, which is a fundamental source of complexity. A graph is defined to be a set of nodes (objects) joined by edges (relationships) and a network is a graph in which the nodes and/or edges have values associated with them. In a food web, for instance, the populations form nodes, and the interactions between them (e.g., predation) form the edges. In a landscape, spatial processes and relationships create many networks. For instance, the nodes might be individual plants and the corresponding edges would be any processes that create relationships between them, such as dispersal or overshading. In an animal social group, the nodes would be individuals and the edges would be relationships such as kinship or dominance.

Nodes that are joined by an edge are called neighbors. The degree of a node is the number of immediate neighbors that it has. A path is a sequence of edges in which the end node of one edge is the start node of the next edge, for example, the sequence of edges A–B, B–C, C–D, D–E forms a path from node A to node E. A cycle is a path that ends where it starts, for example, A–B, B–C, C–A. A network is called connected if, for any pair of nodes, there is always some path joining them (otherwise it is disconnected). The diameter of a network is the maximum separation between any pair of nodes. Clusters are highly connected sets of nodes.

The importance of networks stems from their universal nature. Network structure is present wherever a system can be seen to be composed of objects (nodes) and relationships (edges). Less obvious is that networks are also implicit in the behavior of systems. In this respect, the nodes are states of the system (e.g., species composition) and the edges are transitions from one state to another.

Sometimes, network structure plays a more important part in determining the behavior of a system than the nature of the individual components. In dynamic systems, for instance, cycles are associated with feedback loops. In disconnected networks, the nodes form small, isolated components, whereas in connected networks, they are influenced by interactions with their neighbors. Self-organization in a network can occur in two ways: by the addition or removal of nodes or edges, or by changes in the values associated with the nodes and edges.

Several kinds of network patterns are common and convey important properties.

- A random network is a network in which the nodes are connected at random. In a random network of n nodes, the degrees of the nodes approximate a Poisson distribution, and the average length (L) of a path between any two nodes is given by $L = \log(n)/\log(d)$, where d is the average degree.

- A regular network is a network with a consistent pattern of connections, such as a lattice or cycle.
- Small worlds fall between random networks and regular networks. They are typically highly clustered, but with low diameter. A common scenario is a system dominated by short-range connections, but in which some long-range connections are also present.
- A tree is a connected network that contains no cycles. A hierarchy is a tree that has a defined root node. For instance, the descendents of a particular individual animal (the root of the tree) form a hierarchy determined by birth. Trees and hierarchies are closely associated with the idea of encapsulation.
- A scale-free network is a connected network in which the degrees of the nodes follow an inverse power law. That is, some nodes are highly connected, but most have few (usually just one) connections.

Encapsulation

Encapsulation is the process by which a set of distinct objects combine to act as a single unit. Individual fish, for example, form a school by aligning their movements with their neighbors. Because smaller objects usually merge into larger wholes, encapsulation is often linked to questions of scale. Encapsulation is closely associated with the idea of emergence. The whole emerges when individuals become subsumed within a group in relation to the outside world. There are many examples in ecology. Ecosystems are communities of interacting organisms; populations are groups of interbreeding organisms; and schools, flocks, and herds are groups of animals moving in coordinated fashion. In all of these cases, the individuals may not be permanently bound to the group, unlike cells within the human body. Cellular slime molds present an intermediate case in which cells sometimes act independently but at other times aggregate to form a multicellular individual.

Various ecological theories are based on the assumption that encapsulation plays an important role in ecosystem structure and function. The concept of ecosystem compartments implies that a community is formed of distinct groups (compartments) consisting of mutually interacting species, but the interactions between the groups are limited.

Connectivity and Criticality

Criticality is a phenomenon in which a system exhibits sudden phase changes. Examples include water freezing, crystallization, and epidemic processes. Associated with every critical phenomenon is an order parameter, and the phase change occurs when the order parameter reaches a critical value. For example, water freezes, when its temperature falls to 0 °C. A wildfire spreads when fuel moisture falls below a critical level (else it dies out).

Changes in the connectivity of a network have important consequences and often underlie critical phenomena. When a network is formed by adding edges at random to a set of N nodes, a connectivity avalanche occurs when the number of edges is approximately $N/2$. This avalanche is characterized by the formation of a connected subnet, called a unique giant component (UGC), which contains most of the nodes in the full network. The formation of the UGC marks a phase change in which the network shifts rapidly from being disconnected to connected.

Any system that can be identified with nodes and edges forms a network, so the connectivity avalanche occurs in many settings and is the usual mechanism underlying critical phase changes.

The connectivity avalanche has several important implications. For interacting systems, it means that the group behaves either as disconnected individuals, or as a connected whole. Either global properties emerge, or they do not: there is usually very little intermediate behavior. Landscape connectivity provides an important ecological example of critical phase change.

Phase changes in connectivity also underlie criticality in system behavior. The degree of connectivity between states of a system determines the richness of its behavior. Studies based on automata theory show that if connectivity is too low, systems become static or locked in narrow cycles. If connectivity is too high, systems behave chaotically. The transition between these two phases is a critical region, popularly known as the 'edge of chaos'. It has been observed that automata whose state spaces lie in this critical region exhibit the most interesting behavior. This observation led researchers such as James Crutchfield, Christopher Langton, and Stuart Kauffman to suggest that automata need to reside in the critical region to perform universal computation. More speculative is their suggestion that the edge of chaos is an essential requirement for evolvability (the ability to evolve) in complex systems, including living things. Others have proposed that living systems exploit chaos as a source of novelty, and that they evolve to lie near the edge of chaos. These ideas are closely related to self-organized criticality (SOC).

Self-Organized Criticality

SOC is a phenomenon wherein a system maintains itself in a critical or near-critical state. A classic example is the pattern of collapses in a growing pile of sand. Because information theory suggests that systems in critical states are most amenable to information processing and complexity, self-organized criticality has been proposed as a component of collective behavior in ant colonies, societies, ecosystems, and large-scale evolution. SOC is characterized by events whose size and frequency distributions follow an inverse power law. However, it is often difficult to distinguish genuine cases of SOC from simple cause and effect processes that exhibit similar distributions.

For example, ecosystems might tend toward critical states through the following mechanism. If new species or mutations appear in an ecosystem occasionally, then as the variation in the ecosystem increases over time, so does the probability of forming destabilizing positive feedback loops. Such destabilizing interactions could initiate avalanches of extinctions, and the probable size of such avalanches would be related to the preexisting connectivity of the system. In this way, mutation, migration, and extinction could keep the system near the critical region, as the addition of new variation drives the ecosystem out of subcriticality, while extinction avalanches prevent supercriticality. Proponents of this idea point to extinction events, whose distribution follows an inverse power law, as supporting evidence. However, other explanations of this pattern, such as cometary impacts, are also plausible.

Feedback

Feedback is a process in which outputs from a system affect the inputs. Predator–prey systems are examples of negative feedback. For instance, any increase in the size of a predator population means that more prey are eaten, so the prey population decreases, which in turn leads to a decrease in the predator population. Reproduction is an example of positive feedback: births increase population size, which in turn increases the rate of reproduction, which leads to yet more births. Feedback loops arise when a sequence of interactions form a closed loop, for example, A–B–C–A. Feedback loops play an important role in food webs and ecosystem stability. Time delays in the response within a feedback loop often lead to cyclic behavior (e.g., in predator–prey systems).

Both positive and negative feedback are important in self-organization. By dampening changes, negative feedback acts as a stabilizing force. It is one of the principal mechanisms of homeostasis, the maintenance of dynamic equilibrium by internal regulation. In contrast, positive feedback magnifies minor deviations. An example is competitive exclusion: any small decrease in size of a competing population is likely to lead to further decreases, until it dies out (Fig. 2).

Stigmergy

Stigmergy is a form of self-organization that occurs when parts of a system communicate by modifying their environment. Many examples of stigmergy occur in the organization of eusocial insect colonies. For example, in ant colonies, objects such as food, larvae, and corpses are often stored in discrete larders, nurseries, and cemeteries. Models show that this civic order can emerge through interactions between the ants and their environment. In the model, ants pick up objects at random, and may drop them when they encounter similar objects. Over time, this process creates piles of similar objects. Positive feedback causes larger piles to grow at the expense of smaller ones (Fig. 3).

Synchronization

Synchrony can alter system-level behavior by enhancing or dampening nonlinearities. For example, when predator and prey populations are tightly coupled to one another, a stable, negative feedback relationship can result where an increase in prey causes increased predators and a subsequent decrease in prey. In this case, the ecological interaction acts like a thermostat regulating population size. However, if the two populations respond at different rates, oscillations or even chaotic behavior can occur instead. A classic example of such oscillations occurs in the interaction between populations of hares and lynxes in the Arctic Circle.

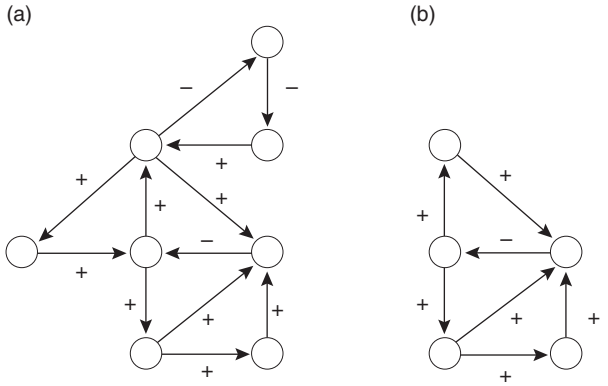


Fig. 2 The role of feedback in self-organization of a food web. In this diagram, circles represent populations and arrows indicate the influence (positive or negative) of one population on another. In a food web, circular chains of interaction between populations form feedback loops, as in the example shown here. (a) The initial food web contains both positive and negative feedback loops. Internal dynamics within the positive feedback loops leads to the local extinction of several populations. (b) The resulting food web contains only negative feedback loops, which stabilize the community.

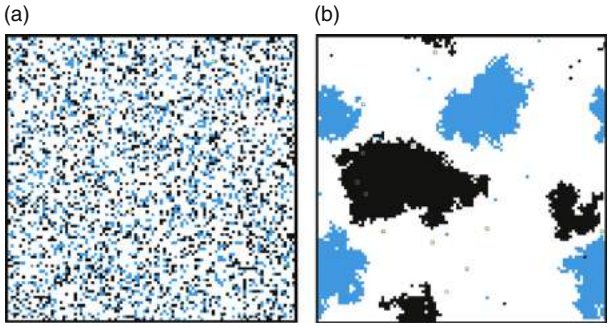


Fig. 3 The emergence of order by stigmergy and feedback in an ant colony. Given a random scatter of objects (a), ants sort objects by picking them up and dropping them again when they find a similar object. This process creates piles, which grow over time. Large piles grow at the expense of smaller ones until only a few large piles remain (b).

Synchronized breeding behavior is common and includes mass flowering in plants, mass breeding in birds, and mass spawning among marine animals such as corals and squid. In these cases, synchrony is usually achieved by individuals responding to a common environmental cue, such as a change in temperature or day length. Synchronized breeding conveys distinct advantages such as maximal exploitation of resources and satiation of predators.

Different species often have co-adapted simultaneous seasonal behavior, such as birds that breed when butterflies emerge. However, both the environmental cues, and the physiological response, may differ among these co-adapted species. For example, great tits time their egg laying by photoperiod. Winter moths are an important food source during the breeding season, and they develop more quickly at higher temperatures. As a result, recent warm springs in Europe caused by climate change have disrupted the synchronization between these species, reducing food availability for nesting great tits and potentially destabilizing populations.

In other cases, synchronous behavior arises through social contagion, where individuals imitate others. The dynamics of such behavior are similar to those seen in epidemiology. Social contagion can lead to coordinated group behavior such as flocking, as well as disparate phenomena such as synchronized flashing in fireflies, and 'fashions' in mate choice among birds and fish. The emergence of synchronous behavior in these cases is highly sensitive to the structure of social networks. Synchrony is easily achieved when networks are highly connected (i.e., individuals can perceive a large number of other individuals, or some individuals have very large influence). However, in loosely connected networks, social contagion can result in asynchronous waves or chaos.

Complex Adaptive Systems

Complex adaptive systems (CASs) consist of diverse, locally interacting components that are subject to selection. Examples include learning brains, developing individuals, economies, ecosystems, and the biosphere. In such systems, hierarchical organization, continual novelty and adaptation, and nonequilibrium dynamics are known to emerge. As a result, the behavior of a CAS is characterized by nonlinearity, historical contingency, thresholds, and multiple basins of attraction. A key question in current CAS research has been the relationship between resilience and criticality. Some authors suggest that a CAS will generally evolve toward self-organized criticality. By being maintained near the edge of chaos, such systems might maximize information processing. In this way, criticality might enhance the ability of CASs to adapt to changing environments and efficiently utilize resources, making systems become more resilient over time.

Artificial Life

The field of Alife uses simulation models to understand biological organization by abstracting crucial features and examining living systems 'as they could be'.

One of the most widespread representations used in Alife models is the cellular automaton (CA). This is a grid of cells in which each cell has a state (some property of interest) and is programmed to behave in identical fashion. Each cell has a neighborhood (usually the cells immediately adjacent to it) and the states of it neighbors affect changes in a cell's state. The most famous example is the Game of Life, in which each cell is either 'alive' or 'dead' at any time. Despite its extreme simplicity, the game showed that large numbers of interactions governed by simple rules lead to the emergence of order within a system. Cellular automata have been used to model many biological and ecological systems. In models of fires, epidemics, and other spatial processes, each cell represents a fixed area of the landscape and the cell states represent features of interest (e.g., susceptible, infected, or immune organisms in an epidemic model).

Other prominent ALife models include Tom Ray's Tierra model, which demonstrated adaptation within self-reproducing automata. Craig Reynolds' boids model demonstrated that flocking behavior emerges from simple interactions between individuals. James Lovelock's Daisyworld model showed the potential for biotic feedback and adaptation to stabilize the biosphere.

Self-Organization in an Ecological Setting

Social Groups

Relationships between individuals create several kinds of organizations within groups of animals.

Coordination between moving animals leads the formation of groups. Examples include swarms of insects, flocks of birds, schools of fish, and herds of mammals. Coordinated group movements, even in very large groups, can be achieved by individuals obeying simple rules, such as 'keep close, but not too close, to your neighbors' and 'head in the same general direction as your neighbors.'

Several mechanisms that channel aggressive behavior create social organization. In social animals, dominance hierarchies reduce the potential costs of conflict over mates and food. Adominance hierarchy emerges when interactions between individuals result in physiological and behavioral changes: for example, 'winning' a contest may elevate testosterone, causing increased dominance behavior, and evoking submissive behavior from individuals who have been less successful in the past. In this way, coherent transitive hierarchies can emerge even when all individuals were initially equal. Similarly, territoriality reduces the costs of conflict over resources by partitioning a landscape among a population. Territoriality often generates spatial patterns, such as regular distances between nests in seabird colonies. In this case, the distance between nests is defined by the maximum area that a sitting bird can defend without abandoning her nest. More complex coordinated group behaviors can emerge when individuals take on different tasks and roles within groups. For example, within ant and termite colonies, individuals can develop into a variety of castes, each with distinct roles such as foraging, nest defense, and nursing young. In honeybees, individuals take on different roles at different life stages.

In some cases, upper limits exist on the size that social groups can attain and depend on interactions between the animals. In apes, for instance, where social bonds are maintained by grooming, troop sizes tend to be 30–60 individuals. Larger troops tend to fragment. Among humans, social groups are usually much larger. The anthropologist Robin Dunbar argues that this is a consequence of speech providing more efficient social bonding than grooming, leading to a natural group size of 100–150 individuals.

In most cases, group size may be the outcome of several interacting ecological and social factors. For example, although lions hunt cooperatively, prides and hunting groups are usually larger than is optimal for hunting efficiency. Lionesses cooperate to defend cubs against infanticidal males by forming crèches. In addition, hunters are vulnerable to attack by larger groups, and territories are more effectively defended by larger prides.

The origin of cooperation among groups of cells and organisms can also be examined from the perspective of self-organization. The paradox of the evolution of cooperation is that (by definition) selfish individuals outcompete altruists, and therefore in a population of self-replicators, a selfish mutant should always spread at the expense of altruists. Nonetheless, altruism does occur among humans and cooperative behavior is often seen among animals. Such cooperative behavior can self-organize when the network structure that governs interactions among individuals results in the same individuals encountering one another repeatedly (e.g., when individuals are fixed in space, so that their only interactions are with their neighbors), or when their reproductive fate is very closely tied to that of others (as is the case for cells within a multicellular organism). Experimentally, the evolution of cooperation has been induced in bacterial populations by production of adhesive, causing individual cells to clump together. Cooperation can also evolve in marginal environments, where the evolutionary impact of competition between individuals is outweighed by the need to survive. Experimental studies of bacteria in marginal environments show that complex spatial patterns and signaling behaviors can emerge as a result of this selection. In theoretical models, the inclusion of policing behavior (punishing nonconformists) can also enforce high levels of cooperation even when interactions occur at random in large societies.

Persistence and Stability in Ecosystems

One of the most puzzling topics in systems ecology is how ecosystems emerge that are at once complex and stable. Field studies suggest that the most complex (diverse) ecosystems are also the most stable. However, this observation runs counter to expectation from systems theory. It shows that the more components a dynamic system has, the more likely it is that a destabilizing interaction (such as a positive feedback loop) will cause it to collapse and lose species. Consequently, systems theory suggests that simpler ecosystems should be more stable than complex ones. The paradox implies that the complex, stable ecosystems seen in nature are not random assemblages. Self-organization in this case involves removal of destabilizing positive feedback loops.

Communities versus Assemblages

The question of how important self-organization is in ecosystems has long been debated in ecology. Are ecosystems communities of co-adapted species, or are they simply random assemblages? Some early theorists, such as Clements, believed that the groups of species found together were specialized for living together, whereas others, such as Gleason, stressed the importance of chance and individuals.

The idea of succession concerns the patterns and processes involved in community change, especially after disturbance. A form of self-organization often associated with succession is facilitation. That is, plants and animals present in an area can alter the local environment, thereby facilitating the appearance of populations that replace them. After a fire, for example, a forest will regenerate with herbs and shrubs growing back almost immediately. The first trees to reappear will be 'pioneer' (disturbance) species, which

disperse well, grow fast, and can tolerate open, exposed conditions. These trees create shade and leaf litter, which favor slow-growing, shade-tolerant trees.

Recent theoretical work (such as Hubbell's neutral theory of biodiversity and biogeography) emphasizes the role of chance and spatial dynamics in generating ecological patterns. In these models, self-organization is trivial because all individuals and species are effectively identical, and species abundances are driven by random birth, migration, and death processes. Both neutral and self-organizing models have been successful in explaining real relative abundance and species–area curves.

Food Webs

Species interactions lead to the flow of material within an ecosystem. For animals the most common processes are eating, respiration, excretion, and egestion. For plants, they are root uptake of water and nutrients, respiration, and photosynthesis. The outputs of material from one organism often become inputs to other. This focus on 'what eats what' led Elton to identify several patterns, notably the food chain and the food web, the food cycle, the ecological niche, and the pyramid of numbers.

Self-organization in ecosystems is evident in the structure in food webs, networks that describe trophic interactions among species. Within food webs, specific patterns of interaction may be prevalent. These patterns, termed ecological motifs, are thought to represent especially stable interactions. The concept of keystone species supposes that certain species play a crucial role in maintaining the integrity and stability of an ecosystem.

Analysis of food webs suggests that a small-world structure is common. That is, most species interact with only a small number of other species, but the connectivity of the web as a whole is maintained by a few species that interact with a large number of others. This observation provides a theoretical basis for the idea of keystone species. Functionally, small world networks are thought to be robust to random loss of nodes (e.g., species), but vulnerable to attacks that target their highly connected nodes (e.g., keystone species).

Spatial Patterns and Processes

Spatial processes lead to the formation of distribution patterns. Seed dispersal, for instance, often produces concentrations of seedlings around parent plants and leads to the formation of clumped distributions. When local dispersal is combined with patchy disturbance, such as fire, the result is a distribution composed of patches. When combined with environmental gradients, such as soil moisture, local dispersal can produce zone patterns, with different species dominating different areas (**Fig. 4**).

Fragmentation is one of the most important consequences of landscape connectivity. When the density of (randomly located) objects in a landscape falls below a critical density, they are mostly isolated individuals. When the density exceeds the critical threshold, they become connected. The density at which the critical threshold occurs depends on the size of the neighborhood of the objects. There are many cases where landscape connectivity plays an important role. Epidemic processes require a critical density of resources to spread. Instances include disease outbreaks (susceptible individuals), fire spread (fuel), and invasions of exotic plants (suitable sites). Populations become fragmented if individuals cannot interact with one another. For instance, in wet years the water bodies of central Australia are essentially connected for water birds, which can fly from one body to another almost anywhere in the continent. In dry years, however, many water bodies shrink or dry up and become too widely separated for birds to migrate between them (**Fig. 5**).

Self-Organization in the Biosphere

Arguably the most ambitious ecological theory based on self-organization is the Gaia hypothesis, which postulates that the biosphere itself evolves to a homeostatic state. Lovelock suggested the Daisyworld model as an illustration of how this process

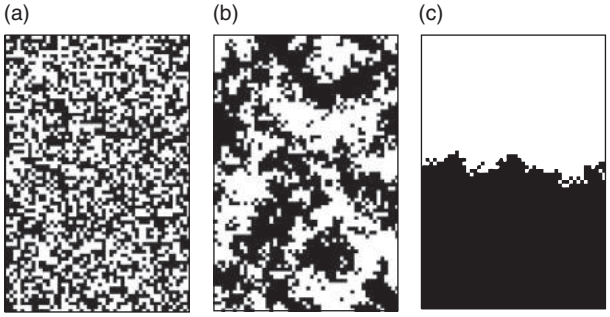


Fig. 4 Emergence of spatial patterns from dispersal. This CA model shows the hypothetical distributions of two plant populations that result in three different scenarios. (a) Global dispersal, in which seeds can spread anywhere, results in random distributions of plants. (b) Dispersal from local seed sources leads to clumped distributions. (c) The combination of local dispersal and environmental gradients (from top to bottom) creates vegetation zones.

Fig. 5 Critical phase changes in connectivity within a fragmented landscape. In this CA model, grid cells represent sites in a landscape. Gray and black cells represent vegetation and white cells have no cover. The black cells show examples of patches of vegetation sites that are connected, for example, by spread of a fire ignited in the center of the grid. Notice that only a small increase in the density of covered sites makes the difference between subcritical and supercritical.

might occur. On the hypothetical Daisyworld, black and white daisies compete for space. Although both kinds of daisies grow best at the same temperature, black daisies absorb more heat than white daisies. When the Sun shines more brightly, heating the planet, white daisies spread, and the planet cools again. When the Sun dims, the black daisies spread, warming the planet. In this way, competitive interactions between daisies provide a homeostatic mechanism for the planet as a whole.

The idea behind Gaia is that ecosystems will survive and spread more effectively if they promote the abiotic conditions required for their own persistence. If so, ecosystems might gradually evolve to be increasingly robust, and if this happened on a global scale, then the biosphere itself might behave as a self-regulating system. However, evidence for Gaian processes in real ecosystems remains tenuous and their theoretical plausibility is disputed.

Evolution

Self-organization may play a prominent role in evolution, especially in the context of landscapes, which regulate interactions between individuals. One consequence is the evolution of cooperation in marginal and viscous habitat networks, whereas randomly interacting populations are more dominated by intraspecific competition and therefore more likely to behave selfishly.

Landscape structure influences genetic diversity and speciation. In connected landscapes, genes flow freely throughout a species and speciation is inhibited. However, in fragmented landscapes, a species breaks into isolated subpopulations. Fragmentation increases the risk of inbreeding and loss of genetic diversity in these subpopulations. Divergence between population fragments may also underlie adaptive radiations, in which many novel species suddenly emerge simultaneously.

As species adapt to their environment, they are often faced by tradeoffs in allocating resources for different purposes. These tradeoffs can lead to the evolution of distinct morphs within a species, or to speciation. For example, many mangrove species face a conflict between salt tolerance and competitive ability. Mangroves grow in estuaries, where salinity varies along the gradient between land and sea. Mangroves growing landward will be under strong selection for competitive ability, while those growing closer to the sea require better salt tolerance. The tradeoff, combined with local seed dispersal, can generate discrete banding patterns in the distribution of mangrove species, where each species is displaced by a more salt-tolerant one closer to the sea.

Contingency also plays a large part in the organization of spatial distributions. Spatial dominance occurs when a particular species is overwhelmingly abundant in a local environment. In this situation, the species can resist invasion, even by a superior competitor, because its propagules are much more numerous locally than those of any other population. For the same reason, a mutation that enables a species to exploit a novel environment may result in it permanently excluding potential competitors from that environment, even after they have evolved similar adaptations.

Practical Considerations

The insights provided by theories of self-organization have many practical implications, both for ecology and for conservation. The sharp end of the conservation debate often hinges on the question of which areas and which sites to conserve. If ecosystems consist of random collections of species, then one site in a landscape is as good as another. All that matters is to preserve representative populations of each species. However, if the ecosystems consist of self-organized communities, in which the species are adapted to depend on one another for survival, then whole communities need to be conserved.

Closely related to the above issue is that the tendency for randomly constructed food webs to be unstable raises questions about the long-term viability of artificially created communities in which translocated species are introduced into new areas. Self-organization is evident even in artificial ecosystems. In biosphere 2, for instance, a closed, experimental environment designed to emulate natural ecosystems, the environment was found to favor species that collect more energy and internal processes led to unexpected problems, such as runaway depletion of oxygen levels.

The need to understand self-organization is important when considering altered ecosystems. For instance, it is usually not possible to carry out experiments to determine the long-term effects of current ecological management practices such as translocation of populations, controlled burning or allocation of reserves and wilderness areas. This problem makes simulation

modeling a potentially crucial tool of ecological theory and practice. New methods of field observation are also appearing. For instance, the need to understand landscape fragmentation has led to studies of connectivity in landscapes, both field based, and using data from remote-sensing and geographic information.

See also: Terrestrial and Landscape Ecology: Ecological Engineering: Design Principles

Further Reading

Ball, P., 1999. The Self-Made Tapestry: Pattern Formation in Nature. Oxford: Oxford University Press.
Camazine, S., Deneubourg, J.-L., Franks, N.R., *et al.*, 2003. Self-Organization in Biological Systems. Princeton: Princeton University Press.
Green, D.G., Klomp, N.I., Rimmington, G.R., Sadedin, S., 2006. Complexity in Landscape Ecology. Amsterdam: Springer.
Holland, J.H., 1996. Hidden Order: How Adaptation Builds Complexity. New York: Addison-Wesley.
Levin, S.A., 1998. Ecosystems and the biosphere as complex adaptive systems. Ecosystems 1 (5), 431–436.
Patten, B.C., Fath, B.D., Choi, J.S., 2002. Complex adaptive hierarchical systems – Background. In: Costanza, R., Jørgensen, S.E. (Eds.), Understanding and Solving Environmental Problems in the 21st Century. London: Elsevier, pp. 41–94.
Prigogine, I., 1980. From Being to Becoming. New York: Freeman, (ISBN 0-7167-1107-9).
Rohani, P., Lewis, T.J., Gruenbaum, D., Ruxton, G.D., 1997. Spatial self-organization in ecology: Pretty patterns or robust reality? Trends in Ecology and Evolution 12 (8), 70–74.
Solé, R.V., Levin, S., 2002. Preface to special issue: The biosphere as a complex adaptive system. Philosophical Transactions of the Royal Society of London B 357, 617–618.
Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. Nature 393 (6684), 440–442.

Systems Ecology☆

Todd M Swannack, U.S. Army Engineer Research and Development Center, Vicksburg, MS, United States

Introduction

This article presents a scientific overview of systems ecology. The first section defines systems ecology and provides a brief history discussing the development of the discipline. The second section defines both complex systems in general and complex ecological systems and further describes several properties of ecological systems. The concluding section provides a brief overview of the systems approach used to analyze ecological systems.

Systems Ecology

Systems ecology is the study of the complex ecological systems using the systems approach. Ecological systems are thermodynamically open, hierarchical, self-organized, and self-regulating. The systems approach, including systems analysis and simulation, is characterized by the conceptualization, quantification, evaluation, and application of a model representing the ecological system of interest. These models can Systems ecology often, but not necessarily, focuses on the development of quantitative models to simulate the dynamics of ecosystem-level processes such as energy flow and nutrient cycling. However, the distinguishing feature of systems ecology is its holistic approach to questions dealing with complex ecological systems, regardless of their specific nature.

Systems ecology is a relatively new field within ecology. Until the middle of the 20th century, ecology was mainly a descriptive science and progressed by traditional, reductionistic methods. Organisms were often the focus of ecological studies. During this time, many ecological processes were identified as relationships between organisms (e.g., pairwise Lotka–Volterra competition equations). In 1935, the concept of an ecosystem as a self-sustaining, functional unit was introduced, but this concept was not built upon significantly until after World War II. Drawing from other fields such as engineering, communications, information theory, and systems sciences, ecologists began to recognize that ecological systems could and should be treated as unique, functional entities. A new definition for ecosystems emerged in 1995—an ecosystem is a complex of ecological communities and their environment, forming a functional whole in nature. This holistic approach recognized the irreducibility and complexity of natural systems. Due to the interactions among many different abiotic and biotic components, natural systems are inherently complex. This complexity makes it virtually impossible to obtain a complete picture of a natural system by reductionistic methods alone. Methods that synthesize analytical results were needed to illuminate the properties of natural systems. Incorporating the philosophy and techniques from general systems theory, ecologists now have the tools and ability to study complex ecological systems. This holistic approach emphasizes synthesis, that is, putting analytical results gathered from the reductionist approach, together to form a picture or model of natural systems.

Complex Ecological Systems

A system consists of a defined set of objects that interact in space and time. Systems are organized collections of interrelated physical components characterized by a boundary and functional unity. Systems are subjective entities created by humans for specific purposes, generally to do work, answer questions, illustrate theory, or explain the natural world. The complexity of any system increases as the number of system components and connections among components increases. An important property of a system is that the interactions among components create emergent properties, specific to that system. These properties result from the interactions among the components, are unique to the system itself, and are more than the collective sum of the properties of the components of the system. In complex systems, cause and effect often are not closely linked in time and space. Systems properties can be studied using the principles from general systems theory.

Ecological systems are functional, complex systems, which can sustain life and are composed of both living and nonliving components from a specific environment. The size and scale of any specific ecological system depends on the point of view of the researcher and are not specified until the objectives of the study are identified.

☆*Change History*: March 2018. Todd M. Swannack (original author) did the updates. The following sections were updated/added. Network connectivity was added. Figure 2 is a new figure. Communication, confusion of a concept, difficulty picturing the system, conflicts with preexisting models sections are new. Further reading was updated.

This is an update of T.M. Swannack and W.E. Grant, Systems Ecology, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 3477–3481.

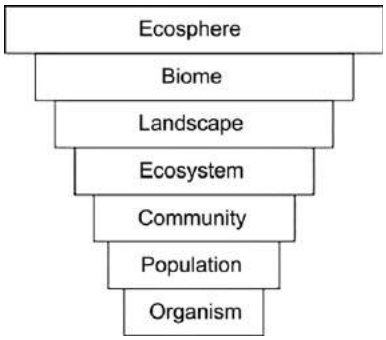


Fig. 1 Most commonly recognized ecological hierarchy. The organism is the simplest level within this hierarchy while the ecosphere is the most complex.

Open Systems

Ecological systems are thermodynamically open. Solar energy enters the biosphere and biological and ecological processes use this energy, which eventually leaves the system as unusable heat. There are two common measures of the flow of energy in a natural system, emergy and exergy. Emergy is the sum of one type of energy required to create an energy flow of another type, usually referenced to solar radiation. Exergy is a measure of the potential of a system to cause change as the system approaches equilibrium. Both emergy and exergy attempt to quantify the energy hidden in the organization and construction of living organisms. Both energy and material (e.g., nutrients in a nutrient cycle or individuals in a population) flow through ecological systems. Several properties of ecological systems result from being open systems. These properties include hierarchical organization, self-organization, and self-regulation via feedback.

Hierarchical Organization

Ecological systems are hierarchical in nature—at each level, components interact with the physical environment (matter and energy) and produce functioning systems. The most recognized ecological hierarchy, from the simplest to the most complex, is illustrated in Fig. 1. This hierarchy organizes the living systems on the planet into seven levels. Each level is composed of the levels below (i.e., populations are composed of organisms, communities are composed of populations, ecosystems are composed of communities, etc.). Energy and material can flow among levels. It is important to note that ecological systems need not be confined to the traditional ecosystem level of the hierarchy. An ecological system can be any system with biotic and abiotic components. Further, most ecological system are hierarchical and Fig. 1 may not explain all ecological phenomena well—for example, nutrient cycling does not fit well within this hierarchy; however, this is a fundamental ecological process that transcends the levels of organization diagrammed in Fig. 1. The complexity of ecological systems is a result, in part, from the interactions across these levels. At any given level in the hierarchy, interactions among components of lower levels produce emergent, irreducible properties that cannot be predicted from studying individual components of these lower levels. Ecological interactions occur across hierarchical levels, so systems ecologists have to deal explicitly with issues of scale, both spatial and temporal. In addition to emergent properties, there are functions that operate across all levels (e.g., diversity, evolution, energetics, regulation). These transcending functions can operate the same way in all levels (e.g., energetics) or can operate differently across levels—for example, evolution at the organism level involves genetic mutations and other genomic interactions and this process indirectly affects other levels of the hierarchy. At broader levels, systems principles apply, like the maximum power principle, which states evolution favors systems that maximize the flow rate of useful energy. The hierarchical structure of ecological systems creates a great deal of complexity which necessitates a holistic approach to elucidate system properties.

Order implies that the relatedness among system components is not a result of random processes. The more information required to describe a system, the higher the degree of randomness, indicating a lower degree of order. At the lower levels of the hierarchy, order is shaped, in part, by evolutionary processes occurring at the organism level. Adaptation at the individual level cascades through the hierarchy—for example, individuals evolve adaptations so they can compete (or avoid competing) with other species; competitive interactions among populations affect the spatial distribution of the populations in the community, which affects the community distribution in the landscape, etc. Evolution, along with other ecological processes, confers order within ecological systems.

Self-Organization

The interactions of components of ecological systems are nonrandom. These nonrandom interactions produce order and, as a result, ecological systems self-organize. Self-organization is when systems composed of many parts organize, achieving a

stable state in the absence of external perturbations. These systems can only maintain themselves by having a constant flow of energy (i.e., these systems are not in thermodynamic equilibrium). Ecological succession is an example of a self-organizing ecological system. During succession, the system builds up biomass and complexity. This process, termed ascendency by Ulanowicz, is the tendency for self-organizing systems to develop complexity in network flows and biomass. Ascendency is one way to measure the self-organizing complexity of ecological systems.

Self-Regulation

Self-regulation is another fundamental property of ecological systems. This concept was adapted from the field of cybernetics. Control is generally in the form of a feedback loop where part of the output of a system feeds back into the system as input. Negative feedback (when the output counteracts the input) is necessary for regulating a system. The energy required for negative feedback can be small compared to the total energy flowing through a system. Low-energy feedback components are common in natural systems. In addition to feedback, systems possess resistance and resilience, which measure the ability of a system to resist and recover from perturbations, respectively. Self-regulation implies that there is some level of stability for a given system. In natural systems, there is not a fixed level of stability, instead the interactions of material cycles and energy flow, along with feedback mechanisms, generate a self-correcting homeorhesis (i.e., evolutionary stability of a system flow). These homorhetic controls keep the variability of the system's energy and material flows altering in the same way they have done in the past.

Network Connectivity

Ecosystems can be viewed as networks because the components of the system have direct transactions between them (e.g., predator–prey, competition, nutrient sequestration, among many others). While these interactions are discrete events, when viewed as a whole, these transactions link the direct and indirect components in an interconnected web. This web represents the network structure. System level properties emerge from the network, so viewing a species or event in isolation limits the ability to observe the higher-level properties. Networks are covered in detail in the Network section of this encyclopedia. Systems analysis and network science provide the analytical tools that embrace the holistic qualities of ecosystems.

The Systems Approach

Ecological systems are inherently complex and due to their emergent properties, most ecological systems are irreducible. These two factors reveal the necessity for a systems science to analyze and interpret ecological systems. Systems ecologists, in general, are interested in the quantification of system properties. There are several approaches to the study of ecosystems:

1. analytical studies,
2. comparative studies,
3. experimental studies, and
4. modeling or simulation studies.

Analytical studies collect pieces of information and attempts are made to integrate and synthesize these into a complete picture of the system. In comparative studies information is collected for a few components of several different systems, and these components are compared across system types. Experimental studies manipulate entire ecological systems to identify patterns and processes. Modeling of ecological systems involves abstracting the real system into either a verbal description, which is often represented as a conceptual model or a series of mathematical equations (quantitative model).

Ecological models represent formal descriptions of the ecosystem of interest. Models do not contain every element of the real system but include the characteristic features that are essential to the structure and function of the system within the context of the objectives of the study. Simulation is the process of using models to represent the behavior of the real-world system. Both models and simulation are invaluable tools for studying ecological systems. Models can be used to reveal systems properties, test hypotheses, and illuminate areas within a system that require more analytical research.

Each of these methods has strengths and weaknesses; however, they all attempt to achieve an understanding of an entire ecological system. Each method views the system as a functional unit. The systems approach, a collection of techniques used to identify and elucidate systems properties, is often used to analyze the properties of ecological systems. Regardless of the specific details of each method, in general, there are essentially five stages of systems analysis: (1) define and conceptualize the system, (2) quantify the relationships among system components by creating a mathematical model of the system, (3) evaluate the usefulness of the model, (4) apply the model toward the objectives of the study, and communication (Fig. 2). The process is inherently iterative as systems are redefined as more understanding of the complexities and interactions emerge.

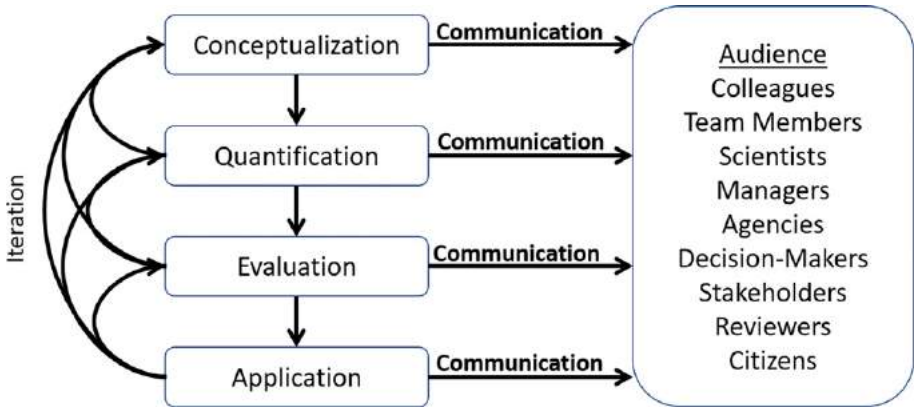


Fig. 2 Diagram of the modeling process. The process flows from conceptualization through application with communication being provided to an audience throughout the modeling process.

Conceptualization

Defining, or bounding, the system of interest is arguably the most critical stage of systems analysis. The idea is to develop a conceptual diagram of the real-world system that represents only the necessary components (i.e., enough of the components to reveal the system-level properties) and how they are related to each other. Important components are often identified by analytical, comparative, or manipulative studies. These components and their relationships are represented in a systems language using symbols that indicate the specific nature of the relationships. Two of the most common systems languages are Forrester diagrams and the energy circuit language developed by H.T. Odum. Each language uses a specific set of symbols, with each symbol indicating a specific function within the system. The major difference between the two systems languages is that Forrester diagrams allow materials and information to flow from one part of the diagram to another, while the energy circuit language is strictly used for representing how energy flows from place to place.

Quantification

The second stage of systems analysis involves translating the qualitative relationships diagrammed in the conceptual model into a series of mathematical equations. Depending on the system of interest, different types of mathematics can be used to solve the equations. Difference equations, differential equations, and matrix algebra are commonly used. The type of mathematics used depends on the objectives of the study. Due to the sheer number of components in ecological systems, analytical solutions rarely exist. Often, simulation is used as a tool to provide numerical solutions for complex models. The series of equations is solved at each time step over the entire period of simulated time. Equations are parametrized from either quantitative data gathered from analytical, comparative, or manipulative studies, or qualitative data which can come from the literature or from expert opinion.

Evaluation/Validation

The third stage of systems analysis is to evaluate the usefulness of the quantitative model in meeting the objectives of the study. This stage, commonly referred to as model validation, is the subject of much debate. This process considers the quantitative representation of the system from an array of distinct aspects, from comparing the results of the model to real-system observations to determining how sensitive model predictions are to changes in specific parameter values.

Application

The final stage of systems analysis is applying the quantitative model for its intended purpose. This stage involves designing and simulating the same experiments with the model that would be conducted in the real system. These results would be analyzed in the same way as empirical results.

Communication

As more complex models are used for large-scale planning and decision making, there is a need to clearly communicate the steps of model development to the appropriate audience. Model documentation should be focused on transparency, replicability, and

scientific defensibility. There are three main obstacles modelers face when communicating a model: confusion regarding a concept, difficulty picturing the system, and conflicts with preexisting cultural or mental models.

Confusion of a Concept

Developing a common vernacular with the desired audience is paramount for successful communication. Many disciplines use the same terminology, yet the underlying connotations of the terms can have different meanings. Overcoming this obstacle requires that modelers ask their audience how they interpret these terms. Similarly, modelers can develop a glossary of definitions for the audience. A common practice to clear up areas of misunderstanding is to develop an elucidating explanation that provides an example of the concept that lists the concept's critical features, and provides a variety of examples and nonexamples. This technique provides the audience with the opportunity to practice distinguishing the features that separate the example and nonexample.

Difficulty Picture of the System

Ecological systems models can be complex and can contain multiple feedback loops, threshold effects, and time lags. An audience can easily become confused when modelers obfuscate the concept by explaining too many details too quickly. In order to develop a useful picture of the overall system, modelers need to provide the audience with forming a general impression (i.e., big picture) of the system that they can grasp the systems components and interactions. Visual aids that depict the step-by-step construction of the model are useful for this task. Describing how the model was put together can be an onerous task, but it is crucial because it lends the process and the model more transparency. There is always more to describe about a model, but a clear presentation is the result of careful editing for a target audience.

Conflicts With Preexisting Models

Every person's intuitive understanding how the world works is framed by that person's specific life history. When a model directly conflicts with audience members' logic or beliefs, the cognitive dissonance disrupts their receptiveness to the assumptions upon which the model was built. It is difficult to overcome, but the use of transformative explanations that acknowledge the audiences preexisting notions of a system, then provide an alternative explanation, can be useful. Effective transformative explanations do not dismiss or insult preexisting models, but rather they help provide the audience with alternative perspectives for interpreting familiar systems.

Iterative Nature of Systems Analysis

The five stages of systems analysis are highly interconnected. This systems approach is a heuristic process. Often components have been overlooked or misrepresented, requiring returning to an earlier step, often conceptualization or quantification.

Synthesizing analytical results illuminates system properties. Often this synthesis identifies where more analytical research is required. The new results will then be incorporated into the synthesis, which in turn will identify the further need for more analytical data. Analytical results are needed to provide components for synthesis, and synthesis identifies the priorities for analysis. Likewise, communicating with interested parties throughout the process can result in components be reconceptualized, which leads to a new synthesis of the information.

Summary

Ecological systems are open systems and exhibit organized complexity. They are complex because they are composed of many interacting components whose interactions are characterized by time lags, thresholds, feedback, and indirect effects. Ecological systems vary in size and scale, depending on the processes being studied or the questions being asked. Typical reductionistic methods do not illuminate system properties. In order to properly study ecological systems, methods that embrace the system as a unit are required. The systems approach is characterized by the conceptualization, quantification, evaluation, application, and communication of a model representing the ecological system of interest. This approach, including systems analysis and simulation, is necessary because the cause–effect relationships within complex ecological systems can only be identified by examining the system as a single, functionally integrated entity.

Further Reading

von Bertalanffy, L., 1969. General system theory: Foundations, development, applications. New York: George Braziller.
Forrester, J.W., 1961. Industrial dynamics. Cambridge: MIT Press.
Grant, W.E., Swannack, T.M., 2008. Ecological modeling: A common-sense approach to theory and practice. Hoboken: John Wiley & Sons.
Jørgensen, S.E., Fath, B.D., 2011. Fundamentals of ecological modeling, 4th edn Amsterdam: Elsevier.

Jørgensen, S.E., Fath, B., Bastianoni, S., *et al*., 2007. A new ecology: Systems perspective. Amsterdam: Elsevier.
Kitching, R.L., 1983. Systems ecology. St. Lucia: University of Queensland Press.
Likens, G.E., 1985. An ecosystem approach to aquatic ecology: Mirror lake and its environment. New York: Springer.
Meadows, D.H., 2008. Thinking in systems: A primer. Hartford: Chelsea Green Publishing.
Odum, H.T., 1971. Environment, power, and society. New York: Wiley.
Odum, E.P., 1977. The emergence of ecology as a new integrative discipline. Science 195, 1289–1293.
Odum, H.T., 1983. Systems ecology. New York: Wiley.
Odum, H.T., 2000. Modeling for all scales: An introduction to systems simulation. San Diego: Academic Press.
Odum, E.P., Barrett, G.W., 2005. Fundamentals of ecology, 5th edn Belmont, CA: Thompson Brooks/Cole.
O'Neill, R.V., DeAngelis, D.L., Waide, J.B., Allen, T.F.H., 1986. A hierarchical concept of ecosystems. Princeton, NJ: Princeton University Press.
Patten, B.C., 1971. Systems analysis and simulation in ecology. vol. I. New York: Academic Press.
Patten, B.C., 1972. Systems analysis and simulation in ecology. vol. II. New York: Academic Press.
Patten, B.C., 1975. Systems analysis and simulation in ecology. vol. III. New York: Academic Press.
Patten, B.C., 1976. Systems analysis and simulation in ecology. vol. IV. New York: Academic Press.
Patten, B.C., Jørgensen, S.E., 1995. Complex ecology. Upper Saddle River, NJ: Prentice Hall.
Solé, R., Bascompte, J., 2006. Self-organization in complex ecosystems. Princeton, NJ: Princeton University Press.
Tansley, A.G., 1935. The use and abuse of vegetational concepts and terms. Ecology 16, 284–307.

Systems Ecology: Ecological Network Analysis☆

Brian D Fath, Towson University, Towson, MD, United States; International Institute for Applied Systems Analysis, Laxenburg, Austria
Ursula M Scharler, University of KwaZulu-Natal, Durban, South Africa

Introduction

Ecological Network Analysis (ENA) uses network theory to study the interactions between organisms or populations within their environment. ENA follows along the synecology perspective introduced by EP Odum which is predominantly concerned with interrelations of material, energy, and information among system components. Bernard Patten was the originator of the environ analysis approach in the mid-to-late 1970s and along with colleagues expanded the analysis to reveal many insightful, holistic properties of ecosystem organization. During about the same time, Robert Ulanowicz developed several system's measures for networks motivated by information theory to analyze the system flows. This approach has been applied in ecological and socio-economic settings.

ENA starts with the assumption that a system can be represented as a network of nodes (compartments, vertices, components, storages, objects, etc.) and the connections between them (links, arcs, flows, etc.). In ecological systems, the connections are usually based on the flow of energy or matter (water, nutrients, etc.) between the system compartments, and in socioeconomic networks of money, trade, water, or social interactions. If such a flow exists, then there is a direct transaction between the two connected compartments. These direct transactions give rise to both direct and indirect relations between all the objects in the system. Network analysis provides a systems-oriented perspective because it is based on uncovering patterns and relations among all the objects in a system. Therefore, it gives a view on how system components are tied to a larger web of interactions.

Theoretical Developments of Ecological Network Analysis

The development of Environ Analysis was motivated by Patten to attempt to answer the question, "What is environment?". In order to study environment as a formal object, a system boundary is a necessary condition to avoid the issue of infinite indirectness, because in principle, one could trace the environment of each object back in history to the big bang origins. The realization of a boundary is, in fact, one of the three foundational principles in his seminal paper introducing the environ theory concept (Patten, 1978). The necessary boundary demarcates now two environments, the unbound external environment, which indeed includes all space–time objects in the universe, and the second internal, contained environment of interest. This quantifiable, internal environment for each system object is termed "environ," and is the study of Environ Analysis. An object's environ stops at the system boundary, but as ecosystems are open systems, they require exchanges across the boundary into and out of the system.

☆*Change History*: April 2018. Brian D. Fath and Ursula M. Scharler. The new article is a combination of two articles in the earlier edition. Previously there was one article Ecological Network Analysis, Ascendancy by Scharler and one titled Ecological Network Analysis, Environ Analysis by Fath.

Therefore, input and output boundary flows are necessary to maintain the system functioning at far from thermodynamic equilibrium. Objects and connections that reside wholly in the external environment are not germane to the analysis.

Another foundational principle of environ analysis theory is that each object in the system itself has two "environs" one on the receiving end and one on the generating end of interactions in the system. In other words, an object's input environ includes those flows from within the system boundary leading up to the object, and an output environ, those flows emanating from the object back to the other system objects before exiting the system boundary. This alters the perception from internal–external to receiving–generating. Thus, the object, while distinct in time and space, is more clearly embedded in and responsive to the couplings with other objects within the network. This shifts the focus from the objects themselves to the relations they maintain; or from parts to processes (or what Ilya Prigogine called from Being to Becoming).

The third foundational principle is that individual environs (and the flow carried within each one) are unique such that the system comprises the set union of all environs, which in turn partition the system level of organization. This partitioning allows one to classify environ flow into what have been called different modes: (1) boundary input; (2) first passage flow received by a compartment from other compartments in the system [i.e., not boundary flow, but also not cycled flow (in other words first time flow reaching a compartment)]; (3) cycled flow that returns to a compartment before leaving the system; (4) dissipative flow in that it has left the focal compartment not to return, but does not directly cross a system boundary (i.e., it flows to another within system compartment); and (5) boundary output. The modes have been used to understand better the general role of cycling and the flow contributions from each object to the other, which has had application in showing a complementarity of several of the holistic, thermodynamic-based ecological indicators (see Fath et al., 2001).

The link to thermodynamics has been incorporated into several other ecosystem measures and methodologies, used to calculate structural and functional systems indicators from the number and weight of links, or from the flows and associated biomass, or size, of nodes. Many of these have their roots in macroecological considerations, and were conceived by the desire to quantify the growth, succession and development of ecosystems (e.g., Ulanowicz, 1986), within the framework of "process ecology." Process ecology focuses less on the biomass (storage) of species, which often in conjunction with abundance is the main focus of ecologists. Rather, processes such as the biomass (or energy, nutrients, etc.) fluxes between species (nodes), recycling of material, decomposition, or production are the main focus. These lead to emerging properties on the ecosystem level, which links it to macro-ecology.

Data Requirements and Community Assembly Rules

On one level Ecological Network Analysis could be referred to as a holistic/reductionistic approach. It is holistic because it considers simultaneously the whole influence of all system objects, yet it is reductionistic in that the fine details of all transactions are entailed in the analysis. In other words, it is the opposite of a black box model. The network data requirements are considerable, which include the complete flow–storage quantities for each identified link and node (note flow and storage are interchangeable as determined by the turnover rate). Data can be acquired from empirical observations, literature estimates, model simulation results, or balancing procedures, when all but a few are unknown. While there is difficulty in obtaining complete network datasets, a number of ecosystem network models are available from various databases [see EcoPath (http://sirs.agrocampus-ouest.fr/EcoBase/), Ulanowicz (https://www.cbl.umces.edu/~ulan/ntwk/network.html), or Borrett (http://people.uncw.edu/borretts/research.html)].Due to this lack of requisite data for fully quantified food webs, researchers have developed community assembly rules that are heuristics to construct ecological food webs. Assembly rules are in general a set of rules that will generate a connectance matrix for a number of N species. Common assembly rules that have been developed are random or constant connectance, cascade, niche, modified niche, and cyber-ecosystem each with its own assumptions and limitations (see Halnes et al. (2007) for a review of these methods). In all but the last case, the assembly rules construct only the structural food web topology. The cyber-ecosystem methodology (Fath, 2004) also includes a procedure for quantifying the flows along each link. It uses a meta-structure of six functional groups: Producer (P), Herbivore (H), Carnivore (C), Omnivore (O), Detritus (D), and Detrital Feeders (F), within which random connections link species based on these definitional constraints.

Methods and Sample Network

To demonstrate a few network analysis methodologies it is best to proceed with an example. Consider the network in Fig. 1, which has five compartments or nodes (x_i, for $i = 1$ to 5). Compartments are connected by transaction of the energy–matter substance flowing between them. These pair-wise couplings are the basis for the internal network structure. A structural connectance matrix, or adjacency matrix, **A**, is a binary representation of the connections such that $a_{ij} = 1$ if there is a connection from j to i, and a zero otherwise (Eq. 1).

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \tag{1}$$

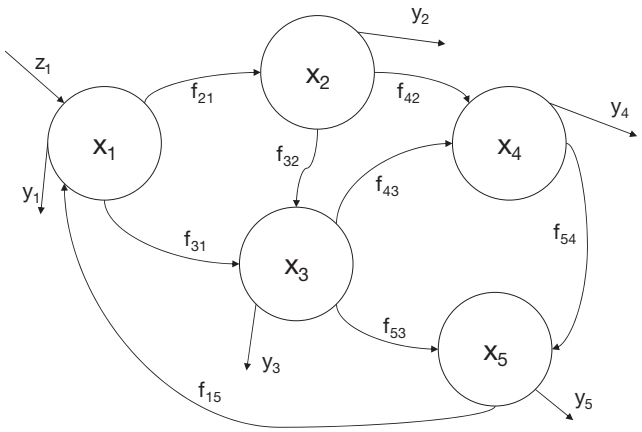


Fig. 1 Sample Network with five compartments use to demonstrate Environ Analysis notation and methodology.

Storage and flows must have consistent units (although it is possible to consider multi-unit networks). Typically, units for storages are given in amount of energy or biomass per given area or volume (e.g., g/m^2), and units for flows are the same but as a rate [e.g., g/(m^{2*}day)]. The intercompartmental flows for Fig. 1 are given in the following flow matrix, **F**:

$$\mathbf{F} = \begin{bmatrix} 0 & 0 & 0 & 0 & f_{15} \\ f_{21} & 0 & 0 & 0 & 0 \\ f_{31} & f_{32} & 0 & 0 & 0 \\ 0 & f_{42} & f_{43} & 0 & 0 \\ 0 & 0 & f_{53} & f_{54} & 0 \end{bmatrix} \tag{2}$$

Note that the orientation of flow from j to i is used in Environ Analysis because that makes the direction of ecological relation from i to j. For example, if i preys on j, the flow of energy is from j to i. To the contrary, other types of ecological network analyses direct the flow of energy from i to j when predator j feeds on i. It is therefore important to check the matrix setup before embarking on an analysis. All compartments experience dissipative flow losses (y_i, for $i = 1$ to 5), and here the first compartment receives external flow input, z_1 (arrows starting or ending not on another compartment represent boundary flows). For this example, these can be given as:

$$y = [y_1 \;\; y_2 \;\; y_3 \;\; y_4 \;\; y_5] \tag{3}$$

and

$$z = \begin{bmatrix} z_1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{4}$$

In this approach, there is no distinction between the types of outflow; whereas with other approach it is split into useable export and respiration loss. Total throughflow of each compartment is an important variable, which is the sum of flows into, $T_i^{in} = z_i + \sum_j^n f_{ij}$, or out of, $T_i^{out} = y_i + \sum_j^n f_{ji}$ the ith compartment. At steady state, compartmental inflows and outflows are equal such that $dx_i/dt = 0$, and therefore, incoming and outgoing throughflows are equal also: $T_i^{in} = T_i^{out} = T_i$. In vector notation, compartmental throughflows are given by:

$$T = \begin{bmatrix} T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \end{bmatrix} \tag{5}$$

The sum of flows in a network can also be depicted as total system throughput, which is defined as the sum of flows in the system including ingoing and outgoing boundary flows, and all internal flows from i to j, TSTput $= \sum_{j=1}^n \sum_{i=1}^n f_{ij} + \sum_{i=1}^n E_i + \sum_{i=1}^n R_i + \sum_{j=1}^n z_j$ where E is the export of usable material, and R the respiration losses across the system boundary.

This basic information regarding the storages, flows, and boundary flows provides all the necessary information to conduct Ecological Network Analysis. Environ Analysis has been classified into a structural analysis—dealing only with the network topology, and three functional analyses—flow, storage, and utility—which require the numerical values for flow and storage in

the network (Table 1). Some of the analyses have been developed in tandem (e.g., utility analysis and mixed trophic impacts analysis) and their differences and similarities are pointed out below. In addition, an illustration of ascendency analysis follows.

The technical aspects of environ analysis are explained in detail elsewhere, so rather than repeat those here, the remainder of the entry highlights some of the important results from environ analysis. But first, one issue that must be covered is the way in which network analysis identifies and quantifies indirect pathways and flow contributions. Indirectness originates from transfers or interactions that occur nondirectly, and are mediated by other within system compartments. These transfers could travel two, three, four, or many links before reaching the target destination. For example, the flow analysis starts with the calculation of the nondimensional flow intensity matrix, **B**, where $b_{ij} = f_{ij}/T_j$. The generalized, **B** matrix corresponding to Fig. 1 would look as follows:

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 & b_{15} \\ b_{21} & 0 & 0 & 0 & 0 \\ b_{31} & b_{32} & 0 & 0 & 0 \\ 0 & b_{42} & b_{43} & 0 & 0 \\ 0 & 0 & b_{53} & b_{54} & 0 \end{bmatrix} \tag{6}$$

These values represent the fraction of flow along each link normalized by the total throughflow at the donating compartment. This elements give the direct, measurable flow intensities (or probabilities) between any two nodes j to i. To identify the flow intensities along indirect paths (e.g., $j \rightarrow k \rightarrow i$), one need only consider the matrix **B** raised to the power equal to the path length in question. For example, $\mathbf{B}^2$ gives the flow intensities along all paths of length 2, $\mathbf{B}^3$ along all paths of length 3, etc. This well-known matrix algebra result is the primary tool to uncover system indirectness. In fact, it turns out that due to the way in which the **B** matrix is constructed, all elements in $\mathbf{B}^m$ go to zero as $m \rightarrow \infty$. Therefore, it is possible to sum the terms of $\mathbf{B}^m$ to acquire an "integral" flow matrix (called **N**), which gives the flow contribution from all path lengths.

$$N = \mathbf{B}^0 + \mathbf{B}^1 + \mathbf{B}^2 + \mathbf{B}^3 + \ldots = \sum_{m=0}^{\infty} \mathbf{B}^m = (\mathrm{I} - \mathbf{B})^{-1} \tag{7}$$

where $\mathbf{B}^0 = \mathrm{I}$, the identity matrix, $\mathbf{B}^1$ the direct flows, and $\mathbf{B}^m$ for $m > 1$ are all the indirect flows intensities. Note, that the elements of **B** and **N** are nondimensional; to retrieve back the actual throughflows, one need only multiply the integral matrix by the input vector: $\mathrm{T} = \mathbf{N}z$. In other words, **N** redistributes the input, z, throughout each compartment to recover the total flow through that compartment. Similarly, one could acquire any of the direct or indirect flows by multiplying $\mathbf{B}^m z$ for any m.

A similar argument is made to develop integral storage and utility matrices.

$$\textit{Storage}: \quad Q = P^0 + P^1 + P^2 + P^3 + \cdots = \sum_{m=0}^{\infty} P^m = (I - P)^{-1} \tag{8}$$

$$\textit{Utility}: \quad U = D^0 + D^1 + D^2 + D^3 + \cdots = \sum_{m=0}^{\infty} D^m = (I - D)^{-1} \tag{9}$$

where $p_{ij} = (f_{ij}/x_j)\Delta t$, and $d_{ij} = (f_{ij} - f_{ji})/T_i$.

Network Properties

Patten has developed a series of "ecological network properties" which summarize the results of environ analysis. These have all been described in the literature (for an overview, see Patten, 2016; Jørgensen et al., 2007). The properties have been used to assess the current state of ecosystem networks and to compare the state of different networks. Furthermore, while interpreting some of the properties as ecological goal functions, it has been possible to identify the structural or parametric configurations that positively affect the network property values as a way to detect or anticipate network changes. For example, certain network alterations, such as increased cycling, lead to greater total system energy throughflow and energy storage, so one could expect that if possible ecological

Table 1 Basic methodologies for Network Environ Analysis

Structural analysis	*Functional analyses*
Path Analysis: a_{ij} Enumerates pathways in a network (connectance, cyclicity, etc.)	*Flow analysis*: $b_{ij} = f_{ij}/T_j$ Identifies flow intensities along indirect pathways *Storage analysis*: $c_{ij} = f_{ij}/x_j$ Identifies storage intensities along indirect pathways *Utility analysis*: $d_{ij} = (f_{ij} - f_{ji})/T_i$ Identifies utility intensities along indirect pathways

networks are evolving or adapting to such configurations. This leads to a new area of research on evolving networks. In this section, a brief overview is given for four of these properties: dominance of indirect effects (or nonlocality), network homogenization, network mutualism, and environs themselves.

Dominance of Indirect Effects

This property compares the contribution of flow along indirect pathways with those along direct ones. Indirect effects are any that require an intermediary node to mediate the transfer and can be of any length. The strength of indirectness has been measured in a ratio of the sum of the indirect flows intensities divided by the direct flow intensities:

$$\frac{\sum_{i,j=1}^{n}(n_{ij}-b_{ij}-\delta_{ij})}{\sum_{i,j=1}^{n}b_{ij}} \tag{10}$$

where δ_{ij}, the Kronecker delta, $=1$ if and only if $i=j$ and is 0 otherwise. When the ratio is greater than one, then dominance of indirect effects is said to occur. Analysis of many different models has shown that this ratio is often greater than one, revealing the nonintuitive result that indirect effects have greater contribution than direct effects. Thus, each compartment influences each other, often significantly, by many indirect, nonobvious pathways. The implications of this important result are clear in that each compartment is embedded in and dependent on the rest of the network for its situation, thus calling for a true systems approach to understand such things as feedback and distributed control in the network.

Network homogenization

The homogenization property yields a comparison of resource distribution between the direct and integral flow intensity matrices. Due to the contribution of indirect pathways, it was observed that flow in the integral matrix was more evenly distributed than that in the direct matrix. A statistical comparison of resources distribution can be made by calculating the coefficient of variation of each of the two matrices. For example, the coefficient of variation of the direct flow intensity matrix **B** is given by:

$$CV(G)=\frac{\sum_{j=1}^{n}\sum_{i=1}^{n}\left(\bar{g}_{ij}-g_{ij}\right)^2}{(n-1)\bar{g}} \tag{11}$$

$$CV(B)=\frac{\sum_{i,j=1}^{n}(\bar{b}_{ij}-b_{ij})^2}{(n-1)\bar{b}} \tag{12}$$

Network homogenization occurs when the coefficient of variation of **N** is less than the coefficient of variation of **B** because this says that the network flow is more evenly distributed in the integral matrix. The test statistic employed here looks at whether or not the ratio $CV(\mathbf{B})/CV(\mathbf{N})$ exceeds one. The interpretation again is clear that the view of flow in ecosystems is not as discrete as it appears because in fact the material is well-mixed (i.e., homogenized) and has traveled through and continues to travel through many, if not, most parts of the system.

Network mutualism

Turning now to the utility analysis, the net flow, utility matrix, **D**, can be used to determine quantitatively and qualitatively the relations between any two components in the network such as predation, mutualism, or competition. Entries in the direct utility matrix, **D**, or integral utility matrix, **U**, can be positive or negative ($-1 \le d_{ij}, u_{ij} < 1$). The elements of **D** represent the direct relation between that (i, j) pairing; for the example in Fig. 1, this produces the following:

$$\mathbf{D}=\begin{bmatrix} 0 & -\frac{f_{21}}{T_2} & -\frac{f_{31}}{T_3} & 0 & \frac{f_{15}}{T_5} \\ \frac{f_{21}}{T_2} & 0 & -\frac{f_{32}}{T_3} & -\frac{f_{42}}{T_4} & 0 \\ \frac{f_{31}}{T_3} & \frac{f_{32}}{T_3} & 0 & -\frac{f_{43}}{T_4} & -\frac{f_{53}}{T_5} \\ 0 & \frac{f_{42}}{T_4} & \frac{f_{43}}{T_4} & 0 & -\frac{f_{54}}{T_5} \\ -\frac{f_{15}}{T_5} & 0 & \frac{f_{53}}{T_5} & \frac{f_{54}}{T_5} & 0 \end{bmatrix} \tag{13}$$

The direct matrix **D**, being zero-sum, always has the same number of positive and negative signs.

$$\operatorname{sgn}(\mathbf{D}) = \begin{bmatrix} 0 & - & - & 0 & + \\ + & 0 & - & - & 0 \\ + & + & 0 & - & - \\ 0 & + & + & 0 & - \\ - & 0 & + & + & 0 \end{bmatrix} \tag{14}$$

The elements of **U** provide the integral, system-determined relations. Kazanci and Adams (2017) recently showed that convergence of the D matrix can always be achieved. Continuing the example, and now including flow values derived from 10% transfer efficiency along each link ($b_{ij} = 0.10$, if $a_{ij} = 1$, and $b_{ij} = 0$ otherwise), we get the following integral relations between compartments:

$$\operatorname{sgn}(\mathbf{U}) = \begin{bmatrix} + & - & - & + & + \\ + & + & - & - & + \\ + & - & + & - & - \\ + & + & + & + & - \\ + & + & + & + & + \end{bmatrix} \tag{15}$$

Unlike, the direct relations, this is not zero-sum. Instead, we see that there are 17 positive signs (including the diagonal) and 8 negatives signs. If there are a greater number of positive signs than negative signs in the integral utility matrix, then network mutualism is said to occur. Network mutualism reveals the preponderance of positive mutualistic relations in the system. Specifically, here, we can identify two cases of indirect mutualism, seven of exploitation, and one competition (Table 2).

Mixed Trophic Impacts Analysis (MTI)

Utility analysis (UA) and MTI have been developed in parallel. Essentially both analyses strive to elucidate the type of interactions between two nodes, both along direct and indirect pathways. Their similarities and differences have been defined in Scharler et al. (2009), and below is given a brief outline. At the start of the analyses, the orientation of fluxes between nodes differs in the **F** matrix between UA and MTI analysis, as pointed out above. Next, the imports across the system boundary are taken into account in the UA analysis (D^0) but not in the MTI analysis (Q^0 omitted), leading to more negative numbers along the diagonal in the M matrix of the MTI analysis. The signs of the relations over direct and indirect effects are used in both the MTI and UA analyses as qualitative descriptions of node relation. In UA, however, the fractions of the matrix are also re-dimensionalized by their total throughflow. All respiration flows are excluded in the calculation of impacts in the MTI analysis (Ulanowicz and Puccia, 1990). In UA they are included in the total system throughflow for each compartment. Lastly, the impacts to detritus are taken as zero in the MTI analysis, as the detritus nodes do not compete for these flows in contrast to living nodes, which actively impact their source nodes. In the UA analysis, on the other hand, a detritus node has a negative impact on the source node. This variation in weighting may give a reason for the different outcomes. Bearing these differences in mind, applying an MTI analysis to the example illustrated for the UA analysis above yields the following results.

In the MTI approach the positive effect of a prey on its predator is expressed as:

$$g_{ij} = f_{ij} / \sum_{k=1}^{n} f_{kj}, \tag{16}$$

where k represents all diets of j, and g_{ij} ranges from 0 to 1. The negative impact of the predator on its prey, is defined as:

Table 2 Direct and integral relations in sample network from Fig. 1

Direct	*Integral*
$(sd_{21}, sd_{12}) = (+, -)$ → exploitation	$(sd_{21}, sd_{12}) = (+, -)$ → exploitation
$(sd_{31}, sd_{13}) = (+, -)$ → exploitation	$(sd_{31}, sd_{13}) = (+, -)$ → exploitation
$(sd_{41}, sd_{14}) = (0, 0)$ → neutralism	$(sd_{41}, sd_{14}) = (+, +)$ → mutualism
$(sd_{51}, sd_{15}) = (-, +)$ → exploited	$(sd_{51}, sd_{15}) = (+, +)$ → mutualism
$(sd_{32}, sd_{23}) = (+, -)$ → exploitation	$(sd_{32}, sd_{23}) = (-, -)$ → competition
$(sd_{42}, sd_{24}) = (+, -)$ → exploitation	$(sd_{42}, sd_{24}) = (+, -)$ → exploitation
$(sd_{52}, sd_{25}) = (0, 0)$ → neutralism	$(sd_{52}, sd_{25}) = (+, -)$ → exploitation
$(sd_{43}, sd_{34}) = (+, -)$ → exploitation	$(sd_{43}, sd_{34}) = (+, -)$ → exploitation
$(sd_{53}, sd_{35}) = (+, -)$ → exploitation	$(sd_{53}, sd_{35}) = (+, -)$ → exploitation
$(sd_{54}, sd_{45}) = (+, -)$ → exploitation	$(sd_{54}, sd_{45}) = (+, -)$ → exploitation

$$h_{ij} = f_{ij} / \sum_{m=1}^{n} f_{im}, \tag{17}$$

where m represents all consumers of i (note, in the original MTI literature this matrix is called f_{ij}, but is renamed here h_{ij} to avoid confusion with the notation defined above).

The direct net impact (q_{ij}) of a network compartment on another is depicted as the difference of the positive impact the prey has on the predator (g_{ij}) minus the negative impact the predator has on its prey (h_{ij}):

$$q_{ij} = g_{ij} - h_{ji} \tag{18}$$

The indirect interactions (M) are calculated, similar to the Utility Analysis, by summing the powers of the matrices describing indirect steps over various path lengths. Note that imports across the system boundary are not part of the MTI analysis.

$$M = Q + Q^2 + Q^3 + Q^4 + \ldots \tag{19}$$

where Q, as defined in Eq. (18), is the matrix of the direct impacts of i upon j (path length of one), Q^2 the indirect impacts of i upon j over a path length of 2, Q^3 the indirect impacts of i upon j over a path length of 3, etc. Assuming the infinite power series converges, Eq. (19) can be written using the commonly known closed form solution as:

$$M = (I - Q)^{-1} - I \tag{20}$$

For the above example, the Q matrix of direct interactions, and M matrix for the integrated interactions are as follows:

Direct (Q)	*Integral (M)*
$(sd_{12}, sd_{21}) = (+, -)$ → exploitation	$(sd_{12}, sd_{21}) = (+, -)$ → exploitation
$(sd_{13}, sd_{31}) = (+, -)$ → exploitation	$(sd_{13}, sd_{31}) = (+, +)$ → mutualism
$(sd_{14}, sd_{41}) = (0, 0)$ → neutralism	$(sd_{14}, sd_{41}) = (+, +)$ → mutualism
$(sd_{15}, sd_{51}) = (-, +)$ → exploited	$(sd_{15}, sd_{51}) = (-, +)$ → exploited
$(sd_{23}, sd_{32}) = (+, -)$ → exploitation	$(sd_{23}, sd_{32}) = (-, -)$ → competition
$(sd_{24}, sd_{42}) = (+, -)$ → exploitation	$(sd_{24}, sd_{42}) = (+, -)$ → exploitation
$(sd_{25}, sd_{52}) = (0, 0)$ → neutralism	$(sd_{25}, sd_{52}) = (+, +)$ → mutualism
$(sd_{34}, sd_{43}) = (+, -)$ → exploitation	$(sd_{34}, sd_{43}) = (-, -)$ → competition
$(sd_{35}, sd_{53}) = (+, -)$ → exploitation	$(sd_{35}, sd_{53}) = (+, +)$ → mutualism
$(sd_{45}, sd_{54}) = (+, -)$ → exploitation	$(sd_{45}, sd_{54}) = (+, -)$ → exploitation

Environ analysis

This property mentioned here is the signature property, the quantitative environ, both in the input and output orientation. Since each compartment has two distinct environs there are in fact $2n$ environs in total. The output environ, **E**, for the ith node is calculated as:

$$E = (B - I)\hat{N}_i \tag{21}$$

where $\hat{N}_i$ is the diagonalized matrix of the ith column of **N**. When assembled, the result is the output oriented flow from each compartment to each other compartment in the system and across the system boundary. Input environs are calculated as:

$$E = \hat{N'}_i(B' - I) \tag{22}$$

where, $b'_{ij} = f_{ij}/T_i$, and $N' = (I - B')^{-1}$. These results comprise the foundation of Network Environ Analysis since they allow for the quantification of all within system interactions, both direct and indirect, on a compartment-by-compartment basis.

Ascendency analysis

The ascendency indicator was developed to quantify both the growth and development of ecosystems. Describing two attributes of ecosystems, ascendency consists of the total system throughput describing growth, and the average mutual information (AMI) describing how constrained fluxes are as they move from node to node (Fig. 2).

The total system throughput ($T..$) is described above, and a description of the AMI follows here. The degree of constraint of a flux between two nodes is described from its flow weight (T_{ij}), by the proportional weight of all flows entering the sink node in relation to total system flows ($T_j/T..$) and lastly by the proportion T_{ij} takes of all flows leaving i ($T_{ij}/T..$).

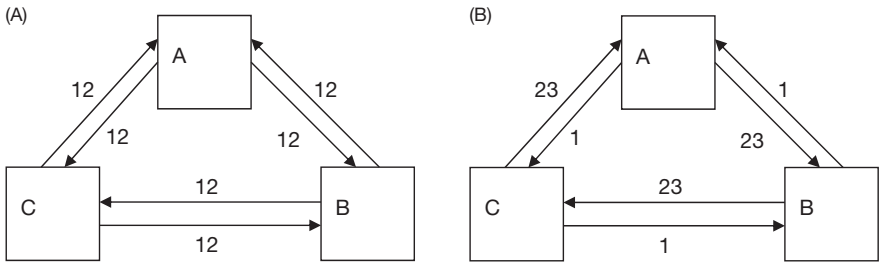


Fig. 2 (A) Hypothetical unconstrained network. Low AMI. (B) Hypothetical constrained network. Higher AMI.

Average Mutual Information

The probability that a quantum of material flows along the highly frequented routes is higher compared to a network where all routes transfer the same amount of material. Conversely, the probability that a quantum of material flows along the less frequented routes will be lower compared to a network where all routes transfer the same amount of material. Such a change in probability can be quantified with the help of information theory. Information is defined as the agent which causes a change in probability. Ulanowicz uses the term information to describe "the effects of that which imparts order and pattern to the system."

The most indeterminate, or least constrained, network is one where all compartments are connected with each other and where, in proportion to the compartmental throughput, equal amounts of material flow along the ingoing and outgoing pathways. Quantifying the information which is gained by transferring material along more and less frequented routes thus gives a clue about the unevenness of material flowing along pathways. The change in probability from a situation where a quantum of material flows along an equiprobable pathway (Fig. 2) and along a pathway which is not equiprobable (Fig. 2B) is calculated using conditional probabilities.

Since, from an ecological network point of view, joint and conditional probabilities refer to transfers of material from compartment i to compartment j, the above formula can be rewritten as

$$\mathrm{AMI} = K \sum_{i,j} \left(\frac{T_{ij}}{T_{..}} \right) \log \left(\frac{T_{ij} T_{..}}{T_{i.} T_{.j}} \right) \tag{23}$$

where the joint probability of a quantum of material $(p(a_i, b_j))$ flowing from species i to species j can be denoted as $T_{ij}/T_{..}$, remembering that the events in an events table are material flows in a system. $T_{..}$ is the total system throughput, or the sum over all combinations of T_{ij}.

Ascendency

The scalar constant, k, has been retained throughout all calculations. To be able to combine growth and development into one single index, k is substituted by the total system throughput or TST in order to scale the AMI to the size of the system in question. The resulting index is called Ascendency and is denoted by

$$A = TST \sum_{i,j} \left(\frac{T_{ij}}{T_{..}} \right) \log \left(\frac{T_{ij} T_{..}}{T_{i.} T_{.j}} \right) \tag{24a}$$

or

$$A = \sum_{i,j} T_{ij} \log \left(\frac{T_{ij} T_{..}}{T_{i.} T_{.j}} \right) \tag{24b}$$

Besides indirect mutualism there are a number of influences that can change the ascendency of a system. These influences are thought to not have a favored direction of change, whereas indirect mutualism is believed to drive development toward increased ascendency. Mutualism is furthermore not a result of events elsewhere in the system's hierarchy but can arise at any level. Previously, it was theorized that in the absence of overwhelming external disturbances, the ascendency of a system has a propensity to increase, that is, both activity (TST) and structure (AMI) increase. More recent research (e.g., Goerner et al., 2009; Ulanowicz et al., 2009) has shown that there are limits to the increase in ascendency (see section "Window of Vitality" below).

In theory, ascendency is higher when pathways are less in numbers (more specialization) and more articulated (few pathways transport most of the material). The highest theoretical value of ascendency is achieved when all players in the system have one input and one output only, and are thus joined in one big single loop. This configuration mirrors highest specialization, and in this case AMI = H. This situation cannot be achieved in real systems, due to reasons discussed below under the heading "Overhead" and "Window of Vitality."

Development Capacity

As mentioned above, the limit to development is set by Shannon's diversity index pertaining to the material transfers or flows. MacArthur applied Shannon's diversity index to the material flows in an ecosystem to arrive at a measure for the diversity of flows, H:

$$H = -k \sum_{i,j} \left(\frac{T_{ij}}{T_{..}}\right) \log \left(\frac{T_{ij}}{T_{..}}\right) \tag{25}$$

where k is a scalar constant, and $T_{..}$ is the total system throughput (TST), or the sum over all combinations of T_{ij}.

H can, like the AMI, be multiplied by TST to scale the diversity of flows to the system in question. TST $\times$ H is called the development capacity, or limit for development, C:

$$C = -TST \sum_{i,j} \left(\frac{T_{ij}}{T_{..}}\right) \log \left(\frac{T_{ij}}{T_{..}}\right) \tag{26a}$$

or

$$C = -\sum_{i,j} T_{ij} \log \left(\frac{T_{ij}}{T_{..}}\right) \tag{26b}$$

The initial complexity, H, consists of two elements. One is the AMI, describing the information gained by reducing the uncertainty in flow probability. It is an index of the organized part of the system. The other is the residual uncertainty, or H_c (also called conditional diversity). Thus, $H = \text{AMI} + H_c$.

H_c or Overhead

The overhead represents the unorganized, inefficient and indeterminate part of the flow structure and is considered an insurance for the system. Should the system become overly organized (high ascendency), it will also be prone to perturbations. The overhead is split into four components: overhead due to imports, exports, respiration and internal pathways.

The combined overhead is denoted by:

$$H_c = -k \sum_{i,j} \left(\frac{T_{ij}}{T_{..}}\right) \log \left(\frac{T_{ij}^2}{T_{i.} T_{.j}}\right) \tag{27}$$

Scaling H_c to the system by replacing k with by TSTput yields

$$\Phi = -\sum_{i,j} T_{ij} \log \left(\frac{T_{ij}^2}{T_{i.} T_{.j}}\right) \tag{28}$$

The relationship between C, A and Φ so becomes $C = A + \Phi$.

Similar terms can be derived for the flows that are imports, respiration, and exports.

The formula for the overhead on imports is as follows:

$$\Phi_I = -\sum_{j=1}^{n} T_{0j} \log \left(\frac{T_{0j}^2}{T_{0.} T_{.j}}\right) \tag{29}$$

where imports are assumed to originate from the environment labeled compartment 0.

The overhead on exports is denoted by:

$$\Phi_E = -\sum_{i=1}^{n} T_{i,\,n+1} \log \left(\frac{T_{i,\,n+1}^2}{T_{i.} T_{.,\,n+1}}\right) \tag{30}$$

where exports are assumed to flow into a fictitious compartment $n+1$.

The overhead on dissipation is:

$$\Phi_D = -\sum_{i=1}^{n} T_{i,\,n+2} \log \left(\frac{T_{i,\,n+2}^2}{T_{i.} T_{.,\,n+2}}\right) \tag{31}$$

where respiration is assumed to flow into a fictitious compartment $n+2$.

Redundancy

The fourth part of the overhead is that of internal transfers and represents the extent of pathway redundancy. There are disadvantages to the system in maintaining redundant, or parallel pathways. For once, there can be an increase in dissipations, if transfers occur not only along one most efficient route, but along more than one route. Also, the resource transferred along different parallel pathways might not always end up at the right time at the consumer.

An obvious advantage of parallel pathways is the insurance of having more than one route of transfer in case of disturbances of other routes. Redundancy is denoted by:

$$R = -\sum_{i=1}^{n}\sum_{j=1}^{n} T_{i,j}\log\left(\frac{T_{i,j}^2}{T_{i.}T_{.j}}\right) \tag{32}$$

Window of Vitality

The Window of Vitality describes a trade-off between organization and resilience in ecosystems through a new metric created from the measures of AMI and Overhead. As such it describes the trade-off in a system between its redundancy (high number of pathways with more uniform flow) and its degree of constraint (articulated pathways with more asymmetric flow). This new indicator has first been defined as "fitness for change" (Ulanowicz, 2009, Ulanowicz et al., 2009), consequently as sustainability (Goerner et al., 2009), and then called system robustness (e.g., Mukherjee et al., 2015; Kharrazi et al., 2013). It is derived by multiplying the ratio of A/C (or AMI/H) by the Boltzmann measure of disorder ($-k\log(a)$, Ulanowicz, 2009):

$$Robustness = -\alpha\log(\alpha)$$

where $\alpha = \text{AMI}/H$.

Work on empirical ecosystem networks revealed that the robustness of ecosystem peaks at intermediate values of AMI/H, which has been termed the "The Window of Vitality." This window describes an optimum balance between redundancy and efficiency in a network. On either side of the curve, where AMI (efficiency) is high or low relative to H (redundancy), the robustness values are comparatively low. Intermediate constrained networks are therefore thought to be of optimal robustness and sustainability (Ulanowicz, 2009; Ulanowicz et al., 2009; Goerner et al., 2009; Fig. 3).

Summary

A practical objective of ecological network analysis in general is to trace material and energy flow–storage through the complex network of system interactions, and has been a fruitful way of holistically investigating ecosystems. In particular, a series of "network properties" such as indirect effects ratio, homogenization, mutualism and ascendency have been observed using this analysis, which consider the role of each entity embedded in a larger system and so tie processes between nodes with their emerging properties describing ecosystem behavior.

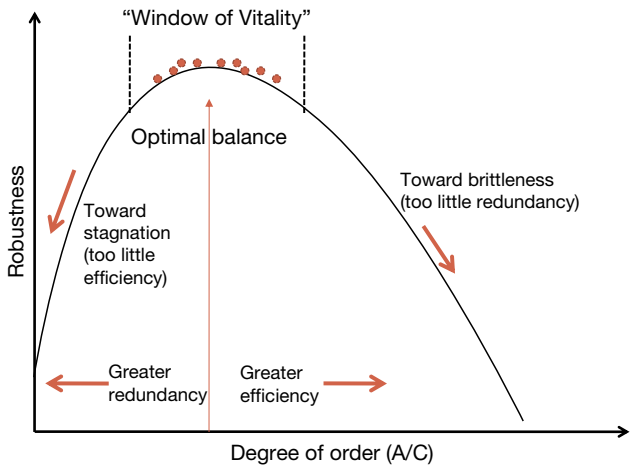


Fig. 3 Robustness as a function of efficiency and resilience. X-axis: $\alpha = \text{AMI}/H$. Y-axis: $-\alpha\log(\alpha)$. Modified from Goerner, S. J., Lietaer, B., and Ulanowicz, R. E. (2009). Quantifying economic sustainability: Implications for free-enterprise theory, policy and practice, *Ecological Economics* **69**, 76–81, with permission.

References

Fath BD (2004) Network analysis applied to large-scale cyber-ecosystems. *Ecological Modelling* 171: 329–337.

Fath BD, Patten BC, and Choi JS (2001) Complementarity of ecological goal functions. *Journal of Theoretical Biology* 208(4): 493–506.

Goerner SJ, Lietaer B, and Ulanowicz RE (2009) Quantifying economic sustainability: Implications for free-enterprise theory, policy and practice. *Ecological Economics* 69: 76–81.

Halnes G, Fath BD, and Liljenström H (2007) The modified niche model: Including a detritus compartment in simple structural food web models. *Ecological Modelling* 208: 9–16.

Jørgensen SE, Fath BD, Bastianoni S, Marques JC, Müller F, Nielsen SN, Patten BC, Tiezzi E, and Ulanowicz RE (2007) *A New Ecology: Systems perspective*. Amsterdam: Elsevier 275 pp.

Kazanci C and Adams MR (2017) Ecological utility theory: Solving a series convergence issue. *Ecological Modelling* 358: 19–24.

Kharrazi A, Rovenskaya E, Fath BD, Yarime M, and Kraines S (2013) Quantifying the sustainability of economic resource networks: An ecological information-based approach. *Ecological Economics* 90: 177–186.

Mukherjee J, Scharler UM, Fath BD, and Ray S (2015) Robustness indicators for aquatic ecological networks—A static model approach. *Ecological Modeling* 306: 160–173.

Patten BC (1978) Systems approach to the concept of environment. *Ohio Journal of Science* 78: 206–222.

Patten BC (2016) The cardinal hypotheses of Holoecology: Facets for a general systems theory of the organism–environment relationship. *Ecological Modelling* 319: 63–111.

Scharler UM, Brian D, and Fath BD (2009) Comparing network analysis methodologies for consumer–resource relations at species and ecosystems scales. *Ecological Modelling* 220: 3210–3218.

Ulanowicz RE (1986) *Growth and development: Ecosystems phenomenology.* New York: Springer Verlag.

Ulanowicz RE (2009) The dual nature of ecosystem dynamics. *Ecological Modelling* 220: 1886–1892.

Ulanowicz RE and Puccia CJ (1990) Mixed trophic impacts in ecosystems. *Coenosis* 5: 7–16.

Ulanowicz R, Goerner S, Lietaer B, and Gomez R (2009) Quantifying sustainability: Resilience, efficiency and the return of information theory. *Ecological Complexity* 6: 27–36.

Further Reading

Dame RF and Patten BC (1981) Analysis of energy flows in an intertidal oyster reef. *Marine Ecology Progress Series* 5: 115–124.

Fath BD and Patten BC (1998) Network synergism: Emergence of positive relations in ecological systems. *Ecological Modelling* 107: 127–143.

Fath BD and Patten BC (1999) Review of the foundations of network environ analysis. *Ecosystems* 2: 167–179.

Fath BD, Jørgensen SE, Patten BC, and Straškraba M (2004) Ecosystem growth and development. *Biosystems* 77: 213–228.

Higashi M and Patten BC (1989) Dominance of indirect causality in ecosystems. *American Naturalist* 133: 288–302.

Patten BC (1981) Environs: The superniches of ecosystems. *American Zoologist* 21: 845–852.

Patten BC (1982) Environs: Relativistic elementary particles or ecology. *American Naturalist* 119: 179–219.

Whipple SJ and Patten BC (1993) The problem of nontrophic processes in trophic ecology: Towards a network unfolding solution. *Journal of Theoretical Biology* 163: 393–411.

Thermodynamics in Ecology

Jae S Choi, Bedford Institute of Oceanography, Dartmouth, NS, Canada

Glossary

Disorder/order In a thermodynamic context, the logarithm of the number of microstate configurations possible (n) in a *phase space* of some model-based representation of a system (also called "ensembles" in quantum mechanics). The greater this number, the greater the "*disorder*" and the lower this number the lower the "*disorder*." The logarithm of n is proportional to the *entropy* of the system in physical systems and is related to it by Boltzmann's constant, $k = 1.380 \times 10^{-23}\,\mathrm{J\,K^{-1}}$, such that *entropy* $= k\,\log(n)$. It has, therefore, a very specific meaning and definition and should not be conflated with more common usage of the term, such as for example the "order" in phylogenetic trees.

Dissipative structures Energy or mass transforming processes that have some identifiable or repeatable structure, pattern or organization. Common examples are convection cells (e.g., Rayleigh–Bénard convection), chemical oscillations such as the Belousov–Zhabotinsky and the Briggs–Rauscher reactions, predator–prey interactions (e.g., Lotka–Volterra model), eddies and vortices in turbulent energy cascades in the hydrodynamic flow of water, smoke and air, and metabolic cycles in living organisms (Kolmogorov, 1941; Glansdorff *et al.*, 1974). Their importance in the thermodynamic context is due to the causal relationship between dissipation of energy *and* the consequent local decrease in *entropy*. These processes are interpreted as being caused by symmetry-breaking instabilities attributed to linear and near-linear *nonequilibrium thermodynamics*. That is, "the flow of energy through a system organises it."

Emergy Energy embodied by some "product" or "service." The concept is attributed to Odum (1996) and is an attempt at a valuation of energy-systems using a common currency (solar energy equivalents in Joules, or sej) by accounting for the total amount of energy required to make an organism or product, including all intermediate steps, starting from solar energy. It is, therefore, an attempt to define a biologically relevant form of *exergy* (often *Life Cycle Analysis* in engineering applications) in a complex system setting where there is a multiplicity of energy and matter flows and cycles. It resembles in some ways the efforts of ecological economics to put a monetary value to "ecosystem services." Closely associated with the concept of emergy is *transformity*, the relative efficiency of emergy production attributable to a species or process. A key hypothesis associated with emergy is the "*maximum empower principle*" which is a variation of Lotka's (1922) "*maximum power principle*" (Odum and Pinkerton, 1955; Washida, 1995) but stated in terms of *emergy*.

Environ Attributed to Patten (1978), an algebraic representation of nested networks across levels of hierarchic (network) and holarchic (inward-outward) organization. It is used to describe the incoming and outgoing environments of *holons* circumscribed by the boundaries of open systems in which holons are component parts. Environ theory thus defines two system-bounded environments associated with each holon within the system, and is mathematically an ecological extension of economic input–output analysis.

Entropy The energy *irreversibly* lost from a system to some reference state due to inefficiencies in matter and energy transformations.

Exergy Energy available for a given process ("work") after boundary conditions are specified. The concept is generally used in simpler engineering problems, accounting for energy efficiency and quality (coal vs. electricity) where boundary conditions, including the reference state, can be clearly demarcated. The focus is, therefore, upon net boundary (aggregate) inputs–outputs of e(n/x)ergy. In ecological systems, as many different forms of energy exist which are transformed at wide ranges of space, time and organizational scales, the boundary conditions are difficult (impossible?) to specify objectively. The concept of *emergy* (below) tries to address this difficulty.

Laws of thermodynamics A coherent understanding of the empirically observed *thermodynamic* states of matter was developed by Carnot (1824), Thomson (1849), and Clausius (1850). It was a monumental feat in that only a few variables of state were shown to be sufficient to describe the behavior of matter even if they are composed of an enormous number of atoms and molecules. Shortly thereafter, Boltzmann (1884) and Gibbs (1902) developed a coherent and complementary *theoretical (mechanical, statistical)* description of the component parts that was connected to the *thermodynamic* variables of state through the medium of probabilistic inference. Thereafter, a nonstatistical, axiomatic and geometric formulation of these laws were made by Carathéodory (1909) and again as recently as Lieb and Yngvason (1999). Landsberg's (1972) representation of these laws are amongst the most concise and clear:

- 0th law: empirical temperature exists.
- 1st law: internal energy exists. It is also known as the principle of energy conservation: it is not created nor destroyed, only transformed.
- 2nd law: entropy and absolute temperature exists. It revolves about the principle of entropy increase, where every energy transformation results in some form of unrecoverable loss.
- 3rd law: systems cannot attain a temperature of 0 Kelvin.
- 4th law: extensive (global) and intensive (local) variables exists for equilibrium states and at least some subset of near-equilibrium/nonequilibrium states. [This law is attributable to Landsberg (1972). While not controversial, it is not universally acknowledged as a "law." Nonetheless, it is important in the Onsager/Prigogine/

Landsberg/Katchalsky tradition and merits emphasis as it is often not appreciated or understood.]

Nonequilibrium thermodynamics Thermodynamic principles as applied to systems that are not fully equilibrated, that is, a system with gradients of energy and matter flow. Usually these are observed in systems *open* to matter and energy exchange, however, *closed* systems can also exhibit internal gradients (e.g., due to transient phenomena).

Phase space A high-dimensional space in which each axis corresponds to one of the coordinates required to fully specify the state of a dynamical system. The canonical example is of m ideal particles. As mass and momentum in each of three dimensions are known to completely specify the knowledge of the system, the phase space representation would have a dimensionality of $6m$ and a point in this space corresponds to a full description of the state of the system at a given moment. The logarithm of the number of such points possible is proportional to the hypervolume and the statistical *disorder* and physical *entropy* of the system. In ecology, if one simplistically assumes the number of individuals in each species represents the fundamental variables describing an ecosystem and so fully specifying it, then the system at a given moment would be "completely" described by a point in this m-dimensional hypervolume. The logarithm of the number of possible such points for a given ecosystem would be proportional to the statistical *disorder* of the ecosystem. This "model" is simple in that the "species" category aggregates information about individuals and therefore ignores qualities such as age, sex, genetics, behavior, social structure, spatial interactions, ontogeny, phylogeny, etc.

Introduction

Thermodynamic principles are fundamental pillars of modern science. They permeate through all physical, chemical, and biological processes. Ecosystems, spanning these physical, chemical and biological processes, are thus equally subject to these universal energy and entropy principles. Though traces of this awareness are present as early as in the writings of Heraclitus ($\sim$500 BCE) where he describes life as being change and flux, a formal awareness of the connection between modern notions of life and thermodynamics might be best attributed to Spencer (1864) who suggested that "evolution is an integration of matter and concomitant dissipation of motion; during which the matter passes from an indefinite, incoherent homogeneity to a definite, coherent heterogeneity; and during which the retained motion undergoes a parallel transformation." Thus, shortly after Carnot (1824), Thomson (1849) and Clausius (1850) had developed the groundbreaking laws of thermodynamics and just before Boltzmann (1884), Planck (1901) and Gibbs (1902) had developed the *theoretical* models to provide a statistical and mechanistic basis to these laws, there was an appreciation for the connection between energy and matter transformation and life in ecological or at least natural history circles.

An overview of some of the main thermodynamic concepts in ecology can be found in Fath *et al.* (2001, 2004), Yen *et al.* (2014) and Chapman *et al.* (2016), while Schneider and Sagan (2005) and Jørgensen (2012) provide accessible historical overviews. What is immediately evident from the literature is that many of the landmark ecological studies such as those by Lotka (1922), Elton (1927), Hutchinson (1959), Lindeman (1942), von Bertalanffy (1950), Whittaker (1953), Odum and Pinkerton (1955), Margalef (1963), MacArthur and Pianka (1966), Odum (1969), Patten (1978), Ulanowicz and Hannon (1987), Kay (1991), Schneider and Kay (1994), Dewar and Porté (2008), Kleidon (2010) and many others have been trying to make refinements to this connection between ecology and thermodynamics. In fact, in some ways, they are almost unavoidable as even the most basic attempts at an empirical accounting or modeling of matter and energy flows or metabolic considerations necessarily begins by invoking conservation of energy and matter, and the implicit estimation of irreversible losses of energy from the study system.

While the empirical/phenomenological/macroscopic notion of energy is measurable and familiar, the notion of *entropy*, the irreversible loss of energy, is less so. This is because *entropy* requires an indirect accounting to measure what was lost and an explicit identification of where the loss went (i.e., the reference or ground state). Unfortunately, it is challenging to unambiguously specify where a system started and ended, and what the key energy and matter transformation processes may be. This is because in most ecological systems there is matter and energy cycling and leaving and entering a system at all sorts of space and time scales and with implied spacetime lag related nonlinearities. This difficulty has lead some such as Meixner (1973) to suggest entropy as being an ill-defined quantity in any system open to material and energy fluxes, that is, any system that is not fully at equilibrium. See Landsberg (1972) for a contrary perspective in fluid dynamics.

Order Principles

Order to Disorder: Entropy

A model-based, statistical–mechanical representation complementary to *entropy* was developed by Boltzmann (1884). The logarithm of the number of possible configurations or "microstates" in the high-dimensional *phase space* representation of a system was demonstrated to be proportional to the *entropy* of the system; the proportionality factor is Boltzmann's constant (Planck, 1901).

Boltzmann (1884) and Gibbs (1902) suggested that this increase in the number of microstates in phase space can be interpreted as the amount of randomness or variability in potential microstates, or more simply, "*disorder*."

For example, imagine an isolated container filled with n molecules of some gas, identical in all respects including, velocity and momentum. As it would take much energy and effort to force the molecules to stay in such an untenable state, the *entropy* associated with this configuration can be considered low. As it has only one microstate in the $6n$-dimensional phase space (two variables of state $\times$ 3 dimensions $\times$ n molecules), the *microstate randomness or disorder* of the system can also be considered low. Releasing the constraints that forced it to stay in this state, the *entropy* will increase to some larger value, as the potential energy is converted into kinetic energy and *entropy*. Simultaneously, the movements and collisions of the gas molecules will result in a spread of the velocity and momentum distributions with each collision until eventually this distribution no longer changes. The number of possible *microstate* representations in phase space would have increased to a maximal level of randomness or *disorder* corresponding to the increase in *macrostate entropy*.

Similarly, an internal combustion engine takes the chemical potential energy of dense long-chained hydrocarbons and in the process of using it, converts it into dispersed gases, heat and intermediately fractured molecules (*entropy* increase). Given the physical–chemical model of long-chained hydrocarbons and associated bond energies, the fuel goes from a structured to a disaggregated state; that is, the number of possible microstates in *phase space* has increased and so is more variable or *disordered*. Lasers are another example which follows along similar lines: filtering and reflection of light to create coherent wavelengths and amplitudes (*ordered*) which upon application to burn through an object causes heat (*entropy* increase) and decoherence of wavelengths and amplitudes (*disorder*). As a consequence of this relationship between *entropy* and *disorder*, the 2nd law entropy principle is also known as an "*order*→*disorder*" principle.

Note that herein, we reserve the word *entropy* to represent the thermodynamic (phenomenological) quantity and distinguish it from the words *disorder* and *order* for the information-theoretic (statistical) analog of entropy based upon the microstates of some theoretical statistical mechanical model representation of the system. In the literature, the word *entropy* has been used to mean both, often causing misinterpretations. Further, we reiterate that in this context, the words *disorder* and *order* have the above unique meanings (in phase space) and must not be conflated with the more common, lay usage of the terms.

Order to Order: Biology and Ecology

Schrödinger (1944) continued in this vein and used this more familiar concept of *order* to focus attention upon biological systems. He noted, in particular, that living systems were operating primarily with a principle that seemed quite distinct from the "*order*→*disorder*" principle associated with the 2nd law. For example, he described DNA as an "aperiodic crystal" of extremely high information content (all the information required to create and maintain life) which manages to replicate itself with extreme stability; DNA mutation rates are much lower than might be expected simply due to thermal effects in chemical reaction systems. We can simplistically represent the DNA of an individual in a phase space with dimensionality equal to the number of base pairs. (This implies a model that assumes each base pair acts independently, which is of course incorrect.) With each replication from parent to child the amount of order remains roughly constant and the range of variability at each base pair constrained to be within a range not too dissimilar from the previous generation. Thus the range of possible microstates are narrow, stable and robust.

Heuristically, biological and ecological systems seem to be operating upon an "*order*→*order*" principle, at least in the short-term. This stability is due to chemical activation energies and chemical bond stability; self-correcting enzyme systems and proteins; and of course Darwinian selection. Indeed, stable life cycles, developmental processes, canalization, phenotypic plasticity, ordered metabolic pathways such as the Krebs cycle, predator–prey interactions, host–parasite interactions, habitat-niche considerations, animal migration patterns, species distributions in space, all represent examples of intricately patterned stable or constrained processes of "*order*→*order*." They are the mainstay of biology and ecology: feedback cycles and their interrelationships (Yodzis, 1981).

The term, *dissipative structures* (Kolmogorov, 1941; Glansdorff *et al.*, 1974), is often used in thermodynamic studies to describe these stable, "*order*→*order*" systems that maintain their internal structure/order, at the cost of elevated energy degradation to their environment: literally pumping ordered energy in and entropy out. Other examples of this have been analyzed including: Rayleigh–Bénard convection cells, the Zhabotinsky reactions and Turing waves. Living systems, in the act of growth and maintenance, tap into these external matter and energy flows to maintain local order.

Disorder to Order: Ordered Complexity

When considering the steps that might be required to lead to the origin of life, from simple chemical compounds to more complex matter/energy cycles or "hypercycles" and eventually the evolution of replication mechanisms such as RNA, DNA, proteins and membrane based compartmentalization in the backdrop of strong energy gradients, they clearly suggest the operation of yet another principle, "*disorder*→*order*" (Lotka, 1922; Schrödinger, 1944; von Bertalanffy, 1950; Eigen and Schuster, 1979; Wicken, 1980; Wiley and Brooks, 1982; Schneider and Kay, 1994; Schneider and Sagan, 2005; England, 2013). They suggest that living systems seem to increase in complexity over evolutionary time: the sheer number, types and hierarchical organization of structures and metabolic pathways is high relative to our primordial ancestors that were presumably much

simpler and had shorter snippets of (D/R)NA, and in terms of the metabolic and organizational structures that they encoded. With this increase in complexity, there is seemingly a corresponding increase in *order* in the biota, due to the additional effort required to maintain this complex and often intricate biological structure. In other words, the action of Darwinian selection in the short-term seems to be to stabilize ("*order*→*order*") and in the long-term to increase order ("*disorder*→*order*"); where the latter occurs presumably through the accumulation of strategies for survival in the face of a relentlessly variable abiotic and biotic environment (van Valen, 1976; Bell, 1982).

This incongruence (some have used the word "antithetic") with the 2nd law has even been suggested by many of the above authors as being a means of defining or distinguishing between life and nonlife. Indeed, Lovelock (1972) used this logic to advantage and suggested the "Gaia hypothesis," that the presence of atmospheric homeostasis (i.e., a nonequilibrium stable state) can be used as an indicator of the presence of life on other planets and is now used as an organizing principle of planetary system models (Kleidon, 2010). Similar kinds of thermodynamic inference have been made for ecological succession, the stage-like patterns of colonization and abundance in various landscapes and increase in order over ecological time scales (Odum and Pinkerton, 1955; Odum, 1969; Margalef, 1963; Gladyshev, 1978; Johnson, 1981; Kay, 1991) and embryonic and developmental patterns (Zotin and Zotina, 1967; Lurié and Wagensberg, 1979; Briedis and Seagrave, 1984; Aoki, 1989; Holdaway *et al.*, 2010).

Synthesis of Order Hierarchies

The presence of the above three different *order* principles has precipitated a wide, sometimes confusing discussion of their relevance in living systems. What has been learned through these discussions is that it is absolutely necessary to specify the frames of reference by which we mean: (1) the focal energy and matter transformation processes; (2) the constraints such as boundary conditions, ground state; (3) the spacetime scale; and (4) whether it is a *thermodynamic* (*entropy*) or *information-theoretic* analysis (microstates of a model representation or *disorder*). To illustrate the above, we use the simple metaphor of water flowing downhill (**Fig. 1**).

In the frame of reference of the universe, *entropy* is always increasing and this is always true relative to the presumably low entropy state of the initial Big Bang. The downward flow of water can be seen as an analog of the universal constraint of the 2nd law, that is, the principle of "*order*→*disorder*." The "ordered" potential energy of water at elevation, "ordered" by solar and wind energy, is converted into kinetic energy, which is dissipated as heat, noise, erosion, and an admixture of molecular configurations until it eventually reaches the lowest elevation possible, where it is in the highest possible *entropy* state (the equilibrium "reference" state in this artificial example; **Fig. 1**A).

Under certain circumstances, however, the flow of water might seem to persist for a short period, such as for example, in eddies near rapids and falls. In the frame of these smaller windows (scales) of time and space, these cycling structures can seem stable and self-perpetuating: that is, "*order*→*order*" (**Fig. 1**B). Though of course, it is only a small volume that becomes part of the eddy and the bulk of the water continues to flow downhill. And so the question is: what are these special "certain circumstances"? As biologists and ecologists we endeavor to detail all these special "certain circumstances" and preconditions to stable material and energy flows, be they at subcellular, cellular, organismal, population, community, ecosystem or earth-system levels of organization.

In ecology, there is an *implicit* understanding that these matter and energy flows are only borrowed by biota, taking in a small fraction of energy and matter and so slowing down its eventual passage by redirecting them into organisms and ecosystems by

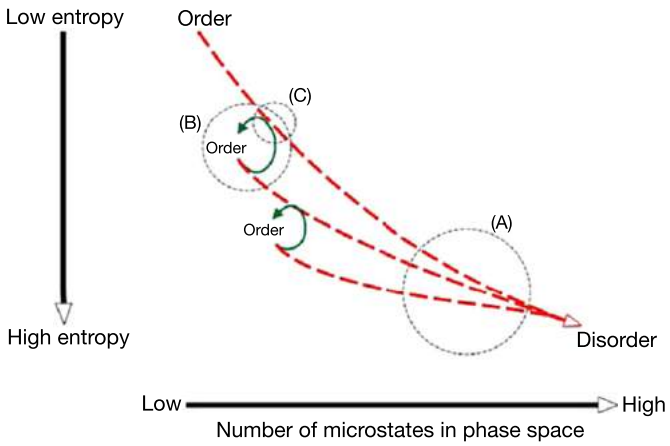


Fig. 1 Heuristic representation of the principles of (A) "*order*→*disorder*": the Second law principles such as in simple heat engines (*red-dashed lines*); (B) "*order*→*order*": homeostatic behavior in biochemical systems, stable ecological systems, genetic stability (*green self-loop*, or "dissipative structure"); and (C) "*disorder*→*order*": embryogenesis, succession, biodiversity increase over evolutionary time scales (transition from *red-dashed lines* to *green loops*). The *circles* show the focal processes for each order principle. Note the dissipation-cascade from top to bottom as each dissipative structure creates other dissipative structures, analogous to an energy-cascade of turbulent dynamics, or food-web dynamics.

growth and reproduction (biomass, chemical bonds), before their final release back into the entropic flow, "down the river" in the form heat, noise, breaking of molecular bonds, etc. Seldom are the latter entropic flows and constraints the explicit focus in biological and ecological studies. There is a matter-of-fact, a priori awareness that they are systems open to energy and matter flows. Instead, ecologists tend to focus preferentially upon the more "interesting" features of stable interactions and negative feedback dynamics.

Nonetheless, these more "interesting" eddy-like matter/energy interactions facilitate the overall boundary losses of energy and create *entropy* by "exporting" it, for example, as waste heat, via respiration, transpiration, mixing of degraded molecules, etc. As such, they act as (free) energy dissipating structures (or the eddies in the river metaphor) that borrow free energy and so helps to dissipate the energy more rapidly than if there had not been a dissipative structure. Lindeman (1942) was amongst the first to explicitly estimate such losses; his and subsequent estimates by Odum (1956), Odum (1969), Mann (1972) and others suggest ecological (trophic) efficiency to be generally quite low, ranging from 1% to 25% and about 10% on average. That is, entropic losses actually dominate most energy transformations in an ecological setting. Indeed from satellite-based radiation estimates, for many plants, >99% of incoming solar energy can be lost due to evapotranspiration and respiration (Schneider and Kay, 1994; Schneider and Sagan, 2005) though of course these processes are functional in that they are used to draw water and minerals from the soil. Only a very small fraction of inputs are aggregated and sequestered as biomass and with it, an associated local decrease in *entropy*. Yet even with such low efficiencies, biota continue to manifest and exert their global influence over geological time scales (Lovelock and Margulis, 1974).

Finally, if the frame shifts to ever smaller spacetime scales, we can focus upon the transition from downward flow to a nondownward flow of water, that is, to the initiation of an eddy (**Fig. 1**C). In this extreme close-up frame, a small fraction of the flow of water might actually seem to be reversing direction and moving uphill ("*disorder*→*order*"), though of course the bulk of the water does continue to flow downhill ("*order*→*disorder*"). The empirical observations in hydrodynamics suggest that stable laminar flow changes into nonlaminar flow once frictional (dissipative) forces increase in importance due to higher kinetic energy. These fluctuations selectively increase and if the higher kinetic energy conditions are sustained, alternate stability regimes can manifest via bifurcation and so in switching to an alternate stability regime, they seem to "self-organize." Empirically, such observations of switching behavior are also documented in ecology and evolution, often associated with feedback mechanisms altering due to major changes in external conditions and/or new biochemical, or species-interactions due to speciation or range expansions/contractions or trophic cascades causing new stable associations.

In the most general context of thermodynamics, Onsager (1931a, b) was one of the first to ask the question: How does this "reversal" begin? His answer was what are known as the Onsager reciprocal relations. The ramifications of his work are still being worked out in various quarters of science (Prigogine, 1955; Glansdorff *et al.*, 1974), including ecology. We will turn to this important development in section Extensive thermodynamics, as it serves as the bridge between the established principles of "*order*→*disorder*" and "*order*→*order*."

Thermodynamics and Ecology

To review, there are two complementary approaches to thermodynamics: (1) an empirical (phenomenological) analysis of *thermodynamic entropy* which is a physical description of a system's macrostate; and (2) an analysis of *information-theoretical order* that is contingent upon the *mechanical* model description of a system's possible microstates in phase space. The field of ecology, like most other scientific domains, looks toward thermodynamics as an ideal of how good science "should" be conducted and has tried to formulate both *empirical* principles with *theoretical* mechanistic/statistical processes. The yearning, of course, is to discover equally all encompassing *empirical* laws or at least some predictive rules of thumb which may serve as a scaffolding for further development of the *theoretical* underpinnings of component processes (Peters, 1991).

Information-Theory

The *information-theoretic* mode of analysis is extremely powerful as *order* is a concept that is readily understood in a mechanical context. However, this strength is dependent upon the completeness of the underlying model representation of the system. There is, therefore, a model-dependent subjectivity to the concept of *order* that is inherently sensitive to the structure and parameterizations of the model. In representing the well-understood molecular motion of physical systems, the underlying model representation has been demonstrated to be very close to reality. When minor divergences were found, further research demonstrated it was due to new physical processes that were previously not considered (e.g., quantum effects; Schrödinger, 1944, Jaynes, 1957).

Unfortunately, there is as yet, no mechanistic model that can be said to represent an ecosystem in any but the most simplistic of terms. The usual choices are some variations of a Lotka–Volterra formulation (e.g., Michaelian, 2005; Chakrabarti and Ghosh, 2009) or the even more aggregate models of first order input–output analysis (Patten, 1978). The utility of defining the microstate variables in the phase space of an incomplete model is not clear. As there exists subjectivity in what the categories are in a model and parameterization of these dynamics, the information content (i.e., *order*) also becomes a subjective concept.

Undeterred, information-theoretic measures of the topology of feeding networks were proposed by Ulanowicz and Hannon (1987) and Ulanowicz *et al.* (2006) as a first approximation to these microstate variables. Jørgensen *et al.* (1995) tried to define this *order* on the basis of the number of genes in various organisms with the assumption that this number would represent the

information content and, therefore, the *order* associated with an organism; but there was no mechanistic model representation of an organism's dynamics associated with gene number and confusingly called this information-theoretic measure, "exergy." Wicken (1980) and Wiley and Brooks (1982) focused upon similar approaches to describe information-theoretic components of phylogenetic complexity; again, a mechanistic model relating phylogenetic dynamics to a phase space representation was lacking.

Though direct application of the *information-theoretic* approach has been controversial, there have been a number of related developments that have impacted ecology. Jaynes (1957) was a strong proponent of using statistical mechanics and information-theory (Shannon, 1948; Wiener, 1949) to a number of different problems of inference that have been variously known as *maximum entropy production* (MEP), *maximum entropy* (MaxEnt) and *maximum relative entropy* (MaxREnt) depending upon the source literature (e.g., Yen *et al.*, 2014; Dewar, 2010). The approach has been used successfully in biodiversity–productivity modeling (Dewar and Porté, 2008); species distribution modeling (Elith *et al.*, 2011); climate modelling (Paltridge, 2001; Martyushev, 2010); as a methodological approach for choosing minimally biased priors for Bayesian inference; and as a foundation for "stochastic processes" and "random fields" approaches in problems of statistical inference (Besag, 1974; Pickens and Iliopoulos, 2005). Even the widely used AIC and related families of information-theoretic evaluation of models have a basis in information theory and MEP arguments. Dewar (2010) and Meysman and Bruers (2010) discuss the complexity of the connection between *information-theoretic* representation of *disorder* with the *extensive* thermodynamic concept of *entropy* (see below), using the stability properties of first order chemical reaction/chemostat-type models. While Nicolis and Nicolis (2010) warn of the naïve and problematic use of MEP.

Extensive Thermodynamics

The *extensive* thermodynamic approach is perhaps best exemplified by the engineering tradition which traces its origin directly to the founding work of Carnot (1824). (*Extensive* simply means an *aggregate*, such as aggregate entropy or aggregate free energy of a whole system.) The aggregate free energy available to do some form of work is the state variable of interest, so much so that it has even been given a name, "*exergy*." Engineering applications usually focus of upon systems whose bounds can be defined explicitly (e.g., some machine or production process) and revolve around a few well defined forms of energy and matter transformation, and an explicitly defined reference/ground state. These conditions permit repeated measurements of exergy and entropy, and transfer efficiencies can be calculated to permit cost-benefit or efficiency analysis of the whole system.

Following in this tradition, Kay (1991), Aoki (1993), Schneider and Kay (1994), Müller (1998), Schneider and Sagan (2005), Jørgensen (2012) and many others have focused upon an analysis of (*extensive*) *exergy* and in particular the role of *exergy* gradients. The core message in this tradition is that, "nature abhors an e(n/x)ergy gradient and that any and all mechanisms will be enabled to reduce their magnitude" (Schneider and Kay, 1994). In this *extensive* tradition, Aoki (1993, 1995) estimated whole system entropy dynamics from energy balance of lakes; while Schneider and Kay (1994) used satellite-based measurements of evapotranspiration as a proxy of entropy production and related this to vegetation cover. These *extensive* approaches, however, suffer a difficulty similar to those faced in defining *entropy* in a complex, open, nonequilibrium system with many mass–energy transformation processes (see section Introduction). There can be many exergy gradients involved (geothermal, tidal, solar radiation, geochemical, chemical, biological), operating at various spacetime scales, often in coupled and cyclic manners with ill-defined boundaries and reference states. That which is a resource for one organism can be not relevant for another or even a toxin. Complex relationships between entropy production and system stability have been shown that suggests a naive application of the *extensive* thermodynamic approach can be problematic (Meysman and Bruers, 2010).

To overcome or at least to acknowledge these challenges, Odum (1996) suggested the concept of "embodied energy" (*emergy*). Emergy attempts to trace and unroll all energy transformation steps starting from solar radiation to the component of interest and expresses it in terms of solar energy units. It is, therefore, a concept complementary to the exergy approach, with the exception of trying to refine the measurement of free energy in these more realistic and complex situations. Emergy is, however, *not* a thermodynamic variable of state and so one cannot connect it directly to thermodynamic theory although it may be amenable to analysis via the least action principle and the concept of path integrals (Vanriel and Johnson, 1995; Martyushev, 2010). Indeed, when Patten's (1978) algebraic approach of unraveling the energy transformation cycles as a first order linear expansion of input–output relations (*environ analysis*) was undertaken, these concepts were shown to be complementary, at least in the sense of dimensional analysis (Fath *et al.*, 2001, 2004).

Intensive Thermodynamics

In contrast to the above *extensive* (aggregate) thermodynamic properties, Onsager (1931a, b) and Prigogine (1955) focused upon *intensive* thermodynamic properties of systems of small spacetime elements. This is a subtle but significant distinction in that an assumption of linearizability of *local* forces and fluxes is reasonable for *intensive* variables but it is not necessarily a reasonable assumption for *extensive* variables. The reason for this is that an arbitrarily small spacetime element can be specified such that it is in approximate steady state with its immediate neighborhood (e.g., Oster *et al.*, 1971; Landsberg, 1972). Indeed, this is the basis for most spacetime discretization schemes; and even applicable in problems with hydrodynamic turbulence, where eddies form at many scales, with the energy cascading through these eddies to ever smaller scales until molecular dissipation occurs via diffusion (near Kolmogorov length scales; Kolmogorov, 1941). Landsberg (1972) called such quasi-steady state elements, "local semi-stable

equilibria with small fluctuations from reference states" and even elevated this distinction to the level of a 4th law of thermodynamics (see Glossary). Irrespective, this is an important distinction that seems to have been lost in the literature which tends to focus primarily upon the *extensive* (e.g., exergy) and *information-theoretic* (e.g., MEP) traditions.

Onsager focused upon the frame of reference of open, nonequilibrium thermodynamic systems that are said to be in the "linear range." Such systems are generalization of well-known phenomena such as Fick's law of diffusion, Fourier's law of thermal conduction and Ohm's law of electrical current. Beyond these wide-ranging connections, there were two substantive results that need to be emphasized. The first result being the presence of *reciprocal relationships*. That is, the gradient of one energy transformation process can affect the speed of another seemingly unconnected transformation process. A well-known, concrete physical example of this *reciprocity* of forces and fluxes is the *Seebeck effect* where electrons flow in the presence of a temperature gradient which serves as the basis of a thermocouple; and the inverse flow of heat with application of electricity, known as the *Peltier effect* which serves as the basis for a thermistor. For thermal–electromagnetic coupling, they are known as the *Nernst–Ettingshausen effects*. The same reciprocal coupling exists in chemical reactions and cell membrane dynamics (Oster *et al.*, 1971; Glansdorff *et al.*, 1974; Mikulecky, 1985).

In an ecological context, this is relevant in that feeding networks may be expected to demonstrate complex indirect effects between seemingly unconnected energy transformation processes such as respiration, photosynthesis, evapotranspiration, growth, or in relationships between organisms via competition, predation, consumption, mutualism, etc., in the presence of biochemical–physical gradients. And indeed, there are correlations between many of these metabolic processes (Kleiber, 1947; von Bertalanffy, 1950) and consumer dynamics. Another example can be seen in the complex benthic–pelagic coupling of lakes as a function of nutrient, oxygen and pH gradients (Regier and Kay, 1996) or the related concept of the marine biological carbon pump (i.e., production vs. carbon gradients; Eppley and Peterson, 1967) and even the relationship between evapotranspiration and sediment mineral profiles (gradients) in tropical versus temperate environments. A number of other potential couplings were also identified by Odum and Pinkerton (1955).

Onsager also demonstrated a second groundbreaking result: any variations (i.e., fluctuations) in these forces or fluxes will have a tendency to diminish and reach some steady state. This is a similar concept to the expectation of gradient reduction in the *extensive* thermodynamic schools, however, it is often misunderstood or ignored by the *extensive* thermodynamic schools. The significant difference being the explicit expectation of a local "steady state" or "homeostasis" in the *intensive* approach. At an organismal level, this expectation of homeostasis is not considered surprising by the literature as it is a necessary constraint for the continuity of life and so it not contested. However, at the population, ecological, evolutionary or even planetary levels of organization, arguments relating to homeostasis has usually been met with skepticism and ridicule, and pejoratively labeled as "vitalism," "Lamarckism," "Panglossianism" or "teleological" (e.g., Tansley, 1935; Gould and Lewontin, 1979; Dawkins, 1986; Mayr, 1991; Grimm and Wissel, 1997). The importance, in this context, is that this thermodynamic principle provides a general directionality of change that is clear without resorting to teleology, an expectation that is based upon the simple and universally accepted directionality of time.

This expectation for dynamical systems to evolve in a direction of *local intensive entropy* decrease was given a stronger basis by the formal demonstration that the second derivative in time of *local intensive entropy* can be considered a *local* Lyapunov function (Prigogine, 1955). This means that *local intensive entropy* acts as a *local* extremum principle; that is, there is a strong expectation for *local* homeostasis. Prigogine extended this analysis to chemical reactions and used the term, "dissipative structures," to denote the *locally* structured flows of matter and energy (reactions) that results in a low *local* entropy density by "exporting" the entropy to the external environment, and so respecting the overall expectations of the *global extrinsic* (aggregate) *entropy* to increase over time. The interpretation is that these hierarchical *cascades* of dissipative structures (**Fig. 1**C; Choi and Patten, 2001) represent *local* departures from equilibrium, that operate in a manner analogous to the cascade of energy in turbulent phenomena (Kolmogorov, 1941). (The Onsager relations represents an approximation of *Le Chatelier's principle* stated in terms of *entropy*. More correctly, the diagonals of Onsager's phenomenological coefficients represent the first order linear effects aligned with *Le Chatelier's principle and the off-diagonals the linearized indirect/cross effects*; von Bertalanffy, 1950.)

Attempts were made to extend this Lyapunov stability analysis into the nonlinear realm using variational principles by Nicolis and Prigogine (1989). But not surprisingly, there are no guarantees that universal stability criteria exist in nonlinear systems; they tend to have complex stability regimes or *attractors* (Byers and Hansell, 1996). Interestingly, Crooks (1999) and England (2013) suggest some statistical generalizations may still be possible in highly nonequilibrium conditions, as entropy production can be linked to the Markov transition probabilities in a statistical-fluctuation process representation of biochemical reactions (i.e., in an information-theoretic sense). But again, we would need a strong or at least believable probabilistic ecological model before this approach can be attempted.

There are, therefore, many deep connections between Onsager's relations to a variety of well-known physical and chemical processes. The question remains, how do they connect to biological and ecological processes. This has been the focus of von Bertalanffy (1950), Odum and Pinkerton (1955), Oster *et al.* (1971), Prigogine *et al.* (1972), Mikulecky (1985) and Choi *et al.* (1999). An ecologically relevant measure of this is heat production or the related processes that generates this heat: respiration (biochemical assays or allometrically estimated from size structure; Choi *et al.*, 1999). However, as Onsager's development is for *intensive* thermodynamic variables, this respiration needs to be normalized by the relevant biovolume or biomass. As such, it has been argued and demonstrated that *local intensive entropy* production is most readily approximated as the respiration/biomass ratio (also known as "Least specific dissipation"; Choi *et al.*, 1999). The word, "*specific*" dissipation was used to denote the *local intensive* nature of these thermodynamically related variables associated with Onsager's analysis. Indeed, the dimensional relationships

between these processes (biomass-storage and respiration-loss) and their underlying efficiency considerations (i.e., biological allometry; Kleiber, 1947, von Bertalanffy, 1950) provide the critical constraints that connect the *intensive* to the *extensive* thermodynamic variables (Choi *et al.*, 1999; Fath *et al.*, 2004). The study of these *intensive* thermodynamic processes in ecosystems shows much promise due to their strong connections with the ideas of stability and integrity in systems that are the product of numerous processes, spanning wider and wider ranges in spacetime and organizational scales such as environmental degradation, resource availability and rapid environmental change (Choi and Patten, 2001).

Finally, in a characteristically forward thinking manner, Odum and Pinkerton (1955) tried to generalize from the *intensive* to the *extensive* by examining thermodynamic efficiencies. Using linear nonequilibrium thermodynamics as a basis (i.e., Onsager, 1931a, b; Prigogine, 1955), they noted the quadratic form of power output as a function of efficiencies in Onsager's approximations. The presence of these power maxima were suggested to substantiate Lotka's (1922) heuristic arguments for the expectation of maximum power throughput in systems operating under Darwinian selection (and potentially the least action principle). (Odum's (1996) empower hypothesis represents a natural extension of these ideas in his attempt to deconstruct ecosystems in a path-integral type analysis.)

Summary

It is perhaps not surprising that there have been numerous misunderstandings about the relevance and utility of thermodynamics to ecological/biological systems, given the number of thermodynamic traditions and sheer range of phenomena to which thermodynamics is applicable: classical states of matter and molecular considerations; statistical (quantum) mechanics; chemical oscillations; electromagnetic phenomena; hydrodynamic phenomena including climate models; cell membranes; organismal growth and development; population, community, ecosystem and planetary systems analyses; engineering analyses of production systems, engines, agricultural production, etc. However, thermodynamics continues to have a critical role to play in the future evolution of the field of ecology. It provides a strong and systemic foundation. Irrespective of the thermodynamic tradition one follows, they converge upon very similar ideas. One must, however, explicitly define this tradition and the spacetime frame of observation and analysis. In the ecological context, we are now at a stage where Carnot began measuring efficiencies and Boltzmann tried to understand which microstates are relevant. There is much to do and a great richness to be expected as a deeper appreciation of thermodynamic constraints develops and guides us to construct more synthetic ecological concepts and models.

See also: Ecosystems: Steppes and Prairies; Swamps

References

Aoki, I., 1989. Entropy flow and entropy production in the human body in basal conditions. Journal of Theoretical Biology 141, 11–21.
Aoki, I., 1993. Inclusive Kullback index—A macroscopic measure in ecological systems. Ecological Modelling 66, 289–299.
Aoki, I., 1995. Entropy production in living systems: From organisms to ecosystems. Thermochimica Acta 250, 359–370.
Bell, G., 1982. The masterpiece of nature: The evolution and genetics of sexuality. Croom Helm Ltd.
von Bertalanffy, L., 1950. The theory of open systems in physics and biology. Science 111, 23–29.
Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society. Series B (Methodological) 36, 192–236.
Boltzmann, L., 1884. Ableitung des Stefanschen Gesetzes, betreffend die Abhängigkeit der Wärmestrahlung von der Temperatur aus der elektromagnetischen Lichttheorie. Wiedemann's Annalen 22, 291–294.
Briedis, D., Seagrave, R.C., 1984. Energy transformation and entropy production in living systems I. Applications to embryonic growth. Journal of Theoretical Biology 110, 173–193.
Byers, R.E., Hansell, R.I.C., 1996. Implications of semi-stable attractors for ecological modelling. Ecological Modelling 89, 59–65.
Carathéodory, C., 1909. Untersuchung über die Grundlagen der Thermodynamik. Mathematische Annalen 67, 355–386.
Carnot, S., 1824. Reflexions sur la puissance motrice du feu. Paris: Bachelier, (English translation in Fox, R. (1986) *Reflexions on the motive power of fire*. Manchester University Press.).
Chakrabarti, C.G., Ghosh, K., 2009. Non-equilibrium thermodynamics of ecosystems: Entropic analysis of stability and diversity. Ecological Modelling 220, 1950–1956.
Chapman, E.J., Childers, D.L., Vallino, J.J., 2016. How the second law of thermodynamics has informed ecosystem ecology through its history. Bioscience 66, 27–39.
Choi, J.S., Patten, B.C., 2001. Sustainable development: Lessons from the paradox of enrichment. Ecosystem Health 7, 163–178.
Choi, J.S., Mazumder, A., Hansell, R.I.C., 1999. Measuring perturbation in a complicated, thermodynamic world. Ecological Modelling 117, 143–158.
Clausius, R., 1850. Uber die bewegende Kraft der Wärme und die Gesetze, welche sïch daraus für die Wärmelehre selbst ableiten lassen. Annalen der Physik Und Chemie 79, 368–397. (English translation in Kestin, J. (1976) *Second law of thermodynamics*. Wiley, 346 p.).
Crooks, G.E., 1999. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. Physical Review E 60, 2721–2726.
Dawkins, R., 1986. The blind watchmaker: Why the evidence of evolution reveals a universe without design. New York: WW Norton & Company.
Dewar, R.C., 2010. Maximum entropy production and plant optimization theories. Philosophical Transactions of the Royal Society of London B: Biological Sciences 365, 1429–1435.
Dewar, R.C., Porté, A., 2008. Statistical mechanics unifies different ecological patterns. Journal of Theoretical Biology 251, 389–403.
Eigen, M., Schuster, P., 1979. The hypercycle: A principle of natural self-organization. New York: Springer-Verlag.
Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J., 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17, 43–57.
Elton, C., 1927. Animal ecology. London: Sidgwick and Jackson.
England, J.L., 2013. Statistical physics of self-replication. The Journal of Chemical Physics 139.121923(1–8).
Eppley, R., Peterson, B., 1967. Particulate organic matter flux and planktonic new production in the deep ocean. Nature 182, 677–682.

Fath, B.D., Patten, B.C., Choi, J.S., 2001. Complementarity of ecological goal functions. Journal of Theoretical Biology 208, 493–506.
Fath, B.D., Jørgensen, S.E., Patten, B.C., Straškraba, M., 2004. Ecosystem growth and development. Biosystems 77, 213–228.
Gibbs, J.W., 1902. Elementary principles in statistical mechanics, developed with especial reference to the rational foundation of thermodynamics. New York: Charles Scribner's Sons.
Gladyshev, G.P., 1978. On the thermodynamics of biological evolution. Journal of Theoretical Biology 75, 425–441.
Glansdorff, P., Nicolis, G., Prigogine, I., 1974. The thermodynamic stability theory of non-equilibrium states. Proceedings of the National Academy of Science, USA 71, 197–199.
Gould, S.J., Lewontin, R.C., 1979. The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. Proceedings of the Royal Society of London, Series B: Biological Sciences 205, 581–598.
Grimm, V., Wissel, C., 1997. Babel, or the ecological stability discussions: An inventory and analysis of terminology and a guide for avoiding confusion. Oecologia 109, 323–334.
Holdaway, R.J., Sparrow, A.D., Coomes, D.A., 2010. Trends in entropy production during ecosystem development in the Amazon Basin. Philosophical Transactions of the Royal Society of London B: Biological Sciences 365, 1437–1447.
Hutchinson, G.E., 1959. Homage to Santa Rosalia, or why are there so many kinds of animals. American Naturalist 93, 145–159.
Jaynes, E.T., 1957. Information theory and statistical mechanics. Physical Review 106, 620–630.
Johnson, L., 1981. The thermodynamic origin of ecosystems. Canadian Journal of Fisheries and Aquatic Science 38, 571–590.
Jørgensen, S.E., 2012. Integration of ecosystem theories: A pattern. vol. 3. Dordrecht: Springer Science & Business Media.
Jørgensen, S.E., Nielsen, S.N., Mejer, H., 1995. Emergy, environ, exergy and ecological modelling. Ecological Modelling 77, 99–109.
Kay, J.J., 1991. A non-equilibrium thermodynamic framework for discussing ecosystem integrity. Environmental Management 15, 483–495.
Kleiber, M., 1947. Body size and metabolic rate. Physiological Reviews 27, 511–541.
Kleidon, A., 2010. A basic introduction to the thermodynamics of the earth system far from equilibrium and maximum entropy production. Philosophical Transactions of the Royal Society of London B: Biological Sciences 365, 1303–1315.
Kolmogorov, A.N., 1941. Local structure of turbulence in an incompressible fluid for very large Reynolds numbers. Doklady Academy of Science USSR 31, 301–305.
Landsberg, P.T., 1972. The fourth law of thermodynamics. Nature 238, 229–231.
Lieb, E.H., Yngvason, J., 1999. The physics and mathematics of the second law of thermodynamics. Physics Reports 310, 1–96.
Lindeman, R.L., 1942. The trophic-dynamic aspect of ecology. Ecology 23, 399–418.
Lotka, A.J., 1922. Contribution to the energetics of evolution. Proceedings of the National Academy of Science, USA 8, 147–151.
Lovelock, J.E., 1972. Gaia as seen through the atmosphere. Atmospheric Environment 6, 579–580.
Lovelock, J.E., Margulis, L., 1974. Atmospheric homeostasis by and for the biosphere: The Gaia hypothesis. Tellus, Series A 26, 2–10.
Lurié, D., Wagensberg, J., 1979. Non-equilibrium thermodynamics and biological growth and development. Journal of Theoretical Biology 78, 241–250.
MacArthur, R.H., Pianka, E.R., 1966. On the optimal use of a patchy environment. American Naturalist 100, 603–609.
Mann, K.H., 1972. Ecological energetics of the seaweed zone in a marine bay of the Atlantic Coast of Canada. I. Zonation and biomass of seaweeds. Marine Biology 12, 1–10.
Margalef, R., 1963. On certain unifying principles in ecology. The American Naturalist 898, 357–374.
Martyushev, L.M., 2010. The maximum entropy production principle: Two basic questions. Philosophical Transactions of the Royal Society of London B: Biological Sciences 365, 1333–1334.
Mayr, E., 1991. The ideological resistance to Darwin's theory of natural selection. Proceedings of the American Philosophical Society 135, 123–139.
Meixner, J., 1973. The entropy problem in thermodynamics of processes. Rheologica Acta 12, 465–467.
Meysman, F.J., Bruers, S., 2010. Ecosystem functioning and maximum entropy production: A quantitative test of hypotheses. Philosophical Transactions of the Royal Society of London B: Biological Sciences 365, 1405–1416.
Michaelian, K., 2005. Thermodynamic stability of ecosystems. Journal of Theoretical Biology 237, 323–335.
Mikulecky, D.C., 1985. Network thermodynamics in biology and ecology: An introduction. In: Ulanowicz, R.E., Platt, T. (Eds.), Ecosystem theory for biological oceanography213. Canadian Bulletin of Fisheries and Aquatic Sciences, pp. 163–175.
Müller, F., 1998. Gradients in ecological systems. Ecological Modelling 108, 3–21.
Nicolis, C., Nicolis, G., 2010. Stability, complexity and the maximum dissipation conjecture. Quarterly Journal of the Royal Meteorological Society 136, 1161–1169. doi:10.1002/qj.642.
Nicolis, G., Prigogine, I., 1989. Exploring complexity: An introduction. WH Freeman.
Odum, H.T., 1956. Primary production in flowing waters. Limnology and Oceanography 1, 102–117.
Odum, E.P., 1969. The strategy of ecosystem development. Science 164, 262–270.
Odum, H.T., 1996. Environmental accounting: Emergy and decision making. New York: Wiley, p. p. 370.
Odum, H.T., Pinkerton, R.C., 1955. Time's speed regulator: The optimum efficiency for maximum power output in physical and biological systems. American Scientist 43, 321–343.
Onsager, L., 1931a. Reciprocal relations in irreversible processes. I. Physical Review 37, 405–426.
Onsager, L., 1931b. Reciprocal relations in irreversible processes. II. Physical Review 38, 2265–2279.
Oster, G., Perelson, A., Katchalsky, A., 1971. Network thermodynamics. Nature 234, 393–399.
Paltridge, G.W., 2001. A physical basis for a maximum of thermodynamic dissipation of the climate system. Quarterly Journal of the Royal Meteorological Society 127, 305–313.
Patten, B.C., 1978. Systems approach to the concept of environment. The Ohio Journal of Science 78, 206–222.
Peters, R.H., 1991. A critique for ecology. New York: Cambridge University Press.
Pickens, J., Iliopoulos, C.S., 2005. In: Markov random fields and maximum entropy modeling for music information retrieval.ISMIR 2005 6th International Conference on Music Information Retrieval Online Proceedings, pp. pp. 207–214.
Planck, M., 1901. Ueber das Gesetz der Energieverteilung im Normalspectrum. Annalen der Physik 4, 553–563. English translation. https://web.archive.org/web/20081217042934/http://dbhs.wvusd.k12.ca.us/webdocs/Chem-History/Planck-1901/Planck-1901.html
Prigogine, I., 1955. Thermodynamics of irreversible processes. New York: Wiley, p. p. 147.
Prigogine, I., Nicolis, G., Babloyantz, A., 1972. Thermodynamics of evolution. Physics Today 25, 23.
Regier, H.A., Kay, J.J., 1996. An heuristic model of transformations of the aquatic ecosystems of the Great Lakes-St. Lawrence River Basin. Journal of Aquatic Ecosystem Stress and Recovery (Formerly Journal of Aquatic Ecosystem Health) 5, 3–21.
Schneider, E.D., Kay, J.J., 1994. Life as a manifestation of the second law of thermodynamics. Mathematical and Computer Modelling 19, 25–48.
Schneider, E.D., Sagan, D., 2005. Into the cool: Energy flow, thermodynamics, and life. Chicago, IL: University of Chicago Press.
Schrödinger, E., 1944. What is life? The physical aspect of the living cell. Cambridge: Cambridge University Press.
Shannon, C.E., 1948. A mathematical theory of communication. Bell System Technical Journal 27, 379–423.
Spencer, H., 1864. The principles of biology. vol. 1. London: Williams and Norgate.
Tansley, A.G., 1935. The use and abuse of vegetational terms and concepts. Ecology 16, 284–307.

Thomson, W., 1849. An account of Carnot's theory of the motive power of heat; with numerical results deduced from Regnault's experiments on steam. Transactions of the Royal Society of Edinburgh 16, 541–574.
Ulanowicz, R.E., Hannon, B.M., 1987. Life and the production of entropy. Proceedings of the Royal Society of London Series B 232, 181–192.
Ulanowicz, R.E., Jørgensen, S.E., Fath, B.D., 2006. Exergy, information and aggradation: An ecosystems reconciliation. Ecological Modelling 198, 520–524.
van Valen, L., 1976. A new evolutionary law. Evolutionary Theory 1, 1–30.
Vanriel, P., Johnson, L., 1995. Action principles as determinants of ecosystem structure: The autonomous lake as a reference system. Ecology 76, 1741–1757.
Washida, T., 1995. Ecosystem configurations consequent on the maximum respiration hypothesis. Ecological Modelling 78, 173–193.
Whittaker, R.H., 1953. A consideration of climax theory: The climax as a population and pattern. Ecological Monographs 23, 41–78.
Wicken, J.S., 1980. A thermodynamic theory of evolution. Journal of Theoretical Biology 87, 9–23.
Wiener, N., 1949. Extrapolation, interpolation, and smoothing of stationary time series. New York: Wiley.
Wiley, E.O., Brooks, D.R., 1982. Victims of history—A nonequilibrium approach to evolution. Systematic Biology 31, 1–24.
Yen, J.D.L., Paganin, D.M., Thomson, J.R., MacNally, R., 2014. Thermodynamic extremization principles and their relevance to ecology. Austral Ecology 39, 619–632.
Yodzis, P., 1981. The stability of real ecosystems. Nature 289, 674–676.
Zotin, A.I., Zotina, R.S., 1967. Thermodynamics aspects of developmental biology. Journal of Theoretical Biology 17, 57–75.

Further Reading

Matsuno, K., 1978. Evolution of dissipative systems: A theoretical basis for Margalef's principle of ecosystems. Journal of Theoretical Biology 70, 23–31.
Nicolis, G., Prigogine, I., 1977. Self-organization in non-equilibrium systems. Wiley.
O'Neill, R.V., DeAngelis, D.L., Waide, J.B., Allen, T.F.H., 1986. A hierarchical concept of ecosystems. In: Monographs in population biology. Princeton, NJ: Princeton University Press. 253 pp.

Agriculture Models☆

James C Ascough II†, Lajpat R Ahuja, Gregory S McMaster, Liwang Ma, and Allan A Andales, USDA-ARS, Fort Collins, CO, United States

Introduction

The inherent complexity of agroecosystems (Fig. 1) requires multidisciplinary analyses from a wide range of scientific disciplines to better understand agronomic production issues and solve complex, real-world problems in agriculture. For example, increased use of water, fertilizers, and pesticides necessary for expansion of global agricultural production has resulted in damage to agroecosystems including excessive runoff and leaching of agricultural chemicals, decline in soil organic matter, and increases in soil salinity and wind/water erosion. Other challenges threaten the viability and sustainability of agroecosystems, including issues related to global climate change and market-based global commercialization and competition. The solution or mitigation of these multifaceted and multidimensional problems requires continual improvement and changes in agricultural management, and selection of agronomic production systems using a whole-system approach. Whole-system approaches were developed specifically to support interdisciplinary studies (with the goal of solving significant agronomic, ecological, and environmental problems that require systemwide integration and quantification of knowledge), and their development and use have increased strongly in the past decade.

An important component for analyzing and assessing whole-system interactions in agronomic systems is the knowledge derived from agricultural model simulations. Simulation models of agroecosystems have evolved into highly useful tools for evaluating and quantifying the effects of management practices, crops, soils, water, and climate on sustainability of both agricultural production and the surrounding environment. Furthermore, agricultural models can serve as strategic and tactical guides for planning and assessment, and help transfer new technologies to various location-specific environments (e.g., climate, soils, production systems) within regions or countries. In this article, the current state of agricultural models and their applications for the above purposes are reviewed and current technologies in agricultural model use (related to modular model development, state of the art interfaces, scaling issues, integrated assessment (IA), and field research integration) are presented. In addition, research needs for agricultural system models are discussed, and conclusions offered as to what the future holds for agricultural system modeling.

Types of Agricultural Models

Several approaches are available for modeling dynamic agroecosystems. They vary from those using process-based or mechanistic-type models to those using statistically based or regression-type models. Mechanistic models may be defined by differential equations (ordinary or partial) or by a combination of differential equations and algebraic equations describing process interactions between the components and states of an agroecosystem (based on underpinning physical, chemical, and/or biological relationships). Regression models on the other hand are based on statistical analysis, that is, identifying a regression-type relationship correlating the inputs (i.e., soil, crop, management, environmental) to the outputs (quantifiable states of the system). Both statistical and mechanistic models use linear and nonlinear mathematical equations to describe the quantitative responses of various processes, and both types of models have advantages and disadvantages. The main advantage of the mechanistic-type models is that they relate to the description (physical, chemical, and/or biological) of the agricultural system, typically have wider generality of application, and can provide more insight into the response of different underlying processes to system management. Disadvantages of the mechanistic-type models are that they are often more complex than their counterpart empirical ones, and are more difficult to parametrize and evaluate. Empirical models provide a concise description about the observations on which they are based and do not need to rely on lower-level system attributes; however, they do not imply causality or even knowledge about underlying processes (although they may provide some insight into these). In addition, extrapolation beyond the data range from which the empirical model was derived involves caution. In this article, we focus on the larger and more complex mechanistic models that typically can simulate a wide range of soil–plant–atmosphere processes and thus have applicability to a diverse range of agroecosystems.

☆*Change History:* March 2018. Todd M. Swannack updated References. No other changes were made.

This is an update of J.C. Ascough II, L.R. Ahuja, G.S. McMaster, L. Ma and A.A. Andales, Agriculture Models, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 85–95.

†Deceased.

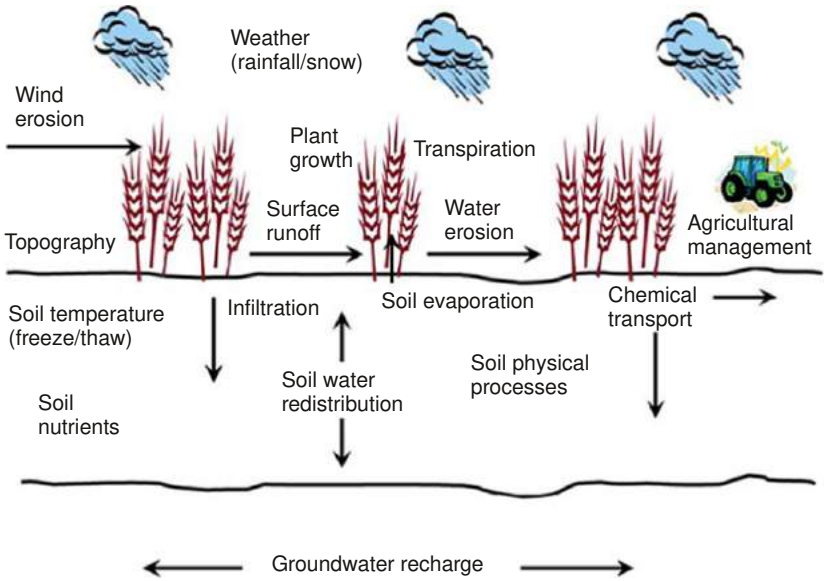


Fig. 1 Major processes in agricultural systems.

History of Agricultural Models

Agricultural modeling started in the early 20th century with the development of regression models exhibiting little attention to causative linkages and mechanistic understanding of the processes involved. This was largely due to limited knowledge concerning soil–plant–atmosphere processes and their response to the complex interactions between various factors in agroecosystems, and also to limited computing technologies available to integrate the processes. Since the early regression model phase, agricultural process modeling and system integration have undergone many years of development and evolution. Before the 1970s, considerable modeling work was undertaken for individual processes of agroecosystems and a foundation for system simulation was built. For example, in soil water movement, models and theories were developed in the areas of infiltration and water redistribution, soil hydraulic properties, tile drainage, and solute transport. In plant–soil interactions, models and theories were developed for evapotranspiration, photosynthesis, root growth, plant growth, and soil nutrients. In Europe, the development of mechanistic process-oriented models in plant biology began with the pioneering work of de Wit on quantitative relationships in the photosynthetic process, through development of a formula calculating photosynthesis of a closed crop surface. Further research modeling light interception and photosynthesis in plant canopies followed. These results, together with knowledge of the effect of soil moisture on uptake and transpiration rates, permitted a comprehensive approach to the study of the soil–plant–water–atmosphere system and its adequacy for plant growth. Lists of authors on early manuscripts suggest a great deal of interaction between crop simulation groups in Europe and the US.

There has been a rich history in the development of agricultural models since that time. The 1980s saw the beginning of the development of whole-system agricultural models. Much of this work began in The Netherlands, at Wageningen. The generic crop model SUCROS for potential production simulation was developed, which formed the basis for recent Wageningen crop models such as WOFOST, MACROS, and ORYZA. For water- and nitrogen-limited production situations, model components were added to SUCROS resulting in models such as ARID CROP, SAHEL, and PAPRAN. Several other important agricultural models developed in the 1980s were CREAMS (soil erosion by water), GOSSYM (cotton crop growth), EPIC (water quality with integrated economics), GLYCIM (soybean crop growth), PRZM (pesticide transport), NTRM (nutrient cycling), and GLEAMS (water quality). System models of even greater mechanistic complexity and level of detail continued to be developed in the 1990s and 2000s including WEPP (soil erosion by water), WEPS (soil erosion by wind), ALMANAC (multispecies crop growth), CropSyst (crop growth), APSIM (agricultural production), OPUS (water quality), SHOOTGRO (wheat growth), DAISY (nutrient cycling), RZWQM (water quality), SWAT (basin-scale water quality), and CSM (crop growth). Some agricultural models have also been linked to a decision support system (DSS) framework—these include GPFARM and DSSAT, which envelopes the CSM (formerly the CERES and CROPGRO model families). Agricultural system research and modeling are now being promoted by several international organizations, such as International Consortium for Agricultural Systems Applications (ICASA) and other professional societies.

In summary, agricultural modeling has evolved into detailed dynamic simulations of most physical, chemical, biological, biochemical, and biophysical processes in agroecosystems and their response to climate and crop management inputs. The advent of high-speed computers and other advances in computing technology are partly responsible for this rapid evolution. The enormous time and effort scientists have invested in simulating agroecosystems are reflected in the seemingly endless number of models cited in literature review articles. Cultivar selection, water and nutrient input optimization, planting date selection, and

water quality management are just a few of the areas of agricultural model application in agroecosystem management across multiple scales.

Agricultural Model Applications

In this section, we provide a brief sampling of applications of agricultural models in research and strategic management. This range of applications is by no means exhaustive as agricultural models are used for innumerable functions and purposes.

Managing Crop Production and Water Quality

A common application of cropping system models is to simulate various management options in different agroecosystems to predict effects on crop production. Crop growth models are also used in yield-gap analyses to quantify potential yield compared with actual yield; that is, the models are used to estimate potential yields at multiple locations to determine how genotype $\times$ environment $\times$ management interactions can be optimized. Simulating differences in yield among varieties is typically by using the appropriate genetic/phenotypic coefficients to characterize each crop variety. The cropping system models (CSM) within the DSSAT framework and the APSIM model have been extensively used for this purpose; for example, the DSSAT framework has been used to derive optimum combinations of management and planting dates for various crops. Another common application of crop models is simulating interactions between crop yield and levels of agricultural inputs such as irrigation water or nitrogen (N) fertilizer. Because agricultural models keep track of the water balance and N amounts in the system as well as the estimated uptake by crops, the models can help estimate the proper amounts and timing of irrigation or N fertilization.

Other major applications in crop production modeling are assessing the sustainability of existing cropping systems and potential for yield improvement through introduction of alternative crops in a rotation, especially in a dryland agriculture context. For example, models have been used in conjunction with experimental field data to show that cropping intensification more effectively utilized available soil moisture and increased overall system productivity compared to the prevalent rotation. In addition, an important requirement for the sustainability of current agricultural systems is the mitigation of adverse environmental effects. Intensive crop production has been recognized as a significant nonpoint source of water contaminants. A major concern is the movement of nitrate (NO_3-N), phosphorus, and agricultural chemicals (e.g., pesticides and herbicides) from agricultural fields into surface and groundwater bodies. Agricultural system models that have the capability to simulate transformations and movement of agricultural chemicals (e.g., GLEAMS, SWAT, and RZWQM) have been extensively used to assess the interactions between water quality and crop production management.

Evaluating Soil–Plant–Atmosphere Parameters and Processes

Agricultural management practices (e.g., tillage and reconsolidation; no-tillage and surface residues; plants and crop rotations; irrigation, manure, and fertilization practices; and grazing management) are major sources of temporal variability of soil properties and processes. Changes in soil properties and processes, in turn, impact soil water, mass transport, plant growth dynamics, and the environment. Weather-related factors such as freezing–thawing and wetting–drying may modify the management effects. Numerous field and agricultural modeling studies have shown evidence of the significant management effects on soil–water–nutrient–plant properties and processes. Important areas where agricultural models have been used to help investigate management effects on soil properties and processes include: (1) predicting effects of tillage and natural reconsolidation; (2) predicting surface roughness and detention storage effects; (3) quantifying the effects of wheel-track compaction; and (4) quantifying long-term no-tillage and crop residue effects, including the impacts of macropores and residue cover on infiltration.

Another extremely important area where agricultural models have been used to study changes in properties at the soil–plant–atmosphere continuum is the influence of roots on soil structure and macroporosity. The magnitude and distribution of root growth in a soil profile vary widely between and within plant species. The depth and temporal pattern of root growth also depends on soil properties such as soil bulk density, temperature, water content regime, salinity, and nutrient deficiencies, which change with depth in heterogeneous layered soils. The distribution of root growth with depth and time determines the distribution of water and nutrient uptake from the soil. This, in turn, influences water, chemical, and heat movement in the soil. Additional areas where agricultural models are being used at the soil–plant–atmosphere continuum include quantification of water and nutrient uptake by roots, and evaluation of transpiration and carbon dioxide fluxes at the canopy–atmosphere interface under water stress.

Assessing Climate Change Effects in Agroecosystems

In evaluating future agroecosystem sustainability, a main concern is the effect of climate change (e.g., increased atmospheric CO_2, elevated air temperature, increased/decreased water availability) on agronomic production. Many of the crop production models discussed previously have been used to investigate potential impacts of climate change on yield. Agricultural modeling is a practical approach to studying this global phenomenon as it is difficult to wholly quantify the interactions between climate change effects and crop production based on limited plot-scale experiments or controlled-environment studies. Another major

international issue with respect to global climate change is to improve our understanding of how agroecosystem soils and management contribute to climate change through emission of greenhouse gases such as CO_2, CH_4, N_2O, NO, and NH_3. Land-use changes are believed to account for about 8% and other agriculture sources about 15% of the anthropogenic greenhouse emissions. Several agricultural models (e.g., Century and DNDC) are being utilized to devise and evaluate management practices that will minimize greenhouse gas emissions and increase C sequestration, such as no-tillage and residue management, adding legumes into rotations, and optimal timing of manure applications. Changes in climate modify the soil environment, especially soil water and temperature and also a number of concurrent processes dependent on these including evapotranspiration, runoff, and erosion. Current global climate change models have predicted an increase in frequency of extreme events such as droughts and heavy rainfalls, but it is not known how these changes will affect agricultural production in different ecosystems around the world. Furthermore, it is unclear how extreme climate events will change worldwide soil resources—for example, changes in soil properties from accelerated soil erosion by wind and water. Agricultural models are increasingly needed to evaluate the magnitude of these influences in different agroecosystems and locations around the world, and to devise strategies for mitigating potential adverse effects. For example, agricultural models are just now being used to develop special management practices during drought conditions. Drought-mitigation strategies may include shift in production among regions and changes in crops, cultivars, and management practices, such as crop rotations and water conservation measures.

Tools for Technology Transfer and Decision Support

Agricultural models have been commonly used to extend the results of experimental research to other soil types, climates, and management conditions outside the experimental design. For example, they have been used for extrapolating limited duration experimental results to variability in climatic conditions across longer periods of time (e.g., 25–50 years), and to extreme climatic conditions (e.g., droughts or flooding) not encountered during the study period. Agricultural models have proven to be useful tools for in-depth analysis of problems in management, environmental quality, global climate change, and other ecological issues, and can thus be a basis for policy or regulatory use. Agricultural models also function as decision aids in choosing best management practices for long-term sustainable production, as well as helping to guide site-specific management on agricultural landscapes and within-season dynamic management in response to spatially variable soil moisture and weather conditions. Many natural resource DSSs have an agricultural model at their core, but are also supported by soil, climate, and management databases, environmental and economic analysis packages, user-friendly interfaces to check default data or enter site-specific data, and graphical visualization of simulation results. An example is the design of the USDA-ARS GPFARM DSS (Fig. 2). GPFARM is a whole-farm DSS for strategic planning and evaluation of cropping systems, range–livestock systems, and integrated crop–livestock farming options for production, economics, and environmental impacts.

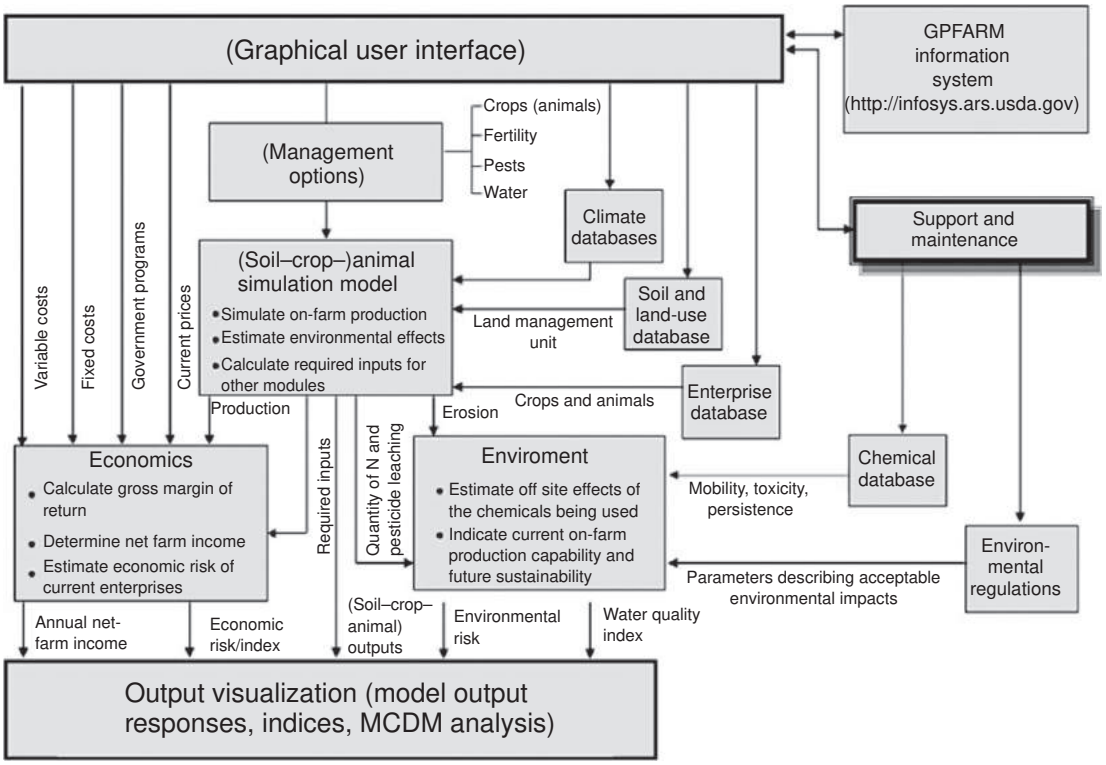


Fig. 2 Schematic of major GPFARM DSS components.

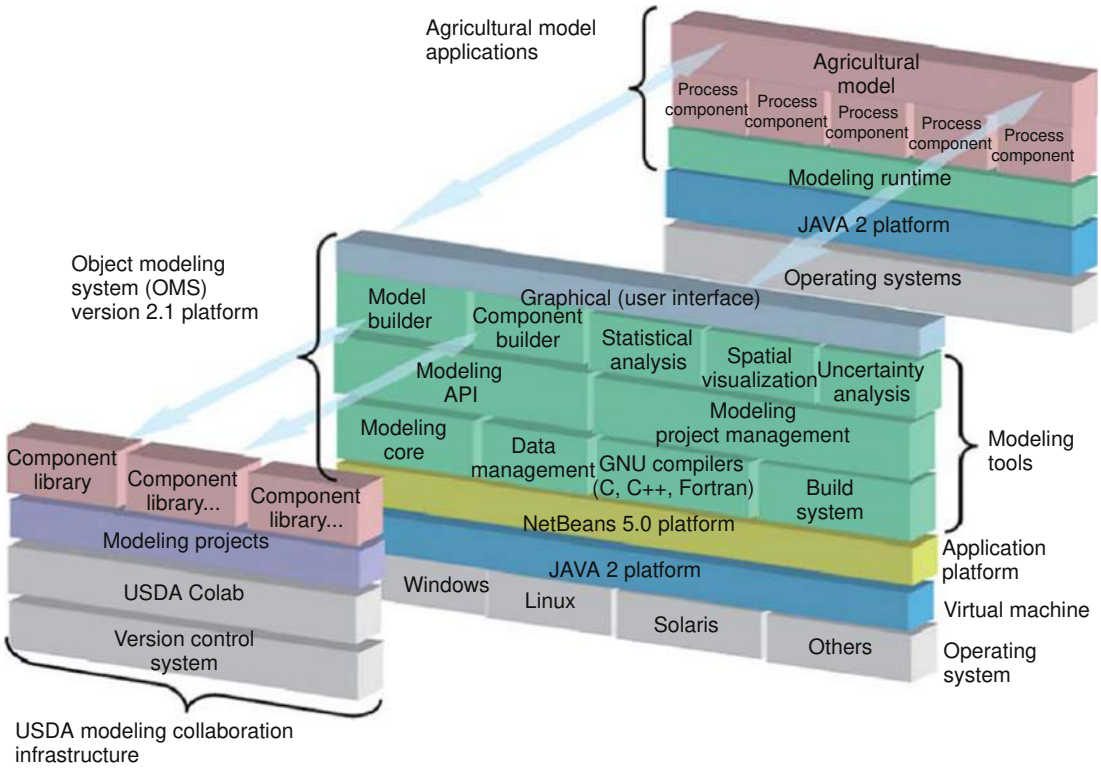


Fig. 3 Object Modeling System (OMS) 2.1 platform with linkages to a collaborative modeling infrastructure and agricultural modeling applications.

Complex and highly detailed process-level agricultural models are generally too difficult for consultants, producers, or policymakers to directly use. An alternative approach is to create an integrated research-information database as a DSS core in place of a simulation model. An agricultural system model, evaluated against available experimental data, is used to generate production and environmental impacts of different management practices for soil types, weather conditions, and cropping systems outside the experimental limits. The model-generated information is then combined with experimental data and long-term management experience of farmers and field professionals to create a database. These databases are often then combined with a socioeconomic analysis package or other tools (e.g., multiobjective decision analysis) in order to conduct a tradeoff analysis between conflicting objectives such as economic return and environmental quality. Overall, this type of approach is very flexible in generating site-specific management recommendations and avoids the problem of having to interpret complicated model output.

Current Technologies for Agricultural Model Development and Use

Modular Components and Frameworks

Agricultural model developers have recognized the value of a modular style of model development from the beginning; for example, the early mechanistic models from the "school of de Wit" were designed with a high degree of modularity in mind. Modular approaches to agricultural system model development are advantageous in that they: (1) facilitate substitution and reusability of different model components, (2) facilitate documentation and sharing of code, (3) allow linkage of components written in different programming languages, (4) allow greater flexibility in model updates and maintenance, (5) enhance collaboration opportunities between different model development groups, and perhaps most importantly, and (6) provide a cost reduction in model development and maintenance. In general, however, modular programming techniques have been slow to filter down to the agricultural modeling community where different approaches to model development in the various disciplines (e.g., crop growth, nutrient cycling, chemical transport, erosion, greenhouse gas emissions) have resulted in the proliferation of code and simulation models as described above. Most agricultural models are still monolithic, difficult to update as new knowledge becomes available, and are not extensible if new problems and modeling needs arise. Notable exceptions are the DSSAT 4.x CSM and the APSIM modeling platform. The basis for the DSSAT CSM design is a modular structure in which components are structured to allow for easy replacement or addition of modules and there are common components for soil water, soil nitrogen, weather, and competition for light and water among the soil, plants, and atmosphere. The APSIM modeling framework has the ability to integrate models derived in disparate research efforts via the implementation of a "plug-in-pull-out" approach; for example, the user can configure a model by choosing a set of submodels from a suite of crop, soil, and utility

modules. Thus, any logical combination of modules can be specified by the user "plugging-in" required modules and "pulling out" any modules no longer required.

Recently, as computer scientists trained in modern software engineering methods have entered the natural resource, environmental science, and ecology disciplines, the APSIM "plug-and-play" approach has been taken one step further with the development of formalized modular modeling frameworks. These frameworks bring together suites or libraries of modules, have architectures designed to fit well with the basic and natural structures of environmental problem situations, and can maintain reusability and compatibility of both science and auxiliary (e.g., parameter estimation, output visualization) components. A range of modeling frameworks with different capabilities currently exist, including the Interactive Component Modeling System (ICMS), Tarsier, the Spatial Modeling Environment (SME), The Invisible Modelling Environment (TIME), the European Open Modeling Interface and Environment (OpenMI) initiative (taking place under the Fifth European Framework Program project HarmonIT), and the Object Modeling System (OMS). The OMS framework development effort is interesting in that:

- a component or module can function as a whole model, a physical process, or a simple equation; that is, there is no assumption about component complexity level, whereas many other frameworks only support the connection of existing models in a loosely coupled fashion;
- spatial and temporal modeling entities are integrated parts of the OMS framework and not specific to an application domain; and
- OMS model development projects are tightly integrated with a web-based collaboration environment (i.e., the USDA Collaborative Software Development Laboratory) which makes it easier for model developers to efficiently collaborate on the design and development of application tools.

Major components within the OMS 2.1 platform are shown in Fig. 3. The goal of the majority of framework development efforts is to allow the agricultural, environmental, and ecological modeling community to focus more on the science module library and the system being modeled, thereby allowing core development, interpretation, and management requirements to be addressed more fully. Modularization is a key concept to simulation model development. The component-oriented and modular approach of the modeling frameworks and the modules/models implemented in them should provide a basis for more efficient and collaborative model development in the future.

Geographic Information System-Based and Visual Modeling Interfaces

With rapid improvements in computer technology and computational power, there are an increasing number of agricultural system models supported by Windows interfaces containing geographic information system (GIS) functionality to perform high-resolution simulation (e.g., multiple spatial land units within a field or watershed) of various biological, physical, and chemical processes in the soil–plant–atmosphere continuum. Most linkages between GIS and various models have been performed for models characteristically applied at larger scales (e.g., SWAT, WEPP-Watershed, and AGNPS), although field-scale models have been linked to GIS as well (e.g., GLEAMS and EPIC). Two important factors are commonly considered when developing high-resolution agricultural modeling tools: (1) agricultural system models require a large number of input data sets representing the initial conditions of the field or watershed; and (2) agricultural system models typically incorporate a large number of model parameters to quantify the underlying physical and biogeochemical processes. Thus, these tools should be capable of managing large and complex data sets while maintaining an acceptable level of user-friendliness. GIS technology has been used extensively for agricultural modeling including data preparation, model parameter extraction, and model results visualization. Research efforts have generally used one of the following approaches for linking GIS and simulation models: (1) embedding GIS functionalities into the model; (2) embedding

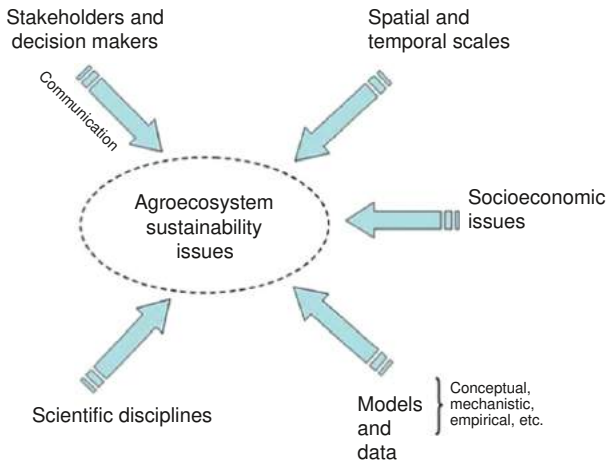


Fig. 4 Types of integration to address agroecosystem sustainability issues.

a model into a GIS; (3) loose coupling—that is, GIS is used to generate model input files and display model results; and (4) full coupling of GIS and a model. The most common approach for developing GIS-based agricultural modeling interfaces is loose coupling—examples of this include GIS linkages of the AGNPS, SWAT, and AGWA models. In general, the database (including multiple GIS layers) is used as a data-collection source, and GIS tools are used for extraction (in many cases automated) of necessary data from the GIS data layers including land-use areas, reach networks, soil information, and elevation. The extracted data are then passed to the model interface, read and processed, and model input files created for simulation. Many of the GIS-based interfaces for agricultural system models are based on the ESRI ArcView 3.x (Avenue) programming language. Ongoing research efforts for many agricultural system models are tied to new interface development under the ArcGIS 9.x (Visual Basic, VB · NET, Python, JavaScript) framework, and to system (software) updating to enhance Internet accessibility.

Traditionally, agricultural system models have been developed and implemented in a code-based modeling approach, that is, constructing a model as a sequence of lines of code in a common programming language such as Fortran, Visual Basic, C, C++, or Java. However, it is becoming increasingly more common to construct and link model components in a visual or 'icon-based' modeling environment. Proprietary visual modeling environments (e.g., Stella (isee systems, Inc.), Vensim (Ventana Systems, Inc.), POWERSIM (Powersim Software AS), and Model Builder (ESRI, Inc.)) are systems where models are constructed by assembling and linking icon-based model components using a sophisticated button- and menu-driven Graphical User Interface. There are three advantages in using these systems for agricultural model development: (1) they typically contain a large number of built-in functions (e.g., mathematical, logical, and statistical including risk and Monte Carlo analysis) that greatly simplify model development and evaluation; (2) they are easy to learn, intuitive to use, and familiar to users of Windows-based software; and (3) they can readily import and export data, generate graphs and tables, and work well together with other commercial applications through dynamic linking with spreadsheets and other software programs. In some respects, the icon-based modeling systems have many of the same capabilities of the more comprehensive modular frameworks described above; however, they are somewhat constrained in their flexibility and effectiveness in producing complex and multifaceted models. These limitations have been somewhat alleviated through a hybrid linkage of code- and icon-based modeling approaches. This concept has been successfully demonstrated by using Vensim to develop a new seed bank module and linking the module (through dynamic linked library support) to the APSIM farming systems model.

Scaling Issues

Scaling is an important issue for agricultural system models because results from fields or portions of a field often need to be extrapolated to another scale. A significant amount of new research in the refinement of agricultural system models is currently centered on determining scale-appropriate parameters for different model components. Generally, scaling-up involves both averaging spatial variability in parameters within a simulation unit, where the averaging process can be highly nonlinear, and incorporating process interactions beneath the simulation scale such that effects manifested at the scales of interest are captured by the upscaled parameters. For field and larger scales, uncertainty is most often due to unaccounted spatial variability of model parameters and processes within a simulation unit (typically assumed homogeneous) and errors in estimating so-called "effective parameters." For highly nonlinear soil-hydrologic processes, strictly speaking, there are no unique effective parameters; however, for practical purposes effective parameters may be calibrated to obtain a selected output variable.

For modeling and managing complex landscape and climate variability across multiple scales, ongoing research is helping to quantify the variability of soil parameters over space and time within a simulation unit using available spatial information about the causative factors, including physical soil properties and surrogate data such as terrain attributes. Management effects on soil hydraulic properties are being considered as well. For a given parent material, climate, biological factors, and time, topography is an important factor that has been shown to cause spatial variability of soil properties. Topographic data can now be rapidly and accurately measured at fine spatial intervals. An important question currently being researched is: can a set of topographic attributes in a given management system be related to spatial variability of soil properties, soil water content, and crop yield, and also used for upscaling?

New physically based methods of scaling up results from plots to field, farm, and watershed scales are also being developed. Historically, scaling of agricultural and landscape variables has been explored using a combination of empirical (data-based statistical) and theoretical (conceptual and numerical) methods. Both approaches have been used to explore scaling of soil properties and processes, field-scale relationships between grain yield, soil moisture, and topographic attributes, and theoretical scaling of infiltration using generated soils and rainfall patterns. Empirical methods provide evidence of real-world scaling behaviors that are useful for parameter estimation at different scales, while theoretical methods provide insights into process interactions in space and time that are currently infeasible to measure. Using measured spatial soil patterns in detailed spatial models allows one to simulate explicit interactions over space and time. Understanding and quantification of this information is being used to scale up responses over variable agricultural landscapes. In this way, the scale-dependence of agricultural management and conservation practices can be incorporated into larger scale (i.e., watershed and basin) models.

Integrated Assessment Approaches

As shown above, the historical motivation for understanding and modeling the dynamics and behavior of agricultural systems was primarily to assess and predict future food production and supply. More recently, however, it has become evident that in addition to functioning as entities for production, agricultural systems may also either damage or provide ecological goods and social

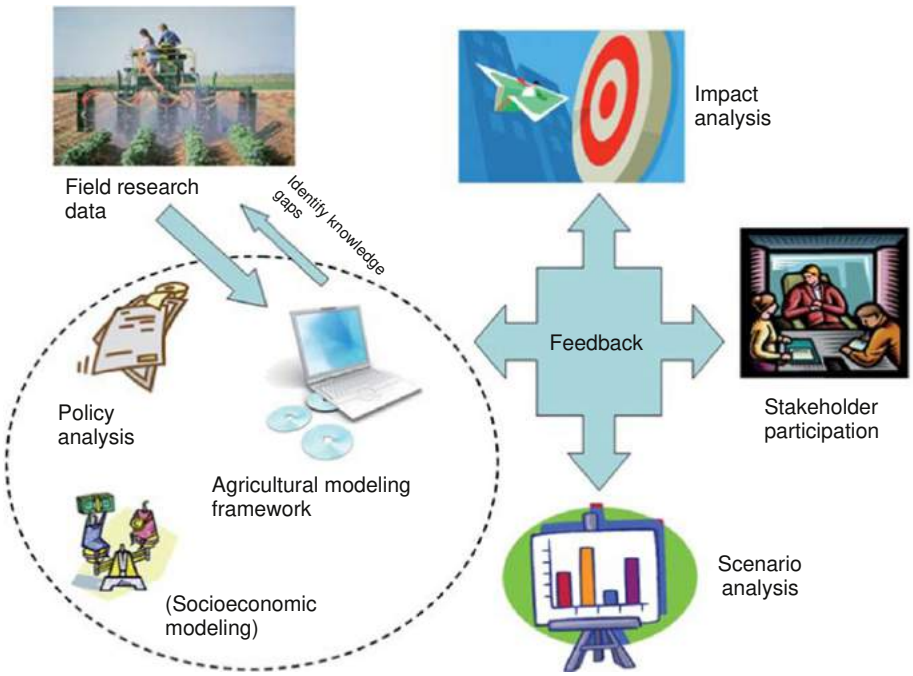


Fig. 5 Interactions among field research, modeling frameworks, and other components for integrated agroecosystem assessment.

welfare. Meeting sustainability challenges for agricultural system management requires approaches that assess socioeconomic concerns and environmental impacts integratively. IA is being increasingly used to integrate the numerous dimensions surrounding agroecosystem management including the consideration of multiple issues and stakeholders, the key disciplines within and between the human and natural sciences, multiple and cascading scales (both spatially and temporally) of agricultural system behavior, models of the different agricultural system components, and multiple databases (Fig. 4). Although there appears to be no universally agreed upon definition in the literature of what constitutes IA, there seems to be widespread agreement that IA:

- is a feedback-driven, interdisciplinary, and participatory (i.e., stakeholder involvement) process;
- is an iterative process of investigation and recommendation that stresses the importance of communication from scientist to decision-makers;
- explicitly accommodates linkages between the natural and human environment, and between research and policy; and
- uses the latest scientific tools including computer models, systems simulation, remotely sensed data, and other forms of information technology to assemble, integrate, and synthesize data from a wide range of sources and across a wide range of spatial and temporal scales.

Agricultural systems around the globe are continuously changing as a result of population demographics, climate fluctuations, and introduction of new agrotechnologies. There is consensus that modeling tools are needed to support sustainability within various agricultural sectors, and even more importantly to enhance the contribution of agricultural systems to sustainable development of societies at large. The integrated assessment modeling (IAM) process attempts to integrate various types of models (e.g., qualitative, quantitative, data, decision support) into an IA framework. More importantly, in current IAM approaches the earlier forms of systems modeling are being replaced with new integrated models that incorporate a three-pronged approach that considers ecological, social, and economic values when addressing sustainable usage of agricultural resources. Examples of this are recent IA analyses of climate change impacts on whole-farm systems. In these analyses, whole-system agricultural modeling frameworks were combined with a stakeholder-driven participatory process in order to assess potential effects of future climate change on agroecosystem land-use and management patterns. The System for Environmental and Agricultural Modelling; Linking European Science and Society Integrated Framework (SEAMLESS-IF) is an example of an IA tool that uses sustainability indicators (economic, environmental, and social) and agricultural systems evaluation (quantitative models, tools, and databases) to assess and compare alternative agricultural and environmental policy options. A review of the literature shows an increasing number of IAM exercises for solving agricultural problems; however, data availability, uncertainty characterization, and software platform development are key issues requiring additional research in the future.

Field Research and Modeling Integration

An important step toward improving model efficacy and usability is to further integrate system modeling with field research. An agricultural system involves complex interactions among several different components and factors (Fig. 1). These interactions need

interdisciplinary field research and quantification with the help of conceptual and process models. Integration of system models with field research has the potential to significantly enhance efficiency of agricultural research and raise agricultural science and technology to the next higher level. The integration will benefit both field research and models in the following ways:

- promote efficient and effective transfer of field research results to different soil conditions, climatic conditions, and alternative cropping and management systems outside the experimental design;
- encourage a whole-systems approach to field research that examines all major component interactions and facilitates better understanding of cause-and-effect relationships and quantification of experimental results; and
- assist field researchers to focus on identified fundamental knowledge gaps to make field research more efficient, and provide needed field evaluation and improvement of agricultural models before delivery to potential users.

A desirable vision for agricultural research and technology transfer is to have close integration between new field research, agroecosystem simulation models, and other components for integrated agroecosystem assessment (Fig. 5). After an agricultural system model has undergone thorough evaluation and both modelers and field scientists are satisfied with the results, it should be advanced to the application stage (with the goal of further model improvement through exposure to differing field conditions).

Future Research Needs for Agricultural System Models

As the development of agricultural models has progressed, controversy has not been lacking. Traditionalists felt that any attempt to simulate highly complex biological systems using mathematical algorithms in computers was bound to fail. Others argued that agricultural models should have a largely heuristic role in research rather than use as predictive tools. Further controversy regarding agricultural models stems from problems of complexity, testability, and parametrization. A commonly accepted truism is that agricultural models should be no more complex than the level of theory and measurements used to build them, and that components of models should also be balanced in terms of levels of complexity. Although this concept is extremely hard to quantify, it is important to note that this definition of model complexity is related to the kinds of questions the model is designed to answer rather than how much empiricism the model contains or the overall number of parameters.

What is incontrovertible is that the use of agricultural models to solve significant issues related to the economic and environmental sustainability of agroecosystems is on the rise rather than on the decline. Furthermore, the collective experiences of agricultural model developers and users show that, even though they are far from perfect, agricultural system models can be very useful in guiding field research, aiding technology transfer, and generating credible assessments of various impacts (e.g., sustainability, management, environmental impacts, climate change) on farming systems. However, a number of needs remain to be addressed that could improve agricultural system models and their application. Important issues include:

1. better quantification of how abiotic factors (e.g., water, temperature, light, nutrients) affect plant growth for both species and subspecies, and most importantly, genotype $\times$ environment interactions;
2. relationships among plant growth and other biotic factors (e.g., weeds, insects, diseases), including quantification of the competitive component to allow for better estimation of the effect of dynamic variations in plant population on growth and partitioning and the competitive aspects of crop–weed and crop–pest interactions;
3. development of comprehensive and common shared experimental databases based on existing standard experimental protocols, with measured values related to modeling variables so that conceptual model parameters can be experimentally verified;
4. improved collaboration between agricultural model developers and field scientists for appropriate experimental data collection, and for evaluation and application of models;
5. better methods of determining model parameters for different spatial and temporal scales, and for scaling and aggregating simulation results from plots to fields and larger scales;
6. continued development and improvement of environmental modeling frameworks to encourage replacement of monolithic agricultural models with modular component-based modeling tools where (a) each model component can be independently tested, improved, and easily substituted; (b) model components can vary with the scale of application; (c) hierarchical parameter estimation from varying degrees of input information is a component of the model; and (d) assembled agricultural systems models are kept compact and easy to use by customizing them to specific problems and regions; and.
7. better coordination of international efforts is needed in the future to improve agricultural system modeling and to encourage model developers and field scientists to work on identified knowledge gaps and research priorities.

Finally, while many important challenges and opportunities exist in model development, the greatest challenge facing the practitioners of agricultural system modeling in the future may center on demonstrating relevance to real-world decision-making rather than on building more accurate or comprehensive models.

Summary and Conclusions

This article describes the potential (and often realized) benefits of developing and implementing agricultural models for solving a broad range of agroecosystem research problems. It provides some examples of agricultural model applications in research and

management, presents various technologies surrounding current agricultural model development and use, and also identifies limitations and knowledge gaps in agricultural models where further improvements could be made. A remaining question to be answered is "what does the future hold?" Certainly there will be continuing (if not greater) attention to aspects of agricultural system models less related to agronomic production. The existing strong demand for agricultural models to address soil and water resource concerns (e.g., limited water issues, nutrient and organic matter cycling, soil salinity) and wider issues of sustainable agriculture in the face of global climate change (e.g., greenhouse gas emissions, wind and water erosion) will intensify. Environmental modeling frameworks will continue to mature, and development of large monolithic agricultural models will essentially disappear to be replaced by component-based models that link, as required, independent modules representing problem-specific processes of interest. The use of agricultural models in IA will continue to grow. Agricultural models provide objective tools to determine biophysical consequences of resource management options at various scales. To truly have a whole-systems approach, biophysical simulation and assessments need to be complimented by socioeconomic analyses (involving decision-makers and other stakeholders) before substantial benefits are realized. Finally, in order to further improve process representation in agricultural models (in whatever context they are used), model developers and field scientists in various scientific disciplines must continue to enhance their collaborative efforts.

Even with current shortcomings, agricultural models are very useful tools in synthesizing and quantifying scientific knowledge and experimental results across different climates, soils, and agricultural management systems at different locations across field, farm, and regional scales. The sustainability of agroecosystems depends on the maintenance of the ecological, social, economic, biological, and physical components that comprise the system. The high level of integration of these components implies that any evaluation of agroecosystem sustainability must consider the dynamics of these multiple factors. Agricultural models have become, and will remain in the future, important tools to facilitate assessment of agroecosystem sustainability.

See also: Conservation Ecology: Conservation Biological Control and Biopesticides in Agricultural. Ecological Complexity: Thermodynamics in Ecology. Ecological Processes: Grazing. Ecosystems: Agriculture Systems. General Ecology: Plant Ecology; Plant Physiology. Global Change Ecology: Sustainable Cropping Systems. Terrestrial and Landscape Ecology: Integrated Farming Systems

Further Reading

Ahuja, L.R., Andales, A.A., Ma, L., Saseendran, S.A., 2007. Whole-system integration and modeling essential to agricultural science and technology for the 21st century. Journal of Crop Improvement 19 (1/2), 73–103.
Ahuja, L.R., Ma, L., Howell, T.A., 2002. Agricultural system models in field research and technology transfer. Boca Raton, FL: CRC Publishers/Lewis Press.
Ahuja, L.R., Ma, L., Timlin, D.J., 2006. Trans-disciplinary research critical to synthesis and modeling of agricultural systems. Soil Science Society of American Journal 70, 311–326.
Anbumozhi, V., Reddy, V.R., Lu, Y.C., Yamaji, E., 2003. The role of crop simulation models in agricultural research and development: A review. International Agricultural Engineering Journal 12 (1 and 2), 1–18.
Belcher, K.W., Boehm, M.M., Fulton, M.E., 2004. Agroecosystem sustainability: A system simulation model approach. Agricultural Systems 79, 225–241.
Bland, W.L., 1999. Toward integrated assessment in agriculture. Agricultural Systems 60, 157–167.
Hunt, L.A., White, J.W., Hoogenboom, G., 2001. Agronomic data: Advances in documentation and protocols for exchange and use. Agricultural Systems 70, 477–492.
Jones, J.W., Hoogenboom, G., Porter, C.H., *et al*., 2003. The DSSAT cropping system model. European Journal of Agronomy 18, 235–265.
Jones, J.W., Keating, B.A., Porter, C.H., 2001. Approaches to modular model development. Agricultural Systems 70, 421–442.
Keating, B.A., McCown, R.L., 2001. Advances in farming systems analysis and intervention. Agricultural Systems 70, 555–579.
Kropff, M.J., Bouma, J., Jones, J.W., 2001. System approaches for the design of sustainable agro-ecosystems. Agricultural Systems 70, 369–393.
Matthews, R., Stephens, W., 2002. Crop-soil simulation models: Applications in developing countries. Wallingford, UK: CABI Press.
Peart, R.M., Curry, R.B., 1998. Agricultural systems modeling and simulation. New York: Dekker.
Shaffer, M.J., Ma, L., Hansen, S., 2001. Modeling carbon and nitrogen dynamics for soil management. Boca Raton, FL: CRC Publishers/Lewis Press.
Stoorvogel, J.J., 1995. Integration of computer-based models and tools to evaluate alternative land use scenarios as part of an agricultural systems analysis. Agricultural Systems 49, 353–367.
Teh, C., 2006. Introduction to mathematical modeling of crop growth: How the equations are derived and assembled into a computer program. Brown Walker Press.
Thornley, J.H., France, J., 2007. Mathematical models in agriculture: Quantitative methods for the plant, animal and ecological sciences. Cabi.
Van Ittersum, M.K., Leffelaar, P.A., van Keulen, H., Kropff, M.J., Bastiaans, L., Goudriaan, J., 2003. On approaches and applications of the Wageningen crop models. European Journal of Agronomy 18, 201–234.
Wallach, D., Makowski, D., Jones, J.W., Brun, F., 2013. Working with dynamic crop models: Methods, tools and examples for agriculture and environment. Academic Press.
Whisler, F.D., Acock, B., Baker, D.N., *et al*., 1986. Crop simulation models in agronomic systems. Advances in Agronomy 40, 142–208.
Wildi, O., 2017. Data analysis in vegetation ecology. CABI.
Wu, L., McGechan, M.B., Watson, C.A., Baddeley, J.A., 2005. Developing existing plant root system architecture models to meet future agricultural challenges. Advances in Agronomy 85, 181–219.

Big Data for Ecological Models

Marin M Kress, US Army Corps of Engineers, Engineer Research and Development Center, Vicksburg, MS, United States

Glossary

Crowdsourcing The act of utilizing the general public to achieve a goal, through active or passive participation examples include raising money, solving puzzles, or calculating traffic speed from cell phone data.

Extensible Extensible is something designed to accommodate future changes, such as a computer programming language designed so that future capabilities can be added.

Machine readable Computer files that are formatted to be ingested and understood primarily by a machine but which may be further processed to be human-readable.

Metadata Information about the dataset, separate from the data itself. Metadata may include elements such as instrumentation or software used in data generation, study duration, spatial extent, file type, keywords, and point of contact.

Remote sensing An observation made without interacting with the subject being observed, such as a photograph of a landscape

Introduction

There is an increasing call for ecological models to be parameterized with long-term datasets, and for modelers to archive their model runs and code to facilitate scientific-defensibility, transparency, and repeatability. These data archives are often created ad-hoc, without standards for the metadata or storage format. In a world full of linked search results and an ever-expanding number of journals and publications trying to gather the right information for your research can seem like an endless rabbit hole of possibilities. Proper data archiving and sharing can improve search results by connecting existing ecological data with those searching for it. To facilitate such archiving, sharing, and discovery it is useful to understand basic technical terms and practices relevant to these efforts. This entry covers concepts relevant to ecological data archiving including metadata, big data, machine readable data and common formats, and major ecologically relevant archive sources, with specific focus on those sources commonly used in ecological modeling. Although ecologists may not envision themselves as computer scientists it will behoove any researcher to understand the basics of the technological standards which underpin so much of modern research sharing. Digital records are not as permanently accessible as they may seem, and the ongoing shift from hardcopy observations to computer-generated and stored results requires careful thinking about how future researchers will be able to access historical files. Technologies that were once seen as revolutionary options for long term information storage (e.g., microfiche, laserdisc) may be less accessible than paper records in future due to lack of playback devices. Ultimately, when it comes to proper archiving and sharing the best advice is to think about the end before you begin collecting data.

Quick Reference Recommendations

The following is a brief list of recommendations related to ecological data archiving and sharing. The remainder of this article provides further explanation behind these recommendations:

- Identify funding agency requirements for data management and archiving before writing a grant proposal.
- Look for an established archive to deposit your data where it will be discoverable and accessible to the public with no further assistance from you. There may be a fee to deposit data.
- Think about metadata and record it as data is generated, rather than trying to reconstruct it after the project is over.
- Complete all metadata fields provided by the archive, and take advantage of space for keywords or tags.
- When the data set is complete, save it in both human-readable and machine-readable file formats for future use.
- If timestamps are used in your data include the local time zone or describe the time zone used in the metadata.

Big Data and Data Science

Ecologists and modelers who do not consider themselves "interdisciplinary data scientists" may not be aware of the contributions that their findings can make to other fields of scholarship. The rapid increase in computing availability, internet connectivity, the number of electronic records, and the changing nature of available digital content has ushered in an era of "big data" (Kress 2016).

This article uses the term big data to refer to any single dataset containing over 10 million records, although this threshold may shift with time. Big data can be any kind of large database, for example, millions of health records, aggregated news articles, or financial transaction records. In ecological sciences, big data could include genetic records across phyla, or sensor records of animal movements, or physical conditions such as temperature. Large quantities of high resolution photographic files can also grow in size to quickly match the data processing needs associated with other big data databases.

Big data is often described in terms of four elements: volume, variety, velocity, and veracity (IBM 2015). Volume refers to the scale of data, which could come from internal or external sources, on a global scale some estimate that 2.5 quintillion bytes of data are created each day (IBM 2015). Variety of big data refers to the different forms and sources, such as transaction data, social media, sensors, and mobile devices—including new products such as wearable wireless health monitors aimed at the general public (IBM 2015). This expanded volume and variety of data may also be generated at a faster velocity than previously seen because of computing and connectivity advances. For example, the New York Stock Exchange captures 1 terabyte of trade information during each trading session, a data stream of interest to both regulators and market analysts. Finally, veracity, or uncertainty, is an important element of big data (IBM 2015). With any data generation there is the potential for errors to enter a data stream. For datasets that grow rapidly and are continuously analyzed there is the potential for undetected data distortion to become magnified and for errors to propagate through dependent systems. IBM asserts that veracity of data is a significant issue and that poor data quality costs the US economy around $3 trillion per year (IBM 2015).

Big data refers not just to datasets, but also requirements for handling data in ways that go beyond the capabilities of traditional statistical software packages (Lohr 2013) or even personal computers. For example, at the time of writing in 2016 a single file in the current widely used version of the Microsoft Excel® software program can contain slightly over 1 million records (Microsoft Corporation 2016). New computer programs have been developed in response to the computational demands of big data analysis such as Apache™ Hadoop® (The Apache Software Foundation 2015), other hardware and software processing tools will continue to develop in response to changing technology and user needs. The rise of big data and expansion of other new sources of data (including digitization of historical paper records) does more than simply provide more data points, this availability can spur new questions about the world and the development of research subdisciplines. Finding, accessing, and organizing data across multiple disciplines is facilitated through best practices in data archiving and sharing. The next section describes important technical standards used in online data sharing.

Machine Readable Data

In scientific research there is no substitute for well-planned and carefully collected observational data. Such observations may be recorded with pen and paper, but increasingly, observations of the world around us are coming from digital instruments. Digital instruments may be associated with proprietary software packages, however many proprietary software packages offer output capabilities into standard open file formats. It is these standard open digital file formats, especially formats for data designed to be shared with a wide audience via the internet, that are the focus of this article.

Standard, nonproprietary, files formats are key to open data and data sharing. Readers are likely to be familiar with widely used electronic file formats such as Portable Document Format (PDF) or the graphics standard from the Joint Photographic Experts Group (JPEG or JPG). These types of files can be opened by many computer programs, and many programs can create files into these formats. However, files that are widely digitally accessible and designed for human readability may not be machine readable. Archiving and sharing required both human readability and machine readability, both terms are defined below.

Human readable: file formats that a human can read and interpret (but often a machine cannot), such as hand written notes, photographs, or PDF files.

Machine readable: file formats intended to be processed by another machine, organized in a specific structured format.

Machine readable data is critical to web services, which are the backbone of any sort of interactive website. Everything from online shopping to location-enabled restaurant recommendations through an app on your phone uses web services. Web services also run background processes you might not be aware of such as speed tracking that contributes to traffic forecasting. Web services matter to ecologists because anyone wanting to share real-time monitoring data will have to establish some type of web service to do so. A brief list of terms relevant to web services and data sharing is included in the **Table 1**. Creating or querying large databases on the internet will likely involve the use of these tools (either directly or behind the scenes).

How would an ecologist interface with these technological elements? In one case an ecologist might transfer field notes and data into a spreadsheet, save the file in a CSV format, and then post it to a website with an existing API that sends alerts when new files are available. In another setting, a researcher might sequence DNA, then create a file from a standard template with the DNA sequence and metadata, then submit the file to an archival institution which has established processes for querying their archive based on keywords and metadata.

Nontraditional Data Sources: Social Media and Crowdsourcing

Rooted in observations of the natural world by scientists, ecological research has expanded to take advantage of social media platforms and the power of crowdsourcing to promote broad participation in observations of natural phenomenon. One

Table 1 Commonly used abbreviations in online data sharing

Abbreviation	*Full name*	*Uses*
RDF	Resource Description Framework	An open standard data language to represent information as a web resource that can be linked to other web resources.
XML	eXtensible Markup Language	A language used to improve metadata accessibility. Proper XML use results in tags that can extracted by a computer, aiding information discovery.
JSON	JavaScript Object Notation	Machine readable data format, used to represent simple data structures and associative arrays. Used for data shared by 'syndication' in regularly updated feeds.
RSS	Really Simple Syndication	A programming language that contains both content and metadata, used in online "syndication" processes. Based on XML or RDF.
CSV	Comma Separated Value	An open file format for storing numerical and text data in a plain-text format.
API	Application Programming Interface	The way that machine readable information is made directly available to other machines. Used for programs that request information over an internet connection.

dictionary defines "crowdsourcing" as the "practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community rather than from traditional employees or suppliers." (Merriam-Webster Incorporated 2015) With the spread of the Internet, and Internet-connected smartphones, the ability for spatially distant groups to share information in near-real time is enormous. In addition to Internet connectivity, many mobile communication devices include the ability to share georeferenced information (i.e., latitude and longitude) along with an observation record (e.g., photo or social media posting) directly from the device itself. In some cases mobile phones are able to act as sensors without conscious action by their owner, a functionality already utilized by some companies to generate location-specific crowdsourced observations (Kress 2016). For example, any mobile phone with the Google Maps application and GPS-location services enabled will report movement speeds back to Google, which continuously combines data from millions of users and projects it through the Google Maps application in the form of color-coded street overlays (Barth and Google Inc. 2009). Examples of crowdsourced ecology research projects are listed below.

1. *Encyclopedia of Life*: This project began with the idea to provide a webpage for every species on earth, and seeks to bring together information from trusted resources such as museum, professional societies, and expert scientists into a massive database (Encyclopedia of Life Contributors 2015).
2. *Wildbook for Whale Sharks*: This photo-identification library allows anyone to upload photographs of whale sharks and associated location information for processing and photo matching by a trained researcher. The information provided by contributors is used to identify individual animals and help track the movements and distribution of the world's largest fish (Holmberg 2016).
3. *iNaturalist*: This free mobile phone application allows users to take observations and photographs of any animal or plant in the wild, then upload and share that information to a central website. The organization compiles and shares information with scientific repositories and makes data, and open source code, available to anyone (iNaturalist 2016).

Due to the potentially enormous numbers of individually-generated records from social media postings or georeferenced Internet search queries, there is interest in using these sources to monitor near real-time events, including predicting or tracking disease outbreaks or other ecologically relevant phenomenon.

Crowdsourced Data, Big Data Hubris, and Algorithm Dynamics

Researchers who rely on crowdsourced data or datasets of opportunity should be aware of fundamental limitations in the explanatory power of such data and the risk of "big data hubris." Big data hubris is the "assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis" because "quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data" (Lazer et al. 2014). The caution has even greater relevance when large data sets are generated through opaque processes that a researcher may not be able to account for within their analysis. For crowdsourced data from private sources, such as online search engines, it is also possible that algorithm dynamics may play a role in influencing any records generated. "Algorithm dynamics" refers to the changes made by software engineers to improve the commercial service being used and changes in behavior of consumers using the service (Lazer et al. 2014). The act of suggesting additional search terms, or using an "autocomplete" function in a text field, can influence user behavior. Lazer et al. (2014) note that "search behavior is not just exogenously determined, it is also endogenously cultivated by the service provider". For example, the act of presenting results in a certain order means that they have the potential to be seen by more people and receive more exposure, creating a positive feedback loop (Kress 2016).

Crowdsourced data from social media might be an accurate reflection of posted content within a single social media site, but such data should not be confused with a being an accurate representative sample of anything beyond a narrowly defined universe of subjects. For example, users of an insect-photo observation website might be more attracted to larger colorful species such as

butterflies, generating fewer records for smaller or more camouflaged species. Active participation rates for various online social media platforms may vary between demographic groups. However, other types of crowdsourced data do not require active participation or content generation, but instead only the use of a certain level of technology (i.e., smartphones) and so may draw from a larger population (Kress 2016).

Section Summary

Crowdsourced and social media data, including big data derived from these sources, can be seen as both rich and dangerous. Data culled from social media offers the potential of highly detailed temporal and spatial topical records that could provide fascinating social insights at low cost, but also of meaningless correlation on a grand scale. The difference between reported data (subject to human discretion) and recorded data (generated through established processes by sensors) will continue to be an important distinction. Ecological researchers who draw upon these sources for their research should document known limitations of such data within the metadata.

Interdisciplinary Health Research Using Ecological Data

Archived ecological data may contribute, and has contributed, to other fields of study such as medicine. Types of data that were not specifically designed for health-related research are regularly being used to investigate environmental health questions. To illustrate this point three examples of environmental health research that drew upon ecological data are summarized below.

Rift Valley Fever Predictions Based on Climate Anomalies

Anyamba et al. (2009) investigated the historical relationship between climate anomalies and Rift Valley fever (RVF) outbreaks in the Horn of Africa to develop a predictive model. Remotely-sensed ecological data relating to mosquito habitat was combined with epidemiological information on previous RVF outbreaks in the regional to develop early warning risk maps (Anyamba et al. 2009). Their results demonstrate the feasibility of combining remote-sensing data and historical epidemiology data to make near-term predictions about disease risk for a virus whose spread is closely coupled with regional environmental conditions (Kress 2016).

Linking Cholera Dynamics in Bangladesh to Environmental Influences

To examine the links between cholera risk and environmental forcing in Bangladesh, Koelle et al. (2005) tied together ecological data related to rainfall and river discharge, spatial data on flood extent, and epidemiological data on previous cholera cases to model the influence of climate anomalies on local cholera dynamics (Koelle et al. 2005). The question of climate anomalies and cholera disease ecology is an example of untangling the interaction between aquatic environmental conditions, disease vector biology, seasonal factors that affect large scale human behavior, and human population-level immunity using both traditional and non-traditional data (Kress 2016).

Predicting Kawasaki Disease in Japan from Regional Air Parcel Movements

Rodó et al. (2014) investigated possible airborne causative agents of Kawasaki Disease (KD) in Japan. They combined epidemiological data, regional wind pattern data, regional ecological land cover data, and microbial profiling of tropospheric and ground aerosol samples collected at times when air was coming from region identified as the possible source of the KD trigger (Rodó et al. 2014). The researchers found that air parcels associated with higher incidence of KD in Japan had previously moved over intensively cultivated croplands for corn, rice, and wheat in northeastern China during a time when the ground was frozen. By combining remote sensing records, field measurements of the airborne, and spatially-based epidemiological data, this research has suggested new avenues of investigation to understand and predict changes in KD risk (Kress 2016).

Section Summary

Revealing linkages between large-scale physical environmental phenomenon (such as sea-surface temperature anomalies, wind patterns, or rainfall levels) and local conditions is increasingly important in solving complex problems involving multiple states or countries (Kress 2016). The examples cited above illustrate the value of archived ecological data when applying a synthesis approach to complex research questions.

Ecology Data in the Public Domain but not Discoverable Online

The potential audience for ecological data is wide, spanning basic science, commercial operators with business interests, and the public health and governance community (Kress 2016). In addition to academic or private research efforts, public government entities engage in a variety of ecology-related research. However, not all data that falls within the public domain is automatically available online. In the United States data is generally considered to be in the public domain if it is the product of official government work, or work paid for with public funds. In 2013 the Presidential Executive Order 13642 "Making Open and Machine Readable the New Default for Government Information" directed US federal government agencies to release data to the public in ways that "make the data easy to find, accessible, and usable" (Obama 2013). This generally means publishing the data online. Data discovery is facilitated through the Data.Gov website, which provides an every-expanding catalogue of federal datasets data available to the public. At the time of writing there were over 190,000 datasets with searchable descriptions and web links available to the public through Data.Gov (US General Services Administration 2016).

For local and state-level offices with limited resources it may not be standard practice to post all public data online. In these cases it may be necessary for researchers to contact public agencies directly to request that they share data files directly. One example of such public domain ecological data is the state-sponsored water sampling of Boston Harbor and Massachusetts Bay. The Massachusetts Water Resources Authority (MWRA) has made certain water quality monitoring data (e.g., bacteria counts, nutrient observations) from Boston Harbor available for select observations starting in 1989 (Massachusetts Water Resources Authority 2015b). However, at the time of writing, similar water quality monitoring results for the larger Massachusetts Bay area are not posted online. Instead, researchers may contact the MWRA to initiate a discussion with staff members who can assist them in their research and provide files extracted from the main database held by the MWRA. Although this is just one example, similar situations likely exist at other state and local government agency offices which may hold ecologically-relevant data.

Notable Public Data Archives

The results from purpose-designed experiments generating direct observations still constitute the highest tier of scientific data, complementing such experimental data are environmental monitoring data which may reveal changes over long time periods (Kress 2016). Magnifying the value of the cumulative efforts of individual scientists, laboratories, and institutions are national and international repositories for data or publications. These data repositories may be managed or funded by nonprofit organizations, academic institutions, government agencies, or some combination thereof. Major examples of such databases are described below, these examples were chosen because of their stability and public accessibility. Some of the example archives are focused on providing public access to government-funded research products, while others allow submissions from the broader research community. Scientists who might wish to deposit their findings in such public archives should consult the associated format requirements early on in their research to make the final archiving process easier.

Ecology and Environment Databases

These databases examples span multiple environment types and may include both biotic and abiotic environmental data.

1. *Ecological Society of America Data Registry*: This registry describes data sets on ecology and environmental topics from articles published in the journals of the Ecological Society of America (Ecological Society of America 2015).
2. *Integrated Ocean Observing System (IOOS®)*: A regional-national partnership for sharing ocean, coastal, and Great Lakes data on topics including wave heights, sea level, wind, temperature, salinity, and dissolved oxygen levels. IOOS is a member of the Global Ocean Observing System (GOOS) coordinated through the United Nations (National Oceanic and Atmospheric Administration 2014). Within the United States there are multiple regional contributors to ocean and weather monitoring data streams ranging from the Northeastern Regional Association of Coastal and Ocean Observing Systems (NERACOOS) to the Pacific Islands Ocean observing System (PacIOOS) (Northeastern Regional Association of Coastal and Ocean Observing Systems 2014).
3. *TRY Plant Trait Database*: The Max Planck Institute for Biogeochemistry in Germany hosts this international database developed by scientists of morphological, anatomical, biochemical, physiological, or phenological features of plants, with many geo-referenced records (Boenisch and Kattge 2014).
4. *National Centers for Environmental Information*: A project of the US National Oceanic and Atmospheric Administration, this website provides public access to climate and historical weather data, along with geophysical, oceanographic, and coastal data (National Oceanic and Atmospheric Administration, 2016). This source publishes data that spans "from the depths of the ocean to the surface of the sun and from million-year-old sediment records to near real-time satellite images." (National Oceanic and Atmospheric Administration, 2016).
5. *Water Information System for Europe*: A joint product of the European Commission and the European Environment Agency, this website connects users with water-related information including water statistics, forecasting models, and policy information across Europe (European Commission and European Environment Agency 2016).

6. *National Water Information System*: Produced by the US Geological Survey, this archive is the principal repository of water resources data for the United States. Through the public website users can access data on streamflow, floods, drought, groundwater, and aquatic habitats (US Geological Survey 2016).
7. *Australian Soil Resource Information System*: This government-run website provides information on soil resources in Australia in a consistent format. Records include information on soil depth, water storage, permeability, fertility, carbon and soil erodibility (Commonwealth Scientific and Industrial Research Organisation 2013).
8. *Long Term Ecological Research Network*: The Long Term Ecological Research Network was established with support from the National Science Foundation with the goal of understanding "a diverse array of ecosystems at multiple spatial and temporal scales." Observational and experimental data from over 25 sites across North America, Tahiti, and Antarctica are available through an online data portal (Long Term Ecological Research Network 2013).

Multi-topic Databases

These multi-topic databases may be useful to researchers in a wide variety of fields, from ecology and biology to computer science to history.

1. *US Census Bureau*: This website provides historical demographic, economic, health, housing, and other official statistical data for the United States. Products may combine categorical data with spatial detail at the level of census block groups (ranging from 600 to 3000 people) (US Census Bureau and US Department of Commerce, 2015; US Census Bureau 2012). The level of detail published by the US Census allows for nuanced spatial analysis over time.
2. *Google Trends*: This website from Google displays stories that are "trending" based on user-entered search terms in the free Google search engine. Topics can be filtered by categories such as "Business" or "Sci/Tech" and by country. Within the United States, Google Trends displays a map of interest by region (state level) for an individual story (Google Inc. 2015).
3. *Dryad Digital Repository*: Dryad is a nonprofit long-term repository for data used in international scientific and medical literature, including data in the form of text, spreadsheets, video, photographs, and software code (Dryad 2015a). Datasets deposited in Dryad are free to use and citable in new publications (Dryad 2015b). Each Dryad data package received a unique Digital Object Identifier that can be used when citing or locating data. Datasets are free to use but there is a small charge for depositing data packages with Dryad (Dryad 2015a).

Health and Medicine Databases

These databases are relevant to health and medicine researchers, topics range from basic biology to clinical specialties. There are numerous websites and databases that serve specific molecular biology topics.

1. *PubMed*: PubMed is a database of over 24 million citations from biomedical literature, it is managed by the US National Institutes of Health (NIH, a government entity) (US National Library of Medicine 2015).
2. *GenBank®*: An annotated genetic sequence database of all publicly available DNA sequences maintained by the NIH since 1982. GenBank releases a public update every 2 months and, as part of the International Nucleotide Sequence Database Collaboration, exchanges data with the DNA DataBank of Japan and the European Molecular Biology Laboratory. Each nucleotide sequence uploaded to GenBank receives a unique Accession Number, as of mid-2015 GenBank has archived over 100 million sequence records representing over 100 billion nucleotide bases (National Center for Biotechnology Information 2014).
3. *Foodborne Outbreak Surveillance System (FOSS) Online Database*: This database is run by the US Centers for Disease Control and Prevention (CDC). FOSS receives reports from state, local, and territorial public health agencies about recorded foodborne illnesses (Centers for Disease Control and Prevention 2013).
4. *HealthData.Gov*: This website is run by the US Department of Health & Human Services (HHS) and aims to make data from the HHS agencies (including the CDC, FDA, and NIH) easily available and accessible to the public. This evolving website aims to make all the data it serves up to be machine-readable, downloadable, and accessible via application programming interfaces (US Department of Health and Human Services 2015).
5. *ClinicalTrials.Gov*: This website is maintained by the National Library of Medicine in the United States. The archive contains information about publicly and privately supported medical studies in human volunteers in the United States. Both interventional studies (aka clinical trials) and observational studies are included in the records (National Library of Medicine 2016).

Environmental Health Databases

These databases contain information relevant to environmental topics with a close relationship to human health.

1. *Enhanced Infectious Diseases (EID2) database*: This database, funded by the European Union and hosted at the University of Liverpool, contains data on pathogenic organisms and the country in which they may occur, lists of carrier organisms, genetic sequences, and publication links (University of Liverpool 2015).

2. *Center for Coastal Monitoring and Assessment National Status & Trends Database (NS&T)*: Run by the US National Oceanic and Atmospheric Administration (NOAA), the NS&T is comprised of three nationwide programs, Benthic Surveillance (discontinued in 1993), Mussel Watch and Bioeffects that are designed to describe the current status of, and detect changes in, the environmental quality of US estuarine and coastal waters through environmental monitoring, assessment and related research. Starting in 1986, the Mussel Watch program is the longest running continuous contaminant monitoring program in US coastal and Great Lakes waters (National Oceanic and Atmospheric Administration 2012).
3. *National Pollutant Discharge Elimination System Permit (NPDES) Database*: A product of the US Environmental Protection Agency this database provides spatial information about facilities holding NPDES permits, including the associated latitude and longitude data (US Environmental Protection Agency 2015).

Remote Sensing Data Sources

Some national space agencies are public sources of satellite remote sensing data and model output products derived from that data. Other remote-sensing data is available from private entities.

1. *US National Aeronautics and Space Administration (NASA)*: NASA provides satellite remote sensing data from multiple spacecraft and instruments sources with varying temporal scales, spatial scales, and image resolution. Topic areas include global precipitation, thermal anomalies, ocean color, land cover and vegetation, and snow and sea ice cover (National Aeronautics and Space Administration 2014).
2. *European Space Agency*: The European Space Agency provides public data related to radar imagery, radar altimetry, optical/multispectral radiometry, atmospheric data, and gravimetric data from multiple missions (European Space Agency 2015).
3. *Japan Aerospace Exploration Agency*: The Japan Aerospace Exploration Agency provides data to the public from satellites and probes. Observation data from multiple satellites provide remotely sensed records on topics such as greenhouse gases, typhoons, and precipitation (Japan Aerospace Exploration Agency, 2016).

The expanding use of small unmanned aerial vehicles (UAVs) (e.g., drones, quad-copters, low orbit weather balloons) with the capacity to carry photographic and sensory equipment is rapidly changing the amount of remotely sensed data available to researchers (Botanical Society of America 2016). As this field develops new repositories may emerge with their own sets of metadata standards.

The data archived in each source listed above requires specialized knowledge to interpret. For example, foodborne illness outbreaks are a different type of data than land cover type images, but when combined they might provide scientific new insights. Proper metadata documentation may help with the data discovery and hypothesis generation process. The need for individuals or specialized data science teams that can combine and utilize diverse data types may grow as society poses research questions related to multiple disciplines (Kress 2016).

Database of Researchers

This article focuses on data archiving and sharing, but researchers should also be aware of ORCiD (2016), an organization which will provide them with a unique number to associate with their name on publications. The purpose of a unique identifying number is for scientists to keep track of their intellectual contributions over time even as their name, affiliation, or field of research changes. Some academic journal publishers and funding agencies are requiring ORCiD identification codes from researchers as part of their normal administrative process (ORCiD 2016). While the practice of tying a unique number to an individual's publication record is clearly beneficial to people with commonly used names or names translated from other languages, it will also serve to keep a consistent record across publications with varying author name formats.

The Variety of Data Sources Consulted for a Single Project

Well documented, and discoverable, ecological data sources can contribute to multiple interdisciplinary or cross-disciplinary studies. For example, the **Table 2** from a doctoral dissertation in the field of ocean and human health shows the wide variety of sources consulted in a single research effort. Ecology data (water samples for microorganisms, nutrient data, physical observations) was combined with human demographic data, geological data (river flow), and public health data (beach water quality testing) to investigate relationships between human inputs to coastal waters and their influence on potential marine-sourced risks to human health from those same coastal waters (Kress 2016).

Chapter Summary

This article discussed the changes in the type and amount of data available to ecological modelers, and the technological standards that are important to web-based data sharing processes. Data accessibility changes include increasing numbers of observations

Table 2 Example of sources used in a single environmental health research project

Source name	*Source type*	*Data types*	*Sampling frequency*	*Source*
US Census, Decennial Census	Federal government	Population, age, sex, housing units, household income, other demographic data	Every 10 years, entire United States	US Census Bureau and US Department of Commerce (2015)
US Census, American Community Survey	Federal government	Housing stock, wastewater treatment, other demographic data	Every year, approximately 1 in 38 US households receive the survey	US Census Bureau (2014)
Dog population in coastal cities bordering, Massachusetts Bay	Compiled by author via phone and email survey in 2012	Number of dogs registered to town, some estimates of unregistered dogs	One time survey	Unpublished data
US Geological Survey, River discharge data	Federal government	Average daily discharge rate at 8 stations along waterways, no data on Cape Cod	Continuous, June–August records 2007–14	US Geological Survey (2015a), and US Geological Survey (2015b)
Northeastern Regional Association of Coastal and Ocean Observing Systems	Nonprofit organization, partnership	Ocean and weather conditions from buoys in the northeast. Air temperature, water temperature at multiple depths, salinity, chlorophyll, turbidity, wind direction, current direction	Minute by minute, but reports of daily averages are available. Annual data acquired for of 2000–14.	Northeastern Regional Association of Coastal and Ocean Observing Systems (2014)
NOAA Buoy Station BHBM3, Boston Harbor	Federal government, data also served via NERACOOS	Air temperature, water temperature	6-min intervals	National Oceanic and Atmospheric Administration (2015)
NOAA National Climatic Data Center	Federal government	Precipitation, air temperature to tenth of degree, average daily wind speed	Daily average	National Oceanic and Atmospheric Administration and National Centers for Environmental Information, 2015)
Massachusetts Department of Public Health, Beach Water Quality Testing	State government	*Enterococcus* sampling results	Weekly, monthly, or daily during summer bathing season depending on location and previous test results	Massachusetts Department of Public Health (2012)
Massachusetts Water Resources Authority	State government	Boston Harbor bacteria counts	Varies, weekly or monthly	Massachusetts Water Resources Authority (2015b)
Massachusetts Water Resources Authority	State government	Boston Harbor nutrient data: Ammonium, Nitrate + nitrite, Total Kjeldahl Nitrogen, Phosphate, Total phosphorus, Chlorophyll *a*, Phaeophytin	Varies, weekly or monthly, Acquired data spanning 1992–2014	Massachusetts Water Resources Authority (2015b)
Massachusetts Water Resources Authority	State government	Ammonium, Nitrate + nitrite, Phosphate, Total P/N, Particulate P/N/C, Chlorophyll *a*, Silicate, Salinity, zooplankton	Varies, weekly or monthly, Acquired data spanning 1995–2014	Massachusetts Water Resources Authority (2015a)
Massachusetts Water Resources Authority	State government	Beach water quality, bacteria counts and precipitation in the form of rainfall	Spring to fall. Daily during summer bathing season	Massachusetts Water Resources Authority (2015b)
Massachusetts Water Resources Authority	State government	*Pseudo-nitzschia* species count data	Approximately monthly from 1992 to 2014 (date range varies by station)	Massachusetts Water Resources Authority (2015c)
US Environmental Protection Agency	Federal government	Location and information for facilities within a National Pollutant Discharge Elimination System (NPDES) permit	Monthly	US Environmental Protection Agency (2015)
Massachusetts Department of Public Health	State government	Enteric diseases diagnosed in the Commonwealth of Massachusetts	Annually, with 1 + year lag for public release	Massachusetts Department of Public Health (2013)

from multiple sources, and the increasing speed of data generation from traditional sources as well as new sources like social media platforms or mobile phone-based applications. For ecological data and models to have the greatest impact, they should be properly documented, archived, and shared through an online platform. Dataset discoverability and usability, scientific defensibility, and transparency are all improved using clearly documented metadata and open file formats that facilitate file sharing across software programs.

See also: Aquatic Ecology: Ecosystem Health Indicators—Freshwater Environments. Ecological Complexity: Ecological Data Archiving and Sharing. Ecological Data Analysis and Modelling: Mediated Modeling and Participatory Modeling; Statistical Inference

References

Anyamba, A., Chretien, J.P., Small, J., Tucker, C.J., Formenty, P.B., Richardson, J.H., Britch, S.C., Schnabel, D.C., Erickson, R.L., Linthicum, K.J., 2009. Prediction of a Rift Valley fever outbreak. Proceedings of the National Academy of Sciences of the United States of America 106 (3), 955–959.

Barth, D., and Google Inc. 2009. The bright side of sitting in traffic: Crowdsourcing road congestion data. http://googleblog.blogspot.com/2009/08/bright-side-of-sitting-in-traffic.html (accessed September 19, 2015).

Boenisch, G., and Kattge, J. 2014. TRY Plant Trait Database: About. https://www.try-db.org/TryWeb/About.php (accessed May 31, 2015).

Botanical Society of America. 2016. Drones take off in plant ecological research: New review explores how to speed up and scale up plant research with aerial robotics. https://www.sciencedaily.com/releases/2016/10/161031113546.htm (accessed November 25, 2016).

Centers for Disease Control and Prevention. 2013. The Foodborne Outbreak Online Database (FOOD Tool). http://www.cdc.gov/foodsafety/fdoss/data/food.html (accessed May 31, 2015).

Commonwealth Scientific and Industrial Research Organisation. 2013. Australian Soil Resource Information System. http://www.asris.csiro.au/index.html.2016.

Dryad. 2015a. Dryad: Frequently Asked Questions. http://datadryad.org/pages/faq (accessed October 10, 2015).

Dryad. 2015b. Dryad: The organization: Overview. http://datadryad.org/pages/organization (accessed October 10, 2015).

Ecological Society of America. 2015. Ecological Society of America Data Registry. http://data.esa.org/esa/style/skins/esa/index.jsp (accessed May 31, 2015).

Encyclopedia of Life Contributors. 2015. Encyclopedia of Life. http://eol.org/about (accessed May 31, 2015).

European Commission, and European Environment Agency. 2016. The Water Information System for Europe. http://water.europa.eu/.2016.

European Space Agency. 2015. ESA Earth Online. https://earth.esa.int/web/guest/data-access (accessed July 13, 2015).

Google Inc. 2015. Google Trends. https://www.google.com/trends/ (accessed July 13, 2015).

Holmberg, J. 2016. Wildbook for Whale Sharks. http://www.whaleshark.org/ (accessed December 7, 2016).

IBM. 2015. The Four V's of Big Data. http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg (accessed July 12, 2015).

Japan Aerospace Exploration Agency. 2016 Observation/Research Result Database. http://global.jaxa.jp/projects/db/index.html (accessed December 12, 2016.

Koelle, K., Rodó, X., Pascual, M., Yunus, M., Mostafa, G., 2005. Refractory periods and climate forcing in cholera dynamics. Nature 436 (7051), 696–700.

Kress, Marin M. 2016. Identification and use of indicator data to develop models for marine-sourced risks in Massachusetts bay. PhD Thesis, University of Massachusetts, Boston.

Lazer, D.M., Kennedy, R., King, G., Vespignani, A., 2014. The parable of Google Flu: Traps in big data analysis. Science 343 (6176), 1203–1205.

Lohr, S., 2013. The origins of "Big Data": An etymological detective story. The New York Times. 02/04/2013.

Long Term Ecological Research Network. 2013. The Long Term Ecological Research Network. https://lternet.edu/ (accessed December 18, 2016).

Massachusetts Department of Public Health. 2012. BeachWaterQualityData_2003-2011.xls [MS Excel file]. Massachusetts Department of Public Health, Boston, MA.

Massachusetts Department of Public Health. 2013. Enteric disease in Massachusetts: 1999–2013. http://www.mass.gov/eohhs/docs/dph/cdc/foodsafety-enterics-state-totals.pdf (accessed March 17, 2015).

Massachusetts Water Resources Authority. 2015a. 1995-2015_F22_F23.xlsx [MS Excel file]. Massachusetts Water Resources Authority, Boston, MA.

Massachusetts Water Resources Authority. 2015b. Boston Harbor and Massachusetts Bay: Water Quality Data. http://www.mwra.state.ma.us/harbor/html/wq_data.htm (accessed May 17, 2015).

Massachusetts Water Resources Authority. 2015c. pseudonitz_1992–2014.xlsx [MS Excel file]. Massachusetts Water Resources Authority, Boston, MA.

Merriam-Webster Incorporated. 2015. Crowdsource: Definition. http://www.merriam-webster.com/dictionary/crowdsourcing (accessed March 8, 2015).

Microsoft Corporation. 2016. Excel specifications and limits. https://support.office.com/en-nz/article/Excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3.

National Aeronautics and Space Administration. 2014. NASA's Earth Observing System. http://eospso.gsfc.nasa.gov/content/nasa-earth-science-data (accessed May 31, 2015).

National Center for Biotechnology Information. 2014. GenBank Overview. http://www.ncbi.nlm.nih.gov/genbank/ (accessed May 31, 2015).

National Library of Medicine. 2016. ClinicalTrials.Gov Background. https://clinicaltrials.gov/ct2/about-site/background (accessed December 12, 2016).

National Oceanic and Atmospheric Administration. 2012. NS&T Program Download Page. http://ccma.nos.noaa.gov/about/coast/nsandt/download.aspx (accessed June 24, 2015).

National Oceanic and Atmospheric Administration. 2014. Integrated Ocean Observing System: About (IOOS). http://www.ioos.noaa.gov/about/welcome.html (accessed August 8, 2015).

National Oceanic and Atmospheric Administration. 2015. National Data Buoy Center: Station BHBM3–8443970—Boston, MA. http://www.ndbc.noaa.gov/station_page.php?station=bhbm3 (accessed May 25, 2015).

National Oceanic and Atmospheric Administration. 2016. National Centers for Environmental Information. https://www.ncdc.noaa.gov/.2016).

National Oceanic and Atmospheric Administration, and National Centers for Environmental Information. 2015. National Centers for Environmental Information: Climate Data Online Search. http://www.ncdc.noaa.gov/cdo-web/confirmation (accessed December 31, 2014).

iNaturalist, 2016. iNaturalist: About. http://www.inaturalist.org/pages/about (accessed December 7, 2016).

Northeastern Regional Association of Coastal and Ocean Observing Systems. 2014. NERACOOS: Data & Tools. http://neracoos.org/datatools (accessed May 25, 2015).

Obama, B., 2013. Executive order 13642—Making open and machine readable the new default for government information. Washington, DC: Federal Register, http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government

ORCiD 2016. ORCiD. http://orcid.org/ (accessed November 25, 2016).

Rodó, X., Curcoll, R., Robinson, M., Ballester, J., Burns, J.C., Cayan, D.R., Lipkin, W.I., *et al*., 2014. Tropospheric winds from northeastern China carry the etiologic agent of Kawasaki disease from its source to Japan. Proceedings of the National Academy of Sciences of the United States of America 111 (22), 7952–7957.

The Apache Software Foundation. 2015. Apache™ Hadoop®. https://hadoop.apache.org/ (accessed January 13, 2016).

University of Liverpool. 2015. Enhanced Infectious Diseases (EID2) Database. http://www.zoonosis.ac.uk/eid2 (accessed May 31, 2015).

US Census Bureau. 2012. Geographic Terms and Concepts—Block Groups. http://www.census.gov/geo/reference/gtc/gtc_bg.html (accessed July 2, 2015).

US Census Bureau. 2014. American Community Survey: About. http://www.census.gov/acs/www/about_the_survey/american_community_survey/ (accessed May 24, 2015).

US Census Bureau and US Department of Commerce. 2015. US Census Bureau Homepage. http://www.census.gov/# (accessed March 13, 2015).
US Department of Health and Human Services. 2015. HealthData.Gov: About. http://www.healthdata.gov/content/about (accessed July 13, 2015).
US Environmental Protection Agency. 2015. Envirofacts: Data Downloads: Custom Search—PCS. http://www.epa.gov/enviro/facts/datadownloads.html (accessed September 25, 2015).
US General Services Administration. 2016. Data.Gov: About. https://www.data.gov/about (accessed December 18, 2016).
US Geological Survey. 2015a. National Water Information System. http://waterdata.usgs.gov/nwis (accessed May 24, 2015).
US Geological Survey. 2015b. National Water Information System: Mapper. http://maps.waterdata.usgs.gov/mapper/index.html (accessed May 24, 2015).
US Geological Survey. 2016. Water Resources of the United States. http://water.usgs.gov/data/.2016.
US National Library of Medicine. 2015. PubMed. http://www.ncbi.nlm.nih.gov/pubmed (accessed May 31, 2015).

Further Reading

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Proceeding of the Second International Symposium on Information Theory. 267–281.
Biological and Chemical Oceanography Data Management Office, and National Science Foundation. 2015. NSF Two Page Data Management Plan. http://www.bco-dmo.org/nsf-two-page-data-management-plan (accessed December 4, 2016).
Bowen, R.E., Depledge, M.H., Carlarne, C.P., Fleming, L.E. (Eds.), 2014. Oceans and human health: Implications for society and well-being. Oxford, England: Wiley.
Boyd, D., Crawford, K., 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, communication & society 15 (5), 662–679.
Burnham, K.P., Anderson, D.R., 2002. Model selection and multimodel inference: A practical information-theoretic approach (2nd edition). New York: Springer.
Hampton, S.E., Strasser, C.A., Tewksbury, J.J., Gram, W.K., Budden, A.E., Batcheller, A.L., *et al*., 2013. Big data and the future of ecology. Frontiers in Ecology and the Environment 11 (3), 156–162.
Heffernan, J.B., Soranno, P.A., Angilletta, M.J., Buckley, L.B., Gruner, D.S., Keitt, T.H., *et al*., 2014. Macrosystems ecology: Understanding ecological patterns and processes at continental scales. Frontiers in Ecology and the Environment 12 (1), 5–14.
Hendler, J., and Pardo, T. A. 2012. A Primer on Machine Readability for Online Documents and Data. https://www.data.gov/developers/blog/primer-machine-readability-online-documents-and-data (accessed November 30, 2016).
Schimel, D., Keller, M., 2015. Big questions, big science: Meeting the challenges of global ecology. Oecologia 177 (4), 925–934.
Soranno, P.A., Schimel, D.S., 2014. Macrosystems ecology: Big data, big ecology. Frontiers in Ecology and the Environment 12 (1), 3.
Whitlock, M.C., 2011. Data archiving in ecology and evolution: Best practices. Trends in Ecology & Evolution 26 (2), 61–65. doi:10.1016/j.tree.2010.11.006.

Relevant Websites

http://datadryad.org/ —Dryad Digital Repository.
http://data.esa.org/ —Ecological Society of America: Data Registry.
https://lternet.edu/ —Long Term Ecological Research Network.
http://www.ncbi.nlm.nih.gov/genbank/ —National Center for Biotechnology Information: GenBank.
https://www.ncdc.noaa.gov/ —National Oceanographic and Atmospheric Administration: National Centers for Environmental Information.
http://orcid.org/ —ORCiD.
http://water.usgs.gov/data/ —US Geological Survey: Water Resources of the United States.
http://www.ncbi.nlm.nih.gov/pubmed —US National Library of Medicine: PubMed.

Biogeochemical Models☆

Joyita Mukherjee, Krishna Chandra College, Hetampur, West Bengal, India
Santanu Ray, Visva – Bharati University, Santiniketan, India

Introduction

The field of biogeochemistry deals with the effect of biological organisms on the chemistry of the Earth. Since there are numerous living organisms, all of which affect the chemistry of their environment in multiple ways, biogeochemistry is a large subject area covering many processes. One process that receives a lot of attention is the emission of CO_2 by humans and the associated global increase in atmospheric CO_2. However, biogeochemistry also deals with the effect of other organisms on the global carbon cycle, like the conversion of CO_2 to organic carbon by marine phytoplankton. Biogeochemical processes can also be of subglobal scale, like the respiration of O_2 by bacteria at the bottom of a lake. Biogeochemistry encompasses all types of chemicals. Nitrification, the conversion of ammonia to nitrate, by bacteria in soils is a biogeochemical process. So is the reduction of sulfate to sulfide by bacteria in groundwater. Also, the chemistry of minor elements is included, like the methylation of mercury by bacteria in sediments. Further, although the field of biogeochemistry traditionally focuses on naturally occurring elements and compounds, it also includes the effect of organisms on the chemistry of manmade chemicals, like the biodegradation of polychlorinated biphenyls (PCBs) by bacteria.

Biogeochemical models are abstract and simplified representations of biogeochemical processes. This includes qualitative models in the form of narratives or diagrams that describe how a process works and convey mechanistic information. It also includes quantitative models in the form of mathematical equations that predict chemical concentrations and fluxes. Quantitative biogeochemical models are used as research tools to test hypotheses, and as management tools to evaluate "what if" scenarios (e.g., nutrient load reduction to prevent eutrophication of lakes). In application, biogeochemical models are typically smaller components of larger models that describe the biogeochemical cycling of elements at various scales ranging from a small volume of soil to the globe. As such, biogeochemical models have to be compatible with other models, including physical models that describe the transport of chemicals in the environment (advection, diffusion, and dispersion) and ecological models that describe population dynamics.

The purpose of this article is to present an overview of biogeochemical modeling. A thorough review of this article would necessitate covering the effect of every biological organism on the chemistry of all affected compounds, which is beyond the scope of this article. Therefore, this article focuses on how the transformation of chemicals by organisms is modeled in general, with applicable references to actual processes. Microorganisms constitute the bulk of the biomass and they have a higher turnover rate than organisms at higher trophic levels. They are therefore generally considered to be the main drivers of biogeochemistry and this article will focus on them. Also, consistent with the scope of this encyclopedia, the article focuses on naturally occurring substances, rather than manmade ones (e.g., PCB). Often, the effect of organisms on chemistry is indirect (e.g., via the redox potential), but that is, strictly speaking, a chemical problem, and this article therefore focuses on the direct effect of organisms on the chemistry. First, basic modeling approaches are reviewed, including conceptual and descriptive models, mechanistic chemistry- and biology-type models, and empirical models. Then the integration of ecology and biogeochemistry models is discussed, including their role, methods of integrating them, examples of integrated aquatic and terrestrial models, and the past, present, and future of those models. Following that, is a description of one of the grand challenges of biogeochemical modeling, the scaling problem. Then, the transformation of arsenic by phytoplankton is presented as a case study.

Modeling Approaches

Conceptual and Descriptive Models

The simplest type of model, and therefore often a starting point in a biogeochemical modeling study, is a qualitative, conceptual or descriptive model. This type of model can be effectively communicated using diagrams, like that shown for mercury methylation in Fig. 1. The purpose of this model is to describe the requirements for mercury methylation, and the model is simplified for that purpose. Water, for example, is also part of the overall reaction, but it is omitted from the model because it typically does not affect the process. Also, no information on what occurs inside of the bacteria, the biochemical reaction(s), is included in the model. Despite (and maybe because of) the simplifications, the model conveys important aspects of the process for this particular

☆*Change History*: March 2018. J Mukherjee and S Ray updated (1) A section on Ecology and Biogeochemistry models has been added, (2) Two biogeochemical models on carbon and nitrogen cycle following process based dynamic modeling have been added as examples, (3) A paragraph on application of biogeochemical cycles added and (4). Relevant references has been added.

This is an update of F.L. Hellweger, Biogeochemical Models, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 386–396.

Fig. 1 Conceptual model of mercury methylation by sulfate-reducing bacteria.

Box 1 Narrative presentation of a model for dissimilatory Fe(III) reduction

The oxidation of detritus to CO_2 by dissimilatory Fe(III) reduction is a multistep process. First, the complex organic compound is hydrolyzed to smaller soluble compounds (e.g., amino acids, fatty acids). Then, those compounds are metabolized to acetate by fermentative microorganisms. Finally, the acetate is oxidized to CO_2 and the Fe(III) is reduced to Fe(II) by iron-reducing bacteria.

purpose. Another way to communicate a qualitative model is in the form of a narrative, as shown for dissimilatory iron reduction in Box 1.

Chemistry-Type Models

Qualitative models are useful, but often quantitative predictions are needed. One common approach to quantitative biogeochemical modeling is to apply the concepts of chemistry. That is, organisms are quantified as concentrations and their effect on chemistry is considered a reaction. There are a number of ways organisms can be incorporated into chemical models, as described in this section.

Organism Is Ignored

The simplest way to model the effect of organisms on chemistry is to simply ignore them, or to not explicitly recognize them or their action in the model. Nitrification, for example, is a two-step process carried out by nitrifying bacteria. First, *Nitrosomonas* converts ammonia (NH_3) to nitrite (NO_2^-) and then *Nitrobacter* converts nitrite to nitrate (NO_3^-):

$$2NH_3 + 4O_2 \rightarrow^{Nitrosomonas} 2NO_2^- + 2H^+ + 2H_2O$$
$$+O_2 \rightarrow^{Nitrobacter} 2NO_3^- + 2H^+ + 2H_2O$$

As shown in the equation, the reaction is mediated by and requires the two organisms. However, in the natural environment, nitrifying bacteria are often present in sufficient quantities, and therefore they are typically not included in rate expressions:

$$\frac{d[NH_3]}{dt} = -k[NH_3]$$

where k (day^{-1}) is the first-order reaction rate constant for the nitrification process. Sometimes the effect of O_2 is included in the rate expression by modifying the rate constant k as a function of $[O_2]$.

Organism Is Included as Reactant or Product

Another way to model the effect of biological organisms on chemistry is to consider them a chemical molecule that participates in a reaction as a reactant or product. Photosynthesis, for example, is often represented using the following reaction:

$$106CO_2 + 122H_2O + 16HNO_3 + H_3PO_4 \rightarrow \underbrace{(CH_2O)_{106}(NH_3)_{16}(H_3PO_4)}_{A} + 138O_2$$

where the molecule A is a simplified chemical-type representation of algae, also called Redfield molecule. This type of representation is useful for composition analysis. It says, for example, that the N:P ratio of phytoplankton is 16:1, which can be used to determine which one of these nutrients will run out first and end up limiting primary production.

Such organism molecules can also be included in kinetic rate expressions. For growth on a substrate or nutrient, it is often observed that the specific growth rate is proportional to the substrate concentration at low substrate concentrations, meaning the substrate is rate limiting. However, at higher substrate concentrations the growth rate is limited by other factors, like the rate of processing the substrate. This type of behavior can be simulated using the well-known Monod model:

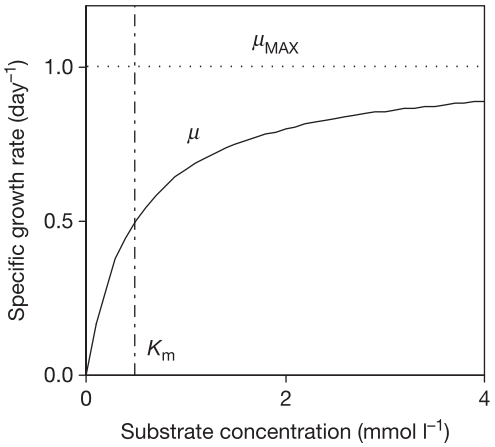


Fig. 2 Monod growth model. $\mu_{MAX} = 1.0\ \text{day}^{-1}$, $K_m = 0.5\ \text{mmol L}^{-1}$.

$$\frac{d[A]}{dt} = \mu_{MAX}\frac{[HNO_3]}{K_m + [HNO_3]}[A]$$

where $[A]$ (mmol L^{-1}) is the phytoplankton concentration, $\mu_{MAX}(day^{-1})$ is the maximum specific growth rate (when $[HNO_3] \gg K_m$), and K_m (mmol L^{-1}) is the half-saturation constant. The Monod model has a hyperbolic shape as illustrated in Fig. 2. The amount of nutrient consumed per biomass synthesized can be calculated using a yield coefficient

$$\frac{d[HNO_3]}{dt} = \frac{-1}{Y}\frac{d[A]}{dt}$$

where Y (1 mol A/16 mol HNO_3) is the yield coefficient.

Organism Is Included as Catalyst

Another method is to consider organisms as a mediator in a chemical reaction without being a reactant or product, which is called catalyst in chemistry and enzyme or biological catalyst in biology. The nitrifying bacteria discussed above can be considered catalysts for the nitrification reaction, although the rate expression does not explicitly recognize that. The general sequence of an enzyme-mediated reaction is

$$S + E \leftrightarrow SE \rightarrow P + E$$

First the substrate S combines with the enzyme E in a reversible reaction to form the complex SE. Then, the SE reacts to form one or more products P and E in an irreversible reaction. The mechanistic rate expression for enzyme kinetics is the Michaelis–Menten equation:

$$\frac{d[S]}{dt} = V_{MAX}\frac{[S]}{K_M + [S]}$$

where V_{MAX} (mmol L^{-1} day^{-1}) is the maximum reaction velocity, and K_M (mmol L^{-1}) is half-saturation constant. The parameters V_{MAX} and K_M are related to the enzyme concentration and the rate constants of the individual reactions. The Michaelis–Menten equation has the same hyperbolic shape as the Monod equation illustrated in Fig. 2.

Biology-Type Models

In some cases treating organisms as chemical molecules is overly simplistic and introduces excessive error into the model. Then, a more explicit representation of their effect on chemistry is needed. Organisms only directly affect the chemistry of their environment by removing (uptake) or adding (excretion) chemicals from or to their environment. These processes can be passive or active, as discussed in this section.

Passive Uptake and Excretion

Substances continuously diffuse in and out of organisms across the cell membrane, with a net transport in the direction of decreasing concentration. Movement through cell membranes can be complicated by multiple layers and binding sites, and therefore models often assume simple diffusion across one layer. For that case, the transport rate is proportional to the concentration gradient across the cell membrane:

$$V = P_m A(S_{IN} - S_{OUT})$$

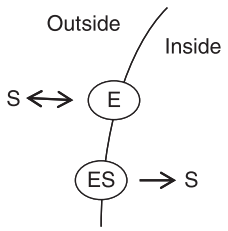


Fig. 3 Enzyme-mediated uptake.

where V (fmol cell^{-1} day^{-1}) is the transport rate, P_m (m day^{-1}) is the membrane permeability coefficient, A (m^2 cell^{-1}) is the cell surface area, and S_{IN} and S_{OUT} (fmol m^{-3}) are the chemical concentrations inside and outside of the cell, respectively. The intracellular concentration can be related to the cell quota (fmol cell^{-1}) using the cell volume (m^3 cell^{-1}). For chemicals that speciate, the model applies to the same species, and care should be taken when the speciation chemistry is significantly different inside and outside of the cell.

Active Uptake and Excretion

Organisms can also actively take up and excrete substances using transport sites (enzymes) on the cell membrane. This process can move chemicals against a concentration gradient, in which case it requires an external input of energy. The nutrient phosphate, for example, is taken up by an active uptake process. Cadmium is an example of a toxic chemical that is excreted from cells using an active excretion process. The sequence of events is similar to that of enzyme kinetics described above, and consists of a two-step process, illustrated in Fig. 3 for uptake. First, the chemical binds reversibly to the transport site. Then, it is transported into the cell in an irreversible reaction. This is typically modeled using the Michaelis–Menten equation:

$$V = V_{\mathrm{MAX}} \frac{S}{K_M + S}$$

where V (fmol cell^{-1} day^{-1}) is the transport rate, V_{MAX} (fmol cell^{-1} day^{-1}) is the maximum transport rate, S (mmol L^{-1}) is the chemical concentration, and K_M (mmol L^{-1}) is half-saturation constant. Active uptake systems are typically designed for and specific to a chemical. However, under certain circumstances other chemicals can be transported by the uptake system by mistake. The equation can be modified for the case where different substances (e.g., phosphate and arsenate) are taken up by the same transport system (competitive inhibition), and for the case where an internal compound slows the reaction (noncompetitive inhibition), using methods from enzymology. In addition, kinetic data sometimes reveal multiple uptake systems for one compound, and the ability of organisms to switch systems on/off.

Empirical Models

The chemistry- and biology-type models discussed above are based on theoretical or known relationships, like molecular diffusion in the case of passive excretion, and those models are therefore classified as mechanistic. Empirical modeling is an alternative approach, based entirely on data. That is, the model is constructed with the objective of reproducing an observed pattern, and little or no attention is paid to the mechanistic correctness of the model equation(s). If a mechanistic understanding of every detail of a process is known, there is really no reason to adopt the empirical approach. However, when a process is too complex and unknown, the empirical approach may be the only viable alternative. Unfortunately, this is the rule rather than the exception in biogeochemistry and many operational models therefore contain at least some empirical components. Consider, for example, the temperature dependence of the endogenous respiration rate of akinetes (resting stage cells) of the cyanobacterium *Anabaena circinalis* shown in Fig. 4. If we do not know the mechanism(s) responsible for the observed pattern, we may simply accept that fact and construct an equation that fits the data. The line in Fig. 4 corresponds to a simple two-part equation with a slope of 0.0055 below 30°C and -0.0035 above that. Although this modeling exercise did not further our mechanistic understanding of the endogenous respiration process, it did provide us with a simple means of predicting the rate, which may be useful and needed as a component of a larger mechanistic ecological model or for management purposes.

Since we have already admitted that our model has no mechanistic basis, the form of the equation (e.g., linear, exponential) is not important, and other empirical approaches are available (e.g., neural networks) that are not based on equations at all. However, it is generally accepted that models should be as simple and with as few parameters as possible. Also, since empirical modeling is based entirely on data, it is generally considered to be less transferable to conditions outside of those used to develop the model. So we should not apply our empirical model for akinete respiration to temperatures higher than 45°C, other experimental conditions (e.g., higher/lower nutrient concentration), or other species (e.g., *Anabaena flos-aquae*).

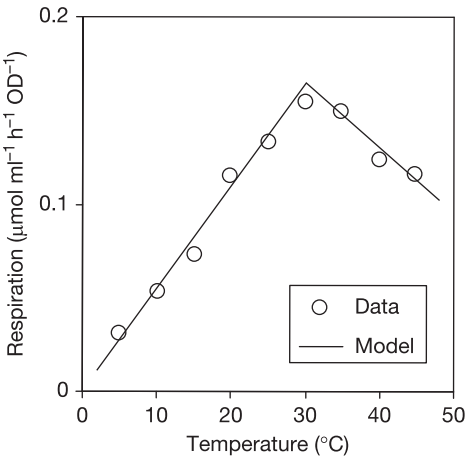


Fig. 4 Temperature dependence of the endogenous respiration rate of *Anabaena circinalis* akinetes. Data from Fay, P. (1988). Viability of akinetes of the planktonic cyanobacterium *Anabaena circinalis*. *Proceedings of the Royal Society of London B* **234**, 283–301.

Integrated Ecology and Biogeochemistry Models

Biogeochemical models are most useful when integrated into larger biogeochemical cycling models, which requires integration with physical and ecological models. Here, the integration with ecological models is discussed in some detail.

Interactions of benthic pelagic coupling are another facet of biogeochemical modeling. Processes and interactions between water column and sediment have long been neglected in the study of biogeochemical modeling. In last few decades, general models included the relationship of fluxes among particulate deposition and bottom-water composition to sediment–water exchange of oxygen and nutrients (e.g., Jahnke *et al.*, 1982; Rabouille and Gaillard, 1991; Ruardij and Van Raaphorst, 1995; Soetaert *et al.*, 1996a), alkalinity and total inorganic carbon (e.g., Boudreau, 1987; Archer, 1991; Hales *et al.*, 1994; Jahnke *et al.*, 1994, 1997; Cai *et al.*, 1995) and silica (e.g., Vanderborght *et al.*, 1977; Schink and Guinasso, 1980; Boudreau, 1990).

Generally, there should be no constraint to combine these models to water column biogeochemical models. However, research on deep sea sediment has got much attention and studies are restricted to steady state condition. A number of non-steady-state models have also become accessible. These models are capable to represent the dynamics of fluxes of sediment–water exchange and coupling to water column models (e.g., Rabouille and Gaillard, 1990; Sayles *et al.*, 1994; Boudreau, 1996; Soetaert *et al.*, 1996b; Jahnke, 1998; Wijsman *et al.*, 2002).

Role of Ecological Models in Biogeochemical Research

Ecological models play an important role in biogeochemical research, as illustrated in **Fig. 5**. Often, the starting point of biogeochemical research projects are field studies where certain spatial or temporal patterns are noticed. Those studies prompt controlled laboratory experiments that often more clearly demonstrate functional relationships and provide kinetic data on processes. Based on those studies, biogeochemical process models are developed. In this sequence of events, the laboratory experiments and process model development were motivated by the desire to understand the field data, and they do provide a qualitative understanding of the field data. However, a more quantitative understanding of the field data is often desired, which can be achieved by integrating the biogeochemical process model into a full ecological model and used it to simulate the field data. This, in turn, will identify further knowledge gaps and the cycle continues. Therefore, ecological models, and their integration with biogeochemical process models, are essential for biogeochemical research.

For example, the study of carbon and nitrogen cycle in the mangrove estuarine system (Mukherjee *et al.*, 2013; Mandal *et al.*, 2009) are illustrated here.

Both the studies follow the process-based numerical simulation modeling approach to understand the dynamics of the state variables and calibration and validation to the models have been applied using data collected from the field.

The transport of organic matter (OM) from terrestrial to aquatic system is a key link in the global biogeochemical cycle (Smith and Hollibaugh, 1993). One of the major transferable organic carbon reservoirs on the planet is dissolved organic carbon (DOC) pool in the oceans. Coastal waters and the wetlands can serve as source and sink of carbon, as well; thus, act as crucial link in the cyclic pathway. The rate and direction carbon flux among the carbon pools are also important (Mitsch and Gosselink, 2000). Particulate and dissolved organic matter (POM and DOM) from mangroves can provide energy and nutrients to the heterotrophic communities of adjoining estuarine and marine ecosystems (Odum and Heald, 1975). Through the process of photosynthesis carbon is trapped into the ecosystem. Eventually, following various biological and physical processes this carbon is transferred to the soil all the way through litter fall, root turnover or death of plants and hence, make up for the substrate for the formation of soil organic carbon (Kirschbaum, 2000).

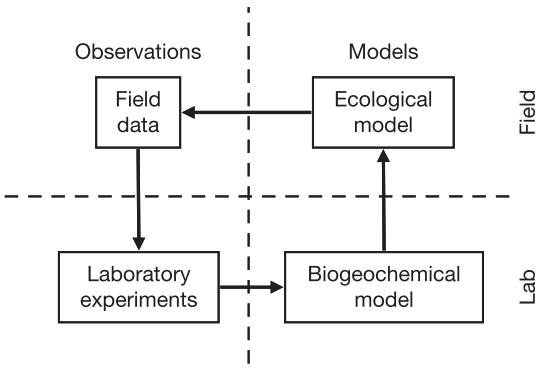


Fig. 5 Role of ecological models in biogeochemical research.

Ionized dissolved inorganic carbon (*DIC*) includes dissolved carbon dioxide (DCO_2), dissolved bicarbonte (*DBC)*. The organic pools take in dissolved organic carbon (*DOC*) and particulate organic carbon (*POC*) (Wetzel, 2001).

These models are the endeavors to assess the role of input of nitrogen and carbon from Sundarban mangrove litter to adjacent Hooghly estuary. For this purpose, two models were constructed to understand the dynamics of nitrogen and carbon up to the formation of dissolved inorganic carbon *(DIC)* and dissolved inorganic nitrogen (*DIN*) respectively from mangrove litterfall via the formation of detritus. The studies looked after following objectives:

(i) Experimentally determining different forms of nitrogen and carbon present in litter, soil and estuarine water.
(ii) Finding out the transformation rates from one form to another for nitrogen and carbon, mineralization rate of detritus and input rate of *DIN* and *DIC* to the adjacent estuary.
(iii) To construct the conceptual and quantitative models of nitrogen and carbon dynamics, that is, from litter nitrogen and litter carbon to dissolved inorganic nitrogen and dissolved organic carob in estuary via formation of detritus and its contribution to grazing food chain.
(iv) Assessment of important environmental factors presiding over litter nitrogen and carbon dynamics.
(v) Determination of sensitive parameters for nutrient dynamics.

Experiments have shown that the contribution of dissolved inorganic nutrients (nitrogen and carbon) to the Hooghly–Matla estuary from adjacent Sundarban mangrove forest is reliant on high litter production and breakdown of detritus. Litter degradation and subsequent uptake of nutrients by the autotrophs are the source of energy in this system. Benthic macrofauna, mainly gastropods, anaemons, bivalves, crabs, polychaetes, nemerteans, plays a noteworthy role by taking first step in degradation of dead organisms and also consumes about 61% and 39% of primary and secondary production respectively (Ghosh, 2001). Litter decomposition leads to the formation detritus which supports detritus food chain. Complete degradation and decomposition of mangrove litter contribute to the inorganic nutrient pool. This is vital to augment the growth of phytoplankton and other higher plants which in turn, enhance the production of zooplankton and other aquatic fauna of higher trophic levels in the estuary.

Decomposed dead, decaying parts of plants and animals enrich the *SOC* and *SON* pool. Increased temperature in premonsoon accelerates microbial activity which in turn results higher values of SIC and *SIN* in premonsoon and the opposite condition is seen in postmonsoon.

Extent of DCO_2 flux is chiefly regulated by two processes—photosynthetic uptake and respiratory release which are in turn, reliant upon various factors like allochthonous input of labile organic matter, water residence time, sunlight availability, rates of community metabolism, temperature and nutrient load (Ahad *et al.*, 2008; Smith and Hollibaugh, 1993). The pathways of the carbon and nitrogen cycle are shown in **Figs. 6** and **7**.

Flux, distribution and fate of dissolved organic and inorganic nutrients are pivotal because these nutrients take part in primary production and global geochemical cycling (Mukherjee *et al.*, 2013). Dissolved organic nutrients are significant machinery to provide nutrients for both bacterial and phytoplankton productions (Seitzinger and Sanders, 1999). It is found from this study that exudation, a portion of primary production of phytoplankton has a recognized role in *DOC* dynamics. Some studies have revealed a positive correlation between river discharge and *DOC* concentration and an increase in *DOC* due to leaching from soil and plant litter during periods of high discharge (Goni and Gardner, 2003). The study also shows that leaching rate of *SOC* is a sensitive parameter in this estuarine region. *DOC* derived from allochthonous input accounts for more than 90% of organic carbon in the water column and is a key substrate for metabolism in the system (Cole *et al.*, 2002). Biological activities thoroughly link *DOC* and *POC* pools. According to the present model, benthic detritivorous animals have important influence upon the composition and dynamics of *POC* pool. Benthic invertebrates residing in forest beds and estuary nourish upon the decaying leaves and facilitate in degradation process. Microorganisms colonize on the incompletely decomposing leaves and dead bodies of the soil and aquatic animals. Invertebrates and detritivores mostly regulate the *POC* pool by their activities. Recalcitrant portion of POC is assumed to export during tidal flush.

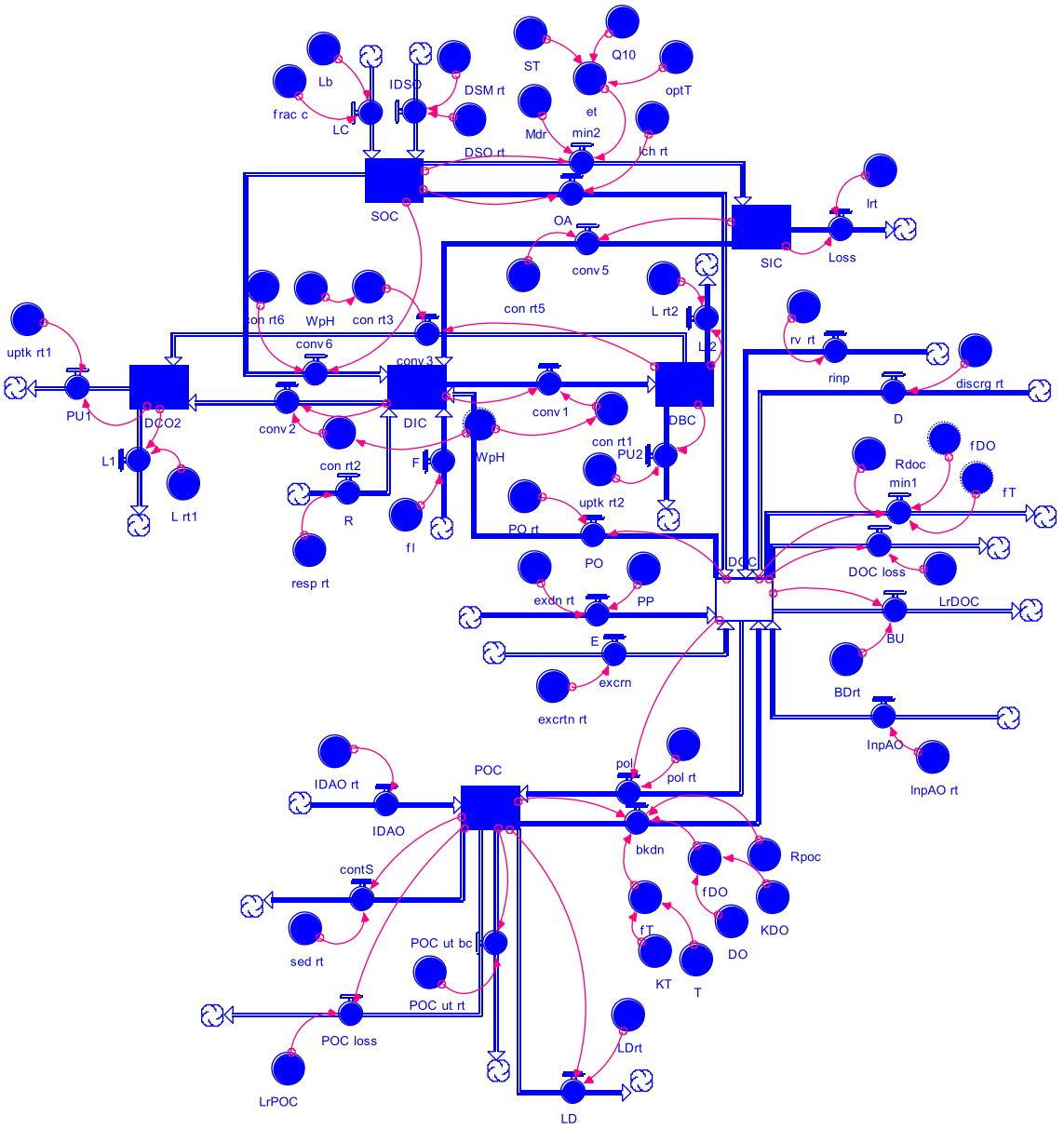


Fig. 6 Conceptual model of carbon dynamics of Hooghly–Matla estuarine system using STELLA 6.0 software. BDrt, bacterial DOC uptake rate; bkdn, breakdown of POC to DOC; BU, uptake of DOC by bacteria; con rt1, conversion rate of DIC to DBC; con rt2, conversion rate of DIC to DCO_2; con rt3, conversion rate of DBC to DCO_2; con rt5, conversion rate of SIC to DIC; con rt6, conversion rate of SOC to DCO_2; conv1, conversion of DIC to DBC; conv2, conversion of DIC to DBC; conv3, conversion of DBC to DCO_2; conv5, conversion of SIC to DIC; conv6, conversion of SOC to DCO_2; contS, contribution of carbon from POC pool to sediment; D, groundwater discharge of DOC; DBC, dissolved bicarbonate pool; DCO_2, dissolved free carbon dioxide pool; DIC, dissolved inorganic carbon pool; discrg rt., groundwater discharge rate of DOC; DO, dissolved oxygen; DOC, dissolved organic carbon pool; DOC loss, loss of DOC from the system; DSM rt., input rate of dead soil microflora to SOC pool; DSO rt., input rate of dead soil organisms to SOC pool; E, exudation of phytoplankton; excrn, excretory loss of aquatic organism that contribute to DOC pool; excrtn rt., excretion rate of aquatic organisms; exdn rt., exudation rate of phytoplankton; frac C, carbon fraction of litter; F, flux of carbon dioxide across air water interface; fl, flux of CO_2; IDAO, input from death of aquatic organism to POC; IDAO rt., rate of input from death of aquatic organism to POC; IDSO, input from death of soil organism; InpAO, input from aquatic organism to DOC; InpAO rt., rate of Input from aquatic organism to DOC; KDO, half-saturation constant for oxygen; KT, temperature coefficient; Lb, litter biomass; LC, litter carbon; lch rt., leaching rate of organic acid from soil; LD, POC loss due to detritivory; LD rt., rate of POC loss due to detritivory; Loss, loss of SIC from the system; LrDOC, loss rate of DOC from the system; LrPOC, loss rate of POC from the system; lrt, loss rate of SIC from the system; L rt1, loss rate of DCO_2 from the system; L rt2, loss rate of DBC from the system; mdr, mineralization rate of SOC to SIC; min1, mineralization of DOC; min2, mineralization of SOC to SIC; optT, optimal temperature in soil; OA, contribution of organic acid from SOC to DOC pool; PO, photo oxidation of DOC; PO rt., photo oxidation rate of DOC; POC loss, POC loss from the system; POC ut bc, utilization of POC by bacteria; pol, polymerization of DOC into POC; pol rt., polymerization rate of DOC to POC; PP, primary productivity; PU1, uptake of DCO_2 by phytoplankton; PU2, uptake of DBC by phytoplankton; Q10, temperature factor for soil mineralization; R, community respiration; Rdoc, mineralization rate for DOC; resp. rt., community respiration rate; rinp, riverine input of DOC; Rpoc, mineralization rate for POC to DOC; rv rt., rate of riverine input of DOC; sed rt., sedimentation rate of POC; SIC, soil inorganic carbon pool; SOC, soil organic carbon pool; ST, temperature of soil; T, temperature of water; uptk rt1, uptake rate of DCO_2 by phytoplankton; uptk rt2, uptake rate of DBC by phytoplankton; WpH, pH of water.

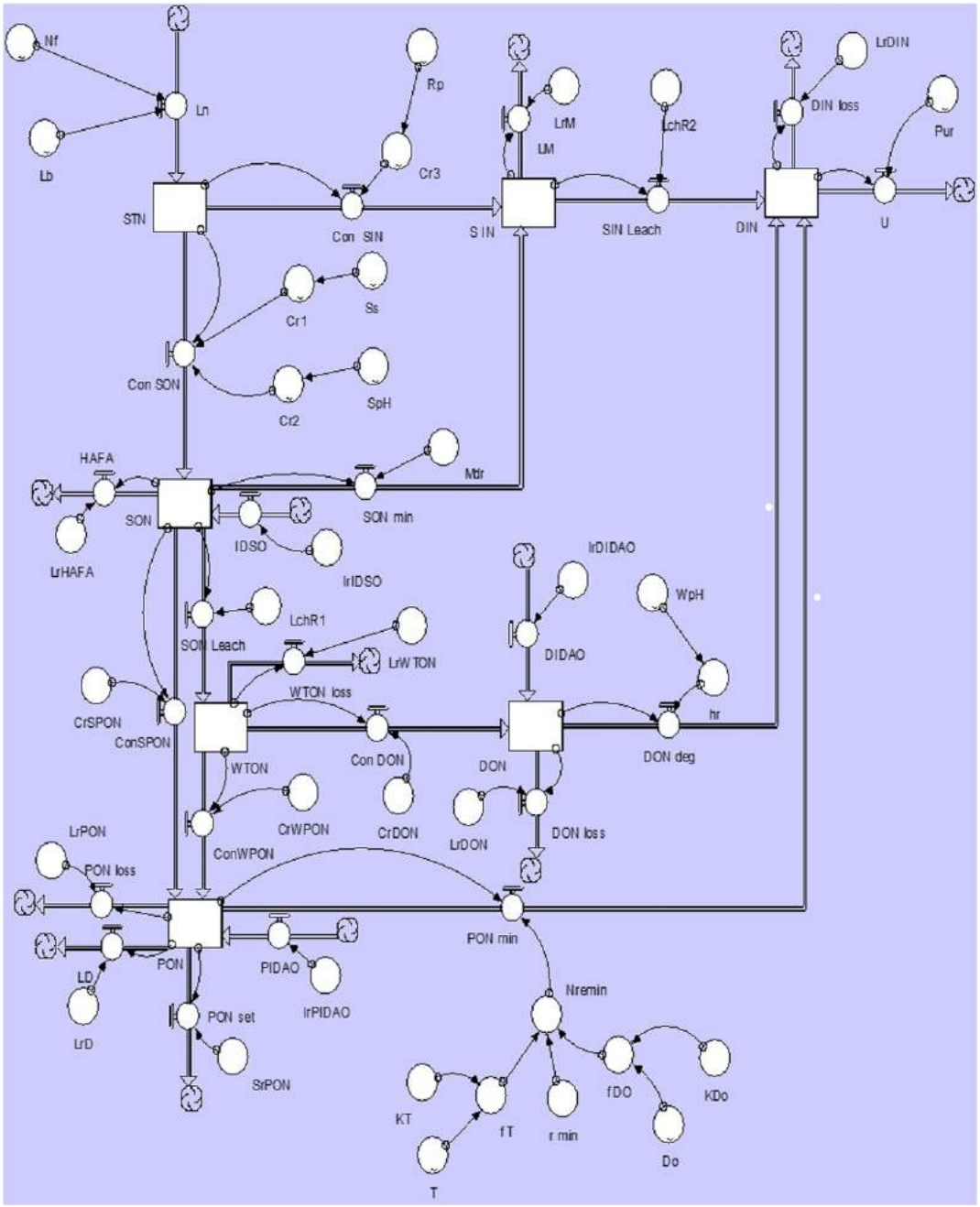


Fig. 7 Conceptual model of the contribution of dissolved inorganic nitrogen (*DIN*) from the litterfall of adjacent mangrove forest using STELLA 6.0 software.

In this estuarine system, the existence of *SIN* mainly depends upon temperature, redox potential of soil and microbial activity. *DON* undergoes degradation and forms *DIN* through pH dependent hydrolysis and the process is dependent upon water temperature and dissolved oxygen. Organic load from the *SON* pool enriches *DON* pool which is been utilized by bacteria. *DIN* is higher in monsoon and lower in premonsoon. In general, mineralization process is governed by dissolved oxygen and water temperature.

A huge data collected through the JGOFS and NASA ocean- color satellites programs are useful for testing and validating hypothesis integrated in model on ocean biogeochemistry. A coupled model of global ocean describing the general circulation, biogeochemistry and radiation of oceans were validated using the satellite data sets. In the model biogeochemical processes were determined through the dynamics of circulation and turbulence, availability of solar irradiance and the exchanges among multiple phytoplankton functional groups (diatoms, chlorophytes, cyanobacteria, and coccolithophores) and four nutrients (nitrate, ammonium, silica, and dissolved iron) in order to analyze its efficiency. In this way, shortcomings in our knowledge and assumptions could be identified, by taking into consideration the global ocean biogeochemical models in a holistic approach (Gregg *et al.*, 2003).

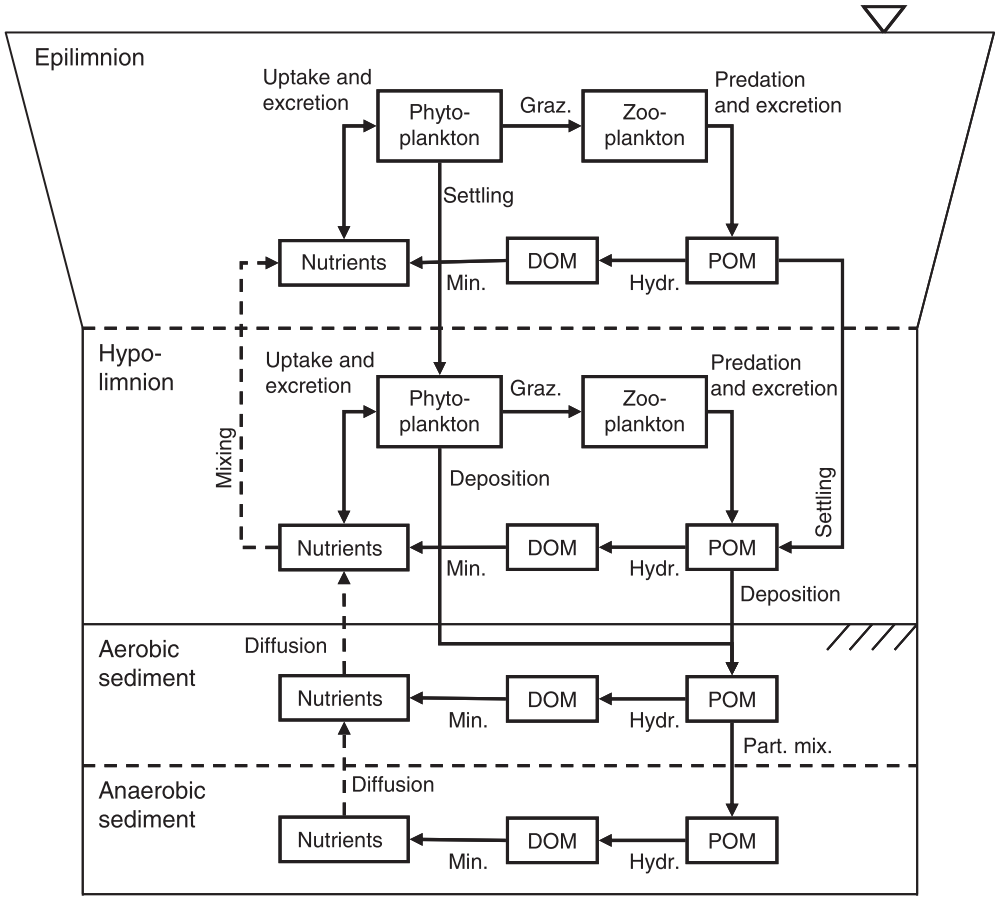


Fig. 8 Schematic of a typical integrated ecological and biogeochemical model. Some components (e.g., herbivorous and carnivorous zooplankton), reactions (e.g., DOM excretion), and transport pathways (e.g., phytoplankton mixing) are omitted from this illustration for simplicity. POM, particulate organic matter; DOM, dissolved organic matter; Graz., grazing; Min., mineralization; Hydr., hydrolysis; Part. Mix., particle mixing.

Integrating Ecological and Biogeochemical Models

The degree of complexity to be incorporated into the ecosystem model serves as equilibrium between superfluous detail and unjustified simplification (Flynn, 2001). Now, there lies serious need to recognize ecosystem model structures and formulations that are geographically transferable, that is, are pertinent over a number of varied ecosystems. There remain a trade-off between the complexity and realism of a model and to the extent to which it can be controlled with the given data (Friedrichs *et al.*, 2007).

The integration of ecological and biogeochemical models can be challenging, because they typically use different modeling approaches. Although ecological models have traditionally used a population-level approach, individual-based approaches are becoming more common. In individual-based models (IBMs) (also called agent-based models, ABMs), the individual members of the population are simulated separately. Each wolf or moose is an independent entity, that moves, eats, reproduces, dies, etc., and a population-level behavior emerges as a result of the action of individuals. This is in contrast to population-level models that modify population-level properties, like the total number of wolfs and moose, directly. Chemical and biogeochemical models typically use a population-level modeling approach, although there is now a movement of individual-based modeling for microorganisms, like algae and bacteria.

Example 1: Aquatic Environment

An important practical management problem that has led modelers to construct linked ecological–biogeochemical models is cultural eutrophication, the excessive growth of algae due to anthropogenic input of nutrients. The cause of eutrophication is typically increased levels of nutrients (phosphorus, nitrogen), which are chemical quantities—a chemical modeling problem. However, the problem manifests itself by increased number of algae, the dynamics of which are often controlled by zooplankton—an ecological modeling problem. Models that address this problem have traditionally been constructed by extending the chemistry concepts to phytoplankton and zooplankton. That is, the algae are quantified as concentration and their growth is conceptualized as a chemical reaction between them and the nutrients. A typical flow diagram for a lake eutrophication model is presented in Fig. 8, which shows

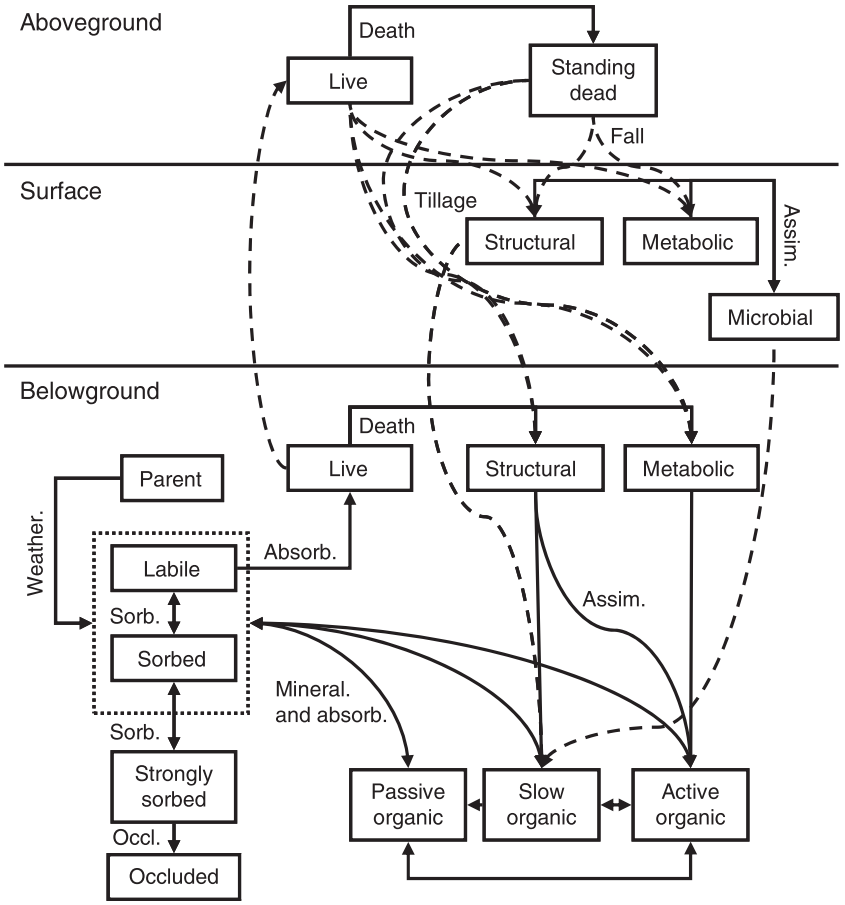


Fig. 9 Schematic of the CENTURY integrated ecological and biogeochemical model for the terrestrial environment. Some transport pathways (e.g., leaching, harvesting) are omitted for clarity.

the state variables (boxes) and processes (arrows) in the various spatial compartments. The spatial segmentation consists of two layers in the water column and two layers in the sediment bed. In the water column, the surface layer (epilimnion) is separated from the bottom layer (hypolimnion) by the seasonal thermocline. In the sediment bed the aerobic layer is separated from the anaerobic layer by the depth of oxygen penetration. The model accounts for nutrients, phytoplankton, zooplankton, particulate organic matter (POM), and dissolved organic matter (DOM). The number of state variables is typically larger than the number of boxes shown in Fig. 8. That is because the models typically track various elements (C, N, P, and Si) individually in each of the components. For P, for example, the state variables can be the concentrations of PO_4 (nutrient), phytoplankton P, zooplankton P, particulate organic P (POM), and dissolved organic P (DOM). Dissolved oxygen and sulfide are also often simulated. Important transport pathways include phytoplankton and POM settling from the epilimnion to the hypolimnion where they decay to DOM and then nutrients. The nutrients are mixed back into the epilimnion when the lake overturns. Phytoplankton and POM also deposit to the sediment bed, where they decay to DOM and nutrients, which diffuse out of the sediment. Important reactions include phytoplankton uptake and excretion of nutrients. The phytoplankton are grazed by zooplankton, which results in the phytoplankton biomass being assimilated by the zooplankton or excreted as detritus (POM). Zooplankton die by predation from higher organisms, which also produces POM. POM hydrolyzes to DOM, which mineralizes to nutrients.

Note that this model includes a zooplankton state variable, which is unusual. Most operational eutrophication models do not explicitly consider zooplankton, but rather implicitly include their effect on the algae by assigning a seasonally varying grazing rate. This is due to the functional complexity of zooplankton, which can, for example, enter stages of diapause or dormancy at various stages of their life cycle that can last from a month to over a decade.

Example 2: Terrestrial/Soil Environment

Linked biogeochemical and ecological models also exist for the terrestrial environment. The development and application of those models is motivated by the desire to understand how terrestrial ecosystems respond to changes in management (e.g., crop rotation, fertilization) and/or climate (e.g., increased CO_2, temperature). A typical P flow diagram for a terrestrial model is presented in Fig. 9. Live plant P is divided into above and below (i.e., roots) ground pools. Upon death above ground, P is moved to a standing dead pool, which can fall to become surface litter in structural or metabolic pools (different decay rates). A surface microbial pool

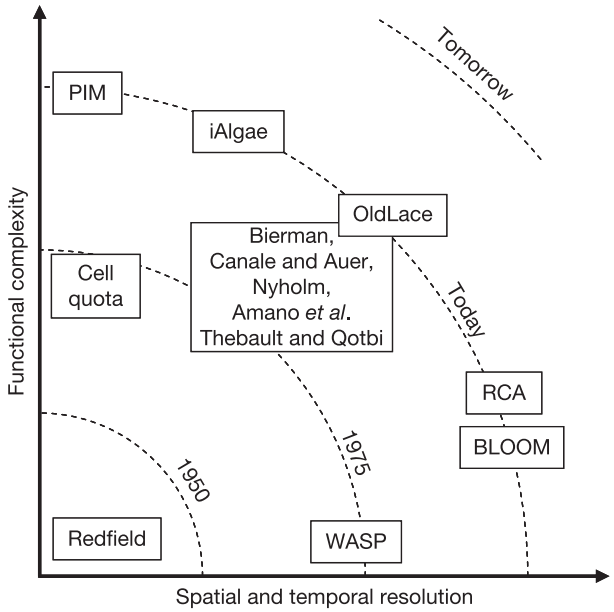


Fig. 10 Resolution and complexity of integrated ecological and biogeochemical models. For simplicity only selected models are included. More detailed information on the specific models is included in the references listed in the References and Further reading section.

is associated with the surface litter. Upon death below ground, P is moved to belowground, structural or metabolic pools. Other soil organic matter is divided into active, slow, and passive pools with different decomposition rates.

Various inorganic forms of P are simulated. Labile P is in equilibrium with sorbed P, and sorbed P is in equilibrium with strongly sorbed P, which in turn is lost to occluded P. Weathering of parent material P (e.g., apatite) results in labile and sorbed P.

Consider, for example, a potential history of a P molecule leached from the parent material. It enters the labile/sorbed pool, where it may become absorbed by the roots of a tree, transported above ground and incorporated into a branch. When the tree dies, it may stand for some time (standing dead), but eventually it will fall and the P molecule may become part of the surface structural pool. If it is not decomposed by surface microbes, it will become part of the belowground, slow organic pool. Below ground it may be absorbed by microbes and enter the active organic pool, which may die and release the molecule back to the labile/sorbed pool.

Past, Present, and Future of Integrated Ecological and Biogeochemical Modeling

This section reviews the past, present, and future of integrated ecological and biogeochemical models. The Redfield relation introduced above can be considered to be the simplest integrated ecological/biogeochemical model. Following the work of Redfield, significant improvements were made to this model in basically two dimensions: (1) spatial and temporal resolution and (2) functional complexity, as illustrated in **Fig. 10**. A similar perspective on the advancement of ecological modeling in Saginaw Bay is shared by V. Bierman:

> Model development is proceeding along two parallel pathways. The first of these involves the development of research-oriented process models, which include biological and chemical detail but which, for simplicity, do not include any spatial detail. The second pathway involves the development of an engineering-oriented water quality model that mimics, as closely as practicable, the actual physical system, including spatial detail. At any given point in time, the water quality model will contain those chemical and biological processes that have previously been investigated and developed using the spatially-simplified model. There is constant feedback between these two pathways and constant interaction between the entire modeling effort and an ongoing sampling effort on Saginaw Bay.

Application of biogeochemical models has a reasonably wide range to solve various problems (Jørgensen and Bendoricchio, 2001); such as, optimization of biological treatment plant (Snape *et al.*, 1995), to recover ground water contamination (covered in National Research Council, 1990), to sort out acidification problem (Alcamo *et al.*, 1990), Forest growth and yield (Vanclay, 1994), air pollution (Gryning and Batchvarova, 2000; Baldasano *et al.*, 1994), to optimize agriculture (France and Thornley, 1984).

It is useful to mention a number of models in **Fig. 10**, because they constitute significant milestones. The introduction of the WASP (Water Quality Analysis and Simulation Program) model represented a significant improvement in model resolution. Further advances in resolution, and the state of the science of today, are represented by the RCA (Row-Column Aesop) and BLOOM models. These models are typically set up with a high spatial and temporal resolution (thousands of

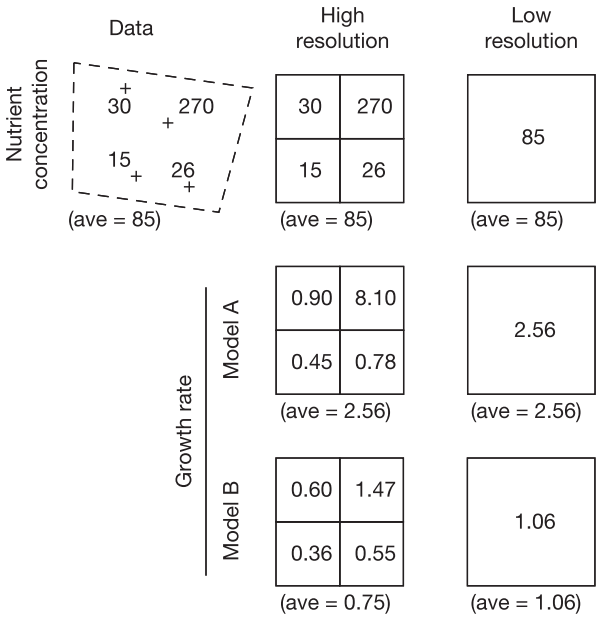


Fig. 11 Illustration of scaling problem. Nutrient concentration, S (nmol L^{-1}); specific growth rate, μ (day^{-1}); model A: $\mu = \mu_{MAX}S/K_m$; model B: $\mu = \mu_{MAX}S/(K_m + S)$; maximum specific growth rate, $\mu_{MAX} = 1.8\ day^{-1}$; half-saturation constant, $K_m = 60$ nmol L^{-1}.

mass balance compartments). Functional model complexity varies mainly in the way nutrient uptake and cell composition are simulated. At the lowest level, there is net uptake (uptake–excretion) of nutrients and a fixed "Redfield" cell composition. Departures from this model were motivated by the realization that phytoplankton composition is variable, and as a result various "variable stoichiometry" or "variable composition" models were developed. At an intermediate level of complexity one variable (e.g., total algal P) is used to describe the composition of the algae. Those models are commonly called "cell quota" models, where the cell quota is the mass of nutrient per cell. At a high level of complexity, models explicitly account for uptake and excretion and various species of the nutrients (e.g., PO_4, polyphosphates, etc.) and reactions in the algal cells. The phosphate interaction model (PIM), for example, has three state variables for intracellular P (soluble inorganic P, polyphosphate, structural and soluble organic P). Variable phytoplankton composition models have been integrated with spatially and temporally explicit models, as exemplified by the Bierman and other similar models. Increasing the spatial and temporal resolution and functional complexity of models can be problematic and for that reason, IBMs, like iAlgae, are being constructed as a potential alternative to the traditional population-level models. Individual-based modeling of algae and bacteria is a current research topic and will likely be a significant factor in future modeling in the area of integrated ecological and biogeochemical modeling.

Grand Challenge: Saling Problem

An important problem in ecological and biogeochemical modeling is related to scale. To illustrate this "scaling problem," consider the illustrative case of nutrient-limited grass growth on a field. The soil nutrient concentration was measured at four locations, roughly equally spaced and representative of an equally large portion of the field, as shown in **Fig. 11** (top left). Two different models are used to estimate the growth rate. Model A is linear and has the form $\mu = \mu_{MAX}S/K_m$, and model B is the nonlinear Monod equation introduced above. Both models are applied at two different scales or resolutions. The high-resolution application (middle column) has four segments each corresponding to one measurement, and the low-resolution application (right column) has one segment with a nutrient concentration equal to the average of the four measurements. When model A is applied, the growth rates are different, but average out to the same value, regardless of the resolution of the model (middle row). This is an important point: linear models are scale insensitive and the resolution can be chosen freely to suit other needs, like the availability of input data, desired resolution of output data, computing resources, etc. The output from model B for the low- and high-resolution applications does not average out to the same value (bottom row). That is because nonlinear models are scale sensitive. Applying them to the same data at different resolutions will produce different results, and the model resolution can not be chosen freely. The underlying mathematical theory is known as "Jensen's inequality."

In future, the scaling problem will become increasingly important as (1) the spatial and temporal resolution of data increases as a result of advances in in situ and remote-sensing technology, and (2) process models are becoming increasingly functionally complex and nonlinear. The heterogeneity of terrestrial and soil environments is well known, but increasingly recognized as

Box 2 Descriptive model of arsenic transformation by phytoplankton

The transformation of arsenic by phytoplankton is linked to the uptake of phosphate. Algae actively take up As(V) ($AsO(OH)_3$) because they cannot differentiate it from phosphate ($PO(OH)_3$). However, because As(V) is toxic, the algae has to detoxify it, which is done by reduction to As(III), methylation to MMA and DMA, and excretion. The end product of the overall transformation reaction is a function of the phosphorus nutrient status of the algae. Under P-limited conditions the algae take up As(V), reduce it to As(III), methylate it to MMA and DMA, and then excrete it as DMA. Under P-replete conditions the algae upregulate their phosphate transport system (luxury uptake), and since As(V) is taken up by the phosphate transport system, it is also taken up at higher rates. The reduction to As(III) is fast, but the methylation is slower, causing As(III) to build up in the cell and be excreted into the medium.

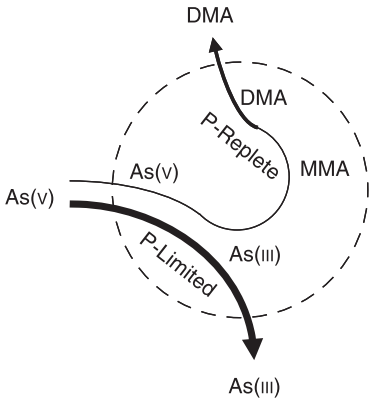


Fig. 12 Conceptual model of arsenic transformation by phytoplankton.

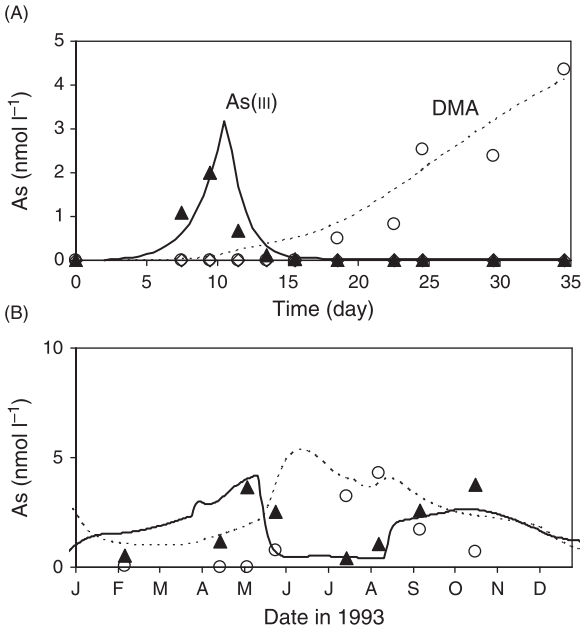


Fig. 13 Simulated and observed transformation of arsenic by phytoplankton in (A) a laboratory batch experiment and (B) Lake Biwa, Japan.

important in the aquatic environment. Two-dimensional imaging fluorometers capture spatial phytoplankton variability of almost an order of magnitude at subcentimeter scale. Moored sensors record fluctuations in phosphate concentration of over an order of magnitude within the course of a day. Phytoplankton models simulate intracellular speciation and transformation among multiple nitrogen species. Models consider luxury uptake of nutrients and trace elements, their intracellular transformation and

excretion. These two trends in spatial and temporal resolution and functional complexity are on a direct "collision course" with the scaling problem. This is one of the grand challenges of integrated ecological and biogeochemical modeling.

Case Study: Arsenic Transformation by Phytoplankton

Recent models (Aumont and Bopp, 2006; Le Quéré *et al.*, 2005; Moore *et al.*, 2004) make out between different plankton functional types (silicifiers, N-fixing plankton, calcifiers, zooplankton of different size-classes) and nutrient-limited conditions (Fe).

Transformation of arsenic by phytoplankton constitutes an interesting case study of coupled ecological and biogeochemical modeling. Arsenic can exist in a number of species, including arsenate (As(V)), arsenite (As(III)), methylarsonate (MMA) and dimethylarsinate (DMA). Under oxygenated conditions As(V) is the only thermodynamically stable form, and the other species spontaneously demethylate and oxidize to As(V). However, early field observations in the Pacific Ocean revealed that, although As (V) is the predominant form, other species are present at relatively high concentrations, meaning that there has to be a process continuously producing them. Algae were identified as being responsible for the transformation. In addition, field and laboratory data revealed that the end product of the transformation reaction varies and is a function of the growth rate and/or nutrient status of the algae. A model was proposed for the transformation of arsenic by phytoplankton, which is presented in Box 2 and **Fig. 12**.

A quantitative biogeochemical model has been developed for transformation of arsenic by phytoplankton using the concepts presented above. The model simulates uptake using the Michaelis–Menten equation modified for competitive inhibition and upregulation (luxury uptake). The model was calibrated to laboratory data. The results, presented in **Fig. 13**A, illustrate that the model captures the major temporal patterns in the data, including the production of As(III) early and DMA later in the experiment. Then, the model was integrated with an ecological model and used to simulate arsenic speciation in a lake. The results, presented in **Fig. 13**B, illustrate that the model captures much of the major temporal patterns in the field data, including spring and fall increases of As(III) and higher DMA in the summer.

See also: Ecological Data Analysis and Modelling: Carbon Biogeochemical Cycle and Consequences of Climate Changes. General Ecology: Ecological Stoichiometry: Overview. Global Change Ecology: Biogeocoenosis as an Elementary Unit of Biogeochemical Work in the Biosphere; Material and Metal Ecology; Microbial Cycles; Sulfur Cycle; Nitrogen Cycle; Oxygen Cycle; Phosphorus Cycle

References

Ahad, J.M.E., Barth, J.A.C., Ganeshram, R.S., Spencer, R.G.M., Uher, G., 2008. Controls on carbon cycling in two contrasting temperate zone estuaries: The Tyne and Tweed, UK. Estuarine, Coastal and Shelf Science 78, 685–693.

Alcamo, J., Shaw, R., Hordijk, L. (Eds.), 1990. The rains model of acidification. Dordrecht, The Netherlands: Kluwer, p. 366. IIASA, Laxenburg, Austria.

Archer, D., 1991. Modeling the calcite lysocline. Journal of Geophysical Research 96, 17037–17050.

Aumont, O., Bopp, L., 2006. Globalizing results from ocean in situ iron fertilization studies. Global Biogeochemical Cycles 20.GB2017. https://doi.org/10.1029/2005GB002591.

Baldasano, J.M., Brebbia, C.A., Power, H., Zannetti, P., 1994. Computer simulation air pollution II. vols. 1 and 2. Southampton and Boston: Computational Mechanics Publications, 588 pp + 560 pp.

Boudreau, B.P., 1987. A steady-state diagenetic model for dissolved carbonate species and pH in the pore waters of oxic and suboxic sediments. Geochimica et Cosmochimica Acta 51, 1985–1996.

Boudreau, B.P., 1990. Modeling early diagenesis of silica in non-mixed sediments. Deep Sea Research 37, 1543–1567.

Boudreau, B.P., 1996. A method-of-lines code for carbon and nutrient diagenesis in aquatic sediments. Computational Geosciences 22, 479–496.

Cai, W.J., Reimers, C.E., Shaw, T., 1995. Microelectrode studies of organic carbon degradation and calcite dissolution at a California continental rise site. Geochimica et Cosmochimica Acta 59, 497–511.

Cole, J.J., Carpenter, S.R., Kitchell, J.F., Pace, M.L., 2002. Pathways of organic carbon utilization in small lakes: Results from a whole-lake 13C addition and coupled model. Limnology and Oceanography 47, 1664–1675.

Flynn, K.J., 2001. A mechanistic model for describing dynamic multi-nutrient, light, temperature interactions in phytoplankton. Journal of Plankton Research 23 (9), 977–997.

France, J., Thornley, J.H.M., 1984. Mathematical models in agriculture. Butterworths. p. 333.

Friedrichs, M.A., Dusenberry, J.A., Anderson, L.A., Armstrong, R.A., Chai, F., Christian, J.R., Doney, S.C., Dunne, J., Fujii, M., Hood, R., McGillicuddy, D.J., 2007. Assessment of skill and portability in regional marine biogeochemical models: Role of multiple planktonic groups. Journal of Geophysical Research, Oceans 112 (C8),

Ghosh, P.B., 2001. Role of macrofauna in energy partitioning and nutrient recycling in a tidal creek of Sunderbans mangrove forest, India. In: Kumar, A. (Ed.), Ecology and ethology of aquatic biota. New Delhi: Daya Publishing House, pp. 90–97.

Goni, M.A., Gardner, L.R., 2003. Seasonal dynamics in dissolved organic carbon concentrations in a coastal water-table aquifer at the forest-marsh interface. Aquatic Geochemistry 9, 209–232.

Gregg, W.W., Ginoux, P., Schopf, P.S., Casey, N.W., 2003. Phytoplankton and iron: Validation of a global three-dimensional ocean biogeochemical model. Deep Sea Research Part II: Topical Studies in Oceanography 50 (22), 3143–3169.

Gryning, S.E., Batchvarova, E., 2000. Air pollution modeling and its applications XIII. In: Proceedings from a conference held in Varna, Bulgaria, 1998. Dordrecht: Kluwer Academic, p. 810.

Hales, B., Emerson, S., Archer, D.E., 1994. Respiration and dissolution in the sediments of the western North Atlantic: Estimates from models of in situ pore water oxygen and pH. Deep Sea Research 41, 413–436.

Jahnke, R.A., 1998. Geochemical impacts of waste disposal on the abyssal seafloor. Journal of Marine Systems 14, 355–375.

Jahnke, R.A., Emerson, S.R., Murray, J.W., 1982. A model of oxygen reduction, denitrification, and organic matter mineralization in marine sediments. Limnology and Oceanography 27 (4), 610–623.

Jahnke, R.A., Craven, D.B., Gaillard, J.F., 1994. The influence of organic matter diagenesis on $CaCO_3$ dissolution at the deep-sea floor. Geochimica et Cosmochimica Acta 58, 2799–2809.
Jahnke, R.A., Craven, D.B., McCorkle, D.C., Reimers, C.E., 1997. CaCO3 dissolution in California continental margin sediments: The influence of organic matter remineralization. Geochimica et Cosmochimica Acta 61, 3587–3604.
Jørgensen, S.E., Bendoricchio, G., 2001. Fundamentals of ecological modeling, 3rd edn. Oxford: Elsevier.
Kirschbaum, M.U.F., 2000. Will changes in soil organic carbon act as a positive or negative feedback on global warming? Biogeochemistry 48, 21–51.
Le Quéré, C., Harrison, S.P., Prentice, I.C., Buitenhuis, E.T., *et al*., 2005. Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. Global Change Biology 11, 2016–2040.
Mandal, S., Ray, S., Ghosh, P.B., 2009. Modelling of the contribution of dissolved inorganic nitrogen (DIN) from litterfall of adjacent mangrove forest to Hooghly–Matla estuary, India. Ecological Modelling 220, 2988–3000.
Mitsch, W.J., Gosselink, J.G., 2000. The value of wetlands: Importance of scale and landscape setting. Ecological Economics 35, 25–33.
Moore, K.J., Doney, S.C., Lindsay, K., 2004. Upper ocean ecosystem dynamics and iron cycling in a global threedimensional model. Global Biogeochemical Cycles 18.GB4028. https://doi.org/10.1029/2004GB002220.
Mukherjee, J., Ray, S., Ghosh, P.B., 2013. A system dynamic modeling of carbon cycle from mangrove litter to the adjacent Hooghly estuary, India. Ecological Modelling 252, 185–195.
National Research Council, 1990. Ground water models, scientific and regulatory applications. Washington, D.C.: National Academy Press, p. 303.
Odum, W.E., Heald, E.J., 1975. The detritus based foodweb of an estuarine mangrove community. In: Cronin, L.E. (Ed.), Estuarine research. New York, USA: Academic Press, pp. 265–286.
Rabouille, C., Gaillard, J.-F., 1990. The validity of steady-state flux calculations in early diagenesis: A computer simulation of deep-sea silica diagenesis. Deep Sea Research 37, 625–646.
Rabouille, C., Gaillard, J.F., 1991. Towards the EDGE: Early diagenetic global explanation. A model depicting the early diagenesis of organic matter, O_2, NO_3, Mn and PO_4. Geochimica et Cosmochimica Acta 55, 2511–2525.
Ruardij, P., Van Raaphorst, W., 1995. Benthic nutrient regeneration in the ERSEM ecosystem model of the North Sea. Netherlands Journal of Sea Research 33 (3), 453–483.
Sayles, F.L., Martin, W.R., Deuser, W.G., 1994. Response of benthic oxygen demand to particulate organic carbon supply in the deep sea near Bermuda. Nature 371, 686–689.
Schink, D.R., Guinasso, N.L., 1980. Processes affecting silica at the abyssal sediment–water interface. In: Biogéochimie de la Matière Organique à l'Interface Eau-Sédiment Marin. Colloques Internationaux du CNRS No. 293., pp. 81–92.
Seitzinger, S.P., Sanders, R.W., 1999. Atmospheric inputs of dissolved organic nitrogen stimulate estuarine bacteria and phytoplankton. Limnology and Oceanography 44, 721–730.
Smith, S.V., Hollibaugh, J.T., 1993. Coastal metabolism and the oceanic organic carbon cycle. Reviews of Geophysics 31, 75–89.
Snape, J.B., Dunn, I.J., Ingham, J., Presnosil, J.E., 1995. Dynamics of environmental bioprocesses. New York and Basel: VCH. Weinheim, p. 492.
Soetaert, K., Herman, P.M.J., Middelburg, J.J., 1996a. A model of early diagenetic processes from the shelf to abyssal depths. Geochimica et Cosmochimica Acta 60 (6), 1019–1040.
Soetaert, K., Herman, P.M.J., Middelburg, J.J., 1996b. Dynamic response of deep-sea sediments to seasonal variations: A model. Limnology and Oceanography 41 (8), 1651–1668.
Vanclay, J.K., 1994. Modelling forest growth and yield. Wallingford: Cab International, p. 312.
Vanderborght, J.P., Wollast, R., Billen, G., 1977. Kinetic models of diagenesis in disturbed sediments: 1. Mass transfer properties and silica diagenesis. Limnology and Oceanography 22, 787–793.
Wetzel, R.G., 2001. Limnology: Lake and river ecosystems, 3rd ed. San Diego, Calofornia: Academic Press.
Wijsman, J.W.M., Herman, P.M.J., Middelburg, J.J., Soetaert, K., 2002. A model for early diagenetic processes in sediments of the continental shelf of the Black Sea. Estuarine, Coastal and Shelf Science 54 (3), 403–421.

Further Reading

Andreae, M.O., 1979. Arsenic speciation in seawater and interstitial waters: The influence of biological–chemical interactions on the chemistry of a trace element. Limnology and Oceanography 24, 440–452.
Bashkin, V.N., 2003. Modern biogeochemistry. Dordrecht: Springer.
Bierman Jr., V.J., 1976. Mathematical model of the selective enhancement of blue-green algae by nutrient enrichment. In: Canale, R.P. (Ed.), Modeling biochemical processes in aquatic ecosystems. Ann Arbor: Ann Arbor Science, pp. 1–29.
Blasco, F., Weill, A., 1999. Advances in environmental and ecological modeling. Oxford: Elsevier.
Chapra, S.C., 1997. Surface water-quality modeling. Boston: McGraw-Hill.
Di Toro, D.M., 2001. Sediment flux modeling. New York: Wiley-Interscience.
Droop, M.R., 1968. Vitamin B12 and marine ecology. IV. The kinetics of uptake, growth, and inhibition in *Monochrysis lutheri*. Journal of the Marine Biological Association of the UK 48, 689–733.
Fay, P., 1988. Viability of akinetes of the planktonic cyanobacterium *Anabaena circinalis*. Proceedings of the Royal Society of London B 234, 283–301.
Fenchel, T., King, G., Blackburn, H., 1998. Bacterial biogeochemistry—The ecophysiology of mineral cycling. London: Academic Press.
Franks, P.J.S., Jaffe, J.S., 2001. Microscale distributions of phytoplankton: Initial results from a two-dimensional imaging fluorometer, OSST. Marine Ecology-Progress Series 220, 59–72.
Hasegawa, H., Sohrin, Y., Seki, K., *et al*., 2001. Biosynthesis and release of methylarsenic compounds during the growth of freshwater algae. Chemosphere 43, 265–272.
Hellweger, F.L., Lall, U., 2004. Modeling the effect of algal dynamics on arsenic speciation in Lake Biwa. Environmental Science and Technology 38, 6716–6723.
Hellweger, F.L., Farley, K.J., Lall, U., Di Toro, D.M., 2003. Greedy algae reduce arsenate. Limnology and Oceanography 48, 2275–2288.
HydroQual, 2001. HydroQual addendum to: Bays eutrophication model (BEM): Modeling analysis for the period. NJ: HydroQual Mahwah, pp. 1992–1994.
Johnson, L., 2002. 2002 annual report. Chemical sensor program. Moss Landing, CA: Monterey Bay Aquarium Research Institute (MBARI).
Jørgensen, S.E., 2007. Ecological modeling—International Journal on Ecological Modelling and Systems Ecology. Oxford: Elsevier.
Lajtha, K., 2007. Biogeochemistry—An International Journal. Dordrecht: Springer.
Lovely, D.R., 1991. Dissimilatory Fe(III) and Mn(IV) reduction. Microbiological Reviews 55 (2), 259–287.
Macalady, J.L., Mack, E.E., Nelson, D.C., Scow, K.M., 2000. Sediment microbial community structure and mercury methylation in mercury-polluted clear lake. *California*. Applied and Environmental Microbiology 66 (4), 1479–1488.
Metherell, A.K., Harding, L.A., Cole, C.V., Parton, W.J., 1993. CENTURY Soil organic matter model environment. Technical documentation. Agroecosystem version 4.0. In: Great Plains System Research Unit Technical Report No. 4. Fort Collins, CO: USDA-ARS.
Schlesinger, W.H., 2005. Biogeochemistry. Oxford: Elsevier.
Schnoor, J.L., 1996. Environmental modeling—Fate and transport of pollutants in water, air and soil. New York: Wiley.

Schulze, E.-D., Heimann, M., Harrison, S., *et al*., 2001. Global biogeochemical cycles in the climate system. London: Academic Press.
Silver, S., Phung, L.T., 2005. A bacterial view of the periodic table: Genes and proteins for toxic inorganic ions. Journal of Industrial Microbiology and Biotechnology 32 (11 – 12), 587–605.
Sohrin, Y., Matsui, M., Kawashima, M., Hojo, M., Hasegawa, H., 1997. Arsenic biogeochemistry affected by eutrophication in Lake Biwa, Japan. Environmental Science and Technology 31, 2712–2720.
Thomann, R.V., Mueller, J.A., 1987. Principles of surface water quality modeling and control. New York: Harper Collins.
Di Toro, D.M., Fitzpatrick, J.J., Thomann, R.V., 1981. *Water Quality Analysis Simulation Program (WASP) and Model Verification Program (MVP)—Documentation*, for US EPA, Duluth, MN, Contract No. 68-01-3872. Westwood, NY: Hydroscience.

Carbon Biogeochemical Cycle and Consequences of Climate Changes☆

Vladimir N Bashkin, Institute of Physico-Chemical and Biological Problems of Soil Science RAS, Moscow, Russia; Institute of Natural Gases and Gas Technologies—Gazprom-VNIIGAZ, Moscow, Russia

Introduction

Cyclic processes of exchange of carbon mass are of particular importance for the global biosphere, both in terrestrial and oceanic ecosystems especially owing to the close connections to the global climate changes.

This element is distributed in the atmosphere, water and land as follows. According to existing data there are 6160×10^9 tons or 1.4×10^{16} mol of CO_2 in the atmosphere (1680×10^9 tons of C). A major source of atmospheric carbon dioxide is respiration, combustion, and decay, compared with oxygen, whose main source is photosynthesis. In its turn, an important sink of CO_2 is photosynthesis (about 66×10^9 tons/year or 1.5×10^{15} mol/year). Since carbon dioxide is somewhat soluble in water ($K_H = 3.4 \times 10^{-2}$ mol/L/atm), exchange with the global ocean must also be considered. The approximate global balance of atmosphere–ocean water exchange is 7×10^{15} mol/year (308×10^9 tons/year) being taken up and 6×10^{15} mol/year (264×10^9 tons/year) being released in different parts of the oceanic ecosystem. The residence time of CO_2 in atmosphere is about 2 years, which makes the atmospheric air quite well mixed with respect to this gas. However, a more recent analysis shows that the terrestrial ecosystems have much stronger sinks of carbon dioxide uptake.

In the global ocean, along with occurrence in living organisms, carbon is present in two major forms: as a constituent of organic matter (in solution and partly in suspension) and as a constituent of exchangeable inorganic ions HCO_3^-, CO_3^{2-}, and CO_2.

$$CO_2(g) \leftrightarrow H_2CO_3(aq) \leftrightarrow -H^+/+H^+ \leftrightarrow HCO_3^-(aq) \leftrightarrow -H^+/+H^+ \leftrightarrow CO_3^{2-} \leftrightarrow +Ca^{2+} \leftrightarrow CaCO_3(s)$$

The amount of CO_2(aq) in the oceans is sixty times that of CO_2(g) in the Earth's air, suggesting that the oceans might absorb most of the additional carbon dioxide being injected at present into the atmosphere. However, there are some drawbacks restricting this process. First of all, CO_2 uptake into surface oceanic waters (0–100 m) is relatively slow ($t_{1/2} = 1.3$ years). Secondly, these surface waters mix with deeper waters very slowly ($t_{1/2} = 35$ years). Consequently the surface oceanic waters have the capacity to remove only a fraction of any increase in the anthropogenic CO_2 loading (Fig. 1).

The known analytical monitoring data obtained over many years at the Mauna Loa Observatory in Hawaii, a location far from any anthropogenic sources of carbon dioxide pollution, show a pronounced 1 year cycle of CO_2 content (Fig. 2).

One can see the peak about April and then through around October each year. These data indicate that the content of carbon dioxide in the Earth's atmosphere is not perfectly homogeneous. Some explanations would be of interest to understand this figure.

Hawaii is in the Northern Hemisphere where the photosynthetic activity of vegetation is maximal in summer time (May–September). In this period CO_2 is removed from the air a little bit faster than it is added. The reverse situation occurs during the winter. This is a reasonable explanation and accordingly the monitoring stations in the South Hemisphere show the highest concentration of CO_2 in October, and the lowest in April (see http://mlo.hawaii.gov).

A gradual increase in the partial pressure of carbon dioxide over the last decades is clearly pointed out from Fig. 2. The value of $p(CO_2)$ was ca. 315 ppmv in 1958, it had reached 350 ppmv in 1988, ab. 370 in the beginning of 21th century and ab. 400 at present. Accordingly, this trend can give a doubling of carbon dioxide content in the Earth's atmosphere sometime during the end of the 21st century and this seems reasonable prediction.

Here we should refer to the opinion of some other authors who have argued that increased CO_2 levels in the atmosphere may be a consequence of atmospheric warming, rather than the cause. The statistical analysis of various authors (see "Further Reading" section) led to the conclusion that, although there is a correlation between $p(CO_2)$ and global temperatures, the changes in $p(CO_2)$ appear to lag behind the temperature change by ca. 5 months. A possible explanation, if this trend is proved correct, would be that natural climatic variability like the solar activity alters the temperature of the global Ocean, which contains about 90% of total CO_2

☆*Change History:* February 2018. V. Bashkin made edits throughout the text. Figures 2 and 3 have been replaced with a more up to date versions.

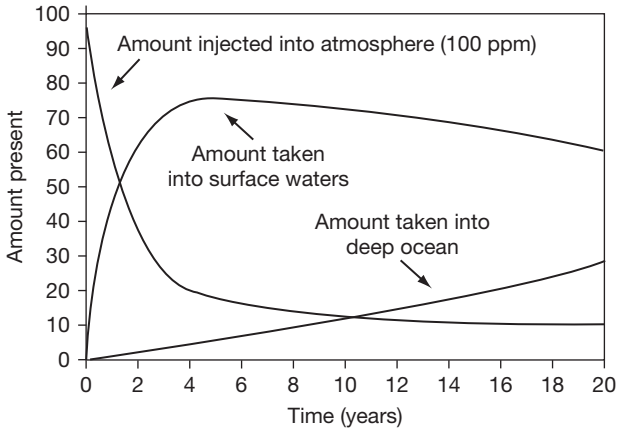


Fig. 1 Calculated uptake of CO_2 from atmosphere to the surface and deep oceanic waters. From Holsten, K. (1992). The global carbon cycle. In: Butcher, S. S., Charlson, R. J., Orians, G. H. & Wolfe, G. V. (eds.) *Global biogeochemical cycles*, pp 239–316. London: Academic Press.

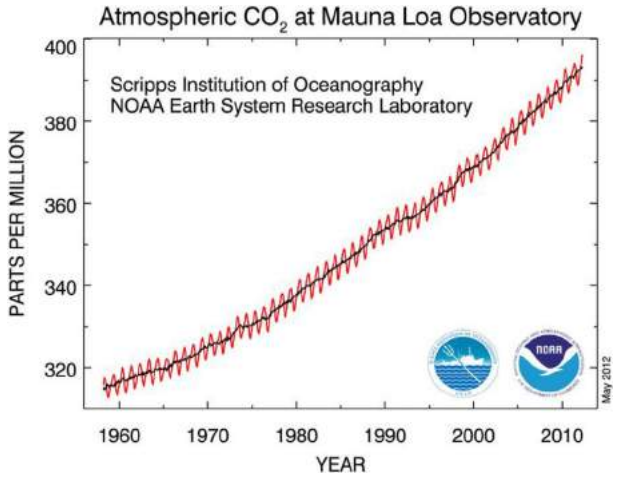


Fig. 2 Observations of CO_2 concentration at the Mauna Loa Observatory for the period of 1958–2012.

mass. In turn, this leads to increase of CO_2 flux from the warmer oceanic water to atmosphere in accordance with the Henry law. Moreover, now the solar activity is decreasing and accordingly the increase $p(CO_2)$ might be changed to the opposite process.

Turnover of Carbon in the Biosphere

As it has been pointed out earlier, the terrestrial ecosystems are the main sink of carbon dioxide due to the photosynthesis process. The present bulk of living organisms is confined to land, and their mass (on dry basis) amounts to 1880×10^9 tons. The average carbon concentration in the dry mater of terrestrial vegetation is 46% and, consequently, the carbon mass in the land vegetation is about 865×10^9 tons.

In accordance with various estimates, the oceanic biomass of photosynthetic organisms contains 1.7×10^9 tons of organic carbon, C_o. In addition, we have to include a large number of consumers. This gives 2.3×10^9 tons of C_o. Totally, the oceanic organic carbon is equal 4.0×10^9 tons or about 0.5% from that in land biomass.

Moreover, a substantial amount of dead organic matter as humus, litter fall and peat is also present in the terrestrial soil cover. The mass of forest litter is close to 200×10^9 tons, mass of peat is around 500×10^9 tons and that of humus, 2400×10^9 tons. Recalculation of this value for organic carbon amounts to 1550×10^9 tons.

However, the greatest amount of carbon in the form of hydrocarbonate, HCO_3^-, ($38{,}600 \times 10^9$ tons) is contained in the ocean, 10 times higher than the total carbon in living matter, atmosphere, and soils.

Thus, in the terrestrial ecosystems the least amount of carbon is monitored in living biomass, followed by dead biomass and atmosphere.

Table 1 Mass distribution of carbon in the Earth's crust

Earth's compartments	*Mass, 10^{18} tons*	*Average concentration, %*			*Mass, 10^{15} tons*				
		CO_2	*C_c*	*C_o*	*CO_2*	*C_c*	*C_o*	*$C_c + C_o$*	*Ratio of C_c/C_o*
Total Earth's crust	28.5	1.44	0.38	0.07	409	108	20	128	5.4
Continental type	18.1	1.48	0.40	0.08	267	72	14	86	5.1
Including									
Sedimentary layer	1.8	9.57	2.61	0.50	177	48	9	57	5.3
Granite layer	6.8	0.81	0.22	0.05	55	15	3	18	5.0
Basalt layer	9.4	0.37	0.10	0.02	35	9.4	1.9	11	5.0
Sub-continental type	4.3	1.37	0.36	0.07	58	16	3	19	5.3
Oceanic type	6.1	1.35	0.36	0.05	82	21	3	24	7.0
Earth's sedimentary shell	2.4	12.4	3.37	0.62	297	81	15	96	5.4
Phanerozoic sedimentary deposits	1.3	15.0	4.08	0.56	194	53	7	60	7.5

Table 2 The major global carbon reservoirs

Reservoirs	*C, 10^9 tons*
Atmosphere, CO_2	1680
Global land	
Vegetable biomass prior to human activity (estimates)	1150
Present natural vegetable biomass	900
Soil cover	
Forest litterfall	100
Peat	250
Humus	1200
Total	1550
Ocean	
Photosynthetic organisms	1.7
Consumers	2.3
Soluble and dispersed organic matter	2100
Hydrocarbonate ions in solution	38,539
Total	40,643
Earth's crust	
Sedimentary shell, C_o	15,000,000
Sedimentary shell, C_c	81,000,000
Continental granite layer, C_o	4,000,000
Continental granite layer, C_c	18,000,000
Total	118,000,000
Total present global C mass	118,044,773

The mass distribution of carbon in the Earth's crust is of interest for understanding of the global biogeochemistry of this element. These values are shown in Table 1. One can see that carbon from carbonates (C_c) is the major form. The C_c/C_o ratio is about 5 for the whole Earth's crust as well as for its main layers (sedimentary, granite, and basalt) and crustal types: continental, sub-continental and oceanic. However, for the latter this ratio is higher.

The sedimentary layer of the Earth's crust is the main carbon reservoir. The C_c and C_o concentrations in the sedimentary layer are by an order of magnitude higher than in granite and basalt layers of lithosphere. The volume of sedimentary shell is about 0.10 from the crust volume, however, this shell accounts for 75% of both carbonate and organic carbon. Dispersed organic matter contains most of the C_o mass. Localized accumulation of C_o in oil, gas, and coal deposits are of secondary importance. It has been estimated that the oil/gas fields amount of 200×10^9 tons of carbon, and the coal deposits contains 600×10^9 tons, totally 800×10^9 tons. This is by three orders of magnitude less than the carbon mass of dispersed organic matter in the sedimentary shell. The general carbon distribution between reservoirs is shown in Table 2.

Thus, there are two major reservoirs of carbon in the Earth: carbonate and organic compounds. It should be stressed that both are of biotic origin. Non-biotic carbonates, for instance, from volcanoes, are the rare exception of the rule. A connecting link between the carbonate and organic species is CO_2, which serves as an essential starting material for both the photosynthesis of organic matter and the microbial formation of carbonates.

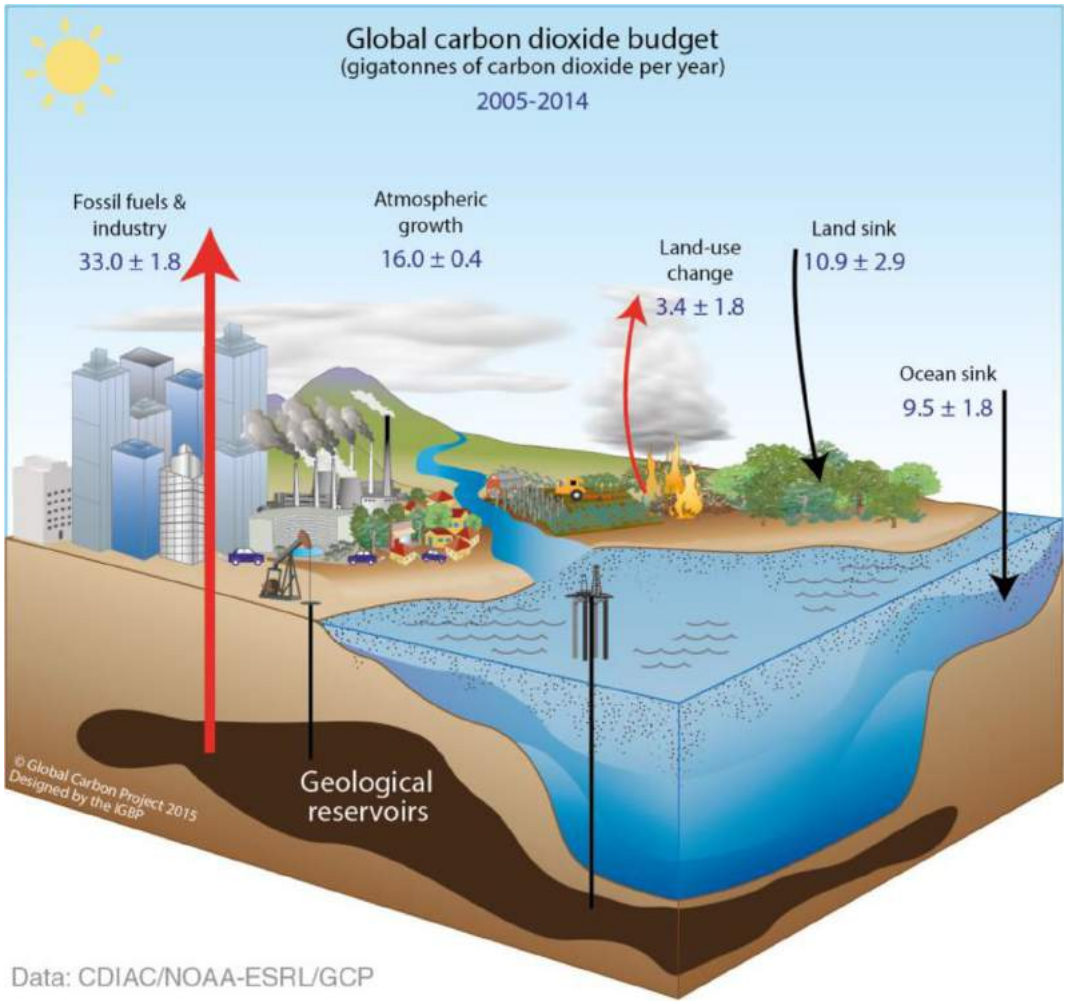


Fig. 3 The global carbon cycle, showing the reservoirs (in 10^9 tons per year) relevant to the anthropogenic perturbation as annual averages over the period 2005–14.

Atmospheric CO_2 provides a link between biological, physical, and anthropogenic processes. Carbon is exchanged between atmosphere, the ocean, the terrestrial biosphere, and, more slowly, with sediments and sedimentary rocks. The faster components of the cycle are shown in Fig. 3.

The component cycles (Fig. 3) are simplified and subject to considerable uncertainty (compare with Table 2, for example). In addition, this figure presents average values. The riverine flux, particularly the anthropogenic portion, is currently very poorly qualified and is not shown here. While the surface sediment storage is approximately 150×10^9 tons, the amount of sediment in the bioturbated and potentially active layer is of order 400×10^9 tons. Evidence is accumulating that many of the key fluxes can fluctuate significantly from year to year (e.g., in the terrestrial sink and storage). In contrast to the static view conveyed by figures such as this one, the carbon system is clearly dynamic and coupled to the climate system on seasonal, inter-annual and decadal time scale.

The carbonate formation and photosynthesis have to be considered as two general processes in the global activity of living matter over geological history of the Earth. The C_c-to-C_o mass ratio may specify the "growth limit" of living matter at sequential stages of Earth's geological history over the period of 3.5–3.8 billion years. This ratio tends to decrease regularly with the last 1.6 billion years. The C_c/C_o ratio was 18 in the sedimentary layers of the Upper Proterozoic period (1600–750 million years); that of the Paleozoic (570–400 million years), 11; of the Mesozoic (235–66 million years), 5.2, and of the Cainozoic (66 million years to the present), 2.9. The never interrupted increase in the relative content of organic matter in the ancient stream loss provides evidence for a progressively increasing productivity of terrestrial photosynthetic organisms. This provides also the proof for growing importance of global terrestrial ecosystems in fixation of CO_2. Apparently, the increasing productivity of land vegetation would be the major sink of CO_2 under the increasing content of this green-house gas in the atmosphere, however, the role of increasing input of nitrogen, for instance, with atmospheric deposition, has to be considered. Moreover, both carbonate formation and the photosynthesis of organic matter share in the common tendency for removal from the atmosphere of CO_2 continually supplied from the mantle.

Table 3 Net primary production of the Earth's major ecosystems

Global ecosystem zone	*Area, 10^6 km^2*	*Plant mass, 10^9 tons*	*C-NPP, 10^9 tons*
Polar	8.1	13.8	1.3
Coniferous forests	23.2	439.1	15.2
Temperate	22.5	278.7	18.0
Subtropical	24.3	323.9	34.6
Tropical	55.9	1347.1	102.5
Total land	133.9	2402.1	171.6
Lakes and rivers	2.0	0.04	1.0
Glaciers	13.9	0	0
Total continents	149.3	2402.5	172.6
Oceans	361.0	0.2	60.0
Earth total	510.3	2402.7	232.6

Consequently these processes take part in the global mechanisms for maintaining the present low concentration of carbon dioxide in the Earth's gas shield, which is an essential parameter in the greenhouse effect.

Carbon Fluxes in Terrestrial Ecosystems

All three CO_2-controlling processes (ocean soaking, photosynthesis and carbonate formation) play an important role in maintaining equilibrium in the biosphere–atmosphere–hydrosphere system. The photosynthetic process is of great importance for living plants and microorganisms. The difference between total photosynthesis and respiration processes is defined as "net primary production," NPP. The global NPP distribution in the Earth's major ecological zones is shown in Table 3.

Oceans, despite their much larger surface area, contribute much less than half of the global NPP. The reason is related to highly nutrient deficiency in surface waters, which limits the photosynthesis process. Oceanic production is mainly concentrated in coastal zones, especially where upwelling of deep water brings the nutrients (P and N, of major interest) into the surface layer, 0–100 m. On land the photosynthetic process is also often limited by nutrient deficit, however, the influence of water storage and low temperature plays more important role. That is why subtropical and tropical ecosystems contribute much more to global NPP than their proportional share.

The amount of annually decaying organic matter is the subject of speculation. However, some estimates might be done. For instance, in terrestrial ecosystems only, the humus accumulation of carbon in soils is about 70% of the total accumulation of CO_2 in the atmosphere. We may presume therefore that the stable long-lived humic compounds acquire some 30% of carbon annually from the dead organs of plants, and the complete renewal of humus in soils extends over period of $0.3–1.0 \times 10^3$ years. The variance depends on the moisture and temperature conditions in the region of question.

Terrestrial biomass is divided into a number of sub-reservoirs with different turnover times. Forest ecosystems contain 90% of all carbon in living matter on land but their NPP is only 60% of the total. About half of the primary production in Forest ecosystems is in the form of twigs, leaves, shrubs, and herbs that only make up 10% of the biomass. Carbon in wood has a turnover time of the order of 50 years, whereas these times for carbon in leaves, flowers, fruits, and rootlets are less than a few years. When plant material becomes detached from the living plant, carbon is moved from phytomass reservoir to litter. "Litter or litterfall" can refer to a layer of dead plant material on the soil surface. A litter layer can be a continuous zone without sharp boundaries between the obvious plant structures and a soil layer containing amorphous organic carbon. Decomposing roots are a kind of litter that seldom receives a separate treatment due to difficulties in distinguishing between living and dead roots. Total litter is estimated as 60×10^9 C and total litterfall as 40×10^9 tons C/year. The average turnover time for carbon in litter is thus about 1.5 years, although for tropical ecosystems with mean temperature above 30°C, the litter decomposition rate is greater than the supply rate and so storage is impossible. For colder climates, NPP exceeds the rate of decomposition in the soil and organic matter in the form of peat is accumulated. The total global amount of peat might be estimated at 165×10^9 tons C. Average temperature at which there is a balance between production and decomposition is about 25°C.

Humus is a type of organic matter in terrestrial ecosystems that is not readily decomposed and therefore makes up the carbon reservoir with a long turnover time (300–1000 years). An assessment of the various carbon pools for a temperate grassland soil is presented in Fig. 4.

The undecomposed litter (4% of the soil carbon) has a turnover time measured in tens of years, and the 22% of the soil carbon in the form of fulvic acids is intermediate with turnover times of hundreds of years. The largest part (74%) of the soil organic carbon (humic acids and humins) also has the longest turnover time (in thousands of years).

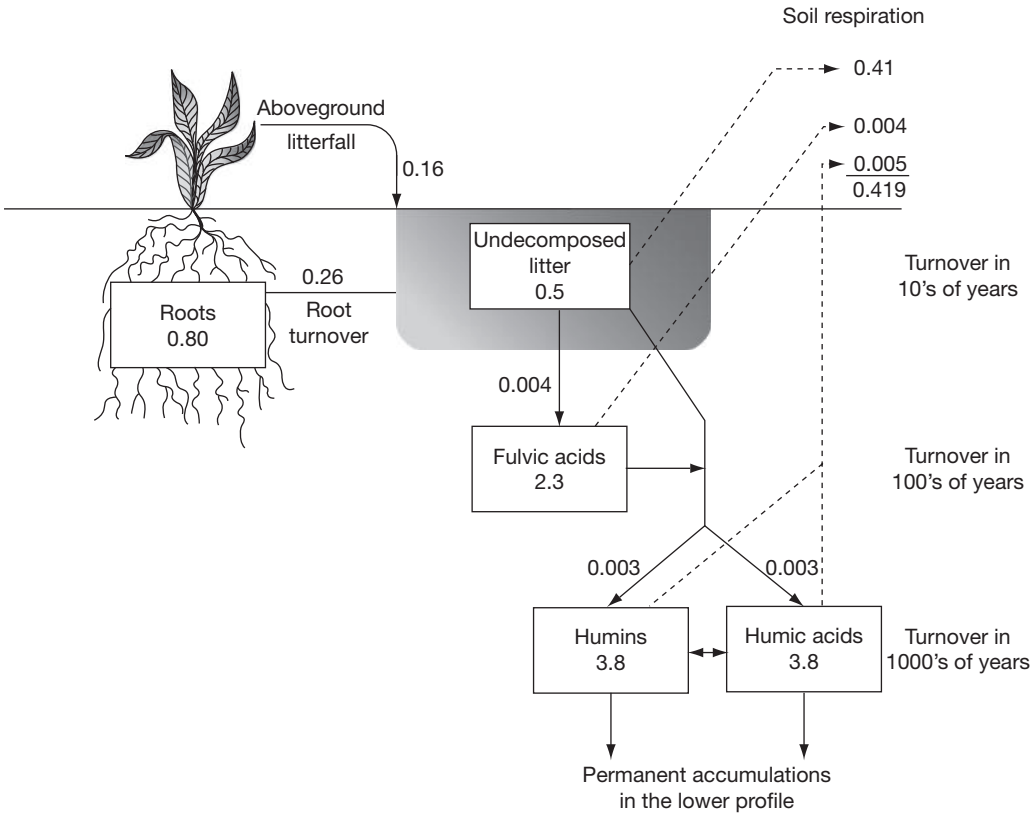


Fig. 4 Detrital carbon dynamics for the 0–20 cm layer of chernozem grassland soil. Carbon pool (kgC/m^2) and annual transfers (kgC/m^2/year) are shown. Total profile content down to 20 cm is 10.4 kgC/m^2.

Comparison of Carbon Biogeochemical Processes in Terrestrial and Aquatic Ecosystems

The synthesis and degradation of organic matter in the ocean are significantly distinct from those in terrestrial ecosystems. The phytoplankton provides for a larger part of photosynthetic organic matter. The dry mass of phytoplankton is three orders of magnitude less than global terrestrial mass, whereas the annual production is only about three times smaller. This can be related to the much faster life cycles of plankton organisms in comparison with the terrestrial vegetation.

Let us consider the renewal of terrestrial and oceanic organic matter. The terrestrial biomass might be assessed as 2400–2500 $\times$ 10^9 tons of dry organic matter and annual production as 170–175 $\times$ 10^9 tons. These values present a period of 13–15 years for complete renewal of organic matter. In the oceans, the problem is much more complicated. The various authors give 8–10-fold discrepancy in the existing estimates of phytoplankton productivity and biomass. It is estimated also that phytoplankton mass cycle takes 1–2 days to be completed. Taking this into account, we can reasonably consider that the renewal of the total biomass in the global ocean takes about 1 month. Based on modern assessments, the annual production of photosynthesis varies from 20–30 $\times$ 10^9 to 100 $\times$ 10^9 tons of organic carbon and the average values are 50–60 $\times$ 10^9 tons C_o. Furthermore, we can hypothesize that the plankton-synthesized organic matter is almost completely assimilated in subsequent upper food webs. Thus, the organic precipitation would not exceed 0.1 $\times$ 10^9 tons. These calculations present the annual uptake of terrestrial and oceanic living organisms of about 440 $\times$ 10^9 tons CO_2 or 120 $\times$ 10^9 tons C_o. Most of this amount recycles into the ocean and atmosphere.

Carbon Dioxide Interactions in Air-Sea Water System

The interaction between carbon dioxide in the atmosphere and the hydrosphere is the principal factor for understanding large carbon biogeochemical cycles. As it has been mentioned above, the gases of the troposphere and the surface layer of the ocean persist in a state of kinetic equilibrium.

Compared with the atmosphere, where most carbon is presented by CO_2, oceanic carbon is mainly present in four forms: dissolved inorganic carbon (DIC), dissolved organic carbon (DOC), particulate organic carbon (POC), and the marine biota itself.

DIC concentrations have been monitored extensively since the appearance of precise analytical techniques. When CO_2 dissolves in water it may hydrate to form H_2CO_3(aq), which, in turn, dissociates to HCO_3^- and CO_3^{2-}. This process depends on pH and specification is shown in Fig. 5.

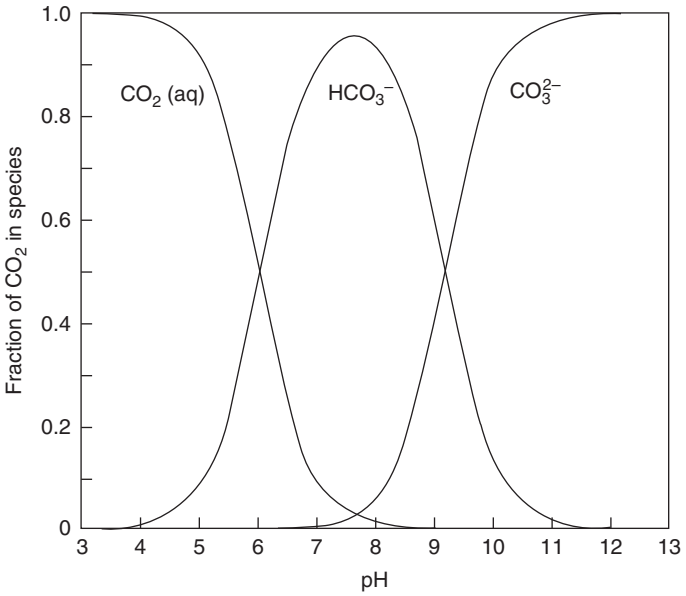


Fig. 5 Distribution of dissolved carbon species in seawater as a function of pH at 15°C and a salinity of 35‰. Average oceanic pH is about 8.2.

The conjugate pairs responsible for most of the pH buffer capacity in marine water are HCO_3^-/CO_3^{2-} and $B(OH)_3/B(OH)_4^-$. Although the predominance of HCO_3^- at the oceanic pH of 8.2 actually places the carbonate system close to a pH buffer minimum, its importance is maintained by the high DIC concentration (~2 mm). Ocean water in contact with the atmosphere will, if the air-sea gas exchange rate is short compared to the mixing time with deeper water, reach equilibrium according to Henry's Law. At the pH of oceanic water around 8.2, most of the DIC is in the form of HCO_3^- and CO_3^{2-} with a very small proportion of H_2CO_3. Although H_2CO_3 changes in proportion to CO_2 (g), the ionic form changes little as a result of various acid–base equilibrium.

From chemical aqueous carbon specification, the alkalinity, Alk, representing the acid-neutralizing capacity of the solution, is given by the following equation:

$$Alk = [OH^-] - [H^+] + [B(OH)_4^-] + [B(OH)_3] + 2[CO_3^{2-}]$$

Average DIC and Alk concentrations for the World's Ocean are shown in Fig. 6.

With an average DIC of 2.35 mmol/kg seawater and the world oceanic volume of 1370×10^6 km^3, the DIC carbon reservoir is estimated to be $37{,}900 \times 10^9$ tons C. The surface waters of the World's ocean contain a minor part of DIC, $\sim 700 \times 10^9$ tons C. However, these waters play an important role in air-deep water exchange (see above).

Oceanic surface water is everywhere supersaturated with respect to the two solid calcium carbonate species calcite and aragonite. Nevertheless, calcium precipitation is exclusively controlled by biological processes, specifically the formation of hard parts (shells, skeletal parts, etc.). The very few existing amounts of spontaneous inorganic precipitation of $CaCO_3$(s) come from the Bahamas region of the Caribbean.

The detrital rain of carbon-containing particles can be divided into two groups: the hard parts comprised of calcite and aragonite and the soft tissue containing organic carbon. The composition of the soft tissue shows the average ratio of biophils as P:N:C:Ca:S = 1:15:131:26:50, with C_c:C_o ratio as 1:4.

The estimation of C_c and C_o mass annually eliminated from the biogeochemical cycles in ocean is a very uncertain task. The carbonate–hydrocarbonate system includes the precipitation of calcium carbonate as a deposit:

$$\begin{array}{lc} \text{Atmosphere} & CO_2 \\ & \uparrow\downarrow \\ \text{Surface ocean layer} & H_2O \leftrightarrow H_2CO_3 \leftrightarrow H^+ + HCO_3^- \leftrightarrow H^+ + CO_3^{2-} + Ca^{2+} \\ & \uparrow\downarrow \\ \text{Deep ocean water} & CaCO_3 \end{array}$$

The binding of carbon into carbonates is related to the activity of living organisms. However, the surface runoff of Ca^{2+} ions from the land determines the formation of carbonate deposits to a significant degree. The Ca^{2+} ion stream is roughly 0.53×10^9 tons/year, which can provide for a $CaCO_3$ precipitation rate of 1.33×10^9 tons/year. This would correspond to the loss of 0.57×10^9 tons CO_2, or 0.16×10^9 tons C from the carbonate–hydrocarbonate system.

The surface runoff from the World's land plays an important role in the global carbon mass exchange. The continental runoff supply of HCO_3^- is 2.4×10^9 tons/year, that is, 0.47×10^9 tons/year for carbon. Besides, the stream water contains dissolved organic matter at 6.9 mg/L, which makes up to an annual loss of 0.28×10^9 tons/year. The average carbon concentration of suspended insoluble organic matter in the stream discharge is 5 mg/L, which gives the loss of about 0.2×10^9 tons/year. Most of

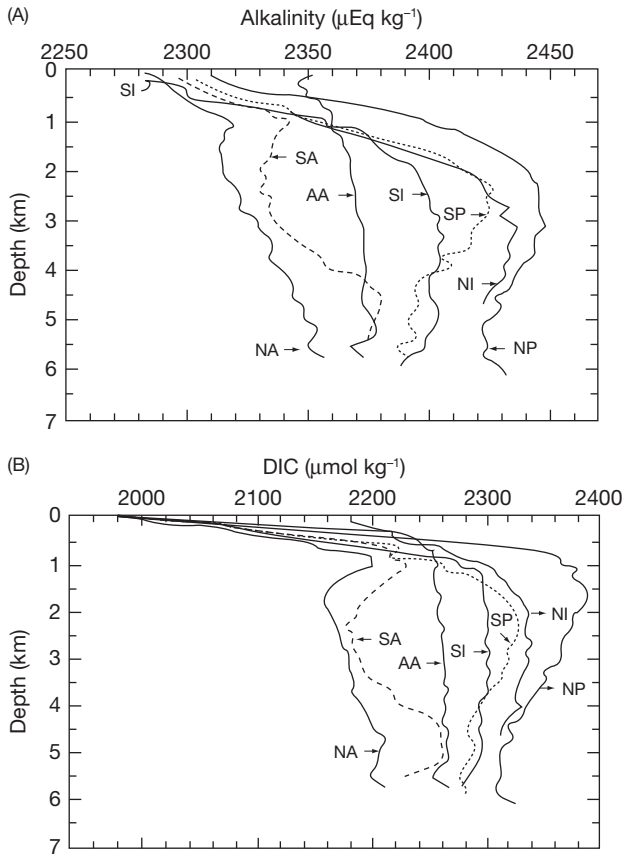


Fig. 6 The vertical distribution of alkalinity (A) and dissolved inorganic carbon in the World's Ocean. Ocean region are shown as NA (North Atlantic), SA (South Atlantic), AA (Antarctic), SI (South Indian), NI (North Indian), SP (South Pacific), and NP (North Pacific).

this mass fails to reach the open ocean and becomes deposited in the shelf and the estuarine delta of rivers. We can see that equal amounts of C_c and C_o (0.5×10^9 tons for each) are annually lost from the World's land surface.

The formation of carbonates and the accumulation of organic matter are not confined solely to ocean; these processes occur also on the land. The mass of carbonates annually produced in the soils of arid landscapes appears to be high enough.

Global Carbon Fluxes

Two large cycles determine global dynamics of carbon mass transport in the biosphere. The first of these is provided for by the assimilation of CO_2 and decomposition of H_2O through photosynthesis of organic matter followed by its degradation to yield CO_2. The second cycle involves the uptake—release of carbon dioxide by natural waters via chemical reactions of CO_2 and H_2O leading to build-up of a carbonate–hydrocarbonate system. The cycles are intimately related to the activity of living matter. The living matter of the biosphere, the global water cycle, and carbonate–hydrocarbonate system regulate the cyclic mass exchange of carbon between atmosphere, land, and ocean. These global carbon fluxes are shown in Table 4.

A specific feature of these two major biogeochemical cycles of carbon is their openness, which is related to the permanent removal of some carbon from the turnover as a dead organic matter and carbonates. The carbon burial in the sea deposits is of great importance for biosphere development.

There is a suggestion that the alteration of glacial and interglacial periods in the Pleistocene was mainly due to fluctuations of CO_2 in the atmosphere. It may be hypothesized that the spread of land ice and the drastic reduction of forest areas with their typically high biomass were favorable to an elevated content of carbon dioxide in the atmosphere and the subsequent climatic warming up. In its turn, the resulting contraction of glacial areas and reforestation was attended by an increased CO_2 uptake from the atmosphere and by its binding to the biomass and soil organic matter. The resulting effect was a gradual cooling and the onset of a new glaciation followed by reduction of forest areas and repetition of the whole cycle.

The role of carbon dioxide in the Earth's historical radiation budget merits modern interest in arising atmospheric CO_2. There are, however, other changes of importance. The atmospheric methane concentration is increasing, probably as a result of increasing

Table 4 Fluxes of carbon in the biosphere

Fluxes	*C, 10^9 tons/year*
World's ocean	
Turnover of planktonic photosynthesis organisms	50
CO_2 uptake by ocean	30
CO_2 release by ocean	30
C_o deposited in precipitation	0.08
C_c deposited in precipitation	0.16
World's land	
Biological cycle (photosynthesis–degradation of organic matter)	85
HCO_3^- ion mass exchange between land and troposphere	
Supply to troposphere	0.136
Rainfall washout from troposphere	0.139
Stream loss of	
DIC	0.47
DOC	0.28
POC	0.20
Transport of oceanic air-borne HCO_3^- ions to land	0.003

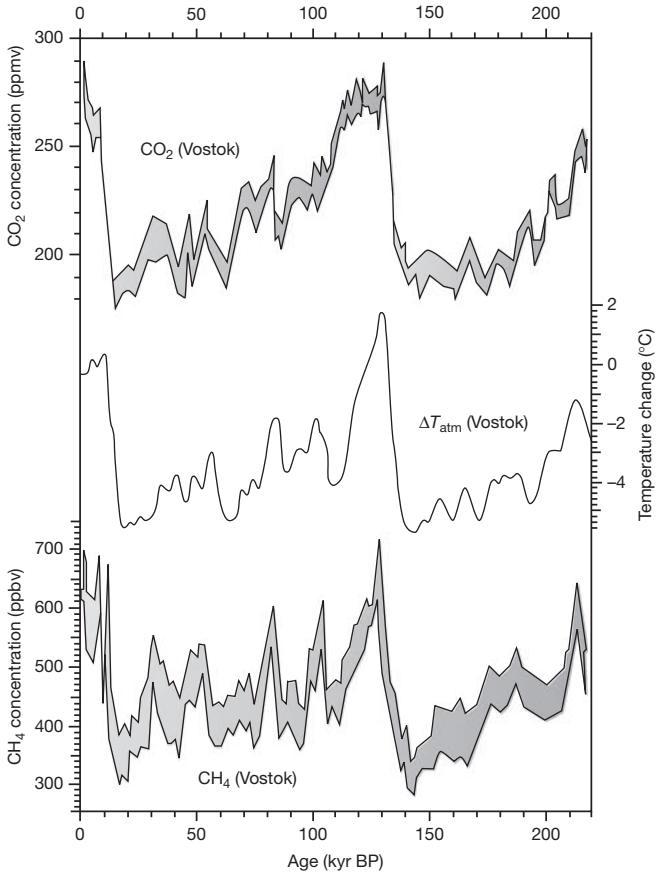


Fig. 7 Temperature anomalies and methane and carbon dioxide concentrations over the past 220,000 years as derived from the ice core records at Vostok, Antarctica.

cattle population, rice production, and biomass burning. Increasing methane concentrations are important because of the role they play in stratospheric and tropospheric chemistry. Methane is also important to the radiation budget of our planet.

Analyses of ice cores from Vostok, Antarctica, have provided new data on natural variations of CO_2 and CH_4 levels over the last 220,000 years. The records show a marked correlation between Antarctic temperature, as deduced from isotopic composition of the ice and the CO_2/CH_4 profiles (Fig. 7).

Clear correlations between CO_2 and global mean temperature are evident in much of the glacial-interglacial palaeo-record. This relationship of CO_2 concentration and temperature may carry forward into the future, possibly causing significant positive climatic feedback on CO_2 fluxes.

Global Climate Changes and Critical Loads of Sulfur and Nitrogen at the European Ecosystems

Global biogeochemical cycle of carbon and its alterations seemly attracted the great attention in the mass media due to CO_2 increase in the atmosphere is closely related to the various changes in the Earth biosphere. These changes include both climate changes and environmental pollution. Here we should mention a project on the integrated assessment of regional air pollution and climate change in Europe, AIR-CLIM project, which inter alia examined whether climate change will alter the effectiveness of agreed-upon or future policies to reduce regional air pollution in Europe. Climate changes and emission abatement strategy can be estimated using the calculations of pollutants critical loads and their exceedances (see "Further Reading" section).

A critical load has been defined as the maximum input of pollutants (sulfur, nitrogen, heavy metals, POPs, etc.), which will not introduce harmful alterations in biogeochemical structure and function of ecosystems in the long-term, that is, 50–100 years according to present knowledge. Starting from this general definition, methodologies have been developed during the 1990s by the working group on effects (WGE) under the Long-Range Transboundary Air Pollution (LRTAP) convention for calculating and mapping critical loads in Europe. This has been used by European countries to calculate critical loads of pollutants for various ecosystems (forests, surface waters, semi-natural vegetation).

Transboundary air pollution by sulfur, nitrogen, heavy metals and persistent organic species is not the only environmental problem calling for internationally agreed abatement policies. In developed countries, it is the issue of climate change, which currently attracts most attention and resources, and negotiations under the framework convention on climate change (FCCC) trying to come up with equitable mitigation policies. To date, climate change policies are mostly discussed in isolation. However, any measures taken (or not taken) to slow down global climate changes are likely to have an impact on other environmental problems.

Eight scenarios for different combinations of future green house gas (GHG), sulfur and nitrogen emissions, covering the years 1990–2100, were developed during the AIR-CLIM project (Table 5).

To assess the risk of ecosystem damage due to a given scenario, critical loads have to be compared with the resulting deposition patterns. Within the integrated assessment framework of AIR-CLIM, deposition fields due to emission scenarios are computed with the source-receptor matrices (SRMs) derived from the EMEP long-range atmospheric transport model. The SRMs derived for the meteorological years 1985–1996 were averaged to minimize the effects of inter-annual variability. With the aid of these SRMs, the sulfur and nitrogen ($NO_x + NH_3$) emissions of the European countries the respective depositions in every grid cell are computed. If the depositions are greater than critical loads, we say the critical loads are exceeded. While in the case of a single pollutant the exceedance can be defined in an obvious manner, for example, $Ex(Ndep) = Ndep - CLnut(N)$, there is no unique exceedance (i.e., amount of deposition to be reduced to reach non-exceedance) in the case of acidifying N and S.

Fig. 8 depicts the temporal development of the percentage of forest area for which critical loads of acidity and nutrient nitrogen are exceeded under the eight AIR-CLIM scenarios. The area for which critical loads are exceeded declines under all scenarios, starting from 41% for acidity critical loads and 75% for nutrient N in 1990. The speed decrease after 2010, however, differs between the two sets of scenarios, with larger decreases in the B1-set. In all the cases, the A1-P and the B1-450-A scenarios are the least and most stringent one, respectively, with the other scenarios giving intermediate results. The most striking conclusion is that acidification (almost) ceases to be a problem, with exceedance percentages in 2100 between 4.7% (A1-P) and 0.7% (B1-450-A). In drawing this conclusion it has to be borne in mind that considering the in-grid variability of deposition (e.g., by reducing the grid size) would certainly lead to higher exceedances. Furthermore, areas that cease to be exceeded at some point in time are not at once without the risk of adverse effects. The recovery of the chemical, and especially the biological, status of the soil is delayed due to finite buffers, which have to equilibrate with the lower deposition. Only dynamic models can provide estimates of the times needed for a full recovery.

Table 5 Overview of AIR-CLIM scenarios

Scenario	*Green house gas policies*	*SO_2/NO_x policies*
A1-P	None	Present policies
A1-A	None	Advanced policies
A1-550-P	To achieve 550 ppm CO2 stabilization	Present policies
A1-550-A	To achieve 550 ppm CO2 stabilization	Advanced policies
B1-P	None	Present policies
B1-A	None	Advanced policies
B1-450-P	To achieve 450 ppm CO2 stabilization	Present policies
B1-450-A	To achieve 450 ppm CO2 stabilization	Advanced policies

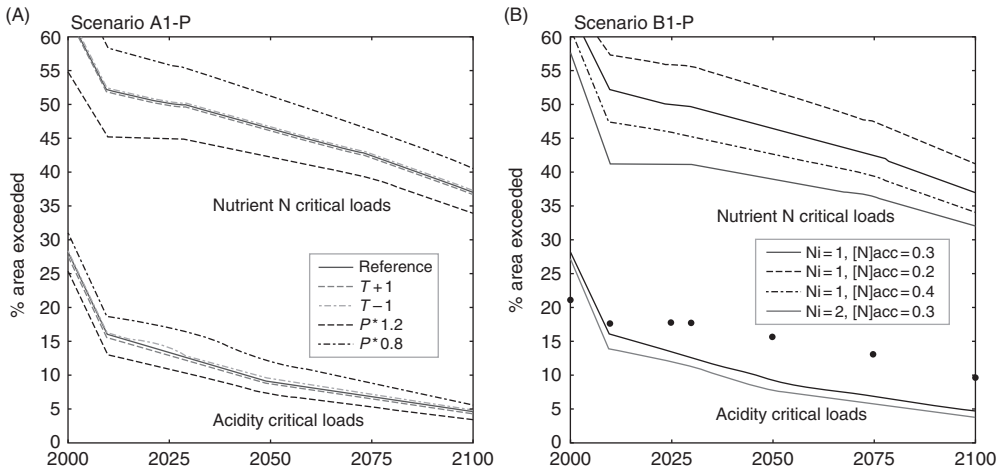


Fig. 8 Temporal development of the percentage of forest area for which the critical loads of acidity and nutrient nitrogen are exceeded for the four scenarios in the A1-set (*left*) and for corresponding four scenarios in the B1-set (*right*).

Eutrophication, on the other hand, continues to be a widespread problem, even under the most stringent scenario, which brings the exceedance hardly down to 15% of the forest area. This confirms the conclusion that nitrogen will be the main pollutant in need of future mitigation.

Summary

The changes in the global biogeochemical carbon cycle become more and more obvious however the relative contributions of natural and anthropogenic activities are still uncertain. However, the role of carbon dioxide in the Earth's historical radiation budget merits modern interest in arising atmospheric CO_2. The relative changes include both climate changes and environmental pollution.

Critical loads have been widely used to formulate European emission reduction policies for sulfur and nitrogen. The critical load values depend on the ecosystem characteristics that might be altered due to climate changes. An investigation of the impact of different scenarios of climate change on critical loads and their exceedances is of both scientific and political interests.

The recent estimates have shown that the acidity critical loads will be exceeded only in small parts of Europe under all scenarios. It should be borne in mind, however, that: (a) non-exceedance does not mean immediate recovery, and (b) higher resolution deposition fields, capturing some of their small-scale variability, would certainly lead to more widespread exceedances.

Eutrophication, on the other hand, will continue to be a problem even under the most stringent scenario. This confirms the important and increasing role nitrogen plays in environmental problems, both in its oxidized and reduced forms. Thus, research should focus on the effects of nitrogen in the environment, especially under conditions of climate change, whereas, policies should concentrate on further reductions of nitrogen emissions. This not only reduces acidification and eutrophication, but also helps curbing the formation of tropospheric ozone.

Further Reading

Achard F and House JI (2015) Reporting carbon losses from tropical deforestation with pan-tropical biomass maps. *Environmental Research Letters* 10: 101002.

Assmann KM, Bentsen M, Segschneider J, and Heinze C (2010) An isopycnic ocean carbon cycle model. *Geoscientific Model Development* 3: 143–167. https://doi.org/10.5194/gmd-3-143-2010.

Barnola J-M, Pimienta P, Raynaud D, and Korotkevich TS (1991) CO_2-climate relationship as deduced from Vostok ice core: A re-examination based on new measurement and on a re-evaluation of the air dating. *Tellus* 43B: 83–90.

Barrett K and Berge E (eds.) (1996) *Transboundary Air Pollution in Europe. EMEP/MSC-W report 1/1996.* Oslo, Norway: Norwegian Meteorological Institute.

Bashkin VN (2006) *Modern biogeochemistry: Environmental risk assessment*, 2nd edn. Dordrecht, The Netherlands: Springer.

Bashkin V (ed.) (2016) Biogeochemical technologies for managing pollution in polar ecosystems. In: *Environmental Pollution*, Vol. 26, Switzerland: Springer.

Bashkin VN and Park S-U (1998) *Acid deposition and ecosystem sensitivity in East Asia.* New York: Nova Science Publishers, Ltd. p. 427.

Bauer JE, Cai W-J, Raymond PA, Bianchi TS, Hopkinson CS, and Regnier PAG (2013) The changing carbon cycle of the coastal ocean. *Nature* 504: 61–70.

Dlugokencky, E. and Tans, P.: Trends in atmospheric carbon dioxide, National Oceanic & Atmospheric Administration, Earth System Research Laboratory (NOAA/ESRL), available at: http://www.esrl.noaa.gov/gmd/ccgg/trends, last access: 8 August 2014.

FAO (2010) *Global Forest Resource Assessment 2010.* p. 378.

Ilyina T, Six K, Segschneider J, Maier-Reimer E, Li H, and Núñez-Riboni I (2013) The global ocean biogeochemistry model HAMOCC: Model architecture and performance as component of the MPI-earth system model in different CMIP5 experimental realizations. *Journal of Advances in Modeling Earth Systems* 5: 287–315.

Le Quéré C, Moriarty R, Andrew RM, Canadell JG, Sitch S, Korsbakken JI, Friedlingstein P, Peters GP, Andres RJ, Boden TA, Houghton RA, House JI, Keeling RF, Tans P, Arneth A, Bakker DCE, Barbero L, Bopp L, Chang J, Chevallier F, Chini LP, Ciais P, Fader M, Feely RA, Gkritzalis T, Harris I, Hauck J, Ilyina T, Jain AK, Kato E, Kitidis V, Klein Goldewijk K,

Koven C, Landschützer P, Lauvset SK, Lefèvre N, Lenton A, Lima ID, Metzl N, Millero F, Munro DR, Murata A, Nabel JEMS, Nakaoka S, Nojiri Y, O'Brien K, Olsen A, Ono T, Pérez FF, Pfeil B, Pierrot D, Poulter B, Rehder G, Rödenbeck C, Saito S, Schuster U, Schwinger J, Séférian R, Steinhoff T, Stocker BD, Sutton AJ, Takahashi T, Tilbrook B, van der Laan-Luijkx IT, van der Werf GR, van Heuven S, Vandemark D, Viovy N, Wiltshire A, Zaehle S, and Zeng N (2015) Global carbon budget. *Earth System Science Data* 7: 349–396. https://doi.org/10.5194/essd-7-349-2015.

Posch M (2002) Impacts of climate change on critical loads and their exceedances in Europe. *Environmental Science & Policy* 5: 307–317.

Rödenbeck C, Bakker DCE, Metzl N, Olsen A, Sabine C, Cassar N, Reum F, Keeling RF, and Heimann M (2014) Interannual Sea-air CO_2 flux variability from an observation-driven ocean mixed-layer scheme. *Biogeosciences* 11: 4599–4613. https://doi.org/10.5194/bg-11-4599-2014.

Shevliakova E, Pacala S, Malyshev S, Hurtt G, Milly P, Caspersen J, Sentman L, Fisk J, Wirth C, and Crevoisier C (2009) Carbon cycling under 300 years of land use change: Importance of the secondary vegetation sink. *Global Biogeochemical Cycles* 23. GB2022. https://doi.org/10.1029/2007GB003176.

Climate Change Models[☆]

Andrey Ganopolski, Potsdam Institute for Climate Impact Research, Potsdam, Germany

Introduction

This article presents an overview of the climate change, climate modeling, and future climate predictions. Of the subsequent sections, the first section discusses the concept of natural and anthropogenic climate change and the mechanisms causing climate variability. The second section, based on the results of paleoclimatological studies and observational data, presents an overview of past climate changes. The third section presents the scientific basis and the existing types of climate models. The fourth section discusses the prediction of future climate change based on the simulations performed with climate models and potential impact of global warming on natural ecosystem and human life.

Climate and Climate Change

Definition of Climate and Climate Change

Traditionally, the term "climate" referred to the averaged weather conditions, such as the mean July temperature or annual precipitation. For many applications it is important to know not only the averaged characteristics, such as monthly averaged temperature, but also different measures of variability (statistics), like the interannual variability of the precipitation or a number of extreme weather events. Thereby, more precisely the term climate is characterized as "the statistical description (of the climate system) in terms of the mean and variability of relevant quantities." Temporal evolution of the climate characteristics beyond the timescale of individual weather events is named "climate variability" while statistically significant trend of climate state on longer timescales (decades and more) is referred to as "climate change." For example, North Atlantic Oscillation (NAO) or El Niño/Southern Oscillation (ENSO) represents climate variability, while dramatic climate variations over past several million years associated with waning and waxing of the Northern Hemisphere ice sheets are the example of climate change. Pronounced trends of many climate characteristics recorded during the 20th century are primarily attributed to human influence on climate and hence represent anthropogenic climate change.

Natural Climate Variability and Anthropogenic Climate Change

It is known that climate has changed in the past and will change in the future under the influence of numerous natural factors such as changes of solar luminosity and orbital parameters of the Earth, volcano eruptions, changes in the atmospheric composition and Earth's geography. Apart from the response to change of external and internal factors, the climate experiences natural fluctuations,

[☆]*Change History*: March 2018. Todd T. M. Swannack updated references.

the so-called internal climate variability. It arises from the fact that the climate system is a strongly nonlinear system and even under constant external conditions posses permanent secular variations of the state variables. The most well known example is the instability of large-scale atmospheric circulation which gives rise to permanent generations of cyclonic and anticyclonic eddies which affect the weather over the globe. Interaction between several components of the climate system can also lead to development of more or less periodic self-sustained climate oscillations. The most known example is ENSO which originates from the interaction between the atmosphere and the ocean in the Tropics and affects a large part of the globe.

Natural climate factors, such as changes in the Earth's orbital parameters or volcanic eruptions, cause additional variations of climate state. Together with internal climate variability, they produce natural climate variability. The term "anthropogenic climate change" refers to the part of climate change attributed to all aspects of human activity, such as emission of greenhouse gases and aerosols, and land-cover changes. Anthropogenic climate change is added on the top of natural climate variability and separation of anthropogenic and natural climate variability still represents a formidable challenge since the magnitude of both types of climate variability are comparable.

Climate Forcings

There are many processes inside and outside of the Earth system which affect climate. On the timescales from billion years to decades, the climate is affected by changes in solar luminosity, composition of the atmosphere, volcanic activity, changes in the position of the continents, and variations of the Earth's orbital parameters. Changes in the internal and external (for the Earth system) factors which directly affect climate are often referred to as "climate forcing." Climate forcing can be quantified in term of "radiative forcing," defined as an energy imbalance imposed on the climate system by changes of given factor. For example, a doubling of CO_2 concentration causes globally averaged imbalance at the top of the troposphere of about 4 W m^{-2}. Radiative forcing is a convenient measure to compare climate impact of different factors, such as changes in concentration of greenhouse gases and aerosols but not all climate forcings can be easily expressed in term of radiative forcing. For example, variations of the Earth orbital parameters cause large changes in seasonal distribution of incoming solar radiation but their globally averaged direct effect on energy balance of the planet is rather small. However, due to a number of strongly nonlinear climate feedbacks associated with the ice sheets and greenhouse gases, the variations in the Earth's orbital parameters caused significant climate changes over the past several millions of years.

While some of climate forcings, such as the Earth orbital parameters or solar luminosity, are external for the Earth system, changes in atmospheric composition on the timescales of thousand years are internal ones and represent internal climate feedbacks to external forcing. For example, the growth of large ice sheets under varying Earth orbital parameters caused widespread cooling, which, in turn, leads to an enhanced carbon uptake by the ocean and a lowering of atmospheric CO_2 concentration. This, in turn, cools climate additionally and facilitates further growth of the ice sheets.

Since the last century human activity became an important factor of climate change. The most important anthropogenic climate forcing is the change in atmospheric composition of the so-called greenhouse gases: CO_2, CH_4, N_2O, and others. Apart from that, burning of fossil fuel and forest leads to an increase of atmospheric concentration of several types of aerosols. It is believed that the net effect of anthropogenic aerosols is cooling and hence aerosols partly compensate the warming effect of greenhouse gases. Changes in land cover (land use), primarily via deforestation of the large part of the continents, also affect climate. The direst effect of deforestation is a cooling, but since deforestation also contributes to an increase of CO_2 concentration, the sign of temperature changes related to deforestation depends on the regions. Atmospheric pollution by several chemicals also affects tropospheric and stratospheric concentration of ozone. The former is increasing under the influence of anthropogenic factors while the latter is decreasing which contributes to the development of the so-called "ozone hole" that represents the direct danger for human health.

Past Climate Changes

Climate Change on Geological Timescale

Paleoclimate records provide rich information about temporal evolution of climate on different timescales. Reconstructions of past climate changes are now extensively used as a test bed for the climate models to assess their performance for the climate conditions different from the present one. Past climate changes have also been used to assess climate sensitivity to change in the atmospheric CO_2 concentration. At last, past climate changes clearly demonstrate a strongly nonlinear response of the climate system to gradual changes in external forcing and hence indicate a possibility of the existence of some thresholds, crossing of which could lead to irreversible climate change.

On the geologic timescales (tens and hundreds of million years), paleoclimate records provide a strong support to the dominant control of the atmospheric CO_2 concentrations on the Earth's climate, although other factors, such as changes in Earth's geography, also played an important role. Over the past tens of million years climate progressively cooled, and some 3 million years ago a periodic widespread glaciation of the Northern Hemisphere began. The cycles of waning and waxing of the Northern Hemisphere ice sheets became progressively stronger and longer with time (Fig. 1A). The last million years were dominated by 100,000-years cyclicity, the nature of which is still not fully understood. However, a strong coupling between ice volume and CO_2 concentrations (Fig. 1B) suggests that the latter represent an important feedback in the climate system amplifying and shaping the glacial cycles.

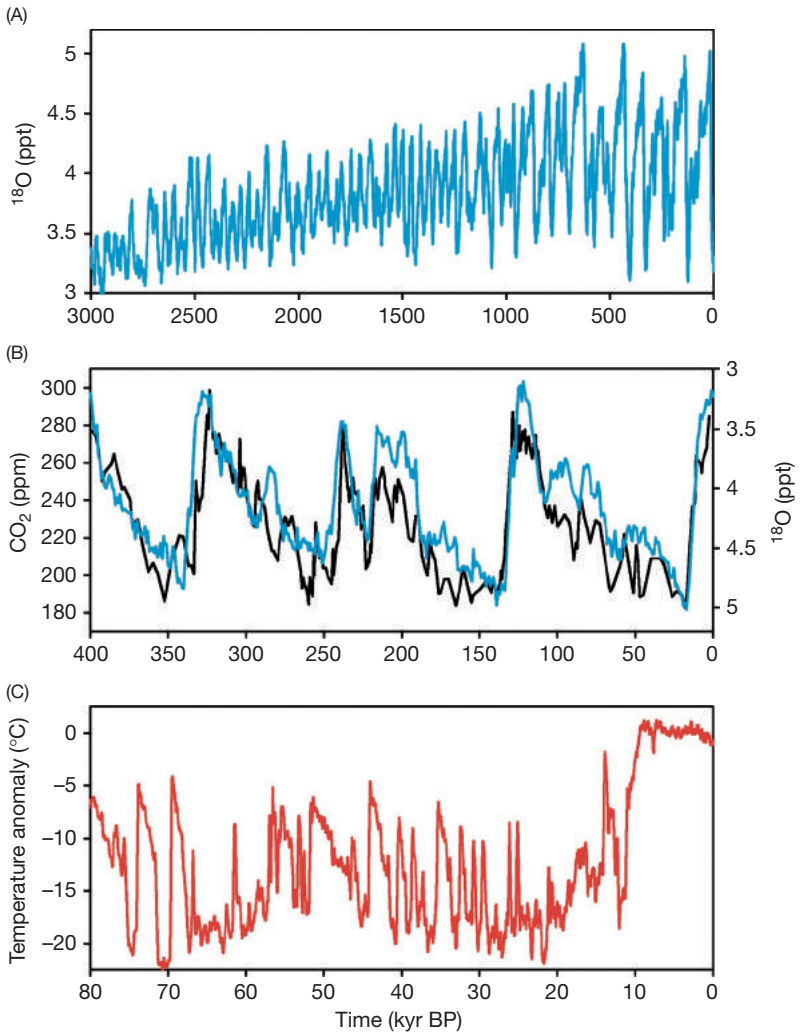


Fig. 1 Past climate changes on different timescales. (A) Temporal variations of ^{18}O isotope concentration in marine sediments over the past 3 million years. This isotope represents a proxy for the global ice volume. (B) Variations of CO_2 concentration in ppm (*black line*) versus variations of global ice volume (*blue line*) over the past 400,000 years. (C) Greenland temperatures from the GRIP ice core during the last 80,000 years. Ice volume data are from Lisiecki, L.E. and Raymo, M.E. (2005). A Pliocene–Pleistocene stack of 57 globally distributed benthic $\delta^{18}O$ records. *Paleoceanography* **20**, https://doi.org/10.1029/2004PA001071; concentration of CO_2 is from Petit, J.R., Jouzel, J., Raynaud, D., et al. (1999). Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* **399**, 429–436; and Greenland temperature record is based on Daansgard, W., Claussen, H.B., Gundestrup, N., et al. (1982). A new Greenland deep ice core. *Science* **218**, 1273–1277.

During the peak of the last glacial cycle, about 21,000 years BP, the globally averaged temperature was about 5°C below present, with a large portion of this cooling explained by a lowering of the concentration of the major greenhouse gases. In particular, atmospheric CO_2 concentration at that time was only two-thirds of its preindustrial value and almost one-half of the current CO_2 concentration.

Abrupt Climate Changes

Analysis of the Greenland ice cores and the North Atlantic marine sediment cores performed in the early 1990s revealed that climate changes in this region were everything but gradual. Reconstruction of Greenland temperature during the last glacial cycle shown in Fig. 1C reveals pronounced instability of the climate system on millennial timescale. Most of this temperature record with a notable exception for the last 10,000 years is punctuated by numerous rapid warming events known as Dansgaard–Oeschger events. These events are characterized by extremely rapid warming with the magnitude exceeding 10°C, that is, more than a half of the glacial–interglacial temperature variations observed in Greenland. Although the strongest climate signal associated with Dansgaard–Oeschger events were recorded in Greenland and in the Northern Atlantic, synchronous climate variations have been found in many other places around the world. Analysis of paleoclimate data and model simulations suggests that these abrupt

climate changes are related to the rapid reorganizations of the Atlantic thermohaline circulation and events of massive iceberg discharge into the Atlantic Ocean from surrounding ice sheets.

Due to very different climate conditions of the glacial age as compared to the present ones, abrupt climate changes which occurred in the past cannot be considered as a direct analog for the future "greenhouse" world but they represent an important evidence for the potential instability of two components of the climate system—the thermohaline ocean circulation and the ice sheets. These two components of the climate system are considered by many experts as the prime suspects for dangerous and irreversible future climate changes.

Climate Change During 20th Century

Historical data based on meteorological observations over the last hundred years clearly show a warming trend on the global and hemispheric scales as well as over most regions and for all seasons. In accordance with observational data, the globally averaged surface temperature has increased by *c*. 0.6°C during the 20th century (Fig. 2). In some regions, especially over the continents in the middle and high latitudes of the Northern Hemisphere, the temperature has risen much faster than the globally averaged.

Although observed temperature rise represents the most direct manifestation of climate change during the last century, other climate characteristics related to the hydrological cycle, cryosphere, and extreme weather events also have experienced detectable trends. These changes are not only important as additional indicators of global warming, but also because they affect natural and

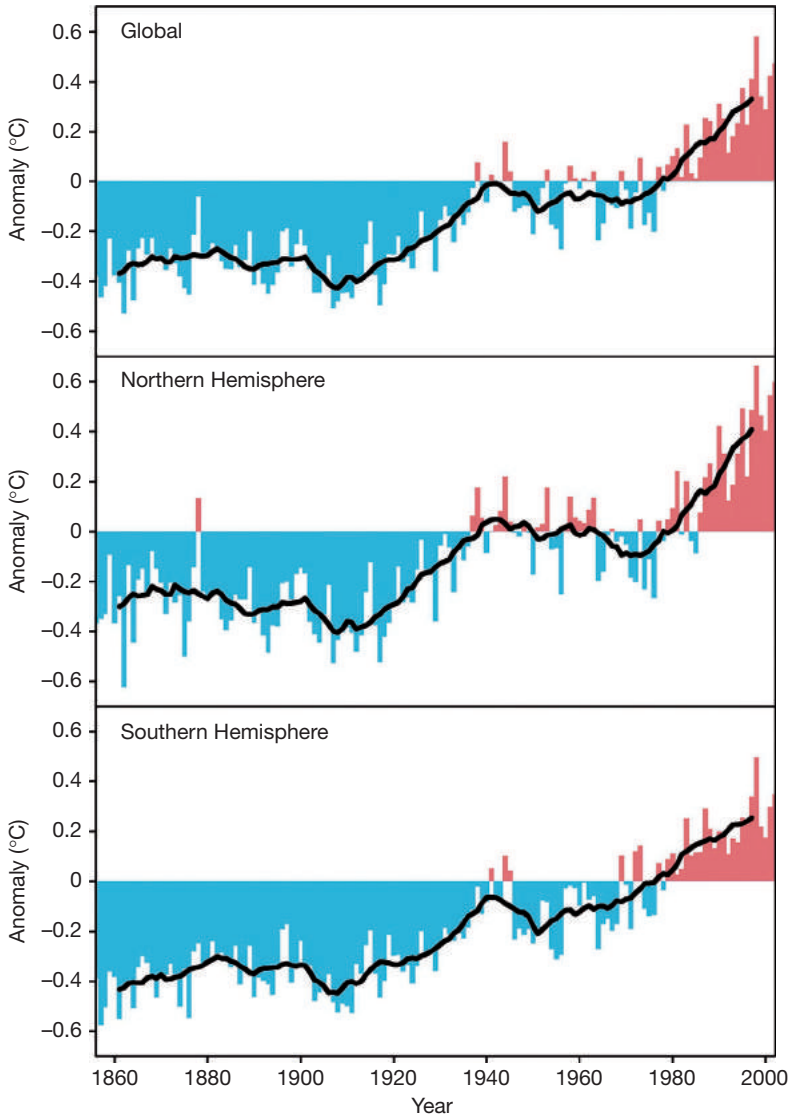


Fig. 2 Global Northern and Southern Hemisphere surface air temperature anomalies over the last 140 years. Color bars represent annual data, while solid line shows 10-year running mean. Temperature anomalies are calculated relative to the averaged period 1950–1979. The data are from the HadCRUT2 database (http://www.cru.uea.ac.uk/cru).

anthropogenic components of the Earth system. In particular, the observations show that the total precipitation over the land area has increased by about 2% during the 20th century, and at high latitudes in the Northern Hemisphere this increase was as large as 5%–10%. However, the increase of precipitation was not uniform during the past several decades. For example, Northern Africa suffered long and devastating droughts.

Global warming leads to a gradual "shrinking" of the cryosphere, that is, a reduction of snow and sea-ice cover. The data show a decrease in snow and sea-ice areas in the Northern Hemisphere during the 20th century. Not only is sea-ice area decreasing, but the ice layer is becoming considerably thinner. These trends are statistically significant and consistent with the results of model simulations. Mountain glaciers are probably the most sensitive to climate changes. It is known that glaciers worldwide have been retreating since the beginning of the 20th century. Because the mass balance of glaciers is affected both by temperature and precipitation, in some areas, where increase of precipitation dominated, some glaciers advanced in recent decades, but the overall number of retreating glaciers is much higher than the advancing glaciers.

Climate Models

The Scientific Basis of Climate Modeling

The most advanced climate models describe the main physical processes in the climate system (atmosphere, ocean, and land surface) and are based on a set of equations for energy, momentum, and mass conservation. Even after being considerably simplified, the governing equations of climate models can only be solved numerically and thereby the development and broad applications of climate models began after the advent of sufficiently powerful computers. Climate models simulate a large set of physical characteristics of the climate system, such as atmospheric and ocean circulation, radiative fluxes, temperature, cloudiness, precipitation, snow and sea-ice cover. Although fundamental physical processes in the atmosphere, the ocean, and on land surface can be described by well-known laws of physics, the enormous complexity of the climate system and limitations of modern computers do not allow to design climate models solely based on the first principles. For example, microphysics of individual water droplets in clouds is well understood but not only individual water droplets, also individual clouds, cannot be resolved in climate models since their spatial resolution is about 100 km at best. Thereby, the description of many important processes in the climate system is based not on the physical laws but on the so-called parametrizations, that is, rather simple, often semiempirical, submodels of individual processes. With the progress in geosciences and growing computer power, these parametrizations become more sophisticated and realistic but the use of different parametrizations in climate models still remains the major source of uncertainties in climate predictions.

Types of Climate Models

There are several types of models used at present for climate research (see also Table 1). The most complex and realistic climate models are the so-called coupled general circulation models (GCMs) of the atmosphere and the ocean. These models are based on the most comprehensive set of dynamical and thermodynamical equations and they describe a large set of relevant processes in the ocean and the atmosphere. The name "general circulation" reflects the fact that these models, unlike more simple models, simulate three-dimensional circulation of the atmosphere (wind speed) and the ocean (current velocity). Climate GCMs originated from the weather prediction models and some models can be used both for weather and climate predictions. The main difference is how the models are used. The weather prediction is aimed on simulations of temporal development of individual weather systems, such as cyclones and anticyclones. Due to chaotic nature of weathers, an accurate prediction of meteorological conditions is only possible on the timescale order of 1 week. On a longer timescale, even small differences in the initial conditions result in large differences in simulated fields. The aim of climate modeling is to simulate climate, that is, the averaged weather conditions. To obtain a sufficiently accurate climate state (i.e., statistics of weather), an averaging of simulated meteorological conditions over at least several decades is required. Current generation of coupled GCMs employs spatial discretization with the resolution of about several hundred kilometers and both the atmosphere and the ocean are divided in vertical direction by several dozens of unevenly spaced

Table 1 Comparison of different types of climate models

Types of climate or Earth system models	*Spatial aggregation and resolution*	*Resolved timescales*	*Number of variables*	*Examples*
Simple models	Box type, 1-D[a] or 2-D	1–10^9 year	1–10	Budyko–Sellers energy balance model, Stommel ocean box model
EMICs	2.5-D or 3-D spatial resolution: 500–1000 km	1 day–10^6 year	10–50	CLIMBER-2, LOVECLIM, UVic
GCMs	3-D spatial resolution: 100–300 km	1 min–10^3 year	>100	CCM3, ECHAM-5 HadCM3

[a]*n*-D: *n*-dimensional model.

levels. Since atmosphere and the ocean are characterized by a number of fast processes, a relatively short time step of numerical integrations—from minutes to hours—is required to guarantee numerical stability and accuracy. This makes coupled GCMs extremely computationally expensive tools which require the use of the most powerful computers.

Another extreme in the spectrum of climate models is represented by simple climate models. Such models describe only a very limited subset of the processes in the climate system and, usually, they employ a very coarse spatial resolution. A prominent example of simple climate models is an energy balance model developed in the mid-1960s. This model is based on one equation for the energy balance of the climate system and it simulates only atmospheric temperature. Atmospheric circulation in this model is parametrized as a large-scale horizontal diffusion. In spite of their simplicity, this class of climate models still remains a useful tool for the analysis and better understanding of some important aspects of climate dynamics, especially related to its nonlinear aspects.

At last, a new type of climate models, the so-called models of intermediate complexity, emerged in the recent decade. These models are aiming on closing a wide gap between simple climate models and GCMs. Design of the models of intermediate complexity represents a compromise between a high degree of complexity required to realistically simulate climate and the necessity to reduce computational cost to perform long-term simulations. Unlike simple climate models, models of intermediate complexity are able to simulate a much large set of climate characteristics, often comparable with GCMs, but due to considerable simplifications of the governing equations and, usually, a much coarse spatial resolution, models of intermediate complexity are suitable for much longer simulations than GCMs. This makes models of intermediate complexity very useful for the study of past climate changes and for long-term (thousand years and longer) future climate predictions.

From Climate Models to Earth System Models

Initially, climate models described only physical processes in the atmosphere, the ocean, and on land surface (Fig. 3A). However, future climate predictions also require modeling of geochemical, biochemical, and biological processes. Indeed, future changes in the chemical composition of the atmosphere, which is the primary cause of anthropogenic climate change, is not known and due to a number of important feedbacks between physical and biogeochemical processes in the climate system, they have to be modeled in consistent matter. For example, at present only one-third of anthropogenic carbon dioxide remains in the atmosphere while the rest is partitioned between oceanic and terrestrial carbon uptakes. This implies that both terrestrial and marine carbon cycles have to be properly modeled. Moreover, terrestrial vegetation not only plays an important role in the carbon cycle but also provides an additional positive climate feedback by alternating surface albedo and other surface characteristics. At the same time, simulation of other important components of anthropogenic changes, such as methane, ozone, and aerosols, requires a detailed description of the atmospheric chemistry. When the simulations on the timescales longer than hundred years are concerned, especially during the glacial age, the dynamics of the ice sheets (both terrestrial and shelf ice) also have to be simulated. Incorporating of all these components into climate models (Fig. 3B) represents an important step from climate models to the comprehensive Earth system models. Currently, a number of the Earth system models based both on GCMs and models of intermediate complexity are used for variety of climate change studies.

Validation and Application of Climate Models

Before application to future climate predictions, it is important to test model ability to simulate observed climate state and past climate variability. Current generation of climate models demonstrates a considerable skill in simulation of different atmospheric and oceanic characteristics as well as their interannual and interdecadal variability, such as tropical climate variability associated with ENSO. This is, however, only the first step in validation of the climate models. Another important step in climate models validation is testing of their ability to simulate different climates and climate change known from observations. For example, a pronounced global warming trend during the 20th century is successfully simulated by climate models when changes in all major climate forcings, both natural and anthropogenic, are prescribed. Paleoclimate reconstructions of past climates present another

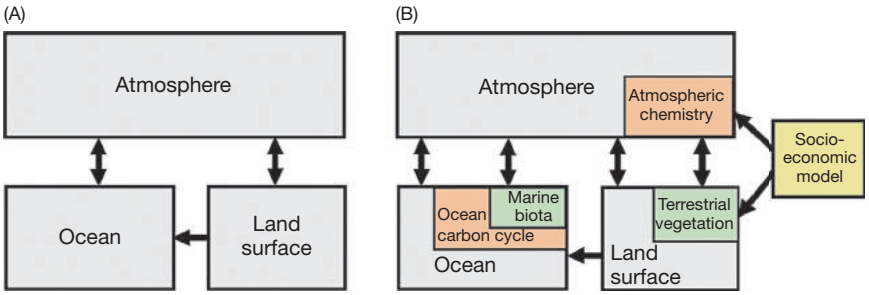


Fig. 3 The principal structure of the coupled climate models (A) and Earth system models (B). *Gray* shading represents physical components; brown, geochemical components; and green, ecological components of the model.

important opportunity to test models under climate conditions different from the present one. For example, climate of the last glacial maximum about 21,000 years ago is relatively well studied and all necessary boundary conditions and atmospheric composition are known for this time with sufficient accuracy. Comparison of model simulations of the last glacial maximum with numerous available paleoclimate reconstructions shows that climate models are able to reproduce major aspects of the glacial climate reasonably well, although some discrepancies between data and model simulations remain and still have to be explained.

Future Climate Change Predictions

Climate Change Prediction and Related Uncertainties

Prediction of future climate change is solely based on the results of computer climate models simulations. In spite of considerable progress in climate modeling achieved in recent decades, large uncertainties remain in predictions of future climate change. These uncertainties stem from several sources. First, to simulate future climate changes, climate models require scenarios for future concentration of atmospheric greenhouse gases, aerosols, and land-cover changes. The latter, in turn, are based on predictions of economic, demographic, and technological development which are rather uncertain. This is why a set of plausible scenarios for anthropogenic climate forcing is produced based on very different assumptions about future socio-economical development. As an example, the upper and lowest projected CO_2 emission at the end of 21st century differs by factor of 6 for the latest set of emission scenarios produced by Intergovernmental Panel on Climate Change (Fig. 4A). These uncertainties in emission scenarios result in a considerable range of uncertainties in the future rise of atmospheric CO_2 and global temperature (Fig. 4B and C).

Another important source of uncertainties in future climate predictions is due to a poor understanding of the radiative properties of atmospheric aerosols. Concentration of several types of aerosol is strongly affected by anthropogenic activity. Among them are sulfate, organic carbon, black carbon (soot), and mineral dust. Some of these aerosols cause cooling, while others cause warming of the Earth's surface. In addition, aerosols affect the optical properties of clouds and hydrological processes in the atmosphere (so-called indirect effect of aerosols). Uncertainties in the direct and indirect effects of aerosols remain very large. Furthermore, unlike the well-mixed greenhouse gases, spatial and temporal aerosol distribution is extremely heterogeneous.

Additional source of uncertainties in future climate predictions arises from the interaction between climate and biosphere. It is still not well understood how natural ecosystems will respond to the combination of climate changes and rise of CO_2 concentration. Some modeling results suggest that terrestrial biosphere under global warming conditions can turn from a sink of carbon, as it is the case at present, to a considerable additional source of CO_2, hence amplifying global warming. Additional methane release from the northern wetlands could also contribute to the amplification of climate change.

At last, different climate models simulate substantially different responses to the same anthropogenic forcing. The globally averaged equilibrium surface temperature response to the doubling of CO_2 concentration is used as a benchmark for the climate model sensitivity to changes of the greenhouse gases concentration. This characteristic, called "climate sensitivity," falls into the broad range between 1.5°C and 4.5°C for different climate models. The reason for such large differences in the climate sensitivity is primarily attributed to the uncertainties related to climate feedbacks, such as water vapor, cloud, and surface albedo feedbacks. Since it is not known which of the climate models is the most accurate for future climate prediction, the whole range of model results has to be used to assess the possible range of uncertainties. Moreover, a possibility remains that the actual climate change may go above the envelope of current climate model simulations.

Predicted Climate Change in the 21st Century

Results of climate model simulations performed for a whole spectrum of possible greenhouse gases and aerosols emission scenarios indicate that globally averaged surface air temperature will rise till the end of the 21st century by additional 1–6°C compared to the present one (Fig. 4). This temperature rise, however, will not be uniform and the warming trend over the continents and high latitudes is expected to be much stronger than the averaged one (Fig. 5A). In the Southern Hemisphere, where the ocean area is much larger than in the Northern Hemisphere, warming will occur at a lower rate.

Another important aspect of global warming is change in the hydrological cycle. All climate models predict an increase in the globally averaged precipitation due to global warming; however, simulated regional patterns of precipitation changes are much less robust and show a low correlation between different models. This is related to a strong spatial variability of precipitation and the large number of factors affecting precipitation changes. Still, some common features can be derived from model simulations (Fig. 5B). In particular, most of the climate models predict the largest increase in precipitation in the equatorial region and middle latitudes, while in the subtropics they predict the precipitation to remain unchanged or even to decrease. Due to the increased contrasts between the land and ocean temperatures, climate models predict considerable intensification of the Asian and African monsoons.

Results of model simulations and analysis of paleoclimatological data suggest that the climate system represents a strongly nonlinear object and its response to gradual changes in external and internal forcing may not necessarily be smooth and reversible. There are several components of the Earth system which are suspected for such strongly nonlinear behavior. One of them is the Atlantic thermohaline circulation which is known to be sensitive to variations in the freshwater flux and did experience major reorganizations during the glacial age. Some model simulations indicate that global warming and associated changes in the hydrological cycles and melting of the Greenland ice sheet can cause a complete shutdown of the Atlantic thermohaline circulations

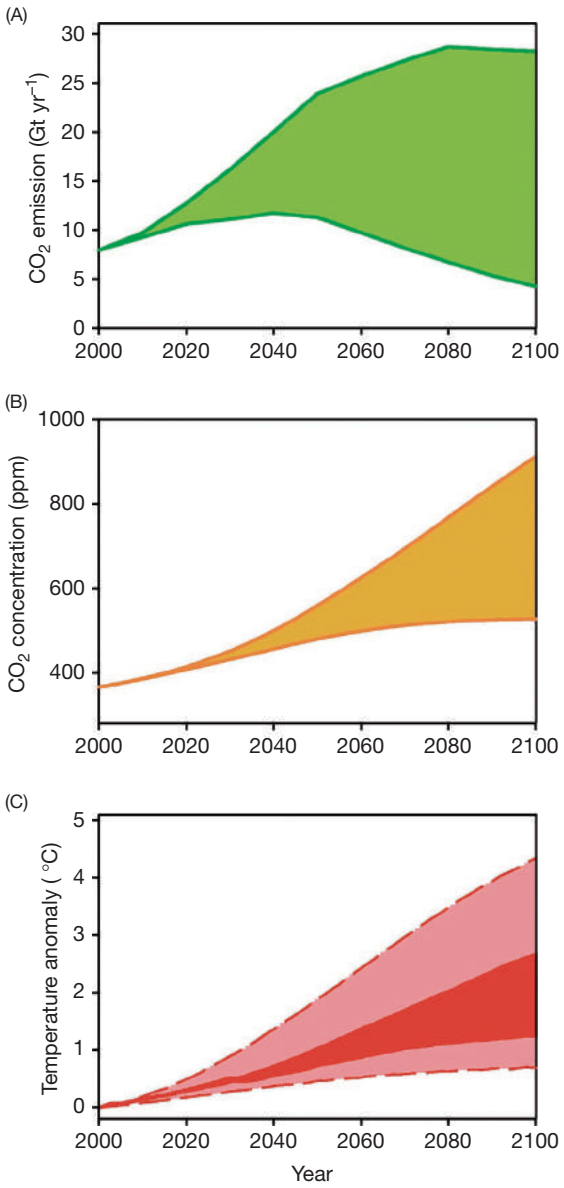


Fig. 4 The range of SRES emission scenarios and corresponding CO_2 concentration and globally averaged temperature changes simulated with the earth system model of intermediate complexity CLIMBER-2. (A) Envelope of the SRES CO_2 emission scenarios for the next 100 years in Gt C $year^{-1}$. (B) Simulated CO_2 concentration scenarios corresponding to SRES emission scenarios in ppm. (C) Simulated range of global temperature changes. *Dark red* area corresponds to the range of emission scenarios with the same climate model parameters. Light red area represents the combined range of uncertainties related to emission scenarios and different climate models. SRES emission scenarios are from http://sres.ciesin.org.

and, as a result, severe and abrupt changes in the regional climate and sea level. The shutdown of the Atlantic thermohaline circulation would also have a pronounced effect on marine ecosystems.

Another component of the climate system which can respond strongly to the future climate change is the West Antarctic ice sheets. There is a possibility that global warming can destabilize the West Antarctic ice shelf, which in turn may trigger abrupt destabilization of the grounded ice sheet. The latter will led to an additional sea-level rise of up to 5 m within several centuries. Due to very complex and still not well understood dynamics of the West Antarctic ice sheet, it is impossible so far to quantify a probability of such collapse, but recent disintegrations of smaller ice shelves and associated accelerations of adjacent ice streams add to the concern about potential instability of the West Antarctic ice sheet.

Although it is still impossible to predict all important consequences of anthropogenic climate change, there is a growing consensus that global warming above several degrees represents a "dangerous climate change," which could pose a danger of severe and irreversible consequences for the human civilization.

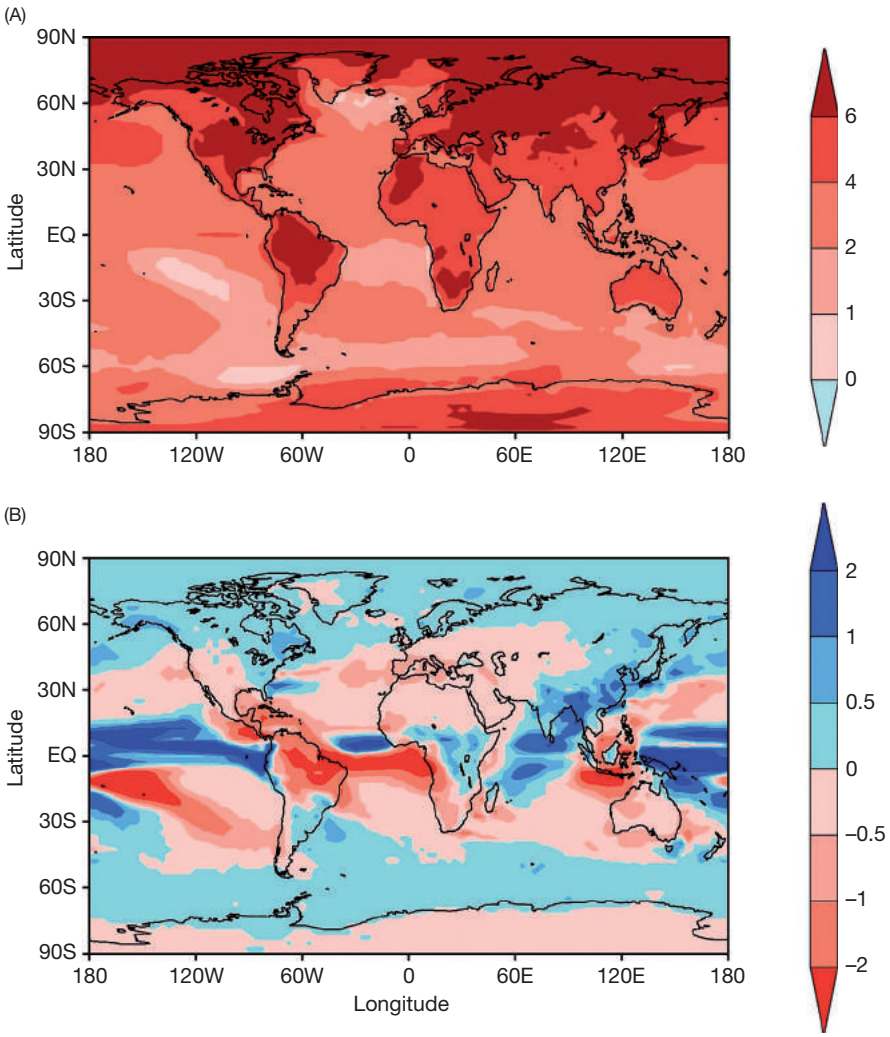


Fig. 5 Mean annual temperature (A) and precipitation (B) changes around the year 2080 compared to the year 1990, simulated with the HadCM3 coupled climate model for SRES A1FI scenario. Temperature changes are in °C. Precipitation changes are in mm per day. The data are from the IPCC data distribution center (http://www.ipcc-data.org).

Climate Impact on Natural Ecosystems and Human Society

Global warming and associated changes in the hydrological cycle and sea-level rise are expected to cause serious negative impact on natural ecosystems, human health, and economy. It is predicted that climate change will disrupt ecosystems and will result in loss of species diversity, as many species will be not be able to adapt to rapidly changing environmental conditions. Some ecosystems, such as tropical montane, mangrove forest, and Arctic ecosystems, are likely to disappear because warmer climate or sea-level rise will not support them. In the high latitudes, warming will cause degradation of permafrost and an increase of methane release from wetlands. Because methane is the next important greenhouse gas after CO_2, this will also amplify global warming.

Simulations with coupled climate model indicate that during 21st century soil moisture in summer will decrease considerably over the large portion of Europe and United States. This could have potentially serious negative impact on natural vegetation and agriculture, and lead to an increase of forest fire frequency. Combination of warming and changes in hydrological cycle will have serious impact on water resources in many regions. Already now one-third of global population is living in water-stressed countries. Unmitigated global warming will considerably increase the number of people exposed to water stress. At the same time, increased probability of extreme weather events, such as catastrophic floods, heat waves, and more devastating hurricanes, are expected to increase the death rate associated with natural disasters.

Sea-level rise will have a profound negative socio-economic impact by increasing the risk of coastal flooding and causing the loss in coastal wetlands. In particular, the estimates show that unmitigated global warming could increase annual number of people in coastal storm surges by factor 10 already in the year 2080. Another potential health impact of global warming is related to the increase of the area where climate is suitable to malaria transmission. Currently, distribution of malaria is limited to the Tropics but global warming could considerably extend this area, which will lead to an increase in the number of people exposed to malaria.

Among recently recognized aspects of rising of atmospheric CO_2 concentration is the acidification of the ocean. The observation and modeling results indicate that carbon dioxide emission from human activity has already led to a reduction of the averaged pH of surface seawater of 0.1 units and pH will fall additionally by 0.5 units by the year 2100. This could lead to mass extinction of coral and some plankton species causing disruption of the entire marine food chain.

Summary

It is established that climate has changed considerably in the past under the influence of natural internal and external factors. However, the recent climate trend revealed by direct instrumental measurements cannot be explained by natural factors alone and a considerable portion of recent climate changes with a very high degree of confidence has to be attributed to the human activity, primarily the emission of greenhouse gases. Growing concern about future climate change stimulated development of climate models, the only tool available for future climate predictions. Current generation of climate models demonstrates a considerable skill in simulation of modern climate and past climate changes. This substantially enhances the confidence in the models ability to provide a reliable picture of the future greenhouse world. Predictions of future climate changes made with numerical climate models clearly demonstrate that unmitigated fossil fuel combustion will lead to an accelerated global warming which represents a serious threat for the well-being of the future generations.

Further Reading

Allen MR and Ingram WJ (2002) Constraints on future changes in climate and the hydrologic cycle. *Nature* 419: 224–232.

Beniston M, Stephenson DB, Christensen OB, Ferro CA, Frei C, Goyette S, et al. (2007) Future extreme events in European climate: An exploration of regional climate model projections. *Climatic Change* 81(1): 71–95.

Claussen M, Mysak LA, Weaver AJ, et al. (2002) Earth system models of intermediate complexity: Closing the gap in the spectrum of climate system models. *Climate Dynamics* 18: 579–586.

Cox PM, Betts RA, Jones CD, Spall SA, and Totterdell IJ (2000) Acceleration of global warming by carbon cycle feedbacks in a 3D coupled model. *Nature* 408: 184–187.

Daansgard W, Claussen HB, and Gundestrup N (1982) A new Greenland deep ice core. *Science* 218: 1273–1277.

Drake JB (2014) *Climate modeling for scientists and engineers*. vol. 19. Philadelphia: SIAM.

Gill SE, Handley JF, Ennos AR, and Pauleit S (2007) Adapting cities for climate change: The role of the green infrastructure. *Built Environment* 33(1): 115–133.

Hallegatte S (2009) Strategies to adapt to an uncertain climate change. *Global Environmental Change* 19(2): 240–247.

J.T. Houghton. Climate change 2001: The scientific basis 2001 Cambridge University Press: Cambridge

Jones PD and Mann ME (2004) Climate over past millennia. *Reviews of Geophysics* 42: RG2002.

Lisiecki LE and Raymo ME (2005) A Pliocene–Pleistocene stack of 57 globally distributed benthic $\delta^{18}O$ records. *Paleoceanography* 20: PA1003.

McGuffie K and Henderson-Sellers A (2001) A forty years of numerical climate modeling. *International Journal of Climatology* 21: 1067–1109.

Neelin JD (2010) *Climate change and climate modeling*. Cambridge: Cambridge University Press.

Pal JS, Giorgi F, Bi X, Elguindi N, Solmon F, Gao X, et al. (2007) Regional climate modeling for the developing world: The ICTP RegCM3 and RegCNET. *Bulletin of the American Meteorological Society* 88(9): 1395–1410.

Petit JR, Jouzel J, Raynaud D, et al. (1999) Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature* 399: 429–436.

Rahmstorf S (2002) Ocean circulation and climate during the past 120,000 years. *Nature* 419: 207–214.

Rockel B, Will A, and Hense A (2008) The regional climate model COSMO-CLM (CCLM). *Meteorologische Zeitschrift* 17(4): 347–348.

Schellnhuber HJ (2006) *Avoiding dangerous climate change*. Cambridge: Cambridge University Press.

Seager R, Ting M, Held I, Kushnir Y, Lu J, Vecchi G, et al. (2007) Model projections of an imminent transition to a more arid climate in southwestern North America. *Science* 316(5828): 1181–1184.

Relevant Websites

cru, n.d., cru.uea.ac.uk. http://www.cru.uea.ac.uk—Climatic Research Unit, University of East Anglia.

ipcc, n.d., ipcc-data.org. http://www.ipcc-data.org—The IPCC DATA Distribution Centre.

sres, n.d., sres.ciesin.org. http://sres.ciesin.org—Special Report on Emissions Scenarios, CIESIN.

Conceptual Diagrams and Flow Diagrams☆

Alexey Voinov, University of Technology Sydney, Sydney, NSW, Australia

Glossary

Conceptual model Presentation of reality in a form of concepts (entities) and interactions (influences) between them.
Mental model These are cognitive representations of external reality. Creating mental models is our only way to explore the world around us. Note that our mental representations of reality are always simplifications, since they do not, cannot and should not take into account all the possible details.
Rich pictures One of the popular graphic ways of presenting a mental model. In addition to concepts and influences it uses icons, images, graphics, text and lines to explain how we think the system works.
System dynamics A modeling approach based on ordinary differential equations used to present how systems change in time. It assumes that reality can be described in terms of stocks (certain amounts of material, energy, money, biomass, individuals, etc.) and flows between them.

Introduction

In most cases he modeling process starts with a conceptual model. A conceptual model is a qualitative description of the system. A good conceptual model is half the modeling effort. To create a conceptual model we need to study the system and collect as much information as possible about the system itself, and about similar systems studied elsewhere. When creating a conceptual model we start with the goal of the study and then try to explain the system that we study in terms that would match the goal. In designing the conceptual model we decide what temporal, spatial and structural resolutions are needed for our study to reach the goal. Reciprocally, the conceptual model becomes important in part to define the goal. In many cases the goal of the study is quite vague, and it is only after the conceptual model is created that the goals of modeling can become clear. Modeling is an essentially iterative process. We cannot prescribe a sequence of steps that takes us to the goal. It is an adaptive process when many times the target is adjusted and moved as we go, depending both on our modeling progress and on the external conditions that may be changing the scope of the study. Building a good conceptual model is an important step on this path.

Temporal Domain

In the temporal domain we first figure out the specific rates of the main processes that we are to model and decide for how long we want to observe the system. If there is little change registered over the study period, the model may not need to be dynamic. It may be static and focus on other aspects of the system. If temporal change is important, we need to identify how this change occurs. In reality, time is continuous. However in some cases it may be useful to think of time as discrete and describe the system using the event-based formalism.

We should start thinking about the appropriate resolution of our temporal model. This resolution is dictated by the goal, which tells us how often we need to update the model to match the expected temporal detail of the study: is once a year is enough as when we model forests, or we need to track the dynamics every 1/100th of a second as when we model the movement of a fly wing. An example of a conceptual diagram depicting the sequence of events in a landscape model is presented in Fig. 1.

☆*Change History*: March 2018. Alexey Voinov introduced small edits in the text of the article including citations.

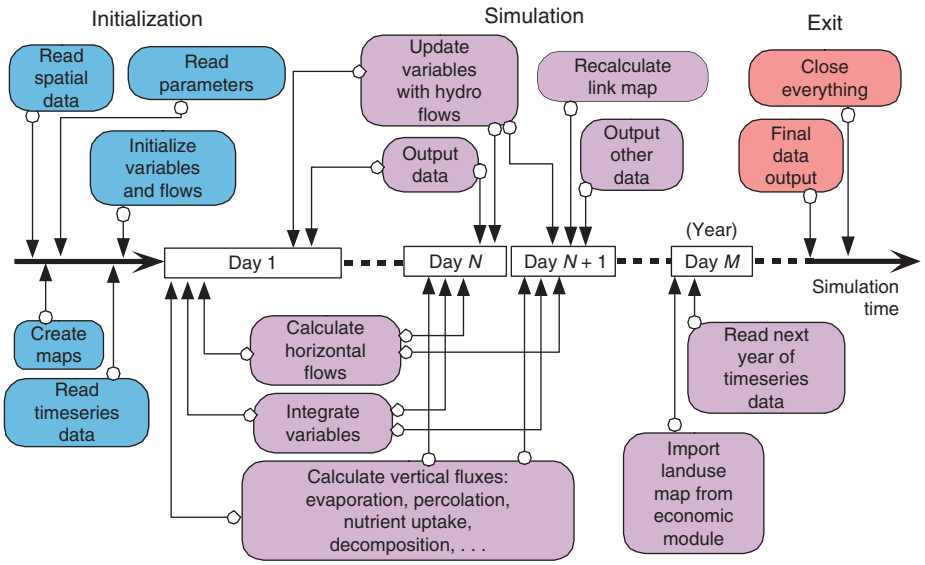


Fig. 1 A conceptual diagram for time in a landscape model. Note that while many processes actually occur at the same time, we need to sequence them when putting in a model. Sequencing of processes is important and may produce quite different results depending upon which processes are executed in the model first, and which ones—next.

Spatial Domain

In the spatial domain we are to make similar decisions about representation of space in the model. Is there much spatial heterogeneity in the system to justify a spatially explicit description? Or the system can be considered spatially homogeneous? Or there are large segments in which the system parts are uniform and the whole system can be described as a collection of these spatially homogeneous compartments. How big are these compartments? Can we describe the system spatially with a map of a 1:1,000,000 scale or should it be 1:100?

In all these decisions diagrams can be helpful to understand the system and to communicate our understanding to others. The picture in Fig. 2 describes how a lake ecosystem is to be modeled in space. It shows that there are large parts of the system that may be considered homogeneous and presented by some average values. The geometry of the lake is described; the central deeper part is considered as a separate compartment and is subdivided in depth. The shallow litoral segments are assumed to be entirely mixed to the bottom and are therefore described by one layer.

Structural Domain

In the structural domain we decide how to represent the structure of the system. In an empirical "black box" model, we are not concerned with what happens inside the system and how. The internal structure in this case is not analyzed, and we describe the system by finding an appropriate function that translates inputs into outputs. This is usually done by statistical methods. In contrast, in a process-based model we describe the structure of a system, which is best done by a diagram, representing the major components of the system: variables, forcing functions, control functions. When deciding about the model structure it is important to match the structural complexity with the goals of the study, the available data and the appropriate temporal and spatial resolution. For example if we are modeling fish populations (Fig. 3), that grow over several years, there would be little use in considering the

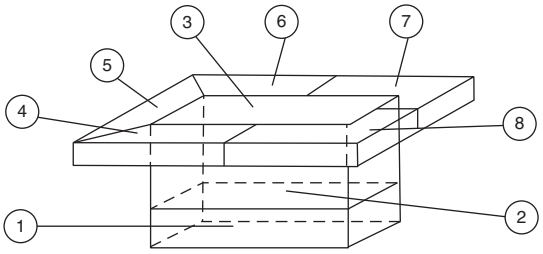


Fig. 2 Spatial representation of a lake described as a diagram. Segment (1)—deep hypolimnion zone, (2)—metalimnion, (3)—epilimnion, (4–8)—shallow litoral zone.

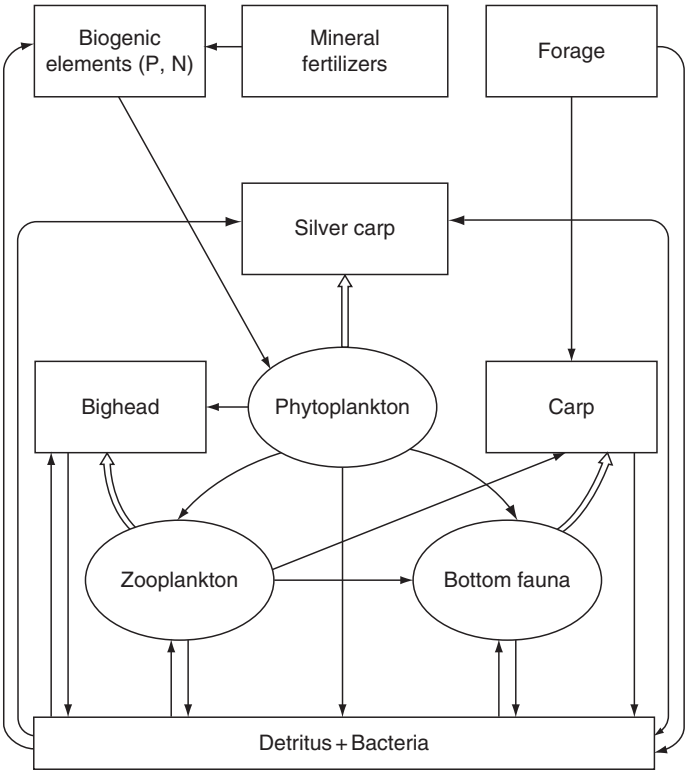


Fig. 3 Structural representation of a fishpond ecosystem.

dynamics of bacterial processes, that have a specific rate of hours. In this case we may probably think of the bacterial population as already at equilibrium, quickly adapting to any changes occurring in the system in the "fish time," which is weeks, or months. We may still want to consider the bacterial biomass for mass balance purposes, but in this case it makes perfect sense to aggregate it with the detrital biomass.

However certain fast processes may have a detrimental effect upon the system. For example, it is well known that fish kills may occur during nighttime in the early morning hours, when there is still no photosynthesis, but only respiration. As a result oxygen content may fall below certain threshold levels. The oxygen concentrations in this case vary from hour to hour, whereas fish biomass changes much slower. We might want to consider oxygen as part of the system to make sure that we do not miss such critical regimes.

In a lake ecosystem model in Fig. 4 in addition to trophic relations certain spatial properties are depicted. The diagram shows how the model structure is presented in the three vertical segments that describe the pelagic part of the lake. In the upper part three phytoplankton groups (A1, A2, and A3) are present, they are food for zooplankton (Z) and fish (R). Various forms of nutrients (organic and inorganic nitrogen and phosphorus) are supplied by decomposition of detritus. In the bottom segments, there is no biota, only nutrients and detritus.

When making all these decisions about the model structure, its spatial and temporal resolution, we should always keep in mind that the goal of any modeling exercise is to simplify the system, to seek the most important drivers and processes. If the model becomes too complex to grasp and to study, its utility drops. There is little advantage in substituting one complex system that we do not understand by another complex system, which we also do not understand. Even if the model is simpler than the original system, it is useless if it is still too complex to shed new light and to add to the understanding of the system.

Forrester put some formalism into the diagrams using a series of simple symbols or icons associated with various processes. The two main ones were level and rate (Fig. 5). Levels were used to identify storages of material or energy, while rates were controlling the flows between them. Using this formalism Forrester created complex models for such systems as cities, industries and even the whole world. Similar formalism was later used in several modeling software packages.

Odum created another set of symbols to model systems based on the energy flows through them. He called them energy diagrams, and used six main icons shown in Fig. 6. All systems were described in terms of energy, assuming that for all variables and processes we can calculate the "embodied" energy. In this case energy works as a general currency to measure all processes and "things."

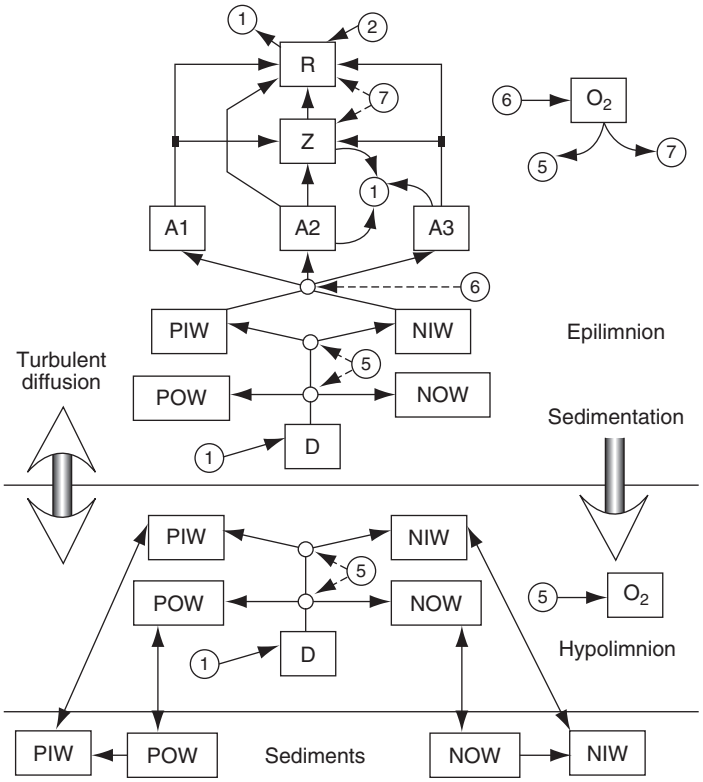


Fig. 4 A conceptual model of a lake ecosystem, which includes elements of spatial and structural description.

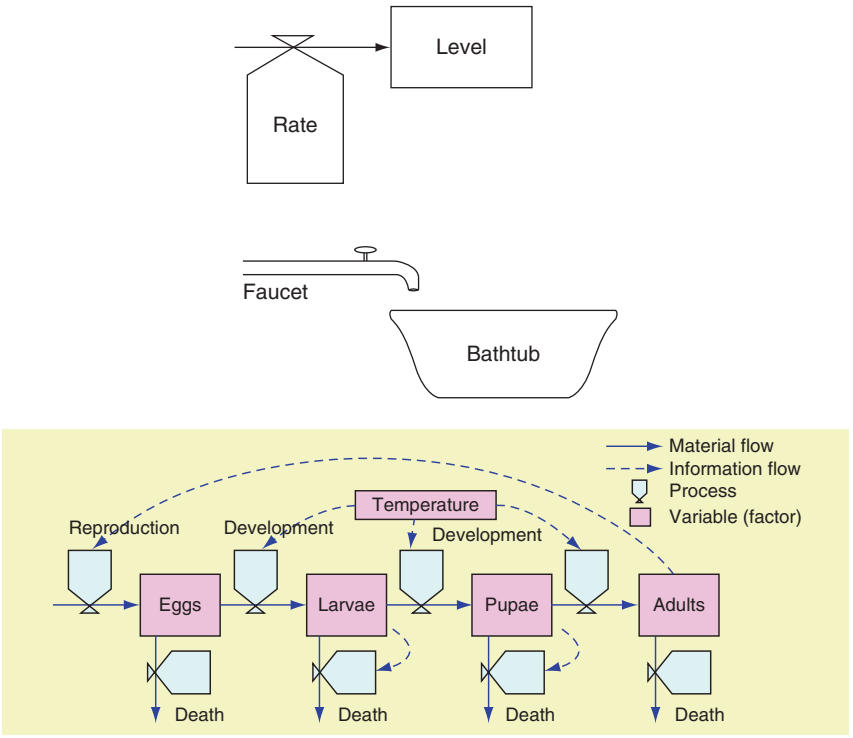


Fig. 5 Forrester's formalism for conceptual diagrams. The rate and the level are two main icons that can be used to put together more complex diagrams such as the one for the insect population. From: http://www.ento.vt.edu/~sharov/PopEcol/lec1/struct.html.

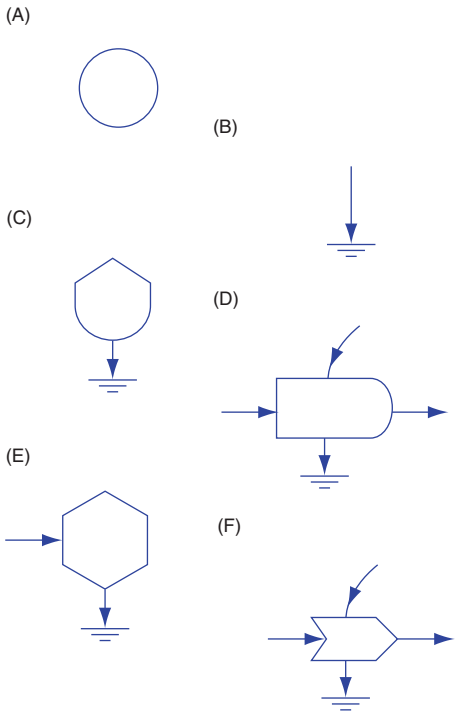


Fig. 6 Odum's formalism for energy-based conceptual diagrams. (A) Source of energy, (B) sink (loss of energy from system), (C) storage tank, (D) production unit (takes in energy and information to create other quality of energy), (E) consumption unit, and (F) energy mixer of work gate.

Modeling Software

In many software packages conceptual diagrams are used as the interface to input the model. For example, one of the reasons that systems dynamics software, such as Stella, became so popular in modeling is because they are also handy tools to put together conceptual diagrams, and, moreover, these diagrams can be then converted into numeric computer models. Fig. 7 presents a sample conceptual model for a river system put together in the Stella interface. It describes the river network as a combination of subwatersheds, river reaches and reservoirs. The Stella interface can be used as a drawing board to put together various conceptual diagrams and discuss them with other people in a process known as participatory modeling. In this case the major value of the interface is that one can easily add or delete variables and processes and immediately see what impact this may have on the model performance. The model itself becomes a tool for deliberations and consensus building.

Very similar diagrams can be put together using other systems dynamics software such as Madonna, Vensim, Powersim or Simile. In these software packages the "stock-and-flow" formalism is used to describe the system. The diagrams are also known as flow diagrams because they represent how material flows through the system.

A somewhat different formalism is used in such packages as GoldSim, Simulink, and Extend. Here we have more flexibility in describing what we wish to do in the model and the model does not need to be presenting only the stocks and flows. Groups of processes can be defined as submodels and encapsulated into special icons that become part of the icon set used to put together the diagrams. As usually we get more functionality and versatility at the expense of a steeper learning curve and higher complexity of design.

Yet another option in building conceptual diagrams is given by the Universal Modeling Language (UML), which is a standardized specification language for object modeling. It is designed as a diagrammatic tool that can be used to build models as diagrams, which can be then automatically converted into a number of object-oriented languages, such as Java, C++, Python, etc. In this case you are actually almost writing computer code, when developing the conceptual model. Once again, even more universality and almost infinite flexibility is achieved at a price of yet more efforts you will need to spend mastering the tool. Fig. 8 presents a sample conceptual diagram created in UML to formulate an agent-based model of a landscape used by sheep farmers, foresters, and National Park rangers that are interacting on very different temporal and spatial scales with different development objectives (sheep production, timber production, nature conservation).

There are several types of diagrams that you can create using UML. One of them is the activity diagram, which describes the temporal dimension of your model. The class diagram presented in Fig. 8 in a way corresponds to the structural dimension, but may also have elements of the spatial representation, like in the Lake model in Fig. 5. Most of software tools designed to create UML diagrams, such as Visual Paradigm would also provide code generators that would convert your UML diagram into computer code in a language of your choice.

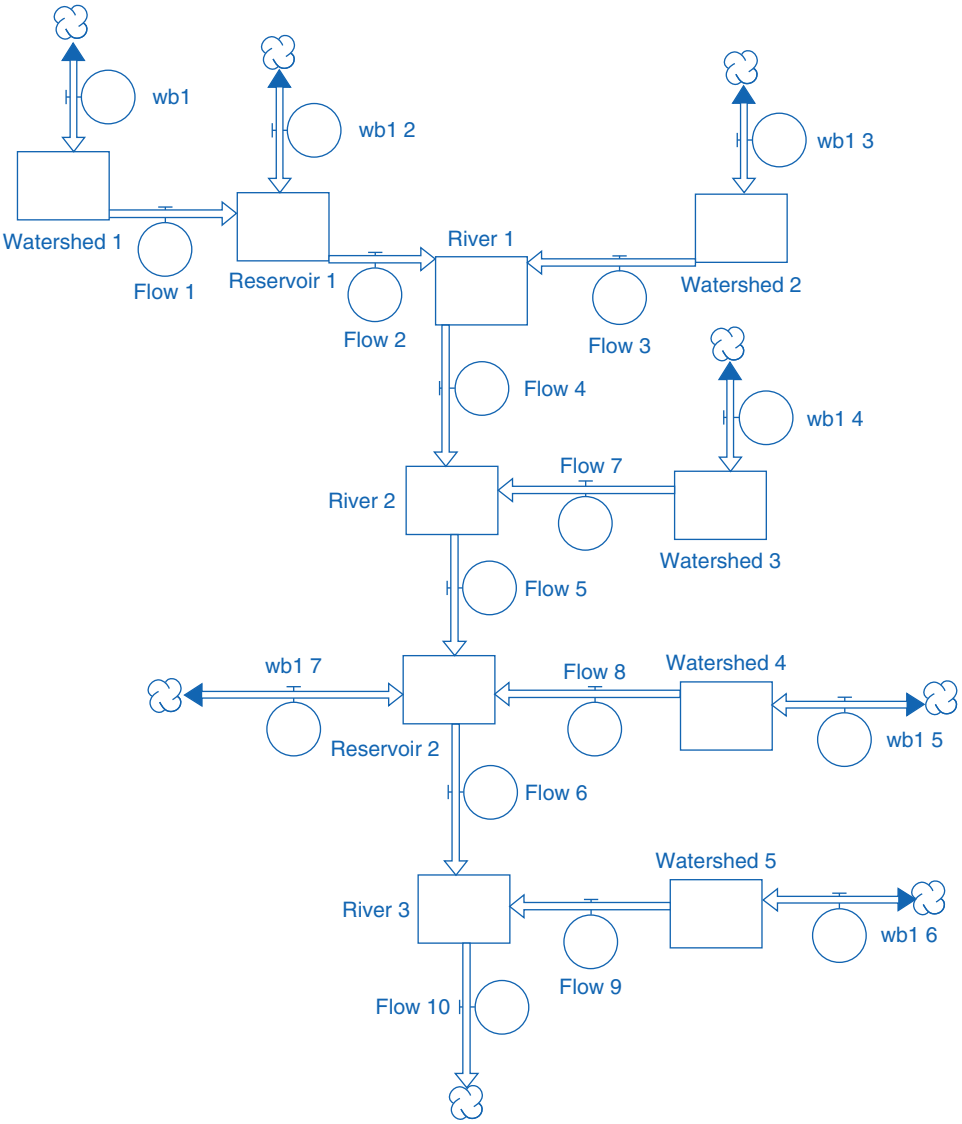


Fig. 7 Using Stella to create conceptual diagrams as stock-and-flow representations of processes in systems.

More recently there have been attempts to standardize the conceptual, diagrammatic representation of system using domain ontologies. A domain ontology represents a certain domain, ecosystem, or part of an ecosystem by defining the meaning of various terms, or names as they apply to those ecosystems. The idea was to define all the various components of ecosystems and present their interactions in a hierarchical way, so that once you need to model some part of the world you could pull out the appropriate set of definitions and connections and have your conceptual model. Several formal languages have been proposed to describe such ontologies. Among them OWL is probably best known, and designed to work over the World Wide Web. So far it is yet to be seen how these ontological approaches will be accepted by the modeling community. Like with other attempts to streamline and automate the modeling process, we may be compromising its most essential part, that is, the exploration and research of the system, its elements and processes, at the level of detail needed for a particular study goal. Any attempt to automate this part of the modeling process may forfeit the exploratory part of modeling and may diminish the new understanding about the system that the modeling process usually offers.

Conclusions

Conceptual diagrams are powerful modeling tools that help design models and communicate them to stakeholders in case of a collaborative, participatory modeling effort. In most cases building a conceptual diagram is the first and very important step in the modeling process.

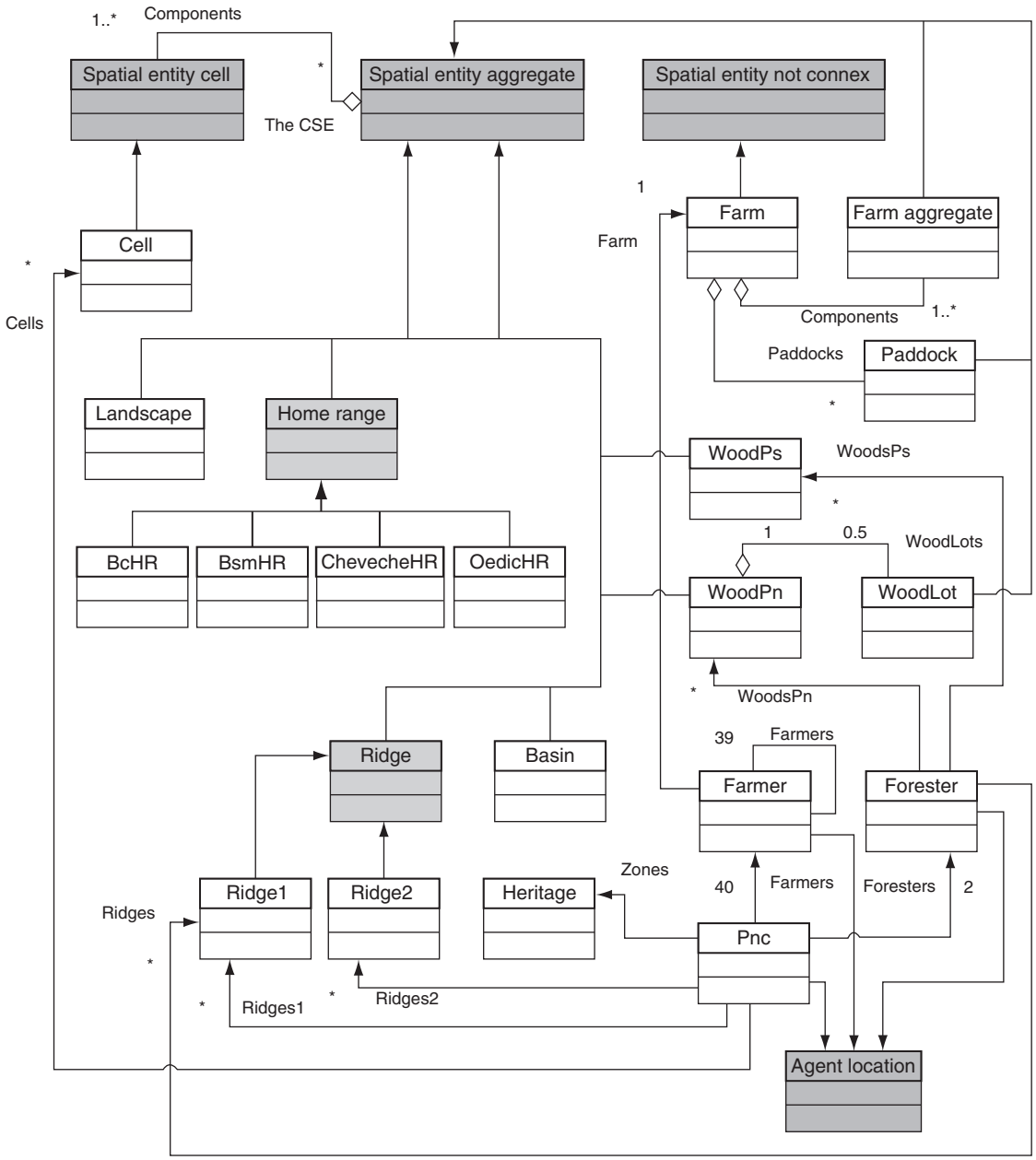


Fig. 8 A UML class diagram of a system can be used both as a conceptual diagram and as a way to program the model. From: http://jasss.soc.surrey.ac.uk/6/2/2.html.

Further Reading

Argent RM, et al. (2014) Best practice in conceptual modelling for environmental software development. In: Ames DP, Quinn NWT, and Rizzoli AE (eds.) *International environmental modelling and software society (iEMSs), 7th Intl. Congress on Env. Modelling and Software, San Diego, CA, USA*, p. 10.
Forrester J (1973) *World dynamics*. Waltham, MA: Pegasus Communications.
Odum HT (1996) *Environmental accounting: EMERGY and environmental decision making*. New York: Wiley.
Voinov A (2008) *Systems science and modeling for ecological economics*. Amsterdam: Academic Press.

Relevant Websites

sparxsystems, n.d., http://www.sparxsystems.com.au/UML_Tutorial.htm.
ian.umces, n.d., http://ian.umces.edu/learn/conceptual_diagrams/.
visual-paradigm, n.d., http://www.visual-paradigm.com/product/vpuml/.

Ecological Models: Individual-Based Models☆

Volker Grimm, Helmholtz Centre for Environmental Research—UFZ, Leipzig, Germany

Introduction

Individual-based models (IBMs) describe individual organisms as autonomous, unique entities. Some IBMs deal with quite simple individuals, which are characterized by just their position. Other IBMs include the individual's adaptive decisions of what to do next, which are based on features of the individuals, for example, age, sex, size, social rank, energy reserves, memory, and of their biotic and abiotic environment. But why should ecological models be based on the representation of individuals in the first place? Models always have to simplify, so why do not we ignore individuals, their variability, and their behavior, and rather consider average individuals as in classical theoretical population ecology?

There are three main reasons why it can be necessary to represent individuals in ecological models: (i) Individual variability. Individuals usually are different, even if they are of the same age. If resources like food or space are scarce, individuals that larger, stronger, or have more experience than others may have a competitive advantage. Moreover, during their life cycle individuals not only change in size but often also in food requirements, behavior, trophic interaction etc. (ii) Local interactions. Most mathematical population models assume global interactions, that is, all individuals interact with all other individuals, but real interactions are local, which can be important. For example, the global density of individuals may be low enough to provide, on average, enough food or space for each individual, but local density may in some places be much higher than global density. (iii) Adaptive behavior. Individuals seek to maximize fitness, that is, survive and produce as many offspring as possible that reproduce themselves. To achieve this aim, they adapt their behavior to the current state of themselves and their biotic and abiotic environment (Fig. 1). Behavior not only includes movement, foraging, mating, etc., but also physiology, growth, development, and life history. Adaptive behavior is very likely to affect or even determine system-level properties of populations, communities, and ecosystems. For example, herbivores like elks change their foraging behavior in the presence of predators, for example, wolves, by selecting other types of habitat. This in turn affects the structure and dynamics of the vegetation and of entire herbivore community, etc.

These three aspects of individual-based ecology are hard to deal with mathematically, for example, using calculus. Therefore IBMs have to be formulated as simulation models that are run on computers. Computers with enough power for running IBMs are generally available since the end of the 1980s, which is the time when individual-based modeling became a declared branch of ecological modeling. Nowadays the IBM approach is widely used in ecology but the term individual-based is used in a very broad sense. Some IBMs include only one individual-based aspect, for example the discreteness of individuals, or local interactions, but otherwise are very similar to classical mathematical models, whereas other IBMs include detailed descriptions of the individuals' fitness-seeking behavior.

In the following we first describe the elements of IBMs, which have to be specified by the modeler. Then we explain a general modeling strategy that can be used to optimize the complexity of IBMs and to deal with uncertainty in model structure and parameters (pattern-oriented modeling). Then we describe how IBMs are analyzed, and finally we briefly present three example models and give an outlook on the future of individual-based modeling. The emphasis of the following is thus more on what IBMs are and how they are developed and analyzed then on what so far has been achieved with the individual-based approach.

☆*Change History:* March 2018. V Grimm updated Synopsis, keywords, Introduction, "Elements of individual-based models," "Implementing individual-based models," Examples, "The future of individual-based modelling," references. No changes to figures and tables.

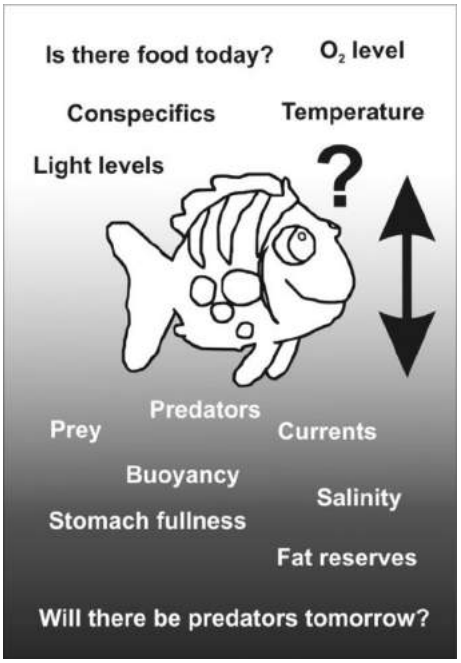


Fig. 1 Real individuals have to make adaptive decisions all the time, i.e., decisions that minimize their risk of starving or being eaten, and maximizing the chance to reproduce. For example, for many fish species vertical migration is a key behavior. Moving up usually means better conditions for feeding, but also higher risk of predation. The very decision made by an individual depends on many internal and external variables, i.e., we assume the fish knows, or at least has estimates, of these variables Modified after Strand, E. (2003). *Adaptive models of vertical migration in fish*. Doctoral Thesis. University of Bergen. Bergen, Norway.

Note that IBMs are referred to as "agent-based models" (ABMs) in disciplines that deal with human agents, for example, economics, social sciences, and demography. In these disciplines emphasis of ABMs always was on the agent's adaptive decisions, whereas most IBMs in ecology so far focus on individual variability and local interactions. But behavioral decisions are increasingly considered also in ecology, so no distinction between "individual-based" and "agent-based" should be made in the future. Likewise, "multiagent systems" are agent-based models which have their roots in computer science and artificial intelligence, but if they are used to tackle ecological problems, they are not fundamentally different from IBMs or ABMs.

Elements of Individual-Based Models

A mathematical model consists of a set of equations, which are formulated in the general language of mathematics. For IBMs, equations and the language of mathematics are only applicable to submodels that describe certain processes, but the entire model consists of further elements, which are listed in the following. This list describes the decisions a modeler has to make while developing and formulating an IBM, and the details of each of these decisions can significantly influence the behavior of the IBM. In 2006, a general format for describing IBMs was proposed, which is now widely used in ecology and increasingly also in other disciplines, the ODD (=overview, design concepts, details) protocol. ODD cannot only be used to describe an existing model, but also as a hierarchical checklist for compiling and designing IBMs. It consists of seven numbered elements, with the first three providing an overview, the fourth being a list of relevant design concepts, and remaining ones providing details of model initialization, data input, and submodels.

1. Purpose

   What is the specific purpose of the model, and what patterns and criteria are used to decide that a model is realistic enough for this purpose?
2. Entities, State Variables, and Scales

   IBMs contain at least two types of entities: individuals and their environment. In the simplest case, there is only one type of individuals, for example one plant species, and a homogeneous environment in which the individuals are located. More complex IBMs consider more than one type of individual, for example a plant species, a herbivore, and a predator, and a more complex environment, for example a heterogeneous landscape consisting of different types of habitat.

   The entities of the IBM are characterized by a set of state variables or behavioral attributes. Individuals might have one or more state variables, for example, numbers representing position, age, sex, size, stage, rank, condition, energy reserves, memory of suitable habitat, etc., or being risk prone, highly mobile, or aggressive. It dependson the question addressed with the model

which state variables are included, but the general guideline is to keep the representation of the individual as simple as possible. The environment often is represented as a grid of square grid cells, which represent spatial units of the landscape and may be characterized by the state variables habitat/nonhabitat, cover, biomass, moisture, temperature, soil type, exposure to wind or predators, etc.

A further important design decision of an IBM is its spatial and temporal resolution and extent. The spatial resolution is determined by the size of the grid cells, which in ecology can range from cm^2 to km^2, and spatial extent is set by the number of grid cells, for example, 100×100 cells. The temporal resolution is given by the length of a time step, which often is a year, a week, a day, or even less than an hour. Different parts of a year may also be represented with different resolution. The temporal extent is set by the time horizon considered, for example, 100 years.

3. Process Overview and Scheduling

   Processes cause changes of the state variables and, in case of individuals, also of the number of entities. Examples of processes in IBMs include energy budgets, growth, dispersal, foraging, habitat selection, mating, habitat dynamics, disturbance events, phenology, mortality, reproduction, etc. The modeler has to decide which process to represent explicitly, and in which detail. This decision is linked to the decisions on state variables and scales. If, for example, temporal resolution is 1 year, the process of foraging may not need to be represented explicitly. If, on the other hand, habitat selection in response to short-term environmental changes is considered important for the problem addressed with the model, a temporal resolution of 1 day or even less might be required, and in turn an explicit representation of the individual's decision where to move in the next time step.

   If processes are not considered important enough to be included explicitly, they usually are represented by constant parameters. Mortality, for example, can explicitly depend on foraging and interaction with predators; or, mortality is represented by a constant parameter, which in the model is implemented as the probability of dying in a certain time step.

   The submodels implementing the model processes are often formulated as a combination of mathematical expressions and IF-THEN conditions. Growth, for example, can be represented by a growth equation, whereas habitat choice will require probabilistic IF-THEN rules: "IF the best habitat within my perception range has higher quality then my current habitat AND IF predation risk is low THEN with a certain probability I will move to the new habitat."

   All design decisions of the modeler regarding entities, processes, scheduling etc. are experimental, that is, they have to be tested and analyzed and, as a result, usually be modified.

   A further important design decision of the modeler is the scheduling of the processes: who does what at what time in what order? IBMs are implemented as computer simulations, so processes cannot, as in reality, run in parallel, but only one after the other. The sequence in which processes are executed can be decisive for the resulting dynamics. An IBMs schedule describes the actions of the IBMs and how they are executed. An action is a list of model entities, the processes performed by these entities, and the order in which the entities are processed. For example, the action "feeding" may be defined as the list of all individuals which feed, one after the other in a fixed order, in their habitat cell. Or, feeding might be an action where all individuals first move to the neighbor habitat which has most food and then feed and where the individuals are processed in a random order.

   Fixed schedules, which are used in most IBMs, define a single order in which events always occur, that is, a cycle which is repeated every time step. Dynamic schedules as in discrete-event simulation allow the order to be changed while the model executes. For example, an individual that has just had an interaction with a predator and survived may put itself on a higher rank in the model's schedule, which would mimic its fleeing behavior. Flow charts, which are often used to illustrate how a model works, usually refer to model processes itself, for example dispersal, but not necessarily to the models schedule and actions. Table 1 shows a typical schedule of an IBM.

4. Design Concepts

   Many IBMs are designed ad hoc, without reference to any general theoretical or conceptual framework. Therefore, general concepts for the design for IBMs have been formulated, which are mainly taken from the research on Complex Adaptive Systems. These design concepts do not require that models necessarily have a certain structure, but their purpose is to make design decisions consciously. The most important design concepts are related to adaptive behavior: emergence versus imposed properties, adaptation (does the model explicitly include behavior decisions), fitness (if adaptation is included, what fitness measures are used by the individuals), what do individuals know, and how do they predict the consequences of their decision alternatives? Further design concepts include: basic concepts, interaction, stochasticity, and collectives.

   Particularly important for model testing and analysis is the design concept "observation." To test whether an IBM's implementation is correct and to understand how system-level dynamics emerge from the individuals' interactions, specific observation tools are needed. These tools include graphical displays of patterns over space and time and the option to execute the model step by step and watch how state variables change. Some software platforms for IBMs like Swarm, Repast or NetLogo provide so-called monitors: on a graphical interface single individuals can be selected and their state variables displayed and manipulated (Fig. 2). Further observation tools are graphical and file output of summary statistics, for example average abundance, point-pattern characteristics of the individual's spatial distribution, or time series characteristics. Observation tools are not part of the model itself, but their appropriate choice determines what we can learn from an IBM.

5. Initialization

   The outcome of an IBM simulation can critically depend on the initial number and state of the model entities. It is therefore important to test different initializations and to carefully document those which were used for the results presented in a publication. Sometimes, dependence on initial conditions is part of the question, for example if we want to understand whether a small, re-introduced population will establish. More often, however, we are interested in results which do not depend on initial

Table 1 Scheduling of the coyote model of Pitt and coworkers as an example of how the schedule of IBMs is organized into actions and IF-THEN rules

Pack actions (executed by all packs):
- Check whether both an alpha male and an alpha female are present
- If both alphas exist, and it is April, produce offspring: create pups, the number of which is stochastic but also depends on pack size, and add them to the pack
- Check whether either alpha is replaced
- If it is December, and there is a contender (another adult of the same sex in the pack), both the male and female alpha coyotes are at risk of being replaced
- Replacement is a stochastic function with the probability of being replaced increasing with the alpha's age
- If replacement occurs, the alpha becomes a transient and the contender becomes the new alpha
- Update the dispersal probability of each member according to its age and pack size
- Force death of pups less than 2 months old if the pack has no adults

Pack member actions (executed by all individual coyotes that belong to a pack):
- If the age of 2 months is attained, leave the den
- If the age of 2 months is attained, change from pup to beta adult
- Update the age-dependent mortality probability and determine whether death occurs
- If individual is a beta less than 2 years old, determine whether it leaves the pack, according to its dispersal probability

Transient coyote actions (executed by all individuals not belonging to packs):
- Update individual's mortality probability, depending on the total number of transients
- Determine whether death occurs

Pack alpha replacement actions (executed by packs that lack a alpha individual):
- If there are beta individuals of the appropriate sex in the pack, promote the oldest beta to alpha
- Otherwise, select a transient of the appropriate sex and promote it to alpha
- If there are no available transients, select a beta from another pack

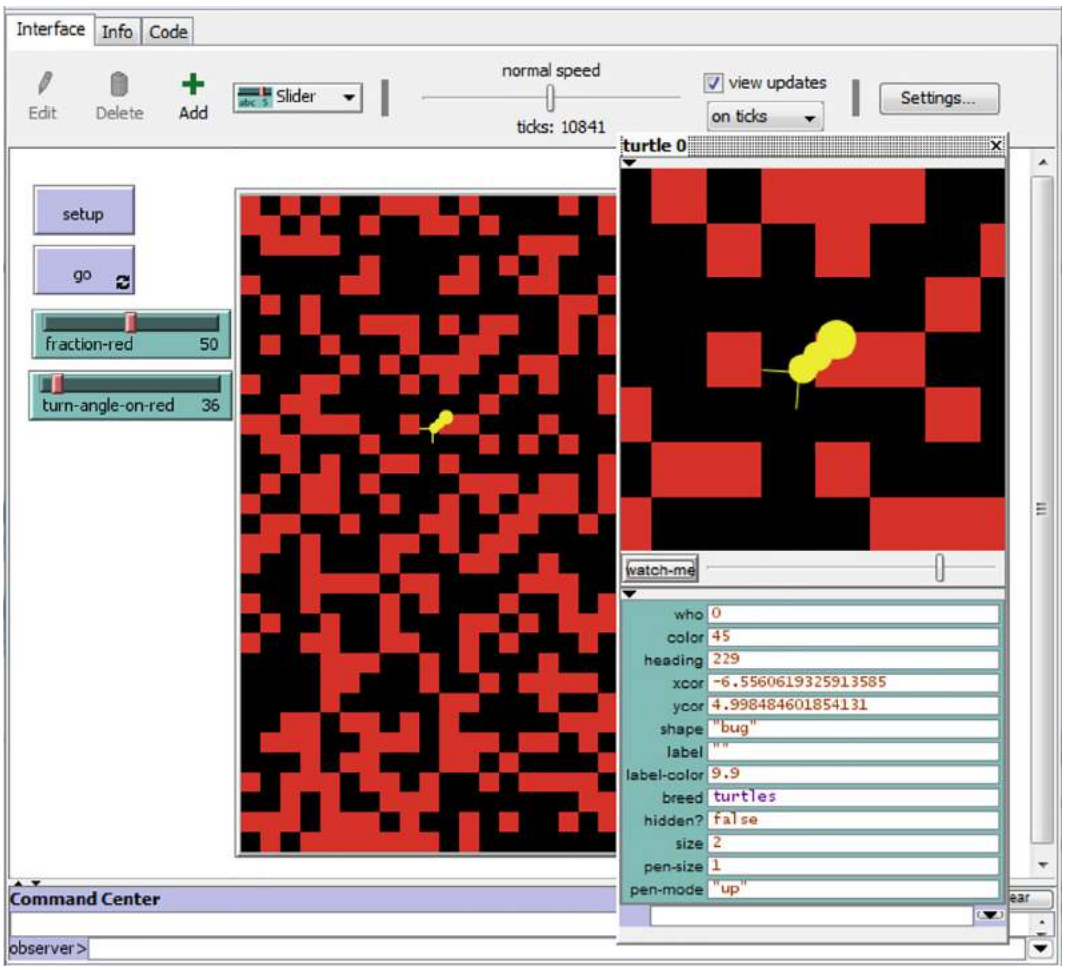


Fig. 2 Screenshot of the interface of a typical NetLogo implementation of an IBM. The interface consists of the grid representing the space, which consists of grid cells (patches), buttons to setup and run the model, sliders to change parameters. Further elements that are added by drag and drop are plots, output windows for text, etc. Individual patches and agents (=individuals) can be inspected (see window labeled ‚turtle 0′), i.e., all their built-in and user-defined state variables displayed while the model is executed. State variables can also be modified interactively. Besides of the Interface tab, the Info tab contains a verbal description of the model, and the Code tab the program, which is using a programming language, NetLogo, that contains numerous commands for dealing with patches and agents (see NetLogo web pages). Note that typical NetLogo applications include more than two types of patches and many agents that interact with each other and the environment, i.e., the patches.

conditions. To achieve this, we can prerun the model until the distribution of the state variables becomes quasi-stationary and evaluate the model only from this time on.

6. Input Data

   IBMs may also be driven by environmental variables, which are not generated by the model itself, but read as an input to the model from external files. Annual rainfall, for example, may be calculated by a rainfall submodel, or, if long-term rainfall data are available, they are taken from an input file. All inputs to an IBM must be carefully documented and, if possible, made available.

7. Submodels

   The processes listed in the ODD element 3, process overview and scheduling, are implemented in submodels which are described here in all detail, using the same submodel names as in the overview and the program implementing the model. "In all detail" means that all information should be provided to allow for a complete re-implementation of the model, so that replication, the cornerstone of the scientific method, is also guaranteed for IBMs. This implies that here also all parameter values used should be listed in a table.

   For complex models, the Submodels section often gets too long to be included in a journal article or book chapter, so that the full ODD has to be provided in an electronic supplement, while the main text only includes a summary description of the Overview part and the most important submodels.

Optimizing Model Complexity and Dealing With Uncertainty

IBMs are more complex than most mathematical population models, because they consider local interactions of individuals that are different and autonomous. This fact has led to the stereotype that IBMs are so complex that they are as hard to understand as nature itself, but this certainly is not true because even the most complex IBMs still are simplified representations of reality. Nevertheless, complexity is a challenge. Many IBMs seem to be more complex than really needed; specific techniques to test and analyze complex IBMs are not routinely used, which severely limits the potential for general insights. However, the general strategy of pattern-oriented modeling (POM) has been formulated which allows to optimize model complexity. POM is also useful for dealing with another important challenge of individual-based modeling: uncertainty of model structure and parameters. How can we know which process to include in a model and how to represent this process, for example dispersal? And how can we determine uncertain parameters?

POM is based on the notion that patterns observed in real systems are indicators of the system's internal organization. If we are able to understand how these patterns emerge then we learned about key processes that determine the systems structure and drive its dynamics. A pattern is defined as anything beyond random variation, or any signal beyond noise. Some patterns are striking, like cycles in the abundance of small mammals, outbreaks of forest insects, or patchy and wave-like spatial patterns. Other patterns are less obvious but nevertheless contain information about the system's key processes: patterns in age and size structure, typical recovery dynamics after disturbance events, the fact that certain system-level state variables like biomass or number of species remain within certain limits, etc.

Patterns are the key to decode the internal organization of any system: atomic spectra helped decoding the structure of atoms, the red shift in the light of galaxies helped formulating the big bang theory, etc. With complex systems, however, that consist of many interacting building blocks, single patterns are not sufficient to narrow down our theories about the system to one single model. For example, there are at least eight models that explain the cycles of small mammals in the boreal zone, and more than 20 theories that explain why ecosystems near the equator have much more species than in other zones.

Thus, for complex systems we need multiple patterns observed at different scales and hierarchical levels. For example, Chargaff's rule of DNA base pairing was not sufficient to decode the structure of DNA. Two additional patterns helped to narrow down the model of the structure to the double helix: patterns from x-ray diffraction of DNA and tautomeric properties of the purine and pyrimidine bases.

Patterns for Model Structure

The most important guide for designing a model is the question addressed. This question, or problem, allows us to decide, in an experimental way, which element of the real system to represent in the model and at which resolution. With POM, in addition we are asking: what patterns can we observe? If we agree that a certain pattern is typical, or even essential for describing the system's identity, we should chose a model structure that in principle allows the same pattern to emerge in the model. This means in particular that we have to include the state variables of the pattern. If, for example, the pattern is spatial, we need to include spatial variables in the model; if the pattern is in size distributions, we need to include size as an individual's state variable; if the pattern is a response to a disturbance event, for example, a drought, we need to include soil moisture as an environmental state variable, etc. Patterns thus make the choice of the model structure and complexity less arbitrary and more directly linked to the internal organization of the real system. The task of the modeler then is to check whether the model is able to reproduce the observed patterns. If not, key processes might be still be missing in the model.

Patterns for Testing Alternative Theories

The first milestone in developing an IBM is a first version of the model that, without too much fine-tuning of parameters, largely behaves like the real system. For example a population which largely has the right size and spatial distribution. The question then is, how can we find the best representation of a certain process, for example habitat selection, mate choice, or life history? The answer is to follow the scientific method of strong inference and formulate alternative models, or theories, of the process and then to check how good the entire IBM, including one of the alternative theories, is capable of reproducing a set of observed patterns. In this way model development becomes less ad hoc, more rigorous, and it becomes easier to communicate why a certain model formulation has been chosen.

Patterns for Reducing Parameter Uncertainty

With IBMs, easily too many parameters are too uncertain, or even completely unknown, to use the model for any practical purpose. It has also been argued that the complexity of IBMs leads to error propagation: even moderate uncertainties of individual parameters would multiply to such a large uncertainty at the system level that the model becomes useless. If, however, the IBM was designed according to POM, it should be structurally realistic, that is, reflect the key structures of the real systems and allow for independent predictions for model validation. This structural realism makes it possible to determine entire sets of unknown parameters by inverse modeling: the model is fitted to the entire set of observed patterns. Technically, this is straightforward: for each unknown parameter, a range and a number of values within this range is specified. Then the set of all possible parameter combinations is constructed, which can be very large. Specific techniques, like Latin Hypercube Sampling, exist for reducing the number of parameter sets, but typically thousands of parameter sets need to be analyzed. The model is run for each parameter set and checked if it is able to reproduce a certain pattern, otherwise the parameter set is discarded. This is repeated for two or more patterns. Eventually, typically less than 50 parameter sets remain that are able to reproduce all patterns simultaneously (Table 2). It has been shown in several examples that uncertainty in model output of this remaining parameter set is largely reduced, so that the model can even be used to support management decisions.

Analyzing Individual-Based Models

IBMs have to be constructed in an iterative way. The first model version should be very simple, for example by making the environment constant, or consider identical individuals, or by representing a certain process by a constant parameter only, etc. This first, or Null, version of the model is deliberately oversimplified. It serves the purpose to start the iterative process of model development as soon as possible. This is done by providing a first set of tools for analyzing the model, for example graphical output, summary statistics, etc. Once we have these observation tools implemented, we can start refining the model, testing its implementation, and comparing it to observed patterns. Analyzing and developing the model then is a time consuming and complex task, but it can be performed as rigorous as real experiments and allows us to understand the relative importance of different processes and how individual behavior is related to system-level properties, and vice versa. In the following, three main elements of analyzing IBMs are explained in more detail.

Currencies

IBMs contain much more information about the state of all the individuals and their environment than we can process. We thus need to aggregate this information into one or more aggregated system-level state variables. We need to define currencies (often also referred to as "indices") that allow us to characterize a single run of the model over a certain time horizon by, ideally, one single number. This number is then used as a currency to compare different parameterizations or formulations of the model. For example, in models addressing extinction risk of small populations, the risk of extinction over a certain time horizon is such a currency; or, if we want to compare model output to an observed time series, we can use root mean square deviation as a currency. In general, any quantitative measure of how good a certain observed pattern is reproduced is a good currency. Often, it is not clear from the outset which currency allows for deepest insights, so currencies are experimental and have to be tested and we usually we will need to consider a set of currencies simultaneously.

Simulation Experiments

Once we have a first version of the model that runs over the time horizon of interest and have both first observation tools and currencies implemented, we can start doing science with the model by performing controlled simulation experiments. As in real experiments, we try to create situations where all but one parameter is kept constant; we try to design experiments whose results are easy to predict just by reasoning. If our predictions fail, we either design even simpler experiments, or we modify our predictions, etc. It is thus useful to consider an IBM as a virtual laboratory. As in real laboratories, we build up understanding step by step. We might, however, not necessarily end up with a full comprehensive understanding of how the model system works. We should, however, be able to say under which conditions the model produces certain types of outputs.

Table 2 Pattern-oriented parameterization of an IBM describing brown bears spreading from Slovenia into the Alps

Pattern description	*Patterns*	*Number of model parameterizations in agreement with observed pattern*
No filter	0	557
Density of females in transition area	1	506
Bear observation in central Austria	2	138
Bear observation in the Carnic Alps	3	154
Bear observation in the Karawanken	4	180
Census time series of females with cubs	5	12
	2 + 3 + 4	13
	5 + 1	10
	2 + 3 + 4 + 1	11

Five different observed patterns are used as filters to reduce the number of possible parameter sets. Note that pattern 5 is as good a filter as patterns 2, 3, and 4 in combination.

Modified after Wiegand, T., Revilla, E., and Knauer, F. (2004). Dealing with uncertainty in spatially explicit population models. *Biodiversity and Conservation* **13**, 53–78.

Robustness Analysis

With many IBMs, we are interested in understanding one fundamental property of real systems, for example persistence of a population, diversity of a community, or coexistence of different life forms, for example trees and grass in savannas. The first task of model development and analysis is to reproduce the desired property with the model. But a model might produce a certain phenomenon only for a very restricted range of parameter combinations, which would make it unlikely that the real system has that property for the same reasons as the model. What we want to achieve are robust explanations that do not depend too much on details of the IBMs formulation and parameters (the same holds for any type of simulation model). For example, IBMs reproducing schooling behavior of fish turned out to be quite robust, even in a quantitative way, to details of how the interactions of neighbor fish is described, as long as the principle mechanism was included that the influence of neighbor fish is averaged in some way. Similarly, complex IBMs of tropical and temperate forests turned out to be quite robust to changes in many model parameters. This robustness reflects internal feedbacks, for example between mortality, recruitment, and dynamics of gaps in the canopy. Such feedbacks are likely to play a similar role in real forests.

Robustness analyses thus means to test the robustness of key model results to changes in model structure and parameters, and also to identify thresholds, for example critical parameter values, or key mechanisms that have to be in the model in order to get the desired model output. Robustness analyses thus helps to check how much of the model's complexity really is needed for explaining the real system's internal organization.

Implementing Individual-Based Models

Implementing an IBM as a computer program, including tools for observation and analyses of a large number of parameterizations and formulations, can be quite challenging. Many tools exist in computer science that support the development and maintenance of complex software, but ecologists usually have no training in computer science. Most developers of IBMs thus are writing their software from scratch, without using concepts and tools of, for example, object-oriented programming. This makes the resulting software prone to errors and its development very inefficient.

There are mainly two alternatives to programming from scratch: (i) Software libraries. The modeler still has to use a certain all-purpose programming language but can utilize many predefined elements that support the implementation of IBMs. Examples of such libraries include Swarm (based on Objective C), Repast (Java), and Mason (Java). These libraries are supported by active user communities and their developers, but often are lacking a comprehensive documentation and tutorials. The learning curve for beginners is quite steep, but for the experienced they provide a very powerful framework for implementing IBMs. (ii) Modeling environments. They consist of menus or simplified programming languages that are easy to use and learn and allow to very quickly developing prototype IBMs. Examples include CORMAS, and NetLogo. NetLogo is freely available on the internet, is well-documented, comes with a good tutorial and many example models, and is actively maintained by its developers. Comparisons with Swarm and Repast showed that NetLogo is much less limited in performance and scope than one might expect due to NetLogo's history as a teaching environment. Being first designed for education and later being recommended for beginners in modeling, NetLogo is now increasingly used for entire modeling projects, sometimes of high complexity. In social simulation NetLogo has become the dominant software platform, while in ecology its use is increasing, a trend that is fostered by textbooks that were published since 2012.

Modern personal computers usually have enough memory and power for developing and running IBMs, but vast analyses of parameter space, for example for parameterization, may need the combined power of PC clusters. For analyzing the output of IBMs, many modelers are using other software packages, for example R, Matlab, Mathematica, Excel, SPSS, etc. For NetLogo, also direkt links to R exist.

Examples

Three example IBMs are briefly described. The coyote model is an example of an IBM that does not explicitly include adaptive behavior. Rather, individuals behave according to empirically determined rates or probabilities. The shorebird model, in contrast, explicitly represents adaptive behavior: each bird is, on a time scale of several hours, making decisions where to move and feed in the habitat. Gap models, finally, have a very long and successful history in ecological modeling of forests.

Coyote Model

The coyote model of W. C. Pitt and coworkers is a good example of an IBM that describes a species with a quite complex social behavior but nevertheless is quite simple. The purpose of the model is to support management decisions. Individuals have the state variables sex, age, social status (alpha, beta, pup), and the group (pack) they belong to. The model considers packs but not territories, and is thus not spatially explicit.

The most important rules of the coyote model are for individuals leaving the group and density-dependent mortality and reproduction. Coyotes between 1.5 and 2 years old have a probability of leaving their pack that is proportional to the square of pack size. Coyotes that leave their pack enter a pool of transients. Mortality of transients increases with the total number of transients and thus tends to be higher than that of pack members. The number of offspring produced is assumed to be density-dependent, that is, to decrease with pack size.

The coyote model was implemented using the Swarm software library. This had the advantage that Swarm's framework for implementing an IBM's schedule could also be used for communicating this schedule (Table 1). This makes is easier to understand and re-implement the model. The schedule also shows that the coyote model consists of actions that are very simple. The complexity of IBMs is more in its implementation and analysis, not necessarily in its formulation.

The model was verified by using five currencies: mean pack size, proportion of transients, average offspring survival rate, average litter size, and proportion of females breeding. Model prediction matched observations surprisingly good, even without fine-tuning of parameters that were taken from the literature. An important insight gained from the coyote model was that the transients are buffering population dynamics. On the one hand they limit, due to their density-dependent mortality, population growth. On the other hand they buffer the loss of alpha individuals.

Shorebird Models

The shorebird models of J. D. Goss-Custard, R. A. Stillman, and coworkers are good examples of IBMs that needed to include adaptive behavior because empirical model rules would not have been sufficient. The IBMs were developed to predict the impact of land reclamation, resource harvesting, and recreation on the winter mortality of species of shorebirds and waterfowl, for example the oystercatcher (*Haematopus ostralegus*) in the Exe estuary in England. The IBMs had to predict the effect of new environmental conditions, for which no empirical rules or data were available. The models had thus to operate on basic principles, that is, physiology and fitness-seeking feeding behavior that is based on adaptive decisions.

The tidal-flat habitat is divided into discrete patches, which vary in their exposure and their quantity and type of food. During each time step birds choose where and on what to feed, or whether to roost. Times steps typically represent 1–6 h. The bird's state variables include foraging efficiency, dominance, location, diet, assimilation rate, metabolic rate, and amount of body reserves. Key environmental inputs to the models are the timings of ebb and flow and temperature, which both affect feeding and the amount of food needed to survive.

A main behavioral process of the model is interference competition (e.g., food stealing), which is related to the individual's dominance status and to local bird density on the patches. The submodels describing the bird's decision where to move, what to eat, and how much time to spend feeding, are based on first principles from optimal foraging theory and game theory. The individuals are assumed to always try and maximize their own chance of survival.

Model predictions were compared with many observed patterns, and after several iterations of the modeling cycle, patch selection, prey choice, and the proportion of time spent feeding were accurately predicted for many species and sites. In one case, the increase in winter mortality due to land reclamation was known from observations. The model was parameterized for the preimpact situation, then run for the situation with reduced feeding area and the increase in winter mortality determined. The match of observed and predicted increase in winter mortality was almost perfect. It could also be shown that the model if it had existed at the time the land reclamation took place, could have been used to recommend a certain mitigation measure that was under discussion but not realized because it was unclear whether it could really compensate the loss of original feeding areas.

The first shore-bird model of the group of Goss-Custard and Stillman needed several years for implementation, parameterization, and testing, but subsequently simpler and more flexible models were developed, that could be used for management support within one to 2 years. Since then, the model has been used for a suite of species of shorebirds, waterfowl, and other bird species at more than 20 different sites all over Europe.

Gap Models of Forests

Gap models describe the change of species composition on a gap in the forest that was created by death of a canopy tree. Typical gap size in early gap models is 0.01 ha. The entire forest is assumed to be an ensemble of gaps with temporally and spatially

independent dynamics. The purpose of gap models is to understand long-term species composition and succession and how they depend on environmental variables. Gap models were thus developed by ecologists, not forest managers.

The representation of the individuals, the trees, is extremely simple: they have only one state variable, trunk diameter at breast height (dbh), which is a standard measure of size in forestry. Tree height is calculated from dbh using an empirical relationship. The structure of gap models is extremely simple: each tree grows according to a sigmoidal potential growth curve. Potential annual growth is then reduced by multipliers which reflect the influence of competition and environmental factors. Competition is only considered vertically and calculated from the vertical distribution of light that is determined by the trees existing in the gap.

Thus, each time step of the model, which usually represents a year, first the vertical light profile is calculated, and then the growth increment. Different species are characterized by different parameters of the growth equation and the relationship between dbh and height. Mortality depends on the growth rate: trees that do not grow for a certain time span, that is, are suppressed by larger trees, will die sooner or later.

Starting from the pioneering model JABOWA, more than 30 gap models have been developed for a wide range of forest types and questions. More recent gap models try to be more realistic in some way, for example by including spatial interactions between neighboring gaps. The great success of gap models has three main reasons: they are conceptually very simple, their growth equations are relatively easy to parameterize, and they make important testable predictions regarding species composition and dynamics of real forests. Modern gap models, which are more complex but also more flexibe, include representations of photosynthesis and thereby link forest models to the global carbon cycle. Using information from satellite images it is nowadays possible to model, for example, the entire Amazon rain forest, taking into account each individual tree.

The Future of Individual-Based Modeling

IBMs are a flexible and powerful tool and very likely to lead to important insights into how system-level properties of ecological systems, for example stability properties (e.g., resilience), emerge from the interactions of the individuals among each other and with their environment. The approach poses also new challenges, in particular optimization of model complexity, coping with uncertainty in model structure and parameters, formulating the models according to a unifying framework, and implementing, testing, and communicating IBMs according to general standards.

IBMs will continue to be used both in a more pragmatic and a more paradigmatic way. Pragmatic IBMs usually do not refer to adaptive behavior and often are designed to be as compatible with more simple mathematical models as possible. There are certainly many questions where this type of IBM is sufficient, because, as we have seen in the coyote model and gap models described above, for many questions we do not need to refer to adaptive behavior explicitly.

Paradigmatic IBMs, however, are based on the assumption that adaptive behavior, that is, the simple fact that individuals adapt their behavior to their current situation and that they are seeking to maximize fitness, is key to understanding most, if not all phenomena at the system level. Paradigmatic IBMs are seen as the nucleus of an Individual-based Ecology which links individual behavior, including life history and phenotypic plasticity, to system-level structures and dynamics. Next-generation IBMs are based on first principles, use information on trait distributions and remotely-sensed data, and are tested and parameterized with multiple patterns, observed at different scales and levels of observation. They make testable predictions about functional relationships, and they are designed to link theory and application via high levels of structural realism.

Further Reading

Botkin DB, Janak JF, and Wallis JR (1972) Some ecological consequences of a computer model of forest growth. *Journal of Ecology* 60: 849–873.
Grimm V, Berger U, DeAngelis DL, Polhill G, Giske J, and Railsback SF (2010) The ODD protocol: A review and first update. *Ecological Modeling* 221: 2760–2768.
Grimm V and Railsback SF (2012) Pattern-oriented modeling: A 'multi-scope'for predictive systems ecology. *Philosophical Transactions of the Royal Society B* 367: 298–310.
Grimm V and Berger U (2016) Structural realism, emergence, and predictions in next-generation ecological modeling: Synthesis from a special issue. *Ecological Modeling* 326: 177–187.
Liu J and Ashton PS (1995) Individual-based simulation models for forest succession and management. *Forest Ecology and Management* 73: 157–175.
Pitt WC, Box PW, and Knowlton FF (2003) An individual-based model of canid populations: Modeling territoriality and social structure. *Ecological Modeling* 166: 109–121.
Railsback SF and Grimm V (2012) *Agent-based and individual-based modeling: A practical introduction*. Princeton, N.J: Princeton University Press.
Rödig E, Cuntz M, Heinke J, Rammig A, and Huth A (2017) Spatial heterogeneity of biomass and forest structure of the Amazon rain forest: Linking remote sensing, forest modeling and field inventory. *Global Ecology and Biogeography* 26: 1292–1302.
Stillman RA and Goss-Custard JD (2010) Individual-based ecology of coastal birds. *Biological Reviews* 85: 413–434.
Strand E (2003) *Adaptive models of vertical migration in fish*. Doctoral thesis, Bergen, Norway: University of Bergen.
Wiegand T, Revilla E, and Knauer F (2004) Dealing with uncertainty in spatially explicit population models. *Biodiversity and Conservation* 13: 53–78.
Wilensky U (1999) *Center for Connected Learning and Computer-Based Modeling*. NetLogo, http://ccl.northwestern.edu/netlogo/. Northwestern University, Evanston, IL.

Ecological Models: Model Development and Analysis☆

Serena H Hamilton, Edith Cowan University, Joondalup, WA, Australia
Susan J Powell, Murray-Darling Basin Authority, Canberra, ACT, Australia
John P Norton and Anthony J Jakeman, The Australian National University, Canberra, ACT, Australia

Introduction

A model is a simplified and imperfect representation of a system, describing only the features essential for the model's purpose. It is crucial for model acceptance and credibility that the modeling process is transparent and follows best practice. The development of a model requires clear agreement on the model purpose and context, involving not only the modelers but also the stakeholders in its use. The process of achieving agreement is iterative, with continual refinement of the purpose and objectives as the modelers and stakeholders learn more about the system to be modeled and the scope for using the model. From that point conceptual models can be developed to guide the choice of model features and families and to help determine how model structure and parameter values are to be found. Performance criteria can then be developed, geared to the model purpose and context, the model structure and the available data. Once the model has been constructed, it must be subjected to calibration, quantification of uncertainty, testing and evaluation of its effectiveness. At any point in the model development it may be necessary to revisit and revise earlier steps as new information unfolds.

This article discusses a 10-step modeling approach (see "Further Reading") in the context of two ecological models developed for a wetland. Modeling should be viewed as a process, and its effectiveness for learning and supporting decision making hinges on how the model is developed, analyzed and reported.

Case Study: Gwydir Wetlands

The Gwydir wetlands comprise a large terminal floodplain wetland complex covering 3000 km^2 in the Murray-Darling Basin in New South Wales, Australia (Fig. 1). The development and regulation of the Gwydir River system, including large headwater storage, diversions, and extractions for irrigated agriculture, has reduced the extent and duration of flooding of the wetlands. Rainfall varies from over 800 mm/year in the upper parts of the catchment to less than 450 mm/year over the wetlands, while potential evapotranspiration can exceed 1400 mm/year. The Gwydir wetlands support a diverse flora and fauna community, including rare, vulnerable and endangered species and migratory bird populations. Part of the wetlands is listed as a Ramsar site of international importance.

Models that represent the ecological response to flows and flooding improve water management by providing a tool to assess management scenarios and facilitate an informed compromise between environmental and human water needs. Here we discuss two ecological models developed in parallel for the Gwydir wetlands. One model represented bird, fish, and vegetation response to flow at a species level; this will be referred to herein as the species model. The other model represented broader landscape unit vegetation productivity response to flooding; this model will be referred to as the vegetation productivity model. Although both were coupled to a water balance model, the two ecological response models were built using different modeling approaches. The models will be described to illustrate the 10-step good-practice framework for model development (see Further Reading).

Step 1: Define the Model Purpose

Modeling involves the systematic organization of data, knowledge and assumptions to fulfill a specific purpose. Models can be used to:

- improve understanding of the system;
- elicit, review and represent knowledge and reveal system properties;
- reveal weaknesses in our knowledge and set research priorities;
- generate and test scientific hypotheses;
- provide a focus for discussion of a problem or simulate further questions about system behavior; and
- forecast or predict outcomes under a range of management or environmental scenarios.

The crucial question in the wetlands case study is "What flow regimes are required to support known ecological assets and maintain the ecological values of the wetland system?" From a water management perspective, this is akin to asking how

☆*Change History*: March 2016. S Hamilton updated all sections including the reference section, deleted Figures 3 and 6 (original version) and added Figure 4 (revised version).

This is an update of S. Powell, J.P. Norton and A.J. Jakeman, Model Development and Analysis, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 2402–2410.

 doi:10.1016/B978-0-12-409548-9.11118-2

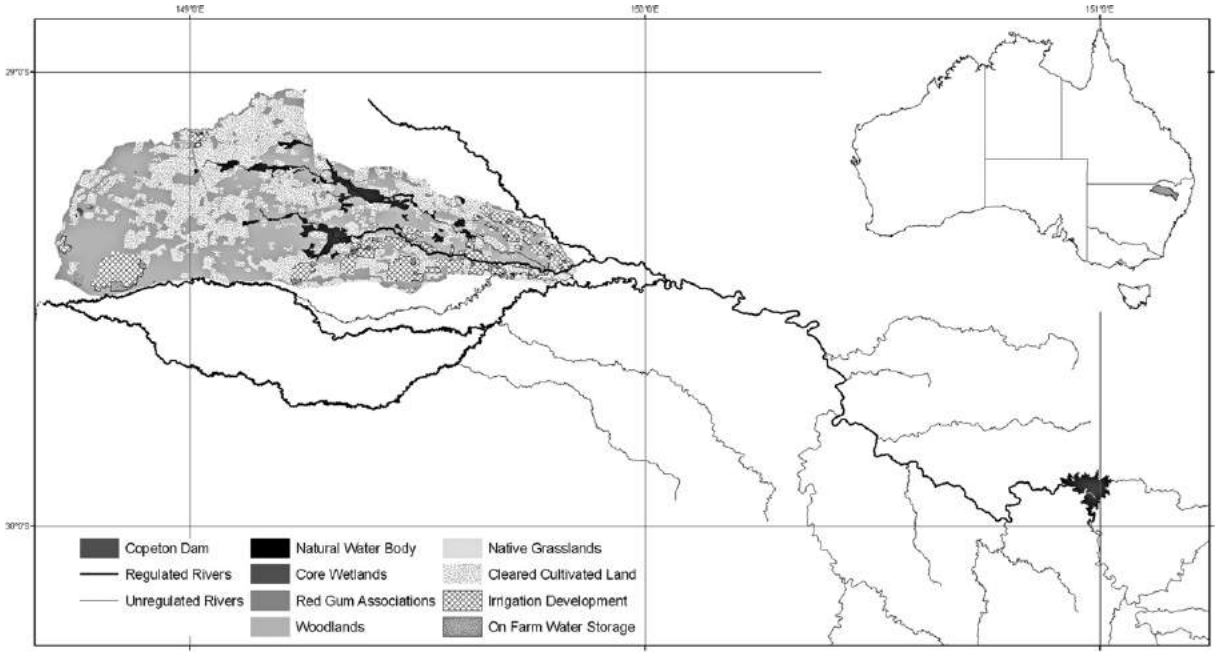


Fig. 1 Location and major features of the Gwydir wetlands.

environmental water can be delivered for greatest ecological benefit. The vegetation productivity model was developed to improve understanding of the flood dynamics of the wetlands and assess possible impacts of inflow and climate change scenarios on vegetation. The species model was developed as part of a decision support system (DSS) for managers of environmental flows into wetlands. The ecological component of the DSS, that is, the species model, was developed to predict the response of various wetland attributes (bird, fish, and vegetation) to environmental flows.

For both models, stakeholders were identified early in the process to ensure relevancy of the model. Stakeholders included representatives of a range of government departments, local landowners, independent scientists, and other interest groups. Funding organizations, technical experts and data suppliers may also be considered as stakeholders. Model end users, in particular, need to be involved in the modeling process from the beginning to make sure they share a common understanding with modelers with regards to the purpose and design of the model.

Ultimately stakeholders may want the model to forecast outcomes (e.g., flood extent and duration, vegetation response) for a range of environmental flow scenarios. The spatial and temporal resolutions must be decided relative to that purpose. Water for the environment is in limited supply and delivery can be both technically and politically difficult. The ability to model a range of water-delivery scenarios provides greater assurance of beneficial outcomes while observing practical constraints.

There are other outcomes of the modeling exercise benefiting all stakeholders, such as better understanding of the system. Partial answers, at least, are obtained for broad questions such as

- How do floods behave and how does flooding affect the vegetation and other ecological communities?
- What are the interactions between flooding, vegetation response and fauna, and which of these processes are of interest to water managers?

For both wetland models, model development also provided a focus for discussion of problems associated with the wetland ecosystem. It enabled a range of stakeholders to define the system and its problems, incorporate prior knowledge and concentrate on the problems rather than being distracted by matters outside their boundaries.

The modeling steps should be revisited if necessary. For example, the first version of the vegetation productivity model revealed gaps in knowledge of flood patterns, vegetation response to flooding and triggers for waterbird breeding, and highlighted the need for further research. The results of that research should allow better definition of the model type, structure, and complexity. This is a cycle of iteration: tentative choices of model scope, type, structure, and resolution determine data needs, which when filled allow refinement of those choices (which may then reveal further needs). The process also includes revision of stakeholder expectations as what is practicable becomes clearer.

Step 2: Model Context

The scope of both models were defined in consultation with stakeholders. For the species model, this was done through a workshop with natural resource managers, scientists, and local landholders. The model process for the vegetation productivity model was

introduced to stakeholders through meetings, position papers, and one-to-one discussion to gain acceptance and elicit advice on specific questions and model boundaries. Questions to which the stakeholders expected answers from the model included

- How much water is required to inundate specific areas (ecological "assets" including Ramsar sites, bird breeding areas, specific vegetation communities and water holes) for a specified length of time?
- What was the flooding pattern and vegetation response in each "asset" prior to and after river regulation?
- Is it just a matter of water volume entering the system, or does timing (for example the daily flow pattern) influence the flood pattern and vegetation response?

Model development included the opportunity for stakeholders to continue to refine their objectives through meetings, focus-group discussions and individual responses.

Spatial and temporal boundaries and scales were also considered at this stage. Selection of appropriate scale depends on factors including the processes of interest and the variability of attributes across time and space. Vegetation response to inundation can be measured in days to weeks, while ecological community structure is a product of longer-term flow and flood patterns (frequency and depth). Spatially the smallest vegetation community covers an area of approximately three square kilometers, but other ecological assets can be smaller. There is also the question of whether to start small (spatially) and build up the model or start larger and refine the model.

Resources in people, time and effort available for the modeling must be identified. The vegetation productivity model project had a timeframe of 3 years, with one researcher. The financial resources available for that project precluded additional staff, expensive or extensive monitoring programs or expansion of the project significantly beyond the identified scope. The species model was developed under a larger project, however the ecological model for Gwydir wetlands was just one of several components in the project. Therefore resources for developing the ecological model were also constrained. Given these constraints, the species model was based on existing models and also designed to be adaptable so that new data and knowledge can be readily incorporated into the model when made available. The species models were based on the likely habitat condition for select species of management interest.

Step 3: Conceptualize System, Data and Prior Knowledge

A conceptual model is an important step in model development, and may be a useful tool on its own. It is used as an abstraction of reality in ecosystems to delineate the level of organization that best meets the objectives of the model. It is not only a list of state variables and forcing functions of importance to the ecosystem and the problem at hand; it also shows how these components are connected by processes. A range of conceptual tools is helpful in this process as summarized in **Table 1**.

With the model objectives and context in mind, the concept for the vegetation productivity model was initially based on a semi-lumped water balance approach as shown in **Fig. 2**. This concept defines the most important drivers of the flood patterns and identifies spatially distributed components that represent different flood dynamics and ecological responses. This water balance component was then linked to vegetation response.

A number of possible concepts were considered for representing vegetation response. Important vegetation species may be modeled in detail if there is sufficient information about their requirements. The flood dependence of each life stage (e.g., germination, establishment, growth, reproduction, death) could be represented and some measure of success devised, based on successful recruitment to the seed-bank or sufficient storage in rhizomes, for example. This approach requires two temporal scales to be considered: the "within-event" scale where day-to-week inundation patterns and vegetation response are important, and multiple "events" which influence the functional groups that respond to many events, ultimately influencing the community structure of the system. The information can be incorporated into a response model as a surrogate for whole-community response, or in conjunction with the functional-group approach. A simpler approach was adopted in the study, based on a surrogate of primary productivity—the normalized difference vegetation index (NDVI). This approach quantified the vegetation (community) response through green-up, maturity, senescence and dormancy. Applying this concept to the study area, multitemporal remote sensing analysis of a flood event clearly demonstrated the vegetation response pattern; not only detecting the green-up, maturity

Table 1 Conceptualization methods

Method	*Description*
1. Word models	A purely verbal description
2. Picture models	A pictorial representation of the system
3. Box models	Boxes represent components; arrow represent processes
4. Input/output models	Same as box models except values for input to and output from the boxes are added
5. Matrix conceptualization	Using matrices to assess possible interactions between components
6. Forrester diagrams	Symbolic language representing variable, parameters, sources/sinks, flows and rate equations
7. Computer flow chart	Flow chart to show the sequence of events in a process
8. Signed digraph models	Plus and minus signs used to represent positive and negative reactions between system components in a matrix
9. Energy circuit diagrams	Designed to give information on thermodynamic constraints, feedback mechanisms and energy flows

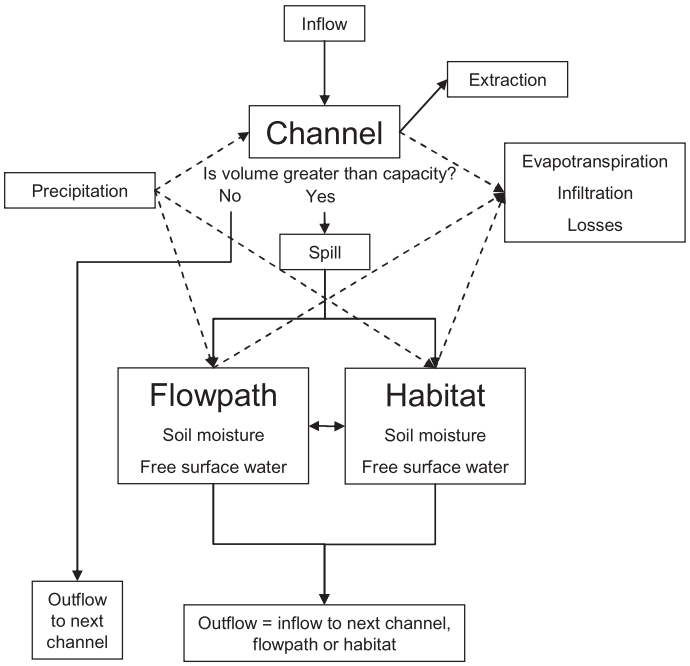


Fig. 2 Conceptual flowchart of channels, flow paths and habitats that can be used to create a semi-distributed model of flooding, soils moisture and vegetation response across a floodplain system.

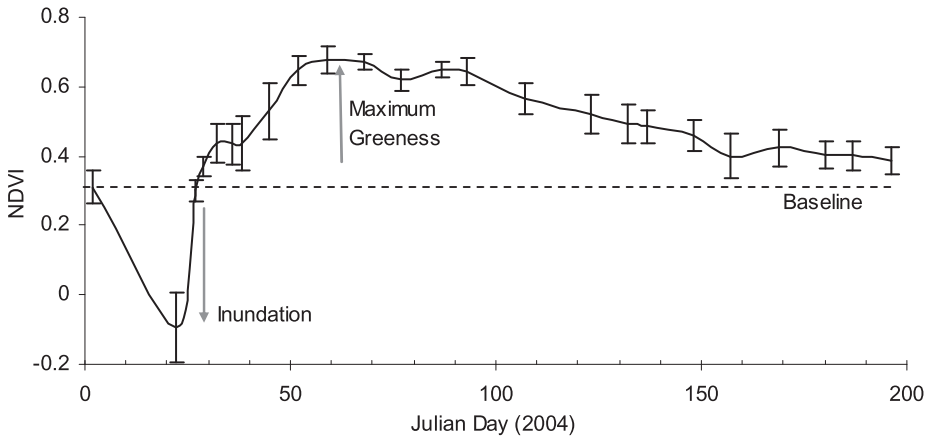


Fig. 3 Vegetation response curve (mean, $\pm$ standard deviation) for a Ramsar site. The initial flooding is shown as normalized difference vegetation index (NDVI) values of less than 0 on day 22. Response to maximum greenness is rapid.

and senescence stages, but also the initial flooding of the area (Fig. 3). Linking the phenology over a range of floods to vegetation functional groups provided the basis for better understanding vegetation community response.

For the species model, the conceptualization step was divided in two stages. Firstly, conceptual models capturing the ecological assets, interactions and drivers of the system were developed by the key stakeholders and DSS end users through workshops. These workshops also identified potential scenarios to be modeled and existing data sets. Secondly, the conceptual models were reviewed by the modelers and updated as appropriate for the DSS. The final conceptual framework (Fig. 4) contained two major components: the hydrological component and the ecological component. More detailed conceptual models were developed for waterbirds, fish, and vegetation. The conceptual models were a useful basis for focusing the scope of the model.

Step 4: Select Model Features and Families

The selection of model features and families depends on the items of interest and the form of model output. Model families and features include:

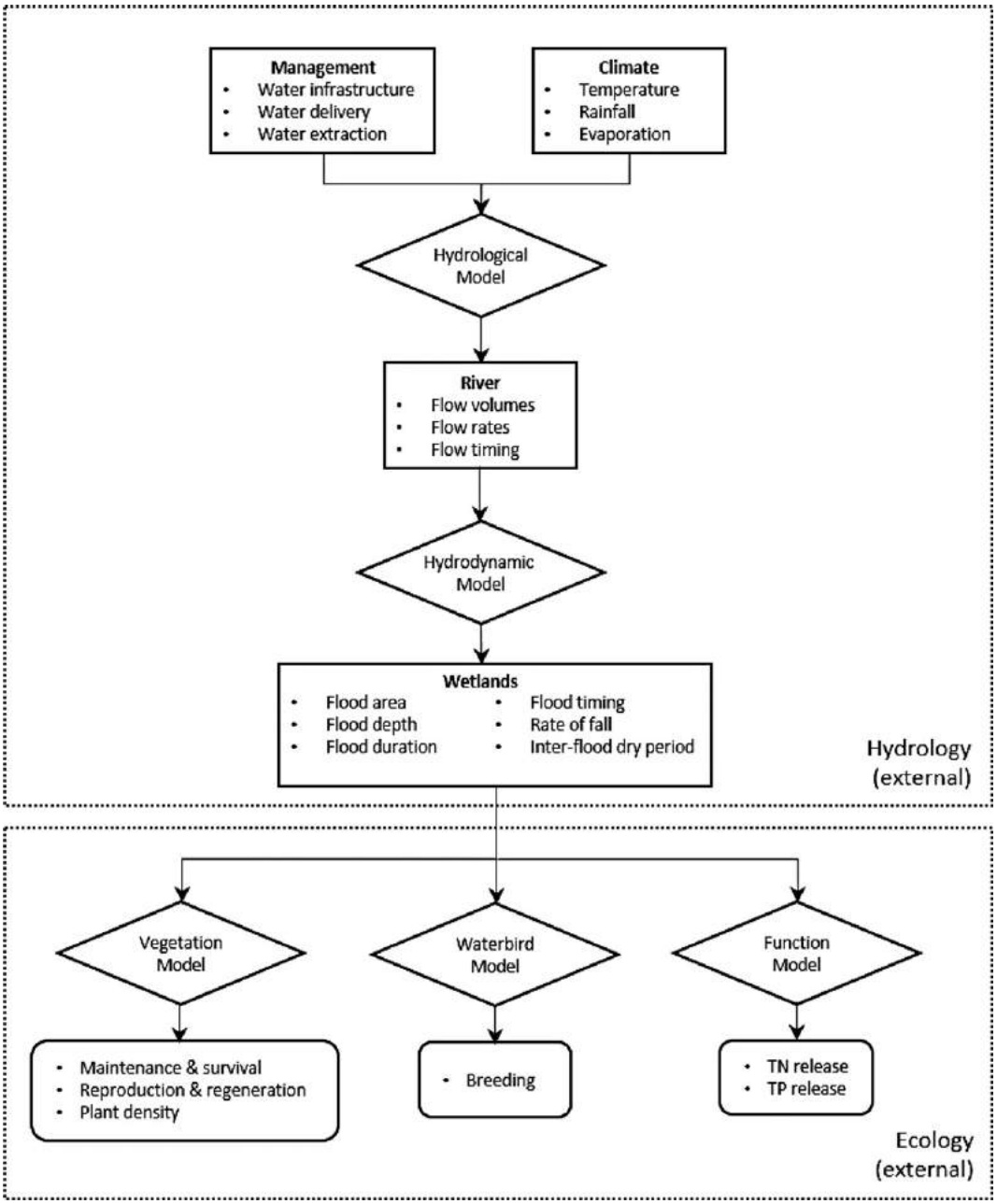


Fig. 4 Conceptual framework of the Gwydir wetlands Decision Support System (DSS). (Adapted from Fu, B., Merritt, W., Pollino, C.A. and Jakeman, A.J. (2010). Decision support system for the Gwydir wetlands—Phase 2. Final report for NSW Department of Environment, Climate Change and Water.)

- *empirical, data-based, statistical models* such as parametric or nonparametric time-series models, regressions and their generalizations such as autoregressive moving-average exogenous models, power laws and neural networks. Such models have detailed structure and parameter values determined exclusively by observational data, rather than selected in advance on the basis of prior scientific knowledge, expert judgment, or custom;
- *stochastic, general-form but highly structured models* which can incorporate prior knowledge, for example, state-space models and hidden Markov models;
- specific *theory-based or process-based models* (usually deterministic, that is, not probabilistic and thus not directly allowing for uncertainty), as often used in environmental physics and economics, for example, specific types of partial or ordinary differential or difference equations;
- *conceptual models* based on assumed structural similarities to the system, for example, Bayesian networks, compartmental models, cellular automata;

- *agent-based models* allowing locally structured emergent behavior, as distinct from models representing regular behavior that is averaged or summed over large parts of the system;
- *rule-based models*, for example, expert systems, decision trees;
- a spectrum of models which represent dynamics, that is, time-spread responses to the inputs at any given instant. This spectrum spans instantaneous (static, nondynamical, algebraic) models, discrete-event and discrete-state models, lumped but continuous-valued dynamical models, and distributed and delay-differential models with infinite state dimension; and
- a corresponding spectrum of spatial treatments, comprising nonspatial, "region-based" or "polygon-based" spatial, and more finely (in principle continuously) spatially distributed models.

For the vegetation productivity model example, the relationships between flow and flood dynamics and flooding and vegetation response to flooding were based on regression analysis. A regression-based model was deemed sufficient, given the good fit found between observed variables. Spatially the model was semilumped into landscape units of similar behavior. The model was to provide a daily time series estimate of productivity as its output.

Bayesian networks were selected as the modeling approach for the species model. In Bayesian networks, linked variables are used to describe cause and effect, and relationships are described probabilistically. This type of model was deemed suitable due to their flexibility in representing components of varying detail, their representation of uncertainty and their capacity to use a broad range of data to populate the model, including expert elicitation. One major limitation of Bayesian networks is their poor representation of spatial and temporal dynamics; this was overcome by linking the networks to a spatially distributed, dynamic hydrological model and dividing the wetland area into "storages," the unit at which outputs were reported.

Step 5: Choose How the Model Structure and Parameter Values Are to Be Found

In finding the structure, prior science-based theoretical knowledge might be enough to suggest the form of the relations between the variables in the model. In the flood dynamics component of the vegetation productivity model, there was insufficient data to allow empirical modeling from scratch, so existing water-balance principles were used. Other parameters in the model were estimated by optimizing the fit of model outputs to observations such as measured water depth at a location, and remote sensing that provided NDVI response as a surrogate for productivity. This approach was used with caution, as experience showed that the river system had been manipulated during past floods, with channel structures altered during the course of a flood or vegetation cleared, grazed, or burnt.

If theoretical understanding is one of the model objectives, and some prior knowledge of system processes exists, then an approach that does not determine model structure solely according to fit to observed data should be favored. Such an approach is very likely to produce a model structure which is an uneconomical summary of the behavior observed in the data, as the structure is partly or wholly dictated by prior knowledge (an empirical modeler would say "prejudice"). Such prior fixing of model structure can impose realistic constraints on possible behavior and may make interpretation of the model parameters much easier, but there is often a conflict between making the structure reflect what is known in advance and making it identifiable (through testable parameter values) from observations. In the vegetation productivity modeling project several parameters were estimated using expert opinion. Experts in the wetland vegetation and ecology were called upon to check vegetation response parameters and model structures. There is little safeguard against misjudgment on their part, however.

Degree of spatial aggregation in the vegetation productivity model was determined by a mapping approach based on identified ecological "assets," major vegetation communities and remote-sensing analysis to identify major flow paths and flood areas. Landscape units that were relatively homogeneous (in their flood regime and vegetation response), but as large (lumped) as practicable to match the objectives of the research and the resolution of the input data were identified and used in the model.

For the species model, the model structures for the waterbird, fish, and vegetation submodels were based on multiple sources of knowledge and data. The fish model structure was developed in collaboration with domain experts, and estimates of the relationship between flow, spawning and recruitment were derived from monitoring data, literature values and expert judgment. The vegetation and waterbird response models were primarily based on an existing database containing the water and habitat requirements of biota in the Murray-Darling Basin. Model structure and parameters can be drawn from various sources including expert opinion where data do not exist.

Step 6: Choose the Performance Criteria and Parameter Estimation Techniques

The criteria by which the model performance is judged should reflect the desired properties of the estimates. This is particularly important in this case study, which must gain the acceptance of a group of nonmodeler stakeholders. Demonstrated lack of bias is important, as there will be significant input from stakeholders who may be perceived to have particular viewpoints or desired outcomes. Stakeholder confidence in the model is best achieved through acceptable prediction performance. This may be a challenge in view of the dynamics of the system; a solution might be to discuss the results with the stakeholders, employing collective memory of past events as well as hard data.

To test for over-parameterization, analysis of the sensitivity of the model outputs to the parameters is useful. It can be performed on the individual components (e.g., the flood dynamics and the vegetation response), but ideally should be performed

on the integrated model. Sensitivity assessment will also help to identify critical parameters whose values may need refining, scope for further lumping, nonlinearities which affect the nature of the responses, and behavior inconsistent with expert knowledge.

Step 7: Identify the Model Structure and Parameter Values

The final model structure should balance sensitivity with complexity and represent the dominant responses of the system at the time and spatial scales of concern. The structure should also ensure that system descriptors such as numbers of variables and processes are aggregated where this makes the representation more efficient. As discussed in step 5, aggregation may be spatial or temporal, or it may be in the way in which ecological response is modeled, focusing on landscape units (or other functional groups) or total productivity rather than individual species. Alternatively, a few specific species may be modeled as indictors of wetland health. Finally, the structure should not be over-flexible as that may result in unrealistic behavior, ill-conditioning and poor identifiability (inability to find well defined parameter values). This should be tested in step 8 if performance criteria are well chosen and verification properly carried out.

Step 8: Verification

Verification of the model structure and parameterization ensures that the model adequately reproduces the observed behavior with regard to the original purpose and context. Common metrics used to measure model performance include root mean square error and coefficient of determination (r^2), however such criteria on their own may be insufficient. Fig. 5 illustrates a simple comparison of outputs from the vegetation productivity model (the inundation depth) against the observed depth at one of the wetland sites. Here measures of model fit such as the root mean square error would indicate a poor fit, but the errors are largely due to mistiming. To address this, model outputs were generated for both daily and 16-day time steps, with the latter producing better model fit. Nonetheless matching of extent, rates and pattern of response may be considered acceptable for the purpose. Testing should also examine the robustness of the model outputs to insignificant changes to data and assumptions.

Assumed physical properties should be plausible, defensible and consistent with prior knowledge (if we are clear about what prior knowledge is genuinely known and what is assumed). In the water-balance component of the vegetation productivity model, for example, assumptions about the behavior of soil moisture were made according to plausible physical processes, but need to be tested and modified as required. Soil moisture and inundation over a floodplain can be modeled but there was no observed dataset to verify. Instead, stakeholders examined the results in the light of personal experience and judged the results to be plausible and consistent with their expectations.

It is also important that the model is tested against statistical knowledge and assumptions, for example, that residuals do not disagree significantly with statistical assumptions, such as absence of systematic structure or significant correlation with the inputs. It is desirable to confirm that parameter estimates have converged, although with short or sparse data sets this may not be possible as was the case in the species model. Excessive variation of parameters with time or location may expose shortcomings of the model structure or the observations. As discussed previously, the entire modeling process must conform with the purpose and context of the model; the verification step is no exception. At this stage, the assumptions and boundaries within which the model seems valid must be clearly established. In the absence of sufficient data for testing, qualitative assessment including peer review can be a useful form of verifying model behavior.

Step 9: Quantify the Uncertainty

Primary sources of uncertainty in a model include errors and finite sample size in the observations, experimental or subjective error in supplied parameter values, and approximation error in the model algorithms and structure. This last source includes both error deliberately incurred in exchange for model simplicity or reduced data needs and, importantly, intrinsic variability in the processes

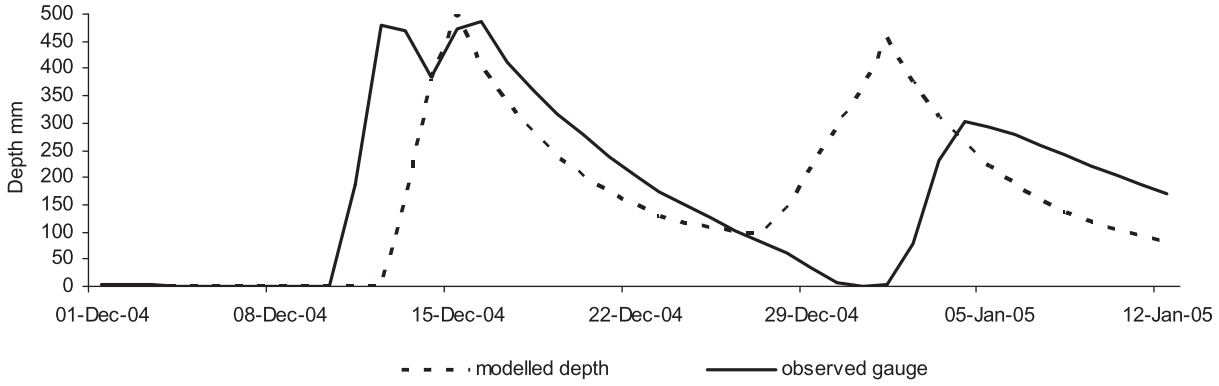


Fig. 5 Modeled and observed depth of inundation for two vegetation communities.

modeled, due to finer-grained processes for which there is no realistic prospect of modeling. In the present examples, this source is prominent and ineradicable.

Bayesian networks explicitly express uncertainties through the use of probabilities, although no distinctions are made between the different uncertainty types and sources. Uncertainties, represented as distributions of probabilities across the states of nodes (variables) in the model, are propagated through the network to the final output. For example, an output of the waterbird model can include the likelihood (%) that good habitat conditions are provided for breeding for a specific species, under specified climate and environmental flow scenarios. These uncertainty values can be valuable for managers, who can use the probability values to weigh up risks of undesirable outcomes associated with management alternatives.

For other model structures such as stochastic models, it is possible to incorporate quantification of uncertainty within the model structure itself. However for model structures such as that used for the vegetation productivity study, uncertainty testing is a separate item. To establish which variables and what types of uncertainty are significant for the model purpose, the results of sensitivity assessment have to be examined, together with estimates of the parameter uncertainties and unmodeled inputs. Some parameter-estimation algorithms (e.g., least squares and its recursive generalizations) provide estimates of parameter uncertainty in the form of error covariances, and can account for observation errors and responses to unknown inputs as "noise" with assumed or estimated statistical properties. Alternatively, fitted parameters can be estimated from different sections of the records and the variation in the parameter values and the output residuals assessed (or approximate probability densities found by resampling, as in bootstrapping). Cross-validation by comparing residuals across various subsamples of the records (see below) is an instance of this process.

Structural uncertainty is often overlooked in modeling, and is potentially large in ecological models partly due to the need to drastically simplify reality. Structural uncertainty is one type of uncertainty not accounted for in the probability values of Bayesian networks. Ideally for any ecological model, one would consider a number of model structures (e.g., based on different hypotheses) to assess the uncertainty associated with the model structure.

Step 10: Evaluate and Test the Model

Finally the model can be evaluated on an independent data set with different input series to test the predictive performance (or cross-validated using a range of subsamples of the original data, in which case steps 9 and 10 are intimately linked). The practical difficulties mentioned in steps 5 and 6, arising from channel or vegetation manipulation in past floods altering the behavior modeled, are typical. Changes in the system (as distinct from its inputs) are common in environmental modeling yet hard to represent, being often episodic and not unambiguously identifiable from the primary records. Auxiliary sources of information, for example, aerial photographs at long intervals, vegetation monitoring and the memories of stakeholders, may be critical in identifying and understanding the changes.

In the context of these case study models, some uncertainties are not readily characterized, especially those stemming from omission or over-aggregation of significant behavior, or incomplete observation records. Moreover, the main performance criteria of the model include its effectiveness as a guide to what water flow regimes will achieve the required ecological values and its value in improving understanding of the effects of floods and regulated flows. Neither is adequately measured by exclusively statistical or other formal means. Consequently, an important approach to model testing (although not the only one) is to look for, explain and if possible rectify anomalies in the outputs produced for realistic input datasets. For example, dummy input flow and climate series may be constructed to test the flood dynamics. The opinions of expert stakeholders can be sought on the plausibility of the responses at selected locations (e.g., flood plains), in crucial periods and overall. Where the results are implausible, the model must be reexamined. Conversely, results accepted as plausible may raise confidence in the fitness of the model for evaluating environmental flow delivery scenarios.

Conclusions

The case studies demonstrate how a best-practice modeling framework helps ensure that modeling is well considered, well documented and transparent. The approach increases the evidence on which the model development may be accepted by the stakeholders. Regardless of a model's final success in relation to its predictive purposes, the 10-step process should ensure benefits to all involved, in the form of better insight into the system, the data and the scope and limitations of modeling.

Requirements for transparency may be summarized as:

- clear statement of the model objectives and end user requirements;
- documentation of the nature, scope and quality of the data;
- a strong rationale for the choice of model families and features;
- justification of the methods and criteria used in calibration and parameter estimation, including readiness to acknowledge, critically, informal or unorthodox methods and criteria where circumstances require them;
- thorough analysis of the performance relative to the resources and application;
- documentation of the model's utility, assumptions, accuracy, limitations, and need and potential for improvement; and
- adequate reporting of the above to inform criticism and review of the model.

See also: Ecological Data Analysis and Modelling: Big Data for Ecological Models; Statistical Inference; Conceptual Diagrams and Flow Diagrams; Sensitivity, Calibration, Validation, Verification

Further Reading

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Jakeman, A.J., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B., Andreassian, V., 2013. Characterising performance of environmental models. Environmental Modelling and Software 40, 1–20.

Chen, S.H., Pollino, C.A., 2012. Good practice in Bayesian network modelling. Environmental Modelling & Software 37, 134–145.

Fu, B., Merritt, W., Pollino, C.A. and Jakeman, A.J. (2010) Decision support system for the Gwydir wetlands—Phase 2. Final Report for NSW Department of Environment, Climate Change and Water.

Fu, B., Pollino, C.A., Cuddy, S.M., Andrews, F., 2015. Assessing climate change impacts on wetlands in a flow regulated catchment: A case study in the Macquarie Marshes, Australia. Journal of Environmental Management 157, 127–138.

Jakeman, A.J., Post, D.A., Beck, M.B., 1994. From data and theory to environmental system model: The case of rainfall runoff. Environmetrics 5, 297–314.

Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. Environmental Modelling & Software 21, 602–614.

Jorgensen, S.E., Bendoricchio, G., 2001. Fundamentals of ecological modeling. Amsterdam: Elsevier.

Kelly, R.A., Jakeman, A.J., Barreteau, O., Borsuk, M.E., ElSawah, S., Hamilton, S.H., Henriksen, H.J., Kuikka, S., Maier, H.R., Rizzoli, A.E., van Delden, H., Voinov, A.A., 2013. Selecting among five common modelling approaches for integrated environmental assessment and management. Environmental Modelling and Software 47, 159–181.

Letcher, R.A., Jakeman, A.J., 2003. Application of an adaptive method for integrated assessment of water allocation issues in the Namoi River catchment, Australia. Integrated Assessment 4, 73–89.

Merritt, W., Powell, S., Pollino, C., Jakeman, T., 2010. IBIS: A decision support system for managers of environmental flows into wetlands. In: Saintilan, N., Overton, I.C. (Eds.), Ecological response modeling in the Murray-Darling Basin. Melbourne: CSIRO Publishing, pp. 85–102.

Norton, J.P., Brown, J.D., Mysiak, J., 2006. To what extent, and how, might uncertainty be defined? Integrated Assessment Journal 6, 83–88. available under.http://journals.sfu.ca/int_assess/index.php/iaj/article/view/9/195

Powell, S.J., Letcher, R.A., Croke, B.F.W., 2008. Modelling floodplain inundation for environmental flows: Gwydir wetlands, Australia. Ecological Modelling 211, 350–362.

Söderström, T., 2000. Model validation and model structure determination. Circuits, Systems, and Signal Processing 21, 83–90.

Forest Models☆

Guy R Larocque, Canadian Forest Service, Quebec, QC, Canada

Introduction

Many models that simulate the dynamics of forest ecosystems have been developed during the last few decades. The term dynamics is meant to represent the attributes of forest ecosystems that describe the changes occurring over time, including growth, productivity, yield, net primary productivity, biomass turnover, or succession. Forest ecosystem models are used as research tools to better understand the mechanisms that govern tree growth and ecosystem dynamics and as decision-making tools to predict the growth of unmanaged and managed forest ecosystems, plan forest management activities, and predict the effects of disturbances. The majority of forest ecosystem models have a relatively complex structure to describe the complexity of the processes that govern the dynamics of forest ecosystems. Forests are characterized by the presence of complex interactions among processes and environmental factors that are regulated by many nonlinear feedback mechanisms.

Forest ecosystem models can be classified into three broad categories. The first category consists of growth and yield models, also known as empirical growth models, and they are used to predict tree and stand productivity. The second category includes process-based models, also known as mechanistic models, and they focus on the modeling of ecophysiological processes that govern the behavior of forest ecosystems. The third category consists of gap models, also known as forest succession models. In this article, the main characteristics of these three categories of forest models are discussed.

Growth and Yield Models

Forests have been managed more or less intensively for the last few centuries to ensure a sustainable supply of stemwood for the production of different goods, including lumber, pulp and paper, or fuelwood. The development of forest management plans requires estimates of forest productivity. Stand tables have been developed for many decades to provide estimates of forest productivity over the life of forest stands, mostly in terms of stem volume. They provide relatively accurate estimates of stand productivity, but only if the growing conditions of forest stands do not change appreciably. As soon as growing conditions change or silvicultural treatments are performed, stand tables become less accurate or reliable. The advent of computers has made it possible to develop sophisticated models to simulate tree and stand growth. One of the reasons that justified the development of such models was to provide forest managers with quantitative tools flexible enough to predict the outcomes of silvicultural treatments. Growth and yield models are developed using stand inventory data collected by forest agencies (e.g., government forestry departments or forest companies), usually through monitored long-term sample plot networks. Predicted state variables include tree or stand attributes, such as stemwood volume or mean diameter at breast height (dbh) at different time cycles (e.g., years) along the development pathways of forest ecosystems. Statistical methods, such as ordinary least squares or maximum likelihood, are used to estimate the parameters of the relationships. Usually, the accuracy and precision of model parameters increase with the availability of historical measurements (tree and stand data collected at different ages). In the literature, the term empirical model is also used to design a growth and yield model. A common definition of an empirical model is a model derived from empirical knowledge obtained from experience or experimentation. Empirical models do not focus on explaining the underlying mechanisms, even though they may have a theoretical foundation or predict biologically consistent patterns of tree and stand growth. Growth and yield models can be classified into two large subcategories: whole-stand models and single-tree models.

Whole-Stand Models

Whole-stand models predict the growth of stand attributes of forest ecosystems, such as basal area, stand volume, or stand density. Basal area is the summation of the cross-sectional area at breast height (usually 1.3 m aboveground) of individual tree stems appropriately weighted to reflect a particular unit area (e.g., 1 ha). Stand volume (e.g., $m^3\ ha^{-1}$) is the summation of the volume of individual tree stems within a forest ecosystem and stand density is the number of trees per unit area (e.g., hectare). An example of a model form to predict basal area over age is

$$\Delta B = \alpha_1 + \alpha_2 T + \alpha_3 B_T + \alpha_1 S + \varepsilon \quad (1)$$

$$B_{T+1} = B_T + \Delta B \quad (2)$$

where ΔB is basal area increment rate (e.g., $m^2\ ha^{-1}\ year^{-1}$), T stand age (year), B_T basal area ($m^2\ ha^{-1}$) at age T, S site index, and

☆*Change History:* March 2018. Todd M. Swannack updated References.

This is an update of G.R. Larocque, Forest Models, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 1663–1673.

ϵ the error. Site index, a measure that estimates the potential productivity of a forest ecosystem, consists of the average tree height at a reference age. For instance, in North America the most common reference age is 50 years and average tree height is based on the average height of the hundred largest trees in dbh per hectare. In this example, basal area at age $T + 1$ is computed by adding ΔB to basal area at age T.

The derivation of whole-stand models to predict the changes in stand attributes over time has been an active field of research for the last few decades. Thus, many models with different forms of mathematical relationships were derived, including linear as well as nonlinear relationships. For instance, the nonlinear Chapman–Richards model has been frequently used in the development of whole-stand models. Different variants were also developed, such as the integration of dbh distribution classes, the derivation of distribution functions to represent the dbh distribution, or the use of transition matrices.

A particular type of whole-stand model that has received much attention is the self-thinning model. The self-thinning model was derived for different types of plant communities, including herbaceous plants and tree species. Self-thinning can be defined as the mortality that occurs in even-aged monospecific ecosystems due to competition among plants. As stand density increases within a forest ecosystem, tree mortality is more likely to occur due to the increase in competitive stress. Indeed, as plant size increases, the intensity of competitive stress accentuates, resulting in mortality. The foundation of the self-thinning model is based on a relationship proposed by Yoda and colleagues in the early 1960s in which the logarithm of mean plant biomass is inversely related to plant population density:

$$\log(\overline{w}) = \log\ (\alpha) + \sigma\ \log\ (N) \tag{3}$$

where $\overline{w}$ is average plant mass, N stand density, α a species-specific parameter, and σ the slope. The work by Yoda and collaborators is largely cited in the literature where the origin of the self-thinning model is discussed. However, similar relationships were also derived by different authors. For instance, Reineke derived in 1933 a stand-density index based on a relationship between stand density and average diameter. Several studies conducted for different types of plant communities, including forest ecosystems, concluded that the slope of the self-thinning model was $-3/2$. The self-thinning model assumes that the species-specific constant and the slope define a maximum limit of plant yield in terms of biomass or size for a given density. This limit is in fact an upper boundary of density-dependent mortality that occurs under competitive conditions in even-aged forest ecosystems. For forest stands, the self-thinning model was widely applied and average plant mass was replaced in many instances by average stem volume. Further research resulted in the development of models of stand-density management diagrams to identify different species-specific development phases related to the intensity of competition. For instance, zones were defined to indicate the initiation of competition, which could be associated with the onset of crown closure, or the lower limit of competition-induced mortality, or to better highlight the asymptotic limit of mean biomass or size as a function of density.

Individual-Tree Models

The dbh, height, basal area, biomass, or volume growth rate of individual trees is predicted by deriving models that represent inter- or intraspecific competitive interactions that occur among a subject tree and its neighbors within a forest ecosystem. The majority of individual-tree models are based on the derivation of competition indices, which are generally classified into spatial and nonspatial indices. This long-used classification distinguishes competition indices that require tree growth data and information about the physical location of individual trees from those that only require tree growth data. Different terms have been used in the literature to define both types of models based on the use of competition indices. For models based on competition indices requiring spatial information, the following terms have been frequently used: single- or individual-tree distance-dependent models, single-tree spatial models, space-dependent models, or spatially explicit models. The terms single- or individual-tree distance-independent models, single-tree nonspatial models, or space-independent models have been used for models based on competition indices that do not use spatial information.

Space-Dependent Models

The derivation of space-dependent models is based on the use of competition indices that integrate spatial information on stand structure to describe the intensity of competition that a subject tree experiences from neighboring trees. The level of detail in the description of competitive interactions varies among competition indices. A common classification of spatial competition indices includes four major groups: (1) amount of overlap between the zone of influence of a subject tree and the zones of influence of competitors, (2) absolute or relative size of competitors adjusted by distance or a distance-based factor, (3) crown interference effect, and (4) potential available area. For the first type of competition index, the zone of influence represents the area within which trees compete for site resources and is a function of the crown width of an open-grown tree of the same dbh to represent the maximum zone of influence of a subject tree and each competitor. Examples of the first three types of competition indices are presented in **Table 1**.

The majority of space-dependent models that were developed using spatial competition indices were derived using multiple linear regression with the following general form:

$$\Delta G_i = \beta_1 + \beta_2 G_i + \beta_3 \mathrm{CI}_{ic} + \beta_4 X + \varepsilon \tag{4}$$

where ΔG_i is the growth rate in dbh, height, tree basal area, volume or biomass of the subject tree; G_i the initial size of the subject tree in terms of dbh, height, tree basal area, volume, or biomass; CI_{ic} the competition index that describes the intensity of competition

Table 1 Examples of three types of spatial competition indices to model competitive interactions among individual trees within a forest ecosystem

Competition index[a]	*Type*	*References*
$\sum_{c=1}^{N_i}\left[\left(\frac{ZO_{ic}}{ZA_i}\right)\left\{\left(\frac{D_c}{D_i}\right)^T\right\}\right]$	Zone of influence	Bella (1971)
$\sum_{c=1}^{N_i}\frac{D_c}{D_i d_{ic}}$	Size/distance of competitors	Hegyi (1974)
$\sum_{c=1}^{N_i} h_c - \left[a\frac{D_{ic}^2}{H_d^{0.25}} + bh_i\right]$	Size/distance of competitors	Vettenranta (1999)
$\sum_{c=1}^{N_i}\left(\frac{CSA_{ic}}{H_i}\right)\left(\frac{BA_i}{BA_c}\right)$	Crown interference	Hatch *et al.* (1975)

[a] a and b, constants; CSA_{ic}, exposed surface of the crown of the subject tree that is not shaded by the presence of competitors; D_c, diameter at breast height (dbh) of a competitor tree; D_i, dbh of a subject tree; d_{ic}, distance between a subject tree and a competitor; N_i, number of competitors for a subject tree; h_c, height of a competitor; h_i, height of a subject tree; h_d, dominant height of a forest stand; H_i, distance between breast height and the base of the crown of a subject tree; T, constant that accounts for species-specific tolerance; ZA_i, area of the zone of influence of a subject tree; ZO_{ic}, area of overlap between the zone of influence of a competitor and a subject tree.

between a subject tree i and the neighboring competitor c, β_n the parameters, and ε the error. X represents any other stand or tree variable that may be included alone or in combination with other variables. For instance, these variables may consist of site index to represent the effect of site quality, age, or basal area or number of trees per unit area to account for the effect of stand density.

Space-Independent Models

Many single-tree space-independent models (i.e., without reference to spatial information on individual trees) were developed using the multiple linear model form listed above (see Eq. 4) to predict individual-tree growth rate in terms of dbh, basal area, height, or volume. Predictor variables consist of tree and stand attributes that may include initial tree size, age, a nonspatial competition index, or other variables representing the effect of stand density or site quality. The derivation of nonlinear models has also been quite common in the last three decades, probably because advances in computer power have facilitated the use of complex algorithms for the derivation of the parameters of nonlinear models. A typical form of an individual-tree nonlinear model is

$$\Delta G_i = \left(\gamma_1 X^{\gamma_2}\left(1 - e^{\gamma_3+\gamma_4 CI_{ic}+\gamma_n Z}\right)\right) + \varepsilon \tag{5}$$

where ΔG_i is the growth rate in dbh, height, basal area, volume, or biomass of the subject tree; CI_{ic} the nonspatial competition index that describes the intensity of competition between a subject tree i and the neighboring competitor c, γ_n the parameters; and ϵ the error. X and Z represent tree or stand variables that influence tree growth.

The effect of competition is commonly modeled by computing nonspatial competition indices that represent the effect of the presence of competitors on the growth of individual trees. Simple competition indices based on the sum of the size of competitors in terms of diameter or basal area have been used quite extensively. Relatively complex competition indices have also been derived (**Table 2**). The effect of the presence of competitors on the growth of the subject tree was also modeled by computing indices or functions that considered in more detail the social status of individual trees within stands, such as the sum of the basal area, the crown surface or volume of the trees within the stand that are greater than the subject tree, percentile in dbh, or tree basal area distribution.

One of the space-independent models that has received much recognition in North America over the last few decades is the Forest Vegetation Simulator (FVS) model. It has a long history of development and calibration for different tree species, forest types, and regions in the United States and Canada. Several variants were developed for specific geographic areas. An important feature of FVS is the development of various model components to simulate different types of silvicultural treatments. FVS has its foundation in the Prognosis model, which was originally developed by Albert R. Stage at the USDA Forest Service in the early 1970s. Several modifications have been made or new components have been implemented over the last few decades to expand its capabilities, including the integration of components from other models, such as TWIGS. The model components of FVS were designed such that simulations could be conducted by using forest inventory or basic stand data to characterize the initial conditions of the forest stands for which growth predictions are desired. For instance, the basic form of the model to predict diameter growth rate of individual trees may include the following independent variables, in addition to a competition index: current tree dbh, stand basal area, site index, or crown ratio (the ratio of crown length to stem height). For some variants, the representation of the effects of site fertility on tree growth was performed by integrating variables describing ecological characteristics, such as elevation or habitat type. Other FVS components include a mortality model to predict individual-tree death, an establishment model to predict regeneration, and other submodels to predict the impacts of disturbances on tree and stand growth, such as insect infestations, diseases, or fire.

Table 2 Examples of nonspatial competition indices to model competitive interactions among individual trees within a forest ecosystem

Competition index[a]	*References*
$\frac{\sum_{c=1}^{N} D_c^2}{D_i^2}$	Stadt *et al.* (2002)
$\sum_{c=1}^{N} \frac{\pi D_c^2}{4}$	Steneker and Jarvis (1963)
$\frac{1-p_c}{\sqrt{10000/N_d}}$	Schröder and von Gadow (1999)
$\frac{D_i\left(g_i+\sum_{c=1}^{N_5} g_c\right)}{D_i g_i+\sum_{c=1}^{N_5} D_c g_c}$	Rouvinen and Kuuluvainen (1997)

[a] D_c, diameter at breast height (dbh) of a competitor tree; D_i, dbh of a subject tree; H, the dominant stand height (m); N, number of competitors around a subject tree; N_D, number of stems per hectare; N_5, number of competitors within a 5 m distance around a subject tree; g_i, basal area of a competitor tree; g_i, basal area of a subject tree; p_c, basal area percentile of a competitor tree.

Process-Based Models

Process-based models, also known as mechanistic models, are structurally designed to describe and simulate the ecophysiological mechanisms in forest ecosystems that govern their growth, productivity, or development. The ecophysiological mechanisms include the processes that regulate carbohydrate production (such as photosynthesis), drive water movement (such as evapotranspiration), or control the synthesis of hormones or the cycle of nutrients (such as nitrogen mineralization). Typical process-based models include descriptions of the effects of environmental factors and their interactions on the processes, including photosynthetic active radiation (PAR) conditions, soil nutrient levels, and climatic conditions. Process-based models are developed by assuming that forest productivity, which can be defined as the allocation of carbohydrates in the different ecosystem compartments (stems, branches, foliage, roots, understory vegetation), and transfer of carbon compounds and nutrients among compartments (e.g., transfer from litter to soil organic matter (SOM)) can be defined and divided into specific interdependent pools and that their properties and behavior can be characterized independently of their functional environment.

Compared with growth and yield models, process-based models are generally considered as powerful tools that contribute to better understanding the processes that regulate the behavior of forest ecosystems, as they are based on the description of cause–effect relationships. Hence, it is largely believed that they have more potential than growth and yield models to predict forest behavior when environmental conditions change significantly, such as the effects of climate change or particulate pollution. For instance, the models that describe the mechanisms involved in the carbon cycle can be used to simulate the changes in the different carbon pools of forest ecosystems under changing conditions of atmospheric carbon and climate. Thus, if the descriptions of the mechanisms are based on realistic mathematical representations, it may be possible to simulate different scenarios of climate change to predict if the CO_2 uptaken by the vegetation may be transferred into pools that may sequester carbon for more or less long periods (e.g., carbon compounds in the SOM) or release it in CO_2 through respiration. Given the amount of details usually included and the number of parameters that must be calibrated, their predictions may be realistic, but not necessarily accurate. On the other hand, the developers of growth and yield models aim at deriving as accurately as possible predictive models using statistical methods on historical data sets, not at explaining causal relationships. Growth and yield models are generally considered "statistically accurate" when environmental conditions do not change appreciably.

Process-based models confer important advantages, in addition to the prediction of the effects of disturbances. Their development, validation, and application may significantly reduce the number of experimental studies and analytical efforts in forest ecosystems. If time and resources were available to undertake experiments under all possible environmental conditions for both unmanaged and managed forest ecosystems, it would be possible to derive the necessary empirical knowledge that would contribute to developing statistical models using historical data sets. However, this approach can be cost-prohibitive for the study of forest ecosystems. For instance, the evaluation of the response of a tree species under all possible combinations of climatic conditions, soil characteristics (e.g., texture, drainage, nutrients), and management scenarios is theoretically possible, but practically and economically unfeasible. The problem becomes cumbersome for the study of mixed forest ecosystems because the interactions that must be considered become more complex and their number increases exponentially. Process-based models may also contribute to predicting the effects of different scenarios of disturbances for which experimentation may be difficult, if not impossible, to achieve using traditional field experimental studies. A good example is the study of the effect of global change on the dynamics of forest ecosystems. Even though experimental sites can be established, such as the Free Air CO_2 Enrichment (FACE) facilities, they can only be implemented under a small range of environmental conditions because of limited resources. In addition, it is not evident that this type of experimental approach appropriately captures the "longevity factor" of forest ecosystems. (FACE facilities consist of experimental sites established in terrestrial plant ecosytems that contain the technology to simulate realistic growing conditions of elevated atmospheric CO_2 concentration for the vegetation.)

Several process-based models have been developed in the last few decades by different groups of modelers (Table 3). They differ in the type and number of simulated processes and level of details, as indicated by the differences in the characteristics of the models provided as examples in Table 3. Among the numerous models, FOREST-BGC, developed by Steven Running and collaborators at the University of Montana, drew much attention. In a way, FOREST-BGC established a benchmark for process-based models of forest ecosystems because several processes were modeled, including several feedback mechanisms, and remote-sensing data were used as input to simulate ecosystem dynamics over large forest regions. FOREST-BGC simulates the main processes in the carbon, nitrogen, and water cycles in forest ecosystems (Fig. 1). For the carbon cycle, model components are included to describe photosynthesis, maintenance and growth respiration, litterfall, litter decomposition, root turnover, and above- and belowground carbon allocation. The nitrogen cycle is tightly coupled with the carbon cycle. Foliage nitrogen retranslocation, soil nitrogen mineralization, and the effect of nitrogen on photosynthesis and leaf growth are represented. Evaporation and evapotranspiration are simulated. Input climatic variables include daily air temperature, solar radiation, precipitation (water or snow), and humidity. FOREST-BGC includes both a daily and yearly time cycle. While the main physiological processes, including photosynthesis, maintenance respiration, evaporation, evapotranspiration, and litter mineralization, are predicted daily, the partitioning of above- and belowground carbon and nitrogen is predicted yearly. Carbon and nitrogen pools are defined for the foliage, stems, roots, and soil. A model like FOREST-BGC can simulate the amounts of carbon sequestered by forest ecosystems. The development of FOREST-BGC was continued and subsequently evolved into descendants, BIOME-BGC and BGC++.

CENTURY is another process-based model that has been widely used for different types of ecosystems. Even though there are model components to simulate vegetation development, CENTURY focuses on the soil carbon, nitrogen, phosphorus, and sulfur cycles. In fact, the physiological processes that occur in the vegetation are not represented in detail in CENTURY. On the other hand, the model components that simulate the processes occurring in the soil are rich in details. SOM is divided into active, slow,

Table 3 Examples of process-based models for forest ecosystems developed in the last three decades

Model	*Cycles*[a]	*Main processes*[b]	*References*
FOREST-BGC	C, N, W	P, I, R_m, R_g, L_f, L_d, F, C_a, C_b, C_s, N_l, N_m, E, N_u	Running and Coughlan (1988) and Running and Gower (1991)
BIOME-BGC	C, N, W	P, I, R_m, R_g, L_f, L_d, F, C_a, C_b, C_s, N_l, N_m, E, N_u	Hunt *et al.* (1996)
CASTANEA	C, W	P, I, R_m, R_g, Rs, Ph, Ca, Cb, L_f, Ld	Dufrêne *et al.* (2005)
CABALA	C, N, W	P, I, R_m, R_g, E, N_m, L_f, C_a, C_b, N_f, N_u	Battaglia *et al.* (2004)
3PG	C, W	P, I, L_f, F, R_t, C_a, C_b, E	Landsberg and Waring (1997)
CENTURY	C, W, N, P, S	L_f, L_d, E, N_m, N_u, F, R_m, R_s, P_u, S_u, N_c, C_a, C_b	Parton *et al.* (1983, 1987)[c]
EFIMOD 2	C, N	I, C_i, C_n, L_f, L_d, F, C_a, C_b, Cs, N_f, N_m	Komarov *et al.* (2003)

[a]C, carbon cycle; N, nitrogen cycle; P, phosphorus cycle; S, sulfur cycle; W, water cycle.

[b]C_a, aboveground carbon allocation; C_b, belowground carbon allocation; C_i, competition for light; C_n, competition for nitrogen; C_s, soil carbon mineralization; E, evapotranspiration; F, fine-roots turnover; I, photosynthetic active radiation interception; L_d, litter decomposition; L_f, litterfall; N_l, foliage nitrogen retranslocation; N_m, soil nitrogen mineralization; N_u, nitrogen uptake; P, photosynthesis; pH, phenology; P_u, phosphorus uptake; R_g, growth respiration; R_m, maintenance respiration; R_s, soil respiration; R_t, total physiological respiration (maintenance + growth respiration); S_u, sulfur uptake.

[c]These two references are related to the original description of CENTURY. More information about recent developments can be found at www.nrel.colostate.edu/projects/century/.

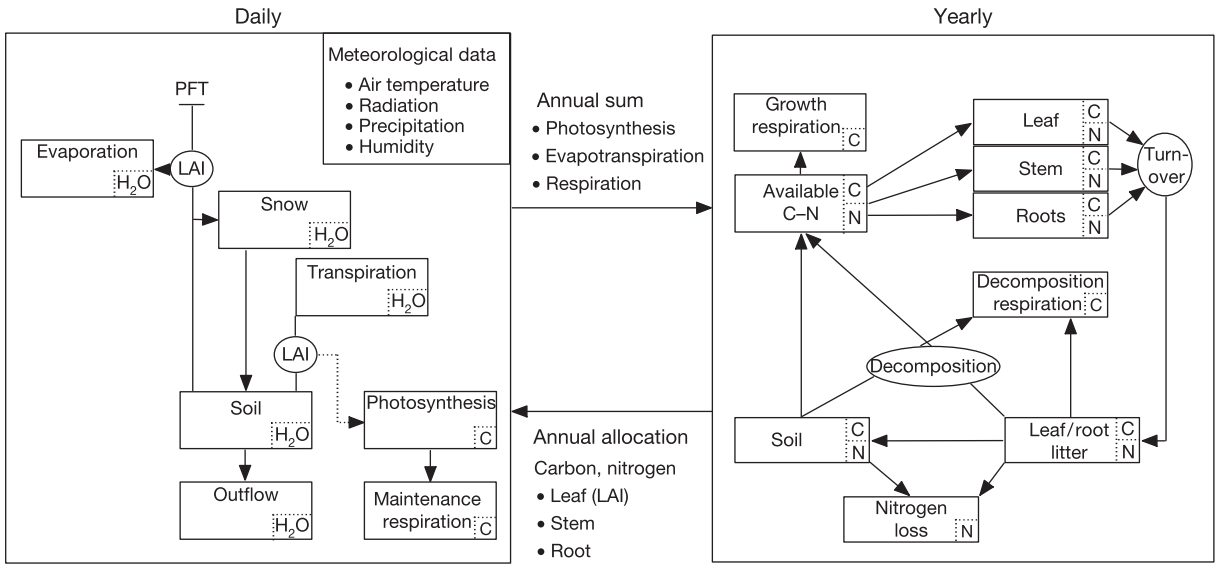


Fig. 1 Schematic diagram that illustrates the compartments and daily and yearly components of FOREST-BGC. C, carbon; N, nitrogen; LAI, leaf area index. Adapted from Running, S.W. and Coughlan, J.C. (1988). A general model of forest ecosystem processes for regional applications. Part I: Hydrologic balance, canopy gas exchange and primary production processes. *Ecological Modeling* **42**, 125–154, with permission from Elsevier.

and passive pools. These three SOM pools differ in the type of carbon compounds they contain and mineralization rate. For instance, the mineralization rate of the active pool may vary from a few months to a few years, while the mineralization rate of the passive pool may be as long as 1000 years, if not greater. The mineralization rates of the SOM pools may be affected by temperature and soil texture and humidity. CENTURY is one of the detailed process-based models that contain numerous pools with complex interactions. For instance, the forest production model allocates the carbon and nitrogen to large wood, foliage, fine branches, and coarse and fine roots. For each of these pools, there are also equivalent pools that represent the amount of dead material. Each type of plant residue is divided into structural and metabolic pools, which are defined on the basis of their lignin and nitrogen content. Within each SOM pool, each time there is mineralization, there is also microbial respiration.

The literature on process-based models has increased sharply in the last few decades. New models were presented to address specific questions, suggest different or advanced modeling approaches or functional relationships, or simply introduce innovative structures to improve the representation of the complexities of the interactions among the processes and effects of environmental factors. Despite the significant evolution, some basic model frameworks remained relatively constant. For instance, for the modeling of photosynthetic rate, the model developed by Farquhar and colleagues in the early 1980s were integrated into many process-based models of forest ecosystems that included the representation of photosynthetic rate. Two basic relationships in Farquhar's model have been widely applied. The regulation of net photosynthetic rate, A_n (μmol CO_2 m^{-2} s^{-1}), is limited by the ribulose-bisphosphate (RUBP) concentration:

$$A_n = J\frac{C_i - \Gamma}{4(C_i + 2\Gamma)}R_d \tag{6}$$

or the activity of ribulose-bisphosphate carboxylase/oxygenase (RuBisCO) at saturating RUBP:

$$A_n = V_c \max \frac{C_i - \Gamma}{C_i + K_c(1 + O/K_o)} - R_d \tag{7}$$

where J (μmol m^{-2} s^{-1}) is the potential electron transport rate, C_i the CO_2 intercellular concentration (μmol mol^{-1}), Γ the CO_2 compensation point (μmol mol^{-1}), R_d the dark respiration rate (μmol CO_2 m^{-2} s^{-1}), V_{cmax} the RuBisCO potential capacity for CO_2 fixation (μmol m^{-2} s^{-1}), K_c the Michaelis–Menten coefficient for carboxylation (μmol mol^{-1}), O the O_2 intercellular concentration (μmol mol^{-1}), and K_o the Michaelis–Menten coefficient for oxygenation (mmol mol^{-1}).

As previously mentioned, the process-based models differ in the amount of details included with respect to input variables and modeled processes. For instance, FOREST-BGC and BIOME-BGC require daily meteorological data (minimum and maximum air temperatures, precipitation, relative humidity, and solar radiation), while CENTURY requires monthly meteorological data (precipitation and mean monthly minimum and maximum temperatures). For process-based models on soil processes, the soil may be considered as a single unit or may be subdivided into horizons. The modeling of PAR interception and distribution within the canopy, including the effect on photosynthetic rate, is another model component that differs among process-based models. For instance, FOREST-BGC considers the canopy as a single three-dimensional layer with homogeneous characteristics ("big-leaf" approach). The stratification of the canopy into at least two sections of sunlit and shaded leaves, the derivation of foliage distribution functions, or the description of the development of individual branches within crowns are "multilayer" approaches used in several models. For instance, the LIGNUM model includes the description of detailed tree architecture, including the development of foliage on individual branches. The multilayer canopy approach implies that the variations in foliage characteristics with canopy position, such as foliage dimensions, area, mass, angle, or nitrogen content, are usually represented and may have a pronounced influence on the degree of precision in predicted PAR interception and photosynthetic rate and on the degree of realism in the scaling of processes. For several tree species, foliage at the top of the canopy is acclimated to full sunlight conditions, while the foliage within the canopy under shade conditions is acclimated to low-sunlight conditions. These differences in acclimation are associated with morphological, physiological, and anatomical characteristics of the foliage. For instance, leaf area or nitrogen concentration may increase from the top to the bottom of the canopy, while leaf biomass and thickness may decrease.

The objectives of the modeling effort must guide the decisions on the complexity of the details to be included, as the number of variables and degree of complexity may require more field measurements for calibration or may influence any scaling exercise. Also, the more variables and parameters in a model the less general it may be. For instance, if a model is developed to examine if nutrient leaching may significantly affect nutrient uptake in soils that contain mineral horizons differing substantially in texture and structure, the subdivision into layers may be entirely justified. However, if the objective of the model is to simulate the productivity of forest ecosystems by including the effect of nutrients in combination with other factors, it may not be necessary to subdivide the soil into different horizons. If a model is intended to simulate how competitive interactions among trees within a forest ecosystem affect carbohydrate partitioning among different sections of individual crowns, stems, and roots, the representation of the variability in foliage characteristics may be important for the modeling of photosynthetic rate. On the other hand, if a model is intended to scale up photosynthetic and respiration rates from the ecosystem to the landscape levels, the use of the "big-leaf" approach may be sufficient.

Gap Models

Gap models, also known as forest succession models, simulate seedling establishment, including their transition to sapling and tree status, and individual-tree growth and mortality. They are very well adapted to simulate the development of mixed uneven-aged forest ecosystems. Several gap models were developed in the last few decades, and the majority of them are descendants of JABOWA, originally developed by Daniel B. Botkin and colleagues in the early 1970s. Among gap models that have received much attention, FORET (developed by Herman H. Shugart in the 1970s and 1980s) and ZELIG and SORTIE (developed by Dean L. Urban and colleagues and Steve W. Pacala and colleagues in the 1990s) are similar in concept to JABOWA. They can simulate the natural course of species replacement for several generations, as well as the succession that is initiated when canopy openings occur in forest ecosystems following the death of a dominant tree. Gap models are a compromise between growth and yield models and process-based models. Biotic and abiotic processes occurring in forest ecosystems are modeled, including competition and the effects of PAR conditions, site fertility, temperature, and water on tree growth and seedling establishment. However, the description of the processes and effects of site factors on ecosystem behavior is much simpler than in process-based models discussed above.

Empirical or semiempirical relationships are used to model the growth of individual trees. For JABOWA-type models, the annual growth rate of individual tree species within a forest ecosystem is modeled using the following basic equation form:

$$\frac{\Delta D_{\text{real}}}{\Delta \text{t}} = \frac{\Delta D_{\text{pot}}}{\Delta \text{t}} \times f(\text{limiting site factors}) \tag{8}$$

where the left-hand term represents the realized annual dbh growth rate, and $\Delta D_{\text{pot}}/\Delta t$ is the maximum or potential dbh growth rate that the tree can achieve under optimal conditions. The relationship for potential growth rate has the following general form:

$$\frac{\Delta D_{\text{pot}}}{\Delta \text{t}} = \frac{GD(1.0 - DH)}{D_{\max} H_{\max}} \times \frac{1}{f(D)} \tag{9}$$

where D is dbh, H stem height, G a growth rate parameter, $D_{\max}$ and $H_{\max}$ the maximum dbh and height that a tree species can reach under optimal conditions, respectively, and $f(D)$ an allometric relationship between dbh and height. The effects of limiting site factors are expressed by functions that describe the influence of abiotic and biotic factors, including PAR, temperature, site fertility, and moisture conditions. These functions constrain the potential growth rate and are based on scalar (0–1) multiplicative relationships. Tolerance or response classes are used to differentiate the response of each species to environmental factors. For instance, ZELIG recognizes five shade-tolerance classes (from tolerant to intolerant), three soil fertility classes (from responsive to stress tolerant), and three soil moisture classes (from drought tolerant to intolerant). The effect of temperature for each species is described by a dimensionless (0,1) response function that scales the growing degree-days computed for a given site relative to the minimum and maximum growing degree-days that exist within the range of distribution of a species. Seedling establishment is modeled stochastically by computing a probability of establishment conditioned by the potential number of seedlings that a site can produce and the prevailing understory PAR conditions, site fertility level, moisture conditions, and growing degree-days. Mortality is also modeled as a stochastic event and may result from natural or stress-related factors.

The advances in computer technology have made it possible to develop integrated modeling tools to simulate forest succession and disturbance at the landscape level. While the JABOWA-type models focus on the simulation of individual-tree development to predict ecosystem dynamics at the forest ecosystem level, forest landscape models simulate the forest dynamics of large regions that include many forest ecosystems. Forest landscape models explicitly consider that ecological processes occur at different scales or units: abiotic and biotic interactions occur among trees within a forest ecosystem as well as among forest ecosystems within a forest region or landscape. Thus, several processes may occur among adjacent or more or less remote forest ecosystems: flow of energy, water and nutrients, species migration or disturbances, such as fire, windthrow, diseases, or insect infestation. One of the objectives in the design of forest landscape models is to integrate the effects of the occurrence of events in individual forest ecosystems on adjacent forest ecosystems. This means that the integration of spatial and temporal complex interactions differing in scale, which may require the use of several variables, must rely on technology, such as geographic information systems for spatialization, in combination with mathematical analysis methods to describe the mechanistic details of the processes. A good example of this relatively recent generation of forest ecosystem models is LANDIS (Fig. 2). This basic diagram illustrates the interactions between the information on species composition and ecological characteristics of the forest ecosystems within a landscape or forest region, the processes that are simulated over time, including species establishment, growth, and death, and the integration of different types of disturbances and management scenarios.

Model Evaluation

Issues or questions related to uncertainty, prediction error, sensitivity, and robustness are becoming more important because models are increasingly used in forest management planning and decision making. Thus, it is important to provide policymakers with quantifiable methods to evaluate the extent to which these issues may affect their judgment or decisions. Models are imperfect representations of reality, which means that their outputs or predictions contain errors or have a degree of uncertainty associated with them. Uncertainty may be caused by model structure, that is, the lack of understanding of biological processes or incorrect mathematical representation, data and parameter estimates, natural variation, and scaling.

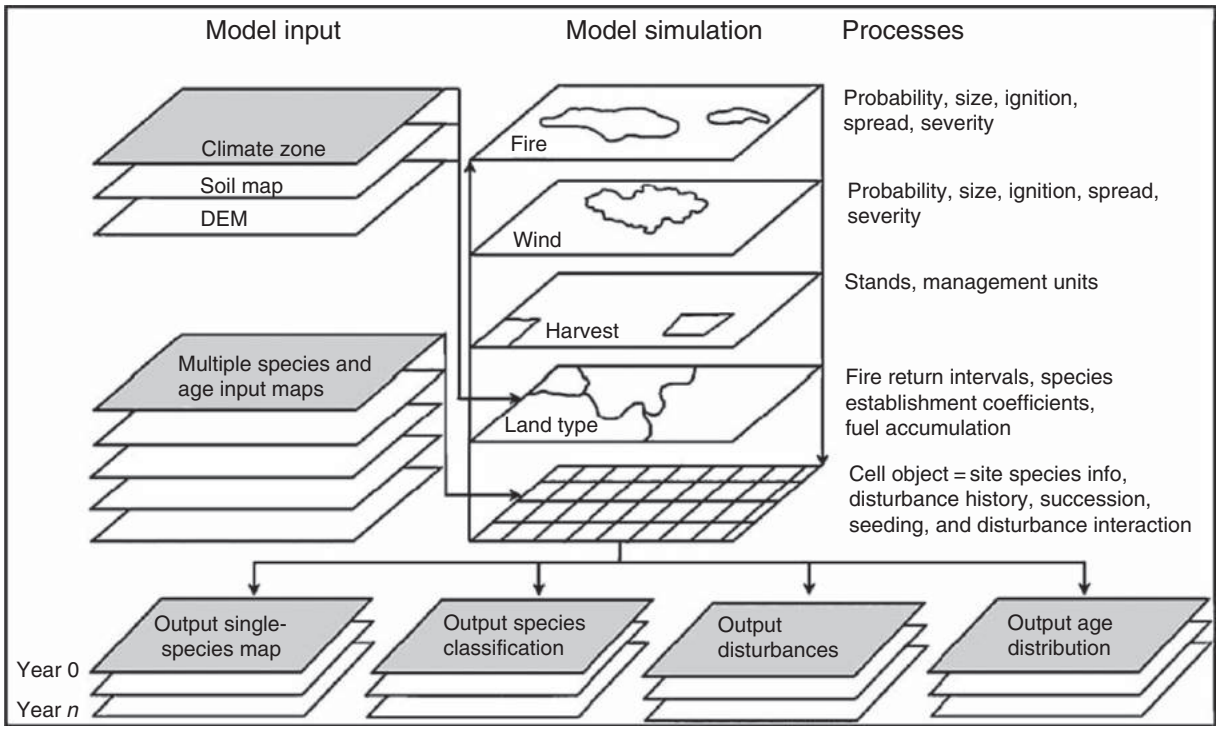


Fig. 2 Basic schematic diagram that illustrates the integrated structure of LANDIS to simulate forest succession and disturbance over landscapes. Reproduced from Mladenoff, D.J. (2004). LANDIS and forest landscape models. *Ecological Modeling* 180, 7–19, with permission from Elsevier.

Uncertainty in the predictions, sensitivity, and robustness have often been evaluated with validation methods. Model validation (the term validation is frequently used interchangeably with verification in the literature, which contributed to maintaining the controversy on what model validation consists of) may include the comparison of model outputs with observations from an independent data set and the examination of the consistency of its logical structure. For instance, the comparison of predictions with observations using historical data is a common validation method. Historical data for forest ecosystems may consist of tree or site variables (e.g., dbh, stem height, foliage, and soil nutrients) measured on the same sample plots at different periods. An essential rule is to conduct the comparison of a model's outputs with data that have not been previously used for its calibration. However, some may argue that this method is simply a statistical validation, and does not allow one to conclude if a model is biologically consistent or realistic. Biological consistency consists in examining the behavior of a model by varying systematically its inputs to represent the variation in widely different initial conditions. This is important for evaluating if the predictions indicate a logically consistent pattern, determining if a model conforms to basic laws of biological growth, or detecting illogical or improper formulation or representation of the underlying processes. For instance, a forest growth model that would predict that the diameter growth of individual trees within an even-aged forest ecosystem increases with increase in stand density (i.e., with the intensity of competition) would be inconsistent, but the statistical fit could be significant. A biological consistency analysis can be conducted simultaneously with a sensitivity analysis, which consists in examining the degree to which the outputs of a model change with variation in input variables or parameters. When the variation in input variables or parameters is extended to their extremes, it allows the modeler to verify the reaction of the model when there is extrapolation.

When a statistical method is used to estimate the parameters of a model, the coefficient of determination (which expresses the proportion of variability in the dependent variable accounted for by the regression on the independent variable(s)) and the mean square error (a measure of dispersion) are usually computed. For more complex models that contain several relationships, other statistics computed from predicted and observed values have been suggested in the forestry literature as quantitative measures to evaluate the performance of growth and yield models in terms of goodness of fit:

$$\text{Mean residual or prediction error} = \sum (y_i - \hat{y}_i)/n \tag{10}$$

$$\text{Root mean square error} = \sqrt{\sum (y_i - \hat{y}_i)^2/n - 1 - p} \tag{11}$$

$$\text{Model efficiency} = \sum (y_i - \hat{y}_i)^2 / \sum (y_i - \overline{y}_i)^2 \tag{12}$$

$$\text{Variance ratio} = \sum (\hat{y}_i - \overline{\hat{y}}_i)^2 / \sum (y_i - \overline{y}_i)^2 \tag{13}$$

where n is the number of samples in the independent data set, y_i are observations, $\hat{y}_i$ predictions, and p the number of independent variables. Several statistical tests exist to validate models, such as the Kolmogorov–Smirnov, χ^2-test, or F-test of a regression.

Summary

Forest models are increasingly being used as research tools to better understand the mechanisms that govern the dynamics of forest ecosystems and as decision-making tools to plan forest management activities and predict the effects of disturbances. Three different types of models were reviewed: growth and yield models, process-based models, and gap models. Each model type is characterized by specific features, including model formulation, complexity of structure, state variables, and required inputs. While growth and yield models aim at predicting tree and stand growth to evaluate the effects of forest management activities, process-based and gap models focus on the simulation of processes and are more flexible for predicting the effects of changing environmental conditions or the natural course of long-term change in species composition. Therefore, when selecting a model type, the intended objectives of a model must be carefully considered.

See also: Ecological Complexity: Thermodynamics in Ecology. Ecosystems: Forest Plantations; Riparian Wetlands. Ecosystems: The Boreal Forest Ecosystem; Ecosystems

References

Battaglia, M., Sands, P., White, D., Mummery, D., 2004. CABALA: A linked carbon, water and nitrogen model of forest growth for silvicultural decision support. Forest Ecology and Management 193, 251–282.
Bella, I.E., 1971. A new competition model for individual trees. Forest Science 17, 364–372.
Dufrêne, E., Davi, H., François, C., *et al*., 2005. Modeling carbon and water cycles in a beech forest. Part 1: Model description and uncertainty analysis on modeled NEE. Ecological Modeling 185, 407–436.
Hatch, C.R., Gerrard, D.J., Tappeiner II, J.C., 1975. Exposed crown surface area: A mathematical index of individual tree growth potential. Canadian Journal of Forest Research 5, 224–228.
Hegyi, F., 1974. A simulation model for managing jack-pine stands. In: Fries, J. (Ed.), International Union of Forestry Research Organizations.Proceedings of Meetings in 1973, Growth Models for Tree and Stand SimulationStockholm Sweden: Royal College of forestry, research note no. 30, pp. 74–90.
Hunt, Jr., E.R., Piper, S.C., Nemani, R., *et al*., 1996. Global net carbon exchange and intra-annual atmospheric CO_2 concentrations predicted by an ecosystem process model and three-dimensional atmospheric transport model. Global Biogeochemical Cycles 10, 431–456.
Komarov, A., Chertov, O., Zudin, S., *et al*., 2003. EFIMOD 2—A model of growth and cycling of elements in boreal forest ecosystems. Ecological Modeling 170, 373–392.
Landsberg, J.J., Waring, R.H., 1997. A generalized model of forest productivity using simplified concepts of radiation-use efficiency, carbon balance and partitioning. Forest Ecology and Management 95, 209–228.
Parton, W.J., Anderson, D.W., Cole, C.V., Stewart, J.W.B., 1983. Simulation of soil organic matter formation and mineralization in semiarid agroecosystems. In: Lowrance, R.R., Todd, R.L., Asmussen, L.E., Leonard, R.A., (Eds.), Special Publication No. 23: Nutrient Cycling in Agricultural Ecosystems. Athens, GA: The University of Georgia, College of Agriculture Experiment Stations, pp. 533–550.
Parton, W.J., Schimel, D.S., Cole, C.V., Ojima, D.S., 1987. Analysis of factors controlling soil organic levels of grasslands in the Great Plains. Soil Science Society of America Journal 51, 1173–1179.
Rouvinen, S., Kuuluvainen, T., 1997. Structure and asymmetry of tree crowns in relation to local competition in a natural mature scots pine forest. Canadian Journal of Forest Research 27, 890–902.
Running, S.W., Coughlan, J.C., 1988. A general model of forest ecosystem processes for regional applications. Part I: Hydrologic balance, canopy gas exchange and primary production processes. Ecological Modeling 42, 125–154.
Running, S.W., Gower, S.T., 1991. FOREST-BGC, a general model of forest ecosystem processes for regional applications. Part II: Dynamic carbon allocation and nitrogen budgets. Tree Physiology 9, 147–160.
Schröder, J., von Gadow, K., 1999. Testing a new competition index for maritime pine in northwestern Spain. Canadian Journal of Forest Research 29, 280–283.
Stadt, K.J., Huston, C., Lieffers, V.J., 2002. A comparison of non-spatial and spatial, empirical and resource-based competition indices for predicting the diameter growth of trees in maturing boreal mixedwood stands. Edmonton, AB: Sustainable Forest Management Network Edmonton Project Report 2002–8.
Steneker, G.A., Jarvis, J.M., 1963. A preliminary study to access competition in a white spruce-trembling aspen stand. Forestry Chronicle 39, 334–336.
Vettenranta, J., 1999. Distance-dependent models for predicting the development of mixed coniferous forests in Finland. Silva Fennica 33, 51–72.

Further Reading

Amaro, A., Reed, D., Soares, P., 2003. Modeling Forest Systems. Cambridge, MA: CABI Publishing.
Barrett, T.M., 2001. Models of vegetative change for landscape planning: A comparison of FETM, LANDSUM, SIMPPLLE, and VDDT. General Technical Report RMRS-GTR-76-WWW. Ogden, UT: USDA Forest Service, Rocky Mountain Research Station.
Botkin, D.B., 1993. Forest dynamics: An ecological model. Oxford: Oxford University Press.
Bugmann, H., 2001. A review of forest gap models. Climatic Change 51, 259–305.
Dixon GE (comp.) (2002) Essential FVS: A User's Guide to the Forest Vegetation Simulator. Internal report. Fort Collins, CO: USDA Forest Service, Forest Management Service Center.
Dixon, R.K., Meldahl, R.S., Ruark, G.A., Warren, W.G., 1990. Process modeling of Forest growth responses to environmental stress. Portland, OR: Timber Press, Inc.
Ehleringer, J.R., Field, C.B., 1993. Scaling physiological processes. Leaf to globe. New York: Academic Press.
Farquhar, G.D., Sharkey, T.D., 1982. Stomatal conductance and photosynthesis. Annual Review of Plant Physiology 33, 317–345.

Farquhar, G.D., von Caemmerer, S., 1982. Modeling of photosynthetic response to environmental conditions. In: Lange, O.L., Nobel, P.S., Osmond, C.B., Ziegler, H., (Eds.), Encyclopedia of plant physiology 12B, Vol. II: Physiological. Plant ecology. Berlin: Springer, pp. 549–587.
Farquhar, G.D., von Caemmerer, S., Berry, J.A., 1980. A biochemical model of photosynthetic CO_2 assimilation in leaves of C3 species. Planta 149, 78–90.
Glenn-Lewin, D.C., Peet, R.K., Veblen, T.T., 1992. Plant succession: Theory and practice. New York: Chapman and Hall.
Homann, P.S., McKane, R.B., Sollins, P., 2000. Belowground processes in forest-ecosystem biogeochemical simulation models. Forest Ecology and Management 138, 3–18.
Hunt Jr., E.R., Lavigne, M.B., Franklin, S.E., 1999. Factors controlling the decline of net primary production with stand age for balsam fir in Newfoundland assessed using an ecosystem simulation model. Ecological Modeling 122, 151–164.
Johnsen, K., Samuelson, L., Teskey, R., McNulty, S., Fox, T., 2001. Process models as tools in forestry research and management. Forest Science 47, 2–8.
Landsberg, J., 2003. Modeling forest ecosystems: State of the art, challenges, and future directions. Canadian Journal of Forest Research 33, 387–395.
Medlyn, B.E., Robinson, A.P., Clement, R., McMurtrie, R.E., 2005. On the validation of models of forest CO_2 exchange using eddy covariance data: Some perils and pitfalls. Tree Physiology 25, 839–857.
Miner, C.L., Walters, N.C., Belli, M.L., 1988. A guide to the TWIGS program for the North Central United States, General Technical Report NC-125. St. Paul, MN: USDA Forest Service, North Central Forest Experiment Station.
Mladenoff, D.J., 2004. LANDIS and forest landscape models. Ecological Modeling 180, 7–19.
Newton, P.F., 1997. Stand density management diagrams: Review of their development and utility in stand-level management planning. Forest Ecology and Management 98, 251–265.
Pacala, S.W., Canham, C.D., Silander, J.A.J., 1993. Forest models defined by field measurements. Part I: The design of a northeastern forest simulator. Canadian Journal of Forest Research 23, 1980–1988.
Perttunen, J., Sievänen, R., Nikinmaa, E., 1998. LIGNUM: A model combining the structure and the functioning of trees. Ecological Modeling 108, 189–198.
Pienaar, L.V., Turnbull, K.J., 1973. The Chapman–Richards generalization of Von Bertalanffy's growth model for basal area growth and yield in even-aged stands. Forest Science 19, 2–22.
Radtke, P.J., Burk, T.E., Bolstad, P., 2001. Estimates of the distributions of forest ecosystem model inputs for deciduous forests of eastern North America. Tree Physiology 21, 505–512.
Reineke, L.H., 1933. Perfecting a stand-density index for even-aged forests. Journal of Agriculture Research 46, 627–638.
Reynolds, J.H., Ford, E.D., 2005. Improving competition representation in theoretical models of self-thinning: A critical review. Journal of Ecology 93, 362–372.
Robinson, A.P., Ek, A.R., 2000. The consequences of hierarchy for modeling in forest ecosystems. Canadian Journal of Forest Research 30, 1837–1846.
Schwalm, C.R., Ek, A.R., 2001. Climate change and site: Relevant mechanisms and modeling techniques. Forest Ecology and Management 150, 241–257.
Shugart, H.H., 1984. A theory of forest dynamics. New York: Springer-Verlag.
Sievänen, R., Nikinmaa, E., Nygren, P., Ozier-Lafontaine, H., Perttunen, J., Hakula, H., 2000. Components of functional–structural tree models. Annals of Forest Science 57, 399–412.
Stage, A.R., 1973. Prognosis model for stand development, Research Paper INT-137. Ogden, UT: USDA Forest Service, Intermountain Forest and Range Experiment Station.
Urban, D.L., Bonan, G.B., Smith, T.M., Shugart, H.H., 1991. Spatial applications of gap models. Forest Ecology and Management 42, 95–110.
Vanclay, J.K., 1994. Modeling forest growth and yield: Applications to mixed tropical forests. Wallingford: CABI.
Yoda, K., Kira, T., Ogawa, H., Ozumi, K., 1963. Self-thinning in overcrowded pure stands under cultivated and natural conditions. Journal of Biology of Osaka City University 14, 107–129.

Relevant Website

nrel.colostate.edu, n.d.—nrel.colostate.edu—CENTURY—National Resource Ecology Laboratory, Colorado State University.

Fuzzy Models☆

R Wieland, Leibniz Centre for Agricultural Landscape Research (ZALF), Muencheberg, Germany

Introduction

In environmental modeling, biologists or agricultural scientists are often able to describe a process quite well, but find it difficult to construct a mathematical model. One such problem is the modeling of alteration in wildlife-habitat quality caused by climate change, pollution, new management methods in agriculture, etc. A further example is the calculation of agricultural yield for different management strategies under changing climate conditions. Such problems often involve uncertainties. Fuzzy-set modeling can help one to cope with this uncertainty.

Fuzzy-set theory and fuzzy-set control theory provide methods to describe:

- the input ranges for the model
- the output ranges for the model
- a set of rules
- a method for defuzzification

To use fuzzy-set models for environmental modeling a method must be provided to enable use of these models in a spatial context. This can be done as part of a geographic information system (GIS) or with a special software toolbox like the matlab fuzzy toolbox (Sivanandam et al., 2006) or the open source software SAMT.

Fuzzy Sets

Fuzzy sets provide a basis for fuzzy modeling. A fuzzy set is an extension of a classical set. In classical set theory an element either belongs or does not belong to a set. In a fuzzy set an element belongs gradually to a set. The fuzzy set theory traditionally uses the

☆*Change History:* April 2013. R Wieland updated Subsection "Adaptation of the rules", and "Dynamic Simulations" sections, and Figures 7 and 8 are new and Table 2 added.

Table 1 Nutrition quality of crops for a crane fuzzy set

	Good	*Medium*	*Bad*
Maize	0.2	0.8	0
Rye	0.4	0.6	0
Wheat	0.6	0.4	0
Rape	0	0.1	0.9
Barley	0.6	0.4	0
...	...	...	...

Greek letter $\mu \rightarrow [0,1]$ as the membership-function. The value between zero and one describes the degree of membership of the element to the set $\tilde{A}$. A fuzzy set is defined by:

$$\tilde{A} = (x, \mu_A(x)) \mid x \in X \tag{1}$$

Every element $x \in X$ gradually belongs to the set $\tilde{A}$. The value $\mu_A(x) = 0$ means that $x \notin \tilde{A}$ and the value $\mu_A(x) = 1$ means that $x \in \tilde{A}$. All values strictly between zero and one characterize the fuzzy members. In other words, if the membership values are restricted to zero and one, the fuzzy set is a classical set.

Another consequence of this definition is that an element x can belong to different fuzzy sets. For example a biologist may express the quality of nutrition of crops for the crane (grus grus) as in **Table 1**.

This estimation could be part of an model of the habitat quality for the crane. Maize with a membership value of 0.8 belongs to the class "medium nutrition quality" and with 0.2 to the class "good nutrition quality" for example. This description can be used to express uncertainty or used in discussing it with other experts.

Combination of Fuzzy Sets

Fuzzy sets may have to be combined with other fuzzy sets. Common operations are "and", "or" and "not". These operations can be defined as:

Definition 1 (Fuzzy AND Min)

$$x \in \tilde{A} \wedge x \in \tilde{B} \Rightarrow \mu_{A \wedge B}(x) = min\{\mu_A(x), \mu_B(x)\} \tag{2}$$

Definition 2 (Fuzzy OR Max)

$$x \in \tilde{A} \vee x \in \tilde{B} \Rightarrow \mu_{A \vee B}(x) = max\{\mu_A(x), \mu_B(x)\} \tag{3}$$

Definition 3 (Fuzzy NOT)

$$x \notin \tilde{A} \Rightarrow 1 - \mu_A(x) \tag{4}$$

These definitions will be used as parts of fuzzy models. Alternative definitions for fuzzy "and" and "or" are

Definition 3 (Fuzzy AND Prod)

$$x \in \tilde{A} \wedge x \in \tilde{B} \Rightarrow \mu_{A \wedge B}(x) = \mu_A(x) * \mu_B(x) \tag{5}$$

Definition 4 (Fuzzy OR Add)

$$x \in \tilde{A} \vee x \in \tilde{B} \Rightarrow \mu_{A \vee B}(x) = min\{\mu_A(x) + \mu_B(x), 1\} \tag{6}$$

These definitions are based on fuzzy t-norm and t-conorm (Hjek, 1998). In fuzzy models the definitions 5 and 6 often show a better behavior than definitions 1 and 2. This will be explained in greater detail below. (Minimum/maximum and/or was used because simple fixed-point micro-controllers are faster that way; floating-point-PCs are not (multiplication is faster than floating-point comparison).)

Fuzzy Sets for Continuous Inputs

In the definition of fuzzy sets the inputs were given by enumeration. Spatial modeling of ecological precesses often involves inputs which continuous variables such as distance or density functions. This continuous values must be "fuzzified", i.e., the modeler

must specify membership function values for the fuzzy sets. Membership functions with triangular or trapezoidal shape are frequently used. A triangular membership function is defined as:

$$\mu_A(x) = \begin{cases} 0 & : & x \leq x_1 \\ (x - x_1)/(x_2 - x_1) & : & x > x_1 \wedge x \leq x_2 \\ (x_3 - x)/(x_3 - x_2) & : & x > x_2 \wedge x < x_3 \\ 0 & : & x \geq x_3 \end{cases}$$

The lower left edge point is x_1, x_2 the point where the triangle is 1, and x_3 the right hand point of the triangle. This membership function describes fuzzy values, e.g., "about 10". (meaning numbers around 10.) In this example $x_1 = 8$ and $x_3 = 12$. The 10 belongs to this set with $\mu_A(10) = 1$, and 9 belongs to this set with a degree $\mu_A(9) = (9-8)/(10-8) = 1/2$.

Another commonly used membership function type has a trapezoidal shape, and is used to describe ranges. A range is a generalization of classical sets. In the middle of a range, between $[x_2, x_3]$ all values belong to the set with a membership of $\mu_A(x) = 1$. But there are some values between (x_1, x_2) and (x_3, x_4) with a strict fuzziness. This type of membership function is often used in habitat modeling. If an effect of disturbance on a habitat has to be modeled, then this range is often helpful. In the center of the range the animal can tolerate the disturbance, but if the disturbance exceeds a determined level, the reaction may be strong. In a formal way this range can be defined as:

$$\mu_A(x) = \begin{cases} 0 & : & x \leq x_1 \\ (x - x_1)/(x_2 - x_1) & : & x > x_1 \wedge x < x_2 \\ 1 & : & x \geq x_2 \wedge x \leq x_3 \\ (x_4 - x)/(x_4 - x_3) & : & x > x_3 \wedge x < x_4 \\ 0 & : & x \geq x_4 \end{cases}$$

The triangular and the trapezoidal membership functions are simple and useful in modeling environmental processes. More sophisticated membership function types like bell shaped functions are possible but often overrated. A trapezoidal membership can approximate a bell-shaped function quite well, but the bell shape-function is itself an approximation. The ultimate objective is to use fuzzy models to transform expert knowledge into formal knowledge that can be used to perform simulations.

Fuzzy Models

Fuzzy models specify rules in terms of fuzzy sets. A fuzzy rule consists of an antecedent (IF-part) and the consequent (THEN-part). The antecedent combines different inputs using the "and" combination. To express "or" combinations, a set of rules with different inputs but the same output is often used. For example:

$$IF((nutrition \in good) \vee (nutriton \in medium)) \Rightarrow quality = good$$

can be expressed with the following two rules:

$$IF(nutrition \in good) \Rightarrow quality = good$$

$$IF(nutrition \in medium) \Rightarrow quality = good$$

This splitting technique leads to easily understood but larger rule sets. The simpler rules are sometimes advantageous for rule inspection in maintenance of fuzzy models.

Output of Fuzzy Rules

The consequent part can be arranged as a bundle of fuzzy sets or more simply as crisp values. (A crisp value means that every output is linked with an real value: $medium \rightarrow 0.8$; $good \rightarrow 0.2$ for example.) It seems rather curious that the consequent part often is implemented using crisp values, but this has a lot of advantages. Some of which are described later, but a major advantage is based on the sematic. The input carries information about the uncertainty. This uncertainty is modeled using fuzzy sets. The output is a desired value, which has to determined by the modeler. The simplest way to do that, is to provide a crisp value for each output. If the output is modeled using fuzzy sets, this implies that the modeler is uncertain about the consequence. But this is often not true. The modeler knows the output but is not able to determine the fuzziness. It is difficult to determine the x_1 and x_3 of a triangle for an output. With other words the modeler is more or less confident about the value or the output, but he can not determine the fuzziness of it.

The modeler has to define the membership functions $\mu_{ij}(x_i)$ (i counts the inputs, j counts the rules) for the inputs, the membership functions for the outputs (or in our case the output values as crisp values) and the rules for combining the input with the output. The membership functions are indexed for every variable and for the linguistic values. For example the input x_1

may have three membership functions (low, medium, high) so the index j will run from 1 to 3. A typical rule set with three inputs and n rules looks like this:

$$\begin{array}{lcl} \text{IF } \mu_{11}(x_1) \wedge \mu_{21}(x_1) \wedge \mu_{31}(x_3) & \Rightarrow & o_1 \\ \text{IF } \mu_{12}(x_1) \wedge \mu_{22}(x_2) \wedge \mu_{32}(x_3) & \Rightarrow & o_2 \\ \cdots & \cdots & \\ \text{IF } \mu_{1n}(x_1) \wedge \mu_{2n}(x_2) \wedge \mu_{3n}(x_3) & \Rightarrow & o_n \end{array}$$

Remark: every rule uses one selected membership function for every input. The outputs o_k must not be different. This means that rules with different membership functions but the same output exist.

Fuzzy Algorithm

The fuzzy algorithm uses the membership functions and the rules in three steps:

- fuzzification: for every input $\vec{x} = (x_1, x_2, \ldots)$ all membership functions $\mu_{ij}(x_i)$ will be calculated
- inference:
  - assigns a value a_k^m using the minimum or product operator to each rule (k points to the outputs; m points to different rules with the same output k):

$$a_k^m = min\left\{\mu_{1j_m}(x_1), \mu_{2j_m}(x_2), \mu_{3l_m}(x_3)\right\} \quad [7]$$

  - selects the the best rule from the m rules with same output k using maximum operator

$$a_k = max\left\{a_k^1, a_k^2, \cdots\right\} \quad [8]$$

- defuzzification

Defuzzification

The defuzzification algorithm assigns the fuzzy output of the fuzzy inference to one floating point value. This value can be used for further calculation, for example in a spatial context. The realization of the defuzzification algorithm depends on the character of the output values. For crisp outputs the algorithm is simple and fast:

$$o = \frac{\sum_k a_k \times o_k}{\sum_k a_k} \quad [9]$$

In eqn [9] the activity of a rule a_k and its output value o_k are used to calculate the output o of the fuzzy system. The realization is fast, which is a precondition for the usage in a spatial simulation. In a spatial simulation a huge number (up to some millions) of grid cells must be often calculate.

A further advantage of crisp output values is not so obvious. This simple algorithm can produce linear functions as well as nonlinear functions. Using a fuzzy set as output leads to a nonlinear behavior. Most ecological processes are nonlinear, but sometimes a piecewise linear part is essential. The modeler must be able to decide the functional behavior of the model. A more demanding modeling approach using outputs as fuzzy sets can be used at the expense of a bit of nonlinearity in the model.

When using fuzzy sets as outputs there are some possibilities for defuzzification. One of the most commonly used is the so called "center of gravity" method. To understand this method a closer look at the fuzzy inference is necessary. The inference procedure assigns a value a_k^m to each rule. After the selection of the best value for a specified output a_k (eqn [8]), this value is multiplied to the output. In case of fuzzy set output the multiplication "cuts" the fuzzy set. A result of the inference procedure may look like **Figure 1**.

To determine one floating point value from this area the following equation is used:

$$o = \frac{\int_{x_a}^{x_e} \mu(x) \times x dx}{\int_{x_a}^{x_e} \mu(x) dx} \quad [10]$$

The eqn [10] is a generalization of eqn [9]. The center-of- gravity algorithm is numerically more complex than the simple eqn [9]. Additionally the integral in eqn [10] introduces the aforementioned nonlinearity in the system. (This can be checked using a simple fuzzy model with one fuzzy input and one fuzzy output.)

Influence of Fuzzification Method to System Behavior

Many different fuzzification methods are quoted in the literature (Hjek, 1998). Two frequently used methods were introduced above. These methods influence system behavior. That means not only an effect on output value but it introduces some

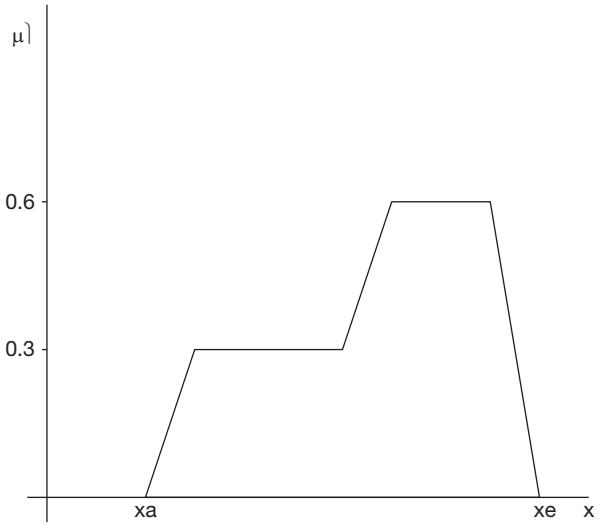


Figure 1 Center of gravity defuzzyfication.

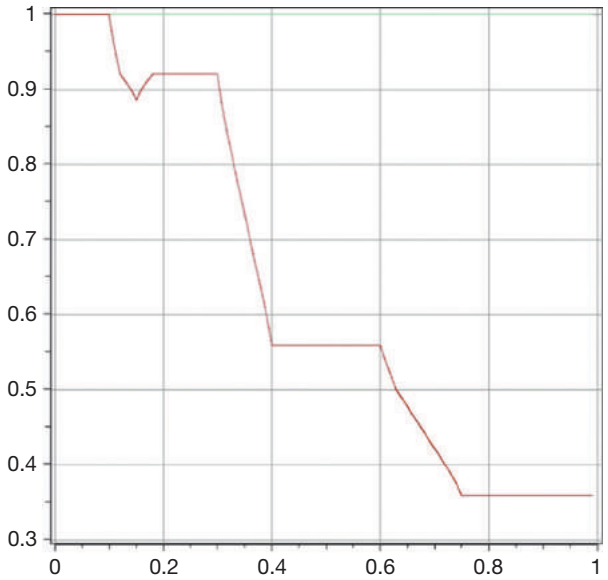


Figure 2 Fuzzy AND Min, x-axis: coverage degree, mean distance to structure = 130 m.

nonlinearity into the system. The FUZZY_AND is often realized as minimum method. This is simple and easy to understand. But sometimes this method introduces some strange behavior in the model. The minimum function cuts the values very strongly. This is sometimes undesirable. The use of the FUZZY_AND Prod is often a better choice. This produces a smoother functional behavior and should be preferred in practical simulations.

This was checked with the fuzzy model for the nutrition quality for the lesser spotted eagle, aquila pomarina, (see below). At first the chart of the model using FUZZY_AND Min was plotted in Figure 2.

Easily visible is the vertex in the upper part of the chart. This can be changed with the use of the FUZZY_AND Prod. This is shown in Figure 3. The last chart looks smoother than the first one.

If the fuzzy model is checked, it can be used as part of a habitat model in a simulation.

Training of Fuzzy Models

A fuzzy system is just a way to formulate heuristically-basis functions for regression and controller functions with a rule table instead of by optimization. The fuzzy modeling approach is a transformation of expert knowledge into a mathematical model. The expert can control this process by defining the membership functions, the outputs and the rules. For modeling tasks like our

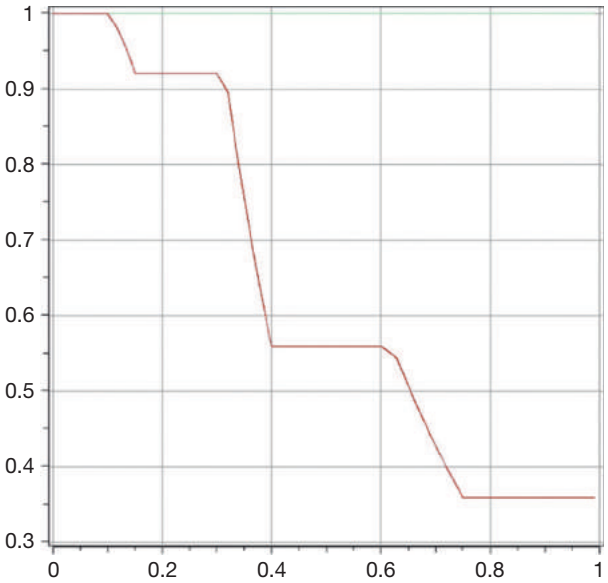


Figure 3 Fuzzy AND Prod, x-axis: coverage degree, mean distance to structure = 130 m.

example is this the only possible way. Usually the expert has lots of experience but little observation data. (Otherwise the modeler should think about a artificial neural network.) When the bird is not as rare as the lesser spotted eagle used in the example below, more data are frequently available. In other investigations the data collection process is part of a project. So the modeler can construct a fuzzy model and wants to enhance it with the collected data set. Often used as an optimization strategy is the minimization of the mean square error (mse) between the modeling output and the observed data:

$$mse = \frac{1}{n}\sum_{i=0}^{n}(y_i - fuzzy(x_{1i},x_{2i},x_{3i}))^2 \qquad [11]$$

y_i is the measured value and the $fuzzy(x_{1i},x_{2i},x_{3i})$ the output of the fuzzy model. There are three possible strategies to train such a fuzzy model:

- to change the membership function for the inputs
- to change the outputs
- to change the rules

Adaptation of Membership Functions

The membership functions are defined by the parameters x_1, x_2, x_3 for triangular shaped membership functions and x_1, x_2, x_3, x_4 for trapezoidal membership functions. These parameters are not independent. Additionally the number of parameters can be large when many membership functions are used. In many simulations it could be shown that this strategy is possible but not very effective and often slow.

Another reason to avoid this is that the expert can see these functions quite clearly. This is caused by the experience with the modeled process. It is not the best idea to start a training procedure by changing this sensitive part of a fuzzy system.

Adaptation of Outputs

The outputs are crisp values. Adaptation of the outputs is simple. The only restriction is that the order of the outputs must not be changed. As training procedure a delta rule can be used (Freeman, 1994). (A delta rule is the standard training rule for artificial neural networks.) The algorithm consists of the following steps:

- determine the active outputs for this input pattern
- calculate the error for every active output
- calculate a delta for these outputs
- change the outputs using the delta rule

All active outputs can be determined using eqn [9]. The error for an output can be calculated as following:

$$error = y_i - fuzzy(x_{1i}, x_{2i}, x_{3i}) \quad [12]$$

With this error the delta can be calculated:

$$\delta_k = a_k * o_k * error * gain \quad [13]$$

The gain is a (small) step rate for the training procedure. The new output follows:

$$o_k = o_k + \delta_k \quad [14]$$

The error determines the direction and strength of the training step. The membership a_k ensures that outputs with small belief are only slightly change.

This training procedure was checked for a yield model. Inputs included soil quality, amount of fertilizer, and water availability. The output was the yield for winter wheat. The fuzzy model consists of 3 inputs with 12 fuzzy sets (5 for soil quality, 4 for the fertilizer, 3 for water). The number of rules equaled 60. The number of training data was 1998. The mean square error of 206.236 before training was reduced to 85.6025 after training. This simple training procedure is effective and very efficient. An alternative training algorithm based on an evolutionary algorithm (Haupt and Haupt, 2004) was implemented but found to be less efficient than the simple delta rule. In contrast to the delta rule the evolutionary algorithm can also be applied to train a rule set.

Adaptation of the Rules

The adaptation of the rules is simple to complete (only a pointer must be changed) but it is a global change in a fuzzy system and must be done with care. (The modeler often wants to control the rules and would not allow an automatically adaptation.) A combination of the outputs is possible and could be used to keep the consistency of the fuzzy system when an output crosses another. To further read about fuzzy adaptation (Nauck et al., 2003), a complete semiautomated training method based on the fuzzy algorithm above is introduced in (Wieland et al., 2011). This method consists of a statistical analysis to support the creation of fuzzy membership functions and an automated rule generation. The big advantage over neural networks (Freeman, 1994) or vector machines (Schlkopf and Smola, 2002) are that the rules can be understood and modified by the modeler. This technique combines learning from data with expert knowledge to a fuzzy model.

Using Fuzzy Models in Spatial Simulation

In a geographic information system the data are stored in form of raster data (uniformly spaced grid cells) or in form of polygons. Grid data cells are more convenient for modeling approaches. An equally spaced grid cell is better suited to make simulation of fluxes with this data model. These simulations are for example important for wind-or water erosion modeling. A fuzzy model needs continuous data as inputs. (An approximation with a discrete set with many members is also O.K.) The spatial information is split into different information layers (maps) each containing a grid. A fuzz modeling approach usually needs two or three maps as inputs to calculate an output grid. The grids contain stored information like soil quality, elevation data, land use data etc. Other data of this type include distances, such as between grid cells and points (nesting places for birds, location of wind energy plant etc.) or between grid cells and linear features (roads, rivers, etc.). Another important example involving continuous data demonstrates the use of the so called moving-window technique, which calculates

$$q_i = f(x, y) = \iint_{A(x, y)} g(x, y) dx dy \quad [15]$$

Here $A(x, y)$ is an area around the point (x, y), and $g(x, y)$ is a function depending on spatial modeling problem. For simple problems, $g(x, y)$ can be the mean or the median for all points in a region $A(x, y)$. This technique produces a spatial abstraction of data at all grid cells. (Not only the value at the point, but also values in the neighborhood of the point are important.)

Before you can start using fuzzy models in spatial context the spatial database must be provided. There are some steps that must be performed:

- select data sources for the project from the GIS data base
- transform the polygons to a grid using an appropriate resolution. (The resolution depends on the modeling task. Every process has its own scale.)
- generate inputs for the fuzzy model using distances and the moving window technique

This task has to be done in a GIS. To apply fuzzy models to this grid data base the modeler should prefer an integrated simulations system. Such a toolbox additionally provides a set of analysis tools. With these tools it is possible to analyze the spatial database, to explain the fuzzy model on concrete points, to control the rule base of the fuzzy model etc. This tools should be also used do perform a validity test, that means the model behavior should be checked against the expected correlations. To further read about a spatial simulation, see (Wieland et al., 2006) and (Mirschel et al., 2006)

Example: Habitat Model of Lesser Spotted Eagle

The lesser spotted eagle is one of the key species in north East Germany. To improve understanding of the fuzzy modeling technique a simplified fuzzy model using the lesser spotted eagle will be explained. (The real model discriminates more details in the habitat structure, but this is not important for understanding.)

Habitat structure

The habitat structure of any animal depends on nutrition, the possibility of reproduction and protection against predators and other dangers. For the lesser spotted eagle this means that:

- When hunting it can either fly or sit and wait. If it hunts in flight, plant cover must not be too dense, so that it can see the prey on the ground. To enable it to sit and wait some "structure elements" to perch on must be available.
- The lesser spotted eagle breeds in a forest. It must be able to reach the hunting area from the nesting place.
- There are no predators in the region, but wind energy plants or roads and railways might be dangerous too.

Fuzzy model nutrition

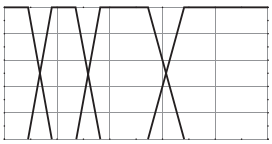
The investigated area is located northeast of Berlin, and measures approximately 18×30 km. Most of the area is used for agriculture but there are also grass lands, forests, lakes and settlements.

Before the fuzzy model "nutrition.fis" can be used the coverage of the soil with plants must be calculated. The coverage depends on the type of crops and therefore on the time of year. The crops covering the soil are shown in Figure 4. As an example the year 1995 and the year day 160 was selected.

Another input for the nutrition part of the habitat model for the lesser spotted eagle is the structure density map. (In this map the mean distance to structure is shown.) A high density of structure can partly compensate a high cover of crop degree because the bird is able to use a structure element (hedge, tree etc.) as hunting point. This input is represented in Figure 5. Remark: this map was produced using moving window technology to generalize from the real structure to a more landscape characteristic.

The output is calculated using a fuzzy model. The two fuzzy inputs degree of coverage and structure density are shown:

Coverage degree [0.000.0.1.000]



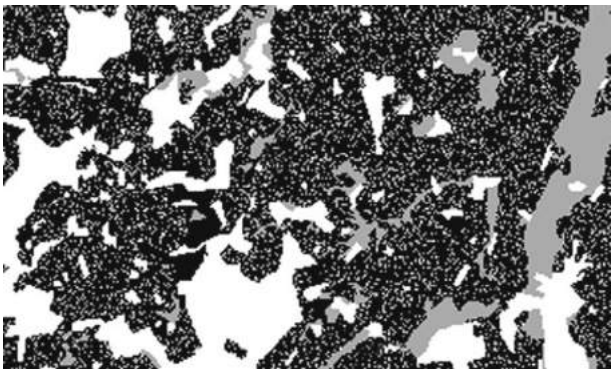


Figure 4 Degree of crop cover, white areas = non food area, gray color: the brighter the color the smaller the coverage degree (the white areas are lakes, forest and settlements).

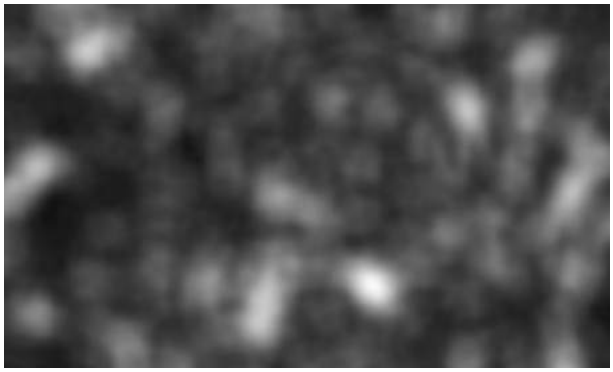
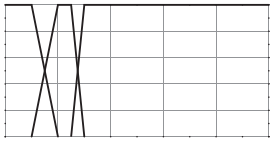


Figure 5 Mean distance to structure elements, the brighter the color the smaller the distance.

Very_small	−0.100	0.000	0.100	0.200
Small	0.100	0.200	0.300	0.400
Medium	0.300	0.400	0.600	0.750
High	0.600	0.750	1.000	1.100

Mean distance to structure [0.0.1000]



High	−0.1	0	100	200
Medium	100	200	250	300
Small	250	300	1000	1001

As outputs crisp values where used:

Very_small	*Small*	*Medium_small*	*Medium*	*Medium_high*	*High*	*Very_high*
0	0.2	0.4	0.5	0.6	0.8	1

The rule base is very simple. It can be presented as a table:

Coverage degree	*Structure*	*Output*	*Numoutput*	*Weight*
Very_small	High	Very_high	1	1

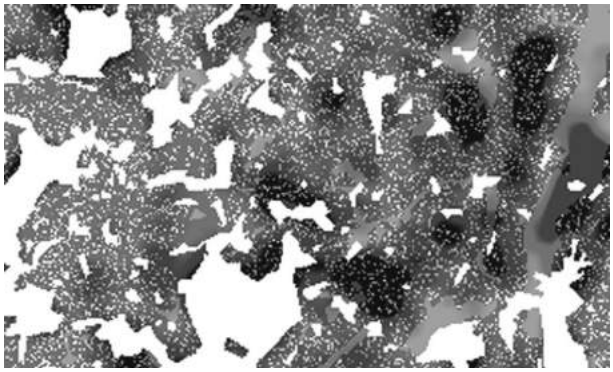


Figure 6 Nutrition quality for the lesser spotted eagle.

Very_small	Medium	Very_high	1	1
Very_small	Small	High	0.8	1
Small	High	Very_high	1	1
Small	Medium	Medium_high	0.6	1
Small	Small	Medium	0.5	1
Medium	High	Medium_small	0.6	1
Medium	Medium	Medium_small	0.4	1
Medium	Small	Small	0.2	1
High	High	Medium_small	0.4	1
High	Medium	Small	0.2	1
High	Small	Very_small	0	1

Sometimes there are contradictory rules. The belief in such a rule can be changed using the weight factor. (This can arise, when more than one expert constructs a rule set.) With this fuzzy model a nutrition quality for lesser spotted eagle can be calculated. The result is shown in Figure 6.

Fuzzy model habitat

The output of the part of the model on nutrition is one input for the calculation of habitat quality. The next input is the distance to the breeding places. These places are located in a forest. The mean distance to the forest can be used to evaluate potential breeding habitat. The last input includes the endangerment caused by wind energy plants. The lesser spotted eagle will avoid areas around wind energy plants. In this sense wind energy plants can shrink the hunting area of the bird. These three inputs are put in a new fuzzy model. The result is the habitat quality of the area. With such a model it is easy to estimate the influence of planned wind energy plants or of some alternative land use modes. The model can be used to optimize the location for new wind energy plants. The big advantage of this modeling approach is that all influences for the habitat quality will be considered. It is not logical to protect the lesser spotted eagle in a region where it can not live.

Simulation of Habitat Quality of the Lesser Spotted Eagle

The habitat model of the lesser potted eagle depends on the degree of vegetation coverage as a dynamic changing value. To determine the habitat quality the simulation has to be repeated over a sequence of points in time (during one breeding session). The resulting habitat quality can be calculated as the minimum for the period which was simulated. (The nutrition condition is especially important in the breeding season from April to June in our region.)

$$q = MIN\{q(t_1), q(t_2), \ldots, q(t_n)\} \quad [16]$$

Dynamic Simulations

Many ecological variables are functions of time. This starts with a simple growth dependency of the coverage degree, as discussed in the example of the lesser spotted eagle. More complicated dependencies include more than one variable, which will be discussed in the Lotka-Volterra model below. In realistic ecological models, the number of time-dependent variables can easily exceed one hundred, as shown in (Abrahamsen and Hansen, 2000). The fuzzy approach can be used to simplify traditional models, making

them more stable and enhancing their accuracy. The most important advantage of dynamic fuzzy models is that ecologists are able to understand and modify all parts of the model. The clear structure of a fuzzy model, consisting of the definition of the membership functions and the rule set, is the key factor for ensuring it is understood. On the other hand, ecologists require a software development system to combine differential equations with fuzzy models. A simple modeling system based on the programming language Python (including "scipy" for solving of differential equations) is available as open source software. This simulator allows a fuzzy model to be analyzed as part of a differential equation. A program combining fuzzy models can then be built as part of a differential equation system with a graphical output to visualize the chart of the solved differential equation. This executable program can be used as the first step for building dynamic fuzzy models. The software is developing very rapidly; efforts are now being made to integrate spatial and dynamic modeling. To find out more about dynamic simulation, see (Deaton and Winebrake, 2000), (Korn, 2013) and (Korn, 2004). An introduction to dynamic fuzzy modeling and how it is applied to landscape modeling is given below.

FUZZY Models as Part of Dynamic Models

In traditional ecological models, differential equations contain algebraic expressions. For example, the well-known Lotka-Volterra model should demonstrate that:

$$\frac{dx}{dt} = \alpha * x - \beta * x * y \quad [17]$$

$$\frac{dy}{dt} = \delta * x * y - \gamma * y \quad [18]$$

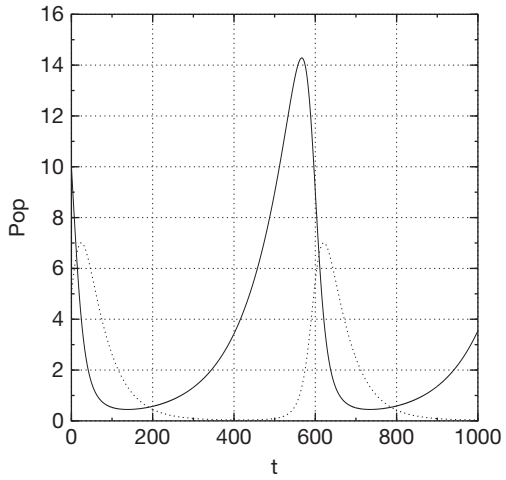


Figure 7 Lotka-Volterra model: prey = black, predator = gray.

Table 2 Subset of fuzzy rules of $f_{predator}(x,y)$

Number	*x (prey)*	*y (predator)*	*Output*
16	High	Zero	sp
17	High	Small	mp
18	High	Medium	hp
19	High	High	sn
20	High	Very_high	mn
21	Very_high	Zero	sp
22	Very_high	Small	mhp
23	Very_high	Medium	hp
24	Very_high	High	mp
25	Very_high	Very_high	sn

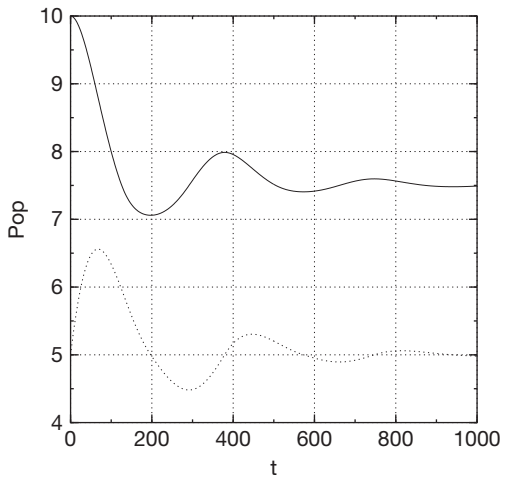


Figure 8 Fuzzy-Lotka-Volterra model: prey = black, predator = gray.

This model describes the interaction between a prey (x) and a predator (y). The prey, which uses a constant nutrition source for growing ($\alpha * x$), is reduced by the predator ($-\gamma * x * y$). The predator depends on the prey ($\delta * x * y$), and is reduced by a constant rate of death ($-\gamma * y$). With the parameters ($\alpha = 0.01$, $\beta = 0.0075$, $\delta = 0.02$ and $\gamma = 0.005$), the starting values for the prey and predator (10.0, 5.0) and the $dt = 0.2$ for 1000 simulation steps produces the following chart with cyclic behavior, see **Figure 7**.

The result of the Lotka-Volterra model appears very "artificial" from an ecological point of view. A stable cycle is very rare in nature. Changing behavior to make it more realistic means changing the model. A fuzzy approach is much easier to understand and adapt here. Firstly, the Lotka-Volterra model is replaced by a set of fuzzy models – one for the prey, the other for the predator.

$$\frac{dx}{dt} = \alpha * f_{prey}(x, y) \quad [19]$$

$$\frac{dy}{dt} = \alpha * f_{predator}(x, y) \quad [20]$$

α is a scaling constant to control the systems behavior. The membership functions are triangular, and are not shown here. However, the rules are interesting. In the fuzzy model $f_{predator}(x,y)$ they combine input x {zero, small, medium, high, very_high} and input y {zero, small, medium, high, very_high} with output {hn = −0.05, mhn = −0.03, mn = −0.02, sn = −0.01, zero = 0.0, sp = 0.01, mp = 0.02, mhp = 0.03, hp = 0.05}. A subset of the rules is shown in **Table 2**.

Rule 25 means that a very_high prey and a very_high predator leads to a small negative (sn) growth of the predator. Ecologists are aware of the fact that, even under optimal conditions, a very large population of predators is often reduced by other factors (by disease, for example). This knowledge can be entered directly into the fuzzy rules. The resulting system behavior is "dampened", making it much more realistic, see **Figure 8**.

As shown in this example, as part of dynamic systems, fuzzy models can make the modeling more flexible, accurate, and easier to understand. A larger application of dynamic fuzzy models to estimate agricultural yield under climate change is given in (Wieland et al., 2013).

Dynamic-System Modeling Part

In order to study how landscape features change over time, the spatial part only requires landscape feature values at fairly widely spaced sampling times (communication times) $t = t0$, $t0 + COMINT$, $t0 + < 2 * COMINT$, . . .; the communication interval *COMINT* could be one day, one month, one year, and so on. The dynamic part, however, can increment the time in smaller steps DT to emulate continuous change. The dynamic part relates current and future data values at each grid point by ordinary differential equations. (We can instead, relate current and future feature values at different sampling times by difference equations, or by combinations of differential and difference equations.)

$$\frac{d}{dt} q_i = f(q_1, q_2, \ldots, p_1, p_2, \ldots, a_1, a_2, \ldots) \forall i \in (1..n) \quad [21]$$

For each grid point, the feature values $q1$, $q2$, . . . *state variables* start with given initial values. $p1$, $p2$, . . . are feature values related to the state variables by *defined-variable assignments*

$$\begin{aligned} p_1 &= g1(p_2,p_3,\ldots;q_1,q_2,\ldots;b1,b2,\ldots;t) \\ p_2 &= g2(p_1,p_3,\ldots;q_1,q_2,\ldots;b_1,b_2,\ldots;t) \\ &\ldots \end{aligned}$$

which may not involve recursive "algebraic loops". $a_1, a_2, \ldots b_1, b_2, \ldots$ are fixed parameter values associated with each grid point. The differential equation system (1) for each grid point is solved by numerical integration to produce time histories of the feature values $q_i = q_i(t)$ and $p_i = p_i(t)$. Such an equation system could, for instance, model the growth of a crop or the population dynamics of competing plant species at a certain point of the landscape.

Summary Fuzzy Modeling

A rule-based system lies at the heart of all fuzzy modeling of ecological problems. The inputs for this system are taken from a GIS describing a landscape or from a dynamic simulation system. The outputs of the model evaluate environmental indicators. The rule set is easy to understand, and should form the basis for discussing the model with other experts. Use of a fuzzy model should be accompanied by a strategy for its further development. This strategy should encompass visualization of the systems behavior, tests on real data sets, sensitivity analysis, etc., including an adaptation using a training algorithm.

The following advice can be offered concerning the development of fuzzy models:

- start with simple models (no more than three inputs, otherwise use two models)
- use triangular or trapezoidal shaped membership functions
- use FUZZY_AND Prod rather than FUZZY_AND Min
- use crisp values as the output type
- use training procedures to enhance the model
- check the model using graphical analysis tools, rule statistics and sensitivity analyses

References

Abrahamsen P and Hansen S (2000) Daisy: an open soil-crop-atmosphere system model. *Environmental Modelling and Software* 15: 313–330.

Deaton M and Winebrake J (2000) *Dynamic modeling of environmental systems*. New York: Springer.

Freeman JA (1994) *Simulation neural networks with mathematica*. Reading, MA: Addison-Wesley.

Haupt RL and Haupt SE (2004) *Practical genetic algorithms*. Hoboken, NJ: Wiley.

Hjek P (1998) *Metamathematics of fuzzy logic*. Dordrecht: Kluwer.

Korn GA (2004) Model-replication techniques for parameter-influence studies and Monte Carlo simu-lation with random parameters. *Mathematics and Computers in Simulation* 67(6): 501–513.

Korn GA (2013) *Advanced dynamic-system simulation model replication and monre Carlo studies*. Hoboken, NJ: Wiley.

Mirschel W, Wieland R, Voss M, Ajibefun I, and Deumlich D (2006) Spatial Analysis and Modeling Tool (SAMT):1 applications. *Ecological Informatics* 1: 77–85.

Nauck D, Borgelt C, Klawonn F, and Kruse R (2003) *Neuro-Fuzzy-systeme*. Wiesbaden: Vieweg.

Schlkopf B and Smola AJ (2002) *Learning with Kernels*. Cambridge, MA: MIT Press.

Sivanandam SN, Sumathi S, and Deepa N (2006) *Introduction to Fuzzy logic using MATLAB*. Berlin: Springer.

Wieland R, Voss M, Holtmann X, Mirschel W, and Ajibefun I (2006) Spatial Analysis and Modeling Tool (SAMT):1 structure and possibilities. *Ecological Informatics* 1: 67–76.

Wieland R, Mirschel W, Groth K, Pechenick A, and Fukuda K (2011) A new method for semi-automatic fuzzy training and its application in environmental modeling. *Environmental Modelling and Software* 26: 1568–1573.

Wieland R, Mirschel W, Nendel C, and Specka X (2013) Dynamic fuzzy models in agroecosystem modeling. *Environmental Modelling and Software* (in print).

Relevant Websites

http://www.esri.com/software/arcgis/.

http://www.zalf.de/en/forschung/institute/lsa/forschung/methodik/samt/Pages/default.aspx.

Grassland Models[☆]

Thorsten Wiegand, UFZ Helmholtz Centre for Environmental Research, Leipzig, Germany
K Wiegand, University of Jena, Jena, Germany
Sandro Pütz, UFZ Helmholtz Centre for Environmental Research, Leipzig, Germany

Introduction

Grasslands

A large proportion of the surface of the Earth is covered with grasslands (Figs. 1 and 2), which are ecosystems where the vegetation component is herbaceous in character, with grasses being predominant. Grasslands can be divided into tropical grasslands that occur in the same regions as savannas, that is, grasslands with scattered individual trees, and temperate grasslands. The major manifestations of temperate grasslands are, for example, the veldts of South Africa, the puszta of Hungary, the pampas of Argentina and Uruguay, the steppes of the former Soviet Union, and the plains and prairies of central North America. Grasslands are important ecosystems because they are frequently used for livestock grazing, they provide important ecosystem services, and may serve as carbon sinks.

Very roughly, grasslands can be classified as natural grasslands where grassland without trees constitutes the vegetation climax, seminatural grasslands which are mostly natural grasslands but modified by low intensity (grazing) management, and man-made grasslands which are either intensively managed natural grasslands that have been substantially altered or secondary grasslands, for example, created by the removal of natural forests for livestock production. The latter is common in temperate, Mediterranean, and tropical regions. In natural grasslands of pristine climax condition, perennial grasses and sedges are dominant and annual grasses are often restricted to locations where perennial plant cover has been disturbed. On the other hand, in seminatural grasslands, grazing often reduces the palatable perennials which are replaced by unpalatable and/or annual grasses.

As one of the world's major ecosystems, grasslands have been an important subject of basic and applied ecological research attempting to understand the ecological processes and factors occurring in grasslands and their effects on grassland dynamics, productivity, and diversity. Grasslands are model ecosystems for basic ecological research to investigate the effects of processes such as competition, seed dispersal, and reproductive strategies on coexistence and diversity. They are of interest in rangeland science where ecological understanding is needed to derive optimal grazing management strategies that maximize fodder or animal production and minimize the risk of degradation. More recently, grassland ecosystems gained interest in climate change ecology because of their importance as carbon sinks.

Overview Over Models of Grasslands

Grasslands are complex ecosystems, and understanding the ecological processes and factors determining its dynamics, productivity, and biodiversity requires use of combined approaches of field measurements, experimentation, data analysis, and modeling. Thus,

[☆]*Change History:* March 2018. Todd M. Swannack updated the references. No other changes were made.

Fig. 1 Examples of grasslands in South Africa (*top*) and Argentina (*bottom*). (*Top*) Experimental plots of a *Themeda triandra—Cymbopogon plurinodis* grassland located at Bloemfontein, South Africa (28°50′ S; 26°15′ E, altitude 1350 m). The left photo shows a nondegraded state dominated by the perennial bunchgrass *T. triandra* and the right photo shows a degraded plot dominated by the stoloniferous perennial *Tragus koelerioides* and the short-lived perennial bunchgrass *Aristida congesta*. (*Bottom*) View of a grass steppe dominated by *Festuca pallescens* ("coiron blanco") that characterize the sub-Andean district of the Patagonian Phytogeographic Province located in a narrow north–south strip between 71° W and 71° 30′ W. (*Top*) Thorsten Wiegand. (*Bottom*) Nestor Fernandez.

ecological models are important tools for ecologists working in grasslands. However, the term "model" has been used in such a variety of contexts that it has become almost meaningless, unless used with some qualifications. In the present context, models may be regarded as a simplified and formalized representation of ecological processes, either using mathematical or computer simulation techniques, which produce, based on a set of assumptions, a quantitative output. Grassland models are as varied as the purposes for which they have been constructed, and cover various spatial and temporal scales and various degrees of detail. Our focus lies on ecological models which are concerned with grassland population and community dynamics with less attention to grassland models of matter and energy flows or biophysical processes. In this article, most grassland models are conceptual, empirical, analytical, or simulation models.

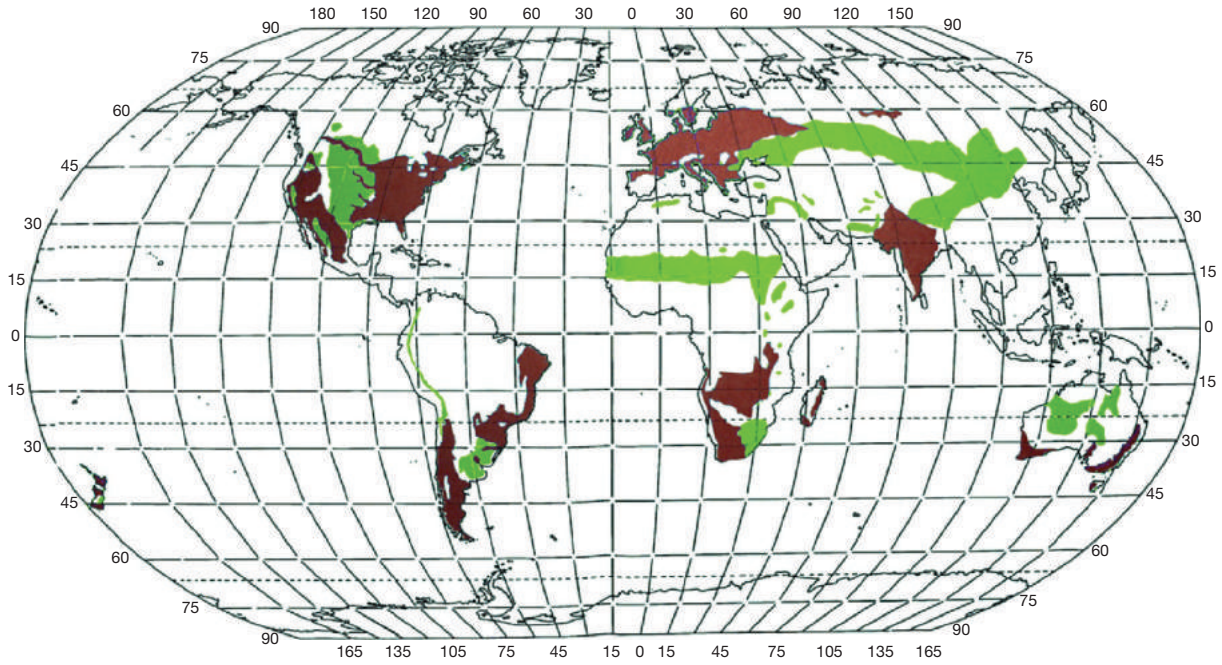


Fig. 2 Grassland vegetation map of the world. *Green*: Regions in which grassland without trees is the vegetation climax over most of the area, and *brown*: man-made grasslands. Adapted from Snaydon, R.W. (1987). Managed grasslands. *Ecosystems of the world*, (vol. 17B). Amsterdam: Elsevier; and (1992). Natural grasslands, In: Coupland, R.T. (ed.). *Ecosystems of the world*, (vol. 8A). Amsterdam: Elsevier.

Modeling Questions and Models

In the following, we present an overview over the most important grassland models, structured by their scientific questions rather than by model type. A summary of these models is given in Table 1. The major variables and properties of ecological grassland models are temporal and spatial grassland dynamics and diversity, grassland productivity, water dynamics, and flow of other matter and nutrients. Grazing is a predominant theme in grassland models.

Grassland Dynamics and Processes of Coexistence

An important topic in grassland research, both from theoretical and applied point of view, is the temporal and spatial dynamics of species abundance and composition. In applied rangeland science, the dominant question is how to manage grasslands for maximal long-term domestic livestock production without degrading the grassland to an extent that would make it unsuitable for further grazing. In applied conservation ecology, the main question is how to promote survival of small and fragmented remnants of natural grasslands. A specific conservation problem in central Europe is that an increasing intensification of agriculture leads at the same time to abandonment of former secondary grasslands maintained by grazing. From a more theoretical perspective, grassland ecologists are interested in the processes and factors that determine the dynamics of grasslands and maintain their biodiversity. Grasslands are good systems to study questions of coexistence and biodiversity in sessile organisms (e.g., vascular plants), because they are relatively easy to monitor, and are less complex than tropical forests or coral reefs but still species rich enough for studying questions of coexistence. Another theoretical question in biogeography is why and under which climatic and environmental conditions grasslands can exist.

Succession in Grasslands

Understanding succession, that is, the change in vegetation of an area over time, is fundamental for the theoretical understanding of grasslands as well for an understanding of the reactions of grasslands to management. Models of grassland succession, especially applied models, are mostly based on the Clementsian theory of ecological succession. Models which explicitly deal with spatial structures are more influenced by the work of A. S. Watt in the 1940s, who proposed that plant communities are composed of a mosaic of patches in different states of a natural cycle of disturbance and regeneration.

Table 1 Overview of different exemplary modeling studies in grasslands

Authors	*Type of model*	*Description*	*Area*
Rangeland models			
Westoby et al. (1989)	Conceptual	Alternative stable states with discontinuous and irreversible transitions, nonequilibrium dynamics	Arid and semiarid rangelands
Noy-Meir (1975)	Nonspatial differential equation	Applies classical predator–prey models to plant–herbivore systems, detects dual stability	
Fernandez-Gimenez and Allen-Diaz (1999)	Conceptual	Assesses the extent to which the current nonequilibrium models of rangeland vegetation dynamics apply	Mongolian rangeland ecosystems
Bosch and Gauch (1991)	Statistical	Description of degradation gradient method for range condition assessment	Grasslands of South Africa
Phelps and Bosch (2002)	Statistical	Degradation gradient method in conjunction with state-and-transition models of rangeland dynamics and condition	Mitchell grasslands, central western Queensland, Australia
Spatially explicit models on succession and disturbance in grasslands			
Coffin and Lauenroth (1989, 1990)	Gap model, simulations	Introduction to STEPPE model, evaluates the effects of disturbances at the scale of a landscape for a semiarid grassland	Semiarid grassland in north-central Colorado, United States
Coffin and Urban (1993)	Gap models, simulations	Compares the STEPPE model to a structurally similar gap model for forest dynamics (ZELIG)	
Lauenroth et al. (1993)	STEPPE, CENTURY, simulations	Couples two models to study interactions between vegetation structure and ecosystem processes	Semiarid grassland in north-central Colorado, United States
Peters (2002)	Gap model, similar to STEPPE, simulations	Studies effects of climatic fluctuations and disturbance on regional patterns of vegetation dynamics at an arid–semiarid grassland ecotone	Chihuahuan desert, central New Mexico, United States
Moloney and Levin (1996)	Spatially explicit, grid-based, simulations	Varies components of disturbance architecture systematically to determine their impact on population dynamics at the scale of the landscape	Jasper Ridge serpentine grassland, United States
Wu and Levin (1994)	Spatially explicit patch-based model	Studies landscape pattern and process dynamics	Jasper Ridge serpentine grassland, United States
Tan and Smeins (1996)	Statistical model, neural networks	Predicts grassland community changes with an artificial neural network model	Grassland communities near Hays, Kansas, United States
O'Connor (1993)	Size–structured matrix models	Investigates how population growth rate depends on factors such as rainfall or grazing	Perennial grasses of two African savannas
Optimal life-history strategies, competition, coexistence, and biodiversity			
Lavorel et al. (1994)	Spatially explicit, two-species, simulation	Spatiotemporal dispersal strategies and annual plant species coexistence in a structured landscape	Species-rich Mediterranean old-fields
Matsinos and Troumbis (2002)	Cellular automaton model	Models competition, dispersal and effects of disturbance in the dynamics of a grassland community	Grasslands in Lesbos, Greece
Schwinning and Parsons (1996)	Spatially explicit cellular automaton model	Extends a pasture model by Thornley et al. in 1995 to study coexistence mechanisms for grasses and legumes including selective grazing and spatial considerations	Perennial rye-grass and white clover communities
Thornley et al. (1995)	Physiological models	Studies complex dynamics in a carbon–nitrogen model of a grass–legume pasture	
Winkler and Fischer (2002)	Grid-based model	Investigates the role of vegetative and seed dispersal within habitats for optimal life histories of clonal plants	
Bolker and Pacala (1999)	Analytical model, moment equations	Aims to understand at a general level how plants coexist in communities	
Spatial structures in arid and semiarid grasslands			
Dunkerley (1997)	Cellular automaton model	Investigates the development of banded vegetation communities in grass- and shrublands	Western NSW grass- and shrublands, Australia
Rietkerk et al. (2004)	Partial differential equations, numerical simulations	Investigates vegetation pattern formation in arid ecosystems and hypothesizes that they may be the result of spatial self-organization, caused by net displacement of surface water to vegetated patches	
Modeling primary production in grasslands			
Le Houèrou et al. (1988)	Regression models	Relationship between the variability of primary production and the variability of annual precipitation in world arid lands	Arid lands of the world
Paruelo et al. (1999)	Regression models and NDVI	Studies how grassland precipitation-use efficiency varies across a resource gradient. Uses 11 temperate grassland sites worldwide, and 19 grassland sites across the central grassland region of North America	Grassland sites worldwide
Paruelo et al. (1997, 2000)	Regression models, NDVI	Estimates ANPP for the central grassland region of the US and subhumid pampa rangelands in Argentina	Central grassland (United States), subhumid pampa (Argentina)
Prince (1991)	Satellite remote sensing	Determines primary production for Sahelian grasslands, 1981–88	Sahelian grasslands

Table 1 (Continued)

Authors	*Type of model*	*Description*	*Area*
Ecosystem modeling and flows of energy and matter in grasslands			
Nouvellon et al. (2001)	Remote sensing, ecosystem model	Couples a grassland ecosystem model for semiarid perennial grasslands with Landsat imagery to calibrate parameters of ecosystem model	Grasslands in southeastern Arizona (United States)
Van Dyne (1972)		Describes organization and management of an IBP "big biology" program	
Coughenour and Chen (1997)	Linked ecosystem model	Assessment of grassland ecosystem responses to atmospheric change using linked plant-soil process models	Grasslands of Colorado and Kansas (United States) and Kenya
Parton et al. (1993)	CENTURY	Parametrizes CENTURY for the world's major grassland types to model biomass and soil organic-matter dynamics in grasslands	Grassland biome worldwide
Gilmanov et al. (1997)	CENTURY	Uses long-term data from several sites to assess the performance of CENTURY	
Hibbard et al. (2003)	CENTURY and matrix transition model	Linked CENTURY to a transition matrix model to simulate the displacement of grassland communities under heavy livestock grazing and climate events	Grassland and thorn woodland in southern Texas (United States)
Paruelo and Sala (1995)	Water balance model	Calculates the amount of water evaporated from the soil and transpirated by the canopy to estimate water losses in a Patagonian steppe	Shrub–grass steppe in Patagonia (Argentina)

Rangeland Models

The range succession model. The range succession model based on the Clementsian theory of ecological succession formed the conceptual framework for most grazing management up to the 1980s. It is supposed that, in the absence of grazing, a rangeland has a single persistent state (the climax), whereas grazing causes continuous and reversible transitions of the grassland state along a single, monotonic gradient between an undisturbed climax and an overgrazed subclimax vegetation state. Therefore, the grazing pressure can be made equal and opposite to the successional tendency, producing an equilibrium in the vegetation at a set stocking rate (Fig. 3A). A sustainable yield of livestock products can be harvested from such an equilibrium. The model recognizes that vegetation is affected when rainfall varies from year to year and supposes that grazing and interannual variation in rainfall cause equivalent changes in the vegetation. Therefore, management should respond to drought by reducing grazing.

State-and-transition models. Over the years, however, substantial empirical evidence accumulated of cases where the assumptions of the range succession model were not met, especially in arid and semiarid environments. To deal with the complex dynamics of semiarid and arid ecosystems, scientists such as M. Westoby, B. Walker, and I. Noy-Meir suggested (by the end of the 1980s) that these ecosystems could be described in terms of discrete states and inter-state transitions (Fig. 3C). Transitions could be triggered by natural events (e.g., rainfall, drought, and fire) or by management actions (e.g., removal of herbivores, altered intensity or timing of herbivory, and burning). Changes in range condition are not unidirectional, but multiple pathways of system transitions to alternative states may exist, depending on the particular sequence of driving events (Fig. 3C).

Such state-and-transition models are valuable tools for describing the structure of the ecosystem to identify irreversible transitions and alternate stable states; however, they provide little information applicable to forecasting and prediction. Additional models are needed to quantify the temporal scales of the transitions, to identify rare events that drive semiarid and arid ecosystems, and to improve the user's understanding of ecosystem dynamics over a long temporal scale.

Degradation gradient models. The degradation gradient model combines in some sense the concepts of Clementsian succession and state-and-transition models (Fig. 3B). It is based on the idea that vegetation compositional changes along a grazing gradient are indicative for the ecological condition. This statistical method was developed in the 1990s by O. J. H. Bosch and H. G. Gauch for the semiarid South African grasslands. Key element of the approach is a classification of the species according to their response to grazing as increaser and decreaser species (Fig. 3B). This classification is derived by means of multivariate statistics. Species composition data are collected from grasslands in various stages of degradation. Based on a series of statistical analyses of these data, the degradation gradient is constructed. The model recognizes that irreversible transitions may exist which are caused by soil loss or major changes in floristic composition. This method is used for range assessment where species composition data of a sample is compared to the gradient.

Spatially Explicit Models on Succession and Disturbance in Grasslands

Disturbance can play a major role in structuring grasslands by producing a spatiotemporal mosaic of patches by locally resetting the successional clock after a disturbance. The spatiotemporal distribution of species within the resulting mosaic depends upon an interaction between species' life-history traits and the spatial and temporal structure of the ecological processes controlling species' distributions. The availability of powerful computers and the advent of spatially explicit simulation models revitalized the

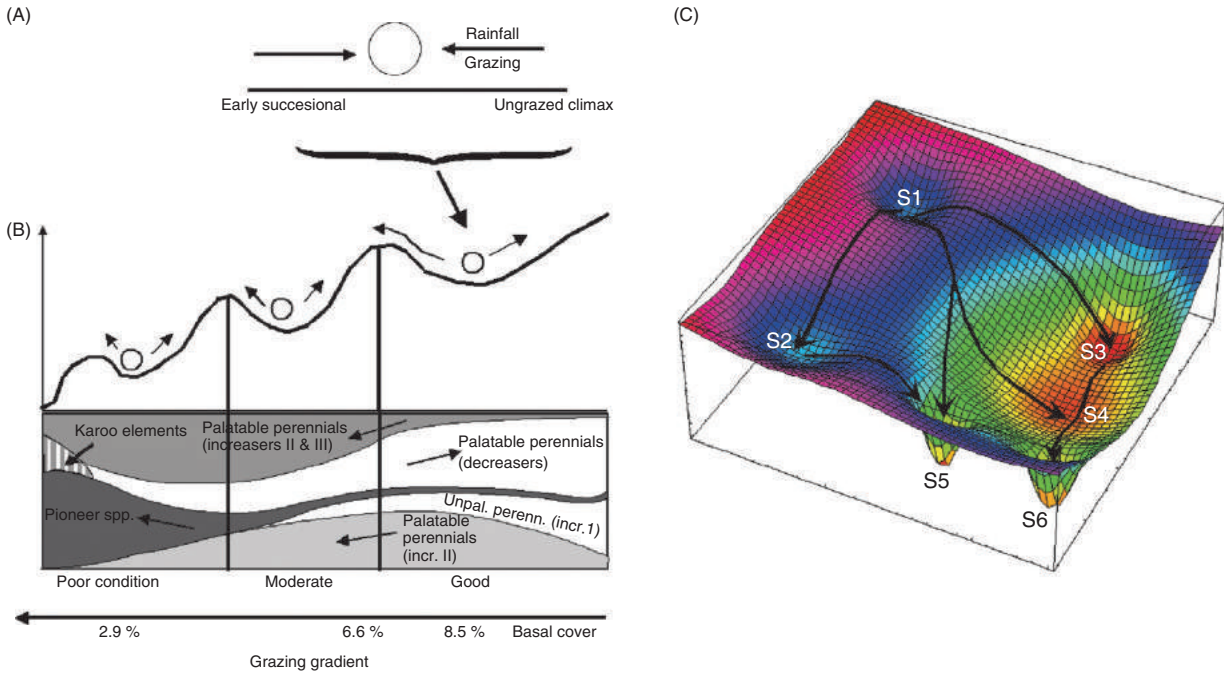


Fig. 3 Illustration of different conceptual rangeland models. (A) Range succession model. Vegetation changes are reversible and unidirectional in response to grazing and rainfall. (B) Degradation gradient model for South African grasslands. The system responds to smaller perturbations (grazing and rainfall) unidirectionally and reversibly as predicted by the range succession. Larger perturbations (drought, overgrazing) may cause the system to cross an irreversible threshold where changes in soil conditions (related to reduced basal cover) and species composition may hinder the system to return. The pathway of degradation is unidirectional. (C) State and transition model. Multiple equilibria and multidirectional pathways of degradation in response to different driving events. S1, S2: Good condition domains; S3, S4: moderate condition domains; and S5, S6: poor condition domain.

conceptual work done by Watts in the 1940s and allowed to elucidate process from patterns. In this line, several spatially explicit plant dynamics models investigate questions related to succession and disturbance.

The Steppe model. D. B. Coffin and W. Lauenroth developed at the end of the 1980s a spatially explicit gap dynamics simulation model (STEPPE) to evaluate the effects of disturbances at the scale of a landscape for a semiarid grassland in north-central Colorado, United States. The approach goes back to spatially implicit gap models of forest succession (see Forest Models). These models simulate succession in a gap left after the death of a large tree. Succession dynamics are then estimated by the average behavior of 50–100 plots of the size of a single large tree. The exact location of each individual is not used to compute competition in these models. However, gap models can be made spatially explicit by linking several individual plots together.

The STEPPE model is a gap model and simulates the establishment, growth, and death of individual grass plants on a small plot ($0.12\ m^2$) through time at an annual time step. Landscapes were simulated either as a collection of independent plots or as a collection of interacting plots. In the 1990s, extending their first approach, W. Lauenroth and colleagues coupled compartment models of nutrient cycling and soil water–plant relation with the STEPPE model. The aim was to understand the interactions between vegetation structure and ecosystem processes across ecosystems. However, the approach of coupling several models was quickly abandoned, probably because of problems with increased model complexity.

A gap model similar to STEPPE was developed at the beginning of the 2000s by D. B. Peters (formerly Coffin). This individual-based, gap dynamics model (ECOTONE) was used to predict the effects of climatic fluctuations on regional patterns of vegetation dynamics, and the effects of disturbance on vegetation dynamics at an arid–semiarid ecotone between shortgrass steppe grassland and a Chihuahuan desert community in central New Mexico, United States.

The Jasper model. K. A. Moloney and S. A. Levin developed in the 1990s a spatially explicit simulation model of a serpentine grassland, focusing primarily on the role of disturbance. The model is hierarchical in design and population dynamics were modeled as occurring within local sites, which were then arranged to form a landscape and interact primarily through seed dispersal. Several components of the disturbance architecture were varied systematically among model runs to determine their impact on population dynamics at the scale of the landscape. Results suggested that predicting the impact of disturbance on ecological communities will require an explicit understanding of at least some aspects of the spatial and temporal architecture of the disturbance regime.

Neural networks. Where long-term site data on species composition and environmental factors are available, an approach using artificial neural networks may be feasible to predict grassland succession. The method takes advantage of the ability of neural networks to learn, recognize, and generalize from patterns contained in the ecological data. In the 1990s, S. S. Tan applied this

approach to a 30-year data set on vegetation changes and climatic factors in grassland communities in Kansas (United States). Their model predicted future community composition using input data on present conditions that have not been used to develop the model. The model performs well for 1–4 year predictions and performance significantly deteriorated from the fifth year on. One disadvantage of the neural networks is that it is purely descriptive and gives little insight into underlying causes and mechanisms of grassland succession. On the other hand, if predictions are required for highly complex systems as they occur in natural rangeland communities, they are an appropriate technique for qualitative analysis and statistical forecasting.

Matrix models. Matrix models of succession are mathematically and conceptually the most straightforward among the succession models. Matrix models are constructed by determining the probability that the vegetation on a local plot will transform to some other vegetation state after a given time interval. To construct the model, the vegetation must be classified into identifiable states. The model consists of a vector representing the state variables of the system and a matrix containing the (usually) constant probabilities of possible state transitions. A question of interest is how a population growth rate depends on factors such as rainfall or grazing. Matrix models provide the means for calculating population growth rates, thereby allowing for an assessment of the influence of different population processes such as seed production and growth and mortality. For example, T. G. O'Connor used in the 1990s stage-structured matrix models to assess the influence of rainfall and grazing on the demography of some African savanna grasses. Results showed that the population growth rates of most species were positively correlated, indicating that an extrinsic force, presumably rainfall, had the greatest effect on population growth.

Optimal Life-History Strategies, Competition, Coexistence, and Biodiversity

A considerable number of models on grassland dynamics and processes are motivated from general theory in ecology and are often put into a spatially explicit context. These models, developing since the mid-1990s, are concerned with tradeoffs in optimal reproduction strategies, and investigate how optimal strategies depend on disturbance and other spatially explicit factors such as a patchy habitat. One early example of this approach is a spatially explicit, two-species simulation model by S. Lavorel to examine the interaction between dispersal, dormancy, and small-scale disturbances on coexistence of two annual plant species in landscapes with varying degree of patchiness. Coexistence patterns depended on the degree of suitability and the patchiness of the landscape, mostly in relation to the interactions between landscape structure and mean dispersal distance.

A model of grassland community dynamics developed in the 2000s by Y. G. Matsinos and A. Y Troumbis aimed to quantify the role of competition and dispersal under disturbance, and investigated the resilience of the communities with respect to gap-creating disturbances that were imposed at different spatial extent. Model results showed that plants with longer seed dispersal distances may have a competitive advantage in their colonization success as compared to the better competitors, especially in the cases of a disturbance-mediated creation of gaps in the landscape. An increase of the species number led to more stable end communities and a higher vegetation cover in the landscape.

A series of papers by S. Schwinning and A. J. Parsons investigated, in the 1990s, coexistence mechanisms for grasses and legumes in grazing systems. This is an important question since legumes fix nitrogen which in turn is beneficial for grasses.

Other questions investigated with spatially explicit models attempt to gain an understanding of clonal growth and ramification of grass tillers in grasslands. For example, E. Winkler and colleagues developed a series of grid-based models of grassland communities to investigate long-term control of species abundances and reproduction strategies in patchy landscapes undergoing disturbances. One example is a study with M. Fischer investigating tradeoffs between sexual and vegetative reproduction of clonal plants. Depending on spatial habitat structure and disturbances, different reproduction strategies lead to different long-term fitness. The model simulated plant population dynamics on a two-dimensional cellular grid consisting of 70×70 square cells. In an extension of this approach, J. Stöcklin and E. Winkler used a spatially explicit, individual-based metapopulation model of *Hieracium pilosella* to examine the consequences of tradeoffs between vegetative and sexual reproduction and between short and far-distance dispersal of seeds. They found that in a spatially heterogeneous landscape, sexual seed production in a clonal plant is advantageous even at the expense of local vegetative growth.

While most of the questions from this section were analyzed by means of grid-based models, advances in analytical modeling made by B. M. Bolker, S. W. Pacala, and others at the end of the 1990s suggested that spatial interactions may be approximated by means of moment equations that describe changes in the mean densities and spatial patterns (covariances) of competing species. The formalism of moment equations is borrowed from physics where it was used to describe phase transitions and is an elegant way to approximate simple spatial dynamics by tracking the dynamics of the first moments (mean densities) and second moments (variances and covariances or spatial covariances) of the spatial distributions of populations. For example, to understand at a general level how plants coexist in communities, Bolker and Pacala studied three different strategies to compete for resources in a spatially variable environment: colonizing new areas, exploiting resources in those areas quickly before other plants arrive, or tolerating competition once other plants arrive. However when adding more realism the moment equations become quickly lengthy, simulations are often required to find an appropriate moment closure.

Spatial Structures in Arid and Semiarid Grasslands

In mesic regions, patchy spatial vegetation patterns arise through the interplay of succession and disturbances. In the past, this has been shown in many studies of patch dynamics for mesic forests and grasslands, for example, summarized in the classic book by S. T. A. Pickett and P. S. White. However, patchy structures are also frequently observed in water-limited arid and semiarid ecosystems,

where distinct banded and spotted vegetation patterns are often found. There are numerous hypotheses on the origin of the distinctive patterning. For example, banded vegetation may be a remnant of more complete vegetation cover diminished by climatic deterioration or by grazing disturbance. Other hypotheses assume these patterns to be natural with downslope water re-allocation from bare areas to vegetated bands as a key-process or are based solely on the intrinsic dynamics of the vegetation without slope-induced anisotropy.

A number of cellular simulation models were developed to investigate the robustness and origin of such patterns. For example, end of the 1990s, D. L. Dunkerley modeled banded vegetation communities in western New South Wales Australian grassland and shrublands to test the hypothesis that water partitioning in spatially unstructured plant communities may lead to the development of banding. The model shows that without any climatic change or external disturbance, strongly developed banding can emerge from an initially random distribution of plants.

Other models used differential equations to search for possible unifying mechanisms to explain these spatial patterns. One hypothesis is that spatial patterns establish themselves through a Turing-like spatial instability depending only on a tradeoff between facilitative and competitive interactions among plants. This hypothesis goes back to the 1950s where A. M. Turing described morphogenesis in chemical systems. These models produce patterns superficially similar to banded and spotted vegetation, which then is taken as evidence for the validity of the underlying hypothesis. However, reproduction of a pattern is not proof that the modeled processes represent the natural processes. Since these top-down models are usually not explicitly related to specific spatial and temporal scales, they are difficult to test and their ecological content remains unclear.

Models on the Impact of Grazing on Grasslands

The impact of grazing on the dynamics and productivity of grasslands has been an important subject of basic and applied ecological research. Theoretical and applied questions are intimately linked since developing sustainable grazing management requires an understanding on the dynamics of the grazing system. Early models on grazing systems were based on the Clementsian theory of ecological succession, and since the 1970s biomass–herbivore grazing systems were modeled in analog to predator–prey models developed one decade earlier. Since the 1980s, however, the equilibrium concept was increasingly challenged, and in the late 1980s nonequilibrium concepts and models emerged which stipulated environmental variation (due to rainfall variability) and spatial heterogeneity. Numerous grazing models have modified the early predator–prey differential equation models, and spatially explicit and rule-based simulation models are increasingly used to analyze specific grazing systems. For a detailed treatment of grazing models, see Grazing Models.

Modeling Primary Production in Grasslands

Regional and global patterns in aboveground net primary production (ANPP) and their determinants have long interested ecologists. Understanding ANPP patterns and controls through time is particularly important in grasslands where grazing is the most important economic activity. Coping with temporal changes in the availability of forage is a prerequisite for the efficient and sustainable use of natural vegetation. More recently, interest in ANPP has intensified as projected global changes in climate, nitrogen deposition, and land use threaten to alter ecosystem carbon and energy flow. Data on primary production are also important to calibrate, parametrize, and evaluate terrestrial biosphere models and for modeling the global carbon cycle.

Regression Models

Early approaches to assess primary production in grasslands, such as the work by H. N. Le Houérou in the 1980s, used empirical models which correlated primary productivity with mean annual precipitation or evapotranspiration. In general, the relationships of production with environmental variables were derived from long-term averages for many sites distributed across environmental gradients (spatial models). ANPP increases linearly along spatial precipitation gradients within the range of 200–1300 mm $year^{-1}$ in North American, South American, and African grasslands.

However, much less is known about the controls of the temporal, inter-annual variation of productivity at a given site (temporal models). Temporal models relating time series of ANPP and annual precipitation for single sites have shown lower slopes (=water use efficiency) and regression coefficients than the spatial models. Additionally, memory and carryover effects, for example, due to storage of carbohydrates in the root system and structural inertia, might play an important role in the functioning of semiarid grasslands. Memory and carryover effects buffer fluctuations in production if wet, productive years alternate with dry, less productive years and amplify fluctuations if wet or dry sequences of several years take place. Identifying and quantifying such memory and carryover effects is an important challenge for global models which mostly use simple linear relationships with precipitation.

Use of Remote Sensing Data

Biomass harvesting is the most common way to estimate ANPP in grasslands, but because of the large effort and the detailed spatial scale, harvesting methods are rather limited in their spatial and temporal extent. Remote-sensing techniques are a fast and nondestructive method for estimating ANPP at a regional scale over longer time periods. NDVI is the most commonly used radiometric index for estimating ANPP in grasslands and the annual summed NDVI can be used as a surrogate for annual ANPP because, in ecosystems dominated by grasses or deciduous life forms, the absorbed photosynthetically active radiation of plant canopies (APAR) and net primary production are directly related. Biomass estimates for grassland using NDVI have been performed, for example, by the group of J. M. Paruelo in the late 1990s and 2000s for the Central Grassland Region of the United States, and for subhumid pampa rangelands and the Patagonia steppes in Argentina. S. D. Prince determined in the early 1990s primary production for Sahelian grasslands.

Additionally, NDVI data allow determination of other important ecosystem characteristics such as the degree of seasonality, and the start and the end of the growing season. Relating NDVI characteristics over regional gradients with climatic and other environmental variables allow for inference on the controls of primary production and ecosystem functioning. Remote-sensing data may also be used to calibrate ecosystem models, for example, by minimizing the difference between the measured and the simulated NDVI. The utility of this approach was demonstrated by a study of Y. S. Nouvellon in the 2000s who coupled a grassland ecosystem model for semiarid perennial grasslands in southeastern Arizona (USA) with Landsat imagery for a 10-year simulation of carbon and water budgets.

Ecosystem Modeling and Flows of Energy and Matter in Grasslands

Although the first interest in ecosystem modeling peaked in the 1970s with the International Biological Program (IBP), it is now having a rebirth with the recent interest in predicting ecosystem effects of global change. In the 1960s and 1970s "big biology" projects were initiated by G. M. Van Dyne to study ecosystems including grasslands. One of the main problems of this big biology project was the attempt to model everything without a clear and focused research question. At the end, the resulting models were nearly as complex as nature itself and they could not be properly analyzed and thus understood. The objective of the IBP grassland simulation model was to simulate biomass dynamics in a variety of grassland types and the response of the system to irrigation, fertilization, and cattle grazing. The model comprises several submodels, that is, abiotic, producers, mammals, grasshoppers, decomposers, nitrogen, and phosphorus. Some of the submodels originally designed to be incorporated into the ELM model never reached this objective but developed a life of their own.

Plant Growth and Production

A number of models, reviewed in detail by J. D. Hanson and colleagues in the mid-1980s, simulated/predicted plant growth and production of grassland ecosystems. For example, the AFRICA model included processes like shoot growth, tillering, root growth, photosynthesis, and nitrogen uptake for single plants. The aim of AFRICA was to model primary production of perennial graminoids and it unites physiological processes and morphometric traits. In a study in the late 1990s, M. B. Coughenour and D.-X. Chen linked models of photosynthesis, plant growth, and biophysical processes with models that simulate water, nutrient, and carbon flows through plant–soil ecosystems. The linked ecosystem model was applied to examine ecosystem-level responses to CO_2, temperature, precipitation, and global-warming scenarios in grasslands of Colorado and Kansas (United States) and Kenya. Using similar ecosystem model approaches, several models have been developed for semiarid perennial grasslands that allow multiyear simulations of plant growth patterns by accounting for carbohydrate storage in root systems and further translocation to aboveground regrowth.

Biochemistry Models

Other models analyzed the soil organic matter dynamics in response to changes in management and climate. These models described the flow of energy and matter (Conceptual Diagrams and Flow Diagrams) in form of balance equations. Following the approach of H. T. Odum, the focus of these models was not on biotic interactions between species but rather on the flows of energy and nutrients, treating plants basically as composers and decomposers.

The most prominent model of this type is the "CENTURY model" (Fig. 4) developed by W. J. Parton in the 1980s. CENTURY is a model of terrestrial biochemistry of grasslands based on the relationships between climate, human management (fire, grazing), soil properties, plant productivity, and decomposition. Studies performed with this model include efforts to link models describing plant and soil responses to the large-scale modeling of global change effects. The model is intended as a generic model whose basic balances of the different flows in grasslands can be calibrated to specific systems. In the 1990s, Parton and colleagues parametrized CENTURY for the world's major grassland types to predict the biomass and soil organic matter dynamics of the grassland biome worldwide.

In the late 1990s, the performance of CENTURY was assessed by T. G. Gilmanov and colleagues using long-term data collected under IBP and at research stations within the former USSR. They found that CENTURY reproduced the seasonal, mid-term, and, in

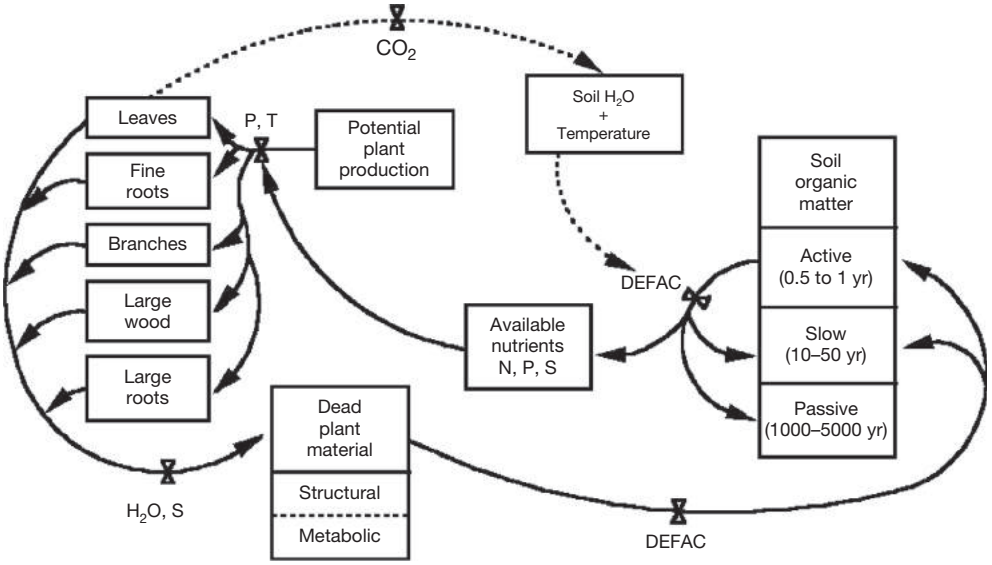


Fig. 4 Flowchart of the CENTURY model. Supplement to Metherell et al., (1993). Century Manual. CENTURY Soil organic matter model environment. Technical documentation. Agroecosystem version 4.0. Great Plains System Research Unit Technical Report No.4. USDA-ARS, Fort Collins, Colorado, USA. Freely available at http://www.nrel.colostate.edu/projects/century/Century_Slides.ppt.

some cases, long-term dynamics in aboveground biomass in a wide range of grassland ecosystems. Model discrepancies were attributed to changes in species composition and short-term responses to intermittent rainfall that are missed by the monthly timestep of the model. In another application, K. A. Hibbard and colleagues assessed in the 2000s the magnitude of changes in plant and soil carbon and nitrogen pools in a subtropical landscape undergoing succession from grassland to thorn woodland in southern Texas (United States). They linked CENTURY to a transition matrix model and parametrized grass and tree production submodels of CENTURY with field data. The Markov transition matrix model simulated the displacement of grassland communities under land-use practices (heavy livestock grazing, no fire) and climate events.

Water-Balance Models

Other models, such as a model of water balance developed by J. M. Paruelo in the 1990s, calculate the amount of water evaporated from the soil and transpirated by the canopy. Typical questions asked with such models are: What are, on a long-term basis, the magnitude of evaporation, transpiration, and deep percolation in a water-limited steppe? How will elevated CO_2 change the fluxes of soil water to the atmosphere and ground water?

References

Bolker B and Pacala SW (1999) Spatial moment equations for plant competition: Understanding spatial strategies and the advantage of short dispersal. *American Naturalist* 153: 575–602.

Bosch OJH and Gauch HG (1991) The use of degradation gradients for the assessment and ecological interpretation of range condition. *Journal of the Grassland Society of southern Africa* 8(4): 138–146.

Coffin DP and Lauenroth WK (1989) Disturbances and gap dynamics in a semiarid grassland: A landscape-level approach. *Landscape Ecology* 3: 19–27.

Coffin DP and Lauenroth WK (1990) A gap dynamics simulation model of succession in a semiarid grassland. *Ecological Modeling* 49: 229–266.

Coffin DP and Urban DL (1993) Implications of natural history traits to system-level dynamics: Comparisons of a grassland and a forest. *Ecological Modeling* 67: 147–178.

Coughenour MB and Chen D-X (1997) Assessment of grassland ecosystem responses to atmospheric change using linked plant–soil process models. *Ecological Applications* 7: 802–827.

Dunkerley DL (1997) Banded vegetation: Development under uniform rainfall from a simple cellular automaton model. *Plant Ecology* 129: 103–111.

Fernandez-Gimenez ME and Allen-Diaz B (1999) Testing a non-equilibrium model of rangeland vegetation dynamics in Mongolia. *Journal of Applied Ecology* 36: 871–885.

Gilmanov TG, Parton WJ, and Ojima DS (1997) Testing the "CENTURY" ecosystem level model on data sets from eight grassland sites in the former USSR representing a wide climatic/soil gradient. *Ecological Modeling* 96: 191–210.

Hibbard KA, Schimel DS, Archer S, Ojima D, and Parton W (2003) Grassland to woodland transitions: Integrating changes in landscape structure and biogeochemistry. *Ecological Applications* 13: 911–926.

Lauenroth WK, Urban DL, Coffin DP, et al. (1993) Modeling vegetation structure-ecosystem process interactions across sites and ecosystems. *Ecological Modeling* 67: 49–80.

Lavorel S, O'Neill RV, and Gardner RH (1994) Spatio-temporal dispersal strategies and annual plant species coexistence in a structured landscape. *Oikos* 71: 75–88.

Le Houèrou HN, Bingham RL, and Skerbek W (1988) Relationship between the variability of primary production and the variability of annual precipitation in world arid lands. *Journal of Arid Environments* 15: 1–18.

Matsinos YG and Troumbis AY (2002) Modeling competition, dispersal and effects of disturbance in the dynamics of a grassland community using a cellular automaton model. *Ecological Modeling* 149: 71–83.

Moloney KA and Levin SA (1996) The effects of disturbance architecture on landscape-level population dynamics. *Ecology* 77: 375–394.
Nouvellon Y, Moran MS, Seen DL, et al. (2001) Coupling a grassland ecosystem model with Landsat imagery for a 10-year simulation of carbon and water budgets. *Remote Sensing of Environment* 78: 131–149.
Noy-Meir I (1975) Stability of grazing systems an application of predator–prey graphs. *Journal of Ecology* 63: 459–482.
O'Connor TG (1993) The influence of rainfall and grazing on the demography of some African savanna grasses: A matrix modeling approach. *Journal of Applied Ecology* 30: 119–132.
Parton WJ, Scurlock JMO, Ojima DS, et al. (1993) Observations and modeling of biomass and soil organic-matter dynamics for the grassland biome worldwide. *Global Biogeochemical Cycles* 7: 785–809.
Paruelo JM and Sala OE (1995) Water losses in the Patagonian steppe: A modeling approach. *Ecology* 76: 510–520.
Paruelo JM, Epstein HE, Lauenroth WK, and Burke IC (1997) ANPP estimates from NDVI for the Central Grassland Region of the US. *Ecology* 78: 953–958.
Paruelo JM, Lauenroth WK, Burke IC, and Sala OE (1999) Grassland precipitation use efficiency varies across a resource gradient. *Ecosystems* 2: 64–69.
Paruelo JM, Oesterheld M, Di Bella CM, et al. (2000) Estimation of primary production of subhumid rangelands from remote sensing data. *Applied Vegetation Science* 3: 189–195.
Peters DPC (2002) Plant species dominance at a grassland–shrubland ecotone: An individual-based gap dynamics model of herbaceous and woody species. *Ecological Modeling* 152: 5–32.
Phelps DG and Bosch OJH (2002) A quantitative state and transition model for the Mitchell grasslands of central western Queensland. *Rangeland Journal* 24: 242–267.
Prince SD (1991) Satellite remote sensing of primary production: Comparison of results for Sahelian grasslands 1981–1988. *International Journal of Remote Sensing* 12: 1301–1311.
Rietkerk M, Dekker SC, Ruiter PC, and van de Koppel J (2004) Self-organized patchiness and catastrophic shifts in ecosystems. *Science* 305(5692): 1926–1929.
Schwinning S and Parsons AJ (1996) A spatially explicit population model of stoloniferous N-fixing legumes in mixed pasture with grass. *Journal of Ecology* 84: 815–826.
Tan SS and Smeins FE (1996) Predicting grassland community changes with an artificial neural network model. *Ecological Modeling* 84: 91–97.
Thornley JHM, Bergelson J, and Parsons AJ (1995) Complex dynamics in a carbon–nitrogen model of a grass legume pasture. *Annals of Botany* 75: 79–94.
Van Dyne GM (1972) Organization and management of an integrated ecological research program. In: Jeffers JNR (ed.) *Mathematical Models in Ecology*, pp. 111–172. Oxford: Blackwell Scientific.
Westoby M, Walker B, and Noy-Meir I (1989) Opportunistic management for rangelands not at equilibrium. *Journal of Range Management* 42: 266–274.
Winkler E and Fischer M (2002) The role of vegetative spread and seed dispersal for optimal life histories of clonal plants: A simulation study. *Evolutionary Ecology* 15: 281–301.
Wu J and Levin SA (1994) A spatial patch dynamic modeling approach to pattern and process in an annual grassland. *Ecological Monographs* 64(4): 447–464.

Further Reading

Clements FE (1936) Nature and structure of the climax. *Journal of Ecology* 24: 252–284.
Hanson JD, Parton WJ, and Innis GS (1985) Plant growth and production of grassland ecosystems: A comparison of modeling approaches. *Ecological Modeling* 29: 131–144.
Innis GS (ed.) (2012), Vol. 26. *Grassland simulation model*. New York: Springer Science & Business Media.
Coupland RT (ed.) (1992) Natural grasslands. *Ecosystems of the World*. In: vol. 8A. Amsterdam: Elsevier.
Nouvellon Y, Rambal S, Seen DL, et al. (2000) Modeling of daily fluxes of water and carbon from shortgrass steppes. *Agricultural and Forest Meteorology* 100: 137–153.
Pickett STA and White PS (1985) *The ecology of natural disturbance and patch dynamics*. New York: Academic Press.
Rietkert M, Boerlijst M, van Langevelde F, et al. (2002) Self-organization of vegetation in arid ecosystems. *American Naturalist* 160: 524–530.
Snaydon RW (1987) *Managed grasslands ecosystems of the world*. Vol. 17B. Amsterdam: Elsevier.
Stöcklin J and Winkler E (2004) Optimum reproduction and dispersal strategies of a clonal plant in a metapopulation: A simulation study with *Hieracium pilosella*. *Evolutionary Ecology* 18: 563–584.
Wan S, Norby RJ, Ledford J, and Weltzin JF (2007) Responses of soil respiration to elevated CO2, air warming, and changing soil water availability in a model old-field grassland. *Global Change Biology* 13(11): 2411–2424.
Watt AS (1947) Pattern and process in the plant community. *Journal of Ecology* 35: 1–22.

Lake Models☆

Peter Reichert and Johanna Mieleitner, Eawag: Swiss Federal Institute of Aquatic Science and Technology, Duebendorf, Switzerland

Introduction

Since the 1970s models have been used to predict the water quality of lakes and reservoirs to support their management. Lake models are also used for testing hypotheses in research. Many different approaches and levels of complexity have been and are used to achieve the different goals.

An overview is given of the basic principles of biogeochemical and ecological lake modeling. We start with a brief discussion of the objectives of ecosystems modeling with a particular emphasis on lake modeling. Then we give an overview of important processes in lakes and discuss major difficulties of describing them in lake models. This is followed by an overview of important components of lake models and how different formulations of these components attempt to overcome the problems mentioned in the preceding section. In the next section, we show mathematical process formulations typically used in the biogeochemical and ecological parts of lake models. Finally, by briefly describing a selection of lake models and their application, we provide an overview of different models currently in use and different modeling strategies applied by different research groups. The article focuses on mechanistic lake modeling because they more comprehensively address the objectives discussed in the next section in comparison to empirical and statistical approaches to lake modeling. Empirical and statistical approaches usually provide less insight into mechanisms in the ecosystem. Nevertheless, also these approaches can be useful for analysis and extrapolation of observed behavior.

Objectives of Lake Modeling

There are three main objectives for constructing and using lake models:

1. *Improving the understanding of lake ecosystem function*: Comparing results of simulations of a lake model with measured data provides a test of the hypotheses formulated in the model. Thus, lake models are ideal tools for quantitative testing of hypotheses about lake ecosystem theories. Furthermore, they provide a link from concentrations to fluxes and transformation rates that are much more difficult to measure than concentrations. The formulation of comprehensive lake models can also lead to the identification of knowledge gaps. Finally, performing model simulations and tests stimulates creative thinking about important mechanisms in lake ecosystems.
2. *Summarizing and communicating knowledge about lake ecosystems*: Lake models are perfect communication tools for exchanging quantitatively formulated knowledge about processes in lakes.
3. *Supporting lake ecosystem management*: Lake models can support lake management by predicting the consequences of suggested (alternative) measures. As both our knowledge and its representation in the models are incomplete, a considerable effort must be on quantifying prediction uncertainty if the models are applied for management purposes.

These objectives are essentially the same as in other fields of environmental modeling. However, lake models had a pioneering role in providing insight into the function of natural ecosystems and in model application for environmental management. The two most important reasons for this pioneering role are (1) the severe eutrophication problems many lakes with excessive nutrient input faced in the 1950s and 1960s, and (2) that already simple one- or two-box phosphorus mass-balance models were able to provide essential insights into these problems.

As in other fields of environmental modeling, the lake model to be used depends on the objective of the study. Typically, models for improving the understanding and communicating knowledge must have a higher structural resolution of model components and processes than models for lake management. For management purposes, getting the important mass fluxes correct is usually more important than providing a detailed insight into the substructures at the trophic levels of the food web. However, knowledge gained from more detailed research models often stimulates the development of simpler management models. Also the method of parameter estimation can depend on the objective of the study. For research purposes parameters are often estimated using frequentist techniques to avoid bias due to subjective prejudices. When the model is used for management purposes, there is usually not enough data available to perform a frequentist parameter estimation. In this case prior knowledge is best combined with empirical data using Bayesian techniques.

☆*Change History:* March 2018. Todd M. Swannack made minor changes to the references.

This is an update of P. Reichert and J. Mieleitner, Lake Models, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 2068–2080.

Processes to Be Represented in Lake Ecosystem Models

The basis of all biogeochemical and ecological lake models are mass balances of nutrients, dissolved oxygen, and organic compounds—particularly aquatic organisms—in the lake. Fig. 1 shows a graphical representation of biological processes in a lake ecosystem that build, at a certain level of simplification or refinement, the basis of most lake ecosystem transformation process models.

By primary production, nutrients are converted into phytoplankton biomass (in shallow lakes, periphyton may also considerably contribute to primary production). This process requires light and produces dissolved oxygen. Herbivorous zooplankton grazes on phytoplankton. Carnivorous zooplankton feeds on herbivorous zooplankton. Omnivorous zooplankton feeds on herbivorous zooplankton and on phytoplankton. All or some of these plankton classes serve as food for planktivorous fish which again are the food source for carnivorous fish. All these grazing and predation activities require dissolved oxygen and lead to release of particulate organic material (fecal pellets and remainings from sloppy feeding), dissolved organic matter (released from broken cells), and nutrients. Death of all organisms transforms them into particulate organic matter. Furthermore, respiration of organisms transforms biomass into nutrients. Particulate organic matter is hydrolyzed to dissolved organic substances which are mineralized into nutrients. These last two processes are of particular importance in the sediment of the lake. In the presence of dissolved oxygen, mineralization is accompanied by dissolved oxygen consumption. In deeper sediment layers, where all dissolved oxygen diffusing into the sediment from the water column is used up, mineralization requires reducing nitrate, manganese oxide, iron hydroxide, or sulfate. Finally, mineralization is also possible by methanogenesis.

Transport processes lead to partial spatial separation of these transformation processes. Fig. 2 gives an overview of the most important transport processes in a lake or reservoir. Depending on the density of the inflow and on stratification of the lake, the inflow enters the lake at a certain depth (with some entrainment of water from the layers above). As the outflow is not at the same level (at the surface for natural lakes and close to the bottom for many reservoirs), this leads to vertical advection of (part of) the water column. In addition, the water column is mixed by turbulent diffusion. During periods of stratification (usually caused by warmer and less dense water layers laying above colder and denser layers), horizontal mixing is usually much faster than vertical

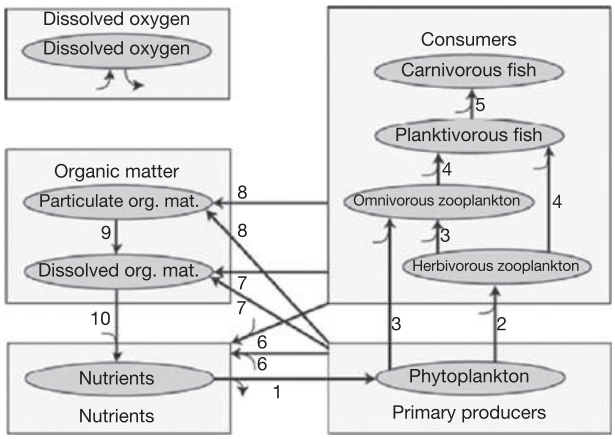


Fig. 1 Overview of important biological processes in the pelagic zone of surfacewaters. Gray ovals represent state variables (chemical compounds or organisms), and arrows represent transformation processes. The following processes are considered: 1, growth of phytoplankton (primary production); 2, growth of herbivorous zooplankton; 3, growth of omnivorous or carnivorous zooplankton; 4, growth of planktivorous fish; 5, growth of carnivorous fish; 6, respiration; 7, release of dissolved organic matter during death, sloppy feeding, and exudation; 8, death; 9, hydrolysis; 10, mineralization. Small arrows indicate oxygen consumption or production.

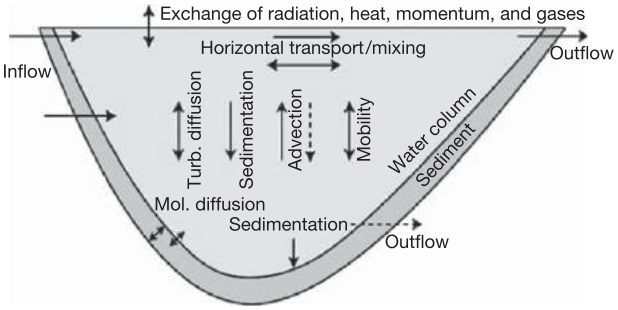


Fig. 2 Important transport and exchange processes in a lake or reservoir.

mixing. Radiation, heat, momentum, dissolved oxygen, carbon dioxide, and molecular nitrogen are exchanged over the lake surface. Due to their mobility, fish, zooplankton, and some phytoplankton species move actively through the water column. Particulate substances are deposited at the surface of the sediment due to sedimentation. Dissolved substances are transported within the sediment porewater and between porewater and lake water by molecular diffusion.

The interaction of transformation and transport processes discussed separately in the previous paragraphs (**Figs. 1** and **2**) often leads to the following typical spatial separation of processes in a lake: primary production of phytoplankton takes place in the upper layer of the lake, where sufficient light is available. This process consumes nutrients. Nutrients are delivered by the inflow and produced by respiration or mineralization either in the epilimnion or in the depth of the lake from where they diffuse to the surface layer. Zooplankton can actively move through the water column. Herbivorous zooplankton feeds on phytoplankton in the surface layer of the lake. Fish dominantly feed on plankton in the surface layer in the pelagial and in the littoral as the light allows them to catch their food. Particulate organic material produced by the organisms is usually sedimenting through the water column much quicker than mineralization takes place. For this reason a large fraction of particulate organic material reaches the sediment where mineralization processes consume dissolved oxygen, nitrate, and other compounds that can be used for the oxidation of organic substances. Due to the small diffusivities in the sediment and often also in the hypolimnion, this leads to large gradients of dissolved oxygen, nitrate, and mineralization products in the sediment and the hypolimnion of the lake. In shallow lakes, benthic organisms and Macrophytes can contribute to substance turnover in a similar way as described here for the pelagial.

A lake ecosystem model should represent the main physical, chemical, and biological processes of the most important substances in the lake and represent the biological communities building the ecosystem.

Difficulties of Lake Ecosystem Modeling

The large number and complicated nature of the processes in a lake ecosystem described in the previous section makes lake ecosystem modeling a very demanding task. However, there are more reasons that even increase the difficulty of building up a lake ecosystem model:

- A lake is a heterogeneous environment consisting of quite different, interconnected habitats (e.g., environmental conditions in the epilimnion, hypolimnion, littoral, and the sediment are significantly different).
- Within each of the trophic levels of the lake ecosystem shown in **Fig. 1**, there are a large number and a high diversity of species that are difficult to describe collectively as well as individually (e.g., there is a huge number of different phytoplankton species which all have the same essential function in **Fig. 1** but differ considerably in their properties).
- The occurrence of many individual species in a lake is highly irregular (some species occur at a high density in particular years and form only a minor contribution to the biomass in other years even under similar driving conditions).
- The properties of species are difficult to extract from lake data and field and laboratory experiments (easily measurable properties such as size and volume are not strongly related to properties relevant for growth and occurrence).
- The species are adaptable so that they can change their properties to some degree in response to the environmental conditions they encountered in the past (e.g., different size, different light dependence, or different elemental composition). This can make even measured properties unreliable when applied to a situation with different environmental condition history.
- There is a strong interaction between mixing on one side and oxygen, nutrient, and biological population dynamics on the other side. These processes are difficult to describe and hard to separate using lake data.
- The interactions between sediment and water column are very important for the lake ecosystem, especially the nutrient release from the sediment. The processes governing these interactions are difficult to quantify.
- Populations of higher organisms have an age distribution. Different age groups of the same species can play entirely different roles in the food web.

Most of these difficulties listed above occur similarly when modeling other ecosystems, particularly aquatic ecosystems. But the generic difficulties related to heterogeneity, adaptability, species richness, and interactions at various levels are even similar when modeling terrestrial ecosystems.

Different lake ecosystem models differ in the degree of simplification or refinement of the food web and transport processes, in the mathematical formulation used for quantifying process rates, and in the way they cope with the difficulties listed above. In the following section, we give an overview of typical approaches grouped according to different components of lake ecosystem models.

Components of Lake Ecosystem Models

In lake ecosystem models different strategies have been applied to deal with the problems described above. The major distinction is the degree of simplification of the physical, biogeochemical, and ecological components of the lake ecosystem and the mathematical formulation of the processes considered in the model. In this section, we give an overview of approaches for physical, biogeochemical, and ecological submodels of lake ecosystem models.

Physical Submodels

Key elements distinguishing physical submodels of lake ecosystem models are the spatial resolution of the model and the description of mixing processes.

Spatial resolution of lake ecosystem models

Already in the early 1970s, it was recognized that simple one-box mass-balance models (mainly for phosphorus contained in phosphate and phytoplankton) are extremely useful for improving the basic understanding of eutrophication processes. Such models represent the whole lake as a mixed reactor and calculate changes in average concentration(s) as a consequence of input, transformation, sedimentation, sediment release, and output. In conjunction with more complex biogeochemical and ecological submodels (see below), such models are still in use today, particularly for shallow lakes with a small degree of stratification.

Thermal stratification suppresses vertical mixing significantly during the summer, when warm, less-dense water layers lay on the top of colder and denser layers. Due to diurnal temperature variations and wind-induced mixing, a frequently mixed surface layer, the epilimnion, builds up on top of the hypolimnion, which is separated from the epilimnion by a zone with a strong temperature gradient and high stability, the metalimnion. Obviously, a simple description of such a system can be obtained with two mixed reactors representing epilimnion and hypolimnion and with an exchange process that is strong during the winter and weak during the summer.

Closing the nutrient cycle in the lake-sediment system requires an extension of the mass balances to the sediment. This can be done by adding an additional mixed reactor describing the sediment, either to a one- or two-box lake model. This makes it possible to describe mineralization of organic particles deposited in the sediment and resulting oxygen consumption and nutrient release. More boxes can be used to describe anoxic and anaerobic mineralization processes in deeper sediment layers.

A better description of mixing in the lake can be achieved by resolving the depth of the lake continuously. Such one-dimensional (1D) lake models are able to resolve the often very strong gradients in vertical concentration profiles of dissolved oxygen, nutrients, and phytoplankton.

Particularly for narrow lakes with a large longitudinal extension and a high throughflow, it may be necessary to resolve the longitudinal dimension as well. This can be the case for reservoirs. Such 2D models are then able to distinguish different vertical profiles along the longitudinal direction of the lake.

The increasing availability of 3D mechanistic mixing models makes it more and more possible to couple such models with biogeochemical and ecological lake models. This leads to a 3D description of all processes in the lake. From the biogeochemical and ecological perspective, such models make it possible to distinguish processes and substance concentrations in the pelagic zone (open water column) from those in the littoral zone (close to the land and sediment) and to resolve concentration gradients across the lake. This can be relevant for the description of increased productivity in the neighborhood of nutrient-delivering inflows.

Description of mixing processes

Options for the description of mixing processes in lakes depend on the spatial resolution of the model described in the previous subsection. Use of box models usually requires an empirical parametrization of mixing processes between the boxes. Models with resolution of the vertical dimension of the lake often also rely on empirical parametrization of mixing processes. Such empirical parametrization of mixing processes is usually calibrated with the aid of temperature profiles in the lake. Such profiles lead to quite reliable estimates of mixing intensity between epilimnion and hypolimnion. However, due to very small temperature gradients, they often do not provide sufficient information for a calibration of mixing processes in the hypolimnion of deep lakes. This leads to the requirement of including profiles of phosphate, nitrate, and/or dissolved oxygen into the calibration process. This can be problematic as there may be nonidentifiability problems between mixing intensity and transformation processes.

Physically based mixing models can be derived for 1D, 2D, and 3D lake models. They describe stratification and mixing caused by heat exchange over the lake surface and by momentum uptake due to wind forcing. Often turbulence due to seiche oscillation is also relevant for lakes. The main advantage of replacing empirical and semiempirical mixing models by such mechanistic mixing and transport models is that this decreases the need for empirical parametrization of the physical part of the ecological lake model. This improves the predictive capability of the models, at least conditionally on assumptions regarding climate forcing.

Biogeochemical Submodels

Biogeochemical submodels differ in the consideration of nutrients, in modeling of the element cycles and exchange with the sediment, and in their description of the mineralization process.

Consideration of nutrients in lake models

The most important elements that can limit phytoplankton growth are phosphorus, nitrogen, carbon, and, for diatoms, silicon. These are usually taken up in the form of phosphate, ammonium or nitrate, carbon dioxide (sometimes also bicarbonate), and silica.

Many lakes are limited by phosphorus during the summer stratification period. This leads to extremely small phosphate concentrations in the epilimnion during summer and makes the consideration of phosphate very important for lake models.

Nitrogen limitation is more difficult to describe than phosphorus limitation, because there are some phytoplankton species that are able to fix nitrogen from dissolved molecular nitrogen. Nitrification of ammonium to nitrate also affects the oxygen budget of the lake. Nitrate is not only important as a nutrient for phytoplankton, it is also important for anoxic mineralization of organic material (primarily in the sediment). This denitrifying mineralization can make the lake a significant sink for nitrate. Quantifying this denitrification capacity of the lake requires consideration of nitrate (and usually also ammonium) in the lake model.

The limiting effect of silica on phytoplankton growth depends on the geology of the watershed (determining silica input) and on the occurrence of diatoms (determining silica consumption). The consideration of silica can be important, if diatoms are distinguished from other functional groups of phytoplankton.

Sometimes also carbon is limiting primary production in lakes. Most phytoplankton species need CO_2 as carbon source. The dependence of the growth rate on the CO_2 concentration varies among species, therefore the depletion of CO_2 can lead to a change in species composition. For example, the dominance of cyanobacteria in hypereutrophic lakes is sometimes caused by CO_2 limitation because cyanobacteria have a very efficient carbon concentration mechanism (CCM). Some species can also use HCO_3^- as carbon source.

Modeling of element cycles and exchange with the sediment

Simple models (with respect to element cycles) treat sedimentation of organic particles as loss from the modeled part of the system and use sediment oxygen demand, phosphate release, and ammonium release as model parameters. With a correct choice of these model parameters, this can lead to reasonable results. However, this decoupling of processes which is very strongly coupled in reality (through mineralization in the sediment) allows an inexperienced model user to work with unreasonable model parameters that violate mass conservation.

For this reason, a more detailed level of description is to model the mass balance of nutrients in the sediment explicitly. This can be done by describing the pools of particulate organic matter, dissolved oxygen, and nutrients in the sediment and calculating sediment oxygen demand and nutrient release by a simple kinetic process.

More detailed sediment models use one or more sediment layers (with different redox conditions) to achieve a more realistic description of the sediment. Particulate organic matter enters the top sediment layer through sedimentation. The mineralization of organic matter to inorganic nutrients in the sediment can be modeled using the same process description as in the water column. The inorganic nutrients produced by mineralization are released into the porewater of the sediment and diffuse into the water column, depending on the concentration difference between the sediment porewater and the water column.

Description of the mineralization process

Among the models which model mineralization explicitly, most do not distinguish hydrolysis of particulate organic matter into dissolved organic matter from mineralization of dissolved organic matter but combine both steps into a single "mineralization" process.

Many models only describe oxic mineralization. Some add denitrifying mineralization, some add further steps with reduction of manganese oxide, iron hydroxide, or sulfate and finally methanogenesis.

Most models parametrize the mineralization process directly without explicitly describing the bacterial community performing the process. This limits the transferability of these models as different bacterial populations in different environments lead to different mineralization rates.

Ecological Submodels

Ecological submodels can be divided into trophic-level models, functional group models, dominant-species models, and adaptive property models. Combinations of some of these model types are also possible.

Trophic-level models

Trophic-level models use the trophic levels of the food web shown in Fig. 1 as state variables without further division. However, some of the trophic levels may be merged and some may be omitted. If higher trophic levels are omitted, their effect on lower levels is considered by increased death rates at the lower levels.

Phytoplankton or periphyton must be considered in each ecological lake model as it is responsible for primary production of biomass out of inorganic nutrients. Phytoplankton consists of hundreds of different species with widely varying properties such as maximum growth rate, edibility, and dependence on light, nutrients, and temperature. In trophic-level models, all these different species are modeled by a single state variable. It seems astonishing that this can work. However, the limitation of primary production by nutrients in many lakes makes production less dependent on formulation and quantification of process kinetics. In such situations, input of nutrients determines production. This may be the explanation why such simple models work astonishingly well.

If zooplankton is considered explicitly it is often modeled as a single state variable or as two state variables representing herbivorous and carnivorous or omnivorous zooplankton. Again, these classes aggregate many different species.

In most ecological lake models, fish are not explicitly modeled. The predation pressure of fish is then quantified by increasing the death rate of zooplankton. To account for changes in predation pressure, a seasonal dependence of such a death rate contribution can be considered.

Functional group models

Functional group models differ from trophic-level models by disaggregating the species within one or several trophic levels into groups with similar properties. Although these groups usually have the same function (given by the trophic level), they are called functional groups.

Criteria for the division of species into functional groups can be (1) ecological properties, such as growth rate, edibility (e.g., size), silica requirement, ability to fix nitrogen, or mobility; (2) taxonomic groups; (3) easily measurable properties, such as maximum extension, volume, etc., which are assumed to correlate with ecological properties. Often different functional groups have similar process formulations, but differ in model parameter values.

Dominant-species models

Dominant-species models make a slightly different approach to model the variability of properties of different species at a given trophic level. Instead of dividing the species at the trophic level exhaustively into functional groups, the key species are modeled individually. If these species can be cultivated in the laboratory, such models have the advantage that essential behavioral parameters such as maximum growth rate or parameters related to nutrient limitation can be measured experimentally. This can decrease the number of calibration parameters of the model significantly. On the other hand, these models have the disadvantage of introducing a large number of additional state variables and still not being able to describe the sum of all species correctly.

A slightly modified version of this type of model uses the properties of a key indicator species to represent functional groups. This leads to a model that combines advantages of the dominant-species model with those of the functional group approach. However, the advantage of using measured properties in the dominant-species approach becomes less strong when applying these to functional groups.

Adaptive property models

Biological species are adaptive. They are able to change their properties, for example, their size, edibility, chlorophyll, or phosphorus content in order to adapt to changing environmental conditions. This is only rarely accounted for in the ecosystem models discussed above (adaptation of phosphorus content is considered in some models). Recently, there have been attempts to include adaptation into aquatic ecosystem models.

One approach is called "structural dynamic models." In this approach, selected properties of species are dynamically changed according to the global criterion of maximization of "exergy." Kinetic parameter values are adapted during the simulation according to exergy maximization.

A second approach identifies "rapid evolution" as the cause for adaptation and (usually) uses an empirical parametrization of this process.

Typical Formulations of Transformation Processes

The temporal change of the concentration, C_X, of a substance or organism, X, in a vertically resolved water column is given by the following differential equation:

$$\frac{\partial C_X}{\partial t} = v_X \frac{\partial C_X}{\partial z} + \frac{\partial}{\partial z}\left(K_z \frac{\partial C_X}{\partial z}\right) + r_X \tag{1}$$

where t is time, z is the vertical coordinate in the lake, v_X is the sum of the advective velocity of the water column and the sedimentation velocity of substance or organism X, K_z is the coefficient of vertical turbulent diffusion, and r_X is the total (net) transformation rate of substance or organism X. The total transformation rate of a substance is the sum of contributions by different processes. The contribution of each process is calculated as the product of the process rate with a substance-specific stoichiometric coefficient. This means that the net transformation rate of substance X is given by

$$r_X = \sum_i v_{i,X} \rho_i \tag{2}$$

where the sum extends over all transformation processes. $v_{i,X}$ is the stoichiometric coefficient of the process i with respect to substance X and ρ_i is the process rate of the process i.

In order to discuss formulations of transformation processes used in the literature, we use a simple, didactic lake model. This lake model contains five state variables: dissolved oxygen, nutrients, phytoplankton, zooplankton, and dead particulate organic material. Compared to Fig. 1, this model aggregates the two functional groups of zooplankton and it omits fish and dissolved organic material. This leads to an aggregation of transformation processes also. We will give formulations of the following processes:

1. Growth of phytoplankton by primary production (gro, ALG)
2. Growth of zooplankton by grazing of phytoplankton (gro, ZOO)

Table 1 Structure of typical formulations of process rates

Process	*Structure of rate formulation*
Growth of phytoplankton	$\rho_{\text{gro, ALG}} = k_{\text{gro, ALG, max, To}} \cdot f_T(T) \cdot f_I(I) \cdot f_N(C_N) \cdot C_{\text{ALG}}$
Growth of zooplankton	$\rho_{\text{gro, ZOO}} = k_{\text{gro, ZOO, max, To}} \cdot f_T(T) \cdot f_{O2}(C_{O2}) \cdot f_g(C_{\text{ALG}}) \cdot C_{\text{ZOO}}$
Respiration of phytoplankton or zooplankton	$\rho_{\text{resp, }i} = k_{\text{resp, }i\text{, To}} \cdot f_T(T) \cdot f_{O2}(C_{O2}) \cdot C_i,\ i = \text{ALG, ZOO}$
Death of phytoplankton or zooplankton	$\rho_{\text{death, }i} = k_{\text{death, }i\text{, To}} \cdot f_T(T) \cdot C_i,\ i = \text{ALG, ZOO}$
Oxic mineralization	$\rho_{\text{miner}} = k_{\text{miner, To}} \cdot f_T(T) \cdot f_{O2}(C_{O2}) \cdot C_{\text{POM}}$

T, temperature; I, light intensity; C_N, nutrient concentration; C_{O2}, dissolved oxygen concentration; C_{ALG}, concentration of phytoplankton; C_{ZOO}, concentration of zooplankton; C_{POM}, concentration of dead particulate organic material; k, specific transformation rates at reference conditions.

3. Respiration of phytoplankton (resp, ALG)
4. Respiration of zooplankton (resp, ZOO)
5. Death of phytoplankton including grazing by zooplankton (death, ALG)
6. Death of zooplankton including predation by fish (death, ZOO)
7. Oxic mineralization of particulate organic material to nutrients including the hydrolysis step to dissolved organic material (miner)

We will now discuss typical formulations of the transformation rates of these processes used in the literature. Transformation rates of substances and organisms are then given by Eq. (2).

Table 1 gives an overview of the structure of typical formulations of these seven process rates. Growth, respiration, and death rates are usually proportional to the concentration of the organism affected by the process. The rate formulation then multiplies this concentration by a specific transformation rate at standard conditions and several modification factors that describe the effect of important influence factors.

Table 2 shows options for the formulation of modification factors used in Table 1 for describing the dependence of process rates on important influence factors.

If several nutrients are limiting, several nutrient limitation terms can be multiplied or the minimum of the limiting factors can be used (Liebig's law).

This short overview should give an idea of how transformation process rates can be formulated. Some models use different or more complicated process formulations or they further divide processes into subprocesses. For example, the growth process of phytoplankton can more realistically be described by a nutrient uptake process into the cell and a growth process on nutrients contained in the cell.

Examples of Ecological Lake Models and Their Application

In this section we give a brief overview of model structures, calibration strategies, and applications of five selected ecological lake models. This overview is far from being complete. Nevertheless, it provides insight into the variety of approaches used in science and management. We will briefly present an aggregated trophic-level lake model (biogeochemical—ecological lake model, BELAMO), two functional group lake models (simulation by means of analytical lake model, SALMO and computational aquatic ecosystem dynamics model, CAEDYM), a dominant-species algal community model (phytoplankton response to environmental change model, PROTECH), and an adaptive property model (structural dynamics model).

BELAMO

The BELAMO represents a relatively simple, aggregated trophic-level lake model with emphasis on biogeochemical cycles rather than ecology. A particular feature is the consideration of closure of element cycles by explicit consideration of mineralization processes in the sediment.

Model overview

BELAMO describes the concentrations of phytoplankton, zooplankton, dissolved oxygen, ammonium, nitrate, phosphate, and degradable and inert dead organic particles in the water column and in the sediment. The model considers growth, respiration, and death of phytoplankton and zooplankton, mineralization, nitrification, and phosphate uptake on sinking particles. The model is 1D and resolves the depth of the lake. The physical processes vertical mixing, advection, sedimentation, mobility of zooplankton, and molecular diffusion in the sediment and across the water sediment interface are considered. Phytoplankton can grow with a variable stoichiometry with respect to phosphorous depending on the phosphate concentration in the water column to describe the low phosphorus content of phytoplankton growing during phosphate-limited periods in summer.

Calibration strategy

BELAMO applications estimate kinetic parameters of transformation processes with the attempt of finding "universal" values across lakes of different trophic state. As all phytoplankton species are aggregated to a single state variable, it is hard to use kinetic

Table 2 Typical formulations of the functions used for describing the dependence of process rates on important influence factors

Term	*Formulations*	*Name/comment*
$f_N(C_N)$	$\frac{C_N}{K_N+C_N}$	Monod
	$\frac{{C_N}^n}{{K_N}^n+{C_N}^n}$	Hill
	$1 - e^{-k_N C_N}$	Exponential
	$\begin{cases} 1 & C_N > 2K_N \\ \frac{C_N}{2K_N} & C_N \le 2K_N \end{cases}$	Blackman
$f_I(I)$	$\frac{I}{K_I+I}$	Monod
	$\left(\frac{I}{I_{opt}}\right)\exp\left(1-\frac{I}{I_{opt}}\right)$	Steele
	$\left(\frac{I}{I_{opt}}\right)^n \exp\left(1-\left(\frac{I}{I_{opt}}\right)^n\right)$	Walker
	$\frac{I}{\sqrt{K_I^2+I^2}}$	Smith
$f_{O2}(C_{O2})$	$\frac{C_{O2}}{K_{O2}+C_{O2}}$	Monod
$f_T(T)$	$\frac{T-T_{min}}{T_{ref}-T_{min}}$	Linear
	$\Theta^{T-T_{ref}}$ with $\Theta = \exp\left(\frac{E}{RT_{ref}T}\right)$	Arrhenius
$f_g(C_{ALG})$	$\frac{C_{ALG}}{K_{ALG}+C_{ALG}}$	Monod
	$1 - \exp(-k_{ALG} \cdot C_{ALG})$	Exponential
	$\begin{cases} \frac{C_{ALG}-C_0}{K_{ALG}+(C_{ALG}-C_0)} & \text{for } C_{ALG} > C_0 \text{ with threshold} \\ 0 & \text{else} \end{cases}$	

parameters measured for selected cultured species in this model. To avoid nonidentifiability problems during the parameter estimation, sensitivity, and identifiability analysis techniques are used.

Model applications

BELAMO so far has been applied to three lakes of different trophic state. It is a research model to summarize knowledge and test hypothesis with a focus on biogeochemical cycles.

Fig. 3 shows measured and calculated profiles of phytoplankton, dissolved oxygen, phosphate, and nitrate in Greifensee.

The profiles clearly demonstrate the yearly cycle of mixing and stratification and its effect on the nutrient cycle. Phytoplankton shows a first maximum in spring, followed by a second in summer. Dissolved oxygen concentrations are nearly uniform in spring; during summer stratification, there is a strong oxygen depletion in the hypolimnion. Phosphate is not limiting phytoplankton growth in spring, but is during the summer and fall months. There is a significant increase in phosphate concentrations in the hypolimnion during the summer and fall months due to mineralization of organic particles in the sediment. Nitrate shows a severe depletion in the deep hypolimnion during the stratification period primarily due to anoxic mineralization in the sediment.

SALMO

SALMO represents a functional group lake model. The emphasis is on a very detailed description of the plankton growth dynamics. Recently SALMO was extended to SALMO-HR (high resolution). In this version the ecological model is coupled to a hydrothermodynamic model of the water column.

Model overview

SALMO describes orthophosphate, dissolved inorganic nitrogen, dissolved oxygen, organic particles, three functional groups of phytoplankton, and zooplankton concentrations in a lake. The processes growth and mortality of phytoplankton and zooplankton and mineralization are considered. Sedimentation of phytoplankton and migration of zooplankton is also modeled.

SALMO was designed to mechanistically describe physical, chemical, and biological processes according to a maximum of generality. The model uses only a small number of state variables but more complex process formulations than other models in order to achieve this goal. Each functional group of phytoplankton is characterized by an indicator species, the properties of which

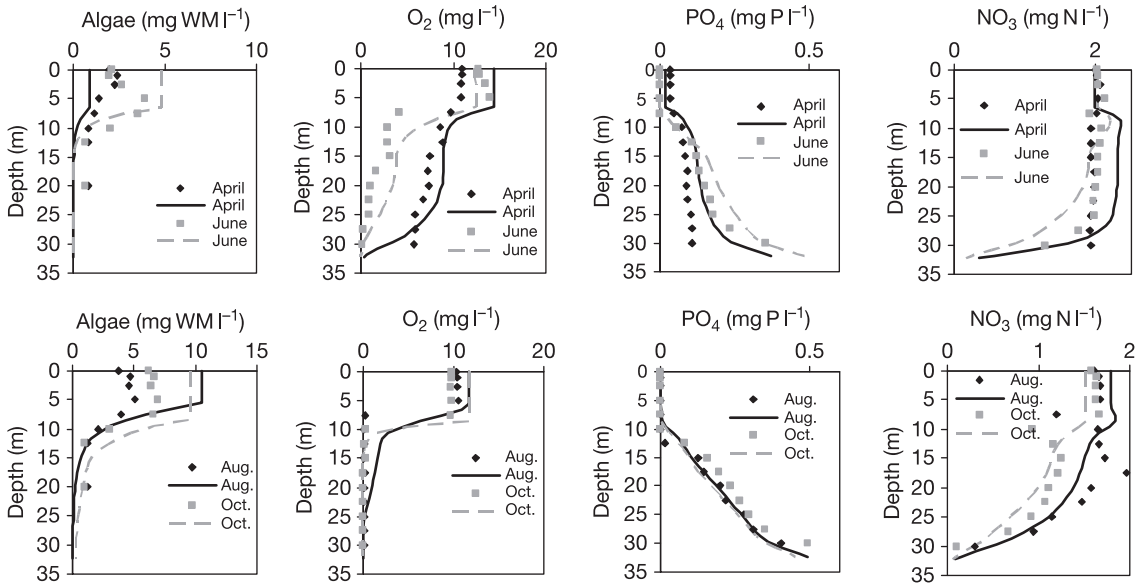


Fig. 3 Lake data (markers) compared with simulation results (lines) of the model BELAMO. Profiles of phytoplankton, dissolved oxygen, phosphate, and nitrate in Greifensee are shown for the year 1989. From Mieleitner, J. and Reichert, P. (2006). Analysis of the transferability of a biogeochemical lake model to lakes of different trophic state. *Ecological Modelling* **194**, 49–61.

were measured or compiled from the literature. Fish are considered implicitly by their predation rate on zooplankton. The nutrient release from the sediment is modeled as an empirical function of oxygen depletion and denitrification.

SALMO describes the water body as two mixed reactors representing the epilimnion and the hypolimnion. The depth of the epilimnion has to be specified as a boundary condition. SALMO-HR uses a very detailed hydrothermodynamic model of the water column.

Calibration strategy

In contrast to most other ecological lake models, the parameters of SALMO are not fitted. Measured values are used for all parameters. The parameter values for phytoplankton growth are determined in the laboratory for key species of each functional group.

This strategy not to calibrate the model has the advantage that the parameters are not adapted to a specific lake at a specific time and the parameters are universal for that reason. This improves the prediction quality and the generality of the model.

Model applications

SALMO is used as a tool to improve the understanding of the ecosystem and to support management decisions. It has been successfully applied to more than 20 lakes and reservoirs of different trophic states and has been used to calculate scenarios for different discharge regimes, climate change scenarios, changing nutrient input, and biomanipulations.

Fig. 4 shows an example of an application of SALMO to the Bautzen Reservoir, Germany. The agreement between calculated and measured concentrations achieved without modifying model parameters is remarkably high. This example shows the interactions between phytoplankton and zooplankton and the depletion of phosphate during the summer.

CAEDYM

The CAEDYM is an ecological model that can be linked to different hydrodynamic models. In our list of example models, CAEDYM represents a functional group lake model of very high degree of resolution of ecosystem variables and processes. This is a chance for a detailed representation of many processes and mass fluxes, but also a challenge with respect to the number of model parameters and to calibration.

Model overview

The ecosystem model implemented in CAEDYM is based on a detailed description of the ecosystem. The user can choose between different ecological configuration options and use a different model for each specific application. CAEDYM can be used for freshwater, estuaries, or coastal waters. The model gives the user a large flexibility in the choice of state variables, processes, and process formulations.

The state variables that can be used include concentrations of dissolved oxygen, ammonium, nitrate, phosphate, silica, dissolved inorganic carbon, quickly and slowly degradable dissolved and particulate organic matter, up to two groups of inorganic suspended

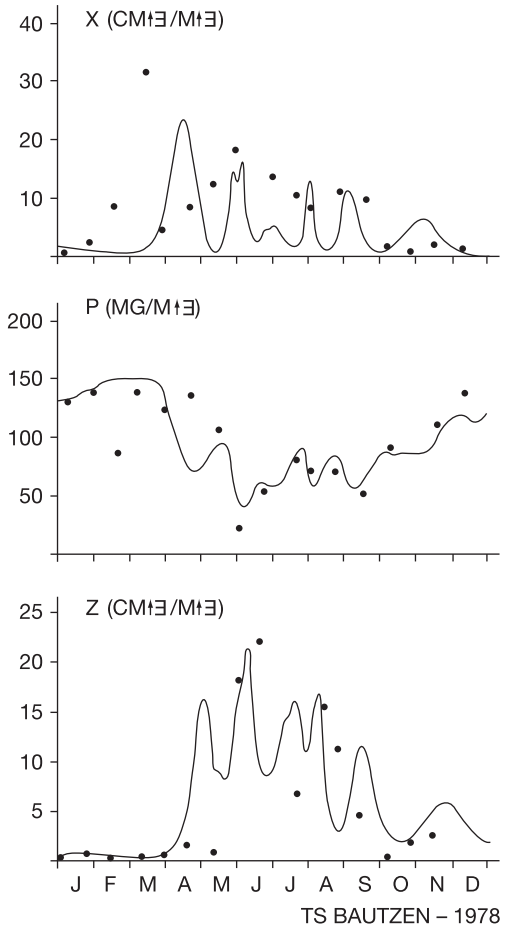


Fig. 4 Application of SALMO to the Bautzen Reservoir. Simulated and measured concentration time series of total phytoplankton biomass (top), dissolved phosphate (middle), and zooplankton (bottom) for the mixed layer of the hypereutrophic Bautzen Reservoir. From Benndorf, J. and Recknagel, F. (1982). Problems of application of the ecological model SALMO to lakes and reservoirs having various trophic states. *Ecological Modelling* **17**, 129–145.

solids, bacteria, up to seven groups of phytoplankton, up to five groups of zooplankton, up to five groups of fish, and pathogens in the water column, up to four groups of benthic macroalgae, seagrass, up to three groups of benthic invertebrates, and up to seven groups of benthic algae and others. The nonliving components in the water column are also modeled in the sediment. Process descriptions for primary production, secondary production, nutrient and metal cycling, and oxygen dynamics and exchange with the sediment are included in the model. CAEDYM can be coupled to 0D, 1D, 2D, and 3D lake hydrodynamics programs. It can be coupled to DYRESM (a 1D hydrodynamic model for lakes and reservoirs) or ELCOM (a 3D hydrodynamic model).

Calibration strategy

CAEDYM studies follow the reductionist approach with a detailed, general lake ecosystem model. Model parameters are fitted, but the attempt is made to find "universal" values that do not depend on the particular application. In typical applications, most parameters are held constant, some are fitted jointly for several systems, and some may need site-specific calibration.

Model applications

CAEDYM has been applied to many lakes and reservoirs. It is used by many research groups and can be downloaded as freeware.

CAEDYM has been used to evaluate different management strategies, to quantify nutrient cycles, and other processes.

Fig. 5 shows an application of the model to the Prospect Reservoir in Sydney, Australia. The simulations qualitatively and quantitatively reproduce the measurements with some problems in the concentrations of phytoplankton.

PROTECH

The PROTECH describes the phytoplankton growth in lakes at the species level. The emphasis of this model is on describing the dynamics of phytoplankton composition in a wide range of different ecosystems.

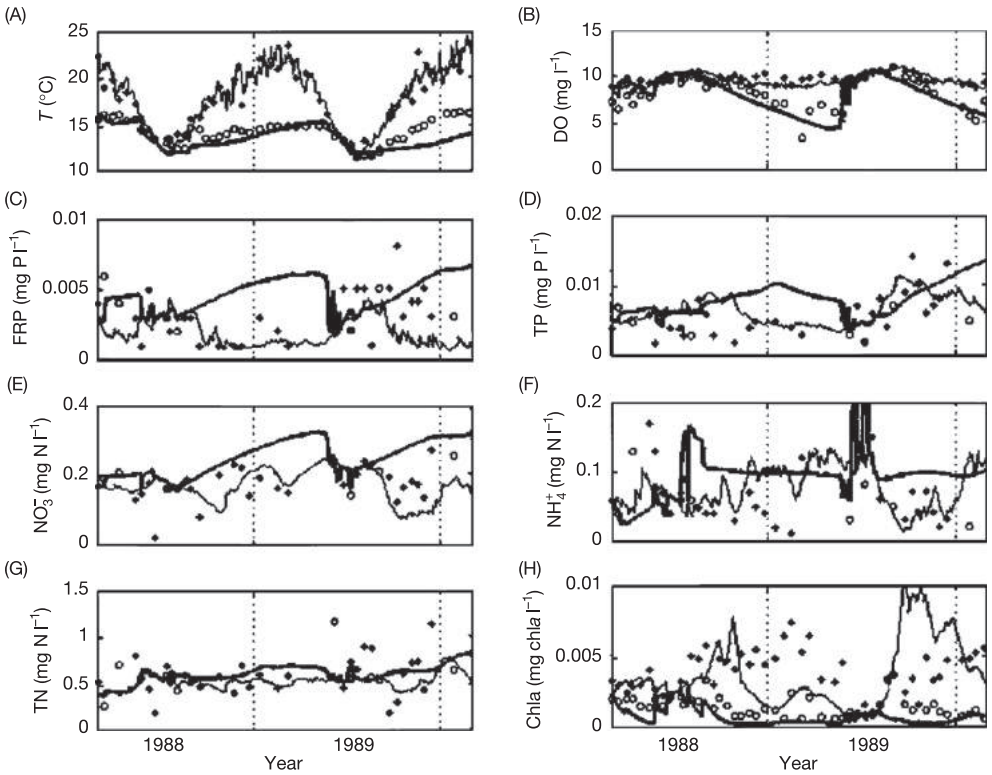


Fig. 5 Simulation results of the model CAEDYM. Comparison of measured time series with 1D simulations of the Prospect Reservoir at 2 m (black symbols, thin line) and 17 m depth (white symbols, thick line). The panels show temperature (a), dissolved oxygen (b), filterable reactive phosphorus (c), total phosphorus (d), nitrate (e), ammonium (f), total nitrogen (g), and chlorophyll *a* (h). From Romero, J. R., Antenucci, J. P., and Imberger, J. (2004). One- and three-dimensional biogeochemical simulations of two differing reservoirs. *Ecological Modelling* **174**, 143–160.

Model overview

PROTECH is designed to make simulations of the dynamic changes in the populations of different species of phytoplankton within a reservoir or lake environment which may be subject to thermal stratification, periodic destratification, and hydraulic exchange.

Chlorophyll *a*, phosphorous, nitrogen, and silica are modeled. The phytoplankton model is very detailed; up to eight species can be selected from a library of 18 phytoplankton species. The effect of zooplankton is described by the death rate of phytoplankton. The maximum growth rate of the different phytoplankton species is calculated using correlations with surface area and volume of the species. Adjustments for temperature dependence, light limitation, and nutrient limitation are made.

The physical model is 1D. It divides the water body into mixed layers.

Calibration strategy

The parameters for the growth of the phytoplankton species are not fitted in PROTECH. However, the choice of considered species is site specific.

Model applications

PROTECH has been applied to different lakes across the world. It has been used to explore ecological theory, to assess the reactions of phytoplankton to changes in temperature and nutrient concentrations, and to support management decisions. It was also coupled to a climate model and to predict the changes in phytoplankton communities due to climate change. Fig. 6 shows a comparison of measured functional groups of phytoplankton with PROTECH simulations. The changes of chlorophyll *a* concentrations of two functional groups are shown. In the spring there is a bloom of *R* species, followed by a period with low phytoplankton concentrations and a bloom of CS species during the summer. The correspondence of the model results with the data is remarkable. In general, the correspondence is better at the functional group level than at the species level.

Structural Dynamic Model

Structural dynamic modeling is an approach that represents an adaptive property model in our list of example models. See more details under Structural Dynamic Models.

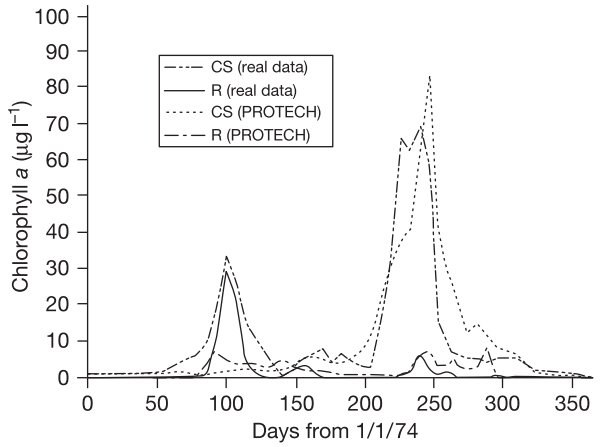


Fig. 6 Time series of data and PROTECH model results for two functional groups of phytoplankton. Phytoplankton is divided into *R*-strategists (ruderals) and a CS-group (intermediate group of competitive *C*-strategists, stress-tolerant *S*-strategists). From Elliott, J. A., Irish, A. E., Reynolds, C. S., and Tett, P. (2000). Modelling freshwater phytoplankton communities: An exercise in validation. *Ecological Modelling* **128**(1), 19–26.

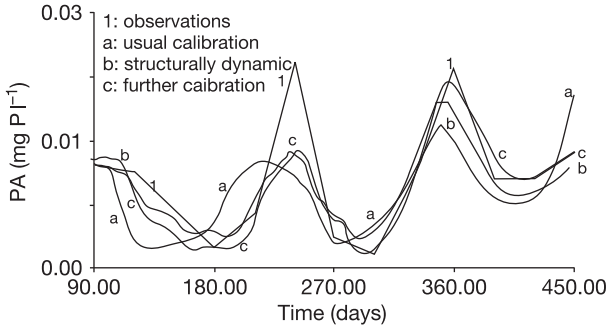


Fig. 7 Time series of phytoplankton. Comparison of data with simulations from Zhang, J., Jorgensen, S. E., Tan, C. O., and Beklioglu, M. (2003). A structurally dynamic modelling—Lake Mogan, Turkey as a case study. *Ecological Modelling* **164**(2), 103–120.

Model overview

Structural dynamics modeling is one of the approaches to consider adaptive processes in ecosystem models. The approach is based on the hypothesis that an ecosystem tends to move away from thermodynamic equilibrium. This is quantified by exergy, defined as the amount of work that a system can perform when it is brought into thermodynamic equilibrium with its environment.

Modeling approach

During the calibration period, some model parameters are adapted dynamically to maximize exergy while reducing the residuals. This leads to time-dependent parameters.

Example application

Fig. 7 shows a comparison of a conventional modeling approach, a structurally dynamic modeling approach, and a further improved simulation with data. It is evident, that the structurally dynamic model fits the data much better than the conventional calibration.

See also: Aquatic Ecology: Abundance Biomass Comparison Method. Conservation Ecology: Trophic Classification for Lakes. Ecological Complexity: Thermodynamics in Ecology. Ecosystems: Freshwater Lakes; Ecosystems

Further Reading

Arhonditsis, G.B., Brett, M.T., 2004. Evaluation of the current state of mechanistic aquatic biogeochemical modeling. Marine Ecology–Progress Series 271, 13–26.

Benndorf, J., Recknagel, F., 1982. Problems of application of the ecological model (SALMO) to lakes and reservoirs having various trophic states. Ecological Modelling 17, 129–145.

Bowie, G.L., Mills, W.B., Porcella, D.B., *et al*., 1985. Rates, constants, and kinetics formulations in surface water quality modeling, 2nd edn. Athens, GA: US EPA Environmental Research Laboratory, EPA/600/3-85/040.

Complex interactions in lake communities. In: Carpenter, S.R. (Ed.), 2012. Springer Science & Business Media. New York.
Chapra, S.C., 1996. Surface water quality modeling. McGraw-Hill: New York.
Droop, M.R., 1973. Some thoughts on nutrient limitation in algae. Journal of Phycology 9 (3), 264–272.
Elliott, J.A., Irish, A.E., Reynolds, C.S., Tett, P., 2000. Modelling freshwater phytoplankton communities: An exercise in validation. Ecological Modelling 128 (1), 19–26.
Fussmann, G.F., Ellner, S.P., Hairston Jr., N.G., *et al.*, 2005. Ecological and evolutionary dynamics of experimental plankton communities. Advances in Ecological Research 37, 221–243.
Håkanson, L., 2004. Break-through in predictive modelling opens new possibilities for aquatic ecology and management—A review. Hydrobiologia 518, 135–157.
Hamilton, D.P., Schladow, D.P., 1997. Prediciton of water quality in lakes and reservoirs. Part I—Model description. Ecological Modelling 96, 91–110.
Imberger, J., Patterson, J.C., Hebbert, B., Loh, I., 1978. Dynamics of reservoirs of medium size. Journal of the Hydraulics Division-ASCE 104 (5), 725–743.
Jørgensen, S.E., Fath, B., 2011. Fundamentals of ecological modelling, 4th edn. Elsevier: Amsterdam.
Jørgensen, S.E., 1999. State-of-the-art of ecological modelling with emphasis on development of structural dynamic models. Ecological Modelling 120, 75–96.
Karlsson, J., Byström, P., Ask, J., Ask, P., Persson, L., Jansson, M., 2009. Light limitation of nutrient-poor lake ecosystems. Nature 460 (7254), 506.
Mieleitner, J., Reichert, P., 2006. Analysis of the transferability of a biogeochemical lake model to lakes of different trophic state. Ecological Modelling 194, 49–61.
Mooij, W.M., Trolle, D., Jeppesen, E., Arhonditsis, G., Belolipetsky, P.V., Chitamwebwa, D.B., *et al.*, 2010. Challenges and opportunities for integrating lake ecosystem modelling approaches. Aquatic Ecology 44 (3), 633–667.
Omlin, M., Reichert, P., Forster, R., 2001. Biogeochemical model of lake Zürich: Model equations and results. Ecological Modelling 141 (103), 77.
Reynolds, C.S., Irish, A.E., Elliott, J.A., 2001. The ecological basis for simulating phytoplankton responses to environmental change (PROTECH). Ecological Modelling 140, 271–291.
Romero, J.R., Antenucci, J.P., Imberger, J., 2004. One- and three-dimensional biogeochemical simulations of two differing reservoirs. Ecological Modelling 174, 143–160.
Scheffer, M., van Nes, E.H., 2007. Shallow lakes theory revisited: Various alternative regimes driven by climate, nutrients, depth and lake size. Hydrobiologia 584 (1), 455–466.
Van Nes, E.H., Scheffer, M., 2005. A strategy to improve the contribution of complex simulation models to ecological theory. Ecological modelling 185 (2-4), 153–164.
Vollenweider, R.A., 1969. Possibilities and limits of elementary models concerning budget of substances in lakes (in German). Archiv für Hydrobiologie 66 (1), 1–36.
Zhang, J., Jorgensen, S.E., Tan, C.O., Beklioglu, M., 2003. A structurally dynamic modelling—Lake Mogan, Turkey as a case study. Ecological Modelling 164 (2-3), 103–120.

Mediated Modeling and Participatory Modeling

Damon M Hall, University of Missouri, Columbia, MO, United States
Eli D Lazarus, University of Southampton, Southampton, United Kingdom
Jessica L Thompson, Northern Michigan University, Marquette, MI, United States

Introduction: Rise of Participatory Modeling

Understanding the dynamics and interactions of social and ecological systems is critical for managing shared resources and human behaviors towards sustainability. Constructing conceptual representations of systems with complex interactions, multiple feedback loops, and nonlinear dynamics in ways that are useful for decision making is difficult. Any abstracted representation of a system must be accurate to technical experts, accessible to stakeholders, and politically acceptable to decision makers if it is to be used to guide management actions. As policy tools intended to change behavior—the way a given natural resource is used, in this case—system representations must be relevant to those living and working within the system, credible to decision authorities, and legitimate to those within and outside of the decision making process (Clark *et al.*, 2016).

When uncertainty is present, significant changes in shared resource use proposed, or when livelihoods are affected by decision outcomes, a democratic process for representing these systems is best to inform decisions. People intuitively construct a practical understanding of how a particular system they live in works, and how to live within it. These first-hand working "mental models" of the social and ecological systems people live in are made often without explicit analysis or even words (Meadows, 2008; Westervelt and Cohen, 2012). Individuals' mental models and shared "cultural models" are influenced by practices of everyday life forming ideas about how systems work (Paolisso, 2002; Özesmi and Özesmi, 2004; Glynn *et al.*, 2017). Tacit lay conceptualizations of ecological functioning are passed down through social learning, business practices, everyday language, and woven into cultural fabrics of everyday living like farming, ranching, irrigating, fishing, hunting, commuting, and others. Cultural models are often expressed piecemeal in ordinary conversation; circulated in story, shared memories, social norms, and common social meanings (Jones *et al.*, 2011). These mental schemas condition how individuals and societies interact with their environment (Butzer and Endfield, 2012). Integrating lay mental and cultural models with scientific and technical models of how ecological systems work for more complete system understanding for deliberation requires a structured process for developing shared models of systems. This process that depends upon communication.

The process of stakeholder and expert groups constructing integrated and collective representations of coupled social-ecological systems is generally called "participatory modeling." Participatory modeling encompasses several stakeholder-based modeling approaches including mediated modeling (van den Belt, 2004), group modeling (Andersen and Richardson, 1997), group model building, collaborative modeling, comodeling (Levrel *et al.*, 2009), companion modeling (Bousquet *et al.*, 2005; Etienne *et al.*, 2011), participatory simulation, and shared vision planning (Voinov and Bousquet, 2010). Here, we use participatory modeling to describe commonalities among approaches.

Participatory modeling is a practice of coordinated communication of actors' working understandings of how a particular system of interest functions and the organization of these communicated conceptualizations into a shared representation or "model." The process combines collaborative learning, public participation, and system dynamics modeling. The model (a product of the process) may take many forms and is at the discretion of the group driven by the purpose, the system, and the participants (Gray *et al.*, 2018). Ultimately, the model—which may be a conceptual model (image, diagram, statement), computer-based model where functional and/or spatial relationships are made explicit, or a series of negotiated scenarios shared in narrative form—serves as a collaborative tool for communicating diverse understandings about the system of interest for some purpose (social learning, collaboration, decision making) (Röckmann *et al.*, 2012; Henly-Shepard *et al.*, 2015; Seidl, 2015). Ultimately, the modeling process yields a shared understanding of the system enabling groups to work together (**Fig. 1**).

The rise of participatory modeling accompanies three trends within environmental policy and academia. First, the decentralization of environmental decision making and policy that acknowledged the shared nature of environmental quality warrants participation of many stakeholders halting decision making by technical experts alone (technocratic rule) (Fischer, 2000). Environmental decision making is moved from the capital buildings where experts reside into communities where people live. Decentralization recognizes both the right of citizens to participate in environmental decisions that affect them as well as the importance of on-the-ground expertise—local knowledge—which citizens possess that can improve policy via contextualization when incorporated into planning (Berkes *et al.*, 2000; Hall *et al.*, 2012). In the United States, this is exemplified by the National Environmental Policy Act (1969), which mandates government agencies to consult affected citizens before carrying out any actions using federal funds that impact shared natural resources.

The second trend is the intellectual and academic development of systems theory elevated by the advancement of computer technology that enables computer programmers to model complex ecological and social systems with multiple variables, feedback loops, nonlinearities, and emergent properties (Forrester, 1971; Meadows *et al.*, 2004; Liu *et al.*, 2007). This precipitated the

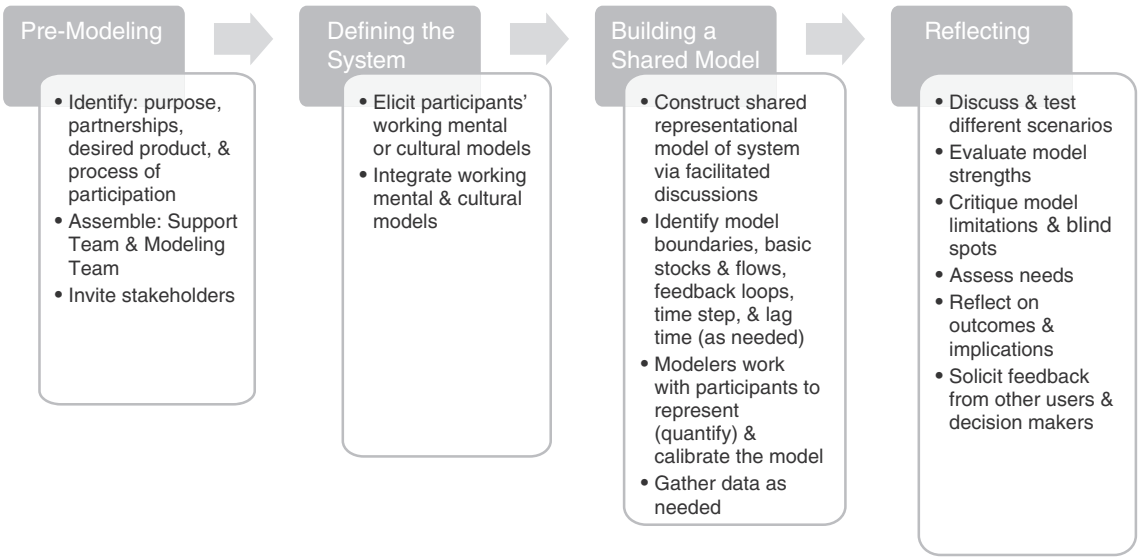


Fig. 1 General modeling process.

development of modeling software (Stella, Fuzzy Cognitive Mapper, GoldSim, POWERSIM, iThink, Simile, Vensim DSS, AnyLogic, others) that facilitated groups' ability to build and animate computer-based systems models collaboratively.

The third trend is the emergence of interdisciplinary research groups like the Resilience Alliance and research fields like sustainability science (Kates *et al.*, 2001; Miller, 2013) explicitly focused on understanding complex systems with the aim of using university research to address pressing sustainability challenges. "Systems thinking" constitutes the central organizing principle of these developing fields (Ostrom, 2009; Wiek *et al.*, 2011). To advance problem solving, requires the integration of social and biophysical sciences (interdisciplinary) and the active participation of public and private sector expertise—stakeholder participation—beyond the university (transdisciplinary) (Gibbons, 1999; Talwar *et al.*, 2011; Mattson *et al.*, 2016). The objective of collaborative problem solving has expedited the maturation of participatory modeling techniques.

Uses of Participatory Modeling

Participatory modeling is used to aid decision making and eventually shape behaviors via informal or formal policy. It is a tool for understanding, conceptualizing, and communicating complex systems to evaluate and improve policy (Schmolke *et al.*, 2010; Voinov *et al.*, 2016). It is increasingly used in environmental management frameworks as it aligns with goals of Ecosystem-based Management, Adaptive Resource Management, and Integrated Water Resources Management (Metcalf *et al.*, 2010; Peterson *et al.*, 2004; Hare, 2011). For example, the European Water Framework Directive calls for integrated water resources management decisions made at the watershed scale and developed with stakeholder participation (Hering *et al.*, 2010)—a good fit for participatory modeling (Carmona *et al.*, 2013; Molina *et al.*, 2011). Participatory modeling has also been use in health care (Vennix, 1999), management (Rouwette and Vennix, 2006; Senge, 2006), and urban water systems (Mirchi *et al.*, 2012).

Models and Participatory Modeling

Models of human–environment dynamics are tools for communicating about complexity (Forrester, 1994). Across all forms, they constitute a way of organizing knowledge (Wierzbicki, 2007). The process of creating a model forces model builders to articulate clearly their held assumptions (Krebs, 2000; Voinov *et al.*, 2014). Effective models describe relationships among system components with a degree of precision not afforded by language alone (Heemskerk *et al.*, 2003; van der Leeuw, 2004).

Models tend to fall into one of two categories (Murray, 2007). Models may be *descriptive*, designed to explain system components and dynamics qualitatively or quantitatively. Models may present a big-picture, simplified, or abstracted perspective of a system for decision groups to improve evaluating problems, assessing management, or testing hypotheses (Grant and Swannack, 2008; Gray *et al.*, 2018). Models may also be used as *predictive* tools, for simulating biophysical systems to forecast trends, evaluate social-ecological interactions under varying management scenarios, to allow comparisons of different technologies on environmental systems, and to anticipate system changes under biophysical or social changes.

It is critical to remember that, regardless of a model's type, a model is a tool—a means to an end—and therefore the purpose of a model should always be made clear throughout a participatory modeling process. Such clarification is especially important in participatory modeling, so that all participants in the problem-solving process work with a shared understanding of the model's

purpose, anticipated uses, and analytical limitations. Clear framing of a model's purpose can thus help address related expectations regarding how the model ultimately informs decision making (Hall *et al.*, 2014).

General Process of Participatory Modeling

The process of modeling is determined by the specifics of the purpose, process of participation, partnerships, and the product (see Gray *et al.*, 2018). Nevertheless, collaborative model building typically involves a series of meetings or workshops designed to elicit and integrate working mental or cultural models, then construct and reflect upon the newly assembled shared representational model. This process constitutes social learning (Pahl-Wostl and Hare, 2004; Hare, 2011) akin to collaborative learning, which combines soft systems methodology and alternative dispute resolution via a series of activities that encourage systems thinking, joint learning, and open communication (Daniels and Walker, 2001).

Stakeholders and experts participate in participatory modeling for various reasons. Stakeholders may simply care about a particular environmental system, depend upon its functioning, want an opportunity to influence policy, or want to ensure that the system is not falsely modeled (Voinov and Gaddis, 2008; Gray *et al.*, 2018). Advice abounds about best practices for involving stakeholders in environmental planning that are transferable to participatory modeling (Forester, 1999; Wondolleck and Yaffee, 2000; Carr, 2015; Hall *et al.*, 2016). However they are included, stakeholders should be involved early and throughout the process (Voinov and Gaddis, 2008; van den Belt and Blake, 2015) because model end-users' "confidence in the model, as well as the overall quality of the model, almost always is higher as the result of early and continued dialogue with potential users" (Grant and Swannack, 2008: 92).

Engaging participants in collaborative model building is administered by a support team. A typical support team consists of: a facilitator who is an impartial party that manages the meetings, guides discussions, summarizes, and brings closure to the meetings; a mediator, an impartial third party who remains involved after the meetings and meets with small groups or individuals between meetings; a modeler (or modeling team) who creates an abstract representation of reality; a recorder who documents the meetings; and a gatekeeper who provides information, advice for the process (Thompson *et al.*, 2010).

Participatory modeling workshops proceed iteratively through a series of phases. The first phase includes problem definition, envisioning, risk assessment, and determining scale issues (van den Belt, 2004). The process includes gathering and eliciting participants' mental and cultural models (Andersen and Richardson, 1997; Luna-Reyes *et al.*, 2006). Elicitation can be direct elicitation of models, such as asking people to diagram their understanding of a system, arrange predetermined variables, or spatially map model components. Or indirect elicitation can be used, where researchers extract system representations from interview transcripts (see Jones *et al.*, 2011 for a review of techniques). During the second phase, participants develop a shared conceptual model of the system, which includes basic stocks and flows, boundaries of the model, feedback loops, and lag time. Data analysis and synthesis occur via facilitated discussions. System scientists or modelers work with the participants to quantify and calibrate the model, gathering data as needed. Finally, participants test their model by trying out different scenarios, and evaluating outcomes (van den Belt, 2004). Some key elements for a workshop series might include integration of ecology, economics, social, and cultural aspects of marsh restoration (see Costanza and Daly, 1992); which requires effective stakeholder participation at the appropriate scale and a linked understanding of past, present, and future relationships (see Peterson *et al.*, 2004; Senge, 2006).

Model selection

Different disciplines and professions construct different kinds of models. Disciplinary training exposes adherents to a selection of models as platforms and software move in and out of popularity. Modelers within the support team will have differing familiarities with different modeling software or limits to software licenses. Despite which modeling platform is selected, the model type, a justification for its selection relative to other modeling approaches, and the objectives for its application to the problem at hand, should be explained clearly, taking into consideration the audience's preexisting experience with modeling. Communicating the implicit assumptions their models use is essential for managing participants' and end-users' expectations (Harte, 2002; Jakeman *et al.*, 2006; Schmolke *et al.*, 2010). The following questions are typically discussed: What is a given model for? What output does it produce and from what input? What is known and unknown? How should its results be interpreted? What insights for decision making, policy, or science can the model yield?

Benefits of Participatory Modeling

In their classic work on conflict resolution, Fischer *et al.* (2011) encapsulated their conflict management process as "one text procedure." Similarly, the participatory modeling process focuses stakeholders with multiple types of expertise on creating "one text"—in this case, a model—to serve as a decision support tool. The modeling practice of focusing a group of experts, citizens, and managers on the activity of building a single text, has a variety of benefits.

Group modeling improves understanding of complex systems. Many persons have expertise used to make decisions across economic, educational, science, cultural, nonprofit, and legal sectors. The activity of model building forces explicit discussions of community values and goals among diverse perspectives into a planning or policy tool. The group activity focused on building a

one-text model forces dialogue to make seemingly incommensurate knowledge commensurate. Groups learn about the system through diverse perspectives; the process elucidates data gaps and information needs (Ritzema *et al.*, 2010). This shared understanding of a complex system increases a group's capacity to collaborate. The resulting model provides ways for groups to explore "what if" scenarios of various decisions and policies.

Participatory modeling can improve environmental policy decisions from stakeholders with local expertise and first-hand experience (Röckmann *et al.*, 2012). Local participants are familiar with local dynamics within regulatory frameworks, organizational cultures, and institutional settings. They remember past efforts to solve problems (Hall *et al.*, 2016). They are capable of developing innovative solutions because of their familiarity with what is politically feasible (Voinov and Gaddis, 2008).

Environmental management agents report having limited knowledge about the publics they serve (Senecah, 2004). Model building can serve as a boundary spanning activity between managers, scientists, and citizens improving relationships within the community via regular engagement and deliberation. Even within systems riddled by conflict, group modeling functions as a "focusing device" which keeps diverse stakeholders focused on the issues and not positions (Fischer *et al.*, 2011). Such collaborative, relationship building among citizens, decision makers, and scientists is valuable in bridging the science-society-policy gaps and mitigating antiintellectualism (Hofstadter, 1963) and distrust of science (Mooney and Kirshenbaum, 2010).

The development of participatory modeling reciprocally benefits the environmental policy and academic trends surrounding its emergence. Fifty years after the National Environmental Policy Act forced US agencies to incorporate the public into environmental decisions that affect them, the intended transition from technocratic management towards citizen-advised forms of resource management is still evolving. Participatory modeling is a tool among many for attaining public participation goals and reflecting upon practices of participation (Seidl, 2015). Further, the practice of participatory modeling advances nascent areas of interdisciplinary research in complex systems thinking and sustainability problem solving through the active development of transdisciplinary collaborative practices.

Risks, Challenges, and Considerations

The assumption within decentralization of environmental policy is that the greater involvement of publics in model construction improves model salience with decision audiences. However, models are political. All models have a backdrop of political gains and losses (Allen *et al.*, 2005; Barnaud *et al.*, 2013). Because any accepted system representation effectively denies alternative representations (Burke, 1966), model conceptions can establish—and decommission—entire paradigms of resource management practices, enacting lasting changes in economic, social, cultural, and ecological functioning (Hall *et al.*, 2014). Such models become perceived in terms of how they affect budget priorities, project approval, and policies (Jacobson and Berkley, 2011; Latour, 2004; Radinsky *et al.*, 2017). Democratic processes of decision making in environmental management arenas do not always come to "successful" ends (Layzer, 2008). The public debate surrounding anthropogenic climate change exemplifies how model-generated predictions can have divisive and significant political, social, and economic impacts (Nisbet, 2009; Fischhoff, 2011). Participatory modeling support teams must remain aware of the impacts of models as selected representations (selections) of reality. Models that do not match—or at least somehow acknowledge—a community's shared values, histories, biases, beliefs, and desired futures are expensive in terms of time and money—litigation, enforcement, education, and other incalculable costs of mistrust. Approaches to modeling that involve and incorporate these innate human aspects of understanding our social and physical contexts may yield more robust decision making instruments—and more sustainable management of natural resources (Glynn *et al.*, 2017).

A model is an abstracted simplification of a system (Allen *et al.*, 2005; Murray, 2007; Grant and Swannack, 2008). While this helps render a complex system more tractable for problem formulation, hypothesis testing, or management, it also forces choices about what can and cannot be modeled. Not all elements can fit within a model. The choices of which variables to include are shaped by epistemic choices as well as technological resources (computational resources, modeler capacities, chosen software). Any variables included in the model must be commensurate with other driving variables and their accompanying dynamics. Finding commensurate variables in social-ecological systems is a difficult enough task for the lone academic. Each component (resource, organism, dynamic relationship) contains implicitly held assumptions and meanings that differ among members of a group. Some social variables lack equivocal units in ecological systems—like trust, power, or influence—and do not conveniently fit into the ecological models (see Hall *et al.*, 2015). These epistemic and technical challenges may appear to participants as political gamesmanship. Furthermore, exploratory models of potential future scenarios—typical of participatory modeling exercises—depend on chains of conditional possibilities; even when they are well informed, such models tend to defy standard quantification of "uncertainty" (for a helpful discussion in ecology, see Higgins *et al.*, 2003). This emphasizes the importance of working with interdisciplinary teams where diversity enables solutions to challenges of incommensurability as well as the importance of communicating the limits of modeling software and the role of models in the ultimate decision making (Gray *et al.*, 2018).

Similarly, not all variables can be conceptually reduced to fit the agreed-upon modeling format. (Again, not all elements can fit within a model.) Modeling requires participants to fit their complex, lived knowledge into scientific discursive forms. Using computer models as tools for insight and prediction always raises questions regarding which scale is the "best" for representing the real system most comprehensively. Even the most reductionist models cannot simulate reality in all its detail, and may trade flexibility for complication. Oppositely, abstracted, extremely simplified models tend to be poorly suited to making credible predictive forecasts for specific settings and contexts, and therefore lack utility for decision makers. These trade-offs require

reflection. Further, a reductionist approach to a resource-management issue, for example, requires a focus on particular resources, and therefore also tends to look for, or contributes to the creation of, specialized institutions for the governance of that particular resource (Buytaert *et al.*, 2014). For some decision makers and stakeholders, the necessity of reduction constitutes an over-simplification at best, or a predetermined decision that undermines the democratic promise of the process. Communicating this part of the process and being aware of how all audiences may perceive these "normal practices" of modeling is paramount.

The process of participatory modeling contains embedded assumptions and values that impact policy outcomes (Paolisso and Trombley, 2017). Social systems are difficult to model. Often the social system is not effectively placed within ecological models (Brulle and Dunlap, 2015). For example, for most trained ecologists humans are external or secondary to more interesting organism-environment interactions. When humans are represented, they often appear as impacts. This may be off-putting for a stakeholder within the system whose livelihood is tied to such impacts. Modelers and support teams should consider the effects on participants of their implicit assumptions resembling characterizations of humans as "bad." This has relational consequences for participatory modeling teams of experts, managers, and citizens. Participants more often volunteering their time and attention. Often the most dedicated stakeholders in attendance, have the most at stake, moreso than the nonlocal scientists, support team, and modelers.

The ownership of the modeling tools may influence participatory modeling outcomes. For example, the technology used to build and run models, the ownership of the computer code, and related institutional structures and arrangements likely have an impact on the process and outputs of participatory modeling (Paolisso and Trombley, 2017).

A wide range of activities and expectations are suggested when members of the public are formally invited to engage in "participatory" decision making, no matter the form (Seidl, 2015). Public participation takes a variety of forms (Arnstein, 1969; Greenwood *et al.*, 1993; Stern and Dietz, 2008) with different levels of participation expected from citizens by environmental agencies or modeling conveners. Most stakeholders participate to have influence in the decision making, yet, in a majority of cases, agencies cannot legally share decision authority (Senecah, 2004). This emphasizes the importance of communicating the terms of participation, clarifying the expectations from participants, planners, the model, and the process (Harrison, 2011; Gray *et al.*, 2018; Radinsky *et al.*, 2017). Clarity will improve stakeholder contributions to the model and planning. This includes clarifying when the desired product is the structured social learning that occurs throughout the process of group modeling rather than any model (Voinov and Bousquet, 2010).

Conclusion

For policy to be meaningful, it must speak to how people experience their world. Participatory modeling offers a way to better represent complex social-ecological systems for decision making: representations that integrate technical discourses with citizens' lived knowledge (Fischer, 2000; Collins and Evans, 2008). Utility in modeling does not derive from accuracy alone. Where human systems and environmental systems are integrated in the same model, and that model is intended to facilitate a common understanding among a broad spectrum of users, the clarity with which a model speaks to both decision makers and the people whose behaviors are the targets of any policy changes becomes a vital quality. Thus, for systems of shared social and ecological resources, a model's eloquence is paramount. Model representations, no matter how accurate, cannot not determine what society should do. That challenge requires engaging citizen expertise.

See also: Ecological Complexity: Citizen Science. General Ecology: Communication. Human Ecology and Sustainability: Socioecological Systems; Ecological Footprint; Ecological Economics 1

References

Allen, T.F.H., Zellmer, A.J., Wuennenbeg, C.J., 2005. The loss of narrative. In: Cuddington, K., Beisner, B.E. (Eds.), Ecological Paradigms Lost: Routes to Theory Change. Hastings, A. (Ed.), Theoretical Ecology Series. New York: Academic Press, pp. 333–370.
Andersen, D.F., Richardson, G.P., 1997. Scripts for group model building. System Dynamics Review 13 (2), 107–129.
Arnstein, S.R., 1969. A ladder of citizen participation. Journal of American Institute of Planners 35 (4), 216–224.
Barnaud, C., Le Page, C., Dumrongrojwatthana, P., Trébuil, G., 2013. Spatial representations are not neutral: Lessons from a participatory agent-based modelling process in a land-use conflict. Environmental Modelling & Software 45, 150–159.
van den Belt, M., 2004. Mediated modeling: A system dynamics approach to environmental consensus building. Washington D.C: Island Press.
van den Belt, M., Blake, D., 2015. Mediated modeling in water resource dialogues connecting multiple scales. JAWRA Journal of the American Water Resources Association 51 (6), 1581–1599.
Berkes, F., Colding, J., Folke, C., 2000. Rediscovery of traditional ecological knowledge as adaptive management. Ecological Applications 10 (5), 1251–1262.
Bousquet, F., Trébuil, G., Hardy, B. (Eds.), 2005. Companion modeling and multi-agent systems for integrated natural resource management in Asia. Los Baños, Philippines: International Rice Research Institute.
Brulle, R.J., Dunlap, R.E., 2015. Sociology and global climate change. In: Dunlap, R.E., Brulle, R.J. (Eds.), Climate change and society: Sociological perspectives. Oxford: Oxford University Press, pp. 1–30.
Burke, K., 1966. Language as symbolic action: Essays on life, literature, and method. Berkeley: University of California Press.
Butzer, K.W., Endfield, G.H., 2012. Critical perspectives on historical collapse. Proceedings of the National Academy of Sciences 109, 3628–3631.

Buytaert, W., Zulkafli, Z., Grainger, S., *et al*., 2014. Citizen science in hydrology and water resources: Opportunities for knowledge generation, ecosystem service management, and sustainable development. Frontiers in Earth Science 2, 26.
Carmona, G., Varela-Ortega, C., Bromley, J., 2013. Participatory modelling to support decision making in water management under uncertainty: Two comparative case studies in the Guadiana river basin, Spain. Journal of Environmental Management 128, 400–412.
Carr, G., 2015. Stakeholder and public participation in river basin management—An introduction. Wiley Interdisciplinary Reviews: Water 2 (4), 393–405.
Clark, W.C., Tomich, T.P., Van Noordwijk, M., *et al*., 2016. Boundary work for sustainable development: Natural resource management at the consultative group on international agricultural research (CGIAR). Proceedings of the National Academy of Sciences 113 (17), 4615–4622.
Collins, H., Evans, R., 2008. Rethinking expertise. Chicago, IL: University of Chicago Press.
Costanza, R., Daly, H.E., 1992. Natural capital and sustainable development. Conservation Biology 6, 37–46.
Daniels, S.E., Walker, G.B., 2001. Working through environmental conflict: The collaborative learning approach. Westport: Preager.
Etienne, M., Du Toit, D., Pollard, S., 2011. ARDI: A co-construction method for participatory modeling in natural resources management. Ecology and Society 16 (1),
Fischer, F., 2000. Citizens, experts, and the environment: The politics of local knowledge. Durham, NC: Duke University Press.
Fischer, R., Ury, W., Patton, B., 2011. Getting to yes: Negotiating agreement without giving in. New York: Penguin.
Fischhoff, B., 2011. Applying the science of communication to the communication of science. Climatic Change 108 (4), 701.
Forester, J., 1999. The deliberative practitioner: Encouraging participatory planning processes. Cambridge, MA: MIT Press.
Forrester, J.W., 1971. World dynamics. Lawrence, KS: Wright-Allen Press.
Forrester, J.W., 1994. Learning through system dynamics a preparation for the 21st Century. In: Keynote Address for Systems Thinking and Dynamic Modeling Conference for K-12 Education. Concord, MA: Concord Academy. June 27–29.
Gibbons, M., 1999. Science's new social contract with society. Nature 402, C81–84.
Glynn, P.D., Voinov, A.A., Shapiro, C.D., White, P.A., 2017. From data to decisions: Processing information, biases, and beliefs for improved management of natural resources and environments. Earth's Future 5, 356–378.
Grant, W.E., Swannack, T.M., 2008. Ecological modeling: A common-sense approach to theory and practice. Oxford: Blackwell.
Gray, S., Voinov, A., Paolisso, M., Jordan, R., BenDor, T., Bommel, P., Glynn, P., Hedelin, B., Hubacek, K., Introne, J., Kolagani, N., 2018. Purpose, processes, partnerships, and products: Four Ps to advance participatory socio-environmental modeling. Ecological Applications 28 (1), 46–61.
Greenwood, D.J., Whyte, W.F., Harkavy, I., 1993. Participatory action research as a process and as a goal. Human Relations 46 (2), 175–192.
Hall, D.M., Gilbertz, S.J., Horton, C.C., Peterson, T.R., 2012. Culture as a means to contextualize policy. Journal of Environmental Studies and Sciences 2 (3), 222–233.
Hall, D.M., Lazarus, E.D., Swannack, T.S., 2014. Strategies for communicating systems models. Environmental Modelling & Software 55, 70–76.
Hall, D.M., Swannack, T.M., Lazarus, E.D., *et al*., 2015. Integrating social power and political influence into models of social-ecological systems. European Journal of Sustainable Development 4 (2), 61–76.
Hall, D.M., Gilbertz, S.J., Anderson, M.B., Ward, L.C., 2016. Beyond "buy-in": Designing citizen participation in water planning as research. Journal of Cleaner Production 133, 725–734.
Harte, J., 2002. Toward a synthesis of the Newtonian and Darwinian worldviews. Physics Today. 29–34.
Hare, M., 2011. Forms of participatory modelling and its potential for widespread adoption in the water sector. Environmental Policy and Governance 21 (6), 386–402.
Harrison, J.L., 2011. Parsing "participation" in action research: Navigating the challenges of lay involvement in technically complex participatory science projects. Society & Natural Resources 24 (7), 702–716.
Heemskerk, M., Wilson, K., Pavao-Zuckerman, M., 2003. Conceptual models as tools for communication across disciplines. Conservation Ecology 7, 8.
Henly-Shepard, S., Gray, S.A., Cox, L.J., 2015. The use of participatory modeling to promote social learning and facilitate community disaster planning. Environmental Science & Policy 45, 109–122.
Hering, D., Borja, A., Carstensen, J., *et al*., 2010. The European water framework directive at the age of 10: A critical review of the achievements with recommendations for the future. Science of the Total Environment 408 (19), 4007–4019.
Higgins, S.I., Clark, J.S., Nathan, R., *et al*., 2003. Forecasting plant migration rates: Managing uncertainty for risk assessment. Journal of Ecology 91 (3), 341–347.
Hofstadter, R., 1963. Anti-intellectualism in American life. New York: Vintage.
Jacobson, R.B., Berkley, J., 2011. Conceptualizing and communicating ecological river restoration. Stream Restoration in Dynamic Fluvial Systems. 9–27.
Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. Environmental Modelling & Software 21 (5), 602–614.
Jones, N., Ross, H., Lynam, T., Perez, P., Leitch, A., 2011. Mental models: An interdisciplinary synthesis of theory and methods. Ecology and Society 16 (1),
Kates, R.W., Clark, W.C., Corell, R., Hall, J.M., Jaeger, C.C., Lowe, I., McCarthy, J.J., Schellnhuber, H.J., Bolin, B., Dickson, N.M., Faucheux, S., 2001. Sustainability science. Science 292 (5517), 641–642.
Krebs, C.J., 2000. Hypothesis testing in ecology. In: Boitani, L., Fuller, T.K. (Eds.), Research techniques in animal ecology. New York: Columbia University Press, pp. 1–14.
Latour, B., 2004. Politics of nature. Cambridge: Harvard University Press.
Layzer, J.A., 2008. Natural experiments: ecosystem-based management and the environment. Cambridge: MIT press.
van der Leeuw, S.E., 2004. Why model? Cybernetic Systems 36, 117–128.
Levrel, H., Etienne, M., Kerbiriou, C., Le Page, C., Rouan, M., 2009. Co-modeling process, negotiations, and power relationships: Some outputs from a MAB project on the island of Ouessant. Society and Natural Resources 22, 172–188.
Liu, J., Dietz, T., Carpenter, S.R., *et al*., 2007. Complexity of coupled human and natural systems. Science 317 (5844), 1513–1516.
Luna-Reyes, L.F., Martinez-Moyano, I.J., Pardo, T.A., *et al*., 2006. Anatomy of a group model-building intervention: Building dynamic theory from case study research. System Dynamics Review 22 (4), 291–320.
Mattson, P., Clark, W.C., Andersson, K., 2016. Pursuing sustainability: An introduction. Princeton.
Meadows, D.H., 2008. In: Wright, D. (Ed.), Thinking in Systems: A Primer. White River Junction, VT: Chelsea Green Publishing.
Meadows, D., Randers, J., Meadows, D., 2004. Limits to growth: The 30-year update. White River Junction, VT: Chelsea Green Publishing.
Metcalf, S.S., Wheeler, E., BenDor, T.K., Lubinski, K.S., Hannon, B.M., 2010. Sharing the floodplain: Mediated modeling for environmental management. Environmental Modelling and Software 25, 1282–1290.
Miller, T.R., 2013. Constructing sustainability science: Emerging perspectives and research trajectories. Sustainability Science 8 (2), 279–293.
Mirchi, A., Madani, K., Watkins, D., Ahmad, S., 2012. Synthesis of system dynamics tools for holistic conceptualization of water resources problems. Water Resources Management 26 (9), 2421–2442.
Molina, J.L., García-Aróstegui, J.L., Bromley, J., Benavente, J., 2011. Integrated assessment of the European WFD implementation in extremely overexploited aquifers through participatory modelling. Water Resources Management 25 (13), 3343–3370.
Mooney, C., Kirshenbaum, S., 2010. Unscientific America: How scientific illiteracy threatens our future. New York: Basic Books.
Murray, A.B., 2007. Reducing model complexity for explanation and prediction. Geomorphology 90, 178–191.
Nisbet, M.C., 2009. Communicating climate change: Why frames matter for public engagement. Environment: Science and Policy for Sustainable Development 51 (2), 12–23.
Ostrom, E., 2009. A general framework for analyzing sustainability of social-ecological systems. Science 325, 419–422.
Özesmi, U., Özesmi, S.L., 2004. Ecological models based on people's knowledge: A multi-step fuzzy cognitive mapping approach. Ecological Modelling 176 (1), 43–64.
Pahl-Wostl, C., Hare, M., 2004. Processes of social learning in integrated resources management. Journal of Community & Applied Social Psychology 14 (3), 193–206.

Paolisso, M., 2002. Blue crabs and controversy on the Chesapeake Bay: A cultural model for understanding watermen's reasoning about blue crab management. Human Organization 61 (3), 226–239.
Paolisso, M., Trombley, J., 2017. Cognitive, material and technological considerations in participatory environmental modeling. In: Gray, S., Paolisso, M., Jordan, R., Gray, S. (Eds.), Environmental modeling with stakeholders: Theory, methods, and applications. New York: Springer International Publishing pp, pp. 3–23.
Peterson, T.R., Kenimer, A., Grant, W.E., 2004. Using mediated modeling to facilitate collaborative learning among residents of the San Antonio Watershed, Texas, USA. In: Mediated Modeling: A Systems Dynamics Approach to Environmental Consensus Building. Washington D.C.: Island Press, pp. 136–163.
Radinsky, J., Milz, D., Zellner, M., *et al.*, 2017. How planners and stakeholders learn with visualization tools: Using learning sciences methods to examine planning processes. Journal of Environmental Planning and Management 60 (7), 1296–1323.
Ritzema, H., Froebrich, J., Raju, R., Sreenivas, C., Kselik, R., 2010. Using participatory modelling to compensate for data scarcity in environmental planning: A case study from India. Environmental Modelling & Software 25 (11), 1450–1458.
Röckmann, C., Ulrich, C., Dreyer, M., *et al.*, 2012. The added value of participatory modelling in fisheries management–what has been learnt? Marine Policy 36 (5), 1072–1085.
Rouwette, E.A., Vennix, J.A., 2006. System dynamics and organizational interventions. Systems Research and Behavioral Science 23 (4), 451–466.
Schmolke, A., Thorbek, P., DeAngelis, D.L., Grimm, V., 2010. Ecological models supporting environmental decision making: A strategy for the future. Trends in Ecology & Evolution 25, 479–486.
Seidl, R., 2015. A functional-dynamic reflection on participatory processes in modeling projects. Ambio 44 (8), 750–765.
Senecah, S., 2004. The trinity of voice: The role of practical theory in planning and evaluating the effectiveness of environmental participatory processes. In: Depoe, S.P., Delicath, J.W., Aepli Elsenbeer, M. (Eds.), Communication and public participation in environmental decision making. Albany: State University of New York Press, pp. 13–33.
Senge, P.M., 2006. The fifth discipline: The art and practice of the learning organization. New York: Broadway Business.
Stern, P.C., Dietz, T., 2008. Public participation in environmental assessment and decision making. Washington DC: National Academies Press.
Talwar, S., Wiek, A., Robinson, J., 2011. User engagement in sustainability research. Science and Public Policy 38 (5), 379–390.
Thompson, J.L., Forster, C.B., Werner, C., Peterson, T.R., 2010. Mediated modeling: Using collaborative processes to integrate scientist and stakeholder knowledge about greenhouse gas emissions in an urban ecosystem. Society and Natural Resources 23 (8), 742–757.
Vennix, J.A., 1999. Group model-building: Tackling messy problems. System Dynamics Review 15 (4), 379–401.
Voinov, A., Bousquet, F., 2010. Modelling with stakeholders. Environmental Modelling & Software 25, 1268–1281.
Voinov, A., Gaddis, E.J.B., 2008. Lessons for successful participatory watershed modeling: A perspective from modeling practitioners. Ecological Modelling 216 (2), 197–207.
Voinov, A., Seppelt, R., Reis, S., Nabel, J.E.M.S., Shokravi, S., 2014. Values in socio-environmental modelling: Persuasion for action or excuse for inaction. Environmental Modelling & Software 53, 207–212.
Voinov, A., Kolagani, N., McCall, M.K., *et al.*, 2016. Modelling with stakeholders–next generation. Environmental Modelling & Software 77, 196–220.
Westervelt, J.D., Cohen, G.L., 2012. Ecologist-developed spatially dynamic landscape models. New York: Springer.
Wiek, A., Withycombe, L., Redman, C.L., 2011. Key competencies in sustainability: A reference framework for academic program development. Sustainability Science 6, 203–218.
Wierzbicki, A.P., 2007. Modelling as a way of organising knowledge. European Journal of Operational Research 176 (1), 610–635.
Wondolleck, J.M., Yaffee, S.L., 2000. Making collaboration work: Lessons from innovation in natural resource management. Washington D.C.: Island Press.

Further Reading

Burke, N.J., Joseph, G., Pasick, R.J., Barker, J.C., 2009. Theorizing social context: Rethinking behavioral theory. Health Education Behavior 36, 55–70.
Castelletti, A., Soncini-Sessa, R., 2007. Bayesian networks and participatory modelling in water resource management. Environmental Modelling & Software 22 (8), 1075–1088.
Chen, H., Chang, Y.C., Chen, K.C., 2014. Integrated wetland management: An analysis with group model building based on system dynamics model. Journal of Environmental Management 146, 309–319.
Cobb, A.N., Thompson, J.L., 2012. Climate change scenario planning: A model for the integration of science and management in environmental decision-making. Environmental Modelling & Software 38, 296–305.
Gray, S., Gray, S., De Kok, J.L., *et al.*, 2015. Using fuzzy cognitive mapping as a participatory approach to analyze change, preferred states, and perceived resilience of social-ecological systems. Ecology and Society 20 (2),
Gray, S., Paolisso, M., Jordan, R., Gray, S. (Eds.), 2016. Environmental modeling with stakeholders: Theory, methods, and applications. New York: Springer.
Gray, S., Paolisso, M., Jordan, R., Gray, S. (Eds.), 2017. Environmental modeling with stakeholders: Theory, methods, and applications. New York: Springer.
Hall, D.M., Lazarus, E.D., Swannack, T.S., 2014. Strategies for communicating systems models. Environmental Modeling & Software 55, 70–76.
Henriksen, H.J., Rasmussen, P., Brandt, G., Von Buelow, D., Jensen, F.V., 2007. Public participation modelling using Bayesian networks in management of groundwater contamination. Environmental Modelling & Software 22 (8), 1101–1113.
Howick, S., Eden, C., Ackermann, F., Williams, T., 2008. Building confidence in models for multiple audiences: The modelling cascade. European Journal of Operational Research 186 (3), 1068–1083.
Liu, J., Hull, V., Batistella, M., *et al.*, 2013. Framing sustainability in a telecoupled world. Ecology and Society 18 (2),
Mooney, H.A., Duraiappah, A., Larigauderie, A., 2013. Evolution of natural and social science interactions in global change research programs. Proceedings of the National Academy of Sciences 110 (Supplement 1), 3665–3672.
Özesmi, U., Özesmi, S.L., 2004. Ecological models based on people's knowledge: A multi-step fuzzy cognitive mapping approach. Ecological Modeling 176 (1), 43–64.
Paolisso, M., 2002. Blue crabs and controversy on the Chesapeake Bay: A cultural model for understanding watermen's reasoning about blue crab management. Human Organization 61 (3), 226–239.
Voinov, A., Bousquet, F., 2010. Modeling with stakeholders. Environmental Modeling & Software 25, 1268–1281.
Voinov, A., Gaddis, E.J.B., 2008. Lessons for successful participatory watershed modeling: A perspective from modeling practitioners. Ecological Modeling 216 (2), 197–207.
Voinov, A., Kolagani, N., McCall, M.K., *et al.*, 2016. Modeling with stakeholders–next generation. Environmental Modeling & Software 77, 196–220.
Vugteveen, P., Rouwette, E., Stouten, H., van Katwijk, M.M., Hanssen, L., 2015. Developing social-ecological system indicators using group model building. Ocean & Coastal Management 109, 29–39.
IAP2, 2007. IAP2 Spectrum of Public Participation. International Association for Public Participation. Available at: http://c.ymcdn.com/sites/www.iap2.org/resource/resmgr/imported/IAP2%20Spectrum_vertical.pdf.

Relevant Website

http://learningforsustainability.net/participatory-modelling/—Participatory model building.

Metapopulation Models☆

Ilkka Hanski† **and Otso Ovaskainen,** University of Helsinki, Helsinki, Finland

Glossary

Detectability The probability of observing an individual or local population that is actually present during the time of a survey.

Extinction debt The number of species that will eventually go extinct due to habitat loss and fragmentation but have not yet had time to do so.

Extinction threshold The critical amount of habitat loss and fragmentation exceeding which leads to metapopulation extinction.

Metapopulation A collection of local populations connected by dispersal.

Patch occupancy model A metapopulation model that is based on classifying patches as empty or occupied and thus ignoring their population sizes.

Introduction

Most landscapes are complex mosaics of many types of habitat. From the viewpoint of a particular species living in such a landscape, only some habitat types, called suitable habitat, provide the resources that are necessary for population growth. The remaining landscape, often called the (landscape) matrix, can only be traversed by migrating individuals. Often the suitable habitat occurs in discrete patches (also called habitat fragments). Individual habitat patches may be occupied by a local population of the focal species, but many patches are likely to be unoccupied at a particular point in time, because a local population went extinct in the past or the patch appeared in the landscape only recently due to, for example, successional changes. The currently unoccupied patches may become colonized in the future. The set of local populations inhabiting the network of habitat patches is called a metapopulation. The local populations are typically connected to other populations by some degree of migration, but how strong this coupling is depends on the structure of the landscape (how far apart the habitat patches are located) as well as on the powers of migration of the species.

Mathematical metapopulation models are used to describe, analyze, and predict the dynamics of metapopulations in fragmented landscapes. This article presents an overview of different kinds of metapopulation models, with an emphasis on spatially realistic models, which can be applied to real metapopulations for the purposes of research, management, and conservation.

Classification of Metapopulation Models

There is no sharp distinction between models for single populations and metapopulations. Classic single-population models assume that individuals interact equally with all other individuals (panmictic population structure). In the other extreme are classic metapopulation models, such as the Levins model described in this article, which assume a set of dynamically independent local populations. But there are also intermediate models. A model involving the spatial coordinates of individuals and limited spatial range of interactions and movements can be viewed as a detailed population model, or it can be considered as a metapopulation model at the landscape level. Such individual-based models for continuous space may exhibit spatial variation in habitat quality and thereby offer a very general framework for population modeling, but their potential to address relevant questions about metapopulation ecology remains largely unstudied and they are not covered in this article. Other models include a description of the genetic structure of the metapopulation, and such models may be used to analyze evolutionary processes in metapopulations. These models too are beyond the scope of this article, which is restricted to ecological metapopulation models.

Stochastic Versus Deterministic Models

Metapopulation dynamics are influenced by four kinds of stochasticity (Table 1): demographic and environmental stochasticity affecting separately each local population, and extinction–colonization and regional stochasticity affecting the entire

☆*Change History:* March 2018. Otso Ovaskainen introduced small edits in the text of the article including citations, added the section "Individual-Based Metapopulation Models", and added a note about imperfect detectability under the section "Spatially Realistic Metapopulation Models".

This is an update of I. Hanski, Metapopulation Models, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 2318–2325.

†Deceased.

 doi:10.1016/B978-0-12-409548-9.11119-4

Table 1 Four types of stochasticity affecting metapopulation dynamics

Type of stochasticity	*Entity affected*	*Correlation among entities*
Demographic	Individuals in local populations	No
Environmental	Individuals in local populations	Yes
Extinction–colonization	Populations in metapopulations	No
Regional	Populations in metapopulations	Yes

metapopulation. Notice that the latter two forms of stochasticity are analogous to demographic and environmental stochasticity. Metapopulation dynamics are inherently stochastic, because population extinction and colonization are stochastic events, and real metapopulations are additionally affected by regional stochasticity.

Nonetheless, stochasticity is often ignored and instead a deterministic model is constructed and analyzed. An important difference between stochastic and deterministic models is that the former predict a smaller or greater time to metapopulation extinction, whereas the latter predict that for some parameter values the metapopulation persists (has a positive equilibrium) while for other parameter values it goes extinct. For large metapopulations the essential behavior of the metapopulation is well captured by a deterministic model. Namely, if the deterministic model predicts that the metapopulation will persist, the stochastic model will predict that the time until metapopulation extinction is very long.

Discrete-Time Versus Continuous-Time Models

Population models can be constructed by assuming discontinuous changes in population size at discrete time intervals, for instance once a year, or continuous temporal changes in population size. Though in reality the sizes of all populations change continuously, the breeding seasons may be so short that a discrete-time model is warranted, particularly if population regulation influences only certain age classes. In discrete-time metapopulation models with population turnover, extinctions and colonizations occur at time intervals that agree with the life history of the species, for example, colonizations follow the season during which migration occurs.

Number of Populations and Description of Landscape Structure

How many populations does the metapopulation consist of? How is the landscape structure represented in the model? Models may consider only two local populations connected by migration, which represents a minimal metapopulation but is nonetheless sufficient to address some general questions about the role of migration in local dynamics. In the other extreme, other models assume an infinite number of habitat patches to simplify model analysis; these models are typically focused on the processes of population extinction and colonization. The treatment of space can be implicit, explicit, or realistic. In the first case, the model includes no information on the spatial locations of habitat patches and local populations, which implies that all local populations are equally coupled to each other. The archetypal spatially explicit model assumes a regular lattice, where lattice cells represent habitat patches. Finally, spatially realistic models involve a description of habitat patches in terms of their spatial coordinates, areas, qualities, and so forth, in a manner that allows the application of the models to networks of real habitat patches.

Five Kinds of Metapopulation Models

Two-Population Models

The simplest metapopulation model extends models for single populations to two populations coupled by migration. For instance, local dynamics may be modeled with the familiar logistic model, and migration by assuming that a fraction m of individuals migrates to the other population. Migration may greatly influence the size of the metapopulation, depending on the form of population regulation, the difference in the carrying capacities in the two patches, and mortality during migration. If local dynamics are modeled with discrete-time models, which are inherently less stable than continuous-time models, more complex dynamics may emerge. Migration may now stabilize local populations that exhibit complex dynamics in the absence of migration, but migration may also amplify population fluctuations, all depending on the details of the model and the exact parameter values. Even limited amount of migration may bring population fluctuations into synchrony.

The two-population models may be extended to n populations, but typically at the cost of the analysis being restricted to simulations. n-population models have been used to study metapopulation viability for conservation. Such models are appealing

to ecologists and conservationists because of their apparent realism, but because of the typically large number of untested structural model assumptions and unmeasured parameter values one cannot place much confidence on model predictions.

The two-population and *n*-population models have been used to study source-sink population dynamics. Some interesting results include the possibility of sink populations enhancing metapopulation stability when source populations exhibit large fluctuations leading to high risk of extinction. A metapopulation consisting of independent sink populations with temporal variation in growth rate may persist even if long-term growth rate is negative in each population in the absence of migration. Migration among such populations enhances metapopulation growth rate by spreading the risk of locally bad period among many independent populations.

Levins Model

The Levins model has a special significance for metapopulation ecology, as it was with this model that Richard Levins introduced the metapopulation concept in two papers published in 1969 and 1970. The Levins model represents a completely different approach to metapopulation modeling in comparison with the two-population model. The assumptions of the Levins model are: (1) The suitable habitat occurs in infinitely many patches that are equally large and of the same quality. (2) The patches have only two possible states, occupied versus empty, and hence the Levins model is an example of patch occupancy metapopulations models, in which local dynamics are ignored. (3) Local extinctions and colonizations occur independently in different patches. (4) All local populations are equally connected to other populations and patches, which is another way of saying that the model is spatially implicit.

With these assumptions, the size of the metapopulation can be described by the fraction of the currently occupied patches, denoted by $p(t)$. Temporal changes in the value of $p(t)$ are given by

$$\mathrm{d}p(t)/\mathrm{d}t = cp(t)(1 - p[t]) - ep(t) \quad (1)$$

where c and e are colonization and extinction rate parameters. The first term describes the rate of colonization of currently unoccupied patches (fraction 1-p) of all patches, while the second term gives the rate of extinction of currently occupied patches (fraction p). Local extinctions may be due to demographic and environmental stochasticity, but at the metapopulation level local stochasticity is assumed to translate to a constant extinction rate parameter. Note that because the existing local populations are assumed to be identical, they all contribute equally to the rate of colonization, and hence the colonization term is proportional to p as well as to 1-p.

The equilibrium metapopulation size is given by $p^* = 1\text{-}\delta$, where $\delta = e/c$ is called the extinction threshold. In the Levins model a species that can invade a currently unoccupied patch network has a positive equilibrium size in the network, and vice versa (Fig. 1**A**). This is not so in more complex models that may have alternative stable equilibria (see section titled "Population size-structured models"). The basic reproductive number R_0 in the Levins model is given by $R_0 = 1/\delta$, which has to exceed unity for the species to be able to spread into an empty patch network.

The Levins model is a deterministic approximation of a model known as the stochastic logistic model, which is an example of stochastic patch occupancy models. The stochastic logistic model assumes a finite network of n patches. If a patch is occupied, it is assumed to go extinct with a fixed rate $E = e$, while the colonization rate of an empty patch is assumed to depend on the fraction of occupied patches, $C = ck/n$, where k is the number of currently occupied patches. These assumptions define a.

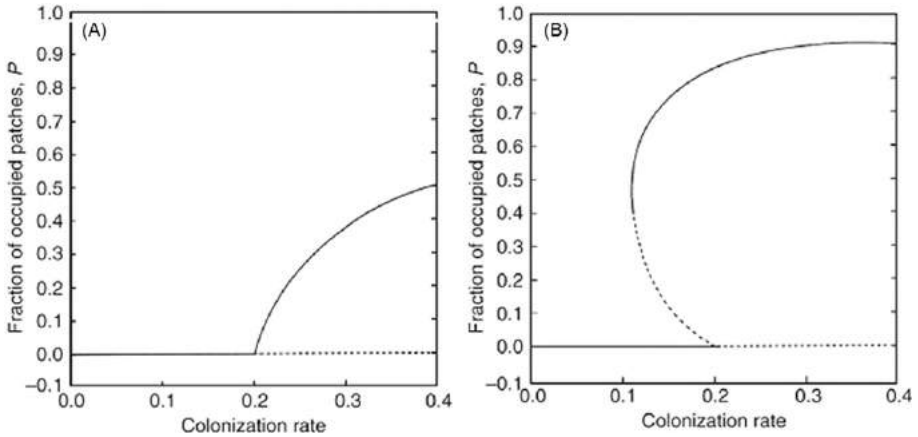


Fig. 1 This figure shows the metapopulation equilibria in relation to increasing colonization rate. Locally stable equilibria are shown by continuous line, unstable equilibria by broken line. (A) In the Levins model, the metapopulation may invade a patch network at the extinction threshold, at which point metapopulation extinction ($p^* = 0$) becomes an unstable equilibrium and only the positive equilibrium is stable. (B) In population size-structured models with strong rescue effect, there are two alternative stable equilibria for some parameter values, in which a metapopulation that is originally sufficiently large may persist in a network that it cannot invade from small size (in the example colonization rate between 0.11 and 0.2). From Hanski, I. (1999). *Metapopulation Ecology.* Oxford: Oxford University Press.

Markov process with metapopulation extinction as an absorbing state, that is, sooner or later the entire metapopulation will go extinct. A key prediction of the model is the mean time to extinction, T. Using a diffusion approximation to analyze the model, which is justified when the number of patches is large, we obtain:

$$T = \sqrt{\frac{2\pi}{n}} \frac{e^{-(n-1)p*}}{p^{*2}(1-p^*)^{n-1}} \tag{2}$$

where p^* is the size of the metapopulation at quasi-equilibrium. Fig. 2 gives the number of patches n that the network must have to make T at least 100 times as long as the expected lifetime of a single local population. For metapopulations with large p^* a modest network of around 10 patches is sufficient to allow long-term persistence, but for rare species (say $p^* < 0.2$) a large network of $n > 100$ is needed for long-term persistence.

The stochastic logistic model includes extinction–colonization stochasticity but no regional stochasticity, which leads to correlated extinctions and colonizations. In the presence of regional stochasticity the mean time to metapopulation extinction does not increase exponentially with increasing n as predicted by Eq. (2) but as a power function of n, the power decreasing with increasing correlation in extinction and colonization rates. This result is analogous to the effects of demographic and environmental stochasticities on the lifetime of single populations.

Population Size-Structured Models

The stochastic logistic model extends the Levins model by incorporating extinction–colonization stochasticity in the model. Population size-structured models represent another type of extension. These models are deterministic and retain the assumption of infinitely many identical and equally connected patches, but the models include a description of local dynamics and migration in the same manner as the two-population models. The simplest structured model divides existing local populations into just two classes, small and large, which is sufficient to model one important new process in comparison with the Levins model: migration from existing large populations (which send out many migrants because they are large) may increase the growth rate in small populations and thereby rescue them from local extinction. At the metapopulation level, the rescue effect leads to the novel prediction of alternative stable equilibria, one of which is metapopulation extinction, the other one a positive metapopulation size (Fig. 1**B**). Therefore, a metapopulation may persist in a network that cannot be invaded from small metapopulation size.

Spatially Realistic Metapopulation Models

An important simplification of the models that have been discussed so far is that they involve no description of the landscape structure, hence it is not possible to apply the models in a quantitative manner to real metapopulations. Real habitat patch networks consist of a finite number of patches that typically differ in their size and quality, and in their spatial connectivities to existing populations, which affect colonization rate.

To model metapopulations in such heterogeneous patch networks, we extend the stochastic logistic model to situations in which the extinction and colonization rates (in continuous-time models) or probabilities (in discrete-time models) are specific to particular habitat patches. This leads to heterogeneous stochastic patch occupancy models. Below, a deterministic approximation of the full stochastic model is discussed, which leads to a family of models with a clear relationship to the Levins model and to models that are helpful for practical applications.

In continuous time, the deterministic model is defined by a set of n equations for a network of n patches,

$$\mathrm{d}p_i(t)/\mathrm{d}t = C_i(\mathbf{p}[t])(1 - p_i[t]) - E_i(\mathbf{p}[t])p_i(t) \tag{3}$$

where $\mathbf{p}$ is a vector of probabilities p_i of each of the n patches being occupied, and C_i and E_i are patch-specific colonization and

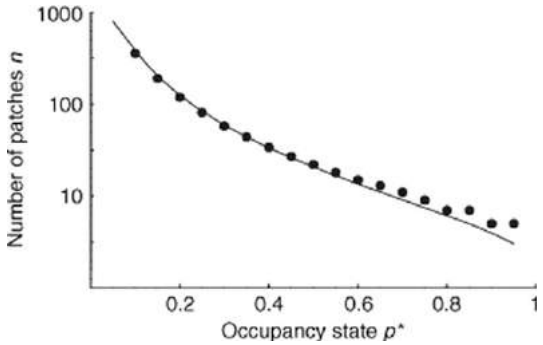


Fig. 2 The number of habitat patches needed to make the mean time to metapopulation extinction at least 100 times longer than the mean time to local extinction. The *dots* show the exact result based on the stochastic logistic model, while the line is based on the approximation given by Eq. (2). From Ovaskainen, O. and Hanski, I. (2003). Extinction threshold in metapopulation models. *Annales Zoologici Fennici* **40**, 81–97.

extinction rates. To link colonization and extinction rates to the structure of the fragmented landscape, we make assumptions of how these rates depend on the properties of the habitat patches, leading to spatially realistic metapopulation theory.

Generally, the extinction risk decreases with increasing patch area A_i, because large patches, when occupied, tend to have large populations with a small risk of extinction. A reasonable parametric assumption is

$$E_i = e/A_i^{\zeta_{ex}}$$

where e and ζ_{ex} are two parameters. The colonization rate increases with connectivity to existing populations, with connectivity defined as a measure of the expected rate of migration from all possible source populations to the focal habitat patch. A reasonable parametric assumption of connectivity is

$$S_i = A_i^{\zeta_{im}} \sum_{j \neq 1} \exp\left(-\alpha d_{ij}\right) A_j^{\zeta_{em}} p_j \tag{4}$$

Thus connectivity of patch i is obtained as a sum of contributions from all other patches from which migrants may arrive. The contribution of patch j increases with its area A_j, scaled by power ζ_{em}; with decreasing distance between patches i and j, with $1/\alpha$ giving the average migration distance; and with the probability of occupancy of patch j, as migrants can only arrive from occupied patches. Immigration to patch i may depend on its own area, scaled to power ζ_{im}. The colonization rate is then defined as $C_i = cS_i$.

The full dynamics of the model is described by the system of n coupled Eq. (3), but its essential behavior can be approximated by just one equation,

$$\mathrm{d}p_\lambda/\mathrm{d}t = c' p_\lambda (1 - p_\lambda) - e' p_\lambda \tag{5}$$

Here metapopulation size is measured as the weighted average of the patch-specific occupancy probabilities p_i, $p_\lambda = \Sigma V_i p_i$, where the weights V_i satisfy $\Sigma V_i = 1$, and they are derived from the relative contributions of individual patches to the capacity of the landscape to support a metapopulation. Large and well-connected patches have larger values of V_i than small and poorly connected patches, because large patches have populations with small risk of extinction and send out many migrants to colonize new patches, especially if they are located close to the other patches. Note that this model is structurally identical with the Levins model, though metapopulation size is measured in a different manner and the colonization and extinction parameters are defined as $c' = c\lambda_M/\omega$ and $e' = e/\omega$, where $\omega = \Sigma V_i A_i$. The new quantity here is λ_M, which is called the metapopulation capacity of the fragmented landscape. Mathematically, λ_M is the leading eigenvalue of a "landscape" matrix, which is constructed with assumptions about how habitat patch areas and connectivities influence extinctions and colonizations. To compute λ_M for a particular landscape one needs to know the scale of connectivity, set by parameter α in Eq. (4) for connectivity, the scaling of immigration, emigration, and extinction rates by patch area (ζ_{im}, ζ_{em}, and ζ_{ex}), and the areas and spatial locations of the habitat patches. The size of the metapopulation at equilibrium is given by

$$p_\lambda^* = 1 - \delta/\lambda_M \tag{6}$$

An attractive feature of the spatially realistic models is the possibility of estimating the values of model parameters with empirical data on the structure of the patch network (patch areas, spatial coordinates, and possibly other information) and spatiotemporal pattern of patch occupancy. Data should preferably be available for several years, including many extinction and colonization events. Having estimated model parameters based on data from a particular landscape, one may predict metapopulation dynamics and persistence in other landscapes, for instance to guide management and conservation actions. The models are most appropriate for highly fragmented landscapes, where all habitat patches are small or relatively small and the respective populations have hence a substantial risk of extinction.

When fitting a metapopulation model to data, it is important to account for the detectability, for example, the possibility that a local population was present but not observed. Failing to account for detectability can lead to biased estimates of colonization and extinction rates. For example, consider a situation in which a local population was actually present in three consequtive years, but it was not observed in the second year. If not accounting for imperfect detectability, the data would suggest that both an extinction and a colonization took place, inflating the estimates of both the extinction rate parameter e and the colonization rate parameter c.

Individual-Based Metapopulation Models

The models described in the previous section "Spatially Realistic Metapopulation Models" classify the patches simply as occupied or empty. While such models can provide many kinds of insights on metapopulation ecology, they lack an explicit connection to individual-level processes. For example, in Eq. (4) the connectivity (and hence colonization rate) of a patch is assumed to increase with the area A_j of a potential source patch, raised to the power of ζ_{em}. The scaling of the patch area with ζ_{em} models two very different proceses at the same time. First, the size of a local population is likely to increase with patch area, making a positive contribution to ζ_{em}. Second, an individual present in a patch is less likely to emigrate from a large patch than from a small patch, making a negative contribution to ζ_{em}. The value of ζ_{em} estimated through a patch occupancy model combines the influences of these two processes, thus failing to separate their roles. More mechanistic insights on connectivity and other components of metapopulation dynamics can be obtained with the help of individual-based models.

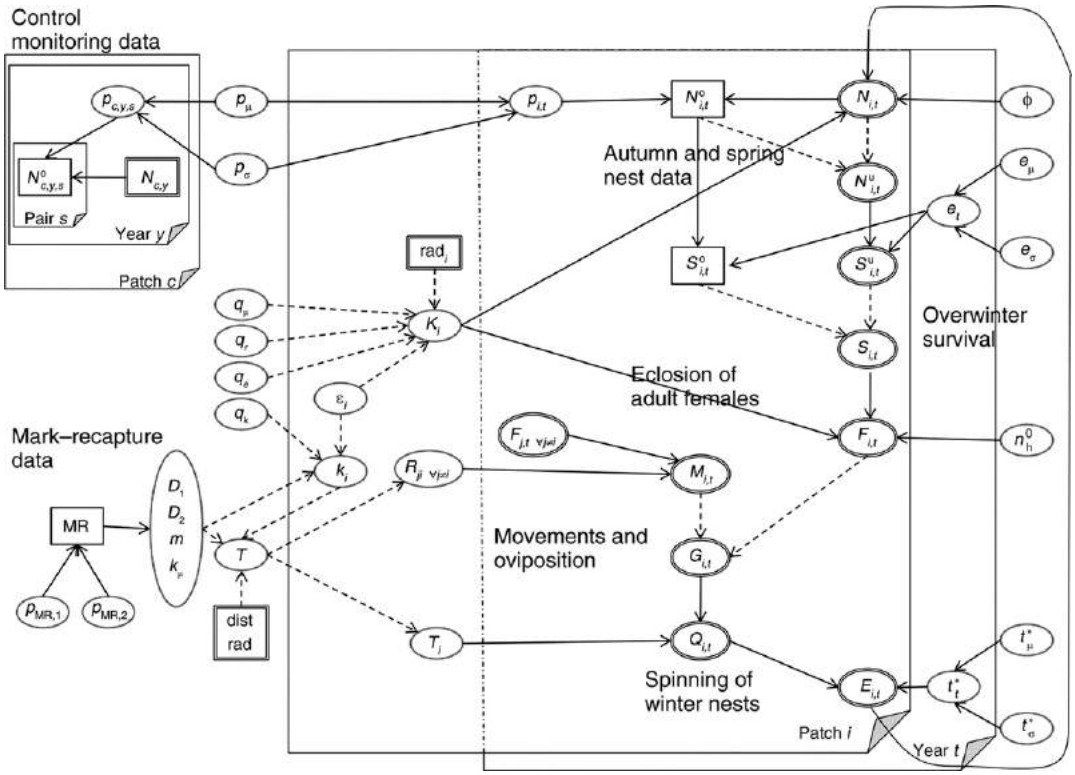


Fig. 3 A graphical description of an individual-based metapopulation model. The quantities in the double ellipses describe the state of the metapopulation (e.g., number of individuals present in each patch), and the quantities in the single ellipses are the parameter values that are estimated. In the single boxes are the observations used to estimate the model, and in the double boxes are the fixed quantities (patch areas and locations). The *full arrows* correspond to stochastic relationships and the *dashed arrows* to deterministic relationships. Imperfect detection is accounted for through an observation model, which describes the probability of observing individuals in mark-recapture studies (used to parameterize the movement part of the model) as well as annual surveys (used to parameterize the demographic part of the model). From Harrison, P.J., Hanski, I. and Ovaskainen, O. (2011). Bayesian state-space modeling of metapopulation dynamics in the Glanville fritillary butterfly. *Ecological Monographs* **81**, 581–598.

If sufficiently rich data are available, it is possible to construct individual-based versions of a spatially realistic metapopulation models and estimate their parameters. Such models are necessarily rather complex, as they generate metapopulation dynamics from individual-level behavior. For example, in the butterfly metapopulation model illustrated in Fig. 3, the individuals move according to a diffusion model both within and between the habitat pacthes, and the emigration rate is controlled by the relucatance of the individuals to leave a patch when encountering its boundary. The demographic part of the model involves the oviposition of eggs by the females, and the survival of the eggs through larvar stages until the eclosion of adult butterflies that form the next generation.

Given the complexity of individual-based models compared to that of patch occupancy models, it is natural to ask what are the additional insights that individual-based models may bring about. First, if parameterized with sufficiently rich data, individual-based models can make more accurate predictions than patch occupancy models, and thus be useful for example, in the conservation context. Second, individual-based models can help to gain a mechanistic understanding on how metapopulation dynamics are generated by the individual-level processes. For example, unlike patch occupancy models, invidual-based models do not involve phenomenological colonization and extinction rate parameters. Instead, a colonization takes place when a migrating individual moves to an empty patch and reproduces there, whereas a local extinction takes place when the last individual in a patch dies or emigrates. Individual-based metapopulation models also work as a natural starting point for studying evolutionary metapopulation dynamics.

Models for Dynamic Landscapes

The vast majority of metapopulation models assume that the landscape structure remains unchanged, only the population sizes and the spatial pattern of habitat occupancy changes. This is a reasonable assumption for many but not all species and landscapes. In particular, metapopulation structures are common in many successional habitats, in which changing habitat quality is an important reason for local extinctions. Human-caused habitat loss and fragmentation are often so fast that one cannot assume

metapopulations to occur at quasi-equilibrium with respect to the current landscape structure, and we may ask how long it takes for a metapopulation to reach the new quasi-equilibrium (which may be metapopulation extinction) following a change in landscape structure. This latter question will be addressed in the section "Transient Dynamics And Extinction Debt" below.

One may extend a stochastic patch occupancy model to include turnover in habitat patches. A general formulation for the stationary probability of occupancy of patch i in a network of n patches is given by

$$J_i = \frac{C_i}{C_i + E_i} \tag{7}$$

J_i is often called the incidence of occupancy, and the model an incidence function model. Assuming that existing patches disappear (which increases the extinction probability of the respective local population) and new patches appear in the landscape, the incidence function may be written as

$$J_i = \frac{C_i - C_i(1 - C_i - E_i)^{age}}{C_i + E_i} \tag{8}$$

where *age* is the age of the patch. Note that initially the patch is always unoccupied ($J_i = 0$ when $age = 0$), while the incidence in old patches (*age* large) approaches the incidence given by Eq. (7). Given data on the ages of the patches in addition to data on patch areas and connectivities and the incidences of occupancy, this model may be parametrized in the same manner as other stochastic patch occupancy models. The size of the metapopulation at quasi-equilibrium is given by the sum of the patchspecific incidences. Given that E_i in Eq. (8) is greater than in Eq. (7), it is clear that landscape dynamics reduce metapopulation size.

Metapopulation Models for Two or More Species

Like all species, species living in fragmented landscapes interact with other species. Metapopulation models have been constructed for competing species and for prey and predator inhabiting the same patch network. This has led to several important results. For instance, a species with a high colonization rate may coexist at the landscape level with a species that is a superior competitor locally, essentially because the former species finds a temporary refuge in those patches that have not yet been colonized by the superior competitor. Similarly, a prey that is driven to extinction locally may persist in a network of habitat patches if it has a sufficiently high colonization rate. Spatially restricted range of movements and ecological interactions may give rise to strongly aggregated spatial distributions of interacting species in the absence of any environmental heterogeneity, which has been studied with continuous-space and discrete-space (lattice) models. The latter are conceptually related to metapopulation models, especially if one assumes that lattice cells (representing habitat patches) are occupied by local populations rather than by single individuals. Spatial pattern formation due to ecological interactions is less likely to be a dominant feature in metapopulation models for fragmented landscapes, because the (fixed or dynamic) spatial variation in habitat quality strongly constrains (meta) population dynamics.

Applications to Landscape Management and Conservation

Habitat Loss and Extinction Threshold

Loss and fragmentation of natural habitats due to human land use is the most important reason for the current catastrophically high rate of loss of biodiversity on Earth. Population viability depends, among other things, on the environmental carrying capacity and hence on the total amount of habitat available. Additionally, the spatial configuration of habitat may influence viability, because most species have limited migration ranges and hence not all habitat in a highly fragmented landscape is readily accessible. Metapopulation models have been used to address the population dynamic consequences of habitat loss and fragmentation.

In the Levins model, habitat loss has been modeled by assuming that fraction 1-h of the habitat patches becomes unsuitable for occupancy and reproduction, while fraction h remains suitable. abitat loss reduces the colonization rate to $cp(h\text{-}p)$, because the model assumes that fraction 1-h of the migrants land on unsuitable habitat and perish (the model thus most literally applies to species with long-range passive migration and no habitat selection). The species persists in a patch network if h exceeds the threshold value δ set by the extinction-proneness and colonization capacity of the species. An interesting implication of this result is that, at equilibrium, the fraction of suitable but unoccupied patches ($h\text{-}p^*$) is constant and equals the amount of habitat at the extinction threshold ($h = \delta$). This is a potentially helpful result, because it suggests a way of measuring the value of the extinction threshold given knowledge of h and p^*. The model is however exceedingly simple and hence not really suitable for quantitative predictions.

The spatially realistic metapopulation models combine the metapopulation perspective of the Levins model with a description of the spatial distribution of habitat in a fragmented landscape and how the landscape structure influences the extinction and colonization processes. In the model described by Eq. (5), the metapopulation capacity λ_M replaces the fraction of suitable patches h in the Levins model, and the threshold condition for metapopulation persistence is accordingly given by $\lambda_M > \delta$. The novelty here is that λ_M takes into account not only the amount of habitat in the landscape but also how the remaining habitat is

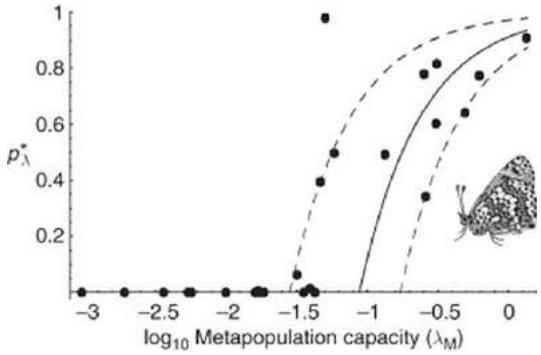


Fig. 4 Metapopulation size of the Glanville fritillary butterfly (*Melitaea cinxia*) as a function of the metapopulation capacity in 25 habitat patch networks. The *vertical axis* shows the size of the butterfly metapopulation based on a survey of habitat patch occupancy in 1 year. The empirical data have been fitted by a spatically realistic model. The result provides a clear empirical example of the extinction threshold. From Hanski, I. and Ovaskainen, O. (2000). The metapopulation capacity of a fragmented landscape. *Nature* **404**, 755–758.

distributed among the individual habitat patches and how the spatial configuration of habitat influences extinction and colonization rates and hence metapopulation viability.

The metapopulation capacity can be computed for multiple landscapes and their relative capacities to support viable metapopulations can be compared even without knowledge of the threshold value δ: the greater the value of λ_M, the better (Fig. 4). One complication is regional stochasticity, which can only be included in the calculations if there is empirical knowledge about the spatial scale and the strength of stochasticity. Without regional stochasticity, it always helps to have as short distances as possible between habitat patches. With regional stochasticity, metapopulation viability is likely to be maximized by intermediate distances between the habitat patches: long enough to reduce the correlation in local dynamics, but not too long to reduce migration too greatly.

Transient Dynamics and Extinction Debt

A change in the structure of a fragmented landscape, for instance a reduction in the area of some habitat patches, will influence the patch-specific extinction and colonization rates, but it takes time before the metapopulation has reached a new quasi-equilibrium following environmental change. The length of the transient period is longer when the change in landscape structure is greater, when the rates of extinction and colonization are lower, and when the new quasi-equilibrium following environmental change is located close to the extinction threshold. The latter result has important implications for conservation. Species that have become endangered due to recent changes in landscape structure are located, by definition, close to their extinction threshold, and hence the length of the transient period in their response to environmental change is predicted to be long. This means that we are likely to underestimate the level of threat to endangered species, because many of them do not occur at quasi-equilibrium with respect to the present landscape structure but are only slowly declining due to past habitat loss and fragmentation.

Considering a community of species, the term extinction debt refers to situations in which, following habitat loss, the threshold condition is not met for some species, but these species have not yet had time to go extinct. More precisely, the extinction debt at a given point in time is the number of extant species that are predicted to go extinct in the future because the threshold condition is not satisfied for them.

Reserve Selection

Setting aside a sufficient amount of habitat as reserves is essential for conservation of biodiversity. Reserve selection should be made in such a manner that a given amount of resources for conservation makes a maximal contribution toward maintaining biodiversity. In the past, making the optimal choice of reserves out of a larger number of potential reserves was typically done by selecting reserves that together would include the largest number of species, without any consideration for the long-term viability of the species. More appropriately, we should ask the question which choice of reserves maintains the largest number of species to the future, taking into account spatiotemporal dynamics of species and possibly and preferably also predicted changes in climate and land use. Metapopulation models can be incorporated into analyses that aim at providing solutions to such questions.

Summary

Metapopulations are assemblages of local populations inhabiting networks of habitat patches in fragmented landscapes. The local populations are coupled by migration among the populations. Metapopulation models are used to describe, analyze, and predict the dynamics of metapopulations. Models have been constructed to investigate the consequences of migration on the pattern and

synchrony of local dynamics, the processes that may lead to spatial pattern formation due to ecological interactions, the persistence of metapopulations in patch networks in a stochastic balance between local extinctions and colonizations, and the response of metapopulations to changing landscape structure. Metapopulation models for highly fragmented landscapes predict an extinction threshold, a critical amount and configuration of habitat that is necessary for long-term metapopulation persistence. The models thus predict that with increasing habitat loss and fragmentation, species will go extinct before all habitat has been lost. Spatially realistic metapopulation models combine a description of landscape structure with a model of the extinction and colonization processes. These models can be parameterized with empirical data and applied to real metapopulations for the purposes of research, management, and conservation.

See also: Behavioral Ecology: Dispersal–Migration. Conservation Ecology: Source–Sink Landscape. Ecological Processes: Succession and Colonization. Evolutionary Ecology: Colonization

Further Reading

Hanski, I., 1999. Metapopulation ecology. Oxford: Oxford University Press.
Hanski, I., Gaggiotti, O.E. (Eds.), 2004. Ecology, genetics, and evolution of metapopulations. Amsterdam: Elsevier.
Hanski, I., Ovaskainen, O., 2000. The metapopulation capacity of a fragmented landscape. Nature 404, 755–758.
Hanski, I., Ovaskainen, O., 2002. Extinction debt at extinction threshold. Conservation Biology 16, 666–673.
Harrison, P.J., Hanski, I., Ovaskainen, O., 2011. Bayesian state-space modeling of metapopulation dynamics in the Glanville fritillary butterfly. Ecological Monographs 81, 581–598.
Levins, R., 1969. Some demographic and genetic consequences of environmental heterogeneity for biological control. Bulleting of the Entomological Society of America 15, 237–240.
Levins, R., 1970. Extinction. Lecture Notes in Mathematics 2, 75–107.
MacKenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G., Franklin, A.B., 2003. Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. Ecology 84, 2200–2207.
Ovaskainen, O., Hanski, I., 2002. Transient dynamics in metapopulation response to perturbation. Theoretical Population Biology 61, 285–295.
Ovaskainen, O., Hanski, I., 2003. Extinction threshold in metapopulation models. Annales Zoologici Fennici 40, 81–97.
Ovaskainen, O., Meerson, B., 2010. Stochastic models of population extinction. Trends in Ecology & Evolution 25, 643–652.
Verheyen, K., Vellend, M., Van Calster, H., Peterken, G., Hermy, M., 2004. Metapopulation dynamics in changing landscapes: A new spatially realistic model for forest plants. Ecology 85, 3302–3312.

Model Types: Overview☆

Sven E Jørgensen, Copenhagen University, Copenhagen, Denmark
Todd M Swannack, Texas State University, San Marcos, TX, United States

Introduction: Classification of Ecological Models

An overview of the available ecological models is best provided by presenting different applied classifications of ecological models. Five classifications will be presented, namely:

1. What is modeled: matter, energy, or population
2. Classification of all the models in nine different pairs of models (this classification involves as much as $2^9 = 512$ classes)
3. Type of model employed (11 types given)
4. Type of system modeled; and
5. Type of problem modeled

Modeling Matter, Energy, or Populations

Table 1 shows what is modeled, the organization, and the pattern.

Model Pairs

Models can be classified by nine model type pairs, shown in Table 2. It implies that all models belong to one of 512 classes, corresponding to all combinations of the nine pairs. A model could, for instance, to illustrate this classification, be a research model, that is, deterministic, and at the same time a compartment model, that is, dynamic, nonlinear, causal, distributed, holistic, and spatial; or it could be a management model, that is, stochastic, and at the same time a compartment model, that is, dynamic, linear, lumped, holistic, and nonspatial.

Eleven Different Model Types/Tools

Ecological modeling, as a tool in environmental management, rose in popularity significantly during the 1970s. In general, three model types were applied at that time, namely population dynamic models with or without age structure represented by matrices, biogeochemical or bioenergetic dynamic models based on differential equations, and static models, corresponding to all differential equations, were zero. The latter type was mainly applied to describe an extreme or average situation.

It was, however, acknowledged that other model types to solve more comprehensive modeling problems were urgently needed. The needs can be formulated as questions:

1. Ecosystems are middle number systems, in the sense that the number of components is orders of magnitude smaller than the number of atoms in a system. All the components are different and that is often important for a proper description of the ecosystem reactions to consider the differences in properties among individuals.
2. Ecosystems and the species that occupy them are adaptable and may change their properties to meet the changes in forcing functions. Furthermore, the species may be replaced by other species better fitted to the combinations of forcing functions. How to account for these changes? Even the interactions and connections among components of the ecosystem (i.e., the ecosystem network) may be changed if more biological components with different properties are replaced by other species. How to account for these structural changes?
3. Can we model a system that has only uncertain observations/data?
4. How to account for stochastic forcing functions and processes?
5. How to develop models for a heterogeneous data set, that is, based on observations from many different ecosystems?
6. How to develop models of ecosystems, when our knowledge is mainly based on a theory or rules/properties/propositions?

☆*Change History*: March 2018. Todd M. Swannack prepared the update. The article was significantly revised from the previous version. The major changes include removing unnecessary text, deleting one table, and adding new model types. The reference section was updated, and a new section on Integrated/Hybrid models was added.

This is an update of S.E. Jørgensen, Model Types: Overview, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008.

Table 1 Classification of models

Modeled/measured	*Organization*	*Pattern*	*Model type*
Number of individuals	Conservations of genes	Life cycles	Biodemographic
Energy	Conservations of energy	Energy flow	Bioenergetic
Mass or concentration	Conservations of mass	Element flow	Biogeochemical

Table 2 Classification by model pairs

Pair 1: Is the model applied for research or management?
Research models
Management models
Pair 2: Is the model deterministic or stochastic?
Deterministic models
Stochastic models
Pair 3: Does the model apply matrices or differential equations?
Matrix models
Compartment models
Pair 4: Are the variables dependent or not on time?
Dynamic models
Static models
Pair 5: Are the equations linear or nonlinear
Linear models
Nonlinear models
Pair 6: Is the model based on causality or is no causality included?
Causal models
Black box models
Pair 7: Are the parameters (the properties of the state variables) dependent on time and/or space or constant?
Distributed models
Lumped models
Pair 8: Is a reductionistic or holistic model approach applied?
Reductionistic models
Holistic models
Pair 9: Is the model considering spatial distribution?
Spatial models
Nonspatial models
Pair 10: Are the equations solved numerically or analytically?
Numerical models
Analytical models
Pair 11: Are the model results discrete or continuous?
Discrete models
Continuous models

7. How to consider toxic substances in the environment? Does development of a toxic substance model require a special model type?
8. How to describe the spatial distribution, which is often crucial for the understanding of ecosystem processes?

Model Types/Tools Available Today

Since the genesis of ecological modeling in the late 1970s, a number of new model types have been developed, due in part to the need to model more complex environments, more integration among disciplines, and significant increases in computing power. Another question is "to what extent have these new model types been applied in ecological modeling?"

The general trend in ecological modeling has been toward models increasing in complexity, with the classical model types, particularly biogeochemical models and population dynamic models, being integrated into more complex model frameworks that combine multiple model types. That is, the principles underlying those models are being applied as useful tools in the development of several different model types. Statistical models are not covered in this section statistics is considered as a tool that can be applied in ecological modeling to give better process description. If a model is based entirely on statistics, it is a so-called black box model, because it has no causality. This section presents different types of ecological models with a strong foundation in the science of ecology.

For historical context, for the period 1975–82, the applications of biogeochemical and bioenergetic dynamic models and population dynamic models are dominant. Fuzzy models, spatial distribution models, structurally dynamic models (SDMs), and models using catastrophe theory were used in the period 1975–82 but their application in ecological modeling was very modest and they were new and untested tools in ecological modeling. Particularly, SDMs, artificial neutral network (ANN), and individual-based models (IBMs) were more extensively applied from 2000 forward, due to increased modeling software availability and a larger need for those types of models in environmental applications. The use of static models, like food web models, is due to wide application software availability that allows modelers to run large, complex simulations in a relatively short period of time—these models are particularly useful for fisheries and various other types of aquatic ecosystems.

Dynamic Models—Biogeochemical and Bioenergetics

Biogeochemical and bioenergetic models generally apply differential equations to express the dynamics. Changes in state variables are expressed as the results of the incoming minus the outgoing substances and the model is therefore based on conservation principles. The process equations are based usually on causality, but can in principle also be a result of a statistic analysis of data. The model type has some clear advantages that make it attractive still to use this model type for the development of many models.

The advantages are as follows:

- most often based on causality;
- based on mass or energy conservation principles;
- easy to understand, interpret, and develop;
- software is readily available (e.g., systems dynamics software);
- easy to use for predictions.

The disadvantages are as follows:

- difficulty developing and parameterizing with heterogeneous data;
- requires relatively good data;
- is difficult to calibrate when it is complex and contains many parameters;
- cannot account for adaptation and changes in species composition.

The advantages and disadvantages define the so-called area of application: for description of the state of an ecosystem, when a good data set is available. A developed model may be applied on different ecosystems of the same type, although calibration and validation should always be carried out for each case study. The model will often but not always take many processes and several state variables into account and require therefore in most cases a good data set. The model type has been extensively applied in environmental management as a powerful tool to understand the reactions of ecosystems to pollutants and to set up prognoses.

Static Models

The model type is a biogeochemical or bioenergetic dynamic model where all the differential equations are set to zero to obtain the values of the state variables corresponding to the static situation. The model provides a *snapshot in time* view of the system, but does not predict or project dynamics forward.

The advantages are as follows:

- requires smaller databases than other types;
- is excellent to give a worse-case or average situation;
- the results are easily validated (and verified).

The disadvantages are:

- does not give any information about dynamics and changes over time;
- prediction with time as independent variable is not possible;
- can only give average or worse-case situations.

This model type will often be used when a static situation is sufficient to give a proper description of an ecological system or to take environmental management decisions.

Population Dynamic Models

This model type is rooted in the Lotka–Volterra model that was developed in the 1920s. Numerous papers have been published about the mathematics behind this model and a number of deviated models. The mathematics of these equation systems is not

very interesting from an ecological modeling point of view, where the focus is a realistic description of ecological populations. Population dynamic models may include age structure, which in most cases is based on matrix calculations.

Population dynamics models can be classified as many different types of models, including individual-based, matrix, statistical, population viability analyses, analytical, among others.

The advantages are as follows:

- fitted to follow the development of a population;
- age structure and impact factors can easily be considered;
- easy to understand, interpret, and develop;
- most often based on causality.

The disadvantages are as follows:

- the conservation principles are sometimes not applied;
- limited to population dynamics;
- require a good database;
- difficult in some situations to calibrate;
- require a relatively homogeneous database.

This model type is typically used to keep a track on the development of a population. The most applied unit is the number of individuals or amount of biomass. Effects of toxic substances on the development of populations can easily be covered by increasing the mortality and decreasing the growth correspondingly. The model type is extensively used in fisheries management, biological opinions of threatened and endangered species, and national parks.

Structurally Dynamic Models

In these models, the parameters, corresponding to the properties of the biological modeling components, change over time to account for adaptation and changes in species composition. It is possible either to use knowledge or artificial intelligence to describe the changes in the parameters. Used most often, however, is a goal function to find the changes of the parameters. Eco-exergy has often been used as goal function in SDMs. By minor changes of the parameters it may be due to adaptation to the changed conditions, but for major changes it is most probably a change in the state variables (i.e., a shift in the species composition), that is causing the changed parameters. It may also be possible to use the approach to a major change in the ecological network, although no reference to this application of the structurally dynamic modeling approach is yet available.

This model type has the following advantages:

- able to account for adaptation;
- able to account for shift in species composition;
- can be used to model biodiversity and ecological niches;
- the parameters determined by the goal functions (a) do not need to be calibrated and (b) are relatively easy to develop and interpret.

The disadvantages of this model type are as follows:

- selection of a goal function needed;
- usually computer time consuming;
- information about structural changes required;
- no available software, programming often needed.

This model type should be applied whenever it is known that structural changes take place. It is also recommended for models that are used in environmental management to make prognoses resulting from major changes in the forcing functions (impacts).

Fuzzy Models

Fuzzy models may either be knowledge based (called the Mamdani type) or data based (called the Sugeno type). The Mamdani-type models are based on a set of linguistic expert formulations, and they are applied when no data are available. The Sugemo type applies an optimization procedure and it is applied when only uncertain data are available.

This model type has the following advantages:

- can be applied on a fuzzy data set;
- can be applied on semiquantitative (linguistic formulations) information;
- can be applied for development of models, where a semiquantitative assessment is sufficient.

This model type has the following disadvantages:

- can hardly be used for more complex model formulations;
- cannot be used where numeric indications are needed;
- fuzzy models based on data are black box models;

This model type should be applied when the data set is fuzzy or only semiquantitative expert knowledge is available, provided of course that the semiquantitative results are sufficient for the ecological description or the environmental management.

Artificial Neural Networks

These types of models can give relationships between state variables and forcing functions based on a heterogeneous database. It is a black box model and is therefore not based on causality; but it gives in most cases very useful models, that can be applied for prognoses, provided that the model has been based on a sufficient big database that allows to find the relationships and to test it afterward on an independent data set.

ANNs including self-organizing maps have the following advantages:

- may be used where other methods must give up;
- easy to apply;
- give a good indication of the certainty due to the application of a test set;
- can be used on a heterogeneous data set;
- give a near-optimum use of the data set.

The disadvantages can be summarized in the following points:

- no causality unless algorithms are introduced or a hybrid between ANN and a normal model is applied;
- cannot replace biogeochemical models based on the conservation principles;
- the accuracy of predictions is sometimes limited.

The advantages and disadvantages of this model type indicate where it would be advantageous to apply ANN, namely where ecological descriptions and understandings are required on the basis of a heterogeneous database, for instance data from several different ecosystems of the same type. It is also often applied beneficially when the database is more homogeneous, for instance, when the focus is on a specific ecosystems, although the modeler should seriously consider to use biogeochemical dynamic models due to their causality. ANN is, however, faster to use and the time-consuming calibration that is needed for biogeochemical models is not needed.

Individual-Based Models (IBMs) and Cellular Automata

This model type can be regarded as a reductionistic approach, as the system-level properties emerge from the interactions of low-level agents (individuals) in the model. Deriving the properties of a system from the properties and interactions among elements of the system. The model type was developed to focus on exploring the theories that ecosystem properties emerge because of the interactions among the individuals (either intra- or inter-specific) within the system. Within this model type, each individual can have different properties, for instance, individuals may have different sizes, which gives a different combination of properties as it is known from the allometric principles. The right combination may be decisive for growth and/or survival in certain situations, as it is known by all modelers. Consequently, a model without the differences among individuals may give a different result.

Cellular automata are systems of cells interacting in a simple way but displaying complex overall behavior. They are usually characterized by a few salient features. Cellular automata form a class of spatio-dynamical models where time, space, and states are discrete. The genesis of IBMs came from the cellular automaton approach, (although there are IBMs that are not cellular automaton models, and cellular automaton models that are not IBMs).

Advantages of this type of model are as follows:

- able to account for individuality;
- able to account for adaptation within the spectrum of properties;
- software is available, although the choice is more limited than for biogeochemical dynamic models;
- spatial distribution can be covered.

The disadvantages are as follows:

- if many properties are considered, the models get very complex;
- requires a large amount of data to calibrate and validate the models.
- Models require rigorous evaluation and can be difficult to communicate to non-modelers

As mentioned under the characteristics above, we know that the individuals have different properties and that may sometimes be crucial for the model results. In such cases, the IBMs are absolutely needed and the cellular automata can often be considered a proper ecological modeling approach.

Spatial Models

As the individual differences may be crucial for the model results, the spatial differences of the forcing functions, nonbiological state variables, and biological state variables may be decisive for the model results, too. Furthermore, it may be required to obtain model results that reveal the spatial differences, because they may be needed to understand the ecological reactions or to make a proper environmental management. Models that give the spatial differences must of course also consider the spatial differences in the processes, forcing functions, and state variables. It can therefore be concluded that there is an urgent need for inclusion of the spatial differences in ecological models.

There are a number of possibilities to cover the spatial differences in the development of an ecological model. It is not possible to cover them all; but as mentioned under IBMs, cellular automata may be used in this context. Geographic information system (GIS) is another possible approach that can also be considered a convenient method to present the model results. For aquatic ecosystems, the ultimate spatial model would give a three-dimensional (3-D) description of the processes, forcing functions, and state variables and is often a question about a good description of the hydrodynamics.

Spatial models offer the following advantages:

- cover spatial distribution, that is often of importance in ecology;
- the results can be presented in many informative ways, for instance, GIS.

The disadvantages are as follows:

- require usually a huge database, giving information about the spatial distribution;
- calibration and validation are difficult and time consuming;
- a very complex model is usually needed to give a proper description of the spatial patterns.

The spatial models are applied whenever it is required that the results include the spatial distribution, because it is decisive or the spatial distribution is crucial for the model results.

Ecotoxicological Models

Ecotoxicological models are in principle not representing a model type, as biogeochemical models or population dynamic models are applied widely in ecotoxicology. It is, however, preferable to treat ecotoxicological models as a separate model type, because

1. Our knowledge of the parameters is limited and estimation methods are therefore needed and have been developed.
2. Due to the use of safety factors and the limited knowledge of the parameters, ecotoxicological models are often simple. In particular, the so-called fugacity models illustrate this feature.
3. They include often an effect component.

The advantages of this model type are as follows:

- It is tailored to ecotoxicological problems.
- It is in most cases simple to use.
- It includes often an effect component or can easily be interpreted to quantify the effect.

The disadvantages are as follows:

- The number of parameters needed to develop models for all toxic substances is very high and we know only at the most 1% of these parameters.
- It implies that we need estimation methods that inevitably have a high uncertainty. The model results have therefore also a high uncertainty.
- Inclusion of an effect component requires knowledge of the effect, which is also limited.

The area of application is in this case obvious: to solve ecotoxicological research and management problems and perform environmental risk assessment for the application of chemicals.

Stochastic Models

This model type is characterized by an element of randomness. The randomness could be the forcing functions, particularly the climatic forcing functions, or it could be the model parameters. The randomness is in both cases caused by a limitation in our

Table 3 Model types applicable to different data sets

Description of data set	*Model type recommended*
High quality, homogeneous	Biogeochemical/bioenergetic dynamic models; or population dynamic models
High quality, heterogeneous	ANN
Only typical or average values	Static models
Uncertain data	Fuzzy models
No data, only rules	Fuzzy models, rule-based models
Important to utilize data set	ANN

knowledge. We can for instance not know the temperature the fifth of May next year at a given location, but we know how the normal distribution of the temperature has been for instance the last hundred years and can use the normal distribution to represent the temperature on this date. Similarly, many of the parameters in our models are dependent on random forcing functions or on factors that we hardly can include in our model without making it too complex. A normal distribution of these parameters is known and by use of Monte Carlo simulations based on this knowledge, it is possible to consider the randomness. By running the model several times, it becomes possible to obtain the uncertainty of the model results. A stochastic model may be a biogeochemical/bioenergetic model, a spatial model, a structural dynamic model, an IBM, or a population dynamic model. There are no differences among these model types on how a model can be made a stochastic model.

This model type has the following advantages:

- able to consider the randomness of forcing functions or processes;
- the uncertainty of the model results is easily obtained by running the model many times.

This model type has the following disadvantages:

- it is necessary to know the distribution of the random model elements;
- has a high complexity and requires much computer time.

It is recommended to apply stochastic models whenever the randomness of forcing functions or processes is significant.

Integrated/Hybrid Models

Hybrid, or integrated, models can in principle be developed by any combination of two of the previously listed ten model types. The underlying concept for integrated modeling is that combining models from different disciplines captures the strengths of each model within the modeling suite (e.g., integrating a hydrodynamic model developed by a hydro-dynamicist and a population dynamics model developed by an ecologist). Similarly, integrating models reduces the need to develop new models for every situation. Model integration is a considerable undertaking. Common issues include: discipline-specific approaches for dealing with spatio-temporal dynamics, underlying assumptions of each model may be incompatible, communicating among the disciplines can take time, input/output requirements are must be designed a priori.

Which Model Types are Recommended to Solve Which Problems?

As mentioned in the introduction, new model types were provoked by the model problems that became clear in the early to mid-1970s when ecological modeling started to be applied more extensively as a tool in environmental management. Biogeochemical/bioenergetic dynamic models and population dynamic models had some shortcomings which the ecological modelers have tried to solve for the last 30 years by development of new model types. Today, the shortcomings have been at least partially eliminated. Further improvements will be possible by development of hybrid models or even new model types, but a solution to the problem formulated in the introduction is available today. Consequently, it is possible to indicate which model type is the best choice in a given model situation, which is defined by the data, and the combination of problem and system. It is possible to indicate with the 11 model types in hand which solution should be used when we know the (1) the available data sets and (2) the combination of problem and system. "Which model type should be applied in which context?" is answered in **Tables 3** and **4**, covering respectively the different data sets and different problem/system combinations.

Which Environmental Problems Have Been Modeled?

Table 5 gives an overview of the environmental problems that have been modeled. Notice that many models have been focusing on a problem for a specific ecosystem and are therefore included in both table overviews. There are of course also models that focus on a specific problem without considering the ecosystem, for instance, fishery, the greenhouse effect, and acid rain, and

Table 4 Model type applicable to different problems/systems

Problem/system	*Model type to be recommended*
Distribution of matter or energy[a]	Biogeochemical/bioenergetic dynamic models
Development of populations[a]	Population dynamic models
Toxic compounds: distribution and effect	Ecotoxicological models
Individuality important	IBM
Structural changes may occur	Structurally dynamic models
Forcing functions and/or processes stochastic	Stochastic models
Spatial distribution important	Spatial models

[a]ANN could also be applied if causality is less important and optimum data utilization is important. Hybrid models may be applied to combine causality with good data utilization.

Table 5 Model effort, classification environmental problem

Environmental problem	*Modeling effort*
Oxygen balance and depletion	5
Eutrophication	5
Heavy metal pollution	4
Pesticide pollution	4
Other toxic substances, including the use of models for environmental risk assessment	5
Regional distribution of pollutants	5
Protection of national parks	4
Endangered species	3
Groundwater pollution	5
Green house effects, global warming	5
Acid rain	4
Microclimate changes	3
Ecosystem health assessment	5
Decomposition of ozone layer	4
Health risk assessment	4
Fishery	5
Timber	5
Non-point pollution from agriculture	5
Optimization of an environmental management strategy	3

models that model a specific ecosystem without considering a specific problem, for instance, element cycling in an arctic ecosystem or in a wetland.

See also: Conservation Ecology: Biodiversity Indices. Ecological Complexity: Cellular Automata; Ecological Informatics: Overview; Complex Ecological Networks; Complex Systems; Systems Ecology; Systems Ecology: Ecological Network Analysis. General Ecology: Growth Models; Age-Class Models

Further Reading

Borsuk, M.E., Reichert, P., Peter, A., Schager, E., Burkhardt-Holm, P., Assessing the decline of brown trout (*Salmo trutta*) in Swiss rivers using a Bayesian probability network Ecological Modeling 192 2006 224–244.
Grimm, V., Railsback, S.F., 2013. Individual-based modeling and ecology. Princeton: Princeton university press.
Haining, R.P., 2003. Spatial data analysis: Theory and practice. Cambridge: Cambridge University Press.
Jørgensen, S.E., 1986. Structural dynamic models. Ecological Modeling 31, 1–9.
Jørgensen., S.E., Integration of ecosystem theories: A pattern 3rd edn. Vol. 2002 Kluwer Academic Publishers Dordrecht, The Netherlands 428 pp.
Jørgensen S. E., and B. D. Fath. 2011. Fundamentals of ecological modeling 4th edn. 2001 Elsevier Amsterdam 628 pp.
Radtke, E., Straskraba, M., Self-organizing optimization Ecological Modeling 9 1980 247–268.
Recknagel, F., 1997. ANN predicting the occurrence of blue-green algae. Hydrobiologia 349, 47–57.

Salski, A., Holsten, B., 2006. A fuzzy and neuro-fuzzy approach to modeling cattle grazing on pastures with low stocking rates in Middle Europe. Ecological Informatics 1 (3), 269–276.

Christensen V and Pauly D (eds.) (1993) Trophic models of aquatic ecosystems. In: *ICLARM Conference Proceedings*, 26, p. 390, Manila: ICLARM/International Council for the Exploration of Sea/Danida.

Yager, R.R., Filev, D.P., 1994. Essentials of fuzzy modeling and control. New York: John Wiley & Sons, p. 388.

Yue, T.X., Wang, Y.A., Liu, J.Y., *et al*., 2005. Surface modeling of human population distribution in China. Ecological Modeling 181 (4), 461–478.

Yue, T.X., Liu, J.Y., Jørgensen, S.E., Ye. Q.H., Landscape change detection of the newly created wetland in Yellow River Delta Ecological Modeling 164 2003 21–31.

Modeling Dispersal Processes for Ecological Systems

Adam Duarte, Oregon State University, Corvallis, OR, United States
Ivana Mali, Eastern New Mexico University, Portales, NM, United States

Glossary

Breeding dispersal The unidirectional movement of an individual from one breeding site to an alternate breeding site.
Census Complete count of the number of individuals or species within a sampling unit.
Detection probability (p) The probability a species is detected, given the sampling unit is occupied by the species.
Emigration The process by which an individual disperses out of a population.
Immigration The process by which an individual disperses into a population.
Natal dispersal The unidirectional movement of an individual from its natal site to the area where it breeds for the first time.
Occupancy probability (ψ) The probability a sampling unit contains at least one individual of a species.

Introduction

Dispersal is a vital life-history strategy that has implications for gene flow, resource competition, population dynamics, and the distribution of species (Clobert *et al.*, 2001). Despite its importance, dispersal is also one of the hardest parameters to estimate. Therefore, dispersal processes are often poorly understood, particularly for rare and elusive species. Also, while dispersal has a clear definition within an ecological context (i.e., the unidirectional movement of an individual), dispersal modeling is less definitive because it is largely grounded on the spatial and temporal dimensions of data and the model that is simulating the process. In light of this uncertainty, we dedicated this article to the introduction of some of the commonly used approaches to estimate movement at various spatial resolutions and follow the structure of Kendall and Nichols (2004) by arranging estimation procedures into two broad categories: direct and indirect methods. We also offer a brief overview on how this information can be integrated to model dispersal processes.

Direct Methods

Direct methods to estimate dispersal require the collection of data on individual movements. Such data can arise through the use of mesocosm studies, behavioral studies that observe a focal animal over a period of time, the tracking of a focal animal over time with the aid of technology, and through the individual marking and subsequent recapture/resighting of marked individuals. Each of these approaches comes with a set of constraints (i.e., cost, sample size limitations, etc.) and provides a different resolution of information concerning dispersal processes.

Mesocosms

Mesocosms are controlled experiments that take place in natural environments. Due to logistical constraints, mesocosm experiments centered on mechanisms of dispersal are usually implemented for smaller-sized animals (i.e., invertebrates, fish, amphibians, and reptiles). Despite such constraints, mesocosms can provide the highest resolution dispersal information; for example, describing natal dispersal of juvenile amphibians from ponds into different habitat types, testing the evacuation hypothesis of individuals from disturbed to more suitable habitats, examining how different substrate types and weather patterns affect movement behavior, assessing how presence of predators affect prey dispersal, and documenting the effects of intraspecific competition on movement. Mesocosm studies often involve a physical barrier that outlines the study area beyond which an organism cannot disperse (i.e., an enclosure; **Fig. 1**). Therefore, it is important to define the appropriate scale at which animals are to be monitored and the scale of treatments. Additional barriers are typically constructed between the treatments either with corridors that accommodate dispersal or with passive traps that capture individuals attempting to disperse (i.e., pitfall traps).

The highest resolution data is obtained by continuously monitoring corridors through which individuals move. This can be achieved by using game cameras and other devices such as stationary chip readers that recognize uniquely marked individuals. In addition to evidence of movement, downloaded data provide the date and time an individual moved into a different spatial subunit. Ideally, the system should only consist of marked/known individuals and be able to record movement events with absolute certainty, where detection probability is equal to 1. Generalized linear models with different error distributions (i.e., Poisson, normal, binomial, etc.) can be fit to these data to link explanatory variables to the number of movement events

 doi:10.1016/B978-0-12-409548-9.10573-1

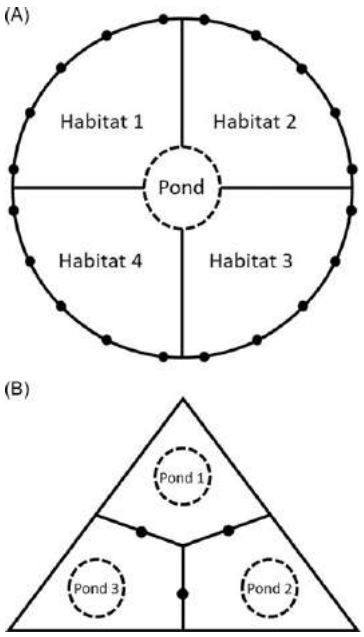


Fig. 1 Two mesocosm experimental designs where *dashed lines* represent ponds, *solid lines* represent barriers to dispersal (i.e., drift or horse fences) and *filled dots* along the barriers present either trap placements (i.e., pitfall traps) or dispersal corridors with some type of monitoring device (i.e., a game camera or chip reader). For example, design A can be used to monitor dispersal of juvenile amphibians into different habitat types while design B can be used to study interpond movement of semiaquatic organisms.

(i.e., count data), a single dispersal event or the chosen terrain to move on during dispersal (i.e., binary [0 or 1] data), the distance moved (i.e., continuous data), etc. In the case where data are collected from the same set of individuals across time, generalized linear mixed models should be fit, where the individual animal identification is treated as a random effect to accommodate potentially nonindependent movement data. That is, an individual is more likely to move the same as itself than it is to share the same behavioral response across multiple individuals, and the analysis should incorporate the hierarchical structure in the data.

Focal Animal Sampling

Focal animal sampling might be one of the oldest methods to monitor movement and consists of direct observations of an individual over a set period of time. Usually if multiple organisms are being observed within a group, each individual is observed during separate time intervals so that the observer is focused on a single animal at a time. However, if an organism is gregarious (i.e., dolphins, elk, etc.), focal-group surveys are conducted where the same set of animals is observed until the end of the survey. A set of behavioral categories to be monitored is usually preestablished. For example, the focus of the study might be distance of focal individual movement between food sites, movement directionality, whether an individual moves independently or as part of a group, or simply total distance traveled.

With technological advancements, research has greatly shifted toward radio telemetry and global positioning system (GPS) devices to monitor focal animals (reviewed in Millspaugh and Marzluff, 2001). Very-high frequency (VHF) radio tracking has become a conventional method to monitor individual movements and requires ecologists acquire a signal via hand-held antenna from a transmitter attached to an animal. Locating marked individuals is done via triangulation (**Fig. 2**), with ecologists traveling either on ground (i.e., afoot, truck, all-terrain vehicle, etc.) or using aerial vehicles (i.e., helicopter, fixed-winged aircraft, etc.). Importantly, aerial vehicles can expand the signal horizon. More advanced GPS tracking technology relies on GPS orbiting satellites to detect signals emitted from a GPS receiver attached to an animal. GPS units can be programmed to record an animal's location at specific time intervals, and data are retrieved either by recapturing the animal to regain access to the GPS receiver or by remote download to a portable hand-held receiver. Finally, satellite-tracking telemetry relies on polar-orbiting Argos satellites to receive signals from platform transmitter terminals (PTTs), which receive the data from GPS satellites. These terminals are located on drifting buoys, ships, or fixed land locations. Argos satellite data are uploaded to several processing centers (i.e., Toulouse, France and Landover, United States of America), which distribute the data worldwide. Therefore, GPS and satellite telemetry are often combined, where an animal is equipped with a GPS/satellite device and the data are transmitted to Argos satellite every few days. While GPS technologies increase the resolution of data and decrease labor requirements, they can substantially increase equipment costs. Ecologists also have to take into account the weight of the transmitter and battery life. In particular, it is recommended that the transmitter is less than 5%–10% of an animal's body mass. Data obtained from focal animal sampling can be used to examine movement rates, migration patterns, and dispersal or space use patterns. Home range can be determined using minimum convex polygon (MCP), bivariate normal probabilistic model, harmonic mean, or kernel density model.

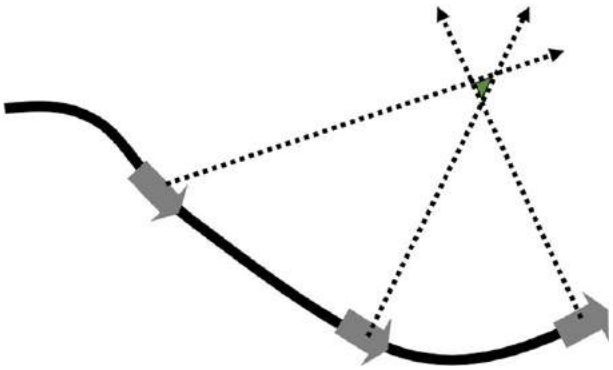


Fig. 2 An example of locating a radio-collared animal via triangulation where the *solid line* represents a surveyor's path, *gray arrows* represent points along the path where a surveyor detected a signal via hand-held antenna, and *dashed lines* represent the direction (i.e., bearing) of the signal from the surveyor's location. The space between where the directional signals intersect, *green triangle*, is the approximate location of the animal.

High-resolution data obtained from GPS devices allow for more sophisticated statistical methods of analyzing daily movement paths and understanding the relationship between movement patterns and environmental factors. First-passage time analysis uses high-resolution data from GPS devices to quantify the time it takes an animal to move through circles with radii that represent evenly spaced points along a movement trajectory (Fauchald and Tveraa, 2003). First-passage times reflect the search effort of an individual at each point. A high first-passage time reflects slower movement representing an encamped area, while a low first-passage time represents faster and straight line movements. To use this method, ecologists have to determine the scale at which the first-passage analysis should be conducted. The results of this technique are dependent on the size of the radius, as an increase in the size of the radius will likely lead to more windings of the pathway within each circle. This increase in tortuosity will probably be much larger in areas with more physical barriers to movement when compared to areas where the animals are able to move unidirectional. Thus, it is assumed that the relative variance of fist-passage times for all points along the path will increase with increasing radii. One approach to determine the appropriate radius (r) is to estimate first-passage times with a range of r values and calculate the variance of log-transformed first-passage times. The r that yields the highest variance represents the scale at which the analysis should probably be conducted.

Brownian bridge movement models (BBMMs) use discrete location data obtained at short time intervals from GPS devices to estimate movement path and model the utilization distribution (*UD*) of individuals (Horne *et al.*, 2007). The BBMM requires sequential location data, estimated error associated with location data, and grid-cell size assigned for the output *UD*. It is assumed that location errors correspond to a bivariate normal distribution and movement between successive locations is random but conditional on the starting and ending locations. The BBMM is dependent on the time between locations, the distance between locations, and the Brownian motion variance (σ_m^2). Brownian motion variance is related to the animal's mobility and represents the diffusiveness or irregularity of movement. This parameter is estimated from the trajectory, which is based on an average of all available movement data. While BBMMs assume homogeneous movement behavior across individual animals, dynamic Brownian bridge movement models (dBBMMs) relax this assumption by allowing σ_m^2 to vary along a path, which enables the estimator to accommodate behavioral heterogeneity in σ_m^2 (Kranstauber *et al.*, 2012). These models can be used to quantify circadian patterns of habitat use and correlated movement patterns among habitat types.

Multistate Models

Multistate models enable ecologists to estimate stochastic transitions between states (Schwarz, 2005). Here, states denote discrete classifications that describe the current condition of an animal. These states can represent an individual's reproductive status, sex, size, disease status, geographic location, occupied habitat type, and much more. When estimating movement parameters, geographic states are typically the states of interest; however, combinations of multiple types of states can also be modeled. Estimating transition probabilities among geographic states requires that animals are monitored in different sites or sample units. Although this model is typically fit to capture–recapture/resight data where recapture/resight probabilities are explicitly modeled to account for imperfect detection, it can also be fit to "known fate" data from VHF or GPS devices. This approach operates under a few assumptions. Specifically, individuals have homogeneous recapture/resighting and survival probabilities, marks are not lost or overlooked, sample occasions are instantaneous or occur over short time periods and individuals are released immediately, all emigration out of the sample units is permanent (but see Kendall and Nichols, 2002), the fate of individuals is independent from each other, states are recorded without error (but see Conn and Cooch, 2009), all individuals in state r at time t have equal movement probabilities, movement and recapture/resight probabilities are only related to the current state and not related to the past location of the individual (but see Hestbeck *et al.*, 1991), and all individuals transition at the same time. It is also worth noting that if permanent emigration out of the sample units is prevalent the survival parameter represents apparent survival, not

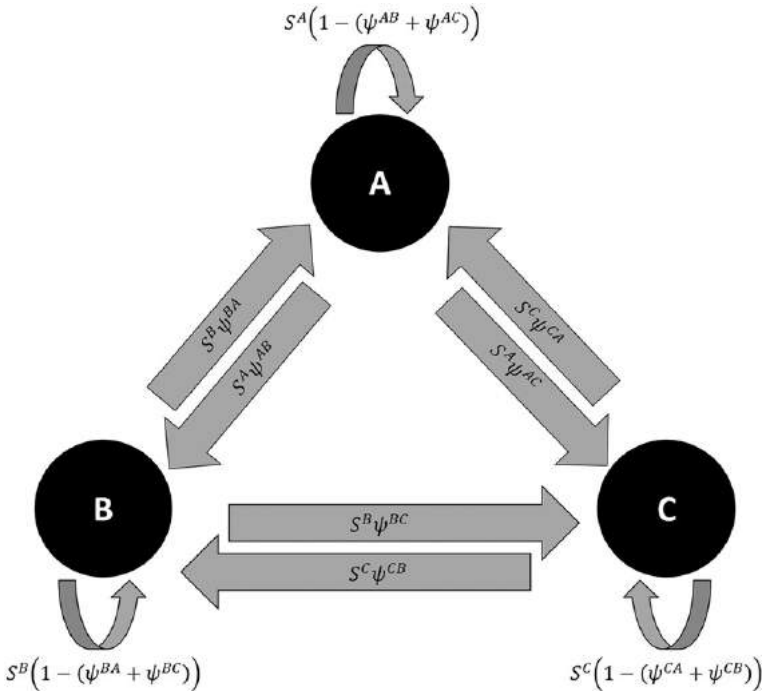


Fig. 3 Graphical depiction of the state transitions for a multistate model with only 3 states. Note that the model also estimates the probability of recapture/resight in each state.

true survival. Apparent survival is the probability that an individual survives and does not permanently emigrate out of the sample units from one occasion to the next.

Here, we focus on a simple example where individuals are sampled on 3 occasions and classified into 1 of 3 states on each occasion, representing sample units A, B, and C (**Fig. 3**). Note that an individual can only belong to 1 state at a time. Let us say you captured and uniquely marked an individual in sample unit C on the first occasion, you reencountered the same individual in sample unit A during the second occasion, and you also reencountered the same individual during the third occasion but the individual was in sample unit B. That individual would have an encounter history of "C, A, B", and the probability statement for this encounter history is $\Pr(h = \text{CAB}) = S^{\text{C}}\psi^{\text{CA}}p^{\text{A}}S^{\text{A}}\psi^{\text{AB}}p^{\text{B}}$, where S^{C} is the probability of surviving through the interval between the first and second occasion, ψ^{CA} is the probability of transitioning from state C to state A, p^{A} is the probability of recapturing/resighting the individual at sample unit A, S^{A} is the probability of surviving through the interval between the second and third occasion, ψ^{AB} is the probability of transitioning from state A to state B, and p^{B} is the probability of recapturing/resighting the individual at sample unit B. What if an individual is captured and uniquely marked on the first occasion in sample unit B, not encountered during the second occasion, and reencountered in sample unit A during the third occasion? Well, the encounter history for that individual is "B, 0, A", and the probability statement for this encounter history is $\Pr(h = \text{B0A}) = S^{\text{B}}(1-(\psi^{\text{BA}}+\psi^{\text{BC}}))(1-p^{\text{B}})S^{\text{B}}\psi^{\text{BA}}p^{\text{A}}+S^{\text{B}}\psi^{\text{BA}}(1-p^{\text{A}})S^{\text{A}}(1-(\psi^{\text{AB}}+\psi^{\text{AC}}))p^{\text{A}}+S^{\text{B}}\psi^{\text{BC}}(1-p^{\text{C}})S^{\text{C}}\psi^{\text{CA}}p^{\text{A}}$, where the terms are defined above. We can see how the complexity of the probability statement for the encounter history increases dramatically when individuals are not encountered on every occasion because we have to consider each potential pathway when going from state B in occasion 1 to state A in occasion 3 (see **Fig. 3**). We can learn a few things by looking at these probability statements. First, we see that the survival between occasions is dependent on the conditions the individual experiences at the state prior to the gap between occasions. That is, if an individual transitions from state A to state B the survival parameter is linked to the location the individual came from (i.e., state A) and is not related to the location the individual travels to (i.e., state B). The individual has to survive and then decide to move or not. Second, we see that the transition probabilities must sum to 1 with respect to each state. That is, if the probability of transitioning from state A to B (ψ^{AB}) is 0.45 and the probability of transitioning from state A to C (ψ^{AC}) is 0.15, the probability of remaining in state A (ψ^{AA}) is $1-(0.45+0.15)$ or 0.40. Thus, in practice we typically estimate the last transition using subtraction. Each of the probability parameters (i.e., S, ψ, and p) can be related to explanatory variables using separate logistic regressions to better understand what factors are related to survival, movement, and recapture/resight probability. One thing to consider, however, is that the number of estimated transition probabilities increases substantially with an increasing number of potential states. Specifically, the number of transitions to be estimated is the number of states squared. So, even when only considering 3 states there are 9 transition probabilities to be estimated. Therefore, this model requires a relatively large amount of data and for the data to contain observations of individuals that move between the states to estimate all of the parameters accurately. If data are robust enough to support fitting these models, multistate models are extremely versatile and have been applied across disciplines.

The above approach enables ecologists to estimate the probability of individuals moving between states or locations. However, it may be of interest to use a 2-step approach in order to estimate the probability of leaving a sample unit (π) and the probability of settling in a recipient sample unit given the individual survived and left the sample unit it was located in during the previous occasion (μ). In fact, Grosbois and Tavecchia (2003) proposed the "$\pi\mu$-parameterization" of multistate models to do just that. Using the example from above, we can rewrite the probability statements for the respective encounter history of "C, A, B" and "B, 0, A" as $\Pr(h = \text{CAB}) = S^{C}\pi^{C}\mu^{CA}p^{A}S^{A}\pi^{A}\mu^{AB}p^{B}$ and $\Pr(h = \text{B0A}) = S^{B}(1 - \pi^{B})(1 - p^{B})S^{B}\pi^{B}\mu^{BA}p^{A} + S^{B}\pi^{B}\mu^{BA}(1 - p^{A})S^{A}(1 - \pi^{A})p^{A} + S^{B}\pi^{B}(1 - \mu^{BA})(1 - p^{C})S^{C}\pi^{C}\mu^{CA}p^{A}$, where the terms are defined above. We see that the probability statements are virtually identical and survival is still dependent on the conditions the individual experiences at the state prior to the gap between the occasions. The difference is that the ψ parameter is replaced by π and μ (or just π when the individual does not emigrate), and the probability of leaving a sample unit and settling in another sample unit can each be related to different explanatory variables using separate logistic regressions. Since π and μ are both probabilities, they are each constrained to sum to 1. That is, π is the probability of emigrating from a given sample unit and $1 - \pi$ is its complement (i.e., site fidelity). Since there are only 3 sample units in this example, if an individual transitions from state B to state A the probability is μ^{BA}, and if an individual transitions from state B to state C the probability is $1 - \mu^{BA}$. Grosbois and Tavecchia (2003) noted that if π and μ are interpreted at the population scale, instead of individual scale, they represent emigration and immigration probabilities, respectively. Importantly, the $\pi\mu$-parameterization of multistate models contains the same set of assumptions as the original ψ-parameterization described above.

Spatial Capture–Recapture Models

Spatial capture–recapture models are a type of capture–recapture model that explicitly model the movement of individuals. Notably, the original focus of spatial capture–recapture models was to derive more accurate density estimates for animal populations by incorporating spatially explicit individual encounter history data and estimating the number of individual activity centers within an area of interest (reviewed in Royle *et al.*, 2014). Ergon and Gardner (2014) extended spatial capture–recapture models to incorporate capture–recapture data collected within fixed trapping grids using the robust design (**Fig. 4**). The robust-design spatial capture–recapture model treats individual activity centers during the first season as independent and uniformly distributed. During the following season, however, change in an individual's activity center is modeled using trigonometric functions that assume dispersal occurs in a random direction and at a random distance. Using the model they developed, ecologists can estimate dispersal distances. Moreover, the estimated survival parameters are more biologically meaningful because the approach allows for better separation of emigration and survival—see the above section concerning apparent survival. Schaub and Royle (2014) approached the apparent survival "problem" differently, but still used concepts from the spatial capture–recapture literature. Specifically, they developed a spatial generalization of the Cormack–Jolly–Seber (CJS) model, which allows for the joint estimation of dispersal distance, emigration rate, and survival probability. Importantly, the spatial CJS model does not require fixed trapping locations or data to be sampled using the robust design to be fit. This approach treats dispersal as a random walk and models dispersal variances using a normal distribution, which can then be used to estimate dispersal distances. It should be noted that, like all of the other techniques introduced in this article, both of these approaches require the sample space to be appropriate in order to examine dispersal for a specific study species.

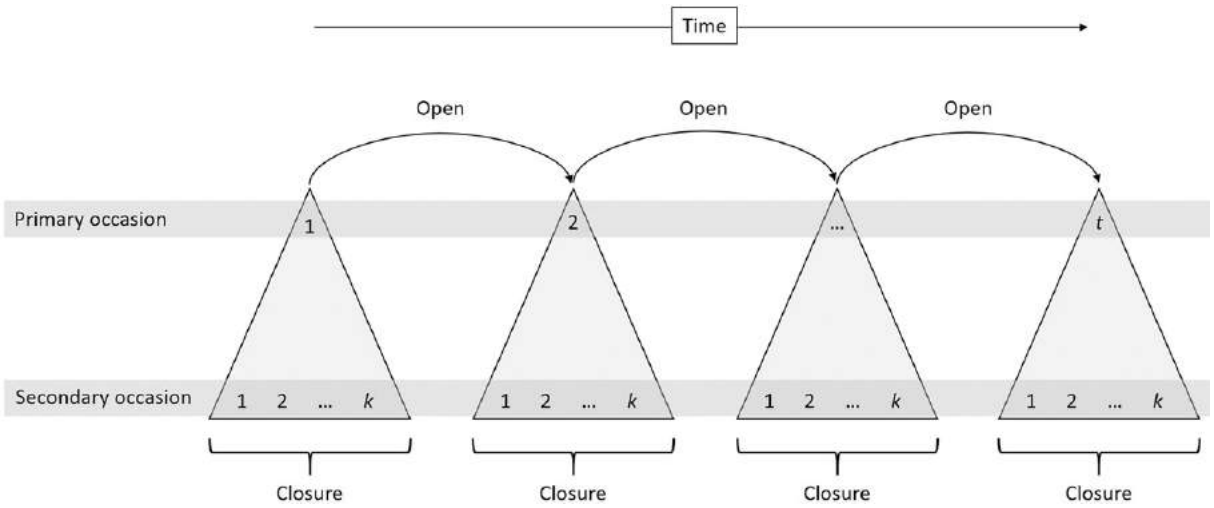


Fig. 4 Illustration of the robust-design sample design. Note that the population is closed (i.e., no loss or addition of individuals) across secondary occasions within a primary occasion, but is open between primary occasions.

Indirect Methods

Indirect methods to estimate dispersal parameters do not require direct data on individuals. Although appealing from a logistical perspective, indirect methods typically do not provide the same resolution of information concerning dispersal processes. Nevertheless, given the logistical constraints associated with most monitoring programs, indirect methods comprise some of the most commonly applied approaches in the field. Moreover, as technology and analytical techniques are further developed and refined, these indirect methods are becoming more useful to understand and model dispersal processes.

Correlated Abundances

Examining for time-lagged correlations in abundance of multiple populations is a method to indirectly study dispersal at large spatial scales (Tittler *et al.*, 2006). If, for example, populations are undergoing source–sink dynamics or asymmetrical dispersal among populations, we can expect abundances to be related across populations over time. That is, if abundance is high in a source population in a given year, we would expect an increase in abundance in the sink population during a subsequent year. Importantly, the time lag in correlations (if present) will vary based on the life history of the study species. The distances over which correlations in abundances occur can be interpreted as the distances over which dispersal occurs. This is perhaps the simplest approach to indirectly study dispersal and, as expected, the information it provides is also coarser than other approaches. Moreover, although simple from a statistical analysis perspective, this approach requires intensive monitoring of multiple populations to reliably estimate abundances and that those data are collected over an extended period of time.

Spatial Dynamic Occupancy Models

Spatial dynamic occupancy models are another indirect method to quantify dispersal. Briefly, dynamic occupancy models leverage replicated detection/nondetection survey data that are collected across seasons using the robust design, where closure within a primary occasion means the occupancy state of a sample unit does not change. The model can estimate the initial occupancy probability (ψ) in the first season of monitoring, detection probability (p) each season, and the occupancy state in subsequent seasons as a result of colonization probability (γ) and local extinction probability (ϵ) as follows:

$$z_{i,1} \sim Bernoulli(\psi_i)$$

$$z_{i,t+1} \sim Bernoulli\left(z_{i,t}\left(1-\varepsilon_{i,t}\right)+\left(1-z_{i,t}\right)\gamma_{i,t}\right) \text{ and}$$

$$\gamma_{i,k,t}|z_{i,t} \sim Bernoulli\left(z_{i,t}p_{i,k,t}\right)$$

where y represents the detection/nondetection data for each site i, during survey k, within season t (i.e., detected [1] or not detected [0]), and z is the latent (partially observed) true occupancy state. Notably, this method accounts for false-negative detections (i.e., imperfect detection) by explicitly modeling the detection process. The probability parameters are linked to potential explanatory variables using separate logistic regressions. In spatial dynamic occupancy models, colonization probability is typically linked to ecological dispersal processes. For example, Risk *et al.* (2011) modeled colonization as:

$$\gamma_{i,t} = \frac{S_{i,t}^2}{S_{i,t}^2 + \delta^2}$$

where δ^2 is a parameter creating a sigmoid-shaped colonization probability function and S is a connectivity metric, which is:

$$S_{i,t} = \sum_{j \neq i}^{m} z_{j,t-1} A_j^{\beta} e^{-\alpha d_{i,j}}$$

where connectivity is related to the area (A) of neighboring occupied sites (i.e., $z_{j,\ t-1}A_j^{\beta}$) and is inversely related to intersite distance between the focal site and its neighbors (i.e., $e^{-\alpha d_{i,\ j}}$). Their approach assumes larger sites have a greater population size and thus contribute more to connectivity. Sutherland *et al.* (2014) used this same rationale, except they calculated colonization and connectivity as:

$$\gamma_{i,t} = 1 - e^{-S_{i,t}}$$

$$S_{i,t} = \delta \sum_{i \neq j} N_{j,t-1}^{*} e^{-\alpha d_{i,j}}$$

where δ is the population-level per capita effective dispersal rate and they replaced $z_{j,\ t-1}A_j^{\beta}$ with the actual juvenile abundances of neighboring sites in the previous season ($N_{j,\ t-1}^{*}$). Chandler *et al.* (2015) approached the problem differently in that colonization was solely related to intersite distance. Specifically, they modeled the probability site i is colonized by at least 1 individual from neighbor j as:

$$\rho_{i,j,t} = z_{j,t-1}\rho_0 e^{-d_{i,j}^2 / 2\sigma^2}$$

where ρ_0 is the baseline colonization probability and σ is the rate of decay in the colonization probability as a function of intersite distance ($d_{i,j}$). This approach assumes a Gaussian kernel for dispersal. The probability that a site is colonized by at least 1 neighbor is then:

$$\gamma_{i,t} = 1 - \left\{ \prod_{j=1}^{M} 1 - \rho_{i,j,t} \right\}$$

where all terms are defined above. Broms *et al.* (2016) modified the base structure of a dynamic occupancy model when developing a spatial dynamic occupancy model. Specifically, their approach models site occupancy after the first season as:

$$z_{i,t+1} \sim Bernoulli\left(z_{i,t}\left(1-\varepsilon_{i,t}\right) + \left(1-z_{i,t}\right) I_{N_{i,t}} \overline{d}_{i,t} + \left(1-z_{i,t}\right)\left(1-I_{N_{i,t}}\right)\gamma_{i,t}\right)$$

where $I_{N_{i,t}}$ indicates whether at least 1 neighbor is occupied (1) or not (0) the previous season and the neighborhood size is specified by the user. Thus, if a site is not occupied it can be colonized by long-distance colonization (γ) if its neighbors are not occupied or by a local neighbor colonization ($\overline{d}$) if at least 1 neighbor is occupied. Neighbor colonization is then modeled as:

$$\overline{d}_{i,t} = 1 - e^{\hat{z}_{N_{i,t}} \log\left(1-d_i\right)}$$

where $\hat{z}_{N_{i,t}}$ indicates if neighbor j is occupied and d_i is the probability site i will be colonized by neighbor j. The approach is flexible as to how d_i is modeled and is largely governed by the objective of the study. Each of these spatial dynamic occupancy models approaches the task of modeling dispersal among occupied and unoccupied sites differently, but we note that each model is parameterized such that if a neighbor is not occupied the previous season it does not contribute to the cumulative colonization probability of a focal site. We also note that each approach requires slightly different data in addition to the typical replicate detection/nondetection data needed to fit traditional dynamic occupancy models.

Integrated Population Models

Integrated population models are an approach to borrow strength across multiple datasets by combining their likelihoods into a single analysis to study population dynamics (reviewed in Schaub and Abadi, 2011). The integration of multiple, but related, datasets within a single analysis enables ecologists to incorporate and estimate complex ecological relationships because relationships among model parameters are constrained to agree with each other since data are fit simultaneously. For example, if data are available to estimate apparent survival, fecundity, and population change for a target population, ecologists can exploit the direct link between the parameters that govern temporal population dynamics (i.e., recruitment, survival, emigration, and immigration) to estimate immigration without having explicit data on movement. That is, if you know how many individuals enter the population through reproduction, how many individuals stay in the population through apparent survival parameters, and how the population abundance changes across time, you can estimate the missing parameter, immigration (**Fig. 5**). Importantly, this approach can also accommodate prior information when using a Bayesian framework if data deficiencies exist for survival or reproduction and the objective is to estimate immigration, albeit the hypotheses that can be evaluated become much more limited. Although these models require a relatively substantial amount of data to be fit, they overcome the logistical challenge associated with estimating dispersal parameters for species that are capable of traveling large distances because only a single target population needs to be monitored.

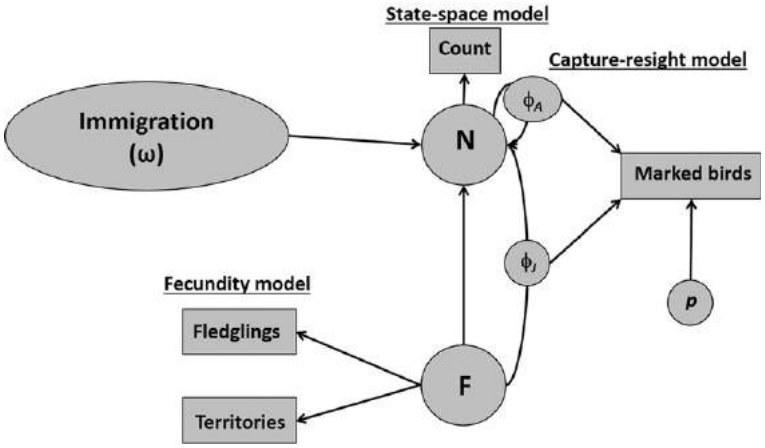


Fig. 5 Graphical depiction of the type of data that can be incorporated within an integrated population model to estimate immigration. Note that this figure depicts a monitoring program that might be implemented for a territorial avian species, but other submodels can be incorporated for different data types and different taxa.

Genetic and Isotope Analysis

Data on gene flow can also be used to infer dispersal. Ecologists can use allozyme, microsatellite, restriction fragment length polymorphism (RFLP), single-nucleotide polymorphism (SNP), and other genotype data to estimate gene flow among populations. There are at least 2 methods of estimating migration parameters. The coalescent approach uses genealogical information contained in deoxyribonucleic acid (DNA) sequences to estimate long-term migration parameters. These traditional indirect estimators of gene flow rely on the assumption that population sizes and migration rates are constant and that populations persist for a time period sufficient to achieve genetic equilibrium. The second approach uses multilocus genotype to estimate short-term migrations (Faubet *et al.*, 2007). This approach uses Bayesian methods to estimate the posterior probability distributions of allele frequencies, migrant proportions, and individual immigrant ancestries. This method also operates under fewer assumptions than the coalescent approach, allowing genotype frequencies to deviate from the Hardy–Weinberg equilibrium. Simulation experiments have shown that with sufficient differentiation among populations and sufficient number of loci, ecologists can derive accurate estimates of recent migration rates using multilocus genotypes (Wilson and Rannala, 2003).

The use of environmental DNA (eDNA) is a relatively novel approach to study dispersal. eDNA is DNA that is collected from the environment, such as soil and water samples, rather than directly from the study species. The presence/absence and even quantitative comparisons of eDNA from a study species in various habitats (i.e., different stretches of the rivers and lakes or freshwater vs. marine environments) can provide information concerning migration routes and potential barriers to movement (Erickson *et al.*, 2016).

The presence of naturally occurring stable isotopes in animal tissues (i.e., carbon [δ^{13}C], nitrogen [δ^{15}N], and deuterium [δD]) can also be used to trace migration (reviewed in Hobson, 1999). Briefly, this approach relies on differential concentration of stable isotopes in different food webs. For example, if differences between food webs exist and they are spatially clustered, organisms moving between geographic regions can carry information on the location of previous feeding. This information can be used to understand movements between different types of habitat (e.g., mesic and xeric, inshore and offshore, marine and freshwater, etc.). Moreover, δD levels in plants and organisms at higher trophic levels vary predictably across continents with growing season precipitation (Cormie *et al.*, 1994). Thus, this stable isotope can be used to study large-scale migration patterns.

Integrating Information to Model Dispersal

Despite the growing number of methods to study dispersal, estimating dispersal is difficult in even the best of cases. Thus, modeling dispersal often requires the integration of any and all available information. Importantly, the objectives of the project should govern what modeling approach is used and, by extension, the data and estimates that are required.

If the objective is to model dispersal corridors, for example, least-cost path (LCP) analysis is a useful modeling approach. LCP analysis uses a geographic information system (GIS) to predict animal movement across landscapes. It involves the use of multiple raster shapefiles that contain information for spatially explicit environmental factors hypothesized to influence movement (i.e., elevation, slope, habitat type, etc.). Combining these raster data using some weighted measure, the user develops a cost or friction surface raster layer, where an individual is more likely to traverse across a cell that contains a lower cell value. The cost surface layer can then be used to identify the most likely (i.e., the least costly) path between multiple locations. LCP analysis is best implemented in combination with the other methods described in previous sections of this article. For instance, information from mesocosms or focal animal sampling can be scaled up to landscape levels using LCP analysis. Also, LCP analysis has been used to explain genetic distances between populations, and ecologists can depict which landscape types are more or less costly for dispersal by identifying genetic immigrant ancestry in different populations (Wang *et al.*, 2009). If there is previous knowledge of established populations, ecologists can study how animals might have dispersed from established populations to newly colonized regions of suitable habitat (LaRue and Nielsen, 2008). Similarly, if Euclidean distance is not as biologically relevant, LCP analysis might also be implemented to develop the intersite distance matrix used when fitting spatial dynamic occupancy models.

When simulating population dynamics using spatially explicit population models, the available information concerning dispersal processes are useful in developing dispersal submodels. For example, Bonnot *et al.* (2011) used the observed distance in correlated abundances reported by Tittler et al. (2006) to specify the dispersal-distance limitation within their simulations. Alternatively, Duarte *et al.*, (2016) used results from genetic and occupancy studies, integrated population models, and capture–recapture analyses to support the use of dispersal submodels without a dispersal-distance limitation and with relatively high dispersal rates compared to previous modeling efforts for their study species. On the other hand, if data are lacking ecologists could also simulate population dynamics with dispersal submodels containing a variety of dispersal-related hypotheses (i.e., density dependence, dispersal cost on survival, no dispersal, dispersal-distance limitations, etc.) to evaluate their influence on projected population dynamics (Horne *et al.*, 2011).

It is important to keep in mind that each of the above approaches provides dispersal information at different resolutions, which may influence how these data are integrated to simulate dispersal in ecological models. For example, spatial dynamic occupancy models differ from the other methods introduced in this article in that the dispersal parameters are related to at least 1 individual colonizing a neighboring site, which is the dispersal probability of the species not an individual within a population. Also, integrated population models provide an estimate of immigration or the number/rate of individuals entering a target population, rather than emigrants leaving a target population or dispersal distances. Again, mesocosm studies are typically conducted at

smaller spatial scales and therefore require the assumption that the observed processes can be scaled up to the larger landscape. Genetic and isotope analyses enable ecologists to have a better understanding of migration rates, how far dispersal occurs and what habitats are used, but the information provided is coarse and can sometimes lack information concerning the mechanisms governing dispersal. Correlation in abundances also provide coarse information concerning dispersal distances, and it should be noted that the time-lagged correlations may be spurious or related to some other population process. Finally, although capture–recapture data and location data from VHF and GPS devices provide high-resolution dispersal information, these approaches can be costly and in some cases not practical at the spatial scales needed to document dispersal events for a study species because of logistical challenges and sample size requirements. This is not to say these approaches do not have merit, but we include this information to highlight that careful consideration should be invested to evaluate which approach or set of approaches best match the objectives and resources of a specific project before implementation.

Ultimately, the use of a sensitivity analysis is paramount to highlight the influence of model assumptions on model results, particularly given dispersal processes have high ecological importance and at the same time dispersal estimates are difficult to obtain. Indeed, it should be standard practice to evaluate the outcome of various dispersal-related rulesets or range in input values, especially if they are not empirically based estimates. Within natural resource management decision making, some rulesets or range of input values may drive the optimal management policies based on simulation results, while others will not affect it at all. Identifying key uncertainties that influence what decision is considered optimal based on simulation results is vital when managers rely on these results to help inform real-world management decision making. Importantly, if modeling efforts are integrated with monitoring programs using an adaptive management framework these uncertainties can be explicitly incorporated into the decision-making process and reduced over time.

See also: Behavioral Ecology: Dispersal–Migration. Behavioral Ecology: Animal Home Ranges. Ecological Processes: Wind Effects

References

Bonnot, T.W., Thompson III, F.R., Millspaugh, J.J., 2011. Extension of landscape-based population viability models to ecoregional scales for conservation planning. Biological Conservation 144, 2041–2053.
Broms, K.M., Hooten, M.B., Johnson, D.S., Altwegg, R., Conquest, L.L., 2016. Dynamic occupancy models for explicit colonization processes. Ecology 97, 194–204.
Chandler, R.B., Muths, E., Sigafus, B.H., Schwalbe, C.R., Jarchow, C.J., Hossack, B.R., 2015. Spatial occupancy models for predicting metapopulation dynamics and viability following reintroduction. Journal of Applied Ecology 52, 1325–1333.
Clobert, J., Danchin, E., Dhondt, A.A., Nichols, J.D. (Eds.), 2001. Dispersal. Oxford, UK: Oxford University Press.
Conn, P.B., Cooch, E.G., 2009. Multistate capture–recapture analysis under imperfect state observation: An application to disease models. Journal of Applied Ecology 46, 486–492.
Cormie, A.B., Schwartz, H.P., Gray, J., 1994. Determination of the hydrogen isotopic composition of bone collagen and correction for hydrogen exchange. Geochimica et Cosmochimica Acta 58, 365–375.
Duarte, A., Hatfield, J.S., Swannack, T.M., Forstner, M.R.J., Green, M.C., Weckerly, F.W., 2016. Simulating range-wide population and breeding habitat dynamics for an endangered woodland warbler in the face of uncertainty. Ecological Modelling 320, 52–61.
Ergon, T., Gardner, B., 2014. Separating mortality and emigration: Modelling space use, dispersal and survival with robust-design spatial capture–recapture data. Methods in Ecology and Evolution 5, 1327–1336.
Erickson, R.A., Rees, C.B., Coulter, A.A., Merkes, C.M., McCalla, S.G., Touzinsky, K.F., Walleser, L., Goforth, R.R., 2016. Detecting the movement and spawning activity of bigheaded carps with environmental DNA. Molecular Ecology Resources 16, 957–965.
Faubet, P., Waples, R.S., Gaggiotti, O.E., 2007. Evaluating the performance of a multilocus Bayesian method for the estimation of migration rates. Molecular Ecology 16, 1149–1166.
Fauchald, P., Tveraa, T., 2003. Using first-passage time in the analysis of area-restricted search and habitat selection. Ecology 84, 282–288.
Grosbois, V., Tavecchia, G., 2003. Modeling dispersal with capture–recapture data: Disentangling decision of leaving and settlement. Ecology 84, 1225–1236.
Hestbeck, J.B., Nichols, J.D., Malecki, R.A., 1991. Estimates of movement and site fidelity using mark-resighting data of wintering Canada Geese. Ecology 72, 523–533.
Hobson, K.A., 1999. Tracing origins and migration of wildlife using stable isotopes: A review. Oecologia 120, 314–326.
Horne, J.S., Garton, E.O., Krone, S.M., Lewis, J.S., 2007. Analyzing animal movements using Brownian bridges. Ecology 88, 2354–2363.
Horne, J.S., Strickler, K.M., Alldredge, M., 2011. Quantifying the importance of patch-specific changes in habitat to metapopulation viability of an endangered songbird. Ecological Applications 21, 2478–2486.
Kendall, W.L., Nichols, J.D., 2002. Estimating state-transition probabilities for unobservable states using capture–recapture/resighting data. Ecology 83, 3276–3284.
Kendall, W.L., Nichols, J.D., 2004. On the estimation of dispersal and movement of birds. Condor 106, 720–731.
Kranstauber, B., Kays, R., LaPoint, S.D., Wikelski, M., Safi, K., 2012. A dynamic Brownian bridge movement model to estimate utilization distributions for heterogeneous animal movement. Journal of Animal Ecology 81, 738–746.
LaRue, M.A., Nielsen, C.K., 2008. Modeling potential dispersal corridors for cougars in Midwestern North America using least-cost path methods. Ecological Modelling 212, 372–381.
Millspaugh, J.J., Marzluff, J.M., (Eds.), 2001. Radio tracking and animal populations. San Diego, CA: Academic Press.
Risk, B.B., De Valpine, P., Beissinger, S.R., 2011. A robust-design formulation of the incidence function model of metapopulation dynamics applied to two species of rails. Ecology 92, 462–474.
Royle, J.A., Chandler, R.B., Sollmann, R., Gardner, B., 2014. Spatial capture–recapture. Waltham, MA: Academic Press.
Schaub, M., Abadi, F., 2011. Integrated population models: A novel analysis framework for deeper insights into population dynamics. Journal of Ornithology 152, 227–237.
Schaub, M., Royle, J.A., 2014. Estimating true instead of apparent survival using spatial Cormack–Jolly–Seber models. Methods in Ecology and Evolution 5, 1316–1326.
Schwarz, C.J., 2005. Multistate models. In: Armstrup, S.C., McDonald, T.L., Manly, B.F.J. (Eds.), Handbook of capture–recapture analysis. Princeton, NJ: Princeton University Press, pp. 165–195.
Sutherland, C., Elston, D.A., Lambin, X., 2014. A demographic, spatially explicit patch occupancy model for describing and predicting metapopulation dynamics and persistence. Ecology 95, 3149–3160.

Tittler, R., Fahrig, L., Villard, M.A., 2006. Evidence of large-scale source–sink dynamics and long-distance dispersal among wood thrush populations. Ecology 87, 3029–3036.
Wang, I.J., Savage, W.K., Shaffer, H.B., 2009. Landscape genetics and least-cost path analysis reveal unexpected dispersal routes in California tiger salamander (*Ambystoma californiense*). Molecular Ecology 18, 1365–1374.
Wilson, G.A., Rannala, B., 2003. Bayesian inference of recent migration rates using multilocus genotypes. Genetics 163, 1177–1191.

Modules and Integrated Modeling☆

Alexey Voinov, University of Technology Sydney, Sydney, NSW, Australia
Paul A Fishwick, University of Florida, Gainesville, FL, United States

Glossary

Integrated model Integrated modeling is the method that is developing to bring together diverse types information, theories, and data originating from scientific areas that are different not just because they study different objects and systems, but because they are doing that in very different ways, using different languages, assumptions, scales, and techniques.

Integronsters Modular constructs that are perfectly valid as software products but ugly and useless as models.

Module A model that is presented as an encapsulated, self-contained independent unit that can be used to construct a more complex models, by linking (coupling) it with other modules.

Introduction

As increasingly more ecological models are developed, it becomes clear that in many cases we are "reinventing-the-wheel." The problem is that models are often developed as nonreusable "one-off" items, or worse, that models are "black boxes" that contain code that by its very nature contains no suitable level of abstraction by which phenomena can be properly understood. One way to avoid that is to build a general ecosystem model, which in theory could eliminate the need for continuous remaking of models for different systems and/or sites. One such model, the general ecosystem model (GEM) has been designed to simulate a variety of ecosystem types using a fixed model structure. Such "generality" logically lead to one of the broader objectives in ecosystem research: with a standard structure for developing a (model) synthesis of a system, comparisons among systems may be facilitated. The model was to be generally applied to ecosystems that range from wetlands to upland forests. It was to provide at least two useful functions in synthesizing our broader understanding of ecosystem properties. One involves using the model as a quantitative template for comparisons of the different controls on each ecosystem, including the process-related parameters to which the systems are most sensitive. Secondly, a simulation model, which is general in process, orientation, and structure, could provide a tool to analyze the influence of scale on actual and perceived ecosystem structure. Object-orientation provides one example of such a structure with ecosystems being natural phenomena for this type of design. Other models, such as CENTURY for example, can claim to be of the same kind of functionality, providing for a wide range of processes that can be parameterized for very different locations and ecosystem types.

However, the general approach turned out to be somewhat insufficient to cover all the possible variety in ecosystem processes and attributes that come into play when going from one ecosystem type to another, and from one scale to another. Modeling is a goal driven process, and different goals in most cases will require different models. There is too much ecological variability to be represented efficiently within the framework of one general model. Either something important gets missed, or the model becomes too redundant to be handled efficiently especially within the framework of larger spatially explicit models. Similarly, when changing scale and resolution, different sets of variables and processes come into focus. Certain processes that could be considered at equilibrium at a weekly time scale need be disintegrated and considered in dynamic at an hourly time scale. For example, ponding of surface water after a rainfall event is an important process at fine temporal resolution, but may become redundant if the time step is large enough to make all the surface water either removed by overland flows, or infiltrated. Daily net primary productivity fluctuations, that are important in a model of crop growth, may be less important in a forest model that is to be run over decades with only average annual climatic data available. Once again the general approach may result in either insufficiency or redundancy.

The modular approach is a logical extension of the general approach. In this case instead of creating a model general enough to represent all the variety of ecological systems under different environmental conditions, we develop a library of modules simulating various components of ecosystems or entire ecosystems under various assumptions and resolutions. In this case the challenge is to put the modules together, using consistent and appropriate scales of process complexity, and make them talk to each other within a framework of a full model. We avoid the "reinventing-the-wheel" by keeping much of the model structure and replacing only the parts that need to be modified under the particular goals of model implementation.

☆*Change History*: March 2018. Alexey Voinov introduced small edits in the text of the article including citations.

This is an update of A.A. Voinov and P.A. Fishwick, Modules in Modeling, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 2419–2425.

 doi:10.1016/B978-0-12-409548-9.11142-X

The modular approach directly leads to integrated modeling, when whole models become modules, or components that can be further reconnected, recoupled to build representations of most complex systems.

Modularity

The concept of modularity gained strong momentum with the wide spread of the object oriented (OO) approach in software development. Engineers and computer designers realized some time ago that it is cheaper and more efficient to build devices made of replaceable units. So if you run out of space on your hard disk you can easily take it out and plug in a new bigger one. Similarly, you can swap your CD reader for a DVD reader. The same kind of functionality came with OO software, where pieces of your code became self-contained and self-sufficient and could be easily plugged into other programs or replaced by other components providing the same or improved functionality.

The next logical step was to apply the same concepts to modeling. But this required specific design criteria and rules for building and maintaining models. The features of *decomposability* and *composability* are the most important ones. The decomposability criterion requires that a module should be an independent, stand-alone submodel that can be analyzed separately. On the other hand the composability criterion requires that modules can be put together to represent more complex systems.

Decomposability is mostly attained in the conceptual level, when modules are identified among the variety of processes and variables that describe the system. There is a lot of arbitrariness in choosing the modules. The choice may be driven either by purely logical, physical, ecological considerations about how the system operates, or by quantitative analysis of the whole system, when certain variables and processes are identified as rather independent from the other ones.

The composability of modules is usually treated as a software problem. That aspect is resolved by use of wrappers that enable modules to publish their functions and services using a common high-level interface specification language (the federation approach). The other alternative is the design of model specification formalism that draws on the object-oriented methodology and embeds modules within the context of a specific modeling environment that provides all the software tools essential for simulation development and execution (the specification approach). In both cases as models find themselves in the realm of software developers the gap between the engineering and the research views on models and their performance starts to grow.

From the software engineering viewpoint the exponential growth of computer performance offers unlimited resources for the development of new modeling systems. With the advent of the Internet the vision was to assemble models from building blocks connected over the Web and distributed over a network of computers. New languages and development tools started to appear to facilitate this process, in many cases even faster than their user-communities managed to develop.

On the other hand from the research viewpoint, if a model is to be a useful simplification of reality it should enable a more profound understanding of the system of interest. It is more important as a tool for understanding the processes and systems, than for merely simulating them. In this context there is a more limited demand for the overwhelming complexity of modeling systems. The existing software may remain on the shelf if it does not really help understand the systems. This is probably especially pertinent to models in biology and ecology, where in contrast to physical science or engineering, the models are much more loose and tend to "black-box" much of the underlying complexity due to the difficulty of parameterization and simulation of all the mechanisms from a first-principal basis. They may require a good deal of analysis, calibration, and modifications, before they may be actually used. In this case the focus is on model and module transparency and openness. For research purposes it is much more important to know all the nuts and bolts of a module to use it appropriately. The "plug-and-play" feature that is so much advocated by some software developers becomes of lower priority. In a way it may even be misleading, creating the illusion of simplicity of model construction from prefabricated components, with no real understanding of process, scale, and interaction.

Major requirements for a modular model are as follows:

- *Expandability*. The modules should be designed in such a way that new modules could be easily added and existing modules modified.
- *Scalability*. There should be some clear idea of scale attached to each module. Either a module is designed only for a specific scale, and this scale is clearly identified, or the scale is incapsulated in the module so that it can adjust depending on the scale used in other modules.
- *Transparency*. Modules should be easy to explore and understand. This is a prerequisite of them being reused. Documentation is crucial.

Object-Oriented Modeling

Object orientation is a style of modeling and software engineering. In a general sense, OO takes specific metaphors associated with how physical objects are characterized, and uses these metaphors as a tool for model design, and ultimately programming design and implementation. This physical metaphor has a natural fit within a topic such as ecosystems since the spatial extent of the ecosystem as well as all species residing in that space can be seen as "objects." One develops an "object-oriented model" and then locates a suitable programming language that embodies OO as its primary architecture. The OO approach uses two fundamental ideas: *encapsulation* and *inheritance*. With encapsulation, it is possible to locate computer code and data "within" an object so that

the code and data become addressable and located by first querying the object for its contents. Code within an object is termed a "method" or "behavior" and data is an "attribute." With inheritance, objects become leaf-nodes on a tree whose root represents the most abstract category (or "class") and whose children represent subcategories (or "subclasses").

An example of encapsulation and inheritance can be seen by taking the domain of fish. The shark and stonefish are two types of "fish." One then can define an OO class called "Fish" and two subclasses called "Shark" and "Stonefish." One aspect of building this hierarchy is to take advantage of the process of inheritance, but before discussing that we first need to employ encapsulation in this example. A fish is composed of a skeleton, a central-nervous system, skin, and internal organs, among many other parts. These parts make up a hierarchy that defines encapsulation. It is reasonable to imagine that to obtain attribute information on the type and format of a skeleton, one would get this by going directly to a specific, unique fish—perhaps a specific shark we'll call shark_1032 for lack of a better identifier. Inside of the object shark_1032, we can find all information that this object encapsulates: the size of the object, its skeletal components, and so forth. Now, getting back to the issue of inheritance, it seems clear from the biological taxonomy of fish that both sharks and stonefish both have skeletons. So, we can move the data structure that contains the data for a skeleton and place it in the "Fish" class. The classes Shark and Stonefish inherit all methods and attributes of Fish. Often, this process is termed inheritance of derived classes (Shark, Stonefish) from the base class (Fish).

So, not only does OO design support a new way of thinking about modularity, it supports encapsulation of attributes and methods as well as ways in which these are "moved" (i.e., inherited-by) other components. It is worth closing this section with a description of how OO differs from prior ways of thinking about program design as well as where all of this new thought will lead in the future. Prior to OO, while most real world scenarios were described using words such as "objects" and "attributes," programs were defined in terms of procedures that encapsulated information in the form of simple and complex data. The computer language FORTRAN is a good example of this: the language was composed of functions and subroutines, and data were stored either globally (i.e., for all subroutines to use) or locally (i.e., within that particular subroutine). Procedural approaches such as the one used in FORTRAN are still used in OO (i.e., a method); however, the key difference in the evolution of programming is the introduction of additional encapsulation of objects, either of the physical variety (i.e., "shark_1032") or the conceptual variety (i.e., "budget").

Examples

Flexible modeling system (FMS)

FMS is a software framework developed by the Geophysical Fluid Dynamics Laboratory (GFDL), which is a NOAA climate modeling center at Princeton. It supports the efficient development, construction, execution, and scientific interpretation of atmospheric, oceanic, and climate system models. It is an outgrowth of the MOM family of climatic models, with the latest one—MOM4. The goal is to provide the international climate research community with a repository for robust and well-documented methods to simulate the ocean climate system. Researchers are invited to support the existing modules and provide various modules that are absent from MOM4, yet may enhance the simulation integrity (e.g., a new physical parameterization or new advection scheme) or increase the model's functionality.

FMS comprises the following:

1. A software infrastructure for constructing and running models. This infrastructure includes software to handle parallelization, input, and output, data exchange between various model grids, orchestration of the time stepping, makefiles, and simple sample run scripts. This infrastructure should largely insulate FMS users from machine-specific details.
2. A standardization of the interfaces between various component models.
3. Software for standardizing, coordinating, and improving diagnostic calculations of FMS-based models, and input data preparation for such models. Common preprocessing and postprocessing software are included to the extent that the needed functionality cannot be adequately provided by available third-party software.
4. Contributed component models that are subjected to a rigorous software quality review and improvement process.
5. A standardized technique for version control and dissemination of the software and documentation.

FMS is a software framework. The FMS developers make it clear that their system does not include the determination of model configurations, parameter settings, or the choice of modules. The development of new component models is a scientific concern that is outside of the direct purview of FMS. Nonetheless, infrastructural changes to enable such developments are within the scope of FMS. The collaborative software review process of contributed models is therefore an essential facet of FMS. The quality review and improvement process includes consideration of (A) compliance with FMS interface and documentation standards to ensure portability and interoperability, (B) understandability (clarity and consistency of documentation, comments, interfaces, and code), and (C) general computational efficiency without algorithmic changes.

As a software framework, it has certain clear requirements that contributed code must meet:

1. Clean modular Fortran 90 code that minimally touches other parts of the model
2. Satisfaction of the FMS code specifications outlined in the FMS Developers' Manual
3. Compatibility with the MOM4 test cases
4. Thorough and pedagogical documentation of the module
5. Comments within the code emulating other parts of the model

Object modeling system (OMS)

OMS is the outgrowth of the modular modeling system (MMS) developed by G. Leavesley and his colleagues in USGS. OMS is developed by USDA and is described as a framework for modeling that can be used to develop, support, and apply any dynamic model, but specifically it is focused in the environmental and natural-resource management disciplines. It uses a module library that contains modules for simulating a variety of physical processes. These are primarily water, energy, chemical, and biological processes. A model is created by selectively coupling appropriate modules from the library to create a suitable model for a desired application. When existing modules do not provide appropriate process algorithms, new modules can be developed.

The conceptual framework for OMS has three major components: preprocess, model, and postprocess. A system supervisor, in the form of a window-based graphical user interface (GUI), provides user access to all the components and features of OMS. There are versions that work under UNIX and Windows operating systems. The GUI provides an interactive environment for users to access model-component features, apply selected options, and graphically display simulation and analysis results.

The *preprocess component* includes the tools used to input, analyze, and prepare spatial and time-series data for use in model applications. A goal in the development of the preprocess component is to take advantage of the wide variety of existing data-preparation and analysis tools and to provide the ability to add new tools as they become available. The time-series and other data that are needed to run the model have to be prepared as a single flat ascii file. Procedures are being developed to interface models with a variety of commercial and user-defined data bases, such as SQL type data bases (Oracle and Ingres) and for the HEC-DSS database. NetCDF is another data format that is supported. These are being used in real time applications with the Bureau of Reclamation and the Natural Resources Conservation Service.

The *model component* is the core of the system and includes the tools to selectively link process modules from the module library to build a model and to interact with this model to perform a variety of simulation and analysis tasks. The module library contains a variety of compatible modules for simulating water, energy, and biogeochemical processes. Several modules for a given process may be present, each representing an alternative conceptualization or approach to simulating that process. OMS requires rewriting modules in Java or C# to be then inserted into the system library.

Modules are located in both read-only directories, where tested, documented, and approved code reside, and in user-defined work directories where new modules are being developed. The user selects and links modules from these directories to create a specific model using an interactive, graphical, model-builder tool. Modules are linked by coupling the outputs of user-selected modules to the required inputs of other user-selected modules. Tools are provided to display a module's input requirements and to list all modules available that will satisfy each of these inputs. When the inputs for all modules are satisfied, a model is complete.

When a model is executed, the user is interfaced with the model through a series of pull-down menus in the GUI, which provide the links to a variety of system features. These include the ability to (1) select and edit parameter files and data files, (2) select a number of model execution options such as a basic run, an optimization run, or a sensitivity analysis run, and (3) select a variety of statistical and graphical analyses of simulation output. During a basic run, up to four graphical display windows can be opened to display any of the variables that have been declared in the model modules. As many as 10 variables can be displayed in each window and plotted results can be output in HPGL or PostScript formats either to a digital file or to a printer.

The *postprocess* component provides a number of tools to display and analyze model results, and to pass results to management models or other types of software. Model output can also be directed to user-specific analysis programs using an ascii flat-file format. Some postprocessing capabilities interact directly with the model component. The parameter-optimization and sensitivity-analysis tools are provided to optimize selected model parameters and evaluate the extent to which uncertainty in model parameters affects uncertainty in simulation results. A geographic information system (GIS) interface is developed to provide tools for the analysis and manipulation of spatial data in the preprocess, model, and postprocess components of OMS. Pre- and postprocessing interfaces are being developed as generic interfaces to support a variety of applications, such as the Arc/Info GIS package. Another candidate support package is the Geospatial Library for Environmental Modeling (GEOLEM).

While the GUI is targeted at a broad range of model users, any updates of the system, and additions of new modules or interfaces require good programming skills.

Community surface dynamics modeling system (CSDMS)

Funded by the National Science Foundation, the University of Colorado-Boulder is developing a modular modeling system to advance fundamental earth system science, and repositories for data, models and numerical tools, and educational use. They are leveraging the Department of Energy's Common Component Architecture (CCA) used by the National Laboratory system, and interacting with other efforts, including OpenMI and OMS, to lay out their system. The core of the system is the basic model interface (BMI), a library specification to simplify the coupling of models. In this context an interface is a named set of functions with prescribed function names, argument types, and return types. The BMI functions make the model self-describing and fully controllable by a modeling framework or application. By design, the BMI functions are straightforward to implement in any language and use only simple (universal) data types. While the CSDMS model coupling framework supports C, C++, Fortran, Java, and Python, a BMI can described for any language. CSDMS provides example bindings for BMI in each of the above languages. Also by design, the BMI functions are noninvasive. This means that a BMI-compliant model does not make any calls to other components or tools and is not modified to use any framework-specific data structures. BMI therefore introduces no dependencies into a model and the model can still be used in a stand-alone manner.

BMI specifies:

- Model control functions
- Model information functions
- Time functions
- Variable information functions
- Variable getter and setter functions
- Model grids

In a prefabricated model, the issues of scale consistency are taken care of by the model developers beforehand. Now with the modular approach, the challenge of combining the modules in such a way that they match the complexity of the modeled system and are mutually consistent becomes the task of the library user. Once again this added concern is the price that is paid for the added flexibility and optimality of the resulting models. In theory, we can envision modeling systems that would keep track of the scales and resolutions of the various processes involved, and automatically allow links with only such modules that would match these scales. In practice, with all the complexity and uncertainty associated with ecological and socio-economic systems, it may still be a while until such modeling tools appear. There is always a risk of creating so-called "integronsters" when modules are coupled in a way that may be correct from the formal point of view of units and variables, but makes little sense from the point of view of matching reality. Model transparency is a very important prerequisite of modularity, especially if the modules are to be used in a research context.

Community Modeling

The modular approach is usually a result of collaboration between different groups of modelers as long as they can agree to subscribe to the same set of rules or specifications. At the same time, the modular architecture can significantly empower this kind of community modeling, calling for new members joining the group and contributing their resources in mutually acceptable formats.

MMS, for example, began as a cooperative research effort between the U.S. Geological Survey (USGS) and the University of Colorado's Center for Advanced Decision Support for Water and Environmental Systems (CADSWES). Later on interest in the MMS concepts was expressed by many other national and international agencies and organizations. Agreements established with several of these groups have provided new ideas for system enhancement and the contribution of resources, in terms of money and/or people, to add these enhancements to the system. In addition, these groups continue to contribute their modeling expertise to the system by converting their models to MMS modules and by providing test sites for system evaluation and development. Other partners contributing to the MMS development include the U.S. Bureau of Reclamation (BOR), the Electrical Power Research Institute (EPRI), the TERRA Laboratory, which is a joint Agricultural Research Service (ARS)-Forest Service (FS)-USGS consortium that was formed to facilitate the development of decision support systems for terrestrial ecosystem problems, and others.

The FMS is mostly developed under the auspice of GFDL at NOAA, however its code is provided at GForge (after 2002, before that it was developed as a SourceForge project), which means that it is open for a broad community of programmers and modelers to contribute modules and interfaces.

However in most cases we are still talking about fairly closed communities built around a set of rules and ideas that define the system architecture. The modular systems remain mostly a product of the core group of developers who subscribe to the same more or less limited set of standards and specifications. Apparently a truly flexible and modular system, widely accepted by a range of model development and application organizations, is yet to be developed.

Participatory and Collaborative Modeling

Modular model architecture becomes especially attractive when models are part of a collaborative, participatory modeling process. In this case stakeholders are engaged in the modeling process, and the model itself becomes a tool for deliberations, joint knowledge building, understanding, and decision making. Modularity is very promising in this context because it offers much transparency to the process and allows "on-the-fly" modifications to the overall modeling structure to accommodate the needs and desires of the stakeholder community.

The idea of collaborative modeling is vital to the future success of the modeling enterprise: providing ways in which different scientists can work together, perhaps within a shared space. Some new technologies on the horizon offer possibilities for collaborative modeling. In particular the domain of multiuser games and environments allow an arbitrary number of participants to interact with shared objects. These environments are "object-oriented" by their very nature since "in world," one has an inventory composed of objects and each object contains scripts used to identify behaviors that the object can adopt. However it is still important to realize that modular collaborative modeling goes beyond the software challenges. In addition to new software tools it requires acceptance of new research paradigms promoting open source and open model development, data sharing, and participatory modeling efforts.

See also: Ecological Data Analysis and Modelling: Mediated Modeling and Participatory Modeling; Ecological Models: Model Development and Analysis; Parameterization; Model Types: Overview; Sensitivity, Calibration, Validation, Verification; Structural Dynamic Models

Further Reading

Belete, G.F., Voinov, A., Laniak, G.F., 2017. An overview of the model integration process: From pre-integration assessment to testing. Environmental Modelling & Software 87, 49–63. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1364815216308805.

Fishwick, P., 1995. Simulation model design and execution: Building digital worlds. Upper Saddle River, NJ: Prentice Hall.

Fishwick, P., Sanderson, J., Wolff, W., 1997. A multimodeling basis for across-trophic-level ecosystem modeling: The Florida Everglades example. Transaction of the Society for Computer Simulation International 15 (2), 76–98.

Laniak, G.F., *et al*., 2013. Integrated environmental modeling: A vision and roadmap for the future. Environmental Modelling & Software 39, 3–23. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1364815212002381 Accessed 31 January 2013.

Leavesley, G.H., Markstrom, S.L., Restrepo, P.J., Viger, R.J., 2002. A modular approach to addressing model design, scale, and parameter estimation issues in distributed hydrological modeling. Hydrological Processes 16, 173–187.

Peckham, S.D., Hutton, E.W., Norris, B., 2013. A component-based approach to integrated modeling in the geosciences: The design of CSDMS. Computational Geosciences 53, 3–12.

Reynolds, J.F., Acock, B., 1997. Modularity and genericness in plant and ecosystem models. Ecological Modelling 94 (1), 7–16.

Silvert, W., 1993. Object-oriented ecosystem modeling. Ecological Modelling 68, 91–118.

Voinov, A., Cerco, C., 2010. Model integration and the role of data. Environmental Modelling & Software 25 (8), 965–969. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1364815210000435 Accessed 9 August 2010.

Voinov, A., Shugart, H.H., 2013. "Integronsters", integral and integrated modeling. Environmental Modelling & Software 39, 149–158. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1364815212001697 Accessed 31 January 2013.

Voinov, A., Fitz, C., Boumans, R., Costanza, R., 2004. Modular ecosystem modeling. Environmental Modelling and Software 19 (3), 285–304.

Voinov, A.A., *et al*., 2010. A community approach to Earth systems modeling. EOS Transactions of the American Geophysical Union 91 (13), 117–124.

Relevant Websites

http://wwwbrr.cr.usgs.gov/mms/—Modular Modeling System.
http://www.uvm.edu/giee/LHEM/—Library of Hydroecological Modules.
http://www.gfdl.noaa.gov/fms/—Flexible Modeling System.
http://csdms.colorado.edu/wiki/Main_Page—Collaborative Surface Dynamics Modeling System.

Parameterization☆

Alexey Voinov, University of Technology Sydney, Sydney, NSW, Australia

Glossary

Boundary conditions These describe the spatial and temporal boundaries of a system. For a spatially homogeneous system we have only *initial conditions*, that describe the state of the variables at time $t = 0$, when we start the model. For spatially distributed systems we may need to define the conditions along the boundary, as well as the geometry of the boundary itself. These are known as *boundary conditions*.

Constants or parameters in a narrow sense These are measured or calibrated quantities that describe the rates of processes in a process-based model. In empirical models these are the constants that are statistically derived based on empirical observations. We may want to distinguish between real constants, such as gravity, g, and, say, half-saturation coefficient, K, in the Michaelis-Menten function. While both of them take on constant values in a particular model run, g will be always the same from one run to another, but K may change quite substantially while we calibrate the model. Even if K comes from observations, it will normally be measured with certain error, so the exact value will not be really known.

Control functions Control functions are also parameters, except that they are allowed to change to see how their change affects systems dynamics. It is like tuning the knob on a radio set. At every time the knob is dialed to a certain position, but you know that it may vary and will result in different performance of the system.

Forcing functions These are parameters that describe the effect of the outside world upon the system. These may change in time or space, but they do not respond to changes within the system. They are external to it, driven by processes in the higher hierarchical levels. Climatic conditions, rainfall, temperature certainly affect the growth of tomatoes in the garden, but the tomatoes hardly affect the temperature or the rainfall patterns. If we build a model of tomato growth, the temperature will be a forcing function.

Parameterization is the process of finding the right parameter values for a model. In process based modeling it is assumed that parameters have some meaning, some physical or ecological sense. This also means that for most of such parameters some maximal and minimal values can be produced, limiting the size of the space when choosing parameter values. Empirical models are certainly also driven by parameters, but here parameters are simply values that are calibrated and cannot be measured in experiments. There are also no restrictions on the domains from which the parameter values can be chosen.

Note that in some texts parameters will be assumed only in the narrow sense of coefficients that may sometimes change, like the growth rate or half-saturation values. However this may be somewhat confusing, since forcing functions are also such parameters if they are fixed. Suppose we want to run a model with temperature held constant and equal to the mean over a certain period of time, say the 6 months of the growth season for a crop. Then suppose later on we want to feed into the model the actual data that we have measured for temperature. So now it is no longer a constant but changes every day according to the recorded time series. Does this mean that temperature will no longer be a parameter? For any given moment it will still be a constant. It will only change from time to time according to the data available. It would make sense to still treat it as a parameter, except now it will be no longer constant but will change accordingly.

Suppose now that we approximate the course of temperatures by a sine function with some constants that control the amplitude and the period. Now temperature will no longer be a parameter. Instead temperature will be an intermediate variable, while we will have two new parameters in the sine function that specify temperature—one parameter ($B = 4$) will make the period equal to 6 months, the other parameter (A) will define the amplitude and make the temperature change from a minimal value (0) to the maximal value (40, if $A = 20$) and back over this period of time, like in the function:

$$\text{Temperature} = A^{*}\text{Sin}\,(t^{*}B^{*}\pi/365 + 3^{*}\pi/2) + A.$$

Here t is time, π—is a constant $\pi = 3.14$, A, and B are parameters. If we change B, $B = 2$, then the period will change from 6 to 12 months.

For process based models, there may be a number of ways to determine model parameters, including the following.

1. Measurements in situ are probably the best way since they define the value of exactly what is assumed in the model. However such measurements are the most labor and cost intensive, and they also come with large margins of error. Besides in many cases such measurements may be risky or not be possible at all, if a parameter represents some aggregated value or an extreme

☆*Change History*: March 2018. Alexey Voinov introduced small edits in the text of the article including citations.

This is an update of A.A. Voinov, Parameters, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 2638–2639.

 doi:10.1016/B978-0-12-409548-9.11120-0

condition, that may not occur in reality (say, the maximal temperature for a population to tolerate—it may differ from one organism to another, and such conditions may be hard to find in reality).

2. Experiments in lab (in vitro) are usually performed when in situ experiments are impossible. Say we take an organism and expose it to high temperatures to find out the limits of its tolerance. We can create such conditions artificially in a lab, but we cannot change the temperature for the whole ecosystem.
3. Values from previous studies found from literature, web searches or personal communications. If data are available for similar systems it certainly makes sense to use them. However always keep in mind that there are no two identical ecosystems, so most likely there will be some error in the parameters borrowed from another case study.
4. Calibration (ref to the entry on calibration and validation). When we know what the model output should look like, we can always tweak some of the parameters to make the model perform best. For empirical models this and literature search are the only ways to identify parameter values.
5. Basic laws, say conservation principles and therefore mass and energy balances
6. Allometric principles, stoichiometry, and other chemical, physical etc. properties. Basic and derived laws may help establish relationships between parameters, and therefore identify at least some of them based on the other ones already measured or estimated.

Note that in all cases there is a considerable level of uncertainty present in the values assigned to various model parameters. Further testing and tedious analysis of the model is the only way to decrease the error margin and deal with this uncertainty.

With proliferation of various sensors and automatic monitoring, we are increasingly entering the era of Big Data when more information about the system can be inferred by only looking at its overall performance paying less attention to the system structure, the processes that drive the system, and the parameters that control those processes. It is yet to be seen if this kind of modeling will replace the more theory rooted process based models that have been previously used in population ecology and ecosystem studies.

See also: Ecological Data Analysis and Modelling: Statistical Inference; Ecological Models: Model Development and Analysis; Model Types: Overview; Sensitivity, Calibration, Validation, Verification; Structural Dynamic Models

Further Reading

Jorgensen, S.E. (Editor-in-chief) Friis,Henriksen,Jorgensen,Mejer, 1978. Handbook of environmental data and ecological parameters. Vaedose, Denmark: ISEM. 1162 pp.
Jørgensen, L., Jørgensen, S.E., Nielsen, S.N., 2000. Ecotox. CD software. Amsterdam: Elsevier.
Refsgaard, J.C., 1997. Parameterisation, calibration and validation of distributed hydrological models. Journal of Hydrology 198 (1–4), 69–97. doi:10.1016/S0022-1694(96)03329-X.http://www.sciencedirect.com/science/article/pii/S002216949603329X

Relevant Website

https://en.wikipedia.org/wiki/Parametrization

Sensitivity, Calibration, Validation, Verification☆

Alexey Voinov, University of Technology Sydney, Ultimo, NSW, Australia

Glossary

Calibration Changing model parameters or functions in attempt to match a certain existing data set. Calibration can be conducted as a trial-and-error process by manually changing certain model parts, or it can be formulated as an optimization task, when the computer will minimize the difference between data and model output by adjusting model parameters.

Sensitivity analysis Analyzing how the model reacts to changes in model parameters. The model output may change quite considerably when modifying certain parameters, while other parameters may have very little impact on it. This helps understand which parameters and processes are most important and deserve more attention when calibrating the model.

Validation Comparing model output to some independent data set, which has not been used previously for model calibration.

Verification Carefully checking the model for internal inconsistencies, errors and bugs.

Introduction

Once the model is formulated and formalized, some rigorous model testing is in order. The major steps that are usually assumed are called sensitivity analysis, calibration, validation and verification. While the first two can be well formalized and are quite standard in any modeling effort, the validation and verification stages are designed to assess the level of "truth" that the model delivers, and therefore tend to be more vague and controversial. There has been a good deal of discussion about what a good model is, and whether it is feasible at all to claim that the model is true in any sense. One can argue that for an open system, where conditions constantly change it is not even possible to design a model that would represent reality, since the reality is constantly changing with additional factors brought in all the time. The model then can only represent the situation that it has been designed for, and is very much limited by the conditions and factors that were included.

Nevertheless some model testing is definitely in order and some models are still better than others. In spite of much leeway in the definitions of what a good model is, the model testing and analysis is an important stage, which can tell us much about the system, even if it does not really tell us how "true" the model is.

Sensitivity

If no analytical analysis is possible, we have to turn to numerical methods. The numerical solution of the model requires all parameters to take on certain values, and as a result is dependent on parameters that have been specified. These include coefficients, or constants, initial conditions, forcing function, and control parameters. Some parameters do not matter much. We can vary them quite significantly, but will not see any large changes in the model dynamics. However other parameters may have a very significant effect on the model performance. Even small changes in their values result in dramatically different solutions.

Analyzing model performance under various conditions is called *sensitivity analysis*. If we start modifying a parameter and keep rerunning the model, instead of a single trajectory, we will be generating a bunch of trajectories (**Fig. 1**). Similarly we can start changing the initial conditions or even some of the formalizations in the process descriptions. By comparing the model output we get an idea of the most essential parameters or factors in the model. We will also get a better feeling of the role of individual parameters and processes in how the model output is formed, what parameters affect what variables, and in which ranges the parameters may be allowed to vary. This is very important because in contrast to an analytical solution, where we could find an equation relating model output to the input parameters, with numerical models we do not have any other way to learn what is the connection between the various parameters and the model output, except than rerunning the model.

Eventually when we get sufficient confidence in the model performance and collect evidence of the model being actually adequate to the system that it represents, we can further sensitivity analysis to the point, when we make conclusions about the sensitivity of the original system to certain processes and factors. It will be then those processes that should get the most attention in experimental research, and which may become important management tools if we intend to modify the system behavior to match certain criteria.

☆*Change History*: March 2018. Alexey Voinov introduced small edits in the text of the article including citations.

This is an update of A.A. Voinov, Sensitivity, Calibration, Validation, Verification, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 3221–3227.

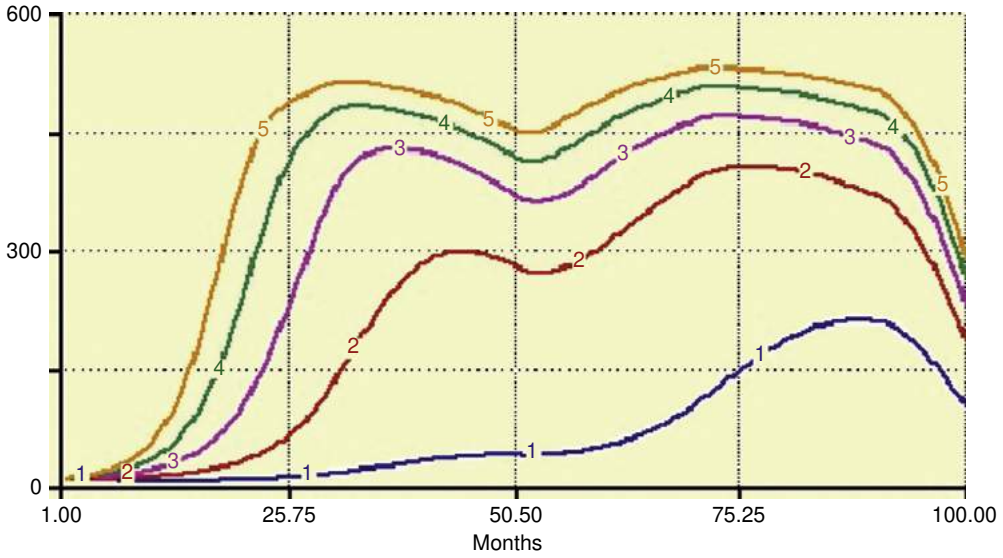


Fig. 1 Analysis of sensitivity in a simple population model.

A full sensitivity analysis of a model is a difficult task, since changing parameters one by one and in combinations may produce entirely different results. But even a partial analysis that will look at some parameters is certainly better than nothing. It will also be of great help for the next step of model analysis, which is calibration.

Calibration

The next thing you will want to do analyzing the model is to compare its output to the other data that is available about the system. In many cases we may have better data about the dependent variables in the model than data about the independent variables or parameters. For instance, USGS provides quite extensive data sets for water flows measured over a network of river gages. For a stream hydrology model that is to produce river flow dynamics, we will most likely have quite good information about the flows, but poor data about the hydrologic coefficients, such as infiltration, transpiration and evaporation rates, etc. By solving an inverse problem we will be determining the values of parameters such that the model output will be as close as possible to the observed data. This process of model refinement in attempt to match a certain existing data set is called *model calibration*. We compare the model output to the data points, and start changing the model parameters or structure in order to get a fit as close as possible.

Sensitivity analysis may have already informed us what are the parameters that need be modified to produce a certain change in model trajectories. Now we actually change them in such a way that the trajectory of the model output matches the data "close enough". How close?—It depends upon the level of confidence in the data we have and upon the goals of the study. It also depends on the model that we built. Sometimes we find very difficult to produce the change in the model output that is needed to get the trajectories closer to the data points. Sometimes it is simply impossible, and we have to find other ways to fix the model, either digging into its structure, or realizing that we have misinterpreted something in the conceptual model, or the time or space scales. Modeling is an iterative process and it is perfectly fine to go back and reevaluate our assumptions and formalizations.

Note that the data set used for calibration, in a way, is also a model of the process observed. The data are also a simplification of the real process and they may also contain errors, they are never perfect, and besides they have been collected with a certain goal in mind, which does not necessarily has to match the goal of the newly built numerical model. We may call these monitoring results an experimental model or a *data model*. In this process of calibration we are actually comparing two models, and modifying one of them (simulation) to better match the other one (data).

When comparing models, it makes sense to think of a measure of their closeness, or a measure of the fit of the simulation model to the data model. This comparison is important for both calibration and further testing of the model (validation, verification). In all these analyses we would want to see how far the model results deviate from the other information we have about the system (both qualitative and quantitative). We may call this measure the *error model*. There may be very different ways to represent this error, but they all have in common one feature, which is that they represent the distance between two models, in this case, the data model and the formal model. The very simplest error model is "eyeballing", or visual comparison. That is when you simply look at the graphs and decide whether they are close enough or not.

However, this may become difficult as we get closer to the target, or when the graph is closer in one time range for one set of parameters, and closer in a different time range for another set of parameters. In those cases visual comparisons can fail. Mathematical formulas can then become useful. One simple formula for the error model is:

Table 1 Available variable tests in the MPI software package

Test	*Description*
BOUNDS	Percentage of points falling into a reference interval
WBOUNDS	Like BOUNDS, weighted according to distance of outliers from nearest limit of interval
CINT	Proportion of points falling into 95% confidence interval of reference data
WCINT	Like CINT, weighted according to distance of outliers
THEIL	Coefficient of inequality between paired data
DBK	Result of simultaneous test of slope = 1, intercept = 0 in linear regression of observed vs. simulated data
STEADY	Steady-state, done as piecewise linear regression and test of slopes
INCREASE	Monotonic increase, done as piecewise linear regression and test of slopes
DECREASE	Monotonic decrease, done as piecewise linear regression and test of slopes
TREND	Known trend, intended as slope of linear regression line
FREQ	Compares the structure of autocorrelation of simulated and observed data to identify common frequencies of oscillation, or looks for specified periods in the simulated data
ERRCOMP	Concordance between the simulated data error composition and specified admissible percentages of mean, variance, and random error

$$E = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{y_i^2}$$

where y_i are the data points, and x_i are the values in the model output that correspond in time or space to the data points.

Very often the metric used to compare the models is the Pearson moment product correlation coefficient,

$$r = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

or the R^2 *value*, which is equal to r^2.

There may be many other ways to estimate the error model. For example a Model Performance Index was proposed that incorporates some 12 metrics to estimate the deviation between the two time series (**Table 1**).

There are many other statistical tools that are available, say, in Excel, in R or in other statistical software packages that may be further used for a refinement of these comparisons.

There is a difference in calibrating empirical and process based models. In empirical models, we entirely rely on the information that we have in the data sets. We come up with some type of equation and then quite mechanically adjust the parameters in attempt to reproduce the data as well as possible. All the information we know about the system is in the data, and the parameters usually can take any values as long as the error model is minimal.

In process based models calibration is different since we are restricted by the ecological, physical or chemical meaning of the parameters that we change. Besides there are usually some estimates for the size of the parameters: they are rarely precisely measured but at least the order of magnitude or a range is usually known. Moreover there are other factors that may play a role, such as confidence in the available estimates for the parameter; sensitivity of the model to a parameter, etc. These are important considerations in the calibration process.

At the bottom of any calibration we have an optimization problem: we seek a minimum for the error model. In most cases, we have certain parameters which values are known and others that are only estimated within a certain domain of change. We call the latter ones—"free" parameters. They are the ones to change in the model in order to minimize the value of the error. To perform optimization we first formulate a *goal function* (also called *objective function*). Then we try to make this function as large or as little as we can by changing different parameters that are involved. In case of calibration the goal function is the error model $\mathbf{E = f(P, C, R)}$, described as a function of the parameter vector **P**, the vector of initial conditions **C** and the vector of restrictions **R**. We then try to find a minimum:

$$\mathbf{\min E}$$

over the space of the free parameters **P** and initial conditions **C**, making sure that the restrictions **R** (such as a requirement that all state variables are positive), hold. There is rarely a model that would allow this task to be solved analytically. It is usually a numerical procedure that requires certain fairly complicated software to be employed.

There are different ways to solve this problem. One approach is to do it manually with the so-called *trial and error* method or *educated guess* approach. The model is run, then a parameter is changed, then the model is rerun, output is compared, the same or another parameter is changed, and so on. It may seem quite tiresome and boring, but actually this process is extremely useful to understand how the system works. Playing with the parameters you learn how they affect output (as in the sensitivity analysis stage), but you also understand the synergetic effects that parameters may have. In some cases you get quite unexpected behavior, and it takes some thought and analysis to explain yourself how and why the specific change in parameters had this effect. If you

cannot find any reasonable explanation, chances are that there is a bug in the model. A closer look at the equations may solve the problem: something may have been missed, or entered with a wrong sign, or some effect was not accounted for.

In addition to the educated guess approach, there are also formal mathematical methods that are available for calibration. They are usually based on the solution of the so-called optimization problem.

Some modeling systems have the functionality to solve the optimization problem and do the curve fitting for models. One such package is *Madonna*. One big advantage of Madonna is that it can also take Stella equations almost as is and run them under its own shell. Madonna also has a nice graphic user interface of its own. So you may as well start putting your model together directly in Madonna, if you expect some optimization to be needed.

The calibration problem may not have a unique solution. There may be several parameter vectors **P** that deliver the same or almost the same minima to the optimization task. In that case it may be unclear what parameters to choose for the model. Other considerations and restrictions may be used to make the decision.

If we have done our best finding the values for all the parameters in the simulation model and yet still the error is inappropriately large, this means that something is wrong in one of the models that we are comparing. Either the conceptual model needs to be revised (the structure changed or the equations modified), or the chosen scales were incorrect and we need to reconsider the spatial or temporal resolution. Alternatively, the data is wrong, which also happens quite often and cannot be dismissed.

To conclude, there are different ways to describe systems by means of models. There are different models that may be built. *The process of adjustment of one model to match the output from another model is called calibration*. This is probably the most general definition. In most cases we would speak of calibration as the process of fitting the model output to the available data points or "curve fitting". In this case it is the data model that is used to calibrate the mathematical model.

Note that there is hardly any reason to always give preference to the data model. The uncertainty in the data model may be as high as the uncertainty in the simulation model. The mathematical model may in fact cover areas that are not yet presented in data at all. However in most cases we will have data models precede mathematical models, and, at least initially, assume that the data models convey our knowledge about the system.

Empirical models are entirely based on data models, they may be considered as "extensions" of the data models. They attempt to generalize the data available and present them in a different form. The process-based models, in addition to knowledge about the modeled system, may also employ information about similar systems studied elsewhere or they may incorporate theoretical knowledge about processes involved. In a way these process-based models can be even better than the data available for the particular system that is modeled. Therefore we may hope that process based models will be performing better outside of the data domains that were used for their calibration. So perhaps it will be easier to apply process-based models to other similar systems, than empirical models, which would require a whole new calibration effort.

Testing

Now we have a simulation model that represents the data set close enough. Does this mean that we have a reliable model of the system, which we can use for forecast or management? Did we really capture the essence of the system behavior, do we really understand how the system works or we have simply tweaked a set of parameters to produce the needed output? Are we representing the system and the processes in it, or, as in empirical models, we only see an artifact of the data set used?

We build process-based models with the presumption that they describe the guts of the system and therefore are general enough to be reapplied in different conditions, since they actually describe how the system works. That would be indeed true if all the parameters in the process formulations could be measured in experiment and then simply substituted into the model. However usually this data is nonexistent or imprecise for all of the parameters.

The solution we found was to approximate the parameter values based on the data we had about the dynamics of state variables, or flows. That was the model calibration procedure. We were solving an inverse problem: finding the parameters based on the dynamics of the unknowns. This would be fine if we could really solve that problem and find the exact values for the parameters. However, in most cases that is also impossible, and, instead, we are finding approximate solutions that come from model fitting. But then how is this different from the fitting we do when we deal with empirical models? In that case, we also had a curve equation with unknown coefficients, which we determined empirically by finding the best combination of parameters that made the model output as close as possible to the data.

The only difference is that instead of some kind of generic equation in the empirical models (say, a polynomial of some form), in process-based models we have particular equations that have some ecological meaning. These equations that display certain behavior by themselves, no matter what parameters are inserted. A polynomial can generate pretty much arbitrary dynamics as long as the right coefficients are chosen. But, say, a classic predator–prey system will always produce a certain type of dynamics, no matter what coefficients we insert. So we may conclude that to a large extent, we are building a good model as long as we chose the right dynamic equations to describe our system.

On top of the basic dynamic equations we overlay the many other descriptions for the processes that need to be included in the model. These may be the limiting factors, describing the modifying effect of temperature, light or other external conditions. There may be some other details that we wish to add to the system. However if these processes are not studied in an experiment, and if the related coefficients are not measured, their role in the model is not any different than that of the coefficients that we have in an

empirical model. In both cases we figure out their values based on a time-series of model output, in both cases the values are approximate, and uncertain. They are only as good as they are the best ones found: we can never be sure that a better parameter set does not exist.

So the bottom line is that there is a good deal of empiricism in most process based models, and the more parameters we have estimated in the calibration process, the more empiricism is involved, the less applicable the model will be in situations outside the existing data range. How can we make sure that we have really captured the essence of the system dynamics, and can reproduce the system behavior beyond the domain that we have already studied?

To answer all these questions, the model needs to undergo a process of vigorous testing. There is and probably will never be a definite procedure for model testing and comparisons. The obvious reason is that models are built for various purposes; their goals may be very different. Moreover these goals may easily change when the project is already underway. There is no reason why goal setting should be left out of the iterative modeling process. As we start generating new knowledge and understanding with a model, its goals may very well modify. We may start asking new questions and will need to modify the model, even while it has not been yet brought to perfection.

Besides, most of the ecological systems are open, which makes their modeling similar to shooting at a moving target. While we study the system and built a model of it, it already evolves. It evolves even more when we start administering control, when we try to manage the ecosystem. As a result models can very well become obsolete even before they are used to produce results. We are modeling the system as it was until a year ago, but during the last year because of some external conditions (say, global climate change) the system has already evolved and the model is no longer relevant.

Nevertheless there are several procedures of model testing that became part of good modeling practice and should be certainly encouraged. Ironically in various applications you may find the names for these processes used interchangeably, which can only add to the confusion. Model testing is probably a more neutral and general term.

One way to test the model is to compare its output with some independent data set, which has not been used previously for model calibration. This is important to make sure that the model output is not an artifact of the model formalization, and that the processes in the model indeed represent reality, and are not just empirical constructs based on the calibrated parameters. This process is called validation. There is no agreed procedure of model validation (verification in some texts), especially when models become complex and difficult to parameterize and analyze.

One way is to run the model for spatial or temporal domains that were not used for building the model. We can run the model for places and time periods, for which we either did not have data, or deliberately set that data aside and have not used it for model calibration. We may have the luxury to wait until the new data sets are acquired, making our predictions first and then comparing them to what we measure. Or we set aside a part of the data set that is already available and pretend that we do not know it while constructing the model. Then, when the model is built and calibrated based on the remaining data, we will want to bring the other portion of data into light and see if we have equally well matched this other data set. This time we do not do any calibration, we do not tweak model parameters or functions, we only compare and estimate the error model. If the error is small, we may conclude that the model is good, and may have certain confidence in applying the model for future predictions.

In reality, unfortunately, it rarely happens like this. First of all, the temptation is too strong to use all the data available when building the model. As a result we usually do not have sufficient data sets for a true validation. Besides, even when the validation is undertaken, in most cases it proves to be less accurate than the calibration and therefore the researcher is likely to jump into model modifications and improvements to make the validation result look better. However this immediately defeats the purpose of validation. Once you started using the validation data set for model adjustments you have abandoned your validation attempts and went back to further calibration.

Actually this became quite standard in many on-going modeling projects and is called data assimilation. Special procedures are designed to constantly update and improve models based on the incoming flow of new experimental data. This becomes crucial for complex open system (which is most usually the case for ecological and socioeconomic systems), which is always changing and evolving. As a result the data set considered for calibration and collected during one period may not be representing quite the same system as the one that produced the other data set that is intended for validation. We might be calibrating a model of one system, and then trying to validate the same model but for a different system.

Another important step in model analysis is *verification*. A model is verified when it is scrupulously checked for all sort of internal inconsistencies, errors and bugs. These can be in the equations chosen, in the units used, or in links and connections. These can be simply programming bugs in the code that is used to solve the model on the computer. They may be conceptual, when wrong data sets are used to drive the model. Once again, there is hardly a prescribed method to weed them out. Just check and recheck. Run the model and rerun it. Test it and test again.

One efficient method of model testing is to run the model with extreme values of forcing functions and parameters. There are always certain ranges where the forcing functions can vary. Suppose we are talking about temperature. Make the temperature as high as it can get in this system, or as low as it can be. See what happens to the model. Will it still perform reasonably well? Will the output stay within certain plausible values? Or will the model crash? If so, try figure out why. Is it something you can explain? If yes, then probably the model can be still salvaged and you may simply need to remember that the forcing function should stay within certain allowed limits. If the behavior cannot be explained, keep digging—most likely there is something wrong.

Another important check is based on first principles, such as mass and energy conservation. Make sure that there is a mass balance in the model, that nothing gets created from nowhere and nothing is lost.

The bottom line of all this testing is that there is no perfect model. It is hardly possible to get a perfect calibration, and the validation results will likely be even worse. No matter how long you spend debugging the model and the code there will always be another bug, another imperfection. Does this mean that this is all futile? By no means! As long as you get new understanding of the system, as long as the model helps communicate understanding to others and helps manage and control the system—you are on the right path, and the efforts will be fruitful. *Any model that is useful—is a good model.*

See also: Ecological Data Analysis and Modelling: Statistical Inference. Ecological Models: Model Development and Analysis; Model Types: Overview; Structural Dynamic Models. General Ecology: Principal Components Analysis

Further Reading

Bennett, N.D., *et al*., 2013. Characterising performance of environmental models. Environmental Modelling & Software 40, 1–20. Available at: http://www.sciencedirect.com/science/article/pii/S1364815212002435. Accessed 18 December 2012.

Beven, K., 1993. Prophecy, reality and uncertainty in distributed hydrological modelling. Advances in Water Resources 16, 41–51.

Haefner, J.W., 1996. Modeling biological systems: Principles and applications. New York: Chapman and Hall, 473 pp.

Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. Environmental Modelling & Software 21 (5), 602–614.

Loehle, C., 1987. Errors of construction, evaluation, and inference: A classification of sources of error in ecological models. Ecological Modelling 36, 297–314.

Oreskes, N., Shrader-Frechette, K., Belitz, K., 1994. Verification, validation and confirmation of numerical models in the earth sciences. Science 263, 641–646.

Rykiel, E.J., 1996. Testing ecological models: The meaning of validation. Ecological Modelling 90, 229–244.

Spatial Models and Geographic Information Systems☆

Arnab Banerjee and Santanu Ray, Visva-Bharati University, Santiniketan, India

Introduction

Models are representation of investigated objects for purposes of description, explanation, simulation, or forecast. Spatial models are complicated by the fact that they include information about position, possible topological connections and attributes of the recorded objects. Spatial modeling of ecological phenomena has always been an important issue in ecology. The mutual influence of patterns and processes of ecosystems are manifested in the spatial distribution of ecosystems at different scales, which has long been the main component of spatial models. A spatial model is a mechanism for assembling spatial knowledge from a range of sources and presenting conclusion based on that knowledge in readily used form. Spatial modeling is the process of constructing a model incorporating space, which can be identified into areal interpolation and surface modeling.

Areal interpolation is the transformation of data between different sets of areal units. The set of zones, for which data are available, are termed source zones. The second set of zones, for which estimates need to be derived, are termed target zones. The third set of zones, for which auxiliary information can be incorporated in the interpolation process, are termed control zones. The methods of areal interpolation based on alternative hypotheses include radially symmetric kernel functions, maximally smooth estimation, piecewise approximation, uniform target-zone densities and uniform control-zone densities.

Surface modeling is the process of numerically representing a planetary surface (Earth or other planets) by grids with known coordinates in an arbitrary coordinate system. Surface modeling is aimed at formulating an ecological object in a grid system, in which each grid cell contains an estimate of the ecological object that is representative for that particular location. Representing data in grid form have the following advantages: (1) regular grid can be easily reaggregated to any areal arrangement required; (2) producing ecological data in grid form is one way of ensuring compatibility between heterogeneous data sets; (3) data in grid form make multiresolution and multisource information fusion easier; and (4) converting data into grid form can provide a way of avoiding some of the problems imposed by artificial political boundaries.

Spatial modeling occurred in the 1960s with general availability of computers, but the tools offered by current geographical information systems (GIS) have appeared to be of little interest for spatial modeling because GIS has been restricted to producing cartographic products rather than spatial models.

☆*Change History*: April 2018. A Banerjee and S Ray updated this article. Figs. 1–4 are newly added, sections "Introduction, Geographical Information System, Historical Development of GIS, Direct Modeling Problem, GIS Data Type, Summary (highly modified)"were updated and "Traditional Approaches, Some Recent Applications of Spatial Modeling (and corresponding subsections), Further Reading" were new sections.

Geographical Information System

There are many definitions of GIS. For instance, GIS can be defined as a powerful set of tools for collecting, storing, retrieving at will, transforming and displaying spatial data from the real world (Burrough, 1986); as an information technology storing, analyzing, and displaying both spatial and nonspatial data (Parker, 1988); a set of computer-based systems for managing geographic data and using them to solve spatial problems (Lo and Yeung, 2003) and also GIS can be defined as a tool for performing operations on geographic data that are too tedious or expensive or inaccurate if performed by hand (Longley et al., 2011). In general, GIS integrates hardware and software to capture, organize, and display geographically referenced data thus allowing for inquiring, interpreting, and imagining the data through visualization that helps in understanding relationships or patterns in the form of maps, charts, and reports.

Historical Development of GIS

The origin of GIS as a computer-dependent application can be traced back to the 1940s and the 1950s when successful data storage, management and processing techniques were successfully implemented. In the early 1960s, R. F. Tomlinson conceptualized the first GIS—Canada GIS—to address the needs of land and resource information management of the federal government of Canada, which became operational much later in 1971. In 1963, H. H. Fisher took advantage of computational techniques to make simple maps by printing statistical values on a grid of plain paper including a set of modules for analyzing data, manipulating them to produce choropleth or isoline interpolations, with the results displayed in many ways using overprinting of line-printer characters to produce suitable gray scales. In spite of the adaptation to advanced computer techniques in 1960s, the cartographers were mostly limited to automatic drafting of maps; fundamental attitudes of traditional cartography gained no evident benefits from the new computer technology.

In the early 1970s, the Swedish Land Data Bank was developed to automate land and property registration; the Local Authority Management Information System and the Joint Information System were developed in Britain and used by local governments to control and monitor land use. By the late 1970s, there had been considerable investments in the development and application of computer-assisted cartography, with hundreds of new computer programs and systems being developed for various mapping applications. Parallel developments in automated data capture, data analysis and presentation in related fields have resulted in emergence of general purpose GIS with the primary focus being map data processing while spatial analysis functionality was rather limited.

Topological principles in cartography (Corbett, 1979)—a milestone of GIS development—by which geographical data can be stored in a simple structure that is capable of representing structure (what they are), position (where they are) and spatial association with one another. In 1982, the release of vector-based ArcInfo GIS software package by Environmental Systems Research Institute, allowed storage of graphic data in topological structure and attributing the same in tabular structure. By the late 1980s, many other GIS software packages were developed by using a similar data model.

In the 1990s, with advances in operating systems, computer graphics technology, data management, increased computer–human interaction and improved graphical user interface design, GIS became multiplatform applications that could be utilized on different classes of computers as stand-alone applications and as time-sharing systems. Development of Web GIS has allowed expensive data and software to be shared. Standardization in interfaces between data programs and other programs has made it much easier to provide the functionality for handling large amounts of data. GIS has been considerably developed in many aspects. Much knowledge on how to set up computer mapping and GIS projects efficiently has emerged. The basic functionality required for handling spatial data has been widely accepted. Although these advances have promoted the considerable development of GIS and the development of GIS has entered the "age of geographical information infrastructure," the basic spatial models used in modern GIS are little different from those of 20 years ago.

Shifting of focus towards emerging conservation paradigms from single species to multispecies approaches all focus on applications of interdisciplinary methodologies including the incorporation of GIS in ecological modeling. Various applications of GIS technologies have shifted the traditional approaches and uses of this to a newer horizon.

The restrictive effects of space–time constraints (Kwan, 2001) on accessibility research originally proposed by (Hägerstraand, 1970) was later incorporated as space–time prisms (Lenntorp, 1976) mapping onto geographical space (Burns, 1979; Villoria, 1989; Dijst and Vidakovic, 2000). Development of space–time accessibility measures utilizing the GIS-based computational strategies and techniques (Kim and Kwan, 2003; Kwan, 1999, 2010; Miller, 2010) and subsequently the development of action space models and detailed examinations of spatial variation of opportunity based choice (Dijst et al., 2002) are two examples of such applications of temporal-spatial framework.

Spatial autocorrelation in ecology is not infrequent and it serves as the base assumption for a multitude of ecological theories and models (Legendre and Fortin, 1989) describing the positive autocorrelation in species abundance. Two different approaches incorporate space into ecological analysis (Legendre, 1993) which are: "raw data approach" that describe the relation of a species with its environment (partial regression—univariate when considering individual species) or "constrained ordination" for community analysis (multivariate case) (Borcard et al., 1992; Legendre and Legendre, 1998). A combination of ecological distances between sampled data with spatial distribution (obtained through GIS technique interpolation) can be represented by a "matrix approach" that combines the distribution of species with its environment and geographical distances and boundaries.

The crucial role of remote sensing in mapping the variations of areal extent and spatial patterns of mangroves has been observed in the late 1990s and also in early 2000s (Heumann, 2011) allowing for detailed characterization of the same in order to understand and contribute to proper management strategies. In spite of the vast applicability of remote sensing, it had been—in this case—primarily targeted to map the areal extent and pattern change at a local scale. The potential for improved accuracy of classification of land cover as well as characterization of various aspects including leaf/foliage area, canopy structure, species composition and abundance were shown through wide scale applications of GIS in the late 2000s and early 2010s owing to the rapid development of newer sensors and systems (Gillespie et al., 2008; Wooster, 2007).

New systems like LiDAR systems (IcsSAT/GLAS), satellite sensors like VHR systems (very high resolution), for example, Quickbird, IKONOS, GeoEye-1, Worldview-2, ALSO PRISM, and synthetic aperture radar satellites (ALOS PALSAR, ASAR ENVISAT, and also Radarsar Satellites), hyper-spectral airborne visible/infrared imaging spectrometer (AVRIS), TOPSAR and AIRSAR (polymeric SAR) as well as various commercial wave-form LiDAR systems, all have contributed to recent developments of remote sensing technologies and applications. With the potential capability of improved accuracy of classification, species detection and estimation of foliage arrangement (canopy height, leaf coverage, etc.) these newer techniques have contributed to the rapid development in the field of GIS and its application in ecology (Heumann, 2011).

Recent advances in remote sensing techniques and sensor technologies and improved geographical imaging techniques have demonstrated plausible and feasible accurate classification techniques that were previously unavailable with only the handful of traditional techniques at the disposal of the researchers. Newer improved data types can now be gathered with these modern approaches and they can be overlapped with some of the traditional findings in order to facilitate exploitation of various new alleys that were otherwise previously limited to simple cartographical applications. While such data fusion has been achieved in relation to mangrove ecology (Wang and Sousa, 2009) and some more recent implementation to map spatial patterns of carbon metabolism and its relation to landscape indices (FRAC—fractional dimensional index) combining them with mathematical statistics (Xia et al., 2017), these paths can be ventured into more deeply to gain further insight into additional elaborate uses of GIS and relating that to answer an even broader spectrum of ecological problems. In spite of the advances and increase in free access to imagery and data storage, over the recent years, intensive involvement, training and accessibility to required technological infrastructure is required to enable this technology to be used and developed further to address more specific and intricate ecological questions.

Error Problem of GIS

Errors are ubiquitous in current GIS. The word, error, can be used in a variety of senses. Unwin, 1995 defined it as the difference between reality and representation of the reality. Defined in this way, error is related to accuracy. Accuracy is the degree with the values or descriptions of the real-world features that they represent.

When field data are incorporated into a GIS, a common mistake is to assume that the error can be simply equated to the measurement error at the sampled points and quoted as a simple global statement, which does not address the spatial variation. Monckton, 1994 criticized that the spatially uniform error assumption was untenable. A complete specification of the error should include not only the spatial field of its mean but also its variance and spatial dependence. The problems of error and uncertainty in field models have been proved to be much hard-addressed in principle using well-developed theory.

The process of integrating remote sensing data into a GIS usually includes five steps, viz. data acquisition, data processing, data analysis, error assessment, and final product presentation. Error may accumulate throughout the process in an additive or multiplicative fashion. Data acquisition error may be from geometric aspects, sensor system, platforms, ground control, and/or scene considerations. Data processing error may be caused by geometric rectification such as resampling and data conversion such as from raster to vector format and from vector to raster format. Data analysis error may be from classification systems, data generalization, and quantitative analysis of relationships between data variables and the subsequent inferences that may be developed. Error of error assessment might be mainly produced by expression of locational accuracy, discrete multivariate and reporting standards. Final product presentation error includes attribute error and spatial error that may be introduced through the use of base maps with different scales, different national horizontal datum in the source materials and different minimum mapping units which are then resampled to a final minimum mapping unit.

The errors can be distinguished into inherent ones and operational ones. The inherent error is the error present in source documents. The operational error is accumulated through data capture and manipulation functions of a GIS. The inherent errors include errors from sampling and attribute errors in data source. Operational errors include positional errors and identification errors. Positional errors stem from inaccuracies in the horizontal placement of boundaries and identification errors occur when there is a mislabeling of areas on thematic maps. Spatial models as simulations of the real world often simplify the complexity of the real world and are therefore obviously open to errors. The inherent errors can be propagated through the simulation process and become manifest in the final products. Although there are many types and sources of error and uncertainty in geographical data and their processing, the problem is not simply technical and it arises from an evident inability of GIS. Integration of data from different sources and indifferent formats, at different original scales, plus inherent errors, can yield a product of questionable accuracy. Manipulation of thematic overlays within GIS to derive model variables are susceptible to inherent and operational errors, from which results may have such error margins as to be useless for specific applications. Any decision based on such products would thus be flawed.

Multiscale Problem

Ecosystems are constantly changing, not only over space but also in time. The understanding of spatial and temporal processes and their interrelations is central to the understanding of the complex behavior of real ecosystems. Relevant processes might span over several temporal and spatial scales. Therefore, tools for modeling such processes should also be able to operate on diverse spatial and temporal scales.

The issues of ecological modeling are involved at various scales. At each scale, a set of spatially explicit indicators needs to be identified to characterize the extent, pressures, condition, trends and scenarios of ecosystem types and land-use patterns as well as the underlying structural features of ecosystems. For any size patch of the Earth's surface that we choose to define as an ecosystem, there will be a set of factors external to the ecosystem that influence how it functions and in turn, there will be flows of material and energy that extend beyond the ecosystem. The larger the scale, the more inclusive it is of these flows of material and energy. However, studies undertaken at larger scales lose the site specificity that policymakers often need. In other words, there is no single scale at which we can obtain a full understanding of ecosystems.

Scale issue is an inherent part of ecology. While in the early 1950s to the early 1970s, many ecologists tried to incorporate scale in environmental biology in the early 1970s to the 1980s, many ecologists focused increasing attention on the problem of spatial scale. In the 1990s, scale problem became the central problem in ecology, for unifying population biology and ecosystem science, and marrying basic ecology and applied ecology.

The explosion of interest in scale has created many methods for scaling. For instance, interpolation brings multiple phenomena measured at different resolutions into a common coordinate grid with a single size. Multiple-variable scaling method simultaneously examines each variable at different scales. Spatially explicit models are simply maps of actual or simulated phenomena to demonstrate scale-sensitive issues. Fractal geometry is used to treat the dependence of various phenomena on scales. Resampling techniques are used to frame samples within a hierarchical framework to assess how scale and sequence of assembly affect ecosystem characteristics. Geostatistical techniques employ knowledge of the spatial covariance to produce a spatial model. Neural models are developed to test scale effects resulting from changes in grain size and spatial structure. Hierarchy theory is employed to address issues of spatial scale, which implies that an ecosystem is composed of interacting components and is itself a component of a larger system. However, they are not generalized in GIS as module.

Real-Time Problem of GIS

Time can be characterized as the fourth dimension of the physical space–time continuum. From the human point of view, a concrete system can move in any direction on the spatial dimension, but only forward on the temporal dimension. Static objects can be defined as objects that do not change in a short time period. GIS systems generally deal with static information. However, in many situations, the information in GIS applications does change dynamically. Quite often, it is desirable to combine static information with dynamic information. Studies show that major impediments to the analysis of spatial data arise from a lack of well-documented methods in terms of error accumulation; errors that may occur due to the static representation of dynamic ecosystem components suggest that a real-time method must be developed. Methods of assessing the accuracy of dynamic images are also inadequate and must be further researched.

Real time means momentary, that is, the same moment as it happens. In real-time systems, this implies momentary updates. However, it is impossible to get momentary updates, there is always some delay. The acceptable delay length for a real-time system depends on how dynamic the processes are and how time-critical the decisions are. Rapid development of computing technology in recent years has enabled real-time spatial analysis and real-time data visualization to become realizable, although current GIS software and interfaces do not encompass the set of technical and real-time functions.

GIS provides powerful functionality for spatial analysis, data overlay, and storage. These spatially oriented systems lack the ability to represent temporal dynamics and their concepts of ecosystems are static. In other words, GIS prefers a static view and generally lacks the representation of dynamics. The current generation of commercial GIS is unable to facilitate real-time decision making without significant modifications or integration with external models.

In general, the technology of digital geographies has found that the representation of change in time is extremely hard to handle. GIS today remains a technology for static data, which is a major impediment to its use in spatial modeling.

Direct Modeling Problem

Some researchers have noted that GIS has been restricted to producing cartographic products rather than spatial modeling. GIS was conventionally developed using a hybrid approach that handled graphical and descriptive geographical data separately. This georelational data model was the norm for GIS implementation until the late 1990s. GIS was usually used as means of overlaying maps. Almost all mathematical models are too complex to be run directly from present state-of-the-art GIS. They often run outside the GIS. In these cases, GIS is used to supply the input data at an appropriate resolution and to display the results graphically in combination with other relevant spatial data.

Integration of GIS and simulation models can be categorized into loose coupling and deep coupling. Most integration is in the loose coupling category that integrates GIS with simulation models through exchanging data files. This approach often requires human intervention, which can become a barrier in automating the operation process. The deep coupling approach

links GIS and simulation models with a common user interface, in which GIS and simulation models can remain, in fact, separate systems.

A management system of ecological modelbase (MSEM) with 3055 models has been developed in order to find a solution for direct modeling problem. Usually, there are two ways to develop MSEM. One is the model management techniques, including database approach, structured modeling approach, object-oriented approach, and knowledge-based approach. Another one is model management in GIS software such as ModelBuilder in ArcGIS. The first way can efficiently manage the models, but requires great code creation to handle spatial data. The second way can utilize GIS to manipulate spatial data, but it does not support building complex mathematical models. Therefore, an object-oriented framework for MSEM is developed, in which models are abstracted to model class and model instance. Model class and model instance are represented as objects. Spatial data and mathematical equation are parsed by Model Engine that is composed of mathematical library and SMTS (Satellite Modem Termination System) component. Integration of SMTS and MSEM would solve the direct modeling problem existing in current GIS. SMTS finds solutions to the error problem, real-time problem, direct modeling problem, and multiscale problem of the current GIS. However, SMTS involve huge computation cost and very slow computational speed because it must solve a partial differential equation set for simulating each lattice of a surface, which has limited wider application of SMTS.

GIS Data Type

One aspect of GIS that is of utmost importance is the type of data that can be generated for the specific purpose that a study demands. These data may be of topographical or topological nature. The topographical data or topography is generated through digital elevation models (DEM) that can describe some spatial information and topological data or simply topology, that can use terrain attributes for describing spatial distribution.

Fig. 1 below shows the representation of a contour map (structure above) that is used for describing the surface topology and topography (structure below).

GIS data can be of the following types, viz. raster or grid based data, triangular irregular networks or vector- or contour-based line networks. Fig. 2 below shows examples of (A) Grid and (B) TIN representation of topographical data.

The grid based approach utilizes grid cell or raster information and is made up of regularly spaced lines that represent a collection of small rectangles containing dots that characterizes the central coordinate—based upon which, each of these rectangles (area) are defined. An example of this is the raster-based GIS of the Geographic Resource Analysis Support System or GRASS (Pentland and Cuthbert, 1971).

TIN or triangular irregular networks (Fig. 2 b) relies upon determination of significant peaks and valley points that are translated into a collection of irregularly spaced points connected by lines producing a patchwork of triangles (planar facets) based on several algorithms. Delaunay triangles are one of the most widely used techniques and ARCINFO is one of the most commonly used commercial systems alongside ADAPT—Areal Design and Planning Tool (Grayman et al., 1975).

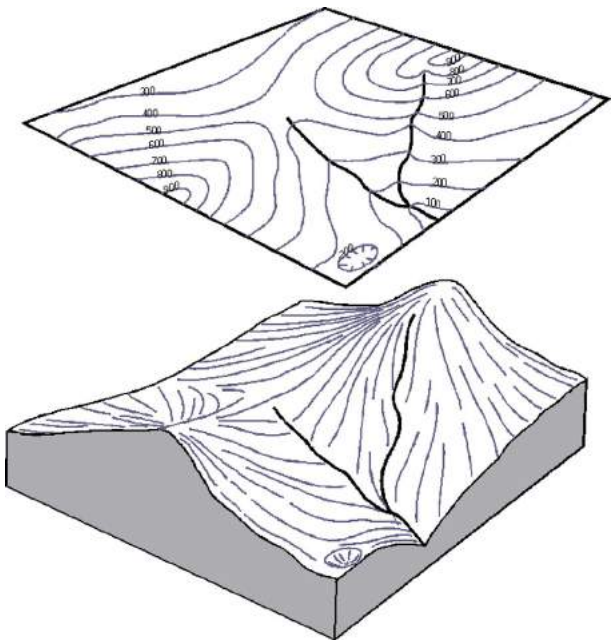


Fig. 1 Example of contour map showing topographic map and corresponding surface. Source: www.cita.utoronto.ca/~murray/GLG130/Exercises/EXE2html.html.

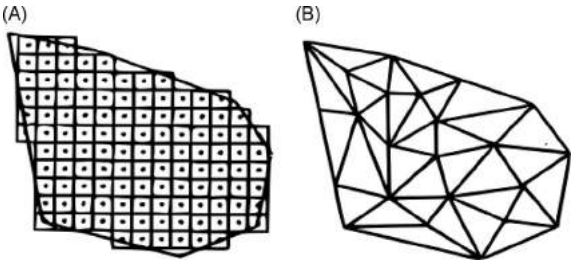


Fig. 2 Representation of Grid (A) and TIN (B) (DeVantier and Feldman, 1993).

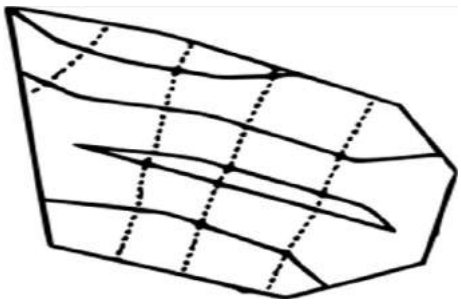


Fig. 3 Representation of vector-based or contour-based diagram (DeVantier and Feldman, 1993).

A contour- or vector-based line network diagram (Fig. 3 below) on the other hand uses digital representation of a point-to-point set (vectors) that can be stored as a digital line graph or DLG. These require a greater magnitude of data as compared to the above two methods but are useful in presenting inherent important attributes (Moore et al., 1991, 1988).

Traditional Approaches

Traditional approaches such as aerial photography (AP) and legacy high resolution systems like Landsat and SPOT are the most commonly used approaches (Newton et al., 2009) for mapping and assessment of mangroves.

Fine grain AP can be successfully used to detect and map individual species as shown by the work of Dahdouh-Guebas et al. (2006). Limited areal extent, relatively high costs of data acquisition over large geographic areas and the possible inconsistencies seen in data collected such as uneven brightness and parallax distortion are the main limitations of using AP. However, these drawbacks can be overcome by the use of satellite based remote sensing. In fact, high resolution satellite imagery (i.e., spatial resolution between 5 and 100 m) such as Landsat (MSS, TM, or ETM+), SPOT (HVR, HRVIR, or HRG), ASTER, or IRS (1C or 1D) have been used by Béland et al. (2006), Benfield et al. (2005), Al Habshi et al. (2007), and so on.

Techniques such as visual interpretation, hybrid classification, unsupervised/supervised classification, and so on have been used to detect and delineate mangroves.

Application of multispectral imagery that includes preprocessing steps such as spectral transformations such as principal components analysis (PCA) or tassel-cap transformation (Crist and Cicone, 1984) or spectral vegetation indices such as normalized difference vegetation index (NDVI) or simple ratio (SR) is another common approach for the classification of mangroves.

In recent times authors such as Gillespie et al. (2008) and Wooster (2007) developed new types of satellite sensors such as very high resolution (VHR) systems (e.g., Quickbird, IKONOS, GeoEye-1 Worldview-2, and ALOS PRISM), synthetic aperture radar systems (e.g., ALOS PALSAR, ASAR ENVISAT, and the Radarsat satellites), and LiDAR systems such as IceSAT/GLAS. Airborne sensors like the hyperspectral airborne visible/infrared imaging spectrometer (AVIRIS), TOPSAR, and AIRSAR (Polarmetric SAR) and various commercial wave-form LiDAR systems have been used to demonstrate the potential for satellite-based sensors. New analysis techniques like the object-based image analysis (OBIA), and image texture metrics, such as lacunarity, use spatial information to improve image classification that has applications in both modern and traditional remote sensing imagery.

Some Recent Applications of Spatial Modeling

Depending upon the capabilities of current GIS systems and owing to the vast availability of data, question arises as to how these are to be represented while evaluating accessibility. Methodological progresses alongside the evolution of theoretical arguments relate to acute perception and explanation of underlying dimensions about the investigated entity thus presenting refinements in the

analysis of spatial interactions and also various facets of its applications. With the increasing dissemination of GIS techniques and advances in statistical methodologies, these advances are becoming more accessible to the common mass (Fischer and Reismann, 2002; Congdon, 2000).

Transportation

Several workers have dealt with applications of accessibility research in transportation and network analysis through the use of matrices measuring connectivity. For example, evaluation of public transit by the use of location–allocation models alongside coverage models focuses on ensuring increased coverage as well as time efficiency and spacing (Wirasinghe and Ghoneim, 1981; Gleason, 1975). Such accessibility research delves in examining not only the geographical distribution and movement of individuals, spatial interaction and population potential, but also patterns of friendship and transmission nets (Cohen and Barabási, 2002; Kwan et al., 2003).

Mangrove Ecosystem

Remote sensing plays a crucial role in the study of mangrove ecology, its management and conservation through the mapping of areal extent and recognizing pattern changes at local scales. Mangroves have unique features like wide dispersal, fast rates of growth, where light acts as a limiting resource and the terrestrial trees have uniform crown shape and prolonged flowering period. These systems provide an wide array of ecosystem services like carbon sequestration, biodiversity support, filtering out pollution and also have the potential to reduce impacts of natural calamities like tsunamis and hurricanes (Alongi, 2002). Hence several studies over the years have been dedicated towards the successful and proper understanding of such systems. Remote sensing plays a critical role in tracking anthropogenic deforestation, impact of natural calamities, effects of conservation projects like reforestation initiatives and also coastal dynamics (Giri et al. 2007; Doyle et al., 2009; Al Habshi et al. 2007; Sirikulchayanon et al., 2008; Lee and Yeh, 2009).

Mapping and Extent

Newer types of imagery like VHR, Hyperspectral Imagery, SAR, etc. helps in overcoming the limitations of multispectral remote sensing in terms of spatial resolution or spectral resolution of sensors, or the inability of optical sensors to penetrate cloud cover.

- VHR imagery like Quickbird or IKONOS is able to reduce the number of mixed pixels.
- Hyperspectral imagery such as HYPERION can potentially detect fine differences in spectral signatures.
- SAR imagery from sensors such as Radarsat or ASAR ENVISAT can penetrate cloud cover.

Even though VHR is capable of and has been used to map mangrove extent, authors such as Giri et al.(2007) and Howari et al. (2009) have used less expensive and coarser resolution imagery over a large area and checked its accuracy using VHR to map a smaller geographic area.

Approaches such as an object-based image analysis (OBIA) or a data fusion to integrate different types of data have been recently developed to improve the accuracy of mapping the extent or detecting the changes over time of mangrove. As its name suggests OBIA uses objects (a group of pixels) instead of individual pixels for image analysis. The pixels are grouped based on image properties or GIS data through an image segmentation process.

Species Composition

Based on biotic and abiotic factors, species in a mangrove exhibit strong zonation patterns which can be used as indicators of geomorphic and environmental changes (Souza Filho and Paradella, 2005). VHR and hyperspectral imagery are among a number of other methods that has been used recently to map mangrove species. Sensors such as Quickbird and IKONOS are used almost exclusively along with satellite-based VHR because of their long-mission life and substantial archived imagery. The spectral information available from Quickbird and IKONOS is limited to the blue, green, red and near-infrared bands which is similar to those of Landsat TM or ETM+. However, due to the very high spatial resolution, the number and effect of mixed pixels may be reduced, which will provide sufficient details for image structure analysis in order to determine canopy structure. Though both sensors are useful for mapping species, IKONOS panchromatic and multispectral data outperformed Quickbird data for texture analysis and MLC, respectively (Wang et al., 2004). Although field studies are limited, lab experiments by Vaiphasa et al. (2005, 2007) indicate that discrimination between multiple species is possible with the inclusion of a genetic algorithm to find the hyperspectral channels that can differentiate between all the species present.

Leaf Area and Canopy Closure

Evapotranspiration, carbon cycling, habitat conditions and forest health can be assessed from bio physical parameters such as leaf area and canopy closure (Kercher and Chambers, 2001; Kovacs et al., 2008). Empirical relationships between ground based

measurements and VHR spectral vegetation indices or SAR backscatter has been utilized to estimate leaf area index (LAI). Kovacs et al. (2004) found strong significant relationships between LAI of red and white mangroves and the simple ratio (SR) and the normalized difference vegetation index (NDVI) using IKONOS; both the indices produced similar results. Spectral vegetation indices from Quickbird sensor produced results similar to the previous IKONOS studies (Kovacs et al., 2009). A stronger relation was observed by Kovacs et al. (2008) between cross polarimetric C-band SAR data and LAI ($r^2 = 0.82$) than the VHR spectral relationships from their earlier studies.

Height and Biomass

Tree and forest biomass estimates can provide a lot of information about the carbon storage and cycling in forests (Litton et al., 2007). Biomass can be estimated

- *Directly* using PolSAR
- *Indirectly* using VHR, SAR Interferometry (InSAR), stereo imagery or LiDAR

Studies by authors such as Lucas et al. (2007), Mougin et al. (1999) demonstrate the potential of SAR to estimate canopy characteristics.

The values and differences of horizontal, vertical and cross-polarizations are used as SAR signal by PolSAR methods to relate with different forest components.

Best estimates of tree height and above ground biomass are given by P-band PolSAR; however, HV polarization of L-Band SAR performs quite well.

Fatoyinbo et al. (2008) and Simard et al.(2008, 2006) demonstrated that Shuttle Topographic Radar Mission (SRTM)—a globally available InSAR digital surface model, provides reasonable estimates of mangrove canopy heights. Though SRTM DSM can be calibrated using field measurements (Fatoyinbo et al., 2008; Simard et al., 2008), vertical canopy structure can be better characterized by air borne LiDAR (Simard et al., 2006) or space-borne LiDAR from IceSAT/GLAS (Simard et al., 2008).

Health Service Access

Extensive use of GIS has been focused on in the health sector for more than a couple of decades to acquire information on the spatial patterns of disease and their correlation with environmental factors, patterns of health service provisions and planning new facilities for making these services more available to the common mass, through various techniques like spatial clustering, and standard GIS functionality like buffering, overlay analyses and network analyses (relying of catchment generation at physical distances of providers and recipients and also the distribution of patients) (Higgs and Gould, 2001; Gatrell and Senior, 1999). Fig. 4 (Higgs, 2004) shows the regions of accessibility to healthcare facilities in Wales and is an example of how to use these kind of study data to potentially improve management, utilization and proper upgradation of such services.

Summary

Simultaneous treatments of issues arising in the representation, selection of methodology and last but not the least—application of GIS techniques have not yet been fully resolved and thus present some form or the other dependability issues that hinder a wider acceptance and successful utilization of GIS technology.

Recent advances have demonstrated an improvement of accuracy with respect to classification, estimation, mapping and measurements. Constant development of comparatively newer imagery techniques like, for example, VHR and SAR generates newer types of data that can stand out on its own as well as play along nicely with some more orthodox traditional approaches. Like VHR and SAR, another method that outperforms its predecessor is OBIA—generating much more accurate data rather than the pixel-based, raster form classification of the past. Classification on the basis of hierarchical rule based systems can be used in conjunction with this new methodology.

Advanced Land Imager (ALI) similar to Landsat TM and ETM sensors with additional blue, NIR and SWIR bands and also HYPERION (hyperspectral HYPERION has 220 bands in the visible, NIR and SWIR spectra) on the EO-1 platform can produce images that—though coarser—can be used in conjunction with Landsat-compatible sensors to detect changes, mapping individual species, estimate photosynthetic activities of a forest and also determine its health.

A good number of traditional GIS methods for conventional ecosystem studies have yet to be explored more deeply in order to expand the range of their usage; for example, spectral unmixing techniques that are used in terrestrial forest studies can be extended to mangrove systems as well in order to get a better idea and explanation of the role of background members (soil and water).

Technological advancements have led to the launch of more than a few up-to-date cutting-edge sensors like ALOS PALSAR, PRISM, Radersat-2, Worldview-2, and some others, that presents newer opportunities to map and present data with additional resolution that allows a researcher to delve deep into newer frontiers that were previously unexplored due to lack of technology or due to the excessive cost involved in such methodology.

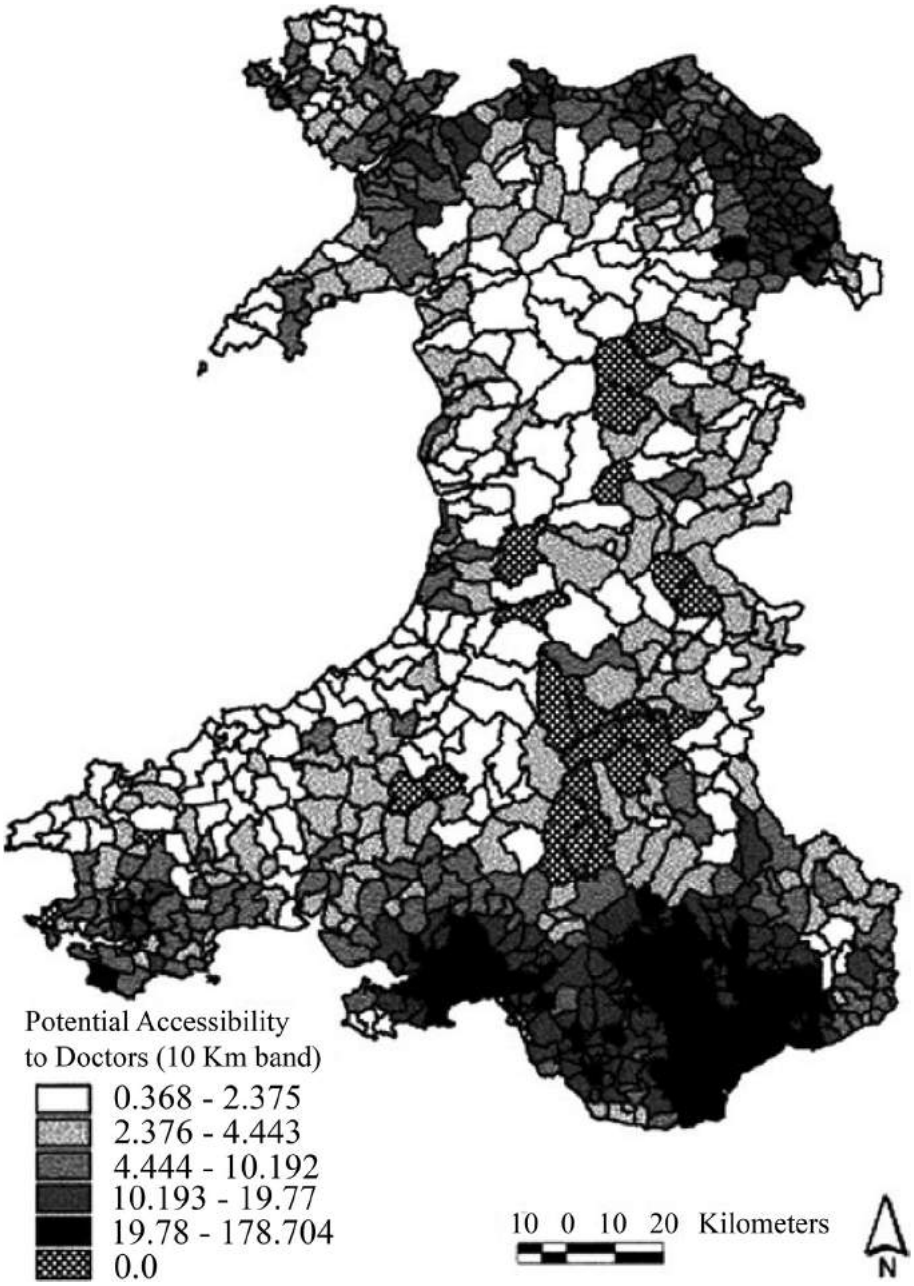


Fig. 4 Potential accessibility to healthcare services in Wales. From Higgs, G. (2004). A literature review of the use of GIS-based measures of access to health care services. *Health Services and Outcomes Research Methodology*, **5**(2), 119–139.

Similarly, newer and advanced VHR sensors like GeoEye-1, having four multispectral bands and higher resolutions provides fresher perspectives and improved opportunities allow further investigation into ecosystems using image texture.

Climate change models has seen the development and use of STMS and YUE-HASM techniques that have considerably reduced computational time and have also improved accuracy and this might be advanced to meet the real time problems of direct modeling and also multiscale problems.

These technological advancements coupled with improved data integration techniques has improved classification accuracy though further in-depth research should be focused in these directions so as to eliminate erroneous occurrences and also to increase efficiency in storing, processing data and analyzing the outputs. Further improved model building strategies and rigorous application of the same is required to expand the horizon on this frontier of research.

References

Al Habshi A, et al. (2007) New mangrove ecosystem data along the UAE coast using remote sensing. *Aquatic Ecosystem Health and Management.* 10: 309–319. Available at: http://www.tandfonline.com/doi/abs/10.1080/14634980701512525.

Alongi DM (2002) Present state and future of the world's mangrove forests. *Environmental Conservation* 29(3): 331–349.

Béland M, et al. (2006) Assessment of land-cover changes related to shrimp aquaculture using remote sensing data: A case study in the Giao Thuy District, Vietnam. *International Journal of Remote Sensing* 27(8): 1491–1510. Available at: http://www.tandfonline.com/doi/abs/10.1080/01431160500406888.

Benfield SL, Guzman HM, and Mair JM (2005) Temporal mangrove dynamics in relation to coastal development in Pacific Panama. *Journal of Environmental Management* 76(3): 263–276. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0301479705001027.

Borcard D, Legendre P, and Drapeau P (1992) Partialling out the spatial component of ecological variation. *Ecology* 73(3): 1045–1055.

Burns LD (1979) *Transportation, temporal, and spatial components of accessibility.* Lexington, MA: Lexington Books.

Burrough PA (1986) Principles of geographical information systems for land resources assessment. *Geocarto International* 1(3): 54.

Cohen EA and Barabási A-L (2002) Linked: The new science of networks. *Foreign Affairs* 81(5): 204. Available at: http://www.jstor.org/stable/10.2307/20033300?origin=crossref.

Congdon P (2000) A Bayesian approach to prediction using the gravity model, with an application to patient flow modeling. *Geographical Analysis* 32(3): 205–224.

Corbett, J.P., 1979. Topological principles in cartography, US Department of Commerce, Bureau of the Census.

Crist EP and Cicone RC (1984) Application of the tasseled cap concept to simulated thematic mapper data. *Photogrammetric Engineering Remote Sensing* 50(3): 343–352.

Dahdouh-Guebas F, et al. (2006) Capacity building in tropical coastal resource monitoring in developing countries: A re-appreciation of the oldest remote sensing method. *International Journal of Sustainable Development & World Ecology* 13(1): 62–76. Available at: https://www.tandfonline.com/doi/full/10.1080/13504500609469662.

DeVantier BA and Feldman AD (1993) Review of GIS applications in hydrologic modeling. *Journal of Water Resources Planning and Management* 119(2): 246–261.

Dijst M and Vidakovic V (2000) Travel time ratio: The key factor of spatial reach. *Transportation* 27(2): 179–199.

Dijst M, de Jong T, and van Eck JR (2002) Opportunities for transport mode change: An exploration of a disaggregated approach. *Environment and Planning B: Planning and Design* 29(3): 413–430.

Doyle TW, Krauss KW, and Wells CJ (2009) Landscape analysis and pattern of hurricane impact and circulation on mangrove forests of the Everglades. *Wetlands* 29(1): 44–53.

Fatoyinbo, T.E. et al., 2008. Landscape-scale extent, height, biomass, and carbon estimation of Mozambique's mangrove forests with Landsat ETM + and shuttle radar topography mission elevation data. Journal of Geophysical Research: Biogeosciences, 113(G2), p.n/a-n/a. Available at: http://doi.wiley.com/10.1029/2007JG000551.

Fischer MM and Reismann M (2002) A methodology for neural spatial interaction modeling. *Geographical Analysis* 34(3): 207–228. Available at: http://doi.wiley.com/10.1111/j.1538-4632.2002.tb01085.x.

Gatrell A and Senior M (1999) Health and health care applications. In: *Geographic information systems: Principles and applications*, pp. 925–938. London: Wiley.

Gillespie TW, et al. (2008) Measuring and modelling biodiversity from space. *Progress in Physical Geography* 32(2): 203–221. Available at: http://journals.sagepub.com/doi/10.1177/0309133308093606.

Giri C, et al. (2007) Monitoring mangrove forest dynamics of the Sundarbans in Bangladesh and India using multi-temporal satellite data from 1973 to 2000. *Estuarine, Coastal and Shelf Science* 73(1–2): 91–100. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0272771407000029.

Gleason JM (1975) A set covering approach to bus stop location. *Omega* 3(5): 605–608.

Grayman WM, et al. (1975) Land-based modeling system for water quality management studies. *Journal of the Hydraulics Division* 101(5): 567–580.

Hägerstraand T (1970) What about people in regional science? *Papers in Regional Science* 24(1): 7–24.

Heumann BW (2011) Satellite remote sensing of mangrove forests: Recent advances and future opportunities. *Progress in Physical Geography* 35(1): 87–108. Available at: http://journals.sagepub.com/doi/10.1177/0309133310385371.

Higgs G (2004) A literature review of the use of GIS-based measures of access to health care services. *Health Services and Outcomes Research Methodology* 5(2): 119–139.

Higgs G and Gould M (2001) Is there a role for GIS in the "new NHS"? *Health and Place* 7(3): 247–259.

Howari FM, et al. (2009) Field and remote-sensing assessment of mangrove forests and seagrass beds in the northwestern part of the United Arab Emirates. *Journal of Coastal Research* 251: 48–56. Available at: http://www.bioone.org/doi/abs/10.2112/07-0867.1.

Kercher J and Chambers J (2001) Parameter estimation for a global model of terrestrial biogeochemical cycling by an iterative method. *Ecological Modelling* 139(2–3): 137–175. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0304380001002344.

Kim HM and Kwan MP (2003) Space-time accessibility measures: A geocomputational algorithm with a focus on the feasible opportunity set and possible activity duration. *Journal of Geographical Systems* 5(1): 71–91.

Kovacs JM, et al. (2004) Estimating leaf area index of a degraded mangrove forest using high spatial resolution satellite data. *Aquatic Botany* 80(1): 13–22. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0304377004000841.

Kovacs, J.M. et al., 2008. The use of multipolarized spaceborne SAR backscatter for monitoring the health of a degraded mangrove Forest. Journal of Coastal Research, 241, pp.248–254. Available at: www.bioone.org/doi/abs/10.2112/06-0660.1.

Kovacs JM, et al. (2009) Evaluating the condition of a mangrove forest of the Mexican Pacific based on an estimated leaf area index mapping approach. *Environmental Monitoring and Assessment* 157(1–4): 137–149. Available at: http://link.springer.com/10.1007/s10661-008-0523-z.

Kwan MP (1999) Gender and individual access to urban opportunities: A study using space–time measures. *The Professional Geographer* 51(2): 211–227.

Kwan MP (2001) Cyberspatial cognition and individual access to information: The behavioral foundation of cybergeography. *Environment and Planning B: Planning and Design* 28(1): 21–37.

Kwan MP (2010) Space-time and integral measures of individual accessibility: A comparative analysis using a point-based framework. *Geographical Analysis* 30(3): 191–216. Available at: http://doi.wiley.com/10.1111/j.1538-4632.1998.tb00396.x.

Kwan MP, et al. (2003) Recent advances in accessibility research: Representation, methodology and applications. *Journal of Geographical Systems* 5(1): 129–138.

Lee TM and Yeh HC (2009) Applying remote sensing techniques to monitor shifting wetland vegetation: A case study of Danshui River estuary mangrove communities, Taiwan. *Ecological Engineering* 35(4): 487–496.

Legendre P (1993) Spatial autocorrelation: Trouble or new paradigm? *Ecology* 74(6): 1659–1673.

Legendre P and Fortin MJ (1989) Spatial pattern and ecological analysis. *Vegetatio* 80(2): 107–138.

Legendre, P. & Legendre, L., 1998. Numerical ecology, 2nd edn. Available at: http://linkinghub.elsevier.com/retrieve/pii/B978044453868050016 2%0Ahttp://biol09.biol.umontreal.ca/PLcourses/Statistical_tests.pdf.

Lenntorp B (1976) Paths in space-time environments: A time-geographic study of movement possibilities of individuals. *Environment and Planning A* 9: 961–972.

Litton CM, Raich JW, and Ryan MG (2007) Carbon allocation in forest ecosystems. *Global Change Biology* 13(10): 2089–2109. Available at: http://doi.wiley.com/10.1111/j.1365-2486.2007.01420.x.

Lo CP and Yeung AKW (2003) *Concepts and techniques in geographic information systems: Laboratory manual.* Upper Saddle River: Pearson Prentice Hall.

Longley PA, et al. (2011) *Geographical information systems and science.* New York: Wiley. Available at: http://www.jstor.org/stable/215736?origin=crossref.

Lucas RM, et al. (2007) The potential of L-band SAR for quantifying mangrove characteristics and change: Case studies from the tropics. *Aquatic Conservation: Marine and Freshwater Ecosystems* 17(3): 245–264. Available at: http://doi.wiley.com/10.1002/aqc.833.

Miller HJ (2010) Measuring space-time accessibility benefits within transportation networks: Basic theory and computational procedures. *Geographical Analysis* 31(1): 1–26. Available at: http://doi.wiley.com/10.1111/j.1538-4632.1999.tb00408.x.

Monckton CG (1994) An investigation into the spatial structure of error in digital elevation data. *Innovations in GIS* 1: 201–211.
Moore ID, O'Loughlin EM, and Burch GJ (1988) A contour-based topographic model for hydrological and ecological applications. *Earth Surface Processes and Landforms* 13(4): 305–320.
Moore ID, Grayson RB, and Ladson AR (1991) Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. *Hydrological Processes* 5(1): 3–30.
Mougin E, et al. (1999) Multifrequency and multipolarization radar backscattering from mangrove forests. *IEEE Transactions on Geoscience and Remote Sensing* 37(1): 94–102. Available at: http://ieeexplore.ieee.org/document/739128/.
Newton AC, et al. (2009) Remote sensing and the future of landscape ecology. *Progress in Physical Geography* 33(4): 528–546. Available at: http://journals.sagepub.com/doi/10.1177/0309133309346882.
Parker HD (1988) The unique qualities of a geographic information system: A commentary. *Photogrammetric Engineering and Remote Sensing* 54(11): 1547–1549.
Pentland RL and Cuthbert DR (1971) Operational hydrology for Ungaged streams by the Grid Square technique. *Water Resources Research* 7(2): 283–291.
Simard M, et al. (2006) Mapping height and biomass of mangrove forests in Everglades National Park with SRTM elevation data. *Photogrammetric Engineering & Remote Sensing* 72(3): 299–311. Available at: http://openurl.ingenta.com/content/xref?genre=article&issn=0099-1112&volume=72&issue=3&spage=299.
Simard M, et al. (2008) A systematic method for 3D mapping of mangrove forests based on shuttle radar topography mission elevation data, ICEsat/GLAS waveforms and field data: Application to Ciénaga Grande de Santa Marta, Colombia. *Remote Sensing of Environment* 112(5): 2131–2144. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0034425708000357.
Sirikulchayanon P, Sun W, and Oyana TJ (2008) Assessing the impact of the 2004 tsunami on mangroves using remote sensing and GIS techniques. *International Journal of Remote Sensing* 29(12): 3553–3576. Available at: http://www.tandfonline.com/doi/abs/10.1080/01431160701646332.
Souza Filho PWM and Paradella WR (2005) Use of RADARSAT-1 fine mode and Landsat-5 TM selective principal component analysis for geomorphological mapping in a macrotidal mangrove coast in the Amazon region. *Canadian Journal of Remote Sensing* 31(3): 214–224. Available at: http://www.tandfonline.com/doi/abs/10.5589/m05-009.
Unwin DJ (1995) Geographical information systems and the problem of "error and uncertainty" *Progress in Human Geography* 19(4): 549–558. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-0029518803&partnerID=40&md5=08f5962e94ccb2bfff1ec6b55edee768.
Vaiphasa C, et al. (2005) Tropical mangrove species discrimination using hyperspectral data: A laboratory study. *Estuarine, Coastal and Shelf Science* 65(1–2): 371–379. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0272771405002179.
Vaiphasa C, et al. (2007) A hyperspectral band selector for plant species discrimination. *ISPRS Journal of Photogrammetry and Remote Sensing* 62(3): 225–235. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0924271607000512.
Villoria OG (1989) *An operational measure of individual accessibility for use in the study of travel-activity patterns*. The Ohio State University.
Wang L and Sousa WP (2009) Distinguishing mangrove species with laboratory measurements of hyperspectral leaf reflectance. *International Journal of Remote Sensing* 30(5): 1267–1281.
Wang L, et al. (2004) Comparison of IKONOS and QuickBird images for mapping mangrove species on the Caribbean coast of Panama. *Remote Sensing of Environment* 91(3–4): 432–440. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0034425704001129.
Wirasinghe SC and Ghoneim NS (1981) Spacing of bus-stops for many to many travel demand. *Transportation Science* 15(3): 210–221.
Wooster M (2007) Remote sensing: Sensors and systems. *Progress in Physical Geography* 31(1): 95–100. Available at: http://journals.sagepub.com/doi/10.1177/0309133307073889.
Xia L, et al. (2017) Analyzing the spatial pattern of carbon metabolism and its response to change of urban form. *Ecological Modelling* 355: 105–115. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0304380016307591.

Species Distribution Modeling

Adam Duarte and Steven L Whitlock, Oregon Cooperative Fish and Wildlife Research Unit, Oregon State University, Corvallis, OR, United States
James T Peterson, U.S. Geological Survey, Oregon Cooperative Fish and Wildlife Research Unit, Oregon State University, Corvallis, OR, United States

Glossary

Census Complete count of the number of individuals or species within a sampling unit.
Detection probability (p) The probability a species is detected, given the sampling unit is occupied by the species.
Occupancy probability (ψ) The probability a sampling unit contains at least one individual of a species.
Species distribution The spatial arrangement of a species across a landscape.
Species distribution model A spatially explicit, quantitative model that relates occurrence data to landscape characteristics for a focal species.

Introduction

Understanding factors governing spatiotemporal heterogeneity in the distribution of species is a fundamental task in ecology. Quantifying how species occurrence is related to environmental factors using species distribution models (SDMs) enables ecologists to make model-based predictions of species occurrence (Fig. 1) and contribute to the fields of biogeography, evolutionary ecology, invasive species ecology, conservation biology, and natural resource management. Excellent reviews of the ecological and evolutionary processes captured by SDMs are available in the peer-reviewed literature (see Elith and Leathwick, 2009 and citations therein). However, given the utility of SDMs in ecology, it is not surprising that multiple approaches are utilized to leverage different types of data to develop SDMs. In this article, we offer a brief overview of some of the more widely used approaches to develop SDMs, and we organize these methods by the type of data required to fit these models, including presence-only data, presence/absence data, and replicate detection/nondetection data.

Models for Presence-Only Data

The minimum information required for developing an SDM is a series of locations where a species is known to be present, which is often referred to as presence-only data. Presence-only data typically arise when a random subset of the total number of potential sample units is surveyed and only successful species encounters (i.e., positive detections) are recorded. This is a typical data collection process when data are gathered from natural history museum specimens, herbarium records, and opportunistic encounters. Presence-only data can be related to landscape-scale environmental predictors or covariates to develop SDMs when

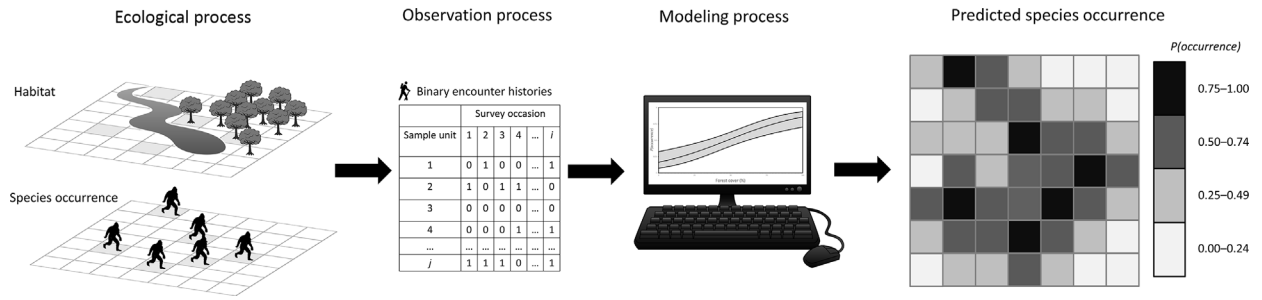


Fig. 1 Illustration of the different components involved when developing species distribution models.

these data are not compromised by sample selection or survey bias. That is, these data need to be collected using a random spatial sample design and patterns in species' detection probabilities need to be constant across sample units—or contain only random heterogeneity—to result in unbiased estimates of species occurrence.

MaxEnt

MaxEnt (e.g., Phillips et al., 2006) is probably the most widely used software package to develop SDMs from presence-only data, despite that it was often referred to as a "black box" method among ecologists for years. The lack of clear understanding was probably related to the program being originally developed in a less familiar machine-learning framework. Fortunately, MaxEnt has since been described in great detail using more familiar statistical terminology (see Elith et al., 2011), providing some clarity to the behind-the-scenes processes. Briefly, users input grid-based spatial environmental data and presence-only records into MaxEnt. The program randomly generates background locations, although this can be further specified outside of the default settings. These background locations may or may not overlap with the presence-only data based on random chance and allow the program to identify the full suite of environmental conditions a species could potentially occupy within the area of interest. MaxEnt examines all pairwise combinations of environmental covariates and the location points, both observed presence and background points, to estimate the conditional probability density of covariates at each observed presence location ($f_1(z)$) and the unconditional probability density of covariates at each background point location ($f(z)$). MaxEnt minimizes the relative entropy between the two probability densities as:

$$f_1(z) = f(z)e^{\eta(z)}$$

where $\eta(z)$ represents a log-linear model that contains a vector of environmental covariates and constrains $f_1(z)$ to sum to 1. MaxEnt has proven to be a useful program to develop SDMs from presence-only data. As Merow et al. (2013) noted, a major limitation of MaxEnt applications is that users rarely deviate from the default settings to ensure the application matches the nuances of a particular dataset or species. In response to this issue, they provided ecologists with a thorough overview of the program options and how to decide among model settings. Such information should be reviewed prior to fitting models in MaxEnt. It is also worth noting that estimates from MaxEnt cannot be reliably interpreted as occurrence probability, but instead are interpreted as indices of habitat suitability.

MaxLike

While MaxEnt has been widely used among ecologists, indices of habitat suitability may not be reliable proxies of occurrence probabilities. As an alternative, Royle et al. (2012) developed MaxLike. MaxLike also relates presence-only data to a vector of environmental covariates. What distinguishes MaxLike from MaxEnt is that MaxLike uses a likelihood-based analysis to fit a logit-linear model to link environmental covariates to presence locations, does not generate background location points, and the output estimates are directly interpretable as occurrence probabilities. Interestingly, they demonstrated that estimates from MaxLike are consistent with the estimates of logit-linear models that are fit to presence/absence data (see "Logistic regression model" section), although they noted MaxLike requires datasets with larger sample sizes. Indeed, smaller sample sizes will result in larger uncertainties (i.e., standard deviations) in the model parameter estimates (Fitzpatrick et al., 2013). The relationship between sample sizes and the precision of the estimates naturally leads one to ask how large the sample size needs to be in order to produce accurate estimates of occurrence probabilities when fitting models in MaxLike. To answer this question, Merow and Silander (2014) simulated presence-only data with variable sample sizes, and their simulations suggested MaxLike estimates were biased and imprecise when samples sizes fell below approximately 1000. Nevertheless, when rescaling the values to relative occurrence probabilities their simulations suggest that MaxLike produced accurate estimates for much smaller sample sizes (approximately 200).

Fitting Models for Presence-Only Data

Given their availability, it is not surprising presence-only data are frequently used to develop SDMs. The application of MaxEnt and MaxLike to analyze presence-only data is continuously increasing in the SDM literature, and it should be noted that the use of

MaxEnt versus MaxLike is somewhat of a contentious issue. In particular, the statistical assumptions imbedded in each of the methods are a primary focus in debates (see Royle et al., 2012; Phillips and Elith, 2013). From a practical perspective, comparisons between the estimates from MaxEnt and MaxLike have been conducted. Similar to Royle et al. (2012), Fitzpatrick et al. (2013) found MaxLike outperformed MaxEnt based on several evaluation metrics. In contrast, Merow and Silander (2014) found a great deal of agreement between MaxEnt and MaxLike after rescaling the estimates from both approaches to relative values. It seems that if data are collected in a manner in which statistical assumptions are met and sample sizes are sufficient, both of these approaches have merits when developing SDMs from presence-only data. Therefore, we encourage ecologists to review these papers, and others, if making the choice between the two methods. Regardless, we stress that presence-only models notoriously produce biased SDMs if sample selection or survey biases are existent in the data and not accounted for during analyses. Unfortunately, these issues are prevalent and often ignored when developing SDMs from presence-only data. We also stress that absence data, if available, should not be discarded in order to use the modeling approaches introduced in this section. If information on the presence and absence of a species is available, ecologists should consider fitting the models described in the subsequent sections.

Models for Presence/Absence Data

Presence/absence data are also commonly used to develop SDMs. These data transpire when both successful and unsuccessful species encounters are recorded during surveys of a random subset of the total number of potential sample units or by generating pseudo-absence data points and combining them with presence-only data. Similar to models for presence-only data, the approaches described in this section are also only valuable to the development of SDMs when the data are not compromised by sample selection or survey bias. Thus, if presence-only data were collected in such a way that they are compromised by these biases, the newly created presence/pseudo-absence data do not overcome these issues. Indeed, there is no shortcut around establishing a robust sample design prior to data collection to develop reliable model-based predictions for SDMs.

Regression Models

Regression models are a useful approach to develop SDMs from presence/absence data. In linear regression models, a response variable is modeled as a linear function of one or more explanatory variables and parameters. Linear models take the basic form:

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_Z x_{Z,i} \ldots$$

where Y is the value of the response (a.k.a. dependent) variable, x_1 and x_Z are the values of the explanatory variables for observation i, β_0 is the intercept, and β_1 and β_Z are the model coefficients. There are several types of linear models that are broadly defined as generalized linear models (GLMs) and the use of each type depends on the characteristics of the response to be modeled. Species presence and absence are binary responses (i.e., 1 or 0), so fitting presence/absence data with a GLM requires the use of a Bernoulli error distribution and the predicted response is a probability of presence or occurrence. Although linear models require the likely violated assumption of perfect (100%) detection, the linear modeling concepts discussed in this section form the basis for more sophisticated models that explicitly model species detection and occurrence (see "Models for Replicate Detection/Nondetection Data" section).

Logistic regression model

Linear regression approaches that model species probabilities of occurrence must produce predictions that are bounded by zero and one (i.e., probabilities cannot be negative and cannot exceed one). To keep probabilities within these bounds, GLMs with a Bernoulli error distribution generally employ one of two link functions: the log-log link and the more widely used logistic link that takes the form:

$$\eta = \ln\left(\frac{p}{1-p}\right)$$

where η is the log odds of presence, ln is the natural logarithm, and p is the probability of presence. The linear logistic regression model (a.k.a. logit-linear model) is then:

$$\eta_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_Z x_{Z,i}$$

where the terms are described above. For logit-linear models, the linear combinations of variables are on a log-odds scale and require the use of odds ratios (e^{β_Z}) or scaled odds ratios ($e^{\beta_Z x'_Z}$, where x'_Z is a unit scalar) to interpret the magnitude of the effect size of each parameter (Hosmer and Lemeshow, 1989). To estimate presence probabilities ($\hat{p}$) with logit-linear models, the predicted log odds of presence ($\hat{\eta}$) are calculated using the linear combination of parameter estimates and corresponding explanatory variables and back transformed using the inverse logit link function as:

$$\hat{p} = \frac{1}{1+e^{-\hat{\eta}}}$$

where the terms are defined above. The logit link and logit-linear models are used in a variety of SDMs as shown in the following sections.

Generalized additive model

Generalized additive models (GAMs) combine the properties of two estimators, GLMs and additive models (Hastie and Tibshirani, 1990). The logit-linear GAMs estimate the log odds of a linear combination of variables as:

$$\eta_i = \beta_0 + f_1(x_{1,i}) + \ldots + f_Z(x_{Z,i})$$

where η_i and β_0 are defined above and f_Z is an unspecified smoothing function for explanatory variable Z that can be either parametric or nonparametric. Parametric GAMs generally employ penalized splines for the smoothing function, which can be thought of as a piecewise continuous polynomial function that is delimited by knots. Nonparametric GAMs typically use kernel smoothers to approximate the relation between the log odds and the explanatory variable and also require multiple parameters to define $f_Z(x_{Z,i})$. Thus, each smoothing function in a GAM can use multiple parameters to define the relationship between the log odds and each explanatory variable (Z), whereas a GLM uses a single parameter for each Z (Fig. 2). This suggests that GAMs can account for potential nonlinearities in the estimated relationships; however, it also means that GAMs have a tendency to overfit the data, which reduces the accuracy of the models when applied to data not used to fit the model (i.e., out-of-sample prediction).

Generalized linear mixed models

The logit-linear GLM and GAM are appropriate for examining the relationship between the response and explanatory variables when all observations (i) are independent. Nonindependent errors can arise in any number of ways but, in the context of SDMs, it is often related to spatial and temporal dependence (a.k.a. autocorrelation). For instance when samples are collected within different geographic areas (e.g., bird species detected in multiple riparian sample plots nested within multiple watersheds), samples from the same geographic area may be more similar to each other due to measured or unmeasured factors and may therefore be dependent. Statistical dependence can arise from a variety of factors, including the fact that the sample units share the attributes of the area they are nested within (e.g., same elevation or aspect) and simply due to the fact that they may share a common history. Nonindependent errors can lead to biased low estimates of variance, biased measures of model fit, biased parameter estimates, or all of the above. The incorporation of the error structure (i.e., the dependence) into a linear model is required to account for nonindependent errors. To illustrate, we begin with the logistic regression model from "Logistic regression model" section and assume that J groups (e.g., forested stands) were randomly selected from the entire population of groups and multiple sample units (e.g., amphibian traps) are randomly selected therein. The relationship between the response and explanatory variables in group (j) can be represented by:

$$\eta_i = \beta_{0j} + \beta_{1j}\, x_{1,i,j} + \ldots + \beta_{Zj} x_{Z,i,j}$$

where the variables are defined above. The parameter estimates (i.e., the β_{Zj}) can be treated as fixed, in which their value is assumed equal across groups (e.g., $\beta_{01} = \beta_{02} = \beta_{03}$), which is equivalent to the logistic regression model in "Logistic regression model" section. The parameters can also be treated as randomly varying in which their values differ among groups (e.g., $\beta_{01} \neq \beta_{02} \neq \beta_{03}$). Thus, models with randomly varying parameters differ from the GLM in that each group can have a unique parameter (Fig. 3). Randomly varying parameters are usually assumed to be normally distributed among groups, but other distributions, such as the gamma, can be used to approximate the variability among groups. This means that GLMs with randomly varying parameters can contain mixtures of statistical distributions, which give rise to their name generalized linear mixed models (GLMMs; Stroup, 2012).

Given the GLMM above, the influence of group-level characteristics (i.e., explanatory variables) on the response can be examined by modeling the group-specific parameters (i.e., β_{Zj}) as a function of the group characteristics:

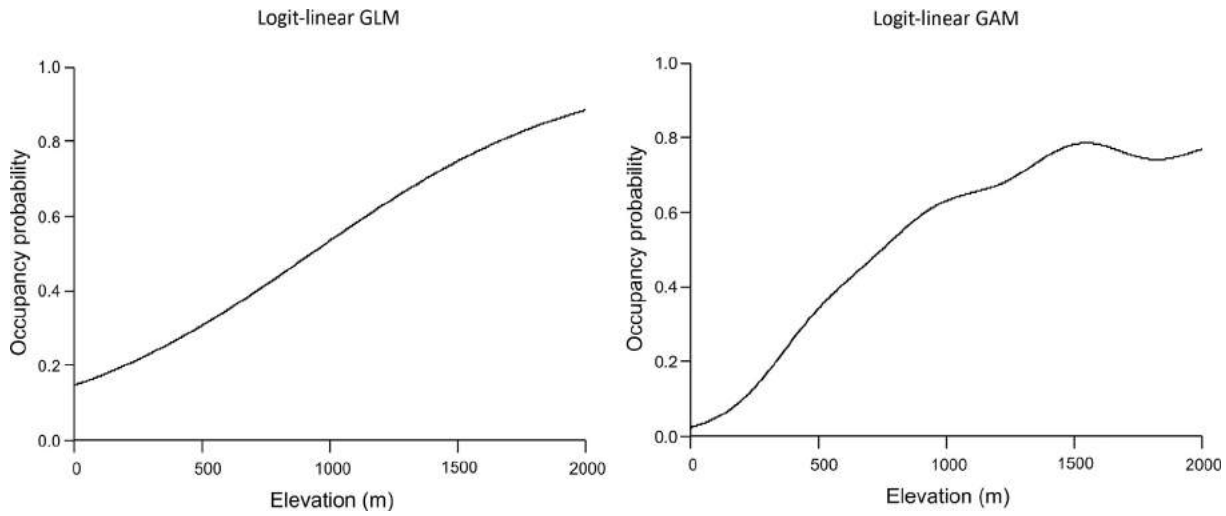


Fig. 2 Comparative illustration between a logit-linear generalized linear model (GLM) and a logit-linear generalized additive model (GAM).

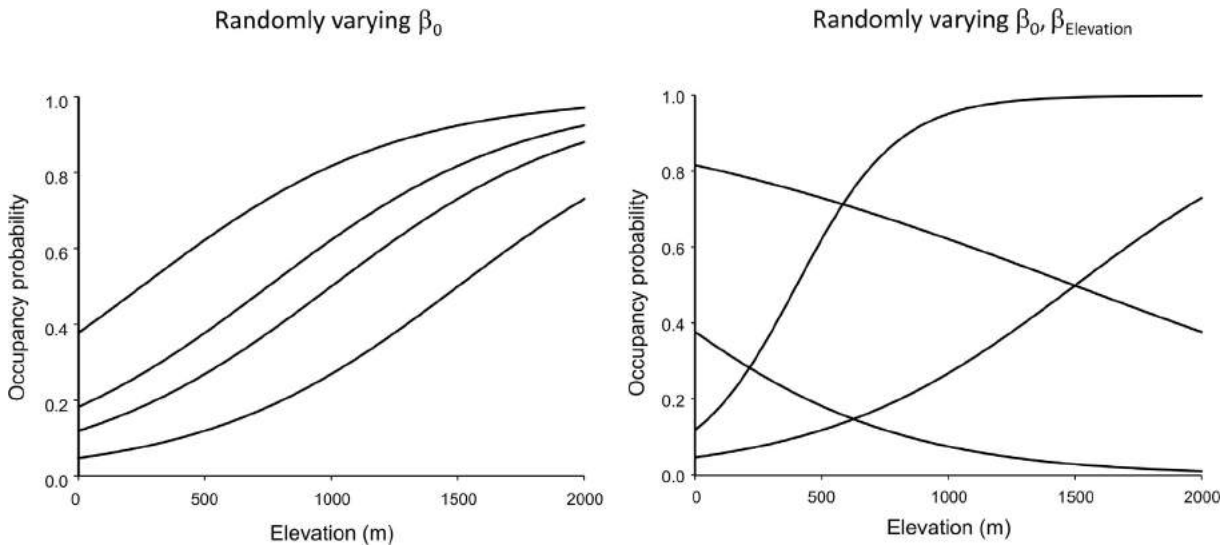


Fig. 3 Illustration of a generalized linear mixed model (GLMM) with randomly varying intercepts (*left panel*) and randomly varying intercepts and model coefficients (*right panel*).

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_{1,j} + \ldots + \gamma_{0S} W_{S,j} + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_{1,j} + \ldots + \gamma_{1S} W_{S,j} + u_{1j}$$
$$\beta_{qj} = \gamma_{q0} + \gamma_{q1} W_{1,j} + \ldots + \gamma_{qS_q} W_{S_q,j} + u_{qj}$$

where $\gamma_{00} \ldots \gamma_{qS_q}$ are the fixed effects, $u_{0j}, \ldots u_{qj}$ are the random effects, and W_{Sqj} are the group-level explanatory variables for group j. The random components $u_{0j}, \ldots u_{qj}$ represent the unique effect associated with each group that is unexplained in the model and is typically assumed to be normally distributed, $N(0, \tau_{qq})$.

The fixed parameters of a logit-linear GLMM can be interpreted using odds and scaled odds ratios. The random effects, however, are more difficult to interpret because of the logit link function and mixture of statistical distributions. One approach is to interpret the fixed effects by conditioning on the random effects (Goldstein, 1995) or averaging over all possible values of the random effects via a "population average" or "marginal" model (Diggle et al., 1994). Alternatively, a median odds ratio (MOR) can be used as a point estimate of the magnitude of the random effect in terms of odds. It has the desirable property of being directly comparable to fixed effects (i.e., odds ratios) and is estimated as:

$$MOR = med\left\{e^{|wv|}\right\}$$

where *med* denotes the median of the distribution and $w = \sqrt{2\tau}$, τ is the variance of the random effect, and v is normally distributed with a mean of 0 and variance of 1 (Larsen et al., 2000).

Occurrence probabilities can be estimated with logit-linear GLMs using the predicted log odds and inverse logit link function as described above. However, the group-specific random effects u_{qj} should be used to predict the log odds for groups that were included in the model fitting dataset. For observations from groups not included in modeling fitting (i.e., out-of-sample prediction), the fixed effects should be used to predict the expected probability of occurrence, but the variation among groups (i.e., the random effect variances, τ_{qq}) should be included in measures of uncertainty (e.g., confidence limits). The latter is best implemented in a Bayesian modeling framework.

Tree-Based Methods

Predictions of species occurrence from tree-based methods are dissimilar to the methods discussed in the previous sections in that they are not generated from an explicit model that was fit to a dataset. Instead, tree-based methods predict occurrence probabilities from an algorithm that has been trained by a dataset through a process known as "learning." Tree-based methods come from the realm of machine learning, which is a relatively young field that combines elements of computer science and statistics (Hastie et al., 2009). This approach relies heavily on the application of algorithms, usually for extracting information from large datasets. A practical difference between machine-learning approaches and the traditional statistical methods common to ecology is in the role that the analyst plays in dictating the potential functional relationships between explanatory and response variables. The predominant approach to data analysis in ecology involves first explicitly defining a model or set of models (each of which is linked to an a priori hypothesis), and then fitting those models to data. Conversely, the ecologist who applies a tree-based method chooses the explanatory variables to include, sets several tuning parameters, and the algorithm automatically identifies influential functional

relationships between explanatory and response variables. Unsurprisingly, there is some philosophical debate between advocates of the two methods, which was briefly introduced in "Fitting Models for Presence-Only Data" section. Arguments in favor of tree-based SDMs are that these methods can reveal patterns ecologists did not initially consider (e.g., thresholds and nonlinear responses; Merow et al., 2014), do not require distributional assumptions (i.e., nonparametric), and commonly outperform traditional regression models based on out-of-sample cross-validation (Elith and Leathwick, 2009). Proponents of the traditional statistical approaches assert that failure to predefine model structures can result in spurious conclusions regarding the importance and effect of habitat variables (Burnham and Anderson, 1998). The case has also been made that better out-of-sample prediction does not imply that the model is generalizable (Wenger and Olden, 2012). Debates aside, a simpler deciding factor for whether to apply tree-based methods versus regression methods is the relative importance that one places on prediction versus explanation of species-habitat associations for a particular study (Elith and Leathwick, 2009). Tree-based models are better for prediction of species occurrence at a particular place and time, but interpreting the specific effect of habitat variables is more difficult.

Classification trees

One of the simplest tree-based methods is a classification tree (De'ath and Fabricius, 2000). Classification trees are used to predict the class (i.e., a categorical value) to which a response belongs based on the values of the associated explanatory variables. SDMs typically use classification trees (rather than regression trees that predict continuous responses) because the desired prediction is whether or not the species is present in a defined area. A classification tree is built by sequentially identifying values of explanatory variables that split responses into the most homogenous possible groupings, which are known as nodes. The algorithm decides on the variable to split and the placement of the split by maximizing a value known as the Gini index. The Gini index is a measure of the purity of each grouping that results from the splits, where a grouping that contains only a single class is considered pure. The index also rewards splits that divide the data into larger groupings. The algorithm carries out additional splits until the data are partitioned such that all of the terminal nodes contain responses from only a single class (Fig. 4). Once a tree is created, it can be used to classify independent data with the same explanatory variables. Although, in practice the distal nodes of the tree are usually "pruned" or "trimmed" to avoid overfitting. While the process of performing binary partitions of data appears simple, the resulting classification tree can represent elaborate nonlinear relationships and interactions between explanatory variables and responses. Also, classification trees do not produce parameters that describe the shape of the functional relationships in the same manner as a traditional regression model (e.g., intercepts and coefficients). Although classification trees have been used to develop SDMs, more advanced tree-based methods that have better predictive performance are typically the focus in the SDM literature. These newer tree-based methods improve upon the classification tree approach by generating multiple classification trees with some built-in randomization of variables to minimize poor predictions caused by overfitting. Approaches that work by aggregating information across multiple trees are referred to as ensemble methods and two widely implemented methods that fall into this category are random forests ("Random forests" section) and boosted trees ("Boosted trees" section).

Random forests

Random forests are an extension of classification trees and have been increasingly used to develop SDMs in the last decade (Cutler et al., 2007). The algorithm relies heavily on bootstrap aggregation, which is commonly referred to as "bagging" in the machine-learning literature. Bootstrapping is the practice of iteratively sampling from the observed data with replacement to create

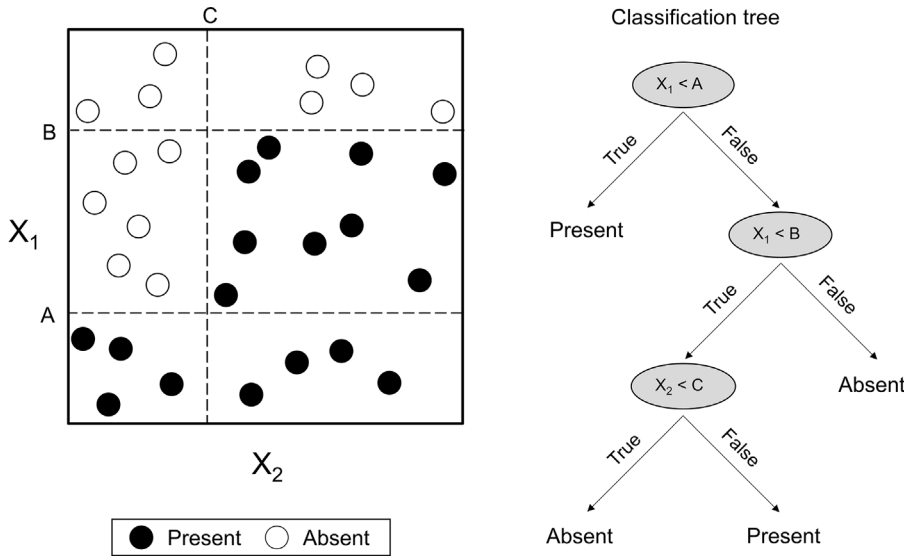


Fig. 4 Illustration of a classification tree for predicting the presence and absence of a species based on two explanatory variables (X_1 and X_2). The data are shown on the *left panel* and the tree on the *right panel*. Letters A, B, and C correspond to the probable locations of sequential splits for this dataset.

a series of artificial datasets or bootstrap samples. The variation among bootstrap samples approximates sample variation of the full dataset, so it is as though each bootstrap sample were an entirely new dataset generated by the same random process that created the full dataset. The first step of the random forest algorithm is to generate a large number (i.e., 100s or 1000s) of bootstrap samples. Unique observations that are included in each bootstrap sample are the "in-bag" data and the observations not sampled are the "out-of-bag" data. The second step is to generate a full classification tree for each of the in-bag datasets, using only a subset of the explanatory variables (e.g., the square root of the total number of variables). The individual classification trees are used to predict their corresponding out-of-bag samples. Predictions of out-of-bag observations across all trees are accumulated and error rates are computed for each observation. The random forest's predicted value for a given observation is the class that was predicted out-of-sample at the highest frequency for that observation. Given the number of trees involved in the prediction process, the results of a random forest are visualized and summarized in a manner that differs from classification trees. The overall influence of explanatory variables within the model is assessed using a variable importance statistic, which is calculated through permutation and is essentially the relative loss in classification accuracy that would occur if the explanatory variable were absent from the dataset. Another method for interpreting and visualizing the results of a random forest is through the use of partial dependence plots. Partial dependence plots illustrate the functional relationships between one or two explanatory variables and the response, averaging over all other explanatory variables. Notably, relationships depicted in these plots can sometimes appear jagged and discontinuous relative to those estimated by regression models (Fig. 5). The complex functions depicted in these plots are a result of the many latent interactions among variables in the random forest. The nature of these interactions can sometimes be difficult to disentangle without creating a large number of plots, and even when the relationships are apparent it can be difficult to ascribe biological meaning to the functional relationships that are identified. The ability to characterize many latent interactions is what gives the random forest algorithm the flexibility to make good predictions, but these latent interactions can also make it difficult to make inferences regarding effects of explanatory variables on species occurrence.

Boosted trees

Boosted regression/classification trees are another ensemble method for predicting the probability of occurrence for species across landscapes (De'ath, 2007; Elith et al., 2008). Unlike the random forest algorithm, the boosted classification tree approach does not draw bootstrap samples. Instead, this approach adds trees to the forest one by one. The initial tree attempts to classify the entire dataset, then the second tree is added in an attempt to correct misclassifications, then the third tree is added to correct remaining misclassifications, and so on. The three tuning parameters that are used to control the operation of the algorithm are the number of trees, the learning rate, and the tree complexity (a.k.a. depth). The number of trees dictates the stopping point of the algorithm. Learning rate dictates the contribution of each tree to the linear combination of trees and is typically set between 0.0001 and 0.1, where a lower learning rate requires a greater number of trees to achieve good predictive performance. Usually tree complexity is doubled whenever the learning rate is halved. The tree depth controls the number of nodes in each classification tree, which affects the order of interactions that can be modeled. For example, trees with a complexity of 1 allow only a single split, which means that the algorithm will only consider additive effects. The recommended method for determining tuning parameters is to run the algorithm on random subsets of data, while systematically applying different combinations of settings and evaluating how well the withheld data are predicted in each scenario. The optimal settings are those which jointly minimize the misclassification rate. Once the preferred settings are identified, the tailored algorithm can be fit to the full dataset and interpreted. Interpretation of boosted regression tree output also consists of estimating variable importance statistics and partial dependence plots.

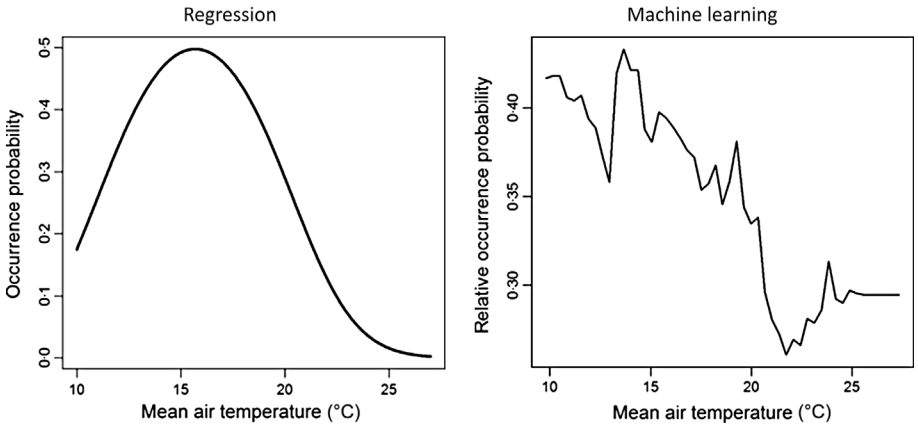


Fig. 5 Illustrating the different shapes of occurrence-habitat functions predicted from a regression model and a machine-learning algorithm. The *left panel* shows the predicted probability of occurrence from a generalized linear mixed model (GLMM) and the *right panel*, a partial dependence plot from a random forest. Figure from Wenger, S. J. and Olden, J. D. (2012). Assessing transferability of ecological models: An underappreciated aspect of statistical validation. *Methods in Ecology and Evolution* **3**, 260–267.

Models for Replicate Detection/Nondetection Data

Although the SDM approaches discussed so far are useful, they all have a critical weakness in that they do not account for imperfect detection. Detection is considered perfect if ecologists are guaranteed to detect a species in a sample unit, given that the sample unit is occupied. When monitoring a species or community a census is rarely, if ever, conducted, and encounter data both at the individual and species level are biased by imperfect detection. Therefore, the terms "presence-only" and "presence/absence" are misnomers. As the desire to acknowledge imperfect detection escalates, these terms have generally, and more accurately, been replaced by "detection-only" and "detection/nondetection." There is not a straightforward way to disentangle the occurrence state and measurement errors (i.e., false-positive and false-negative detections) using the approaches introduced above. Fortunately, imperfect detection can be statistically accounted for through the use of occupancy models. Occupancy models explicitly link the occupancy state and the observation process, allowing for the disentanglement of these complicated processes. Here, we offer a brief introduction to single-season, multiseason, and dynamic occupancy models. A thorough overview of these models and their extensions can be found in MacKenzie et al. (2017).

Occupancy models contain a few assumptions that should be listed. First, the occupancy state cannot change within a season, but it can change across seasons. Second, detection probabilities must be independent across sample units and survey occasions if occupancy is to be estimated. Third, this approach assumes that heterogeneity in occupancy and detection probability can be explained using covariates or is constant. Last, the occupancy models introduced herein assume false-positive detections do not occur, but see Chambert et al. (2015). A balanced sample design is not required to fit these models, meaning the sample units can be surveyed a variable number of times; however, a greater number of replicate surveys (both spatial and temporal) tend to lead to more accurate estimates.

Single-Season and Multiseason Occupancy Models

Single-season occupancy models were first described by MacKenzie et al. (2002) and Tyre et al. (2003). They realized that if detection/nondetection data across multiple sample units were collected in replicates within a timeframe that the occupancy state within sample units was static, information concerning both occupancy probability (ψ) and detection probability (p) is contained in the species encounter history. Occupancy probability is the probability that a randomly selected sample unit within the larger study area (i.e., the collection of sample units) is occupied by a species, and detection probability is the probability of obtaining a positive detection for that species, given the sample unit is occupied. For example, let us say a sample unit was surveyed three times and the recorded encounter history was "101". Given the species was encountered during at least one survey, we know the sample unit was occupied by the species during all surveys. Therefore, we can be certain that the encounter history reflects the species was detected during the first and third surveys, but not detected during the second survey despite the sample unit being occupied (i.e., a false-negative detection) or $\Pr(h = 101) = \psi p_1(1 - p_2)p_3$. However, what if the encounter history was "000," indicating the species was not detected in sample unit during all three surveys? In this case, it is possible the sample unit was occupied but we failed to detect the species or that the sample unit was unoccupied during the study period. This can be written as $\Pr(h = 000) = \psi(1 - p_1)(1 - p_2)(1 - p_3) + (1 - \psi)$. These processes can be linked hierarchically within occupancy models using coupled Bernoulli processes:

$$z_i \sim \text{Bernoulli}(\psi)$$

$$y_{i,j}|z_i \sim \text{Bernoulli}(z_i p)$$

where y represents the encounter history data for sample unit i and survey j, and z is the latent (partially observed) occupancy state for sample unit i. These equations tell us that whether we detect the species or not on each survey is conditional on the occupancy state of the sample unit and detection probability. Also, the occupancy state for each sample unit is linked to some probability of occupancy. The above example assumes both occupancy and detection probabilities are constant, which is hardly ever the question of interest. This can be modified, however, by relating these probabilities to explanatory variables using separate logit-linear models (see "Logistic regression model" section). Importantly, identical explanatory variables are not fit for each regression, but they can share a subset of explanatory variables. Also note that variation in occupancy is restricted to sample unit covariates since it is static within a season, whereas variation in detection can vary by sample unit and survey. If sample units are surveyed over multiple seasons, one approach to fitting a multiseason occupancy model is to treat each season's data as independent and simply add a covariate for occupancy probability and, perhaps, detection probability that indicates the season the surveys took place. This approach is useful in that it allows ecologists to borrow strength across all of the data, rather than fitting separate single-season occupancy models to each season's data. Importantly, the replicate detection/nondetection data must be collected using a robust design to apply this approach, where the occupancy state does not change within a season (Fig. 6).

Dynamic Occupancy Models

Sometimes our interests are focused on separating patterns from process, where we are less concerned with what is correlated with occurrence and are more interested in what factors govern distributional changes across time. A dynamic occupancy model is one approach to achieve this objective (MacKenzie et al., 2003; Royle and Kéry, 2007). Dynamic occupancy models are similar to the

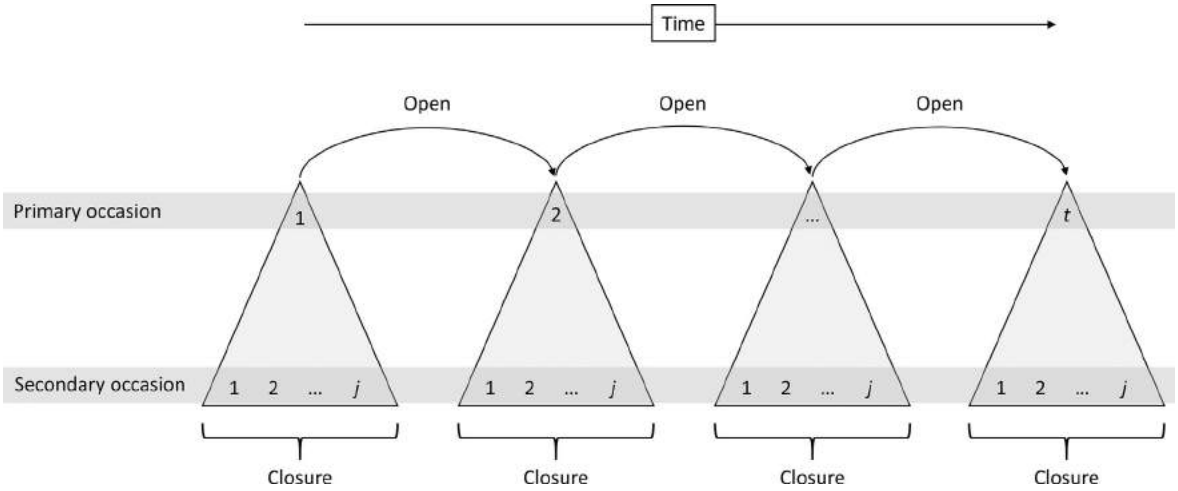


Fig. 6 Illustration of the robust design sample design. Note that the occupancy state of a sample unit does not change across secondary occasions within a primary occasion, but the sample unit can undergo local extinction or colonization between primary occasions.

occupancy models described in the "Single-Season and MultiSeason Occupancy Models" section (in fact they are an extension), except occupancy following the first season uses a first-order Markov model to accommodate temporal autocorrelation in the data. That is, it is likely that at a given sample unit the occupancy state in season $t + 1$ is related to the occupancy state in season t. Like simpler multiseason occupancy models, data must be collected using the robust design to fit this model (see Fig. 6). Dynamic occupancy models contain a couple of extra parameters when compared to simpler occupancy models. In particular, the model estimates the initial occupancy probability (ψ) in the first season, detection probability (p), and the occupancy state in subsequent time periods as a result of colonization (γ) and local extinction (ε) probabilities. For consistency and clarity, here is a dynamic occupancy model written in algebra:

$$z_{i,1} \sim \text{Bernoulli}(\psi)$$

$$z_{i,t+1} \sim \text{Bernoulli}(z_{i,t}(1-\varepsilon) + (1-z_{i,t})\gamma)$$

$$y_{i,j,t}|z_{i,t} \sim \text{Bernoulli}(z_{i,t}p)$$

where the terms are defined above. We see the difference here is the inclusion of a season (t) index and the explicit link between occupancy states across seasons. Similar to the simpler occupancy models described above, all four of these parameters (i.e., ψ, p, ε, and γ) can be related to explanatory variables using separate logit-linear models. Also, occupancy probability in season $t + 1$ can be estimated as $\psi_{t+1} = \psi_t(1-\varepsilon) + (1-\psi_t)\gamma$. It should be noted that local extinction probability is sometimes exchanged for local persistence probability (ϕ). In such cases, "$(1-\varepsilon)$" is replaced by "ϕ" in the equations above. The decision to model local extinction or local persistence is largely dependent on the story you want to tell. That is, are you interested in what causes populations to wink out (a.k.a. local extinction) or what leads local populations to remain intact (a.k.a. local persistence) across seasons? Both approaches are useful ways to describe occupancy dynamics.

Going Forward

Each of the approaches introduced in this article has contributed significantly to our ability to model the distribution of species, and extensions to these methods are being developed rapidly. Notably, there are an increasing number of studies that model spatially explicit abundances (reviewed in Kéry and Royle, 2016) or integrate multiple data sources (e.g., Coates et al., 2016; Fletcher et al., 2016) to develop SDMs, which were not introduced herein. Given the growing number of tools within an ecologist's toolbox, choosing among the different approaches can be daunting, but there are a few things to consider when choosing among the available methods. Axiomatically, fitting a more complicated model requires more robust data (i.e., datasets with larger sample sizes) to support the estimation of a greater number of parameters. Therefore, we stress that the question of interest is the most important thing to consider when choosing among the methods, and the simplest approach that can achieve the desired objective is often the best approach. Moreover, it is beneficial to have objectives in mind before data collection takes place. Otherwise data in hand will dictate what analyses are possible, which may or may not align with the objectives. Similarly, establishing objectives before data collection takes place allows ecologists to simulate monitoring data under various sample designs and establish more efficient and effective monitoring programs. After all, the reliability of an SDM is directly tied to the quality of the data that was used to develop it, no matter what analytical method is used. In our view, analytical methods that account for imperfect detection are always preferred over the alternative methods, if possible, to explicitly and reliably model the state of interest (i.e., species

occurrence). Complicating the estimated relationships with measurement errors, such as false-negative detections, can lead to uninterpretable results that provide little usefulness within ecological research. Finally, each of these methods contains a set of assumptions that should be carefully considered prior to data analysis. The approaches we present herein do not form an exhaustive list and we are unable to cover each approach in great detail. Nevertheless, we hope this article proves to be a useful primer of widely used methods to model the distribution of species and encourages readers to explore these topics in greater detail.

Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government. The Oregon Cooperative Fish and Wildlife Research Unit is jointly sponsored by the U.S. Geological Survey, the U.S. Fish and Wildlife Service, the Oregon Department of Fish and Wildlife, Oregon State University, and the Wildlife Management Institute.

References

Burnham KP and Anderson DR (1998) *Model selection and multimodel inference: A practical information-theoretic approach*, 2nd edn New York: Springer.

Chambert T, Miller DAW, and Nichols JD (2015) Modeling false positive detections in species occurrence data under different study designs. *Ecology* 96: 332–339.

Coates PS, Casazza ML, Ricca MA, Brussee BE, Blomberg EJ, Gustafson KB, Overton CT, Davis DM, Niell LE, Espinosa SP, Gardner SC, and Delehanty DJ (2016) Integrating spatially explicit indices of abundance and habitat quality: An applied example for greater sage-grouse management. *Journal of Applied Ecology* 53: 83–95.

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, and Lawler JJ (2007) Random forests for classification in ecology. *Ecology* 88: 2783–2792.

De'ath G (2007) Boosted trees for ecological modeling and prediction. *Ecology* 88: 243–251.

De'ath G and Fabricius KE (2000) Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology* 81: 3178–3192.

Diggle PJ, Laing K-Y, and Zeger SL (1994) *Analysis of longitudinal data*. Oxford: Oxford University Press.

Elith J and Leathwick JR (2009) Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics* 40: 677–697.

Elith J, Leathwick JR, and Hastie T (2008) A working guide to boosted regression trees. *Journal of Animal Ecology* 77: 802–813.

Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, and Yates CJ (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions* 17: 43–57.

Fitzpatrick MC, Gotelli NJ, and Ellison AM (2013) MaxEnt versus MaxLike: Empirical comparisons with ant species distributions. *Ecosphere* 4(5): 55.

Fletcher RJ Jr., McCleery RA, Greene DU, and Tyre CA (2016) Integrated models that unite local and regional data reveal larger-scale environmental relationships and improve predictions of species distributions. *Landscape Ecology* 31: 1369–1382.

Goldstein H (1995) *Multilevel statistical models*, 2nd edn. New York: Halstead Press.

Hastie T and Tibshirani T (1990) *Generalized additive models*. London: Chapman and Hall.

Hastie T, Tibshirani R, and Friedman J (2009) *The elements of statistical learning: Data mining, inference, and prediction*, 2nd edn. New York: Springer.

Hosmer D and Lemeshow S (1989) *Applied logistic regression*. New York: Wiley & Sons.

Kéry M and Royle JA (2016) *Applied hierarchical modeling in ecology: Analysis of distribution, abundance and species richness in R and BUGS, volume 1: Prelude and static models*. San Diego: Academic Press.

Larsen K, Petersen JH, Budtz-Jorgensen E, and Endahl L (2000) Interpreting parameters in the logistic regression model with random effects. *Biometrics* 56: 909–914.

MacKenzie DI, Nichols JD, Lachman GB, Droege S, Royle JA, and Langtimm CA (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology* 83: 2248–2255.

MacKenzie DI, Nichols JD, Hines JE, Knutson MG, and Franklin AD (2003) Estimating site occupancy, colonization and local extinction when a species is detected imperfectly. *Ecology* 84: 2200–2207.

MacKenzie DI, Nichols JD, Royle JA, Pollock KH, Bailey LL, and Hines JE (2017) *Occupancy estimation and modeling: Inferring patterns and dynamics of species occurrence*, 2nd edn. San Diego: Academic Press.

Merow C and Silander JA Jr. (2014) A comparison of Maxlike and Maxent for modeling species distributions. *Methods in Ecology and Evolution* 5: 215–225.

Merow C, Smith MJ, and Silander JA Jr. (2013) A practical guide to MaxEnt for modeling species' distributions: What it does, and why inputs and settings matter. *Ecography* 36: 1058–1069.

Merow C, Smith MJ, Edwards TC, Guisan A, McMahon SM, Normand S, Thuiller W, Wüest RO, Zimmermann NE, and Elith J (2014) What do we gain from simplicity versus complexity in species distribution models? *Ecography* 37: 1267–1281.

Phillips SJ and Elith J (2013) On estimating probability of presence from use-availability or presence-background data. *Ecology* 94: 1409–1419.

Phillips SJ, Anderson RP, and Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modeling* 190: 231–259.

Royle JA and Kéry M (2007) A Bayesian state-space formulation of dynamic occupancy models. *Ecology* 88: 1813–1823.

Royle JA, Chandler RB, Yackulic C, and Nichols JD (2012) Likelihood analysis of species occurrence probability from presence-only data for modeling species distributions. *Methods in Ecology and Evolution* 3: 545–554.

Stroup WW (2012) *Generalized linear mixed models: Modern concepts, methods and applications*. Boca Raton: CRC Press.

Tyre AJ, Tenhumberg B, Field SA, Niejalke D, Parris K, and Possingham HP (2003) Improving precision and reducing bias in biological surveys: Estimating false-negative error rates. *Ecological Applications* 13: 1790–1801.

Wenger SJ and Olden JD (2012) Assessing transferability of ecological models: An underappreciated aspect of statistical validation. *Methods in Ecology and Evolution* 3: 260–267.

Statistical Inference

Daniel M Wolcott, University of Central Missouri, Warrensburg, MO, United States
Adam Duarte, Oregon State University, Corvallis, OR, United States
Floyd W Weckerly, Texas State University, San Marcos, TX, United States

Glossary

Bias Processes or procedures that consistently result in measurements that differ from the true or correct value.
Inductive reasoning Making a scientifically valid conclusion about a large unknown entity from only a part of the unknown entity.
Maximum likelihood estimation The omnibus estimation approach that assumes large sample size.
Precision The repeatability of a measurement taken multiple times.
Probability The science of uncertainty with values that range from 0 to 1, inclusively.
Repeated measures Observations that are repeatedly measured through time and when subjected to different treatments.
Restricted maximum likelihood estimation Using maximum likelihood estimation on small data sets where estimation relies on using some of the observed data.
Statistics The science of analyzing data.

Introduction

Simply put, statistical methods are used by ecologists to detect patterns in data. If patterns are detected, then the ecologist might also want to predict. Prediction is particularly useful for visualizing the estimated relationships.

Fundamental to the correct application of statistical methods, is an understanding of words and terms such as observation, variable, sample, random sample, population, statistics, parameters, and sample error. An observation is the basic unit. By definition an observation is independent, meaning that each observation has its own information unconnected to the information in any other observation. Replicates or experimental units are synonymous with observations. Characteristics measured from observations are called variables. It is the numerical values of variables that are actually analyzed, and it is convenient to classify variables to assist in choosing the appropriate statistical method, which is largely governed by whether a variable is categorical or numeric. Numeric variables are further classified into continuous or discrete because of different mathematical characteristics (Fig. 1). Continuous variables have infinite possible values between end points and discrete variables are restricted to integer values (e.g., 0, 1, 2, etc.). The population (usually denoted as N) is all the observations of interest and can be quite large or infinite in size. Because it is often not practical to measure variables from all observations of a population, a random subset of observations (usually denoted as n) is selected. The random subset of observations that are drawn from the population is the sample. To, hopefully, obtain a sample that is a microcosm of the population, observations should be selected without any known bias (i.e., a random sample). For example, if we wanted to know if male and female deer were different in body mass at the Kerr Wildlife Management Area, the observation would be the individual deer, the variables measured from each deer would be sex and body mass, the sample would be the collection of deer that were randomly selected and measured, and the population would be all the deer on the Kerr Wildlife Management Area. Clear definitions of these basic terms are needed as the foundation to understand statistical inference.

The essence of statistical inference is making a conclusion about the large unknown and unknowable (population) from the known (random sample). As such, inductive reasoning is required so that a valid conclusion is obtained. Integral to making statistical inferences is accommodating sample error—which is from the sampling process—and using distributions (Fig. 1). The practical consequence is that sample statistics (e.g., mean and variance) almost always will not be identical, but are hopefully close to, their corresponding parameters (population mean and variance). In other words, the data analyst strives for statistical estimates that are precise and unbiased. Using a simple example, a valid inference about a population mean can be made from a 95% confidence interval of a sample mean. The confidence interval is patently estimated to accommodate sample error, and it is a representation of the uncertainty associated with the point estimate. Notably, the interval is more likely to correctly inform about the population mean when the sample data is unbiased and sufficiently large.

Distributions

There are a number of distributions that are used in statistical inference to model uncertainty. We describe six distributions that are often used (Fig. 1). The binomial and Poisson are discrete distributions. Values of these distributions have a lower limit of zero and upper limits that can be quite large. The values are integers. The binomial distribution is one of the most widely

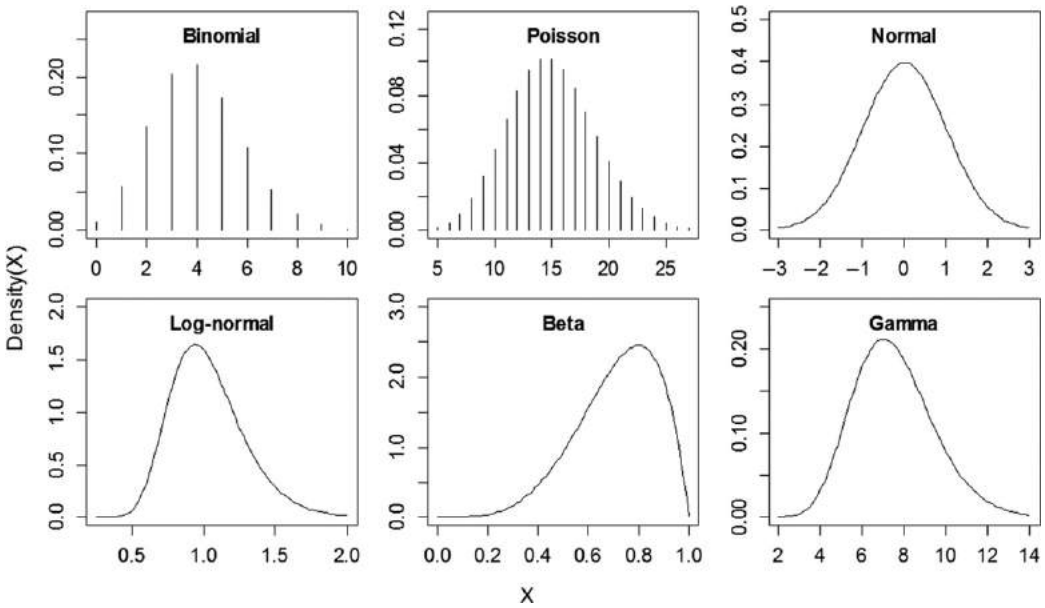


Fig. 1 Examples of six distributions used in statistical inference. Each distribution can have a variety of shapes. Two of the distributions are discrete and the remaining four are continuous. Because the discrete distributions can only have integers for *x* values, vertical lines were used instead of a smooth curve.

implemented distributions for ecological studies focused on the distribution and abundance of animal populations. The key property of the binomial are trials, which by definition are independent. Each trial has two mutually exclusive outcomes. Often, one of the outcomes is of interest and labeled a success. For example, a binomial distribution can be used to describe whether an individual female deer (the trial) gave birth to one or more young in a year (the success) or not. With knowledge about the probability of success it is possible to use the binomial distribution to predict the probability of success for a given set of scenarios. The Poisson distribution is sometimes labeled the distribution of rare events. Imagine you are counting fish at a weir and it is likely that there will be thousands of fish that pass through the weir in a day. One of your tasks is to note if the observed fish are tagged, but there might be three tagged fish counted in a day. For the binomial distribution, the probability of success would be very small and not a convenient application because of the extra variation in the data from the few successes relative to the number of trials. A convenient alternative is the Poisson distribution. The defining feature of the Poisson distribution is the equality between the mean and variance. This distribution is frequently used for modeling population abundances because there usually is small variance among counts of a small population and large variance among counts of an abundant population.

The remaining four distributions are continuous distributions. The normal or Gaussian distribution is a widely used distribution for at least three main reasons. One reason has to do with symmetry. If you superimpose a vertical line through the mean or the peak of the distribution, the right half is a mirror image of the left half. Second, values of the distribution can be negative or positive. Third, the bell-shape of the distribution is useful for describing the sampling process and it matches the distribution of numerous ecological patterns. The normal distribution can be described from knowing the mean and standard deviation. The remaining three continuous distributions are useful in particular settings. The lognormal distribution is a right skewed distribution that becomes more symmetrical with larger means. The lognormal is useful when populations grow exponentially or when the mean is close to zero but the values are constrained to be positive. The beta distribution ranges from 0 to 1 and has two shape parameters that are used to calculate mean and variance. Because the beta distribution ranges from 0 to 1, it is suited for modeling probabilities or proportions. Last but not least is the gamma distribution. The gamma distribution ranges from 0 to values that are quite large. It is a distribution of waiting times until a certain number of events occur. For example, the length of time it takes for three deer to die. The mean and variance are calculated from a shape parameter and a length per event or scale parameter.

Regression Models

General linear models are popular statistical methods (Bolker, 2008). Three conditions (assumptions) define a general linear model, which should be met in the response variable: independence, normality and homoscedasticity (similar variances). It is important to note that the assumption of normality is often misunderstood as a requirement for data to be distributed normally. Instead, normality refers to an error distribution (assessed from residuals, value of variable minus mean of variable) (see Kéry and Hatfield, 2003).

Table 1 Findings of the same general linear model expressed in two ways, as a two-sample *t*-test and as a simple linear regression

Two-sample t-*test*							*Simple linear regression*				
Sex	*n*	*Mean*	*Variance*	*t*	*df*	*P*	*Coefficient*	*Estimate*	*SE*	*t*	*P*
Female	51	2.64	0.39	**2.51**	122	**0.007**[a]	Intercept	2.64	0.09	28.78	
Male	73	2.94	0.46				Sex	0.30	0.12	**2.51**	**0.007**

[a]The *P* can be interpreted as, if you decide to conclude that males are indeed larger than females, there is a 0.007 probability that you are incorrect. Because of the small probability, the conclusion is that males are heavier than females.

The data set is birth mass (kg) of female and male white-tailed deer born to 2 year old mothers fed a pelleted-diet rich in nutrients. For the simple linear regression sex was coded 0 for females and 1 for males. Note the *t* and *P* values (in bold font) of the two-sample *t*-test and sex coefficient of the regression are identical. Because I expected males to be heavier than female a one-tailed alternative hypothesis was used instead of the default two-tailed alternative.

Examples of general linear models include *t*-tests, analysis of variance (ANOVA), simple and multiple regression. Simple regression has one predictor variable (sometimes referred to as an explanatory variable or independent variable) and multiple regression has two or more predictor variables. The choice of a general linear model depends on the number of groups or samples and the number of variables measured from observations. It might also be supposed that method of choice is dictated by whether you need to test for differences among means or estimate relationships. Yet, these boundaries that seemingly determine the kind of method to use are actually blurred (Draper and Smith, 1998). Realizing how the same data set can be analyzed by more than one general linear model is useful to grasping that categorical as well as numeric variables can be analyzed in a regression analysis. To illustrate, we compare findings from a two-sample *t*-test and a simple linear regression. The data was birth mass of female and male white-tailed deer (*Odocoileus virginianus*) born to 2-year-old mothers on a high nutritional plane. Mothers only gave birth to a single young (Wolcott *et al.*, 2015). Our goal was to test whether male birth mass was greater than female birth mass. In size-dimorphic ungulates, adult males are larger than adult females and the difference can manifest at birth (Wolcott *et al.*, 2015). Findings from the two-sample *t*-test indicated that males, on average, were heavier than females (**Table 1**). The *t* value was 2.51 and with 122 degrees of freedom the *P* was 0.007. A *P* is the probability of a type I error, which occurs when you conclude that there are differences when there actually are no differences. The critical or α value for a *P* (the cutoff in which you decide there is a difference or not) that is most often used in ecology is 0.05; however, 0.1 and 0.01 are also used. The other kind of error is a type II error, which is failing to detect differences that truly exist, a possibility that seems remote in this example due to the relatively large sample size, substantial size dimorphism and low variances between the sexes. To conduct a simple linear regression on this data, the categorical variable sex was coded as a dummy variable (Draper and Smith, 1998). Here, we arbitrarily assigned 0 to females and 1 to males. In the parlance of regression, the predictor was sex and the response variable was birth mass. Because females were coded 0, the intercept of the regression is the estimate of mean birth mass of females (**Table 1**). The estimated slope measures the difference in means between male and female birth masses because males were coded 1. On average, male birth mass was 0.30 kg heavier than females. If 0.30 is added to 2.64, the intercept, you arrive at the mean birth mass of males (2.94 kg). Thus, sometimes the distinction between two-sample *t*-test and simple linear regression is in name only.

A large part of the popularity with general linear models is the versatility to summarize complex relationships. When the response variable is continuous and there are multiple categorical, numeric, or both kinds of predictors; multiple regression is appropriate. To illustrate, we estimated relationships between body mass, age and liver mass of male white-tailed deer during the mating season (Parra *et al.*, 2014). Deer age was either young (1.5–2.5 years-old, $n = 26$) or prime-aged (4.5–6.5, 13). Body mass was continuous and age was categorical. During the mating season, prime-aged, but not young, males engage in physically demanding activities to attain copulations. The consequence is that energetic demands usually exceed energy intake in prime-aged males. To meet the energy demands, prime-aged males mobilize adipose tissue, which is a function of the liver that leads to a heavier liver mass because the workload of the liver increases (Parra *et al.*, 2014). There is a complication, however, in that prime-aged male deer are larger than young deer and also require larger livers simply due to their larger body size. Thus, the need to include body mass as a predictor to control for this effect. The estimated multiple regression revealed that body mass was positively related to liver mass and that prime-aged males had heavier livers (**Fig. 2**). Accounting for body mass, prime-aged males had livers that were, on average, 0.21 kg heavier than young males. The R^2, adjusted for number of predictors, (0.75) indicated that about 75% of the variation in liver mass was accounted for by the regression. In addition to versatility, general linear models are appealing because they can be used to predict and scatterplots help to visually display complex relationships in data.

Hierarchical Models

In many ecological studies, efficient sampling design often entails repeated or clustered measurements of subjects or observations. These types of sampling designs inherently create hierarchical structure within the data, which are often correlated spatially and temporally. Correlations between measurements violate one of the foundational assumptions that each observation is

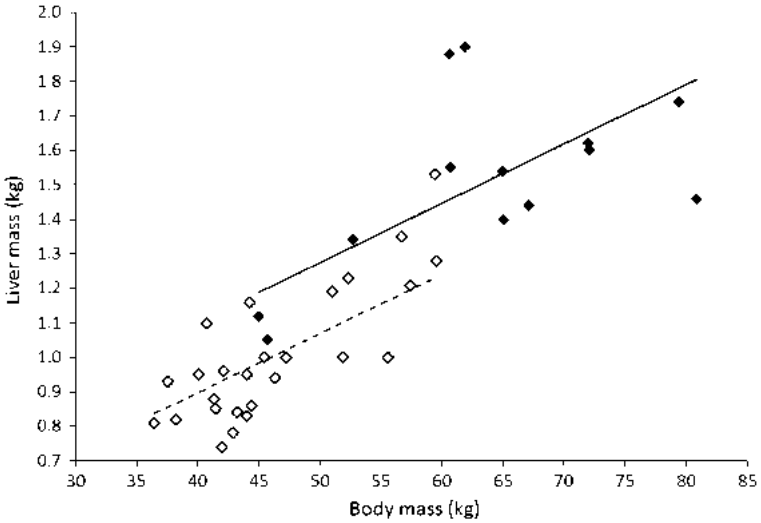


Fig. 2 Scatterplot of data and relationships between body (minus rumen-reticulum and liver organ masses) and liver masses of young (1.5–2.5 years of age, open diamonds, dashed line) and prime-aged (4.5–6.5, filled diamonds, solid line) male, white-tailed deer. The general linear model or multiple regression was: Liver mass $= 0.21 + 0.02$*body mass $+ 0.21$*age. Body mass and age were influential ($t \geq 2.7$, df $= 36$, $P \leq 0.011$). Young males were coded 0 and prime-aged males were coded 1. The adjusted R^2 was 0.75.

independent and identically distributed. Repeated measures analyses have been devised to account for the issue of dependence among observations, however, many of these analyses are not flexible about missing data (i.e., unequal sample size among levels) or uneven temporal spacing of repeated measurements. Fortunately, mixed-effect models account for these issues and are also more easily interpretable than most classical repeated measures analyses.

Mixed-effect modeling is a powerful tool that uses both fixed and random variables within the same model. Whether a variable should be considered as fixed or random is largely dependent on the question of interest and the number of levels within the variable. If the number of levels for the variable is representative of the population and the question of interest involves finding differences between levels, utilizing the variable as a fixed component is the recommended approach. If the levels of the variable are a randomly selected subset of all available levels and the question of interest involves understanding the variation within the population, treating the variable as a random variable is the recommended approach. Understanding the desired objective of the research question is of great importance when considering whether a variable should be treated as fixed or random. Further, while variables treated as fixed can be of several types (e.g., continuous, categorical, etc.) random variables must be categorical with a large enough number of levels to achieve an accurate representation of the variance in the population (usually >4 but preferably >10 levels, Zuur *et al.*, 2007).

Mixed-effect models demonstrate tremendous flexibility in how the random components can be modeled. For example, consider a study in which 10 individuals were repeatedly sampled across time to measure the influence of a continuous predictor (X) on the response variable (Y). We could simply regress X against Y at a cost of three parameters (one intercept, one slope of X, and residual error); however, we know that repeatedly sampling 10 individuals is most likely creating correlations within the residual error because individuals are more likely to be similar to themselves than the other nine individuals. We could account for this, in a fixed effect only framework, by incorporating the individual identification as a categorical covariate. Inclusion of this covariate would create a regression line with a different intercept but same slope for each one of the sampled individuals, which increases the number of parameters drastically to 12 (10 unique intercepts, 1 slope for all sampled individuals, and residual error). The inclusion of an individual-specific variable quickly inflates the number of parameters, which penalizes the ability to detect differences through the loss of degrees of freedom. Instead, we could include the individuals as a random effect when regressing X against Y. This essentially produces the same outcome as the above fixed effect only model except it only costs us four parameters (one intercept, one random effect to estimate variation of individuals around the intercept, one slope, and residual error). With this mixed-effect model we are able to reduce the residual error by accounting for the variation created by sampling the 10 individuals. Mixed-effect models are even more powerful than this simple problem. Random effects can be modeled to assess variation in intercepts only (like the above example), slopes only, intercepts and slopes, and even residual error itself. We strongly encourage the reader to continue to explore the many possible ways that mixed-effect modeling can be used to partition variance structures (see Pinheiro and Bates, 2000, and Zuur *et al.*, 2007 for more in-depth examples).

Invariably, mixed-effect model development will require some form of selection process to assess which fixed variables and random components are necessary to accurately describe reality, within a modeling framework, without overfitting (i.e., include too many predictors and random components) the data. The generally accepted approach consists of a top-down strategy beginning with a saturated model that includes all fixed factors of interest as well as any interactions between fixed factors that are thought to be biologically relevant (Zuur *et al.*, 2007). The next step is to assess which random error structure is most optimal (e.g.,

random intercept, slopes, or intercept and slopes, etc.) using restricted maximum likelihood estimation. Once the optimal random error structure is defined for the saturated model, assess which fixed variables should be kept to form an optimal model using maximum likelihood estimation. Finally, conclusions from the optimal model, using the estimated parameters and standard errors, should be presented and discussed using restricted maximum likelihood estimation to avoid biased variance estimates.

When comparing models with the same fixed effects and different random effects through hypothesis testing, it is important to note that significance values are only approximations, which should be assessed with caution if the value is on the cusp of the selected α level. This approximation is a consequence of two-tailed hypothesis testing during model comparison. Since variances, which are centered on zero, can only be positive values, a two-tailed hypothesis test inaccurately assesses the range of variance values yielding a significance value that is generally inflated higher than expected. There is limited concern for significance values that are very large or very small as compared to the α level; however, the concept of "testing on the boundary" should be considered when the significance value nears the value for α.

Bayesian Inference

Up until now we have described the process of statistical inference using what is known as a classical or frequentist paradigm. However, the use of Bayesian inference within ecological studies has increased in the past decade, so a chapter focused on statistical inference would be inadequate without some mention of it. For an in-depth description of the background information and application of Bayesian inference, we recommend readers consult Link and Barker (2010) and Hobbs and Hooten (2015). It should be stated that there is no such thing as a "Bayesian model" despite the term used within the peer-reviewed literature. As Kéry and Schaub (2012) discuss, ecologists develop models to represent hypotheses they want to explore in data, and they can choose to fit the model(s) to the data in either a Bayesian or frequentist framework. Indeed, the regression models described above as well as many complex hierarchical models can be fit under both paradigms (Royle and Dorazio, 2008; Kéry and Royle, 2016). Furthermore, when prior distributions are specified as diffuse or noninformative in Bayesian statistics there is a great deal of agreement in the output estimates from the two paradigms (Kéry, 2010).

So what is the difference in the two paradigms and why would you choose to fit models using one paradigm versus the other? In a frequentist paradigm, parameters are treated as fixed and unknown constants, and inferences are based on the notion of repeated experiments or the frequency of hypothetical replicates. Here, ecologists make probability statements about the data and the probability is based on the hypothetical replicates. Thus, a confidence interval is interpreted as "If I performed this experiment an infinite number of times, X% of the parameters would fall within this range." In contrast, under the Bayesian paradigm all parameters are treated as random variables and the unknown quantities are modeled using the statistical distributions introduced above, among others. Here, our knowledge of parameters is updated using data, rules of conditional probability, and Bayes' theorem. In other words, prior (knowledge) distributions and data are used to obtain or yield posterior distributions to make inferences. This is usually accomplished with the aid of simulation methods, such as Markov chain Monte Carlo (MCMC) algorithms, to characterize the posterior distribution of parameters. Because parameters are treated as random variables, under the Bayesian paradigm credible intervals, which are the Bayesian analog of confidence intervals in frequentist statistics, are interpretable as "I am X% confident the parameter is within this range." Although appealing from an interpretability standpoint, in practice this difference is of little consequence as confidence intervals and credible intervals are typically applied in the same manner.

As we mentioned earlier, Bayesian inference can accommodate previous knowledge from other studies via informative priors so that inferences are based on both the current and previous studies. Also, in Bayesian inference all results are exact even for small datasets because Bayesian analyses do not rely on asymptotic approximations that are used in frequentist analyses (Conroy and Peterson, 2013). This is particularly appealing because data are often limited in ecological studies. The BUGS language that is often used to fit models using a Bayesian analysis, offers a framework for better understanding the structure of statistical models and by extension a more transparent structure for accounting for sources of variation in the data. Once ecologists have gained the skills to develop simpler models in BUGS, more complicated models are a relatively simple extension that use the same rulesets. Thus, more complicated models can be fit using a Bayesian framework relatively easy, including models in which there is no frequentist method available. This has led to the recent increase in the use of integrated analysis of multiple but related datasets within the ecological literature. These integrated analyses represent an important advancement in the ecological sciences because they allow for the development of tailored analyses for specific datasets, which, in turn, allow for the incorporation and estimation of complex ecological relationships. Similarly, derived parameters (i.e., estimates that are calculated from other estimated parameters) that propagate the full uncertainty of estimates can be calculated in a straightforward manner using BUGS. In the frequentist paradigm, deriving fitted models requires using the delta method or the omission of parameter uncertainty.

It should be noted that there are also drawbacks to fitting models in a Bayesian framework. First, the BUGS programming language can admittedly be intimidating when getting started, particularly if you are not well acquainted with statistical distributions. Nonetheless, we stress that if fitting more complex models is a necessary objective, the time investment to overcome the initial learning curve is well worth it. Also, it should be noted that the number of iterations required during the MCMC simulations to achieve convergence when fitting more complex models to large data sets can take hours to weeks, whereas the same model using a frequentist analysis may be much quicker. We view the frequentist and Bayesian paradigms as useful tools for statistical inference and stress that ecologists do not have to choose one or the other. Instead, we recommend ecologist consider the pros and cons of each approach as they relate to the objectives of a particular study and choose the best tool for the task.

Model Selection

To aid in statistical inference, models are developed to mimic the underlying distribution of a population using empirical data. In an ecological context, most studies are considered to be observational with limited ability to set up control–treatment experiments that include proper randomization and replication. Since it is rarely possible to create a model that describes a population with complete accuracy, a suite of variables that are hypothesized to be responsible for some portion of the observed phenomenon are measured and multiple competing models are developed. Because of this, it is necessary to choose a model that best approximates the underlying distribution of the sampled population, without overfitting the available data, through a model selection process.

The coefficient of determination (R^2), as an example and demonstrated above, is a useful tool to assess how well a model replicates observed data. Although a useful tool to assess how well a single model explains the observed data, it is not useful for comparison between competing models. The R^2 value will almost always increase with each additional inclusion of a predictor variable, which leads to the selection of a complicated model that probably over fits the observed data. To account for this issue, various stepwise procedures have been developed to assess how strongly each variable influences the predictive power of a model using hypothesis tests.

Hypothesis testing procedures are popular techniques for model selection and include forward selection, backward elimination, or stepwise processes. The forward selection process begins with the most simplified (null) model and compares the null model to another model that includes a single added variable. These two models are then compared using an F-test. If the F statistic is larger than the set threshold, the model with the added variable has a stronger predictive power than the null model and is considered the null model thereafter. The procedure repeats itself by including additional variables and comparing the newest model to the previously selected model until a candidate variable no longer increases the F statistic above the accepted threshold. The backward elimination process is similar except that it begins with a global model that contains all predictor variables. This model is tested against another model that is reduced by a single variable with a subsequent F-test assessing if the variable should be included or not. Stepwise procedures were developed as an extension of forward and backward selection processes to remedy flexibility issues after a variable is considered in the analysis. In the stepwise process, all variables are considered, at each step, for inclusion or omission from the model even if they were previously included or excluded in a previous model. While these procedures are widely used it is important to note that they are not without flaws. Specifically, there is no clear criterion for model choice and the final model choice is usually a good model but not always the optimal model. An optimal model does not under- or over fit the data. Indeed, using forward and backward processes on the same dataset can often lead to the selection of very different models.

An information-theoretic approach to model selection seeks to find the optimal model that bests describes the relationship between the response variable and a set of predictors. Unlike hypothesis testing, it uses a criterion, involving the log likelihood (a maximum likelihood measure of model fit), to assess the distance (i.e., information lost) between a model and the observed data. Instead of focusing on the significance of individual predictor variables within a model, information-theoretic model selection assesses the likelihood of a model as a whole. The intent of multimodel inference in an information-theoretic framework is to be a confirmatory analysis rather than an exploratory analysis so the process of data dredging (assessing all possible models) is strongly discouraged (Burnham and Anderson, 2003). Instead, the goal is to develop models with sound biological reasoning for comparison through multimodel inference to determine the best approximating model given the data through strength of evidence. Yet, information-theoretic approaches have been applied to traditional hypothesis testing procedures such as the forward selection, backward elimination and stepwise processes because of simplicity and flexibility.

One of the most common forms of information-theoretic model selection in ecology, from the frequentist perspective, is the Akaike's Information Criterion (AIC). The criterion is simply defined and calculated as "deviance plus $2K$," where deviance is a measure of the information loss between the model and the observed data (-2 times the log likelihood) and $2K$ is a penalty for increasing model complexity (K is the number of parameters estimated in the model). A second-order bias correction (AICc) has also been developed for use when sample size is low (generally $n/K < 40$). Since AIC and AICc values converge as sample size increases, the use of AICc in all cases is common. Multimodel inference from AIC is very simple because the model with the lowest AIC value is the best approximating model. For competing models (models with <2 AIC value difference from the best model), it is possible to model average parameter estimates, however, parsimony should also be considered before model averaging competing models. Burnham and Anderson (2003) demonstrate the process, flexibility, and advantages to AIC model selection, which includes valuable resources necessary to calculate evidence ratios between models and the assessment of relative variable importance within the selected model. We advise reading this resource for further information.

An information-theoretic approach to model selection is also possible from the Bayesian perspective. The most commonly used criterion for Bayesian inference is the Deviance Information Criterion (DIC) because it is effective for models with informative prior information and is readily accessible as a function in the BUGS software. Although DIC has been widely used over the past decade, it has come under increasing criticism for several reasons (e.g., model selection is not always consistent and tends to produce over-fitted models). A relatively new addition to the expanding field of information criterion is the Watanabe–Akaike information criterion or widely applicable information criterion (WAIC). It has been shown to be asymptotically equal to Bayesian cross-validation procedures and is the most fully Bayesian process because it uses the full posterior distribution (rather than using a point estimate like AIC and DIC). WAIC is still not without limitations and has some difficulties with spatially structured data. For Bayesian inference, cross-validation is still the recommended method for model validation with WAIC used as a more computationally expedient option (Gelman *et al.*, 2014).

See also: Ecological Data Analysis and Modelling: Ecological Models: Model Development and Analysis; Model Types: Overview; Sensitivity, Calibration, Validation, Verification; Structural Dynamic Models. General Ecology: Principal Components Analysis

References

Bolker, B.M., 2008. Ecological models and data in R. Princeton, USA: Princeton University Press.
Burnham, K.P., Anderson, D., 2003. Model selection and multi-model inference: A practical information-theoretic approach, 3rd ed. New York, USA: Springer Science & Business Media.
Conroy, M.J., Peterson, J.T., 2013. Decision making in natural resource management: A structured, adaptive approach. Hoboken, USA: Wiley-Blackwell.
Draper, N.R., Smith, H., 1998. Applied regression models. New York, USA: John Wiley & Sons.
Gelman, A., Hwang, J., Vehtari, A., 2014. Understanding predictive information criteria for Bayesian models. Statistics and Computing 24, 997–1016.
Hobbs, N.T., Hooten, M.B., 2015. Bayesian models: A statistical primer for ecologists. Princeton, USA: Princeton University Press.
Kéry, M., 2010. Introduction to WinBUGS for ecologists: A Bayesian approach to regression, ANOVA, mixed models and related analyses. New York, USA: Academic Press.
Kéry, M., Hatfield, J.S., 2003. Normality of raw data in general linear models: The most widespread myth in statistics. The Bulletin of the Ecological Society of America 84, 92–94.
Kéry, M., Royle, J.A., 2016. Applied hierarchical modeling in ecology: Analysis of distribution, abundance and species richness in R and BUGS, volume 1: Prelude and static models. San Diego, USA: Academic Press.
Kéry, M., Schaub, M., 2012. Bayesian population analysis using WinBUGS: A hierarchical perspective. New York, USA: Academic Press.
Link, W.A., Barker, R.J., 2010. Bayesian inference with ecological applications. New York, USA: Academic Press.
Parra, C.A., Duarte, A., Luna, R.S., Wolcott, D.M., Weckerly, F.W., 2014. Body mass, age, and reproductive influences on liver mass of white-tailed deer (*Odocoileus virginianus*). Canadian Journal of Zoology 92, 273–278.
Pinheiro, J.C., Bates, D.M., 2000. Mixed-effects models in S and S-Plus. New York, USA: Springer Verlag.
Royle, J.A., Dorazio, R.M., 2008. Hierarchical modeling and inference in ecology: The analysis of data from populations, metapopulations and communities. New York, USA: Academic Press.
Wolcott, D.M., Reitz, R.L., Weckerly, F.W., 2015. Biological and environmental influences on parturition date and birth mass of a seasonal breeder. PLoS One 10.e0124431
Zuur, A.F., Ieno, E.N., Smith, G.M., 2007. Analyzing ecological data. New York, USA: Springer Science & Business Media.

Structural Dynamic Models☆

Arnab Banerjee, Nabyendu Rakshit, and Santanu Ray, Visva-Bharati University, Santiniketan, India

Introduction

Ecosystems are well-established network organizations consisting of various kinds of biotic and abiotic components. The main basis of network establishment and its subsequent maintenance is solar energy that is trapped and stored into potential energy by plants and is subsequently utilized by heterotrophs via grazing and detrital pathways in ecological networks.

Owing to the dynamic nature of storage and utilization of solar energy, an ecosystem continuously evolves in order to modify, regulate and change to accommodate for maximum efficiency. It is thus very important to apprehend this aspect in order to understand ecosystem functioning and hence arise the need for mathematical models describing the structural dynamics of ecosystems.

Ecosystems are constantly affected by various environmental factors (forcing functions) such as nutrient loading, natural calamities, anthropogenic activities, etc. that affect an ecosystem continuously and force it move toward thermodynamic equilibrium. In order to cope with this ever changing scenario, an ecosystem may undergo minor to extensive structural changes so as to maintain its organization and efficiency. Structural dynamic modeling (SDM) is a tool for predicting or assessing such dynamic changes in an ecological organization that is able to continuously optimize system properties (parameters) in response to such dynamic fluctuations (Jørgensen and Bendoricchio, 2001).

Structural dynamic models are radically different and more advanced than previously used dynamic models (which were much more rigid) and have been referred to as fifth generation modeling tool (Jørgensen and Bendoricchio, 2001). In the past few decades, a good number of studies have been conducted on the structural dynamic models with specific underlying goal functions. Recently, there has been increasing concern for environmental remediation technologies on the basis of ecological methodology, which strongly depends on the ecological structure and also promotes the study on structural dynamic models.

Deterministic mass balanced models have been utilized in order to achieve a better understanding and interpreting the underlying cause–effect (feedback) mechanisms in ecological observations. These models range from simpler N-P-Z models having nutrient, phytoplankton and zooplankton as state variables to complex physico-biological coupled models incorporating time-series data with the intention of describing the ecological organization and structure, optimization of exergy as goal function with implications of phytoplankton and zooplankton body size, etc. (Banerjee et al., 2017a; Bierman Jr et al., 1994; Hofmann and Ambler, 1988; Kremer and Nixon, 2012; Mandal et al., 2009; O'Connor, 1981; Ray et al., 2001a).

In this article, methodology of modeling approaches for SDM have been introduced, especially focusing on the holistic approaches as well as classical bottom-up-style modeling approach. Structural dynamic models have also been described in the backdrop of ecological exergy and ascendancy as goal functions followed by several case studies.

Feedbacks in Ecosystem

A large number and wide variety of feedback mechanisms are in constant action in any ecosystem at any time that fluctuates in strength and timescale. Keeping in mind the fact that a model is just a replica of the concerned system it is impossible to include all

☆*Change History*: March 2018. Arnab Banerjee, Nabyendu Rakshit and Santanu Ray modified sections Abstract, Keywords, Introduction, Feedbacks in Ecosystem, Experiments and Observations, Holistic Approach, Eco-exergy or Exergy, References and Further Readings. The section that have been introduced are Structural Dynamic Modelling Approach (majorly rewritten as a new section), Bottom-up Approach, Goal Functions and Ascendancy.

feedbacks into a singular model. Therefore, in order to formulate the flows in an ecosystem, it becomes necessary to understand what kinds of feedbacks are dominant in the targeted ecosystem and discuss how to incorporate these feedbacks into the model. Modeling of an ecosystem structure is the finding out of the dominant feedback mechanisms in the ecosystem and mathematical description of these dominant feedbacks in the model. These feedbacks with widely different action mechanisms; especially, the response timescale, is one of the important points to describe the structural dynamics, varies greatly from case to case (Jørgensen and Bendoricchio, 2001).

For example, in an aquatic ecosystem, over a short timescale, growth of phytoplankton is controlled by available nutrient concentrations dissolved in water and/or the intracellular nutrient concentration. When the concentration of a nutrient is far less relative to requirement (e.g., half-saturation constant of the phytoplankton for the concerned nutrient), the growth rate of the phytoplankton is far below the maximum growth rate. On the other hand, if the concentration is enough (and there are no other limitation, such as toxic substances), phytoplankton grows in accordance to their maximum growth rate. This is an inherent property of phytoplankton growth and most lake ecological models include these feedbacks. However, more long-term feedbacks—such as biochemical adaptation, biological adaptation, species selection, selection of predator–prey relationship, etc. are sometimes required to be introduced through structural dynamic models. Furthermore, the emergence of new species may be required in a structural dynamic model which is used to assess the very long-term ecological succession.

Hence modeling strategies should be changed depending on the feedback timescales which are introduced into a model. Particularly, the long-term feedbacks tend to be important in the structural dynamic models.

Structural Dynamic Modeling Approach

The ability of ecosystems to replace currently constituent species (components) with others (better fitted) that can be utilized for successful optimization, can be reflected by constructing complex models that encompass all of the actual constituent components, on the basis of experimental observations or literature information, for the entire period that the model embodies. This is a major aspect in modeling studies and is the most advanced tool for development of more realistic and scientific ecological models refining the structure of a model by accumulating information on all components and their interactions. This classic approach of SDM methodology includes the "bottom-up approach" and here the model structure is determined based on experimental and observational knowledge.

However, incorporation of a larger set of model parameters that require calibration and validation introduce high uncertainties in the models. In that context, another approach that estimates the dynamics of the system from a comprehensive viewpoint—ecological goal functions—have been introduced to overcome the cumbersome difficulties of tedious parameter calibration. This relatively new approach is called as the "holistic approach" and not many structurally dynamic models are present which account for goal functions of system performance. Of the several goal functions proposed for such holistic modeling, exergy, emergy and ascendancy principles have been widely used by several authors.

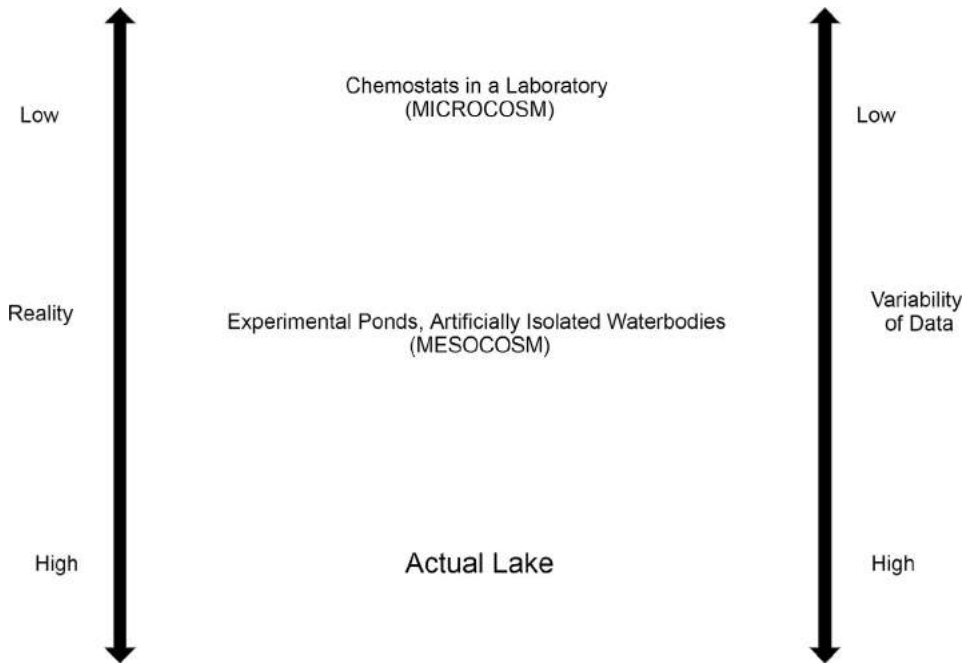


Fig. 1 An example of mass balance in a lake.

Experiments and Observations

In many cases, most of data used to develop of a model are obtained from the experiments and past observations or from published literature. In some cases, however, newly designed experiments and observations are implemented for the particular requirements of the modelers. For adequate designing of experiments and observations, it is desirable for modelers to understand the feature of the experiments and observations of the target ecosystems.

In the model of lake ecosystems, for example, various scales of experiments and observations can be designed and implemented as described by Nielsen (1995). These size variations and properties are shown in Fig. 1. The experiments and observation size can be roughly divided into three classes: chemostats (flasks, small-scale incubators, etc.) in a laboratory (the so-called microcosm); experimental pond, artificially isolated water bodies, etc. (the so-called mesocosm) and the observations of actual lakes.

The chemostat experiments are useful to obtain the fundamental property of growth kinetics such as maximum growth rate, half-saturation constant of nutrient uptake, respiration rate, etc., under well-controlled environmental factors such as irradiation and temperature. In these experiments, single species or a few species of organisms are incubated under various conditions (e.g., incubation of *Microcystis* sp. under various phosphorus concentrations) and the time variations of the biomasses of organisms are measured.

Organisms can be incubated in the mesocosm experiments at the same time. The time variations of relatively small organisms (e.g., phytoplankton and zooplankton) under artificially controlled environments (e.g., fish biomass, irradiance) are observed. The information about the structural dynamics in a water ecosystem during the relatively short periods can be obtained. The fluctuations of environmental factors are relatively easy in the mesocosm experiments. Whole mass balance of an ecosystem can be approximately obtained under nearly natural conditions. Mesocosm experiments have been attracting increasing attention in practice of structural dynamic modeling. Of course, it is desirable to install the mesocosm into or nearby the actual lake in question.

Larger-scale and long-duration information, of course, should be obtained from the actual lake and consist of the most important information. However, in many cases, data may not be newly collected especially for the purpose of the structural dynamic modeling (secondary data from literature). Therefore, the sampling points and sampling intervals are generally limited. As a result, it essentially has high variability.

Therefore, these data may be combined to model the structural dynamics of the lake (see the next section).

Bottom-Up Approach

The bottom-up approach is one of the most fundamental approaches to modeling the structural dynamics of an ecosystem where the structure of an ecosystem network is determined on the basis of experimental and observational knowledge. In practice, some of the important and dominant components from ecological viewpoint are selected and introduced in the model. The kinetics of material flows is also formulated in the same manner. In this approach, the selection of adequate experiments and observations and a reasonable processing method of the obtained data are important.

An Example of Bottom-Up Approach

An example of bottom-up approach of structural dynamic modeling has been presented here in pertaining to a lake ecosystem using experimental and observational data.

At the first stage, the flow diagram of the model (Fig. 2) (inclusive of multispecies details of phytoplankton and zooplankton, multigrowth stages of zooplankton, etc.) is determined. Of course, most aspects of concern must be included. In this stage, the structure of the food-web network can be determined on the basis of the correlations between organisms obtained from the above-mentioned mesocosm experiments, or actual lake observations.

In the next stage, the kinetic equations and their parameters for each organism are determined, based on the knowledge from literature and/or chemostat experiments. After complete formulation of the model for the concerned system, parametric sensitivity

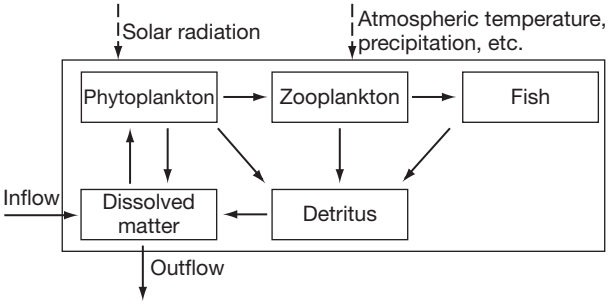


Fig. 2 Size variations of some experiments and observations and their properties for the modeling of a lake.

analysis have to be performed and the sensitive (critical) parameter set is to be elucidated. This is followed by sensitive parameter set calibration within certain ranges on the basis of the data from mesocosm observations. Calibration ranges of the parameter set are also determined based on the knowledge from the literature and/or the chemostat observations. Note that the kinetic parameters of natural organisms are essentially distributed. From this circumstance, the probability distributions of kinetic parameters are assumed in some cases. Of course, in these cases, the calculated values also have probability distributions.

Then, in the final stage, the model is extrapolated to the subject lake by considering its physiographic characteristics and environmental factors. The model validation is performed using the actual observation data. Of course, in most cases, relative minor calibration is required in this stage.

However, as mentioned above, several organisms should be included in a model to describe the structural dynamics in an ecosystem. Thus, as many organisms (state variables) as possible (if possible all) should be included in the model. It is also clear that this leads to an increase not only in the number of parameters that should be calibrated, but also in the whole uncertainty of the model. Furthermore, there is also a certain limitation of kinetic data. This is a fundamental drawback of the bottom-up and structure refinement modeling approach. To reduce this drawback, some rationalization methods such as likelihood test can be applied to modify the models. Furthermore, descriptions of the long-term feedbacks such as various adaptations, selections and emergence of new species, as described before, are essentially difficult or impossible in this modeling procedure. This is also why a holistic approach, described in the next section, is now desired, especially for the description of long-term structural dynamics.

Holistic Approach

It is obvious that a food-web network in an ecosystem is very complicated in most cases. Furthermore, the food selectivity of a predator changes due to a change in the availability of prey (if a species is in a starvation state, it may eat a wide range of preys), or change in environmental condition. Properties of the biotic components (organisms) are fundamentally varied, for example, it is often observed that same species have different properties depending on the location of ecosystem; and many organism groups have age distribution and as a result, the body size and their growth kinetics vary. Rational modeling of these properties is the key to describe the structural dynamics in an ecosystem.

From this point of view, a holistic approach to construct structural dynamic models is receiving more attention today. The concept of holistic approach for ecological modeling is discussed in depth mainly by Jørgensen and colleagues in a series of studies that appeared in the journal *Ecological Modelling* and elsewhere. The most important point of holistic approach is that the kinetic parameters of the model are assumed to change depending on the time elapsed and the parameter values are calibrated to be consistent with the ecological principles. To estimate the degree of consistency with the ecological principles, an index called "goal function" is introduced in the approach.

Goal Functions

If any ecosystem face threating conditions such as environmental disaster or anthropogenic disturbance it cannot escape from that condition, but instead it has to adapt itself to those conditions in an efficient manner. The maximum efficiency regards to moving system far from thermodynamic equilibrium. This property indicates the goal functions of any ecosystem which they try to achieve in order to maximize their performance. For example exergy (means the amount of stored workable energy), ascendency (means the degree of system organization), entropy (its minimum value indicated highest efficiency of system performance).

Only a few structurally dynamic models have been introduced till date which accounts for goal functions of system performance.

For example, optimization of exergy (maximum stored workable energy) as a goal function for system performance indicator has been studied in detail by Santanu Ray and colleagues (Banerjee et al., 2017a; Ray et al., 2001a). Here they have studied the implications of body sizes of phytoplankton and zooplankton as an important parameter for total system dynamics by optimizing exergy and also the effect of phosphorus input corresponding different type of system such as oligotrophic, mesotrophic and eutrophic on its dynamics; also the variations in exergy content of a reservoir system following hypothetical perturbations scenarios (representing realistic situations) have been studied. In addition to exergy, "ascendency" as a goal function or indicator of system performance under the above prevailing conditions have also been studied (Ray et al., 2001b). Ascendency quantifies the growth and development of an ecosystem as a product of total system throughflow and the mutual information inherent in the pattern of internal system flows. For a detailed description of these goal functions, the reader is referred to literature.

In the following section, exergy and ascendancy are briefly reviewed as a goal function for ecological structural dynamic modeling.

Eco-Exergy or Exergy

Exergy describes the maximum work which can be produced from a system under a given environment. This concept is commonly used in process engineering to estimate (or design) various energy systems such as co-generation systems. Exergy is one of the most widely used goal functions in the structural dynamic modeling. Jørgensen suggested a typical procedure of the holistic approach of structural dynamic modeling (Fig. 3). In this approach, exergy is used as a goal function.

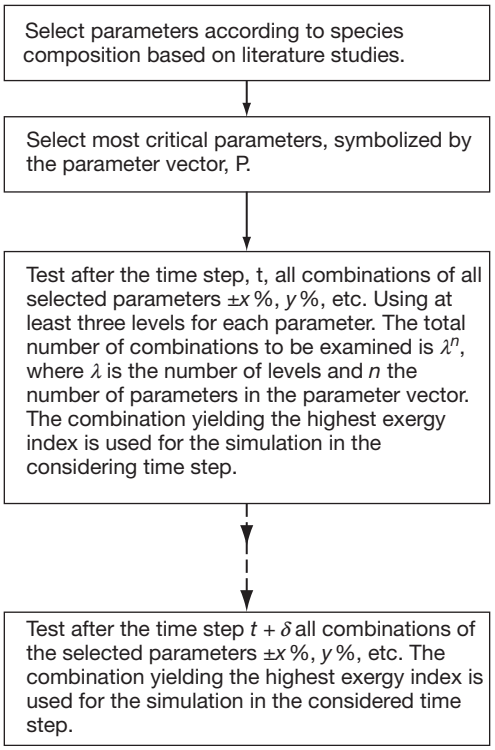


Fig. 3 A typical holistic approach of structural dynamic modeling using exergy as goal function. Adapted from Jørgensen, S.E. and Fath, B.D. (2004) Modelling the selective adaptation of Dawin's Finches. *Ecological Modelling* **176**, 409–418, with permission from Elsevier.

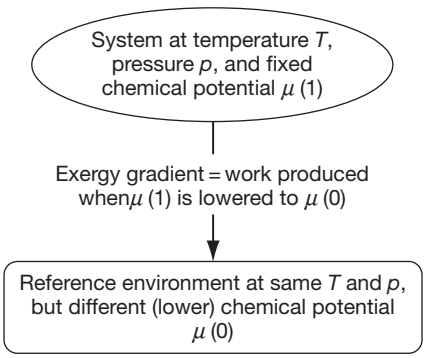


Fig. 4 Idea of the ecological exergy. Adapted from Jørgensen, S.E., Patten, B.C., and Straškraba, M. (2000). Ecosystem emerging: 4. growth. *Ecological Modeling* **126**, 249–284, with permission from Elsevier.

In relation to different goal functions, exergy has been used most widely in ecological models. Exergy has some advantage compared to others: it is related to state variables, which are easily determined or measured. It can be considered as fuel for any system to convert energy and matter in metabolic processes. The concept of exergy on thermodynamical aspect has been described in detail in previous studies, where detailed exergy calculation and future scope of study on various nonlinear effects like structural instability, dynamic chaos etc. are thoroughly discussed (Mandal et al., 2009).

The exergy concept is useful to assess the thermodynamical state of a system, and is extended to the ecosystem analysis, where it is called ecological exergy. The idea of the ecological exergy is depicted in Fig. 4. An ecosystem that has temperature T, pressure p, and chemical potential $\mu(1)$ is assumed to exist under a reference environment that has temperature T, pressure p, and chemical potential $\mu(0)$ (only chemical potential is different). The exergy gradient can be identified as the work produced, and the gradient is calculated by the difference in chemical potential. This is the principle of the ecological exergy.

Ecological exergy and its application for structural dynamic modeling have been studied in detail by Jørgensen and colleagues. The principle of the application of exergy is that the ecosystem tends to develop with maximizing the exergy to keep the organization far from the thermodynamic equilibrium. According to Jørgensen, the conceptual diagram of the change in the ecological exergy can

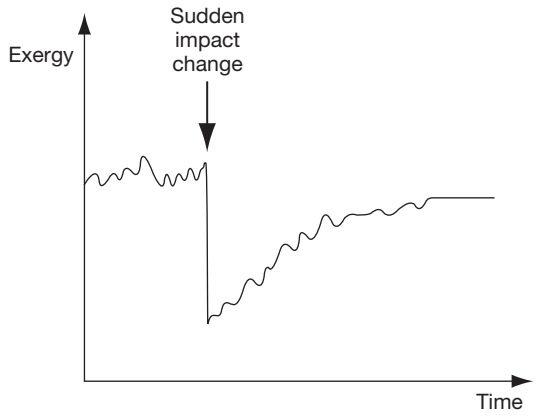


Fig. 5 Conceptual diagram of the trends in the ecological exergy changes. Adapted from Jørgensen, S.E. (1986). Structural dynamic model. *Ecological Modeling* **31**, 1–9, with permission from Elsevier.

be depicted as in Fig. 5. One of the main purposes of the structural dynamic modeling is to predict these organizational trends as a result of some impact, as in Fig. 5. Thus, a modeling approach based on exergy (or some other goal functions) is thought to be required.

One of the advantages of using exergy as a goal function is that it is easily calculated from the state variables in an ecological model. The ecological exergy is defined, with some approximations as

$$Ex = RT \sum_{i=0}^{N} c_i \left(\mu_i - \mu_{i,0}\right)$$

where Ex is the ecological exergy, R is the gas constant, T is the absolute temperature, c_i is the concentration of ith component in the ecosystem, μ is the chemical potential, and $(\mu_i - \mu_{i,0})$ is the difference in chemical potential of ith component between the ecosystem and its corresponding thermodynamic equilibrium under given environment. For a detailed description of the derivation of this equation, the reader is referred to works by Mejer, Jørgensen, and others.

Some applications of exergy for actual structural dynamic modeling, especially for the water ecosystems, can be found in the literature. These outcomes show the reasonability of the utilization of ecological exergy compared with the phenomena observed in the actual ecosystems.

Ascendency

Ascendency quantifies the growth and development of an ecosystem as a product of total system throughflow and the mutual information inherent in the pattern of internal system flows. Total system throughflow indicates the amount of energy or materials flowing into the system and ascendency quantify both the amount of energy flowing through the system and how much it retained into system at a given time. Maximum ascendency value indicates the highest system performance and maturity. Studies related to ascendancy fluctuations in relation to autotroph and fish biomass perturbation scenarios and also in relation to keystone species variations have also been reported (using mass balanced models) previously (Banerjee et al., 2017a,b).

Applications of Structural Dynamic Models

One of the essential objectives of the structural dynamic models is to assess the dynamics of the ecosystems. Currently, one of the most important issues in global environment is the extinction of rare and endangered species. Accordingly, the concern for the importance of biodiversity has been growing today. These topics can be dealt with in the structural dynamic models. In this circumstance, the expectation from the structural dynamic models will increase more and more in the future.

On the other hand, some environmental purification and remediation methods using ecological techniques have been brought to public attention. These are especially, applied for water treatment, or remediation of water environments. Some of them require the ecological structural dynamic models as a tool for the assessment of the methods. Biomanipulation, which has been brought to public attention, is one of the better examples of the remediation of water environment using the ecological technique. Biomanipulation aims to prevent the unusual growth of phytoplankton as a result of eutrophication in a lake. The basic concept of biomanipulation is that if the effective grazing of phytoplankton by zooplankton is achieved in a lake, the unusual phytoplankton growth is suppressed at certain levels of nutrient loadings. To construct this adequate ecological structure, the biomass of planktivorous fish (which is a predator of zooplankton) should be suppressed and the biomass of piscivorous fish (which is a

predator of the planktivorous fish) should be maintained. To do so, planktivorous fish is removed from a lake and/or piscivorous fish is introduced into a lake in the implementation of biomanipulation. For more detailed information about the biomanipulation, the reader is referred to the literature. The success/failure of biomanipulation strongly depends on the structural change in lake ecosystem after the manipulations. Especially, the succession of zooplankton is thought to be the most important. In general, in cases where biomanipulation is a success, the large-size zooplankton such as *Daphnia* sp. dominates. It is clear that the large-size zooplankton can essentially graze phytoplankton of various sizes at a high rate and is effective in suppressing the unusual growth of phytoplankton. For successful biomanipulation, therefore, some kinds of zooplankton, which graze harmful phytoplankton such as *Microcystis* sp., should dominate. To do so, sufficient prediction of zooplankton succession should be required prior to the implementation of biomanipulation. That is why the analysis of structural dynamics in the lake is strongly required for the implementation of biomanipulation, and the structural dynamic models become a strong tool to deal with this subject.

Summary

The complex but precise ecosystem networks have been attracting the attention of many researchers for a long time and, as a result, many useful ecological models have been proposed. In this circumstance, the concern with structural dynamic models of ecosystems has been growing from the viewpoints of not only mathematical ecology but also environmental technology.

The structural change in ecosystems is fairly complicated, because there exists a great number of components and their feedbacks, as described in the text. The holistic approach for the structural dynamic models, which is also described in the text, is one of the keys to solve the problem. A continuous research on the feedback mechanisms in ecosystems and the holistic approach would strengthen the methodology of the structural dynamic modeling for various ecosystems.

In the future, more knowledge about the structural changes of various ecosystems should be clarified. Structural dynamic models become a good tool to summarize the knowledge. To elucidate the fundamentals underlying the ecosystem functions, analysis through the structural dynamic models will also be required.

References

Banerjee A, Chakrabarty M, Rakshit N, Mukherjee J, and Ray S (2017a) Indicators and assessment of ecosystem health of Bakreswar reservoir, India: An approach through network analysis. *Ecological Indicators* 80: 163–173.

Banerjee A, Scharler UM, Fath BD, and Ray S (2017b) Temporal variation of keystone species and their impact on system performance in a South African estuarine ecosystem. *Ecological Modelling* 363: 207–220.

Bierman VJ Jr., Hinz SC, Zhu D-W, Wiseman WJ Jr., Rabalais NN, and Turner RE (1994) A preliminary mass balance model of primary productivity and dissolved oxygen in the Mississippi River Plume/Inner Gulf Shelf Region. *Estuaries* 17: 886–899.

Hofmann EE and Ambler JW (1988) Plankton dynamics on the outer southeastern U.S. continental shelf. Part II: A time-dependent biological model. *Journal of Marine Research* 46: 883–917.

Jørgensen SE and Bendoricchio G (2001) *Fundamentals of ecological modelling*, 3rd ed. Amsterdam: Elsevier.

Kremer JN and Nixon SW (2012) *A coastal marine ecosystem: Simulation and analysis*. Springer Science & Business Media.

Mandal S, Ray S, Roy S, and Mandal S (2009) The concept of exergy and its extension to ecological system. In: Pélissier G and Calvet A (eds.) *Handbook of exergy, hydrogen energy and hydropower research*, pp. 1–14. New York, NY, USA: Nova Science.

Nielsen SN (1995) Optimization of exergy in a structural dynamic model. *Ecological Modelling* 77: 111–122. https://doi.org/10.1016/0304-3800(93)E0088-K.

O'Connor DJ (1981) Modeling of eutrophication in estuaries. In: *Estuaries and nutrients*, pp. 183–223. Springer.

Ray S, Berec L, Straškraba M, and Jørgensen SE (2001a) Optimization of exergy and implications of body sizes of phytoplankton and zooplankton in an aquatic ecosystem model. *Ecological Modelling* 140: 219–234.

Ray S, Berec L, Straskraba M, and Ulanowicz RE (2001b) Evaluation of system performance through optimizing ascendency in an aquatic ecosystem model. *Journal of Biological Systems* 9: 269–290.

Further Reading

Bendoricchio G and Jørgensen SE (1997) Exergy as goal function of ecosystems dynamic. *Ecological Modelling* 102: 5–15.

Jørgensen SE (1999) State-of-the-art of ecological modelling with emphasis on development of structural dynamic models. *Ecological Modelling* 120: 75–96.

Jørgensen SE and Fath BD (2004) Modelling the selective adaptation of Darwin's Finches. *Ecological Modelling* 176: 409–418. https://doi.org/10.1016/j.ecolmodel.2003.12.048.

Jørgensen SE and Mejer H (1977) Ecological buffer capacity. *Ecological Modelling* 3. https://doi.org/10.1016/0304-3800(77)90023-0.

Jørgensen SE and Mejer H (1979) A holistic approach to ecological modelling. *Ecological Modelling* 7: 169–189.

Jørgensen SE and Nors Nielsen S (1994) Models of the structural dynamics in lakes and reservoirs. *Ecological Modelling* 74: 39–46.

Jørgensen SE, Nielsen SN, and Mejer H (1995) Emergy, environ, exergy and ecological modelling. *Ecological Modelling* 77: 99–109.

Jørgensen SE, Patten BC, and Straškraba M (2000) Ecosystems emerging: 4. Growth. *Ecological Modelling* 126: 249–284.

Marques JC and Jørgensen SE (2002) Three selected ecological observations interpreted in terms of a thermodynamic hypothesis. Contribution to a general theoretical framework. *Ecological Modelling* 158: 213–221.

Marques JC, Pardal MÅ, Nielsen SN, and Jørgensen SE (1997) Analysis of the properties of exergy and biodiversity along an estuarine gradient of eutrophication. *Ecological Modelling* 102: 155–167.

Nielsen SN (1992) Strategies for structural-dynamic modelling. *Ecological Modelling* 63: 91–101.

Shapiro J (1990) Biomanipulation: The next phase—Making it stable. *Hydrobiologia* 200–201: 13–27.

Visualization as a Tool for Ecological Analysis

S Kyle McKay, U.S. Army Engineer Research and Development Center, New York, NY, United States

Glossary

Information visualization "The processes of producing visual representations of data and the outputs of that work. Information visualisation aims to enhance one's ability to carry out a task by encoding often highly abstract information into a visual form. Visualisations can be static, or interactive and dynamic, and hosted in a variety of media (e.g., journal poster, website, or software)" (McInerny *et al.*, 2014).

Visual analytics "The science of analytical reasoning facilitated by interactive visual interfaces" (Thomas and Cook 2005 in Keim *et al.*, 2008) or "combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision-making on the basis of very large and complex data sets" (Keim *et al.*, 2008).

Visualization "A method of computing, [which] transforms the symbolic into the geometric, enabling researchers to observe their simulations and computations. Visualization offers a method for seeing the unseen. It enriches the process of scientific discovery and fosters profound and unexpected insights" (McCormick *et al.*, 1987).

Introduction

Visual exploration of empirical, experimental, or model data is a powerful tool for increasing understanding of complex, long-term, and variable data sets common in ecology (Keim *et al.*, 2008; Fox and Hendler, 2011; McInerny *et al.*, 2014). Data visualization methods are well-studied in fields of visual analytics, information visualization, computer graphics, and scientific communication (e.g., McCormick *et al.*, 1987; Tufte, 2001; Keim *et al.*, 2008; Aigner *et al.*, 2011). Ecologists informally use visualization to parse data sets, guide analyses, and explore new ideas, but the field rarely acknowledges formally the role of visualization in ecological analysis and synthesis.

Large data sets are increasingly available in ecology (e.g., stream gage networks, high resolution sensor networks, large-scale remote sensing), and effective visualization techniques will be crucial to rapidly and efficiently understand and communicate these observations (Michener and Jones, 2012). Visualization cannot substitute for more rigorous quantitative and statistical methods (Garbrecht and Fernandez, 1994). However, visual exploration takes advantage of the capacity of the human eye to rapidly detect and discern visual patterns (e.g., color, shape, grouping), when presented effectively (McCormick *et al.*, 1987; Keim *et al.*, 2008; Fox and Hendler, 2011; Healey and Enns, 2012).

Given the breadth of ecological data types, formats, volumes, and analytical needs, innumerable data visualization approaches are potentially pertinent to the ecological community of practice. Rather than undertake a foolhardy review of these methods, the objective of this article is to highlight the value of visualization as a component of ecological analysis and synthesis and to present a variety of key issues that must be addressed in the selection and application of a visualization approach. The fields of visual analytics, information visualization, computer graphics, and scientific communication provide a rich body of literature on the subject, and this article serves only as an entry point for uncovering the seemingly endless body of data visualization approaches. To this end, data visualization examples are presented relative to four common ecological applications: data exploration, experimental analysis, numerical model output and evaluation, and ecological decision-making. The article concludes with a set of questions to guide ecologists in the selection and application of a visualization approach.

Reviewing Data Visualization Via Case Study

Ecological data visualization is inherently specific to a problem, purpose, or question. For instance, three questions about the management of an invasive riparian plant would drive an analyst to explore vastly different visual media: What is the plant's current extent (may lead to a map)? What environmental conditions influence the current distribution (may lead to a scatterplot between variable-x and plant density)? Does chemical-y effectively control the invasive plant (may lead to a barplot of mortality relative to treatment and control groups)? This pedestrian example is merely intended to suggest that visualizations are akin to other ecological analysis tools; the method must befit the need. Because of this intimate connection to applications, case studies are used herein to review common issues in visualization of complex ecological data sets. These examples often omit ecologically and analytically relevant details in the interest of focusing of key aspects of the visual approach. Case studies were selected to present a diversity of ecological applications and highlight crucial considerations for the visual presentation. Many potentially interesting visualization approaches were not considered (e.g., interactive graphics, animations) due to the constraints of the two-dimensional, print medium (See section Selecting a Visualization Method).

Data Exploration

Often ecological data are collected over long time scales to understand trends, patterns, and variability associated an ongoing process or system. Additionally, the resolution of these data streams is increasing as sensor networks improve and computational power increases. This often leads to the need to analyze patterns in large, complex data sets, and visualization provides an ideal tool for exploration of large data sets. For example, the Luquillo Long-Term Ecological Research Program in eastern Puerto Rico has trapped freshwater shrimp in tropical streams weekly for more than 20 years (Crowl *et al.*, 2017). Here, catch per unit effort from a single pool (Pool 0 in Quebrada Prieta) and species (*Atya lanipes*) from 1993 to 2014 are examined through the lens of multiple time series visualization methods (**Fig. 1**).

Many visualization analyses begin with traditional plotting methods such as line, scatter, or bar plots, and these simple figures can prove extremely valuable, particularly if a visual benchmark is used to call attention to a specific aspect of the data (e.g., the mean and standard deviation are shown in the upper left figure). As with quantitative methods, a variable may be transformed to highlight important nonlinearity in the distribution of a variable, but for visualization transformations should be conducted by way of altering the axis not the data to maintain natural units for the analyst. For instance, shrimp catch data are logarithmically

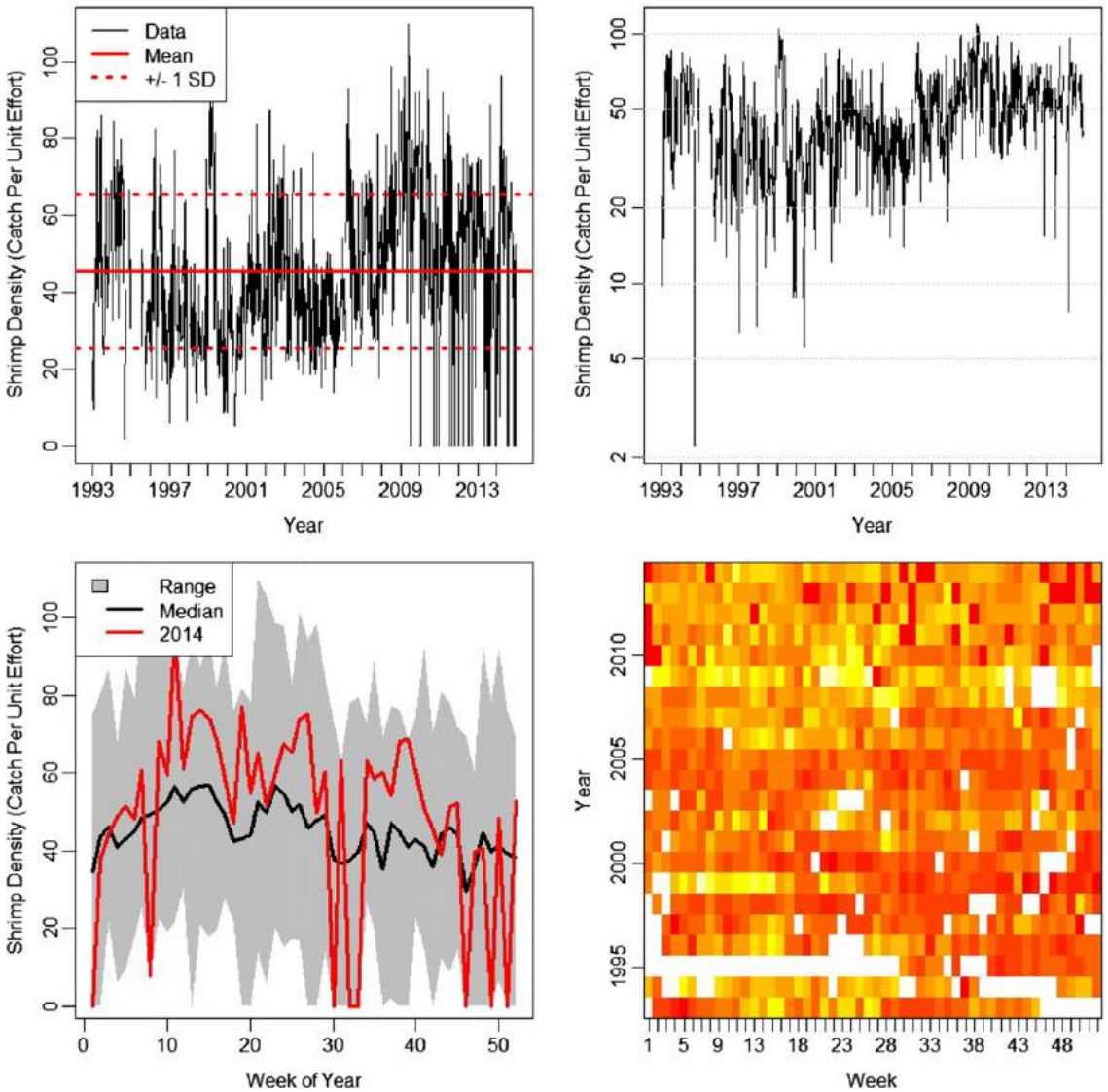


Fig. 1 Multiple visualization methods applied to a long-term, ecological time series. Data represent weekly freshwater shrimp (*Atya lanipes*) trapping densities collected at pool-0 of the Luquillo Long-Term Ecological Research (LTER-LUQ) site in Puerto Rico: (top-left) line plot with linear scale and reference points, (top-right) line plot with logarithmic scale, (bottom-left) envelope plot summarizing 2014 data relative to long-term observations, (bottom-right) a raster plot using color rather than location to represent density (red = high, yellow = low, white = missing).

transformed in the upper right figure, but the original units for catch per unit effort (CPUE) are preserved as a mechanism for maintaining understandability in the plot rather than *log* (*CPUE*). As a result of transformation, a more pronounced pattern of long-term catch declines and variability from 1995 to 2005 is shown, which is followed by a period of relative stability and high catch from 2005 to 2014. Ecological time series often exhibit periodic or cyclical patterns (e.g., diel movement, seasonal rainfall), and periodicity can provide a mechanism for compressing data for visualization. For instance, an envelope plot is used to provide historical context of the data (e.g., the range of observations and median for each sample week) for understanding a particular component of the data set (e.g., 2014 highlighted here; lower left figure). Although these line-based methods can lead to important observations, large data sets are challenging to communicate through this media because data are compressed into a finite plotting space. A variety of time series methods exist for overcoming these weaknesses, and the reader is encouraged to explore these approaches (Aigner *et al.*, 2011). As one example, a raster-based summary (lower right figure) presents the shrimp catch data set, and new observations may be made that were unapparent in other visuals such as data gaps (white cells), interannual variability (e.g., large-scale catch reduction across 2009), and massive intraannual variability (e.g., sequential high and low catch values in 2012).

Experimental Data

Experiments often help distinguish causal mechanisms in ecological systems, and visualization can help identify potential patterns, guide quantitative analyses, and facilitate the presentation of findings. More detailed guidance on the appropriate use of visualization in experimental analysis is addressed elsewhere (e.g., Weissgerber *et al.*, 2015), and only a few common issues are illustrated here. For instance, McKay *et al.* (2017a) used a tethered flight mill to study immune trade-offs with flight effort in North American monarch butterflies (*Danaus plexippus*). The experiments collected flight distance and duration data for individual monarchs in reproductively active summer conditions and reproductively inactive fall conditions for a maximum of four, 60-min flight trials conducted over sequential days. While additional analyses were originally pursued, the simplest form of the raw flight data are used here to address key aspects of experimental data visualization (**Fig. 2**).

The type of data generally serves as the first distinguishing feature driving a particular plotting approach with continuous and categorical data types providing the primary points of differentiation (e.g., 0.1, 0.8, and 1.2 and treatment-A and treatment-B, respectively; Weissgerber *et al.*, 2015). Typically, continuous variables lend themselves well to scatter and line plots, while categorical variables lend themselves to bar and box plots (top and bottom figures, respectively). Data grouping can also lead to alternative observations within a continuous variable. For instance, the top figures address the same data with all trials and individuals lumped into a single data set (left) and data parsed by individual monarch across four trials (right), and the two plots give the reader an alternative understanding of the amount of uncertainty and individual variation within the data. Generally speaking, a set of visualizations that effectively embed more data and information are preferred to those that present a single view of data. For instance, in the bottom figures, the bar plot conveys the general effect of the two seasonal treatments based on the mean observation, but the box plot provides additional information about the distribution of outcomes between the treatments (e.g., the lower tail of the duration data are affected more by season, nonnormality of the distribution is more apparent, the complete range of the data are shown).

Ecological Model Development

In addition to empirical approaches, visualization can inform many of the steps in an ecological model development process of conceptualization, quantification, evaluation, and application. **Fig. 3** presents a diversity of examples of visual approaches used to inform multiple types of ecological models. While not explicitly data visualization, conceptual representations of ecological models can guide the development of a model, structure data input–output, and serve as a mechanism for communication in an interdisciplinary team. For example, a conceptual model helped structure the development of an ecological model for quantifying the benefits of restoration actions in the Proctor Creek Watershed in Atlanta, Georgia (upper left, McKay *et al.*, 2017b), and the model was subsequently used by the restoration team to catalog potential restoration actions (orange boxes), input variables (white boxes), summary variables (gray boxes), and categorical outputs (yellow boxes). Visualization can also inform the evaluation of ecological models during calibration, verification, or testing phases. For instance, Shrestha (2016) developed a watershed-scale hydrologic model and applied separate portions of the streamflow gaging record to calibrate and verify model outcomes (upper right). Joint visualization of these phases, along with reference points of perfect prediction and $\pm 20\%$ error, allowed simultaneous consideration of the relative value of predictions and avoided systematic bias or error (i.e., calibration and verification have similar distributions of observed and predicted values). As ecological models become more sophisticated, so too must the visualization approaches for verification. For instance, Swannack *et al.* (2009) developed a model of Houston toad (*Bufo houstonensis*) movement through a complex, patchy landscape, and model outcomes were verified by a "Turing test," where a local subject matter expert familiar with toad movement was asked to select between observed and modeled landscape utilization patterns. Model results may also be summarized with visualizations in the application phase of ecological model development, but the model type and audience will play a large role in the selection of the visual approach. For instance, a population model of an oyster reef network was used to develop a network representation of the connectivity between reefs as well as a chord diagram summarizing key ecological aspects of connectivity such as the proportion of larvae moving to and from a given reef (middle row;

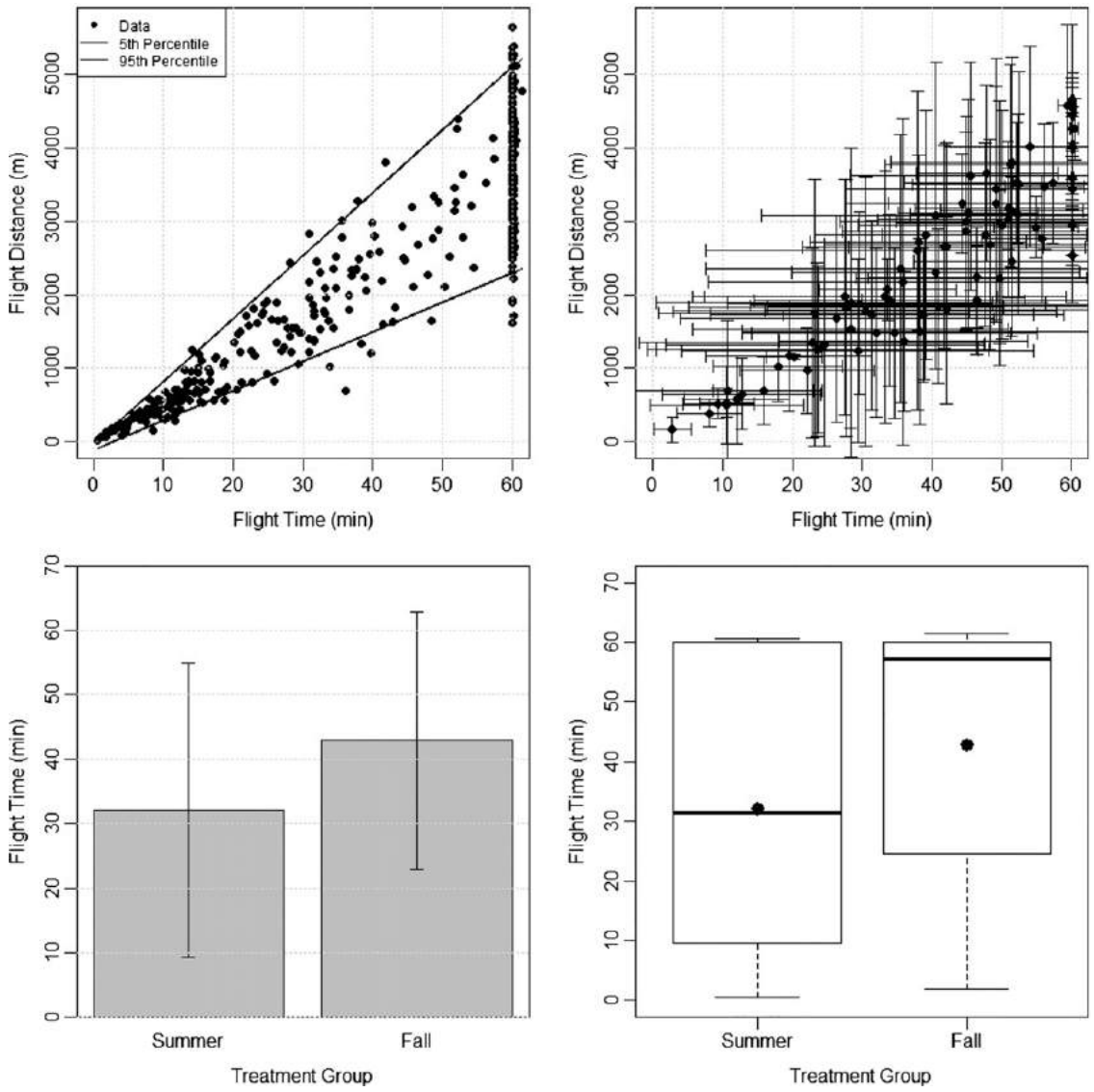


Fig. 2 Visualization of experimental results for monarch butterfly (*Danaus plexippus*) flight trials (McKay *et al.*, 2017a): (top-left) flight data from individual trials with a quantile regression bracketing outcomes, (top-right) flight data summarized by individual butterfly across multiple trials, (bottom-left) bar plot summarizing mean flight duration $\pm$ one standard deviation by treatment, and (bottom-right) a boxplot summarizing the effect of an experimental treatment across the entire sample population where the black point is the mean, black line is the median, box extent is the interquartile range, and whiskers represent maximum and minimum values.

Kjelland *et al.*, 2015). Conversely, the time series visualization methods described above were used to present a set of water management simulations to inform local officials of the magnitude of hydrologic change associated with a municipal water supply (bottom row). While ecological models are often system- or question-specific, visualization methods are agnostic to application, but selecting an appropriate visualization requires careful consideration of the step of the modeling process, goals of the visualization, and structure of the ecological output.

Informing Management Decisions

Visualization methods inform not only the ecological analyses themselves, but also the communication and condensation of those data for management and policy decision-makers (McInerny *et al.*, 2014). For example, an interagency team of federal, state, local, and nonprofit groups is partnering to restore the highly urbanized Proctor Creek watershed in Atlanta, Georgia. Visualization methods have served a vital role in facilitating knowledge transfer between the partners, and here two particular applications are highlighted that benefited from the visualization tools discussed in the article (**Fig. 4**). Urban watersheds represent diverse landscapes with many competing social, ecological, and economic objectives, and stakeholders often hold different relative value across those objectives. In this project, structure decision-making methods are being applied to examine the relative difference in

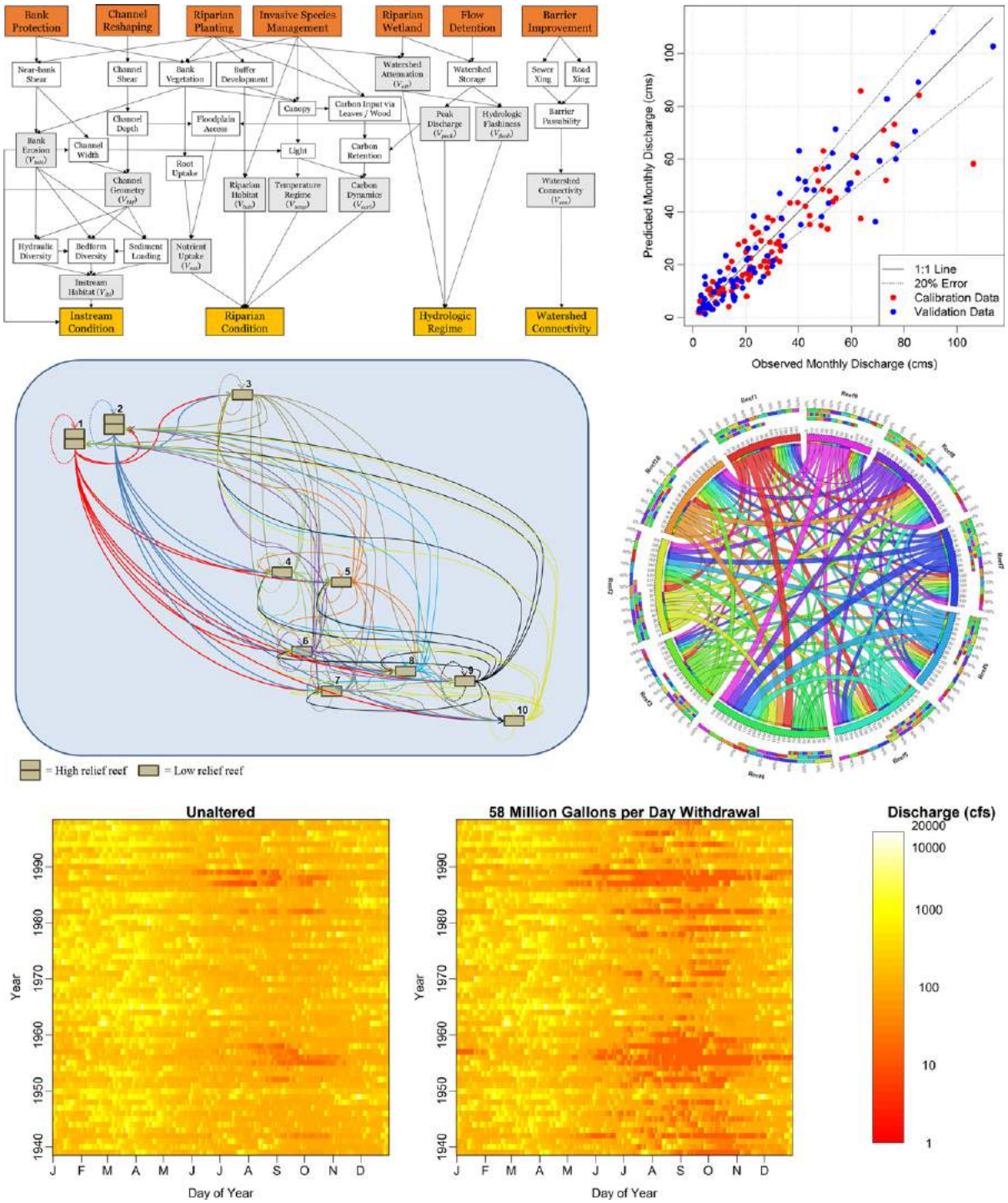


Fig. 3 Visualizations of ecological model results informing all aspects of the modeling process: (top-left) conceptual model of a multivariate model for stream restoration used in Proctor Creek, Atlanta, Georgia, (top-right) hydrologic model outcomes for both calibration and verification analyses (Shrestha, 2016), (middle) Oyster reef larvae dispersal network for a given year in a metapopulation dynamics model illustrating proportion of larvae originating from a given reef and going to other reefs, as well as the proportion of larvae from other reefs going to a given reef (Kjelland *et al.*, 2015), and (bottom row) raster-based time series of river discharge with and without water withdrawal.

stakeholder values across eight primary objectives. While not explicitly ecological data, the family of visualization methods presented in this article were used to summarize these data for communication among these divergent groups. Over 85 stakeholders were asked to distribute 100 "points" across each of the eight objectives, which were subsequently summarized in visualizations which show each stakeholder's values individually (top-left) as well as the aggregate value of all stakeholders (top-

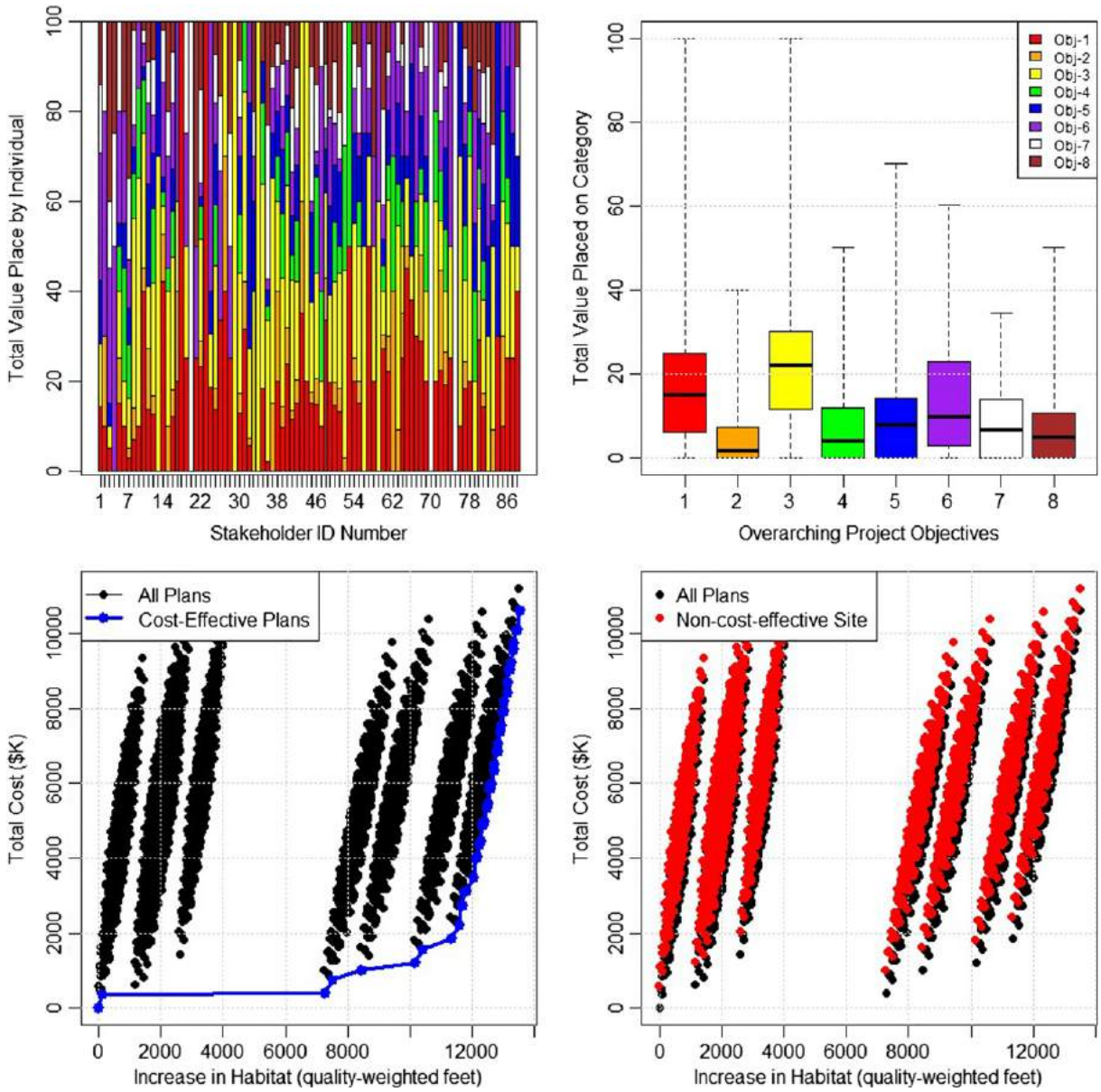


Fig. 4 Sample visualizations used to inform stream restoration decision-making in Proctor Creek, Atlanta, Georgia, United States: (top-left) distribution of values across 8 management objectives for 88 staekholders, (top-right) summary of the distribution of stakeholder values across objectives, (bottom-left) trade-off diagram highlighting sets of restoration actions that provide the most cost-effective restoration schemes, and (bottom-right) trade-off diagram highlighting plans including a particular site which shows that all sets incorporating this location are suboptimal (i.e., there are black points below and to the right of all red points).

right). In these figures, color is used to call information to the eye of the user and create a visual summary of the data (Healey and Enns, 2012), while still showing the range of values. For one of these objectives, an ecological model was developed to quantify the expected change in stream condition for a variety of potential restoration actions at 13 locations (8192 combinations of restoration actions watershed-wide). Model outputs were condensed into trade-off diagrams for decision-makers (bottom row), but the figures were then used to highlight particular aspects of the decision. For instance, suites of restoration actions were identified that were cost-effective (i.e., maximum habitat gain per cost and minimum cost per habitat gain), and color was used to call attention to these for decision-makers (bottom-left). Conversely, restoration sites were identified that were never included in cost-effective actions, which were also summarized with visualization (bottom-right). In communicating for decision and management, visualization must be careful to accurately represent information without imposing a judgment (McInerny *et al.*, 2014). For instance, color selection often implies preference, which can be used appropriately to guide decision-makers away from ineffective actions based on ecological evidence, but analysts must be careful not use visualization to impose their judgment on complex environmental decisions.

Selecting a Visualization Method

The vast array of options for data visualization can make choosing a technique challenging. The following questions are proposed to help ecologists navigate these methods. These questions are merely intended to structure thinking on method selection and are presented in an approximate priority of importance based on the key issues visualization (Aigner *et al.*, 2007, 2011; Kelleher and Wagener, 2011). Importantly, multiple methods can (and should) be applied simultaneously to explore a data set, emphasize specific components of those data, highlight different elements of variability, and guide quantitative analyses.

(1) What is the best method for communicating a message? Purpose and objectives should always drive the selection of a visualization method (Kelleher and Wagener, 2011). For instance, a simple line plot may be sufficient for communicating the time of sampling relative to recent environmental disturbances. However, a pixel-based approach might be more effective for communicating long-term trends (e.g., drought periodicity; **Fig. 1**). In particular, methods might change relative to the components of a data set that are of interest (e.g., extremes vs. central tendencies; **Fig. 2**).

(2) What are the relevant scales? Relevant ecological temporal and spatial scales range from minutes to decades and centimeters to thousands of kilometers in both length and interval. For many ecological processes, valuable understanding can be gained from the presentation of data in the historical context of a long-term record or large spatial domain. The resolution of the data relevant to a particular problem also influences the efficacy of a particular visualization method.

(3) How are data distributed? Ecological data often occur over many orders of magnitude (e.g., changes in streamflow, boom-and-bust of a population). Visualization can often be made more effective by rescaling figures, but care should be taken to maintain understandability in units (e.g., natural units of a process, not transformed units; **Fig. 1**).

(4) What are the constraints of the visualization environment? Selection of methods is influenced by not only the need to present data effectively, but also the limits of the presentation medium (Aigner *et al.*, 2011). The spatial extent, resolution, and use of color in a figure are often limited by screen or page size and the medium of interest (e.g., journal article vs. presentation vs. interactive on-line infographic). This article has focused on static presentation of a single time series, but the capacity to animate or interact with figures can increase visualization options significantly.

(5) What tools and expertise are available? A large variety of software is available to visualize data (e.g., Microsoft Excel, MATLAB, R Statistical Software, Geographic Information Systems). While all of these tools can present data in multiple formats, not all visualizations are easily conducted in all programs. In addition to availability and capability of the software, there may be personnel limitations in terms of expertise or time availability, which should not be overlooked. Herein, the freely-downloadable, R statistical software package was used to develop all figures (R Development Core Team 2016 Version 3.3.2), and code is available upon request from the author.

Conclusions

As ecological analysis has increased in sophistication and rigor, so too have visualization approaches (Kelleher and Wagener, 2011). However, broad adoption of many visualization methods has lagged. Large data, complex, and long-term data sets are becoming increasingly common in ecology (Michener and Jones, 2012). Using a variety of case studies, this article has highlighted a few (of many) techniques for visualizing ecological data and provided a set of criteria for guiding analysts to an appropriate technique. Visualization cannot substitute for rigorous quantitative analyses, but it can inform the analyst, guide the analyses, and facilitate communication (McCormick *et al.*, 1987).

See also: Ecological Data Analysis and Modelling: Mediated Modeling and Participatory Modeling. Statistical Inference. General Ecology: Communication. Principal Components Analysis. Human Ecology and Sustainability: Ecological Systems Thinking

References

Aigner, W., Miksch, S., Muller, W., Schumann, H., Tominski, C., 2007. Visualizing time-oriented data: A systematic view. Computers & Graphics 31, 401–409.
Aigner, W., Miksch, S., Schumann, H., Tominski, C., 2011. Visualization of time-oriented data. London: Springer-Verlag.
Crowl, T., Covich, A.P., Melendez-Colom, E., Perez-Reyes, O., 2017. Shrimp populations in Quebrada Prieta (pools 0, 8, 9, 15) (El Verde). Luquillo Long-term Ecological Research Site. http://luq.lternet.edu/data/luqmetadata54 Accessed 13 March 2017.
Fox, P., Hendler, J., 2011. Changing the equation on scientific data visualization. Science 331, 705–708.
Garbrecht, J., Fernandez, G.P., 1994. Visualization of trends and fluctuations in climatic records. Water Resources Bulletin 30 (2), 297–306.
Healey, C.G., Enns, J.T., 2012. Attention and visual memory in visualization and computer graphics. IEEE Transactions on Visualization and Computer Graphics 18 (7), 1170–1188.
Keim, D., Andrienko, G., Fekete, J.D., Gorg, C., Kohlhammer, J., Melancon, G., 2008. In: Kerren, A., Stasko, J.T., Fekete, J.-D., North, C. (Eds.), Visual analytics: Definition, process, and challenges. Information visualization: Human-centered issues and perspectives. Berlin: Springer-Verlag, pp. 154–175.
Kelleher, C., Wagener, T., 2011. Ten guidelines for effective data visualization in scientific publications. Environmental Modelling & Software 26, 822–827.
Kjelland, M.E., Piercy, C.D., Lackey, T., Swannack, T.M., 2015. An integrated modeling approach for elucidating the effects of different management strategies on Chesapeake Bay oyster metapopulation dynamics. Ecological Modelling 308, 45–62.

McCormick, B.H., DeFanti, T.A., Brown, M.D., 1987. Visualization in scientific computing. Computers and Graphics 21 (6), 1–14.
McInerny, G.J., Chen, M., Freeman, R., Gavaghan, D., Meyer, M., Rowland, F., Spiegelhalter, D.J., Stefanr, M., Tessarolo, G., Hortal, J., 2014. Information visualisation for science and policy: Engaging users and avoiding bias. Trends in Ecology & Evolution 29 (3), 148–157.
McKay, A.F., Ezenwa, V.O., Altizer, S., 2017a. Unravelling the costs of flight for immune defenses in the migratory monarch butterfly. Integrative and Comparative Biology 56 (2), 278–289.
McKay, S.K., Pruitt, B.A., Zettle, B., *et al*., 2017b. Proctor Creek ecological model (PCEM) phase 2: Benefits analysis. ERDC EL-TR. Vicksburg, Mississippi: U.S. Army Engineer Research and Development Center.
Michener, W.K., Jones, M.B., 2012. Ecoinformatics: Supporting ecology as a data-intensive science. Trends in Ecology & Evolution 27 (2), 85–93.
Shrestha, S., 2016. Impact of wood pellet production on water availability: A case study from Northeast Oconee River Basin in Georgia. Athens, Georgia: Master's Thesis, University of Georgia.
Swannack, T.M., Grant, W.E., Forstner, M.R.J., 2009. Projecting population trends of endangered amphibian species in the face of uncertainty: A pattern-oriented approach. Ecological Modelling 220, 148–159.
Tufte, E.R., 2001. The visual display of quantitative information. Cheshire, Connecticut: Graphics Press.
Weissgerber, T.L., Milic, N.M., Winham, S.J., Garovic, V.D., 2015. Beyond bar and line graphs: Time for a new data presentation paradigm. PLoS Biology 13 (4), doi:10.1371/journal.pbio.1002128.

Relevant Websites

www.visual-literacy.org/periodic_table/periodic_table.html—A Periodic Table of Visualization.
http://www.informationisbeautiful.net/—Information is beautiful.
http://survey.timeviz.net/—Survey of methods for time series visualization.
http://www.creativebloq.com/design-tools/data-visualization-712402—"The best 38 tools for data visualization".

Watershed Models☆

Vojtech Novotny, Northeastern University, Boston, MA, United States

Introduction

Since 1960s computerized watershed models have been used by scientists to simulate the hydrological erosion and deposition throughout the watershed and nonpoint pollution loads from the watershed. Recently, ecological watershed models and applications have been developed and are finding their way into practice. The term "watershed modeling" implicitly describes a category of geographical models that simulate movement of water and associated processes that change the quantity and quality of water. This type of modeling may differentiate from other geographical ecological modeling such as the models of forest ecology, impact of drought conditions on crops and crop yields, habitat suitability for fauna and flora, etc.

A "watershed" is a geographical unit contributing flow to a location on a receiving water body. The watershed is bordered by a "watershed divide" that surrounds the watershed, connecting the highest points. The area behind the divide contributes surface and shallow groundwater flows to another receiving water body. However, in many watersheds deeper groundwater flows may not follow the surface geography of the watershed and groundwater contribution to flow may originate from recharge areas beyond the divide, or the groundwater flow originating within the watershed may discharge into another water body.

The category "ecological watershed modeling" includes mostly mathematical computer models. Fig. 1 presents general categories of models for watershed modeling. Physical watershed models, common in the 1940s, still do exist and are used by laboratories but will not be extensively covered herein. In most cases these models have small ecological relevance and have been primarily used for scaled-down development, testing, and simulation of floods, overland flow, rainfall/runoff transformation, and erosion. The largest physical model covering an area of many hectares was built during the 1940s in Vicksburg (Mississippi, USA). This model represented the Mississippi River Valley watershed from Minneapolis to New Orleans and was used for predicting flood flows and flood propagation. Smaller "watershed" models, as small as a platform (Fig. 2), were built and are still used by scientists to study overland flow, washoff of particulates that accumulate on impervious surfaces, and erosion and elutriation of pollutants from soils. Physical model categories are distinguished from physically based models which is a model category synonymous to deterministic–mechanistic computer models.

Distinction between small physical watershed models and small pilot watershed plots may be sometimes fuzzy but one can presume that the former physical models are built in laboratories and are detached from the natural systems while the latter are a part of nature equipped by monitoring instrumentation. Analog models are rare and have been typically used for modeling groundwater aquifer processes.

Today, watershed models are mathematical computer models and range from the typical desktop or laptop computers to high-speed large computers. However, the capacity of the current desktop or laptop computers can accommodate most watershed models. Ideally, ecological models should include three spheres: (1) atmosphere; (2) terrestrial, including shallow underground and riparian zones; and (3) aquatic freshwater and oceanic systems (Fig. 3). In this article the focus is on modeling terrestrial processes that affect water quality and biotic integrity of aquatic systems.

☆*Change History:* March 2018. Todd M. Swannack updated References.

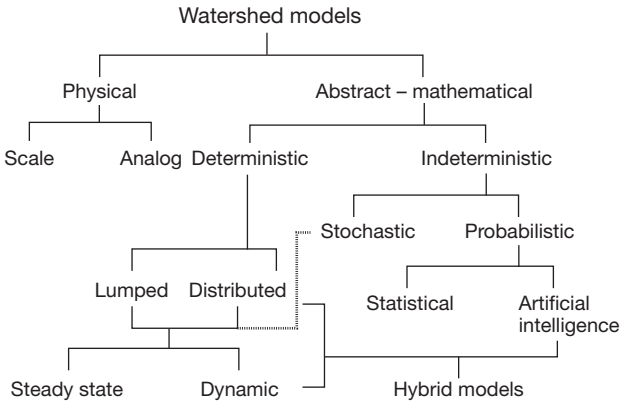


Fig. 1 Categorization of watershed models.

Fig. 2 A platform size physical watershed model (Research Station of the US Department of Agriculture in Tombstone, Arizona).

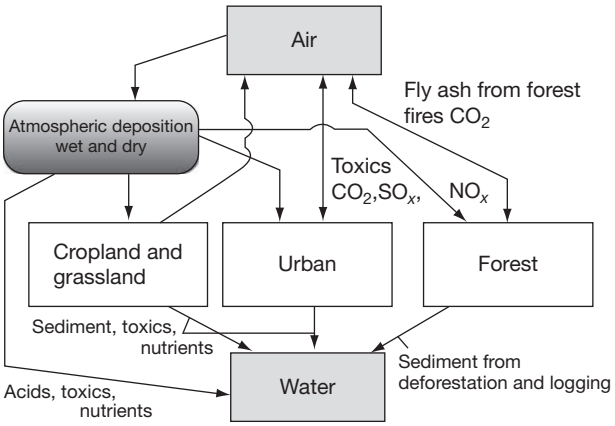


Fig. 3 Spheres of ecological watershed modeling.

Types of Models

The computer watershed models can be separated into two distinct categories (Fig. 1):

1. deterministic, also called mechanistic, models and

2. indeterministic models that include probabilistic models and artificial intelligence (e.g., artificial neural net or genetic algorithm models).

Deterministic (mechanistic) models are generically a priori developed from an assembly of known processes, for example, hydrologic processes included in rainfall/runoff transformation, accumulation of pollutants in the watershed, vegetation growth, and hydrological and pollution impact, etc. The model components are then tested by special field monitoring and in the final stage the entire model is then calibrated and verified by field data for modeling a specific watershed. Indeterministic models are typically developed a posteriori from measured data. The data assemblies should be large enough because the basic premise of these models is that the models are developed from the data (data mining) and the previous knowledge about the processes is only basic. Hybrid models, however, combine the knowledge-based deterministic features with data mining and model creation by artificial intelligence models. Stochastic models yield both the mean and probability distribution of outputs. The most common stochastic modeling is Monte Carlo application, where a deterministic model is run thousands times with randomly generated (based on their probability distribution) inputs and model parameters to yield a probability distribution of the model output.

Deterministic models do not consider random variables and for each unique set of input data they produce fixed repeatable results. A mechanistic model is a representation of the physical, biological, or mechanistic theory governing the system; in contrast, a statistical model accounts for the statistical fitting of equations to the available data. Stochastic models use distribution for each variable and parameter to generate random variables of model inputs and system parameters.

Every watershed model can be represented by a box (the watershed) concept (Fig. 4). The term "black box" signifies that not much is known a priori about the system and the model is developed from the monitored data.

Deterministic (Mechanistic) Watershed Models

Watershed Loading Models

Since the 1960s, scientists focused on putting the knowledge of hydrological processes constituting rainfall–river flow transformations into a watershed hydrological model. The most known and the first comprehensive watershed model was the Stanford Watershed Model conceived by Professor Linsley and his graduate students. The model was sound and the concepts and the basic structure of the model are valid today and incorporated into modern hydrologic watershed models. Hydrology and hydrological processes are the backbone of all deterministic watershed models. However, ecological models represent and simulate more than hydrology and have been developed and are used for simulating erosion and sediment movement, pollutant loads, crop and forest growth and resulting loads of nutrients, spread of forest fires, surface–groundwater interactions, drainage, etc.

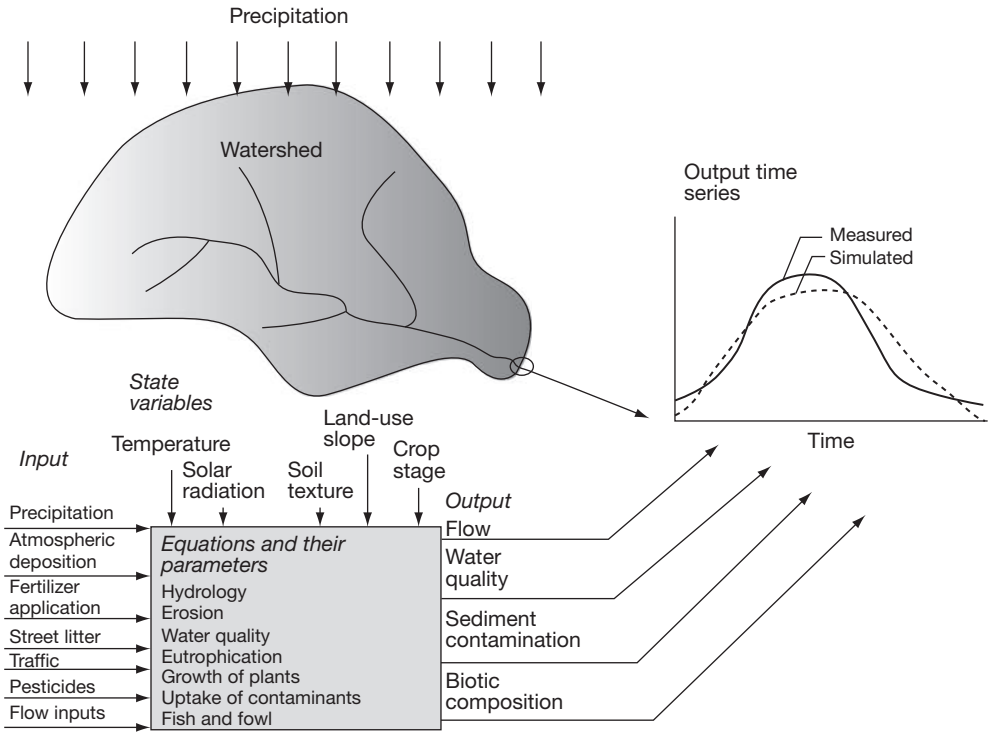


Fig. 4 Black box representation of watershed models.

A detailed deterministic or stochastic ecological model has five components: (1) inputs or forcing functions, (2) state variables, (3) mathematical equations, (4) parameters of the models, and (5) constants. Inputs can be controllable or uncontrollable. The model, like the real system, produces outputs to various inputs and the outputs also reflect the changes in the system itself. The variables describing the system are called "state variables." The distinction between inputs and state variables is sometimes fuzzy but, typically, watershed size, soil composition and erodibility, land-use distribution, watershed configuration with slope are examples of state variables while rainfall, atmospheric deposition, temperature, humidity, and solar radiation are inputs. Most watershed models are driven by precipitation which is an uncontrollable input, while fertilizer application is a controllable, managerial input. The foundation of the model is in equations that describe the input to output transformation. The parameters of each equation may have ranges and the proper value of the parameter must be typically established by calibration and verification of the model. The equations also contain constants and thresholds. A threshold is a constant that activates or terminates a process described by a particular equation or a submodel. For example, many biodegradation processes cease when the temperature is near or below freezing or a priority pollutant is not toxic below a certain concentration.

There are two approaches to modeling watershed processes: "lumped parameter" and "distributed parameter models" (Fig. 5). Lumped parameter models can be both stochastic and deterministic. Distributed parameter models are mostly deterministic. The lumped parameter models treat the watershed (subwatershed) as a homogenous unit. Because of a significant variability of the parameters, even within a relatively small area (e.g., less than 1 ha), the various characteristics of the watershed are lumped together by an empirical equation, and the final form and magnitude of the parameter are simplified to represent the computational watershed as one homogenous unit. The computational units can be vertically compartmentalized to represent surface–ground zones interactions. Water and mass flow from one compartment can overflow into another adjoining compartment. For lumped parameter models the input–output relationship can be represented as

$$\mathbf{Y} = \mathbf{\Phi}(\mathbf{X}) \tag{1}$$

where **X** is the input vector or matrix (single or multidimensional), **Y** is the output vector or matrix (single or multidimensional), and Φ is the multidimensional transfer function.

According to this definition, a lumped watershed model is a multidimensional transfer function. The transfer function Φ does not represent a simple multiplication. In most general terms, a transfer function is a functional response of the system to a pulse (in dynamic models) or a step (in steady-state models) input. A well-known unit hydrograph incorporated in many lumped parameter

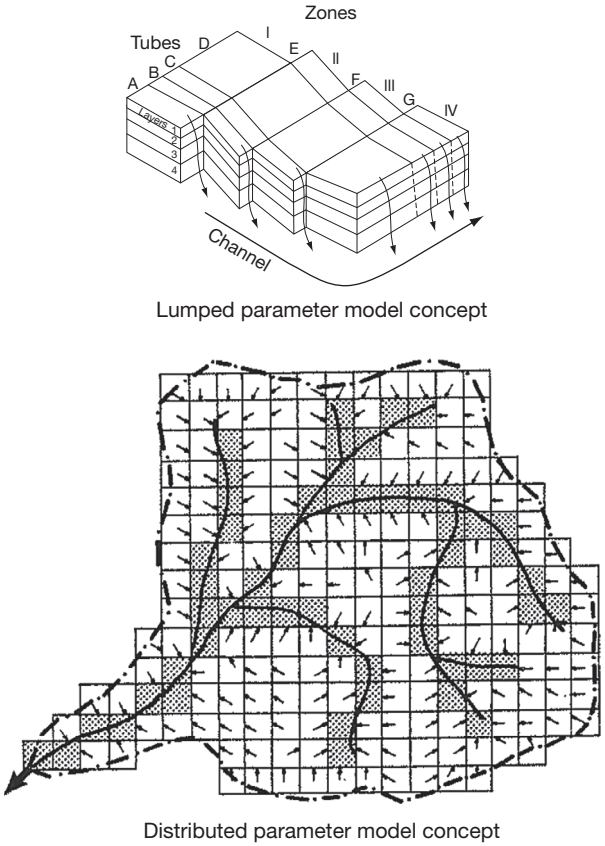


Fig. 5 Lumped (vector) and distributed (raster) parameter concepts for watershed modeling.

hydrologic models is a watershed flow response to a short duration (pulse) unit rainfall while the "rational (Lloyd–Davies) formula" is the watershed flow response to a step (continuous after initiation) uniform.

In the distributed parameter models the watershed is divided into computational subunits that are much smaller than those of the lumped parameter. The size of the subunits may range from 1 ha to more than 1 km^2. The subunit shapes are either rectangular (square) or triangular or they may follow some morphological or hydrological feature of the watershed. Each unit may receive external and internal (from surrounding units) input and is homogenous. The general mass-balance equation for any component within the computational unit is

$$\frac{dS}{dt} = X - Y \pm G \tag{2}$$

where S is the mass storage of the compound in the computational element, G is the sinks (losses) and gains of the compound within the element, X is the external input to the element, and Y is the output.

The major difference between the lumped parameter and distributed parameter models is the mathematical formulation of the mass balance in each unit that for distributed parameter models is generally expressed by a differential mass-balance equation. The differential mass-balance equations for each computational unit are then solved simultaneously in a small computational time interval, Δt, which ranges from minutes for the hydrologic calculation of a small watershed, to a day or more for quality constituents or crop modeling in larger watersheds. Another difference between the lumped and distributed parameter models is the availability of the output. In lumped parameter models the output is available only at the watershed or subwatershed outlet. In distributed parameter models, results are available for each computational unit which makes distributed parameter models attractive for multidimensional displays in the raster geographical information systems (GIS) modeling environment.

Watershed computer models come in all sizes, from 1 ha uniform experimental modeling plots or hydrological subunits to regional large models with a size of thousands of km^2. Large watershed models are dominated by channel storage while smaller watersheds are sensitive to precipitation events and are dominated by overland flow.

Models can be designed or run on an "event or continuous" basis. Event models simulate the response of a watershed to a single large rainfall or another major input (e.g., forest fire). The output from event models is a hydrograph of loads of a constituent from the watershed. The principal advantage of event modeling over continuous modeling is that it requires less meteorological data. The disadvantage of even modeling is that it requires a definition of the "design" storm and antecedent state of the system prior to the event.

Continuous (dynamic) process modeling sequentially simulates processes such as precipitation, water and pollutant storage and movement within the watershed toward the watershed outlet, runoff components, evaporation and transpiration, uptake of pollutant and nutrients by vegetation and their release into aquatic systems, and erosion. The output is the time series of flow and quality constituents from the watershed. Continuous modeling provides time series of the output (flow, concentrations, soil and vegetation contamination). These time series can be statistically analyzed but with caution and warning. The simulated time series are not the "true" time series because the deterministic model inherently does not include random components contained in every time series of natural data. The models are calibrated to simulate local means but not extremes. Thus the statistics of exceedances or comparisons of the time series of outputs generated by deterministic models with ambient standards could be misleading. This problem could be overcome with Monte Carlo modeling methodology which would then categorize such applications as stochastic modeling.

Traditional deterministic watershed models are mostly linear. Linearity means that the equations in the model compartments are either algebraic or differential linear equations. If the mathematical formulation describing the process is nonlinear, it is often linearized within the short computational interval. Under the assumption of linearity, the principles of proportionality and superposition apply. The principle of proportionality implies that if the input is multiplied by a constant the output is also multiplied by the same constant. The superposition means that if two separate inputs to the model are added, the output of the model is the addition of the separate outputs. At the end of twentieth century and thereafter the focus of model developers shifted to development of nonlinear models.

Components of the Deterministic Watershed Ecological Models

Deterministic watershed models are compartmental models, that is, they describe and simulate the transfer of water and biomass from one compartment of the ecosystem to another. The ecological models are almost always dynamic; they calculate mass balances in specific time intervals that do not have to be identical for each compartment. For example, water balance and overland water routing can be done in time intervals of an hour or less, while the growth of plants or mass balances of water quality and algae in a receiving lake can be accomplished in time intervals of weeks or months. The models simulate the fate of water and its quality (chemical) constituents and also their direct and indirect effects on the resident organisms. The mass balance can be calculated sequentially or concurrently. Watershed ecosystem models employ a compendium of submodels (Fig. 6).

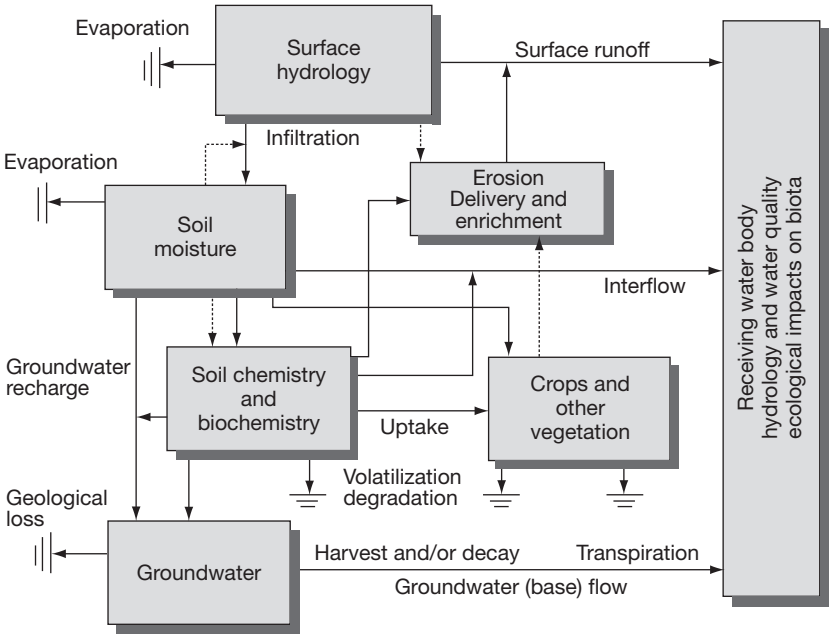


Fig. 6 Suite of basic submodels in watershed modeling. *Solid line arrows* represent transport processes for water and contaminants (including sediment and nutrients), *dot arrows* are impact effects.

Hydrological Component

This component is the backbone of most deterministic watershed ecological models that perform transformation of precipitation into water flow. The watershed water-balance models are either built in a "lumped subwatershed format" (vector representation in GIS) where computational subunits have more or less uniform characteristics or in a "distributed format" (raster representation in GIS) (Fig. 5). The water movement throughout the watershed is calculated in the Lagrangian coordinate system of fluid mechanics in which Z is vertical, and X and Y are lateral axes. X is in the main directions of flow, that is, it may not be perfectly horizontal. The hydrological mass balance is carried out vertically in three zones: (1) on the surface, (2) in the top soil, and (3) in the groundwater zone. Each zone represents a storage component and the mass balance according to Eq. (2). There are many literature sources dealing with hydrology and hydrological modeling (see section on "Further Reading").

The "surface hydrology" modeling component performs the following calculations:

Snowmelt calculation. Based on the ambient temperature, this component makes a decision whether the precipitation is in a form of liquid rain or snow. It then calculates snow accumulation and melting.

Initial subtraction from rain or snowmelt. Water that is intercepted by vegetation, surface roughness, or terrain is returned back to the atmosphere by evaporation.

Infiltration: Infiltration into soil depends on the top soil texture, moisture content, and freezing.

Excess or net rain. This occurs when the precipitation (rain or snowmelt) at the soil surface exceeds the sum of infiltration and surface subtraction. The surface water component accepts solids with associated contaminants liberated from top soil by erosion and leaching. Surface runoff is excess rain transformed by overland routing. Surface runoff is the most polluted component of flow in the receiving water bodies.

Top soil layer water balance. This has infiltration as input and evaporation, vegetation water uptake (transpiration), and groundwater recharge as outputs. Groundwater recharge depends on saturation permeability of subsoil. 'Interflow' in the top soil occurs when the top soil layer is saturated and infiltration exceeds the downward movement of water through the subsoil.

Groundwater recharge. This is a residual of infiltration into top soil storage evapotranspiration. 'Groundwater flow' is a near-horizontal gravity movement of water from the recharge area to surface outlets into the receiving waters. A smaller portion of groundwater is lost to deep groundwater zones and to anthropogenic withdrawals.

Interaction of Contaminants With Soils and Surface Vegetation

This component simulates retention and attenuation of chemicals on soil particles of top soil and on dust and dirt particles that accumulate on impervious surfaces of street gutters and other impervious areas. Chemical contaminants are either hydrophilic that are mostly dissolved or dissociated in pore water of the soil and in groundwater, or hydrophobic that have a high affinity for adsorption on soil particulate matter or precipitate in soil. Nitrate ((NO_3^-)) is an example of a strongly hydrophilic compound that moves readily with soil water into groundwater recharge and with groundwater. Many pesticides and phosphate fertilizers are mostly hydrophobic and remain adsorbed onto soil particles or precipitate into solid forms, such as most toxic metals at a higher

pH. However, a small portion of these pollutants can exist in the dissolved form in the soil pore water. The relationship between the concentration of the chemical in pore water of soil and the total concentration is

$$C_s = \frac{C_T}{\theta + \Pi m_{ss}} \tag{3}$$

where C_s is the concentration of the soluble contaminant fraction in the pore water of soil ($\mu g\, L^{-1}$), C_T is the total contaminant concentration in the soil volume, ($\mu g\, L^{-1}$), θ is the volumetric water content of the soil, Π is the partition coefficient of the contaminant on soil particles ($1\, g^{-1}$), and m_{ss} is the soil density in $g\, L^{-1}$.

Other isotherms used in modeling specific pollutants are Langmuir isotherm for phosphorus or Freundlich equations for some toxic compounds.

The soil component separates the contaminants into dissolved and adsorbed (precipitated) fractions. Also some contaminants can undergo biochemical transformation (e.g., nitrification, biochemical breakdown) that often depends on the redox status of the soil which is affected by soil moisture.

Crop/vegetation growth: This component is important for agricultural or forest growth models. Plants withdraw soil moisture, use for buildup of organic plant matter and release some of the moisture to the atmosphere as transpiration. The withdrawal is limited by the wilting point which is soil moisture content held by soil so tightly that plants cannot overcome it by their suction. With water, plants can get their nutrients and also pick up contaminants. Plants can only take up the dissolved mobile fraction. The adsorbed immobilized fraction is not available for uptake.

Soil Erosion Component

Erosion moves sediment and contaminants adsorbed on or associated with the sediment from the source area to the receiving water body. In modeling, sediment is routed overland by surface runoff; however, not all eroded sediment will reach the receiving water body. The attenuation of the sediment loads and associated contaminants during overland flow is expressed by a fuzzy parameter called the delivery ratio, DR, or

$$Y = \sum_{i=1}^{N} DR_i E_i A_i \tag{4}$$

where Y is the total sediment yield from the watershed, E_i is the erosion rate from the segment I, and A_i is the area of the segment.

Erosion rate depends on the energy of the rainfall to dislodge soil particles from the soil. In addition, erosion is activated only when excess rain is generated that will become surface runoff. Hence, erosion is linked to the surface hydrology component. The most widely used erosion model was developed by Wischmeier and Smith. The model is more than 40 years old and has been modified several times but the fundamentals are sound. The erosion equation is

$$E = (R)(K)(LS)(C)(P) \tag{5}$$

where E is the calculated soil loss in $t\, ha^{-1}$ for a given storm or season or a year, R is the rainfall energy factor that has the same units as R, K is the soil erodibility factor, LS is the slope/length factor, C is the cropping management or vegetation cover (protection) factor, and P is the erosion control factor.

The magnitude methods of calculating the factors are presented in many publications. The "universal soil loss equation" considers the impact of "best management practices" on the loads sediment and associated pollutants.

Examples of Watershed Loading Models

HSP-F. This lumped parameter model evolved from the Stanford Watershed Model. It is now included in the US EPA's watershed modeling suite "BASINS" (better assessment science integrating point and nonpoint sources). The model has several components. The components used for modeling hydrology and pollutant loadings are "the agricultural runoff management model" (ARM) and urban nonpoint simulator (NPS). Both are driven by the hydrological component and the main difference between the two submodels is the soil and groundwater modeling in ARM, and impervious surface hydrology, pollutant accumulation, and washoff incorporated in NPS. By including a channel component (RCHRES) HSP-F has a capability of integrating the land and receiving water quality modeling. The result of the modeling is a time history of flow and water quality at any point in the watershed.

Agricultural nonpoint source pollution model (AGNPS). This was developed by the US Department of Agriculture. The primary emphasis of the model is on nutrient and sediment, and on comparing of various best management practices on the pollutant loading.

Soil and water assessment tool (SWAT). This is a distributed parameter version of previous models "CREAMS" developed by a team of agricultural researchers lead by Knisel and a follow-up model with a groundwater component "GLEAMS." SWAT is spatially distributed, up to several hundred of subbasins, and the subbasins can interact. The SWAT hydrology is based on the water-balance equations and uses National Resources Conservation Service (formally soil conservation service) runoff equation expanded to also consider soil and groundwater movement. It can also simulate irrigation and channel transmission. Erosion is calculated by the

modified universal soil loss equation (MUSLE) developed by Williams and includes overland and channel sediment routing. Arnold incorporated into SWAT model GIS linkage, advanced visualization tolls, and statistical analysis of outputs.

Receiving Water Quality Models

In deterministic ecological watershed modeling, each compartment should be described by a model. Models of hydrology, contaminant transformation, and movement throughout the watershed provide estimates of the water and contaminant loads to the receiving water body. The format of the loads is time-variable hydrographs or histograms of the contaminants (plots of the pollutant contaminant concentrations vs. time), daily loads, or seasonal (annual) loads for larger impoundments with a longer residence time. A riparian zone model can be also considered to provide information on attenuation of pollutant loads (Fig. 7).

Because the endpoint of the modeling effort is the ecology of the water body that receives the loads from the watershed and from the point sources, a receiving water quality model should be the last component of the modeling effort. Fig. 7 shows a schematic of the ecological receiving water body model. Typically, the receiving water model is a stand-alone model which receives the inputs from the watershed loading model by the file transfer, or the receiving water body model is a part of the overall modeling package (e.g., HSP-F, BASINS).

In, general, an ecological model is an expanded version of a deterministic water quality model with added flora and fauna components. The first ecological models linked with water quality were models simulating growth of algal population which signifies eutrophication. These models can be steady state such as QUAL 2-E and its derivatives, or dynamic, such as WASP4. Deterministic modeling of aquatic flora is more demanding, more complex, and less reliable. An extensive review of models for receiving waters and wetlands is in a publication by Straškraba listed in the section on "Further Reading."

The key water quality components depicted on Fig. 6 that impact the aquatic flora and fauna are "organic matter" (both particulate and dissolved), "suspended sediment," "nutrients" (nitrogen and phosphorus), "toxic (priority) pollutants," and "dissolved oxygen." Nutrients and toxic compounds in the water column are both particulate (part of suspended sediment) and dissolved. Only dissolved and dissociated nutrients and contaminates are available for uptake by flora and fauna populations.

An important component of the ecological receiving water model is benthic sediment. Legacy pollution is stored in the sediments as well as organic matter that is slowly anaerobically decomposed. The benthic decomposition releases methane, ammonium, and phosphates. The rate of release depends on the redox status of the interstitial sediment–water layer. If this layer is aerobic, oxides reduce the release of phosphorus, and ammonium is oxidized to nitrate and methane to carbon dioxide. DïToro with coworkers proposed a "sediment oxygen demand" (SOD) model that considers oxidation of methane and ammonium in the interstitial layer. Nitrate, on the other hand, because of its concentration gradient, infuses partially back to the lower anoxic layers of the sediment where it is reduced to nitrogen gas. This simultaneous nitrification/denitrification results in a loss of nutrient nitrogen from the water body.

The processes described and simulated by the ecological models are numerous and described throughout the encyclopedia. They include biodegradation of organic matter imposing demand on oxygen resources in the water body, adsorption–desorption equilibriums of toxic compounds with the particulate organic matter and suspended sediment, uptake of dissolved nutrients and toxics by flora and fauna that impacts their growth, growth and dieoff modeling for flora and fauna, dissolved oxygen-balance

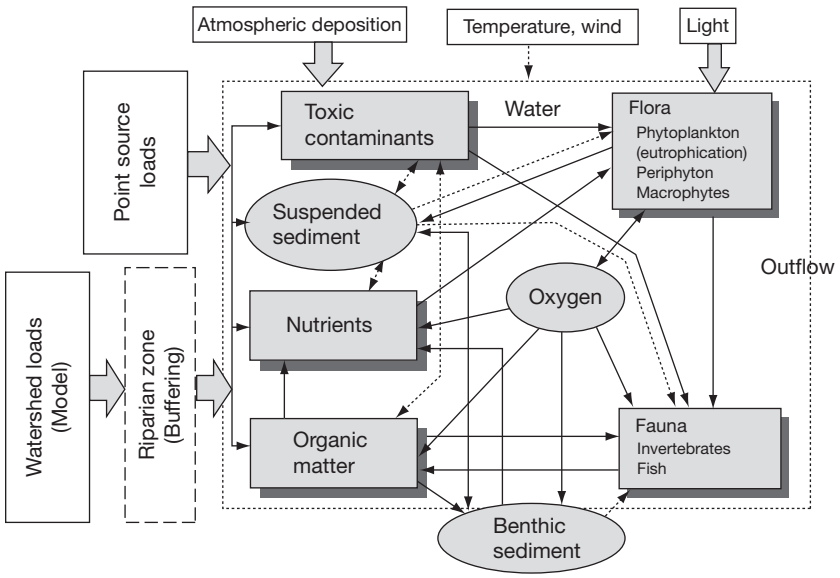


Fig. 7 Schematics of the water–sediment ecological model.

model, settling and scour of cohesive sediments, and nitrification. Many reactions and processes are temperature dependent. There are several links of water quality constituents to the flora and fauna that are not yet quantitatively understood. For example, one of the parameters that most affect the integrity of fish and the macroinvertebrate population is "embeddedness" which describes the amount of clay embedding the gravel and rocks in the bottom substrate. High embeddedness interferes with spawning of fish. This parameter is observational, that is, predictive models were not available at the time of writing this article. It is related to the amount of suspended solids load of the receiving water body and its geomorphic and hydraulic parameters.

Example of Ecologic Water Body Models

AQUATOX is the latest in a series of models developed by Richard Park, starting with the aquatic ecosystem model CLEAN that was subsequently improved and released by the US Environmental Protection Agency in 2005. AQUATOX simulates transfer of biomass, energy, and chemicals from one compartment of the ecosystem to another in daily intervals. The model includes several interrelated components of aquatic ecosystems, including multiple species of phytoplankton, periphyton, and submerged aquatic vegetation, planktonic and benthic algae, forage, game and bottom fish, nutrients and dissolved oxygen, organic and inorganic sediments, and toxic organic chemicals. AQUATOX is now an extension of US EPA's BASINS model that provides linkages to GIS data, HSP-F, and SWAT simulation models.

Regression-Based Statistical Models

Regression allows modeling the dependence of (usually) one response variable on one or more predictor variables. The simple models are based on a stepwise progression of univariate and multivariate analysis. Simple linear regression involves discovering the equation for a line that best fits the given data. These models are easier to test in replication and cross-validation studies. Furthermore, they are less costly to put into practice in predicting and controlling the outcome in the future. In watershed ecological modeling, multiregression models have been used to find relationships between the key chemical (e.g., nitrate concentrations in the receiving water bodies) or biotic (e.g., indices of biotic integrity (IBI), diversity indices) and key watershed stresses such as percent of polluting land uses (imperviousness, agriculture), landscape parameters (slope), habitat quality, etc. A growing number of studies have established relationships between some landscape metric and a biological endpoint.

Principal Component Analysis Multiregression Models

More recent improvements of multiregression statistic such as "principal component analysis" (PCA) can alleviate problems with cross-correlations between multiple inputs. PCA enables grouping of a large number of multiple input parameters that may be cross-correlated into fewer independent variables. The work and paper by Yuan and Norton described extraction of a multi-parameter model from measured biotic integrity (response endpoint) and physical/habitat and chemical stressors (inputs) in watersheds in western Ohio.

As any multiple-regression method, PCA can only analyze single-parameter output matrix, for example, an overall IBI value or a single nutrient concentration versus multiple-parameter input vectors. However, IBI (fish, macroinvertebrates) are composites of and calculated from multiple metrics. Extracting simultaneously knowledge on the metrics relation to the input parameters and to the overall IBI would be very tedious and would have to be done individually, metric by metric.

Canonical Correspondence Analysis

Canonical correspondence analysis (CCA) and similar correspondence analysis models are also special cases of multivariate regression described extensively in a monograph by P. Legendre and L. Legendre (see the section titled "Further Reading"). CCA is a direct gradient technique that can, for example, relate species composition directly and intermediately to the input environmental variables. CCA combines correspondence analysis (CA) with multiple regression, whereby the measured endpoints (dependent parameters) are related to measured environmental, landscape, and habitat stresses as shown in works of Ter Braak and Palmer. CCA and CA are weighted-average ordination techniques that provide simultaneous ordering of sites and species, rapid and simple computation, and very good performance when species have nonlinear and unimodal relationships to environmental gradients. CCA allows simultaneous plotting of output species variables and their site scores in an ordination diagram known as joint plot. Fig. 8 shows the two-axis plot of the 25 variables affecting the IBI in Ohio streams. The plotting of variables when appropriate scaling is used is represented by an arrow (vector) in the direction of the output variables or clusters and the length of the arrow represents a measure of significance of the variable. Furthermore, the direction of the arrows indicates not only how closely the input variable is correlated with the output species variable, but also how closely the input variables are correlated to each other. For two arbitrarily chosen input variables, if their arrow vectors are close to each other or on the same line they are closely correlated. If they are normal (90 degree angle) to each other they are uncorrelated. CCA is a powerful method for the multivariate exploration of large-scale data. CCA preserves the chi-square (χ^2) distance between the rows and columns of the contingency table. Palmer documented that CCA is a weighted-average ordination technique that provides simultaneous ordering

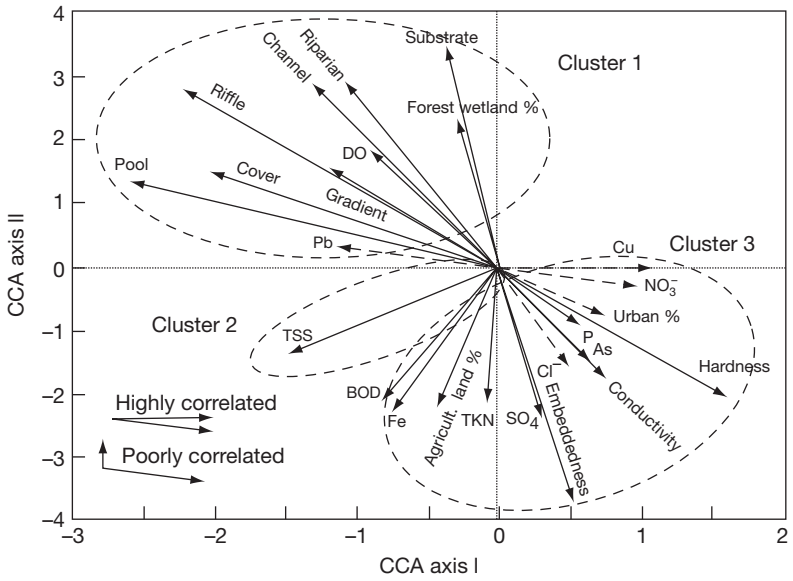


Fig. 8 CCA two-axis plot of the watershed stressors on the metrics of the index of biotic integrity (IBI) derived from the data collected in Ohio. Cluster 1 contains sites with good integrity, Cluster 2 is intermediate, and Cluster 3 has inferior sites. The *arrows* represent the cluster-dominating parameters. The two axes represent about 50% of variability. Replotted from Virani, H., Manolakos, E., and Novotny, V. (2005). Self-organizing feature maps combined with ecological ordination techniques for effective watershed management. *Technical Report # 4*. Boston, MA: Center for Urban Environmental Studies, Northeastern University. http://www.coe.neu.edu/environment.

of sites and species, rapid and simple computation, and very good performance when species have nonlinear and unimodal relationships.

The vectors of stressors on Fig. 8, taken from the work by Virani and coworkers, were combined with the clusters of metrics of the IBI to determine the cluster-dominating parameters. Clustering of metrics was accomplished by "self-organizing mapping" (invented in Finland by Kohonen) by the unsupervised learning of the ANN modeling. It can be seen that the inferior Cluster 3 sites were dominated by pollution caused by pollutants while the superior quality Cluster 1 sites were dominated by habitat parameters.

The explosive growth of information technology has given a major boost to the development and implementation of modeling strategies in watershed management and restoration. GISs are becoming important components of simulation models and decision systems, increasingly being used to store both georeferenced data and associated attributes. GIS technology has played critical roles in all aspects of watershed management, from assessing watershed conditions through modeling impacts of human activities on water quality and to visualizing impacts of alternative management scenarios. GIBSI is one such integrated modeling system comprising of physically based simulation models (hydrological, soil erosion, agricultural–chemical transport, and water quality), a relational database management system and a GIS.

21st Century Developments in Watershed Modeling

Novel techniques such as fuzzy logic, artificial neural networks, genetic algorithms, or ABM are increasingly tested against ecological data. The advantage of these approaches is the self-selection of the critical model inputs based on their data-driven approach.

The "land transformation model" (LTM) described by Pijanowski and coworkers employs GIS and ANNs to describe the influence of landscape changes on ecosystem integrity of large areas. The LTM currently employs a multiplayer perceptron (MLP) neural net topology with one or two hidden layers; each layer has at least the same number of nodes as the number of input vectors. ANNs are used to learn the patterns of development in the region and test the predictive capacity of the model, while GIS is used to develop the spatial, predictor drivers and perform spatial analysis on the results.

Guertin and coworkers have been developing new tools such as global positioning systems (GPS) and remote sensing to inventory and monitor watershed characteristics. The tools ArcPad, developed in the ArcPad environment, implements GISs, mobile computing systems, satellite and aerial images, and network interconnection in the frame of standard ecological methodology. All these tools provide a better understanding of how ecological systems are managed on various levels of the ecological research.

Hybrid ABMs recognize the fact that organisms, including humans, do not behave in an ecological system uniformly *en masse* but their growth, movement, and dieoff are a result of outside stimuli, system thresholds, and their own individualities that have a great degree of randomness. Such models can trace animals, algal species, fish, etc. Because of the large number of individuals (e.g., the

number of cyanobacteria in an algal bloom is 10^4–10^4 mL^{-1}), a "superagent" represents thousands of similar species who are followed by the model throughout their life cycle.

The reasons for the new look at watershed modeling and the new direction of modeling development are

1. Key processes of hydrological and ecological systems to be modeled are nonlinear while the current deterministic watershed models are mostly linear (described by linear differential equations) or the equations that originally were not linear were simplified to linear representations. While a nonlinear watershed model can be deterministic, the nonlinearity complicates the model structure and solutions of the equations.
2. The typical impacts of contaminants and other stressors on flora and fauna have thresholds, that is, a minimal or even beneficial (e.g., nutrients, some metals) concentration effect on the biota below the threshold limit and nonlinear detrimental effect on biota beyond the toxicity threshold. This leads to sudden shifts in species occupying the terrestrial and aquatic environments and clustering. Ecological clustering describes a similar and steadier response (resilience in species population or diversity) of the aquatic and terrestrial biota to smaller perturbation in input stresses. Clusters may also be sensitive to different stressors as shown in Fig. 8 for the metrics of the fish IBI. Most deterministic models, as stated before, are proportional, that is, they result in a corresponding change in populations when any input changes unless a threshold (an "if–then") block is incorporated into the model.
3. Large databases of ecological (e.g., fish, macroinvertebrate, phytoplankton, periphyton numbers, composition, and IBI), chemical, and habitat quality data have been collected by the many agencies. The mega databases containing hundreds of thousands of measurements over larger, often multinational regions cannot be modeled by classic deterministic models that are mostly suitable for smaller watersheds. The models and the relationship must be retrieved by "data mining."
4. Complex interdisciplinary and interuniversity collaborative monitoring and model development projects are being promoted by funding agencies both in the US and European Community. In the US, the National Science Foundation program CLEANER (Collaborative Large-scale Engineering Analysis Network for Environmental Research) and CUAHSI (Consortium of Universities for the Advancement of Hydrologic Sciences, Inc.) plan developing a dual purpose Water and Environmental Research Systems (WATERS) Network which will consist of highly instrumented field facilities for acquisition and analysis of environmental data. The system will also provide environmental cyberinfrastructure for data archiving and networking among community members, and information technology for engineering modeling, analysis, and visualization of data.

Further Reading

Abbott MB and Refsgaard JC (1996) *Distributed hydrological modeling*. Kluwer Academic Publishers: Netherlands.

Allan JD (2004) Landscapes and riverscapes: The influence of land use on stream ecosystems. *Annual Review of Ecology Evolution and Systematics*. 35: 257–284.

Ambrose RB Jr., Wool TA, Connolly JP, Schranz RW, and WASP4 (1990) *A Hydrodynamic and Water Quality Model, Model theory, User's guide, and Programmer's guide*. Athens, GA: Environmental Research laboratory, US Environmental Protection Agency.

Arnold JG, Srinivasan R, Muttiah RS, and Williams JR (1998) Large scale hydrologic modeling and assessment Part I: Model development. *Journal of American Water Resources Association* 34(1): 73–89.

Arnold JG, Srinivasan R, Muttiah RS, Allen PM, and Walker C (1999) Continental scale simulation of the hydrologic balance. *Journal of American Water Resources Association* 35(5): 1037–1052.

Band LE, Peterson DL, Running SW, et al. (1991) Forest ecosystem processes at the watershed scale: Basis for distributed simulation. *Ecological Modelling* 56: 151–176.

Brown LC and Barnwell TO Jr. (1987) The enhanced stream water quality models QUAL2E and QUAL2E-UNCAS. In: *Documentation and User Manual, Report EPA/600/3–87/007*. Athens, GA: US Environmental Protection Agency.

Carpenter S, Brock W, and Hanson P (1999) Ecological and social dynamics in simple models of ecosystem management. *Conservation Ecology* 3(2): 4.

Brooks KN, Ffolliott PF, Gregersen HM, and Thames JL (1991) *Hydrology and the management of watersheds*. Ames, IA: Iowa State University Press.

Chapra SC (1997) *Surface water quality modeling*. McGraw Hill: New York.

Chow VT, Maidment SR, and Mays KW (1988) *Applied hydrology*. New York, NY: McGraw-Hill.

Clarke RT (1998) *Stochastic processes for water scientists: Developments and applications*. New York: Wiley p. 183.

Crawford NH and Linsley RK (1966) Digital simulation in hydrology: Stanford Watershed Model IV. In: *Stanford University Department of Civil Engineering Technical. Report 39*. Palo Alto, CA: Stanford University.

DeBarry PA (2004) *Watersheds: Processes, assessment and management*. New York, NY: Wiley.

DiToro DM (2000) *Sediment flux modeling*. Wiley Interscience: New York.

DiToro DM, Paquin PR, Subburamu K, and Gruber D (1990) Sediment oxygen demand model: Methane and ammonia oxidation. *Journal of Environmental Engineering* 116(5): 945–986.

Duan Q, Sorooshian S, Gupta HV, Rousseau AN, and Turcotte R (2003) *Calibration of watershed models, water science and application series*. vol. 6. American Geophysical Union. ISBN: 0-87590-355-X.

Folke C, Carpenter S, Elmqvist T, et al. (2002) *Resilience and sustainable development—Building adaptive capacity in a world of transformation*. Stockholm: The Environmental Advisory to the Swedish Government.

Folke C, Carpenter S, Walker B, et al. (2005) Regime shifts, resilience, and biodiversity in ecosystem management. Annual review of ecology. *Evolution and Systematics* 35: 557–581.

Goldman SJ, Jackson K, and Bursztynsky TA (1986) *Erosion and sediment control handbook*. New York, NY: McGraw-Hill.

Guertin DP, Miller SN, and Goodrich DC (2000) Emerging tools and technologies in watershed management, USDA Forest Service Proceedings RMRS-P-13. In: *Proceedings of the Conference on Land Stewardship in the 21st Century*, pp. 194–204. AZ: *The Contributions of Watershed Management* Tucson.

Gunderson L (1999) Resilience, flexibility, and adaptive management—Antidote for spurious certitude? *Ecology* 3(1): 7.

Inamdar S (2006) Challenges in modeling hydrologic and water quality processes in riparian zones. *Journal AWRA* 42(1): 5–14.

Johanson RC, Imhoff JC, Kittle JL, and Donigian AS (1984) *Hydrologic Simulation Program-Fortran (HSPF): User's Manual, Release 8, EPA 600/3–84-006*. Washington, DC: US Environmental Protection Agency.

Jørgensen ES and Fath B (2011) *Fundamentals of Ecological Modeling*, 4th edn. Amsterdam: Elsevier p. 530.

Karr J, Fausch KD, Angermeier PR, Yant R, and Schlosser IJ (1986) *Illinois Natural History Survey Special Publication No. 5: Assessment of Biological Integrity in Running Water: A Method and its Rationale*. IL: Illinois Natural History Survey Champaign. 28.
Kendall BE, Briggs CJ, Murdoch WW, et al. (1999) Why do populations cycle? A synthesis of statistical and mechanistic modeling approaches. *Ecology* 80: 1789–1805.
Knisel WG (1980) *CREAMS:* A field scale model for chemicals, runoff, and erosion from agricultural management systems. In: *Conservation Research Report No. 26*. Washington, DC: USDA.
Kohonen T (1990) *Self-Organizing Maps*, 2nd edn. Berlin: Springer.
Law AM and Kelton WD (1982) Simulation Modeling and. Analysis. In: *McGraw-Hill*, p. 400. New York: NY.
Legendre P and Legendre L (1998) *Numerical Ecology*. Amsterdam: Elsevier Science BV.
Lehr JH and Keeley J (2005) *Water encyclopedia*. Hoboken, NJ: Wiley.
Leonard RA, Knisel WG, and Smith DA (1987) GLEAMS: Groundwater loading effects of agricultural management systems. *Transactions ASAE* 30(5): 1402–1418.
Linsley RK and Crawford NH (1960) Computation of a synthetic streamflow record on a digital computer. *International Association of Scientific Hydrology. Publication* 51: 526–538.
Maidment DR (1993) *Handbook of Hydrology*. New York, NY: McGraw-Hill.
L. Matejicek. Development of software tools for ecological field studies using ArcPad ESRI 24th Annual International User Conference San Diego, California 2003 12.
Muradian R (2001) Ecological thresholds: A survey. *Ecological Economics* 38: 7–24.
Negev M (1967) Sediment model on a digital computer. In: *Department of Civil Engineering Technical Report 76*. Palo Alto, CA: Stanford University.
Novotny V (2003) *Water quality: Diffuse pollution and watershed management*. New York, NY: Wiley.
Novotny V and Chesters G (1981) *Handbook of Nonpoint Pollution: Sources and Management*. New York, NY: Van Nostrand-Reinhold.
Novotny V and Olem H (1994) *Water Quality: Prevention, identification and Management of Diffuse Pollution*. New York, NY: Van Nostrand-Reinhold Publisher and Wiley.
Palmer MW (1993) Putting things in even better order: The advantage of canonical correspondence analysis. *Ecology* 74: 2215.
Pijanowski B, Brown D, Shellito B, and Manik G (2002) Using neural networks and GIS to forecast land use changes: A land transformation model. *Computers, Environment and Urban Systems* 26(6): 553–575.
Pijanowski BC, Gage SH, Long DE, and Cooper WE (2000) A land transformation model for the Saginaw Bay Watershed. In: Sanderson J and Harris L (eds.) *Landscape ecology: A top down approach*, pp. 183–198. Boca Raton FL: CRC Press.
Ramaswami A, Milford JB, and Small MJ (2005) *Integrated environmental modeling: Pollutant transport, fate, and risk in the environment*. Hoboken, NJ: Wiley.
Recknagel F (2001) Applications of machine learning to ecological modeling. *Ecological Modelling* 146: 303–310.
Roni P and Beechie T (2012) *Stream and watershed restoration: A guide to restoring riverine processes and habitats*. New York, USA: John Wiley & Sons.
Rousseau AN, Mailhot A, Turcotte R, et al. (2000) GIBSI: An integrated modeling system prototype for river basin management. *Hydrobiologia* 4223: 465–475.
Salomons W and Förster U (1984) *Metals in the hydrosphere*. Springer: Berlin.
Schnoor JL (1996) *Environmental modeling: Fate and transport of pollutants in water, air, and soil*. Wiley: New York.
Singh VP (1995) *Watershed Models*. Boca Raton, FL: CRC Press.
Soil Conservation Service (1986) *Soil conservation service urban hydrology for small watershed—TR55, 1986*. Washington, DC: USDA.
Straškraba M (1995) Models for reservoir, lakes and wetlands. In: Novotny V and Somlyódy L (eds.) *Remediation and Management of Degraded River Basins*. Berlin: Springer.
Ter Braak CJF (1986) Canonical correspondence analysis: A new eigenvector method for multivariate direct gradient analysis. *Ecology* 67: 1167–1179.
Ter Braak CJF (1987) The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio* 69: 69–77.
Ter Braak CJF (1994) Canonical community ordination part I: Basic theory and linear methods. *Écoscience* 1: 127–140.
US Environmental Protection Agency AQUATOX (2005) *(*Release 2.1*):* Modeling Environmental Fate and Ecological Effects in Aquatic Ecosystems, *EPA-823-F-05-007, Office of Water*. Washington, DC: USEA.
Virani H, Manolakos E, and Novotny V (2005) Self organizing feature maps combined with ecological ordination techniques for effective watershed management. In: *Technical Report # 4*. Boston, MA: Center for Urban Environmental Studies, Northeastern University. http://www.coe.neu.edu/environment.
Walker B, Carpenter S, Anderies J, et al. (2002) Resilience management in social-ecological systems: A working hypothesis for a participatory approach. *Conservation Ecology* 6(1): 14.
Wigmosta MS, Vail LW, and Lettenmier DP (1994) A distributed hydrology-vegetation model for complex terrain. *Water Resources Research* 30: 1665–1680.
Williams JR (1975) Sediment routing for agricultural watersheds. *Water Resources Bulletin* 11(5): 965–974.
W.H. Wischmeier, D.D. Smith. Predicting Rainfall-Erosion Losses from Cropland East of Rocky Mountains, USDA Agricultural handbook No. 282.1965. USDA Washington, DC.
Young R, Onstad CA, Bosch DD, and Anderson WP (1987) AGNPS: Agricultural non-point source pollution model: A watershed analysis tool. In: *USDA-Agricultural Research Service, Conservation Research Report 35*. Washington, DC: Department of Agriculture.
Young R, Onstad CA, Bosch DD, and Anderson WP (1989) AGNPS: A nonpoint-source pollution model for evaluating agricultural watersheds. *Journal of Soil and Water Conservation* 44(2): 168–173.
Yuan LL and Norton SB (2003) Assessing the relative severity of stressors at a watershed scale. *Environmental Monitoring and Assessment* 98: 323–349.

Acidification

A Lükewille, Norwegian Institute for Air Research (NILU), Kjeller, Norway
C Alewell, University of Basel, Basel, Switzerland

Introduction

Acidification processes in soils, freshwaters, and oceans are natural processes in geological time frames. However, anthropogenic activities on planet Earth have considerably accelerated acidification by enhancing natural processes as well as by changing dynamics, balances, and pathways.

Acidifying substances can have ecosystem external natural sources such as volcanism, dimethyl sulfide (C_2H_6S) emissions from oceans, or, to a minor extent, sulfide emissions from freshwater wetlands. However, most important are anthropogenic emission sources, mainly fossil fuel combustion processes (e.g., public power plants, industry, and traffic) and agriculture. Emissions of SO_2 and NO_x to the atmosphere increase the natural acidity of rainwater due to the formation of H_2SO_4 and HNO_3, both being strong acids. Furthermore, NH_3 emissions mainly from agricultural activity (volatilization from fertilizers and animal liquid manure) trigger acidification processes in soils. After deposition to ecosystems the conversion of NH_4^+ to either amino acids or to NO_3^- in soils is connected to the production of acidifying H^+ ions.

Since the end of the nineteenth century, industrialized regions of the world have been confronted with the consequences of acidic atmospheric deposition, `acid rain'. There was, and still is, substantial concern about the environmental impacts of air pollution at the local, regional, and global scale. `Acid rain' has threatened vegetation, wildlife, soil biology, and human health, caused damage to materials, and changed the chemistry of soils and waters.

Anthropogenic land-use changes and use of fossil fuels have further led to dramatically increasing atmospheric CO_2 concentrations worldwide. CO_2 is absorbed by oceans and reacts with seawater to form H_2CO_3. Acidification of oceans has adverse effects on marine organisms using $CaCO_3$ in seawater to construct their shells and skeletons (e.g., corals and calcareous phytoplankton).

Acidifying compounds can be carried by winds over long distances and affect ecosystems in pristine areas located hundreds or thousands of kilometers away from pollutant sources. Terrestrial and freshwater ecosystems affected by acidification are usually located in regions where precipitation inputs exceed evapotranspiration, that is, where water percolates through the soil and bedrock.

Acidification is the result of a sensitive (un-)balance between ecosystem internal and external H^+ sources and internal H^+ sinks of different capacities and reaction rates. Acidification processes in terrestrial and aquatic ecosystems can have natural and/or anthropogenic causes; natural internal or external activities can drive these processes in the ecosystem. These processes and their consequences are discussed in this article.

Acidification Processes in Soil and Bedrock

An acid is a compound which has the capability to release H^+ ions (**Box 1**). High concentrations of H^+ (low pH values) can attack natural materials such as limestone, soil minerals, and living tissues or man-made materials or artwork such as cement, concrete, metal surfaces, or sculptures.

Acidification of soils is a natural process on geological timescales. In general, soil acidification can be described as a two-step process:

1. The slow gradual depletion of nutrient cations, that is, the leaching of Ca^{2+}, Mg^{2+}, K^+, bases (HCO_3^-, CO_3^{2-}, etc.).
2. Their replacement by `acidic' $H^+$, Al, Fe, and Mn ions or complexes. While $H^+$ is mainly supplied by internal ecosystem processes or by atmospheric deposition, the `acidic' metal cations are released from the bedrock by mineral weathering.

Intensive agriculture and forestry can lead to high ecosystem internal H^+ production (see below). Many man-made landscapes originate from extensive land-use activities. One example is the heath lands in northwestern Europe, where human pasture, field, and forest

Box 1 H^+ ion concentration or pH The pH scale is logarithmic, neutral water has a pH of 7.0. In the absence of strong acid anions such as SO_4^{2-} and NO_3^-, pure rainwater has a pH of 5.6–5.7. This means that `clean' rain in equilibrium with atmospheric carbon dioxide (CO_2) is acid.

Each whole unit on the pH scale represents a multiplication factor of 10. Thus, water with a pH of 5.0 is 100 times more acidic than water with a pH of 7.0.

 doi:10.1016/B978-0-444-63768-0.00251-1

management over centuries have led to soil acidification and erosion. Plant material was removed by grazing. Sparsely growing trees were used in salt refineries and as firewood. The humus layer including the ground vegetation was removed and interspersed in stables. The mixture of soil and dung was used for manuring fields at locations different from the areas where organic material had been removed.

Application of dung, liquid manure, or compost can compensate (part of) nutrient losses. Agriculture and partly also forestry often apply multinutrient mineral fertilizers containing lime (H^+ buffering $CaCO_3/MgCO_3$).

Most crystalline shields and noncarbonated sedimentary rocks can be considered as being sensitive to acidification by `acid rain'. Areas where acidification has been an issue are major parts of northern Europe, northeastern USA, eastern Canada, and parts of China. Due to rapid increases in acidifying emissions potential future problem areas could be Nigeria, India, Venezuela, Southern Brazil, and Southeast Asia.

Hydrogen Ion (H^+) Sources to Soils

Several biologically mediated processes lead to ecosystem internal H^+ production ($=H^+$ sources), while atmospheric deposition or mineral fertilizer applications are external H^+ sources.

Nutrient cation uptake and the consequences of biomass export

The majority of nutrients needed for plant growth exist as cations (Ca^{2+}, Mg^{2+}, K^+, Na^+, NH_4^+, Fe^+, etc.). Fewer nutrients and their less amounts are taken up as anions (NO_3^-, HPO_3^-, SO_4^{2-}, etc.). The latter implies that vegetation assimilates an excess of non-N nutrient cations over anions. To compensate for electroneutrality, plants release either weak organic acids or H^+ to the soil solution for each positively charged ion taken up by the roots (e.g., one H^+ in the case of K^+ uptake and two H^+ for each Mg^{2+} ion). As a result the pH of the soil solution near the root surface (the rhizosphere) can drop considerably during the growing season. However, if no plant material is removed from the system, nutrient cations return to the soils during decomposition, which is an H^+-consuming process. Thus, without biomass export, plant uptake has no long-term effect on acidification.

In a growing forest the consequence of nutrient cation uptake poses a net production of H^+ in the soil solution over decades, because nutrients can be stored in the biomass and humus layer for a relatively long time periods (Fig. 1). However, organic matter in natural ecosystems is usually exchanged in cycles, that is, when a forest or part of a forest dies, assimilated nutrients are released again via decomposition.

Thus, excess uptake of positively charged nutrients by plants affects soil acidity in the long term only if plant material is removed (Fig. 2). This removal can be driven by harvesting grain crops in agriculture, by removing cattle, which have converted part of the plant material they have eaten into body tissue, or by using timber and firewood in forestry.

Decomposition, root respiration, and the production of carbonic acid

Microbial degradation (decomposition) of organic material and root respiration lead finally to relatively high CO_2 concentrations in the soil air (high CO_2 partial pressure). A greater part of this CO_2 resolves in soil water and forms carbonic acid (Box 2). A consequence is that waters percolating through soils (or bedrocks) contain usually substantially higher concentrations of H_2CO_3 than rainwater or surface waters. The major anion produced by H_2CO_3 dissociation is HCO_3^- (Box 2).

Under natural conditions, the deprotonation of H_2CO_3 is the most significant H^+ source in acidifying soils down to pH $\geq$ 5 (note that the pK_a of H_2CO_3 is 6.46 preventing a decrease of pH below 5). Thus, CO_2 is the major agent of CO_3^{2-}, mineral

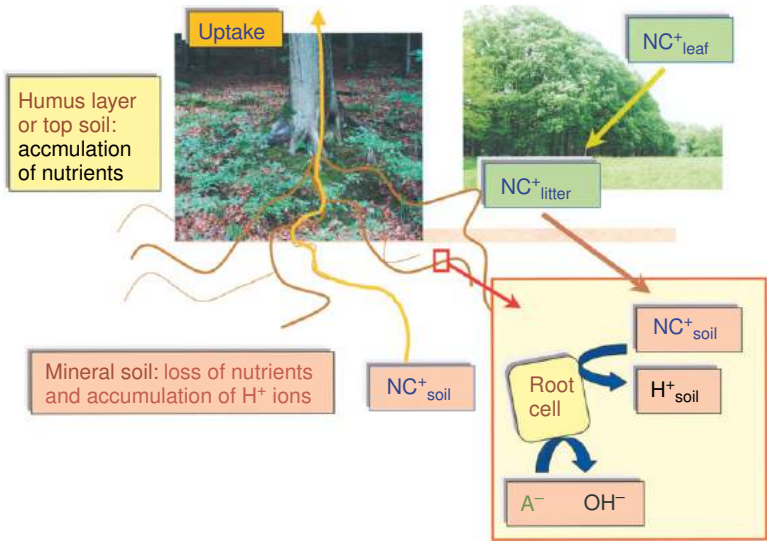


Fig. 1 Nutrient uptake. NC^+, nutrient cations; A^-, anion.

Box 2 Dissociation of carbonic acid, formation of hydrocarbonic acid and H^+

$$H_2CO_3 \leftrightarrow H^+ + HCO_3^- \quad [1]$$

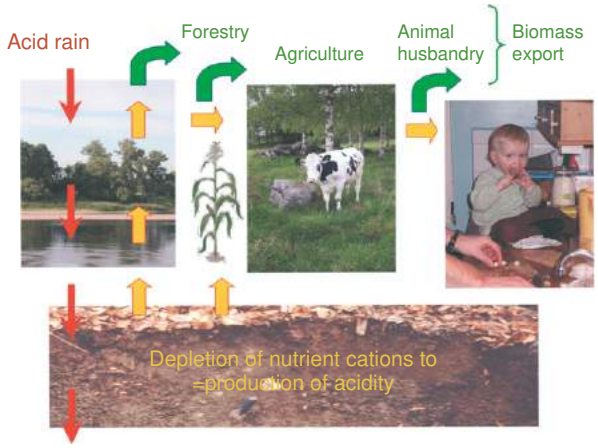


Fig. 2 Biomass export.

Table 1 pK_a values of some important inorganic and organic acids

Acid	*Formula*	pK_{a1}	pK_{a2}	pK_{a3}
Sulfuric acid	H_2SO_4	−3	1.92	
Nitric acid	HNO_3	−1.32		
Oxalic acid	$(COOH)_2$	1.23	4.19	
Phosphoric acid	H_3PO_4	2.12	7.21	12.67
Formic acid	HCOOH	3.75		
Acetic acid	CH_3COOH	4.75		
Carbonic acid	H_2CO_3	6.46	10.25	
Humic and fulvic acids	Complex organic molecules in soil solution and freshwaters; pK_a between 3 and 8			

weathering, and natural acidification (see the section entitled `Hydrogen ion (H^+) sinks'). Below pH 5, production organic acids drive natural acidification.

Soil organic matter and the production of soil organic acids

Soil organic matter consists of carbohydrates, which contain acidic groups (e.g., carboxyl, carbonyl, or hydroxyl). An increase in soil organic matter is in itself a potential source of acidity, as also the application of dung or liquid manure. However, organic matter contains only weak acids, that is, in contrast to strong acids such as H_2SO_4 they do not dissociate completely but release only a portion of their H^+. This proportion varies according to the H^+ concentration in the solution. The lower the pH, the fewer the H^+ ions released (and the more the acid groups protonated).

The deprotonation of dissolved organic acids can be described by dissociation constants (pK_a values). The lower the pK_a, the stronger is the acid (**Table 1**). Dissolved organic acids are ubiquitous in soils and can deprotonize depending on the pH values.

Transport of organic anions causes soil acidification in deeper soil horizons (a process called podsolization) and waters. In general, organic anions can be rapidly degraded to CO_2 by microbial activity and they are important components of groundwater or stream water acidification only in fens or bogs.

H^+ turnover within the nitrogen cycle

The N cycle (**Fig. 3**) is connected to major H^+ turnover processes in soils. Nitrogen is one of the major nutrients, and N turnover exceeds the turnover of all other nutrients and trace elements quantitatively, with the exception of carbon. Because inorganic N can occur as a cation or an anion in soils, the influence on H^+ budgets caused by N turnover is complex (**Box 3**).

Decomposition of N-containing organic material is usually followed by the oxidation of NH_4^+ (nitrification), which is connected to the production of 2 mol of H^+ for each NH_4^+ molecule (**Box 3**).

Nitrogen is an important nutrient as it is part of proteins and nuclides in living organisms. The assimilation of NH_4^+ during the production of amino acids produces 1 mol of H^+ per mole of NH_4.

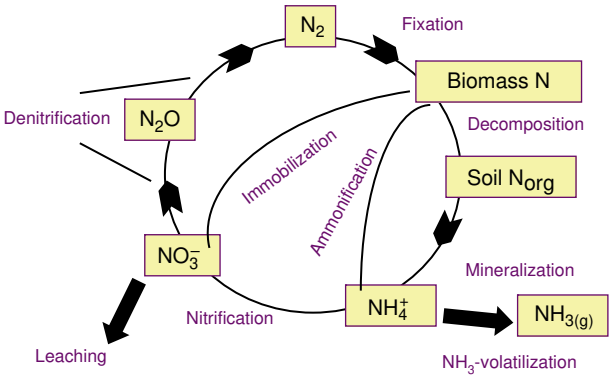


Fig. 3 Nitrogen cycle.

Box 3 Proton sources and sinks within the nitrogen cycle (blue, H^+ sink; red, H^+ source) N_2-fixation:

$$N_2 \rightarrow N_{org} \text{No ions in solution, thus no } H^+ \text{turnover} \quad [1]$$

Ammonification:

$$R-C-NH_2 + H_2O + H^+ \rightarrow NH_4^+ + R-C-OH \quad [2]$$

Denitrification:

$$2NO_3 + 3CH_2O + 1/2O_2 + 2H^+ \rightarrow N_2 + 3CO_2 + 4H_2O \quad [3]$$

NO_3^- uptake:

$$NO_3^- + H^+ \rightarrow N_{org} \quad [4]$$

NH_4^+-uptake and assimilation:

$$NH_4^+ + R-C-OH \rightarrow R-C-NH_2 + H_2O + H^+ \quad [5]$$

Nitrification:

$$NH_4^+ + 2O_2 \rightarrow H_2O + NO_3^- + 2H^+ \quad [6]$$

NH_3-volatilization:

$$NH_4^+ + OH^{-a} \rightarrow NH_3 + H_2O \quad [7]$$

[a]Note that consumption of OH^- is equivalent to production of H^+ and vice versa.

Adding N as NH_4^+ fertilizer will cause acidity ($(NH_4)_2SO_4$ or NH_4NO_3). If the NH_4^+ added is converted to NO_3^- and leached out of the soil, then a very rapid rate of acidification occurs. If plants take up the NH_4^+, then an intermediate rate of acidification occurs. Fertilization with NH_3 can actually be a neutral process (**Box 4**). However, fertilizing NH_3 is bound to extreme rates of volatilization and will increase local and regional N deposition dramatically.

Another way for N into the ecosystem is the fixation of N_2 from the atmosphere by bacteria. While there are only few free living species, most N-fixing bacteria live in a symbiosis with plants, for example, within a legume nodule where the very stabile molecular N_2 is converted into a form available for plants to use. N_2 fixation involves no H^+ transfer. Only after ecosystem internal mineralization of organic N to NH_4^+ or NO_3^-, will fixed N_2 become involved in H^+ transfer. If N accumulates in the ecosystem, it usually does so in soil organic matter. Besides leaching and harvest as an N loss to ecosystems, N can volatilize into the atmosphere as gaseous N compound via denitrification (bound to consumption of H^+) or volatilization (production of H^+).

To conclude, a disruption of the N cycle by either biomass export (harvest) or fertilization has major consequences concerning the H^+ balance and thus the acidification of soils.

Box 4 Acid production due to nitrogen fertilization (negative for H^+ consumption, positive for H^+ production)

Application of ammonium (e,g., as$(NH_4)_2SO_4$)
NH_4^+ uptake and assimilation: $+1$ mol H^+ per NH_4^+
Nitrification: $+2$ H^+ mol H^+ per NH_4^+
If produced NO_3 is taken up by plant: -1 mol H^+ per NO_3^-
Application of NH_4NO_3
NH_4^+ uptake and assimilation: $+1$ mol H^+ per NH_4^+
Nitrification: $+2$ H^+ mol H^+ per NH_4^+
If produced or applied NO_3 is taken up by plant: -1 mol H^+ per NO_3^-
Application of Ammonia (NH_3)
Dissolution (re-volatilization): -1 mol H^+ per NH_3
Nitrification of produced NH_4^+: $+2$ H^+ mol H^+ per NH_4^+
If produced NO_3^- is taken up by plant: -1 mol H^+ per NO_3^-

Oxidation of reduced compounds

Increasing water saturation promotes anoxic conditions in soils, and microorganisms can use NO_3^-, SO_4^{2-}, Fe, Mn, and CO_2 as electron acceptors instead of O_2. During such reduction processes, H^+ is consumed and alkalinity is generated. Up to 70% of the impacted acidity can be neutralized in forested freshwater wetlands by Fe and SO_4^{2-} reduction alone.

Conversely, during the oxidation process of (prior) reduced compounds, H^+ is released. For example, if soils or waters contain a substantial amount of reduced Fe and have a low buffering capacity, the pH of the solution may fall from a value of about 6–7 to 2–3 caused by the oxidation of formerly reduced Fe compounds. Concerning SO_4^{2-} reduction, FeS and FeS_2 are the most important products. If conditions stay anoxic over a longer time period, reduced S species might be incorporated into the organic substance and thus stored long term, resulting in an equally long-term alkalinity generation.

Drainage of valley floors and thus exposure to air (O_3) causes reduced compounds to re-oxidize and release substantial amounts of acid. In contrast, wetland soils and riparian zones may act as long-term sinks for deposited H^+, SO_4^{2-}, and NO_3^-, depending on soil characteristics, climatic parameters, and the composition of the soil microbiota.

Atmospheric deposition of acidifying compounds

Acidifying pollutants are deposited into ecosystems as follows:

1. directly as gases and aerosols to vegetation or other surfaces (dry deposition, especially NH_4^+);
2. as rain or snow (wet deposition); and
3. via impaction and sedimentation of fog or droplets to various surfaces (occult deposition).

High acidification rates occur in forested coniferous sites (compared to deciduous sites) due to more efficient scavenging of acidifying pollutants from the atmosphere especially during wintertime. The acidification rate caused by acid deposition is in the range 0.8–7 kmol ha^{-1} yr^{-1} due to the combined effects of HNO_3, H_2SO_4, HCl, and NH_4 deposition.

Hydrogen Ion (H^+) Sinks

In natural ecosystems weathering of minerals counteracts acidification, that is, acts as H^+ sink. Thus, main sinks in ecosystems are geochemical buffering reactions in soils and bedrocks, and only a minor fraction is buffered in waters. The so-called acid neutralization capacity in soils and bedrock (ANC_{solid}) can be defined as the sum of all unprotonated buffering substances. Thus, acidification is always accompanied by a decrease in ANC_{solid} over time. It is important to note that this decrease in ANC_{solid} is irreversible (against the background of our human calculation of times).

Besides the capacity, that is, the total buffering pool of a soil or bedrock, the geochemical reaction rate of the buffering substances are a crucial factor determining how much of the acidifying compounds are neutralized over a certain period. This rate can, for example, be estimated as kilomole charge per hectare per year ($kmol_c$ ha^{-1} yr^{-1}).

Ecologically effective are, last but not the least, the concentrations of certain ions in the soil/bedrock solution or freshwaters. Such intensity parameters can be measured as concentrations (e.g., pH, Mg^{2+}, or Al^{3+}; in mol_c per liter) or as base cation saturation of exchanger complexes in soils (in %).

In theory, many of the geochemical buffering processes are equilibrium reactions. However, the loss of ions with percolating water leads to permanent disequilibria. The latter has important implications for the expected reversibility of soil acidification under decreasing deposition regime. In North America and Europe, soil acidification is irreversible as long as the supply of weathering products from bedrock is smaller than the loss of weathering products due to the combined effects of natural and anthropogenic acidification processes. It is highly unlikely that the latter two will become smaller than the supply by weathering products because natural acidification processes already exceed buffering by weathering processes in most systems.

Box 5 Dissolution of calcium carbonate

$$CaCO_3 + H^+ \Rightarrow HCO_3^- + Ca_2^+ \quad pK_{a1} = 6.46$$
$$HCO_3^- + H^+ \Rightarrow CO_2 + H_2O \; pK_{a2} = 10.25$$

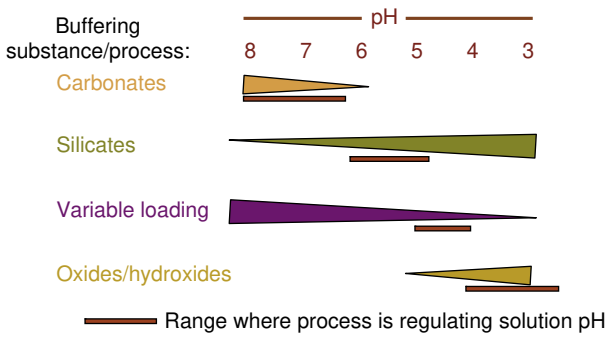


Fig. 4 Buffering systems.

Carbonate dissolution

The H^+ ions are buffered via the dissolution of Ca(or Mg)CO_3 in soils and bedrock (**Box 5**) as long as soils or bedrocks contain accessible carbonate.

The pH values in the soil solution are quasiconstant and stay above pH 6.2 (**Fig. 4**). The buffering rate is high, for example, 2 $kmol_c\ ha^{-1}\ yr^{-1}$ at a water percolation rate of 200 $l\ m^{-2}$ and a CO_2 partial pressure of 0.3 kPa CO_2. The CO_3^{2-} buffering is usually an irreversible one-way reaction resulting in the loss of Ca^{2+} and HCO_3^- from soils and bedrock.

Silicate weathering

The H^+ ions are buffered by the (relatively slow) weathering of primary silicate minerals (e.g., **Box 6**). The soil solution stays in the pH range of 6.2–5.0 (**Fig. 4**), and the rate lies between 0.2 and 2 $kmol_c\ ha^{-1}\ yr^{-1}$. Compared to the accumulated H^+ production or input rates in ecosystems affected by acid deposition (see the section titled `Hydrogen ion (H^+) sources in soils') this rate is rather low.

Silicate weathering results usually in the irreversible destruction of clay minerals, the release of exchangeable cations, and Al ions to the soil solution. Dominant anions in the solution are HCO_3^- and organic anions.

Further, some clay mineral crystals can dissolve completely, leading to high Al concentrations in the solution (**Box 7**).

Exchanger with variable loading

The H^+ ions exchange against base cations bound to clay minerals and oxides or organic matter (pH range 5–4.2; **Fig. 4**; **Box 8**).

Exchangeable cations are lost with the percolating water. The buffering capacity depends on the absolute cation exchange capacity and on the percentage of saturation of the exchanger complex with base cations. The buffering rate is very high (fast reaction). The Ca^{2+} ion is usually the dominant cation. In systems influenced by `acid rain', HCO_3^- is replaced by the anions SO_4^{2+} and NO_3^-. In naturally acidic ecosystems, for example, in bogs, organic anions are dominant.

Amorphous hydroxides or oxides of aluminum and iron

The so-called Al and Fe buffer ranges can be described by the equilibrium reactions shown in **Box 9**.

The H^+ ions are bound in water (H_2O), and soluble Al (or Fe) ions (ion complexes) emerge in the soil solution. The neutralization capacity depends on the reactive amount of Al or Fe (hydr)oxides. Buffering rates are high, and Al ions (or Fe ions) become the predominant cations in the soil solution. In ecosystems affected by acid deposition, SO_4^{2-} and NO_3^- are the predominant anions.

Acidification of Groundwater, Freshwaters, and Oceans

Acidification and Buffering in Ground- and Freshwaters

In the absence of strong acid anions such as SO_4^{2-} and NO_3^-, pure rainwater has a pH of 5.6–5.7. The latter is caused by the equilibrium with atmospheric CO_2 (see the section titled `Introduction').

Box 6 Weathering of primary silicate minerals

$$-\mathrm{SiOM} + \mathrm{H^+} \Rightarrow -(\mathrm{SiOH}) + \mathrm{M^+}\,(\mathrm{M^+} = \text{exchange ablemetal cation})$$

Box 7 Complete dissolution of clay mineral crystals

$$\underset{\text{(feldspar)}}{\mathrm{KAl_3Si_3O_{10}(OH)_2}} + 10\mathrm{H^+} \Rightarrow \mathrm{K^+} + 3\underset{\text{(silica)}}{\mathrm{Al^{3+}}} + 3\mathrm{H_4SiO_4}$$

Box 8 Exchange of H^+ ions against base cations (clay minerals, oxides, organic matter)

$$\mathrm{M^+(soil)} + \mathrm{H^+} \leftrightarrow \mathrm{H^+(soil)} + \mathrm{M^+}$$
$$(\mathrm{M^+} = \text{exchangeable metal cation})$$

or

$$\mathrm{R-COO-M^+} + \mathrm{H^+} \leftrightarrow \mathrm{R-COOH} + \mathrm{M^+}$$
$$(\mathrm{p}K_\mathrm{a}\text{dependent; p}K_\mathrm{a}3-8)$$

Box 9 Aluminum hydroxides, iron oxides and hydroxides

$$\mathrm{Al(OH)_3} + 3\mathrm{H^+} \leftrightarrow \mathrm{Al^{3+}} + 3\mathrm{H_2O} \quad \text{pH range : } 4.2-3.2$$
$$\mathrm{FeOOH} + 3\mathrm{H^+} \leftrightarrow \mathrm{Fe^{3+}} + 2\mathrm{H_2O} \quad \text{pH range :} < 3.2$$

Box 10 Acid neutralization capacity in freshwaters – ANC_{aqua}

$$\mathrm{ANC_{aqua}} = 2\mathrm{CO_3^{2-}} + \mathrm{HCO_3^-} + \mathrm{OH^-} - \mathrm{H^+} - \mathrm{Al^{3+}} - \mathrm{Al(OH_x^{y+})}$$

Constraint of electroneutrality:

$$\mathrm{HCO_3^-} + \mathrm{OH^-} + \mathrm{Cl^-} + \mathrm{NO_3^-} + 2\mathrm{SO_4^{2-}} = \mathrm{H^+} + \mathrm{Al^{3+}} + \mathrm{Al(OH_x^{y+})} + 2\mathrm{Ca^{2+}} + 2\mathrm{Mg^{2+}} + \mathrm{K^+} + \mathrm{Na^+}$$

or:

$$\mathrm{HCO_3^-} + \mathrm{OH^-} - \mathrm{H^-} - \mathrm{Al^{3+}} - \mathrm{Al(OH_x^{y+})} = 2\mathrm{Ca^{2+}} + 2\mathrm{Mg^{2+}} + \mathrm{K^+} + \mathrm{Na^+} - \mathrm{Cl^-} - \mathrm{NO_3^-} - 2\mathrm{SO_4^{2-}}$$

Reuss' and Johnson's definition of ANC_{aqua} (easy to measure):

$$\mathrm{ANC_{aqua}} = 2\mathrm{Ca^{2+}} + 2\mathrm{Mg^{2+}} + \mathrm{K^+} + \mathrm{Na^+} - \mathrm{Cl^-} - \mathrm{NO_3^-} - 2\mathrm{SO_4^{2-}}$$

A major part of rainwater reaching the groundwater and/or surface waters percolates through soils and bedrock. A small portion of (acidified) rainwater is directly deposited into lakes and streams. Areas highly affected by water acidification are small watersheds with shallow soil cover, rapid flushing rates, and slowly weathering bedrock, such as granite and quartzite. These types of soil and bedrock do not contain unstable or readily soluble minerals such as $CaCO_3$ and $MgCO_3$, which are very effective in neutralizing the acids (see the section titled `Carbonate dissolution').

The natural buffering system in surface waters is provided by HCO_3^-, released by the weathering of soil/bedrock minerals, and by the balance between dissolved atmospheric CO_2 and CO_2 from respiration/decomposition processes (**Box 2**). As for soils, freshwater and groundwater acidification can be defined as a decrease in acid neutralization capacity (ANC_{aqua}). Because of the electroneutrality constraint in solutions, ANC_{aqua} can also be defined as the sum of all `base' cations minus the sum of all `strong' acid anions (**Box 10**).

In acidified surface waters, the pH and/or ANC_{aqua} have fallen significantly below pre-industrial levels ($ANC_{aqua} < 0.10$ $mmol_c$ of HCO_3^- per liter). Elevated levels of Al compounds and low pH values usually accompany low ANC_{aqua}. One visible sign of acidification is that the water becomes clearer because humus substances that normally color the water precipitate out together with Al (or Fe) compounds. Biomass production (algae and bacteria) and decomposition slow down, and organic matter such as leaves often collect on the lake or riverbeds.

Acidification of Oceans

Of all CO_2 emitted globally due to land-use changes, fossil fuel burning, and cement production in the past 200 years, only about half has remained in the atmosphere. Besides terrestrial plants, oceans have taken up considerable parts of it. A concern of many scientists is that rising levels of atmospheric CO_2 are causing an increasing acidification of oceans. The latter is due to the equilibrium between atmospheric CO_2 and the CO_2 dissolved in seawater. Dissolved CO_2 forms H_2CO_3 (**Box 2**), leading to increasing acidity. H^+ reacts with the CO_3^{2-} ion to form HCO_3^-, that is, there is also a carbonate buffer in seawater.

CO_2 is absorbed at the sea surface, which is thus most affected. Marine organisms that produce $CaCO_3$ shells live above the so-called `saturation horizon', where $CaCO_3$ does not readily dissolve. Increasing CO_2 concentrations decrease the saturation state of $CaCO_3$ making structures of $CaCO_3$ vulnerable to dissolution.

Organisms containing shells or plates of $CaCO_3$ fall to the sea floor when they die. $CaCO_3$ is thus abundant in sediments and interacts with seawater. Slow mixing throughout the oceans, mixing necessary to bring up compounds from the oceans' sediments to buffer the increased ocean surface chemistry, causes a delay in $CaCO_3$ dissolution. Warming of oceans as a consequence of global warming may even further reduce this mixing rate with deeper waters.

Consequences of Acidification

The effects of natural acidification processes and the acceleration by human land use and `acid rain' are not the main focus of this article. However, the most serious consequences are summarized in the following sections.

With decreasing acid deposition since the 1980s in Europe and North America, the impacts of acidification and eutrophication are foreseen to show a decrease, as a consequence of biodiversity showing some recovery. However, a full return to pre-pollution conditions is not to be expected, because of changes in competition patterns and distribution of species. The introduction – whether voluntary or accidental – of species alien to European ecosystems or to other regions of Europe represents an increasing risk, favored by globalization of trade, exchange, and transports. Thus, even if chemical parameters (e.g., nutrient status, acidity) return to pre-pollution condition, indigenous species might not be successful in competing with well-established alien species.

Consequences of Soil Acidification

The monitoring of soil acidification induced by `acid rain' is difficult because of the long time frames involved and the parallel development of natural acidification processes. Both natural and anthropogenic-driven soil acidification can result in extremely low base saturation, low pH, and low ANC_{aqua}. Anthropogenic soil acidification due to `acid rain' is characterized by high soil solution concentrations of SO_4^{2-} and NO_3^-. In contrast, the anions responsible for H^+ and aluminum leaching during natural acidification processes are organic anions and HCO_3^-. High N and/or SO_4^{2-} deposition have resulted in N and SO_4^{2-} accumulation in forest soils, decreases in forest soil pH, and leaching of base cations.

Despite the lack of evidence for direct effects, there is no doubt that 'acid rain' has complex negative effects on forest ecosystems in central Europe and northeastern America. More recently, serious effects have also been recognized in China, particularly in the industrialized regions such as the Sichuan Province now exploiting nearby extensive coal deposits of high S content.

In many areas, leaching of base cations from soils has led to nutrient deficiency, especially of Mg^{2+} and K^+. Today, many forest soils of central Europe have low base saturation and low pools of exchangeable nutrient cations. The Mg-deficient nutrition of trees (yellowing of needles) is a widespread phenomenon and has been related to decreased concentrations of Mg^{2+} in the soil solution.

Besides loss of nutrient cations, the buffering of H^+ from deposition causes increased levels of Al in soil solutions, which might have detrimental effects on tree root growth and nutrient uptake. In water culture experiments, the ratio of Ca/Al and Mg/Al, rather than the Al concentration itself, largely determines the deleterious effect on roots.

Consequences of Freshwater Acidification

Acidification is largely a problem in naturally nutrient-poor lakes and streams. Some organisms are directly sensitive to low pH values, for example, shell-bearing organisms such as mollusks, mussels, and many crustaceans, including crayfish (dissolution of $CaCO_3$). Acidification and the elevated levels of dissolved Al ions in water can have direct harmful effects on the eggs and fry of many fish species (e.g., salmons, trout, and perches). Further, the number of phytoplankton species falls dramatically in acidified waters.

Al compounds can precipitate in the gills of adult fish and lead to suffocation (mainly `mechanical' effect). The disappearance of fish species or other animals can have an influence on the food chains in a river or lake, and can thus have dramatic effects on the ecosystem structure as a whole. Fish-eating birds, such as divers, merganser, and osprey, are put under pressure, while insect-eaters, such as goldeneye, are favored.

Another effect can be shifts in plant communities. Certain mosses are favored by acidification. In nutrient-poor lakes, plant species such as shore weed or water lobelia, living in shallow waters close to lake shores, are overgrown by bog mosses.

Last but not the least, `acid rain' can also have negative effects on human health, either directly (e.g., SO_2) or indirectly through ground-level ozone formation followed by NO_x emissions. Further, the acidification of groundwater and drinking water supplies can be a direct human health hazard through high NO_3^- concentrations or indirectly through metals (Al, Cd) released from the soils and/or water pipes.

Consequences of Eutrophication in Ecosystems

Atmospheric N deposition very often leads to an excess of N in terrestrial and aquatic ecosystems, leading in general to increased growth. However, imbalances can be the consequence: tree crowns can grow faster than the root systems (increased risk of desiccation). Furthermore, depletion of other nutrients such as Mg^{2+} or phosphorous in acidified soils can counteract plant growth enhanced by excess N supply or can lead to nutrient imbalances and plant instability in spite of enhanced growth. Excess N input in forest ecosystems can change the occurrence of mycorrhizal fungi living in symbiosis with trees and support nutrient uptake by plant roots.

In some relatively nutrient-rich freshwaters atmospheric N deposition can contribute to eutrophication. However, usually phosphorus is the nutrient that limits growth in freshwater ecosystems.

Consequences of Ocean Acidification

Ocean acidification could have profound effects on ocean ecosystems' structure, food chains, population dynamics, and nutrient cycles. It is so far unknown how, for example, tropical and subtropical coral reefs and fisheries will respond to this man-made acidification. Corals, calcareous phytoplankton, mollusks, and other marine organisms use $CaCO_3$ in seawater to construct their shells and skeletons. Some shallow-water animals, which play a vital role in releasing nutrients from sediments, also calcify. In a more acidic environment, it becomes difficult to secrete $CaCO_3$, leading to slower growth rates and more fragile skeletal structures. How this will affect ecosystem community structure and the marine food web is unclear at present.

See also: Aquatic Ecology: Acidification in Aquatic Systems. General Ecology: Ecological Effects of Acidic Deposition. Global Change Ecology: Sulfur Cycle

Further Reading

Alewell, C., 2002. Acid input into the soils by acid rain. In: Rengel, Z. (Ed.), Handbook of Soil Acidity. New York: Dekker, pp. 83–115.
Alewell, C., Armbruster, M., Bittersohl, J., *et al.*, 2001. Are there signs of acidification reversal after two decades of reduced acid input in the low mountain ranges of Germany? Hydrology and Earth System Sciences 5, 367–378.
Emmett, B.A., Hudson, J.A., Coward, P.A., Reynolds, B., 1994. The impact of riparian wetland on streamwater quality in a recently afforested upland catchment. Journal of Hydrology 162, 337–353.
Gilliam, J.W., 1994. Riparian wetlands and water quality. Journal of Environmental Quality 23, 896–900.
Haugan, P.M., Turley, C., Poertner, H.O., 2006. Effects on the marine environment of ocean acidification resulting from elevated levels of CO_2 in the atmosphere. Trondheim: DN-utredning, 2006-1.
Krug, E.C., Frink, C.R., 1983. Acid rain and acid soil: A new perspective. Science 221, 520–525.
Lükewille, A., Bredemeier, M., Ulrich, B., 1993. Input–output relations of major ions in European forest ecosystems. Agriculture, Ecosystems, and Environment 47, 175–184.
Nilsson, I.S., Miller, H.G., Miller, J.D., 1982. Forest growth as a possible cause of soil and water acidification: An examination of the concept. Oikos 39, 40–49.
Reuss, J.O., Johnson, D.W., 1986. Ecological Studies 59: Acid Deposition and the Acidification of Soils and Waters. Berlin: Springer.
Steinberg, C.E.W., Wright, R.F. (Eds.), 1994. Acidification of Freshwater Ecosystems, Implications for the Future. Chichester: Wiley.
Stoddard, J.L., Jeffries, D.S., Lükewille, A., *et al.*, 1999. Regional trends in aquatic recovery from acidification in North America and Europe. Nature 401, 575–578.
Ulrich, B., 1986. Stability, elasticity and resilience of terrestrial ecosystems under the aspect of matter balance. In: Schulze, E.-D., Zwölfer, H. (Eds.), Ecological Studies 61: Potentions and Limitations of Ecosystem Analysis. Berlin: Springer, pp. 11–47.
Van Breemen, N., Driscoll, C.T., Mulder, J., 1984. Acidic deposition and internal proton sources in acidification of soils and waters. Nature 307, 599–604.
Van Breemen, N., Mulder, J., Driscoll, C.T., 1983. Acidification and alkalinization of soils. Plant and Soil 75, 283–308.

Allometric Theory: Extrapolations From Individuals to Ecosystems

George B Arhonditsis, Yuko Shimoda, and Noreen E Kelly, University of Toronto Scarborough, Toronto, ON, Canada

Abbreviations

AIC	Akaike information criterion
c	Abundance of consumers/predators
E	Activation energy, eV (electron volts, 1 eV = 23.06 kcal mol^{-1}, = 96.49 kJ mol^{-1})
ES	Ecological stoichiometry
F	Consumer per capita feeding rate
hr	Hour
h	Handling time that predators require to digest resources, s (expressed in seconds)
I	Individual metabolic rate
ind	Individual organisms
k	Boltzmann constant, 8.62×10^{-5} eV/K (see also activation energy)
KH_p	Parameter that represent half saturation constant for phosphorus uptake by primary producer, μmol P L^{-1}
M, m	Mass of organisms, g or kg
MTE	Metabolic theory of ecology
N	Resource abundance in individuals
NH	Parameter represent that half saturation constant for nitrate uptake by primary producer, μmol N L^{-1}
PPMR	Predator-prey mass ratio
QP_{max}	Maximum internal phosphate cell quota, fmol P $cell^{-1}$
QP_{min}	Minimum internal phosphate cell quota, fmol P $cell^{-1}$
R	Rate of metabolism, such as rate of respiration, excretion
r	Abundance of resource/prey
SA	Surface area of microorganisms cell, μm^2
$SA{:}V$	Surface area to cell volume ratio of individual organism's cell, μm^{-1}
T	Temperature, K
V	Body size as volume, cell volume μm^3
VP_{max}	Maximum phosphorus uptake rate by primary producer, μg P μm^{-3} h^{-1}
W	Organism body size, such as body mass, weight, volume, body length or height and surface area (SA)
wR	Mass specific rate of metabolism
Y	Biological attributes, such as growth rate, metabolic rate, physiological rate
μ_{max}	Parameter that represent maximum growth rate of organisms, day^{-1}

Glossary

Ecosystem model A mathematical representation of the interactions among biological, chemical and physical components in an ecological system.

Epilimnion The upper most layer of lakes when the difference in density due to the warmer surface water create stratification in the water column.

Eukaryotes Eukaryotes whose cells have a nucleus and organelles enclosed within membrane, whereas prokaryotes are unicellular organisms with no membrane-bound organelles.

Functional groups A specific group of organisms (or chemical compounds) that share similar attributes.

Half saturation constant for nutrient uptake The nutrient concentration at which uptake rate of a nutrient is half of the maximum potential rate.

Harmful algal blooms Excessive growth of phytoplankton bloom in aquatic environments dominated by species that cause toxic or harmful effects on people, fish, shellfish, marine mammals and birds.

Homeotherm An organism that maintains their thermal homeostasis. Some rely on internal metabolic processes as a heat source (endotherms), while others maintain their body temperature by behavioral thermoregulation.

Kleiber's rule A rule defining that organism's metabolic rates scale to the ¾ power of the animal's mass.

Lotka-Volterra models A model that describes predator-prey relationships with a pair of differential equations as function of growth rate of prey, mortality of prey and predator, predator's search/handling efficiency of prey and time.

Meroplankton Organisms that spend only part of their life (early stage, larval) drifting in pelagic environment and spend their adult life in the benthic community.

Parameterization The process of deciding and defining the parameters necessary for a defensible specification of a model.

Poikilotherm An organism whose internal temperature varies considerably as a consequence of variation in the ambient environmental temperature. Many ectotherms, organisms in which internal physiological sources of heat are of relatively limited importance in controlling body temperature. Such organisms rely on environmental heat sources, which permit them to operate at optimal metabolic rates.

Ecological stoichiometry The balance of multiple chemical elements (mainly carbon, nitrogen, phosphorus) and energy in ecological interactions. It particularly deals with the disequilibrium existing between the nutrient requirements of a consumer and the nutrient availability present in their resources (either mineral for autotrophic organisms or organic for heterotrophic organisms).

 doi:10.1016/B978-0-12-409548-9.10576-7

Occam's Razor Occam's (or Ockham's) razor is a principle attributed to the 14th century logician and Franciscan friar William of Ockham. The most useful statement of the principle for scientists is "when you have two competing theories that make exactly the same predictions, the simpler one is the better."

Fatty acids A fatty acid is a long hydrocarbon chain with variant length and degrees of unsaturation that terminates with a carboxylic acid group. Fatty acids with more than one double bonds are referred to as polyunsaturated fatty acids (PUFA); PUFAs with more than 20 carbons are referred to as highly unsaturated fatty acids (HUFAs). HUFAs are nutritionally critical molecules that animals cannot synthesize but can obtain them through intake of plants.

Introduction

Allometry, also known as biological scaling, describes the dependence of a biological variable on an organism's body mass, size, or shape. Originally used to describe the scaling relationships between body size and metabolic rates, allometric relationships can be expanded into a broader context to include morphological (e.g., total body length with body mass of invertebrates), physiological (e.g., metabolic rates with body size among mammals) or ecological traits (e.g., egg size with survival rates of immature stages in butterflies). Allometry offers the foundation for the development of scaling relationships that capture the variation in physiological mechanisms, individual behaviors such as locomotion or dispersal, as well as spatial distributions, population dynamics, and evolutionary patterns. In principle, allometric relationships stipulate that an easily identifiable predictor, the body size, can provide a reliable estimate of a given biological parameter (Fig. 1).

Most allometric relationships are presented as a simple power function of the form:

$$Y = aW^b \tag{1}$$

where W is the organism's body size, Y is a biological attribute, a and b are the experimentally derived constant and scaling exponent, respectively. Often log-log transformation is used to linearize the relationships, and empirically derive values for the coefficients. The logarithmic-scaled equation is thus:

$$\log Y = \log(a) + b \bullet \log W \tag{2}$$

Numerous allometric equations have been developed to estimate biological attributes as a function of body mass/size, or other morphological features, such as length (L), volume (V), surface area (SA), carbon content, or length: height ratio. The selection of the predictor is generally limited by its measurement precision and practicability for a particular organism. For example, the cell volume of microalgae is relatively easy to measure but the vacuoles may occupy most of the cell volume, thereby introducing considerable discrepancy from the actual algal biomass. While fresh biomass is the critical input for the majority of existing

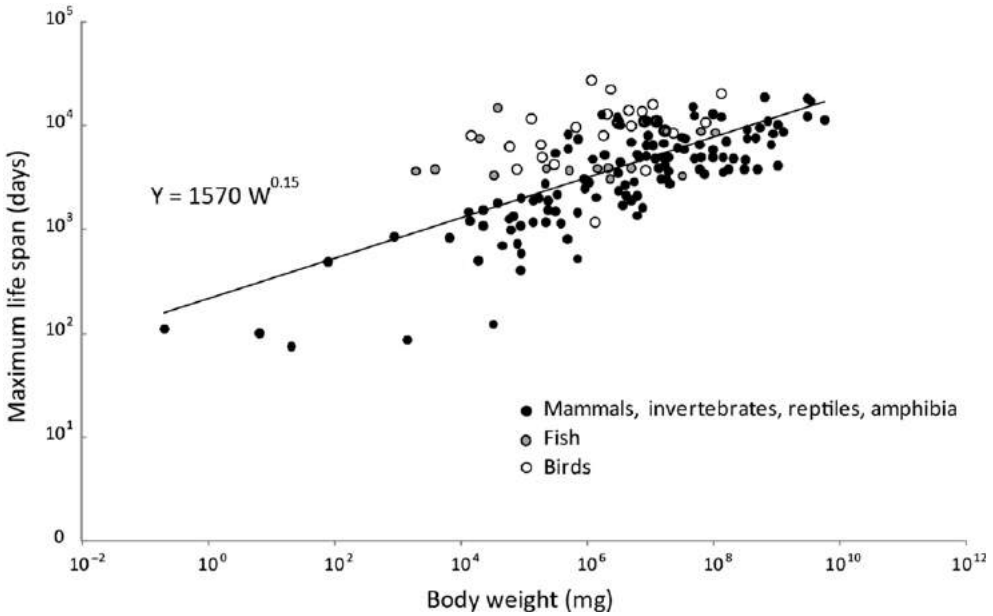


Fig. 1 Maximum life span for a range of animals against adult body weight (mg). Modified with permission from Blueweiss, L., Fox, H., Kudzma, V., Nakashima, D., Peters, R., Sams, S., (1978). Relationships between body size and some life history parameters. *Oecologia* 37, 257–272. https://doi.org/10.1007/BF00344996.

allometric equations, it is often difficult to obtain reliable measurements for many microorganisms. Other allometric relationships are based on the length as a proxy variable of the body size to estimate other morphological characteristics (e.g., stem basal diameter for the below-above ground tree biomass, the length-weight relationships of zooplankton) or organismal size-at-age (e.g., size at maturity). Surface area is often used in the allometric equations of nutrient kinetics and toxicology to quantify the transport of compounds through the cell membrane. In some cases, the ratio to the cell volume (*SA*:*V*) has also been used to account for the effects of morphological traits of unicellular organisms. For example, the uptake of chemical compound is largely regulated by the number of uptake receptors at the cell membrane relative to its cell volume, i.e., species with higher *SA*:*V* demonstrate higher nutrient uptake rates and therefore obtain an advantage when competing for a limiting resource.

Examples of Allometric Principles

Metabolism collectively represents the set of processes in living cells by which energy is provided for vital processes and new material is assimilated in order to maintain life. Metabolism thus determines the demands that organisms place on their environment for all resources. The overall rate of these metabolic processes (often measured as respiration rate; **Table 1**) sets important constraints on the allocation of resources to all components of fitness. Different facets of metabolic activity can be expressed through this strategy. The most common is the basal metabolic rate, which reflects the rate of a fasting, inactive individual; the field metabolic rate characterizing the rate of a free-living individual in its environment; the maximal metabolic rate describing the metabolism of an individual at maximum aerobic activity.

Early research demonstrated that metabolic rate (R, in Watts) could be predicted from fresh body mass (W, in kg) for vastly different groups of organisms (**Fig. 2**A). The fitted allometric relationships for each group all scaled to the ¾ power:

$$R_{homeotherms} = 4.1W^{0.751} \tag{3}$$

$$R_{poikilotherms} = 0.14W^{0.751} \tag{4}$$

$$R_{unicells} = 0.018W^{0.751} \tag{5}$$

The common power in Eqs. **(3)**–**(5)** implies that the allocation of energy and materials to metabolism follow a similar pattern across most organisms. The positive exponent demonstrates that larger organisms within each metabolic class have a higher metabolic rate than do smaller organisms, while a value <1 predicts that metabolic rate rise more slowly as the body size increases. Adjusting these allometric relationships to be expressed as *mass-specific* metabolic rates (wR, in Watts kg^{-1}) yields scaling exponents of the $-¼$ power:

$$\begin{aligned} wR_{homeotherms} &= R_{homeotherms}/W \\ &= 4.1W^{0.751-1} \\ &= 4.1W^{-0.249} \end{aligned} \tag{6}$$

$$wR_{poikilotherms} = 0.14W^{-0.249} \tag{7}$$

$$wR_{unicells} = 0.018W^{-0.249} \tag{8}$$

Thus, the rate of energy expenditure per unit mass declines with increasing body size, while the cost of maintaining a given biomass is less for larger animals than smaller ones.

Subsequent research using this simple allometric equation has been successful in describing the relationship between metabolic rate and body mass for diverse groups of organisms, all characterized by scaling exponents of $\sim$¾. This recurring relationship was adopted as *Kleiber's rule* and its wide applicability suggested that it may be a rare example of a general biological law. However, despite the presence of plausible theoretical explanations for its ubiquity (e.g., processes that control chemical reactions within cells and/or designs of resource distribution networks with functional similarities to all organisms), the debate whether the true value of the scaling exponent b should be set at ¾, ⅔, or whether it is variable, has continued. For example, a recent survey of 642 published regressions of (laboratory-measured) metabolic rates in marine invertebrates found a wide range of scaling exponents for metabolic rates (**Fig. 2**B), with similar species exhibiting scaling exponent values from 0.75 up to (or greater than) 0.9. This finding suggests that the ¾ power law is not universal. While *Kleiber's rule* is a valid statistical generalization, it is important to note that $b = ¾$ is an approximation, rather than the "true" value of the scaling exponent for all allometric equations of metabolic rates.

As metabolism represents the total energetic cost of an organism's biological processes, size-related changes in most other biological functions should parallel the scaling of metabolism. To illustrate this scaling for physiological, morphological, and life-history rates, a series of allometric relationships is presented in **Table 1** for various groups of marine and freshwater zooplankton. Zooplankton is considered as an ideal organism to develop allometric relationships, due to their numerical abundance and ease of sampling in aquatic environments, while their relatively short generation times are conducive to laboratory studies. Furthermore, the trophic linkages between primary producers and zooplankton are arguably the most important in aquatic food webs, as their interactions control the flow of energy to higher trophic levels.

Grazing (ingestion) sets an upper limit to all other physiological rates, and is one of the most significant interactions between an organism and the surrounding environment. Zooplankton grazing can exert significant control on phytoplankton biomass and

Table 1 Representative listing of physiological, morphological, and life-history allometric relationships measured for marine and freshwater zooplankton from the 1970s to 2016. *n* = number of individuals used in the regression analysis; r^2 = coefficient of determination; f = freshwater; *m* = marine; *NA* = not available

Taxon or species	*Allometric relationship*		n	*Allometric equation:* $\log_{10}Y = a + b \bullet \log_{10}W$				*Reference*
	Y	**W**		*Intercept (a)*	*Slope (b)*	r^2	P	
Physiology								
Zooplankton (m)	Respiration rate ($\mu L\ O_2\ ind^{-1}\ h^{-1}$)	Dry mass (mg)	42	0.21	0.90	0.54	<0.0001	Hébert *et al.* (2016)[a]
Zooplankton (f)	Respiration rate ($\mu L\ O_2\ ind^{-1}\ h^{-1}$)	Dry mass (mg)	17	1.24	0.82	0.94	<0.0001	Hébert *et al.* (2016)[a]
Daphnia pulicaria (f)	Ingestion rate ($\mu g\ C\ ind^{-1}\ h^{-1}$)	Body length (mm)	128	−1.067	2.739	0.747	NA	Carotenuto and Lampert (2004)
Cladocerans (f)	Filtering rate ($mL\ day^{-1}$)	Body length (mm)	519	0.896	2.403	0.867	NA	Knoechel and Holtby (1986)
Daphnia (f)	Clearance rate ($mL\ ind^{-1}\ h^{-1}$)	Body length (mm)	30	0.210	2.83	0.98	NA	Demott (1982)
Zooplankton (f, m)	P excretion rate (mmol $P\text{-}PO_4^{3-}$ $ind^{-1}\ h^{-1}$)	Dry mass (mg)	47	0.56	0.70	0.72	<0.0001	Hébert *et al.* (2016)[a]
Zooplankton (f, m)	P excretion rate ($\mu g\ P\ day^{-1}$)	Body size (μg)	462	−1.65	0.54	0.63	<0.001	Wen and Peters (1994)
Zooplankton (f, m)	N excretion rate ($\mu g\ N\ day^{-1}$)	Body size (μg)	574	−1.38	0.67	0.72	<0.001	Wen and Peters (1994)
Zooplankton (f, m)	N excretion rate (mmol $N\text{-}NH_4$ $ind^{-1}\ h^{-1}$)	Dry mass (mg)	71	2.50	0.84	0.73	<0.0001	Hébert *et al.* (2016)[a]
Morphology								
Zooplankton (m)	Body dry mass (mg)	Body length (mm)	37	−3.910	2.791	0.94	<0.001	Hébert *et al.* (2016)[a]
Zooplankton (f)	Body dry mass (mg)	Body length (mm)	148	−4.814	2.075	0.74	<0.001	Hébert *et al.* (2016)[a]
Cladocerans (f)	Dry weight (μg)	Total length (mm)	283	0.994	2.1	0.84	NA	Peters and Downing (1984)
Life-history								
Copepods (m) (broadcast spawners)	Weight-specific fecundity (day^{-1})	Female body weight(μg C)	35	−0.474	−0.262	0.32	<0.001	Kiørboe and Sabatini (1995)
Copepods (m) (sac-spawners)	Weight-specific fecundity (day^{-1})	Female body weight(μg C)	10	−0.850	−0.260	0.72	<0.001	Kiørboe and Sabatini (1995)
Zooplankton (f, m) (flagellates, ciliates, rotifers,meroplankton larvae, copepods, cladocerans)	Growth rate (hr^{-1})	Body volume (μm^3)	69	−0.52	−0.21	0.69	<0.01	Hansen *et al.* (1997)
Copepods,cladocerans, rotifers (f, m)	Generation time ($^\circ C\ day^{-1}$)	Dry body mass (μg)	111	2.26	0.21	0.72	<0.001	Gillooly (2000)

[a]Regression coefficients estimated using natural instead of $\log_{10}$ transformation.

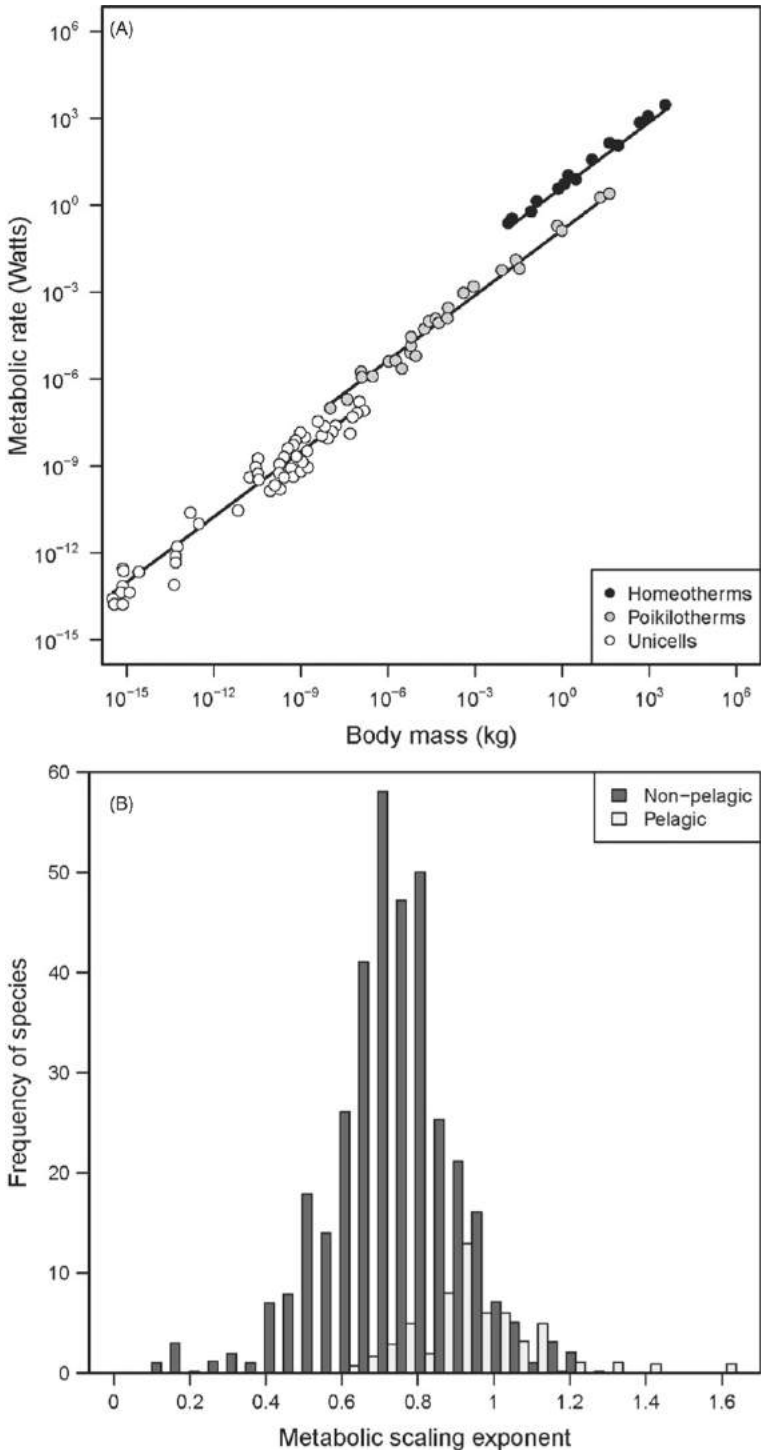


Fig. 2 (A) Standard metabolic rates of homeotherms, poikilotherms, and unicells (Eqs. (1)–(3) in text). (b) Frequency distributions of scaling exponents from regressions of log metabolic rate vs. log body mass of pelagic and nonpelagic marine invertebrate species. (A) Modified from Peters, R. H. The ecological implications of body size. Cambridge, UK: Cambridge University Press, 1983. (B) Modified with permission from Glazier, D.S. (2005). Beyond the "3/4-power law": Variation in the intra- and interspecific scaling of metabolic rate in animals. *Biol. Rev.* **80**, 611–662. https://doi.org/10.1017/S1464793105006834.

species composition. The allometric scaling of ingestion (measured as ingestion, filtration, or clearance rates) with zooplankton body size (measured as body length) suggests that larger zooplankton graze particles at a higher rate than smaller ones (**Table 1**). Similarly, a unit increase in zooplankton body size results in 2.4–2.8 increase in ingestion rate (**Table 1**), and therefore larger zooplankton have a greater impact on energy transfer through the food web.

In aquatic ecosystems, zooplankton regenerate mineral nutrients (e.g., nitrogen, phosphorus) via their excretion; thus, changes in the rate of materials recycled can directly impact the abundance and composition of primary producers. Body size has a large effect on excretion rates in zooplankton, where scaling exponents <1 indicate that larger zooplankton excrete at a lower rate for their size than smaller taxa (Table 1). The less than proportional increase in excretion rate is linked to metabolism. Interestingly, the scaling exponents can be higher or lower than ¾ depending on the nutrient recycled, as nitrogen excretion rates scale to the 0.67–0.84 power of body mass, while phosphorus excretion rates can be slightly lower ($b = 0.54$–0.70) (Table 1).

The measurement of zooplankton biomass, together with the productivity assessment, is an important component for estimating standing stocks, as well as community structure and dynamics. Only two quantities are required to calculate the production of a population, since multiplying the number of individuals by their average mass yields an estimate of the total population biomass. The production rate is then the amount of biomass accumulated by a population per unit of time. The most common way to estimate average mass of zooplankton involves prediction of animal weight from body length with allometric equations (Table 1). These relationships suggest that there is a greater than proportional increase in zooplankton weight with a corresponding increase in body size. The variation (in scaling exponents) among allometric regressions may relate to environment influences at the study site in question, such as latitude, temperature, and/or food availability.

Body size imposes constraints on the life-history (e.g., growth, fecundity, survival) of organisms. Quantifying the variation of the relationship between life-history processes and body size can offer insights into the natural selection pressures. For example, examining rates of fecundity, mortality, and developmental time as a function of (adult) body mass is critical for understanding zooplankton evolution. Generation time (time span from egg to maturity) is a critical determinant of the rate of population growth in zooplankton, and scales positively to body mass close to the ¼ power (Table 1).Weight-specific fecundity rates scale close to the $-¼$ power for both broadcast and sac-spawning marine copepods, although the intercepts differ, suggesting that weight-specific fecundities are $\sim$2.5 times greater for broadcast spawners than sac-spawners, which in turn may relate to their different life-cycle strategies (Table 1). Maximum specific growth rates appear to scale to the $-¼$ power, declining with increasing body volume across a wide range of marine and freshwater zooplankton groups (e.g., dinoflagellates, other flagellates, ciliates, rotifers, meroplankton larvae, cladocerans, and copepods), suggesting that small zooplankton have higher mass-specific metabolic rates and grow faster than large ones (Table 1).

The predictive power of scaling laws to life-history processes depends on whether natural selection can alter the scaling exponent value. If the scaling exponent varies with environment, then this reflects the strong role of selection on this exponent. Returning to the large survey of allometric scaling exponents presented earlier, pelagic (open-water) species had significantly greater mean scaling exponents ($b = 0.947$) than those of nonpelagic species ($b = 0.744$) (Fig. 2B). While sampling error can explain some of this variation, it may also reflect real biological differences in metabolic rates across taxa with diverse body characteristics and widely separated phylogenies.

Recent Advances in the Application of Allometric Principles

Metabolic Theory of Ecology

Metabolic and other process rates are strongly affected by both body size and temperature. More recent theoretical advances have combined first principles of allometry and biochemical kinetics to develop the metabolic theory of ecology (MTE). MTE uses scaling functions to incorporate the effects of body size and temperature on individual metabolic rates, which in turn modulate the performance of individual organisms and subsequently the ecology of populations, communities, and ecosystems. The joint effects of body mass (M), and temperature (T, in K) on individual metabolic rate (I), is given as:

$$I = i_0 M^{3/4} e^{-E/kT} \tag{9}$$

where E is the activation energy (0.6–0.7 eV), k is the Boltzmann constant (8.62×10^{-5} eV/K), and i_0 is a normalization constant independent of body size and temperature. Across diverse groups of organisms (from unicellular microbes to the largest vertebrates and trees), this relationship predicts a 100,000-fold variation in metabolic rates over 20 orders-of-magnitude in body size, while temperature predicts a $\sim$30-fold variation in metabolic rates over the temperature range of normal activity for most organisms (0°C–40°C). The dependence on mass may be a consequence of the scaling of resource supply and exchange surfaces in branching hierarchical networks, while the dependence on temperature reflects its impact on biochemical reaction rates. For ectotherms, this is equivalent to ambient temperature, while for endotherms, this temperature is high (35°C–40°C) and mostly temperature independent.

Taking the logarithm of both sides and rearranging terms yields a linear relationship:

$$\ln\left(IM^{-3}/4\right) = -E(1/kT) + \ln(i_0) \tag{10}$$

By incorporating the logarithm of mass raised to the ¾ power, the metabolic rate has been "mass-corrected," and the predicted scaling is incorporated into the y-axis of bivariate plots. This mass-corrected relationship predicts that the (natural logarithm of) whole-organism metabolic rate is a linear function of inverse absolute temperature; the slope gives the activation energy of

metabolism, E, while the intercept represents the (natural logarithm) normalization constant, $\ln(i_o)$. To isolate the effects of body mass, metabolic rates can be "temperature-corrected" using the Boltzmann-Arrhenius factor ($e^{-E/kT}$), to yield:

$$\ln\left(Ie^{E/kT}\right) = \left({}^{3}/_{4}\right)\ln\left(M\right) + \ln\left(i_0\right) \tag{11}$$

This equation predicts a linear relationship between the (natural logarithm) temperature-corrected metabolic rate and (natural logarithm) body mass. Comparing these relationships to empirical data collected from a wide variety of taxa (e.g., endotherms, fish, amphibians, reptiles, invertebrates, unicellular organisms, and plants) demonstrated that the observed slope of mass-corrected metabolic rates for all groups fell within the predicted range of 0.6–0.7 eV (Fig. 3A). In comparison, temperature-corrected metabolic rates for all groups clustered closely around a common allometric scaling relationship with an exponent of 0.71, which is close to the ¾ power predicted from theory, although the intercepts (normalization constants) vary among groups (Fig. 3B).

By isolating the effect of mass (or alternatively temperature), the MTE can be used to investigate other biological rates, which are predicted to scale as $M^{-1/4}$, and biological times, which are expected to scale as $M^{1/4}$. For example, the MTE has been used to predict ontogenic development as a function of body mass and temperature. Using zooplankton eggs reared in the laboratory and fish eggs collected from the field, plots of temperature-corrected hatching rate (day^{-1}) versus body mass (g) were well fit by straight lines with similar slopes very close to the predicted $-¼$ power (-0.26 and -0.22, respectively). The MTE has also been incorporated into allometric modeling to predict excretion rates of invasive fish species at different temperatures. Accounting for the role of temperature is critical when predicting seasonal excretion rates in freshwater ecosystems. Consistent with the MTE, mass-specific nutrient excretion rates decreased with increasing fish size, such that smaller fish generally excreted more nutrients per gram of body mass than larger fish, and were greater in summer than winter.

The MTE framework has been extended to enable predictions regarding population growth. Metabolic rates of organisms determine their rates of growth and reproduction, which in turn fuels population growth. Population growth rate is often measured as net outcome of the maximal growth rate (the capacity of a population to reproduce at maximum rate when resources are not limiting), and the rate of turnover at steady state (where the total number of individuals in the population does not change over time). Allometric scaling relationships of temperature-corrected maximal growth rate as a function of body size for a wide variety of taxa, from unicellular eukaryotes to vertebrates, are suggestive of a single line with a slope of -0.23, across 12 orders-of-magnitude of variation in body size. The rate of population turnover, and thus birth and death rates, scaled similarly. For example,

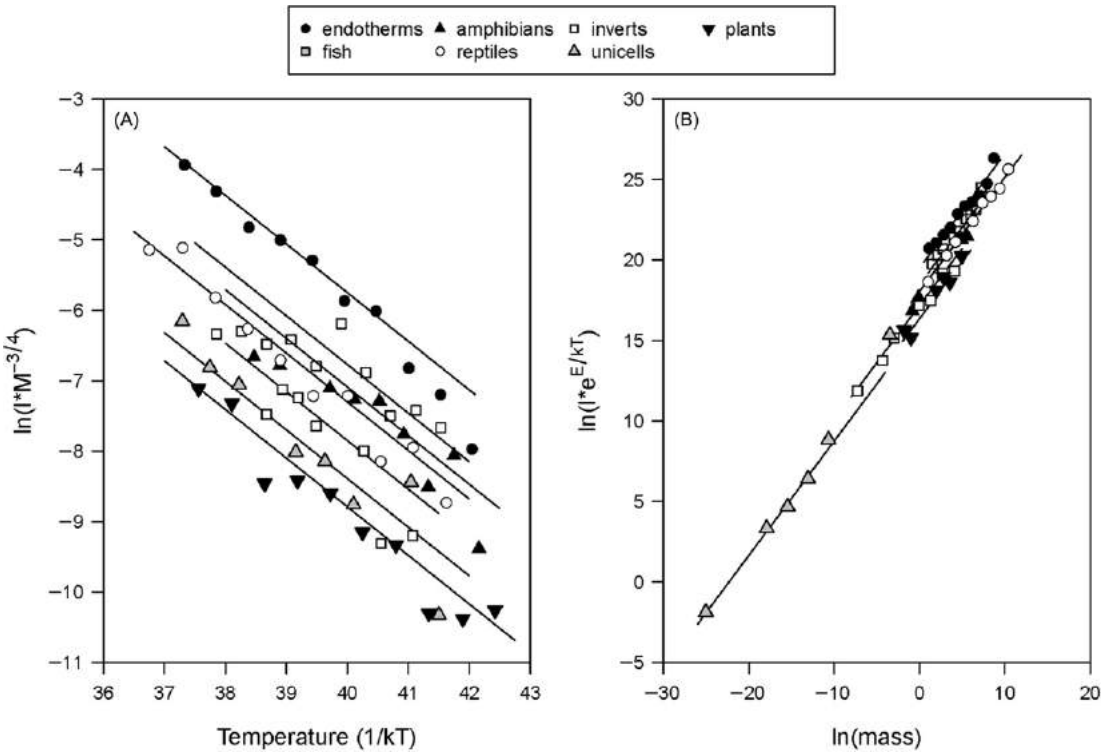


Fig. 3 Allometric scaling relationships of (A) mass-corrected metabolic rate, $\ln(I \cdot M^{-¾})$ (in Watts $g^{-¾}$), or (B) temperature-corrected metabolic rate, $\ln(I \cdot e^{E/kT})$ (in Watts), as a function of body mass, $\ln(M)$ (in grams) for endotherms, fish, amphibians, reptiles, invertebrates, unicellular organisms, and plants. k = Boltzmann constant; T = absolute temperature (in K); E = activation energy (in eV). Modified with permission from Brown, J. H., Gillooly, J. F., Allen, A. P., Savage, V. M. and West, G. B., (2004). Toward a metabolic theory of ecology. *Ecology*, **85**, 1771–1789. https://doi.org/10.1890/03-9000.

an allometric equation predicting temperature-corrected instantaneous mortality rate for marine fish in the field yields a straight line and a scaling exponent of -0.23, statistically indistinguishable from the predicted $-¼$ power.

The MTE framework has also been extended to link the processing of energy and elements at the individual level to the flux, storage, and turnover of these elements at the ecosystem level. For example, the constraints that body size and temperature place on C dynamics at all levels of biological organization (e.g., cellular organelles to the biosphere) can be combined to create a model that relates the global C cycle directly to the C flux, storage, and turnover in individual organisms. Compilations of data from major biomes that include forests, grasslands, tundra, and oceans, have demonstrated that allometric relationships between C storage in plant communities as a function of average plant size scale as a 0.24 power, while C turnover expressed as a function of average plant size yielded scaling exponents of -0.22. Both empirically-estimated scaling exponents were very close to their predicted quarter-power scaling values.

To recap, energy and mass are not distinctly different ecological currencies that operate independently of each other to shape ecosystem structure and dynamics. At all levels, from individual organisms to ecosystems, fluxes, reserves, and transformations of energy and materials are constrained by the biochemical and physiological constraints of metabolism. Within this context, the metabolic theory appears to explain much of the variability in process rates and somatic reserves. However, it is important to note that although metabolism is one of the most integrative processes in biology, building connections from molecules to ecosystems, metabolic theory cannot account for all important patterns and processes. The existence of residual variation around the predictions of the metabolic theory underscores the importance of other variables and processes not considered by the contemporary paradigm.

Ecological Stoichiometry and Metabolic Theory of Ecology

Ecological stoichiometry (ES) theory aims to quantify how variations in the balance of biologically important elements impact, and in turn are impacted by, organisms and their environment. Integration of the theory of ES with MTE may offer a useful framework to link the dynamics of energy and mass across different levels of biological organization. The ES and MTE models are founded upon a series of principles that link the energetics and stoichiometry at the level of cellular organelles with individual-level energetics and stoichiometry, and ultimately with higher-order ecosystem processes. The four major principles proposed state that:

1. Links between the fluxes of energy and materials are based on the kinetics and elemental compositions of processes and subcellular structures;
2. Biomass is comprised of metabolic and structural pools, which can have distinct allometric and elemental signature;
3. Metabolic rate (and its determinants) govern the fluxes of energy and elements at the organismal level;
4. The storage, flux, and turnover of energy and mass in a biological community can be estimated by summing across individuals within that community.

Patterns in rates of consumption by herbivores across freshwater, marine, and terrestrial ecosystems, at the individual and population level, have been used to validate joint ES–MTE predictions. Under the combined ES–MTE theory, per-capita rates of herbivory are expected to:

a. increase with body size;
b. increase with ambient temperatures for ectotherms, whereas for endotherms, either be independent of temperature or decrease with increasing temperature if high metabolic costs occur at low temperatures; and
c. increase with increasing stoichiometric mismatch.

In contrast, population-level rates of herbivory are expected to:

d. be independent of body size;
e. increase with increasing ambient temperature for both *endo*- and ectotherms; and
f. decrease with increasing stoichiometric mismatch.

Empirical data have rendered support to the predicted metabolic and stoichiometric constraints on herbivory at individual and population levels (**Fig. 4**). At the individual level, the body size of herbivores appears to be an important factor of the per-capita consumption rates for large gradients of body size, ranging from small zooplankton to large mammalian herbivores, with larger species consuming more biomass per-capita. Consumption rates in individual endotherms declined at high ambient temperatures, but increased for individual ectotherms. Stoichiometric mismatch had a small, positive effect on per-capita consumption rates. At the population level, consumption rate was invariant with body size, increased with ambient temperature for both ectotherms and endotherms, and declined with increasing stoichiometric mismatch (**Fig. 4**). Interestingly, examining per-capita consumption rates within ecologically similar groups (e.g., taxa with similar sizes and ecological roles) reduced the explanatory power of body size, while stoichiometric mismatch emerged as a more influential factor. Furthermore, MTE-related variables (body size, temperature) are more closely related to per-capita consumption rates, while ES-related variables (stoichiometric mismatch) appear to shape population-level rates. Overall, these results suggest that the integration of ES-MTE theories offer a microscopic-to-macroscopic strategy that can explicitly relate the energetics and stoichiometry of individuals, communities and ecosystems to subcellular structures and processes.

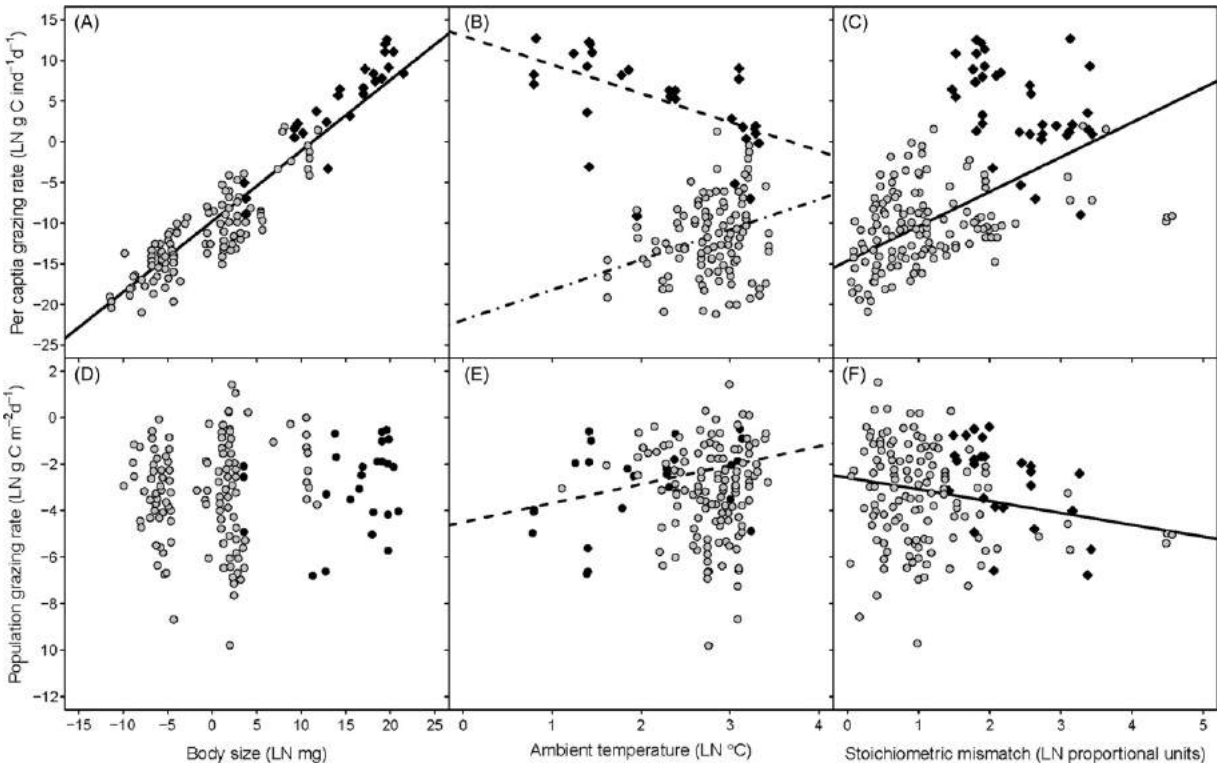


Fig. 4 Per-capita (top panels) and population-level (bottom panels) rates of herbivory as functions of body size, ambient temperature, and stoichiometric mismatch between prey and consumers. *Gray circles* = ectotherms; *black diamonds* = endotherms. Modified with permission from Hillebrand, H., Borer, E.T., Bracken, M.E. S., Cardinale, B.J., Cebrian, J., Cleland, E.E., Elser, J.J., Gruner, D.S., Harpole, W.S., Ngai, J.T., Sandin, S., Seabloom, E.W., Shurin, J.B., Smith, J.E., Smith, M.D. (2009). Herbivore metabolism and stoichiometry each constrain herbivory at different organizational scales across ecosystems. *Ecology Letters*, **12**, 516–527. https://doi.org/10.1111/j.1461-0248.2009.01304.x.

Food-Web Interactions

The allometric approach has been primarily considered in the context of organismal physiology and less so to elucidate ecosystem processes. More recently, there has been a growing emphasis on the idea that functional traits, as developed through natural selection, can directly affect the intra-specific variations and may ultimately determine demographic performance, spatial distribution (i.e., ecological niche), population dynamics, and food-web organization. In particular, the body size is one of the most fundamental functional traits that shapes predator-prey interactions and may conceivably modulate other ecological processes (e.g., sedimentation, nutrient recycling, foraging, and migration).

Consumer-resource foraging interaction has long been identified as one of the dominant forces that connect individual ecosystem components (i.e., species, populations, trophic groups). Foraging interactions are traditionally characterized by two types of functional response curves: hyperbolic (type II) and sigmoid (type III) curves, where a consumer's per capita feeding rate (F) increases with the resource abundance (N). Originally established by Holling (1959), foraging ecology has yielded many variations of the functional response models. One of the generalized models describes consumer's feeding rate as:

$$F = \frac{bN^{1+q}}{1 + bhN^{1+q}} \tag{12a}$$

Per capita feeding rate is regulated by the time required to kill, ingest, and digest a resource (handling time, h), as well as the hunting efficiency representing the rate that a resource is captured by a consumer (bN), where b is a coefficient for hunting efficiency. The scaling exponent (q) dictates the response curve type, whose value switches from the hyperbolic type II ($q=0$) into the sigmoid type III ($q>0$) functional response. Although this simple concept around the functional response models has provided mechanistic understanding of consumer-resource interactions, it has failed to characterize more dynamic interactions in natural communities. For example, a predatory spider in central Europe did not follow any of the expected response curves, displaying nonlinear and/or nonmonotonic patterns under conditions of abundant food availability.

In recent years, a size-based perspective has been increasingly recognized as an alternative approach to estimating density-dependent foraging interactions. In size-based functional response models, handling time and/or capturing rate (i.e., hunting efficiency) are expressed as a function of the consumer's body size. For example, handling time (h) in Eq. **(12a)** can be replaced by:

$$h = h_0 m_r^{\alpha_r} m_c^{\alpha_c} \tag{12b}$$

where h_0 is a constant, m_c, m_r and α_c and α_r are body masses (m) and allometric exponents (α) of consumer/predator c and resource/prey r, respectively.

Originally considered as a constant in functional-response models, capture rate, also known as attack rate, is known to have its own unique body-size dependency. Predator-prey mass ratio (PPMR) has been identified as an indicator to depict body size constraints on the capacity of predators to efficiently utilize excessively large prey. On the other hand, a predator does not fully benefit by targeting excessively small prey because of its limited nutritional value relative to the energy expense for the consumption or handling of the prey. PPMR has been used to measure the strength of the trophic interaction between predators and prey, and to illustrate potential shifts in the energy flows and the reliance on specific resources. In general, capturing rate has a hump-shaped response pattern with PPMR, maximizing at intermediate/moderate PPMR, while energy flows at high and low PPMR may not sustain the predator's biomass or could even risk their survival (**Fig. 5**). This hump-shaped relationship between capture rate and PPMR is consistent with optimal-foraging and niche theory, and has broad generality across different species in various habitats.

To accommodate body-mass constraints on the capture rate, hump-shaped relationships with PPMR can be further expressed as a combined equation comprising a power-law relationship with prey body mass and an exponential Ricker function for the optimal foraging body-mass ratio. The capture coefficient (b) in Eq. **(12a)** is then replaced by an allometric scaling relationship, where b_0 is a constant, β_r the exponent for the scaling of m_r, and ϵ is a constant for the range of the optimal foraging body-mass ratio:

$$b = b_0 m_r^{\beta_r} \frac{m_c}{m_r} e^{\epsilon \frac{m_c}{m_r}} \tag{12c}$$

Efforts to develop a generalized functional-response modeling framework have further advanced our mechanistic understanding of consumer-resource interactions in natural food-web dynamics. Counter to early consumer-resource models, such as the Lotka-Volterra models, where all parameter sets are independently assigned only to fit data points, allometric scaling models ensure that parameters lie within biologically plausible ranges based on their size related capacity. Size-based functional response models highlighted that the coexistence of consumer-resource is restricted within specific body-mass ranges, which in turn regulates the resilience of consumers in the food-web and the broader stability of biotic communities. The consideration of body-mass ratio can also be useful in parameterizing more complex population models. Nonetheless, like any other type of models, allometric scaling models require caution in their use; as the definitions of certain facets of trophic interactions, like the hunting efficiency, are somewhat ambiguous and their complex characteristics (i.e., multiple food-sources, mobility of organisms) are essentially habitat/community-specific.

Ecosystem Models

Ecosystem models extend the application of consumer-resource interactions to community-level processes, whereby ecologists attempt to reproduce the interplay among organisms and their surrounding physico–chemical environment. Integration of size-based characteristics into ecosystem models may refine the description of community-level processes, such as species seasonal succession and flow of energy and/or matter across trophic levels. Body-size patterns may be more important, or at least more obvious, in aquatic ecosystems than in terrestrial habitats for several reasons. The majority of autotrophic organisms in aquatic ecosystems are very small and grazed by larger consumers, and thus the relative size ratio of consumer to resource more consistently manifests itself than in terrestrial environments. There are significant operational and technical advantages in the collection of datasets from aquatic environments. Reflecting upon these factors, there are a number of allometric equations developed for aquatic organisms and (not surprisingly) many of the existing size-based ecosystem models have been developed for aquatic environments.

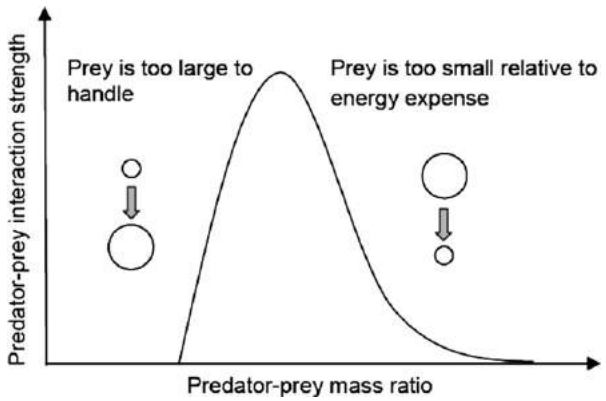


Fig. 5 Conceptualization of the relationship between predator-prey mass ratio and strength of their interaction. Circles represent the relative mass of predators (top) to the mass of prey (bottom). Modified from Nakazawa, T. (2017). Individual interaction data are required in community ecology: A conceptual review of the predator–prey mass ratio and more. *Ecological Research* **32**, 5. https://doi.org/10.1007/s11284-016-1408-1.

The basic concept of size-based ecosystem modeling is that the functional characteristics of the biological components of the studied system can be described by allometric equations. That is, model parameters associated with organismal physiological rates, such as maximum growth rate, nutrient kinetics, and basal metabolism, are determined by the empirically-derived relationships with their body size (i.e., mass, volume, length). Shimoda *et al.* (2016) employed several allometric equations in an existing aquatic biogeochemical model to describe the physiological processes of multiple phytoplankton functional groups, and thus predict how can the morphological features (i.e., cell volume, surface-to-volume ratio, and shape) influence the response to external perturbations, interspecific competition, and ultimately the seasonal composition of algal assemblages (Fig. 6A). For example, maximum growth rate μ_{max} (day^{-1}) was replaced by $\mu_{max}=10^{0.54}V^{-0.15}$, where V denotes algal cell volume (μm^3) (Fig. 6B). Nutrient kinetics were also replaced by several allometric equations such as: half saturation constant for nitrate uptake (NH:$\mu mol\ N\ L^{-1}$), $NH=10^{-0.72}V^{0.52}$, half saturation constant for phosphorus uptake (KH_p:$\mu mol\ P\ L^{-1}$), $KH_p=10^{-1.5}V^{0.53}$ (Fig. 6C), maximum phosphorus uptake rate (VP_{max}:$\mu g\ P\ \mu m^{-3}h^{-1}$), $VP_{max}=10^{-10.7}SA/V^{1.7}$ where SA/V denotes algal cell surface-to-volume ratio (μm^{-1}), maximum internal phosphorus quota (QP_{max}:$fmol\ P\ cell^{-1}$), $QP_{max}=10^{-0.29}V^{0.767}$, minimum internal phosphorus quota (QP_{min}:$fmol\ P\ cell^{-1}$), $QP_{min}=10^{-1.04}V^{0.714}$.

The allometric configuration of the process-based model allowed to realistically reproducing the observed phosphate, total phosphorus, nitrate, total ammonia, total nitrogen, chlorophyll *a*, and total zooplankton biomass patterns in the Hamilton Harbor, Ontario, Canada. Consistent with empirical evidence, the allometric-scaled ecosystem model showed that small algal species have a distinct competitive advantage in summer epilimnetic environments across the range of cell volume and nutrient loading conditions examined; especially, when they are characterized by higher optimal temperature for growth. The same study

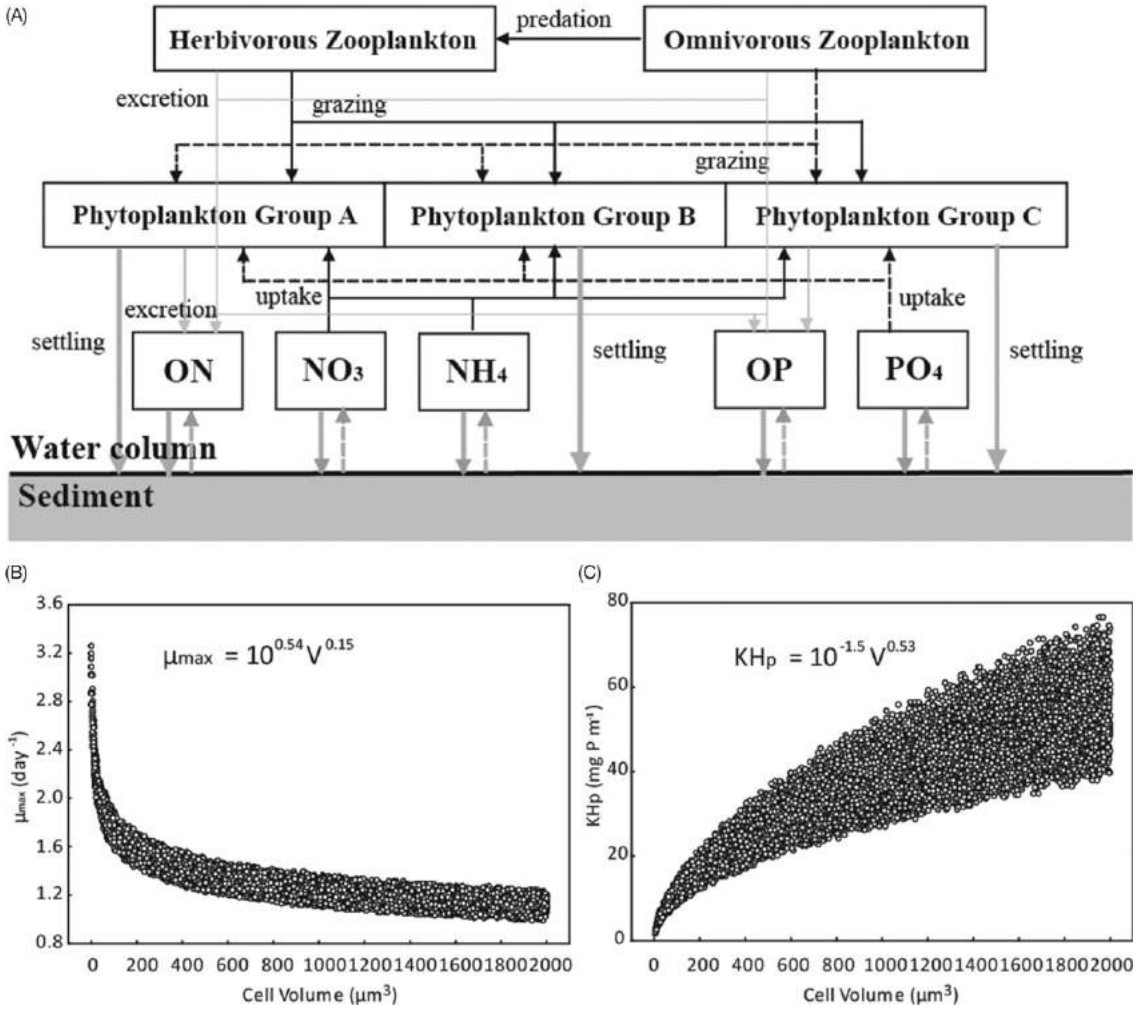


Fig. 6 (A) Conceptual diagram of an aquatic ecosystem model. Parameter values were calculated with allometric equations linking phytoplankton (B) maximum growth rate and (C) half saturation constant for phosphorus uptake with the cell volume. In Shimoda *et al.* (2016), the cell size variability for a given functional groups was assigned a range between 0.1 and 2000 μm^3 (*x*-axis). The parametric uncertainty (as seen in the distributed dots in B and C) of the allometric regressions can be propagated with Monte Carlo simulations through the ecosystem model. (B-C) Shimoda, Y., Yerubandi, R., Watson, S., Arhonditsis, G.B. (2016). Optimizing the complexity of phytoplankton functional group modelling: An allometric approach. *Ecological Informatics* 31, 1–17.

also showed that intense herbivory rates act as a "safety valve" and effectively control the standing biomass of phytoplankton species that can otherwise realize high growth rates under the conditions typically prevailing in the end-of-summer epilimnetic environments. By contrast, when the summer community is released by the zooplankton grazing, the exceedance of critical phytoplankton biomass levels and the likelihood of harmful algal blooms are determined by the multitude of factors that shape inter-specific competition patterns (e.g., relative abundance of competing species, nutrient uptake kinetics). One of the future challenges associated with the allometric approach to plankton modeling involves the characterization of dependence of prey-predator relationships as a function of the interplay among prey morphology, nitrogen, phosphorus, highly unsaturated fatty acids, and other potentially important metabolic congeners through the grazers' digestive tracks.

From a technical standpoint, one of the benefits of the allometric approach to ecosystem modeling is that the characterization of simulated biotic compartments is no longer based on adjustable parameters, often treated as "properties of convenience" for fitting models to the observed data, but instead their morphological features are treated as the common denominator that influences the corresponding physiological rates. In a broader context, this practice may be one of the ways to address the problem of complex over-parameterized models and improve our ability to set quantitative (or even qualitative) constraints while ensuring satisfactory model performance. Model parametric uncertainty is more effectively delineated; namely, the literature-based ranges typically assigned to the calibration parameters can be replaced by the parameter standard error values and/or the estimates of residual variability of allometric equations, which in turn collectively reflect how well does the regression line match the original empirical physiological rates, the variability of the predictor (morphological) variable used to develop the allometric equation as well as the sample size. Using suitable uncertainty analysis techniques, these error estimates can then be propagated through our ecosystem models, whereby we can effectively quantify the degree of confidence in model predictions (**Fig. 6**C).

The allometric approach to ecosystem modeling also offers a different perspective on the optimization of future data collection. Model calibration is not solely perceived as a typical inverse solution exercise, constantly inviting the collection of data on model outputs and subsequently readjusting the parameters to match measurements and predictions. Instead model parameter estimation requires a more robust experimentation focused on the development (or further refinement) of the causal description of model parameters based on the morphological features of the biotic components modeled. Moreover, depending on the nature of the dataset used for the allometric regressions (e.g., marine vs. freshwater algae), the proposed method allows the potential users to delineate the application domain more easily and determine to what extent a particular model has local or universal use.

Cautions, challenges, and future prospects

Allometric theory meets a series of fundamental criteria to be considered as a "good theory." It is extremely simple, quantitative, and most of the dependent and independent variables are easily defined. Empirical relations capture the characteristics of a wide range of organisms, and provide reasonably accurate predictions of many biological processes. Thus, allometric equations have been used to describe biological and ecological processes, ranging from the micro- to macroscale, such as the effects of drugs and other substances on the physiological responses of humans/animals, the estimation of fish stocks, and the prediction of life-stage specific population size in ecosystems, to name a few.

The popularity of the allometric theory, however, appears to have slowed down after its peak around 1970–1980s and much less development and examination of the corresponding equations has been seen in recent literature. It has been argued that its simplicity and generality (supposedly criteria for a good theory) paradoxically represent a "double-edged sword" making it less appealing for research, because any new regressions merely provide validation of the theory. There are also a number of critical viewpoints presented in the literature. First, the precision of the empirical relations is compromised by our desire to achieve generality. It is obvious that capturing the entire range of a physiological characteristic using a regression with only two parameters is nearly impossible. Even though the development of taxa-specific allometric regression models is logical, as phylogenetically neighboring taxa share common physiological traits, other important factors, such as resource availability, may not yield commonality in response to similar selective pressures. For example, the distribution patterns of maximum potential growth rate of phytoplankton vary not only among genera, but also by habitat type (**Fig. 7**). There is still significant space for advancing allometric theory using scientific creativity and technological innovation.

One of the most significant, yet not frequently explored facets of allometry, involves the scatter around each regression (e.g., model residuals) representing the variability among organisms, missing ecological functions (i.e., adaptation), and measurement errors. Many of the empirical equations used are based on small sample sizes (low degrees of freedoms) and/or capture a fairly narrow range of body/cell sizes typically encountered in natural ecosystems. All these factors inflate the magnitude of the uncertainty (confidence, predictive) bands of allometric equations. Rather than perceiving this error as a weakness of the allometric approach, recent viewpoints claim that it offers an excellent piece of information to conduct rigorous uncertainty analysis in complex process-based models and a solid foundation to draw probabilistic inference on important ecological questions.

Many of the existing equations have been derived in experimental controlled settings (i.e., laboratory, mesocosms), and do not necessarily represent the response of organisms that may be observed in the natural world. Unless allometric equations predict maximum potential metabolic/physiological rate under resource saturated environment, more accurate representation of biological traits must be tested with free-living organisms. Although validation of theoretical relationships and existing models is an integral part of science, few studies have attempted to provide a comprehensive review of the existing allometric equations. Often many equations that describe the same phenomenon exist, but objective comparison of these equations has been rarely performed. Further development of new, taxon-specific allometric relationships, expansion of size range for the existing relations may improve their credibility and predictive power.

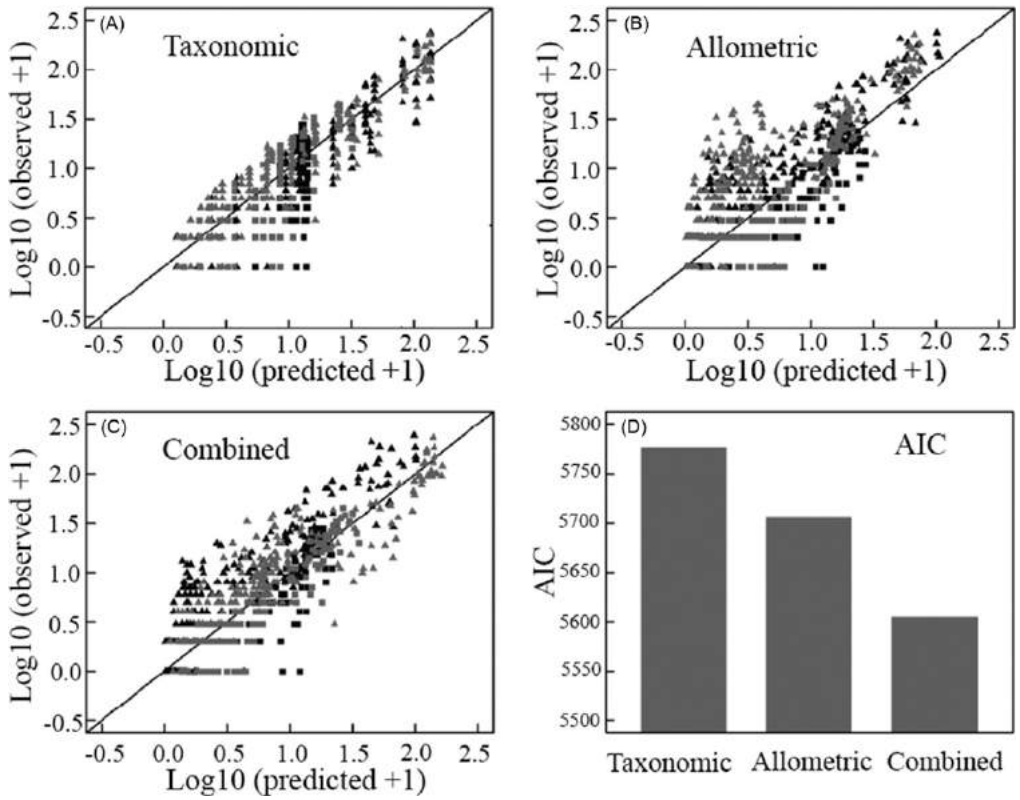


Fig. 7 Goodness of fit comparison of (A) taxonomic, (B) allometric and (C) combined models. The empirically observed values of per capita feeding rates (*y*-axis) are plotted against the values predicted by the models (*x*-axis) using the statistically fitted parameters. (D) AIC of three models. Markers represents: centipedes feeding on fruit flies (*black squares*); centipedes feeding on springtails (*black triangles*); spiders feeding on fruit flies (*gray squares*); spiders feeding on springtails (*gray triangles*). Modified with permission from Rall,B. C., Kalinkat, G., Ott, D., Vucic-Pestic, O., Brose, U. (2011). Taxonomic versus allometric constraints on non-linear interaction strengths. *Oikos*, **120**(4), 483–492. https://doi.org/10.1111/j.1600-0706.2010.18860.x.

While Occam's razor is (and should be) the cornerstone of any ecological modeling activity, the integration of process-based models and empirical parameter specification, founded upon the basic allometric concepts, offers an appealing prospect. The development of predictive ecological frameworks that are based on our best mechanistic understanding of biotic processes and ecosystem feedback loops, yet remain within the bounds of data-based parameter estimation and therefore can accommodate rigorous error analysis has both methodological and ecophysiological advantages. Size structure of biotic communities is an important regulatory factor of the biogeochemical fluxes and energy transfer via the food webs that ultimately affects system productivity. The improvement of empirical description of ecological parameters could reconcile the debate regarding the need to balance between simplicity and realism in predictive ecology.

See also: Ecological Data Analysis and Modelling: Grassland Models. Ecosystems: Ecosystems

References

Carotenuto, Y., Lampert, W., 2004. Ingestion and incorporation of freshwater diatoms by *Daphnia pulicaria*: Do morphology and oxylipin production matter? Journal of Plankton Research 26 (5), 563–569.

Demott, W.R., 1982. Feeding selectivities and relative ingestion rates of *Daphnia* and *Bosmina*. Limnology and Oceanography 27 (3), 518–527.

Gillooly, J.F., 2000. Effect of body size and temperature on generation time in zooplankton. Journal of Plankton Research 22 (2), 241–251.

Hansen, P.J., Bjørnsen, P.K., Hansen, B.W., 1997. Zooplankton grazing and growth: Scaling within the 2-2,-μm body size range. Limnology and Oceanography 42 (4), 687–704.

Hébert, M.-P., Beisner, B.E., Maranger, R., 2016. A meta-analysis of zooplankton functional traits influencing ecosystem function. Ecology 97 (4), 1069–1080.

Holling, C.S., 1959. The components of predation as revealed by a study of small-mammal predation of the European pine sawfly. The Canadian Entomologist 91 (5), 293–320.

Kiørboe, T., Sabatini, M., 1995. Scaling of fecundity, growth and development in marine planktonic copepods. Marine Ecology Progress Series 120, 285–298.

Knoechel, R., Holtby, L.B., 1986. Construction and validation of a body-length-based model for the prediction of cladoceran community filtering rates. Limnology and Oceanography 31 (1), 1–16.

Peters, R.H., Downing, J.A., 1984. Empirical analysis of zooplankton filtering and feeding rates. Limnology and Oceanography 29 (4), 763–784.

Shimoda, Y., Yerubandi, R., Watson, S., Arhonditsis, G.B., 2016. Optimizing the complexity of phytoplankton functional group modelling: An allometric approach. Ecological Informatics 31, 1–17.
Wen, Y.H., Peters, R.H., 1994. Empirical models of phosphorus and nitrogen excretion rates by zooplankton. Limnology and Oceanography 39 (7), 1669–1679.

Further Reading

Allen, A.P., Gillooly, J.F., 2009. Toward an integration of ecological stoichiometry and the metabolic theory of ecology to better understand nutrient cycling. Ecology Letters 12, 369–384.
Belgrano, A., Allen, A.P., Enquist, B.J., Gillooly, J.F., 2002. Allometric scaling of maximum population density: A common rule for marine phytoplankton and terrestrial plants. Ecology Letters 5 (5), 611–613.
Brose, U., 2010. Body-mass constraints on foraging behavior determine population and food-web dynamics. Functional Ecology 24, 28–34.
Brown, J.H., Gillooly, J.F., Allen, A.P., Savage, V.M., West, G.B., 2004. Toward a metabolic theory of ecology. Ecology 85 (7), 1771–1789.
Brown, J.H., West, G.B., Enquist, B.J., 2000. Patterns and processes, causes and consequences. In: Brown, J.H., West, G.B. (Eds.), Scaling in biology. New York, USA: Oxford University Press, pp. 1–24.
da Silva, J.K.L., Garcia, G.J.M., Barbosa, L.A., 2006. Allometric scaling laws of metabolism. Physics of Life Reviews 3 (4), 229–261.
Glazier, D.S., 2005. Beyond the '3/4-power law': Variationin the intra- and interspecific scaling ofmetabolic rate in animals. Biological Reviews 80 (4), 611–662.
Hildrew, A.G., Raffaelli, D.G., Edmonds-Brown, R. (Eds.), 2007. Body size: The structure and function of aquatic ecosystems. Cambridge, UK: Cambridge University Press.
Hillebrand, H., Borer, E.T., Bracken, M.E.S., Cardinale, B.J., Cebrian, J., Cleland, E.E., Elser, J.J., Gruner, D.S., Harpole, W.S., Ngai, J.T., Sandin, S., Seabloom, E.W., Shurin, J.B., Smith, J.E., Smith, M.D., 2009. Herbivore metabolism and stoichiometry each constrain herbivory at different organizational scales across ecosystems. Ecology Letters 12, 516–527.
Kalinkat, G., Schneider, F.D., Digel, C., Guill, C., Rall, B.C., Brose, U., 2013. Body masses, functional responses and predator–prey stability. Ecology Letters 16, 1126–1134.
Morgan, D.K.J., Hicks, B.J., 2013. A metabolic theory of ecology applied to temperature and mass dependence of N and P excretion by common carp. Hydrobiologia 705 (1), 135–145.
Peters, R.H., 1983. The ecological implications of body size. Cambridge, UK: Cambridge University Press.
Rall, B.C., Kalinkat, G., Ott, D., Vucic-Pestic, O., Brose, U., 2011. Taxonomic versus allometric constraints on non-linear interaction strengths. Oikos 120, 483–492.

Ammonification

Nicolas Romillac, Université de Lorraine, Laboratoire Agronomie et Environnement, Vandoeuvre Cedex, France

Nomenclature

DON Dissolved organic nitrogen

MIT Mineralization immobilization turnover

µmol L^{-1}
Micromole per liter

µgN-NH_4^+ h^{-1} g dry $soil^{-1}$
Micrograms of ammonium nitrogen produced per hour per gram of dry soil

Glossary

Aquaporin Membrane proteins facilitating the transport of water through the membrane.

Deamination Reaction of liberation of an amino group ($-NH_2$) from an organic molecule.

Dissimilatory nitrate reduction Biological reduction of nitrate (NO_3^-) to ammonium (NH_4^+).

Nitrification Biological oxidation of ammonium (NH_4^+) to nitrate (NO_3^-).

Rhizodeposition Release of organic and mineral compounds by plant roots in the soil (or other environment).

Transamination Reaction of transfer of an amino group ($-NH_2$) from an amino acid to a ketoacid.

Introduction

Ammonification is the process by which microorganisms present in soil, sediment, or water mineralize low molecular weight, dissolved, organic molecules presenting amine or amide groups (of general formula R-NH_2) and produce ammonium (NH_4^+). Ammonification is the last step of the nitrogen cycle involving an organic compound, and is the intermediary step between the depolymerization of large organic molecules and the nitrification step (**Fig. 1**).

In marine ecology, ammonification is also referred to as ammonium regeneration and ammonium recycling. The term "nitrate ammonification" is sometimes used to refer to the dissimilatory reduction of nitrate to ammonium (e.g., Rysgaard et al., 1996). Nitrate ammonification is beyond the scope of this article.

One should distinguish between gross ammonification, the amount of mineralized organic molecules, ammonification activity, the activity of microbial enzymes responsible for ammonification in a given environment, and net ammonification, the release of NH_4^+ in the environment. While gross ammonification and ammonification activity are well correlated, they are not necessarily correlated to net ammonification, as net ammonification is the result of the co-occurring processes of ammonium uptake and release from microbial cells after ammonification.

Dissolved Organic Nitrogen as a Substrate for Ammonification

Dissolved organic nitrogen (DON), in soils, sediments, seawater, and freshwater, is a mixture of diverse molecules, of which a great part is of unknown chemical structure. DON is composed of both labile and recalcitrant, high molecular weight and low molecular weight molecules (Neff et al., 2003; Jones et al., 2004a).

Only the low molecular weight molecules can be the substrates of ammonification. They include amino acids, urea, amino sugars, and nucleotides. Of these substrates, amino acids are the best studied, and are generally considered as the model substrates for ammonification. An important characteristic of amino acids is their solubility in water or in soil solution, which make them an easily accessible source of carbon and nitrogen for microorganisms (Jones and Kielland, 2012).

Amino acids originate from various sources. They can be released from cells after cell lysis (Miltner et al., 2009), rhizodeposited by plant roots (Jones et al., 2004b), or released by microorganisms in order to adapt their osmotic pressure to a fluctuating

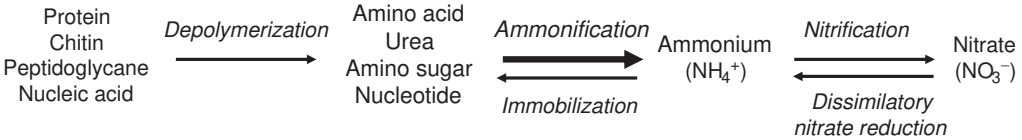


Fig. 1 Ammonification positioned in the nitrogen cycle.

 doi:10.1016/B978-0-12-409548-9.10889-9

environment (Halverson et al., 2000). In water, they can be trapped by humic substances and subsequently released by the action of UV, in a photochemical reaction (Moran and Zepp, 1997).

Free amino acids in soil solution, sea, and freshwater are relatively scarce, as they are rapidly uptaken or mineralized by plants and microorganisms. As an example, Jones et al. (2005) found a half-life of 3.5 h for amino acids in agricultural soils, and Fuhrman and Ferguson (1986) found a turnover of 6 h in winter and 1 h in summer for amino acids in temperate seawaters. In soil solution, amino acids represent 10%–40% of DON and their concentrations range from 0.1 to 50 $\mu mol\ L^{-1}$ (Jones et al., 2002). Amino acids represent 0.3% of the dissolved organic matter in the equatorial Pacific Ocean (Lee and Bada, 1975), and concentrations inferior to 1 $\mu mol\ L^{-1}$ are frequently observed in oceans and estuaries water (Lee and Bada, 1975; Fuhrman and Bell, 1985; Fuhrman and Ferguson, 1986; Coffin, 1989; Pomeroy et al., 1990).

Urea originates from mammal urine and from excretion of uric acid by birds, reptiles, or arthropods. It can be excreted by heterotrophic bacteria, zooplankton, and phytoplankton. It is also a product of the degradation of the amino acid arginine, and a major fertilizer. Urea can move from ecosystems to ecosystems, and thus can leach from soils to fresh and coastal waters, be released in the water from sediments, and originate from atmospheric deposition (Glibert et al., 2006; Geisseler et al., 2010; Solomon et al., 2010).

Amino acids, amino sugars, and nucleotides can also be released from the depolymerization of macromolecules (respectively, proteins and peptides, chitin and peptidoglycane, and nucleic acids) by the action of microbial enzymes (proteases and peptidases, chitinases and *N*-acetylglucosamidase, and nucleases) (Jones et al., 2004a; Roberts and Jones, 2012; Isobe and Ohte, 2014).

Biochemical Pathways of Ammonification

Ammonification can be either extracellular or intracellular (Geisseler et al., 2010). In the extracellular pathway, the breakdown of organic compounds is realized by free extracellular enzymes or by enzymes bound to the outer cell membrane. The released ammonium is subsequently uptaken by the microbial cells. This is classically called the mineralization immobilization turnover (MIT) route. Extracellular enzymes involved in this process are poorly known, with the exception of urease (EC 3.5.1.5) and amino acid oxidases (EC 1.4.3.2).

Urease catalyzes the formation of ammonia (NH_3) and carbamate. Carbamate spontaneously decomposes to NH_3 and carbonic acid. Ammonia then dissolves to form ammonium (NH_4^+), in an equilibrium between ammonia and ammonium controlled by pH.

Microbes can synthetize urease constitutively, but its synthesis is more often regulated by N availability. In this case, the synthesis is repressed by ammonium and activated by presence of urea and by N starvation (Mobley et al., 1995).

Approximately 45%–60% of the soil urease activity is extracellular, while the remaining activity is intracellular (Geisseler et al., 2010). Extracellular ureases are probably bound to soil organic matter, which protects them from degradation by soil proteases. The greatest part of the extracellular ureases in soil appears to derive from the release of intracellular ureases by cell lysis and not from a secretion by living organisms.

Amino acid oxidases are bound to the outer cell membrane of microorganisms and catalyze the deamination of amino acids and produce ketoacids, NH_3 and H_2O_2. They have either broad or strict substrate specificity. In soils, enzymes with a broad substrate specificity seem to be dominant (Nuutinen and Timonen, 2008). Their synthesis is repressed by ammonium, and induced by N starvation and by amino acids in the presence of a carbon source.

On the other hand, microorganisms can also uptake low molecular weight organic compounds, such as amino acids, and mineralize them to ammonium inside the cell. Ammonium can be subsequently excreted from the cell, if in excess. This is called the direct route. This metabolic pathway includes an active transport through the membrane and deamination or transamination of the molecule inside the cell, in order to release ammoniac and ammonium.

In the case of amino acids, the transport systems are located in the cytoplasmic membrane and are usually specific to a group of structurally similar amino acids. Bacteria possess relatively specific transport systems, and up to 12 different transport systems can be found in a cell. Fungi have transport systems with a broader specificity and present fewer different transporters than bacteria (Geisseler et al., 2010). Some transport systems can be constitutively synthetized, but the majority of the systems are regulated by N availability. Ammonium and intracellular amino acids repress the synthesis while starvation of C, N, or S induces the synthesis.

Furthermore, microbes can possess active transport systems for small peptides, up to 600 Da, corresponding to a peptide made of five or six amino acids (Payne and Smith, 1994).

Transport systems for amino sugars are less well known. They are repressed by glucose and induced by their substrate (Geisseler et al., 2010).

In marine waters, urea uptake can contribute to more than 50% of N uptake by phytoplankton (Glibert et al., 2006). In a cell, high affinity active transporters and low affinity passive transporters for urea can coexist. Urea can also enter a cell through aquaporines and urea/amides channels. The synthesis of transport systems for urea is induced by N starvation and is repressed by ammonium (Solomon et al., 2010).

Once in the cell, ammonium can be released from the organic compounds by a great variety of enzymes pertaining to the hydrolases and lyases groups. These include, in the super-family of the amidohydrolases, L-asparaginase (EC 3.5.1.1; substrate: asparagine), L-glutaminase (EC 3.5.1.2; substrate: glutamine), amidase (EC 3.5.1.4; substrate: monocarboxylic acid amides), and urease. The groups also include the hydrolases arylamidase (EC 3.4.11.2; substrate: amides and arylamides), L-arginine deiminase

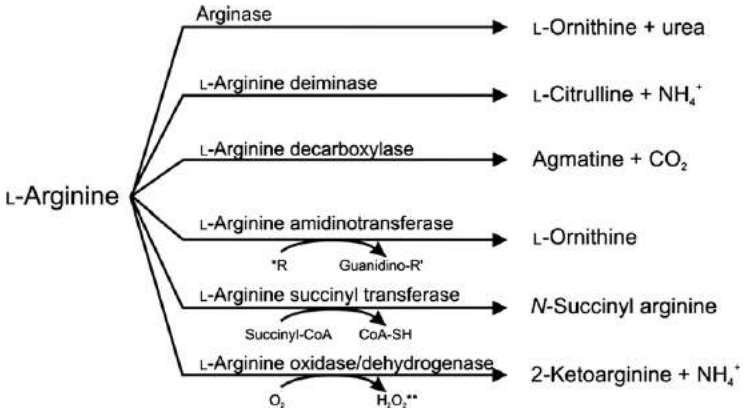


Fig. 2 Pathways of L-arginine degradation in bacterial cells. The arginase, the L-arginase deiminase, the L-arginine oxidase/dehydrogenase and, putatively, the L-arginine decarboxylase pathways, lead to the release of ammonium inside bacterial cells. Reproduced from Schriek, S., Rückert, C., Staiger, D., Pistorius, E. K. and Michel, K. (2007). Bioinformatic evaluation of L-arginine catabolic pathways in 24 cyanobacteria and transcriptional analysis of genes encoding enzymes of L-arginine catabolism in the cyanobacterium Synechocystis sp. PCC 6803. *BMC Genomics* **28**, 1–28.

(EC 3.5.3.6; substrate: arginine) and arginase (EC 3.5.3.1; substrate: arginine), and the lyases L-aspartase (EC 4.3.1.1; substrate: aspartate) and L-histidase (EC 4.3.1.3; substrate: histidine). As an example, at least six pathways are known for the degradation of the arginine, of which at least three lead to the release of ammonium (**Fig. 2**; Schriek et al., 2007).

Microbial Choice Between MIT and Direct Route

The direct route presents the advantage of fulfilling the C and N requirements of the microorganisms at the same time, as the carbon skeleton of the amino acid can be metabolized after deamination, and serve as a source of energy. Amino acids can also be directly incorporated into microbial proteins. On the other hand, the direct route requires the synthetizing of a diversity of amino acid transporters (while ammonium transporters are more likely to be present already in a cell), amino acids profile into the environment does not necessarily matches the microbial needs, and the deamination process can create toxic byproducts. The production of H_2O_2 in the MIT route can also have a role in suppressing other microbial strains. All those elements can favor the MIT route.

Geisseler et al. (2010) developed a conceptual model in order to understand which route will be predominant in soils. According to their analysis, based on the mechanisms of regulation of the two pathways, the MIT route should be favored over the direct route (**Fig. 3**):

- when the N availability is high relatively to the C availability;
- when the availability of ammonium is high relatively to other sources of nitrogen (i.e., amino acids and amino sugars); or
- when the availability of carbohydrates is high relatively to organic compounds rich in nitrogen (amino acids and amino sugars).

Net Ammonification

The net ammonification, that is, the release of ammonium in the extracellular environment, is not directly correlated to gross ammonification, as it is influenced by the degree of retention, uptake, and excretion of ammonium by microbial cells after ammonification. The balance between retention/uptake and excretion of ammonium is likely to be influenced by the N nutritional status of the microbial community (starvation or nitrogen excess), the C:N ratio of the microbial population, and the quantity and chemical composition of dissolved organic matter (Jones et al., 2004a). Typically, bacteria will excrete N for a C:N of the dissolved organic N inferior to 12.5 and fungi for a C:N inferior to 30.3 (Hodge et al., 2000).

Organisms Responsible for Ammonification

It is generally considered that the majority of, if not all, microorganisms (bacteria and fungi) are capable of ammonification. Ammonification enzymes and transport systems are shared by widely diverse microbial species. As an example, more than 50 strains of soil bacteria are capable of arginine ammonification (Alef and Kleiner, 1986). Most heterotrophic bacteria are able to uptake amino acids, as well as a few species pertaining to phytoplankton (Fuhrman and Azam, 1982; Jones et al., 2004a). A wide number of marine cyanobacteria and dinoflagellates, as well as some heterokont unicellular algae, are also able to uptake urea (Glibert et al., 2006). Urease is produced not only by bacteria and fungi, including nitrogen-fixing bacteria colonizing legume nodules (Chuntanom and Pongsilp, 2011), but also by plants and unicellular algae (Solomon et al., 2010; Geisseler et al., 2010).

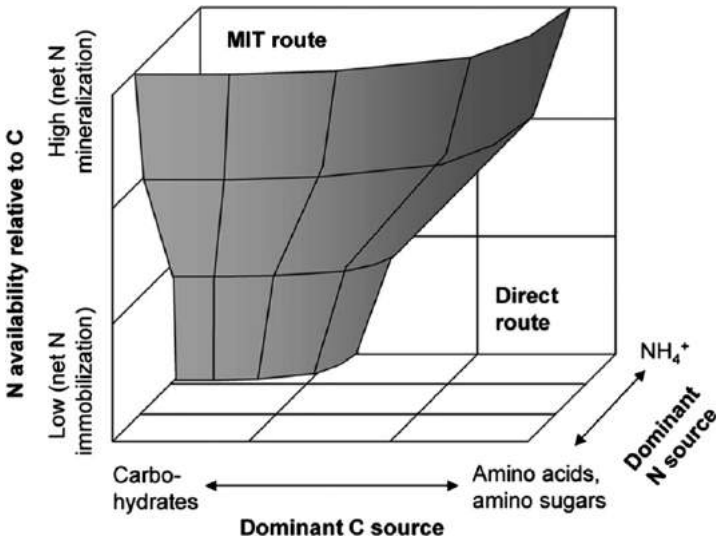


Fig. 3 Conceptual model of the factors affecting the relative importance of the microbial immobilization turnover and the direct route of N uptake by soil microorganisms. Reproduced from Geisseler, D., Horwath, W. R., Joergensen, R. G. and Ludwig, B. (2010). Pathways of nitrogen utilization by soil microorganisms – A review. *Soil Biology and Biochemistry* **42**, 2058–2067. Available at: https://doi.org/10.1016/j.soilbio.2010.08.021, with permission from Elsevier.

Amino acid oxidases are produced by a wide diversity of bacteria and fungi (in both ascomycetes and basidiomycetes phyla, and including some mycorrhiza), found in soil and water, and by some species of phytoplankton (Pantoja and Lee, 1994; Geisseler et al., 2010). This is in contrast, for example, with the enzymes involved in the steps of depolymerization of the nitrogen cycles (proteases, peptidases, nucleases, etc.) which are produced by phylogenetically restrained microbial groups.

Methods to Study Ammonification

Several methods to measure potential ammonification activity have been developed, using L-amino acids as model substrates. The method developed by Alef and Kleiner (1986), based on the arginine amino acid, is the most commonly used, while methods based on histidine (Burton and McGill, 1992), asparagine (Frankenberger and Tabatabai, 1991a), glutamine (Frankenberger and Tabatabai, 1991b), or aspartase (Senwo and Tabatabai, 1996) are also available. Briefly, the arginine method consists of adding a saturating concentration of arginine to a soil slurry, and conducting a short incubation (1–2 h) at normal soil pH and at 22°C. At the end of incubation, soil mineral nitrogen is extracted, using an extractant such as KCl, and ammonium abundance is measured in the extracts. Methods based on other amino acids have similar protocols. Arylamidase, amidase, and urease activity are measured in the same way, using L-leucine, formamide, and urea, respectively, as the substrate (Acosta-Martínez and Tabatabai, 2001; Miller et al., 2001; O'Toole, 1991).

These methods measure the potential activity, that is, the activity of all enzymes in optimal conditions with unlimited substrate amount. However, arginine ammonification is considered a reliable indicator of ammonification, as it is well correlated to the gross N mineralization, measured by isotopic methods (Bonde et al., 2001).

Among other methods, the gross ammonification flux can be measured by isotopic methods using ammonium enriched in ^{15}N (Rosenkranz et al., 2006). Culture media, containing peptone, can be used to count the number of cultivable ammonifying organisms in environmental samples (Chuntanom and Pongsilp, 2011).

Importance of Ammonification for the Nitrogen Cycle

Overall, ammonification does not appear to be a limiting step of the nitrogen cycle in soils. Indeed, the rate of ammonification appears to depend on the rate of depolymerization of macromolecules such as proteins, chitin, and nucleic acids. Depolymerization is probably the limiting step of the nitrogen cycle (Jan et al., 2009). This is consistent with the rapid mineralization of amino acids in agricultural topsoils (average half-life 3.5 h; Jones et al., 2005), in boreal forest soils (half-life 1–6 h; Kielland et al., 2007), and in marine waters (turnover of 1–6 h; Fuhrman and Ferguson, 1986).

The ammonium produced by ammonification is the substrate of the nitrification process, probably the most studied step of the nitrogen cycle. As a consequence, the rate of ammonification tightly controls the rate of nitrification in estuaries and oceans (Ward, 2008; Flindt et al., 1999), and, even if the evidence is scarcer, in some soils (e.g., Papen and Butterbach-Bahl, 1999).

Sensitivity to Environmental Factors

Because of the widespread ability to perform ammonification among bacteria and fungi, ammonification activity does not depend on the composition or diversity of the microbial community. However, it depends on the activity and abundance of the whole community. Arginine ammonification activity is well correlated to indicators of microbial activity, such as respiration (Alef and Kleiner, 1987), dehydrogenase, protease, phosphatase, and β-glucosidase activities (Graham and Haynes, 2005; Dilly et al., 2007; Stark et al., 2008), and indicators of microbial abundance such as microbial biomass C (Lin and Brookes, 1999) or microbial biomass N (Pajares et al., 2009). Then, all factors affecting microbial activity, microbial abundance, and C and N availability should modify the ammonification activity. Furthermore, every perturbation destroying microbial biomass, such as freeze-thawing, drying-rewetting, or grazing by predators, is susceptible to release amino acids from lyzed cells, thus increasing the rate of ammonification (Schimel and Bennett, 2004).

Arginine ammonification is the standard method to measure ammonification activity. As the measurements of arginine ammonification are scarce in waters and sediments, below is a summary of knowledge on the response of soil arginine ammonification to environmental factors, particularly in agricultural soils. In soil, typical values for arginine ammonification range from 1 to 5 $\mu gN\text{-}NH_4^+\ h^{-1}$ g dry $soil^{-1}$ (Dilly et al., 2007; Stark et al., 2007; Pajares et al., 2009; Meisner et al., 2011; Romillac et al., 2015), with extreme values up to 15 $\mu gN\text{-}NH_4^+\ h^{-1}$ g dry $soil^{-1}$ under a lupin crop (Stark et al., 2008).

As ammonification is largely determined by microbial activity and abundance, which are themselves determined by substrate availability, arginine ammonification is correlated to soil organic carbon content (Lin and Brookes, 1999; Pajares et al., 2009; Nourbakhsh and Alinejadian, 2009). Thus, all factors influencing soil organic matter content are likely to modify arginine ammonification. Arginine ammonification is higher:

- under permanent grasslands than under temporary grassland;
- in grasslands than in cultivated fields (Milne and Haynes, 2004);
- under reduced tillage than under conventional tillage (Van Capelle et al., 2012);
- under organic fertilization than under mineral fertilization (Mercik et al., 1995);
- under plants than in bare soil (Dinesh et al., 2004).

The activity also decrease with soil depth (Kandeler et al., 1994) and several studies also found higher activity in organic agriculture fields, compared to conventional agriculture (Haynes and Tregurtha, 1999; Gunapala and Scow, 1998; Stark et al., 2008). All these variations of activity can probably be explained by variations of soil carbon content and microbial biomass size.

Arginine ammonification can also be influenced by the plant species (Dinesh et al., 2004) and the climate. A higher activity has been observed under legume litters than under cereal litters, which can be explained by the higher amino acid content of the legume litter (Dilly et al., 2007). The activity decreases linearly with soil water potential (Chen et al., 2011a, b; Tietema et al., 1992; Low et al., 1997), and is thus higher in humid soils and after rainfalls and lower in summer months (Bonde et al., 2001).

Among other enzymes involved in ammonification, amidase, L-asparaginase, L-aspartase, and L-glutaminase activities are positively correlated to microbial biomass C and N (Dodor and Tabatabai, 2003). Amidase and urease activities are higher in grassland than in cultivated fields. Urease activity is higher in soils under organic fertilization and in natural soils, while amidase is higher in soils under mineral N fertilization (Gianfreda and Ruggiero, 2006). In general, those enzymes' activities increase with liming and decrease with soil acidification; L-glutaminase is particularly sensitive to pH and L-aspartase is far less sensitive (Acosta-Martínez and Tabatabai, 2001).

Consequences for Pollution

Ammonium released from ammonification in soils is retained by the soil constituents, preventing its leaching to water. However, it is readily available for nitrification, which quickly transforms ammonium to nitrate, a nitrogen form highly susceptible to being leached. Ammonium can also be volatilized as ammonia (NH_3), a major source of nitrogen pollution and eutrophication through atmospheric deposition, as well as a source of air pollution through the formation of particulate matter. In agricultural systems where urea or animal dejections are used as fertilizer, a significant part of the ammonium produced by the urease activity is lost to the atmosphere by volatilization (Bussink and Oenema, 1998). Habitually, 20%–30% of the N applied as urea is volatilized (Chadwick et al., 2005). Urease activity and volatilization are also of concern during the storage and composting of manure. This has led to important research to find urease inhibitors applicable in the field. Nowadays, some commercialized urea fertilizers contain urea inhibitors in order to prevent this phenomenon by slowing down the rate of ammonification.

Importance for Plant Nutrition

Ammonification produces ammonium and probably controls the rate of nitrate production by nitrification. It is thus likely to be an important process for plant nutrition. However, ammonification has been poorly studied in this context, probably because this activity is widely distributed among microorganisms and because its rate is controlled by upstream depolymerization.

Plants preferentially uptake N under the nitrate form, but are also able to uptake ammonium and amino acids. Thus, ammonifying microorganisms can be both competitors and providers of nitrogen for plants. The role they will play will depend on the importance of the different forms of nitrogen in soil.

In soils of the tundra, taiga, and alpine ecosystems, where mineral nitrogen is present in low amounts, amino acids are usually the main source of nitrogen for plants (Jones and Kielland, 2012). In soils of the wetlands, deserts, and some temperate or boreal forests, ammonium can be an important source of nitrogen for plants, in addition to amino acids (Schimel and Bennett, 2004). In these conditions, intense competition for N should occur between plants and ammonifying microorganisms. In theory, microorganisms should win competition due to faster growth, higher surface-to-volume ratio, and higher substrate affinities than plants (Hodge et al., 2000). However, plants can obtain N due to the presence of N-rich microsites where nitrogen mineralization is higher than nitrogen retention by microorganisms. Nitrogen can then leak from those microsites and be captured by plants. Another mechanism explaining the success of plants in the face of microbial competition is the longer lifespan of roots compared to microorganisms. The high turnover of microbial cells, which die from predation, starvation, freeze-thawing, or drying-rewetting, releases nitrogen, making it available for plants. Nitrogen acquisition via mycorrhiza can also permit plants to obtain more nitrogen (Hodge et al., 2000; Schimel and Bennett, 2004).

In soils where substantial amounts of nitrate are present, such as agricultural soils and tropical forests, plants are hypothesized to uptake preferentially nitrate; thus ammonifying microorganisms are no longer competitors, but providers of mineral nitrogen (Schimel and Bennett, 2004). In some cases at least, plants growing in those conditions can increase ammonification activity at times when their N requirements are the highest (Romillac et al., 2015).

Apart from competition, plants and ammonifying microorganisms can maintain a relationship of the mutualistic type: plants rhizodeposit significant amounts of amino acids (Jones et al., 2004b), which are substrates for ammonification and can help microorganisms to fulfill their carbon and nitrogen requirements. Rhizodeposition increases the abundance and activity of microorganisms, resulting in higher ammonification and nitrogen mineralization rates, which prove beneficial for plants' nutrition. This is called the "rhizosphere priming effect." The detailed mechanisms of the rhizosphere priming effect are yet to be discussed (Dijkstra et al., 2013; Cheng et al., 2014). In the microbial mining hypothesis, the rhizosphere priming effect results from an increase in the production of extracellular enzymes responsible for depolymerization, the step ahead of ammonification in the nitrogen cycle (Fontaine et al., 2011). In the microbial loop hypothesis, the increasing predation by protozoa and nematodes on the increased microbial biomass accelerates the turnover of microbial organic nitrogen and liberates ammonium available for plants (Clarholm, 1994). Thus, the link between plant rhizodeposition and ammonification rate would be an indirect one.

Conclusion

Ammonification activity is a widely shared activity among bacteria, fungi, and even unicellular algae, making ammonification a quick and non-limiting step of the nitrogen cycle. Therefore, ammonification rate is mainly controlled by the size and activity of the microbial community, and by the availability of DON. It responds quickly to the modification of these parameters, and in particular to modifications of the rate of depolymerization of macromolecules such as proteins, chitin, and nucleic acids.

See also: Aquatic Ecology: Acidification in Aquatic Systems. Ecological Processes: Acidification; Decomposition and Mineralization; Nitrification. Global Change Ecology: Nitrogen Cycle

References

Acosta-Martínez, V., Tabatabai, M.A., 2001. Arylamidase activity in soils: Effect of trace elements and relationships to soil properties and activities of amidohydrolases. Soil Biology and Biochemistry 33, 17–23. Available at:.http://www.sciencedirect.com/science/article/pii/S0038071700001097

Alef, K., Kleiner, D., 1986. Arginine ammonification, a simple method to estimate microbial activity potentials in soils. Soil Biology and Biochemistry 18, 233–235.

Alef, K., Kleiner, D., 1987. Applicability of arginine ammonification as indicator of microbial activity in different soils. Biology and Fertility of Soils 5, 148–151.

Bonde, T.A., Nielsen, T.H., Miller, M., Sørensen, J., 2001. Arginine ammonification assay as a rapid index of gross N mineralization in agricultural soils. Biology and Fertility of Soils 34, 179–184.

Burton, D.L., McGill, W.B., 1992. Spatial and temporal fluctuation in biomass, nitrogen mineralizing reactions and mineral nitrogen in a soil cropped to barley. Canadian Journal of Soil Science 72, 31–42.

Bussink, D.W., Oenema, O., 1998. Ammonia volatilization from dairy farming systems in temperate areas: A review. Nutrient Cycling in Agroecosystems 51, 19–33.

Chadwick, D., Misselbrook, T., Gilhespy, S. et al. (2005). Ammonia emissions and crop N use efficiency. Department for Environment, Food and Rural Affairs, United Kingdom.

Chen, Y., Borken, W., Stange, F.C., Matzner, E., 2011a. Effects of decreasing water potential on gross ammonification and nitrification in an acid coniferous forest soil. Soil Biology and Biochemistry 43, 333–338. Available at.https://doi.org/10.1016/j.soilbio.2010.10.020

Chen, Z.H., Chen, L.J., Zhang, Y.L., Wu, Z.J., 2011b. Microbial properties, enzyme activities and the persistence of exogenous proteins in soil under consecutive cultivation of transgenic cottons (Gossypium hirsutum L.). Plant and Soil 57, 67–74.

Cheng, W., Parton, W.J., Gonzalez-Meler, M.A., *et al.*, 2014. Synthesis and modeling perspectives of rhizosphere priming. New Phytologist 201, 31–44.

Chuntanom, S., Pongsilp, N., 2011. Environmental parameters affecting urease production and ammonification in Phaseolus vulgaris-nodulating rhizobia and Vigna radiata-nodulating rhizobia. International Journal of Microbiological Research 2, 222–232.

Clarholm, M., 1994. The microbial loop in soil. In: Ritz, K., Dighton, J., Giller, K.E. (Eds.), Beyond the biomass. London: Wiley-Sayce.

Coffin, R.B., 1989. Bacterial uptake of dissolved free and combined amino acids in estuarine waters. Limnology and Oceanography 34, 531–542.

Dijkstra, F.A., Carrillo, Y., Pendall, E., Morgan, J.A., 2013. Rhizosphere priming: A nutrient perspective. Frontiers in Microbiology 4, 1–8.

Dilly, O., Munch, J.C., Pfeiffer, E., 2007. Enzyme activities and litter decomposition in agricultural soils in northern, central, and southern Germany. Journal of Plant Nutrition and Soil Science 170, 197–204.
Dinesh, R., Suryanarayana, M.A., Chaudhuri, S.G., Sheeja, T.E., 2004. Long-term influence of leguminous cover crops on the biochemical properties of a sandy clay loam Fluventic Sulfaquent in a humid tropical region of India. Soil and Tillage Research 77, 69–77.
Dodor, D.E., Tabatabai, M.A., 2003. Effect of cropping systems on phosphatases in soils. Journal of Plant Nutrition and Soil Science 166, 7–13.
Flindt, M.R., Pardal, M.Â., Lillebø, A.I., Martins, I., Marques, J.C., 1999. Nutrient cycling and plant dynamics in estuaries: A brief review. Acta Oecologica 20, 237–248.
Fontaine, S., Henault, C., Aamor, A., 2011. Fungi mediate long term sequestration of carbon and nitrogen in soil through their priming effect. Soil Biology and Biochemistry 43, 86–96.
Frankenberger, W.T., Tabatabai, M.A., 1991a. L-Asparaginase activity of soils. Biology and Fertility of Soils 11, 6–12. Available at.https://doi.org/10.1007/BF00335826
Frankenberger, W.T., Tabatabai, M.A., 1991b. L-Glutaminase activity of soils. Soil Biology and Biochemistry 23, 869–874. Available at.http://www.sciencedirect.com/science/article/pii/0038071791900996
Fuhrman, J.A., Azam, F., 1982. Thymidine incorporation as a measure of heterotrophic bacterioplankton production in marine surface waters: Evaluation and field results. Marine Biology 66, 109–120.
Fuhrman, J.A., Bell, T.M., 1985. Biological considerations in the measurement of dissolved free amino acids in seawater and implications for chemical and microbiological studies. Marine Ecology Progress Series 25, 13–21.
Fuhrman, J.A., Ferguson, R.L., 1986. Nanomolar concentrations and rapid turnover of dissolved free amino acids in seawater: Agreement between chemical and microbiological measurements. Marine Ecology Progress Series 33, 237–242.
Geisseler, D., Horwath, W.R., Joergensen, R.G., Ludwig, B., 2010. Pathways of nitrogen utilization by soil microorganisms – A review. Soil Biology and Biochemistry 42, 2058–2067. Available at.https://doi.org/10.1016/j.soilbio.2010.08.021
Gianfreda, L., Ruggiero, P., 2006. Enzyme activities in soil. In: Nannipieri, P., Smalla, K. (Eds.), Nucleic acids and proteins in soil. Berlin, Heidelberg: Springer Verlag, pp. 257–312.
Glibert, P.M., Harrison, J., Heil, C., Seitzinger, S., 2006. Escalating worldwide use of urea – A global change contributing to coastal eutrophication. Biogeochemistry 77, 441–463.
Graham, R.J., Haynes, M.H., 2005. Organic matter accumulation and fertilizer-induced acidification interact to affect soil microbial and enzyme activity on a long-term sugarcane management experiment. Biology and Fertility of Soils 41, 249–256.
Gunapala, N., Scow, K.M., 1998. Dynamics of soil microbial biomass and activity in conventional and organic farming. Soil Biology and Biochemistry 30, 805–819.
Halverson, L., Jones, T., Firestone, M., 2000. Release of intracellular solutes by four soil bacteria exposed to dilution stress. Soil Science Society of America Journal 64, 1630–1637.
Haynes, R.J., Tregurtha, R., 1999. Effects of increasing periods under intensive arable vegetable production on biological, chemical and physical indices of soil quality. Biology and Fertility of Soils 28, 259–266.
Hodge, A., Robinson, D., Fitter, A., 2000. Are microorganisms more effective than plants at competing for nitrogen? Trends in Plant Science 5, 304–308.
Isobe, K., Ohte, N., 2014. Ecological perspectives on microbes involved in N-cycling. Microbes and Environments 29, 4–16.
Jan, M.T., Roberts, P., Tonheim, S.K., Jones, D.L., 2009. Protein breakdown represents a major bottleneck in nitrogen cycling in grassland soils. Soil Biology and Biochemistry 41, 2272–2282. Available at.http://linkinghub.elsevier.com/retrieve/pii/S0038071709002958 (accessed August 27, 2011).
Jones, D.L., Kielland, K., 2012. Amino acid, peptide and protein mineralization dynamics in a taiga forest soil. Soil Biology and Biochemistry 55, 60–69. Available at.https://doi.org/10.1016/j.soilbio.2012.06.005
Jones, D.L., Owen, A.G., Farrar, J.F., 2002. Simple method to enable the high resolution determination of total free amino acids in soil solutions and soil extracts. Soil Biology and Biochemistry 34, 1893–1902.
Jones, D.L., Shannon, D., Murphy, D.V., Farrar, J., 2004a. Role of dissolved organic nitrogen (DON) in soil N cycling in grassland soils. Soil Biology and Biochemistry 36, 749–756.
Jones, D.L., Hodge, A., Kuzyakov, Y., 2004b. Plant and mycorrhizal regulation of rhizodeposition. New Phytologist 163, 459–480.
Jones, D.L., Kemmitt, S.J., Wright, D., 2005. Rapid intrinsic rates of amino acid biodegradation in soils are unaffected by agricultural management strategy. Soil Biology and Biochemistry 37, 1267–1275.
Kandeler, E., Eder, G., Sobotik, M., 1994. Microbial biomass, N mineralization, and the activities of various enzymes in relation to nitrate leaching and root distribution in a slurry-amended grassland. Biology and Fertility of Soils 18, 7–12.
Kielland, K., McFarland, J.W., Ruess, R.W., Olson, K., 2007. Rapid cycling of organic nitrogen in taiga forest ecosystems. Ecosystems 10, 360–368.
Lee, C., Bada, J.L., 1975. Amino acids in equatorial Pacific Ocean water. Earth and Planetary Science Letters 26, 61–68.
Lin, Q., Brookes, P.C., 1999. Arginine ammonification as a method to estimate soil microbial biomass and microbial community structure. Soil Biology and Biochemistry 31, 1985–1997.
Low, A.P., Stark, J.M., Dudley, L.M., 1997. Effect of soil osmotic potential on nitrification, ammonification, N-assimilation, and nitrous oxide production. Soil Science 162, 16–27.
Meisner, A., de Boer, W., Verhoeven, K.J.F., Boschker, H.T.S., Van Der Putten, W.H., 2011. Comparison of nutrient acquisition in exotic plant species and congeneric natives. Journal of Ecology 99, 1308–1315.
Mercik, S., Korschens, M., Bielawski, W., Russel, S., Rumpel, J., 1995. Ammonification, nitrification activity and soil respiration intensity as affected by long-term fertilization and soil type. Annals of Warsaw Agricultural University, Agriculture 29, 27–35.
Miller, M., Nielsen, T.H., Bonde, T., Sørensen, J., 2001. Fluctuations of formamide hydrolase activity in a barley field soil after drying and rewetting. Applied Soil Ecology 16, 159–167. Available at.http://www.sciencedirect.com/science/article/pii/S0929139300001049
Milne, R.M., Haynes, R.J., 2004. Soil organic matter, microbial properties, and aggregate stability under annual and perennial pastures. Biology and Fertility of Soils 39, 172–178.
Miltner, A., Kindler, R., Knicker, H., Richnow, H., Kästner, M., 2009. Fate of microbial biomass-derived amino acids in soil and their contribution to soil organic matter. Organic Geochemistry 40, 978–985. Available at.https://doi.org/10.1016/j.orggeochem.2009.06.008
Mobley, H.L.T., Island, M.D., Hausinger, R.P., 1995. Molecular biology of microbial ureases. Microbiological Reviews 59, 451–480.
Moran, M.A., Zepp, R.G., 1997. Role of photoreactions in the formation of biologically labile compounds from dissolved organic matter. Limnology and Oceanography 42, 1307–1316.
Neff, J.C., Chapin III, F.S., Vitousek, P.M., 2003. Breaks in the cycle: Dissolved organic nitrogen in terrestrial ecosystems. Frontiers in Ecology and the Environment 1, 205–211.
Nourbakhsh, F., Alinejadian, A., 2009. Arginine ammonification and L-glutaminase assays as rapid indices of corn nitrogen availability. Journal of Plant Nutrition and Soil Science 172, 127–133.
Nuutinen, J.T., Timonen, S., 2008. Identification of nitrogen mineralization enzymes, L-amino acid oxidases, from the ectomycorrhizal fungi Hebeloma spp. and Laccaria bicolor. Mycological Research 112, 1453–1464. Available at.https://doi.org/10.1016/j.mycres.2008.06.023
O'Toole, P., 1991. Assay of urease enzyme activity in an ammonium-fixing soil. Communications in Soil Science and Plant Analysis 22, 213–224. Available at.https://doi.org/10.1080/00103629109368409

Pajares, S., Gallardo, J.F., Masciandaro, G., *et al.*, 2009. Biochemical indicators of carbon dynamic in an Acrisol cultivated under different management practices in the central Mexican highlands. Soil and Tillage Research 105, 156–163.
Pantoja, S., Lee, C., 1994. Cell-surface oxidation of amino acids in seawater. Limnology and Oceanography 39, 1718–1726.
Papen, H., Butterbach-Bahl, K., 1999. A 3-year continuous record of nitrogen trace gas fluxes from untreated and limed soil of a N-saturated spruce and beech forest ecosystem in Germany – 1. N_2O emissions. Journal of Geophysical Research-Atmospheres 104, 18487–18503.
Payne, J.M., Smith, M.W., 1994. Peptide transport by microorganisms. Advances in Microbial Physiology 36, 1–80.
Pomeroy, L.R., Macko, S.A., Ostrom, P.H., Dunphy, J., 1990. The microbial food web in Arctic seawater: Concentration of dissolved free amino acids and bacterial abundance and activity in the Arctic Ocean and in Resolute Passage. Marine Ecology Progress Series 61, 31–40. Available at.http://www.jstor.org/stable/24842245
Roberts, P., Jones, D.L., 2012. Microbial and plant uptake of free amino sugars in grassland soils. Soil Biology and Biochemistry 49, 139–149. Available at.https://doi.org/10.1016/j.soilbio.2012.02.014
Romillac, N., Piutti, S., Amiaud, B., Slezack-Deschaumes, S., 2015. Influence of pea root traits modulating soil bioavailable C and N effects upon ammonification activity. Soil Biology and Biochemistry 90, 148–151.
Rosenkranz, P., Brüggemann, N., Papen, H., *et al.*, 2006. Soil N and C trace gas fluxes and microbial soil N turnover in a sessile oak (Quercus petraea (Matt.) Liebl.) forest in Hungary. Plant and Soil 286, 301–322. Available at.https://doi.org/10.1007/s11104-006-9045-z
Rysgaard, S., Risgaard-Petersen, N., Sloth, N.P., 1996. Nitrification, denitrification, and nitrate ammonification in sediments of two coastal lagoons in Southern France. Hydrobiologia 329, 133–141.
Schimel, J.P., Bennett, J., 2004. Nitrogen mineralization: Challenges of a changing paradigm. Ecology 85, 591–602.
Schriek, S., Rückert, C., Staiger, D., Pistorius, E.K., Michel, K., 2007. Bioinformatic evaluation of L-arginine catabolic pathways in 24 cyanobacteria and transcriptional analysis of genes encoding enzymes of L-arginine catabolism in the cyanobacterium Synechocystis sp. PCC 6803. BMC Genomics 28, 1–28.
Senwo, Z.N., Tabatabai, M.A., 1996. Aspartase activity of soils. Soil Science Society of America Journal 60, 1416–1422.
Solomon, C.M., Collier, J.L., Berg, G.M., Glibert, P.M., 2010. Role of urea in microbial metabolism in aquatic systems: A biochemical and molecular review. Aquatic Microbial Ecology 59, 67–88.
Stark, C., Condron, L.M., Stewart, A., Di, H.J., Callaghan, M.O., 2007. Effects of past and current crop management on soil microbial biomass and activity. Biology and Fertility of Soils 43, 531–540.
Stark, C.H., Condron, L.M., Callaghan, M.O., Stewart, A., Di, H.J., 2008. Differences in soil enzyme activities, microbial community structure and short-term nitrogen mineralisation resulting from farm management history and organic matter amendments. Soil Biology and Biochemistry 40, 1352–1363.
Tietema, A., Warmerdam, B., Lenting, E., Riemer, L., 1992. Abiotic factors regulating nitrogen transformations in the organic layer of acid forest soils: Moisture and pH. Plant and Soil 147, 69–78.
Van Capelle, C., Schrader, S., Brunotte, J., 2012. Tillage-induced changes in the functional diversity of soil biota – A review with a focus on German data. European Journal of Soil Biology 50, 165–181. Available at.https://doi.org/10.1016/j.ejsobi.2012.02.005
Ward, B., 2008. Nitrification in marine systems. In: Capone, D., *et al.* (Eds.), Nitrogen in the marine environment. Amsterdam: Elsevier.

Biological Nitrogen Fixation☆

N Rascio and N La Rocca, University of Padua, Padua, Italy

Introduction

Nitrogen is a key element present in many biochemical compounds (such as nucleotide phosphates, amino acids, proteins, and nucleic acids) of living cells. Only oxygen, carbon, and hydrogen are more abundant in the cell. The entry of organic nitrogen in the food chains of natural ecosystems is essentially due to the activity of photoautotrophic organisms (cyanobacteria, algae, and terrestrial plants). These primary producers take up nitrogen from the environment mainly as nitrate, reduce it to ammonia, and then assimilate ammonia into organic compounds to form amino acids. However, this process of assimilatory reduction of nitrate is not the only change that nitrogen undergoes.

In the biosphere nitrogen passes through many forms, ranging from the most reduced, NH_3 (or NH_4^+), to the most oxidized, NO_3^-, in a biogeochemical cycle whose steps depend on both physical and biological events. The processes that involve living organisms (Fig. 1) include the following:

- ammonification, carried out by saprophytic bacteria (e.g., *Clostridium* spp.) and fungi that detach NH_3 from organic nitrogenous compounds and release this reduced form of inorganic nitrogen into the environment;
- nitrification, carried out by chemosynthetic bacteria that draw energy from oxidation of ammonia to nitrite (e.g., *Nitrosomonas* spp.) and of nitrite to nitrate (e.g., *Nitrobacter* spp.):

$$NH_4^+ + 3/2\ O_2 \rightarrow NO_2^- + H_2O + 2H^+ \left(\Delta G^{0,} = -66\ \text{kcal mol}^{-1}\right)$$

$$NO_2^- + 1/2\ O_2 \rightarrow NO_3^- \left(\Delta G^{0,\prime} = -18\ \text{kcal mol}^{-1}\right)$$

- assimilatory reduction of nitrate to nitrite and then of nitrite to ammonia by nitrogen autotrophic organisms:

$$NO_3^- + NAD(P)H_2 \rightarrow NO_2^- + NAD(P)^+ + H_2O$$

$$NO_2^- + 6\ \text{reduced ferredoxins} + 8H^+ \rightarrow NH_4^+ + 6\ \text{oxidized ferredoxins} + 2\ H_2O$$

- ammonia assimilation into aminoacids (see the pathway in the Ammonia Assimilation section) that leads to the recovery of organic nitrogen, thus closing the cycle.

Nonetheless, biological pathways of irreversible nitrogen loss also exist. One of them is carried out by facultative aerobic bacteria. These microorganisms (e.g., *Pseudomonas* spp. and *Alcaligenes* spp.) bring about a process of dissimilatory reduction of nitrate, called denitrification. Under anoxic conditions they activate anaerobic respiration by using nitrate and other oxidized forms of nitrogen instead of oxygen as final electron acceptors of the respiratory chain:

$$2\ NO_3^- + 4e^- + 4H^+ \rightarrow 2\ NO_2^- + 2H_2O$$

$$2\ NO_2^- + 2e^- + 4H^+ \rightarrow 2NO + 2H_2O$$

$$2\ NO + 2e^- + 2H^+ \rightarrow N_2O + H_2O$$

$$N_2O + 2e^- + 2H^+ \rightarrow N_2 + H_2O$$

This leads to volatile nitrogen forms (e.g., N_2O or N_2), which are lost to the atmosphere.

Another widespread process which contributes substantially to the loss of fixed nitrogen as N_2 in different natural environments has recently been discovered. This process is the anaerobic ammonium oxidation (anammox) performed by chemoautotrophic bacteria ("anammox" bacteria, such as *Brocadia anammoxidans* and *Scalindua* spp.) that derive energy for growth from oxidation of NH_4^+ to N_2 in complete absence of oxygen using nitrite as electron acceptor:

$$NH_4^+ + NO_2^- \rightarrow N_2 + 2H_2O\ \left(\Delta G^{0,\prime} = -86\ \text{kcal mol}^{-1}\right)$$

In natural ecosystems, the recovery of nitrogen, necessary to satisfy the nutritional demands of the inhabiting organisms, occurs through biological nitrogen fixation (Fig. 1). This event is of capital importance and consists in the reduction of molecular nitrogen (N_2) to ammonia (NH_3), providing the Earth's ecosystems with about 200 milion tons N per year. It has been estimated that the 80–90% of the nitrogen available to plants in natural ecosystems originates from biological nitrogen fixation.

☆*Change History:* July 2013. N Rascio and N La Rocca updated keywords, introdution, references, further reading. Fig. 1 legend, and cyanolichens, and A swift Ecological Overview of Biological Nitrogen Fixation in Terrestrial and Aquatic Ecosystems sections.

 doi:10.1016/B978-0-444-63768-0.00685-5

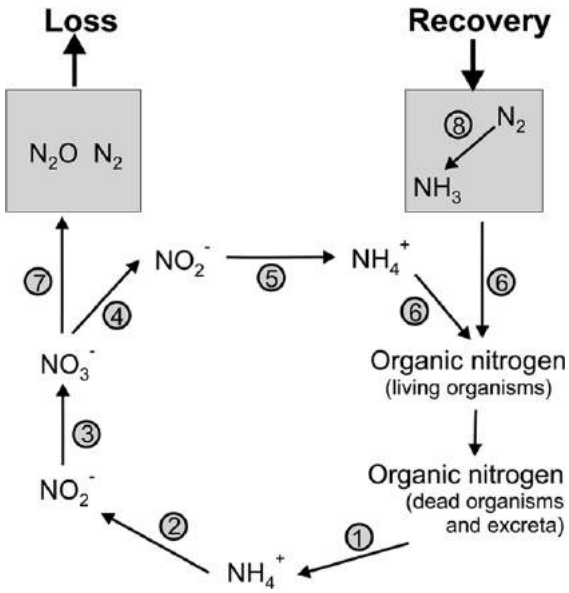


Fig. 1 A simplified scheme of the nitrogen cycle showing the steps carried out by living organisms: (1) ammonification; (2, 3) nitrification; (4, 5) assimilatory reduction of nitrate; (6) ammonia assimilation; (7) denitrification or anammox; and (8) nitrogen fixation.

Table 1 Some examples of organisms that carry out nitrogen fixation

N_2-fixing prokaryotes	*Genera*
Aerobic bacteria	*Azotobacter*
	Azospirillum
Facultative bacteria	*Klebsiella*
	Bacillus
Anaerobic bacteria photosynthetic	*Chromatium*
	Chlorobium
Non-photosynthetic	*Clostridium*
	Desulfovibrio
Cyanobacteria	*Anabaena*
	Nostoc
	Calotrix

Nitrogen-Fixing Organisms

Nitrogen constitutes almost 80% of the atmosphere, but is metabolically inaccessible to plants due to the exceptional stability of the triple covalent bond (N≡N). The ability to catalyze enzymatic reduction of N_2 to NH_3 is limited to a variety of prokaryotes defined as nitrogen-fixing or diazotrophic microorganisms, which are widely distributed in all ecosystems as either free-living organisms or in symbiotic association with a number of different plants. These N_2-fixing prokaryotes can be anaerobic, facultative aerobic, aerobic, photosynthetic, or nonphotosynthetic (**Table 1**). All carry out N_2 reduction by an enzymatic complex termed nitrogenase.

Nitrogenase and Nitrogen Fixation

The complex of nitrogenase (**Fig. 2**) consists of two distinct enzymes: dinitrogenase reductase and dinitrogenase, neither of which has enzymatic activity by itself (Igarashi and Seefeldt, 2003). Dinitrogenase reductase is a dimeric (α_2) Fe-protein of about 70 kDa with a 4Fe–4S cluster, which binds ATP and transfers electrons to dinitrogenase. The latter enzyme is a tetrameric ($\alpha_2\beta_2$) FeMo protein of about 220 kDa. It contains two Mo–Fe–S clusters and a variable number of Fe–S clusters and binds N_2. Both these enzymes are very sensitive to oxygen, which rapidly inactivates them (Dixon and Wheeler, 1986).

The molybdenum requiring nitrogenase (Mo-N_2ase) is the long-studied enzyme complex found in all diazotrophs. However other two alternative nitrogenases have been identified in some free-living bacteria belonging to the genera *Azotobacter* and

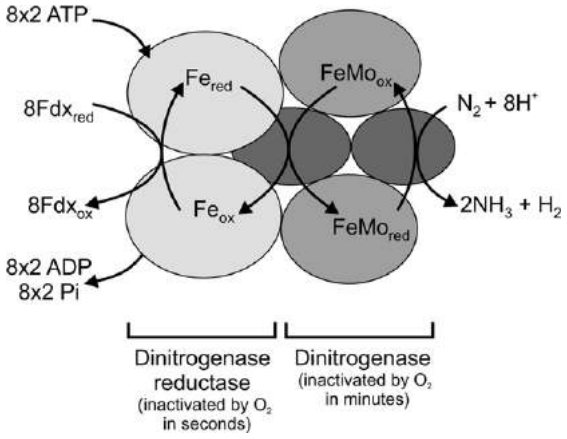


Fig. 2 The enzymatic complex of nitrogenase.

Rhodopseudomonas and in some cyanobacteria. These peculiar diazotrophic organisms can synthesize, in addition to Mo-Fe N_2ase, a vanadium nitrogenase (V-N_2ase) and a nitrogenase which requires only Fe (Fe-N_2ase) (Newton, 2007). The alternative nitrogenases may serve as substitutive pathways for nitrogen fixation in molybdenum deficient conditions. They are structurally similar to the Mo-N_2ase, except for the dinitrogenase components which are composed of $\alpha_2\beta_2\delta_2$ hexamers, they also operate in the same way, have the same requirements for activity and are reactive towards the same range of substrates.

In the nitrogen reduction carried out by the Mo-N_2ase (also referred to as conventional nitrogenase), the oxidized dinitrogenase reductase accepts an electron from a donor (reduced ferredoxin or reduced flavodoxin) and binds two molecules of adenosine 5′-triphosphate (ATP). This binding causes a conformational change of the Fe protein that lowers its redox potential (from -300 to -400 mV). The reduced Fe protein transfers the electron to the oxidized dinitroreductase with concomitant hydrolysis of both ATP molecules. Finally, the FeMo protein carries out the electron (and proton) transfer to the N_2 bound to the MoFe cofactor. Since six electrons are required to reduce N_2 to $2NH_3$, six sequential reduction events occur with the hydrolysis of 2ATP for each electron flowing through the nitrogenase. However, the nitrogenase also recognizes the protons (H^+) in the cell, so that for each N_2 reduced to $2NH_3$, two H^+ ions are reduced to H_2, with the involvement of two additional electrons and the hydrolysis of another 4ATP. Thus, the overall reaction catalyzed by nitrogenase in the diazothrophic organisms is

$$N_2 + 8e^- + 8H^+ + 16ATP \rightarrow 2NH_3 + H_2 + 16ADP + 16Pi$$

Under natural conditions the reduction of protons to hydrogen competes with that of nitrogen to ammonia for the electrons provided to nitrogenase by the donors. This lessens the efficiency of nitrogen fixation and leads to a waste of metabolic energy (ATP). Nevertheless, many nitrogen-fixing organisms have an uptake hydrogenase that reoxidizes H_2 to $2H^+$ and $2e^-$ The activity of this enzyme can greatly increase the efficiency of nitrogen fixation since it leads to ATP recovery by the flow of electrons through a respiratory transport chain, nitrogenase protection against the oxygen poisoning by reduction of O_2 to H_2O, and maintenance of nitrogenase activity by removal of H_2 that inhibits the enzyme.

For nitrogenase to function, low-potential electrons and energy (ATP) are needed. The most common source of electrons is ferredoxin (a small iron–sulfur protein). In phototrophic organisms the reduced ferredoxin can be derived from the photosynthetic electron flow, while in heterotrophic organisms the reduction occurs enzymatically through a pyruvate ferredoxin oxidoreductase that transfers electrons to oxidized ferredoxin from α-ketoacids such as pyruvate and α-ketoglutarate. An analogous reaction, carried out by a pyruvate flavodoxin oxidoreductase, produces reduced flavodoxin, a flavoprotein also used as an electron donor to nitrogenase. Some organisms may generate the required low-potential electrons by other alternative methods.

The source of energy for reduction of N_2 is ATP obtained from different metabolic pathways, according to the diazotrophic organism. In anaerobic phototrophic bacteria, the ATP comes from photosynthesis, while anaerobic heterotrophic organisms gain ATP essentially from fermentations that, due to the scarce oxidative efficiency, force them to consume large quantities of substrates. The aerobic organisms take advantage of the production of ATP through more efficient respiratory processes. Nevertheless, they still require mechanisms that keep oxygen away from the O_2-sensitive nitrogenase. Some of these mechanisms are described in the following sections.

Even if the nitrogen fixation is actually an exergonic process ($\Delta G^{\circ\prime} = -15.2$ kcal mol^{-1}), it has a high demand for energy due to the necessity to overcome unfavorable activation energies. The theoretical cost for reducing one molecule of N_2 with the concomitant reduction of $2H^+$ is 16ATP. However, under natural conditions the cost may be even higher due to the less than perfect efficiency of the process. This high energetic cost makes nitrogen fixation a strictly controlled process, through the modulation of the synthesis and activity of nitrogenase. All the other available forms of nitrogen (such as nitrate, nitrite, or amino acids) inactivate the enzymatic complex and inhibit the expression of the genes (*nif* genes) that code for the nitrogenase components (Helber *et al.*, 1988; Cheng *et al.*, 1999). The reaction products NH_3 and H_2 also cause strong inhibition. Thus, N_2 fixation is an inductive process that diazotrophic organisms activate only in the absence of other more economic nitrogen sources.

Ammonia Assimilation

The assimilation of NH_3 obtained from N_2 reduction occurs principally via the glutamine synthetase–glutamate synthase (GS-GOGAT) pathway (Nagatani *et al.*, 1971). The first enzyme catalyzes the ATP-dependent assimilation of ammonia into glutamine using glutamate as a substrate:

$$\text{Glutamate} + NH_3 + \text{ATP} \rightarrow \text{Glutamine} + \text{ADP} + \text{Pi}$$

The second enzyme catalyzes the reductive transfer of the amide group from glutamine to α-ketoglutarate, forming two molecules of glutamate. The reductants are NAD(P)H or reduced ferredoxin (Fdx_{red}):

$$\text{Glutamine} + \alpha - \text{ketoglutarate} + \text{NAD(P)H (or } Fdx_{red}) \rightarrow 2\text{glutamate} + \text{NAD(P)}^{+} (\text{or } Fdx_{ox})$$

One glutamate serves to keep the pathway going, whereas the other represents the gain in organic nitrogen.

Diazothrophic Bacteria – Plant Symbioses

Nitrogen-fixing microorganisms have been found in roots or other organs of many species of plants with which they establish symbiotic associations. The diazotrophic partners can be aerobic bacteria or, in some cases, cyanobacteria.

Symbiotic associations of current ecological importance for wide diffusion and the large nitrogen supply to the ecosystems are those between N_2-fixing bacteria and roots of higher plants and, in particular, the rhizobia-legume and *Frankia*-dicotyledon symbioses. Another association of particular interest is that established by endophytic diazotrophic bacteria with cereals.

Rhizobia–Legume Symbiosis

Most Leguminosae (about 90%) can establish a symbiotic association with aerobic diazotrophic Gram-negative bacteria commonly referred to as rhizobia. This symbiosis takes place in roots and brings about the formation of nodules in which N_2 fixation occurs. The large contribution made by these symbioses to the nitrogen availability for agronomically important legumes is well known. *Medicago sativa*, for instance, can fix 300 kg N ha^{-1} $year^{-1}$ and *Vicia faba* over 500 kg N ha^{-1} $year^{-1}$. It has been calculated that soybean (*Glycine max*), which is the dominant crop legume, representing 68% of global production, can fix more than 16×10^6 T N annually, corresponding to over 75% of the N fixed by crop legumes (Herridge *et al.*, 2008).

This makes biological nitrogen fixation a major component of sustainable agricultural systems, since it has the potential to greatly limit the use of chemical nitrogen fertilizers.

Numerous species belonging to the family Leguminosae are also abundant in natural ecosystems, such as the forests of tropical regions (e.g., those in Brazil and Guyanas) (Koponen *et al.*, 2003, Kreibich *et al.*, 2006), where they can represent over 50% of all trees. Tropical forests often grow on substrates poor in mineral nutrients, and thus the continuous supply of nitrogen through biological N_2 fixation acquires an essential role to maintain large nitrogen pools in these ecosystems.

The rhizobia forming symbiosis with legume roots belong to five different genera: *Rhizobium*, *Azorhizobium*, *Mesorhizobium*, *Sinorhizobium*, and *Bradyrhizobium*. A given species of bacterium establishes symbiosis with one or few species of legumes (**Table 2**).

This is due to the host-symbiont recognition occurring in the rhizosphere through the exchange of molecular signals. The first event of the root nodule formation is the chemotactic movement of the bacterium toward the root of the host plant in response to chemical attractants, usually specific flavonoids secreted by the root under nitrogen-starvation conditions. Each legume species produces its own particular cocktail of these compounds and this contributes to one of the distinctive characteristics of rhyzobium-legume symbiosis, with a rhizobial species usually infecting a very limited host range (Marks, 2004).

The compounds secreted by the roots induce the expression of host-specific bacterial genes (*nod* genes) coding for Nod factors (lipo-chito-oligosaccharides) that, in turn, induce plant responses and trigger the nodule developmental program (Spaink, 2000).

Root infection starts with the bacterium-induced curling of a root hair, bacterium attachment to the hair surface, cell wall degradation, and formation of the infection thread. This is an internal tubular extension of the hair plasma membrane that carries out the proliferating rhizobia from the root surface into the root cortex. Concomitantly, some cortical cells undergo rapid divisions

Table 2 Some examples of associations between rhizobia and legumes

Rhizobia	*Host plants*
Bradyrizobium japonicum	*Glycine*, *Vigna*
Sinorhizobium meliloti	*Medicago*, *Trigonella*, *Melilotus*
Sinorhizobium fedii	*Glycine*, *Vigna*
Azorhizobium caulinodans	*Sesbania*
Rhizobium leguminosarum biovar. *phaseoli*	*Phaseolus*
Rhizobium leguminosarum biovar. *trifolii*	*Trifolium*
Rhizobium leguminosarum biovar. *viciae*	*Vicia*, *Pisum*, *Cicer*
Mesorhizobium loti	*Lotus*, *Lupinus*, *Anthillis*

that give rise to the nodule primordium. When the branched infection thread reaches target cells within the developing nodule, its tip vesiculates releasing bacteria packaged in a membrane derived from the host cell plasmalemma. The rhizobia undergo some divisions but very soon they stop dividing and differentiate into diazotrophic bacteroids. Bacteroids and surrounding peribacteroid membrane form the symbiosome (**Fig. 3(b)**), which is the site of N_2 fixation. In the mature nodule (**Fig. 3(a)**) specialized structures are developed around the infected tissue: an endodermis and a vascular system continuous with the root stele, and a layer of cells hampering O_2 diffusion to the root nodule interior. Some leguminous species such as soybean, peanut, and bean form spherical determinate nodules with a nonpersistent meristem (**Fig. 3(a)**).

Others, such as pea, clover, and alfalfa, form cylindrical indeterminate nodules with a persistent terminal meristem (**Fig. 4**). The mature determinate nodule contains a homogenous central tissue with cells fully packed with nitrogen fixing bacteroids, whereas in the indeterminate nodule a gradient of developmental states occurs, due to the active meristem that continuously produces cells that become infected with bacteria (Ferguson *et al.*, 2010). At maturity a meristematic zone, an invasion zone with infection threads, a N_2-fixing zone with bacteroids and a senescent zone with degraded bacteroids can be distinguished along the nodule axis (**Fig. 4**).

Different mechanisms take place to obtain the microaerobic environment appropriate for maintaining ATP production in host cells and bacteroids and for preserving, at the same time, nitrogenase activity in N_2-fixing tissue. The first hindrance to the entry of oxygen into the infected cells is the mechanical diffusion barrier in the nodule parenchyma. Moreover, leghemoglobin is synthesized in the cytoplasm of the host cells (Downie, 2005). This oxygen-binding protein plays a major role in delivering oxygen to the bacteroid surface and accounts for the characteristic pink color of N_2-fixing tissue. Efficient bacteriod respiration also restricts oxygen penetration into the cytoplasm and provides nitrogenase with the ATP and reductants required. Finally, in most rhizobia the activity of an uptake hydrogenase is an additional help for protecting nitrogenase against the O_2-poisoning (Ciccolella *et al.*, 2010).

Bacteroids do not have enzymes for the ammonia assimilation. For this reason, the NH_3 obtained from N_2 reduction is released into the root cell where the assimilation occurs via GS-GOGAT pathway (**Fig. 5**). This leads to production of glutamine, glutamate, and, successively, of other nitrogenous transport compounds. Some of these organic compounds are returned to the bacteroids, but most are exported to the plant shoot via xylem.

In order to sustain N_2 fixation, the host plant must supply the bacteroids with a carbon source, which arrives to the root nodule via phloem as sucrose. However, this sugar is metabolized in the host cell and converted to C_4 dicarboxylates, principally malate. The dicarboxylates, in fact, are transported across the peribacteroid membrane, becoming the primary carbon source for the N_2-fixing organisms (White *et al.*, 2007).

Symbiotic nitrogen fixation is crucial to success of legumes, but the plant has to control the number of nodules it forms to balance the nitrogen gains with the developmental costs in order to avoid the severe energy drain that would be imposed by having to many nitrogen fixing nodules. The control of nodule number on the colonized root occurs via a complex systemic mechanism called "autoregulation of nodulation", based on root-derived and shoot-derived signals (Kouchi *et al.*, 2010).

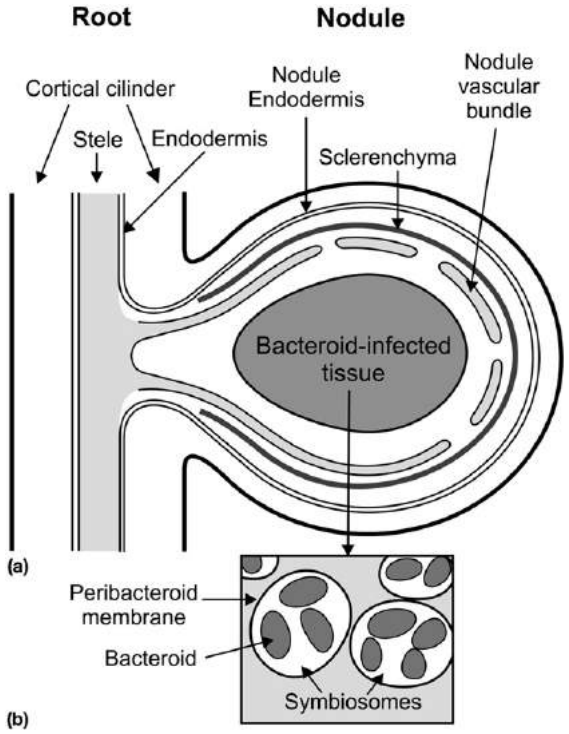


Fig. 3 Schematic drawings of: (a) a determinate root nodule of a rhizobia-legume symbiosis and (b) a part of an infected cell with symbiosomes.

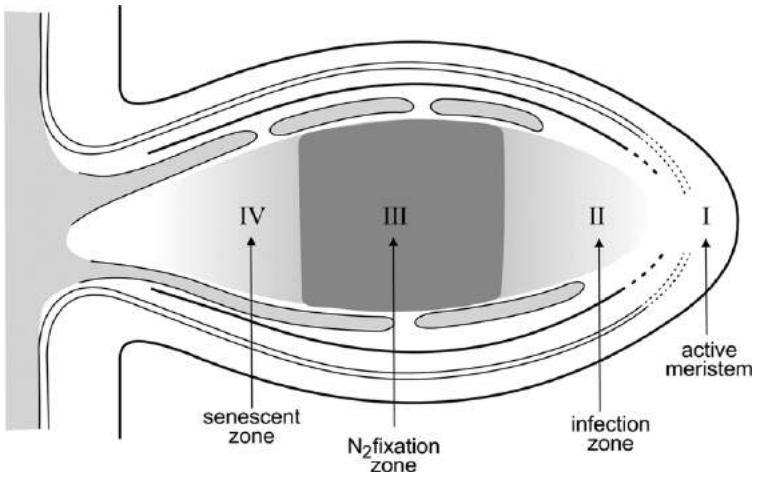


Fig. 4 Schematic drawing of an indeterminate root nodule of a rhizobia-legume symbiosis.

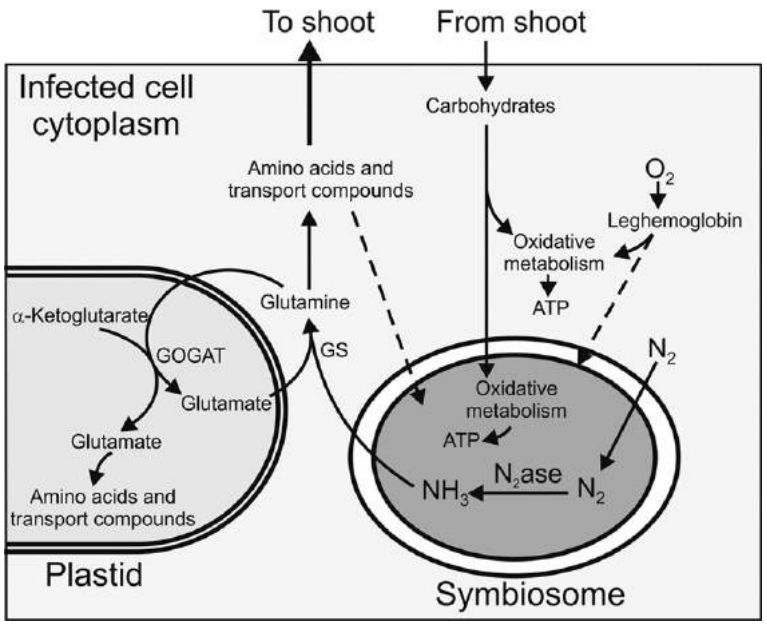


Fig. 5 A simplified diagram showing nitrogen fixation and ammonia assimilation in an infected cell of a legume root nodule. GOGAT, glutamate syntase; GS, glutamine syntetase; N_2ase, nitrogenase.

Furthermore legume regulates the nodule number in response to environmental nitrogen (nitrate or ammonia) availability to preferentially obtain this nutrient from sources energetically favourable relative to the cost of nodulation (Jeudy *et al.*, 2009).

In both determinate and indeterminate nodules bacteroids carry out an optimal N_2-fixation for 4–5 weeks after infection. Beyond this period the bacteroid activity declines and a senescence process occurs in the N_2-fixing zone, leading to degradation of the infected cells. Finally, bacteroids (or some undifferentiated bacteria) are released from decayed nodules to the soil where they can recolonize the rhizosphere resuming the free saprophytic lifestyle.

Frankia-Dicotyledon Symbiosis

The aerobic Gram-positive actinomycetes belonging to the genus *Frankia* are diazotrophic bacteria that are capable of inducing formation of N_2-fixing nodule lobes in roots of many dicotyledonous angiosperms. The plants nodulated by *Frankia* strains are known as actinorhizal plants and include 8 families, 25 genera, and over 200 species, most of which are perennial woody shrubs or trees distributed in all landmasses except Antarctica. The actinorhizal plants share a predilection for marginally fertile soils and the majority are pioneers on nitrogen-poor sites. In addition, many actinorhizal species are able to tolerate environmental stresses such as heavy metals, high salinity, drought, cold, and extreme pH. They inhabit a variety of ecosystems, including coastal dunes,

riparian zones, alpine communities, arctic tundra, glacial tills, and forests. Actinorhizal plants are especially important in high latitude regions, such as Scandinavia, Canada, Alaska, and New Zealand where Leguminosae are absent or rare while actinorhizal plants are abundant and capable of vigorous growth (Wall, 2000). Much of the new nitrogen entering these ecosystems comes from the actinorhizal symbioses that, on the whole, account for over 15% of the biologically fixed nitrogen worldwide.

The filamentous frankiae, besides in symbiotic association with actinorhizal plants, can also occur as free-living diazotrophic organisms (Benson and Silvester, 1993). In pure culture, *Frankia* strains produce extensive hyphae and sporangia. In response to nitrogen deprivation, they also differentiate vesicles, named diazovesicles, which contain nitrogenase and are the site of N_2 fixation. The diazovesicles are encapsulated by a series of laminated lipid layers that are rich in neutral lipids, glycolipids, and hopanoids. This envelope, whose thickness depends on the environmental O_2 concentration, works as an oxygen-diffusion barrier, providing an anaerobic environment for nitrogenase to function inside vesicles (Berry *et al.*, 1993).

The *Frankia* strains that nodulated actinorhizal plants can be phylogenetically distinct in three groups (groups I, II, and III) that infect specific dicotyledon families (**Table 3**).

Actinorhizal plants fall into families of three related orders: Rosales (Rosaceae, Eleagnaceae, Rhamnaceae), Fagales (Betulaceae, Casuarinaceae, Myricaceae) and Cucurbitales (Coriariaceae, Datiscaceae). Together with Fabales (legumes), they form a single "nitrogen-fixing clade" within the angiosperms (Soltis *et al.*, 1995).

In the actinorhizal symbioses, root nodule formation begins with the host-symbiont recognition through the exchange of molecular signals, the knowledge of which is still limited.

However, some findings suggest that the signaling mechanisms of *Frankia*-actinorhizal plants might be similar to those of rhizobia-legumes (Hocher *et al.*, 2011b) and genomic analyses reinforce the hypothesis of a possible single origin for legume-rhizobia and actinorhizal symbioses (Hocher *et al.*, 2011a).

Frankia strains can infect the host root by intracellular or intercellular mechanisms. Intracellular infection, such as that occurring in genera *Myrica*, *Comptonia*, *Alnus*, and *Casuarina*, starts with penetration of bacterial hyphae in a curled root hair. Afterward the hyphae move in cortical cells encapsulated with a layer of plant cell wall material surrounded by host plasmalemma. In intercellular infection, common in genera *Elaeagnus*, *Ceanothus*, and *Cercocarpus*, the bacterial hyphae penetrate between two adjacent rhizoderm cells and progress apoplastically through cortical cells encapsulated in a pectic matrix. Concomitantly, cell divisions induced in the root pericycle give rise to the nodule lobe primordium to which the hyphae move. The mature actinorhizal nodule lobe resembles a modified lateral root with an apical meristem but without a root cap. It shows a central stele with vascular tissues and has *Frankia* hyphae restricted to the cortical cells (**Fig. 6**).

In most actinorhizal symbioses, the N_2-fixing activity of *Frankia* in infected cells is associated with differentiation of diazovesicles whose morphology is strictly controlled by the host plant. As in the free-living frankiae, these vesicles are surrounded by the multilayered lipid envelope and contain nitrogenase. However, in some symbioses (with plants of genera *Myrica*, *Coriaria*, *Comptonia*, and *Casuarina*), the *Frankia* hyphae proliferate without forming vesicles. The mature anatomy of a nodule lobe is reached at about 2 weeks after inoculation while the N_2-fixation can be detected after three weeks (Huss-Danell, 1997).

In infected cells of mature nodule lobes, some mechanisms take place to lower the oxygen tension near the site of the oxygen-intolerant nitrogenase. The first diffusion resistance to oxygen is provided in diazovesicles by the multilayered envelope and a further reduction of the pO_2 is obtained through their high respiration rate. In many nodule lobes devoid of diazovesicles, the infected cells contain high levels of hemoglobins that have homologous sequences to leghemoglobins and are believed to play the same role (Fleming *et al.*, 1987). In these nodules, moreover, a low pO_2 may be maintained by lignification of the host cell walls. Finally, the activity of uptake hydrogenases can also help to protect the nitrogenase against O_2 in both hyphae and diazovesicles of the symbiotic frankiae (Leul *et al.*, 2009). In free-living *Frankia* strains, as in the other free-living diazotrophs, the ammonia produced by N_2 fixation is assimilated by the organism via the GS-GOGAT pathway. On the contrary, these enzymes are differently regulated in the symbiotic frankiae. In diazovesicles of root nodule lobes GS activity is very low and ammonia remains unassimilated (Alloisio *et al.*, 2010). As in rhizobia–legume symbiosis, NH_3 is released into the host cell where its assimilation gives rise to amino acids and other organic nitrogen compounds. Some are furnished to the bacterium, but most of them are transferred to the plant shoot.

Table 3 Association between *Frankia* and actinorhizal plants

Frankia phylogenetic groups	*Plant families*	*Plant genera*
Group I strains	Coriariaceae	*Coriaria*
	Datiscaceae	*Datisca*
	Rosaceae	*Cercocarpus*, *Chamaebatia*, *Dryas*, *Cowania*, *Purshia Ceanothus*
	Rhamnaceae	
Group II strains	Betulaceae	*Alnus*
	Casuarinaceae	*Casuarina*, *Allocasuarina*, *Ceuthostoma*, *Gymnostoma*
	Myricaceae	*Myrica*, *Comptonia*
Group III strains[a]	Elaeagnaceae	*Elaeagnus*, *Hippophaë*, *Shepherdia*
	Rhamnaceae	*Calletia*, *Discaria*, *Kentrothamnus*, *Retanilla*, *Talguenea*, *Trevoa*

[a]Group III strains are more promiscuous and can occasionally inhabit root nodules of Rosaceae, Coriariaceae and *Ceanothus*.

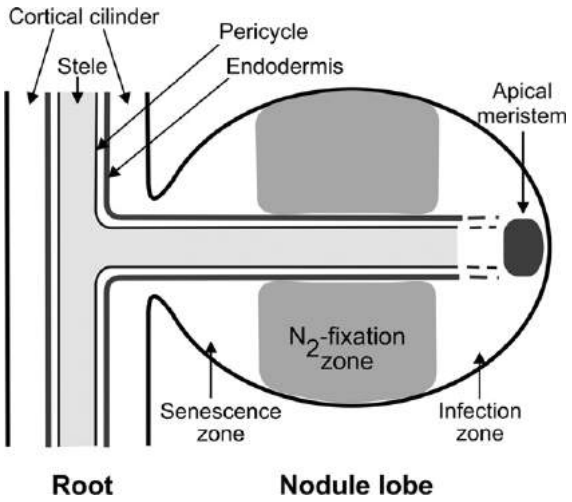


Fig. 6 Schematic drawing of a mature actinorhizal nodule lobe.

The scarcity or lack of GS activity in the diazotrophic symbiont also characterizes the rhizobia legumes as well as some cyanobacterial symbioses such as *Anabaena–Azolla*, showing a remarkable convergence of physiological strategies in the N_2-fixing associations.

The actinorhizal plant must provide photosynthates to the symbiotic bacterium. As in the rhizobia–legume symbiosis C_4 dicarboxylates derived from sucrose metabolism occurring in the host cell are likely to be the carbon sources for *Frankia* strains in actinorhizal symbiosis.

Furthermore, as it occurs in rhizobia–legume symbiosis, the actinorhizal plants can control infection by *Frankia* and regulate number and development of nodule lobes on roots by systemic autoregulatory processes (Wall *et al.*, 2003).

Endophytic Diazotrophic Bacteria–Cereal Association

A recently discovered nitrogen-fixing association is that between grasses, such as sugar cane, maize, rice, wheat, sorghum, and other graminaceous species and endophytic diazotrophic bacteria that can colonize the plant interior without causing symptoms of disease. These bacteria enter the plants at root tips or at the emergence points of lateral roots and penetrate the root cortex, the stelar tissues, and the xylem vessels through which they may migrate toward the shoot (Cocking, 2003). Endophytic diazotrophic bacteria are generally restricted to intercellular spaces and especially to the xylem vessels where the low pO_2 and the high bacterial respiration rate create the microaerobic conditions needed for nitrogenase activity. Some of these diazotrophic microorganisms, such as those belonging to the genera *Acetobacter*, *Herbaspirillum*, *Azospirillum*, and *Azoarcus*, are of extreme interest since they can significantly contribute to the nitrogen requirement of the graminaceous plants (Kennedy *et al.*, 2004). Certain rice varieties, for instance, can obtain over 30% of their nitrogen from these endophytes (James *et al.*, 2002) and some Brazilian sugarcane varieties up to 80%, with a total contribution of more than 170 kg N ha^{-1} $year^{-1}$. (Urquiaga *et al.*, 1992).

Studies of these N_2-fixing associations form a topical field of research whose aim is to explore the possibility of both enhancing the N_2 fixation and extending this efficient system to other cereals. This would greatly reduce the use of nitrogen fertilizers with considerable economic benefits, and, above all, with enormous environmental advantages. Over two-thirds of arable lands, in fact, are dedicated to the growing of cereals, which provide 80% of the food for the world's populations.

Nitrogen Fixation in Free-Living Cyanobacteria

Among the free-living diazotrophs a prevailing interest is that addressed to cyanobacteria. This interest comes from the wide and abundant distribution of these microorganisms in all terrestrial and aquatic ecosystems as well as from their unique photosynthetic metabolism that makes the nitrogen fixation an apparently paradoxical event. Cyanobacteria, in fact, are the only prokaryotes that carry out oxygenic photosynthesis.

A very great number of these microorganisms is able to both fix N_2 under aerobic conditions and produce O_2 by photosynthesis. Filamentous cyanobacteria resolve this oxygenic photosynthesis–diazotrophy paradox by segregating the oxygen-sensitive machinery for N_2 fixation in specialized nonphotosynthetic cells named heterocysts, and by maintaining the oxygen evolving photosynthesis in the neighboring vegetative cells. Thus, the simultaneous operation of the two basically incompatible processes is made possible through their spatial separation.

Nitrogen starvation leads to the appearance at regular intervals along the cyanobacterial filament of heterocysts which function as anaerobic factories for N_2 fixation under external aerobic conditions. The ability to fix N_2 ensues from changes that occur in vegetative cells that differentiate to heterocysts (Fig. 7).

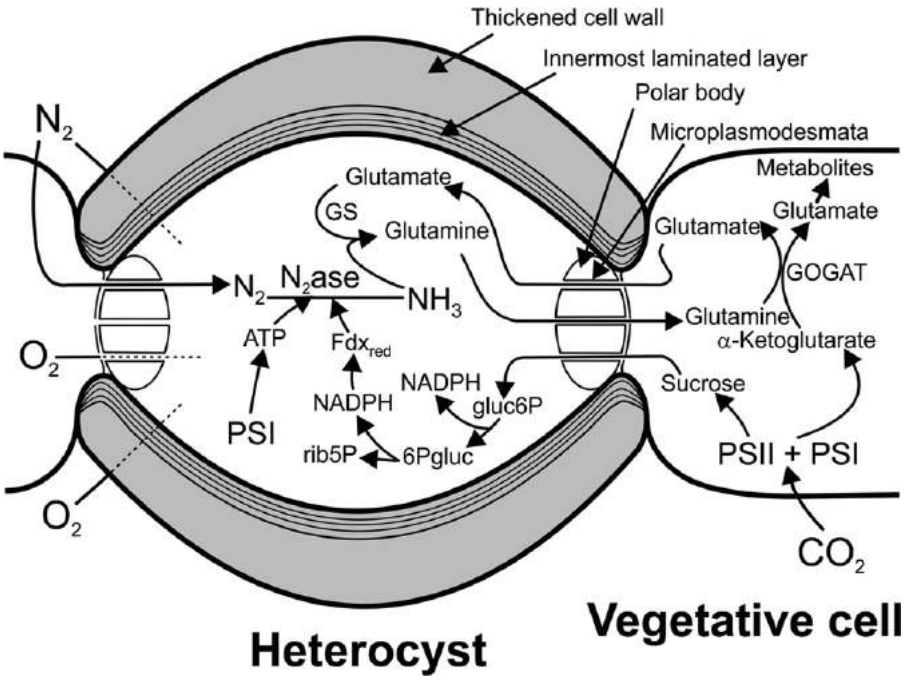


Fig. 7 Schematic drawing of a cyanobacterial heterocyst showing nitrogen fixation and metabolite exchange with the neighboring vegetative cell. Fd_{red}, reduced ferredoxin; gluc6, glucose-6-phosphate; 6gluc, gluconate-6-phosphate; GOGAT, glutamate syntase; GS, glutamine syntetase; N_2ase, nitrogenase; PSI, photosystem I; PSII, photosystem II; rib5P, ribulose-5-phosphate.

First they build up a very thick cell wall with an innermost laminated glycolipid layer, whose function is to provide an O_2 permeability barrier to avoid the inactivation of nitrogenase inside the cell (Nicolaisen *et al.*, 2009). The connections between heterocyst and neighbor cell occur through thin cytoplasmic channels (microplasmodesmata), which traverse the septum separating the two cells and the plag (polar body) of cyanophycin filling the adjacent region (Mullineaux *et al.*, 2008). In addition, the photosynthetic apparatus undergoes a deep reorganization during the heterocyst differentiation: phycobilisomes disappear and the oxygen-evolving photosystem II is totally dismantled, while photosystem I, which produces ATP through cyclic photophosphorylation, persists in thylakoid membranes (Wolk, 1996). The ATP necessary to fulfill the energy demand for nitrogenase activity, in fact, comes in heterocysts from cyclic photophosphorylation, while the reductant for the N_2-fixing enzyme is furnished by neighboring photosynthetic cells probably as sucrose (Cumino *et al.*, 2007). The sugar hydrolysis and glucose oxidation through the pentose phosphate pathway produces NADPH used for ferredoxin reduction.

The heterocysts of free-living cyanobacteria contain high levels of glutamine synthetase (GS), but are deficient in GOGAT that, on the contrary, is active in vegetative cells. Thus, after N_2 reduction, the NH_3 assimilation in glutamine is carried out in the heterocyst while the successive reaction leading to glutamate synthesis occurs in the near vegetative cell into which glutamine moves through microplasmodesmata (Martin-Figueroa *et al.*, 2000).

A major role in protecting nitrogenase against O_2 is played in heterocysts by the thick cell wall which prevents gas diffusion toward the cell. However, it is unlikely that this envelope provides a truly impermeable gas diffusion barrier, since this would also exclude nitrogen from the fixation site. Moreover, gases can reach the heterocyst cytoplasm through the junctions between them and the contiguous vegetative cells which produce O_2 by photosynthesis. Thus, also cyanobacterial heterocysts, as the other aerobic diazotrophic organisms, need systems to remove oxygen that enters the cell. These include enhanced rate or respiration (Valladares *et al.*, 2007), probable presence of hemoproteins in the cytoplasm peripheral region, and activity of an uptake hydrogenase (Tamagnini *et al.*, 2007).

A great ecological interest arose from the unexpected finding that unicellular and nonheterocystous filamentous cyanobacteria are also able to both fix nitrogen and carry out oxygenic photosynthesis. These cyanobacteria, which are very abundant in the phytoplanktonic populations of marine environments, are responsible for most of the photosynthetic organic carbon provided to the ecosystem (Karl *et al.*, 2002), and they may also account for a high percentage of the nitrogen fixed biologically worldwide.

The oxygenic photosynthesis–diazotrophy paradox is resolved by these nonheterocystous cyanobacteria through creative strategies enabling them to separate spatially or temporally the two incompatible processes.

The marine filamentous cyanobacteria of genus *Trichodesmium* fix nitrogen during the light period in a subset of specialized nitrogenase-containing cells, named diazocytes, formed by cell division primarily confined to dark period. In addition, nitrogenase shows a diurnal pattern in which its activity is highest early in the day (Sandh *et al.*, 2009).

A number of unicellular diazotrophic cyanobacteria, such as those of genus *Cyanothece*, exhibit temporal separation of the two physiological processes that should necessarily occur in the same cell. They carry out only the oxygenic photosynthesis during the

day and fix N_2 only during the night, when the photosynthetic O_2 production does not occur. This timing of N_2 fixation is also related to the fact that the nitrogenase is active exclusively during the dark period. Interestingly, the daily oscillation of nitrogenase activity occurs according to an endogenous circadian rhythm (Toepel *et al.*, 2008).

The finding that the nitrogenase of nonheterocystous N_2-fixing cyanobacteria possesses this kind of rhythmic activity was also of great scientific importance since it was the first clearly recognized circadian rhythm in prokaryotic organisms, which led to the backdrop of the former dogma that restricted the biological clocks to eukaryotes (Golden *et al.*, 1997).

Diazotrophic Cyanobacteria–Plant Symbioses

Many species of filamentous N_2-fixing cyanobacteria, in the great majority of cases belonging to the genus *Nostoc*, can form symbiotic associations with a wide range of eukaryotic hosts, among which fungi (cyanolichens), microalgae (diatoms), corals, sponges and numerous plants.

The cyanobacteria-plant symbioses include Bryophyta (liverworts, hornworts, and mosses), Pteridophyta (the genus *Azolla*), gymnosperms (Cycadaceae) and angiosperms (Gunneraceae) as hosts.

The free-living cyanobacteria which can form these symbiotic associations share two major characteristics: they are able to differentiate heterocysts and to produce short motile filaments, known as hormogonia. Hormogonia, which provide a means of dispersal for otherwise immotile cyanobacteria, are the infective agents which enter the host plant tissues (Meeks and Elhai, 2002).

The hormogonia, which are devoid of heterocysts, form in response to chemical signals (HIFs: hormogonia-inducing factors) released from the plant, that also produces chemoatractants to guide them into symbiotic cavities. Subsequent to the infection the host may release signals (HRFs: hormogonia-repressing factors) to prevent formation of further hormogonia and to shift the cyanobiont development towards heterocyst differentiation and N_2 fixation (Rai *et al.*, 2000). The frequency of heterocysts in filaments of the symbiotic cyanobacterium is several folds higher than in the same cyanobacterium in free-living state. However, only a low quantity of nitrogen fixed is retained by the cyanobiont, the remaining being transferred to the host plant as ammonia. Photosynthesis can be highly depresses in cyanobionts, relative to that in free-living strains, but the reduced CO_2 fixation is compensated by sugars derived from the hosts (Meeks, 2003).

Among the numerous cyanobacteria-plant symbioses, particular attention has been paid to that between the cyanobiont *Anabaena azollae* and the leaves of the aquatic fern *Azolla*. The interest for the *Anabaena–Azolla* association is mainly due to the potential use of *Azolla* as fertilizer in rice fields.

Anabaeana–Azolla Symbiosis

Azolla is a small floating fern and is the only known pteridophyte that lives in symbiosis with a diazotrophic cyanobacterium. All the species of the genus harbour in their fronds a filamentous N_2-fixing cyanobacterium until now referred as *Anabaena azollae* (Nostocaceae) (Papaefthimiou *et al.*, 2008).

The *Azolla* sporophyte (generally 0.5–7 cm in lenght but up to 40 cm in *A. nilotica*) consists of a multibranched rhizome generating, on the ventral surface, adventitious roots hanging down into the water to absorb nutrients directly. The rhizome bears small leaves (approximatively 1 mm in length) consisting of an aerial chlorophyllous dorsal lobe and a partially submerged colourless ventral lobe which is cup-shaped to provide buoyancy. Each dorsal lobe contains a specialized cavity housing the cyanobiont permanently.

The mature cavity, ellipsoid in shape, has the interior surface covered by a mucilaginous layer delimited by an envelope, where 2000–5000 cyanobacterial cells are embedded and immobilized. Several trichomes, traditionally called hairs, protrude from the cavity surface into the mucilage layer and create an intimate contact between the symbiotic partners helping in the exchange of metabolites. Thus the leaf cavity can be considered as a natural microcosm with a self organization and an ecological defined structure. It behaves as both the physiological and dynamic interface unit of the symbiotic association where the main metabolic and energetic flows occur (Peters and Perkins, 1993; Rai *et al.*, 2000).

The diazotrophic cyanobacterium *Anabaena azollae* consists of unbranched filaments containing bead-like highly pigmented vegetative cells and lightly pigmented intercalary N_2-fixing heterocysts. In the youngest leaves of the water fern the *Anabaena* filaments lack heterocysts while these gradually increase in frequency, relative to photosynthetic cells, reaching 30%–40% of the cyanobacterial cells in *Anabaena* population of mature leaf cavities. This heterocyst frequency in *Azolla*-associated *Anabaena* (compared to 5%–6% in other free-living species of *Anabaena*) is far higher than necessary to support the fixed N needs of the symbiotic cyanobacterium. Moreover, an overall decrease (up to 70%) in GS activity in the *Azolla* cyanobiont has also been shown, that limits the ammonia assimilation in heterocysts. Thus, some 50%–90% of fixed nitrogen is delivered by *Anabaena* to the fern as ammonia. Translocation of fixed N from the symbiotic environment of mature cavities to other parts of the host plant occurs in the form of amino acids.

Concomitantly with the differentiation of *Anabaena* into a higher proportion of nitrogen-fixing heterocysts, relative to the photosynthetic cells, a reduction of photosynthesis and CO_2 fixation occurs in cyanobacterial filaments, which are supplied with sucrose by the host plant (Chapman and Margulis, 1998).

The *Anabaena–Azolla* symbiosis is perpetual and hereditary and the symbiotic condition can be described as obligate for the cyanobiont whose free-living form is not found in nature.

The symbiosis is maintained during all the life cycle of the pteridophyte throughout both sexual and asexual reproduction without requiring fresh infection from the environment. In contrast with other plant-cyanobacterial symbioses, *Azolla* hormogonium initiation factors (HIFs) are unknown. A colony of *Anabaena* is associated with each fern shoot apex and, as the plant grows, the cyanobacterial filaments are partitioned off into each new leaf.

In the *Azolla–Anabaena* symbiosis, the cyanobiont growth is synchronized with that of the host plant. The growth rates of both partners are highest in the apical parts of the fern and decrease along the axis away from the apex. When growth of *Azolla* stops, cyanobiont cells cease to divide (Lechno-Yossef and Nierzwicki-Bauer, 2002).

The water fern *Azolla* naturally occurs on lake surfaces, slow-moving rivers, canals, ponds, and ditches in warm-temperate to tropical climates, but its world distribution has been enlarged by humans. In fact, the *Anabaena–Azolla* association has been shown to be of major agronomic importance for its potentiality as a biofertilizer to substitute chemical nitrogen compounds. *Azolla* has been used as "green manure" in several countries to fertilize rice paddies and to increase rice yields (van Hove and Lejeune, 2002). *Azolla–Anabaena* is capable of fixing nitrogen at higher rates than legumes and is able to growth successfully in waterlogged habitats having low level of nitrogen. The Asians have recognized benefits of *Azolla* on rice cultivation first, since both rice crop and fern require similar environmental growing conditions. In rice fields, for instance, it can fix over 1 kg N ha^{-1} day^{-1}, providing sufficient nitrogen to allow sustainable rice cultivation. Increase from 14% to 40% in grain yield of rice has been reported with *Azolla* used as dual crop. Moreover, *Azolla* has a high rate of multiplication, which helps in covering very rapidly the surface of water bodies where it is growing. Thus the thick mat produced helps to reduce the volatilization of ammonia in rice fields.

The exploitation of this system in temperate environments is limited due to the *Anabaena–Azolla* sensitivity to low winter temperatures and to alternating day/night temperatures in spring before the rice sowing period (Pabby *et al.*, 2003). Recently, however, *Azolla* strains have been selected as practicable biofertilizer even in the high latitude of the temperate rice areas, such as those of Northern Italy (Bocchi and Malgioglio, 2010).

Besides its extensive use as a N-supplement in rice-based ecosystems, it has also been utilized in other crop cultivations such as taro, wheat, tomato, and banana.

Furthermore, the *Anabaena–Azolla* association is also applied as controlling agent for weeds and mosquitos, due to its ability to cover water surfaces, and for improving water quality for its properties of removing excess quantities of nitrate and phosphorous.

Cyanolichens

Lichens are symbiotic associations between fungi (mycobionts, commonly Ascomycetes) and photosynthetic partners (photobionts) which can be green algae (commonly *Trebouxia* spp.) or diazotrophic cyanobacteria (commonly *Nostoc* spp.). These mycobiont-photobiont symbioses are regarded as mutualistic. The photoautotroph partners provide photosynthetic products and, in the case of cyanobacteria, also organic nitrogen compounds to the fungus. This latter, in turn, provides shelter for the photobionts by enclosing them within the thallus. In this way the photosynthetic organisms, protected from drying and strong sunlight, are allowed to growth in very harsh conditions.

Lichens can tolerate the most extreme environments on Earth. They can live in hot deserts and arctic regions, on steril soils, bare rocks, wood debris and epiphytes on tree trunks, branches and leaves. Recently it has been reported that lichens are able to survive exposure to space conditions and even to adapt to the harsh Mars environment (de Vera, 2012; Raggio *et al.*, 2011). Lichens can reproduce asexually through the dispersal of small fragments of thallus or through propagules (diaspores) tipically containing cells from the symbiotic partners.

Bipartite lichens are symbiotic associations of a fungus with a single photosynthetic partner, which in so-called chlorolichens is a green alga while in lichens defined as cyanolichens is a diazotrophic cyanobacterium. In these latter associations the cyanobacterium may be localized in a distinct layer or dispersed through the thallus. As the sole photobiont, it is responsible for provision of both fixed carbon and fixed nitrogen to the mycobiont. The cyanobacterium maintains high photosynthetic levels and N_2-fixing activity, with heterocyst frequency (7–8%) close to its free-living counterpart.

In tripartite lichens the fungus is associated with both a green alga (phycobiont) and a cyanobacterium (cyanobiont). In these cyanolichens (also referred to as cephalolichens) the phycobiont is widely distributed through the thallus, while the cyanobiont is confined to special structures, named cephalodia. In cephalodia the cyanobiont becomes predominantly heterotrophic and is specialized for nitrogen fixation, as shown by the increase of heterocyst proportion over 30% of cyanobacterial cells. Thus, in these tripartite cyanolichens the fixed carbon is provided to the fungus by the phycobiont, while the cyanobiont supplies it with the fixed nitrogen.

A Swift Ecological Overview of Biological Nitrogen Fixation in Terrestrial and Aquatic Ecosystems

Nitrogen is an essential element for all living organisms and the nitrogen supply is crucial for the ecology and biogeochemistry of terrestrial and aquatic ecosystems. A number of both physical (e.g., the dissolved nitrate exiting with percolating waters) and biological (e.g., denitrification and anammox) processes tend to restrict the biological nitrogen availability in the ecosystems. Apart from a minor contribution by lighting, the biological nitrogen fixation is the process which provides the largest source of new combined nitrogen to natural environments for replenishing the nitrogen losses. However, the fixation rates vary widely

across the different ecosystems where symbiotic and free-living diazotrophic microorganisms can be differently involved in the process. Although the biological fixation is the major nitrogen input into terrestrial and aquatic ecosystems, currently there is only limited understanding about the factors that regulate this process in natural environments.

In arid lands of the world, from arctic ecosystems (Stewart *et al.*, 2011) to hot deserts of Southern Africa (Büdel *et al.*, 2009), much of the nitrogen input via biological nitrogen fixation relies heavily on diazotrophic microorganisms, mostly cyanobacteria, living in biological soil crusts (BSCs). The BSCs, also referred to as cryptogamic crusts, are microbic communities (millimeters to centimeters thick) covering large portions of the dryland soils (more than 70% in untouched deserts of North America). BSCs are composed of fungi, cyanobacteria and microalgae, either as free-living organisms or as partners of lichen symbioses. Also symbiotic cyanobacteria-bryophyte associations can occur in well developed BSCs, which substantially contribute to available nitrogen increase in these extreme environments.

Biological nitrogen fixation is limited in arctic and alpine tundra as well in enormous extended boreal forests of Eurasia and North America. The majority of nitrogen accretion in these environments occurs through cyanobacteria either free-living or symbiotic in cyanolichens or associated with bryophytes. The boreal forests lack significant quantities of plants forming root symbioses with diazotrophic bacteria (referred to as N_2-fixing plants), with the exception of actinorhizal forms, among which species of *Alnus* and *Ceanothus* and of the coushion-forming dwarf shrubs *Dryas*. However, these fast-growing pioneer plants substantially contribute to nitrogen input only in early successional forests that develop after some form of disturbance, such as wildfire, windstorm or logging. In late successional and old-growth forests (more than 100–150 years since the last disturbance), instead, dominant sources of fixed nitrogen are communities of cyanobacteria (*Nostoc* spp.) associated with feather mosses (*Pleurozium schrebery* and *Hylocomium splendens*) that contribute approximatively 2 kg N ha^{-1} $year^{-1}$ (DeLuca *et al.*, 2002). The feather mosses, which account for as much as 95% of ground cover of the boreal forest floor, and their associated cyanobacteria may be the most broadly distributed N_2-fixing association in Earth and are considered the primary source of nitrogen fixation in boreal ecosystems (Gundale *et al.*, 2012) which are the second largest biome in the world.

Many overstory and understory species of N_2-fixing genera (e.g., *Alnus*, *Robinia*, *Ceanotus*, *Myrica*, *Lupinus*) can be present in temperate forests, accounting for high rates of nitrogen input in the environment (even more than 100 kg N ha^{-1} $year^{-1}$). However, like in boreal ones, these plants are prominent only in early successional stage of temperate forests. In mid and late successional forests the N_2-fixer plants become rare, being totally absent in old-growth forests. Failing N_2-fixing plants, the available nitrogen input into the ecosystem relies essentially on epiphytic cyanolichens (Antoine, 2004) and on nonsymbiotic N_2 fixation that can account for over 1 kg N ha^{-1} $year^{-1}$ and increases in magnitude toward mid, late and old-grown successional stages of forest. The nonsymbiotic N_2 fixation is carried out by heterotrophic bacteria living on decomposing leaf litters or woody debris of forest floor, where reduced forms of carbon are available as energy sources for the energetically expensive microbial process (Pérez *et al.*, 2010).

Biological nitrogen fixation rates are very high in tropical rain forests that may fix more N_2 than any other unmanaged ecosystem (even more than 200 kg N ha^{-1} $year^{-1}$). Legumes are abundant in many of these forests, representing in some of them more than 50% of all trees. However, although these plants can be a major source of nitrogen input into the environment, their abundance may lead to overestimate the potential for N_2-fixation. Not all leguminous trees, in fact, are able to establish an effective symbiosis with rhizobia or may not do so under natural conditions (Pons *et al.*, 2007). High legume nodulation levels and N_2 fixation rates occur in recently disturbed forests and in forests subjected to seasonal flooding, which have N-poor soils. Nodulation and N_2 fixation, instead, are virtually absent from N-rich soils of undisturbed late successional and old-growth forests, despite the abundance of N_2-fixing leguminous species. This is due to the ability of many symbiotic legumes to employ a strategy of facultative nitrogen fixation. These plants can induce N_2 fixation in N-poor environments and down-regulate it to low or negligible levels in response to increased soil nitrogen availability (Barron *et al.*, 2011). Therefore, the N_2-fixing trees can act to rapidly redress any local nitrogen deficiencies that develop within tropical landscapes. Despite the lack of nitrogen fixation by tree in N-rich environmental conditions, substantial amounts of nitrogen enter the tropical forest ecosystem through several other diazotrophic organisms. Heterotrophic nonsymbiotic bacteria living on organic litter layers that cover the forest floor can bring about an input of available nitrogen up to 12 kg N ha^{-1} $year^{-1}$. Canopy communities, dominated by autotrophic organisms, such as free-living cyanobacteria, cyanolichens and cyanobacteria-bryophyte associations can also contribute significant fluxes of nitrogen (up to 5 kg N ha^{-1} $year^{-1}$) via biological fixation to the tropical rain forest ecosystems (Reed *et al.*, 2008). Canopy cyanolichens and other epiphyte diazotrophic communities are a source of fixed nitrogen also in tropical dry forests, where the greatest contribution to available nitrogen input is given by actinorhizal tree species (e.g., those of genus *Casuarina*).

Leguminous (e.g., *Mimosa*, *Calliandra*, *Leucaena*, *Prosopis*) and actinorhizal (e.g., *Casuarina*) genera are widespread in xeromorphic and arid woodlands from North and South America to West Africa and *Ceanothus* shrubs are common in European and American mediterranean shrublands. All these N_2-fixing plants can greatly contribute to the nitrogen supply to their ecosystems (Cleveland *et al.*, 1999). However, also in these environments an additional input of fixed nitrogen is provided by other symbiotic and nonsymbiotic diazotrophic communities.

Legumes are not abundant in most grasslands and temperate savannas (from prairies of North America to pampas of South America and steppes of Eurasia) and contribute little available nitrogen to these environments that essentially rely on diazotrophic microorganisms of the soil (cyanobacteria and eterotrophic bacteria) as sources of fixed nitrogen. Conversely, N_2-fixing plants seem to be the largest source of available nitrogen in some tropical savannas. High proportions of legume trees, such as species of *Acacia*, are commonly present in tropical savannas that take up almost the half of the African continent, whereas Cycads (with symbiotic *Nostoc* spp. or *Anabaena* spp. in so called coralloid roots) form large populations in tropical savannas of Northern

Australia. An estimate of biological nitrogen fixation, however, may be difficult in some of these ecosystems, due to the fact that they are often managed and contain N_2-fixing forage plants (Cadish *et al.*, 1994).

Biological nitrogen fixation is an essential process in regulating biological productivity of freshwater and marine ecosystems that are most frequently under N-limited conditions also due to the worldwide presence of anammox bacteria whose activity accounts for up to 50% of the N_2 gas released from these environments (Jetten *et al.*, 2009). The major N_2-fixers in the aquatic ecosystems are cyanobacteria. They are present as planktonic or benthic forms of nonheterocystous filamentous cyanobacteria, unicellular cyanobacteria and free-living and symbiotic heterocystous filamentous cyanobacteria. Recently, a significant role in N_2 fixation has also been assigned to heterotrophic bacteria (Halm *et al.*, 2012).

Nitrogen fixation by planktonic cyanobacteria is rather low in oligotrophic lakes while the process may reach consistent levels in the mesotrophic and eutrophic ones. In these latter ecosystems the N_2 fixation depends on some environmental factors including light intensity and phosphorus concentration. Phosphorus is often the limiting nutrient in lakes and diazotrophic cyanobacteria become abundant in planktonic communities only when phosphorus levels increase, leading to low nitrogen/phosphorus ratios. Benthic diazotrophic cyanobacteria also contribute the input of available nitrogen into lakes. They can develop extensive mats of filamentous species (such as *Anabaena* spp. and *Oscillatoria* spp.) on sediments, or they can occur as epilithic peryphyton forms (such as *Nostoc* spp., *Calothrix* spp., and many others).

Extensive mats of diazotrophic cyanobacteria are also present in N-poor desert streams of western North America (Grimm and Petrone, 1997) which are characterized by warm temperature, high light, slow currents and ample supply of P. These environmental conditions support abundant cyanobacterial populations that can account for very high rates of N_2 fixation (up to 150 mg N m^{-2} day^{-1}).

High current velocity, low light or high turbidity, instead, can limit growth of benthic and planktonic cyanobacteria and N_2 fixation rate in flowing water systems, in most of which these diazotrophic microorganisms do not occur at all. Moreover, great turbulence, low availability of trace elements (in particular molybdenum) and also grazing by zooplankton (that breaks down diazotrophic filaments preventing the accumulation of enough photosynthetic cells to support the energetic requirement by heterocysts) are among the causes that negatively affect growth of cyanobacteria and N_2 fixation in most temperate estuaries (Marino *et al.*, 2006).

N_2-fixing heterotrophic bacteria have never been found in any of these freshwater ecosystems.

Oceans are oligotrophic environments that make up 71% of the Earth's surface. The South Pacific, which is the largest ocean system in the world, and the other tropical and subtropical oceanic gyres are regarded as biological deserts because of the extremely low availability of nutrients and minimum productivity. The only biological source of available nitrogen in all the oceanic ecosystems is the N_2 fixation carried out by different types of cyanobacteria and heterotrophic bacteria (more than 100 million tons N per year).

The most representative diazotrophs of North Atlantic Ocean, also abundant in the Arabian Sea and in the nearby Red Sea, are nonheterocystous filamentous cyanobacteria of genus *Trichodesmium*, which form large surface colonies at water temperature above 25 °C. In North Pacific Ocean and in equatorial Pacific, instead, the dominant diazotrophs are small (<10 μm) unicellular coccoid cyanobacteria (e.g., *Cyanoteche* spp. and *Crocosphaera* spp.) together with heterotrophic bacteria (Halm *et al.*, 2012). Heterocystous filamentous cyanobacteria (e.g., *Nodularia* spp. and *Aphanizomenon* spp.) are the N_2 fixers living in the cooler brackish waters of the Baltic Sea.

Although these free-living cyanobacteria are regarded as dominant diazotrophs in the oceanic environments, other N_2-fixers also deserve consideration. Most notably, cyanobacteria-diatom symbioses capable of high N_2 fixation rates, which are widely distributed through the warm oceans. Some genera of diatoms (e.g., *Rhizosolenia*, *Hemiaulus* and *Guinardia*), which are common members of the phytoplankton communities, form symbiotic associations with heterocystous filamentous cyanobacteria (e.g., *Richelia intracellularis* and *Calothrix* spp.) and represent a significant component of the nitrogen budget in these ecosystems (Foster *et al.*, 2011).

Planktonic diazotrophs are the only N_2 fixers living in open oceans, while in coastal marine environments, such as salth marshers, intertidal zones and coral reefs, benthic filamentous cyanobacteria (e.g., heterocystous *Calothrix* and *Anabaena* and nonheterocystous *Lyngbya* and *Oscillatoria* genera) colonize sediments forming extensive mats that make substantial contributions to nitrogen supply to the ecosystem.

The highest rates of N_2 fixation in coastal environments are exhibited by the coral reef communities. Most of the fixed nitrogen entering the coral reef is provided by both free-living diazotrophs and benthic cyanobacteria that cover large areas of reef substratum. Additionally, several important members of the reef community, including sponges and corals, have also the capability to fix nitrogen through symbiotic associations with N_2-fixing cyanobacteria and eterotrophic bacteria. Corals, in particular, are holobionts, with the coral animals that harbor a variety of microorganisms, including endosymbiotic dinoflagellates (*Symbiodinium* spp.), commonly referred as zooxanthellae, which provide over 95% of fixed carbon to the hosts. In addition, corals harbor endosymbiotic diazotrophic bacteria and cyanobacteria (e.g., *Synechococcus* and *Prochlorococcus* strains) that benefit with fixed nitrogen both zooxanthellae and coral hosts which possess enzymes enabling ammonium assimilation (Yellowlees *et al.*, 2008).

Interestingly, the symbiotic diazotrophic bacteria of many coral reefs, including the Great Barrier Reef, are closely related to bacterial species belonging to the order *Rhizobiales* (Lema *et al.*, 2012). It is still unclear how the coral rhizobia might be protected against high concentrations of oxygen arising from the zooxanthella photosynthesis in host tissues. Possibly the diazotrophic communities are sheltered in oxygen-depleted coral microhabitats.

See also: Ecological Processes: Ammonification. Global Change Ecology: Nitrogen Cycle

References

Alloisio, N., Qeiroux, C., Fournier, P., *et al.*, 2010. The *Frankia alni* symbiotic transcriptome. Molecular Plant-Microbe Interactions 23, 593–607.

Antoine, M.E., 2004. An ecophysiological approach to quantifying nitrogen fixation by *Lobaria oregano*. Bryologist 107, 82–87.

Barron, A.R., Purves, D.W., Hedin, L.O., 2011. Facultative nitrogen fixation by canopy legumes in a lowland tropical forest. Oecologia 165, 511–520.

Benson, D.R., Silvester, W.B., 1993. Biology of *Frankia* strains, actinomycete symbionts of actinorhizal plants. Microbiological Reviews 57, 293–319.

Berry, A.M., Harriott, O.T., Moreau, R.A., *et al.*, 1993. Hopanoid lipids compose the *Frankia* vesicle envelope, presumptive barrier of oxygen diffusion to nitrogenase. Proceedings of the National Academy of Sciences of the USA 90, 6091–6094.

Bocchi, S., Malgioglio, A., 2010. *Azolla-Anabaena* as biofertilizer for rice paddy fields in the Po Valley, a temperate rice area in Northern Italy. International Journal of Agronomy 2010, 1–5.

Büdel, B., Darienko, T., Deutshewitz, K., *et al.*, 2009. Southern African Biological soil crusts are ubiquitous and highly diverse in drylands, being restricted by rainfall frequency. Microbial Ecology 57, 229–247.

Cadish, G., Schunke, R.N., Giller, K.E., 1994. Nitrogen cycling in a pure grass pasture and a grass-legume mixture of a red latosol in Brazil. Tropical Grasslands 28, 43–52.

Chapman, M.J., Margulis, L., 1998. Morphogenesis by symbiogenesis. International Microbiology 1, 319–326.

Cheng, J., Hipkin, C.R., Gallon, J.R., 1999. Effects of inorganic nitrogen compounds on the activity and synthesis of nitrogenase in *Gloeoteche* (Nägeli) sp. ATCC27152. New Phytologist 141, 61–70.

Ciccolella, C.O., Raynard, N.A., Mei, J.H., Church, D.C., Ludwig, R.A., 2010. Simbiotic legume nodules employ both rhizobial exo and endo-hydrogenases to recycle hydrogen produced by nitrogen fixation. PloS One 5, e12094.

Cleveland, C.C., Townsend, A.R., Schimel, D.S., *et al.*, 1999. Global pattern of terrestrial biological nitrogen (N_2) fixation in natural ecosystems. Global Biogeochemical Cycles 13, 623–645.

Cocking, E.C., 2003. Endophytic colonization of plant roots by nitrogen-fixing bacteria. Plant and Soil 252, 169–175.

Cumino, A.C., Marcozzi, C., Barreiro, R., Salerno, G.R., 2007. Carbon cycling in *Anabaena* sp. Pcc 7120. Sucrose synthesis in the heterocysts and possible role in nitrogen fixation. Plant Physiology 143, 1385–1397.

de Vera, J.-P., 2012. Lichens as survivors in space and on Mars. Fungal Ecology 5, 472–479.

DeLuca, T.H., Zackrlsson, O., Nillson, M.-C., Sellstedt, A., 2002. Quantifying nitrogen-fixation in feather moss carpets of boreal forests. Nature 419, 917–920.

Dixon, R.O.D., Wheeler, C.T., 1986. Nitrogen fixation in plants. New York: Chapman and Hall.

Downie, J.A., 2005. Legume haemoglobins: symbiotic nitrogen fixation needs bloody nodules. Current Biology 15, R196–R198.

Ferguson, B.J., Indrasumunar, A., Hayashi, S., *et al.*, 2010. Molecular analysis of legume nodule development and autoregulation. Journal of Integrative Plant Biology 52, 61–76.

Fleming, A.L., Wittenberg, J.B., Wittenber, B.A., Dudman, W.F., Appleby, C.A., 1987. The purification, characterization and ligand-binding kinetics of hemoglobins from root nodules of the non-leguminous *Casuarina glauca-Frankia* symbiosis. Biochimica et Biophysica Acta 911, 209–220.

Foster, R.A., Kuyper, M.M.M., Vagner, T., *et al.*, 2011. Nitrogen fixation and transfer in open ocean diatom-cyanobacterial symbioses. International Society for Microbial Ecology 5, 1484–1493.

Golden, S.S., Ishiura, M., Hirschie, J.C., Kondo, T., 1997. Cyanobacterial circadian rhythms. Annual Review of Plant Physiology and Plant Molecular Biology 48, 327–354.

Grimm, N.B., Petrone, K.C., 1997. Nitrogen fixation in a desert stream ecosystem. Biogeochemistry 37, 33–61.

Gundale, M.J., Nilsson, M., Bansal, S., Jäderlund, A., 2012. The interactive effects of temperature and light on biological nitrogen fixation in boreal forests. New Phytologist 194, 453–463.

Halm, H., Lam, P., Ferdelman, T.G., *et al.*, 2012. Heterotrophic organisms dominate nitrogen fixation in the South Pacific Gyre. International Society for Microbial Ecology 6, 1238–1249.

Helber, J.T., Johnson, T.R., Yarbrough, L.R., Hirschberg, R., 1988. Effect of nitrogenous compounds on nitrogenase gene expression in anaerobic cultures of *Anabaena variabilis*. Journal of Bacteriology 170, 558–563.

Herridge, D.F., Peoples, M.B., Boddey, R.M., 2008. Global imput of biological nitrogen fixation in agricultural systems. Plant and Soil 311, 1–18.

Hocher, V., Alloisio, N., Auguy, F., *et al.*, 2011a. Transcriptomics of actinorhizal symbioses reveals homologs of the whole common symbiotic signalling cascade. Plant Physiology 156, 700–711.

Hocher, V., Alloisio, N., Bogusz, D., Normand, P., 2011b. Early signaling in actinothizal symbioses. Plant Signaling and Behavior 6, 1377–1379.

Huss-Danell, K., 1997. Actinorhizal symbioses and their N_2 fixation. New Phytologist 136, 375–405.

Igarashi, R.Y., Seefeldt, L.C., 2003. Nitrogen fixation: the mechanism of the Mo-dependent nitrogenase. Critical Reviews in Biochemistry and Molecular Biology 38, 351–384.

James, E.K., Gyaneshwar, P., Mathan, N., *et al.*, 2002. Infection and colonization of rice seedlings by the plant growth-promoting bacterium *Herbaspirillum seropedicae* Z67. Molecular Plant-Microbe Interactions 15, 894–906.

Jetten, M.S.M., van Niftrik, L., Strous, M., *et al.*, 2009. Biochemistry and molecular biology of anammox bacteria. Critical Reviews in Biochemistry and Molecular Biology 44, 65–84.

Jeudy, C., Ruffel, S., Freixes, S., *et al.*, 2009. Adaptation of *Medicago trunculata* to nitrogen limitation is modulated via local and systemic nodule developmental responses. New Phytologist 185, 817–828.

Karl, D., Michaels, A., Bergman, B., *et al.*, 2002. Dinitrogen fixation in the world's oceans. Biogeochemistry 57/58, 47–98.

Kennedy, I.R., Choudhuri, A.T.M.A., Kecskés, M.L., 2004. Non-symbiotic bacterial diazotrophs in crop-farming systems: can their potential for plant growth promotion be better exploited? Soil Biology and Biochemistry 36, 1229–1244.

Koponen, P., Nygren, P., Domenach, A.M., *et al.*, 2003. Nodulation and dinitrogen fixation of legume trees in a tropical freshwater swamp forest in French Guiana. Journal of Tropical Ecology 19, 655–666.

Kouchi, H., Imaizumi-Anraku, H., Hayashi, M., *et al.*, 2010. How many peas in a pod? Legume gene responsible for mutualistic symbioses underground. Plant and Cell Physiology 51, 1381–1397.

Kreibich, H., de Camargo, P.B., Moreira, M.Z., Victoria, R.L., Werner, D., 2006. Estimation of symbiotic N_2 fixation in an Amazon floodplan forest. Oecologia 147, 359–368.

Lechno-Yossef, S., Nierzwicki-Bauer, S.A., 2002. Azolla-Anabaena symbiosis. In: Rai, A.N., Bergman, B., Rasmussen, U. (Eds.), Cyanobacteria in Symbiosis. Dordrecht: Kluwer Academic, pp. 179–193.

Lema, K.A., Willis, B., Bourne, D.G., 2012. Corals form characteristic associations with symbiotic nitrogen-fixing bacteria. Applied and Environmental Microbiology 78, 3136–3144.

Leul, M., Normand, P., Sellstedt, A., 2009. The phylogeny of uptake hydrogenases in *Frankia*. International Microbiology 12, 23–28.

Marino, R., Chan, F., Howarth, R.W., Pace, M.L., Likens, G.E., 2006. Ecological constraints of planktonic nitrogen fixation in saline estuaries. I. Nutrient and trophic controls. Marine Ecology Progress Series 309, 25–39.

Marks, J., 2004. The roots of plant-microbe collaborations. Science 304, 234–236.
Martin-Figueroa, E., Navarro, E., Florencio, F.J., 2000. The GS-GOGAT pathway is not operative in the heterocists. Cloning and expression of *glsF* gene from the cyanobacterium *Anabaena* sp. PCC 7120. FEBS Letters 476, 282–286.
Meeks, J.C., 2003. Symbiotic interactions between *Nostoc punctiforme*, a multicellular cyanobacterium, and the hornwort *Anthoceros punctatus*. Symbiosis 35, 55–71.
Meeks, J.C., Elhai, J., 2002. Regulation of cellular differentiation in filamentous cyanobacteria in free-living and plant-associated symbiotic growth states. Microbiology and Molecular Biology Reviews 66, 94–121.
Mullineaux, C.W., Mariscal, V., Nenninger, A., *et al*., 2008. Mechanism of intercellular molecular exchange in heterocyst-forming cyanobacteria. EMBO Journal 27, 1299–1308.
Nagatani, H., Shimizu, M., Valentine, R.C., 1971. The mechanism of ammonia assimilation in nitrogen fixing bacteria. Archives of Microbiology 79, 164–175.
Newton, W.E., 2007. Physiology, biochemistry and molecular biology of nitrogen fixation. In: Bothe, H., Ferguson, S.J., Newton, W.E. (Eds.), Biology of the Nitrogen Cycle. Amsterdam: Elsevier, pp. 109–130.
Nicolaisen, K., Hahn, A., Schleiff, E., 2009. The cell wall in heterocyst formation by *Anabaena* sp. PCC 7120. Journal of Basic Microbiology 49, 5–24.
Pabby, A., Prasanna, R., Singh, P.K., 2003. Azolla-Anabaena symbiosis. From traditional agriculture to biotechnology. Indian Journal of Biotechnology 2, 26–37.
Papaefthimiou, D., Hrouzek, P., Mugnai, M.A., *et al*., 2008. Differential patterns of evolution and distribution of the symbiotic behaviour in nostocacean cyanobacteria. International Journal of Systematic and Evolutionary Microbiology 58, 553–564.
Pérez, C.A., Carmona, M.R., Armesto, J.J., 2010. Non-symbiotic nitrogen fixation during leaf litter decomposition in an old-growth temperate rain forest of Chiloé Island, southern Chile: effects of single versus mixed species litter. Austral Ecology 35, 148–156.
Peters, G.A., Perkins, S.K., 1993. The *Azolla-Anabaena* symbiosis: endophyte continuity in the *Azolla* life cycle is facilitated by epidermal trichomes. II. Reestablishment of the symbiosis following gametogenesis and embryogenesis. New Phytologist 123, 65–75.
Pons, T.L., Perrejin, K., van Kessel, C., Werger, M.J.A., 2007. Symbiotic nitrogen fixation in a tropical rainforest: ^{15}N natural abundance measurements supported by experimental isotopic enrichment. New Phytologist 173, 154–167.
Raggio, J., Pintado, A., Ascaso, C., *et al*., 2011. Whole lichen thalli survive exposure to space conditions: results of Lithopanspermia experiment with *Aspicilia fruticulosa*. Astrobiology 11, 281–292.
Rai, A.N., Söderbäck, E., Bergmann, B., 2000. Cyanobacterium-plant symbioses. New Phytologist 147, 449–481.
Reed, S.C., Cleveland, C.C., Townsend, A.R., 2008. Tree species control rates of free-living nitrogen fixation in a tropical rain forest. Ecology 89, 2924–2934.
Sandh, G., El-Shehawy, R., Díez, B., Bergman, B., 2009. temporal separation of cell division and diazotrophy in the marine diazotrophic cyanobacterium *Trichodesmium erythraeum* IMS101. FEMS Microbiology Letters 295, 281–288.
Soltis, D.E., Soltis, P.S., Morgan, D.R., *et al*., 1995. Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen fixation in angiosperms. Proceedings of the National Academy of Sciences of the USA 92, 2647–2651.
Spaink, H.P., 2000. Root nodulation and infection factors produced by rhizobial bacteria. Annual Reviews of Microbiology 54, 257–288.
Stewart, K.J., Lamb, E.G., Coxon, D.S., Siciliano, S.D., 2011. Bryophyte-cyanobacterial associations as a key factor in N_2-fixation across the Canadian Arctic. Plant and Soil 344, 335–346.
Tamagnini, P., Leitão, E., Oliveira, P., *et al*., 2007. Cyanobacterial hydrogenases: diversity, regulation and applications. FEMS Microbiology Reviews 31, 692–720.
Toepel, J., Welsh, E., Summerfield, T.C., Pakrasi, H.B., Sherman, L.A., 2008. Differential transcriptional analysis of the cyanobacterium *Cyanothece* sp. strain ATCC 51142 during light–dark and continuous-light growth. Journal of Bacteriology 190, 3904–3913.
Urquiaga, S., Cruz, K.H.S., Boddey, R.M., 1992. Contribution of nitrogen fixation to sugarcane: nitrogen-15 and nitrogen balance estimates. Soil Science Society of America Journal 56, 105–114.
Valladares, A., Maldener, I., Muro-Pastor, A.M., Flores, E., Herrero, A., 2007. Heterocyst development and diazotrophic metabolism in the terminal respiratory oxidase mutants of the cyanobacterium *Anabaena* sp. Strain PCC 7120. Journal of Bacteriology 189, 4425–4430.
van Hove, C., Lejeune, A., 2002. Applied aspects of *Azolla-Anabaena* symbiosis. In: Rai, A.N., Bergman, B., Rasmussen, U. (Eds.), Cyanobacteria in symbiosis. Dordrecht: Kluwer Academic, pp. 179–193.
Wall, L.G., 2000. The actinorhizal symbiosis. Journal of Plant Growth Regulation 19, 167–182.
Wall, L.G., Valverde, C., Huss-Danell, K., 2003. Regulation of nodulation in the absence of N_2 is different in actinorhizal plants with different infection pathways. Journal of Experimental Botany 54, 1253–1258.
White, J., Prell, J., James, E.K., Poole, P., 2007. Nutrient sharing between simbionts. Plant Physiology 144, 604–614.
Wolk, C.P., 1996. Heterocyst formation. Annual Review of Genetics 30, 59–78.
Yellowlees, D., Alwyn, T., Rees, V., Leggat, W., 2008. Metabolic interactions between algal symbionts and invertebrate hosts. Plant, Cell and Environment 31, 679–694.

Further Reading

Adams, D.G., Duggan, P.S., 2008. Cyanobacteria-bryophyte symbioses. Journal of Experimental Botany 59, 1047–1058.
Adams, D.G., Bergman, B., Nierzwicki-Bauer, S.A., Rai, A.N., Schüßler, A., 2006. Cyanobacterial plant symbioses. In: Dworkin, M., Falkow, S., Rosengerg, E., Schleifer, K.-H., Stackebrandt, E. (Eds.), The prokariotes: a handbook on the biology of bacteria, Symbiotic associations, biotechnology, applied microbiology, 3rd edn., 1. New York: Springer, pp. 331–363.
Benson, D.R., Brooks, J.M., Huang, Y., *et al*., 2011. The Biology of *Frankia* sp. strains in post-genome era. Molecular Plant-Microbe Interactions 24, 1310–1316.
Berry, A.M., Mendoza-Herrera, A., Guo, Y.-Y., *et al*., 2011. New perspectives on nodule nitrogen assimilation in actinothizal symbioses. Functional Plant Biology 38, 645–652.
Boys, E.S.M., Hamiltion, T.L., Peters, J.W., 2011. An alternative path for the evolution of biological nitrogen fixation. Frontiers in Microbiology 2, 205.
Church, J.G., Short, C.M., Jenkins, B.D., Karl, D.M., Zehr, J.P., 2005. Temporal patterns of nitrogenase gene (*nifH*) expression in the oligotrophic North Pacific Ocean. Applied and Environmental Microbiology 71, 5362–5370.
Daalsgaard, T., Thamdrup, B., Canfield, D.E., 2005. Anaerobic ammonium oxidation (anammox) in the marine environment. Research in Microbiology 156, 457–464.
Fiore, C.L., Jarret, J.K., Olson, N.D., Lesser, M.P., 2010. Nitrogen fixation and nitrogen transformations in marine symbioses. Trends in Microbiology 18, 455–463.
Franche, C., Lindström, K., Elmerich, C., 2009. Nitrogen-fixing bacteria associated with leguminous and non-leguminous plants. Plant and Soil 321, 35–59.
Hedin, L.O., Brookshire, E.N.J., Menge, D.N.L., Barron, A.R., 2009. The nitrogen paradox in tropical forest ecosystems. Annual Review of Ecology, Evolution, and Systematics 40, 613–635.
Herridge, D.F., Peoples, M.B., Boddey, R.M., 2008. Global input of biological nitrogen fixation in agricultural systems. Plant and Soil 311, 1–18.
Hirsch, A.M., Lum, M.R., Downie, J.A., 2001. What makes the rhizobia–legume symbiosis so special? Plant Physiology 127, 1484–1492.
James, E.K., 2000. Nitrogen fixation in endophytic and associative symbiosis. Field Crops Research 65, 197–209.
Johnson, C.H., Mori, T., Xu, Y., 2008. A cyanobacterial circadian clockwork. Current Biology 18, R816–R825.
Jones, K.M., Kobayashi, H., Davies, B.W., Taga, M.E., Walker, G.C., 2007. How rhizobial symbionts invade plants: the *Sinorhizobium-Medicago* model. Nature Reviews Microbiology 5, 619–633.
Kereszt, A., Mergaert, P., Kondorosi, E., 2011. Bacteriod development in legume nodules: evolution of mutual benefit or of sacrifical victims? Molecular Plant-Microbe Interactions 24, 1300–1309.

Kumar, K., Mella-Herrera, R.A., Golden, J.W., 2009. Cyanobacterial heterocysts. Cold Spring Harbor Perspectives in Biology 2, a000315.
Magnani, G.S., Didonet, C.M., Cruz, L.M., *et al*., 2010. Diversity of endophytic bacteria in brazilian sugarcane. Genetics and Molecular Research 9, 250–258.
Markmann, K., Parniske, M., 2009. Evolution of root endosymbiosis with bacteria: how novel are nodules? Trends in Plant Science 14, 77–86.
Massena Reis, V., Baldani, J.I., Divan Baldani, V.L., Dobereiner, J., 2000. Biological dinitrogen fixation in gramineae and palm trees. Critical Reviews in Plant Sciences 19, 227–247.
Murray, J.D., 2011. Invasion by invitation: rhizobial infection in legumes. Molecular Plant-Microbe Interactions 24, 631–639.
Oldroyd, G.E.D., Downie, J.A., 2008. Coordinating nodule morphogenesis with rhizobial infection in legumes. Annual Review of Plant Biology 59, 519–546.
Pawlowski, K., Demchenko, K.N., 2012. The diversity of actinorhizal symbiosis. Protoplasma 249, 967–979.
Pawlowski, K., Sirrenberg, A., 2003. Symbiosis between *Frankia* and actinorhizal plants: root nodules of non-legumes. Indian Journal of Experimental Biology 41, 1165–1183.
Pawlowski, K., Bogusz, D., Ribeiro, A., Berry, A.M., 2011. Progress on research on actinorhizal plants. Functional Plant Biology 38, 633–638.
Perrine-Walker, F., Gherbi, H., Imanishi, L., *et al*., 2011. Symbiotic signaling in actinorhizal symbioses. Current Protein and Peptide Science 12, 156–164.
Reid, D.E., Ferguson, B.J., Hayashi, S., Lin, Y.-H., Gresshoff, P.M., 2011. Molecular mechanisms controlling legume autoregulation of nodulation. Annals of Botany 108, 789–795.
Schumpp, O., Deakin, W.J., 2010. How inefficient rhizobia prolong their existence within nodules. Trends in Plant Science 15, 189–195.
Seefeldt, L.C., Hoffman, B.M., Dean, D.R., 2009. Mechanism of Mo-dependent nitrogenase. Annual Review of Biochemistry 78, 701–722.
Sessitsch, A., Hiwieson, J.G., Perret, X., Antoun, H., Martinez-Romero, E., 2003. Advances in *Rhizobium* research. Critical Reviews in Plant Sciences 21, 323–378.
Sohm, J.A., Webb, E.A., Capone, D.G., 2011. Emerging patterns of marine nitrogen fixation. Nature Reviews Microbiology 9, 499–508.
Taulé, C., Mareque, C., Barlocco, C., *et al*., 2012. The contribution of nitrose fixation to sugarcane (*Saccharum officinarum* L.), and the identification and characterization of part of the associated diazotrophic bacterial community. Plant and Soil 356, 35–49.
Vitousek, P.M., Cassman, K., Cleveland, C., *et al*., 2002. Towards an ecological understanding of biological nitrogen fixation. Biogeochemistry 57, 1–45.
Wasson, M.F., 1999. Global patterns of terrestrial biological nitrogen (N_2) fixation in natural ecosystems. Global Biogeochemical Cycles 13, 623–645.
Zhang, C.C., Laurent, S., Sakr, S., Peng, L., Bédu, S., 2006. Heterocyst differentiation and pattern formation in cyanobacteria: a chorus of signals. Molecular Microbiology 59, 367–375.

Decomposition and Mineralization☆

L Wang, Indiana University-Purdue University, Indianapolis (IUPUI), Indianapolis, IN, USA
P D'Odorico, University of Virginia, Charlottesville, VA, USA

Introduction

Decomposition can be considered the inverse process of production, because it is a metabolic degradation of organic matter (such as plant residues, animal tissues and microbial material) into simple organic and inorganic compounds. Decomposition is essentially a process of breakdown of the carbon skeleton existing in the organic compounds with consequent liberation of energy. Decomposition is an important component in global carbon cycle: without decomposition, the atmospheric CO_2 pool could be depleted literally in one decade based on the current annual rates of net photosynthesis (without considering the effect of biological feedbacks). Moreover, soil organic matter is one of the largest and most dynamic reservoirs of carbon in the global carbon cycle. The carbon stored in soil organic matter is 2400 petagrams (Pg, 10^{15} g), which is almost twice as much the combined amount stored in vegetation (550 Pg) and atmosphere (750 Pg). A better understanding of processes involved in the dynamics of soil organic matter is crucial to predict future changes in atmospheric CO_2 concentrations. Decomposition and the subsequent mineralization are also an indispensable process for sustaining life on Earth, as they are the only processes enabling massive recycling of chemical elements in the biosphere. Most nutrients cycle from an inorganic form in the soil solution to vegetation and back to the soil solution through decomposition and subsequent mineralization. Mineralization is the conversion of nutrients and other substances from an organically bound form to a water-soluble inorganic form. Decomposition and mineralization are closely related processes. In fact, mineralization is often considered as part of the decomposition process, however, decomposition does not always lead to mineralization. Decomposition is associated with carbon cycling whereas mineralization with nutrient cycling. Part of decomposition processes such as fragmentation and chemical alteration could be classified as mineralization, if inorganic nutrients or other simple bases (e.g., $-NH_3$, $(PO_4)^{3-}$) are released from the complex organic compounds during these processes. Mineralization is a vital process in ecosystem dynamics since most plant essential nutrients (such as nitrogen, phosphorus and sulfur) are made available to plant uptake through the process of mineralization (Figure 1).

Mechanisms and Processes

Overview

Decomposition is comprised of a series of interacting physical, chemical and biological processes. In general, three major processes are involved in terrestrial decomposition: leaching, fragmentation and chemical alteration. Leaching is a physical process through which ions (such as K^+, Mg^{2+} and Ca^{2+}) and small water-soluble organic compounds (such as sugars, amino acids and amino sugars) dissolve in water and move out of the decomposed organic material. Leaching could happen even from green leaves still

☆*Change History:* March 2013. L Wang and P D'Odorico updated all parts of the text.

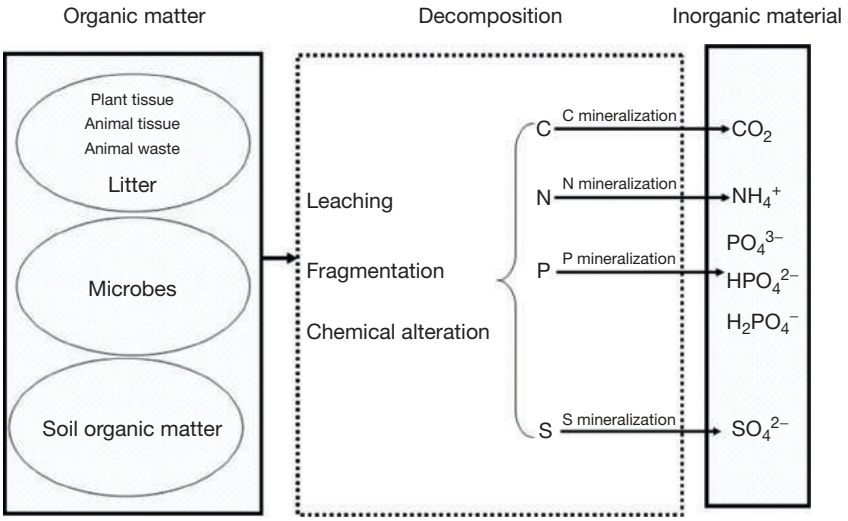


Figure 1 The process of decomposition and mineralization.

attached to the plants. These soluble materials move into the soil matrix where they are taken up by plant roots or soil microbes, adsorbed to soil minerals, or leached and transported through the soil column by water drainage. Leaching losses are greatest in environments with high precipitation and negligible in dry environments. Fragmentation is a physical process through which fresh detritus is broken down into smaller pieces. During fragmentation some chemical bases can break off from organic compounds thereby contributing to nutrient mineralization. Fragmentation also provides more fresh surfaces that can be used by microbial colonization thereby facilitating further decomposition. Biological factors are major contributors of litter fragmentation. Fragmentation is a byproduct of the feeding activities of larger animals and the direct product from the feeding of smaller organisms such as protozoan, potworm and earthworms. Abiotic factors such as freezing-thawing and wetting-drying cycles can also facilitate litter fragmentation. Through chemical alteration, the final stage of the decomposition process, litter fragments are further broken down into simple organic and inorganic compounds. The complete decomposition, i.e., the release of all the energy fixed in organic compounds, may take thousands of years if it happens at all. One of the most commonly known products of incomplete decomposition is fossil fuel, on which our modern societies heavily rely.

Carbon mineralization is about the same as carbon decomposition if we exclude the leaching process. Nitrogen (N) and phosphorus (P) mineralization, however, although related to decomposition, are distinct processes. In particular, P mineralization is less tightly linked to decomposition than N mineralization due to the chemical structure of P-containing compounds. In fact, because the N atoms are directly bonded to carbon skeletons of organic matter (C-N), N is generally released as dissolved organic N (DON) in the breakdown of these skeletons, i.e., in the course of the decomposition process (e.g., during fragmentation or chemical alteration). On the other hand, because P atoms can form ester linkages C-O-P, P can be released (i.e., mineralized) independently of the decomposition of organic matter, i.e., without break down of the carbon skeleton. N mineralization starts with the release of DON associated with decomposition. Both plants (through the mycorrhizal fungi associated with plant roots) and soil microbes can take up DON, although in most cases soil microbes out compete plants for DON uptake. When microbial needs for DON are met, microbes break down the remaining DON – using the energy released by the break down of the carbon skeleton and secrete NH_4^+ to the surrounding soil matrix. This process is known as "N mineralization". When DON is insufficient to meet the microbial N requirement, soil microbes absorb additional N from the pool of inorganic N (e.g. NH_4^+, NO_3^-) in the soil solution, a process known as "immobilization". Immobilization also includes the removal of inorganic N from the soil solution by chemical fixation.

Decomposers

Fungi are an important class of decomposers due to their abundance and ability to decompose rather recalcitrant organic material. In fact, fungi are able to secrete enzymes that are capable of breaking down virtually all classes of plant compounds. Thus, fungi can decompose substrates such as fresh plant litter and some structural materials (e.g. lignin, chitin and keratin) that are initially almost inaccessible to other decomposers. Moreover, fungi account for a large fraction of the soil microbial biomass, as they contribute to about 60–90% of the microbial biomass in forest soils and to 50% in grassland soils. Fungi have extensive hyphae networks, which make it possible to acquire carbon (e.g. from forest litter) and nutrients (e.g. from mineral soil) from different locations. Mycorrhizae are symbiotic associations between plant roots and fungi. Based on conservative estimates, approximately 95% of all vascular plant species have the potential to form this mutualistic association with mycorrhizal fungi. Approximately 70% of all plant families include species, which develop specialized endomycorrhizae called vesicular-arbuscular mycorrhizae (VAM) or just arbuscular mycorrhizae (AM). In these associations, plants receive nutrients from fungi – especially the less mobile groups (e.g.,

phosphates), while providing, in return, fungi with carbohydrates. Mycorrhizal fungi play a role in the decomposition process by breaking down proteins into amino acids. Under certain conditions, mycorrhizal fungi have been found to turn into aggressive decomposers capable of decomposing humus that used to be considered stabilized.

Bacteria are another major group of decomposers. Like fungi, bacteria spores are ubiquitous in air, water, and both dead and live organic matter. There is a wide range of types of soil bacteria. Recent studies have shown that bacteria are able to degrade cellulose/hemicellulose, lignin and even intact fiber walls. Due to their small size and large surface to volume ratio, bacteria are able to rapidly absorb soluble substrates and to reproduce quickly in substrate-rich conditions. In substrate-rich environments such as the rhizosphere or dead animal carcasses, bacteria tend to undergo population "explosions", thereby become the dominant decomposers. These populations collapse as the freely available resources are consumed. Bacteria decomposition seems to be more common in situations where fungi are under stress. Bacteria have also been found to degrade substrates resistant to fungal decay.

Soil fauna can be classified into three categories based on the size – microfauna (less than 0.1 mm), mesofauna (between 100 μm and 2 mm) and macrofauna (between 2 mm and 20 mm), though some researchers adopt slightly different size criteria. Soil fauna used to be considered an important contributor to litter decomposition, however, it was later recognized that, soil microorganisms such as fungi and bacteria are the dominant functional groups of decomposers. In fact, complex organic polymers, such as lignin, can be degraded exclusively by these microorganisms. Soil fauna affect decomposition through the processes of litter fragmentation, bacteria/fungi grazing and soil structure alteration. They graze either directly on microorganisms or on dead organic matter inhabited by bacteria and fungi. At the same time, soil fauna spread the populations and increase the turnover rate of microbial communities, thus enhancing the rates of organic matter decomposition. Therefore, despite its limited direct participation to the decomposition process, the overall influence of soil fauna on the turnover of soil organic matter should not be underestimated.

Temporal and Spatial Patterns

Litter mass decreases exponentially with time in the course of the decomposition process. In general, leaf litter typically loses 30–70% of its mass in the first year and another 20–30% in the following 5–10 years. Only a few of the important chemical changes occurring during decomposition have been completely understood, while most of these reactions still need to be investigated. The organic compounds in plant tissues can be grouped into several broad classes on the basis of their different rates of decomposition. These organic compounds are typically grouped (from fast to slow decomposition) into 1) sugars, starches and simple proteins, 2) crude proteins, 3) hemicellulose, 4) cellulose, 5) fats and waxes, and 6) lignin and phenolic compounds. A three stage model is often used to describe the chemical changes and rate-regulating factors contributing to decomposition. These stages include an early, a late, and a near-humus stage. At the early stage, the decomposition of water-soluble substances and unshielded hemicellulose/cellulose is stimulated by high levels of nutrients; at the late stage, the degradation rate of lignin determines the rate of litter decomposition, while higher N level decrease the decomposition rates; in the near-humus stage, the rate of litter decomposition tends to zero, while the total amount and the lignin content of the residual soil organic material remain constant. An alternative commonly accepted three-phase model of litter decomposition, considers the following phases: 1) leaching of cell solubles; 2) fragmentation and chemical alteration; this second phase occurs slowly and includes most of the fragmentation and alteration of the litter structure; and 3) chemical alteration of litter detritus mixed with mineral soils; in this final phase decomposition occurs at a slower pace.

Climate determines regional patterns of decomposition. In general, regions with higher temperature and water availability tend to experience higher decomposition rates. However, decomposition is limited under water logging conditions typical of wetland environments. At a given location, most decomposition occurs near the ground surface, where soils are substrate-rich, due to the presence of relatively high amounts of leaf and root litter.

Controlling Factors

Decomposition and mineralization rates are controlled by both biotic and abiotic factors. Biotic factors include substrate quantity and quality (i.e., nutrient content) available to decomposers, and the type and size of microbial community; abiotic factors include variables such as soil temperature, moisture, and pH, which determine the soil environmental conditions.

Biotic factors

Substrate quality and quantity

Substrate quantity and quality are major biotic factors controlling the rates of decomposition. Litter quality is determined by three general characteristics: 1) the type of chemical bonds present in the organic compounds, 2) the amount of energy released by their decay, and 3) the size and structure of these compounds and their nutrient content. Glucose and other simple sugars have high carbon quality for microbial decomposer, followed by cellulose and hemicellulose. Lignin a structural compound second only to cellulose in quantitative importance in most plant tissues. Lignin content ranges from 2% to more than 50% in plant dry weight. These polyphenols dramatically slow the decomposition and mineralization rates. The quality of organic matter is generally expressed (especially in biogeochemical models) in terms of the C/N ratios of litter, soil organic matter, and microbial biomass, though the mechanistic role of C/N ratios in the decomposition and mineralization processes is still not completely understood.

Because microbes are generally more N-limited than carbon limited, lower C/N ratios in plant residues usually lead to higher decomposition rates. The C/N ratios of plant residues range from 10:1 to 100:1. The C/N ratios of soil organic matter remain rather constant with a typical value of 10. The C/N ratios of microbial biomass range from 5 (e.g., tropical arable soil) to 10 (e.g., tropical dry forest). Microbial biomass is the most readily decomposable pool of organic material due to the simple structure and high quality of both carbon and nutrients. The microbial decomposition rates are followed by those of plant litter and soil organic matter. In most systems, if the C/N ratio of soil organic matter exceeds about 25:1, net N immobilization (i.e., microbes holding nitrogen instead of releasing it) occurs instead of net mineralization.

Microbial community composition

The composition of microbial communities is another important factor determining the rates of decomposition due to the different type and rate of enzyme production in different microbial communities. These enzymes are major players in break-down of different classes of substrates. In addition, different microbial decomposers have different tolerance to soil moisture and temperature conditions, with consequent effects on the rates of decomposition and mineralization. For example, because fungi are usually less sensitive to water stress than bacteria, they may play a more important role in organic matter decomposition in arid and semi-arid environments.

Abiotic factors

Temperature

Soil temperature may affect microbial activity both directly and indirectly, through its impact on other factors such as soil moisture and litter quantity. Higher temperatures are associated with higher rates of microbial activity. Moreover, changes in soil temperature also affect the microbial community composition. The effect of temperature is also modulated by other factors. For example, in water limited environments higher temperatures do not necessarily lead to higher decomposition rate if the soil is dry. In cold climates temperature fluctuations determine freeze-thaw cycles, which kill soil microbes, release pulses of available nutrients, and lead to higher rates of decomposition and N mineralization in the subsequent growing season.

Soil moisture

Conditions of limited soil water availability reduce the rate of microbial activity due to the emergence of conditions of microbial water stress caused by dehydration, and to the reduction in the size of the water films coating the soil grains. Low moisture contents also limit the mobility and the supply of substrate to the soil microbes by diffusion through the soil solution. In wet soils microbial activity is limited by the low soil aeration and the consequent limitations in the amounts of oxygen available to the decomposition process. Thus, under water-logging conditions typical of wetland soils decomposition is for most part inhibited due to the limited supply of oxygen to the soil microbes and to the limited transport of the CO_2 produced in the decomposition process (soil respiration). The optimal environment for microbial activity is provided by a warm soil that is both moist and aerated soil. These conditions are met with intermediate moisture contents between those of dry and completely saturated soils. The effect of soil moisture is also modulated by other factors: in energy-limited environment, e.g. boreal forest, higher moistures do not necessarily lead to higher decomposition rates. Soil moisture fluctuations, typical of arid and semi-arid environments, are associated with pulses in the rates of decomposition and mineralization, with consequent pulses in the availability of soil mineral nutrients. Other abiotic factors such as solar irradiance, soil pH, and soil physical properties (e.g. clay content) also affect the decomposition rates. For example, it has been found that in arid environments litter decomposition can undergo a process of photodegradation, which provides a shortcut in the cycling of soil carbon.

Dynamics of Selected Ecosystems

Arid and Semi-Arid Environments

Arid and semi-arid lands occupy an increasingly large portion of world's land surface. Decomposition processes in arid and semi-arid environment have some unique characteristics due to the distinct physical environment, e.g. low rainfall and relatively high solar radiation. In these ecosystems 1) leaching is not an important decomposition process as opposed to temperate and tropical environments; 2) fragmentation and chemical alteration may be separated both temporally and spatially; 3) the ultimate location of plant litter is more important than its physical-chemical structure; 4) a variable quantity of the litter input in a desert ecosystem can be buried under the soil surface. The buried litter decomposes more rapidly than surface litter though it goes through the same pathway of chemical transformation as the surface litter; 5) photodegradation may provide an alternative pathway of aboveground litter decomposition. Photodegradation provides a shortcut in the carbon cycle, with a substantial fraction of vegetation carbon being lost directly to the atmosphere without cycling through the soil organic matter pools. Partially due to all the above reasons, compared with wetter environments, the rate of litter decomposition in arid and semi-arid environments is much poorly predicted using abiotic factors such as mean annual precipitation and soil moisture.

Aquatic Ecosystems

Aquatic ecosystems comprise the largest portion of the biosphere and include both freshwater and marine ecosystems. The sources of organic matter in these systems can be both internal (autochthonous) and external (allochthonous). In general, the autochthonous material has higher available N concentration and is structurally easier to decompose than the allochthonous plant residues. Decomposition in aquatic ecosystems follows similar patterns as in terrestrial environments (i.e., it involves leaching, fragmentation and chemical alteration), though with some major differences due to the aquatic environment. A major form of organic matter in aquatic ecosystems is the particulate organic matter (POM). POM can come both from autochthonous and allochthonous sources. The allochthonous sources include terrestrial leaves and small twigs, which are usually colonized by fungi and fragmented by shredders, leading to the formation of POM. Autochthonous POM is derived from the fragmentation of dead organisms and other organic material. POM is partly ingested, digested and mineralized by organisms and micro-organisms before settling on the bottom. The remaining organic matter that reaches the bottom is further broken down by bacteria both through aerobic and anaerobic processes. Another important component of organic matter in aquatic ecosystems is the dissolved organic matter (DOM). Major sources of DOM in the water column are exudates excreted by macroalgae, phytoplankton and zooplankton and autolysis the remains of phytoplankton and zooplankton. DOM is taken up by bacteria and converted into bacterial biomass without undergoing any break-down into inorganic compounds. This bacterial biomass is later consumed by the zooplankton, which, in turn, excretes nutrients in the form of exudates, contributing to a significant portion of the suspended material in the water column. Bacteria, then, take these exudates (even at very low concentration) to obtain both carbon and nutrients, and a new cycle starts. Thus, in contrast to terrestrial ecosystems, bacteria in aquatic systems act as converters rather than as decomposers, whereas phytoplankton and zooplankton play major roles in the release of available nutrients.

Boreal Forests

Boreal forests are among the best studied ecosystems in terms of litter decomposition and mineralization. In these forests – and in forest ecosystems in general – litterfall is the largest source of soil organic material, in that it can account for more than 50% of Net Primary Productivity (NPP). Due to the low energy environment (low temperature and solar radiation), litter decomposition in boreal forests is slow. In these environments the initial leaching from leaf litter is generally slow, while microbial degradation is the major decomposition process. The chemical changes of litter biomass in boreal forests have been well documented. The concentrations of N, P, S, Fe, Pb, Cu and Zn in litter increase with time during decomposition. However, these relationships are empirical and have not been fully explained. The concentration of K normally decreases with time until it reaches a minimum value, and, then, it slowly increases, probably due to the fact that K is the most mobile element among all plant nutrients and its leaching may start as soon as the trees shed their leaves. Mg is another mobile nutrient and its leaching pattern is similar to that of K, though at a slower pace. Ca concentration usually increases in the early stage of decomposition until it reaches a maximum value, and then it decreases. The concentration of Mn, in contrast to most of the other elements, decreases almost linearly throughout the decomposition process.

Anthropogenic Impacts

Human activities significantly modify several aspects of ecosystem function in many environments around the world. Decomposition is unavoidably altered in most ecosystems.

Effects of N Deposition

Dry and wet deposition and fertilization add significant amounts of N both to aquatic and terrestrial ecosystems. In most systems, the increase in soil N content leads to higher levels of foliar N and, hence, to higher litter N contents. The higher substrate quality (high litter and soil N levels) generally result in higher rates of decomposition. In some ecosystems (e.g. Scots pine and Norway spruce forests), however, N fertilization has also been found to increase the amount of lignin in the litter, with consequent reduction in the decomposition rates, suggesting that the effect of N fertilization on decomposition could be complicated and counterintuitive.

Effects of Heavy Metal Pollution

Some heavy metals tend to accumulate in the soil, due to their high affinity both with soil organic matter and with mineral particles. This accumulation eventually exceeds the toxicity threshold tolerable by soil microorganisms and soil fauna – the major drivers of the decomposition process. The direct effect of heavy metal accumulation on plant uptake depends on soil properties: soils with neutral pH and high clay content can immobilize large amounts of heavy metals, while the chemical composition of leaves is not significantly affected by the heavy metals in the soil. However, plants growing on these soils are only temporarily protected against heavy metals until the soil retention potential for these metals is reached. In acidic soils, the low pH increases the solubility of most heavy metals leading to the uptake of heavy metals by vegetation, and to their accumulation in the plant tissues

(e.g., leaves). Because all heavy metals are potentially toxic, there is some concern about the possible increase in heavy metal content in the plant leaves, and its consequent deleterious effects on ecosystem function.

Common Study Methods

Decomposition Methods

The use of litter bags is one of most common field techniques for litter decomposition studies to the point that it has become a sort of standard method in studies on decomposition. This method is used both for the quantitative assessment of litter biomass loss and for studies on its chemical changes. With this method a bag is filled generally with 1–10 g of litter dried at room temperature until a constant moisture level is reached. High temperatures are avoided in the preparation of these litter samples to prevent important changes in microbial community and fiber structure. The bag is first exposed to field conditions for a specific time period, and then it is brought back to the laboratory for reweighing and performing chemical analyses on the remaining litter using techniques such as atomic emission spectrometry (AES), atomic absorption spectrometry (AAS) and inductively coupled plasma spectrometry (ICP). A typical litter bag size is between 10×10 cm and 20×20 cm and it is made of biologically resistant polyester or nylon. Nylon is not used in N studies because this material contains N. Mesh size and incubation time depends on aim of the study and the precision required. Sometimes a certain mesh size is purposely employed to exclude particular groups of soil fauna in order to determine their functional significance to decomposition processes. The number of replicate bags is important for the accuracy in mass loss estimation and the minimum of bags must be sufficient to estimate the decomposition rate constant k adequately. The study methods for woody detritus decomposition are analogous to the litter bags method, though they usually do not use litter confinement. Other methods commonly used to investigate decomposition include microcosm studies, or laboratory and field techniques based on the measurement of concentration and fluxes of soil CO_2 through time. These measurements allow separating the different contributions to soil organic matter loss (e.g., carbon mineralization vs. leaching).

Mineralization Methods

The buried bag technique is a common approach to measure N mineralization. It consists of 1) collecting soil core samples and measuring the initial concentration of inorganic N; 2) reburing subsamples of this core in polyethylene bags for specific period of time, and 3) measuring the inorganic N after incubation. The net rate of N mineralization is calculated by the difference in N concentrations between the two measurements. The buried bag method is relatively simple, cost effective and provides results that can be compared with other studies, due to its widespread use in many different ecosystems around the world. The major problem of this method arises from the disturbance of the soil sample. For example, the method eliminates plant uptake, with a consequent increase in soil inorganic N. In N-limited systems this increase leads to higher microbial N immobilization, which, in turn, results in the underestimation of plant uptake. The mineralization rates of other nutrients essential to plants, such as P and S, can also be measured using buried bags with appropriate adjustments. Several other methods can be used to determine the rates of N mineralization, including the N budget analysis (ideally for large temporal scales), the "super sinks" analysis (e.g. ion exchange resin), the use of substrate analogs, as well as of isotope tracer and dilution measurements. Over the last decade, the general notion of N mineralization has evolved: the concept of N mineralization as the driving process in the N cycle has been replaced by the idea that exoenzyme-driven depolymerization is the rate-limiting step in the generation of bioavailable N. This new N cycling paradigm does not invalidate the traditional methods of net mineralization measurement as fundamental tools in the study of N-cycling, but caution needs to be used in the interpretation of the results. Net mineralization is an indirect indicator of N availability and not the key step of N cycling.

Further Reading

Aerts R (1997) Climate, leaf litter chemistry and leaf litter decomposition in terrestrial ecosystems: A triangular relationship. *Oikos* 79: 439–449.
Austin AT and Vivanco L (2006) Plant litter decomposition in a semi-arid ecosystem controlled by photodegradation. *Nature* 442: 555–558.
Berg B and Laskowski R (2006) *Advances in ecological research for volume 38: Litter decomposition: A guide to carbon and nutrient turnover*. London: Elsevier.
Brady N and Weil RR (2004) *Elements of the nature and properties of soils*, 2nd ed. Upper Saddle River, New Jersey: Pearson/Prentice–Hall.
Chapin FS, Matson PA, and Mooney HA (2002) *Principles of terrestrial ecosystem ecology*. New York: Springer-Verlag.
Michener RH and Lajtha K (2007) *Stable isotopes in ecology and environmental science*. Oxford: Blackwell Scientific Publications.
Parton W, Silver WL, Burke I, et al. (2007) Global-scale similarities in nitrogen release patterns during long-term decomposition. *Science* 315: 361–363.
Sala OE, Jackson RB, Mooney HA, and Howarth RW (eds.) (2000) *Methods in ecosystem science*. New York: Springer-Verlag.
Schimel JP and Bennett J (2004) Nitrogen mineralization: Challenges of a changing paradigm. *Ecology* 85: 591–602.
Schlesinger W (1997) *Biogeochemistry: An analysis of global change*, 2nd ed. New York: Academic Press.
Smith RL and Smith TM (2001) *Ecology and field biology*, 6th ed. Benjamin Cummings.
Vitousek P (2004) *Nutrient cycling and limitation-Hawaii as a model system*. Princeton, New Jersey: Princeton University Press.

Erosion

EJ Comoss and DA Kelly, Bureau of Facility Design and Construction, Harrisburg, PA, USA
HZ Leslie, Bureau of State Parks, Erie, PA, USA

Outline of Erosion

Erosion could be defined as displacement of solids (e.g., soil, mud, rock) by the agents of currents such as wind, water, or ice by downward movement in response to gravity or by living organisms (i.e., bioerosion) (http://en.wikipedia.org/wiki/Erosion). The degree of erosion is accelerated by various sources. Rainfall (i.e., the amount and intensity of precipitation) is the primary cause of erosion and plays a key role for the other agents to intensify the erosion impact on ecosystems. Rate of erosion depends on various environmental factors such as soil texture, gradient of slope, ground cover (e.g., vegetation, land use), and current velocity of streams (http://en.wikipedia.org/wiki/Erosion).

Due to rapid agricultural/industrial development and mismanagement of natural ecosystems (e.g., overuse of trails in parks), erosion (especially soil erosion) has been a global issue in achieving sustainable ecosystem management. Soil erosion is accelerated by water (e.g., rain detaching, transporting soil; **Fig. 1**), wind (**Fig. 2**), or tillage, and affects greatly agricultural areas and the natural environment. Soil erosion is one of the most serious environmental problems and has both on-site and off-site impacts (Favis-Mortlock 2005 in http://www.soilerosion.net/).

The main `on-site' impact occurs at the place where the soil is detached and is presented as the reduction in soil quality resulting from the loss of the nutrient-rich upper layers, and the reduced water-holding capacity. Loss of soil quality is presently a global problem. Soil erosion's most serious impact may well be its threat to the long-term sustainability of agricultural productivity. The `off-site' problem is caused from soil detachment by accelerated water or wind erosion and occurs wherever the eroded soil ends up. The soils could be transported considerable distances (see **Fig. 3**) and may give rise to the `off-site problems'. The transportation of soil will result in accumulation of sediments and agricultural pollutants in watercourses, leading to the silting up of dams, disruption of the ecosystems of lakes, and contamination of drinking water and downstream watercourses (Favis-Mortlock 2005 in http://www.soilerosion.net/).

An Innovative Control of Shoreline Erosion

In addition to the traditional problems of agriculture-related erosion, erosion control for natural ecosystems (e.g., national parks) has been recently a critical issue. In particular, the border area between land and water (e.g., shoreline, riparian zone) is extremely sensitive to erosion. Proper management of erosion in this type of ecotone is essential for achieving sustainability of ecosystems (e.g., biodiversity, efficiency in energy production) and for obtaining maximum socioeconomic benefit from environmental amenity.

In this article, an innovative control of shoreline erosion is introduced to demonstrate how efficient ecosystem management could be optimized between nature conservation and environmental engineering. Protection of recreational beaches along ocean coasts and inland lakes, bays, and inlets, as well as finding a beneficial use for dredge material, has become a sensitive issue to a diverse public. It is often difficult to find a solution that makes good engineering sense while maintaining environmental responsibility.

Current conventional methods used to retard shoreline erosion include the installation of breakwaters, groins, and jetties. Sand replenishment is often used in conjunction with these methods when shorelines are being extended or restored. These techniques, though often functional, are costly and can detract from the natural environment. Dredge material management was viewed as a `necessary evil' associated with overdevelopment of coastal areas in the past. Through the years, hundreds of metric tonnes of sediment have been dredged annually for commercial and recreational purposes and subsequently discharged into land-based disposal facilities or into oceans, estuaries, rivers, and lakes. As the space for disposal facilities reaches capacity, and discharge into water bodies becomes more of an ecological concern, the problem arises as to what to do with this material.

The purpose of this article is to describe in detail how Presque Isle State Park, located along the shoreline of Lake Erie in Pennsylvania, implemented a unique erosion protection project, which also included the beneficial use of dredge material. This low-cost, innovative demonstration project minimized erosion in the lesser-energy zone of Misery Bay in Presque Isle State Park by utilizing native plants, bioengineering, dredge material placement, and nonconventional erosion practices.

Outline of the Park and Erosion

Presque Isle State Park is a 1295 ha migrating sand spit that juts 11.3 km into Lake Erie and is a major recreational landmark for approximately 4 million visitors each year. The park, a National Natural Designated Landmark, is particularly environmentally

 doi:10.1016/B978-0-444-63768-0.00054-8

Fig. 1 Severe soil erosion in a wheat field near Washington State University, USA. Photographer: Jack Dykinga, http://www.ars.usda.gov/is/graphics/photos/k5951-1.htm.

Fig. 2 Wind erosion on Ulen fine sandy loam, Grand Forks, North Dakota, USA. Photographer: Adrian Fox, http://www.nrcs.usda.gov/TECHNICAL/ECS/agronomy/Photo%20File/FrontAgr3.jpg.

sensitive with its constantly evolving shoreline (**Fig. 4**) and the presence of numerous plants recognized as being of exceptional value. Presque Isle is rated as one of the top birding areas in the Northeast, as birds use the distal end of the spit for a resting and feeding area.

Fig. 3 Yangtze River showing the sediment-rich water, the Three Gorges, Hubei Province, China. From http://www.soilerosion.net/.

Protection of the spit has been an ongoing process since 1828. Along the Lake Erie shoreline, a series of conventional erosion control techniques such as groins, bulkheads, seawalls, and beach nourishment have been used with varying degrees of success. Between 1989 and 1992, many of the previous structures were removed and 55 offshore rubble mound breakwaters were constructed. Since completion of the breakwaters, shoreline maintenance has been limited to an annual beach nourishment program. Construction of the breakwaters has decreased sand purchased for annual nourishment by *c.* 85%, from approximately 231 000 m^3 before breakwater installation to *c.* 30 764 m^3 after breakwater construction.

Since 1975, the beaches along the lakeside of the park have been nourished annually; nourishment amounts varied based on fluctuating lake levels and storm severity. The prevailing winds along the lake are from the west, and, as a result, the beach sand is in continual motion as it moves in response to longshore transport. While most of this transient sand is redeposited in offshore bars in the lake, some of the sand is carried around the distal end of the spit into the back bay area. Accumulation of this fine-grained material in the back bay has been a continual problem as these shallow areas become choked with sediment. As a result, the park struggles with the problem of dredging these areas and finding a suitable disposal option for the dredged material.

Historically, protection of the shoreline from erosion had been accomplished along the Presque Isle Bay side of the park by utilizing large stones to riprap the shoreline. Although this process was very effective in preventing shoreline erosion and was quite suitable in the more developed recreational areas of the park, because of its `non-natural' appearance, riprap did not concur with the desired results and appearance specified by the management plan for the low density and natural areas.

As a result of the fragile ecosystem of the spit, specific erosion problems along the bay, and the development of a sand bar within the park's back bay area, the decision was made to seek funds to advance an innovative solution to these problems. With this goal in mind, the Department of Conservation and Natural Resources, Bureau of State Parks – Presque Isle State Park, in conjunction with the Presque Isle Partnership, secured funding via a matching grant from the Great Lakes Commission. The project coordinated efforts between state and federal government units, as well as private, nonprofit volunteer organizations to design, implement, and provide construction services for the project. This project brought forward a concept that provided the park with the protection needed for the infrastructure as well as creating a shoreline appearance that resembled natural shorelines along environmentally sensitive areas of the park. Additionally, the project provided a beneficial use of dredge material from the back bay sand bar.

In order to realize the goals of the project, the decision was made that rather than solely utilizing conventional riprap, the project would incorporate a combination of riprap as well as indigenous vegetation, bioengineering, dredge material, and innovative landscape architecture to retard shoreline erosion along a heavily used multipurpose trail. Completion of this project has provided valuable information to other parks and recreational facilities in the Great Lakes area (especially along bay inlet areas), which are also faced with the challenges of minimizing erosion and sedimentation as well as finding a beneficial use for dredge material.

Problem Areas

There are numerous recreational features within the park. One of these is a 15.4 km multipurpose trail. This trail, designated as a National Recreation Trail, begins at the park entrance and completes a 21.7 km loop throughout the park. This is the most popular trail within the park, and is heavily used by bicyclists, joggers, roller bladers, and is wheelchair accessible. Because of its popularity, protection of the trail from erosion is paramount.

A portion of this trail lies along the southern shoreline of the peninsula within Presque Isle Bay, Misery Bay, Marina Lake, and Thompson Bay; this area had been exhibiting significant erosion. Because this area was adjacent to Presque Isle's ecological reservation area, the standard riprap remedy was not appropriate because it did not match the park's designated management prescriptions for maintaining a natural shoreline appearance.

Fig. 4 Presque Isle State Park. (a) Aerial view of Gull Point and the park from the east. (b) The shore on the bay side, looking north. (a) Photographer: Robert K. Grubbs, http://en.wikipedia.org/wiki/Image:PresqueIsleStatePark.JPG; (b) http://en.wikipedia.org/wiki/Image:PresqueIsleBay-lookingN.JPG.

Another popular tourist attraction is the Perry Monument, dedicated to Commodore Perry. The area surrounding the monument, located along Misery Bay, receives widespread use for shoreline fishing as well as the launching of recreational boats. Normal wave energy along the shoreline of Presque Isle Bay, fluctuating water levels and currents, caused a significant sand bar to develop off the northeast tip of Perry Monument. The sand bar measured approximately 91 m long by 8 m wide by 1.5 m deep (1100 m^3) and severely restricted recreational boat usage. Removal of the sand bar was essential to preserve the recreational activities at the monument.

A small portion of this sand undoubtedly was the beach sand along the Lake Erie side of the spit. The probable source for the remainder of this sand was the erosion of the shoreline around Misery Bay. Historical photos of this area show that the east shoreline of Misery Bay has eroded several hundred feet since the late 1800s. Suitable disposal of this sand would be difficult because of its susceptibility to erosion no matter where it would be placed. Rather than following the standard disposal options of this dredged material, the park wanted to find a constructive use for this sand.

Project Description

The first phase of the project was to remove some of the sand from Perry Monument. After passing the mandated state tests for disposal of dredge material, and in accordance with all applicable permits and regulations, approximately 917 m^3 of material from the sand bar was dredged via a clam shell and placed in an unused gravel parking lot within the park so it could naturally dewater; the water could naturally seep through the gravel into the ground. The remaining sand bar was then graded to provide a suitable launch/mooring area for canoes and shallow boats.

The next phase involved the creation of a stabilized area on the backside of Misery Bay where the multipurpose trail is located adjacent to an ecologically sensitive area of the park (an area where the natural habitat is to be maintained and little to no development is to occur). In this area, significant erosion had occurred, to the point that water was only 3–4.5 m from the trail.

Initially, the project proposed to install 10–25-cm-sized riprap *c.* 7–9 m from the existing shoreline. However, based on design criteria for the worst-case scenario for wave height, which in this area would be 1 m, the decision was made to place 30–61-cm-sized riprap offshore; it was felt that this sized riprap would provide better protection for the shoreline. Placement of this riprap below the water line created an artificial stone shoreline that protected the multipurpose trail by functioning as an erosion/wave energy dissipater.

Once the rock was in place, the third phase began. In this phase, the park placed the dewatered dredge material (sand) over the newly placed riprap, creating a higher-elevation dune line. This subsequently provided a buffer of *c.* 8–9 m between the water and the trail. Next, to enhance the `natural' appearance of the shoreline, randomly spaced downed trees and stumps from the park (25–91 cm in diameter), minus the limbs, were used as timber groins. To function as groins, the tree root bases were anchored behind the riprap in the fill, and the trunks extended out past the riprap and into the water, also serving as sediment catch basins.

After the fill had been placed and prior to planting, the decision was made to use geotextile (comprised of coconut fiber cured to aid in longevity) in conjunction with wattles (poles interwoven with slender branches) in order to augment the vegetative plant rooting and further stabilize the dredged sand. The woven geotextile material is biodegradable, but it was anticipated that the plants should be well established before it decomposed.

Within the fill material, several trenches, parallel to the shoreline, were dug. First, the geotextile was laid in the trench, and then the wattles were placed on top – end to end and parallel to the shoreline, approximately at the average high-water mark. The geotextile was then rolled back over the wattles and staked with live saplings. The wattles and geotextile were then further secured by placing sand on top; the sand helped to anchor the entire apparatus against the wind. The geotextile and wattles provided extra erosion protection to the shore zone area, as well as ensuring a stabilized area in the fine-grained sand for plant rooting.

Prior to planting of the indigenous vegetation, plant community goals were established to ensure that the plants would thrive in the newly created environment. The plant community goals were developed by reviewing historical records of plant community structure in that area, consulting local and regional plant experts, and considering wildlife uses of the site. After the goals were established, a vegetation planting plan was prepared. Final preparation of the site prior to planting included the addition of topsoil to the upper layer of sand, and shaping of the dune line.

The final phase was the vegetative planting. For this phase, local sources of plant material within the park were identified for transplanting onto the dune. These included beach grass (*Ammophilia breviligulata*), Indian sea oats (*Chasmanthium latifolium*), switch grass (*Panicum virgatum*), choke cherry (*Prunus virginiana*), bayberry (*Myrica pensylvanica*), and black oak (*Quercus velutina*). Additionally, driftwood from local sources was collected and was dispersed in the restored area to provide shelter for the young seedlings. Local sources of emergent wetland plants were located, and these were then transplanted into shallow water below the wattle trenches. Transplanted aquatic plants included species that enhanced the establishment of desirable native emergent communities. These species, such as branching bur reed (*Sparganiuma androcladum*), duckweed (*Spirodela oligorrhiza*), and soft-stem bulrush (*Scirpus validus*) were also beneficial to waterfowl by providing a native food source.

After the native species were established, invasive species, such as purple loosestrife (*Lythrum salicaria*) and common reed (*Phragmites australis*), targeted in the Presque Isle Partnership report, were mechanically removed as they were encountered throughout the restored areas. Roundup, a glyphosate, was applied as necessary to eliminate invasive species that could not be controlled by mechanical means. The final objective was to achieve at least 50% vegetative cover in both the shoreline and dune habitats – this goal had been achieved within 6 months of project completion. Annual plane surveys have shown that since project completion, erosion has been reduced by approximately 90%, thus providing protection for the heavily used multipurpose trail.

Summary and Conclusion

Due to rapid agricultural/industrial development and mismanagement of natural ecosystems, erosion has been a global issue in achieving sustainable ecosystem management. Degree of erosion is accelerated by various sources covering rainfall, soil texture, gradient of slope, ground cover, and current velocity of streams. Soil erosion accelerated by water, wind, or tillage affects greatly agricultural areas and the natural environment.

In addition to the traditional problems of agriculture-related erosion, erosion control for natural ecosystems has been recently a critical issue. Especially erosion control in the border area between land and lake such as shoreline is critical for achieving sustainability of ecosystems and for obtaining maximum socioeconomic benefit from environmental amenity.

An innovative control for shoreline erosion was successfully carried out in a state park in the lake area. The dual goal of the project was to combine the beneficial use of dredge material, indigenous plants, and landscaping to reduce sediment loading into Lake Erie, and to protect the recreational aspects of Presque Isle State Park. The completed project has resulted in several additional hectares of stabilized vegetation and has decreased soil and subsequent nutrient runoff from entering Lake Erie. The amount of material removed from the Perry Monument sand bar has facilitated recreational boat usage and shoreline fishing in this area.

Through the years, conventional erosion protection techniques at Presque Isle State Park have been both costly and inappropriate for natural area management. Conversely, this economical project (a total cost of $33 000) has provided a natural and esthetic alternative to conventional shoreline erosion protection. While remaining within standard bureaucratic financial constraints, the project affords a valuable example to other parks and recreational facilities in the similar situations with the challenge of minimizing erosion while maintaining a natural appearance, and finding a beneficial use for dredge material.

Acknowledgments

The authors would like to acknowledge the Great Lakes Commission for awarding a matching grant to accomplish this project. Also, they would like to thank the Buffalo District of the Corps of Engineers for aerial photos, Ernst Seeds for developing the vegetative planting plan, Mercyhurst College in Erie for the actual vegetative plantings, and the Presque Isle Partnership for their volunteer efforts for a variety of tasks related to the project. Finally, thanks are due to the Bureau of Facility Design and Construction engineers and the staff at Presque Isle State Park for their tireless efforts to ensure the success of this project.

See also: Ecosystems: Dunes. General Ecology: Soil Ecology

Further Reading

Boardman, J., Poesen, J., 2006. Soil Erosion in Europe. Chichester, UK: Wiley.
Commonwealth of Pennsylvania, Department of Conservation and Natural Resources, Pennsylvania Bureau of State Parks, Department of Conservation and Natural Resources. Pennsylvania Bureau of State Parks, 1998. Presque Isle State Park Management Plan, vol. VI.
Favis-Mortlock, D.T., Boardman, J., MacMillan, V.J., 2001. The limits of erosion modeling: Why we should proceed with care. In: Harmon, R.S., Doe III, W.W. (Eds.), Landscape Erosion and Evolution Modeling. New York: Kluwer Academic/Plenum Publishing, pp. 477–516.
Montgomery, D.R., 2007. Soil erosion and agricultural stability. Proceedings of the National Academy of Sciences of the United States of America 104, 13268–13272.
Poesen, J., Nachtergaele, J., Verstraeten, G., Valentin, C., 2003. Gully erosion and environmental change: Importance and research needs. CATENA 50, 91–133.
Schmittner, K.-E., Giresse, P., 1999. The impact of atmospheric sodium on erodibility of clay in a coastal Mediterranean region. Environmental Geology 37, 195–206.

Relevant Websites

http://en.wikipedia.org
Entry on Erosion, Wikipedia.
http://www.soilerosion.net
Soil Erosion Site.

Evapotranspiration☆

Jan Pokorny, ENKI, O.P.S., Třeboň, Czech Republic

Introduction

Water is one of the most important limited natural resources. Declining water resources and water quality problems have resulted in dramatic increase in the need for water-conserving methodologies on a field, watershed, and regional scale and this makes efficient use of freshwater resources an obligation of each user. During the 30-year period from 1950 to 1980, the actual level of per capita water supply decreased significantly in many countries due to population increases. It has been projected that in early year 2000 considerably low water availability per capita is anticipated in many regions of the world. As water becomes increasingly scarce and the need becomes more pressing, newer and more complete methods of measuring and evaluating techniques of handling water resources are necessary. In terms of agricultural production, approximately 17% of the cropped area of the world is irrigated and contributes more than one-third of the total world food production. In the United States, about 12% of the cropped area is irrigated and contributes about 25% of the total value of the United States crops. In the United States and around the world, irrigated agriculture uses most of the water withdrawals from the surface and groundwater supplies. Thus, accurate quantification of plant water use (evapotranspiration) is crucial for better management and allocation of water resources.

The process known as evapotranspiration (ET) is of great importance in many disciplines. Accurate quantification of ET in agroecosystems is critical for better planning, managing, and efficient use of water resources, especially in arid or semiarid environments where lack of precipitation usually limits plant growth and yield and negatively affects ecological balances. Quantification of ET is also crucial in water allocation, irrigation management, evaluating the effects of changing land use on water yield, environmental assessment, and development of best management practices to protect surface and groundwater quality.

ET can be defined as the loss of water from the ground, lake or pond, and vegetative surfaces to the atmosphere through vaporization of liquid water. In agroecosystems, ET is the sum of two terms: (1) transpiration, which is water entering plant roots and used to build plant tissue or being passed through leaves of the plant into the atmosphere in the vapor form, and (2) evaporation which is water evaporating from soil and water surfaces, or from the surfaces of plant leaves. Evaporation from buildings, streets, parking lots, etc., after a rain event also contributes to the total ET in the hydrologic cycle.

Evaporation and transpiration processes occur simultaneously and there is no easy method to separate these two processes. Evaporation in the field can take place from crop canopies, from the soil surface, or from a free water surface. When the soil surface is bare, evaporation will take place from the soil directly. In the absence of vegetation, and when the soil surface is subject to radiation and wind effects, evaporation can result in considerable loss of water in both irrigated and nonirrigated agriculture, and other ecological landscapes. In the semiarid and arid western regions of the United Sates, evaporation can be as high as 40% of the total ET.

Transpiration increases with increasing leaf area until complete closure of the canopy occurs. For agricultural crops such maximum transpiration is usually attained at a leaf area index (LAI) of about 3–3.5. In the transpiration process, stomata opening and closure depends on water uptake rate which in turn depends on the density and distribution of roots and their effectiveness to uptake water and nutrients from the soil. Stomata would close when roots cannot uptake water from soil with sufficient rate to keep up with the transpiration. In irrigated agroecosystems, the goal should be decreasing the evaporation component of the total ET for optimum crop production because yield and transpiration are strongly related and evaporation does not have any contribution to the crop growth and yield. Thus, the evaporation falls into the "unbeneficial water use" category.

The ET rate and amount for different vegetation surfaces (i.e., agronomical crops, which are mostly "annual crops" vs. trees and shrubs, which are mostly "perennials") show significant variation from one location to another and are strong functions of climatic, soil conditions, and management practices. For example, the seasonal crop ET for corn (maize, *Zea mays*) can range from 500 to 800 mm depending on climate. The ET for typical alfalfa (*Medicago sativa*) plant can range from 800 to 1600 mm per growing period depending on climate and length of growing period. For a tropical plant such as banana (*Musa* spp.), this value is between 1200 mm in the humid tropics and 2200 mm in the dry tropics. Water requirement of trees can also show wide variations depending on climate, soil type, and root structures. For example, an orange tree (*Citrus aurantium*) can use as much as between 900 and 1200 mm of water per year whereas olive tree (*Olea europea*) can use only between 400 and 600 mm of water per year. The expected water use of natural vegetations can also show significant variation. For example, the seasonal water use of cattail (*Typha*)

☆*Change History*: March 2018. J Pokorný added Sections "Solar Energy Flux Between Sun and Earth," "Main Fluxes of Solar Energy in Landscape," "Cooling and Air-Conditioning Effect of Evapotranspiration," "Effect ET on Local Climate," "Landscape Drying," "Evapotranspiration of Forests (biotic pump)" together with 6 new figures (Figs. 2–4 and 6–8).

This is an update of G. Katul and K. Novick, Evapotranspiration, In Encyclopedia of Inland Waters, edited by Gene E. Likens, Academic Press, Oxford, 2009, pp. 661–667.

 doi:10.1016/B978-0-12-409548-9.11182-0

can range from only 890 mm to as much as 2500 mm. Water use for foxtail (*Lycopodium clavatum*) is about 140 mm and for pine tree (*Pinus*) water use can range from 480 to 1190 mm. Water use of different natural vegetation and agronomical plants are important and necessary for accurate determination of hydrologic balance components.

The Hydrologic Cycle and ET

ET is a major component of the hydrologic cycle. A major proportion of the total precipitation falling on the land surface is returned to the atmosphere by ET. As a global average, 57% of the annual precipitation falling over the land is returned to the atmosphere by ET. ET amounts to about 70% of the annual precipitation of the United States, and more than 90% of the precipitation in the arid and semiarid areas of the western United States. Different components of a typical hydrologic cycle are illustrated in Fig. 1. The hydrologic cycle can be defined as the pathways of water as it moves in its various phases through the atmosphere, to the Earth, over and through the land, to the ocean, and back to the atmosphere. During this cycle, which has no beginning or end, water molecules may assume various states, returning to a hydrologic pathway as new chemical compounds that are mixed with various solid and liquid substances. In the cycle, water evaporates from the oceans, ponds, rivers, and various land surface to become part of the atmosphere; water vapor is transported and lifted in the atmosphere until it condenses and precipitates on the land or oceans. Precipitated water may be intercepted by vegetation, become overland flow over the ground surface, infiltrate into the ground, flow through the soil as subsurface flow, or discharge into streams as surface runoff. In a given watershed, discharge of water is primarily from groundwater withdrawals for irrigation, ET where the water table is near land surface, overland flow (runoff), and seepage to streams and springs where the water table intersects the land surface. Recharge of water is primarily from precipitation; other sources of recharge are irrigation return flow and seepage from streams, canals, and reservoirs. Large amounts of the intercepted water and surface runoff return to the atmosphere through evaporation. Infiltrated water may percolate to deeper soil layers to recharge groundwater, and later emerge in springs, or as seepage into streams, to form surface flow. Finally, this water may flow to the larger rivers and, eventually to the sea and/or evaporate into the atmosphere. Throughout this cycle, water is usually subject to evaporation of one kind. Types of vegetation, management and land use, and climatic conditions significantly affect ET, and therefore determine the amount of water lost through ET from a watershed. In agroecosystems, it is important to have a water balance to protect the sustainability and productivity of the agroecosystems. Water-level declines may result in increased costs for groundwater withdrawals because of increased pumping lift and decreased well yields. Water-level declines also can affect groundwater availability, surface water flow, and near-stream habitat (riparian) areas,

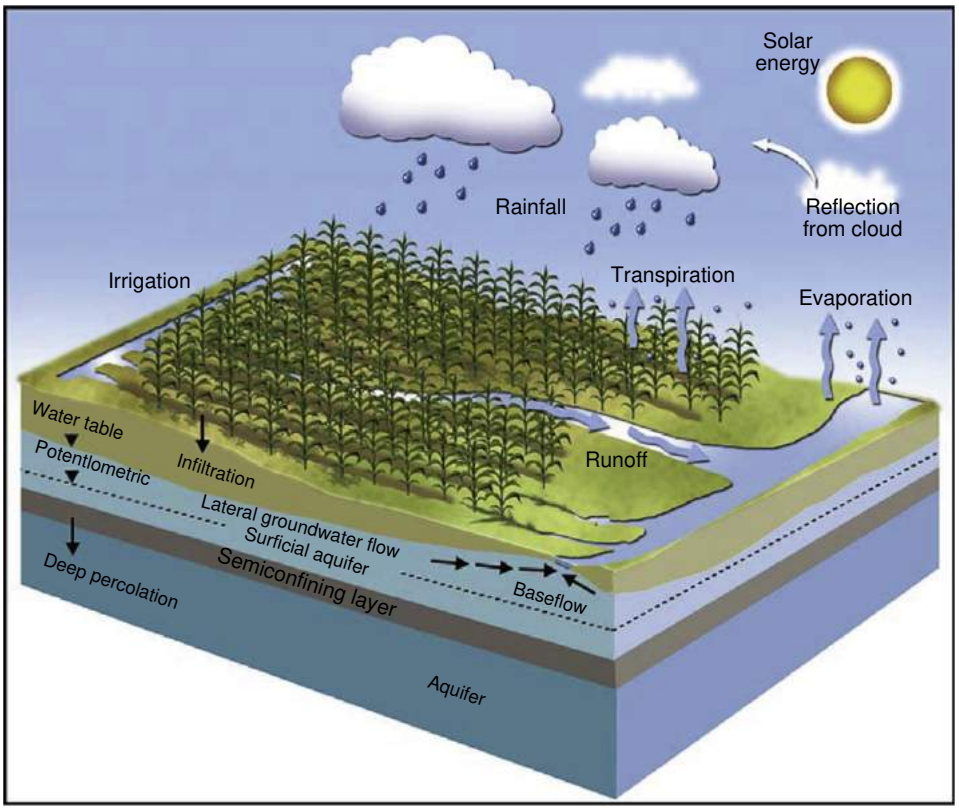


Fig. 1 The hydrologic cycle showing different components of the hydrological process.

and other ecological systems. Therefore developing efficient and effective management strategies is crucial for protecting sustainability of efficient use of water resources, protecting habitat and environment, and preventing ground and surface water degradation.

Solar Energy Flux Between Sun and Earth

For a mean distance between the Sun and the Earth, the intensity of solar radiation incident upon a surface perpendicular to the Sun's rays measured above the atmosphere is approximately 1367 W m^{-2}. This quantity is called the solar constant. The actual direct solar irradiance at the top of the Earth's atmosphere fluctuates during a year from 1412 to 1321 W m^{-2} due to the Earth's varying distance from the Sun. The maximum irradiance on Earth's surface commonly lies between 800 and 1000 W m^{-2} in the tropics and subtropics and during the growing season in temperate zones. This indicates that approximately 25%–40% of energy incident on the upper layer of the atmosphere is reflected, scattered, or absorbed in the atmosphere and does not reach the Earth's surface (Fig. 2).

The amount of incoming energy differs significantly with weather conditions (Fig. 3). The amount of incoming radiation on a clear day (e.g., 8.5 kWh m^{-2} and maximum flux 1000 W m^{-2}) can be an order of magnitude higher than the amount of incoming radiation on an overcast day (e.g., 0.78 kWh m^{-2}, maximum flux 100 W m^{-2}). Part of the energy is reflected straight away after incidence.

The ratio of reflected to incident radiation is called albedo. Dark surfaces such as water, wet soil, and wet vegetation absorb solar radiation whereas light surfaces like snow or sand are more reflective. The sum of incoming radiation minus all outgoing radiation across a unit area of the plane is called net radiation.

Main Fluxes of Solar Energy in Landscape

There is a big difference between the distributions of net radiation in functioning natural ecosystems of high plant biomass well supplied with water versus dry, nonliving physical surfaces (Fig. 4). In ecosystems, net radiation (Rn) is divided in varying proportion into following four parts: latent heat flux (LE), sensible heat flux (H), ground heat flux (G), and storage of energy (S).

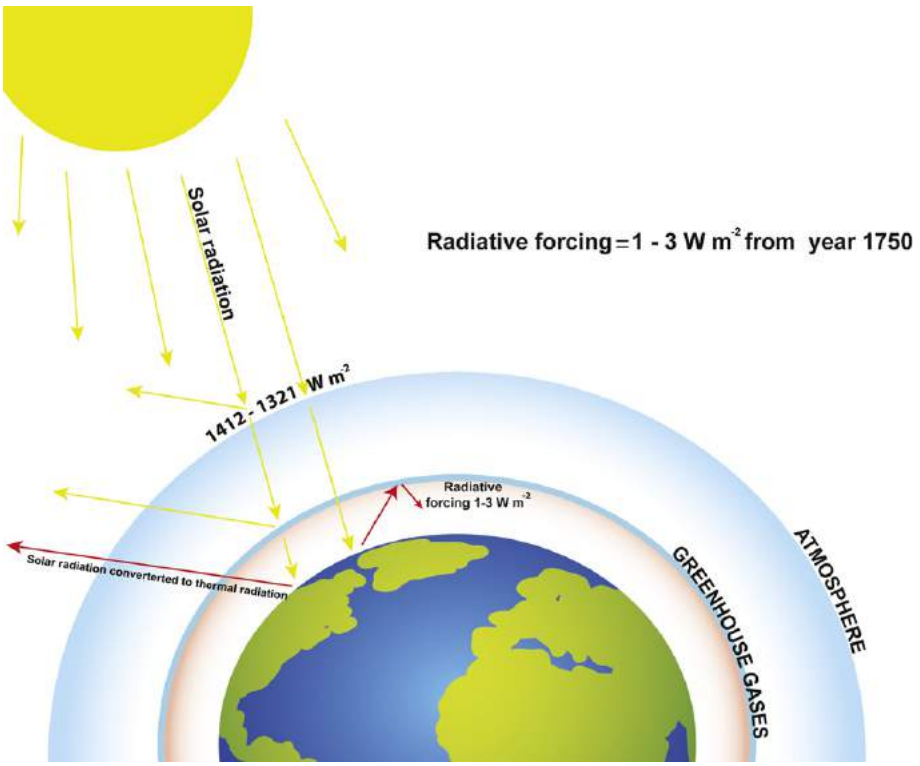


Fig. 2 Energy flux between Sun and Earth. Outer layer of atmosphere gets 1412–1321 W m^{-2} during 1 year. At clear sky up to 1000 W m^{-2} comes to Earth's surface. Radiative forcing due to an increase of concentration of greenhouse gases is 1–3 W m^{-2} from year 1750. ET is most powerful process of distribution of solar energy incoming to Earth's surface.

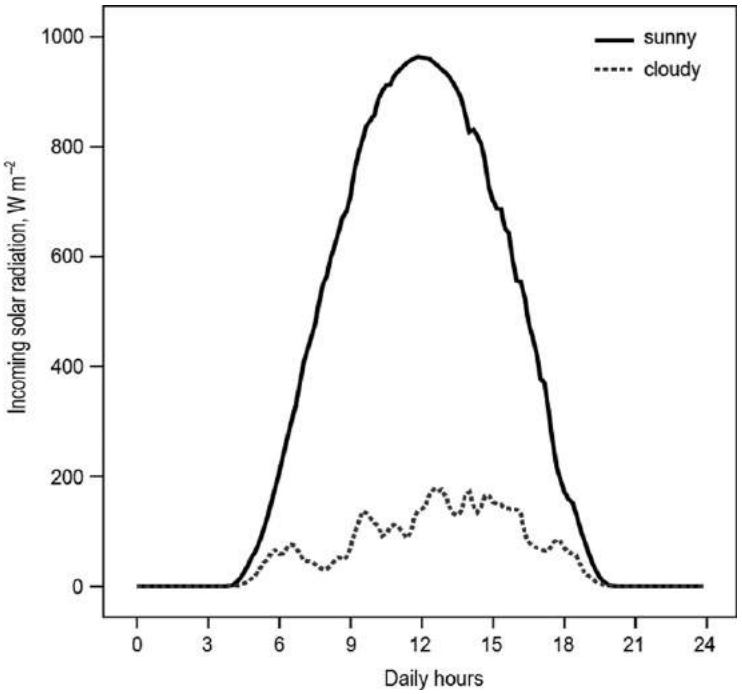


Fig. 3 Daily mean series of incoming solar radiation (W m^{-2}) on five sunny and five cloudy days.

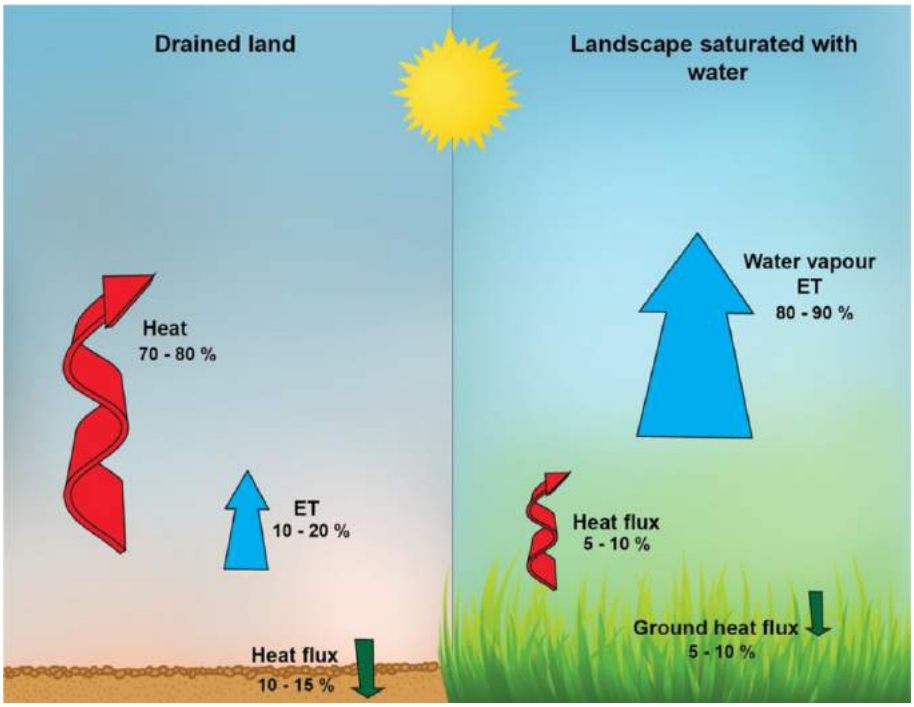


Fig. 4 Evapotranspiration, sensible heat flux, ground heat flux on drained surface, and plant stand well supplied with water.

Latent heat flux (LE) represents the energy that is released or absorbed from the surface during phase transition process. Transition of liquid into a gas phase consumes energy and thus local cooling accompanies it. Latent heat flux is generally referred to as evapotranspiration, which describes the total evaporation from land surface and transpiration by plants. Evapotranspiration from wetlands use several hundred W m^{-2} on a sunny day.

Sensible heat flux (H) represents the sum of all heat exchanges between the surface of landscape and its surroundings by conduction and convection. The proportion of sensible heat in the energy balance of an ecosystem increases when water is not present, since the capacity for evaporative cooling by latent heat is diminished. On dry surfaces, the sensible heat flux may reach values of several hundreds of W m^{-2} at a sunny day.

Ground heat flux (G) is positive when the ground is warming, normally being positive during the day and negative at night. During the plant-growing period in daylight hours, G can reach up to 100 W m^{-2}.

The energy stored in vegetation (S) is the smallest part of Rn. There are two energy sinks within a plant stand: metabolic sink (photosynthesis with consequent biomass production) and a physical sink (heating of the plant material itself). Energy stored flux is a maximum of 30 W m^{-2} on a sunny day, that is, several percent of Rn and usually is neglected in the energy balance calculations.

The transformation of solar energy in landscape is then expressed by the energy balance equation (1):

$$\mathrm{Rn} = \mathrm{LE} + H + G \tag{1}$$

The greatest importance in the transformation of solar energy on the Earth's surface is the latent heat flux of the vapor and the sensible heat flux. The latent heat of vaporization is related to the water vapor, in which, due to the change of phase process, energy is consumed. Phase changes between liquid water and water vapor are linked with the consumption of a large amount of energy. The enthalpy of liquid water is -2.5 kJ g^{-1}. Evapotranspiration (ET) or latent heat flux represents large, invisible fluxes of water and energy in the landscape; the scale of several hundred W m^{-2}. The evaporation of water is not accompanied by temperature increase, because energy is transformed in the change of phase that is acceleration of the kinetic movement of molecules, as a result of which the liquid is converted into water vapor. Evaporation cools the environment. On the contrary, condensation of water vapor (e.g., in the night hours) back to the liquid (e.g., dew formation, frost) releases the stored energy as heat and the surrounding environment is heated.

The sensible heat flux is driven by the temperature differences between the surface and the overlying air. Heat is initially transferred into the atmosphere by conduction. Then, with gradual heating of air, it circulates upwardly through convection. When the surface is warmer than the overlying air, heat will be transferred upwards into the air as positive sensible heat transfer. If the air is warmer than the surface, heat is transferred from the air to the surface creating a negative sensible heat transfer. Sensible heat is therefore part of the energy that warms the environment. We feel increasing surface and air temperature and can measure it with a thermometer.

Ground heat flux is less significant compared to the two previous components of the equation, it accounts for 5%–10% of net radiation. It ranges in summer months from 10 to 100 W m^{-2} for growing crops. The magnitude of this flux depends on the temperature gradient and the thermal conductivity that is affected by the mineral composition of the soil, its texture, and water content.

Bowen ratio is an important variable when evaluating ecosystems in terms of transformation of solar energy. It is defined as the ratio of sensible and latent heat flux. If a great portion of available energy at the surface is transformed into latent heat, the Bowen ratio is less than one. This is usually observed at wet surfaces, vegetation (forests, wetlands) and open water during a day. When major part of available energy becomes sensible heat flux, then Bowen ratio is equal or greater than one. This is typical for sealed, vegetation-free, and dry surfaces.

ET Terminology

Potential ET (ET_p)

Many methods have been developed for direct and indirect measurement of ET. The water loss (evaporative losses) from different surfaces such as turf, bare soil, and water was originally measured in large tanks (lysimeters). The term "lysimeter" was derived from the Greek words "*lysis*" and "*metron*" meaning dissolving and measuring, respectively. The term is applicable to any device utilized that measures the rate, amount, and composition of percolation of water through soil. In a simple term, the lysimeter can be defined as large containers packed with soil located in the field to represent field and environmental conditions, with bare soil or vegetated surfaces (field crops, trees, shrubs, grass, etc.) for measuring the ET of plants or evaporation from bare soil through a mass-balance approach. Lysimeters are expensive and labor-intensive tools to measure evaporation or ET. Thus, other meteorological approaches have been developed over the years to simplify the measurements of ET. One of the commonly used methodologies to determine ET will be discussed later.

The original ET equation was based on evaporation from free water surface as measured with lysimeters. The definition of potential ET that emerged from an earlier work implied a maximum value of ET when there was adequate amount of water to be transpired or evaporated. Formally, potential ET has been defined as "the evaporation from an extended surface of short green crop which fully shade the ground, exerts little or negligible resistance to the flow of water and is always supplied with water." Potential ET cannot exceed free water evaporation under the same weather conditions. However, in the definition of potential ET, the condition of nonlimiting supply of water is never achieved because the resistance of water flow through plants and soil has a finite value greater than zero. Another problem with this definition is the phrase "short green crop." The short green crop has been defined as 8–15 cm tall grass cover, but it has also been defined as a 30–50 cm tall crop of alfalfa. It is important to distinguish

between the short green vegetations because the ET rates from well-watered agricultural crops may be as much as 10%–30% greater than that occurring from short green grass. This dichotomy in the definition of a "short green crop" has led to the use of the term "reference crop ET (ET_{ref})." To eliminate the confusion, in late 1970s and early 1980s, engineers and practitioners introduced and started using the "reference ET" concept rather than "potential ET." The use of the term "potential ET" is diminishing rapidly and the term "reference ET" has been gaining significant acceptance by the water resources community.

Reference Evapotranspiration (ET_{ref})

The reference ET (ET_{ref}) concept was introduced by irrigation engineers and researchers to avoid the confusions that existed in the definition of potential ET. By adopting a reference crop (grass or alfalfa), it became easier and more practical to select consistent crop coefficients and to make reliable actual crop ET estimates in new areas. Introduction of the reference ET concept also helped to enhance the transferability of the crop coefficients from one location to another. Two reference crops have been used to represent the reference ET: grass and alfalfa.

Grass Reference Evapotranspiration (ET_o)

Grass reference ET is defined as "the rate of ET from a hypothetical reference crop with an assumed crop height of 0.08—0.12 m, a fixed surface resistance of 70 s m^{-1}, and an albedo of 0.23, closely resembling the ET from an extensive surface of green grass of uniform height, actively growing, well-watered, and completely shading the ground." In the grass reference ET definition, the grass is specifically defined as the reference crop and this crop is assumed to be free of water stress and diseases.

Alfalfa Reference Evapotranspiration (ET_r)

Alfalfa reference ET is defined as the ET rate from an extensive, uniform surface of dense, actively growing alfalfa, 0.30–0.50 m tall and not short of soil water. In the literature, the terms "reference ET" and "reference crop ET" have been used interchangeably and they both represent the same ET rate from a short, green alfalfa or grass surface. Unlike the potential ET definition, in the alfalfa reference ET definition, an alfalfa crop is specifically noted as the reference crop.

One of the other important differences between potential and reference ET is that the weather data collection site is well defined in the reference ET definition. It is important to note in the reference ET definition that the climate data that are used to estimate reference ET need to be collected in a well-watered and has certain characteristic (reference) environment. Therefore, based on the definition, the weather data for the reference ET estimations should be collected in a well-irrigated and well-maintained grass or alfalfa field. The irrigated grass area of the weather data collection site should be fairly large (e.g., at least 4 ha) to have enough fetch distance between the instrumentation to measure the climatic variables and the edge of the field because the quality of the weather data will ultimately affect the final estimated reference ET value. Enough fetch distance allows the air to travel on the reference crop surface and represent the aerodynamic, humidity, and temperature characteristics of the reference crop before it is sampled at the weather station. In a hot, dry month the average air temperature may be as much as 5°C–6°C higher in a dryland (nonirrigated) area than for a nearby well-irrigated area. The differences in the air temperature will also affect the relative humidity and vapor pressure deficit values, and these differences will ultimately cause differences in the reference ET calculated using the weather data collected from the two sites (dry vs. well-irrigated).

Determination of Crop ET (Plant Water Use) in Agroecosystems Using Climate Variables

In irrigated agroecosystems, a large part of the irrigation water applied to agricultural lands is consumed by evaporation and transpiration. In practice, in field measurements, it is hard to separate evaporation from transpiration, and the two processes are usually considered as one component. Crop ET can be measured directly using precision weighing lysimeters, Eddy correlation system, Bowen ratio energy balance system, atmometers, including evaporation pans, soil water balance by measuring soil water status continuously, etc. However, because direct measurement of crop ET (ET_c) is difficult, time consuming, and costly, the most common procedure is to estimate ET_c using climatic data. Currently, most commonly practiced way of estimating the crop ET rate (or crop water use rate) for a specific crop or vegetation surface requires first calculating reference ET (ET_{ref}) and then applying the crop coefficients (K_c) to estimate actual crop ET (ET_c) as Eq. (2).

$$ET_c = ET_{ref} \times K_c KT_c = ET_{ref} \times K_c \tag{2}$$

where ET_c is the crop ET (crop water use) in units of water depth (inches day^{-1}, cm day^{-1}, or mm day^{-1}), ET_{ref} (ET_o or ET_r) is the reference ET in unit of water depth (inches day^{-1}, cm day^{-1}, or mm day^{-1}) as calculated from the basic weather variables (solar radiation, air temperature, wind speed, and relative humidity) measured with a weather station in reference conditions.

Although the first equation by Penman for potential ET, (ET_p), was introduced almost 60 years ago; it still provides fundamental principles for the calculation and/or modification of ET models today. Numerous methods have been introduced for computing ET_{ref} causing confusion among users, decision- and policymakers as to which method to select for ET_{ref} estimation. Recently, the American Society of Civil Engineers (ASCE) Evapotranspiration in Irrigation and Hydrology Committee established a

Task Committee on "Standardization of Reference Evapotranspiration Calculation." Based on extensive research and data analyses and comparison of lysimeter-measured reference ET across various climates and Task Committee experience, the Task Committee recommended the use of the ASCE-Penman–Monteith (PM) method as the representation for reference ET. A reduced form of the ASCE-PM was used as the basis for "standardized" ET_{ref} computation. Equation parameters differ for hourly and 24-h data. Coefficients and parameters for a taller, rougher crop surface (0.5 m tall, like alfalfa) were also developed. The ASCE standardized ET_{ref} equation based on a surface resistance of 50 s m^{-1} during daytime and 200 s m^{-1} during nighttime provided the best agreement with the full form of the ASCE-PM method applied on a daily basis. The advantages of adapting a specific procedure as a standardized method are (1) it provides commonality to computing ET_{ref}, and (2) the use of a standardized method enhances the transferability of crop coefficients.

The standardized ASCE-PM equation is intended to simplify and clarify the application of the method and associated equations for computing aerodynamic and bulk surface resistance (r_a and r_s, respectively). Equations were combined into a single expression for both grass and alfalfa reference surfaces and for a 24-h or an hourly time step by varying coefficients. Computation of standardized short grass ET_o with a 24-h time step uses a grass height of 0.12 m and an r_s value of 70 s m^{-1}, which is the same as for the FAO56-PM equation. For hourly time steps, r_s is set to 50 s m^{-1} for daytime hours and to 200 s m^{-1} for nighttime hours. The standardized ASCE-PM equation is Eq. (3).

$$ET_{ref} = \frac{0.408\Delta(R_a = G) + \gamma(C_n/(T + 273))\,U_2(e_a = e_a)}{[\Delta + \gamma(1 + C_d\,U_2)]} \tag{3}$$

where ET_{ref} is the standardized reference ET (mm day^{-1} or mm h^{-1}), Δ is the slope of saturation vapor pressure versus air temperature curve (kPa°C^{-1}), R_n is the calculated net radiation at the crop surface (MJ m^{-2} day^{-1} for 24-h time steps or MJ m^{-2} h^{-1} for hourly time steps), G is the heat flux density at the soil surface (zero for 24-h time steps or MJ m^{-2} h^{-1} for hourly time steps), T is the mean daily or hourly air temperature at 1.5–2.5 m height (°C), U_2 is the mean daily or hourly wind speed at 2 m height (m s^{-1}), e_s is the saturation vapor pressure (kPa), e_a is the actual vapor pressure (kPa), $e_s - e_a$ is the vapor pressure deficit (kPa), γ is the psychrometric constant (kPa°C^{-1}), C_n is the numerator constant that changes with reference surface and calculation time step (C_n = 900°C mm s^3 Mg^{-1} day^{-1} for 24-h time steps, and C_n = 37°C mm s^3 Mg^{-1} h^{-1} for hourly time steps for the grass reference surface), C_d is the denominator constant that changes with reference surface and calculation time step (C_d = 0.34 s m^{-1} for 24-h time steps, C_d = 0.24 s m^{-1} for hourly time steps during daytime, and C_d = 0.96 s m^{-1} for hourly nighttime for the grass reference surface), and 0.408 is the coefficient having units of m^2 mm MJ^{-1}. The values of C_n and C_d for the grass and alfalfa reference surfaces for daily and hourly time steps are given in Table 1.

Crop Coefficient Concept

The K_c is the crop coefficient for a given crop and is usually determined experimentally. The K_c values represent the integrated effects of changes in leaf area, plant height, crop characteristics, irrigation method, rate of crop development, crop planting date, degree of canopy cover, canopy resistance, soil and climate conditions, and management practices. Each crop will have a set of specific crop coefficient and will predict different water use for different crops for different growth stages. An example of a K_c curve as a function of days or weeks after planting for a plant for initial, development, mid-season, and end-season stages is given in Fig. 5.

In general, crop growth stages can be divided into four main growth stages: initial, crop development, mid-season, and late season. The length of each of these stages depends on the climate, latitude, elevation, planting date, crop type, and cultural practices. Local field observations are best for determining the growth stage of the crop and adjust the empirical K_c values accordingly. Early in the growing season, during the crop germination and establishment stage, most of the ET occurs as evaporation from the soil surface. As the crop canopy develops and covers the soil surface, evaporation from the soil surface decreases and transpiration component of the ET increases.

Early in the season when plant is small, the water-use rate and K_c value are also small (K_c initial stage) and the crop ET rate increases as the plant develops (Fig. 5). For agronomical plants, the crop ET rate is at the maximum level when plant is fully developed (K_c mid-season). The ET rate decreases again when plant completes development and reaches physiological maturity toward the end of the season (K_c end season).

Table 1 Values for C_n and C_d in Eq. (3)

	Grass reference (ET_o)		*Alfalfa reference (ET_r)*			
Time step	C_n	C_d	C_n	C_d	*Units for ET_o and ET_r*	*Units for R_n and G*
Daily	900	0.34	1600	0.38	mm day^{-1}	MJ m^{-2} day^{-1}
Hourly during daytime	37	0.24	66	0.25	mm h^{-1}	MJ m^{-2} h^{-1}
Hourly during nighttime	37	0.96	66	1.7	mm h^{-1}	MJ m^{-2} h^{-1}

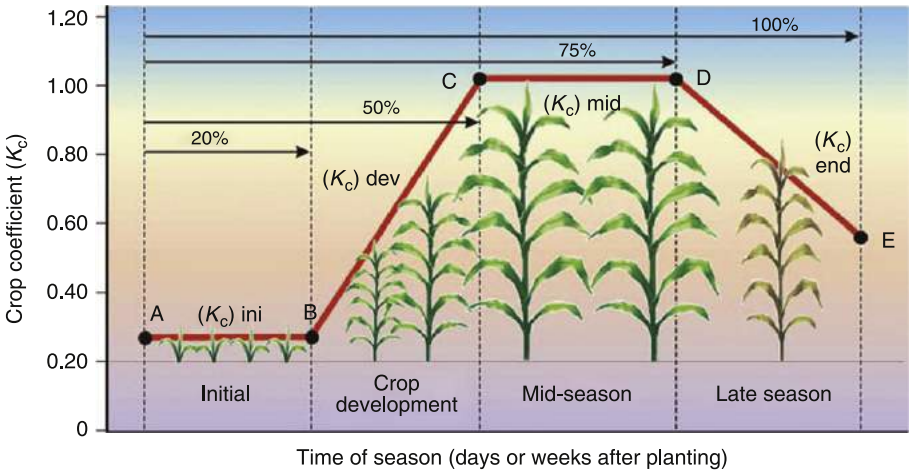


Fig. 5 Schematic representation of increase and decrease in crop coefficient based on different plant development stages.

For perennial crops a similar pattern can occur as the plant starts to develop canopy area, grow new shoots, and develop fruit. The percentage of leaf area, soil water status, and climatic conditions will drive the rate of crop (ET) at a given growth stage. Usually, the maximum canopy cover coincides with the time of year when the solar radiation and temperature are at their peak values (usually mid-season) and the maximum ET therefore occurs during that period. The K_c values for many different crops have been published in numerous literatures.

Cooling and Air-Conditioning Effect of Evapotranspiration

Transpiration by plants can be seen as a water loss in such cases as water scarcity; managers of water reservoirs that supply drinking water would usually see it as a loss. Transpiration is sometimes even called an unavoidable evil, in the sense that water is sacrificed for the sake of enabling intake of CO_2 for photosynthesis. For a plant, however, transpiration is a necessity by which a plant maintains its inner environment within the limit of optimal temperatures. It can be shown on basis of elementary physics that at the level of landscape, evapotranspiration is the most efficient air conditioning system developed by nature. In addition to optimizing temperature, through evapotranspiration plants control the optimum water balance in their root zone. Water, thanks to its high heat carrying capacity, is able to redistribute much of the solar heat energy received by the Earth through the water cycle: by evapotranspiration and condensation. Water evapotranspiration and condensation therefore plays an instrumental role in climate control with regard to temperature distribution in time and space, that is, reducing the peaks and modulating the amplitudes of high and low temperatures on the land surface—making conditions on Earth suitable for life. Just compare daily temperature fluctuation in desert and in forest in the same latitude.

Water has a unique feature. It exists in three aggregate states in our living environment: solid, liquid, and vapor. Phase transition from liquid into vapor is associated with changes of volume (18 mL of liquid forms 22,400 mL of vapor) and consumption or release of energy (0.68 kWh, 2.45 MJ kg^{-1} at 20°C), which is a cooling or heating environment. Water has high heat capacity, so its transformation involves exchange of energy, thus equalizing the temperature differences in time (day and night) and space (between different spaces).

Let us imagine a tree with a crown of 5 m in diameter covers an area of ca. 20 m^2. On a sunny day, at least 150 kWh of solar energy fall on the crown. What happens with this energy? 1% is used for photosynthesis, 10% is reflected in the form of light energy, 5%–10% is released as sensible heat and the same percentages transferred as ground heat flux into soil. The largest percentage enters the process of transpiration whereby water vapor is released from the tree. If a larger tree has a sufficient water supply, it can evaporate more than 100 L of water a day. In order to evaporate 100 L of water, approximately 70 kWh (250 MJ) of solar energy is needed. This energy is hidden in water vapor as latent heat and is released again during the process of condensation to liquid water.

The tree transpired around 100 L of water, thus cooling its environment by c. 70 kWh; during a 10-h period the tree cools its environment with a 7 kWh power output. Energy of 70 kWh did not appear as sensible heat, it stayed in form of water vapor and was released in cool places or during a night. Such a tree has a cooling capacity comparable with several technological air-conditioning system used in households, hotels, offices. Transpiring tree has a double air-conditioning effect: it cools when water evaporates and water vapor passes energy to cool places where latent heat is released when water vapor condensates back to water liquid.

From thermodynamic point of view, trees reduce gradients of energy between the Sun and outer space, they degrade incoming solar radiation through life processes. Sagan, Schneider imagine tree as a giant dissipative structure capturing sunlight and

degrading most of that energy as respiration and "low grade" latent heat via transpiration. Tree is like a giant water fountain spewing water in the form of latent heat. Trees well supplied with water reduce gradients which would realize as strong wind, torrential rain, etc.

Effect of ET on Local Climate

Climate change and global warming are widely believed to be caused only by an increase in CO_2 concentration from 250 to 390 ppm. Novel recent research, however, highlights the dynamic role of water vapor in climate change, with its concentration two orders of magnitude higher than that of other greenhouse gases. The implication of this research is that human landscape management affects the behavior of water vapor and its role in the dissipation of solar energy, in a much more important way than formerly appreciated.

This research has focused on wet meadows in the Czech Republic, which evapotranspirated about 7 mmol m^{-2} s^{-1} (i.e., 126 mg m^{-2} s^{-1}) during a sunny afternoon, converting about 315 W of energy per square meter of its surface into latent heat flux. The wetland, which covered an area of about 4 km^{-2}, evapotranspirated about 500 kg of water per second, which is equivalent to the flow rate of a small river. This invisible stream represents the latent heat flux of approximately 1260 MW. Thus, this ecosystem regulates the temperature through energy and water fluxes with a power equivalent to that of a moderately large power station. In drained or dry landscapes, wetland ecosystems thus act as "wet islands," important both for their conservation value and for their important hydrological function (in addition to their hydrologically dependent nutrient processing).

The drainage of large areas of natural vegetation and the loss of their latent heat function causes surprisingly large amounts of sensible heat to be released into the atmosphere. A drop in evapotranspiration by 1 L m^{-2} (equivalent to about 700 Wh) is capable of increasing the daily flux of sensible heat about 40 times more effectively (by 70 W) than the quoted effect of greenhouse gases [radiative forcing, Intergovernmental Panel on Climate Change (IPCC)]. For example, a drop in evapotranspiration of 1 mm over the territory of the Czech Republic (79,000 km^2) within a single day, releases an amount of sensible heat comparable to the annual production of electric energy from all Czech power plants (about 60,000 GWh).

The Czech study also measured the daily dynamics of radiation surface temperature and air temperature of different land cover types in a temperate, "cultural" landscape and their consequences for the local climate.

Typical rural area with seven localities with different land cover types were chosen in Trebon Biosphere Reserve, Czech Republic, Central Europe. A combined method of airship thermal scanning of Ts (radiation surface temperature) and ground measurement of thermodynamic Ta (air temperature measured in a meteorological screen at 2 m height) was used (**Fig. 6**). The localities differed markedly in both the values and the dynamics of Ts and Ts − Ta. In the early afternoon, the difference in Ts between the different land covers reached almost 20°C. Ecosystems with nonfunctional or no vegetation largely resembled the asphalt surface, whereas ecosystems covered with dense, bushy, or tree vegetation showed relatively well-balanced daily temperature dynamics with low temperature extremes and a slow temperature morning increase or afternoon decrease. Ts − Ta at the peak solar irradiance ranged between −1°C at the forest and 14°C–17°C at the dry harvested meadow and the asphalt surface, respectively (**Fig. 7**). Therefore surface radiation temperature (Ts) can be considered as a measurable indicator of ecosystem and landscape functioning, and the importance of functional vegetation for local climate should also be considered.

Landscape drying: It should be pointed out that air heated by warm surface and ascending into atmosphere contains water vapor. Landscape loses water with the upwards flowing warm air driven by sensible heat. The amount of water in the air transported by sensible heat high into atmosphere can be substantially higher than that released by evapotranspiration. For example, air of 100% relative humidity and temperature 40°C contains 50 g of water vapor in 1 m^3 that is, such air of 20% relative humidity contains still 10 g of water vapor in m^3. Air driven by sensible heat from 1 m^2 at speed 1 m s^{-1} would transport into atmosphere 36 kg water during 1 h. Common value of ET is several millimeters (several liters per m^2 per day). Very high value of ET is about 10 mm. From this point of view, ET can be considered as a process of slowing down evaporation water losses from landscape on regional level. ET binds surplus of solar energy into latent heat of water vapor and reduces release/production of sensible heat, water vapor is not driven up it stays close to the canopy.

Evapotranspiration of forests plays an extensive role in the transport of moisture from ocean in continents. It is evident that annual precipitations are high in continents with large and continuous forest from coast inside of continents (West Africa, Amazonia). The biotic pump theory (Makarieva and Gorshkov) suggests the atmospheric circulation that brings rainfall to continental interiors is driven and maintained by large, continuous areas of forest beginning from coasts. The theory explains that, through transpiration and condensation, forests actively create low pressure regions that draw in moist air from the oceans, thereby generating prevailing winds capable of carrying moisture and sustaining rainfall far within continents. Moreover, considerations of the surface pressure gradients created by the processes of evaporation and condensation, as highlighted in the biotic pump concept, may lead to improved predictions of large-scale climates compared to atmospheric circulation models which only consider temperature effects.

The biotic pump concept explains why moist winds blow readily from ocean to well forested land and how this flow can decline and reverse when forest cover is absent or depleted.

How does the biotic pump work? Natural forests maintain high transpiration fluxes which exceed the evaporation fluxes over the ocean. This moisture condenses as it rises. The resulting low-pressure zone draws in moist air from the ocean. Deforestation reduces evapotranspiration, condensation and hence this pressure difference, thus weakening or removing the coast-to-interior moisture transport (**Fig. 8**).

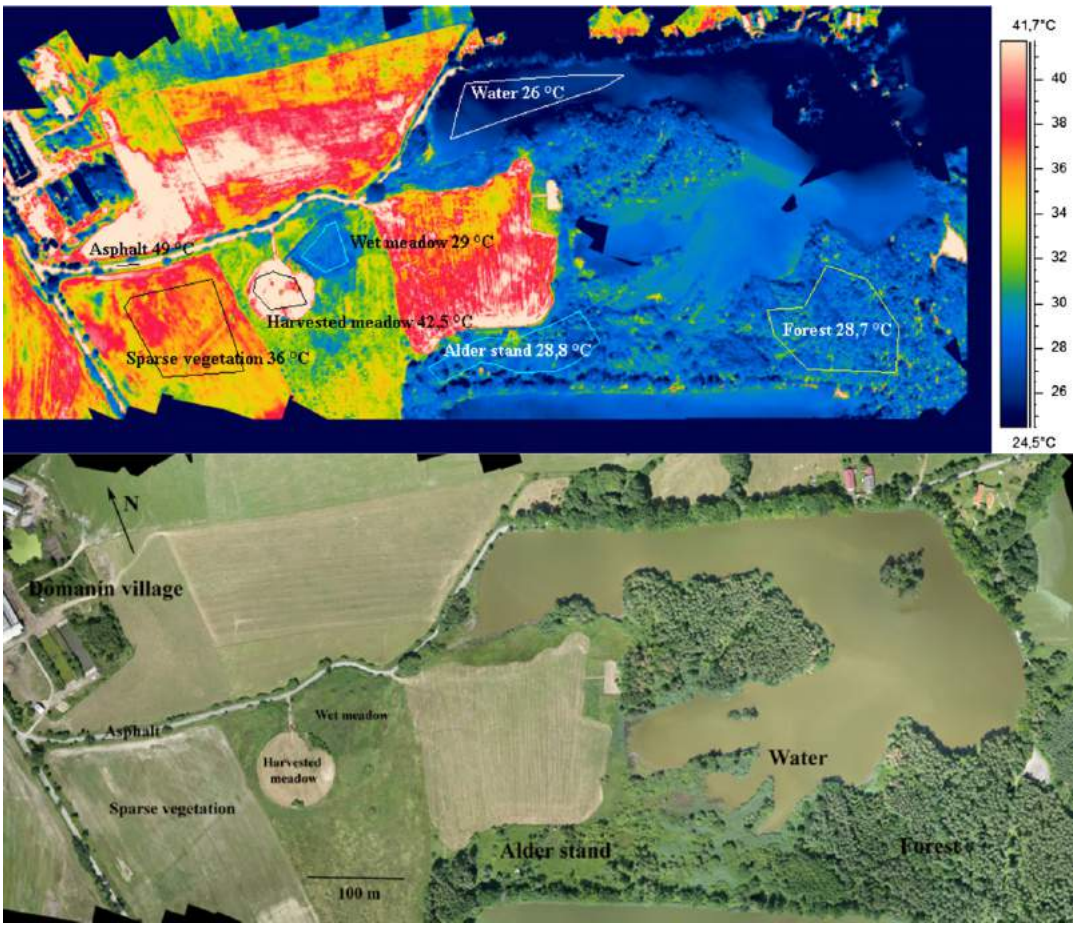


Fig. 6 Surface temperature of a "cultural" landscape on summer sunny day in T◆ebo◆ Biosphere Reserve (Czech Republic) at 2 pm, taken by thermographic and visible cameras carried by an airship.

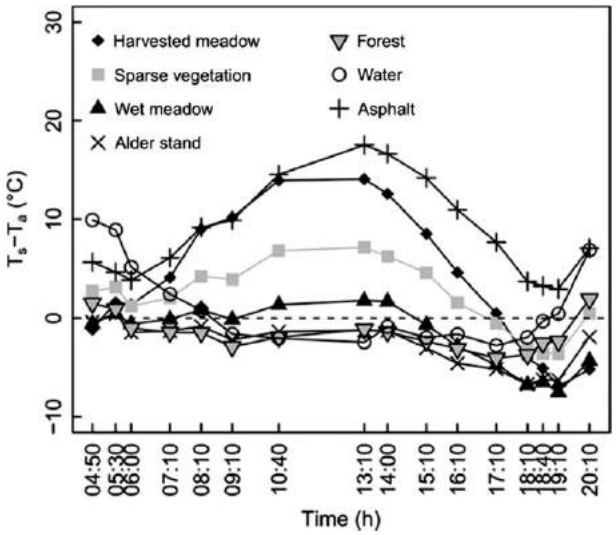


Fig. 7 Temperature differences Ts–Ta between surface Ts and air temperature Ta (at 2 m above ground under white screen) at all the studied localities. With permission from Hesslerová, P., Pokorný, J., Brom, J., Rejšková – Procházková, A. (2013). Daily dynamics of radiation surface temperature of different land cover types in a temperate cultural landscape: Consequences for the local climate. Ecol. Eng. **54**, 145–154. https://doi.org/10.1016/j.ecoleng.2013.01.036.

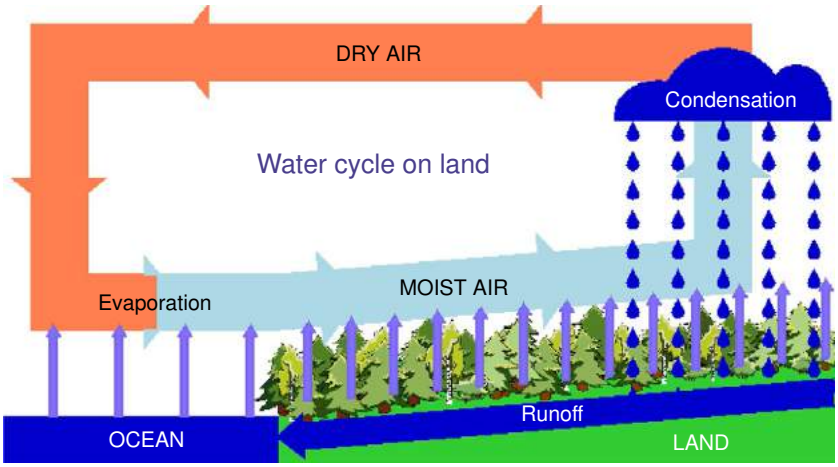


Fig. 8 How biotic pump ecology works.

Reliable rainfall in the continental interiors of Africa, South America, and elsewhere may thus be dependent on maintaining relatively intact and continuous forest cover from the coast.

A corollary of the biotic pump theory has further crucial implications for planetary air circulation patterns: if airflow patterns that move toward continental interiors are dependent upon the presence of forests, then their removal may foretell significant changes or wind pattern reversals. Reforestation and the restoration of degraded forest landscapes on an adequate scale may however reactivate such pumps, returning rainfall to continental interiors.

Acknowledgments

This article is a contribution of the University of Nebraska-Lincoln Extension, Journal Series No. 1037. The author expresses his appreciation to Sheila Smith, illustrator in the Department of Biological Systems Engineering at the University of Nebraska-Lincoln, for her excellent technical assistance in **Figs. 1** and **5**. In updated version figures from work by ENKI were used and results of work supported by TE02000077. Gratitude is expressed to V. Gorshkov and A. Makarieva for picture and text on biotic pump.

See also: General Ecology: Microclimate; Plant Physiology. Global Change Ecology: Energy Flows in the Biosphere; Water Cycle

Further Reading

Aboukhaled, A., Alfaro, A., Smith, M., 1982. Lysimeters. In: FAO irrigation and drainage paper no. 39. Rome: FAO, p. 68.

Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop evapotranspiration—Guidelines for computing crop water requirements. In: FAO irrigation and drainage paper no. 56. Rome: FAO, p. 300.

ASCE-EWRI, 2005. The ASCE standardized reference evapotranspiration equation. In: Allen, R.G., Walter, I.A., Elliot, R.L., *et al.* (Eds.), Environmental and Water Resources Institute (EWRI) of the American Society of Civil Engineers, ASCE, standardization of reference evapotranspiration task committee final report. Reston, VA: American Society of Civil Engineers (ASCE), p. 213.

Burman, R.D., Cuenca, R.H., Weiss, A., 1983. Techniques for estimating irrigation water requirements. In: Hillel, D. (Ed.), Advances in irrigation,vol. 2. Orlando, FL: Academic Press.

Ellison, D., Morris, C.E., Locatelli, B., *et al.*, 2017. Trees, forests and water: Cool insights for a hot world. Global Environmental Change 43, 51–61.

Huryna, H., Pokorny, J., 2016. The role of water and vegetation in the distribution of solar energy and local climate: A review. Folia Geobotanica 51, 191–208.

Irmak, S., Howell, T.A., Allen, R.G., Payero, J.O., Martin, D.L., 2005. Standardized ASCE-Penman–Monteith: Impact of sum-of-hourly vs. 24-hr-timestep computations at reference Weather Station sites. Transactions of the ASABE 48 (3), 1063–1077.

Itenfisu, D., Elliot, R.L., Allen, R.G., Walter, I.A., 2003. Comparison of reference evapotranspiration calculations as part of the ASCE standardization effort. Journal of the Irrigation and Drainage Engineering 129 (6), 440–448.

Johns, E.L., 1989. Water use by naturally occurring vegetation including an annotated bibliography. ASCE Task Committee on Water Requirements of Natural Vegetation Committee on Irrigation Water Requirements. New York, NY: American Society of Civil Engineers (ASCE), p. 216.

Kravčík, M., Pokorný, J., Kohutiar, J., Kováč, M., Tóth, E., 2008. Water for the recovery of the climate—A new water paradigm, People and Water NGO Slovakia. 122 pp.

Makarieva, A.M., Gorshkov, V.G., Li, B.-L., 2009. Precipitation on land versus distance from the ocean: Evidence for a forest pump of atmospheric moisture. Ecological Complexity 6, 302–307.

Mays, L.W., 1996. Water resources handbook. New York: McGraw-Hill.

Penman, H.L., 1948. Natural evaporation from open water, bare soil and grass. Proceedings of the Royal Society of London A 193, 120–146.

Pokorný, J., Brom, J., Čermák, J., Hesslerova, P., Huryna, H., Nadezhdina, N., Rejskova, A., 2010. Solar energy dissipation and temperature control by water and plants. International Journal of Water 5 (4), 311–336.

Rosenberg, N.J., Blad, B.L., Verma, S.B., 1983. Microclimate: The biological environment, 2nd edn New York: Wiley.
Shiklomanov, I., 1993. World fresh water resources. In: Gleick, P. (Ed.), Water in crisis. Oxford/New York: Oxford University Press, pp. 13–24. (Chapter 2).
United States Department of Agriculture, 1982. Food-from farm table 1982 yearbook of agriculture. US, Washington, DC: Government Printing Office.

Evolutionary Ecology: Evolution of Parasitism☆

Gabriele Sorci and Stéphane Garnier, University of Bourgogne Franche-Comté, Dijon, France

Introduction

It is usually believed that all organisms on earth are involved in host–parasite interactions, either as a host or as a parasite. Parasites are not the only organisms that are intimately associated with another individual for their growth, reproduction and survival. The evolution of eukaryotes has been tightly linked with the establishment of permanent and obligate coexistence of genetic entities that were formerly able to live independently from each other. These kind of obligate associations are extremely numerous and can be classified according to the relative net effect that each partner inflicts on the other. The overall spectrum, therefore, goes from associations where both partners benefit from each other presence, to the other extreme where the association comes to the exclusive benefit of one partner at the expense of the other. This latter case is what is commonly called a host–parasite interaction. Here, the parasite entirely depends on the host for its growth, reproduction and survival. The metabolic resources necessary for the vital functions of the parasite are provided by the host, which in turn receives nothing by the parasite. This results in a net negative effect of parasitism on host fitness, as the resources consumed by the parasite are no longer available for the growth, reproduction and survival of the host. In this context, it is easy to understand that parasites are selected to exploit the host in the most effective way and hosts are selected to limit the negative effect of parasites. Therefore, the evolution of parasitism and the exploitation strategies that come along, cannot be easily disentangled from the evolution of host counteradaptations to resist parasitic attacks. This sets the scene for a coevolutionary scenario where hosts and parasites are endlessly selected to respond to the threat provided by the opponent.

Parasitism has evolved from an ancestral nonparasitic form. Although it is difficult (or impossible) to assess the selective pressures that have promoted the shift from a nonparasitic to a parasitic life style, the use of modern phylogenetic tools has provided very interesting insights on the evolutionary history of parasitism. Comparison of closely related extant taxa that differ in their life style (parasitic vs. nonparasitic) also provides an invaluable tool to study the traits (adaptations) that are associated with a parasitic mode of life and to infer the selection pressures that act on these traits.

Parasites are an important factor promoting the evolution of hosts because of their negative effect on them. This negative effect is generally called virulence and as any other traits of parasites, it evolves in response to environmental factors. Understanding how virulence evolves is obviously an important step if we want to fully assess the threat that parasites impose on humans, domestic animals, and wildlife, as well as to guide us to take the proper decisions on how control infectious diseases. In this article, we will go through these different aspects of parasite evolution, starting with a snapshot on the impressive diversity of organisms with a parasitic life style.

Parasite Diversity

Although it is difficult to date precisely the emergence of parasitism in living organisms (and even more difficult within each phylum), paleontological data show that parasitism is a very old life style. Several parasites have been identified in fossils from the Cretaceous period (-135 to -72 MY), and the older traces suggesting parasitism date back to late Cambrian (-497 to -485 MY). The first parasites surely evolved from free-living organisms, simply because no one around means nobody to parasitize. One thing is for sure—there is today a considerable diversity of parasites which can be seen from different angles (taxonomy, life cycles, hosts colonized, transmission modes).

Parasites are widespread among living organisms, as they are thought to represent $>50\%$ of all species on earth. From the simplest viral particle to avian brood parasites, parasitic lifestyle has regularly arisen during evolutionary time. Not all phylogenetic groups have been equally prone to the evolution of parasitism since the proportion of parasitic species is very heterogeneous among taxa. Whereas all viruses are parasites, no parasite is known in echinoderms. Nearly all platyhelminthes are parasites, whereas a unique species of cnidarian is parasitic among the 11,000 species described in this phylum, so far. Finally, some groups contain roughly equal proportion of parasitic and free-living species, such as bacteria, fungi, and nematodes. This heterogeneity raises two main questions. The first one involves the mechanisms responsible for the evolution of this taxonomic diversity, which may result from speciation events (formation of two different parasitic species from a single one) or transitions from free-living stages toward parasitism. Both mechanisms have played a role in parasite history, but they have different implications. Within the context of speciation, parasitic life style is simply inherited from a common ancestor, whereas transition events mean independent acquisitions of parasitism by several species. How many transitions toward parasitism have punctuated the evolution of life? It is impossible to give a precise answer to this question because the number of parasitic lineages which became extinct is unknown.

☆*Change History*: February 2018. Gabriele Sorci & Stéphane Garnier updated section introduction, parasite diversity, competition and further reading.

This is an update of G. Sorci and S. Garnier, Parasitism, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 2645–2650.

 doi:10.1016/B978-0-12-409548-9.11094-2

However, phylogenetic analyses have shown no less than 63 independent transitions in the metazoan phylogeny, which contains >100,000 described species (Table 1). Again, the frequency of these events varies among taxonomic groups. For example, only one transition has occurred in cnidarian, whereas parasitism has evolved several independent times in nematodes.

This leads to the second question. Are some taxonomic groups more prone to evolve toward parasitism? Probably yes. A mutation occurring in a free-living organism might enable it to exploit another organism (which needs to be frequently encountered for an intimate interaction to be established). If the mutation provides the individuals with a slight advantage in term of reproductive success, a parasitic lifestyle will be favored by natural selection. However, it is unlikely that a single mutation would allow a free-living species to exploit a host without preadaptations for survival, feeding or reproduction within the host. It is therefore possible that traits currently associated with successful parasitism are exaptations (secondary functions of preexisting traits not directly linked to parasitism). In agreement with this view, the genomic comparison of parasitic trypanosomatids and a free-living relative showed that many of the unique characteristics of trypanosomatids, thought to be tightly linked with their parasitic lifestyle, are already present in the free-living relative *Bodo saltans*. However, once the progression toward a parasitic lifestyle is engaged, the new parasitic lineage is usually subject to morpho-anatomical (such as reduction/loss of sense or digestive organs), functional or physiological changes, leading to a stronger dependence of the parasite upon its host. It seems, however, that this process is not irreversible because the phylogeny of Diplomonadida (a group of protozoa) suggests that reversal to free-living stages has occurred twice.

Parasitism is also characterized by a great diversity of life cycles. In the simplest case, the parasite only needs one host to complete its life cycle. Several fungi, such as mildew, are plant parasites. Once a spore reaches a leaf, the fungus penetrates the plant, matures and starts to produce and release spores in the environment. These spores have to encounter another leaf to begin a new cycle. Complex life cycles involve two or more (up to four) hosts, each host housing a different developmental stage of the parasite (Fig. 1). Adults of the trematode *Halipegus ovocaudatus* live under the tongue of green frogs where they reproduce sexually. Eggs are released in the water and the emerging parasite needs to pass through three intermediate hosts (a mollusk, a copepod and a dragonfly). If a parasitized dragonfly is eaten by a frog, the cycle is eventually completed.

Table 1 Minimum number of evolutionary transitions to parasitism (sensu stricto) and number of living species in the major groups of metazoan parasites of metazoan hosts

Parasite taxon	*Minimum number of transitions*	*Minimum number of living species*
Phylum Mesozoa	1	>80
Phylum Plathelminthes[a]		
Class Cercomeridea (subclasses Trematoda, Monogenea and Cestoidea)	1	>40,000
Phylum Nemertinea[a]	1	>10
Phylum Acanthocephala	1	>1,200
Phylum Nematomorpha	1	>350
Phylum Nematoda[a]	4	>10,500
Phylum Mollusca[a]		
Class Bivalvia[a]	1	>600
Class Gastropoda[a]	8	>5,000
Phylum Annelida[a]		
Class Hirudinea[a]	3	>400
Class Polychaeta[a]	1	>20
Phylum Pentastomida	1	>100
Phylum Arthropoda[a]		
Subphylum Chelicerata[a]		
Class Arachnida[a]		
Subclass Ixodida	1	>800
Subclass Acari[a]	2	>30,000
Subphylum Crustacea[a]		
Class Branchiura	1	>150
Class Copepoda[a]	9	>4,000
Class Cirripedia[a]		
Subclass Ascothoracida	1	>100
Subclass Rhizocephala	1	>260
Class Malacostraca[a]		
Order Isopoda[a]	4	>600
Order Amphipoda[a]	17	>250
Subphylum Uniramia[a]		
Class Insecta[a]		
Order Diptera[a]	2	>2,300
Order Phthiraptera (suborders Ischnocera, Amblycera and Anoplura)	1	>3,000
Order Siphonaptera	1	>2,500

[a]Denotes taxa containing free-living species.

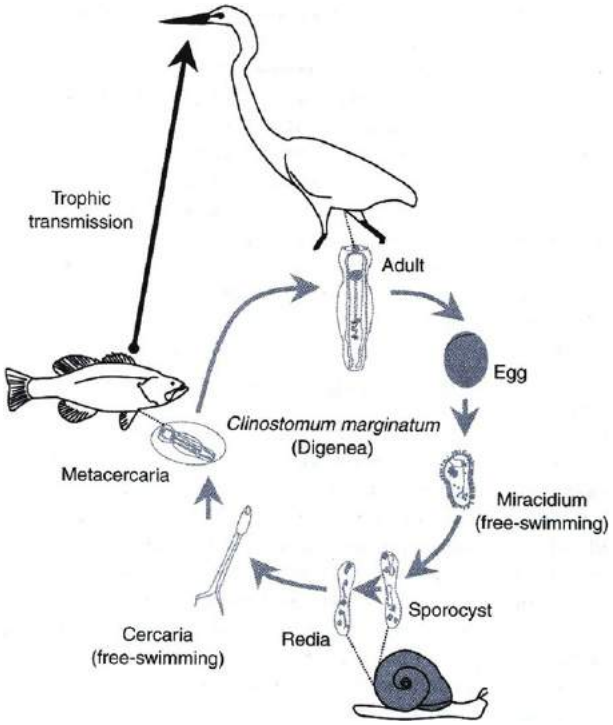


Fig. 1 A complex life cycle. The life cycle of the digenean *Clinostomum marginatum* involves three hosts. Adult worms parasitize the intestine of egrets and other fish-eating birds, where they produce eggs that are dropped into water with the bird's feces. A miracidium hatches out of the egg and swims until it finds a snail and infects it. The miracidium sheds its cilia and develops into a sporocyst, which then produces multiple redia. The redia produce multiple cercaria, which leave the snail and swim until they find a fish to infect, and then develop into metacercaria. Predation of fish by birds facilitates the completion of the parasites life cycle. Reproduced with permission from Thomas, F., Renaud, F. and Guégan, J. F. (2005). *Parasitism and ecosystems*. New York: Oxford University Press.

It seems parsimonious to assume that early evolutionary stages involved simple life cycles, with direct transmission between hosts of the same species. Understanding the evolution of complex, multispecies, life cycles has been more puzzling, as adding intermediate hosts might increase the risk of failing to encounter the right host. Theoretical work has explored the conditions that might have led to the evolution of complex life cycles in helminth parasites with no penetrative infectious stages. Two scenarios, and the benefits associated with each of them, have been put forward. According to the first scenario, transition from a single to a multihost life cycle can be attributed to upward incorporation of a new host which preys upon the original host (**Fig. 2**). In this scenario, benefits for the parasites are avoidance of mortality when the host is eaten by the predator, greater body size at maturity and fecundity. In the second scenario, incorporation comes downwards by adding an host at a lower trophic level (**Fig. 2**). This new host can initially be a paratenic (facultative) host, which later becomes an obligate host, if this enhances transmission rate to the definitive host. Although the addition of a paratenic host in a life cycle is an accidental event, complex life cycles of parasites are probably adaptive responses to the two main hurdles set by their environment, namely the transmission between hosts, and the compatibility between the parasite and the host(s).

Exploitation and Transmission Strategies

The success of a parasitic infection depends on two key steps: the ability of the parasite to establish within the host (to grow and reproduce) and the likelihood of propagules to be transmitted to a novel host. These two steps are associated with a number of adaptations aiming at maximizing parasite fitness. The strategies adopted to exploit the host and to transmit to other hosts can involve different traits. We will briefly discuss some exploitation and transmission strategies encountered in parasitic species.

Compared to most free-living organisms, parasites experience a relatively constant environment: the host. In most cases, the host provides a relatively constant food supply, constant temperature (for endotherms), shelter and protection toward predators. Parasites have adapted to this particular milieu with an array of adaptations ranging from special structures to attach to host structures, to regression of organs no longer needed for a life within a host. The view that the host provides a benign environment for the parasites is, however, too simplistic. Although living in the blood vessels of the host certainly provides shelter and protection, it exposes the parasite to the attack of a particular form of predators: the cells and molecules of the immune system. The immune system is probably the most sophisticated, although not unique, defense mechanism hosts have evolved. Theoretical models have analyzed the impact of host immunity on parasite life histories, using the framework that has been developed to

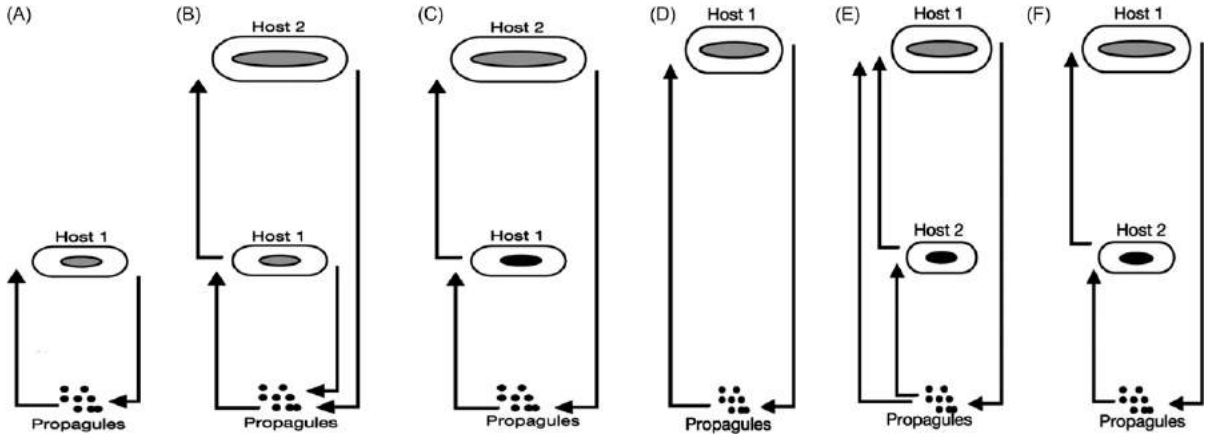


Fig. 2 Transition from a one- to a two-host cycle by upward incorporation of a definitive host (A–C) or by downward incorporation of an intermediate host (D–F). (A and D) The initial life cycle involves one host. (B) Host 1 is frequently ingested by a predator (potential host 2), resulting in a flexible two-host cycle (reproduction of the parasite in both hosts). (C) Reproduction in host 1 is suppressed, leading to a two-host cycle in which host 1 has become an intermediate host. (E) Propagules sometimes enter potential host 2, which can be ingested by host 1 (host 1 can be directly infected or indirectly via host 2). (F) Direct transmission to host 1 may later be lost. The *gray area* indicates adult parasites in definitive hosts; *black areas* indicate immature parasites in intermediate hosts. Reproduced with permission from Parker, G. A., Chubb, J. C., Ball, M. A. and Roberts, G. N. (2003). Evolution of complex life cycles in helminth parasites. *Nature* **425**, 480–484.

model the interactions between prey (here the parasite) and predators (here cytotoxic lymphocytes). These models have shown how pervasive the effect of host immunity can be on the dynamics and the virulence of the infection. Therefore, it is not surprising that parasites have adopted a series of strategies aiming at escaping the host immune response. Viruses, bacteria, protozoa, but also macroparasites such as helminths, can escape the immune response by hiding from or suppressing it. Antigenic variation is one of such strategies, where the same strain of a given microparasite expresses different antigenic epitopes. Antigenic variation allows the parasite to escape the immune response, since epitopes that are first recognized by the immune system are gradually replaced as long as the infection progresses.

As mentioned above, a successful parasite is a parasite that transmits its progeny to other hosts. Thus, transmission is tightly linked to fitness in parasitic species. However, transmission to another host is also associated with high mortality risk. A propagule has to face the hostile external environment during a period that can be quite long, it has to encounter the appropriate host, and finally enter it. Everything else being equal, large fecundities can compensate for the high risk of mortality incurred by propagules during their free-living stage. Several parasite species are well known for their fecundity records with several millions eggs produced during a lifetime. Another way to reduce the hazard of external life and predation is to add an intermediate host and modify its behavior to favor transmission to the final host. This kind of upward incorporation of hosts has been suggested to be at the origin of complex parasite life cycles (see above). Parasites with heteroxenous life cycles, however, face another dilemma. Their cycle is completed only when the intermediate host is eaten by the definitive host. Any mechanisms that make the intermediate host more susceptible to predation directly favor parasite transmission. This process of parasite manipulation (the parasite modifies the morphology and/or behavior of its intermediate host to increase the chance it will be preyed upon) is surprisingly very widespread (it has been reported in several species of protozoa and metazoan parasites) and can take unsuspected forms. Classical examples of parasite manipulation of host behavior include the effect of the digenean *Dicrocoelium dendriticum* on its ant intermediate host. *Dicrocoelium* causes infected ants to climb to the tip of grass blades and stay there waiting for a grazing sheep. As one might expect, sheep are the definitive hosts of *Dicrocoelium*. Other digenean parasites, such as species of the genus *Leucochloridium* are known to alter the shape, size and coloration of the tentacles of the snail they exploit. Modified tentacles strikingly resemble caterpillars which are likely to be detected by birds which are the definitive hosts of the parasite. However, such altered host phenotype facilitates the transmission of the parasite to the definitive host only if predation occurs when the parasite has reached the appropriate transmissible stage. If predation of the intermediate host occurs before the parasite has matured into the transmissible stage, the parasite will die with its intermediate host. It could be expected then that the optimal parasitic strategy to maximize transmission success should be to protect the intermediate host from predation until the parasite reaches the transmissible stage. This early reduced and late increased predation risk has been reported in the tapeworm *Schistocephalus solidus* that exploits copepods as intermediate hosts and fish as final hosts.

Competition

In nature, hosts are frequently infected by multiple parasite species. Depending on environmental conditions and the order of infection, the different species present in the same individual host, at the same time, can interact or not, depending on

whether they have overlapping or nonoverlapping niches (the range of environmental conditions experienced by the species). Two kinds of interactions have received much attention: interspecific competition and host defense mediated interactions.

Opportunities for competition between two species exist when both species exploit the same resource in the host (food, space). The performance (in terms of growth, reproduction, survival) of one species is reduced when the second species is present, either because the amount of available resources has been depleted (competition through exploitation) or because one species prevents the other from exploiting some resources via direct interaction (competition through interference). Asymmetrical competition has been often reported for parasites, when the species A suffers from the presence of species B whereas species A has no effect on species B. This pattern suggests a one-sided interference rather than exploitative competition, but it could also be due to host-mediated effects involving the immune response.

Competition is thought to be one of the main factors driving the assemblage of species within communities. When two parasite species with overlapping realized niche in isolation, co-occur in the same host, the realized niche of at least one of them might be changed (compared to when occurring alone) to minimize competition. This process is called interactive site segregation. There are many examples of site segregation in helminths inhabiting the digestive tract of their hosts, sometimes leading to a more or less regular spreading of species along the digestive tract. If there is some genetic basis for the niche preference, and if the cost of competition (in terms of fitness) outweighs the cost of realized niche change (because of less optimal environmental conditions), then this change may be selected for and become genetically fixed. This can lead to a complete segregation of the fundamental niche, where species have nonoverlapping niches whether alone or together in the same host.

Host immunity can also mediate and affect the outcome of the competition between parasite species. As mentioned above, immunosuppression is a common strategy adopted by parasites to persist within the host. Host immunosuppression can, however, have profound effects on the colonization and the population dynamics of other parasites. The spread of opportunistic diseases following infection with HIV virus is one of the most striking examples of this host immunity mediated interaction. A meta-analysis of published work has shown that resource-based (bottom-up) and predator-based (top-down) mechanisms can actually shape the pattern of coinfection between helminths and microparasites, depending on whether they compete for similar resources or whether they interfere with the host immune system.

Coevolution

It should be clear by now that whatever the parasite trait we consider, its evolution can only be understood in the light of the selection pressures exerted by the environment. A striking and major difference between free-living and parasitic species is that free-living organisms evolve in response to selection pressures exerted by both abiotic and biotic factors, whereas parasites almost uniquely respond to the selection pressure exerted by their hosts (although this does not apply to parasitic species that spend a considerable part of their life outside or not in contact with a host). This means that parasite evolution cannot be envisaged other than in the light of host evolution. In other words, hosts and parasites are involved in a process of coevolution where the emergence and spread of a trait in the parasite (i.e., a trait that confers a better ability to exploit the host) select for a specific response in the host and vice versa the emergence and spread of a trait in the host (i.e., a defense mechanism) select for a specific trait in the parasite.

One particular group of parasites provides, probably, the best illustration of the coevolutionary process. As mentioned above, parasitism in not restricted to microorganisms and invertebrates. Some birds, such as cuckoos and cowbirds, have also adopted a parasitic lifestyle. Of course, cuckoos do not develop inside a host and do not consume host resources as a microorganism would do. Cuckoos and cowbirds exploit a particular resource of the host: parental care. Cuckoos cannot reproduce unless they find an appropriate host (another bird species) that takes care of their eggs and nestlings. To complete its life cycle (producing offspring) a cuckoo needs a host, which fully includes it in the category of parasites. Outside the reproductive season, parasitic cuckoos behave as any other nonparasitic relative. The parasitic behavior is, therefore, clearly restricted to a particular stage of the life cycle of these birds. To be successful, a female cuckoo has to find a suitable host (usually a passerine species) and lay an egg in the nest of the host. If the cuckoo egg is incubated by the female host, a cuckoo nestling will hatch, usually before any other host nestlings. The impact of cuckoo parasitism on host reproductive success is dramatic as the cuckoo nestling ejects all host eggs and nestlings out from the nest, reducing the host brood to a single parasitic nestling. Given the cost for host fitness, it is straightforward to expect that hosts have evolved a set of traits aiming at reducing the risk of brood parasitism and, in turn, brood parasites have evolved a series of strategies to overcome host defense. **Table 2** summarizes the most prominent host adaptations and parasite counteradaptations involved in the coevolutionary process between brood parasites and their hosts. Why is this example particularly relevant to illustrate the coevolutionary process? The specificity (usually brood parasite species or host races exploit a single host species, even though exceptions exist) and the particular nature of the traits involved limit the chances that the presumed adaptation has arisen because of indirect selection exerted by any other source not involved in the interaction.

Even though more diffuse, coevolution is a major and pervasive characteristic of host–parasite interactions. Understanding the evolution of parasitic strategies and lifestyle is, therefore, a particularly tough task as it has and currently still responds to the selection pressures exerted by other living and evolving organisms, the hosts.

Table 2 Adaptations and counteradaptations involved in the coevolutionary process between brood parasites and their hosts

Parasite exploitation strategy	*Host defense*	*Parasite counteradaptation*
Finding a suitable host nest in the appropriate breeding phenology. Laying an egg	Defending the nest against the intruder	Dropping the egg in the nest within few seconds when the host has left the nest unattended
	Spotting any strange egg in the clutch and reject or destroy any egg differing from the other eggs of the clutch	Laying mimetic eggs that match as close as possible host eggs in size and color. Lay thicker eggs that resist dropping in the nest and host puncturing
Monopolize host resources (parental care). Incubation time is shorter in the parasite than the host. As soon as the parasite hatches, it ejects host eggs or kills/outcompetes host nestlings (nestlings of the European cuckoo have evolved a particular structure in the back that, like a spoon, allows them to eject host eggs out of the nest rim). Cuckoo nestlings provide superstimuli (visual and vocal) to the foster parents to obtain sufficient food	Desert the brood[a]	Nestling mimicry[a]

[a]Host defense against parasitic nestlings has been reported for the interaction between the super fairy-wren (*Malurus cyaneus*) and the Horsfield's bronze-cuckoo (*Chrysococcyx basalis*).

See also: Evolutionary Ecology: Allee Effects

Further Reading

Combes, C., 2001. Parasitism. The ecology and evolution of intimate interactions. Chicago, IL: The University of Chicago Press.
Cox, F.E.G., 2001. Concomitant infections, parasites and immune responses. Parasitology 122, S23–S38.
Davies, N.B., 2000. Cuckoos, cowbirds and other cheats. London: T & A D Poyser.
Frank, S.A., 2002. Immunology and evolution of infectious disease. Princeton, NJ: Princeton University Press.
Goater, T.M., Goater, C.P., Esch, G.W., 2013. Parasitism: The diversity and ecology of animal parasites. Cambridge: Cambridge University Press.
Graham, A.L., 2008. Ecological rules governing helminth-microparasite coinfection. Proceedings of the National Academy of Sciences of the United States of America 105, 566–570.
Hammerschmidt, K., Koch, K., Milinski, M., Chubb, J.C., Parker, G.A., 2009. When to go: Optimization of host switching in parasites with complex life cycles. Evolution 63, 1976–1986.
Hughes, D.P., Brodeur, J., Thomas, F., 2012. Host manipulation by parasites. Oxford: Oxford University Press.
Jackson, A.P., Otto, T.D., Aslett, M., Armstrong, S.D., Bringaud, F., Schlacht, A., Hartley, C., Sanders, M., Wastling, J.M., Dacks, J.B., Acosta-Serrano, A., Field, M.C., Ginger, M.L., Berriman, M., 2016. Kinetoplastid phylogenomics reveals the evolutionary innovations associated with the origins of parasitism. Current Biology 26, 161–172.
Leung, T.L.F., 2015. Fossils of parasites: What can the fossil record tell us about the evolution of parasitism? Biological Reviews 92, 410–430.
Parker, G.A., Chubb, J.C., Ball, M.A., Roberts, G.N., 2003. Evolution of complex life cycles in helminth parasites. Nature 425, 480–484.
Poulin, R., Morand, S., 2000. The diversity of parasites. The Quarterly Review of Biology 75, 277–293.
Schmid-Hempel, P., 2011. Evolutionary parasitology. New York: Oxford University Press.
Thomas, F., Renaud, F., Guégan, J.F., 2005. Parasitism and ecosystems. New York: Oxford University Press.
Thomas, F., Guégan, J.F., Renaud, F., 2008. Ecology and evolution of parasitism. New York: Oxford University Press.

Fermentation☆

M Ciani and F Comitini, Università Politecnica delle Marche, Ancona, Italy
I Mannazzu, Università degli Studi di Sassari, Sassari, Italy

Metabolic Biodiversity

In the biological world, the enormous assortment of energy sources can be used in many different ways to support the growth of organisms. The characteristics of the different pathways that can be used to produce the energy that arises from coupled oxidation–reduction reactions and is needed for living processes are listed in Table 1.

The Fermentation Process

When respiration is not possible, due to either a lack of an external electron acceptor or an impairment of the respiratory chain, fermentation is the catabolic pathway that is used for the production of energy from the partial oxidation of glucose or other carbon sources (Figure 1). The oxidation of the substrate, which occurs mainly through the Embden–Meyerhoff and Parnas (EMP) or Entner–Doudoroff (ED) pathways, results in the production of pyruvate, adenosine triphosphate (ATP), and nicotinamide adenine dinucleotide phosphate (NAD(P)H). In the absence of external electron acceptors, the pyruvate or the other organic compounds that derive from the initial substrate reaction undergo reduction, with the regeneration of $NAD^+(P)$. This step is essential for the fermentation process to progress and it leads to the production of waste products (ethanol and organic acids) that are excreted from the cell. ATP is the main product of fermentation, and it is generated by phosphorylation at the substrate level. An exception to this general rule is seen with the fermentation of carboxylic acids. The catabolism of these substrates generates a gradient of H^+ or Na^+ ions across the plasma membrane, and the production of ATP involves the activity of the H^+ or Na^+ membrane ATPases.

While the complete oxidation of 1 mol of glucose to CO_2 through oxidative phosphorylation (respiration) generates up to 38 mol of ATP, fermentation produces only a few moles of ATP (1–3) per mole of glucose. Thus, the recovery of energy from fermentation is rather low compared to that yielded by respiration. Moreover, it varies depending on the initial substrate and the fermentation process itself.

Ecological Distribution of Fermentation Processes

Fermentative metabolism is widespread among living organisms, and the ecology of fermentative processes is particularly complicated due to the ability of different organisms to ferment a plethora of substrates under different environmental conditions. Fermentation is carried out in anoxic environments by strict anaerobic or facultative anaerobic microorganisms, although some microaerophilic or facultative anaerobic microorganisms are also able to carry out fermentation in the presence of oxygen.

☆*Change History:* June 2013. M Ciani introduced small edits in the text of the article, added the sections Controlled Mix Fermentation in Wine, Fermented Vegetables and Fruits through Lactic Acid Bacteria and Fermented Milks: Kefir and Kumis as Matrix of Microbial Interactions, and added Figure 6.

Table 1 Metabolic biodiversity among living organisms

Metabolic diversity	*Characteristics*
Phototrophy	When radiant energy is absorbed by chlorophyll or other similar pigments, resulting in an excitation of the electrons present in the complex and providing oxidation that produces oxygen (oxygenic photosynthesis) or does not produce oxygen (nonoxygenic photosynthesis)
Aerobic respiration	Molecular oxygen is the final acceptor of electrons in a redox reaction and appears in reduced form as water
Anaerobic respiration	Under anaerobic conditions where molecular oxygen is absent or limited, inorganic ions (nitrate, sulphate, carbonate) serve as final acceptors of electrons
Fermentation	An organic compound that is often a metabolic intermediate coming from oxidation of an organic compound serves as terminal oxidant, producing a more reduced organic molecule as the metabolic end product

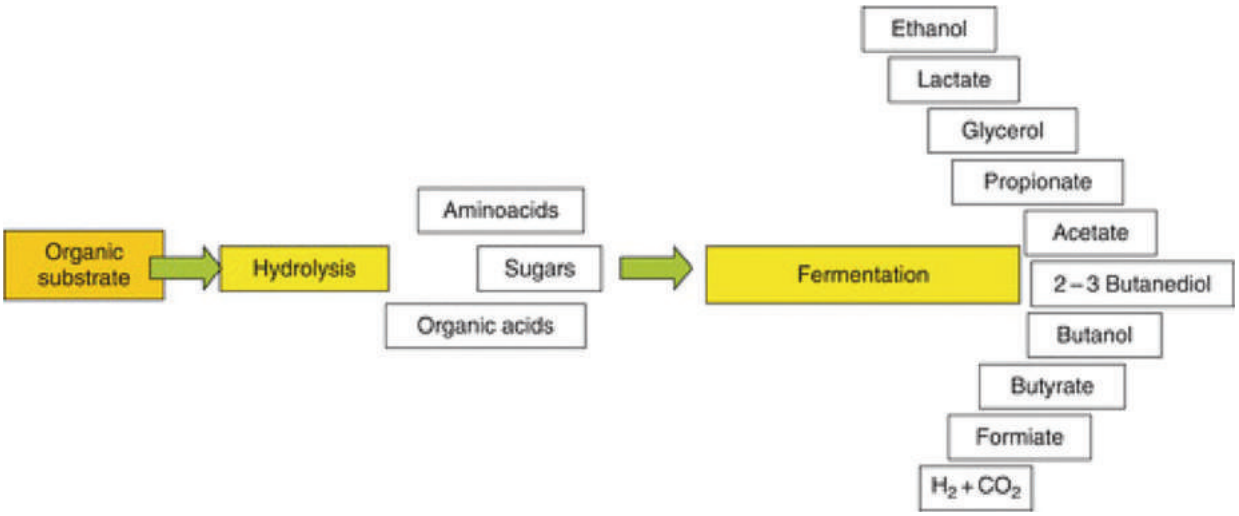


Figure 1 Main end products of the various fermentation processes.

Microaerophilic fermenting microorganisms, such as lactic acid bacteria, colonize habitats that are characterized by low oxygen concentrations or the absence of oxygen (often *c*. 10% of atmospheric levels). These microaerophilic microorganisms show a wide ecological distribution in well-defined habitats, such as the animal and human oral cavity, the gastro-intestinal tract, and feces, as well as in fermented meat, beverages, vegetables (silage, olive brine, sauerkraut), and dairy products. Lactic acid bacteria have limited biosynthetic ability and require pre-formed amino acids, group B vitamins, purines, and pyrimidines. These multiple requirements restrict their growth to habitats where the required compounds are abundant. The fermentation activity of lactic acid bacteria produces large amounts of lactic acid, which can cause a drop in pH to about 4.0, and thus inhibit the growth of most other bacteria and exert an antagonistic effect on spoilage microorganisms and the most common human pathogens. For these reasons, the transformation of food and beverages by fermentation is one of the main modalities for the conservation of food products and the enhancement of their quality.

Besides microaerophilic lactic acid bacteria, anaerobic facultative microorganisms such as yeasts are the main group of microorganisms that are involved in the transformation of fermented food. In contrast to the facultative anaerobic bacteria, which always preferentially carry out respiration as a metabolic pathway in the presence of oxygen, facultative anaerobic yeast can also adopt a fermentative metabolism in the presence of oxygen. This is the case for *Saccharomyces cerevisae*, the most representative yeast used in a wide variety of industrial fermentative processes. *S. cerevisiae* exhibits a specific mechanism of respiro-fermentative regulation that is known as the 'Crabtree effect'. At high sugar concentrations, and even in the presence of oxygen, fermentation overrides respiration. This mechanism results in a high rate of sugar consumption and therefore the colonization of habitats with high sugar content. Consequently, the ecological niche of fermentative yeasts includes sugary fruits, flowers, lymph, and tree exudates.

Anaerobic facultative bacteria represent a group of microorganisms that is broadly diffuse in several habitats. These microorganisms can grow in the presence of oxygen, and they can use fermentative metabolism when oxygen is not available. Anaerobic facultative bacteria such as *Enterobactericeae* colonize the intestinal tract of warmblooded animals, and some species belonging to the genera *Yersinia*, *Salmonella*, and *Shigella* can cause infectious diseases. The habitat of enteric bacteria is very specific. *Escherichia coli*, the most well-known species among the enteric bacteria, is used as indicator of fecal contamination of water and food because of its low survival in other environments. Since enteric bacteria are anaerobic and facultative, they have important roles in the ecology of the gut of warmblooded animals. By consuming the oxygen available in this habitat they maintain the right conditions for the proliferation of other bacteria that constitute the intestinal microflora (lactic acid bacteria, bifidobacteria).

Other facultative or obligate anaerobic bacteria can be found in both soil and water environments, where they have fundamental roles in the transformation of organic substances under anoxygenic conditions. An example is provided by methanogenesis, a bioprocess that occurs in natural environments, such as the rumen, marshes, and the industrial sites of anaerobic wastewater treatment plants. During this process, a first group of fermentative/hydrolytic microorganisms produces low molecular weight organic acids, alcohols, CO_2, and H_2. Another bacterial group, known as the 'syntrophic microorganisms' is very important for the conversion of organic compounds to CH_4 during methanogenesis. In this process, called syntrophy, methanogenic bacteria cooperate with other fermenting microorganisms to produce the substrates that are necessary to realize their specific ecological interactions. The H_2 derived from organic molecules and produced after energetically unfavorable fermentation is consumed by methanogenic bacteria, and the overall reaction produces energy that is used by the syntrophic couple.

Industrial Fermentation

As indicated above, natural fermentation is widely diffuse in several ecological niches where the conditions of anaerobiosis, high concentrations of carbohydrates (the Crabtree effect), or the lack of carbohydrates (fermentation of amino acids) determine the predominant fermenting organisms. All fermentative processes that are today devoted to food and animal feed transformation and preservation have ancient origins and have been traditionally carried out by the microorganisms naturally present in the substrates. The advent of industrialization, the construction of appropriate equipment, and the development of microbiology as an applied science have led the development of the fermentation industry, transforming a great number of the natural fermentation processes into industrial-scale fermentation. These transformations have been applied to wine, beer, distilled beverages and bread industries where alcoholic fermentation is mainly involved, and to the dairy and meat transformation industries, in which lactic fermentation is the main fermentation process. In these fermentation industries, the use of selected starter cultures during different stages of the fermentation processes has become progressively diffuse. The aim of this procedure is to improve the management of the fermentation process through avoidance of the development of spoilage and pathogenic microorganisms, and enhancement of the quality of the final product. Table 2 lists these main fermentation processes that are involved in the food and animal-feed industries.

Alcoholic Fermentation

Alcoholic fermentation is the best known of the fermentation processes, and is involved in several important transformation, stabilization, and conservation processes for sugar-rich substrates, such as fruit, and fruit and vegetable juices. Alcoholic fermentation is carried out by yeasts and some other fungi and bacteria. The first step of the alcoholic fermentation pathway involves pyruvate, which is formed by yeast via the EMP pathway, while it is obtained through the ED pathway in the case of *Zymomonas* (bacteria). In the following step, the pyruvate is decarboxylated to acetaldehyde in a reaction that is catalyzed by the enzyme pyruvate decarboxylase (Figure 2).

The redox balance of alcoholic fermentation is achieved by the regeneration of NAD^+ during the reduction of acetaldehyde to ethanol, which is catalyzed by alcohol deydrogenase. The ATP yield of alcoholic fermentation is 1 or 2 mol of ATP per mole of glucose oxidized via the ED and EMP pathways, respectively. *Zymomonas mobilis* is the most important bacterial species that is able to perform alcoholic fermentation. The habitat of this species is the lymph of tropical trees, such as the palma tree from where it was originally isolated. *Z. mobilis* was proposed and used as a starter for ethanol production at the industrial level, although at present alcoholic fermentation carried out by yeast is better known and has been more thoroughly investigated. Natural alcoholic fermentation of fruit (cacao) and fruit juices (grape must and apple juice) is carried out by different microorganisms that act

Table 2 Examples of industrial fermentation and their producer microorganisms

Fermented food and animal feed	*Microorganisms involved*	*Fermentation*	*Undesired fermentation*
Fermented meat (salami)	Lactic acid bacteria	Lactic	
Fermented milk	Lactic acid bacteria	Lactic	Alcoholic
Fermented milk	Lactic acid bacteria and yeast	Lactic and alcoholic	
Cheese	Lactic acid bacteria; propionic bacteria	Lactic and sometimes propionic	Mixed-acid and butyric acid
Bread and bakery products	Yeast and lactic acid bacteria	Alcoholic and heterolactic	
Beer	Yeast	Alcoholic	Lactic and mixed acid
Wine	Yeast	Alcoholic	Lactic
Olive brines	Lactic acid bacteria	Lactic	Butyric
Silage	Lactic acid bacteria	Lactic	Amino acids Butyric Mixed-acid
Coffee	Lactic acid bacteria and yeast	Alcoholic and lactic	
Cocoa	Yeast and lactic acid bacteria	Alcoholic and lactic	

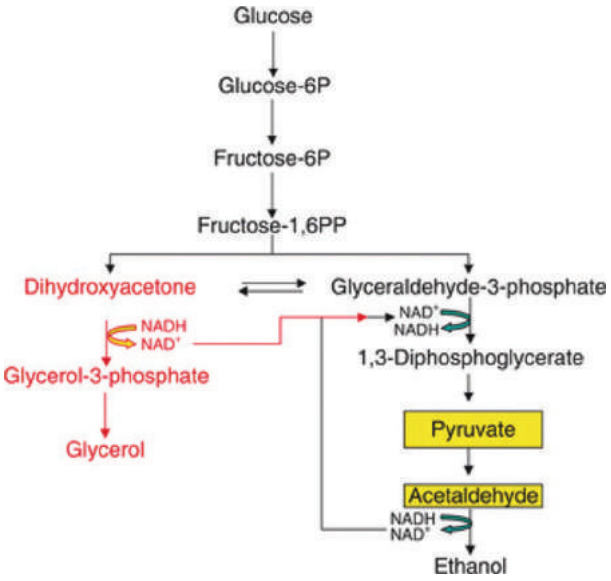


Figure 2 Alcoholic fermentation and deviation from alcoholic to glycero-pyruvic fermentation. Black line: alcoholic pathway; red line: Glycero-pyruvic pathway; yellow box: intermediate metabolites involved in alcoholic fermentation block.

sequentially. The substrate fermentation is first achieved by apiculate yeast (*Hanseniaspora*), which is followed by elliptical yeast (*Saccharomyces*). Other fermenting yeast ascribed to the genera *Candida*, *Kluyveromyces*, *Metschnikowia*, *Pichia*, *Saccharomycodes*, *Torulaspora*, and *Zygosaccharomyces* are also sometimes found during natural alcoholic fermentation.

Controlled Mix Fermentation in Wine

Grape juice fermentation is a complex biochemical process in which wine yeasts play fundamental roles during the transformation of grape sugars into ethanol, carbon dioxide and hundreds of other secondary products. To improve the quality of wine, the inclusion of non-*Saccharomyces* wine yeasts, together with *Saccharomyces cerevisiae* strains as part of mixed and multistarter fermentations, has been proposed as a tool to take advantage of spontaneous fermentation avoiding the risks connected with this last one. Criteria for the selection and development of non-*Saccharomyces* yeasts for wine fermentation have evolved over many years and are largely discussed. On the basis of these criteria yeasts should harbor characters that influence wine quality and affect the performance of the fermentation process, Some characters and properties are associated with the commercial production of these wine yeasts. Within each category, there are properties of varying degrees of significance and importance, some being essential and some being desirable. Possible synergistic and positive interactions between different yeasts might represent a tool for new fermentation technologies. Indeed, a lot of studies demonstrated that in *Saccharomyces*/non-*Saccharomyces* mixed cultures, interactions due to the wide inter-generic metabolic diversity could contribute to improve the final quality of wine.

It is well established that the most important agent of alcoholic fermentation is *S. cerevisiae*, the yeast that is used widely in several fermentation industries (wine, beer, cider, and bread) as a microbial starter. *S. cerevisiae* becomes the dominant species during alcoholic fermentation of fruit and fruit juices because of the strongly selective environment due to the low pH and high sugar and ethanol concentrations, and the anaerobic conditions. The ecological distribution of *S. cerevisiae* and the role of different habitats in the evolution of this species are controversial. Numerous investigations have revealed the low diffusion of this yeast species in natural environments such as soil, fruit surfaces, and tree exudates. On the other hand, *S. cerevisiae* is found widely in wineries and other fermentation plants since it is used to carry out the fermentation processes.

In the winery, several environmental factors, including high ethanol and sugar concentrations, the presence of SO_2, and others, can exert selective pressures on the *S. cerevisiae* population. Following these considerations, *S. cerevisiae* is defined as a domesticated species because of its selection through time in man-made environments.

Glycero-Pyruvic Fermentation

Glycero-pyruvic fermentation is always concomitant to alcoholic fermentation, although it involves a very low percentage of sugar (5–8%). However, under particular fermentation conditions, some osmotolerant yeast species (e.g., *Torulopsis magnoliae*, *Torulopsis bombicola*, and *Candida stellata*) and other fermenting yeast (*S. cerevisiae*) can ferment sugar to produce glycerol and, for example, acetaldehyde, acetic acid, acetoin, 2,3-butanediol and succinic acid, all compounds that can be derived from pyruvate.

The change from alcoholic to glycero-pyruvic fermentation occurs mainly because of the need to regenerate NAD^+ when the reduction of acetaldehyde to ethanol is not possible (Figure 2). This could be due to: (1) the nonavailability of acetaldehyde, if it is

bound by sulfite; (2) the absence or low activity of pyruvate decarboxylase; and (3) the high activity of the aldehyde dehydrogenase enzyme (under alkaline pH) that catalyzes the reaction from acetaldehyde to acetate.

In the past, between the two world wars, glycero-pyruvic fermentation was exploited at an industrial level for the production of glycerol.

Propionic Acid Fermentation

Propionic acid fermentation is carried out by several bacteria that belong to the genus *Propionibacterium* and to the species *Clostridium propionicum*. During propionic acid fermentation, both sugar and lactate can be used as the initial substrate. When sugar is available, these bacteria use the EMP pathway to produce pyruvate; the pyruvate is carboxylated to oxalacetate by methyl malonyl coenzyme-A (CoA) and then reduced to propionate via malate, fumarate, and succinate. The other end products of propionic fermentation are acetic acid and CO_2 (**Figure 3**). In particular, the propionic acid fermentation of 3 mol of glucose produces 4 mol of propionic acid, 2 mol of acetic acid, 2 mol of CO_2, and 12 mol of ATP. When lactate is the initial substrate, propionic fermentation results in the production of 2 mol of propionic acid, 1 mol of acetic acid, and 1 mol of CO_2. In this process, 1 mol of ATP is generated per nine carbons, and because of this, propionic bacteria generally grow very slowly.

The typical natural habitats of *Propionibacterium* are the rumen, the intestinal tract of animals and the skin of mammalians. *Propionibacterium* also colonize cheese during maturation. While the metabolic activity of *Propionibacterium* should be avoided during the maturation of the vast majority of cheese, it is required for the production of some typical products, such as Emmental cheese. In this Swiss-type cheese, two successive fermentations occur. During its manufacture, lactic acid bacteria convert lactose into lactate, and then during ripening, propionic acid bacteria convert the lactate into propionic acid, acetic acid, and CO_2. The CO_2 is responsible for 'eye' formation and the propionic acid promotes the typical nutty flavor of this Swiss-type cheese. The ability to use lactate is particularly relevant for the ecological distribution of propionic acid fermenting bacteria, which can use the final product of lactic acid fermentation.

Amino Acid Fermentation

In the absence of electron acceptors such as oxygen, nitrate, and sulfate, *Clostridium*, *Fusobaterium*, and a few other anaerobes can ferment amino acids. This fermentation occurs during anaerobic and putrefaction processes. The most important mechanism for amino acid degradation is Stickland fermentation; during Stickland fermentation of two amino acids, one serves as the electron donor while the other serves as the acceptor. All amino acids are classified into electron acceptors or donors on the basis of Stickland fermentation, and only tryptophan and tyrosine can behave both as an acceptor and a donor. In addition to decarboxylation of various amino acids by this mechanism, the subsequent reactions yield a variety of products that can have unpleasant odors. In the absence of fermentable carbohydrates and in rich protein substrates, a large number of *Clostridium* species – such as *C. botulinum*, *C. tetani*, and *C. perfringens* – can generate ATP from amino acid fermentation. The ATP yield here is 1 mol per 3 mol of amino acid used, and thus the reaction is highly advantageous for organisms that can grow in rich anaerobic protein environments. The main products of the Stickland reaction are NH_3, CO_2, short-chain fatty acids and H_2, and the minor products are hydrogen sulfide, methyl mercaptane, phenols, and alcohols, which together with fatty acids form a typical putrefying odor. The favorite habitat of *Clostridium* is the soil, in anaerobic scrap previously colonized by aerobic bacteria, and *Clostridium* species can also colonize the mammalian intestine. Moreover, these species can produce infectious diseases, such as botulism caused by

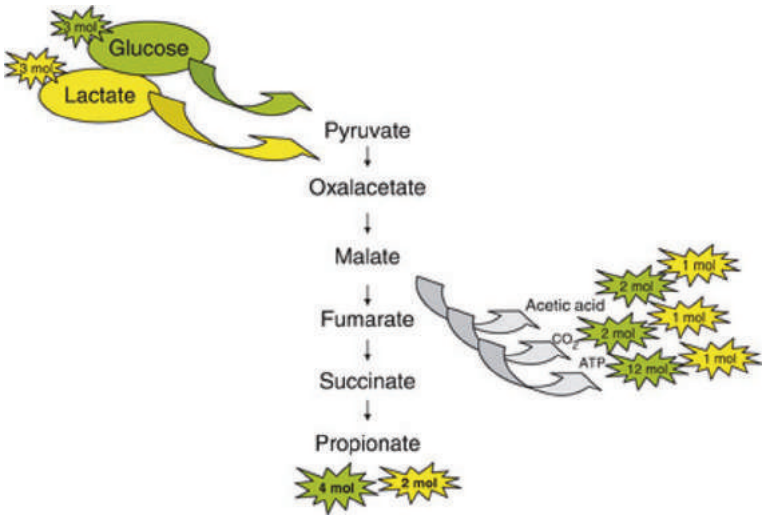


Figure 3 Propionic fermentation and different yields of the final products coming from glucose or lactate as the initial substrate.

C. botulinum, tetanus by *C. tetani*, and gangrene by *C. perfringens*. *Clostridium sporogenes*, a typical soil bacteria, can also participate in peritoneal infection that can occur as a result of the proliferation of food pathogens, intestinal obstruction, or mesenteric thrombosis.

Lactic Acid Fermentation

Depending on the pathway used for glucose oxidation, lactic acid fermentation can result in the production of lactate (homolactic fermentation), lactate, ethanol/acetate, and CO_2 (heterolactic fermentation) or lactate and acetate (the bifid shunt). Lactic acid fermentation is carried out by lactic acid bacteria and bifidobacteria, and also by some species of *Bacillus*, some protozoa and water molds, and the cells of human skeletal muscle when they are subjected to extreme work under oxygen deprivation. Both homolactic and heterolactic fermentation are involved in food and animal feed transformation and preservation. In particular, lactic acid fermentation is mainly responsible for the souring of milk products and is used in the production of yogurt and other fermented milk products (e.g., cheese, buttermilk, and sour cream). However, lactic acid fermentation also occurs during the fermentation of sauerkraut, and in other vegetable and sourdough bread fermentation, and it has important roles in sausage maturation, and silage fermentation and stabilization. The bifid shunt occurs mainly in the human and animal large intestine, where bifidobacteria are among the most abundant of the microbial groups.

Homolactic Fermentation

Homolactic fermentation is carried out by bacteria belonging to the genera *Lactococcus*, *Enterococcus*, *Streptococcus*, and *Pediococcus*, and by some species of the genus *Lactobacillus*. All of these bacteria can convert sugar to mainly lactic acid, via glycolysis. The enzyme lactate dehydrogenase (LDH) catalyzes the last step of this fermentation; in particular, by transferring hydrogen from NADH to pyruvate, LDH leads to pyruvate reduction and NADH reoxidation, with the production of D- or L-lactate. ATP formation is coupled to pyruvate production and the ATP yield of homolactic fermentation is 2 mol per mole of glucose oxidized (Figure 4). The homolactic behavior is not obligatory, but depends on sugar type, rate of the glycolytic flux, and growth conditions. Some homofermentative bacteria can catabolize glucose in a heterofermentative fashion, or carry out mixed-acid fermentation when the glycolytic flux is low, thus decreasing or increasing the ATP yield per mole of glucose, respectively.

Heterolactic Fermentation

Heterolactic fermentation is carried out mainly by bacteria of the genera *Leuconostoc*, *Oenococcus*, and *Weissella*, and by heterofermentative lactobacilli. Obligate heterofermentative bacteria do not perform glycolysis due to the lack of aldolase, the enzyme that breaks down fructose 1,6-bisphosphate into glyceraldehyde 3-phosphate and dihydroxyacetone phosphate. Glucose 6-phosphate is oxidized to 6-phosphogluconate and fermentation is carried out through the phosphoketolase pathway (Figure 4). The final products of this fermentation are lactate, ethanol, and CO_2 but acetate may also be produced. The ATP

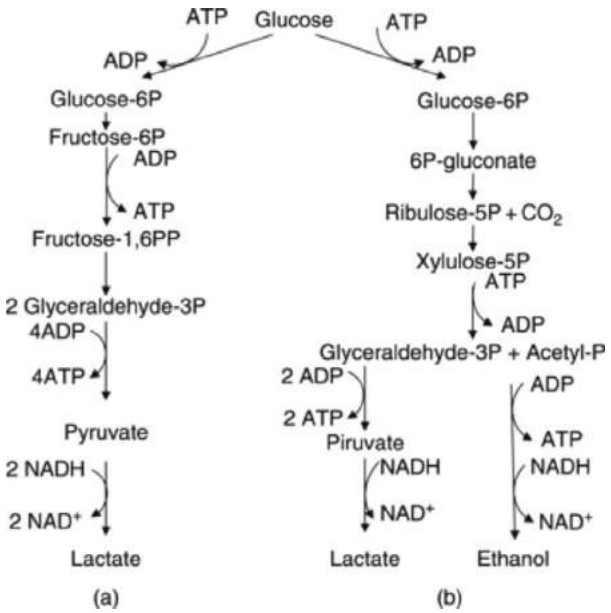


Figure 4 Schematic representation of homolactic (a) and heterolactic (b) fermentations. Pyruvate reduction leads to the regeneration of NAD^+.

yield is 1 mol per mole of glucose; thus, heterolactic metabolism yields less energy than homolactic fermentation. Heterolactic fermentation can also be carried out by facultative homofermentative bacteria.

Bifidobacterium Lactic Acid Fermentation

This type of lactic acid fermentation is exclusive to Gram-positive bacteria belonging to the genus *Bifidobacterium*. These bacteria are found mainly in the intestinal tracts of warmblooded animals and they are recognized as probiotic. In fact, they help in the maintenance of the balanced composition of the intestinal microflora and they exert positive effects on the health and well-being of the host. Bifidobacteria do not have aldolase and glucose-6-phosphate dehydrogenase; they thus ferment hexose via a phosphoketolase pathway that is known as the 'bifid shunt', where the final products are acetic and lactic acids in a molar ratio of 3:2. The key enzyme in the bifid shunt is fructose-6-phosphate phosphoketolase (F6PPK), which converts frutose 6-phosphate into acetyl 1-phosphate and eritrose 4-phosphate (**Figure 5**). The ATP yield is 5 mol per 2 mol of glucose, and it is therefore higher than that of homolactic fermentation.

Fermented Vegetables and Fruits Through Lactic Acid Bacteria

Lactic acid fermentation represents the most suitable approach for increasing the daily consumption of fresh-like vegetables and fruits. Lactic acid bacteria are a small part of the autochthonous microbiota of vegetables and fruits. The diversity of the microbiota markedly depends on the intrinsic parameters such as the pH and aw values of matrix, temperature, and extrinsic parameters of the plant matrix, first of all the interactions between different microorganisms that coexist in the environment.

Besides the above semi-industrial or artisanal products, a large variety of fermented pickles, which represents a culture heritage from the tradition of rural communities, is scarcely documented. Such fermented pickles are very popular in Asian and African countries, where they are fundamental components of the daily diet Under these poor technology conditions, lactic acid fermentation is the simplest, and probably the only, way to preserve fruits and vegetables.

In these natural matrix, spontaneous fermentation was shown to improve some health-promoting features of pickled garlic, especially the concentration of polyphenols and the related antioxidant activity.

Kimchi is the name of various traditional fermented vegetables, which are emblematic of the Korean culture. These beneficial effects are attributed either to functional components (vitamins, minerals, fibre and phytochemicals) or to fermentation by lactic acid bacteria Usually, *Leuconostoc mesenteroides* starts the fermentation but it is suddenly inhibited by the increasing concentration of lactic acid. Acid-tolerant species such as *Lactobacillus brevis* dominate during the middle stage, being replaced by *Lactobacillus plantarum* during late fermentation.

Another example of fermented vegetable regards cucumbers that, fully ripe, is harvested and dipped into brine (with 5–7% of NaCl) inside plastic or glass boxes and subject to spontaneous lactic acid fermentation. During fermentation, lactic acid bacteria synthesize several bacteriocins and liberates antimicrobial peptides, which inhibit spoilage bacteria.

Also fermented capers are typical pickles of Mediterranean countries (e.g., Greece and sud of Italy). Fruits are harvested in June or July, immersed in tap water, and subjected to spontaneous lactic acid fermentation for about 5–7 days at ambient temperature. After that, fermented capers are placed into brine. *Lactobacillus plantarum* is the species more frequently isolated from the brine of capers.

Innovation in food biotechnology plays an important role to improve the nutritional features, possibly also enhancing the hedonistic aspects. Smoothies are an example of this trend to increase the consumption of vegetables and fruits, as an alternative and/or a complement to fresh products. For example, white grape juice and Aloe vera extract are mixed with red (cherries, blackberries prunes and tomatoes) or green (fennels, spinach, papaya and kiwi) fruits and vegetables, and subjected to fermentation with mixed autochthonous starters such as *Lactobacillis plantarum* and *Lactobacillus pentosus* strains.

Functional beverages made with a mixture of rice and barley or emmer and concentrated red grape must are fermented with selected strains of *L. plantarum*, which remain viable throughout storage at 4 °C for 30 days. The use fruit and vegetable matrices as potential non dairy vehicles for delivering probiotic strains is largely assessed. Probiotics have the capacity to grow and to survive, depending on the inherent characteristics of the plant species. Lactobacillus acidophilus, *L. plantarum*, *Lactobacillus casei*, *Lactobacillus rhamnosus*, *Lactobacillus delbrueckii*, *Leuconostoc mesenteroides*, and species of the genus *Bifidobacterium* are mainly used. Compared to cranberry juice, *Lactobacillus* and *Bifidobacterium* strains survive for a longer time in orange and pineapple juices *L. casei*, *L. rhamnosus*, and *Lactobacillus paracasei* remaine viable in orange juice during 12 weeks of storage.

Often lactic acid fermentation occurs spontaneously following protocols of manufacture, which are strongly linked to the culture and traditions of each country. New molecular approaches to study the composition of the microbiota and to select autochthonous starters targeted for different vegetables and fruits have to be encouraged to get new insights and to allow controlled fermentation processes as it was done for other fermented foods (cheeses, sausages, leavened baked goods). Lactic acid bacteria tailored to the various intrinsic and extrinsic environmental conditions completely exploit the potential of vegetables and fruits, which enhances the hygiene, sensory, nutritional and shelf life properties (**Figure 6**).

Mixed-Acid Fermentation

Mixed-acid fermentation is characteristic of the *Enterobacteriaceae* ascribed to the genera *Citrobacter*, *Escherichia*, *Proteus*, *Salmonella*, *Shigella*, *Yersinia*, and *Vibrio*, and to some species of *Aeromonas*; it is also carried out by some anaerobic fungi. This metabolic group includes microorganisms that have a different ecology and impact on human activities. Some of them are members of the normal intestinal microflora of mammals and other vertebrates (*E. coli*) and have a role in the colonization of lignocellulose in the rumen (*Neocallimastix*). Some others are pathogens that are responsible for human and animal diseases, and they can be abundant in

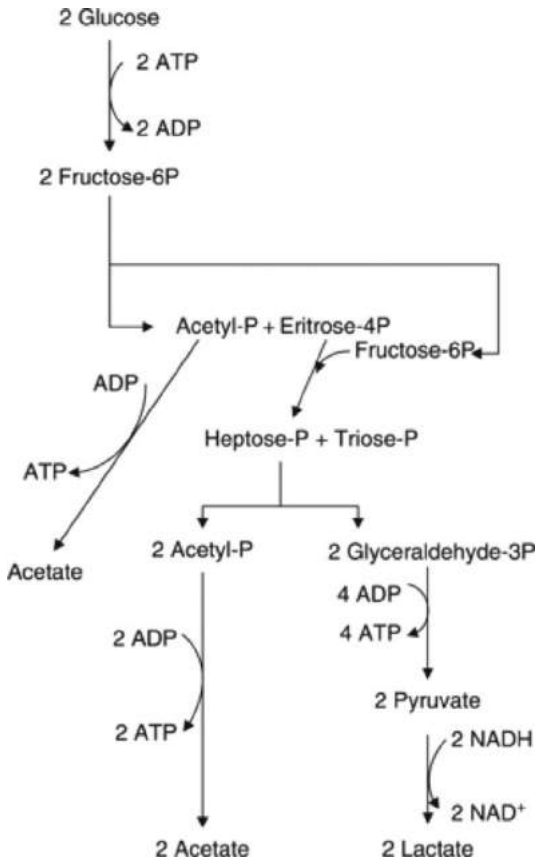


Figure 5 Schematic representation of bifid shunt.

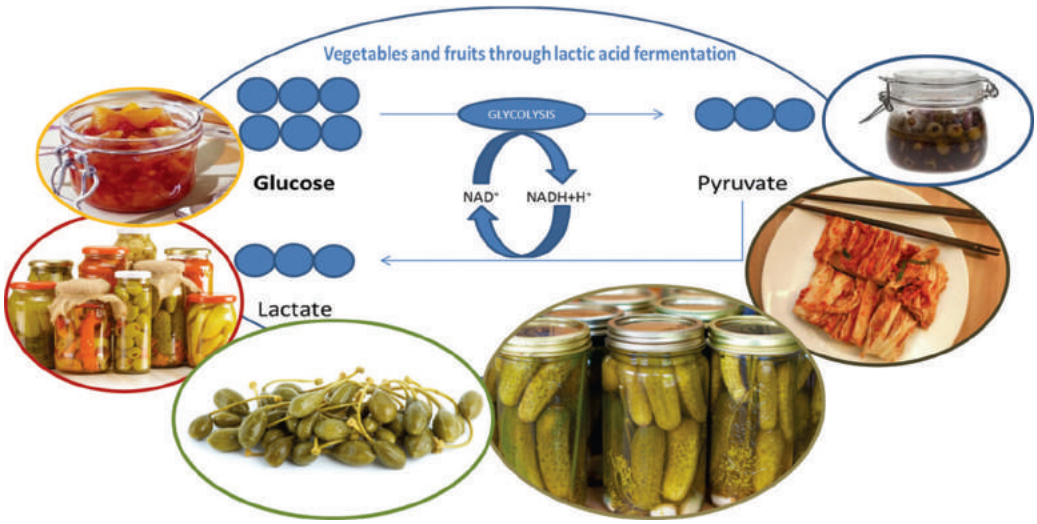


Figure 6 Fermented food through lactic acid fermentation.

aquatic and terrestrial environments. These microorganisms can ferment monosaccharides, disaccharides, polyalcohol, and, less frequently, polysaccharides, via the glycolytic pathway, producing lactic, formic, succinic and acetic acids, and ethanol (Figure 7). The final amounts of each end product vary depending on the microorganism and the growth conditions; however, the ratio of acid to neutral products is 4:1. Mixed-acid fermentation also results in the production of equimolar amounts of CO_2 and H_2 in those bacteria with the formate–hydrogen–lyase complex. Pyruvate formate–lyase and LDH, the enzymes that control entry into mixed-

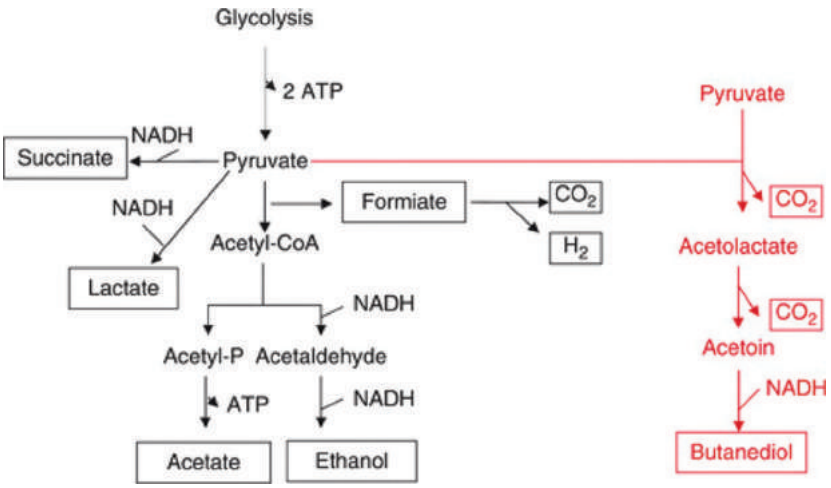


Figure 7 End products of mixed-acid and butanediol fermentation.

acid fermentation, are negatively regulated by oxygen; thus, mixed-acid fermentation requires anaerobic conditions to occur. That is why the natural habitat of microorganism carrying out mixed-acid fermentation is the gastro-enteric apparatus. The ATP yield of mixed-acid fermentation is about 2.5 mol of ATP per mole of glucose.

Butanediol Fermentation

Butanediol fermentation is carried out by members of the genera *Enterobacter*, *Erwinia*, *Hafnia*, *Klebsiella*, and *Serratia*. Most of these bacteria can be found in soil and water (*Enterobacter* and *Serratia*) and they can be plant pathogens (*Erwinia*). Butanediol fermentation produces less acids than mixed-acid fermentation as two molecules of pyruvate are used to produce one molecule of 2,3-butanediol and are therefore not available for the production of acid compounds. Thus, the ratio of acidic to neutral products is 1:6. Moreover, the reactions that lead to the production of 2,3-butanediol involve a double decarboxylation step; thus, butanediol fermentation produces more gas than mixed-acid fermentation, with a ratio of carbon dioxide to hydrogen of 5:1.

Butyric Acid Fermentation

Butyric acid fermentation is characteristic of several obligate anaerobic bacteria that mainly belong to the genus *Clostridium*; by means of glycolysis, these are able to oxidize sugar, and occasionally amylose and pectin, to pyruvate. Pyruvate is in turn oxidized to acetylCoA by the pyruvate–ferredoxin oxidoreductase enzyme system, with the production of CO_2 and H_2. Part of the acetylCoA is converted into acetic acid, with ATP production. The other part generates acetoacetylCoA, which is reduced to butyrylCoA through the production of β-oxybutyrylCoA and crotonylCoA. The transformation of butyrylCoA into butyrate leads to further ATP production. Thus, this fermentation process produces a relatively high yield of energy, with 3 mol of ATP for each mole of glucose. Small amounts of ethanol and isopropanol can also be produced (Figure 8). Butyric fermentation is quite common in silage when the pH is not low enough to ensure the exclusive activity of lactic acid bacteria. The carbon dioxide produced during butyric fermentation also causes an increase in the pH of the silage, thus enhancing further butyric fermentation. Some bacteria, such as *Clostridium acetobutylicum*, produce less acids and more neutral products, thus carrying out acetone butanol fermentation. This fermentation had great importance during World War I due to the need for acetone for the production of munitions.

The main fermentation processes are summarized in Table 3 in terms of energy yield.

Other Fermentation Pathways

The names of some other fermentation pathways derive from the names of the final products. This is the case for caproic, homoacetic, and methanogenic fermentation. For caproic fermentation, *Clostridium kluyveri* is the species that can metabolize acetic acid and ethanol under anaerobic conditions, producing butyric and caproic acids, in a specific controlled mechanism: if acetic acid is present in excess, a considerable amount of butyric acid is formed, while if ethanol is in excess, caproic acid is the main product. These relationships suggest that butyric acid is an intermediate in the synthesis of caproic acid from acetic acid.

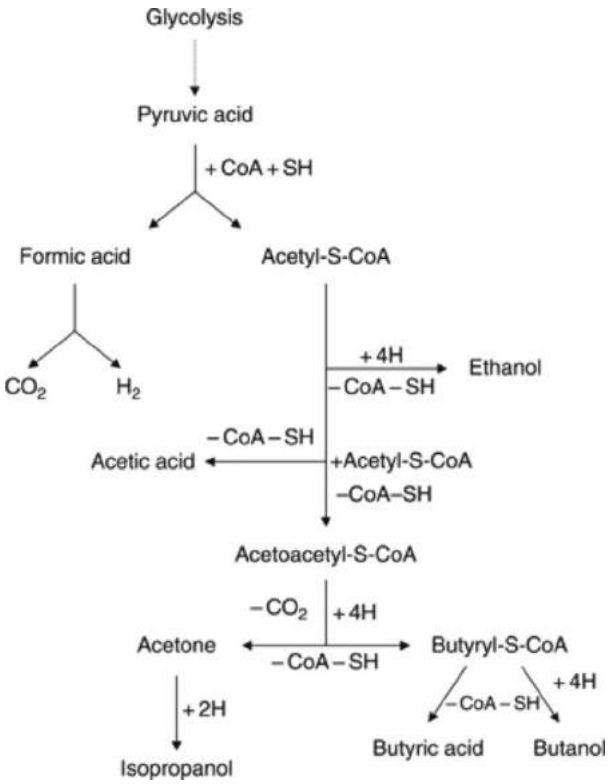


Figure 8 Different metabolic pathways of butyric acid fermentation.

Table 3 Summary of fermentation processes with the corresponding energy yield

Fermentation process	*Energy yield*
Alcoholic fermentation	2 mol ATP/mol glucose
Glycero-pyruvic fermentation	Net ATP production
Propionic acid fermentation (glucose[a])	4 mol ATP/mol glucose
Propionic acid fermentation (lactate[a])	0.3 mol ATP/mol lactate
Amino acid fermentation	0.3 mol ATP/mol amino acid
Lactic acid fermentation (homolactic)	2 mol ATP/mol glucose
Lactic acid fermentation (heterolactic)	1 mol ATP/mol glucose
Bifidobacterium lactic acid fermentation	2.5 mol ATP/mol glucose
Mixed acid fermentation	2.5 mol ATP/mol glucose
Butanediol fermentation	2.5 mol ATP/mol glucose
Butyric fermentation	3 mol ATP/mol glucose

[a]Carbon source.

In homoacetic fermentation, the *Acetobacterium* group converts fructose into acetate, and it appears that the nutritional requirements of this organism are complex. During methanogenesis, methane can be formed via methanogenic fermentation by *Methanosaeta* and *Methanosarcina*, which convert acetate and water into CH_4 and carbonic acid.

In addition, there are some relatively rare fermentation pathways that are carried out by very restricted anaerobic microorganisms. In Table 4, some examples of the fermented substrates, the microorganisms and the biochemistry of the reactions are summarized. Generally, these fermentation pathways are carried out by specialized bacteria that use substrates that cannot be metabolized by other microbiological groups. Nevertheless, during the catabolic process, all of these rare fermentation pathways produce intermediate compounds that have high residual energies and are commonly CoA derived, from which these microorganisms obtain their ATP.

Table 4 Microorganisms and type of substrates metabolized in rare fermentation processes

Substrate	*Microorganism*	*Reaction*
Acetate	*Methanosaeta*	$Acetate + H_2O \rightarrow CH_4 + HCO_3^-$
Ethanol and acetate	*Clostridium kluyveri*	$Ethanol + Acetate + CO_2 \rightarrow Caproate + Butyrate + H_2$
Fructose	*Acetobacterium* spp.	$Fructose \rightarrow Acetate + H_2CO_2 \rightarrow Acetate + H_2O$
Hydroxyhydroquinone	*Pelobacter massiliensis*, *Pelobacter acidigallici*	$C_6H_6O_3 + H_2O \rightarrow Acetate + H^+$
Malonate	*Malonomonas rubra*	$Malonate + H_2O \rightarrow Acetate + HCO_3^-$
Oxalate	*Oxalobacter formingenes*	$Oxalate + H_2O \rightarrow Formaite + HCO_3^-$
Putrescine	Gram + bacteria	$C_4H_{12}N_2 + H_2O \rightarrow Acetate + Butyrate + NH_4^+ + H_2 + H^+$

The first example includes anaerobic microorganisms that can degrade xenobiotic industrial chemical compounds through a combination of co-metabolic steps, which often yield partial degradation, or by serving as growth substrates that are accompanied by mineralization of at least part of the molecule. Indeed, while aerobic microorganisms use oxidative reactions, degradation by anaerobic bacteria takes place by reduction reactions, and they thus degrade aromatic compounds by reductive conversions with the central intermediates that are ready for hydrolytic ring cleavage having a 1,3-dioxo structure. Using an example, including aromatic, chloroaromatic, aliphatic, and chloroaliphatic compounds, one case of anaerobic degradation is presented. The recently isolated fermenting bacterium *Pelobacter massiliensis* is the only strict anaerobe that is known to grow on hydroxyhydroquinone (1,2,4-trihydroxybenzene) as the sole source of carbon and energy, converting it to stoichiometric amounts of acetate. Another example is seen in *Malonomonas rubra*, which is a microaerotolerant fermenting bacterium that can maintain its energy metabolism for growth by decarboxylation of malonate to acetate. *M. rubra* is closely related to the cluster of mesophilic sulfur-reducing bacteria within the delta subclass of *Proteobacteria*, with the fermenting bacterium *Pelobacter acidigallici* and the sulfur reducers *Desulfuromusa kysingii*, *D. bakii*, and *D. succinoxidans* as its closest relatives.

Ecological Niches and Interactions Among Fermenting Microorganisms

Fermentative processes are carried out by different microbial groups that can interact in well-defined habitats for the transformation of the substrate. The balance between the different microbial groups in each specific ecosystem depends on their initial concentrations and on environmental conditions (e.g., temperature, O_2 concentration, pH, and nutrients), and the predominance of one or the other determines the type of fermentation and the characteristics of the final product. Some fermented milks (Kefir, Koumis) and sourdoughs are ecological niches colonized by different fermenting microorganisms, such as lactic acid bacteria and yeasts, which carry out contemporary lactic acid and alcoholic fermentation. Lactic acid and propionic bacteria intervene sequentially during 'eyed' cheese production. Lactic acid bacteria ferment lactose and produce lactic acid during the first stages of cheese making. Lactic acid is then used by propionic bacteria during cheese ripening. Another example of an ecosystem is the intestinal tract of warmblooded animals, where enteric bacteria, lactic acid bacteria, bifidobacteria, and clostridia coexist, all of which are involved in the transformation of the substrate via different fermentation pathways. In this ecological niche, lactic acid and bifidobacteria are involved in the maintenance of the balanced composition of the intestinal microflora. Different fermenting microorganisms, such as lactic acid bacteria and clostridia, are involved in silage fermentation and maturation. Also in this case, lactic acid bacteria should predominate to counteract the development of clostridia and other spoilage microorganisms (**Figure 9**).

Fermented Milks: Kefir and Kumis as Matrix of Microbial Interactions

An example of microbial cooperation is represented by kefir, a typical fermented beverage of Eastern Europe. Kefir is the efficacious consequence of the metabolic activity of several lactic acid bacteria, mainly ascribed to the genus *Lactobacillus*, and of numerous yeast species belonging to the genera *Kluyveromyces*, *Candida*, *Saccharomyces*. This microbial consortium is enclosed in a matrix of proteins and polysaccharides called 'kefiran', which is produced during growth under aerobic conditions. The beverage flavor is due to a mixture of fermentation products such as lactic acid, ethanol, carbon dioxide and acetaldehyde.

Kumis, also spelled kumiss or koumiss is a fermented dairy product traditionally made from mare's milk. The drink is important for the peoples of the Central Asian steppes, of Huno-Bulgar, Turkish and Mongol origin: Bashkirs, Kalmyks, Kazakhs, Kyrgyz, Mongols, Uyghurs, and Yakuts. Kumis is a dairy product similar to kefir, but is produced from a liquid starter culture, characterized by *Lactobacillus leichmannii*, *L. delbrueckii* subsp. bulgaricus and yeasts belonging to *Torulaspora* genera, differently from the solid kefir "grains". Because mare's milk contains more sugars than cow's or goat's milk, when fermented, kumis has a higher alcohol content compared to kefir. Even in the areas of the world where kumis is popular today, mare's milk remains a very limited source. During the fermentation process, lactobacilli acidify the milk, and yeasts turn it into a carbonated and mildly alcoholic drink. In modern controlled production, the initial fermentation takes two- five hours at a temperature of about 27 °C; this may be followed by a cooler aging period. The finished product contains between 0.7 and 2.5% alcohol a very low level of alcohol, comparable to small beer, the common drink of medieval Europe that also avoided the consumption of potentially contaminated water (**Table 5**).

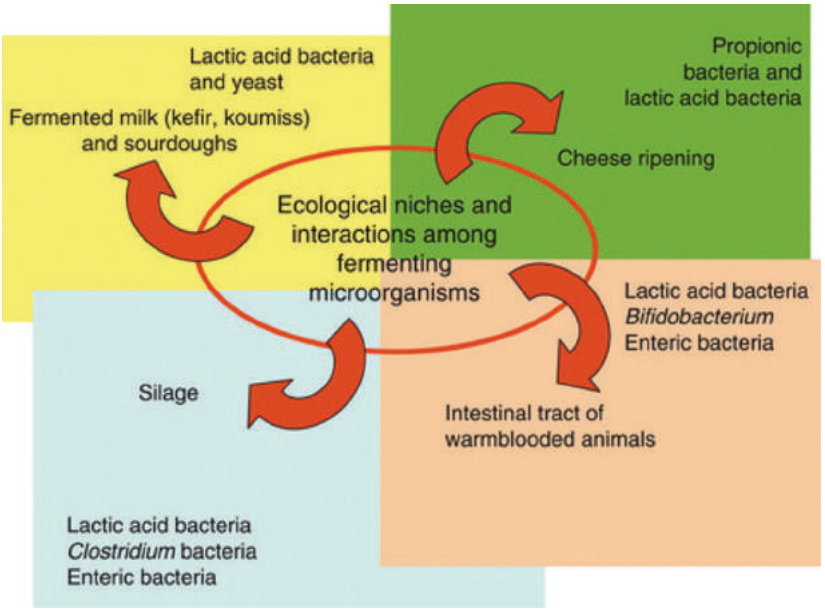


Figure 9 Examples of ecological niches where different fermenting microorganisms coexist.

Table 5 Some of the most popular products obtained by milk fermentation

Food and products	*Raw materials*	*Microorganisms*	*Production site*
Kefir	Cow milk	*Lactococcus lactis*, *L. delbrueckii* subsp. *bulgaricus*, *Kluyveromyces*, *Candida*, *Saccharomyces*	Southeast Asia
Kumis	Crude mare's milk	*Lactococcus leichmannii*, *L. delbrueckii* subsp. *bulgaricus*, *Torulaspora* spp	Russia
Taette	Cow milk	*Staphylococcus lactis* var. *taette*	Scandinavian
Tarhana	Wheat flour and yogurt	Lactic acid bacteria (spontaneous population)	Turkey
Biogurt	Powder milk	*Lactobacillus acidophilus*, *L. lactis*	Worldwide

Further Reading

Bengmark S and Martindale R (2005) Prebiotics and synbiotics in clinical medicine. *Nutrition in Clinical Practice* 20: 244–261.
Bernardeau M, Guguen M, and Vernoux JP (2006) Beneficial lactobacilli in food and feed: Long-term use, biodiversity and proposals for specific and realistic safety assessments. *FEMS Microbiology Review* 30: 487–513.
Ciani M (2002) *Biodiversity and biotechnology of wine yeasts*, 1st edn. Trivandrum: Research Signpost.
Madigan MT, Martinko JM, and Parker J (2003) *Brock biology of microorganisms*, 10th edn. Edinburgh: Pearson Education.
Perry JJ, Staley JT, and Lory S (2002) *Microbial life*, 1st edn. Sunderland, MA: Sinauer Associates.
Stanier RY, Ingraham JL, Wheelis ML, and Painter PR (1988) *The microbial world*, 5th edn. Englewood Cliffs, NJ: Prentice-Hall.
Waites MJ, Morgan NL, Rockey JS, and Higton G (2005) *Industrial microbiology: An introduction*, 3rd edn. Oxford: Blackwell Science.
Zverlov VV, Berezina O, Velikodvorskaya GA, and Schwarz WH (2006) Bacterial acetone and butanol production by industrial fermentation in the Soviet Union: Use of hydrolyzed agricultural waste for biorefinery. *Applied Microbiology and Biotechnology* 71: 587–597.

Grazing☆

Antony J Underwood, University of Sydney, Sydney, NSW, Australia

Grazing in Terrestrial Environments: Why Is the World Green?

Grazing is a widespread and important process, so it has been debated how it comes about that the majority of terrestrial habitats remain dominated by plants. Why have the grazers not removed most of the biomass and cover of their food?

Generally, plants are limited in production by availability of light, which has seasonal patterns in temperate regions and extreme limitation for long periods of the year near the poles. Production of plants is also often reduced or prevented by lack of sufficient water or, in some circumstances, essential nutrients. Plants are not generally limited by their grazers.

To understand why, in general, terrestrial grazers do not keep their food-plants at small abundances therefore requires understanding of what reduces or controls the activities of the grazers. Much discussion of this, as a large-scale, world-wide phenomenon, considers three potential controls. One of these is competition among the grazers. This seems quite unlikely, given that the original phenomenon being explained is the preponderance of plant foods, not their small amounts which would lead to shortages of food. Competition will only occur among herbivorous species when more than one species of grazer requires the same plant foods and the food-resources are in short supply. Competition then occurs among the grazers as they try to gain sufficient resources to breed (adults), to grow (juveniles) or simply to maintain their tissues (all sizes).

Alternatively, abundances or effectiveness of herbivores may be controlled by inclement weather. There is evidence for this for many populations of insects, which do not reach their potential carrying capacity (the size of a population, given a certain availability of resources of food and shelter) because mortality due to bad weather keeps reducing the numbers. Such cases are, however, not considered to be typical. Many types of grazers can deal with bad weather by alterations of activity (e.g., hibernation or estivation) or of reproduction (e.g., by producing offspring during the periods of favorable weather).

The third general mechanism is for populations of grazers to be controlled by their predators (plus pathogens and parasites). This requires that large numbers of herbivores cannot be maintained, even where there is appropriate weather and an abundance of food, because their numbers are culled by their natural enemies.

This led long ago to the conclusion that primary producers-the plants-are largely regulated by availability of and resulting competition for limited resources. Presumably bad weather is also important because many species of plants will not grow outside a specific range of temperature, light-intensity and humidity. Similarly, predators are often limited by availability of their herbivorous prey-suffering from competition when populations of prey are at small abundances. Both are examples of "bottom-up" control. Populations of plants and predators are generally supposed to be regulated by their resources.

In contrast, terrestrial grazing species are thought to be primarily limited by their consumers, which is an example of "top-down" control.

The extent to which such a generality is correct can only be ascertained by examining a large number of experimental studies of the process controlling populations of herbivores, which is beyond the scope of this overview. Nevertheless, it is clear that, in general, plants on land are abundant despite numerous varieties of grazers.

It is evident that when grazers are released from their predators, they can completely destroy their food-plants. Two sorts of examples demonstrate this. Accidental introductions of alien herbivorous species into a new biogeographical region are often followed by outbreaks of the herbivores in extremely large densities, voraciously consuming the plants they use as food. Quite often, the cause of an outbreak is that the natural enemies (predators, pathogens, parasites) of the introduced pest are not introduced with it. The herbivorous species is therefore not held in check and its numbers become excessive.

The other scenario is where human intervention reduces the numbers of insects (or other predators). Spraying forests or farms to remove nuisance herbivores (e.g., larval insects) sometimes also kills other insects which are predatory. Populations of herbivores that are normally killed by these predators then expand and become a new nuisance because they consume trees or agricultural produce. In both cases, the grazers, by being released from their enemies, seriously reduce their food-plants.

There are many exceptions to a general principle that grazers do not control their food-plants. A well-known example is the destruction of plants caused by plague locusts, which break out in massive populations and greatly reduce the amounts of plants in areas where they feed.

The most notable exception is, however, the case of grazing by planktonic animals (see below). In the plankton, grazing generally reduces populations of plants to very small abundances. Given that about 70% of the earth's surface is ocean, the pattern of populations of plants being controlled by grazing is, in fact, more widespread than that described above for terrestrial habitats.

☆*Change History:* March 2018. A.J. Underwood made minor changes to the following sections: Grazing in Terrestrial Environments: Why is the World Green; Modeling Grazing Processes and Defenses Against Herbivores.

This is an update of A.J. Underwood, Grazing, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 1765–1773.

The Paradox of the Plankton

In the open oceans, organisms that float about in the open water, either permanently or during their larval development, consist of micro-algae (the phytoplankton) and zooplankton. In many areas of the world, there is essentially a paradox in that there are very large numbers and standing amounts of biomass of zooplankton, but very little in the way of phytoplankton for them to eat. This comes about because, even though the standing stocks of plants are small, their production provides sufficient food for their numerous consumers. Calculations using the amounts of nutrients in the water and the estimated rate of uptake and productivity of the algae demonstrate that algal biomass can be very small, sometimes as small as 0.5% of the amounts expected. The rest is being consumed by grazing as fast as it is produced.

This has important consequences where there are marked seasonal differences in intensity or period of light. During Spring, light intensity and period of daylight increase in the higher latitudes of the northern hemisphere. Phytoplankton then increase in abundance. In most areas, it takes some weeks for the populations of zooplankton to respond, because their reproduction is slower. Therefore, there is an algal bloom, followed by a decline. The decline is due to two processes, the first of which is grazing by increasing numbers of zooplankton. The second process is availability of nutrients. As phytoplankton increase in abundance, they take nutrients out of the water. When grazed and digested by zooplankton, some of the nutrients are released back into the water, but in smaller amounts. Thus, production of phytoplankton is slowed because there are smaller amounts of nutrients.

This situation is made more complex where there are higher-level consumers, such as fish, that eat the zooplankton. It is then possible for the amount of grazing to be reduced and for the algae to become much more numerous and abundant in the waters. In these circumstances, the control of algal productivity is very much a combination of bottom-up (nutrients control productivity) and top-down (grazers are controlled by predators and therefore do not control the algae, but are themselves top-down controlled). Whichever outcome is going to happen in some area is therefore dependent on the seasonality of availability of light and nutrients, the abundances of zooplankton, the availability and intensity of activities of predators, etc.

Modeling Grazing Processes

There have been many different developments of mathematical models to simulate or analyze grazing and its effects. A compelling modeling framework should include what happens to the plants, what happens to the grazers and, where appropriate, the consequences of predators influencing the numbers or activities of grazers. One example from Caughley (1977) is briefly considered here.

The rate of growth of biomass of a population of vegetation, V, can be modeled as a logistic equation:

$$r_v = r_{mv}(1 - V/K)$$

where K is the carrying capacity, that is, the amount of vegetation in some area when it reaches the maximum that can be sustained by available resources. r_v is the rate of increase, V is the amount of vegetation present and r_{mv} is the intrinsic rate of increase, that is, how much the biomass would increase in the absence of any limit or constraint due to shortage of resources. As V gets closer to K (the biomass of plants increases towards its carrying capacity), there is a decrease in r_v, the rate of growth. At the carrying capacity (K), there is no further increase in the plant population.

Meanwhile, if herbivores are eating plants at rate c_1 (how much is eaten by a single herbivore when food is freely available, i.e., V is large), which is reduced when the amount of food is smaller. This consumption by herbivores is at a rate:

$$c_1\left(1 - e^{-d_1 V}\right)/V$$

where c_1 and V are as above and d_1 is a constant describing how consumption changes from 1 to 0 due to decreased amounts of food.

At any moment, the biomass of plants is therefore changing at a rate:

$$r_v = r_{mv}(1 - V/K) - c_1\left(1 - e^{d_1 V}\right)/V$$

The population of herbivores is, however, also changing at a rate:

$$r_H = -a_1 + c_2\left(1 - e^{-d_2 V}\right)/V$$

where a is the rate of decrease per capita when there is no food; c_2 is a constant describing how the decrease is less when food is available.

All other factors being held constant, the grazer-plant interaction will reach an equilibrium (i.e., at populations V' and H') when:

$$V' = \frac{1}{d_2} \qquad \log_e\left(\frac{c_1}{c_1 - d_1}\right)$$

$$H' = \frac{V' r_{mv}\left(1 - V'/K\right)}{c_1\left(1 - e^{-d_1 V'}\right)}$$

This modeling shows, among other things, that the equilibrium population of plants in a grazed habitat is not affected by the intrinsic rate of increase of the plants (r_m is not in the equation for V'). Instead, increases in r_m lead to greater populations of herbivores (c_2 and/or a_1) leading to changes in equilibrial values of populations of plants and herbivores (c_2 and a_1 are in the equation for V'; V' affects H').

Finally, changes of populations of predators feeding on herbivores are modeled by:

$$r_P = -a_2 + c_3\left(1 - e^{-d_3 H}\right)/H$$

where r_p is the rate of change of the population of predators. H is the size of the population of herbivores (the prey); a_2, c_3, d_3 are as explained above for herbivores, that is, a_1, c_2, d_2.

Predators are also reducing the rate of growth of herbivores which is now:

$$r_H = -a_1 + c_2\left(1 - e^{-d_2 V}\right)/V - fP\left(1 - e^{-d_4 H}\right)/H$$

where f is the rate of consumption per capita of herbivores by predators, when there are no shortages of food for predators, P is the size of the population of predators and d_3 is a constant describing how f is altered as the population of herbivores changes.

This type of modeling can produce a steady-state (an equilibrium) or oscillating populations of plants, herbivores and predators, depending on the values of various constants. It is therefore quite a general tool for analyzing various grazer/plant, predator/grazer interactions and to examine "bottom-up" (V matters most) versus "top-down" (P matters most) controls of the populations of herbivores.

This type of modeling makes many assumptions (e.g., that growth of populations fits a logistic trajectory). It also assumes that the ecology of the grazing system is simple and that there are no complex indirect interactions, multi-species influences of competition among species of grazers, etc.

Grazing and Indirect Interactions Among Species

Indirect interactions occur when the presence or activities of one species influence the interactions among other species. There are, in general, two major types of indirect interactions: (i) an interactive chain or (ii) a modifying interaction. Grazing can be involved in either of these. An interactive chain is the situation where one species of plant (A) is a superior competitor for space or light or nutrients compared with a second plant (B). Left to their own devices, A would outcompete B in any shared patch of habitat. If, however, grazers arrive which preferentially consume the superior competitor, there is a resulting indirect, positive effect on B, the inferior competing plant. Thus, grazers have a direct, negative effect on A, their food and an indirect, positive effect on B because of the reduction in competition.

A modifying interaction occurs when one species indirectly influences the direct interaction between two others. An example involving grazing is the situation where a grazer (A) consumes a plant (B). There are, however, other plants (C) which are well-defended from attacks by grazers, because they have spikes, thorns, etc., which discourage the grazers from foraging near them. Where plant species B and C grow together, C causes an indirect, modifying interaction by reducing the effectiveness of grazing on B and, thus, has a positive, indirect effect on B and, usually, a direct negative effect on the grazers (A), which have their foraging reduced.

An example of an interactive chain is grazing of seaweeds on the rocky shores of New England (USA) by littorinid snails. In some areas, the snails consume green algae, such as *Enteromorpha intestinalis*. On the open shore, these green algae can be eliminated by grazing, resulting in a reduction in the number of species present. In contrast, in shallow rock-pools, the green algae can grow over and reduce the cover of encrusting species of seaweed. The encrusting species are tough and difficult for the snails to eat, in contrast to the soft-bodied *E. intestinalis*. As a result, the effect of grazing on *E. intestinalis* is to increase the survival or abundance of the encrusting species.

This is a classically known result for non-selective grazing or lawn-mowing of terrestrial vegetation. This was noted a very long time ago by Darwin, who commented that mowing lawns causes the persistence of more species of plants than is the case where plants are allowed to grow undisturbed. In the absence of mowing, a few species of plants eventually dominate the space, light and resources in the soil. Competition eliminates the others. Lawn-mowers operate in an unselective manner, cutting back all the species of plants and thereby preventing competitive dominance and maintaining a greater diversity of plants.

Most grazers are more selective in their choice of food, but can have important effects on the diversity of plant species. A long-known example (and one of the early sustained experimental ecological analyses of the influences of grazing) concerns rabbits as grazers in British grasslands. Where rabbits have been experimentally excluded from plots of grasses, the experimental areas became dominated by a reduced number of species. Rabbits were generally consuming competitively dominant plants, such as clover.

This situation is, however, made more complex because grazers such as rabbits can exist in very large numbers. Intensive grazing then occurs and selectivity of species of food breaks down, so that less preferred species are also eaten in large quantities. Under these circumstances, large numbers of rabbits cause reduced diversity, as a direct effect of the increased intensity of grazing.

By the 1920s, it was known that the number of species of plants was strongly related to the intensity of grazing by rabbits. When there is no grazing, the direct influences of competition among plants cause there to be few species present. Under intermediate amounts of grazing, competitively dominant species are consumed selectively and inferior competing species persist.

Diversity is greater. At very great intensity of grazing, many species of plants are consumed, some to extinction, thus reducing the diversity of species.

Such influences on diversity of grazing (or predation or physical disturbances) are these days described by the model of "Intermediate Disturbance". Where grazing or disturbance is mild or non-existent, competition results in loss of species. Where grazing or disturbance is frequent or intense, it causes direct elimination of species. Diversity of species is greatest at intermediate levels of grazing or disturbance, where these act selectively to reduce the abundances of competitively dominating species, thus allowing inferior competitors to survive.

Switching and Facilitation by Grazers: Mechanisms Sustaining Biodiversity

Grazers do not only change the diversity of their food-plants by direct action or by indirect interactions as discussed above. By changing their own behavior, grazers can sometimes ensure that food-plants are not eliminated.

The processes used by herbivores to choose what to eat are complex. Some situations conform to "optimal" selection; the composition of the diet reflects how the grazer may be able to maximize the input of energy for the minimal expenditure of energy used to acquire the food. In some cases, the consumption of different species is altered by the risks imposed in feeding. Venturing into open spaces to gain access to some species of plant may make a herbivore vulnerable to attacks by predators.

There are also defensive systems adopted by plants, that is, modifications to morphology to make tissues tougher to eat or to provide protection in the form of thorns and spines. Chemical defenses consist of noxious or bad-tasting chemicals which herbivores learn to avoid.

It is common for herbivores to display choices or preferences among different potential components of diet. Modification of behavior by individual grazers can make very marked differences to the persistence of diversity of the plants they are eating. A well-known example is switching-changing between different species of food as their relative abundances change because of grazing.

Switching occurs when grazers preferentially consume the more abundant of the available food-species, but change to eat different species when these become more abundant. Thus, certain butterflies (e.g., *Battus philenor*) lay eggs on two different species of plants. The caterpillars then hatch and consume the plant on which they have been laid. The adults tent to seek out the more common of the two plants in disproportionately large relative abundances. Thus, if species A to B are in the ratio 60:40, A will be chosen for oviposition at >60% of all the plants on which eggs are laid. In areas or at times when B are more abundant, the butterflies change to lay eggs on an excessively larger proportion of B. Where a grazer actually consumes the more abundant species in disproportionally excessive amounts, it will reduce the relative abundance of that species. Under these circumstances, the grazers then alter the availability of different components of their diet and switch to whichever has now become more available.

Complex interactions among grazers can also have as an outcome the continued persistence of a suite of grazing species. A particularly well-known example is the set of grazing mammals in the Serengeti Plain in East Africa. There are numerous grazers (>20 major species) of which the three most studied are zebra (200 kg weight), wildebeest (160 kg) and Thompson's gazelle (25 kg). These mammals migrate over very large distances, in a cyclic pattern. All three grazers eat the same grasses, but consume different parts of the plants.

Zebra mostly eat stems and the sheaths of grasses, which are the predominant components of older plants. They obtain adequate nutrient by consuming very large amounts of plant material. Removing these coarser and relatively indigestible components of the grasses stimulates growth of new leaves and makes the food-supply more accessible to wildebeest, which eat grass-sheaths and leaves. The wildebeest, in turn, modify the grasses and also expose various understorey herbs, which are then eaten by gazelles. The gazelles' diet is leaves and herbs, the latter containing the most nutrition.

Experimental enclosures demonstrated that wildebeest can reduce the biomass of plants to about 15% of that in ungrazed areas. This does, however, cause an increase in the production of new growth of leaves and, after the wildebeest move to new areas, there is a great amount of food available to gazelles.

Zebra move into areas where there has been no grazing since the previous year. Their feeding reduces the availability of their food, so they move on to new areas. Grazing by zebras facilitates feeding by wildebeest because the zebra have made the plants more suitable for them. The wildebeest, in turn, consume parts of the plants and then, as their food decreases in abundance, move on to the areas recently fed on by zebra. The wildebeest facilitate feeding by gazelles, by making the latter's food accessible. The gazelle then follow the wildebeest. The entire cycle starts again when zebra return to the first areas, where gazelles ceased feeding months before and where grasses have grown up and are now older and with more stems and fewer leaves.

This cyclic pattern of facilitated grazing maintains the suite of grazers which use different parts of the grasses, in turn, maintaining the biodiversity of large grazing mammals.

Defenses Against Herbivores

Plants show numerous responses to being grazed. Morphological and chemical defenses are the two most commonly studied. Some plants produce thickened bark or tough leaves, making them difficult to chew and thus reducing damage done by grazers. Plants can also produce thorns or spikes, preventing some grazers from being able to climb on them to eat their soft tissues.

Many plants respond to grazing by producing bad-tasting or toxic chemicals. These can dissuade actively foraging herbivores from continuing to feed or can cause them to learn to avoid that plant.

The chemicals used in defense seem to be of two general types. First, there are compounds that are produced in the plant's normal metabolism that have noxious or toxic properties when contacted or ingested by animals. The plants respond to herbivory by accumulating these otherwise waste products, so that they are present in their tissues in sufficient quantities to deter grazers.

Second, the chemicals may be produced specifically in response to attacks by grazers, but would otherwise not be part of normal metabolism. Most of such chemicals would be produced at an energetic cost to the plants.

These compounds are called secondary metabolites. They are found in some species or populations of plants, but not others and are not an absolute necessity for metabolism of all plants. The diversity of types of compounds is enormous and the ecology of chemical defense by plants is a vast topic in its own right, so will only briefly be illustrated here.

A well-known example of mechanical defense is that of oak-trees in Europe. This species is attacked by larvae of >200 species of butterflies and moths, but most attack is in spring when the leaves of the trees are young. The leaves change color and toughness during spring. Experimental rearing of larvae showed greater growth on a diet of young leaves (those first produced in the northern spring) than of those fed on older leaves (those collected in the third and subsequent week of the appearance of leaves). Older leaves suffered from reduced grazing, primarily because they are tougher.

A particularly well-studied example of chemical defense is the production of indole glucosinolate compounds by cabbages when they are attacked by herbivores. When grazed, the plants start to produce more of the defensive chemicals and grazing is subsequently reduced or prevented.

Induction of defensive chemicals rather than the plants always containing them is believed to be a response by the plants to the behavior of specialist species of grazers. Thus, experimental evidence demonstrates that production of glucosinolates causes increased attack by larvae of cabbage-white butterflies and some other species. Producing the chemicals only in response to grazing reduces this risk.

There are other influences on production of chemical defenses. For example, insects are quite often found destructively grazing trees which are normally attacked only in minor ways. One explanation for this is that the production of secondary metabolites costs the plant energy. When plants are stressed by environmental factors such as drought or reduced amounts of nutrients, production of energetically expensive chemical compounds is reduced. Attacks by insects are then much more severe. The interaction of excessive grazing on top of other environmental stresses is then a serious problem for the survival of the plants.

Another aspect of chemical defenses is tri- or multi-trophic defense. This occurs when plants subjected to grazing release chemicals to which other organisms respond. Some of the responding animals are predators, which arrive and start to eat the grazers.

An example of indirect defense is plants that are attacked by spider mites (*Tetranychus urticae*). When the mites begin to eat a plant, it releases volatile compounds from the damaged tissues. These are attractive to predatory mites (*Phytoseiulus persimilis*) which arrive at the damaged plant and proceed to eat the plant-eating mites. Adjacent, unattacked plants can also start to emit the signaling chemicals. The chemicals used for such tritrophic (plant–herbivore–carnivore) interactions are of two different types. Some are specific-they are only produced and released from the plant when it is attacked. Others are the same chemicals normally found in undamaged or mechanically damaged plant-tissue. These compounds are, however, released in much greater quantities when the plants are attacked by herbivores.

A final aspect of defense against grazers concerns the biology of plants that cause predators of their grazers to live permanently with them, thus deterring or reducing the activities of grazers.

Some features of plants that encourage predators to live with them are the construction of sheltering galls or special sized "domatia" which are inhabited by predators or fungal-feeding arthropods. These structures are modified veins of leaves or specialized pits or crevices in the cuticle. They provide shelter from wind and weather and can also prevent attacks by enemies of the predators. The presence of permanent populations of their predators clearly reduces the number and effectiveness of grazers.

Other plants maintain populations of defensive predators because they provide food for them. Acacia ants (*Pseudomyrmex* spp.) in tropical regions of America live in hollow thorns on some species of *Acacia*. The ants feed on sugars, proteins and oils produced by the plants. When ants are experimentally removed from plants, the acacias grow much more slowly and more of them die as a result of grazing. While foraging on the plants, the ants also attack any grazing species that they encounter. They also remove seeds and any other new vegetation that appears on the ground around the base of their host-tree. This potentially reduces competition from surrounding plants, resulting in yet more protection for the host acacias.

Grazing Promoting the Growth of Plants: Examples of Gardening

Some grazing animals are territorial, in that they occupy, patrol and defend a patch of habitat in which they live, feed and, often, breed. In some cases, the direct or indirect consequence of this behavior encourages the growth of particular plants, so that the grazing species is farming or gardening its supply of food.

Examples are widespread, but one example of each of a vertebrate and an invertebrate grazer will illustrate some general features. First, pomacentrid or damsel fishes on coral reefs live in areas that are over-grazed by a large diversity of other fish and many invertebrates. Several species in the family Pomacentridae occupy territories in which they threaten and attack intruding fish. In these territories, there can be extensive growth of algae which are absent from adjacent, undefended areas. The territorial

resident grazes on these algae-a supply of resources created by its activities. Some species of fish select what they allow to grow within their territories. They weed out unwanted species of seaweeds by biting them off the substratum and carrying them out of the territory. As a result, territorial fish can have profound influences on local diversity of algal species.

Despite vigilance and rapid responses to intruders, the occupant of a territory can, however, be overwhelmed by the sheer weight of numbers of intruders if large numbers arrive at the same time. This is almost certainly a strong influence on how large a territory an individual can defend. If a school of grazing fish arrive in a territory, no amount of threat, harassment or attack by the occupant will prevent removal of some of the algal food.

The second example concerns grazing by limpets on rocky shores. Most limpets wander around grazing over the surface of the rock using a special feeding organ or toothed "tongue", called a radula. This scrapes over and into the surface of the rock, removing grains of rock and any attached bacteria or microscopic plants, which are then swept into the animal's mouth. In many parts of the world, grazing by these and other gastropods (snails) removes such a large proportion of the algae that the animals generally suffer from a shortage of food and often show evidence of considerable amounts of competition for the food that is present.

In some areas of the world, such competition has resulted in aggressive behavior to defend a territory. On Californian intertidal shores, the owl-limpet (*Lottia gigantea*) reaches a few cm length as its maximal adult size. It patrols an area of rock about 1000 cm^2 extent. When it encounters an intruding limpet or grazing snail, the resident limpet attacks by ramming the front of its shell against the intruder. Sometimes this dislodges the intruder which is then rolled away by waves. Sometimes, the attack, or even the threat of attack as the resident approaches, causes the intruder to leave the territory.

The outcome is a very reduced number of grazers inside the territories. When a resident limpet is experimentally removed, the numbers of other, smaller grazers increases. Often, a resident that has been removed is replaced by a smaller individual of *L. gigantea* from a nearby area. If an owl-limpet is introduced to an area, it will eventually drive away many of the resident grazers.

The result of this activity is the growth inside the territory of a film of visible macro-algae, which are a major source of food for the limpets. When a territorial limpet is removed, smaller grazers enter the territory and remove all the algal film because they scrape much closer to the surface of the rock.

Grazers and Indirect Effects

Some grazing activity can have profound effects on ecological processes, diversity and variety of co-existing species and general structure of assemblages, involving many other species, not just the plants consumed for food. Such interactive consequences of grazing have led to defining some grazers to be ecosystem engineers – species that modify habitats so that they become more or less suitable for occupation by other species. Engineering species must make or modify patches of habitat to create the environmental conditions not found outside the modified patches.

A particularly large-scale engineering species is the beaver, *Castor canadensis*, which grazes on plants on the riparian zones of rivers. Grazing by beavers has profound effects on the composition of species of plants in these areas because of major reductions in the abundances of preferred food-plants.

Importantly, however, beavers also dam a stream creating a pond. In some areas, the dam and pond are maintained for <10 years, during which period the surrounding vegetation is substantially modified by the grazers. The dams trap sediments which store nutrients. The banks of the areas in which beavers forage are generally devoid of woody vegetation, allowing more light than would occur where there were no beavers.

When the beavers abandon a pond to seek new supplies of food elsewhere, the dams eventually break down and large meadows of grasses or swamps dominated by alder (*Alnus incana*) develop. These engineered habitats can persist for at least 50 years before the habitat eventually reverts to the unmodified woodland.

Beavers, as grazers, have clear direct effects on the assemblages of plants where they are actively foraging. Because of their engineering (dam-building), they also create ponds with nutrient-rich sediments and plenty of light. These habitats can continue to exist for decades after the beavers cease their activity.

There are also examples of indirect effects of invertebrate grazers that can cause profound shifts in the species occupying local assemblages. On many rocky shores, intertidal assemblages contain numerous grazers (mostly snails and limpets, but also crabs, chitons, starfish, polychaetes and some fish). Where grazing is sufficiently intense, the grazers remove much, if not all, of the microscopic algal food from the rocks. This food consists of truly microscopic species, such as diatoms. It also contains the spores, gametes and early sporelings of many macroscopic species which would, if not consumed whilst still at unicellular or microscopic sizes, grow up to form erect, upright seaweeds.

Sometimes grazers are absent, as a result of storms or other disturbances, diseases or excessive predation by their own enemies, or simply because of failure to recruit for long enough for the existing adults in a population to die without being replaced. As a result, spores and early reproductive stages of the algae survive in very large numbers compared with the usual grazed condition. The consequence is that these algae then grow up to occupy the space on the shore.

The growth of algae has serious and sometimes long-term consequences for the other species in such an area. Most of the species in the diverse assemblage on intertidal rocky shores require open, relatively bare space to live on or to feed over. For example, barnacles and tubeworms recruit from developing stages in the plankton. These are washed along a coastline by waves, winds and currents, developing to a stage capable of settlement and metamorphosis in a suitable adult habitat. For this to be possible, they need uncluttered space, either bare rock or a surface covered by a biofilm of unicellular and extracellular organic

material. Where the surface is already covered by algae, it is impossible for the sessile species to become established. Even if larvae can settle and metamorphose, they are usually smothered by the algae growing over them, preventing them from growing or feeding.

The grazers themselves are often eliminated by the algae once these grow large enough. Many of the invertebrate grazers are quite unable to consume seaweeds once they are too large. There are some invertebrates, such as sea-urchins, which have no problems, but they are not usually common components of intertidal shores.

Thus, grazing by micro-algal feeders frees space that is then occupied by a range of other sessile space-users and a suite of grazing species. This maintains the diversity of species from many Phyla in the assemblage. Algae grow more quickly in areas lower on the shore, where habitats are subject to greater splash and spray during low tide and longer periods of submersion under water during high tide. As a result, in such areas, grazers are less effective and unable to keep surfaces free from foliose algae. Consequently, there are fewer animals and types of animals in such areas.

In addition to their direct consumption of plants, grazers, through their indirect effects on habitat, contribute a lot to the maintenance of biodiversity in assemblages of other species that are neither the food-plants nor their consumers.

See also: Ecological Data Analysis and Modelling: Grassland Models. Ecological Processes: Succession and Colonization

Further Reading

Branch, G.M., 1984. Competition between marine organisms: Ecological and evolutionary implications. Annual Review of Oceanography and Marine Biology 22, 429–593.
Caughley, G.J., 1977. Analysis of vertebrate populations. London: Wiley Interscience.
Dicke, M., 1999. Evolution of induced indirect defense of plants. In: Tollrian, R., Harvell, C.D. (Eds.), The ecology and evolution of inducible defenses. Princeton: Princeton University Press, pp. 62–88.
Fryxell, J., 1997. Functional responses to resource complexity: An experimental analysis of foraging by beavers. In: Olff, H., Brown, V.K., Drent, R.H. (Eds.), Herbivores: between plants and predators. Oxford: Blackwell Science, pp. 371–396.
Hairston, N.G., Smith, F.E., Slobodkin, L.B., 1960. Community structure, population control, and competition. American Naturalist 94, 421–425.
Hay, M.E., Fenical, W., 1988. Marine plant-herbivore interactions: The ecology of chemical defenses. Annual Review of Ecology and Systematics 19, 111–145.
Hutchinson, G.E., 1961. The paradox of the plankton. American Naturalist 95, 137–145.
Janzen, D.H., 1966. Coevolution of mutualism between ants and acacias in Central America. Evolution 20, 249–275.
Karban, R., Baldwin, I.T., 1997. Induced responses to herbivory. Chicago: University of Chicago Press.
Lubchenco, J., Gaines, S.D., 1981. A unified approach to marine plant-herbivore interactions. I. Population and communities. Annual Review of Ecology and Systematics 12, 405–437.
McNaughton, S.J., 1970. Serengeti migratory wildebeest: Facilitation of energy flow by grazing. Science 191, 92–94.
Sabelis, M.W., Van Baalen, M., Bakker, F.M., Bruin, J., Drukker, B., Egas, M., Janssen, A.R.M., Lesna, I.K., Pels, B., Van Rijn, P.C.J. and Scutareanu, P. (1997). The evolution of direct and indirect plant defence against herbivorous arthropods. In Olff, H., Brown, V.K. & Drent, R.H. (eds) Herbivores: Between plants and predators. pp. 109–166. Oxford: Blackwell Science.
Wootton, J.T., 1994. The nature and consequences of indirect effects in ecological communities. Annual Review of Ecology and Systematics 25, 443–466.

Greenhouse Gases Formation and Emission

Antonio C Barbera, University of Catania, Catania, Italy
Jan Vymazal, Czech University of Life Sciences Prague, Praha, Czech Republic
Carmelo Maucieri, University of Padua, Legnaro, Italy

Glossary

Biogeochemical cycles Natural pathways by which the chemical elements that are present in the organic matter are circulated through biological, geological, and chemical cycles in the biotic and abiotic Earth's compartments.
Global warming A rise in the Earth's standard atmospheric temperature that causes related changes in climate and that may result from the greenhouse gases effect.

Introduction

The rising temperature of earth, known as global warming (GW), is mainly the result of the rise of greenhouse gases (GHGs) concentration in the atmosphere since the beginning of the 20th century, mostly due to anthropogenic activities. Global warming, as we know, is one of the major threats to the environment because of the resulting climate change. Oxygen (O_2) and nitrogen (N_2) are the principal atmosphere gases at concentrations of 21% and 78%, respectively, nevertheless it is thought that they do not absorb or emit thermal radiation. Water vapor and less abundant gases, such as carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O), because of their long atmospheric lives and their relatively high thermal absorption capacities, are generally known as greenhouse gases. Besides water vapor that is not considered to be a cause of man-made global warming because it does not persist in the atmosphere for more than few days, the most important greenhouse gas is CO_2, followed by CH_4 and N_2O which are the other most relevant contributors to GW. Hundred year global warming potential of methane, nitrous oxide and other GHGs is compared to CO_2 in terms of CO_2 equivalence (CO_2eq). This last is a simple way to normalize all greenhouse gases in a standard unit based on the radiative forcing of a unit of CO_2 over 100 years. For example 1 g of methane has a global warming potential 28 times higher than 1 g of CO_2 and so it is expressed as 28 g of CO_2eq. Biological and/or non-biological, as natural and anthropogenic processes are involved in GHG cycling (most carbon (C) and nitrogen (N) cycles).

Soils can be a source or a sink of CO_2, CH_4, and N_2O depending on the specific conditions. Natural soils are mainly a sink for natural GHGs and sequester as much C and N they emit, but due to human activities, mainly agriculture, soils can be mainly a source for GHGs. Infact, intensive agriculture, sustained by mineral fertilizer use, has contributed significantly to the elevation of atmospheric GHGs, including CO_2, CH_4, and N_2O. Rising GHG emissions usually lead to a decrease in soil C. Currently, soil organic C is twice that of all standing crop biomass, making it an extremely important player in the C cycle.

Anyhow, agronomic practices have the potential to reduce agricultural GHG emissions. Main management practices that impact GHG emissions and soil C content include various tillage practices, several N fertilization amounts and typologies (mineral, manure, or a combination of both), the use of cover crops, aeration, and water levels. Moreover, other agriculture GHGs sources are intensive livestock such as cattle intensive production systems and rice paddy fields. Furthemore deforestation, expecially in tropical rainforest, adds more C to the atmosphere than cars and trucks trafic; the reason is that when trees are cut down or dead, they release their carbon content in the environment. Employing best agricultural management practices (BMPs), we can promote the sequestration of CO_2 plus N_2O and the preservation of soil C. Measuring soil C storage and GHG emissions and using them as metrics to evaluate BMPs are vital in understanding agriculture's role in this topic.

Carbon Greenhouse Gases

Carbon Dioxide (CO_2)

CO_2 is a colorless gas with a density about 50% higher than that of dry air. In the CO_2 molecule (molecular weight of 44 g mol^{-1}) carbon atom is covalently double bonded with two oxygen atoms, and it is naturally present in the earth's atmosphere. Because the oxygen of CO_2 molecule forms intermolecular hydrogen bounding with the hydrogen of water, CO_2 is soluble in water, and it occurs in groundwater, rivers and lakes, ice caps, glaciers, and seawater (CO_2 solubility in water is about 1.45 g L^{-1} at 15°C) and in deposits of petroleum and natural gas too. CO_2 is odorless at usual concentrations, however its aqueous solution has, at high concentrations, a sharp and acidic odor and taste.

Among the greenhouse gases, CO_2, as presented in the introduction paraghaph is the reference gas with a global warming potential of 1, lower than all other greenhouse gases. However it contributes >60% to GW due to its huge emission amount. The concentration of CO_2 in the atmosphere has increased from approximately 277 parts per million (ppm) in 1750, at the beginning of the Industrial Era, to 402.8 ± 0.1 ppm in 2016. The increase of atmospheric CO_2 above the preindustrial levels was primarily caused by the release from deforestation and other land-use change activities. While emissions from the burning of fossil fuels started before the Industrial Era, they only became the dominant source of anthropogenic emissions to the atmosphere from around 1920 and their relative share has continued to increase until present.

Sources of atmospheric CO_2 include volcanoes activities, the combustion and decay of organic matter, respiration of aerobic (oxygen-using) organisms; extraction and burning of fossil fuels, clearing of lands, and production of cement and steel an so on by human activities.

These sources are at least partially balanced by a set of physical, chemical, or biological processes, called "sinks," that tend to remove CO_2 from the atmosphere. The main C sink are plants that through the photosynthesis process grab CO_2 from atmosphere.

Plant life is a foundamental natural sink. In the oceans, marine life can absorb dissolved CO_2, and some marine organisms even use CO_2 to build skeletons and other structures made of calcium carbonate ($CaCO_3$).

Organic matter respiration, operated by heterotrophic organisms, is the most important path through which the CO_2, photosynthetically fixed as glucose, comes back to the atmosphere:

$$\text{Glucose} + 6\ \text{Oxygen} \rightarrow 6\ \text{Carbon dioxide} + 6\ \text{Water}$$
$$6(CH_2O) + 6O_2 \rightarrow 6CO_2 + 6H_2O$$

This organic matter degradation occurs in both soils (primarily in the rooting zone) and oceans (within the surface mixed layer). Organic matter respiration and CO_2 release in the atmosphere are controlled in part by the composition of the organic matter and the environmental factors (e.g., temperature, soil moisture, and soil porosity, texture/mineralogy).

In total, more than twice as much C is stockpiled in the world's soil than in the vegetation or atmosphere combined. Considering all the C stored in soil, soil organic carbon (SOC) makes up about 50% of all soil organic matter (SOM).

About a third of the soil organic C occurs in forests, another third is in grasslands and savannas, and the rest is in wetlands, croplands, and other biomes. SOM is composed by soil microorganisms communities (mainly bacteria and fungi), plant and animal tissues, fecal material, and products derived from their decomposition. Soil CO_2 flux is primarily the result of a combination of microbial decomposition of SOM and plant root respiration. The main drivers of soil CO_2 flux are soil temperature, soil moisture, and substrate C availability. Temperature affects CO_2 flux by speeding up the rate of microbial decomposition when soils are warm and water is not a limiting factor. Although rising temperatures cause an increase in CO_2 flux rate from soils, in some parts of the world there are no clear trends of decreasing soil carbon with increasing mean annual temperature. This is due, partly, to competing processes within the system, such as SOC rising due to increased primary productivity that principally occurs through the photosintesis process (better water and nutrient availability), and SOC decreasing by increased respiration processes. While in the short-term, warming depletes SOC, in the long-term, C losses by accelerated microbial respiration may be equalized by increases in C inputs to the soil tied to increased net primary production, as well as any acceleration of soil physico-chemical "stabilization" reactions. Additionally, changes in microbial community composition or declines in the temperature sensitivity decomposition processes may reduce the response of microbial respiration to increasing temperature over time (i.e., thermal acclimation cold soil and air temperatures have the opposite effect on CO_2 flux rate, causing it to slow down). Even though slowed, soil microorganisms maintain both catabolic (CO_2 production) and anabolic processes (biomass synthesis) under frozen conditions. Because of this, gaseous exchange between the atmosphere and soil does not stop even under frozen soil, resulting in the accumulation of CO_2 during winter and its release into the atmosphere during spring thaw events. Another dominant factor controlling the net exchange of GHGs is soil moisture, which can vary dramatically over time and space. The production and transport of GHGs in soil is strongly affected by changes in soil moisture through diel 24-h period cycles, wet-up and dry-down events, management practices, seasonal patterns, and interannual variation in climate. Overall, when water is limiting, plant and microbial availability increase with soil moisture, thereby increasing soil CO_2 flux directly by alleviating plant and microbial desiccation stress and indirectly by increasing substrate availability (via higher rates of plant growth, photosynthesis, belowground C allocation by root exudate) and microbial access to substrate for example, increase C diffusion through soil water. Finally, respiration generally increases with C availability. Plant respiration is largely dependent on C from current photosynthetic activity and, under non-limiting soil temperatures and moisture availabilities, microbial respiration increases with labile C availability. Thus, soils with high organic matter inputs and stocks, like those found near the equator, means greater C substrate availability, which is synonymous with greater flux. Depth and

placement of soil carbon is yet another factor to consider when attempting to make precise conclusions about CO_2 flux. For example, in agroecosystems, the bulk of SOM is within the top 10 cm of the soil surface.

Because of this, temporal dynamics of CO_2 flux are more intimately related to air temperature than to soil temperature. Also, it is known that the respiration rates of many soils are strongly linked with the amount of carbon not intimately associated with minerals. Mineral soil occurs below the litter and organic layer, where soil carbon may be closely associated with mineral particles—accounting for over 60% of carbon in most forest soils, proposing that the decomposition/respiration rate of mineral soil carbon is relatively insensitive to temperature. This is because the carbon located here may be protected from microbial mineralization by stabilization mechanisms, such as occlusion in soil aggregates (physical protection) or interactions with mineral surfaces (chemical sorption to mineral surfaces).

Methane (CH_4)

CH_4 is the second most important greenhouse gas in volume after CO_2, contributing to global warming potential with 28 times more infrared radiative heating effect than CO_2 on a mole-per-mole basis at a 100 year time horizon. Its atmospheric concentration, due to human activity, has increased by a factor of 2.5 since preindustrial times, that is, from 722 ppb in 1750 to 1803 ppb in 2011. Furthermore, after almost one decade of stable CH_4 concentrations since the late 1990s, atmospheric measurements have shown a renewed concentration increase since 2006, probably tied to a rise in natural wetland and fossil fuel emissions.

Generally, CH_4 emissions in atmosphere can be due to both human activities or natural processes from biogenic, thermogenic or pyrogenic sources. Anthropogenic emissions account for 50%–65% of total emissions including coal mining, natural gas use, agriculture, wastewater and waste treatment.

Considering CH_4 budget for the decade of 2000–09 (bottom-up estimates) agriculture and waste sectors (rice, animals and waste) are the second contributors to biogenic CH_4 emissions (with values ranging from 187 to 224 Tg (CH_4) year^{-1}) after natural wetlands emissions that ranged from 177 to 284 Tg (CH_4) year^{-1}.

Soils play an important role in the CH_4 cycle because both methanogenesis (CH_4 production) and methanotrophy (CH_4 oxidation) take place in them. In view of this soils CH_4 fluxes are the net result of the CH_4 production by methanogenesis and CH_4 oxidation by methanotrophy processes.

Methane production

Methanogenesis is operated by strictly anaerobic bacteria which requires negative oxydo-reduction potentials ($Eh < -200$ mV). Methanogens belong to the domain Archaea which have a limited trophic spectrum comprised of a small number of simple substrates: $H_2 + CO_2$, acetate, formate, methylated compounds (methanol, methylamines, dimethylsulphur), and primary and secondary alcohols. This allows to distinguish five trophic groups of methanogens: hydrogenotrophs, formatotrophs, acetotrophs, methylotrophs, and alcoholotrophs. The two major pathways of CH_4 production in most environments where organic matter decomposition is significant are acetotrophy and CO_2 reduction by H_2.

The possible pathway for CH_4 emission from soil are: (i) diffusion of dissolved CH_4 along the concentration gradient, (ii) release of CH_4-containing gas bubbles (ebullition), and iii) transport via the aerenchyma of vascular plants (plant-mediated transport).

The first process, diffusion, takes place because of the formation of a CH_4 concentration gradient from deeper soil layers, where the production of CH_4 is large, to the atmosphere, while oxidation of CH_4 occurs in upper layers. Diffusion is a slow process compared to the other two transport mechanisms, that is, ebullition and plant-mediated transport, but it is biogeochemically important because it extends the contact between CH_4 and methanotrophic bacteria in the upper aerobic layer, promoting CH_4 oxidation.

The second process, ebullition, takes place when CH_4 production is large. Gas bubbles are formed and emigrate to the surface. As this process is fast, CH_4 oxidation is absent or negligible.

The third process, plant mediated transport, takes place through an internal system of continuous air spaces named aerenchyma, a structure which is developed by vascular plants to adapt to flooded environments. The basic function of this structure is to transport the O_2 necessary for root respiration and cell division in submerged organs, but it is also used for CH_4 transportation from the rhizosphere to the atmosphere, bypassing the aerobic, CH_4-oxidizing layers. This process involves two major mechanisms: molecular diffusion and bulk flow. The gradient of CH_4 concentration formed inside the aerenchyma conduits is the driving force for CH_4 diffusion from the peat root zone to the aerial parts of the plant. The other plant-mediated transport mechanism, bulk transportation, involves the migration of CH_4 along the plant, also through the aerenchyma structure, from the leaves to the rhizome and back to the atmosphere through old leaves or horizontal rhizomes connected to other shoots. The driving force for this process is a pressure gradient generated by differences in temperature or water vapor pressure between the internal air spaces in plants and the surrounding atmosphere. This is a very efficient and rapid mechanism of CH_4 transportation and, in consequence, it is responsible for most of CH_4 emissions (>95%) to the atmosphere from rice paddies.

The factors controlling CH_4 production in soil are anaerobic conditions and redox potential, electron acceptors, substrate availability, temperature, diffusion, water availability and water table, soil pH and salinity, fertilizer and manure additions and amendments, trace metals, competitive inhibition, vegetation, plant species and cultivars, and elevated CO_2 concentrations.

Methane oxidation

The major fraction of CH_4 emitted in the atmosphere is submitted to oxidation. Oxidation by OH radicals represent the main sink of atmospheric CH_4 (around 90% of the global CH_4 sink) and it takes place mostly in the troposphere according to the reaction:

$$CH_4 + OH\bullet \rightarrow CH_3\bullet + H_2O$$

The remaining sink is thought to be split roughly equally between the stratosphere (removal by OH and O1[D]) and biological consumption in near-surface soils. Uptake of methane occurs via oxidation by specialized aerobic bacteria, methanotrophs, although CH_4 oxidation can be also monitored in anaerobic conditions.

The aerobic methane oxidation, operated by methanotrophs, can be summarized as follow:

$$CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O$$

Aerobic methanotrophic bacteria can be classified in three groups (type I, II, and X) considering morphological features, membrane structures, guanine and cytosine content, phospholipid fatty acids composition and various other physiological characteristics.

Several mechanisms have been offered to explain anaerobic CH_4 oxidation. First, sulfate-driven anaerobic methane oxidation:

$$CH_4 + SO_4^{2-} \rightarrow HS^- + HCO_3^- + H_2O$$

Other electron acceptors such as oxides of iron, manganese and chrome, could also oxidize methane anaerobically as below reported:

$$CH_4 + 8Fe(OH)_3 + 16H^+ \rightarrow CO_2 + 8Fe^{2+} + 22H_2O$$
$$CH_4 + 4MnO_2 + 8H^+ \rightarrow CO_2 + 4Mn^{2+} + 6H_2O$$
$$CH_4 + 4/3Cr_2O_7^{2-} + 32/3H^+ \rightarrow 8/3Cr^{3+} + CO_2 + 22/3H_2O$$

Another proposed CH_4 oxidation process is the anaerobic oxidation of methane coupled to denitrification with CO_2 release as below summarized:

$$5CH_4 + 8NO_3^- + 8H^+ \rightarrow 5CO_2 + 4N_2 + 14H_2O$$
$$3CH_4 + 8NO_2^- + 8H^+ \rightarrow 3CO_2 + 4\ N_2 + 10H_2O$$

Biological anaerobic CH_4 oxidation is done by a consortium of anaerobic archaea in association with anaerobic bacteria.

Optimal conditions for most of methanotrophs have been found in environments with near neutral pH, temperature in the mesophilic range (ca. 25°C) and low salinity.

Nitrogen Greenhouse Gases

Nitrous Oxide (N_2O)

The N_2O is one of the most important greenhouse gases with a global warming potential, an index of the total energy added to the climate system by a component in question relative to that added by CO_2, for a time horizon of 100 years of 298 $CO_{2(eq)}$. N_2O is also capable of ozone depletion with negative impact on world life due to the lower capability to shield negative fraction of solar radiation (UV). The concentration of N_2O in the atmosphere increased at a rate of 0.73 ± 0.03 ppb year^{-1} over the last three decades, mostly caused by nitrification and denitrification reactions of reactive nitrogen in soils and in the ocean.

The major source of N_2O in the atmosphere is heterotrophic denitrificatrion, that is, reduction of nitrate to nitrogen gas under anoxic conditions:

$$6(CH_2O) + 4NO_3^- \Rightarrow 6CO_2 + 2\ N_2 + 6H_2O$$
$$4(CH_2O) + 4NO_3^- \Rightarrow 4HCO_3^- + 2\mathbf{N_2O} + 2H_2O$$

It is generally agreed that during heterotrophic denitrification, nitrate is reduced to nitrite, which is followed by a stepwise reduction to nitric oxide, nitrous oxide, and finally nitrogen gas as below schematized:

$$2NO_3^- \Rightarrow 2NO_2^- \Rightarrow 2NO \Rightarrow \mathbf{N_2O} \Rightarrow N_2$$

This reaction is irreversible and occurs in the presence of available organic substrate mainly under anoxic or anaerobic conditions, where NO_3^-'nitrogen is used as an electron acceptor in place of oxygen. Nitrate is the first inorganic compound which used by bacteria as electron acceptor after dissolved oxygen depletion from the water-soil system. Denitrification starts at redox potential values of about +220 mV. Environmental factors known to influence denitrification rates include the absence of O_2, redox potential, soil moisture, temperature, pH value, presence of denitrifiers, soil type, organic matter, nitrate concentration and the presence of overlying water.

Facultative heterotrophic anaerobic bacteria are the main actors of heterotrophic denitrification. Most denitrifying bacteria are chemoheterotrophs. They obtain energy solely through chemical reactions and use organic compounds both as electron donors and

as a source of cellular carbon. Denytrifying bacteria are aerobes that substitute nitrate for oxygen as the terminal electron acceptor when there is little or no O_2 available. The genera *Bacillus*, *Micrococcus*, and Pseudomonas are probably the most important in soils; *Pseudomonas*, *Aeromonas*, and *Vibrio* in the aquatic environment.

The quantity of N_2O evolved during denitrification depends upon the amount of nitrogen denitrified and the ratio of N_2 to N_2O produced. The ratio is also affected by aeration, pH, temperature and nitrate to ammonia ratio in the denitrifying system. If the pH is below 4.5, the denitrification rate is relatively slow and only N_2O is produced. At $pH > 5$, N_2 is the main end product of denitrification. It has also been shown that certain amount of N_2O is formed during nitrification, however, it is difficult to distinguish between nitrification and denitrification as sources of N_2O.

Ozone (O_3)

O_3 is a short-lived trace gas that either originates in the stratosphere or is produced in situ by precursor gases (monoxide (CO), CH_4, and non-CH_4 hydrocarbons in the presence of nitrogen oxides (NO_x)) and sunlight. It is a bluish gas with a pungent smell, dipolar and electrophilic molecule capable of very selective reactions, present in the air at concentrations of 0.01–0.05 ppm. In the atmosphere the presence of the O_3 is crucial for any life because it prevents shortwave solar UV radiation in the UVB (λ 280–315 nm) and UVC (λ 200–280 nm) from penetrating the atmosphere and reaching the Earth's surface. On the other hand O_3 exert a greenhouse effect increasing the global warming. As reported by the Intergovernmental Panel on Climate Change in the fifth assessment report: emissions of carbon monoxide and volatile organic compounds, a not well-defined group of hydrocarbons, lead to production of ozone on short time scales. By affecting OH and thereby the levels of CH_4 they also initiate a positive long-term ozone effect. The effects via ozone and CH_4 cause warming, and the additional effects via interactions with aerosols and via the O_3–CO_2 link further increase the warming effect.

Further Reading

Davidson EA and Janssens IA (2006) Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature* 440(7081): 165–173.

Huang W and Hall SJ (2017) Elevated moisture stimulates carbon loss from mineral soils by releasing protected organic matter. *Nature Communications* 8(1): 1774.

IPCC (2013) In: Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, and Midgley PM (eds.) *Climate Change 2013:* The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, p. 1535. Cambridge and New York, NY: Cambridge University Press.

Janzen HH (2004) Carbon cycling in earth systems—A soil science perspective. *Agriculture, Ecosystems & Environment* 104(3): 399–417.

Ji B, Yang K, Zhu L, Jiang Y, Wang H, Zhou J, and Zhang H (2015) Aerobic denitrification: A review of important advances of the last 30 years. *Biotechnology and Bioprocess Engineering* 20(4): 643–651.

Lal R (2004) Soil carbon sequestration impacts on global climate change and food security. *Science* 304(5677): 1623–1627.

Le Mer J and Roger P (2001) Production, oxidation, emission and consumption of methane by soils: A review. *European Journal of Soil Biology* 37(1): 25–50.

Le Quéré C, Andrew RM, Friedlingstein P, Sitch S, Pongratz J, Manning AC, and Boden TA (2018) Global carbon budget 2017. *Earth System Science Data Discussions* 10(1): 405.

Lehmann J and Kleber M (2015) The contentious nature of soil organic matter. *Nature* 528(7580): 60–68.

Maucieri C, Barbera AC, Vymazal J, and Borin M (2017) A review on the main affecting factors of greenhouse gases emission in constructed wetlands. *Agricultural and Forest Meteorology* 236: 175–193.

McDaniel MD, Tiemann LK, and Grandy AS (2014) Does agricultural crop diversity enhance soil microbial biomass and organic matter dynamics? A meta-analysis. *Ecological Applications* 24(3): 560–570.

Megonigal JP, Hines ME, and Visscher PT (2014) *Anaerobic metabolism: Linkages to trace gases and aerobic processes. Treatise on geochemistry.* Amsterdam: Elsevier pp. 273–359.

Rui Y, Murphy DV, Wang X, and Hoyle FC (2016) Microbial respiration, but not biomass, responded linearly to increasing light fraction organic matter input: Consequences for carbon sequestration. *Scientific Reports* 6: 35496.

Serrano-Silva N, Sarria-Guzmán Y, Dendooven L, and Luna-Guido M (2014) Methanogenesis and methanotrophy in soil: A review. *Pedosphere* 24(3): 291–307.

Staehelin J, Harris NRP, Appenzeller C, and Eberhard J (2001) Ozone trends: A review. *Reviews of Geophysics* 39(2): 231–290.

Tomaszewski M, Cema G, and Ziembińska-Buczyńska A (2017) Influence of temperature and pH on the anammox process: A review and meta-analysis. *Chemosphere* 182: 203–214.

Valentine DL (2002) Biogeochemistry and microbial ecology of methane oxidation in anoxic environments: A review. *Antonie Van Leeuwenhoek* 81(1–4): 271–282.

Zhu J, Wang Q, Yuan M, Tan GYA, Sun F, Wang C, et al. (2016) Microbiology and potential applications of aerobic methane oxidation coupled to denitrification (AME-D) process: A review. *Water Research* 90: 203–215.

Gross and Net Production in Different Environments

Martin T Dokulil, University of Innsbruck, Mondsee, Austria

Preamble

This pamphlet intends to provide an overview on definitions, methodologies and results of gross and net production measurements as estimates for different ecosystems, and the globe. The content however, might be biased towards freshwater plankton production due to the long-term professional experience of the author.

Introduction

Measurements of photosynthetic rates and the estimation of ecosystem production have a long tradition. The early history has been elegantly summarized by Macfadyen (1948), the history of plankton productivity was described in detail by Barber and Hilting (2002).

Quantification of ecosystem metabolism was introduced by Juday (1940) and Riley (1944) followed by the foundation of ecological energetics through the classic paper by Lindeman (1942). His transfer efficiency concept between trophic levels of ecosystems was the nucleus which inspired the International Biological Program (IBP) for the decade 1964–74. The "biological basis of productivity" to better understand the dynamics of whole ecosystems was one of the essential objectives of the program. IBP boosted methods, their standardization, and measurements of production at the major trophic levels in land, fresh water and marine ecosystems. Studies were based on biomass, gross production, respiration, net production, and many other variables. The focus on production ecology induced by IBP lasted for about the next 50 years or so (see, e.g., Williams *et al.*, 2002).

The world-wide effort by IBP produced a number of mainly methodological "IBP Handbooks" such as, for example, Vollenweider (1969) and finally culminated in a series of "synthesis volumes" (e.g., Cooper, 1975; Le Cren and Lowe-McConnell, 1980; Westlake *et al.*, 1998). Very few scientists might have recognized during IBP that this decade was perhaps innovative in what is now called "Big data" and certainly was part of "Big Science," phrases which were already reflected by Weinberg (1961, 1967) as outlined and discussed by Aronova *et al.* (2010). These authors also emphasized IBP's key role in acceptability of long-term synoptic data collection. As a follow up of IBP, the Long Term Ecological Research Network (LTER) was initiated in1977. Meanwhile the network produced a series of LTER publications including various aspects of production (e.g., Arnott *et al.*, 2006) as well as standards for measuring primary production for a variety of ecosystems assembled by Fahey and Knapp (2007).

Photosynthesis and Respiration

The dominant autotrophic process in most ecosystems on the earth is photosynthesis. The process is separated into light-dependent reactions and light-independent reactions. In the light energy contained within photons oxidizes water to O_2, produces potential cellular energy in the forms of ATP and NADPH, and generates protons. In the dark reactions which are physically separated from the light-dependent reactions, ATP, NADPH, and protons produced in the light are used to reduce CO_2 to carbohydrates (for more details see, e.g., Falkowski and Raven, 2007; Raich *et al.*, 2014). These two net reactions together form the net reaction for plant photosynthesis:

$$CO_2 + H_2O \rightarrow (CH_2O) + O_2$$

Aerobic respiration generates useful cellular products from carbohydrates and oxygen. In short form:

$$(CH_2O) + O_2 + n\mathrm{ADP} + n\mathrm{Pi} \rightarrow CO_2 + H_2O + n\mathrm{ATP}$$

Although both oxygen evolution and carbon uptake are measures of photosynthesis, both variables originate from two physical separated reactions. Therefore clear statements are essential on what has been measured and which conversions were applied if any. As a recommendation measurements based on oxygen evolution estimate *photosynthetic rate* those based on carbon measure *carbon uptake* and shall be named accordingly.

Productivity or Production

The complex process of photosynthesis in photoautotrophic organisms usually creates new organic material from inorganic elements by harvesting light. This time dependent photosynthetic rate often is termed "primary productivity" or "primary production" (e.g., Falkowski and Raven, 2007). Both terms are often confused and used interchangeably (Thornton, 2012).

 doi:10.1016/B978-0-12-409548-9.10891-7

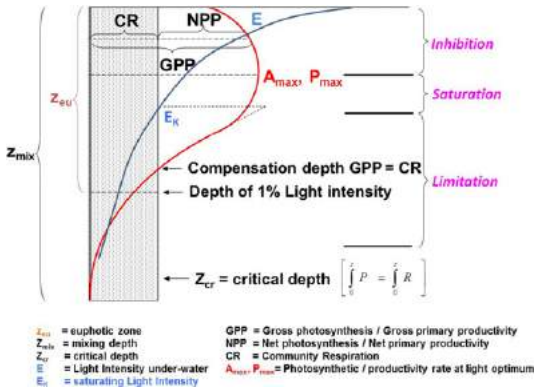


Fig. 1 Conceptual diagram of a schematic vertical depth profile graphically showing definitions of variables and parameters used in production ecology. For abbreviations refer to the legend and the text. Modified from Dokulil, M.T., Kaiblinger, C., (2009). Phytoplankton productivity. In: Likens, G.E. (ed.) Plankton of inland waters, Vol. 1, pp. 210–218. Oxford: Elsevier. https://www.elsevier.com/books/plankton-of-inland-waters/likens/978-0-12-381994-9.

Since there is no generally accepted definition of primary production terms (Underwood and Kromkamp, 1999) various authors have discussed the expressions "productivity" and "production" with the conclusion to abandon one or the other (Williams, 1993a). Until now these concepts are poorly differentiated and often synonymously used. The dilemma can be resolved maybe by adopting a definition from economics where productivity is a *rate* at which something is produced while production is the final product. Accordingly, primary productivity would refer rate per area or depth for a time less than a day, month or year depending on the organism in question. The ultimate end-product of the process is "primary production," the accumulation of organic material leading to growth. Accordingly, primary production will be a derived quantity integrated over area, depth and/or time, usually perhaps a year with units of mass or energy. Productivity can then also be regarded as equivalent to photosynthetic or production rate. Similar differentiation between these terms is not explicitly mentioned but inherent in, for example, Nátr and Lawlor (2005) summarizing photosynthetic plant productivity.

Definition of Major Parameters

The history and development defining production terms was nicely outlined in Williams (1993a). Some of the definitions explained below are graphically presented in **Fig. 1** for a marine/fresh-water vertical depth profile.

Euphotic zone (z_{eu}): Per common definition the zone between the water surface and the 1% light level (LI) where light supports photosynthesis.

Compensation depth (z_c): Determined by the physiology, acclimation or adaptation of the species, population or assemblage where gross photosynthesis is compensated by community respiration. Therefore it might or might not be equivalent or even identical to the euphotic depth as repeatedly assumed. The two terms strictly must be separated due to their different meaning.

Standing crop: All living organisms on a unit area. Equivalent to biomass above-ground.

Biomass (B): Total mass of all living (and dead) material in carbon units (e.g., g C m^{-2}), units of energy (e.g., kJ m^{-2}) or dry organic matter (e.g., tons per hectare) above and below ground.

Gross photosynthesis (GP): The rate of electron equivalents photochemically extracted from the oxidation of water. It resembles gross oxygen evolution rate if no respiration losses can be assumed.

Net photosynthesis (NP): Corresponds to gross photosynthesis minus all losses due to autotrophic respiration.

Photosynthetic efficiency (PE) = 100 × (incident radiation converted to NPP)/(total incident radiation); conversion, for example: 39 kJ per g C.

Photosynthetic quotient (PQ): The volume of oxygen released in photosynthesis as a proportion of the volume of carbon dioxide used in that process.

Primary productivity (PP): Refers to rate at which energy is accumulated by green plant in unit time in the form of organic substance that can be used as food. The rate at which radiant energy is stored by photosynthetic activity of green plants and algae in the form of organic substance is termed primary productivity, because it is the first and most basic form of energy stored in the ecosystem.

Gross primary production (GPP): Is the total amount of CO_2 that is fixed by the plant in photosynthesis or the sun's energy that is assimilated by *total photosynthesis*. It includes the organic matter used up in respiration during the measurement period. It is also called *total assimilation*. Units: g C m^{-2} $year^{-1}$, or kJ m^{-2} $year^{-1}$.

Ecosystem GPP: is then the sum for an entire lake, forest, field, biome, etc. Units: g C $year^{-1}$.

Community respiration (CR): is the amount of CO_2 that is lost from an organism or system from metabolic activity. Respiration can be further divided into components that reflect the source of the CO_2.

R_p = Respiration by plants.

R_h = Respiration by heterotrophs.

R_d = Respiration by decomposers (the microbes).

Net primary production (NPP): Is the net amount of primary production after the costs of plant respiration are included. Therefore, NPP = GPP − R.

Net ecosystem production (NEP): Is the net amount of primary production after the costs of respiration by plants, heterotrophs, and decomposers are all included. NEP = GPP − $(R_p + R_h + R_d)$. Randerson *et al.* (2002) suggested to change the definition of NEP and equate it to the rate of carbon accumulation while Lovett *et al.* (2006) argued to retain the original definition.

Net ecosystem exchange (NEE): Refers to net primary production minus carbon losses due to R_h (Kirschbaum *et al.*, 2001). NEE = NEP = NPP − R_h. Usually used in terrestrial PP.

Net biome exchange/net biome production (NBE, NPB): Refers to the change in carbon stocks after losses due to natural or anthropogenic disturbances. NBE = NEE − L_d or NBP = NEP − L_d with L_d losses by episodic disturbances (Roy and Saugier, 2001).

New production: according to Dugdale and Goering (1967) represents a primary production quantity in the oceans driven by nutrient inputs from outside the euphotic zone, usually nitrate (NO_3). In this sense, nitrogen fixation is new production since the nitrogen source originates from the atmosphere.

Recycled production: Is driven in the oceans by ammonium (NH_3) excreted by grazers within the euphotic zone.

Turnover: Is the ratio of standing crop or biomass to production (Standing Crop/Production) of the system. It is calculated by dividing standing crop or biomass (units of g m^{-2}) by production (units of g/m^{-2} year).

Production to biomass ratio (P:B): Is the amount of biomass replaced per unit time ($year^{-1}$).

Productivity versus irradiance curve and derived parameters are an approach widely used in marine and freshwaters as well as in various terrestrial ecosystems (Dokulil *et al.*, 2005; Dokulil and Kaiblinger, 2009).

Measure Primary Productivity and/or Production

Different techniques will produce slightly different rates of productivity (Bender *et al.*, 1987). As a result of the biases associated with each method no technique can provide "true" rates of primary productivity. Consequently no algorithm or model will estimate "real" production.

The overwhelming quantity of techniques for estimating primary productivity and their manifold modifications are beyond the scope of this article and will therefore not be discussed in detail. Many methodological summaries, reviews and outlines are available such as Newbould (1967), Milner and Hughes (1968), Vollenweider (1969), Cullen (2001), Fahey and Knapp (2007) to name a few. Moreover, instrumentation, methods and techniques for estimation of primary production are constantly evolving.

Three general approaches are possible which will not give the same answer (Marra, 2002):

(a) Measurements of increase in plant biomass—Largely used in terrestrial ecosystems; in marine and freshwater limited by non-discriminatory for algal assemblages; only the changes in chlorophyll-a can be taken as a proxy for autotrophic biomass.
(b) Yield measurements of productivity—The most variable approach to production (see discussion in Marra, 2002).
(c) Rate measurements of photosynthesis/productivity—Extensively used in innumerable alternatives, modifications and refinements.

Primary productivity is typically measured by the following main techniques:

1. *Light and dark bottle method*—Change in oxygen concentration within sealed containers, usually glass bottles (Gaardner and Gran, 1927).
2. *Radioactive tracer LD bottle method*—Incorporation of inorganic carbon as ^{14}C in the form of sodium bicarbonate into organic matter (Steemann-Nielsen, 1951, 1952).
3. Differential measurements of *stable isotopes of oxygen* (^{16}O, ^{18}O and ^{17}O) by, for example, Bender *et al.* (1987) or Luz and Barkan (2000).
4. *Chlorophyll concentration*—Based upon the relationship between chlorophyll concentration and photosynthesis at any given light intensity. If the assimilation ratio and the available light are known gross production can be estimated. Method first used in the sea but later applied to freshwater terrestrial ecosystems (Bot and Colijn, 1996; Gitelson *et al.*, 2006).
5. *Fluorescence kinetics*—Several modifications of measuring modes are in use (Jakob *et al.*, 2005), including passive, in vivo fluorescence and active fluorescence methods (Dokulil and Kaiblinger, 2009).
6. *Stable isotopes of carbon* (^{12}C and ^{13}C)—Measurements of respiration in the light allowing consequently calculation of gross photosynthetic rates (e.g., Carvalho and Eyre, 2012).
7. *Oxygen/argon ratios*—First used by Craig and Hayward (1987) and modified by Reuer *et al.* (2005).
8. *Carbon dioxide flux*—Most useful methods in terrestrial ecosystems for estimating both gross and net primary productivity. Carbon dioxide concentration in incoming and outgoing air is measured in the light and in the dark usually with an infrared gas analyzer (IRGA).
9. *Harvest analysis (standing crop method)*—Several modifications are widely used to estimate PP in terrestrial ecosystems. Vegetation is removed at maximum stand or at periodic intervals. Samples are dried to constant weight or determined as caloric

values. Results obtained are NPP for above ground biomass. Below ground biomass is difficult to estimate particularly in, for example, grass species or trees.

10. *Dimension analysis*—Modified harvest technique estimating standing crop and productivity from the measurement of light, diameter growth and age. Net annual production of wood, bark, leaves, twigs, roots and flower is calculated. All these information are used to calculate production of trees and other vegetation in a sample unit.
11. *Remote sensing*—Continuous monitoring of GPP and NPP possible via satellite imaging spectroradiometer data (Running *et al.*, 2004; Lee *et al.*, 2015). Another possibility is light detection and ranging (LIDAR) with pulsed laser light from aircrafts (Kotchenova *et al.*, 2004; Maselli *et al.*, 2013).

The most important flux of the terrestrial carbon cycle is gross primary production (GPP) of the vegetation. At ecosystem or landscape scales no direct measurement technique for GPP is available (Xiao *et al.*, 2014).

Several basic alternatives exist for marine and freshwater ecosystems (Vernet and Smith, 2007).

1. *Open water* changes in oxygen or carbon dioxide concentration. Diel oxygen techniques to measure lake metabolism are summarized by Staehr *et al.* (2010).
2. In situ enclosures—In containers, commonly glass bottles, which are suspended at in situ depth for a certain time.
3. *Simulated* in situ experiments—Samples from the euphotic zone are filled into containers as above but incubated on the ships deck simulating in situ conditions of light and temperature. A similar technique was adopted for the River Danube (Dokulil and Holst, 1990).
4. *Laboratory incubation* experiments—Productivity is estimated from the response of algae incubated at a range of artificial irradiances and at mean in situ temperature. The resulting photosynthetic versus irradiance behavior (P vs. E curve) is analyzed for certain parameters and finally modeled to represent potential in situ production (e.g., Kabas, 2004). A dilution method was developed by Moigis and Gocke (2003) to estimate primary production of phytoplankton in coastal waters.
5. *Non-invasive, indirect methods*—Besides the open water methods mentioned under point 1 above, these techniques are usually based on fluorescent properties and do not depend on incubation of small samples. Moreover, they have the advantage of being almost instantaneous and closely related to sampling rates of physico-chemical variables (Kaiblinger *et al.*, 2005; Kaiblinger and Dokulil, 2006a,b).

Comparison between different methods measuring marine PP and conversion factors are provided by Regausie-de-Gioux *et al.* (2014) claiming that the ^{18}O technique provides the best estimate of GPP.

Respiration

The "correct" approximation of respiration is of prime importance for the assessment of gross primary production. Respiration in the dark can be unequal to uptake of oxygen in the light due to photorespiration (Peterhansel *et al.*, 2010). Common practice assumed that algal respiration in the oceans was unimportant until a model developed by Langdon (1993) indicated how significant respiration can be and how important is to consider the composition of the algal assemblage. Oceanic respiration has meanwhile identified as a major component of carbon flux in the biosphere or, in other words, whether oceans are net sources or sinks of carbon (Del Giorgio and Duarte, 2002). For a detailed discussion on respiration in aquatic ecosystems refer to Del Giorgio and Williams (2005a,b). Respiration by terrestrial vegetation and net primary production is outlined by Amthor and Baldocchi (2001) and Raich *et al.* (2014). Better separation of autotrophic and heterotrophic respiration is certainly still required (Trumbore, 2006) as has, for example, shown for a tropical forest ecosystem by Chambers *et al.* (2004).

At best respiration can be determined as community respiration (CR) during routine field operations in aquatic ecosystems. Consequently GPP will be overestimated since some of CR will certainly be non-autotrophic. Similarly in situ measurements of respiration in terrestrial vegetation remain largely uncertain which is particularly true for soil respiration affecting the approximation of below-ground GPP. Increasing CO_2 concentrations and warming due to global change may further affect respiration in the near future. These and other aspects of aquatic and terrestrial respiration are detailed in Del Giorgio and Williams (2005a,b). An analysis of ecosystem and background respiration based on free-water high frequency measurements from 25 lakes around the globe revealed GPP and CR positively related to total phosphorus. Respiration was linked to GPP tightly to more weekly depending on trophic level (Salomon *et al.*, 2013).

Although the dark uptake of ^{14}C does not measure respiration, several attempts were made in the past to obtain respiration from carbon uptake experiments (Steemann-Nielsen and Hansen, 1959). More recently, Williams and Lefèvre (2008) critically analyzed assessments of respiration from dark ^{14}C incubations based on the procedure proposed by Marra and Barber (2004). The authors came to the conclusion that the procedure does not exactly measure respiration but might come close when P/R ratios are high. Reviewing net and gross productivity, Marra (2009) concludes that estimates of plankton respiration seem reasonable when compared to the ^{18}O method.

What Does the ^{14}C Method Measure?

The radiocarbon method is one of the most widely used techniques in aquatic systems. Even after more than 50 years of intensive investigations there is still no common consensus as to what the method is measuring (Dring and Jewson, 1982). Measurements

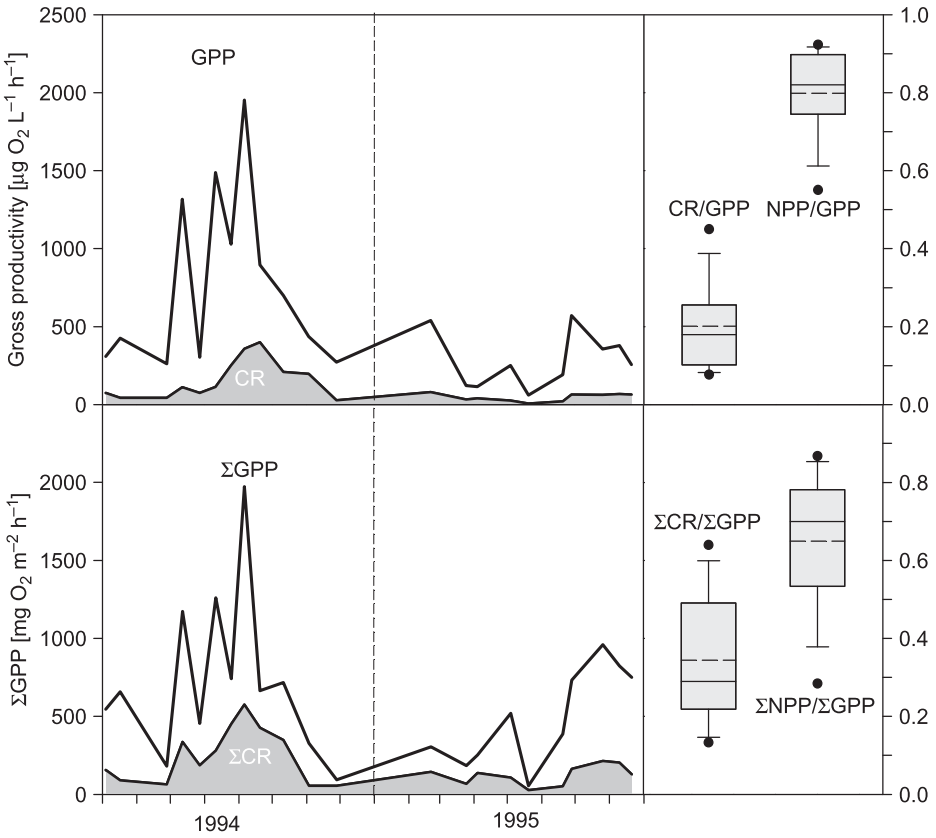


Fig. 2 Seasonal dynamics for the years 1994 and 1995 for gross primary productivity (GPP) and community respiration (CR) at light saturation (upper panel) and as integral (ΣGPP) and (ΣCR) for the water column (lower panel) in the urban lake "Alte Donau" in Vienna, Austria. Box whisker plots for the ratios CR/GPP and NPP/GPP for both panels are added. Continuous line in the box is the median, dashed line is the mean.

and modeled data suggest an imbalance between CO_2 and O_2 dynamics (Williams, 1993b). If internal CO_2 from respiration is taken as a source, then ^{14}C uptake must be smaller than ^{18}O uptake which will then approximate gross photosynthesis while carbon uptake is closer to net productivity. As a consequence, the ^{14}C method approximates gross productivity over very short periods of time when respiration is low and productivity is much higher ($P \gg R$). Whenever production is greater or equal to respiration, ^{14}C uptake approaches net productivity (Marra, 2002; Vernet and Smith, 2007).

Gross and Net Production in Different Environments

A central process regulating the structure and functioning of terrestrial ecosystems is gross primary productivity (GPP). Respiratory processes, supporting plant embolism, loose about half of assimilates derived from GPP back to the atmosphere. What remains is used for new living biomass. Plants also transpire large quantities of water during photosynthesis thus influencing both water and carbon global cycles. Changes in these complex interactions impact local, regional, and global climate. All NPP will finally be released back to the atmosphere as CO_2 because of utilization by heterotrophic organisms.

Aquatic ecosystems and particularly the oceans largely influence the global carbon cycle. The 1–2 Giga tons of autotrophic biomass in the oceans seem small compared to the 600–1000 Giga tons of total terrestrial plant biomass. The importance of the oceans becomes clear when turnover times are compared. Ocean turnover is 0.02–0.06 years while it is 9 to >20 years in terrestrial vegetation (Falkowski and Raven, 2007).

Comparison of different ecosystems can only be achieved if measurements are converted to a common unit. Since terrestrial primary production is usually express per unit area, values from other environments, particularly from aquatic ecosystems need to be converted by integrating vertical profiles over depth.

The time sequence of gross productivity at light optimum and the respective community respiration (CR) are compared to the integral production per square meter (ΣGPP) and ΣCR in an urban lake (**Fig. 2**). Seasonal changes are not very different in their appearance but differ largely in extent and relative importance. Relation of parameters is depicted at the right side of each graph as box-Whisker plots. Contribution of CR to GPP is smaller and varies less than ΣCR/ΣGPP. Accordingly NPP has a much larger share of GPP than does ΣNPP on ΣGPP (**Table 1**).

Table 1 Mean, maximum, minimum, median, 25% and 75% percentiles for the Box plots in Fig. 2

	Mean	*Max*	*Min*	*Median*	P *25%*	P *75%*
CR/GPP	0.20	0.45	0.08	0.18	0.10	0.26
NPP/GPP	0.80	0.92	0.55	0.82	0.75	0.90
ΣCR/ΣGPP	0.34	0.65	0.13	0.29	0.22	0.49
ΣNPP/ΣGPP	0.65	0.87	0.27	0.70	0.53	0.78

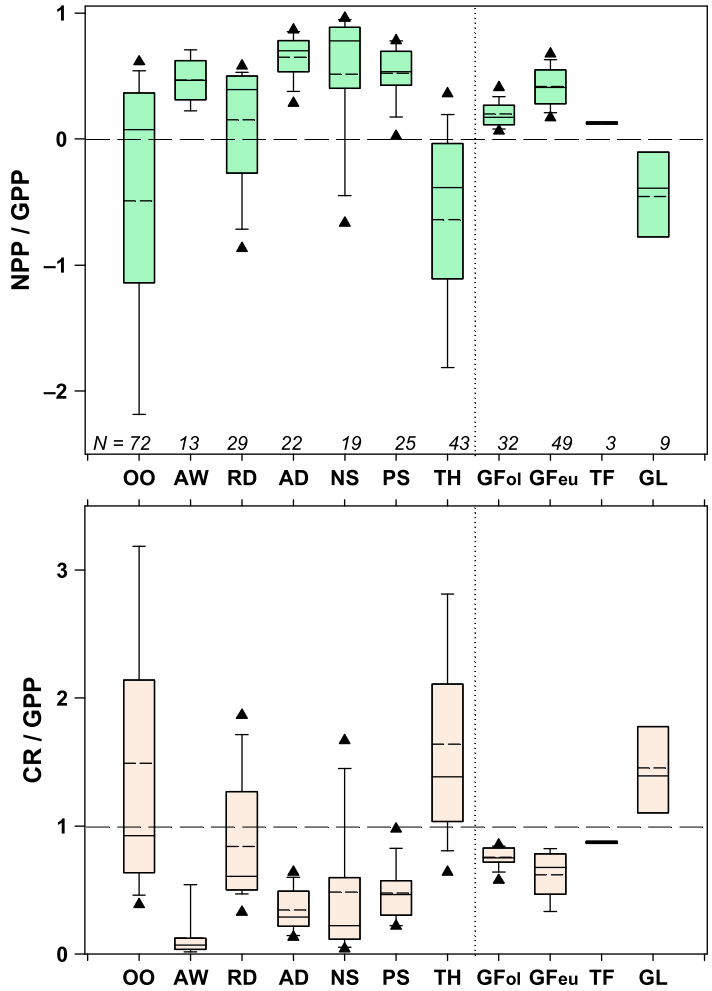


Fig. 3 Relation of NPP to GPP (upper panel) and CR to GPP (lower panel) for different ecosystems as Box–Whisker plots. Continuous line in box is the median, dashed line is the mean. The dashed line in the upper panel at 0 indicates NPP = GPP, the dashed line in the lower panel at 1 indicates CR = GPP. OO—Oligotrophic Ocean from López-Urrutia *et al.* (2006) and Williams (1998), AW—River impoundment "Altenwörth," Austria from Holst and Dokulil (1987), ED—River Danube, data compilation from Dokulil (2014), AD—urban lake "Alte Donau" from Roschitz (1996), NS—Neusiedler See, shallow lake at the Austro-Hungarian border from Dokulil (pers. data compilation), PS—Parakrama Samudra, tropical reservoir in Sri Lanka from Dokulil *et al.* (1983), TH—Tai Hu, sub-tropical shallow lake in east-central China from Liu *et al.* (2011), GF_{ol} and GF_{eu}—Global Forest nutrient poor and nutrient rich, both from Fernández-Martínez *et al.* (2014), TF—temperate forest from Sun *et al.* (2014) and Grassland from Albergel *et al.* (2010). Number of data for each system is indicated in the upper panel. A dotted vertical line separates aquatic from terrestrial ecosystems.

Relations of NPP to GPP and CR to GPP are compared for different ecosystems on a unit area basis in **Fig. 3**. Phytoplankton production data from the open, oligotrophic ocean (OO) indicate high variability in both NPP/GPP and CR/GPP. Negative relations are much more variable than positive ones indicated by the positive median in both cases meaning that over 50% of data are positive (**Table 2**). The debate over the metabolic state of the oligotrophic ocean (if auto- or heterotrophic) is still an open question. Results largely depend on the variability in respiration and certain assumptions over scaling (Serret *et al.*, 2015). The

Table 2 Mean, maximum, minimum, median, 25% and 75% percentiles for the Box plots in Fig. 3

	Mean	*Maximum*	*Minimum*	*Median*	P *25%*	P *75%*
NPP/GPP						
OO	−0.49	0.67	−5.54	0.08	−1.14	0.37
AW	0.47	0.71	0.21	0.47	0.31	0.62
RD	0.15	0.63	−0.93	0.39	−0.27	0.50
AD	0.65	0.87	0.27	0.70	0.53	0.78
NS	0.52	0.96	−0.67	0.78	0.40	0.89
PS	0.52	0.78	0.01	0.54	0.43	0.70
TH	−0.64	0.59	−3.16	−0.38	−1.11	−0.04
GFol	0.20	0.44	0.05	0.17	0.12	0.27
GFeu	0.42	0.68	0.11	0.41	0.28	0.55
TF	0.13	0.14	0.12	0.13	0.12	0.14
GL	−0.46	0.03	−1.08	−0.39	−0.78	−0.10
CR/GPP						
OO	1.49	6.54	0.33	0.92	0.64	2.14
AW	0.13	0.79	0.01	0.07	0.04	0.12
RD	0.84	1.93	0.19	0.61	0.50	1.27
AD	0.34	0.65	0.13	0.29	0.22	0.49
NS	0.49	1.67	0.04	0.22	0.12	0.60
PS	0.48	0.99	0.22	0.47	0.30	0.57
TH	1.64	4.16	0.41	1.38	1.04	2.11
Gfol	0.76	0.85	0.56	0.75	0.72	0.83
GFeu	0.62	0.85	0.33	0.68	0.47	0.78
TF	0.87	0.88	0.87	0.87	0.87	0.88
GL	1.45	2.08	0.97	1.39	1.10	1.78

Abbreviations for ecosystems as in Fig. 3.

authors conclude that oligotrophic oceans are functionally diverse and that variability in respiration needs to be considered before the metabolic state of an ocean area can be defined.

The six specific examples on plankton production from inland waters in Fig. 3 include a river impoundment (AW), the River Danube (RD), a side arm which is now an urban lake (AS), a shallow lake (NS), a tropical reservoir (PS) and a shallow subtropical lake (TH). Five of the examples have largely positive NPP/GPP relations and their medians are all higher than the ocean or the terrestrial examples indicating a high positive net production. Negative ratios are encountered in NS and, to a greater extent, in RD. Interestingly the highly eutrophic shallow system TH is largely heterotrophic possibly due to the shallowness and turbidity. Both mean and median are negative (Table 2). Primary production for freshwater biocenoses other than phytoplankton and in different climatic regions is comparatively assessed in Dokulil (2009).

Terrestrial ecosystem production ratios are provided for four environments, nutrient poor (GF_{ol}) and nutrient rich (GF_{eu}) global forests, temperate forest (TF) and grassland (GL). The forest examples all show les NPP contribution to GPP compared to the inland water examples largely due to a greater share of community respiration. Evident is also that nutrient rich forests have higher NPP/GPP ratios than nutrient poor forests and both are higher than the few ratios available for temperate forests. For more and detailed data refer to, for example, Vogt (1991) and Wehr *et al.* (2016). Grassland vegetation, at least in France, has high CR/GPP ratios and hence all negative NPP/GPP values (Table 2). According to Ma *et al.* (2015) gross primary production was likely overestimated in forest ecosystems because coverage rate has not been taken into account for the past 30 years. Understanding terrestrial carbon fluxes in response to climate change largely depends on understanding and modeling gross primary production (GPP) in forest ecosystems (Chen *et al.*, 2013).

Global Production

Adequate estimation of primary production at global level necessitates an enormous quantity of basic data. The acquisition of information at reasonably high temporal and spatial resolution can be obtained by remote sensing using aircraft or satellite-based technologies. Combined with testing and improving models and algorithms by measurements acquired at ground level, his approach follows ecological principles able to detect even regional disturbances and anthropogenic impacts (Field *et al.*, 1995). Based on these rationales, Field *et al.* (1998) estimated global NPP for terrestrial and marine primary producers at 104.9 Pg (10.5×10^{10} metric tons) C per year. Contributions from land and oceans were roughly equal, 54% and 46% respectively but showed marked heterogeneity.

Turner *et al.* (2005, 2006) report a range of NPP from 80 g C m^{-2} $year^{-1}$ at an arctic tundra site to 550 g C m^{-2} $year^{-1}$ at a temperate deciduous forest site and strong seasonality in GPP. Ground validation and evaluation from satellite agreed closely at temperate deciduous forest, arctic tundra, and boreal forest sites in both NPP and GPP.

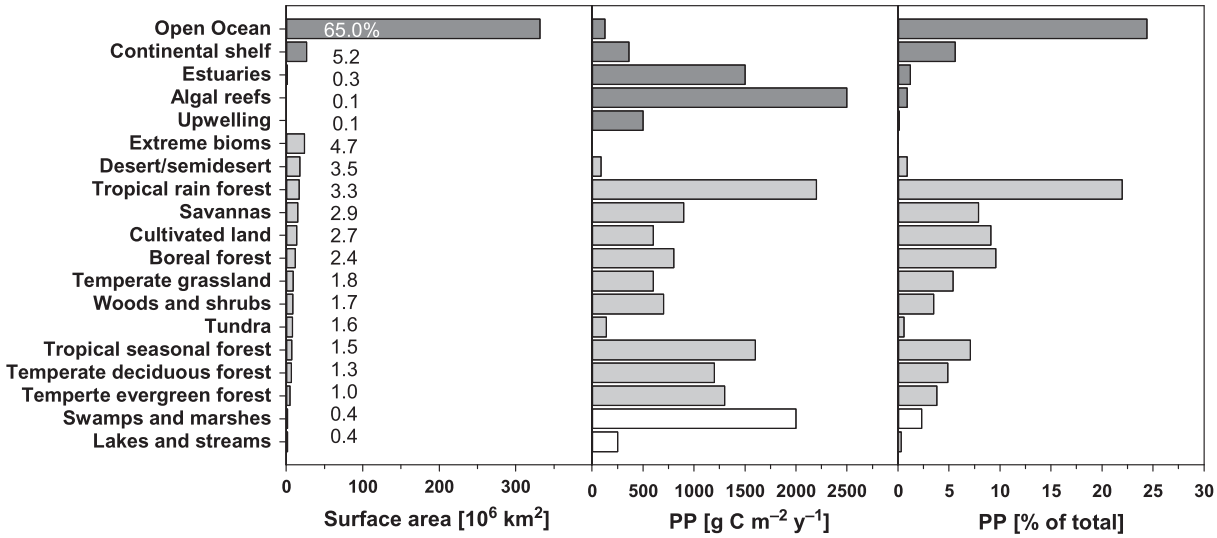


Fig. 4 Surface area for different ecosystems of the earth as 10^6 km^2 with percent contribution added as values (left panel), primary production in g C m^{-2} year^{-1} for these ecosystems (middle panel) and primary production as percent of total world production (right panel). Data modified from Whittaker, R.H., Likens, G.E., 1973. Primary production: The biosphere and man. Human Ecology 1, 357–369.

Table 3 Global NPP, GPP and the NPP/GPP ratio from 2000 to 2003

Year	*GPP*	*NPP*	*NPP/GPP ratio*
2000	108.00	55.92	0.518
2001	110.33	57.57	0.522
2002	107.40	55.36	0.515
2003	107.09	54.63	0.510

Data are in Pg C year^{-1}.

Primary productivity varies widely among different ecosystems of the world. Estuaries, coral reefs, swamps, marshes, tropical and temperate forests are the most productive systems per unit area (**Fig. 4**). The highest productive systems in terms of global NPP are open oceans because of their size. Tropical rain and seasonal forests, and savannas add a high proportion since they are very productive. The contribution of estuaries and coral reefs to global production remains small because of their limited surface area (**Fig. 4**). Similarly, lakes, rivers and streams are of minor importance for global production but are essential for the survival of the world's population.

Improvement and refinements in instrumentation, observational coverage and modeling continuously improved global production estimates over the years. Global assessments are influenced however by the instrumentation and algorithm used but variation between years and within years largely depends on environmental and climatic unpredictability.

Calculations by Zhao *et al.* (2005) using MODIS (moderate resolution imaging spectroradiometer) produced total global GPP and NPP of 109.29 and 56.02 Pg C year^{-1}, respectively for the years 2001–2003. For the same years, global GPP and NPP and their ratio showed minor fluctuations shown in **Table 1** adapted from Zhang *et al.* (2009). Highest global GPP and NPP/GPP ratio occurred in 2001 and the lowest in 2003 (**Table 3**).

Average annual global GPP was calculated as 110.5 $\pm$ 21.3 Pg C for 2000–2003 by to Yuan *et al.* (2010). The humid tropics of Amazonia, Central Africa and South-east Asia had the highest annual production of over 2000 g C m^{-2} since temperature and moisture are optimal for photosynthesis is. Intermediate GPP of 1000–1400 g C m^{-2} occur in temperate regions. Cold and arid regions, where either temperature or precipitation are limiting factors, have lowest GPP of $<$400 g C m^{-2} (Yuan *et al.*, 2010).

Average total global GPP estimated for the decade 2000–2010 using eight biome models was 117 $\pm$ 13 Pg C year^{-1} (Chen *et al.*, 2017). According to the authors Tropical Latin America contributes most to global GPP. The Asian regions are largely linked to the global GPP trend while the Northern Hemisphere dominates seasonal variations in global GPP. The inter-annual variability of global GPP seems likely controlled by the Oceania region.

The global GPP estimates for the period 2000–2016 range from 121.60 to 129.42 Pg C year^{-1} with an increasing rate of approximately 0.39 Pg C year^{-1} (Zhang *et al.*, 2017). Details for each year on continental and global GPP are given in **Table 4** modified from Zhang *et al.* (2017). Inter-annual variation is relatively small on all continents. Maximum and minimum GPP occurs in different years across the globe indicating regional climatic differences. Asia, and Europe had their maxima in 2016 with a clear upward trend on both continents ($r^2 > 0.8$). North America shift upwards with large ups and downs ($r^2 = 0.6$). No trend is

Table 4 Continental and global total GPP for 2000–2016

Year	*Africa*	*Asia*	*Europe*	*North America*	*South America*	*Oceania*	*Global total*
2000	27.43	*29.82*	*8.50*	16.69	33.72	5.44	*121.60*
2001	27.69	30.52	8.64	17.22	33.43	4.96	122.46
2002	27.62	31.64	8.82	*16.46*	34.05	*4.17*	122.76
2003	27.57	31.18	8.54	17.33	34.02	4.29	122.93
2004	27.84	30.86	8.83	17.81	34.61	4.79	124.74
2005	*26.93*	31.00	8.90	17.88	34.27	4.43	123.41
2006	28.62	31.50	8.86	17.37	**34.92**	4.74	126.02
2007	28.20	31.76	9.04	18.03	33.28	4.39	124.70
2008	28.27	31.39	9.1	17.47	33.07	4.52	123.82
2009	28.80	31.82	9.07	17.23	33.92	4.54	125.38
2010	28.30	31.80	8.88	18.14	33.46	5.32	125.90
2011	28.54	32.21	9.43	17.26	33.52	**5.82**	126.79
2012	28.23	32.26	9.11	18.19	32.90	4.98	125.66
2013	29.09	33.00	9.48	17.89	*32.89*	4.48	126.81
2014	**29.41**	33.13	9.67	18.08	33.33	4.73	128.35
2015	28.02	**34.42**	9.79	19.05	33.73	4.42	**129.42**
2016	27.19	33.43	**10.08**	**19.10**	33.10	4.79	127.70
Avg	28.10	31.87	9.10	17.72	33.66	4.75	125.20
Max	**29.41**	**34.42**	**10.08**	**19.10**	**34.92**	**5.82**	**129.42**
Min	**26.93**	**29.82**	**8.50**	**16.46**	**32.89**	**4.17**	**121.60**

Units are in Pg C $year^{-1}$.
Bold numbers in the main table indicate maximum values and italics numbers indicate minimum values.

obvious for Africa and Oceania, and production in South America is declining. As a result the global total GPP is on the rise ($r^2 = 0.85$). In this context, Campbell *et al.* (2017) report a 31% ± 5% growth of GPP during the 20th century. Tropical regions have high annual GPP peaking around the equator. Mean maximum daily GPP creates another peak around 50°N (for both see Fig. 4B and D in Zhang *et al.*, 2017).

Analyzing the average global carbon cascade from GPP to net biome production (NBP) for land and ocean, Woodward (2007) concludes that the NBP of 2–3 Pg C produced per year accounts for about 20%–30% of anthropogenic emissions.

Perspective

Climatic signals such as temperature, light availability and humidity are the major factors controlling GPP and consequently NPP. Further important variables are soil fertility, species composition, and age of the vegetation. Since climate is of prime importance for GPP, climate change is expected to seriously impact global patterns of plant production. Enhanced plant growth is expected in colder climates, declining GPP in regions where water losses increase due to evaporation and perhaps dramatic changes in vegetation because of extreme events such as heavy precipitation, fires, storms and hurricanes. Negative feedback to climate change might thus be provided.

Various ways of anthropogenic intervention in global primary production to mitigate climate change are often discussed such as planting trees to reduce terrestrial carbon emissions or iron fertilization of the ocean to stimulate CO_2-uptake by phytoplankton (Woodward, 2007). These and other interventions into global primary production may have short lived positive effects but are unlikely to produce long term solutions to the problem (Woodward *et al.*, 2009).

See also: Aquatic Ecology: Abundance Biomass Comparison Method. General Ecology: Biomass. Global Change Ecology: Sustainable Cropping Systems. Terrestrial and Landscape Ecology: Agroforestry

References

Albergel, C., Calvet, J.-C., Gibelin, S.-L., Lafont, S., Roujean, J.-L., Berne, C., Traullé, O., Fritz, N., 2010. Observed and modelled ecosystem respiration and gross primary production of a grassland in southwestern France. Biogeosciences 7, 1657–1668. doi:10.5194/bg-7-1657-2010.

Amthor, J.S., Baldocchi, D.D., 2001. Terrestrial higher plant respiration and net primary production. In: Roy, J., Saugier, B., Mooney, H.A. (Eds.), Terrestrial global productivity. Oxford: Academic Press, pp. 33–59.

Arnott, S.E., Magnuson, J.J., Dodson, S.I., Colby, A.C.C., 2006. Lakes as islands: Biodiversity, invasion, and extinction. In: Magnuson, J.J., Kratz, T.K., Benson, B.J. (Eds.), Long-term dynamics of lakes in the landscape. Long-term ecological research on north temperate lakes.Long-term ecological research network series. New York: Oxford University Press, pp. 67–88.

Aronova, E., Baker, K.S., Oreskes, N., 2010. Big science and big data in biology: From the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) network, 1957–present. Historical Studies in the Natural Sciences 40, 183–224.
Barber, R.T., Hilting, A.K., 2002. History of the study of plankton productivity. In: Williams, P.J.l.B., Thomas, D.N., Reynolds, C.S. (Eds.), Phytoplankton productivity. Carbon assimilation in marine and freshwater ecosystems. Oxford: Blackwell Science, pp. 16–43.
Bender, M., *et al*., 1987. A comparison of 4 methods for determining planktonic community production. Limnology and Oceanography 32, 1085–1098.
Bot, P.V.M., Colijn, F., 1996. A method for estimating primary production from chlorophyll concentrations with results showing trends in the Irish Sea and the Dutch coastal zone. ICES Journal of Marine Science 53, 945–950.
Campbell, J.E., Berry, J.A., Seibt, U., Smith, S.J., Montzka, S.A., Launois, T., Belviso, S., Bopp, L., Laine, M., 2017. Large historical growth in global terrestrial gross primary production. Nature 544, 84–87. doi:10.1038/nature22030.
Carvalho, M.C., Eyre, B.D., 2012. Measurement of planktonic CO_2 respiration in the light. Limnology and Oceanography: Methods 10, 167–178.
Chambers, J.Q., Tribuzy, E.S., Toledo, L.C., 2004. Respiration from a tropical forest ecosystem: Partitioning of sources and low carbon use efficiency. Ecological Applications 14, S72–S88.
Chen, G., Yang, Y., Robinson, D., 2013. Allocation of gross primary production in forest ecosystems: Allometric constraints and environmental responses. New Phytologist 200, 1176–1186.
Chen, M., Rafique, R., Asrar, G.R., *et al*., 2017. Regional contribution to variability and trends of global gross primary productivity. Environmental Research Letters 12.105005
Cooper, J.P. (Ed.), 1975. Photosynthesis and productivity in different environments,International biological programme, vol. 3. Cambridge: Cambridge University Press, p. 715.
Craig, H., Hayward, T., 1987. Oxygen supersaturations in the ocean: Biological vs. physical contributions. Science 235, 199–202.
Cullen, J.J., 2001. Primary production methods. In: Steele, J.H., Turekian, K.K., Thorpe, S.A. (Eds.), Encyclopedia of ocean science, 2nd edn. New York: Academic Press, pp. 2277–2284.
Del Giorgio, P.A., Duarte, C.M., 2002. Respiration in the open ocean. Nature 420, 379–384.
Del Giorgio, P.A., Williams, P.J.l.B. (Eds.), 2005a. Respiration in aquatic environments. Oxford: University Press, p. 315.
Del Giorgio, P.A., Williams, P.J.l.B., 2005b. The global significance of respiration in aquatic ecosystems: From single cells to the biosphere. In: del Giorgio, P.A., Williams, P. J.l.B. (Eds.), Respiration in aquatic environments. Oxford: University Press, pp. 267–303.
Dokulil, M.T., 2009. In: Likens, G.E. (Ed.), Comparative primary production,Plankton of inland waters, vol. 1. Oxford: Elsevier, pp. 130–137. https://www.elsevier.com/books/plankton-of-inland-waters/likens/978-0-12-381994-9
Dokulil, M.T., 2014. Potamoplankton and primary productivity in the River Danube. Hydrobiologia 729, 209–227. doi:10.1007/s10750-013-1589-3.
Dokulil, M.T., Holst, I., 1990. Methods of biological sampling. Phytoplankton—photosynthesis. In: Humpesch, U.H., Elliott, J.M. (Eds.), Methods of biological sampling in a large deep river—The Danube in Austria. Wasser und Abwasser 2/90., pp. 17–23.
Dokulil, M.T., Kaiblinger, C., 2009. Phytoplankton productivity. In: Likens, G.E. (Ed.), Plankton of inland waters,vol. 1. Oxford: Elsevier, pp. 210–218. https://www.elsevier.com/books/plankton-of-inland-waters/likens/978-0-12-381994-9
Dokulil, M., Bauer, K., Silva, I., 1983. An assessment of the phytoplankton biomass and primary productivity of Parakrama Samudra, a shallow manmade lake in Sri Lanka. In: Schiemer, F. (Ed.), Limnology of Parakrama Samudra—Sri Lanka. A case study of an ancient man-made lake in the tropicsDevelopments in hydrobiology 12. The Netherlands: Springer, pp. 49–76.
Dokulil, M.T., Teubner, K., Kaiblinger, C., 2005. Produktivität aquatischer Systeme. Primärproduktion (autotrophe Produktion). In: Steinberg, C., Calmano, W., Klapper, H., Wilken (Hg), R.-D. (Eds.), Handbuch angewandte Limnologie, IV-9.2, 21. Erg. Lfg, 4/05, ecomed, Landsberg., pp. 1–30.
Dring, M.J., Jewson, F.H., 1982. What does ^{14}C uptake by phytoplankton really measure? A theoretical modelling approach. Proceeding of the Royal Society London B 214, 351–368.
Dugdale, R.C., Goering, J.J., 1967. Uptake of new and regenerated forms of nitrogen in primary productivity. Limnology Oceanography 12, 196–206.
Principles and standards for measuring primary production. In: Fahey, T.J., Knapp, A.K. (Eds.), In: Long-term ecological research network series. Oxford, New York: Oxford University Press, p. 268.
Falkowski, P.G., Raven, J.A., 2007. Aquatic photosynthesis, 2nd edn Princeton, New Jersey: Princeton University Press, p. 488.
Fernández-Martínez, M., Vicca, S., Janssens, I.A., Sardans, J., Luyssaert, S., Campioli, M., Chapin III, F.S., Ciais, P., Malhi, Y., Obersteiner, M., Papale, D., Piao, S.L., Reichstein, M., Rodà, F., Peñuelas, J., 2014. Nutrient availability as the key regulator of global forest carbon balance. Nature Climate Change 4, 471–476.
Field, C.B., Randerson, J.T., Malmström, C.M., 1995. Global net primary production: Combining ecology and remote sensing. Remote Sensing of Environment 51, 74–88.
Field, C.B., Behrenfeld, M.J., Tanderson, J.T., Falkowski, P., 1998. Primary production of the biosphere: Integrating terrestrial and oceanic components. Science 281, 237–240.
Gaardner, T., Gran, H.H., 1927. Investigations of the production of plankton in the Oslofjord. Rapport Conseil Exploration Mer 42, 3–48.
Gitelson, A.A., Viña, A., Verma, S., Rundquist, D.C., Arkebauer, T.J., Keydan, G.P., Leavitt, B., Ciganda, V., Burba, G.G., Suyker, A.E., 2006. Relationship between gross primary production and chlorophyll content in crops: Implications for the synoptic monitoring of vegetation productivity. Journal of Geophysical Research 111, D08S11. doi:10.1029/2005JD006017.http://digitalcommons.unl.edu/natrespapers/28
Holst, I., Dokulil, M., 1987. Die steuernden Faktoren der planktischen Primärproduktion im Stauraum Altenwörth an der Donau in Österreich. 26. Arbeitstagung der IAD, Passau 1987, BA Gewässerkunde Koblenz, pp. 133–137.
Jakob, T., Schreiber, U., Kirchesch, V., Langner, U., Wilhelm, C., 2005. Estimation of chlorophyll content and daily primary production of the major algal groups by means of multiwavelength-excitation PAM chlorophyll fluorometry: Performance and methodological limits. Photosynthesis Research 83, 343–361.
Juday, C., 1940. The annual energy budget of an inland lake. Ecology 21, 438–450.
Kabas, W., 2004. Die Veränderungen der Primärproduktion in der Alten Donau in den Jahren 1995–2002. Mit einem Methodenvergleich. Ph.D. Thesis, Univ. Vienna, pp. 86.
Kaiblinger, C., Dokulil, M.T., 2006a. Application of fast repetition rate fluorometry to phytoplankton photosynthetic parameters in freshwaters. Photosynthesis Research 88, 19–30.
Kaiblinger, C., Dokulil, M.T., 2006b. Saisonale Unterschiede der photosynthetischen Parameter des Phytoplanktons in Seen: Kurzzeitmessungen mit aktiver Fluoreszenz (FRRF-Technik). Dt. Ges. Limnol. (DGL), Tagungsbericht 2005, Karlsruhe, Werder 2006, pp. 179–182.
Kaiblinger, C., Teubner, K., Dokulil, M.T., 2005. Comparative assessment of phytoplankton photo-synthesis using conventional ^{14}C-determination and Fast Repetition Rate Fluorometry in freshwaters. Verhandlungen Internationale Vereinigung Limnologoe 29, 254–256.
Kirschbaum, M.U.F., Eamus, D., Gifford, R.M., Roxburgh, S.H., Sands, P.J., 2001. In: Definitions of some ecological terms commonly used in carbon accounting.NEE workshop proceedings, 2001, p. 5.
Kotchenova, S.Y., Song, X., Shabanov, N.V., Potter, C.S., Knyazikhin, Y., Myneni, R.B., 2004. Lidar remote sensing for modeling gross primary production of deciduous forests. Remote Sensing of Environment 92, 158–172.
Langdon, C., 1993. The significance of respiration in production measurements based on oxygen. ICES Marine Science Symposia 197, 69–78.
Le Cren, E.D., Lowe-McConnell, R.H. (Eds.), 1980. The functioning of freshwater ecosystems. Cambridge: Cambridge University Press, p. 576.
Lee, Z., Marra, J., Perry, J.M., Kahru, M., 2015. Estimating oceanic primary productivity from ocean color remote sensing: A strategic assessment. Journal of Marine Systems 149, 50–59.
Lindeman, R.L., 1942. The trophic-dynamic aspect of ecology. Ecology 23, 399–417.
Liu, X., Wu, Q., Chen, Y., Dokulil, M.T., 2011. Imbalance of plankton community metabolism in eutrophic Lake Taihu, China. Journal of Great Lakes Research 37, 650–655.
López-Urrutia, A., San Martin, E., Harris, R.P., Irigoien, X., 2006. Scaling the metabolic balance of the oceans. PNAS 103, 8739–8744.
Lovett, G.M., Cole, J.J., Pace, M.L., 2006. Is net ecosystem production equal to ecosystem carbon accumulation? Ecosystems 9, 1–4.

Luz, B., Barkan, E., 2000. Assessment of oceanic productivity with the triple-isotope composition of dissolved oxygen. Science 288, 2028–2031.
Ma, J., Yan, X., Dong, W., Chou, J., 2015. Gross primary production of global forest ecosystems has been overestimated. Scientific Reports 5, e10820.
Macfadyen, A., 1948. The meaning of productivity in biological systems. Journal Animal Ecology 17, 75–80.
Marra, J., 2002. Approaches to the measurement of plankton production. In: Williams, P.J.I.B., Thomas, D.N., Reynolds, C.S. (Eds.), Phytoplankton productivity. Carbon assimilation in marine and freshwater ecosystems. Oxford: Blackwell Science, pp. 78–108.
Marra, J., 2009. Net and gross productivity: Weighing in with ^{14}C. Aquatic Microbial Ecology 56, 123–131.
Marra, J., Barber, R.T., 2004. Phytoplankton and heterotrophic respiration in the surface layer of the ocean. Geophysical Research Letters. LO9314
Maselli, F., Mari, R., Chiesi, M., 2013. Use of lidar data to simulate forest net primary production. International Journal of Remote Sensing 34, 2487–2901.
Milner, C., Hughes, R.E., 1968. Methods for the measurement of the primary production of grassland. In: IBP handbook no. 6., London: Blackwell Scientific Publishing, Scanned and converted to text file by P. Sprott, LTER Network Office 2000. www.coweeta.uga.edu/
Moigis, A., Gocke, K., 2003. Primary production of phytoplankton estimated by means of the dilution method in coastal waters. Journal of Plankton Research 25, 1291–1300.
Nátr, L., Lawlor, D.W., 2005. Photosynthetic plant productivity. In: Pessarakli, M. (Ed.), Handbook of photosynthesis, 2nd edn Boca Raton: CRC Press, pp. 501–524.
Newbould, P.J., 1967. Methods for estimating the primary production of forests. In: IBP handbook no. 2, 2nd edn Oxford, Edinburgh: Blackwell Scientific Publishing. Scanned and converted to a text file by P. Sprott LTER Network Office 2000.https://coweeta.uga.edu/
Peterhansel, C., Horst, I., Niessen, M., Blume, C., Kebeish, R., Kürkcüoglu, S., Kreuzaler, F., 2010. Photorespiration. In: The Arabidopsis book8. e0130 doi:10.1199/tab.0130.
Raich, J.W., Lambers, H., Oliver, D.J., 2014. In: Holland, H.D., Turekian, K.K. (Eds.), Respiration in terrestrial ecosystems, Treatise on geochemistry, 2nd edn, vol. 10. Oxford: Elsevier, pp. 613–649.
Randerson, J.T., Chapin, F.S., Harden, J.W., Neff, J.C., Harmon, M.E., 2002. Net ecosystem production: A comprehensive measure of net carbon accumulation by ecosystems. Ecological Applications 12, 937–947.
Regausie-de-Gioux, A., Lasternas, S., Agusti, S., Duarte, C.M., 2014. Comparing marine primary production estimates through different methods and development of conversion equations. Frontiers in Marine Science 1, 1–14. doi:10.3389/fmars. 2014.00019.
Reuer, M.K., Barnett, B., Bender, M.L., 2005. Marine productivity estimates from continuous O_2/Ar ratio measurements by membrane inlet mass spectrometry. Geophysical Research Letters 32. L19605. doi:10.1029/2005GL023459.
Riley, G.A., 1944. Carbon metabolism and photosynthetic efficiency. American Scientist 32, 132–134.
Roschitz, E., 1996. Sukzession und Produktion in der Alten Donau vor und nach der Sanierung. Diplomarbeit Univ. Wien, 153 S.
Roy, J., Saugier, B., 2001. Terrestrial primary productivity: Definitions and milestones. In: Roy, J., Saugier, B., Mooney, H.A. (Eds.), Terrestrial global productivity. Oxford: Academic Press, pp. 1–6.
Running, S.W., Nemani, R.R., Heinisch, F.A., Zhao, M., Reeves, M., Hashimoto, H., 2004. A continuous satellite-derived measure of global terrestrial primary production. Bioscience 54, 547–560.
Salomon, C.T., Brusewitz, D.A., Richardson, D.C., *et al.*, 2013. Ecosystem respiration: Drivers of daily variability and background respiration in lakes around the globe. Limnology and Oceanography 58, 849–866.
Serret, P., Robinson, C., Arangueren-Gassis, M., *et al.*, 2015. Both respiration and photosynthesis determine the scaling of plankton metabolism in the oligotrophic ocean. Nature Communications 6, 6961. doi:10.1038/ncomms7961.
Staehr, P., Bade, D., Van de Bogert, M.C., Koch, G.R., Williamson, C., Hanson, P., Cole, J.J., Kratz, T., 2010. Lake metabolism and the diel oxygen technique: State of the science. Limnology and Oceanography: Methods 8, 628–644.
Steemann-Nielsen, E., 1951. Measurement of the production oforganic latter in the sea by means of carbon-14. Nature 167, 684–685.
Steemann-Nielsen, E., 1952. The use of radioactive carbon (C^{14}) for measuring organic production in the sea. Journal du conseil/Conseil international pour l'exploration de la mer 18, 117–140.
Steemann-Nielsen, E., Hansen, V.K., 1959. Measurements with the carbon-14 technique of respiretion rates in natural populations of phytoplankton. Deep Sea Research 5, 222–233.
Sun, J., Wu, J., Guan, D., Yao, F., Yuan, F., Wang, A., Jin, C., 2014. Estimating daytime ecosystem respiration to improve estimates of gross primary production of a temperate forest. PLoS One 9. e113512. doi:10.1371/journal.pone.0113512.
Thornton, D.C.O., 2012. Primary production in the ocean. In: Najafpour, M. (Ed.), Advances in photosynthesis—Fundamental aspects. Rijeka: Tech Publ978-953-307-928-8, pp. 563–588. http://www.intechopen.com/books/advances-in-photosynthesis-fundamental-aspects/primary-production-inthe-ocean
Trumbore, S., 2006. Carbon respired by terrestrial ecosystems—Recent progress and challenges. Global Change Biology 12, 141–153. doi:10.1111/j.1365-2486.2005.01067.x.
Turner, D.P., Ritts, W.D., Cohen, W.B., *et al.*, 2005. Site-level evaluation of satellite-based global terrestrial gross primary production and net primary production monitoring. Global Change Biology 11, 666–684.
Turner, D.P., Ritts, W.D., Cohen, W.B., *et al.*, 2006. Evaluation of MODIS NPP and GPP products across multiple biomes. Remote Sensing of Environment 102, 282–292.
Underwood, G.J.C., Kromkamp, J., 1999. Primary production by phytoplankton and microphytobenthos in estuaries. Advances in Ecological Research 29, 94–153.
Vernet, M., Smith, R.C., 2007. Measuring and modeling primary production in marine pelagic ecosystems. In: Fahey, T.J., Knapp, A.K. (Eds.), Principles and standards for measuring primary production. New York: Oxford University Press, pp. 142–174.
Vogt, K., 1991. Carbon budgets of temperate forest ecosystems. Tree Physiology 9, 69–86.
Vollenweider, R.A., 1969. A manual on methods for measuring primary production in aquatic environments. In: IBP handbook no. 12., Oxford and Edinburgh: Published for the International Biological Programme by Blackwell Scientific Publications, p. 213.
Wehr, R., Munger, J.W., McManus, J.B., Nelson, D.D., Zahniser, M.S., Davidson, E.A., Wofsy, S.C., Saleska, S.R., 2016. Seasonality of temperate forest photosynthesis and daytime respiration. Nature 534, 680–683. doi:10.1038/nature17966.
Weinberg, A.M., 1961. Impact of large-scale science on the United States. Science 134, 161–164.
Weinberg, A.M., 1967. Reflections on big science. MIT Press: Cambridge, p. 192.
Westlake, D.F., Květ, J., Szczepański, A. (Eds.), 1998. The production ecology of wetlands. The IBP synthesis. Edinburgh: Cambridge University Press, p. 568.
Williams, P.J.I.B., 1993a. On the definition of plankton production terms. ICES Marine Science Symposia 197, 9–19.
Williams, P.J.I.B., 1993b. Chemical and tracer methods of measuring plankton production. ICES Marine Science Symposia 197, 20–36.
Williams, P.J.I.B., 1998. The balance of plankton respiration and photosynthesis in the open oceans. Nature 394, 55–57.
Williams, P.J.I.B., Lefèvre, D., 2008. An assessment of the measurement of phytoplankton respiration rates from dark ^{14}C incubations. Limnology Oceanography Methods 6, 1–11.
Williams, P.J.I.B., Thomas, D.N., Reynolds, C.S. (Eds.), 2002. Phytoplankton productivity. Carbon assimilation in marine and freshwater ecosystems. Oxford: Bleckwell Scientific, p. 386.
Woodward, F.I., 2007. Global primary production. Current Biology 17, R269–R273.
Woodward, F.I., Badgett, R.C., Raven, J.A., Hetherigton, A.M., 2009. Biological approaches to global environment change mitigation and remediation. Current Biology 19, R615–R623.
Xiao, X., Jin, C., Dong, J., 2014. Gross primary production of terrestrial vegetation. In: Hanes, J. (Ed.), Biophysical applications of satellite remote sensing. Springer remote sensing/photogrammetry. Berlin, Heidelberg: Springer, pp. 127–148.
Yuan, W., Liu, S., Yu, G., *et al.*, 2010. Global estimates of evapotranspiration and gross primary production based on MODIS and global meteorology data. Remote Sensing of Environment 114, 1416–1431.

Zhang, Y., Xu, M., Chen, H., Adams, J., 2009. Global pattern of NPP to GPP ratio derived from MODIS data: Effects of ecosystem type, geographical location and climate. Global Ecology and Biogeography 18, 280–290.
Zhang, Y., Xiao, X., Wu, X., Zhou, S., Zhang, G., Qin, Y., Dong, J., 2017. Data descriptor: A global moderate resolution dataset of gross primary production of vegetation for 2000–2016. Scientific Data 4, 170165. doi:10.1038/sdata.2017.165.www.nature.com/scientificdata
Zhao, M., Heinsch, F.A., Nemani, R.R., Running, S.W., 2005. Improvements of the MODIS terrestrial gross and net primary production global data set. Remote Sensing of Environment 95, 164–176.

Further Reading

Coupland, R.T. (Ed.), 2009. Grassland ecosystems of the world: Analysis of grasslands and their uses,International biological programme, vol. 18. Cambridge: Cambridge University Press, p. 234.
Sakshaug, E., Bricaud, A., Dandonneau, Y., Falkowski, P.G., Kiefer, D.A., Legendre, L., Morel, A., Parslow, J., Takahashi, M., 1997. Parameters of photosynthesis: Definitions, theory and interpretation of results. Journal of Plankton Research 19, 1637–1670.
Whittaker, R.H., Likens, G.E., 1973. Primary production: The biosphere and man. Human Ecology 1, 357–369.

Light Extinction☆

Alberto Barausse, University of Padova, Padua, Italy

Nomenclature

I	Light intensity (or irradiance), W m^{-2}. In the case of PAR, it can be expressed as PPDF (photosynthetic photon flux density, which is the number of photons in the visible range incident per unit time on a surface unit) (mol m^{-2} s^{-1})
I_{in}	Light intensity before crossing a certain medium of finite length (W m^{-2}). In the case of PAR, it can be expressed as PPDF (photosynthetic photon flux density, which is the number of photons in the visible range incident per unit time on a surface unit) (mol m^{-2} s^{-1})
I_{out}	Light intensity after crossing a certain medium of finite length (W m^{-2}). In the case of PAR, it can be expressed as PPDF (photosynthetic photon flux density, which is the number of photons in the visible range incident per unit time on a surface unit) (mol m^{-2} s^{-1})
k	Extinction (or attenuation) coefficient (m^{-1}) (but dimensionless when applied to canopies)
L	Length of the layer of medium crossed by light (m)
$LAI(z)$	The cumulated "Leaf Area Index" from the top of the canopy to distance z (measured along the vertical direction), dimensionless
z	Distance measured along the direction of light (m)

Glossary

Absorption The capture of the energy of a photon by a medium being crossed by light (e.g., gases and aerosols in the atmosphere, water and the suspended/dissolved substances that it contains, leaves, etc.). It is one of the processes causing light extinction.

Extinction coefficient Also known as "attenuation coefficient," it expresses the capability of a given medium to attenuate the intensity of light when crossed by it.

LAI An acronym for "Leaf Area Index," it is an indicator of canopy structure in plants. It is the ratio of the total one-sided green leaf surface to the surface of the ground underneath the canopy (or the projected needle area per unit of ground surface).

PAR An acronym for "photosynthetically active radiation," it is the light which can be used in photosynthesis, approximately in the wavelength range of visible light (400–700 nm).

Scattering The deviation of photons from their original trajectory, due to interactions with the medium being crossed by light (e.g., gases and aerosols in the air, suspended particles in water, etc.), without loss of energy. It is one of the processes causing light extinction.

Introduction

Light is the main source of energy for most ecosystems and a fundamental driving factor in several ecological processes, such as photosynthesis and evapotranspiration. For this reason, understanding and quantifying the light effectively reaching the Earth, its ecosystems and its living organisms can be of primary importance in ecology. Indeed, not all solar radiation is able to reach the Earth's surface, since gases and aerosols forming the atmosphere partially extinguish it. Light penetration is also impeded in aquatic and terrestrial ecosystems, by diverse factors such as water depth and turbidity or the presence of vegetation. This work describes the process of light extinction in air, water bodies, and terrestrial ecosystems, taking a mainly ecological perspective. Simple relationships for quantifying and modeling light extinction are presented, first order kinetic Lambert–Beer law being the main one.

A General Model: The Lambert–Beer Law

The most common relationship for modeling light extinction across a certain medium is Lambert–Beer law, also known as Bouguer's law. It behaves reasonably well for low-concentration media, like gases or low turbidity water. Its mathematical derivation is simple: z axis being along light direction, let us define a thin (infinitesimal) layer of medium of length dz. The

☆*Change History:* November 2017. A Barausse made several small edits throughout the Abstract and the whole text to clarify concepts, to make text smoother and unambiguous, to make symbols more consistent, to eliminate typos and to mention novel concepts. Two recent books have been added to the "Further Reading" section. The Glossary has been added.

This is an update of A. Barausse, Light Extinction, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 2180–2184.

 doi:10.1016/B978-0-12-409548-9.10977-7

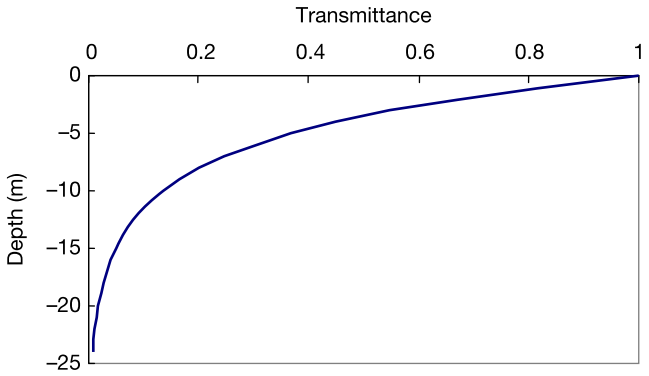


Fig. 1 Light extinction in a certain medium following Lambert–Beer law. The value of the k extinction coefficient used to create the plot is 0.2 m^{-1} (e.g., extinction by water in a coastal sea: for this reason the length of the medium has been plotted as depth).

intensity (or irradiance, W m^{-2}) of light I is diminished passing through the layer by a quantity dI proportional to dz

$$dI = -k \cdot I \cdot dz$$

where k (m^{-1}) is called extinction coefficient and can be determined empirically. It is worth noticing that k is dependent on the type of medium (e.g., clear and turbid waters are characterized by different extinction coefficients), the position in the medium (if heterogeneous), light direction, and (importantly) wavelength.

Rearranging we get,

$$dI/I = -k \cdot dz$$

Integrating over a given layer of finite length L to account for extinction in all the thin layers composing it, and assuming homogeneous extinction properties, it is possible to compute the intensity of the light flux leaving the medium:

$$I_{out} = I_{in} \cdot e^{-k \cdot L}$$

where $I_{in} = I(0)$ and $I_{out} = I(L)$ are the intensities of light fluxes entering and quitting the layer respectively (e.g., if atmospheric extinction is investigated, the flux entering atmosphere and the one finally reaching the ground). The ratio I_{out}/I_{in} is called transmittance (Fig. 1). The quantity $k \bullet L$ is known as "optical depth" and is a measure of the ability of the finite layer to block light.

In stratified systems (e.g., lakes or seas) composed of layers which are heterogeneous with respect to light extinction, in order to obtain more precise results, this relationship can be applied separately to the different layers by using different k values according to layer properties.

Light Extinction in the Atmosphere

Not all the solar radiation approaching the Earth's surface reaches the ground. The atmosphere weakens light intensity in several ways, usually classified as absorption and scattering and summed up as atmospheric light extinction.

Absorption takes place when photons are stopped by atmospheric gases like O_2, N_2, H_2O, CO_2, SO_2, O_3, and N_2O, or by aerosols, which are liquid droplets, solid particles or a mixture of both; they can be due to human activities (e.g., combustion, industrial emissions) or of natural origin (e.g., volcanoes, meteors). In such a case, the photon energy is taken up by gases or particles and can be transformed into heat or radiated.

Photons are scattered when they are deviated without energy loss from their original path by diffusion phenomena due to gases and aerosols, and then no longer reach the ground. Atmospheric scattering is usually divided into Rayleigh scattering, due to particles with diameter smaller than 1/10 of radiation wavelength (mostly oxygen and nitrogen for visible radiation) and Mie scattering, caused by particles with diameter up to 10 times the radiation wavelength. Above this length, a nonselective scattering takes place and geometrical optics laws apply.

Usually, these different extinction phenomena are collectively modeled using Lambert–Beer law. Ignoring Earth curvature effects, light extinction in the atmosphere can be described by the following relationship, which accounts for the fact that solar radiation can hit the atmosphere with a solar zenith angle θ different than zero, and consequently the extinction path can vary in length depending on the hour, day, and latitude:

$$I_{out} = I_{in} \cdot e^{-k \cdot m \cdot L}$$

where I_{out} is the radiation at the ground level, L is the height from the ground at which I_{in} (the incoming radiation) is observed, and m is the optical air mass, defined as the relative length of the path that light has to travel through the atmosphere to reach the ground. Air mass takes the value of one at zenith, and for small zenith angles (up to 60°) it can be computed with good approximation from

$$m = 1/\cos\theta = \sec\theta$$

For larger zenith angles, the effects of refraction, air density, curvature, and nonuniform vertical distribution of substances and temperature are not negligible, so semiempirical formulas derived from field values are used.

The k extinction coefficient can be split into a sum

$$k = (\Sigma)k_i$$

to account for the contributions to extinction of different processes such as absorption and scattering by aerosols, pollution gases, water vapor, and atmospheric gases. As an example, usually, extinction in the visible range due to gases is small compared to the one by suspended particles in inhabited regions.

Light extinction in the atmosphere is strongly wavelength-dependent. For example, greenhouse gases like H_2O, CO_2, N_2O, and CH_4 intensely absorb infrared radiation.

Light Extinction in Water Bodies

Light extinction in aquatic ecosystems can be an important limiting factor for organism growth, particularly for primary producers such as plants and algae, since too strong or too weak light can limit their productivity. Suspended and dissolved substances and water scatter and absorb light; so radiation is attenuated with increasing depth and is often limited only to a superficial layer. Since light used in photosynthesis (PAR or "photosynthetically active radiation") is mainly in the visible range (400–700 nm), henceforth I will refer to PAR when speaking of "light." Anyway, it must be noted that almost all radiation outside that range is extinguished at about 1 m depth. PAR is expressed as W m^{-2} or, more commonly, as PPDF (photosynthetic photon flux density, which is the number of photons in the visible range incident per unit time on a surface unit, μmol m^{-2} s^{-1}).

Light in a water body is attenuated in the so-called euphotic (or photic) zone, often defined as the water layer where photosynthesis can take place or by the depth where 1% of surface irradiance is still to be extinguished, but more precisely limited by the depth where primary producer respiration equals photosynthesis (the so-called compensation depth). The euphotic depth depends on water turbidity and typical values are about a meter or less in very turbid waters like estuaries or some lakes, around 30 m in coastal waters and up to 200 m in clear open ocean waters with low primary productivity.

The attenuation of the intensity of light entering the water surface I_{in} is usually modeled with Lambert–Beer law as a function of depth

$$I(z) = I_{in} \cdot e^{-k \cdot z}$$

where $I(z)$ is light intensity in water at depth z. The extinction coefficient k is associated with a given wavelength and can be determined experimentally by the use of photometers at several depths. It must be noted that most light wavelengths (e.g., red) are filtered after a few meters and then light becomes almost monochromatic (typically blue-green, depending on the presence of suspended and dissolved substances).

Also, k can be computed as the sum of different contributions to extinction, like those of water, suspended solids (e.g., detritus), dissolved material, and living material (e.g., phytoplankton cells). For example, one can consider contributions from water and dissolved matter ($k_{w\text{-}diss}$) and particulate (k_{part}) matter

$$k = k_{w-diss} + k_{part}$$

k_{part} can also be written as a function of particulate matter concentration C:

$$k_{part} = a \cdot C$$

where a is a scalar coefficient to be determined on field. In ecosystems characterized by a shallow euphotic depth or where noticeable algal blooms can be found, phytoplankton concentration usually has dominant effects on light extinction, and it can be useful to express k as a function of such concentration (as measured, e.g., by Chl-a concentration A).

$$k = b + c \cdot A$$

or

$$k = b + c \cdot A + d \cdot A^e$$

where b, c, d, and e are coefficients to be determined experimentally. These equations model the so-called shading and self-shading in phytoplankton, meaning that primary producers attenuate light and can also self-limit their growth because of their increasing biomass blocking PAR. Because of this process, phytoplankton growth as a function of light availability should have a saturation-like behavior (or even a humped, "optimum" behavior, since too high radiation is harmful to algae).

In oligotrophic systems, usually water and dissolved matter have the most significant effects, and b, which is related to background turbidity other than algae and to water effects, is bigger than the other terms.

The extinction coefficient k (m^{-1}) can also be related to Secchi disk depth z_{SD} (m) by the Poole and Atkins equation

$$k \cdot z_{SD} = constant$$

where the proportionality constant must be empirically calculated (it commonly ranges from about 1.7 in clear ocean water to about 1.4 in turbid coastal water). Attention must be paid also to the intrinsic limits of the Secchi disk measure of water

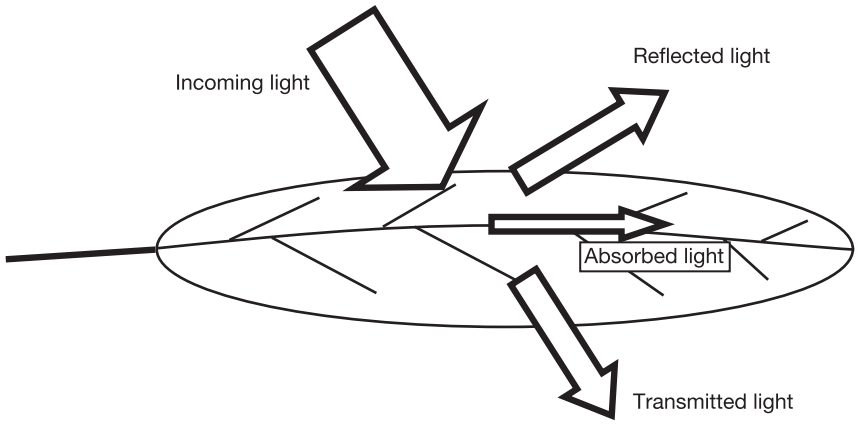


Fig. 2 Conceptual model of the radiation balance for a leaf. The source of incoming radiation can be direct illumination, transmission or reflection.

transparency, defined as the depth at which a standardized white and black disk is no longer visible from the surface: z_{SD} is a simple, cheap, quick and relatively robust measure of water clarity, but it is dependent on observer visual acuity, sun altitude, presence of shadows, and refraction caused by the water surface.

Light Extinction in Terrestrial Vegetation

Light extinction in terrestrial ecosystems is a matter of great interest, as availability of and competition for light can strongly influence plant composition, structure and growth, and consequently shape the productivity and other features of these ecosystems.

In a forest, for example, incoming visible light is mostly attenuated by the upper layer formed by leaves and branches belonging to tall trees (canopy). The underlying, shaded understory and undergrowth vegetation is influenced by this process, and so are the functions it performs, like supporting biodiversity, providing habitat to several animals and nutrient cycling.

The study of light interception and absorption by vegetation is also connected with activities like farming and silviculture, since optimal plant yield depends on whether light can penetrate crops (e.g., those planted in narrow rows to optimize the use of space) and make them grow fast and efficiently. A quantitative knowledge of light extinction can be important to understand and model crop productivity or linked processes like those determining the water balance and the need for irrigation.

Light can penetrate through canopies (this term defines also the upper layer of crops) and reach shaded lower plants and the ground by direct illumination (e.g., through gaps in foliage and canopy), transmission through leaves (with an intensity lower than the incoming one), and reflection by leaves or ground of incoming direct and diffuse radiation. Also, some light is absorbed (e.g., by leaves) or reflected away (Fig. 2).

Given the many, different terrestrial ecosystems influenced by light extinction and their large variability even on short spatial and time scales (e.g., seasonal variability), it is difficult to develop generally applicable models. One of the commonest models used to describe light transmission through a canopy is a modified Lambert–Beer law, obtained assuming a homogeneous (with respect to light attenuation) canopy with small, randomly dispersed leaves:

$$I(z) = I_{in} \cdot e^{-k \cdot LAI(z)}$$

where $I_{in} = I(0)$ is the incoming PPDF (on the top of the canopy, where canopy depth $z = 0$), $I(z)$ is the available PPDF at canopy depth z and $LAI(z)$ is the cumulated "Leaf Area Index" from the top of the canopy to distance z measured long the vertical direction. The Leaf Area Index is the ratio of the total one-sided green leaf surface to the surface of the ground underneath the canopy (or the projected needle area per unit of ground surface). Values are dimensionless and range from 0.2–2.5 (tundra) up to 10 (tropical forest) with higher values for conifers. *LAI* estimates can be achieved by destructive sampling or by using litter traps, allometric relationships (e.g., linking *LAI* to tree diameter or height) and light interception methods (e.g., calculating *LAI* from the inversion of the Lambert–Beer relationship or other models, based on measurements of light transmission through the canopy). Indirect estimates of *LAI* can also be obtained through remote sensing methods such as airborne laser scanning. The extinction coefficient k is species or canopy related and can span from about 0.3 to 2.0. Plants with vertically oriented leaves usually have lower values (e.g., for cereals, $k < 0.5$).

The above relationship linking canopy density and structure to light transmission has showed to be relatively robust with respect to violations of its underlying, often unrealistic assumptions. Anyway, numerous, more complex models describing light attenuation in forest canopies or in single plants can be found in the literature to account for the fact that, for example, many plants do not have a homogeneous spatial structure with respect to light extinction (e.g., leaves are not randomly spatially distributed or their finite size cannot be neglected).

These models can be based on aggregate parameters related to canopy structure, like *LAI*, leaf orientation, and average leaf area density (m^{-1}, total leaf upper surface per unit of volume), can explicitly account for factors like sun position, foliage clumping and distribution, ground and leaf optical properties or diffuse radiation, or can be based on more elaborate approaches, like 3D vegetation models. Also, the importance of the direct evaluation of light extinction by field measurements (e.g., at several canopy depths) is increasing. Data can be collected using different methods, like hemispherical photography, and are also used for obtaining information about canopy structure and for providing inputs to models (e.g., for calibration).

See also: Ecosystems: Tropical Rainforest. General Ecology: Plant Ecology; Plant Physiology

Further Reading

Cescatti, A., 1997a. Modelling the radiative transfer in discontinuous canopies of asymmetric crowns. I. Model structure and algorithms. Ecological Modelling 101, 263–274.
Cescatti, A., 1997b. Modelling the radiative transfer in discontinuous canopies of asymmetric crowns. II. Model testing and application in a Norway spruce stand. Ecological Modelling 101, 275–284.
Chen, J.M., Rich, P.M., Gower, S.T., Norman, J.M., Plummer, S., 1997. Leaf area index of boreal forests: Theory, techniques, and measurements. Journal of Geophysical Research 102, 29429–29443.
Eidels-Dubovoi, S., 2002. Aerosol impacts on visible light extinction in the atmosphere of Mexico City. The Science of the Total Environment 287, 213–220.
Heavens, O.S., Ditchburn, R.W., 1991. Insight into optics. Chichester: John Wiley & Sons.
Hirose, T., 2005. Development of the Monsi-Saeki theory on canopy structure and function. Annals of Botany 95, 483–494.
Kishino, M., Booth, C.R., Okami, N., 1984. Underwater radiant energy absorbed by phytoplankton, detritus, dissolved organic matter, and pure water. Limnology and Oceanography 29, 340–349.
Kirk, J.T.O., 2011. Light and photosynthesis in aquatic ecosystems, 3rd edn. Cambridge: Cambridge University Press.
Palmeri, L., Barausse, A., Jørgensen, S.E., 2014. Ecological processes handbook. Boca Raton: CRC Press.
Preisendorfer, R.W., 1986. Secchi disk science: Visual optics of natural waters. Limnology and Oceanography 31, 909–926.
Thomason, L.W., Herman, B.M., Reagan, J.A., 1983. The effect of atmospheric attenuators with structured vertical distributions on air mass determinations and Langley plot analyses. Journal of the Atmospheric Sciences 40, 1851–1854.

Nitrification☆

BB Ward, Princeton University, Princeton, NJ, USA

Introduction

Nitrification is an essential process in the nitrogen cycle of soils, natural waters, and wastewater treatment systems. It is responsible for the biological conversion of ammonium to nitrate. While both of these compounds are suitable for plant use as nutrients, they behave quite differently in soil systems, and have quite different sources and fates in the marine environment. Ammonium is produced as a waste product from cellular and organismal metabolism, a breakdown product of organic material. It is the preferred nitrogen source for many plants and algae. Nitrate is not only a nutrient, but the substrate for the bacterial process of denitrification, by which nitrate is reduced to dinitrogen gas, N_2. Most plants cannot use dinitrogen gas as a nitrogen source, so denitrification represents a loss term for fixed nitrogen in the ecosystem. Nitrification itself does not directly affect the nitrogen budget, but by linking organic matter decomposition to denitrification, it completes the N cycle.

The significance of nitrification can be summarized in the following list, and the individual items are described in the sections below: (1) transformation of ammonium to nitrate, with implications for the availability of N for plants and algae, (2) production of substrate for denitrification, (3) production of nitrous oxide in aquatic and terrestrial ecosystems, (4) consumption of oxygen in sediments, (5) acidification of the environment.

Nitrifying Microorganisms

Ammonia-Oxidizing Bacteria

Nitrification is performed by two functionally defined groups of microbes, referred to together as nitrifiers. The first group of nitrifiers is the ammonia oxidizers, which oxidize ammonia to nitrite. In most natural waters, ammonium is present predominantly as the positively charged ion, ammonium (NH_4^+), but the enzyme responsible for the first step of the reaction uses the gaseous form, NH_3, which is usually a minor component at equilibrium. We shall use the term ammonium when we are mainly concerned with the form that is important in the environment, and ammonia when referring to the enzymatic oxidation process of the specific substrate. There are two very different groups of ammonia-oxidizing microbes. One is the well-known bacterial group (ammonia-oxidizing bacteria, AOB), which includes a few different kinds of bacteria that all make a living by generating reducing power (ATP) from the oxidation of ammonia and using that energy to fix carbon dioxide (Bock and Wagner, 2006). They are generally considered to be obligate autotrophs, that is, they are unable to utilize or grow on organic carbon to any important extent, and can grow only by fixing their own CO_2 using the Calvin cycle. Ammonia is their only energy source, and their main metabolic product is nitrite. Nitrous oxide is a minor product of ammonia oxidation, and is produced by two different pathways. AOB have been cultivated for over 100 years and their description played an important role in the discovery and early research on chemoautotrophy.

Ammonia-Oxidizing Archaea

A second distinct group of ammonia-oxidizing microbes has only recently been recognized and brought into culture only in 2005 (Konneke *et al.*, 2005). These are not bacteria, but archaea (ammonia-oxidizing archaea, AOA). Like AOB, AOA oxidize ammonia to nitrite and produce nitrous oxide and nitrite from ammonia, but the enzymatic pathways are quite different. AOA are also thought to be predominantly autotrophic, but they fix CO_2 using the 3-hydroxypropionate/4-hydroxybutyrate pathway, rather than the Calvin Cycle (Walker *et al.*, 2010). AOA are abundant in many environments and in the ocean and many terrestrial systems, they far outnumber the AOB. In estuaries, the ratio of AOB to AOB varies widely, with AOB sometimes more abundant.

Although the enzymes and pathways differ for the AOA and AOB, aerobic ammonia oxidation in both groups apparently proceeds by the same stoichiometry:

$$NH_3 + 1.5\,O_2 \rightarrow NO_2^- + H_2O + H^+$$

In addition to the net production of nitrite by the above equation, AOB are also capable of producing nitrous oxide (N_2O) by two distinct pathways. Most AOB investigated to date possess the genes and enzymes necessary for the partial denitrification pathway that reduces nitrite to nitric oxide (NO) and then to N_2O (Casciotti and Ward, 2001, Casciotti and Ward, 2005). The genes involved are homologous to those found in denitrifiers, and the process is often referred to as nitrifier denitrification. The result is the production of N_2O, whereas complete denitrification by the usual denitrifying bacteria produces N_2O only as a

☆*Change History:* March 2013. BB Ward updated the entire article.

transient intermediate. A second pathway produces N_2O from hydroxylamine (NH_2OH), which is an intermediate in the oxidation of ammonia to nitrite by the AOB. The nitrifier denitrification pathway of N_2O production is favored during nitrification at low oxygen concentrations, implying that nitrifiers use this pathway for anaerobic respiration, just as in denitrifiers. AOA also produce N_2O, in approximately the same proportion to NO_2 as in AOB, and with similar isotopic fractionation (Santoro *et al.*, 2011). Nevertheless, the pathways of N_2O production in AOA are unknown, but almost certainly are different from the pathways in AOB. Significantly, AOA do not use NH_2OH as an intermediate, so the production of N_2O from NH_2OH cannot occur in AOA, and AOA do not possess the reductive enzymes used in nitrifier denitrification by AOB.

Anammox Organisms

A third group of bacteria, members of the Planctomycetes phylum, are capable of oxidizing ammonium using nitrite instead of oxygen and producing N_2 instead of nitrite (Kuenen, 2008). This metabolism is strictly anoxic and the process is known as anaerobic ammonia oxidation, or anammox. Anammox organisms are unique in a number of ways; the pathway for oxidation of ammonium is not very similar to that found in AOB or AOA, although the enzymes involved may be evolutionarily related. Hydrazine, more commonly associated with rocket fuel than with biological systems, is an intermediate, while hydroxylamine, an intermediate in AOB, is apparently not involved in anammox. Anammox organisms are strict autotrophs, and apparently use the acetyl-CoA pathway for CO_2 fixation. Their growth is extremely slow, with generation times on the order of 2 weeks. The cells contain an internal membrane-bound 'organelle' called the anammoxosome, in which the anammox reaction is localized. The cell membranes contain unique lipids called ladderanes, after their diagrammatic appearance as a ladder, which form the very dense membrane needed to handle hydrazine as an intermediate, and to prevent its diffusion out of the anammoxosome (van Niftrik and Jetten, 2008). The net reaction for anammox involves a 1:1 combination of ammonium and nitrite in the production of N_2.

$$NO_2^- + NH_4^+ \rightarrow N_2 + 2H_2O$$

Thus, unlike conventional nitrification, anammox results in the loss of fixed nitrogen from the system, and is ecologically equivalent to denitrification, rather than to nitrification. Anammox results in the anaerobic removal of ammonium using nitrite, derived from either aerobic ammonium oxidation or partial denitrification, as the oxidant.

Nitrite-Oxidizing Bacteria

The second functionally defined group of nitrifying microbes is the nitrite-oxidizing bacteria (NOB), which include several genera. The best-known cultivated members, in the genus *Nitrobacter*, are chemolithoautotrophic, like the AOB, using nitrite as an energy source and CO_2 as a carbon source via the Calvin cycle (Bock and Wagner, 2006). However, the lesser know genus, *Nitrospina*, is apparently most abundant in the ocean, and uses the reductive tricarboxylic acid pathway for CO_2 fixation. Many strains are known to possess heterotrophic capabilities and are considered mixotrophic or facultative autotrophs. Although they have limited metabolic capabilities for uptake and degradation of organic molecules, they can supplement their growth with organic carbon and, in some cases, grow slowly in the absence of nitrite when certain organic substrates are present. The oxidation of nitrite is even less energy yielding than ammonia oxidation, so perhaps this ability for heterotrophic growth is not surprising. Aerobic nitrite oxidation proceeds by the following stoichiometry:

$$NO_2^- + 0.5O_2 \rightarrow NO_3^-$$

There are no other pathways, nor any different kinds of bacteria or archaea known to be capable of or involved in nitrite oxidation in the environment. The recent finding of greater diversity among ammonia-oxidizing microbes begs the question, however, of whether additional nitrite oxidation pathways and organisms remain to be discovered.

Heterotrophic Nitrifiers

The ability to nitrify, via pathways involving the inorganic transformations normally associated with the autotrophic nitrifiers described above, or via pathways involving organic intermediates but resulting in the net oxidation of ammonium, has been attributed to some heterotrophic bacteria and fungi. Heterotrophic nitrification does not conserve energy (i.e., is not linked to ATP production) and the rates observed are much slower than rates found in cultivated conventional nitrifiers. Autotrophic nitrifiers are susceptible to inhibition by a number of naturally occurring substances, including secondary metabolites of some trees, for example. AOB are inhibited by acidic conditions, which pertain in some soils. These observations led to the suggestions that heterotrophic nitrification might be particularly important under conditions in some soils that are very unfavorable for known autotrophic nitrifiers. The quantitative importance of heterotrophic nitrification remains uncertain in both aquatic and terrestrial environments.

Ecological Roles of Nitrification

Agricultural and Terrestrial Systems

Nitrogen is the main component of fertilizers applied in many agricultural systems. Addition of N as ammonium is advantageous because it is easily assimilated by plants and, due to its positive charge, it binds to soil particles and is somewhat resistant to loss in runoff. Nitrifying bacteria in the soil can convert the ammonium to nitrate, which is more easily lost in the soil solution, thus reducing the efficiency and increasing the cost of fertilizer application. Nitrification inhibitors are therefore often applied along with fertilizers, to slow down this conversion and increase the amount of N available to the plants.

Not only is the nitrate more susceptible to physical loss from the system, but it is also the substrate for denitrification. If soils become waterlogged to the extent that interstitial spaces become anoxic, denitrifying bacteria present in the soils will switch to anaerobic metabolism, in which they respire oxides of nitrogen, beginning with nitrate, instead of oxygen. Nitrate respiration leads to the removal of nitrate by its conversion to N_2 gas, which is not biologically available to most plants and is lost from the system by evasion. Both ammonia-oxidizing and denitrifying bacteria can carry out the reduction of nitrite to N_2O. For denitrifiers, this is part of the usual pathway from nitrate to N_2 ($NO_3^- \rightarrow NO_2^- \rightarrow NO \rightarrow N_2O \rightarrow N_2$). For AOB, the pathway is analogous but includes only the steps $NO_2^- \rightarrow NO \rightarrow N_2O$ and it is not known what purpose it serves. Most of the N_2O produced by ammonia oxidation is probably produced by AOA via a so far undescribed pathway (Santoro *et al.*, 2011). Especially in low oxygen conditions, substantial nitrogen can be lost as N_2O. Not only is this N lost from the bioavailable pool, but it plays a very important role in the atmosphere as a greenhouse gas. N_2O has a radiative forcing that is on the order of 200 times more potent per molecule than CO_2, the most abundant greenhouse gas. Thus, N_2O fluxes from agricultural systems to the atmosphere are potentially of concern. Management practices that optimize the amount of N added, and the timing and amount of water applied, are important in minimizing N loss both as N_2 and N_2O. Inhibition of nitrification is also widely seen as the main point of control for N loss from agricultural systems, and is practiced by the application of various commercial additives with varying degrees of specificity for nitrification over other microbial processes.

Wastewater Treatment

Waste nitrogen enters the wastewater stream in the form of ammonia, urea, and organic nitrogen. The first steps in wastewater treatment involve oxidative degradation of organic matter, resulting in the conversion of dissolved organic N into ammonium. If released as effluent in this form, the ammonium would fertilize the receiving waters, potentially leading to eutrophication and the growth of undesirable plant life and eventually possibly anoxia. Tertiary wastewater treatment is designed to remove inorganic nitrogen from the stream by nitrification. When carried out by AOB and AOA, this is an obligately aerobic process, and requires aeration to allow the growth and activity of AOB, AOA and NOB to produce nitrate. Water containing the nitrate thus produced is then subjected to an anoxic treatment in which denitrification reduces the nitrate to N_2. In this form, the effluent does not increase the bioavailable N in the receiving waters. Optimization of nitrification in wastewater treatment is the subject of much research, focusing both on the species composition of nitrifying communities in wastewater systems and biofilms, and on the physiology of nitrifiers subjected to the many potentially inhibitory components of wastewater.

Tertiary treatment is an expensive component of wastewater treatment, especially as it involves the cultivation of fastidious microbes and the sequential use of oxic, then anoxic conditions, usually requiring multiple tanks and pumping systems. It is clear, therefore, why the discovery of anammox led to its immediate patenting and a flurry of study on the process. Here is a group of organisms that can oxidize ammonia completely to N_2 in one tank and require nothing but anoxia and a supply of CO_2. In fact, anammox was discovered in an enrichment culture from a wastewater treatment plant, and the real mystery is why it was not found before. This may be due to its slow growth rate; many water treatment plants may have discouraged the development of the anammox process by the timing and conditions of treatment stages. Anammox bacteria have never been grown in pure culture; they apparently require the presence of complex consortia including nitrifiers and or denitrifiers, in order to obtain both ammonium and nitrite in the right proportions. Commercial anammox reactors for wastewater treatment usually involve a mixture of aerobic nitrifiers and anammox bacteria. The nitrifiers produce the nitrite required for anammox, and consume trace oxygen that would inhibit the anammox reaction. Establishment of a stable reactor consortium can require years (Kuenen, 2008).

The Marine Environment

The nitrogen cycle of the ocean is interesting because of the role of N as a limiting nutrient for primary production in the sea, and because the ocean is the ultimate repository for waste from land, in the form of wastewater effluent and natural drainage from rivers. Nitrogen loading in natural waters, from excess fertilizer applications as well as wastewater effluent, has increased in recent decades, such that the impact in coastal waters is now detectable. Nitrification plays a part in both of these processes as described above.

The Ocean Ecosystem

The two primary forms of nitrogen that are available to phytoplankton as nutrients are the same as those used by terrestrial plants, ammonium and nitrate. Ammonium is supplied to surface waters via recycling of organic nitrogen in waste products of grazers and

heterotrophic bacteria. It is rapidly recycled in the euphotic zone (well lit surface layer of the ocean) and usually present in very low concentrations. Nitrate is supplied to the sunlit surface waters by upwelling or wind mixing of deeper waters where nitrate concentrations are generally elevated, or by nitrification of ammonium in near surface waters. Although both ammonium and nitrate are suitable N sources for many phytoplankton, different species of phytoplankton exhibit important differences in their abilities to utilize and grow on them. For example, most clades of the most abundant small phytoplankton, a small cyanobacterium called *Prochlorococcus*, cannot grow on nitrate at all, and some forms cannot utilize nitrite either. Their only inorganic N source is ammonium. *Prochlorococcus* is most important in the oligotrophic central gyres of the oceans, where the mixed layer is so deep that nitrate is rarely mixed into the photic zone; natural selection has evidently led to the loss of genes involved in the assimilation of oxidized nitrogen because they were not useful in this environment.

At the opposite extreme are diatoms, eukaryotic phytoplankton with silicious shells, which are extremely important in upwelling and coastal regimes where nitrate is typically more abundant. Diatoms often show a strong preference for ammonium, presumably because it requires less energy to assimilate than does nitrate. This preference is demonstrated by the observation that in the presence of both ammonium and nitrate, even when the latter is present at much higher concentrations, diatoms will first use up the ammonium before beginning to assimilate nitrate. The irony is, however, that their subsequent growth on nitrate can be much faster than the earlier growth on ammonium. Diatoms can attain very high growth rates on nitrate, and are characteristic bloom formers because they can grow much faster than their grazers can, and thus they avoid predation.

The significance of nitrification in this surface ocean N cycling is the conversion between ammonium and nitrate. The deep ocean has very high nitrate concentrations, while the surface ocean is usually quite depleted in nitrate, and this observation led to the long-held belief that nitrification occurred only in the deep sea. If that were true, then nitrate availability would be controlled by physical processes that somehow mix the deep water up into the sunlit surface zone. In fact, broad patterns of oceanic primary production can be explained at first pass by a consideration of the physical oceanographic constraints on mixing in various regions of the world ocean. It is now known, however, that AOB, AOA and NOB are present in the surface ocean and that the rate of nitrification shows a distinct maximum near the bottom of the euphotic zone. It can be shown in culture, enrichment experiments, and incubations of natural seawater that nitrification, both ammonia and nitrite oxidation, is inhibited by strong light, an observation that likely contributes to the depth distribution of nitrification. Even with maximum nitrification rates near the bottom of the euphotic zone (e.g., at a depth where 5–10% of the surface light intensity penetrates), nitrification can still supply much of the nitrate demand by primary producers. In this situation, nitrate is cycled rapidly too, and contributes to the support of primary production even when present at low levels. The depth distribution of nitrification rates is characterized by a subsurface maximum near the bottom of the euphotic zone, a rapid decrease in rate with increasing depth, and very low rates in the deep ocean (Ward, 2008). The accumulation of nitrate in deep waters is thus explained by its production at very low rates by nitrification and the very slow loss rates; that is, absence of phytoplankton N demand.

The bottom of the euphotic zone is often characterized by a primary nitrite maximum, a small but distinct accumulation of nitrite that usually occurs around the depth of 1% surface light penetration. The origin of this feature has long been debated, and the two main processes thought to be involved are nitrification and nitrate assimilation by phytoplankton. Light plays an important role in both proposed mechanisms. In the case of nitrification, it is proposed that NOB are more sensitive to light inhibition than are the ammonia oxidizers, such that the AOB or AOA are active at slightly more shallow depths than are the NOB (Olson, 1981). Thus, there is a net production of nitrite from ammonium oxidation, but at slightly deeper depths, the NOB are active and remove the nitrite. This is an attractive explanation and is consistent with the widespread occurrence of the primary nitrite maximum at essentially the same relative depth in many parts of the ocean. The alternative explanation is that phytoplankton involved in assimilation of nitrate find themselves severely light, and therefore energy, limited at this depth (Lomas and Lipschultz, 2006). After taking up nitrate, they are unable to obtain the reducing power necessary for its complete reduction to ammonium for incorporation into biomass, and release some of it as nitrite. Both of these mechanisms might result in diel, as well as seasonal, variability in the feature. Both processes probably contribute the primary nitrite maximum, and their relative contributions vary with system and season.

The distributions of AOB, AOA and NOB are now much more well known due to recent advances in enumeration technology. AOA are generally more abundant than AOB and often show a depth distribution that is closely correlated with distribution of ammonium oxidation rate; i.e., low numbers in surface waters, a discrete subsurface maximum and decreasing numbers with depth. The depth distribution of NOB has been investigated very rarely; NOB are generally less abundant that AOA and likely have a similar depth distribution. Isotopic methods for measurement of nitrification rates are not dependent upon the kind of organism responsible for the process, and they generally show that rates decrease rapidly with increasing depth. This is consistent with the general pattern of decreasing biological activity overall with increasing distance from the surface layer.

Oxygen-Depleted Waters and Sediments

In addition to its role in controlling the nitrate distribution in the ocean, nitrification performs the same role in the ocean as it does in agriculture and wastewater treatment, in linking ammonium regeneration to denitrification and thus facilitating the conversion of organic N to N_2 (see Denitrification). In the ocean, denitrification is restricted to sediments and to a few regions of the water column where organic supply and ocean circulation cooperate to limit oxygen concentrations to very low levels. Three such regions account for essentially all of the water column denitrification in the ocean: the eastern tropical North Pacific (the Mexican Margin),

the eastern tropical South Pacific (the Peru upwelling region), and the Arabian Sea (Devol, 2008). These regions are characterized by high surface productivity and limited intermediate water renewal, so that water in the depth interval of about 80–1000 m is very low in oxygen. In this interval, oxygen concentration is low enough that denitrification can occur, and nitrifier denitrification is also enhanced, leading to N_2O production. It is not known how much nitrification and denitrification each contribute to the N_2O flux, but these oceanic regions are responsible for most of the atmospheric N_2O flux from the ocean.

As in other regions of the ocean, nitrification rates are generally highest in the near subsurface region of oxygen minimum zones (OMZs), where there is usually a strong primary nitrite maximum. OMZs are also characterized by the presence of a strong secondary nitrite maximum (**Fig. 1**), usually at the heart of the low-oxygen depth interval where it is assumed that oxygen concentrations are too low to support conventional aerobic nitrification. Nitrite concentration in the secondary nitrite maximum can be much higher than found in the primary nitrite maximum, and is thought to derive from partial denitrification, although an adequate mechanism has never been proved.

Nitrous oxide also has a characteristic distribution in OMZs (**Fig. 1**), typically exhibiting two maxima; one occurs just below the primary nitrite maximum and is often attributed to nitrifier denitrification. The second N_2O maximum occurs below the secondary nitrite maximum and is usually attributed to denitrification. At the depth interval of the secondary nitrite maximum itself, N_2O is at a minimum, and its undersaturation here is attributed to complete denitrification. Although the main processes involved in the maintenance of these characteristic distributions are probably known (i.e., nitrification, denitrification, anammox), the mechanisms by which they are regulated to result in such predictable distributions is unknown.

Anammox does not produce or consume N_2O but can be an important component of the total fixed N loss in OMZ regions, because it contributes to the production of N_2. On a global basis, its contribution to fixed N loss is regulated by the stoichiometry of organic matter that is degraded in the OMZs and averages about 29%, with denitrification responsible for the rest (Dalsgaard *et al.*, 2012).

Coastal regions that receive nitrogen-rich runoff from land are susceptible to nutrient enrichment leading to eutrophication. This can also lead to low-oxygen conditions in the shallow bottom water, and thence to the production of N_2O and fixed N loss via denitrification and anammox, as described for the low-oxygen oceanic regimes. Both coastal and oceanic low-oxygen regions are likely to be sensitive to environmental change brought about by changes in nutrient loading and stratification, both factors that respond to anthropogenic forcing of the atmosphere and land systems. Denitrification, nitrification, and anammox are all sensitive to oxygen concentration, so the extent of oxygen-depleted waters may have a major effect on the overall N budget of the oceans.

Coupled nitrification/denitrification and anammox are also very important in the N cycle of marine sediments. Approximately half the N_2 production via anaerobic organic matter degradation is derived from anoxic sediments, where degradation by aerobic heterotrophs of organic matter settling down from overlying waters leads to consumption of oxygen in the sediments. Ammonium released from both aerobic and anaerobic decomposition of organic matter diffuses into the oxygenated zone of surface sediments and is nitrified to nitrate. Nitrification is often identified as the main sink term for oxygen in oceanic sediments. Conventional nitrification can be coupled to denitrification by the diffusion of intermediates and end products across the oxic/anoxic gradient

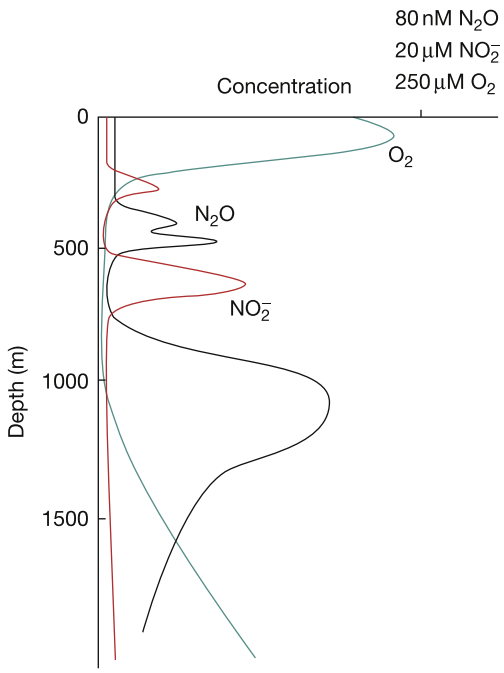


Fig. 1 Characteristic distributions of nitrite (NO_2^-) nitrous oxide (N_2O) and oxygen (O_2) in the water column of an oxygen minimum zone (OMZ) in the ocean. Concentration axis shows the scale for each plotted variable.

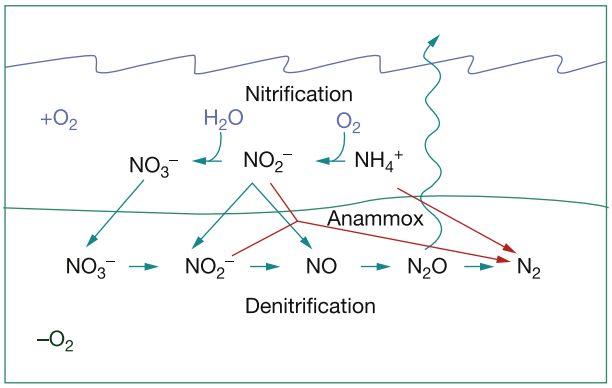


Fig. 2 Interactions between nitrification, denitrification, and anammox across the oxic ($+O_2$)/anoxic ($-O_2$) interface in a marine sediment. In the oxic layer (above the green line) nitrification oxidizes ammonium to nitrate. Nitrate and nitrite diffuse into the anoxic layer where they can be denitrified or used in combination with ammonium for anammox. Both denitrification and anammox are ultimately dependent upon nitrification for the production of oxidized N species, which are reduced to N_2 gas.

near the surface sediment. Similarly, anammox is dependent upon supply of oxidants (nitrate or nitrite) that are ultimately derived from conventional nitrification (Fig. 2). Both conventional nitrification, coupled to denitrification, and anammox are important in sediment systems, and because of the metabolic differences (especially their relationships to oxygen and organic matter) among the organisms responsible, it is likely that their contributions vary greatly among estuarine, shallow, and deep marine systems.

Environmental Factors that Affect Nitrification

Several environmental factors that might control nitrification in various ecosystems have already been mentioned. They include the kinds of things that affect biological processes in general, as well as those particular to the metabolism of nitrifiers: temperature, salinity, light, organic matter concentrations, substrate (ammonium and nitrite) concentrations, pH, and oxygen concentration. A few of the interesting and unique interactions of nitrifiers with their environment are explored below.

Ammonium, the primary substrate for AOB and AOA, is rarely abundant in oxic environments, so the rate of nitrification is likely to be at least partially controlled by substrate limitation. Most cultivated AOB have affinities for ammonium that preclude their being effective competitors for ammonium in the ocean. The one cultivated marine AOA strain, however, has a very high affinity for ammonium (Martens-Habbena *et al.*, 2009), as do natural assemblages (Newell *et al.*, 2013). The absence of ammonium in the deep ocean, where sinking organic material is mineralized to produce ammonium, speaks to the efficiency with which AOA, and to a lesser extent AOB, remove it.

AOB, AOA and NOB require molecular oxygen for their metabolism and thus are restricted to oxic environments. Nonetheless, they seem to prefer and to do quite well under very low oxygen conditions, displaying a microaerophilic lifestyle. Anammox organisms, in contrast, are strictly anaerobic, and while oxygen apparently does not kill them, it does inhibit their activity. Thus, oxygen concentration, in the bulk water of aquatic environments and in the interstices of sediments and soils, is likely a very important variable for regulation of microbial activities and the resulting distribution of nitrogen-cycling processes.

All of the nitrifying microorganisms are predominantly autotrophs, that is, they fix their own carbon from CO_2, and thus do not rely on a supply of organic matter for nutrition. This means that they are not in competition with heterotrophs for the utilization of organic substrates, but rather that they exploit a different niche. This niche involves certain 'sacrifices', in terms of slower growth rates (see Units of Selection). These forms of autotrophic growth are also quite inefficient, due to the low energy yield of the transformations involved. Thus nitrifiers process large amounts of nitrogen in order to obtain the energy required for CO_2 fixation. The molar ratio of N oxidized to C fixed has been estimated at 25–100 for AOB, AOA and NOB, ensuring that their metabolism has a very large effect on the nitrogen cycle, but very little influence on the carbon cycle, where photosynthetic autotrophs are overwhelmingly important.

Light inhibition of nitrifiers is suspected as a mechanism for the formation of the primary nitrite maximum (see above) and it is easily demonstrated in culture that both AOB and NOB are sensitive to light. Cultivated AOA are at least as light sensitive as AOB, perhaps even more so (Merbt *et al.*, 2011). The basis for the light sensitivity of AOB and NOB is assumed to be damage to the many cytochromes that are involved in the energy transduction pathways of nitrification. AOA do contain cytochromes, so light sensitivity must be yet another way in which AOA and AOB exhibit similar physiologies but for different reasons.

Any transformation that involves the production or consumption of hydrogen ions is pH sensitive, and ammonia oxidation is no exception (see Acidification). Oxidation of ammonia by AOB and AOA results in the acidification of the medium. Low pH eventually inhibits both groups in culture, and activity can be restored by pH adjustment. Ammonia oxidation rates in the ocean were reduced by increased acidification, suggesting that nitrification might be affected by ocean acidification due to increased CO_2 concentrations in the atmosphere and ocean (Beman *et al.*, 2011). Short term experiments, however, may overestimate the

response to lowered pH, so the sensitivity of nitrification to long term global change is unknown. It is unlikely that pH is an important controlling variable in the ocean, even in sediments, but pH could be very important in regulation of nitrification in acid soils. While nitrification generally occurs in acid soils, it has proven difficult to obtain acidophilic nitrifiers in culture, leading to speculation about the importance of heterotrophic nitrification in this system. It is now known that many of the kinds of nitrifying bacteria that can be identified by their gene sequences in the natural environment, are not in fact represented in culture collections. Thus it is quite possible that acidophilic autotrophic nitrifiers exist but are resistant to cultivation.

Salinity and temperature do not appear to set any unusual constraints on the range of conditions under which nitrification can occur, and different kinds of nitrifiers appear to have adapted to the wide range of these variables found on Earth. The mechanism by which salinity affects nitrification is not known, but it is clear that salinity is an important determinant of the community composition of nitrifying microbes, if not the net rate of nitrification; that is, different kinds of nitrifiers are adapted to different salinity levels, but nitrification occurs under high as well as low salinity and depends on the presence of different species. Nevertheless, it can be shown that salinity dramatically affects the rate of nitrification, when salinity changes are imposed in an experiment with natural assemblages. Ionic strength effects related to the sorption and availability of ammonium are not sufficient to explain the effects of salinity.

Methods for Measurement of Nitrification Rates

^{15}N Tracer Methods

The best method for quantification of nitrification rates remains the direct ^{15}N tracer approach in which a small amount of either $^{15}NH_4^+$ or $^{15}NO_2^-$ is added to a sample and incubated (Ward, 2010). The labeled product, $^{15}NO_2^-$ or $^{15}NO_3^-$, is then extracted and measured on a mass spectrometer. This approach has suffered in the past from the necessity to enrich the substrate pool by the addition of large concentrations of ^{15}N-ammonia or -nitrite, thus artificially enhancing the observed rate. Improvements in assay techniques and mass spectrometer sensitivity with small N masses have minimized this problem.

The same analytical approaches can be applied in a tracer or isotope dilution format. In the isotope dilution format, label is added to the product pool and its dilution by addition of new product with natural abundance isotopic signature during the incubation provides an estimate of production rate.

The advantages of the direct ^{15}N approaches, compared to inhibitor methods, include shorter, thus less artifactual, incubations, minimal perturbations to *in situ* conditions (ambient light and nutrient conditions can be used), and much greater sensitivity. Inhibitor-based assays have the advantage, however, of requiring simpler less expensive instrumentation, as colorimetric, rather than mass spectrometric, analysis usually suffices.

Inhibitor-Based Nitrification Assays

Inhibitor methods depend on the ability of many compounds to interact specifically with the active site of the ammonia monooxygenase (AMO) or nitrite oxidoreductase (NXR) enzyme. A large number of potential inhibitors has been used for AMO, while chlorate is still the only specific inhibitor reported for nitrite oxidation. Replicate incubations are carried out in the dark with additions of either AMO or NXR inhibitors and the changes in inorganic N concentrations are used to infer nitrification rates. A more sensitive permutation of this approach involves measurement of $^{14}CO_2$ fixation in the presence and absence of inhibitor, where the decrease in the rate of $^{14}CO_2$ assimilation in the presence of nitrifier inhibitor is attributed to nitrification (Rees *et al.*, 2002). A conversion between CO_2 fixation and N oxidation rates is then used to estimate nitrification.

One of the main attractions of the inhibitor approaches is their ease of use and analysis. Scintillation counters are much more common and easier to use than instruments required for stable isotope analysis and inorganic N determinations involve analytical methods that are already standard to most laboratories. Although the inhibitor approaches are usually not appropriate for absolute rate measurements (because of uncertainty in conversion factors and perturbations during incubations) they can be very useful for spatial or temporal comparisons within studies.

See also: Ecological Processes: Ammonification. Global Change Ecology: Nitrogen Cycle

References

Beman, J.M., Chow, C.-E., King, A.L., *et al.*, 2011. Global declines in oceanic nitrification rates as a consequence of ocean acidification. Proceedings of the National Academy of Sciences of the United States of America 108, 208–213.

Bock, E., Wagner, M., 2006. Oxidation of inorganic nitrogen compounds as an energy source. The Prokaryotes 2, 457–495.

Casciotti, K., Ward, B., 2001. Dissimilatory nitrite reductase genes from autotrophic ammonia-oxidizing bacteria. Applied and Environmental Microbiology 67, 2213–2221.

Casciotti, K., Ward, B., 2005. Phylogenetic analysis of nitric oxide reductase gene homologues from aerobic ammonia-oxidizing bacteria. FEMS Microbiology Ecology 52, 197–205.

Dalsgaard, T., Thamdrup, B., Farias, L., Revsbech, N.P., 2012. Anammox and denitrification in the oxygen minimum zone of the eastern South Pacific. Limnology and Oceanography 57, 1331–1346.
Devol, A.H., 2008. Denitrification including anammox. In: Capone, D.G., Bronk, D.A., Mulholland, M.R., Carpenter, E.J. (Eds.), Nitrogen in the marine environment, 2nd edn. Amsterdam: Elsevier.
Konneke, M., Berhnard, A.E., De La Torre, J.R., *et al.*, 2005. Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature 437, 543–546.
Kuenen, J.G., 2008. Anammox bacteria: from discovery to application. Nature Reviews Microbiology 6, 320–326.
Lomas, M.W., Lipschultz, F., 2006. Forming the primary nitrite maximum: Nitrifiers or phytoplankton? Limnology and Oceanography 51, 2453–2467.
Martens-Habbena, W., Berube, P.M., Urakawa, H., De La Torre, J.R., Stahl, D.A., 2009. Ammonia oxidation kinetics determine niche separation of nitrifying Archaea and Bacteria. Nature 461, 976–982.
Merbt, S.N., Stahl, D.A., Casamayor, E.O., *et al.*, 2011. Differential photoinhibition of bacterial and archaeal ammonia oxidation. FEMS Microbiology Letters 327, 41–46.
Newell, S.E., Fawcett, S.E., Ward, B.B., 2013. Depth distribution of ammonia oxidation rates and ammonia-oxidizer community composition in the Sargasso Sea. Limnology and Oceanography. in press.
Olson, R.J., 1981. Differential photoinhibition of marine nitrifying bacteria: A possible mechanism for the formation of the primary nitrite maximum. Journal of Marine Research 39, 227–238.
Rees, A.P., Malcolm, E., Woodward, S., *et al.*, 2002. Size-fractionated nitrogen uptake and carbon fixation during a developing coccolithophore bloom in the North Sea during June 1999. Deep Sea Research Part II: Topical Studies in Oceanography 49, 2905–2927.
Santoro, A.E., Buchwald, C., Mcilvin, M.R., Casciotti, K.L., 2011. Isotopic signature of N(2)O produced by marine ammonia-oxidizing archaea. Science 333, 1282–1285.
Van Niftrik, L., Jetten, M.S.M., 2008. Anaerobic ammonium-oxidizing bacteria: Unique microorganisms with exceptional properties. Microbiology and Molecular Biology Reviews 76, 585–596.
Walker, C.B., De La Torre, J.R., Klotz, M.G., *et al.*, 2010. Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. Proceedings of the National Academy of Sciences of the United States of America 107, 8818–8823.
Ward, B.B., 2008. Nitrification in marine systems. In: Capone, D.G., Bronk, D.A., Mulholland, M.R., Carpenter, E.J. (Eds.), Nitrogen in the marine environment, 2nd edn. Elsevier.
Ward, B.B., 2010. Measurement and distribution of nitrification rates in the oceans. Methods in Enzymology 486, 307–323.

Further Reading

Capone, D.G., Bronk, D.A., Mulholland, M.R., Carpenter, E.J., 2008. Nitrogen in the marine environment, 2nd edn. Amsterdam: Elsevier.
Fenchel, T., King, G.M., Blackburn, T.H., 2012. Bacterial biogeochemistry: The ecophysiology of mineral cycling, 3rd edn. San Diego: Academic Press.
Ward, B.B., 2005. Molecular approaches to marine microbial ecology and the marine nitrogen cycle. Annual Review of Earth and Planetary Science 33.092203.122514.
Ward, B.B., Arp, D.J., Klotz, M.G., 2011. Nitrification. Washington D.C: American Society for Microbiology Press.

Physical Transport Processes in Ecology: Advection, Diffusion, and Dispersion

A Marion, University of Padua, Padua, Italy

Introduction

Most of us experience physical transport of mass and heat in fluids several times in a day. Mass transport is indeed one of the first processes we deal with every day, when we sit in front of our cup of coffee or tea. Our first action is to stir the fluid in the cup, to mix or blend the sugar uniformly in the fluid. We succeed by generating a highly turbulent flow field where mass is displaced very rapidly within the fluid domain. We continually perform countless actions, including breathing, which aim at enhancing or reducing the physical transport of mass in a fluid.

Physical transport of mass and heat in fluids not only occurs in human behavior, but it is also an inherent part of life in all forms, and plays a fundamental role in the fate of organic as well as inorganic matter. In ecological systems, the fate of substances such as nutrients and toxic matter is of fundamental importance and has received increasing attention in the last decades. The aim of these studies is to explain and model how substances move within and across media, typically water or air, and to estimate what concentration the substance may attain in the domain of interest at any given time. The need to understand the complex physical, chemical, and biological interactions in the environment leads to a wide increase of interdisciplinary studies involving physicists, ecologists, biologists, and engineers, to the development of new branches of science, such as biogeochemistry and environmental engineering, and even to the identification of new physical domains of interests, typically at the interface between the elements (water–soil, soil–air, water–air).

The starting point of the discussion on the physical transport of mass is the evidence that, unless temperature vanishes in absolute terms (0 K = − 273 °C), matter moves. Electrons move within atoms, atoms move within molecules, and molecules move within cells and bodies. Even the condition we normally refer to as stillness involves movement, although only at scales that are invisible or imperceptible to our senses. Some examples are trivial: a drop of ink in water is expected to spread in all directions even if the water is motionless! We call this process molecular diffusion – it is the result of small displacements of molecules about their position, a process known as Brownian motion. The displacement of molecules increases with the rise in temperature. While molecules agitate about their local position, we perceive a natural tendency of matter to transfer from regions of high concentration to regions of low concentration. This transfer only stops when the substance is uniformly distributed in the physical domain. However, if Brownian motion is the only cause of mixing, the process is extremely slow and may become negligible when other transport processes are present.

Since ecology is interested in the description of processes at a scale much bigger than the molecular size, a model which describes mass flux as proportional to the gradient of the concentration is sufficient to describe diffusion in still fluids. Such a model, called Fick's law, will be discussed in the section devoted to diffusion. At this point, it is worth noticing that (1) molecular diffusion is active even in the absence of perceivable motion in the fluid, that is, it is the basic process of mass transfer; and (2) diffusion is to be considered an upscaling model rather than a true physical process, as it is a model description of the combined effect of a multitude of individual displacements.

Advection

Advection is defined as the transport of a conserved scalar quantity that is transported in a vector field.

In ecology, the scalar quantity is the mass of transported substance (or its amount per unit volume, called concentration C [ML^{-3}]), while the vector field is the fluid flow field, identified by the three components of a velocity vector $\mathbf{v} = (u, v, w)$ [LT^{-1}] defined at each point as a function of time. In the case of transport of heat, the discussion still holds, as long as the concentration of mass is substituted by the concentration of heat, which is normally expressed proportional to temperature. If the substance behaves like a solute, that is, has the same density as the medium or does not feel significant effect of its buoyant weight, each element of it (molecule or particle) will be displaced along the direction of the local velocity vector following the same path as if it were an element of the medium. This assumption allows the advective transport to be modeled in a relatively simple way. The mass flux Φ[$ML^{-2}T^{-1}$] through a small area normal to direction $\mathbf{n}$ can simply be written as the product of the local concentration times the component of the local velocity vector along $\mathbf{n}$ (Fig. 1):

$$\Phi_n = C\mathrm{v}\cdot\mathrm{n} = Cv_n \qquad [1]$$

In more general terms, the flux vector can be written as

$$\Phi(\mathrm{x}, t) = C(\mathrm{x}, t)\mathrm{v}(\mathrm{x}, t) \qquad [2]$$

where $\mathrm{x} = (x, y, z)$ is the position in Cartesian coordinates and t is time.

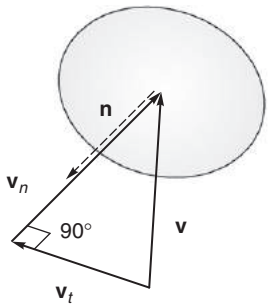


Fig. 1 Schematic of the physical quantities used in eqn [1]. The velocity field generates advective flux through an area only if it has a component in the direction orthogonal to the area, i.e., if $\mathbf{v} \cdot \mathbf{n} \neq 0$.

Deriving a mass balance for a control volume leads to the following differential equation:

$$\frac{\partial C}{\partial t} = -\nabla \cdot \Phi \qquad [3]$$

$$\frac{\partial C}{\partial t} = -\nabla \cdot (C\mathrm{v}) = -\left[\frac{\partial (Cu)}{\partial x} + \frac{\partial (Cv)}{\partial y} + \frac{\partial (Cw)}{\partial z}\right] \qquad [4]$$

where the symbol $(\nabla.)$ indicates the divergence applied to a vector quantity.

The continuity equation shown above can be simplified further if the flow field is solenoidal, that is, the divergence of the velocity is zero $(\nabla.\mathrm{v} = 0)$. This condition applies to incompressible fluids, such as all liquid media found in ecological applications. Under this assumption, the continuity equation simplifies to

$$\frac{\partial C(x,y,z,t)}{\partial t} + u(x,y,z,t)\frac{\partial C(x,y,z,t)}{\partial x} + v(x,y,z,t)\frac{\partial C(x,y,z,t)}{\partial y} + w(x,y,z,t)\frac{\partial C(x,y,z,t)}{\partial z} = 0 \qquad [5]$$

It is simple to show (but omitted here for brevity) that for steady conditions, the solution of eqn [5] is a simple translation of the substance along the paths imposed by the flow field. Although this may not be good representation of the 'diffusive' reality as we experience it, the advective model finds important applications.

Pure advection does not exist alone in nature, as it is always associated at least to molecular diffusion, as discussed above. However, the significance of molecular diffusion in an advective process is negligible in most cases, due to the extremely low value of the diffusive flux. Advection is associated to only molecular diffusion when the flow field is slow, such as in laminar flows. One common application is the flow of water carrying substances in a porous medium, such as an aquifer or the hyporheic zone, as long as the process is modeled at the scale of the pores of the medium. Another useful property of the mass balance equation, valid also when diffusion is significant, is that the center of mass of the substance is subject to advective transport. This is often a sufficient tool to estimate the travel time of a mass of substance between two points of the domain, that is, two locations along a river. The substance disperses in the domain, but its average travel time is always determined by the advective part of the process. Another application is the modeling of the transport of buoyant or heavy substances. These substances are affected by gravity, as their density is either smaller or larger than the density of the medium. Examples are the transport of colloids and suspended solids in water and in air. Particles are subject to the pulsating action of the flow and follow very irregular trajectories. Their behavior is often modeled by the addition of an advective effect, usually defined by a vertical velocity with an intensity dependent on the particle size, shape, and density.

Although the concept of advection and its mathematical description are quite simple, the solution of advective processes of ecological interest in complex flows poses a rather difficult task to modelers. The advection equation is not simple to solve numerically, particularly when strong gradients of concentration (shocks) are treated. Complications arise because numerical methods are indeed curiously affected by a form of diffusion, in this case numerical diffusion rather than the physical one.

Diffusion

One definition of diffusion is the spontaneous spreading of matter, heat, or momentum. Such definition is rather general, and does not differentiate the process based on the cause of the spreading. A key feature of all diffusion processes is the flux of mass from higher to lower concentration.

Diffusion is a spontaneous process that occurs as a result of the second law of thermodynamics, which states that the entropy or disorder of any closed system must always increase with time. The effect of diffusion can only be counteracted by expenditure of external energy. Because substances diffuse from regions of higher concentration to regions of lower concentration, transport occurs only if there is a spatial variation of the concentration. In all cases of diffusion, the net flux of the transported substance is expressed as equal to a physical property (diffusivity, D $[\mathrm{L}^2\mathrm{T}^{-1}]$) multiplied by a concentration gradient:

$$\Phi_n = -D\frac{\partial C}{\partial n} \text{ or } \Phi(\mathrm{x},t) = -D(\mathrm{x},t)\nabla C(\mathrm{x},t) \quad [6]$$

where the symbol (∇) indicates the gradient differential operator.

In molecular diffusion, eqn [6] is written with $D=D_{\mathrm{m}}$ independent of C and is known as Fick's first law.

All diffusion processes are modeled through the diffusion equation, which, in its general form, is a nonlinear partial differential equation obtained combining [3] with [6]:

$$\frac{\partial C}{\partial t} = \nabla\cdot[D\nabla C] = \frac{\partial}{\partial x}\left(D\frac{\partial C}{\partial x}\right) + \frac{\partial}{\partial y}\left(D\frac{\partial C}{\partial y}\right) + \frac{\partial}{\partial z}\left(D\frac{\partial C}{\partial z}\right) \quad [7]$$

When D is constant, as in the case of molecular diffusion, eqn [7] reduces to Fick's second law (or heat equation):

$$\frac{\partial C(\mathrm{x},t)}{\partial t} = D_m\nabla^2 C(\mathrm{x},t) = D_m\left[\frac{\partial^2 C(\mathrm{x},t)}{\partial x^2} + \frac{\partial^2 C(\mathrm{x},t)}{\partial y^2} + \frac{\partial^2 C(\mathrm{x},t)}{\partial z^2}\right] \quad [8]$$

The value of the molecular diffusion coefficient varies according to the combination of solute and solvent. As far as water is concerned, molecular diffusion is easier for polar molecules such as salt and sugar due to the interactions with polar water molecules. The molecular diffusion coefficient is of the order of $10^{-9} - 10^{-8}$ m^2 s^{-1}. Conversely, apolar molecules like oils and proteins are subject to hydrophobic effects and diffuse in water at a lesser rate. The molecular diffusion coefficient for these substances is of the order of 10^{-10} m^2 s^{-1}.

Combined Advection–Diffusion Processes

Whenever the fluid is in motion, advective transport and diffusive transport act simultaneously. The complexity arising for the combined effects can be treated mathematically by simply adding the advective and diffusive fluxes indicated in eqns [4] and [7]:

$$\frac{\partial C}{\partial t} = -\nabla\cdot(C\mathrm{v}) + \nabla\cdot[D\nabla C] \quad [9]$$

For incompressible fluids, eqn [9] becomes

$$\frac{\partial C}{\partial t} + u\frac{\partial C}{\partial x} + v\frac{\partial C}{\partial y} + w\frac{\partial C}{\partial z} = \frac{\partial}{\partial x}\left(D\frac{\partial C}{\partial x}\right) + \frac{\partial}{\partial y}\left(D\frac{\partial C}{\partial y}\right) + \frac{\partial}{\partial z}\left(D\frac{\partial C}{\partial z}\right) \quad [10]$$

Finally, if diffusion is only of the molecular type, eqn [10] reduces to

$$\frac{\partial C}{\partial t} + u\frac{\partial C}{\partial x} + v\frac{\partial C}{\partial y} + w\frac{\partial C}{\partial z} = D_m\left[\frac{\partial^2 C}{\partial x^2} + \frac{\partial^2 C}{\partial y^2} + \frac{\partial^2 C}{\partial z^2}\right] \quad [11]$$

The overall effect of the superposition of advection and diffusion can be visualized as the translation of the center of mass of the substance according to the advective part, while mass diffuses around the center of mass at a rate dictated by the diffusive term, as shown in **Fig. 2**.

Turbulent Diffusion

Molecular diffusion produced by Brownian motion is no longer the dominant diffusion mechanism when the flow velocity becomes fast enough to overcome viscous forces that tend to keep fluid elements aligned along parallel paths. When this threshold is reached, the flow becomes very irregular, and the fluid elements are entrained and transported by eddies which form either from the slowing effect of the bottom and side boundaries or from the disturbances introduced by geometrical irregularities. This type of flow is called turbulent, and it is characterized by an enhanced momentum and mass transfer across the flow field. Diffusion of mass is no longer controlled by Brownian motion, but rather by the continuous displacement of fluid elements in all directions induced by turbulence. While molecular diffusion is isotropic, turbulent diffusion is often different in all directions, as eddies are continuously stretched and deformed by the flow.

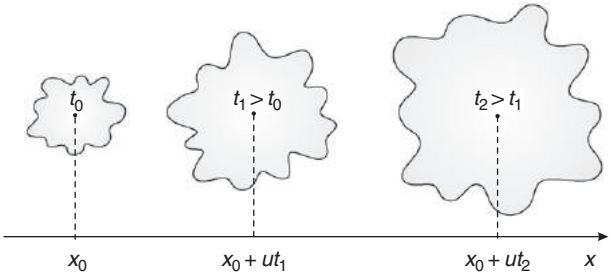


Fig. 2 The combined effect of one-dimensional advection and diffusion on a cloud. The center of mass travels according to the advective process, while the cloud spreads around the center of mass at a rate dictated by diffusion.

Turbulent flows are usually modeled splitting the physical quantities into time-averaged mean values and fluctuations around the mean. After manipulation of the advection–diffusion equation [11], a new mathematical transport term appears, which is the time-averaged product of the fluctuating values of velocity and concentration. If velocity and concentration fluctuations were statistically independent, then these terms would produce no net diffusive mass fluxes. It turns out instead that velocity and concentration irregularities are correlated and that the integral effect over time of turbulent fluxes is always much higher than the fluxes induced by Brownian motion. The mass transport equation can no longer be written as in eqn [11], but it becomes

$$\frac{\partial C}{\partial t} + u\frac{\partial C}{\partial x} + v\frac{\partial C}{\partial y} + w\frac{\partial C}{\partial z} = \frac{\partial}{\partial x}\left(D_x\frac{\partial C}{\partial x}\right) + \frac{\partial}{\partial y}\left(D_y\frac{\partial C}{\partial y}\right) + \frac{\partial}{\partial z}\left(D_z\frac{\partial C}{\partial z}\right) \quad [12]$$

where D_x, D_y, and D_z are called eddy diffusion coefficients, are direction dependent, and so much larger than the molecular diffusion coefficient D_m to replace it in all terms of the equation. The conceptual difference between eqns [12] and [11] is that the value of the coefficients D_x, D_y, and D_z is now determined by the flow regime, that is, they are flow properties, while D_m was independent of the flow and determined only by the combination of solute and solvent. Another difference resides in the fact that eddy diffusivity is scale dependent, while molecular diffusivity is scale independent. Eddy diffusivity typically scales with a 4/3 power of the length scale of the process. This implies that as diffusion makes the substance spread in the domain, diffusivity increases due to the effect of larger eddies that come into play. This dependence is important in large and deep water bodies such as the sea or lakes, where the diffusion process involves several different scales over time. In rivers, instead, the size of the eddies is controlled by water depth and width, and diffusivity is no longer affected by the scale of the process. Typical values of eddy diffusion coefficients are $10^{-6} - 10^{-4}$ m^2 s^{-1} in slow-moving, deep, and stratified water bodies such as lakes and the sea, $10^{-2} - 10$ m^2 s^{-1} for horizontal surface flows such as rivers and channels.

Dispersion

The word dispersion has several meanings. In physics, it is often used to describe the process of separation of the components of waves with different frequency. In chemistry, a dispersion is a system of fine particles uniformly distributed in a medium. Other definitions apply to biological concepts and a few other technical areas. When dealing with mass transport, dispersion is defined as the combined effect of advection and diffusion acting in a flow field with velocity gradients. This is why it is sometimes referred to as 'shear dispersion'.

The presence of velocity gradients on the fate of the substance becomes apparent when spatial averaging of the physical quantities is carried out along with the temporal averaging described in the section on turbulent diffusion. In surface water bodies, it is often convenient to simplify the description of mass transfer by averaging velocity and concentration over the vertical direction (shallow water approach) or over a cross section (unidirectional approach). If quantities are depth-integrated, they are then described only as a function of the two planimetric Cartesian coordinates, that is, $\overline{C}(x, y, t)$ and $\mathbf{v}(x, y, t)$ (the overbar indicates averaged values). Depth averaging is justified when dealing with large rivers, estuaries, and lagoons, using the the evidence that vertical mixing is usually much faster than lateral and longitudinal mixing, due to the limited extension of the domain in the vertical direction. If the physical quantities are averaged over a cross section, their description is limited to the distribution along the longitudinal direction, that is, $\overline{C}(x, t)$ and $\mathbf{v}$(x,t). The one-dimensional approach is justified when the transverse dimensions of the domain are small compared to the longitudinal dimension. This is the reason why the one-dimensional approach is commonly adopted for dispersion processes in rivers and channels.

The comprehension of the physical mechanism leading to dispersion may be improved by the example presented in **Fig. 3**. Case 1 shows a shear flow in a highly turbulent environment. It is a vertical two-dimensional flow, just for the sake of simplicity, which can easily be generalized to a three-dimensional flow. The horizontal arrows represent the time-averaged velocity field. In these conditions, while solute molecules are transported along the flow by the mean velocity, they are displaced vertically (or laterally) by turbulence. Each molecule rapidly samples regions of the domain characterized by high velocity as well as regions characterized by slow velocity. Due to high turbulent mixing, after a sufficiently long time, all molecules are expected to travel a similar distance downstream, at an average velocity close to the mean flow velocity. The overall result is that mass is advected downstream with relatively small dispersion, that is, without spreading along the longitudinal direction. Case 2 shows instead a similar flow field with a much smaller turbulence intensity. The limited mixing does not allow for the displacement of solutes far from their original location. Elements like A are likely to remain in relatively fast flow regions while elements like B are likely to stay in regions of slow advection. The overall result in this case is a rapid increase of the distance between elements over time, that is, rapid dispersive spreading of mass along the flow direction. It is apparent that, when averaged concentrations are used, dispersion is counteracted by turbulent diffusion. The diagram in **Fig. 3** shows the qualitative response of cases 1 and 2 to a slug injection of mass. In case 1, the depth-averaged concentration remains more concentrated while being advected downstream. In case 2, dispersion is much more efficient due to small vertical mixing.

This result, which may appear surprising at first, finds unambiguous confirmation in all dispersion models. If dispersion is modeled over sufficiently long timescales, it can be proven to be well approximated by a Fickian process. Dispersion models are

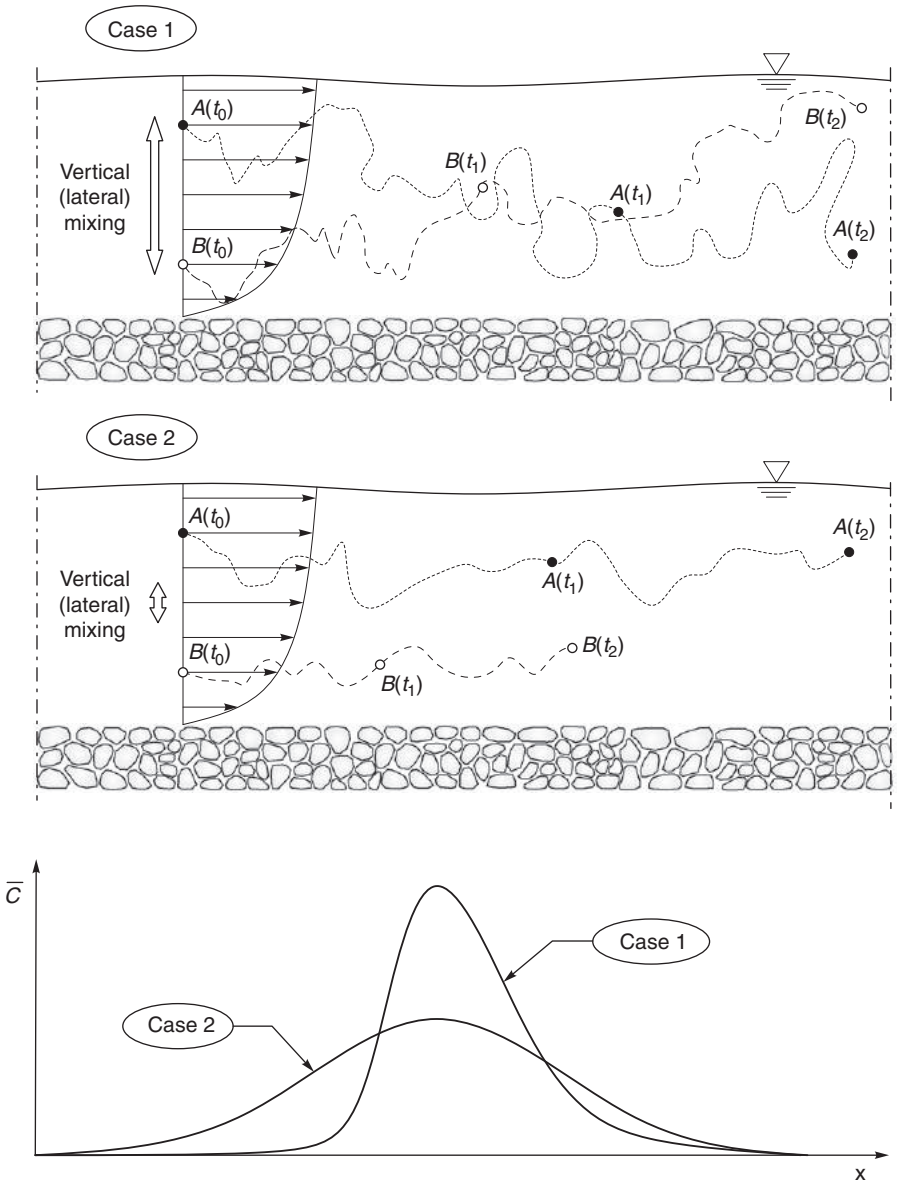


Fig. 3 Schematic of the transport processes affecting shear dispersion. Case 1 shows transport of elements *A* and *B* in a shear flow with high turbulent mixing. Case 2 shows transport of the same elements in a shear flow with low turbulent mixing. The graph shows the response of the two flows to a pulse injection of mass.

therefore mathematically equivalent to advection–diffusion models. For example, in the simplest case of cross-sectional averaging of the physical quantities, the dispersion equation can be reduced to the one-dimensional form:

$$\frac{\partial \overline{C}(x,t)}{\partial t} + \overline{u}(x,t)\frac{\partial \overline{C}(x,t)}{\partial x} = \frac{\partial}{\partial x}\left(D_x^{\mathrm{L}}(x,t)\frac{\partial \overline{C}(x,t)}{\partial x}\right) \quad [13]$$

where D_x^{L} is the longitudinal dispersion coefficient which turns out to be proportional to the spatial variability of the flow field around its mean value and inversely proportional to the turbulent diffusivity. The value of the dispersion coefficient may vary by a few orders of magnitude according to application. The reader is invited to read specialist books and papers to obtain predictions and estimates of the dispersion coefficient.

See also: Behavioral Ecology: Dispersal–Migration. Ecological Data Analysis and Modelling: Modeling Dispersal Processes for Ecological Systems

Further Reading

Elder, J.W., 1959. The dispersion of marked fluid in turbulent shear flow. Journal of Fluid Mechanics 5, 544–560.
Fischer, H.B., 1967. The mechanics of dispersion in natural streams. Journal of the Hydraulics Division, ASCE 93 (6), 187–215.
Fischer, H.B., List, E.J., Koh, R.C., Imberger, J., Brooks, N.H., 1979. Mixing in Inland and Coastal Waters. New York: Academic Press.
Rutherford, J.C., 1994. River Mixing. Chichester: Wiley.
Taylor, G.I., 1954. The dispersion of matter in turbulent flow through a pipe. Proceedings of the Royal Society of London, Series A 223, 446–468.

Predation and Its Effects on Individuals: From Individual to Species

Patricio Lagos, uBiome Inc., Santiago, Chile

Glossary

Aposematism A warning signal present in some species, directed to predators, that informs them about the unpalability or unprofitability of the aposematic individual.

Community Assembly of populations of different species and their interactions, inhabiting the same space and time.

Key species In community ecology, it refers to the species whose effect on the structure of the community is larger than its mere contribution in abundance. The key species is usually a predator that feeds on the dominant competitor (i.e., the individual that dominates in abundance the community in absence of the key species). When a key species is present, the diversity in the community is larger than when it is not present.

Lotka–Volterra equations A set of two models of population dynamics that considers the effect of the predator–prey interaction on the change in abundances of individuals in populations, based on the rates of growth, the effect of predation on prey and how prey consumed contribute to the increase in the number of predators.

Landscape of fear A behavioral trait of an animal, that provides spatially dependent measure of the perceived risk of predation by the individual, and the cost and benefits of foraging or inhabiting the area available.

Population Group of individuals of the same species, occupying the same space and time.

Predation Interaction between two individuals, where one of the individuals benefits from the interaction (i.e., the predator) whereas the other is harmed (i.e., the prey). It includes the carnivory, herbivory, parasitism, parasitoidism, and pathogen–host interaction.

Parasitoidism A particular case of predation, where the adult individual is free-living, but deposits its larvae inside a host. The larvae then feeds on the host and finally kills it.

The study of ecology is, in many aspects, the study of the interactions between individuals. These interactions can differ according to which individuals are positively affected from the interaction and which ones are negatively affected by it. Thus, those cases where both individuals benefit from the interaction are known as mutualism, whereas the case when the interaction is negative for both individuals involved, it is usually known as competition. One important case is positive for one of the individuals, but negative for the other. Some authors refer to these interactions as exploitation, but are more commonly known as predation.

Predation is an interaction between two organisms, where one, the predator, consumes the other the prey, while it is still alive. Predation falls in the category of "exploitation" interactions, where one of the individuals consumes the other. Besides predation, some other examples of exploitation are herbivory (and herbivore that consumes a plant), parasitism (an organism that lives in the tissues of its host, at the expense of it) and parasitoidism (as in parasitism, but the parasite's larvae live in the host's tissues). For some authors, however, predation involves any interaction between species were one consumes partially or totally the other. Here, the focus will be placed in this definition, as it offers a broader spectrum of scenarios to describe the interaction.

Predation, as a consumption interaction, has obvious effects at the individual level, as can reduce the prey's fitness to zero. But does not stop there. In fact, the effects of predation scale up to the population level, determining distinctive patterns in the dynamic of populations. Furthermore, predation plays a role in defining the structure of communities. Lastly, as the predator–prey interactions have strong effects on the fitness of both prey and predators, it is probably the strongest way natural selection has to drive the evolution of the species, determining the frequency of important traits in populations.

Predation at Individual Level

Predator–Prey Interactions: Avoiding Predation

An important part of an animal activities are preventing been consumed by others. Similarly, carnivores spent their day searching for prey to consume. A variety of strategies have evolved for both predator and prey to optimize (i.e., maximizing the benefits and minimizing the costs) either their hunting performance or surviving predation. Looking at the prey's perspective may result very useful in order to understand how important predation is, as we can see how predator push the evolution of species in one or another direction.

Morphological, physiological and behavioral traits are usual targets for natural selection, leading to in many cases, sophisticated mechanisms used by prey to avoid or minimize the risk of predation. The body armor of the armadillo, thorns of the thorny dragon and shells in Mollusca are some examples of how natural selection has favored morphological traits that provide a defense against the attack of predators. Animals that rely on these defense incur in an important energetic cost in order to produce the structures, and energy must be allocated to this rather than growth or reproduction. For example, under risk of predation, tadpoles of the species *Rana dalmatina* develop larger and more muscular tails, enabling them to escape further and faster (Teplitsky *et al.*, 2005). However, tadpoles under risk of predation suffered of reduced rates of growth (Van Buskirk, 2000).

Similarly, different physiological attributes in prey individuals are affected by predation pressure, so much so that the concept of "non-consumptive effects of predation" has been coined. For instance, the resting metabolic rate (i.e., the basal level of metabolism of an organism) has shown to increase in the presence of a predator, or even some predatory cues. This poses a direct cost to prey, which may even cascade up to higher trophic levels (Hawlena and Schmitz, 2010). Similarly, presence of predators triggers a reduction in movement rates and reduction in food consumption in prey individuals (higher giving-up densities), imposing an energetic cost of deploying some anti-predator strategies (Brown, 1988). But this is not only true for animals. In fact, plants' defenses against predators are mostly chemical. Is the well-known case of the wheat Triticum uniaristatum, which produces hydroxamic acid when attacked by Aphids (Abrahamson, 1975). Many plants have trichomes in their leaves, which act as both physical and chemical defense against herbivory (Agrawal and Rutter, 1998).

Furthermore, an important part of the behaviour of most animal species have evolved in order to reduce the risk of predation. Several examples exist of such behavioral traits. Some species will escape when facing a predator, and the cost of escaping is traded-off with the cost of, for example, abandoning a foraging patch. Ydenberg and Dill (1986) proposed the theoretical curves of cost of remaining foraging and cost of escaping for a foraging prey in the presence of a predator. For 30 years the curves of cost remained experimentally unmeasured. Lagos and collaborators (2014) designed and conducted a laboratory experiment where the theoretical curve of cost of remaining foraging was tested for the first time, and matched was predicted by Ydenberg and Dill (1986).

Some species, however, hide in refuges to escape predation. The use of a refuge, is similar to escaping from a predator, as there are cost associated to both remaining hiding (because of the loss of fitness-increasing activities) and leaving the refuge (because of the persistent risk of predation). Martín and López (1999) proposed an economic model, similar to Ydenberg and Dill's (1986) for the optimal time a prey should remain hiding. To the date, however, the curves of cost have not been measured empirically.

Predation at the Population Level

As predators consume prey, they effectively remove individuals from populations. This effect, very important in inter-individual interactions, is also determinant in the abundance of individuals within populations and thus, the dynamics of both predator and prey populations. In the late 1920s, Alfred Lotka and Vito Volterra studied how the rate of population growth in a population changes as function denso-dependent predator and prey factors (Eqs. **(1)** and **(2)**).

$$dN/dt = rN - pNM \tag{1}$$

$$dM/dt = cpNM - dM \tag{2}$$

In Eq. **(1)**, the rate of change in the population of prey over time (dN/dt) is direct function of both the rate of growth of the prey population (rN, where r is the reproductive rate and N is the population size) and inverse function of the effect of the predators (pNM, where p is the rate of predation, N is the prey's population size and M is the predator's population size). Similarly, the rate of change in the predator population over time (dM/dt) is direct function of the number of offspring the predators produce per number of prey consumed ($cpNM$, where c is the rate of prey–predator conversion, p is the rate of predation, N is the prey's population size and M is the predator's population size), and inverse function of the rate of death of predators (dM, where d is the rate of predator's death and M is the predator's population).

The interrelatedness between the dynamics of the populations of both prey and predators can give rise to cyclic dynamics in time, were the growth in the population of prey will be followed by an increase in the population of predators, as prey become more abundant, predators can capture more prey items per unit of time, and the excess of energy available for predators translate into increased reproductive rates. This, however, has a limit, and when the population of predators becomes large enough, their consumption of prey overcomes the rate of growth of prey, leading to a decrease in the size of the prey's population. Having less prey items available for consumption, the number of offsprings produced in the predator population decreases, reducing its size. Under a constant environment and migration rates, this process may lead to a cyclic dynamic that can be sustained in time. This kind of population processes are very well documented, and remarkably examples can be cited. Is the case of the lemming (*Dicrostonyx groenlandicus*), which is preyed on by four different species, the stoat (*Mustela erminea*), the arctic fox (*Alopex lagopus*), the snowy owl (*Nyctea scandiaca*), and the skua (*Stercorarius longicaudus*). In this case, a regular 4-years cycle has been predicted and observed (Gilg *et al.*, 2003). Another noteworthy example of cyclic dynamics the one of the larch budmoth (*Zeiraphera diniana*), which was supposed to be caused by an intra-population feedback. However, Berryman (1996) noted that the regular oscillation of its populations were caused by Diptera and Hymenoptera parasitoid. A parasitoid is a special case of predation, were the free-living adult lays eggs inside another animal. The eggs then hatch, and the larvae feeds on the individual, ultimately killing it.

The effects of predation are not necessarily as strong in the population as they are on individuals though. Two reasons have been proposed to explain this. Firstly, predators usually go after the weakest individuals, as they are easier targets. These individuals are inherently less likely to contribute much to the population, in terms of how many offsprings they leave. Secondly, these targeted prey might well being older individuals, which already reproduced, thus removing these individuals from the populations will not have a deep impact on the overall dynamic of the population. This argument vanishes though if the preferred targets are younglings, and the effect of predation will be stronger in the population when these individuals are consumed, as they have not yet reproduced. But they might not be as likely preyed on, as predators will not profit as much from young individuals, because of their reduced body size.

Predation at Species Level: Evolution of Traits Driven by Predation

Predation acts strongly and negatively on prey's fitness. Therefore, it is only logical that traits that reduce predation (or predation risk) will readily be selected for by natural selection. Many different anti-predator strategies have evolved in order to reduce the risk of predation. And many of these strategies involve using the color or shape to deceive would-be predators. For instance many animals have colors that match their surroundings, reducing the contrast between the animal and their environment. This camouflage makes them virtually invisible in many cases. Such is the case of the Darwin's frog (*Rhinoderma darwinii*), whose brown or green color perfectly matches the colors of the rainforests were this species occurs. Very famous is the example of the peppered moth (*Biston betularia*). This moth has two colors, black and white. Before the industrial revolution, the white ones were more abundant than the black ones, because the trees were they live have light colors, so the black moths contrasted much more than the white ones, making them frequent target of predators. However, with the industrial revolution, the amount of atmospheric CO_2 increased, and the bark of the trees became darker. This change made the dark moth more cryptic now, and the white ones became more frequent target of predation. Nowadays, the black moth is more frequent than the white ones, as the white moth became more conspicuous and is more frequently targeted by predators.

Seemingly paradoxical, it has been proposed that some animals have evolved bright colorations as a response to predators. This is known as aposematism. The idea is that bright colors are used to inform the possible predators about the unpalatability of the prey (Joron, 2009). In fact, aposematic species are usually venomous, or produce some kind of toxin or are somehow distasteful to predators (Mappes *et al.*, 2005). Thus, the bright color in aposematic species actually increase survival, despite the fact that makes them easier to detect.

Over evolutionary time, anti-predator traits might change, as more efficient variant will be naturally selected. However, the same can happen in predators. New variations in hunting strategies will randomly appear in the populations, and those better at foraging or capturing prey will also be naturally selected, leading to further changes in prey anti-predatory traits. And so forth. This reciprocal evolutionary tendency to overcome the enemy's strategies, an arms race of sorts, is known as "coevolution." For instance, there is a group of cacti, the "senita" *Pachycereus schottii* produces a chemical of the alkaloid family, which is toxic, even lethal, for most flies, except for the fruit fly *Drosophila pachea*, which has acquired a resistance to this toxin (Heed, 1978). Another know example of coevolution is the ability of ground squirrels (*Otospermophilus beecheyi*) to resist the venom of rattlesnakes (*Crotalus oreganus*). The squirrels display several anti-predator strategies that seem especially useful against snakes, which suggests a synchronous, joint evolution (Holding *et al.*, 2016).

The relationship between a pathogen and its host is an interaction were one of them, the pathogen, benefits from colonizing the host, whom is negatively affected by the pathogen. Thus, pathogen–host interaction can be considered as a special case of predation. Pathogen–host systems have been very useful to test the hypotheses of coevolution, as the short generational times of most pathogens allow for quick evolutionary changes. For instance, Borghans *et al.* (2004) performed simulations aiming to explain the high polymorphism found in the MHC (major histocompatibility complex), responsible for the immunity in vertebrates. They found that coevolution by itself is capable of generating the variability observed in the MHC. Similarly, *Mycobacterium tuberculosis*, the responsible organism for tuberculosis, has accompanied human ever since they left Africa the first time. In a literature review, Gagneux (2012) showed that *M. tuberculosis* and other closely related pathogens that cause tuberculosis in human have a phylogeographic population structure associated with different human populations. This suggests that coevolution might explain the specificity in the pathogenicity of *M. tuberculosis* with different human populations.

Even though predator and prey change their strategies over time, this does not necessarily mean that their rate of success will increase. This is known as the "Red Queen hypothesis" for coevolution, as a reference to Lewis Carroll novel, where Alice and the Red Queen ran for a long period and did not move from where they stood. In 2011, Morran and collaborators conducted an experiments to test the "Red Queen" hypothesis for coevolution using a bacterial pathogen (*Serratia marcescens*) and its host, *Caenorhabditis elegans*. If the Red Queen hypothesis was correct, it was expected that *C. elegans* would favor outcrossing (i.e., reproducing with another individual) rather than asexual reproduction, and this is exactly what the researchers found, but not just that. In fact, they observed that individuals that selected asexual reproduction were driven to extinction by the pathogen.

Predation at Community Level

A community (i.e., a group of populations inhabiting the same space in the same time) can be described in several ways. One aspect to it is to describe the structure of the community in terms of the interactions between individuals, and ultimately, between populations. Key concepts in community ecology are the transfer of both energy and matter. This transfer occurs mainly by trophic interactions (i.e., interactions were one individual eats the other). The observation of this kind of interactions in a community will quickly reveal a network of interactions were individuals are either producers or consumers. The summary of these trophic interactions is known as a food web. Paine (1969) proposed the idea that one, or a few species can have a disproportionate effect in the structure of their community. Paine called these the "key species." He observed that key predators oppose to the dominant competitors (i.e., those species that dominate the community in terms of abundance). In reducing the abundances of the dominant competitors, new species are now able to colonize the community. This effect of higher trophic levels modifying the structure of the communities is known as top-down regulation.

Paradoxically, the fact that key predators can control the structure of the community by removing the dominant competitor means that a predator's effect on the community can lead to an increase in the community's diversity. This phenomenon has been

observed in several communities. A good example comes from the intertidal communities, were sea stars (*Pisaster* or *Heliaster*) are active predators of many of the inhabitants of the intertidal, mainly bivalves and snails. When sea stars are present in the community, the number of species is high (i.e., high diversity). However, when the sea stars are removed from the community, the structure of the intertidal changes drastically, as the diversity is reduced and the dominant competitor takes over the community (Paine, 1966, 1969). Thus, in this example is easy to see the important effect predation has in determining the structure of the biological communities.

Predators in the communities do not need to actually consume individuals in order to have an effect on the community. In fact, besides the lethal effects, there are also non-consumptive effects of predation on the communities, and individuals will occupy space and time so that the risk of predation is minimized. This idea that predation determines the use of the habitat by individuals in a community is known as the "landscape of fear" (Brown *et al.*, 1999). The concept of the landscape of fear was first empirically observed by Laundré and collaborators (2001), when a population of wolves (*Canis lupus*) was reintroduced in Yellowstone park after 50 years. The researchers studied the anti-predator and foraging behaviors of elks (*Cervus elaphus*) and bisons (*Bison bison*) in areas with and without wolves. They found that in places where wolves were reintroduced, female elk and bisons had higher rates of vigilance and reduced rates of foraging than the same species in areas were wolves had not reached yet. These are results predicted by the idea of changes in the landscape of fear.

See also: Behavioral Ecology: Herbivore-Predator Cycles; Anti-Predation Behavior

References

Abrahamson, W.G., 1975. Reproductive strategies in dewberries. Ecology 56 (3), 721–726.
Agrawal, A.A., Rutter, M.T., 1998. Dynamic anti-herbivore defense in ant-plants: The role of induced responses. Oikos 83, 227–236.
Berryman, A.A., 1996. What causes population cycles of forest Lepidoptera? Trends in Ecology & Evolution 11 (1), 28–32.
Borghans, J.A., Beltman, J.B., De Boer, R.J., 2004. MHC polymorphism under host-pathogen coevolution. Immunogenetics 55 (11), 732–739.
Brown, J.S., 1988. Patch use as an indicator of habitat preference, predation risk, and competition. Behavioral Ecology and Sociobiology 22 (1), 37–47.
Brown, J.S., Laundré, J.W., Gurung, M., 1999. The ecology of fear: Optimal foraging, game theory, and trophic interactions. Journal of Mammalogy 80 (2), 385–399.
Gagneux, S., 2012. Host–pathogen coevolution in human tuberculosis. Philosophical Transactions of the Royal Society B 367 (1590), 850–859.
Gilg, O., Hanski, I., Sittler, B., 2003. Cyclic dynamics in a simple vertebrate predator–prey community. Science 302 (5646), 866–868.
Hawlena, D., Schmitz, O.J., 2010. Herbivore physiological response to predation risk and implications for ecosystem nutrient dynamics. Proceedings of the National Academy of Sciences 107 (35), 15503–15507.
Heed, W. B. (1978). Ecology and genetics of Sonoran desert Drosophila. In Ecological genetics: The interface (pp. 109–126). Springer, New York, NY. ISO 690.
Holding, M. L., Biardi, J. E., & Gibbs, H. L. (2016). Coevolution of venom function and venom resistance in a rattlesnake predator and its squirrel prey. In Proceedings of the Royal Society B: Biological Sciences (vol. 283, No. 1829, p. 20152841). The Royal Society.
Joron, M., 2009. Aposematic coloration. In: Encyclopedia of insects, 2nd edn. New York: Academic Press, pp. 33–38.
Lagos, P.A., Ebensperger, L.A., Herberstein, M.E., 2014. A quantitative test of the "economic" and "optimal" models of escape behaviour. Animal Behaviour 97, 221–227.
Laundré, J.W., Hernández, L., Altendorf, K.B., 2001. Wolves, elk, and bison: Reestablishing the "landscape of fear" in Yellowstone National Park, USA. Canadian Journal of Zoology 79 (8), 1401–1409.
Mappes, J., Marples, N., Endler, J.A., 2005. The complex business of survival by aposematism. Trends in Ecology & Evolution 20 (11), 598–603.
Martín, J., López, P., 1999. When to come out from a refuge: Risk-sensitive and state-dependent decisions in an alpine lizard. Behavioral Ecology 10 (5), 487–492.
Morran, L.T., Schmidt, O.G., Gelarden, I.A., Parrish, R.C., Lively, C.M., 2011. Running with the red queen: Host-parasite coevolution selects for biparental sex. Science 333 (6039), 216–218.
Paine, R.T., 1966. Food web complexity and species diversity. The American Naturalist 100 (910), 65–75.
Paine, R.T., 1969. The Pisaster-Tegula interaction: Prey patches, predator food preference, and intertidal community structure. Ecology 50 (6), 950–961.
Teplitsky, C., Plénet, S., Joly, P., 2005. Costs and limits of dosage response to predation risk: To what extent can tadpoles invest in anti-predator morphology? Oecologia 145 (3), 364–370.
Van Buskirk, J., 2000. The costs of an inducible defense in anuran larvae. Ecology 81 (10), 2813–2821.
Ydenberg, R.C., Dill, L.M., 1986. The economics of fleeing from predators. In: Advances in the Study of Behavior, vol. 16. Orlando: Academic Press, pp. 229–249.

Further Reading

Begon, M., Twonsend, C.R., Harper, J.L., 2006. Ecology. From individuals to ecosystems, 4th edn. Oxford: Blackwell publishing.
Brown, J.S., Laundré, J.W., Gurung, M., 1999. The ecology of fear: Optimal foraging, game theory, and trophic interactions. Journal of Mammalogy 80 (2), 385–399.
Davies, N.B., Krebs, J.R., West, S.A., 2012. An introduction to behavioral ecology, 4th edn. Chichester: Wiley-Blackwell publications.
Molles, M.C., 2005. Ecology, concepts and applications, 3rd edn. New York: McGraw-Hill.
Pianka, E.R., 1978. Evolutionary ecology, 3rd ed. New York: Harper & Row publishers.
Ydenberg, R.C., Dill, L.M., 1986. The economics of fleeing from predators. In: Advances in the Study of Behavior, vol. 16. Orlando: Academic Press, pp. 229–249.
Abrams, P.A., 2000. The evolution of predator–prey interactions: Theory and evidence. Annual Review of Ecology and Systematics 31 (1), 79–105.

Succession and Colonization

Chryssanthi Antoniadou, Eleni Voultsiadou, and Chariton-Charles Chintiroglou, Aristotle University of Thessaloniki, Thessaloniki, Greece

Glossary

Deterministic process A process in which the future state is completely determined by present and past states.

Disturbance Any temporary change in average environmental conditions severely disrupting an ecological population or community, such as fire, flood, ice-scouring; such events create open space, available for colonization.

k-selection The type of natural selection experienced by populations that live at carrying capacity or maximal density in a relatively stable environment.

Metapopulation A group of local populations of a species in patches, inter-linked through dispersal.

Niche models Models based on the niche concept, that is, the ecological space occupied by a species, and the occupation of this species in a community.

Patch dynamics An ecological concept according to which communities consist of a mosaic of patches, that is, discrete areas of variable size, shape, composition, history, and disturbance regimes, making up a landscape.

Resilience The ability of a community to return in its former state after disturbance.

r-selection The type of natural selection experienced by populations that are undergoing rapid population increase under high resource availability.

Stochastic process A process where future state is based on probabilities or randomness.

Supply-side ecology An ecological concept according to which population dynamics are driven by immigration from external sources.

Introduction

Communities are assemblages of living organisms in specific areas or habitats; their structure changes in time and the processes of succession and colonization are recognized as the main driving factors. Community change has important implications for management and conservation and great scientific effort has been historically devoted to ascertain relevant patterns and predict future states.

Colonization: A Precondition for Succession

Colonization (L. *colere*, to inhabit) is a very generic term, applied to a wide spectrum of scientific fields, such as archeology, astronomy, biology, cosmology, history, medecin, paleontology, and parasitology. Concerning biology, the term is generally used in community ecology, describing the process by which a species is spreading into a new area. The process gained publicity under the predictions of the Island Biogeography Theory (MacArthur and Wilson, 1967) and the concept of metapopulation (Levins, 1969), which assumes spatially separated populations that interact through successful immigration. Colonization may occur at a broad range of scales, from microcosmic to macrocosmic. Perhaps the lowest scale refers to the biofilm formation, that is, the settlement of microorganisms on bare substrata after molecular fouling, a foundational process for higher-scale colonization, especially in humid or aquatic environments. Although the term usually refers only to the natural spread of species, it could possibly also apply to the field of invasion ecology (Davis and Thompson, 2000). Thus, as previous authors suggested, colonizers of different types can be distinguished under three simple criteria: dispersal distance, uniqueness, and environmental impact. Species that are nor extending their range of occurrence when colonizing an area, are considered as successional colonizers indifferent of whether they originated from a local species pool and their possible impact on the new environment. Regardless of dispersal distance, species that are novel in an area (range extension) may or may not cause severe environmental impacts, and are characterized as novel, invasive or noninvasive colonizers, respectively.

The combined effect of colonization and extinction is the main driver of ecosystem structure and function; the balance between these two opposite processes determines biodiversity levels and keeps ecological systems in a dynamic, self-organizing, and multistable state. The colonization abilities vary between species mainly according to their life-history traits, but are also affected by abiotic environmental factors. Species dispersal is the first step of colonization whereas the second, and maybe the most decisive step, is establishment and persistence (Lohmus *et al.*, 2014). Dispersal skills differ between species depending on the intrinsic reproductive strategies and motility patterns. Furthermore, the original presence of a species in a habitat, that is, local species pool, and landscape connectivity affect species dispersal. Life history traits also determine establishment success, according to the habitat's specific environmental conditions and presence of other species. The outcome of the above processes will determine species persistence over time. Therefore, the colonization success of a species seems to be a function of time.

Whether the environmental factors prevail in the determination of species colonization success, against the biotic ones, remains unclear. Accordingly, two opposed hypotheses, the habitat filtering metaphor or abiotic filtering and the biotic filtering, have been proposed to highlight main drivers and explain species cooccurrence (Purschke *et al.*, 2013). In the former, particular environmental conditions will screen for functionally similar species, while in the latter functionally dissimilar species will cumulate by biotic screening processes such as the competitive exclusion and resource partitioning. Both above hypotheses premise deterministic processes in contrast to the stochastic view of the 'trait-neutral' approach (Hubbell, 2001, in Purschke *et al.*, 2013). Despite the strong ongoing debate on the stochastic or deterministic nature of colonization and succession, the above theoretical approaches have founded the most classical models explaining succession (i.e., inhibition and facilitation/tolerance models, see below).

Succession: Meaning and Relevant Concepts

The term "succession" refers to linear, directional changes in species populations that start from an initial point and continuously modify community pattern. These changes are generally slow and perceptible over long time, resulting from a dynamic balance of species intrusions and exclusions, producing species substitutions over time. As time proceeds, the biotic community tends to a more stable formation or climax state, according to the conventional succession theories. In the climax state, community structure still varies, but within natural variability confines, finally reaching a dynamic equilibrium point outlined by climatic conditions. Despite the simplistic definition, succession is a particularly complex process, operating in various spatiotemporal scales coupled with environmental variability.

There are three possible different types of succession: (i) the degradative or decompositional succession, where scavengers and other microorganisms assiminate dead organic matter, for example, the insect succession on carrion, being widely applied in forensic medicine; (ii) the allogenic succession, where external, abiotic factors, varying from massive disturbance to temporal changes of environmental factors, modify the geophysical conditions of an area driving community succession, for example, volcanic eruptions or seasonal phytoplankton cycles; and (iii) the autogenic succession, where biotic processes, that is, the organisms themselves, modify existing conditions affecting, positively or negatively, other organisms and themselves, for example, the forest succession. Allogenic and autogenic succession, however, may occur simultaneously and are not mutually exclusive; thus, the above terms may apply to the factors involved in succession rather than succession per se (Tansley, 1935).

All succession types, apart from the degradative one, require a vacant space and are typically separated in (i) primary and (ii) secondary succession according to the initial absence or presence, respectively, of living organisms (or biological legacies, see relevant section of succesional theories, below) in the newly available space. Most sucessions are assigned to the latter case involving recovery of sites after disturbance events. However, in disturbances of extreme magnitude, such as glacial retreats and volcanic eruptions, which totally eradicate all organisms and their propagules, primary succession occurs. Typically, primary succession takes places in much slower rates than the secondary one, due to the complete initial absence of life (sterile substrate). The terms "primary succession" and "secondary succession" predominate in recent terminology, due to the problematic application of the terms "allogenic" and "autogenic," as emphasized above.

Another differentiation, mostly applied on terrestrial ecosystems, is the progressive versus retrogressive succession. In the former case, species diversity and biomass increase over time by contrast to the latter, in which the above biocoenotic variables are decreasing. Progressive is the typical form in most directional succession processes. Very few retrogressive successions have been described, such as the case of coastal sand dunes vegetation.

Ecological succession has three intermingling components: mechanisms (species dispersal and establishment), stages (initial, early, mid, and late) and trajectories or pathways (convergent, divergent, and circular). A key concept in succession is seral stage, that is, each distinct community type within succession, from the very beginning to the dynamic equilibrium state. The occurrence of a species within a seral stage is determined by its life history features and natural changes induced by succession. As a general rule, good colonizers are bad competitors and thus nonsurvivors. Early successional species typically follow r-strategy: rapid growth rate and resource uptake, low mature size/biomass, high reproductive output without parental care, and short lifespan.

A particular type of succession is the cyclic one, where the seral stages rotate in circular way; shrub-dominated communities often undergo cyclic succession as the loss of a canopy-forming species opens space to novel colonizers, including their own juveniles, which will in turn become overstory plants.

Finally, relevant to succession is the concept of chronosequence, based on the fact that most ecological communities are not spatiotemporally homogenous; they rather constitute a mosaic of patches in different successional stages (Menge, 1975), under the driving process of disturbance (Sousa, 1984). Such mosaic patterns, representing different recovery stages, and thus having a successional relationship, are called chronosequences, in contrast to toposequences, where the mosaic aspects are due to topographic divergence.

Theories on Ecological Succession and Driving Mechanisms

Succession theories emerged from the two extreme views of community structure; that is, whether the relationships between populations of living organisms in a specific area or habitat are obligatory or not (McCook, 1994; Krebs, 2009). In the first case, community can be viewed as a kind of "superorganism" where species populations are organized in a complex network, a "web of

life," and accordingly, any change to a population will induce seriatim results reflected to the entire community. In the second "individualistic" view, each species functions under its own rules, and they cooccur simply due to their similar physico-chemical niche requirements; accordingly, the community represents a random collection of individual populations with minimal integration. These opposed views triggered a strong debate among ecologists, but, as often happens, reality seems to lie somewhere between the above extremes.

Several attempts have been made towards understanding the processes of succession (McCook, 1994; Raevel *et al.*, 2012), among which we should note the recent work of Pulsford *et al.* (2016) thoroughly revising successional theories. Most theories are strongly related with post-disturbance events as ecologists have long ago realized the importance of predicting how ecosystems will change after disturbance. Alternative theories are based on niche models and patch dynamics.

The first successional theories emerged from the field of plant ecology, based on two opposed ideas: Cowles' dynamic view of vegetation and Clements' static climax state view. Cowles considered succession as "*a variable approaching a variable, not a constant*" (Cowles 1901 in Pulsford *et al.*, 2016) due to the dynamic nature of communities under a variable abiotic environment and biotic interactions. After his pioneer work on sand dunes, he concluded that succession is a complex dynamic process without a certain end point. Sharply contrasting to the above, Clements visualized succession as a strictly linear determinist process leading to a predefined endpoint, the climatic climax state (Clements 1916 in Pulsford *et al.*, 2016). Clements' degraded the importance of abiotic in favor of biotic processes and gained wide acceptance among his confreres; however, his ideas are no longer in use. Although not directly referring to his work, Gleason (1927) supported Cowles' ideas; he regarded succession as being directed by individual plant characteristics and identified migration and environmental selection as the main driving factors. His work has been overlooked at the time but gained support after 1950; Hutchinson (1951) realized the importance of life-history traits and invented the concept of "fugitive species," that is, species that are good dispersers but bad competitors, thus being the first to establish after disturbance and able to resist further invasion by other fugitives, but outcompeted at the end by slower-dispersing and superior–competitor species. Whittaker (1953) also accepted Gleason's ideas and introduced the pattern climax concept, in which the endpoint is a spatially variable pattern of species. Egler (1954) proposed two models: Relay Floristics, fitting with Clements' ideas, and Initial Floristic Composition as an alternative. In his first model, he suggested a deterministic progression of seral stages (alternation of specific species assemblages) ending to the climax state (Clementsian succession). In his second model, he claims that various species invade a barren area and grow according to their life-history traits, with fast-growing/short-lived species (r-selected) dominating during first stages, but outcompeted by slow-growing, long-lived species (k-selected) as succession proceeds, up to an "equilibrium" state; further invasion may occur at any stage altering the above pathway. He proposed both models as ideal cases, actually functioning in conjuction and influenced by biotic interactions (e.g., predation, competition).

Connel and Slatyer (1977) revised the existing successional theories and suggested three possible models according to the effect of the first colonizers to the next ones: Facilitation, Tolerance and Inhibition (**Fig. 1**). In Facilitation model, only pioneer species are able to initially colonize a barren area; these species condition habitat making it more suitable to other species than to themselves. Hence, species replacement is gradual, ordered and predictable, following a directional process towards an equilibrium state. In Tolerance and Inhibition models, any species is assumed able to colonize a barren area, without the prerequesition of conditioning species; after colonization, the occurrence of a species may be indifferent or prohibitive to other species, respectively. Thus, species replacement is not necessarily ordered and predictable, and succession proceeds without converging towards a specific climax state. The Facilitation model is founded on Clements' succession and subsequent Floristic Relay, whereas the Tolerance and Inhibition models paraphrase Egler's Initial Floristic Composition, by accepting that both early and late successional species are able of initial colonization, the former being inferior competitors according to the first model, whereas any early colonizer persists until dead or damaged, in the latter. Thus, life history traits are identified as the prominent driving factors of succession: species longevity in the Inhibition model and species competitiveness in the Tolerance model. Connel and Slatyer's models have been converted by Fox (1982) to fit animal succession under the Habitat Accomodation model, which relates the structure of vegetation to the post-disturbance recovery of associated fauna. Habitat Accomodation model predicts that an animal species will be established in a defaunated area when habitat conditions (i.e., vegetation) fulfill its requirements; it will persists as long as floristic conditions remain favorable, but it will decline or vanish, when floristic compositions turn out to be adverse,

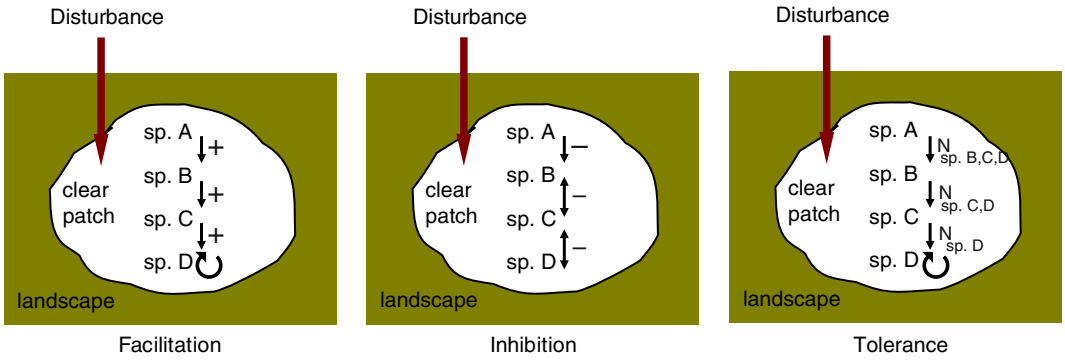


Fig. 1 Schematic representation of Connel and Slatyer's successional model (1977).

getting outcompeted by other better-fitted species. Connel and Slatyer's models are classical in modern ecology (Krebs, 2009). However, they are considered as empirical descriptions of the succession outcome rather than as mechanistic explanations (McCook, 1994).

All successional models presented so far, apart from the Clementsian model, recognize the importance of life history traits. This inspired researchers to explain succession by the individual species response to resource or stress gradients according to life-history traits (McCook, 1994). Among various attempts are the CSR theory (Grime, 1974) and the Logistic Simulation model (Whittaker and Goodman, 1979). The former classifies species in functional types (competitive, stress tolerant and ruderal) and suggests their different contribution under different stress, competition and disturbance levels, as succession proceeds, while the latter substitutes CSR functional types with relevant carrying capacity categories (adversity, exploitation and saturation selection) along environmental gradients (Whittaker and Goodman, 1979).

Succession has been also attributed to limiting resources availability under the predictions of Resource Ratio hypothesis (Tilman, 1985). Synthesizing existing theories, McCook (1994) proposed the "Strategic Resource Allocation" model, in which the specific adaptive strategy of a species is dictated by the particular balance of allocating resources to different physiological functions. Resource and life-trait models have given place to more advanced but similar approaches, grouping species according to shared life-history traits (Raevel *et al.*, 2012). Such examples are the "vital attributes" (Noble and Slatyer, 1980), "functional types" (Gitay and Noble, 1997), and "critical life cycles" (Whelan *et al.*, 2002) concepts. Albeit a promising approach, the definition of these concepts in practice remains challenging, and a general trait-based model is still missing (Pulsford *et al.*, 2016).

Westoby *et al.* (1989) introduced State and Transition models (STMs) to describe ecosystem dynamics, accepting that multiple stable communities can occupy ecological sites. Individual states may be identified by relatively large differences in functional groups and ecological processes and thus, in community structure. Transition between states can be gradual or rushed. STMs have been used to assess ecosystem response to disturbance by predicting the future state of community structure under specific conditions driving the system to one of the possible alternative "equilibrium" states. Thus, it is important to identify thresholds and ecological resilience of individual states to stressors, as well as factors involved in the transitioning of a site between potential alternative states. STMs are advanced succession-retrogression models that accept the existence of multiple equilibria and the return to equilibrium following disturbance (Briske *et al.*, 2005; Bestelmeyer *et al.*, 2017). STMs (**Fig. 2**) can be sequencial, radiated or of maximum connectivity (Phillips, 2011); they provide a promising tool for managers, especially if their predictive power is increased (Pulsford *et al.*, 2016).

A radically different approach to interpret succession is through stochastic models, such as the Markovian chains (Horn, 1975; Van Hulst, 1979). These models predict the transition of community structure from a set of alternative states, where any future state is always dictated by the previous one, to a constant outcome under the same replacement probabilities (following Clementsian successionl view). They may also apply the Neutral metacommunity theory, which considers ecological drift as the main driving factor, to predict transitions (Hubbell, 2001). Appropriate data, however, are generally lacking and these approaches suffer from oversimplifications. Markovian models are non-explanatory, entirely based on empirical estimates of species replacement probabilities, even if proven predictive.

An ultimate interpretation of succession, strongly related to the disturbance theory, is through the concept of biological legacies defined as "*living organisms that survive a catastrophe; organic debris, particularly the large organically-derived structures; and biotically derived patterns in soils and understories*" (Franklin, 1990). Obviously, in most cases, some survivors or residuals remain after disturbance (i.e., biological legacies) and manipulate succession pathways and recovery rate. The impact of biological legacies on subsequent succession by influencing recruitment and accelerating the rehabilitation of ecosystem functioning has been only recently acknowledged (Pulsford *et al.*, 2016), although noted in the early studies.

A carefull reader has already noticed that many of the theories presented above are founded on the same basic ideas. Redundancies in successional theories have been thoroughly revised by Pulsford *et al.* (2016) who synthesized relevant information to propose a framework of a rationalized disturbance theory. Within this framework, a number of mechanisms may account for the transition of a community (represented by a continuum of alternative states) to another state during succession following disturbance. Disturbance activates stabilizing procedures (resilience, biological legacies and inhibitory effects), whereas

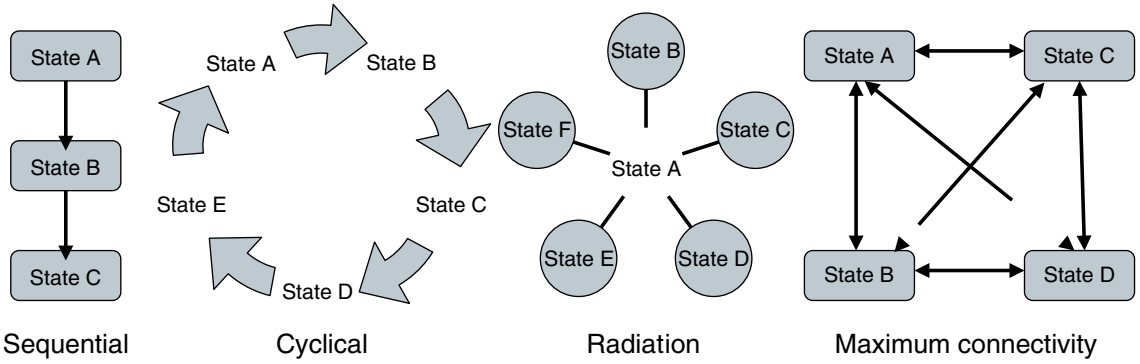


Fig. 2 Graphical representation of State and Transition Models STMs (according to Phillips, 2011).

species traits under varying resource allocation influence possible pathways through pulse events (rapid shift), stochastic drift, facilitation, competition, initial composition and further operation of stabilizing processes. This conceptual framework may assist in the interpretation of successional studies in the future.

General Trends in Ecological Succession

Succession is a central issue in ecology; most relevant studies have addressed various aspects of vegetation dynamics, focusing on major structural components of the communities involved. Through these efforts, and despite the variety of succession types identified and/or the uniqueness of individual successional processes, some common patterns may be observed. When disturbance initiates succession, different pattterns may occur; each possible pattern has its own specific rate and is defined by the outcome of colonization and species interaction processes. Estimation of the successional rate, or otherwise the required time-length for a landscape or seascape to regain its pre-disturbance functions, is critical (Anderson, 2007). Rates are measured through coverage, biomass, species richness, species turnover or by the time required for the community to reach a certain stage. Successional patterns may lead to convergence or divergence towards a "mature," undisturbed community composition in a given area, depending on whether biological processes, such as competition, or biological legacies overcome initial stochasticity of colonization and environmental variability (del Moral, 2007).

Among pioneers, Odum (1969) identified possible ecosystem traits that are expected to vary during succession, according to which alternative community states are conventionally divided into early and late ones. Early successional stages are characterized by the presence of r-selected species (opportunistic, pioneer, fugitive, and early colonizers) having short life cycles, rapid growth rate and small-size, such as grasses. These species are later replaced by larger forms with longer life cycles, that is, k-selected species, such as trees. A shift in energy allocation from reproduction to growth is, therefore, observed during succession. Apart from the very early successional stages, both kinds of species may coexist, depending on resource availability. K- and r-strategies are extreme ends along a continuum and so, many species exhibit intermediate traits. Biomass, coverage, stratification and structural complexity (architectural variation of growth forms) are increasing during succession, overall traping more energy as nutrients are stored as biomass rather than in soil or liquid environment. Foodwebs are becoming more complex with increasing input of litter and organic matter, so that wastes and nutrients are more efficiently recycled. On the other hand, net primary production may decline due to rising respiration and/or nutrient limitation. The effect of macroclimate is gradually reduced in later stages as local conditions are stabilized. In terrestrial succession, the environment becomes more mesic, due to higher soil moisture retention (as soil depth and texture improves), shrinkage of diurnal temperature/humidity fluctuations (as canopy closes) and buffering of seasonal precipitation (Barbour *et al.*, 1999). Species diversity will typically increase during succession from the early to late stages, with the highest values attained in intermediate or more advanced stages and declining towards the "terminal" or dynamic equilibrium state; the same happens with other traits, such as biomass and complexity. Diversity has been traditionally considered as positively related to ecosystem's stability, but such a relationship remains fuzzy. This is partly due to the different definitions of stability, generally falling into two categories, those based on the dynamic stability of the system and those based on its ability to resist change (McCann, 2000). Despite the breadth of definitions, the ecological theory emphasizes equilibrium stability (ability to return to equilibrium after small disturbance) and equilibrium resilience (ability to quickly return to equilibrium after disturbance), which are supposed to increase and decrease, respectively, in late successional stages. However, patterns of community transition and ecosystem development during succession vary (Walker and del Moral, 2011), and there are many exceptions to the above generalized trends, even within a single community type (Barbour *et al.*, 1999). Ecological succession remains a highly complex process challenging the development of generalizable patterns.

Methods to Study Ecological Succession

Succession is an intricate, continuous, physical process extending in long temporal scales, which can be considered as an archetypical evolutionary process (Wurtz and Annila, 2010). This poses a great difficulty for methodological studies that must overcome the serious drawback of the required time-length, usually of eons. Thus, a single ecologist can't follow directly an entire successional process; indeed most studies are restricted to few-year surveys. In general, there are two possible alternatives: (i) aquire field data from permanent plots in a specific site and (ii) aquire field data from plots at different sites of different ages, under the assumption that these sites are ecologically equivalent differing only in age.

Apparently, the first option is the most explicit way to study succession, but very few relevant long-term studies exist. Such attempts require the initiation of a multiannual monitoring survey using a standardized sampling protocol (i.e., repeated measures of permanent plots over decades) strictly adopted by the researchers involved. The outcome is the precise record of change and consequently, a firm documentation of ecological succession. However, this long term commitment is hard to apply in practice and the vast majority of successional studies are either short-term or have followed the second option of using different chronosequences across a landscape. In the latter case, one can use different plots from different patches under different recovery states, across a landscape, and compare community structure to describe succession. The challenge is to establish a series of plots in areas affected by the same disturbance at different but known times. However, this is rarely well-documented and the entire method is much less accurate.

In the past, chronosequence methods were perhaps the only feasible way, but nowadays the advantages of applying direct long-term methods has been widely acknowledged. The use of comparable, standardized, experimental manipulative studies across abiotic and biotic gradients, that is, techniques that allow direct measurements of change and factors, is highly recommened in future efforts for understanding succession (Prach and Walker, 2011).

Ecological Succession on Terrestrial and Inland Water Habitats

Primary Succession Examples

The volcanic eruption of Mount St. Helens (del Moral and Lacher, 2005), the rose of Surtsey island after an underwater volcanic eruption in Iceland (Krebs, 2009), and the glacial retreat at Glacier Bay in Alaska are typical examples of primary terrestrial succession (Barbour *et al.*, 1999). In all such cases, extreme disturbance events cleared extended surfaces from any biological legacies. These catastrophic disturbances have been followed up by ecologists who stressed the importance of soil formation. In Mount St. Helens, the nitrogen-fixing plants (lupines), as controlled by herbivorous insects, proved to be crucial for ecosystem rehabilitation, enriching soil, and allowing establishment of other species. In Surtsey island, sandworts were initially established, followed by lull and a limited number of species, as soil conditions were sterile. Plant colonization accelerated only after the arrival of gulls who transferred seeds but also, and most importantly, due to the enrichment of soil by inputing nutrients from the sea, nitrogen in particular. In Glacier Bay, the constant presence of nitrogen fixing organisms (cyanobacteria substituted then by dwarf-shrub and alder) lead to an exceptionally rapid primary succession; it took only about 200 years after glacial retreat for the development of spruce-hemlock forest, compared to the 1000 years for deciduous forests in Lake Michigan sand dunes, or the 5000 years for moss-birch-tussock grass from glacial debris in Alaska (Barbour *et al.*, 1999). Soil formation, depending on the complex interactions of climate, organisms and time, is the most influencing process driving primary succession (Walker and del Moral, 2011). Nutrients accumulation in soil, nitrogen in particular, determines fertility. Bacterial fixation of nitrogen is especially important; mosses, lichens, fungi, algae, and bacteria (symbiotic bacteria in the roots of vascular plants) are important in soil formation and development after severe disturbances. Animal parts, insect or bird feces (guano) can be also important sources of nitrogen.

Secondary Succession Examples

A well-documented case of secondary succession is the North Carolina Piedmont old-field (Barbour *et al.*, 1999). The Piedmont area supports a patchy landscape of pine and hardwood forests and agricultural fields, and so, abandoned crops, represent ideal systems to study secondary succession (**Fig. 3**). The comparison of community structure through time showed that the cropland was initially colonized by crabgrasses and horseweeds that gave place to white-asters and ragweeds, which in turn gave place to broomsedge and then to pine seedlings. It took about 10 years to reach the stage of young pines canopy combined with broomsedge. Thereafter, seral stages differed according to humidity; in moister sites loblolly pines predominated in contrast to drier sites where shortleaf pines flourished. Gradually, within next 50 years, a hardwood overstory developed in both field types, becoming predominate over pines after about 150 years. However, the effect of disturbances, such as fires periodically occurring every 5–7 years, may maintain old-fields at the pine seral stage, preventing the development of hardwood forests. Old-field succession seems to vary according to the biogeographic area whereas in other cases of secondary succession, a gradual overlapping between seral stages has been observed in contrast to the sharp distinction of seres in Piedmont.

Fig. 3 Generalized patterns of terrestrial succession.

The controlling factor in most secondary succession cases is disturbance; the severity and frequency of disturbances determines available biological legacies initiating community development. Stand-replacing fire promotes succession in boreal forests and several studies documented the replacement of shade-intolerant pioneers by shade-tolerant, late-seral species. However, apart from this simplistic explanatory model, specific-site conditions, initial species composition, structural quality, density dependence, resource levels and intermediate disturbances are also involved leading to converging or diverging successional pathways (Chen and Taylor, 2012).

Successional studies on land ecosystems are typically limited on main structural plants, not taking into account the associated biota; such studies are enclaved in prolonged field observations linked to the generation length of dominant plants, reaching eons for late-successional tree species. On the opposite, in inland aquatic environments, such as lakes, relevant studies are dealing with the annually repeated seasonal succession of plankton communities, which operates in short time intervals, usually of few months. A comprehensive study of planktonic secondary succession, in which population and system level indices were applied, is that of the Lake Constance in Alps (Boit and Gaedke, 2014). This is a large, deep and warm monomictic lake of glacial origin with limited external input, in which a typical shift in community composition along succession was observed passing through predictable stages. The authors observed along the successional stages i) higher energy transfer efficiency across trophic levels, ii) diversification and increasing specialization of consumers on resources exploitation, iii) decrease in total production, iv) increase in functional diversity and complexity of foodwebs. Ongoing grazing pressure, lower prey edibility, lower food quantity and quality, and declining nutrient concentrations were the main successional drivers.

The Special Case of Carrion Succession

Carrion ecology has gained increased attention as ecologists started to realize the importance of carcasses in ecosystem function, nutrient cycling and organic matter input (Michaud and Moreau, 2017). However, many aspects of carrion ecology remain unclear, including the colonization and succession patterns on carcasses. In these processes, arthropods, insects in particular, seem to have the most prominent role. It is generally acceptable that carrion succession is a continuous process with no abrupt changes in faunal composition across sequential stages, representing nonequilibrium systems without a terminal stable state (either monoclimax or polyclimax). Certain flies (families Calliphoridae, Muscidae, and Sarcophagidae) are the pioneer colonizers, thus considered the primary drivers of succession, followed by other arthropod species in a predictable sequence. The patterns of arthropod colonization on carrion is fairly constant at the family level. However, the exact species composition along succession depends on the local colonizers pool, species interactions (trophic interactions in particular, i.e., within or between necrophagous and their predators/parasites, omnivorous or adventine species), and other factors such as habitat, season, climatic conditions and decomposition environment (Sukchit *et al.*, 2015). Despite being a continuous process, carrion succession may be splitted in apparent stages linked both to decay process and succession of specific arthropod groups (**Table 1**); such defined patterns can be used to estimate postmortem interval, which is particularly useful in medico-legal forensic.

Carrion succession patterns often present a typical horseshoe-shaped arch form; faunistic diversity and similarity are low in the early successional stage (fresh stage of decomposition), gain maximum values in midstages (bloating and active decay), decrasing in late seral stages (advanced and postdecay) towards skeletonization (Tabor *et al.*, 2004). Between seral stages, periods of

Table 1 Main arthropod families or faunal groups along succession according to decomposition stage (data from Payne, 1965)

Family or faunal group	*Decomposition stage*					
	Fresh	*Bloated*	*Active decay*	*Advanced decay*	*Postdecay (dry)*	*Skeletonization*
Calliphoridae	*	***				
Centipedes					***	
Cleridae				*	***	
Dermestidae				*	***	
Dipterans				Greatly decreased	Absence of most	
Histeridae		*	***	*	Absence	
Lonchaeidae		*				
Maggots			*			
Millipedes					***	
Mites						*
Muscidae	*	***				
Phoridae			*			
Piophilidae		*	*			
Sarcophagidae	*	***				
Scarabaeidae		*				
Sepsidae			*			
Silphidae		*	*		Absence	
Staphylinidae		*	***	*	Absence	
Trogidae				*		

Fig. 4 Generalized patterns of marine succession using artificial colonization panels.

taxonomic stasis, that is, periods with very little change in faunistic composition, may be also observed. In temperate regions, successional patterns varies according to season, but interannual changes are considered negligible. There is a controversy on the application of classical successional theory on carrion succession. The inhibition model seems inappropriate (Krebs, 2009), while Michaud and Moreau (2017) tested experimentaly the facilitation model and rejected it, as the removal of pioneer species did not prevent colonization by secondary colonizers. The latter authors concluded that although facilitation may stand in some cases, succession seems to be driven by the complicated and, not yet fully understood, biotic interactions (within the arthropod community and between the arthropod and the microbial communities), as influenced by abiotic factors.

Ecological Succession in Marine Habitats

In the marine environment, successional studies started in about the middle of the previous century focusing on the dynamics of structural components, that is, sessile animals, algae, and seagrasses, and conducted from the intertidal to shallow subtidal zone. The basic aim was to understand successional patterns per se, to investigate rehabilitation of marine habitats using mimics, for example, artificial reefs or synthetic corals and seagrasses, or to examine fouling development. Instead of chronosequences, direct data acquisition using submerged panels was the most common option (**Fig. 4**). However, the applied protocols were highly variable hindering the extraction of comparable results and general conclusions, and typically suffer from limited replication and immersion-length (<5 years). In spite of the increasing research efforts, the underlying mechanisms of succession in marine environment remain far from understood.

A comprehensive example of secondary succession in the marine environment (to our knowledge there are no examples of primary succession) is the study of Farell (1991) who experimentally examined patterns in a barnacle-*Pelvetia* intertidal community on the coast of Oregon. The author showed that succession followed spatially the same general sequence of species colonization and substitution, though the rate greatly varied among sites according to the timing of successful *Balanus* recruitment. Species interactions, that is, predation, grazing, and competition for space were identified as the most likely driving factors to successional pathways.

Other studies in the rocky sublittoral zone of the eastern Mediterranean (Antoniadou *et al.*, 2010, 2011) and SE Pacific (Pacheco *et al.*, 2011) identified as a general pattern the increase in biodiversity, abundance and biomass as succession proceeds, with maximum values at the intermediate to late stages, under low recovery rates, whereas potential colonizers arrived from surrounding biota. Convergence towards a common structure as succession proceeds has been suggested, despite the strong initial differences attributed either to the effect of seasonal onset of colonization or stochasticity, and explained by supply-side processes and species interactions. A critical issue was the establishment of conspicuous sessile species that provide physical structure, that is, ecosystem engineers, which contribute to the acceleration of succession. However, Brown and Swearingen (1998), studying intertidal fouling assemblages in Northern Mexico Gulf, claimed for the absence of any unidirectional sequence of succession. As Christensen (2014) states "*We now understand that there is no single unique or unifying mechanism for successional change, that successional trajectories are highly varied and rarely deterministic, and that succession has no specific endpoint.*"

See also: Ecosystems: Estuaries; Freshwater Lakes; Riparian Wetlands. General Ecology: Succession

References

Anderson, K.J., 2007. Temporal patterns in rates of community change during succession. American Naturalist 169, 780–793.
Antoniadou, C., Voultsiadou, E., Chintiroglou, C., 2010. Benthic colonization and succession on temperate sublittoral rocky cliffs. Journal of Experimental Marine Biology and Ecology 382, 145–153.

Antoniadou, C., Voultsiadou, E., Chintiroglou, C., 2011. Seasonal patterns of colonization and early succession on sublittoral rocky cliffs. Journal of Experimental Marine Biology and Ecology 403, 21–30.
Barbour, M.G., Burk, D.J.H., Pitts, W.D., Gilliam, F.S., Schwartz, M.W., 1999. Terrestrial plant ecology, 3rd edn. Canada: Pearson.
Bestelmeyer, B.T., Ash, A., Brown, J.R., Densambuu, B., Fernández-Giménez, M., Johanson, J., Levi, M., Lopez, D., Peinetti, R., Rumpff, L., Shaver, P., 2017. State and transition models: Theory, applications, and challenges. In: Briske, D.D. (Ed.), Rangeland systems. Switzerland: Springer Series on Environmental Management, pp. 303–346.
Boit, A., Gaedke, U., 2014. Benchmarking successional progress in a quantitative food web. PLoS ONE 9.e90404
Briske, D.D., Fuhlendorf, S.D., Smeins, F.E., 2005. State-and-transition models, thresholds, and rangeland health: A synthesis of ecological concepts and perspectives. Rangeland Ecology & Management 58, 1–10.
Brown, K.M., Swearingen, D.C., 1998. Effects of seasonality, length of immersion, locality and predation on an intertidal fouling assemblage in the Northern Gulf of Mexico. Journal of Experimental Marine Biology and Ecology 225, 107–121.
Chen, H.Y.H., Taylor, A.R., 2012. A test of ecological succession hypotheses using 55-year time-series data for 361 boreal forest stands. Global Ecology and Biogeography 21, 441–454.
Christensen, N.L., 2014. An historical perspective on forest succession and its relevance to ecosystem restoration and conservation practice in North America. Forest Ecology and Management 303, 312–322.
Connel, J.H., Slatyer, R.O., 1977. Mechanisms of succession in natural communities and their role in community stability and organization. The American Naturalist 111, 1119–1144.
Davis, M.A., Thompson, K., 2000. Eight ways to be a colonizer; two ways to be an invader: A proposed nomenclature scheme for invasion ecology. Bulletin of the Ecological Society of America 81, 226–230.
del Moral, R., 2007. Limits to convergence of vegetation during early primary succession. Journal of Vegetation Science 18, 479–488.
del Moral, R., Lacher, I.L., 2005. Vegetation patterns 25 years after the eruption of Mount St. Helens, Washington, USA. American Journal of Botany 92, 1948–1956.
Egler, F.E., 1954. Vegetation science concepts. I. Initial floristic composition, a factor in old-field vegetation development. Vegetatio 4, 412–417.
Farell, T.M., 1991. Models and mechanisms of succession: An example from a rocky intertidal community. Ecological Monographs 61, 95–113.
Fox, B.J., 1982. Fire and mammalian secondary succession in an Australian coastal heath. Ecology 63, 1332–1341.
Franklin, J.F., 1990. In: Biological legacies: A critical management concept from Mount St. Helens.Transactions of the 55th North American Wildlife and Natural Resources Conference, Washington, USA.
Gitay, H., Noble, I., 1997. What are functional types and how should we seek them? In: Smith, T.M., Shugart, H.H., Woodward, F.I., (Eds.), Plant functional types: Their relevance to ecosystem properties and global change. Cambridge: Cambridge University Press, pp. 3–19.
Gleason, H.A., 1927. Further views on the succession concept. Ecology 8, 299–326.
Grime, J.P., 1974. Vegetation classification by reference to strategies. Nature 250, 26–31.
Horn, H.S., 1975. Markovian properties of forest succession. In: Cody, M.L., Diamond, J.M., (Eds.), Ecology and evolution of communities. Cambridge: Belknap Press of Harvard University Press, pp. 196–211.
Hubbell, S.P., 2001. The unified neutral theory of biodiversity and biogeography. Oxford: Princeton University Press.
Hutchinson, G.E., 1951. Copepodology for the ornithologist. Ecology 32, 571–577.
Krebs, C.J., 2009. Ecology, 6th edn. Canada: Pearson.
Levins, R., 1969. Some demographic and genetic consequences of environmental heterogeneity for biological control. Bulletin of the Entomological Society of America 15, 237–240.
Lohmus, K., Paal, T., Liira, J., 2014. Long-term colonization ecology of forest-dwelling species in a fragmented rural landscape – Dispersal versus establishment. Ecology and Evolution 4, 3113–3126.
MacArthur, R.H., Wilson, E.O., 1967. The theory of island biogeography. Princeton: Princeton University Press.
McCann, K.S., 2000. The diversity–stability debate. Nature 405, 228–233.
McCook, L.J., 1994. Understanding ecological community succession: Causal models and theories, a review. Vegetatio 110, 115–147.
Menge, B.A., 1975. Ecological implications of patterns of rocky intertidal community structure and behavior along an environmental gradient. In: Costlow, J.D. (Ed.), The ecology of fouling communities. Beaufort, NC, USA: Duke University Marine Laboratory, pp. 155–180.
Michaud, J.-P., Moreau, G., 2017. Facilitation may not be an adequate mechanism of community succession on carrion. Oecologia 183, 1143–1153.
Noble, I.R., Slatyer, R.O., 1980. The use of vital attributes to predict successional changes in plant communities subject to recurrent disturbances. Vegetatio 43, 5–21.
Odum, E.P., 1969. Strategy of ecosystem development. Science 164, 262–270.
Pacheco, A.S., Laudien, J., Thiel, M., Oliva, M., Heilmayer, O., 2011. Succession and the seasonal onset of colonization in subtidal hard-bottom communities off northern Chile. Marine Ecology 32, 75–87.
Payne, J.A., 1965. A summer carrion study of the baby pig *Sus scrofa* Linnaeus. Ecology 46, 592–602.
Phillips, J.D., 2011. Predicting models of spatial change from state-and-transition models. Ecological Modelling 222, 475–484.
Prach, K., Walker, L.R., 2011. Four opportunities for studies of ecological succession. Trends in Ecology and Evolution 26, 119–123.
Pulsford, S.A., Lindenmayer, D.B., Driscoll, D.A., 2016. A succession of theories: Purging redundancy from disturbance theory. Biological Reviews 91, 148–167.
Purschke, o., Schmid, B.C., Sykes, M.T., Poschlod, P., Michalski, S.G., Durks, W., Kuhn, I., Winter, M., Prentice, H.C., 2013. Contrasting changes in taxonomic, phylogenetic and functional diversity during a long-term succession: Insights into assembly processes. Journal of Ecology 101, 857–866.
Raevel, V., Violle, C., Munoz, F., 2012. Mechanisms of ecological succession: Insights from plant functional strategies. Oikos 121, 1761–1770.
Sousa, W.P., 1984. Intertidal mosaic: Patch size, propagule availability and spatially variable patterns of succession. Ecology 65, 1918–1935.
Sukchit, M., Deowanish, S., Butcher, B.A., 2015. Decomposition stages and carrion insect succession on dressed hanging pig carcasses in Nan Province, northern Thailand. Tropical Natural History 15, 137–153.
Tabor, K.L., Brewster, C.C., Fell, R.D., 2004. Analysis of the successional patterns of insects on carrion in southwest Virginia. Journal of Medical Entomology 41, 785–795.
Tansley, A.G., 1935. The use and abuse of vegetational concepts and terms. Ecology 16, 284–307.
Tilman, D., 1985. The resource-ratio hypothesis of plant succession. American Naturalist 125, 827–852.
Van Hulst, R., 1979. On the dynamics of vegetation: Markov chains as models of succession. Vegetatio 40, 3–14.
Walker, L.R., del Moral, R., 2011. Primary succession. In: eLS. Chichester: John Wiley & Sons, Ltd.
Westoby, M., Walker, B., Noy-Meir, I., 1989. Opportunistic management for rangelands not at equilibrium. Journal of Range Management Archives 42, 266–274.
Whelan, R.J., Rodgerson, L., Dickman, C.R., Sutherland, E.F., 2002. Critical life cycles of plants and animals: Developing a process-based understanding of population changes in fire-prone landscapes. In: Bradstock, R.A., Williams, J.E., Gill, A.M. (Eds.), Flammable Australia: The five regimes and biodiversity of continent. Cambridge: Cambridge University Press, pp. 94–124.
Whittaker, R.H., 1953. A consideration of climax theory: The climax as a population and pattern. Ecological Monographs 23, 41–78.
Whittaker, R.H., Goodman, D., 1979. Classifying species according to their demographic strategy. I. Population fluctuations and environmental heterogeneity. The American Naturalist 113, 185–200.
Wurtz, P., Annila, A., 2010. Ecological succession as an energy dispersal process. Biosystems 100, 70–78.

Further Reading

Barbour, M.G., Burk, D.J.H., Pitts, W.D., Gilliam, F.S., Schwartz, M.W., 1999. Terrestrial plant ecology, 3rd ed. Canada: Pearson.
Christensen, N.L., 2014. An historical perspective on forest succession and its relevance to ecosystem restoration and conservation practice in North America. Forest Ecology and Management 303, 312–322.
Connel, J.H., Slatyer, R.O., 1977. Mechanisms of succession in natural communities and their role in community stability and organization. The American Naturalist 111, 1119–1144.
Connel, J.H., Noble, I.R., Slatyer, R.O., 1987. Om the mechanisms producing successional changes. Oikos 50, 136–137.
Drury, W.H., Nisbet, I.C.T., 1973. Succession. Journal of the Arnold Arboretum 54, 331–368.
Hubbell, S.P., 2001. The unified neutral theory of biodiversity and biogeography. Oxford: Princeton University Press.
MacArthur, R.H., Wilson, E.O., 1967. The theory of island biogeography. Princeton: Princeton University Press.
McCann, K.S., 2000. The diversity–stability debate. Nature 405, 228–233.
McCook, L.J., 1994. Understanding ecological community succession: Causal models and theories, a review. Vegetatio 110, 115–147.
Odum, E.P., 1969. Strategy of ecosystem development. Science 164, 262–270.
Prach, K., Walker, L.R., 2011. Four opportunities for studies of ecological succession. Trends in Ecology and Evolution 26, 119–123.
Pulsford, S.A., Lindenmayer, D.B., Driscoll, D.A., 2016. A succession of theories: Purging redundancy from disturbance theory. Biological Reviews 91, 148–167.
Raevel, V., Violle, C., Munoz, F., 2012. Mechanisms of ecological succession: Insights from plant functional strategies. Oikos 121, 1761–1770.
Walker, L.R., del Moral, R., 2011. Primary succession. In: eLS. Chichester: John Wiley & Sons, Ltd.
Wurtz, P., Annila, A., 2010. Ecological succession as an energy dispersal process. Biosystems 100, 70–78.

Ecological Processes: Volatilization☆

Zhi-Qing Lin, Southern Illinois University Edwardsville, Edwardsville, IL, United States

Significance of Biological Volatilization in Biogeochemical Cycling

Volatilization is the transfer of biogenic volatile elements or compounds from soil or water to the atmosphere. As a natural ecological process, biological volatilization plays an important role in the biogeochemical cycling of elements such as arsenic (As), mercury (Hg), and selenium (Se). For example, on a global scale, biogenic volatile Se emission from the ocean to the atmosphere is on the order of $5–8 \times 10^6$ kg per year and approximately 1.2×10^6 kg per year from the terrestrial surfaces. The flux of volatile Se from agricultural lands varies substantially, ranging from 0.1 to 10 mg Se m^{-2} $year^{-1}$. It was also estimated that as much as 2.1×10^7 kg of As could be lost annually through biological volatilization from land surfaces to the atmosphere. The continental volatile As flux is about eight times that of the continental particulate As flux, suggesting that biological volatilization of As be an important pathway in the global biogeochemical cycling of As.

Biogenic volatilization is a biological metabolism process that varies substantially under different environmental conditions. Thus, the biogeochemical cycling of Se, As, or Hg can be complicated because of those variations. Although several elements can be volatilized biologically from water or soil to the atmosphere, Se and As are the only elements that have extensively been studied to date.

Mechanisms of Biological Volatilization

Different elements involve different mechanisms in biological volatilization. For instance, the production of biogenic volatile Se involves the biological methylation of inorganic Se to dimethyl selenide (DMSe) or dimethyl diselenide (DMDSe), while volatilization of Hg results primarily from biochemical reduction of inorganic Hg to elemental Hg. DMSe, DMDSe and elemental Hg can dissipate from water, soil and plants to the atmosphere.

The pathway of Se assimilation and volatilization in higher plants has previously been constructed. The pathway involves various biochemical reduction and methylation processes catalyzed by different specific enzymes. In brief and for illustration, it contains the following major steps in relation to the formation of volatile Se compounds in most higher plants (but excluding Se hyperaccumulator plants), and needless to say, there are some other intermediate Se-compounds during each step of transformation: *Selenate* (SeO_4^{2-}) → *Selenite* (SeO_3^{2-}) → *Selenocysteine* (*SeCys*) → *Selenomethionine* (*SeMet*) → *Dimethyl Selenide* (*DMSe*)↑. In Se hyperaccumulator plants (such as *Stanleya pinnata*), Se volatilization proccess involves a different approach, including *Selenocysteine* (*SeCys*) → *Methylselenocysteine* (*MeSeCys*) → *Dimethyl Diselenide* (*DMDSe*) ↑. The chemical reduction from selenate to SeCys likely occurs at chloroplasts in plant leaves, while the formation of SeMet and methylation of SeMet likely takes place in the cytosol.

With regard to microbial volatilization, the pathway of Se methylation has not clearly been documented. Experimental evidence indicates that assimilatory selenate and selenite reduction to elemental Se may be an important biological transformation process in anoxic environment. Bio-methylation of inorganic Se compounds (like SeO_4^{2-}, SeO_3^{2-}, or Se^0) might occur under aerobic conditions. Dimethyl selenide is the major metabolite of Se volatilization by microbes and many higher plants except for those Se hyperaccumulator plant species. More DMDSe can be produced by Se-hyperaccumulators (such as *Astragalus bisulcatus* and

☆*Change History*: April 2018. Zhi-Qing Lin prepared the update. Table 1 has been updated with new and additional information and the content has been updated with recent research findings.

Stanleya pinnata) because those Se hyperaccumulator plants metabolize SeCys to methylSeCys, and further DMDSe. Trace amounts of other volatile Se compounds such as dimethyl selenone, methane selenol, and dimethyl selenenyl sulfide, have also been observed.

Selenium volatilization results from the Se mass balance between the production of biogenic DMSe and de-methylation of the biogenic volatile compounds in soil and water systems. Microbial de-methylation removes a methyl group from the methylated compounds, and therefore, reduces the amount of volatile Se to the atmosphere. Laboratory experiments have found that high soil moisture and manganese (Mn) contents could enhance the de-methylation process of Se in the soil.

Major Factors Affecting the Process of Biological Volatilization

Plant Species and Soil Microbe Strains

As a biological process, biogenic volatilization of Se or other elements is significantly affected by the presence of different plant species and soil microbes. Plants and microbes take up selenium and metabolize the element into volatile chemical compounds. Both bacteria and fungi are important in Se volatilization in soil and sediment. Crops, including rice, broccoli, and cabbage, have great capabilities to volatize Se from water under laboratory conditions. Previous field studies identified that pickleweed (*Salicornia bigelovii* Torr.), rabbitfoot grass (*Polypogon monspeliensis* L.), and Prince's plume (*Stanleya pinnata* L.) are among the best plant species for Se volatilization from selenate (Table 1). Unlike Indian mustard (*Brassica juncia* L.) for which selenate reduction appears to be rate-limiting in Se volatilization, pickleweed has an enhanced capacity to reduce selenate to organic Se forms. Such chemical transformation of Se in pickleweed is independent of the presence of microbes.

In addition to Se, some Hg resistant bacterial strains (such as *Klebsiella pneumoniae*) are able to reduce ionic Hg to elemental Hg and effectively volatilize Hg compounds in the substrates. Soil microbes (e.g., *Fusarium* sp. and *Pseudomonas alcaligenes*) can also produce volatile As like dimethylarsenate via the reductive methylation process from inorganic or methylated As compounds. Several fungal species (e.g., *Aspergillus glaucum*, *Candida humicola*, *Gliocladium roseum*, and *Penicillium gladioli*) were found to be capable of transforming inorganic As to volatile trimethylarsine (TMA) compound.

Concentrations and Chemical Forms

The Se bioavailability in soil–plant systems determines the level of Se volatilization. The rate of Se volatilization significantly correlates with the Se concentration in the substrate or plant tissues. The pathway of Se assimilation and volatilization suggests that the production of DMSe from selenate be less favorable energetically than selenite or SeMet. Laboratory experiments have demonstrated that volatilization of Se from different chemical forms (including nanoscale elemental Se or SeNPs) varies substantially, with the following order of magnitude: SeMet $\gg$ selenite $>$ selenate $>$ SeNPs. For example, when broccoli plants were supplied with 20 μM Se as selenate, selenite, or SeMet in hydroponic solution, the rate of Se volatilization from SeMet was about 3 times greater than from selenite, or 13 times greater than from selenate. When in bulk form, elemental Se is not water soluble or not bioavailable; however, nanoscale elemental Se becomes partially bioavailable and can be volatilized at low rates in soil–plant systems.

Under aerobic conditions with a high redox potential, selenate is the dominant chemical form of Se in soil. Different chemical forms of Se or other elements can be transformed in plants or by soil microbes. For instance, most of the Se taken up by India mustard plants supplied with SeO_4^{2-} will remain unchanged, but plants supplied with SeO_3^{2-} or SeMet will contain SeMet-like Se compounds.

Table 1 Maximum rates (μg Se m^{-2} day^{-1}) of Se volatilization observed in different soil/sediment-plant systems under field conditions

Plant species	*Upland system*	*Wetland system*
Baltic rush (*J. balticus*)	–	29 ± 5
Cattail (*T. latifolia*)	–	57 ± 12
Cordgrass (*Spartina* spp.)	46 ± 12	80 ± 10
Elephant grass (*P. purpureum*)	29 ± 1	
Indian mustard (*B. juncea*)	<25	
Pickleweed (*S. bigelovii*)	420 ± 91	–
Prince's plume (*S. pinnata*)	1002 ± 509	
Rabbitfoot grass (*P. monspeliensis*)	–	274 ± 100
Saltbush (*A. lentiformis*)	29 ± 13	–
Saltgrass (*D. spicata*)	28 ± 21	38 ± 27
Saltmarsh bulrush (*S. robustus*)	–	13 ± 9

Values are mean and standard deviation ($n = 3$).

Chemical Interaction of Other Elements

Sulfur and Se are chemical analogs (i.e., having similar chemical properties). High concentrations of sulfate in the substrate inhibit Se volatilization from selenate due to their competition for the uptake transporter or the active sites of enzymes that are responsible for biological transformation of selenate to DMSe. For example, the inhibitory effect of sulfate on Se volatilization by selenate-supplied broccoli increases with increasing the ratio of S/Se in plant tissues. When the sulfate concentration increases from 0.25 to 10 mM in the substrate, the rate of Se volatilization can be decreased by 85%, while effects of sulfate on volatilization of Se from selenite or SeMet are not significant.

Similarly, phosphorus and As are also chemical analogs. Although effects of soil phosphate on the accumulation of arsenate in plant tissues have been documented, there are no experimental evidences indicating that phosphate significantly inhibits As volatilization from arsenate.

Soil Organic Matter and Water Moisture

Soil organic matter content is a critical factor for microbial volatilization of Se and other metalloids. The addition of organic materials to soil can provide soil microbes with essential energy and carbon sources to metabolize inorganic Se to DMSe. Soil amendments with different organic materials (such as casein, manure or plant litter) have been evaluated for the enhancement of Se volatilization in soil and sediment. For example, in a soil-pickleweed system, the addition of shoot biomass/fallen litter to the soil surface at a rate that is equivalent to the annual biomass production of $\sim$1.5 kg m^{-2} increased Se volatilization up to 2.2 times of magnitude. Those research findings have resulted in developing a new concept or remediation strategy for Se-contaminated water and soil through biogenic volatilization process.

The soil water content also significantly affects the process of Se volatilization. For example, sequential changes in soil moisture (such as creating drying and wetting cycles) will enhance Se volatilization because such hydraulic fluctuation will increase the decomposition rate of organic matter, Se bioavailability, and microbial activity in soil. Following up to each wetting event in a soil-pickleweed system, Se volatilization increased up to threefolds repeatedly, and the rate of Se volatilization significantly correlated with soil water potential (kPa). While soil water potential decreased from 0 kPa (water saturation) to -25 kPa in the top 5-cm layer of soil, the rate of Se volatilization was reduced by 78%. For microbial volatilization, the optimum soil water content generally ranges from 20% to 70% of the soil water holding capacity.

Temperature, Redox Potential, pH, and Salinity

Other environmental factors, such as temperature and redox potential, will also affect Se volatilization. Biological volatilization process is temperature-dependent. Therefore, the rate of Se volatilization generally increases with increasing air or soil temperature. Laboratory experiments show that the optimal temperature for Se volatilization is approximately 40°C. In the field studies conducted in Central California, high rates of Se volatilization were oftentimes observed during later spring or early summer months. In addition, more volatile Se can be generated by soil microbes under aerobic conditions than under anaerobic conditions. Changes in soil pH will affect the bioavailability of Se in soil. The optimum pH for Se volatilization is around eight in soil or sediment. Few studies have been conducted to determine the impact of soil salinity on Se volatilization, which is likely depending on the type of salinity (i.e., sulfate salinity or chloride salinity) and the magnitude.

Soil Bacteria and Plant–Soil Microbial Interactions

Soil bacteria and fungi can volatilize Se without the presence of vegetation, although the addition of vegetation generally results in higher levels of Se volatilization. It is difficult to determine the extent to which biogenic volatile Se can be produced directly from plants versus soil microbes (i.e., separating plant roots from the soil microbes). Previous studies showed that rhizosphere bacteria enhanced the uptake and accumulation of selenate in plant roots, and therefore, facilitated volatilization of Se from selenate. Indeed, the presence of soil microbes plays a very important role in the high production of volatile Se in soil–plant systems. For example, in the soil-pickleweed system, the amount of volatile Se dissipated directly from shoots accounts for only 10% of the total volatile Se, and the majority of total volatile Se was produced from the soil–root compartment. Pickleweed has the capacity to reduce selenate to SeMet. Thus, the accumulation of SeMet in plant tissues provides a large pool of bioavailable SeMet that can be easily methylated and volatilized by soil microbes in the field. Therefore, high rates of Se volatilization in the pickleweed field likely resulted from the interaction between pickleweed and the species associated soil microbes. The addition of vegetation will not only increase levels of Se volatilization by directly dissipating Se from plants, but also create special habitats (or rhizosphere) for specific soil microbes. Plant roots may also produce specific root exudates that are essential for certain rhizosphere microbes. In addition, annual biomass production of plants will provide soil microbes with C and energy sources in a sustainable manner.

Specific soil bacteria have been isolated for As methylation and volatilization. For example, paddy-soil abundant *Cytophagaceae* spp. has the ability to convert almost half of 10 μM arsenite in the growth medium to trimethylarsine gas during a 24-h time period. The inoculation of this bacterium had greatly enhanced As methylation and volatilization in the soil contaminated with As.

Genetic Modified Plants for the Enhancement of Biological Volatilization

To substantially increase biological volatilization, techniques of genetic engineering have been applied to overcome biological limitations that are associated with the pathway of Se assimilation and volatilization. For example, the overexpression of ATP

sulfurylase (APS) can facilitate the reduction of selenate to selenite, which is commonly rate-limiting with respect to the production of other organic compounds, such as SeCys and SeMet. Similarly, other downstream rate-limiting steps can also be eliminated by the overexpression of other enzymes. India mustard plants that were overexpressed with cystathionine-γ-synthase (CGS) has shown an enhanced efficiency (about two to three times greater) for Se volatilization under laboratory conditions.

In addition to genetically engineered Indian mustard for Se volatilization, other transgenic plants have also been developed to metabolize ionic- and methyl-mercury and to produce elemental mercury (Hg^0) for volatilization. For example, genetic engineering has added a bacterial mercuric ion reductase to *Arabidopsis thaliana* and therefore, the transgenic plant is able to convert Hg cations to elemental Hg and further releases Hg^0 into the air. Some Hg-resistant bacterial strains like *Klebsiella pneumoniae* are able to reduce ionic Hg to elemental Hg and to effectively volatilize Hg compounds from the substrates.

Transfer of Volatile Compounds From Water to the Air

The environmental transfer process of biogenic volatile compounds of Se or other elements is primarily controlled by a physical dispersion process that is generally governed by Henry's law. Henry's law constant (H) describes the distribution of a volatile compound (e.g., DMSe or DMDSe) between gas and water phases at thermodynamic equilibrium. Therefore, the potential for a volatile compound to partition from water to air increases with increasing Henry's law constant. A larger Henry's law constant value (e.g., >0.1) suggests a higher potential for the organic compound to volatilize from water to the air. For example, using the DMSe water solubility of 0.0244 g g^{-1} and the DMSe vapor pressure of 32.03 kPa, Henry's law constant of DMSe at 25°C is calculated as 0.058, suggesting a moderate tendency to partition between water and air. Henry's law constants generally increase with increasing temperature, primarily due to the significant temperature dependency of chemical vapor pressures. A higher temperature will not only elevate plant and microbial activities but also increase Henry's law constant, which, in turn, will increase the flux of Se volatilization from water to the air.

The stability of DMSe in soil is affected by the soil water content. In water saturated soil, DMSe can be partially dissolved in water before its transfer from soil water to the air. Oxidants (such as manganese oxides) in soil solution may convert DMSe to nonvolatile Se compounds. Therefore, the transport of volatile Se compounds in soil significantly relates with the soil water content. The rate of Se volatilization will be likely low in flooded soil or sediment due to DMSe solubilization in soil solution. For example, the flux of Se volatilization from a natural wetland in Southern Switzerland was generally <0.12 μg m^{-2} day^{-1}, along with As of <0.54 μg m^{-2} day^{-1} and S of <37 μg m^{-2} day^{-1}, during the summer season.

Flux Measurements of Biological Volatilization

Volatile Se compounds in the air can be collected by physical adsorption using activated carbon materials or by chemical oxidation with peroxide (H_2O_2) or acid (HNO_3) trap solutions. The alkaline-peroxide trap solution (30% hydrogen peroxide and 0.05 M HaOH, 4:1, v/v) has commonly been used to directly trap or to extract Se from the surface of the activated carbon material. When nitric acid is used as trapping solution, dimethylselenide (DMSe) can be oxidized to dimethylselenoxide (DMSeO) or dimethyldiselenide (DMDSe) to methaneseleninic acid (MSeA). Thus, speciation analysis of volatile Se compounds can be performed using the nitric acid trapping solution. In addition, solid phase microfiber extraction (SPME) has also been used for sampling airborne organic As compounds from soil–plant systems.

The flux of Se volatilization can be determined using different types of sampling chambers that can be open-bottom cubic compartments made of transparent materials such as *Plexglas*. An open-flow sampling system has been commonly selected because it allows for a minimal heat and humidity buildup inside the chamber as well as for a maximum sampling efficiency. Briefly, with an open-flow chamber sampling system, ambient air is continuously scrubbed of volatile Se before entering the chamber through charcoal filters or the trap solution at the inlet port, and Se-free air mass is then mixed with volatile Se produced inside the chamber. The mixed air is drawn under slight vacuum into the outlet port passing through charcoal filters or the trap solution where volatile Se compounds are collected. To further minimize the effects of rising temperature inside sampling chambers on the volatile Se measurements, the sampling chamber can also be equipped with a cooling radiator connected to an evaporative cooler.

The micrometeorological flux measurement technique is an attractive method for the flux measurement of Se (or other elements) volatilization under filed conditions. The technique provides an average integrated flux of volatile Se over a large area by measuring the difference in Se concentrations in ambient air at two different heights above ground. Indeed, this technique determines the flux of volatilization through eddy correlation under naturally occurring environmental conditions. However, it is difficult to apply this sampling technique in practice because it is costly in instrumentation and often associated with interference such as the footprint effect from other sources of volatilization.

Fate of Biogenic Volatile Se in the Atmosphere

Biogenic volatile Se can be transformed into more oxidized and less volatile chemical species in the atmosphere. Volatile Se compounds (DMSe or DMDSe) can react with airborne oxidants such as O_3, OH, and NO_3 radicals, and be converted into aerosols or particulates during a short period of time, for example, ranging from 5 to 6 h in the atmosphere in Central California. The

dispersion of airborne Se in the mixing boundary layer and its subsequent atmospheric long-range transport are highly controlled by local air stability and regional air mass movement pattern. The forward trajectory analysis showed that volatile Se from the Western San Joaquin Valley can be transported out of the Valley within 24 h toward remote mountain areas. In the atmosphere, aerosol- or particulate-associated Se accounts for 75% of total airborne Se, with the remaining 25% being in the vapor form. Therefore, wet deposition is likely the dominant pathway of Se removal from the atmosphere to the land surfaces.

Significance of Biological Volatilization in Phytoremediation of Se Contamination

Phytoremediation is the use of plants and associated microbes to clean up contaminated water and soil. Phytoremediation of Se contamination mainly involves two processes: (1) biological volatilization that removes Se from contaminated soil–plant systems to the atmosphere and (2) accumulation of Se in plant tissues that will be harvested and removed from the contaminated site (i.e., phytoextraction). Phytoextraction can only be effective during the plant growth season, while biological volatilization can be in action continuously during the whole year. Field studies showed that biological volatilization is a significant pathway of Se removal in phytoremediation. For example, in the pickleweed field in the San Joaquin Valley, the annual total Se removal through volatilization can be 5.5 times more than via phytoextraction. Biological volatilization represents an environmentally sound and sustainable biotechnology for the remediation of Se contaminated soil and water. The volatilization process generates DMSe compound that is ~500 times less toxic than the inorganic forms of Se fed to rats. In addition, Se removal via volatilization likely diminishes the amount of Se available for entry into food chains in the contaminated environment.

There are few long-term studies on Se volatilization conducted under field conditions. In a previous constructed treatment wetland study, different wetland cells were built and vegetated with different plant species for the treatment of Se-laden agricultural drainage water (25 μg Se L^{-1}). The plant species included cattail (*Typha latifolia* L.), baltic rush (*Juncus balticus* Willd.), smooth cordgrass (*Spartina alternifolia* Loisel), saltgrass (*Distichlis spicata* (L.) Greene), saltmarsh bulrush (*Scirpus robustus* Pursh), and rabbitfoot grass. Rates of Se volatilization were determined monthly for 2 years in each wetland cell. Significant variation in Se volatilization was observed among the wetland cells or different plant species. The highest average rate over the 2-year period was 33 ± 12 μg m^{-2} day^{-1} in the rabbitfoot grass cell. Selenium volatilization was highest in late spring and early summer during the year in Central California. About 35% and 48% of the Se entering the rabbitfoot grass cell was volatilized during May and June, respectively, whereas only <5% was volatilized in the winter months. Overall, biological volatilization removed 9.4% of the total Se input to the rabbitfoot grass wetland cell to the atmosphere during the 2-year study period. Clearly, to implement the phytoremediation technology, Se volatilization needs to be substantially enhanced and sustained at high levels during the entire year.

Volatilization of organic pollutants is usually defined as a physical process of compounds transferring from water to air phase without involving metabolism. Although biological degradation of organic compounds may result in different intermediate products that could dissipate to the air, such processes are not included in this paper because they do not involve biological methylation and volatilization as the natural ecological process. Studies have also been conducted on biological volatilization of S, mainly with respect to its geochemical cycling. As mentioned above, because S and Se are chemical analogs, the above discussion on Se volatilization is generally applied to the volatilization process of S in the environment.

Further Reading

Cahill TA and Eldred RA (1998) Chapter 30—Particulate selenium in the atmosphere. In: Frankenberger WT Jr. and Engberg RA (eds.) *Environmental chemistry of selenium*. New York, NY: Marcel Dekker, Inc.

Chasteen TG (1998) Chapter 29—Volatile chemical species of selenium. In: Frankenberger WT Jr. and Engberg RA (eds.) *Environmental chemistry of selenium*. New York, NY: Marcel Dekker, Inc.

Frankenberger WT Jr. and Dungan RS (1999) Microbial transformation of selenium and the bioremediation of seleniferous environments. *Bioremediation Journal* 3: 171–188.

Jones L, Sever V, Lin Z-Q, and Bañuelos GS (2014) The source-partitioning of selenium volatilization in soil-*Stanleya pinnata* and *Brassica juncea* systems. In: Bañuelos GS, Lin Z-Q, and Yin XB (eds.) *Selenium in the environment and human health*. Boca Raton, FL: CRC Press.

Lin Z-Q and Terry N (2003) Selenium removal by constructed wetlands: Quantitative importance of biological volatilization in the treatment of selenium-laden agricultural drainage water. *Environmental Science and Technology* (3): 606–615.

Lin Z-Q, Cervinka V, Pickering IJ, Zayed A, and Terry N (2002) Managing selenium-contaminated agricultural drainage water by the integrated on-farm drainage management system: Role of selenium volatilization. *Water Research* (12): 3149–3159.

Losi ME and Frankenberger WT Jr. (1997) Bioremediation of selenium in soil and water. *Soil Science* 162: 692–702.

Pilon-Smits EAH, Banuelos GS, and Parker DR (2014) Chapter 6—Uptake, metabolism, and volatilization of selenium by terrestrial plants. In: Chang AC and Silva DB (eds.) *Salinity and Drainage in San Joaquin Valley, California: Science, Technology, and Policy*. Dordrecht: Springer Science + Business Media.

Ruppert L, Lin Z-Q, Dixon RP, and Johnson KA (2013) Assessment of solid phase microfiber extraction fibers for the monitoring of volatile organoarsinicals emitted from a plant-soil system. *Journal of Hazardous Materials* 262: 1230–1236.

Terry N and Zayed A (1998) Chapter 31—Phytoremediation of selenium. In: Frankenberger WT Jr. and Engberg RA (eds.) *Environmental chemistry of selenium*. New York, NY: Marcel Dekker, Inc.

Terry N, Zayed AM, de Souza MP, and Tarun AS (2000) Selenium in higher plants. *Annual Review of Plant Physiology and Plant Molecular Biology* 51: 401–432.

Vriens B, Lenz M, Charlet L, Berg M, and Winkel LHE (2014) Natural wetland emissions of methylated trace elements. *Nature Communications* 5: 3035.

Zayed A, Pilon-Smits E, de Souza M, Lin Z-Q, and Terry N (2000) Chapter 4—Remediation of selenium-polluted soils and waters by phytovolatilization. In: Terry N and Bañuelos G (eds.) *Phytoremediation of metal-contaminated water and soils*. New York, NY: Lewis Publishers.

Waves as an Ecological Process

CA Blanchette, MJ O'Donnell, and HL Stewart, University of California – Santa Barbara, Santa Barbara, CA, USA

Wave Mechanics

Of the physical process in the ocean, ocean waves are perhaps the most visible to the casual observer. The surface of the ocean is rarely still; slight puffs of wind send small ripples across still water. On exposed coasts, little ripples are usually unnoticed as walls of water rear up from the undulating sea surface and fling themselves against the shore with crashing roars and showers of foam. In shallow habitats, waves can exert an important influence on the distributions of organisms.

Waves on the surface of the ocean get their energy from wind blowing over the top of the water. As air moves over the water, friction causes the water to pull in the same direction, forming small ripples. With time and distance, little ripples become larger ripples and eventually form large waves, which can travel far from the location where they were originally formed. Actually, the idea of a wave traveling needs some explanation: waves are a rhythmic displacement of the surface of the water. The water itself moves in an elliptical pattern, and an individual particle of water returns to almost the same spot at the top of each wave. The waveform itself travels across the surface; what is really moving is the energy of the wave passing through the water. A wave can be described by several parameters: its wavelength, which is the distance between two subsequent peaks; its period, the time it takes for a complete wave cycle; and its amplitude, the vertical distance between the peak and the trough of a wave (**Fig. 1**). Waves travel across the ocean with some speed, known as the celerity. The disturbance of waves does not just occur at the surface; water beneath the surface is also moving in an orbital path, with smaller and smaller orbits until the depth of the water is half of the wavelength.

Once formed, ocean waves can travel long distances with very little dissipation. This means that storms in the middle of the Pacific Ocean send waves to both California and Japan. As they enter shallow water (where the water depth is less than half the wavelength), waves begin to change; the orbital motion of the water interacts with the bottom to become more elliptical. As a result, waves in shallow water undergo a process known as shoaling, becoming taller and steeper. In this region near the shore, waves begin to interact with the communities of organisms living on the bottom, subjecting them to oscillatory water motion. Over soft bottoms, waves will form ripples in the substratum, which help to structure the habitable space.

In very shallow water waves become so steep that they can no longer maintain their shape. The water below the wave is being slowed by drag on the bottom to a greater extent than the water at the top of the wave. At this point, the top portion of the wave will tumble over and begin to break. During breaking, all of the energy of the wave is dissipated in the turbulent motion of water on the shore. Breaking is a violent process, and can result in very high water velocities and high levels of turbulence. Most of the energy of breaking waves is expended in the intertidal zone and the shallow subtidal region. For organisms living in these regions on wave-swept shores, waves are one of the dominant features of their physical environment.

Life in the Wave-Swept Environment

The most direct mechanism by which waves influence populations is by removing or destroying individual organisms that live on wave-swept shores. The velocity of the water beneath shoaling and breaking waves can impose large forces on biological structures in their paths. The primary forces of moving fluid over a stationary object are lift and drag. Both of these forces increase with the velocity of the fluid squared, which means that a small increase in water velocity leads to a large increase in the force that an object experiences. The velocities beneath breaking waves can be very high (investigators have measured velocities in excess of 30 m s^{-1}), as can the forces exerted on organisms living beneath breaking waves. For an organism of the size of a golf ball on a moderately wave-swept shore, this can translate into forces as high as 200 N ($\sim$20 kg) that the organism must resist if it is to remain on the shore.

In coastal ocean regions, waves can impose large hydrodynamic forces on benthic organisms. If the forces imposed by waves are larger than the attachment to the substratum (tenacity) or the mechanical strength of organisms, they may become broken or dislodged (**Fig. 2**). The tenacity of mobile animals is often reduced when they move, raising the possibility that predators and grazers increase their risk of dislodgement when they forage. Breakage and dislodgment does not always lead to death; some organisms can regenerate broken parts or, if dislodged, may be able to re-attach and grow at another location. This may be an effective mechanism of dispersal for organisms that live in extremely wave-swept environments. For a large number of organisms, dislodgment does lead to death and several researchers have been experimenting with predictive models to describe survivorship as a function of wave height. These models are becoming more refined over time as more data become available to strengthen the link between wave height and maximal water velocity.

 doi:10.1016/B978-0-444-63768-0.00303-6

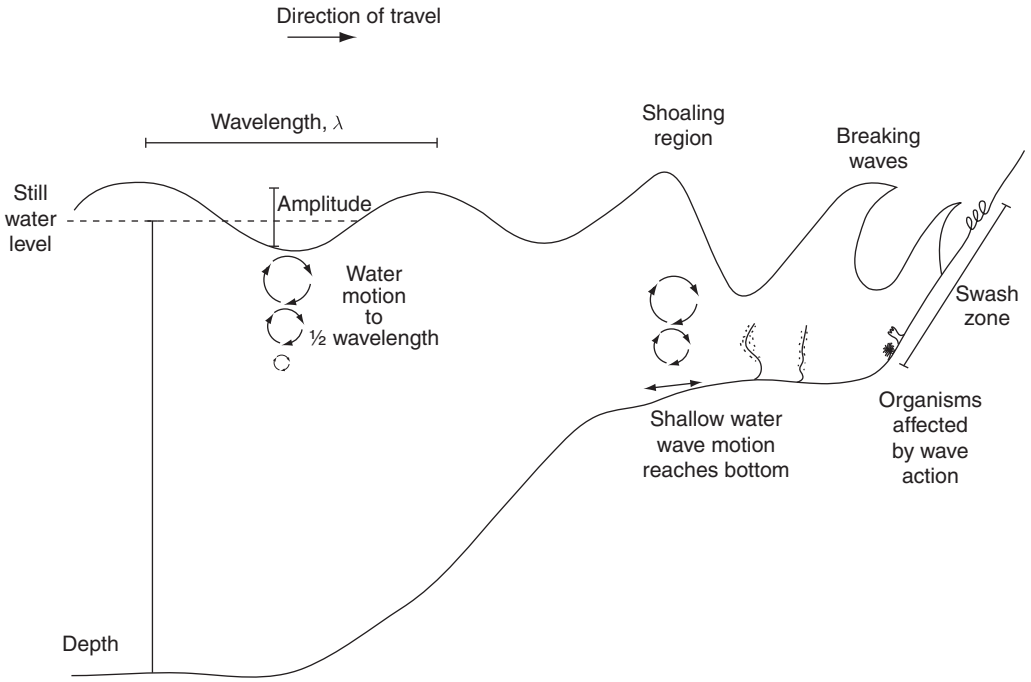


Fig. 1 Anatomy of a wave. Draft by Michael O'Donnell.

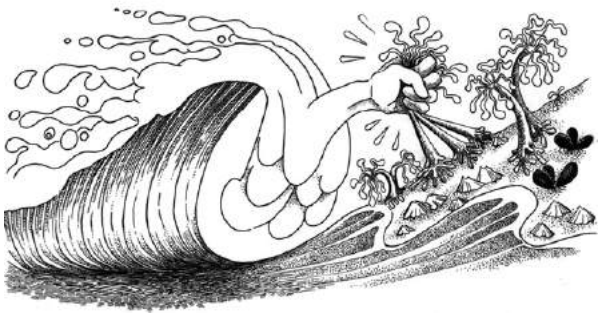


Fig. 2 Cartoon representing the opposing forces of drag imposed by waves resisted by the strength and tenacity of benthic organisms. Cartoon by Jeffrey Harding.

Strategies for Survival in the Wave-Swept Environment

From coral reefs to temperate intertidal rocky shores to sandy beaches, waves shape near-shore marine communities through interactions with individual organisms. The high energy of water moving in waves has important implications for wave-exposed organisms. There are several ways marine organisms deal with waves.

Behavioral Strategies

Mobile organisms in wavy habitats may time their forays into wave-swept areas to calm periods or high tide when the effect of the waves is less intense. For example, crabs on rocky shores walk around and forage between waves but grab on and assume a low, flattened posture that minimizes drag when waves hit. Seabirds run out of the way of waves as they forage in the swash on sandy beaches (Fig. 3), and burrowing invertebrates dive down into sand to find refuge from waves deep down in between sand grains. Not all organisms avoid waves, however. Seals and dolphins can be found surfing in waves and killer whales have been reported to use the froth of breaking waves as camouflage for hunting seals on the beach.

Fig. 3 Seabirds running on beach. Photo courtesy of Callie Bowdish.

Strategies of Sessile Organisms

Benthic sessile organisms that cannot move out of the way of wave action use a number of strategies that enable these organisms to withstand the forces imposed on them by waves.

Strength

Strength is a common strategy for persisting in wave-swept habitats. The strength of the attachment (by holdfasts of algae, byssal threads in mussels, and foot size in snails) of organisms exposed to waves is often higher than in calm habitats. Barnacles, snails, and limpets also have strong shells, which protect them against wave action. By retracting into their hard, strong shells, these animals can protect their soft bodies from the high energy of waves breaking directly on them. On coral reefs, the strength of scleractinian (reef building) coral skeletons enables them to thrive under waves breaking on the reef. The strong, stiff structures of corals enable them to resist forces from all directions imposed by waves. Many corals are strong enough to withstand the hydrodynamic forces that they experience in waves, and the substratum to which corals are attached often breaks before the coral itself.

Flexibility and elasticity

Another strategy for surviving waves is to be flexible. Flexible organisms do not require resources for structural support; therefore, this may be a less costly alternative than strength for survival in wavy habitats. Seaweeds provide a good example of this approach. Flexible algae are easily reconfigured into streamlined shapes by water moving around them. This reduces the force exerted on an alga by decreasing the area exposed to hydrodynamic forces. A flexible organism may also be pushed easily down toward the substratum where flow velocities are reduced by interactions with the benthos. A flexible organism can be reconfigured and pushed in alternating directions by the bidirectional flow of waves. The back-and-forth motion of flexible organisms in waves may also rub them against the adjacent substratum. This may damage the organism itself, but can also dislodge potential neighbors and keep the area around it free from competitors. Back-and-forth motions of flexible organisms in waves can also act to dislodge epibionts such as snails or epiphytic algae. These riders may prey upon the organism or may compete with it, as do epiphytic algae competing for light or nutrients with their host alga.

Another approach to survival in waves is used by the alga *Postelsia palmaeformis*. This alga resembles a small palm tree and stands upright, often sticking up in the air (**Fig. 4**). It thrives on the most wave-exposed areas in temperate intertidal zones, where individuals are knocked over by almost every wave. The stipe (equivalent to terrestrial plant stalk) of *Postelsia* is flexible and highly elastic. In combination with the shape of the stipe, this allows this alga to rebound back into an upright posture after it is pushed over by a wave and the wave has passed. In this way *Postelsia* thalli are reconfigured into streamlined shapes and pushed down toward the substratum where water velocities may be reduced, but they then passively spring back to upright postures by energy stored in their resilient stipes, much the same way an elastic band rebounds back into shape after being stretched. Similarly, other algae rely on buoyancy to return them to their upright postures after waves have pushed them over. Such algae are pushed against the bottom by waves, and then passively return to an upright position by the upward force of buoyancy when the flow slows.

Size and numbers

A flexible organism can survive waves by moving along with the flow. This strategy is effective for long organisms, such as kelps. By being longer than the displacement of the water in a wave, such kelps can avoid experiencing high hydrodynamic forces associated with waves. While the kelp moves along at the same velocity as the water in the wave, it experiences no water motion relative to its surfaces, and therefore no force. When the water in the wave reverses direction, the kelp is passively carried along with the water motion in the other direction. If the kelp never becomes completely strung out in one direction before the water velocity of the wave reverses, it can avoid the high forces associated with waves. In this way, large flexible, weak organisms can persist in areas exposed to waves.

Fig. 4 The sea palm, *Postelsia palmaeformis*. Photo by Carol Blanchette.

Small size is also thought to be a characteristic of organisms exposed to waves and, indeed, many of the organisms persisting in wave-swept habitats are small (e.g., many seaweeds, barnacles, snails, and limpets). Organisms capable of altering aspects of their morphology often have a smaller, more compact shape in wave-exposed than wave-protected habitats. However, the long, flexible kelps and large strong corals provide exceptions to this generality.

Some organisms form dense aggregations and this can help buffer then from the velocities and subsequent forces exerted by waves. Mussels, algae, and other organisms can often be limited in rocky substratum for attachment space. As a result, patches of densely aggregated individuals often form on suitable surfaces on wave-swept shores (Fig. 5). Water motion caught inside aggregations can be damped and slowed. As hydrodynamic forces are proportional to water velocity, this can lead to reduced forces inside aggregations, as interior individuals may be buffered from fast water velocities by neighbors.

The Benefits of Life in a Wave-Swept Environment

Because waves can be so destructive, scientists have devoted a great deal of attention to understanding the strategies that organisms have evolved to avoid or cope with damage. However, not all of the ecological consequences of waves are negative. The vast numbers and diversity of organisms living in wave-swept environments indicate that the tradeoffs required for survival there are worthwhile. Many aspects of biological processes are enhanced by wave action.

Reproduction and Fertilization Success

Waves can have significant effects on the numbers of young organisms that produce and release into a population. Beneath breaking waves, the tumultuous water motion is characterized by high turbulence. This turbulence can have important effects on the success of fertilization, and therefore the number of young produced in broadcast-spawning species. Sexual reproduction via the release of sperm into the water column is widespread among marine organisms. Given the limited swimming capabilities of sperm, if adults are separated by more than a few centimeters, some water motion is required to bring sperm and eggs together. Turbulent mixing due to wave energy can be advantageous to bring gametes into contact; however, turbulence will also influence the dilution rate of the gametes and imposes viscous forces on them. If these forces inhibit the attachment of sperm to eggs or damage the gametes or zygote, the advantages of mixing can be negated. Thus, turbulent water motion can either aid or hinder fertilization depending upon the species and the degree of turbulence.

Dispersal of Young

The vast majority of marine benthic animals have a planktonic larval stage, and the dispersal and settlement patterns of these larvae are important determinants of population dynamics. Larval transport is greatly affected by the near-shore flow regime. In this case, a wave phenomenon known as `internal waves' plays a role. Internal waves are similar to surface waves in being periodic undulations of a fluid interface, but internal waves happen within the ocean, at points where there are sharp gradients in the temperature or salinity of water. Such gradients often occur somewhere in the top 30 m of the ocean. Although internal waves cannot be seen by eye, they can be observed with thermometers mounted in the ocean. Though not as obvious as surface swell, internal waves have important ecological consequences. For example, tidally generated internal waves are accompanied by circulating cells of water near the surface, and on many shores these cells are advected shoreward with the internal waves. Larvae that can swim fast enough or are sufficiently buoyant to stay at the water's surface are concentrated in areas of downwelling between cells and are consequently carried inshore. Internal waves thus provide a mechanism for returning dispersed larvae to the shore where they can settle and recruit into the population. The mass transports and long-shore and rip currents accompanying

Fig. 5 *Turbinaria* aggregation. Photo courtesy of Hannah Stewart.

surface gravity waves provide alternative mechanisms by which larvae can be transported, in this case both on- and offshore. For any of these advective mechanisms, behavioral control by the larva over its position in the water column can affect the direction and rate of transport. Aside from advective transport, the process of turbulent mixing, common to wave-swept shores, may also disperse larvae.

Dispersal of Chemical Cues

Waves provide water motion that can spread settlement cues in the water column, but the back-and-forth motion of water in waves slows the rate at which chemical cues are advected away from their area of origin. As the cue moves in the oscillatory flow of waves, it becomes mixed with the surrounding water. Larvae of a coral-eating nudibranch have a settlement mechanism that takes advantage of this situation. When a larva experiences a settlement cue above a certain concentration, it stops swimming and sinks. Due to the oscillatory nature of water under waves, as the larva sinks it is moved back and forth but lands on the bottom roughly below the position in the water column where it sensed the cue. In this way, the simple behavior of the nudibranch and the water motion of waves provide a mechanism that increases the chances of a larva reaching its desired settlement site in a wavy environment.

Feeding and Nutrient Uptake

Turbulence in the benthic boundary layer is generated by the interaction of flow with the roughness of the substratum and can be augmented by mainstream turbulence from breaking waves. The intensity of this turbulence controls the rate at which suspended food and dissolved nutrients can be delivered to benthic organisms. If turbulent mixing is not sufficiently energetic, food or nutrients may become a limiting commodity. This effect has been demonstrated for populations of mussels in estuaries and for kelps in slow-moving flows. Nutrient limitation in kelps due to insufficient mixing appears to be a problem only at very low velocities.

Ecological Consequences

Disturbance and Patch Dynamics

Probably the most important and well-studied ecological effect of wave action is the effect of wave-induced disturbance to the community. Storms and intense wave action can be a leading agent of disturbance in marine communities. Disturbance plays a central role in nonequilibrium theories of community structure, including the concepts of disturbance theory, patch dynamics, and supply-side ecology. These theories attribute high species diversity and species coexistence to the processes of stochastic recruitment in a heterogeneously disturbed, patchy environment. The `intermediate disturbance hypothesis' proposes that species diversity is highest in communities that are subject to moderate levels of disturbance. Disturbances have been shown to be important in many marine systems. For example, the rate at which waves clear gaps in intertidal mussel beds of the northeast Pacific creates opportunities for other, less competitive, species to settle and grow (**Fig. 6**). Species diversity has been shown to be highest in intertidal boulder fields containing boulders of medium size, which are overturned more frequently than large boulders and less frequently than small boulders. Similarly, the rate at which storm waves cause breakage can have a controlling influence on the structure of coral reef communities.

Fig. 6 Patches created in an intertidal mussel bed by wave disturbance. Photo by Carol Blanchette.

Productivity

Intertidal organisms cannot transform wave energy into chemical energy, as photosynthetic plants transform solar energy, nor can intertidal organisms `harness' wave energy. Nonetheless, example after example finds that communities exposed to high wave action are more productive than similar communities in less wave-exposed areas. Despite severe mortality from wave-driven storms, communities at some wave-beaten sites produce an extraordinary quantity of biological structures per unit area of shore per year. Highly productive organisms such as the sea palm, *Postelsia palmaeformis*, are restricted to wave-beaten sites. Water motion is known to enhance the growth of aquatic organisms. In general, productivity of marine and freshwater plants is higher in moving than in still water, and it has long been known that coral reef growth is most vigorous on those margins of the reef where waves pound hardest. Wave-beaten reef platforms produce four times as much calcium carbonate per square meter per year as do those in protected lagoons. Increased exposure to waves does not always increase productivity. Along the southern coast of Chile, the subtidal kelp *Macrocystis* appears to grow best at intermediate levels of water motion: at the most exposed sites, storm waves tear these kelps away. In the northeastern Pacific, however, intertidal kelps do grow better in wave-beaten places, even though waves select stringently for small size, because winter storms shred the fronds of most kelps, and tear away many kelps and mussels. In general, intertidal zones of the northeastern Pacific are more completely covered by plants and animals the more exposed they are to wave action. The mechanisms for this enhanced productivity are complex and different for different systems. However, the phenomenon is well documented as an important component of how waves influence intertidal communities.

Climate Change and Wave Activity

Will the changing climate of the future be accompanied by changes in the wave climate? Recent investigation has shown wave and storm activity to be highly correlated with the intensity of El Niño-Southern Oscillation events. Although it is unclear what mechanism could account for this relationship, the presence of such a correlation suggests that local wave exposure can be strongly affected by the type of large-scale climate phenomena that are currently the subject of intense predictive efforts. Researchers have also noted an increase in wind stress averaged over large areas of the sea surface for the seas adjoining California, Peru, Morocco, and the Iberian Peninsula. Wind stress is a large contributing factor in the formation of waves. These studies suggest that substantial fluctuations in the severity of the wave climate may be a common phenomenon. The ability to predict ecological effects of waves on species distributions will become increasingly important in the face of climate change. For example, many intertidal species present on the central California coast are at or near the limits of their biogeographic ranges. A shift in the rate of disturbance that results in even a slight shift in the ability of a given species to persist may in this case result in a substantial shift in that species distribution. Thus, the ability to make accurate mechanistic predictions regarding wave-induced disturbance may augment our ability to predict future shifts in species distributions.

See also: Aquatic Ecology: Intertidal Zonation. Ecological Processes: Wind Effects. Global Change Ecology: Energy Flows in the Biosphere

Further Reading

Bascom, W., 1979. Waves and Beaches. Garden City, NY: Anchor Press.
Denny, M.W., 1987. Life in the maelstrom: The biomechanics of wave-swept rocky shores. Trends in Ecology and Evolution 2, 61–66.
Denny, M.W., 1988. Biology and the Mechanics of the Wave-Swept Environment. Princeton, NJ: Princeton University Press.
Denny, M.W., 1995. Predicting physical disturbance: Mechanistic approaches to the study of survivorship on wave-swept shores. Ecological Monographs 65, 371–418.

Denny, M.W., Blanchette, C.A., 2000. Hydrodynamics, shall shape, behavior, and survivorship in the owl limpet, *Lottia gigantea*. Journal of Experimental Biology 203, 2623–2639.
Denny, M.W., Daniel, T., Koehl, M.A.R., 1985. Mechanical limits to size in wave-swept organisms. Ecological Monographs 55, 69–102.
Hadfield, M.G., Koehl, M.A.R., 2004. Rapid behavioral responses of an invertebrate larva to dissolved settlement cue. Biological Bulletin 207, 28–43.
Kampion, D., 1989. The Book of Waves. Niwot: Roberts Rinehart Publishing.
Koehl, M.A.R., 1982. The interaction of moving water and sessile organisms. Scientific American 247, 124–132.
Koehl, M.A.R., 1984. How do benthic organisms withstand moving water? American Zoologist 24, 57–70.
Koehl, M.A.R., Wertheim, A.R., 2006. Wave-Swept Shore, the Rigors of Life on a Rocky Coast. Berkley: University of California Press.
Pedlosky, J., 2003. Waves in the Ocean and Atmosphere: Introduction to Wave Dynamics. Berlin: Springer.
Stewart, H.L., 2006. Hydrodynamic consequences of flexural stiffness and buoyancy for seaweeds: A study using physical models. Journal of Experimental Biology 209, 2170–2181.
Vogel, S., 1996. Life in Moving Fluids. Princeton, NJ: Princeton University Press.

Wind Effects

W Eugster, ETH, Zürich, Switzerland

Introduction

There are two categories of wind effects in ecology: (1) the effect of the vegetation surface on the wind, how it lowers wind speed near the ground, shelters niches from strong winds where small animals and plants can establish and live; and (2) the effect that wind and turbulence excert on many aspects of animal behavior, plant growth and survival, and the overall metabolisms of organisms.

Studies on forest recovery in North America have pointed to the important role of high winds in temperate forests. Although such catastrophes are rare, they could be instrumental in the creation and maintenance of mosaic patterns and hence the diversity of these woods. Some attempts to reconstruct the history of winds during the past glacial maximum (about 18 000 years ago) indicate that tropical storms generating winds of hurricane force were scarser, less intense, and shorter than those of the present day, with important consequences for forest ecology which include the influence on development, structure, and composition of the migrating and reassembling forests of the mid- and higher latitudes. However, direct evidence of the effects of wind on forests is hard to come by, and also other aspects of wind effects in a wide variety of ecologically relevant topics are rather scarcely covered in the scientific literature. Almost all studies reviewed in this article are based on an ecological question that suggest some partially unknown dependence on environmental variables. Mostly, temperature, precipitation, and humidity are considered important variables in such investigations, and wind effects are rather considered a possible or likely additional side effect of the overall ecological process under investigation. It is therefore not surprising that there are almost no systematic studies in the scientific literature that cover all aspects of wind in all details.

Some specific aspects, where the ecological importance of the wind is rather obvious, are covered in much greater detail in other topics. Thus, here we focus on the very general physical relationship between wind (and thus turbulence) and other environmental factors such as thermal heat loss and metabolic rates of organisms, and the exchange of trace gases such as CO_2 shall be addressed before summarizing the most relevant specific aspects of wind effects in the ecological literature.

Mean Wind Speed and Turbulence

Wind is a vector variable, but in many scientific applications only the scalar wind speed is investigated or of interest. Since wind – that is the term for the atmospheric motion over the solid surface of the Earth with respect to the surface itself – is mainly driven by pressure gradients on relatively large scales over the globe, this term is often used as a short-cut for horizontal mean wind speed. Over sufficiently long observation intervals and over large surface areas, there is no net loss or gain of air due to vertical motion; thus, the vertical component of the wind vector is considered to be zero. Hence, the wind vector is approximated as a two-dimensional entity that can be described by the scalar horizontal wind speed and the wind direction. All standard weather stations use this basic concept for recording wind. This is however not necessarily the best possible simplification for small-scale and short-term investigations, and thus has important implications for ecological processes. As an example, three-dimensional wind gusts in autumn can easily pick up fallen leaves from the ground, despite the fact that the leaves are relatively heavy and the mean vertical and horizontal wind speed over an hour or longer may be rather low. But on the timescale from tenths of a second to several minutes, turbulence – which includes such wind gusts – can be the most relevant wind effect. The turbulent timescale typically extends up to 1 h, whereas longer timescales are associated with mean wind effects. There is not a sharp separation between turbulence and mean wind, although a spectral gap between the two timescales has been postulated by some scientists. In reality, there is a confounding effect with the diurnal course of wind speeds that show different and locality-specific conditions during the day as compared to the night.

Laminar Flow and Turbulent Winds

Turbulence is generated inside a laminar flow when there is mechanical friction or thermal convection that perturbs the flow. **Fig. 1** illustrates this for a laminar flow with a certain wind speed that moves from a smooth onto a rough surface. At a certain distance downwind of the leading edge of this increased roughness, the laminar flow becomes chaotic, that is, turbulent. However, this turbulence does not completely reach the surface, a minute laminar surface layer always exists, over the surface of any object, including plant leaves (**Fig. 2**). Turbulent exchange of heat, moisture, CO_2, and other trace gases is by far more efficient than exchange in laminar flows (see below). Once the air is turbulent the flow does not easily become laminar again since the transition from turbulent to laminar flow is not as clearly defined at it is in the opposite direction. The decay of turbulence in the air is subject

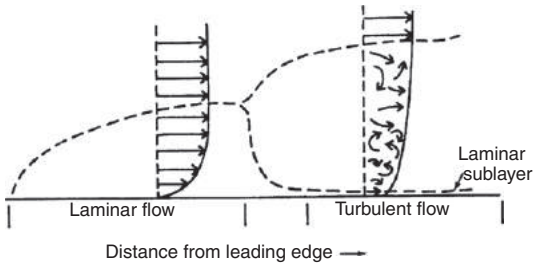


Fig. 1 Transition from laminar to turbulent boundary layer as wind blows over a vegetation surface changing from smooth to rough. Modified from Grace, J., 1977. Plant Responses to Wind. London: Academic Press.

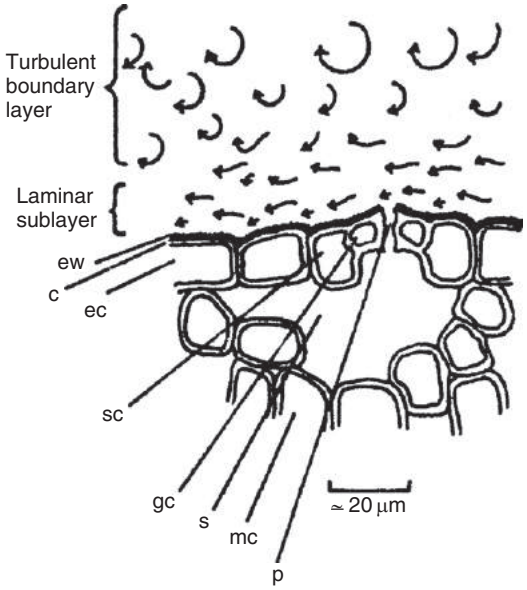


Fig. 2 Diffusion pathways at a leaf surface on a windy day: c, cutin; ec, epidermal cell; ew epicuticular wax; gc, guard cell; mc, mesophyll cell; p, pore; s, substomatal cavity; sc, subsidiary cell. Modified from Grace, J., 1977. Plant Responses to Wind. London: Academic Press.

to the physical rules of turbulent kinetic energy dissipation that ends up in the Brownian motion of single molecules and thus dissipates kinetic to thermal energy.

Wind has a kinetic energy,

$$E_k = \frac{m}{2}V^2 \quad [1]$$

with m the unit mass of the unit air volume, and V the scalar wind speed of the three-dimensional wind vector, that is composed of mean kinetic energy and turbulent kinetic energy. At moderate to high wind speeds the mean kinetic energy is by far greater than the turbulent kinetic energy. In addition, wind carries a momentum,

$$\tau = mV \quad [2]$$

At high wind speeds, especially during storms, a very high kinetic energy (both mean and turbulent components increase with increasing wind speed) may result from the wind, which is responsible for the devastating damages by hurricanes and other storm events, that however are also important for producing gaps (wind throws) in forest ecosystems and thus for these ecosystems' overall life cycle. Turbulent motions that are most relevant at lower wind speeds do not have such a devastating effect when mean wind speed is low. The kinetic energy of the mean wind is what wind mills profit from, and also migratory birds benefit from this component. Near the ground, the vegetation has to absorb both the kinetic energy of the wind and its momentum, but in this case the momentum absorption is by far the more relevant process and as a first approximation only momentum transfer by the vegetation is considered, neglecting the additional effect of energy absorption.

For flying birds it is mostly the kinetic energy of the wind that influences their daily life. It has been shown for Sandwich Terns (*Sterna sandvicensis*) on the isle of Griend, the Netherlands, that their loss of prey to the competing kleptoparasitizing Blackheaded Gulls (*Larus ridbundus*) significantly increases with wind speed. Thus, wind directly reduced the ability of Sandwich Terns to defend

their prey (mostly fish) against attacks of Blackheaded Gulls. This had a negative effect on the amount of food transported to the colony, while kleptoparasitism increased. Therefore, wind speed severely affected energy intake of the chicks and had strong negative effects on chick growth. During the first two weeks posthatching, kleptoparasitism was relatively low and had only small effects on chick growth, even under unfavorable weather conditions. From then on, the negative effects of kleptoparasitism on growth became considerable.

Wind over the Vegetated Surface

Vegetation is the most important interface between the atmosphere and the solid ground in terrestrial ecosystems. On the one hand, vegetation adapts to wind conditions and special plant social community compositions are found in the Arctic and Alpine environments where persistant and strong winds influence exposed locations such as hills, mountains, and crests. On the other hand, vegetation makes the Earth's surface rougher than what would be the case over bare soil (**Table 1**), and thus strongly influences the wind speed (**Fig. 3**) and direction in the atmosphere near the ground. The wind is driven by pressure and temperature differences on large scales, whereas the Earth's surface does not move and becomes stationary under most occasions. Exceptions are very special conditions during hurricane-force winds, and certain exposed locations with corresponding soil conditions, where bluffs are created by the steady wind movement. Under normal conditions, wind speed at some nonzero height above the ground must be zero to fulfill the criterion that the vegetation stays in place. Rough vegetation such as forests excert a much higher roughness ($\sim$1 m) to the atmospheric wind motion than shortcut grass (on the order of millimeters to centimeters; see **Table 1**). Based on this roughness the increase in wind speed with height above the vegetation depends strongly on vegetation type and structure. This vertical wind speed profile tends to increase logarithmically with height above the canopy (**Fig. 3**), and the physical process responsible for this is momentum absorption. Tall vegetation such as forests absorb momentum with their roots, and also in the dynamic motion of the stems. Thus, under strong winds it depends on the rooting type of the tree and the wood quality whether a tree can be uprooted or whether the stem breaks at a certain height above the ground.

Table 1 Relations between canopy heights (m) and aerodynamic roughness length (m) for different vegetation types

Vegetation	*Type*	*Canopy height (m)*	*Roughness length (m)*
Forest	Tropical	32–35	2.2–4.8
	Coniferous	10.4–27.5	0.28–3.9
	Pine	12.4–15.8	0.32–0.92
Woodland	Trees	10–15	0.4
	Savannah	8–9.5	0.4–0.9
Crops	Vines	0.9–1.4	0.023–0.12
	Beans	1.18	0.077
	Corn	0.8	0.064
	Wheat	0.25/0.4/1.0	0.005/0.015/0.05
	Wheat stubble	0.18	0.025
Grass	Thick/thin	0.1/0.5	0.023/0.05
	Sparse	0.025/0.015/0.45/0.65	0.0012/0.002/0.018/0.039
Soil	Bare		0.001–0.01

After Garatt, J.R., 1992. The Atmospheric Boundary Layer. Cambridge: Cambridge University Press.

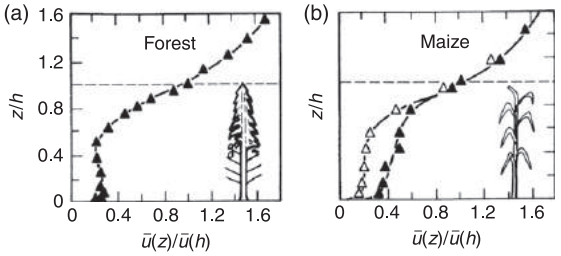


Fig. 3 Wind profiles (a) in a pine forest canopy of 16 m height (h), and (b) in a maize canopy of 2.1 m height. Both profiles show the mean horizontal wind speed, normalized for the wind speed at the top of the canopy ($z = h$). In the case of the forest the specific wind profile inside the trunk space ($z/h < 0.5$) a secondary maximum of the wind speed can be seen. Maize profiles for light winds of 0.88 m s^{-1} (closed triangle) and strong winds of 2.66 m s^{-1} (open triangle) at the top of the canopy are shown. From Raupach, M.R., Thom, A.S., 1981. Turbulence in and above plant canopies. Annual Review of Fluid Mechanics 13, 97–129.

Wind Pruning and Salt Spray

To unroot a tree normally requires strong gusts in heavy mean winds, as for example during storms. If winds are strong and steady, but not very gusty, then the energy may not be sufficient to unroot a tree and thus the wind-pruning effect may shape trees and shrubs (Fig. 4). Along the seacoast of British Guiana, along the subtropical shores of the island and Trinidad and southern California, and the subarctic shores of Hudson Bay and Labrador, the wind is reported to result in pruned trees. It appears that the steady subtropical winds have a similar pruning power as the icy blasts of the subarctic. However, this is not necessarily an effect of the wind alone. It has been argued that the proximity to the sea leads to a high load of salt spray in the wind, and that the toxicity of that salt may be the true ecological factor of wind pruning. Salt spray deposition on young shoots seems to actually kill many of them, thus causing the pruning. This observation is however only weakly based on pH readings along a transect from the shore of the Belcher Islands off the Hudson Bay coast, and it is also noted that the drying effect of the wind, possibly in combination with salt spray and other factors (ice and sand particles), may be as important.

Directional Growth Response

Besides pruning steady winds from a persistent wind direction can lead to directional growth response of trees, it is widely observed in coastal areas, in deep mountain valleys with a well-developed valley wind system (Fig. 5), or on wind exposed crests, rims, and hills. Since wind speeds generally increase with altitude from the lowland to the mountains, it has even been argued that high-altitude plants in wind-swept mountains may be less affected by global warming, and that the spread of lowland plant species into uplands as predicted by some global warming scenarios may be strongly restricted in higher altitudes due to the lack of adaptation of lowland plants to such steady and comparatively strong winds.

Changes in Surface Roughness: The Edge Effect

Sharp edges of vegetation – which are less abundant and less pronounced in natural ecosystems than in anthropogenically disturbed and shaped ecosystems such as agroecosystems and managed forests – are subject to wind effects that depend strongly on the distance to the change in roughness. When the wind first blows over a smooth (e.g., grass) surface and then abruptly has to change to a rough (e.g., agricultural crop or forest) surface, additional turbulence is created within a relatively short distance as wind passes over this roughness change (Fig. 6). This additional turbulence carries extra momentum that has to be absorbed by the vegetation downwind to obtain a new equilibrium with the rougher surface. This leads to the phenomenon that in a wheat field for example there may be a few rows of plants directly at the roughness change that seem quite unaffected even by strong winds, while only 1 m downwind one or several rows may be completely flattend by this additional momentum. In the case of forests, wind throw often excludes the trees at the forest edge, partially due to the same phenomenon. But trees also can adapt to constant wind pressures by building special cells to counteract this pressure. This is best known for trees in mountain valleys and along seashores with persistant and sufficiently strong winds in specific directions. In mountain valleys these are the up-valley (daytime) and down-valley (nighttime) wind directions. Which one is stronger depends on the complex combination of orientation of the valley, length, topographic differences in the surroundings, and more. But by studying trees which are leaning in the direction of the dominant strong winds (Fig. 5), it is easy to determine the locally dominant wind system. Near costs it is the diurnal sea breeze that dominates wind pressure on trees, while the nocturnal land breeze is in most cases much weaker.

On smaller scales, linear landscape elements such as hedgerows, tree lines, and tree lanes are ecologically important surface roughness elements, especially in otherwise rather smooth agricultural landscapes. In the Netherlands, for example, it has been

Fig. 4 Wind-pruning effect on *Metrasideros polymorpha* trees of a cloud forest on the Big Island of Hawaii (left). Under the strong onshore winds, leafs are detached from branches until only clusters of leaves at the outer margin of the tree volume remain (right). Photographs by Werner Eugster.

Fig. 5 Trees growing under conditions with persistently high wind speeds from a specific direction retain their asymmetric shape even when there is no wind. In this example from near Zweisimmen, Switzerland, the daytime up-valley wind (from right to left) shaped the characteristic habitus of these trees. Photograph by Werner Eugster.

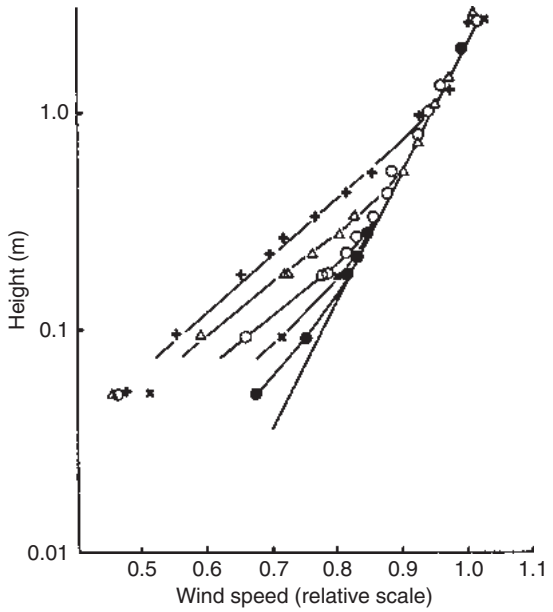


Fig. 6 Effect of a change from smooth to rough terrain. Fetch – that is the distance of uniform surface in the upwind direction – over the rough terrain was: (•), 0.32 m; (×), 1.18 m; (○), 2.32 m; (Δ), 6.42 m; (+), 16.42 m. Modified from Grace (1977) after Bradley (1968), *Quart. J. Roy. Meteorol. Soc.* 94, 361–379.

shown that among other possible functions (orientation clues, foraging habitat) such linear elements provide shelter from wind and/or predators for the two bat species *Pipistrellus pipistrellus* and *Eptesicus serotinus*.

Turbulent Mixing and Trace Gas Exchange

Turbulent exchange is roughly three to four orders of magnitude more efficient than diffusive mixing in a laminar airflow. For trace gas exchange between the atmosphere and the plants, the tiny laminar layer surrounding each leaf (**Fig. 2**) is thus non-negligible. Given this huge difference in effectiveness of turbulent versus diffusive transport a laminar boundary layer of 0.1–1 mm provides a similar resistance against the free exchange of CO_2 between the atmosphere and the plant stomates as does 1 m of turbulent air. In **Fig. 2** it is clearly seen that the laminar layer separates the turbulent atmosphere – where CO_2 is available in vast quantities – from the stomatal opening and the substomatal cavity, the buffer from where CO_2 is used for photosynthesis. Any changes in turbulence, wind speed, and wind direction will also affect the thickness of this laminar boundary layer and thus have an effect on the exchange of trace gases, heat, and momentum between plants and the atmosphere. **Fig. 7** shows that depending on plant leaf shape, the laminar boundary layer and thus the wind speed profile at varying distances from the leaf surface show a relatively large microscale variation.

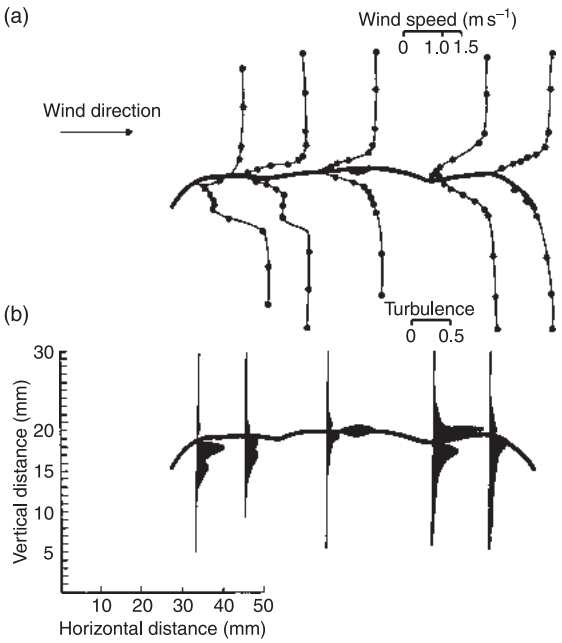


Fig. 7 The boundary layer over a *Populus* leaf. Profiles of (a) mean wind speed and (b) turbulence, shown in transverse sections in a laminar free stream. Modified from Grace, J., 1977. Plant Responses to Wind. London: Academic Press. After Grace, J., Wilson, J., 1976. The boundary layer over a populus leaf. Journal of Experimental Botany 27, 231–241.

On much larger scales, as wind blows over oceans and open water, it induces mixing of the surface layer, thereby enhancing the exchange of gases across the water surface, which is important for the oxygen content in the water and uptake or release of CO_2 and CH_4 to and from water bodies. Similar mixing occurs in the air above the surface which reaerates the plant canopy and essentially is responsible for resupplying photosynthetically active plants with CO_2 from the atmosphere, while at the same time O_2 produced by plants is carried away and mixed into the surface layer of the atmosphere.

Evaporation and Transpiration

A widely investigated topic of wind effects on ecosystems not covered in this article is found in the hydrological and biophysical literature on evaporation of water from ecosystems and transpiration from plants, either as a component of the hydrological cycle (the viewpoint taken by ecohydrologists), or in combination with CO_2 exchange (the ecophysiological viewpoint) for more information.

Dispersal of Pollen, Spores, and Microorganisms

The explosive pollen release from many wind-pollinated plants, particularly tree species with copious pollen production, is triggered by moderately gusty winds. Similarly, spores from the Swiss fern *Asplenium ruta-muraria* are released either by wind-induced shaking of the leaves (ballanemochory) or by the physical energy of impacting raindrops. In some palms (*Chamaedorea pinnatifrons* (Jacq.) Oerst. and *Wendlandiella* sp.) in Peruvian Amazonia the release of the pollen is triggered by movements of insects inside the flowers, and the term `insect-induced wind pollination' has been suggested since these insects do not normally also visit the female flowers of these palms.

On the ground, a wind gust can pick up small dust particles, sedimented pollen and microorganisms such as bacteria and mites from the surface or host organism. Once in the air, moderately turbulent winds are already sufficient to keep such small biotic and abiotic objects aloft. The general concept of updraft of a voluminous body in the atmosphere is described by Stoke's law of sedimentation, where the terminal falling velocity V_t of an object is

$$V_t = \sqrt{\frac{2mg}{c_w \rho A}} \qquad [3]$$

with m the mass (kg m^{-3}), g the gravitational acceleration (≈ 9.81 m s^{-2}), c_w the friction coefficient (≈ 1 for circular bodies, < 1 for aerodynamically formed bodies), ρ the density of the air (≈ 1.2 kg m^{-3} at sea level), and A the projected surface of the body (m^2). Fig. 8 shows the terminal fall velocity for small organisms of 1 μm to 2.5 mm and how typical vertical wind speeds in the air

can counteract the falling of such objects, once they are dispersed in the air. In this respect wind has almost exactly the same effect as the water flow in rivers: under high turbulence and horizontal speeds animals may find sheltered spots where they are not picked up by the motion of water or the air, but once they lose adhesive contact, their body size and weight may be too small to grasp ground again, and they become suspended in the fluid until they happen to end up in a calmer area where their settling velocity is greater than the wind (or water) motion, which allows them to reach the ground surface again. **Fig. 8** shows that bodies with a diameter smaller than ≈ 10 μm are normally too small to return to the ground in the turbulent atmosphere. Thus, for such small bodies, impaction becomes the most relevant process how they can be eliminated from the atmosphere, that is, when they physically hit the surface of a tree or another plant. The sticky stigma of a flower's pistil further helps to capture pollen even when impaction is weak.

Wind pollination is considered inefficient compared with insect pollination. This finding has led to the hypothesis that the rise to dominance of the angiosperms over gymnosperms at evolutionary timescales is due to reproductive innovations, especially those involving coevolution with biotic gene dispersers. This has most likely contributed to the present-day situation that conifers are biogeographically restricted to stressful environments where gymnosperms may suffer a comparative disadvantage if pollinators face persistently high wind speeds.

Influence on Small Animals and Seed Dispersal

Small insects need to adopt to wind speeds. Studies carried out in a wind tunnel indicate that weak winds >0.2 m s^{-1} already have an effect on the flight and landing behavior of the bug *Prostephanus truncatus* (Horn). In the open landscape, mosquitoes (*Anopheles marajoara* in Brazil) can only freely navigate in air with wind speeds below about 0.85 m s^{-1} (3 km h^{-1}). From the aphid parasitoid *Aphidius nigripes*, it is reported that males generally did not reach females at wind speeds of 1.0 m s^{-1}, as the majority of individuals taking flight in the pheromone plume (81.8%) was unable to sustain upwind flight. The general picture is that as wind speed increases these small animals increasingly lose control over their flight trajectory and may no longer target their prey or mate as desired. Swallows, for example, are known to fly close to the ground before thundershowers, where the prefrontal increase in wind speed restricts the activity of small flying insects to the few lowest meters close to the ground. In summer, when mosquitoe abundance is enormous in the Arctic, reindeer and caribou select windy locations for resting and rumination, preferably close to the seashore, on gravel pads in large rivers with a well-developed diurnal valley wind system, or on snow fields with thermotopographic wind resulting from the contrasting surface temperatures between snow and vegetation or rocky surfaces.

Some spider species benefit from this effect by letting themselves drift away – termed ballooning in the scientific literature – to explore new habitats using a short thread that increases their updraft and thus drifting distance at higher wind speeds. Although there are contradictory views between authors about the importance of various environmental factors, it is widely agreed upon the upper wind speed limit of 3 m s^{-1} for ballooning. Most work has been focusing mostly on the meteorological conditions at the time of take-off, whereas literature on the underlying motivation of the spiders and instigation of pre-ballooning behavior (climbing to a prominent point and silk release) is very limited and largely considered supposition by some experts. Since spiders are important polyphagous predators on arable farmland, the high mobility of ballooning species means that they are often the first to arrive in a crop newly infested with pests, and have a role in controlling the outbreak until more specific predators arrive.

In a similar way as ballooning spiders take advantage of the horizontal translocation by wind plants profit from the wind to disperse their seeds. The parachute type seeds of *Asteraceae* and the winged seeds of *Acer*, *Fraxinus*, *Ulmus*, and many coniferous trees are good examples of how plants benefit from available wind to spread out faster than would be possible without the help of the wind. For seeds that do not have wings, hairs, or parachute-type annexes, the Stoke's settling velocity (eqn [3], **Fig. 8**) applies and explains why in general small seeds are wind dispersed because of their long residence

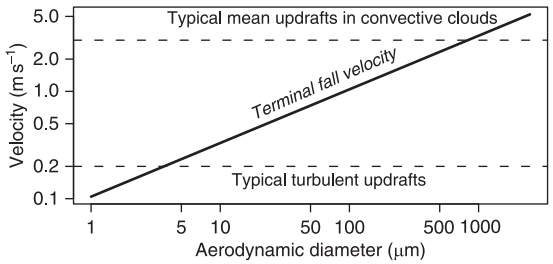


Fig. 8 The terminal fall velocity of ball-shaped objects (pollen, seeds, bacteria, microorganisms) compared to typical updrafts in the turbulent atmosphere (up to ≈ 0.2 m s^{-1}) and typical mean updrafts in convective clouds (thunderstorms). The thick line applies to objects that have a similar density as water. If updrafts are stronger than the terminal fall velocity, then an organism or particle in the atmosphere remains suspended in the atmosphere and will only deposit on obstacles such as trees due to impaction. Larger organisms that are subject to a high terminal fall velocity need active means to keep themselves aloft.

time in the atmosphere (the residence time is inversely proportional to the terminal fall velocity) than large seeds, that lead to small dispersial kernels downwind of the seeder plant unless the seeds are dispersed by animals. A special case exists for vegetation near open water bodies, where large buoyant seeds can float on the water and be dispersed by its currents, such that even large and heavy seeds can be transported over long distance that would not be possible by the wind alone.

Nature has brought about a wealth of shapes and forms of seeds that do not correspond with the simplest version of a spherical with $c_w = 1$. For some plants, specific wind tunnel studies have been carried out to determine the true dispersal capacity of seeds. For six Canadian perennial grassland species with different seed aerodynamic attributes, it was investigated how dispersal distances vary with varying wind speeds and release heights. Dispersal distances of long-range dispersed seeds (99th percentile values) increased exponentially with wind speed. At wind speeds of 14 m s^{-1}, predicted maximum distances were 10–15 m for small and relatively heavy spherical seeds and 20–30 m for large and relatively light cylindrical or disk-like seeds. In the study area, wind gusts >10 m s^{-1} at plant height occur at least annually, and plants of the selected species live up to several decades. This suggests a great potential for long-range dispersal during the lifetime of a plant. It is argued that plants may gain wider dispersal of seeds by increasing the release height (e.g., taller infructescences) and by requiring stronger winds to release seeds (e.g., dispersal in autumn and winter).

Transport distance is one aspect, the other is the release of seeds from the flower heads by wind speed. In a wind tunnel study with flower heads of two thistle species, *Carduus nutans* and *Carduus acanthoides*, with ripe seeds, the effect of laminar versus turbulent flows of increasing velocity was investigated. Seed release increased with wind speeds of both laminar and turbulent flows. However, far more seeds were released, at significantly lower wind speeds, during turbulent flows. In other cases, the seeds are primarily dispersed by the wind, followed by secondary dispersal by rodents living on the ground which collect the seeds and cache them in the soil. Treatment by rodents, primarily yellow pine chipmunks (*Tamias amoenus*), of four species of pine seeds, lodgepole pine (*Pinus contorta*, 8.7 mg seed weight), ponderosa pine (*Pinus ponderosa*, 55 mg), Jeffrey pine (*Pinus jeffreyi*, 157 mg), and sugar pine (*Pinus lambertiana*, 213 mg), that vary in size and weight was studied in the Carson Range of western Nevada. For the species examined, seed size appeared to have had little effect on several other attributes, including mean dispersal distance, substrate choice, and microhabitat choice. It was found that although a larger seed size and weight decreases primary wind dispersibility of pine seeds, the secondary dispersal by scatter-hoarding rodents compensates for poor wind dispersal so that total dispersibility of large-seeded pines is not compromised.

Influence on Bird and Insect Migration

At a much larger scale many long-distance migratory bird species have adopted flight tracks that best profit from large-scale wind fields on the Earth, which saves energy and thus increases the survival rate. On the other hand, migratory birds that are facing strong headwinds, may suffer severe losses if such an occurrence combines with low temperatures, scarce food resources or the like. In general, the nocturnal and diurnal wind directions and speeds are not necessarily the same. Near the ground the so-called low-level jet, a relatively strong wind with its maximum speed at only 100–300 m above the ground surface is active at night, whereas the atmosphere may be calm during the day, and wind speeds are higher aloft.

In a study on the migration patterns and environmental effects on stopover of monarch butterflies at Peninsula Point, Michigan, it was found that wind direction had a significant influence on the number of monarchs recorded on each count over a 7-year period, with higher counts during north winds.

On smaller scales, it has been shown for two dragon fly species, *Pantala hymenaea* and *Pantala flavescens*, in natural flight over a lake at ambient wind speed and direction, that they are able to compensate at least partially for crosswind drift, which shows evidence for use of a ground reference to correct for drift when flying over water, and their ability to cope with much higher wind speeds (5.0 m s^{-1}) than small insects are able to.

No Wind Effect?

Although there are many studies that found ecological effects of wind on animals, plants, and ecosystem processes, it should be remembered that there are other studies that were unable to find such effects. For example, the bat *Pipistrellus pipistrellus* in Oxfordshire did not show an apparent response to wind nor rain in the time spent outside the roost. This is remarkable since it feeds on insects and one might expect a behavior similar to the one known from swallows. Since it is generally difficult to publish negative results in the scientific literature, and moreover wind as a three-dimensional vector variable makes it particularly challenging to derive the relevant information from simple measurements (e.g., if mean wind speed was measured when actually turbulent kinetic energy would have been the variable with higher predictive power), the lack of clear statements on where wind does not have an effect should not come as a surprise.

In forest ecology it has been postulated that there are two main factors why biotic effects of wind have not been well studied: (1) the difficulty of measuring wind in the field and separating its effects from the confounding variables of temperature and humidity; and (2) the expense involved in carrying out wind tunnel experiments in the laboratory.

Wind Chill and Heat Index

The bioclimatic temperature sensed by an organism can differ considerably from the absolute physical temperature that is measured by conventional instruments. For humans, elaborate concepts to compute a wind chill temperature have been established to account especially for the effect of wind. The concept bases on the knowledge that increasing mean wind speeds increase also turbulence, and thus the heat transport away from an organism that has a warmer skin temperature than the atmosphere. Although it is widely known that the ambient moisture or humidity in the air has an additional influence, for the sake of simplicity most approaches only consider wind speed as a specific factor when considering wind chill.

Controlled experiments with humans were carried out to determine the functional relationship between wind chill and preceived temperature. This was not possible with primates, where thermoregulation is known to be an important ecological constraint. Shade temperatures, solar radiation, humidity, and wind speed all serve to alter an animal's `perceived' temperature. In a recent review, three thermal indices currently available were compared. Black bulb temperatures can account for the effect of solar radiation, with wind chill equivalent temperatures and the heat index providing quantifiable estimates of the relative impact of wind speed and humidity, respectively. The authors presented three potential indices of the `perceived environmental temperature' that account for the combined impact of solar radiation, humidity, and wind speed on temperature, and performed a preliminary test of all of the climatic indices against behavioral data from a field study of chacma baboons (*Papio cynocephalus ursinus*) at De Hoop Nature Reserve, South Africa. It was found that the complexity of the interactions among environmental factors that influence thermoregulation in primates will require the development of biophysical models of the thermal characteristics of the species and its environment. Until such models are developed, however, it is concluded that wind chill and heat indices should permit a more detailed examination of the thermal environment, allowing thermoregulation to be given greater precedence in future studies of primate behavior.

Another widely established approach is not to try to compute a bioclimatic temperature or index, but relate the metabolic energy consumption of an animal to environmental factors.

Metabolic Stress by Wind

Small animals can profit from the presence of a laminar sublayer (Fig. 1) even under highly turbulent conditions. Due to the much lower heat exchange in that laminar layer, they may avoid metabolic stress under high winds. This is almost impossible for larger animals, such as breeding arctic shorebirds. It was found that tarsus length in all shorebirds breeding in the Canadian arctic shows an evolutionary response to average metabolic stress encountered across the breeding range, such that birds nesting in metabolically stressful environments have relatively shorter legs. Longer-legged birds living in colder environments will experience greater metabolic costs because their torsos are elevated farther away from the ground's wind-dampening boundary layer. It was suggested that the widely known Allen's rule that relates the metabolic rate of an organism to its volume should be extended: body-supporting appendages of homeotherms may be shorter in colder environments so as to take advantage of a boundary layer effect, thereby reducing metabolic costs.

Another study that investigated the effects of water levels and weather on wintering herons and egrets found that larger and longer-legged species tended to be found in deeper water, although both species frequently were found together in shallow water. Severe weather with high winds caused the birds to suspend foraging and remain sheltered from the wind. Consequently, a higher percentage of smaller heron and egret species did not survive severe storms since searching shelter from wind meant fasting. A 3-day storm period was simulated to lead to $>10\%$ decline in body mass of the smaller herons and egrets.

Wind Throws and Wild Fires

Extreme events with high wind speeds are important in the life cycle of many ecosystems, especially forests. Hurricanes in the tropics, tornadoes, and other windstorms further north and south reshape forest ecosystems via windthrows that eliminate the weakest and thus most often the oldest individuals in the forest canopy. For example, in New England forests, leaning is the most prevalent damage to young stands, whereas breakage and uprooting dominated in older stands. Breaking was slightly more important in older conifer than hardwood stands, comprising 6–14% of the stems and generally occurring 1–5 m from the ground, but numbers vary not only widely between species and stand composition but also among storm events in the same stand.

Since heavy storms are often accompanied with severe lightning strikes, the wind effect can easily be a combination of wind and fire. When a wildfire starts then the wind conditions will strongly determine how quickly the fire advances with the wind, and what damage is done to the ecosystem. In some cases, such as the Bishop pines, the fire even is necessary to free the seeds in the cones and initiate the life cycle of this forest type. In the gaps the new vegetation can resprout, and since more light and precipitation reaches the ground, this provides niches and living space for early successional plants. As a consequence also the fauna may be affected. Organisms with a life size that is much smaller than gaps in forests may only find a suitable ecological niche in the gaps that shift their location over the years. In subalpine forests of the Swiss alps, it was found that the gaps created by windthrows add considerably to the species diversity of macrofungi. Larger animals such as black bears in southeast Alaska were found to react in just an opposite way: 58% of the den sites were found in forests that were most protected from catastrophic storm

effects, and only 6% in forests most exposed to storm damage. These results suggest that the effect of catastrophic windstorm disturbance on overwinter habitat for black bears is the key factor influencing the site selection for black bear dens.

See also: Ecological Processes: Physical Transport Processes in Ecology: Advection, Diffusion, and Dispersion. Ecosystems: Estuaries. General Ecology: Seed Dispersal

Further Reading

Cartar, R.V., Morrison, R.I.G., 2005. Metabolic correlates of leg length in breeding arctic shorebirds: The cost of getting high. Journal of Biogeography 32, 377–382.
de Gayner, E.J., Kramer, M.G., Doerr, J.G., Robertsen, M.J., 2005. Windstorm disturbance effects on forest structure and black bear dens in southeast Alaska. Ecological Applications 15, 1306–1316.
Doutt, J.K., 1941. Wind pruning and salt spray as factors in ecology. Ecology 22, 195–196.
Ellenberg, H., Strutt, G.K., 1988. Vegetation Ecology of Central Europe. Cambridge: Cambridge University Press.
Ennos, A.R., 1997. Wind as an ecological factor. Trends in Ecology and Evolution 12, 108–111.
Foster, D.R., 1988. Species and stand response to catastrophic wind in central New England, USA. Journal of Ecology 76, 135–151.
Garatt, J.R., 1992. The Atmospheric Boundary Layer. Cambridge: Cambridge University Press.
Geiger, R., Aron, R.H., Todhunter, P., 1995. The Climate Near the Ground. Braunschweig: Vieweg.
Grace, J., 1977. Plant Responses to Wind. London: Academic Press.
Grace, J., Wilson, J., 1976. The boundary layer over a populus leaf. Journal of Experimental Botany 27, 231–241.
Hill, R.A., Weingrill, T., Barrett, L., Henzi, S.P., 2004. Indices of environmental temperatures for primates in open habitats. Primates 45, 7–13.
Moore, P.D., 1988. Forest ecology: Blow, blow thou winter wind. Nature 336, 313. 313.
Myers, R.K., van Lear, D.H., 1998. Hurricane–fire interactions in coastal forests of the south: A review and hypothesis. Forest Ecology and Management 103, 265–276.
Raupach, M.R., Thom, A.S., 1981. Turbulence in and above plant canopies. Annual Review of Fluid Mechanics 13, 97–129.
Senn-Irlet, B., Bieri, G., 1999. Sporocarp succession of soil-inhabiting macrofungi in an autochthonous subalpine Norway spruce forest of Switzerland. Forest Ecology and Management 124, 169–175.
Stienen, E.W.M., Brenninkmeijer, A., Geschiere, C.E., 2001. Living with gulls: The consequences for Sandwich Terns of breeding in association with black-headed Gulls. Waterbirds 24, 68–82.
van Dorp, D., van den Hoek, W.P.M., Daleboudt, C., 1996. Seed dispersal capacity of six perennial grassland species measured in a wind tunnel at varying wind speed and height. Canadian Journal of Botany–Revue Canadienne de Botanique 74, 1956–1963.
van Gardingen, P., Grace, J., 1991. Plants and wind. Advances in Botanical Research 18, 189–253.
Vonlanthen, C.M., Kammer, P.M., Eugster, W., Bühler, A., Veit, H., 2006. Alpine vascular plant species richness: The importance of daily maximum temperature and pH. Plant Ecology 184, 1–9.
Weyman, G.S., 1993. A review of the possible causative factors and significance of ballooning in spiders. Ethology, Ecology and Evolution 5, 279–291.
Woodward, F.I., 1993. The lowland-to-upland transition modeling plant-responses to environmental change. Ecological Applications 3, 404–408.

Agriculture Systems

O Andrén, TSBF-CIAT, Nairobi, Kenya
T Kätterer, Department of Soil Sciences, Uppsala, Sweden

Introduction

An agricultural ecosystem is an ecosystem managed with a purpose. This purpose usually is to produce crops or animal products. Agricultural ecosystems are designed by humans, and current agroecosystems are products of a long chain of experimental work. These experiments have been performed by individual farmers as well as research institutions, and when results were positive for the purpose, the methods have been adopted.

The purpose has, however, changed with time. In highly productive regions, for example, Western Europe, the emphasis has changed from maximum productivity to environmental considerations, such as reduction of nutrient losses to groundwater and maintaining an open landscape with high biodiversity, etc. In less-productive regions, where resources such as water or fertilizers are scarce and production is too low to properly feed the farmer, environmental considerations have low priority. This is a major global problem, since this leads to land degradation and even lower production, etc. in a downward spiral.

Agroecosystems are conceptually fairly similar to managed forests and grasslands, and whether extensively cattle-grazed natural grasslands should be included under the category of agroecosystems is a matter of choice in the individual case. Arable land is defined as land that is soil cultivated regularly, but also here the boundaries are not sharp (seminatural grasslands, permanent crops, etc.). At the other end, agroecosystems border horticultural systems, that is, vegetable cropping. Alternatively, horticulture can be viewed as a subset of agriculture. Production of cabbage in a field can be considered as agriculture, but hydroponic (soil-less) production of tomatoes in a greenhouse under artificial light can perhaps not be included. However, in many respects even an artificial ecosystem such as this can be considered as an agricultural ecosystem. It is designed for production of a crop and is just managed to a higher extent than an arable field.

According to FAO statistics for 2002, agricultural ecosystems comprise almost 40% (5 Gha) of the total land area of the Earth. About 11% of the total land area is arable land (cultivated with crops), and approximately 27% of the total land area is under permanent pasture, grazed by cattle, goats, sheep, camels, etc. Clearly, we are actively managing a considerable part of our planet for agricultural purposes, and to this one can add other similar systems, such as intensively managed forest systems (planted and harvested, sometimes fertilized), etc.

Ecological research performed in agricultural systems has many advantages compared with research in most natural ecosystems. For example, there are a number of long-term field experiments running, although originally designed for, for example, crop production response to fertilizer dose, that can give us a 30-year integration of what has happened, for example, to organisms in the soil under different conditions. Further, agricultural fields are `homogenized', that is, trees, larger stones, etc. are removed and regular soil cultivation evens out differences in topsoil properties over time. However, even after many years of cultivation, a fairly high variability in soil properties remain, which is the incentive for `precision farming', where soil and crop properties are measured at high resolution (m^2), and management is based on these measurements. For ecological research, this is an opportunity, since any given hectare will yield numerous observation points, each helping us to answer questions such as: Why does this particular location yield more wheat, or why is more water present at that location?

Another advantage is that agricultural crops often have a short lifespan and a small size, compared with, for example, forest trees. Often, an experiment can be started when the soil is bare, and a single crop can be followed from sowing, through harvest, and finally when the stubble is plowed down at the end of the growing season. This life cycle can take a century for a tree in a northern forest, which, to add insult to injury, also may contain several other plant species. Therefore it is not surprising that a considerable part of modern ecological theory (predator–prey interactions, general soil ecology, above- and belowground plant growth dynamics, organic matter decomposition, nutrient mineralization, etc.) is based on work performed in agricultural land, and that the reluctance of ecologists to work in agricultural systems that was obvious 30 years ago seems to have vanished.

The Agroecosystem

Fig. 1 is an attempt to summarize the characteristics of an agroecosystem as compared to most natural ecosystems. Note that this comparison is between a typical natural ecosystem and a typical, high-production agroecosystem.

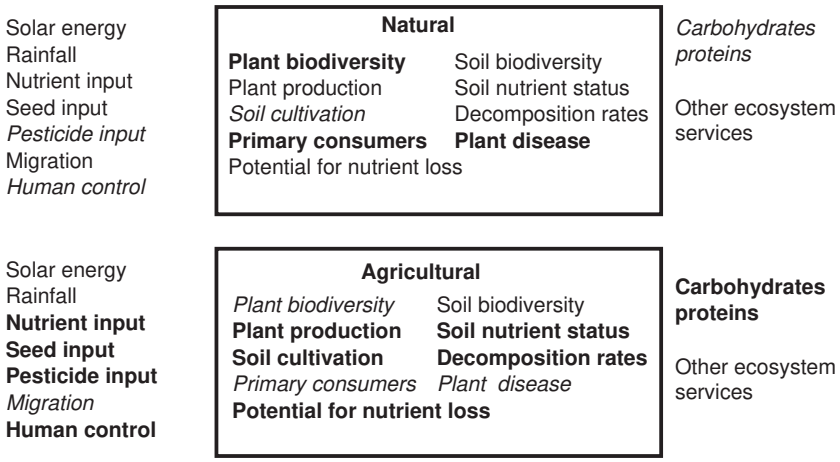


Fig. 1 Similarities and differences between typical natural and high-production agricultural ecosystems. Inputs of energy, mass, and control (left), comparison of selected ecosystem properties (center), outputs (right). Note that cattle, etc. are not included as primary consumers here. (**Bold** = markedly higher value than in the other ecosystem type. *Italic* = very low.)

Abiotic Constraints

Just like natural ecosystems, agroecosystems are constrained by climate and soil properties – maize does not grow in Northern Sweden. However, climate can be modified, that is, in dry climates one can irrigate (with surface- or groundwater), and soil properties can be modified through, for example, liming, organic matter amendments, and fertilization. Too high water tables can be lowered through ditching or tile draining.

Nutrients

Highly productive agroecosystems need high inputs of plant nutrients (nitrogen, phosphorus, potassium, and other elements) to replenish the nutrients removed with the exported products. These inputs can be delivered either as commercial fertilizer, recirculated sewage sludge and ash from garbage burning, or manure from cattle, pigs, poultry, etc. All sources have their advantages and disadvantages. Commercial fertilizers are well defined, low in pollutants such as heavy metals (although exceptions exist), hygienically safe, and are concentrated, easy to transport, and rapidly available to the plant when applied in the field. However, production and long-range transport of fertilizers is energy consuming, and the concentrated product increases the risk for too high doses, leading to environmental pollution. An even greater problem is that a large part of the farmers of the world cannot afford to buy enough fertilizer to maintain soil fertility and obtain good yields. In all, world N fertilizer production in 2001 was slightly less than 90 Mt, very unevenly distributed. In sub-Saharan Africa, only 1.1 kg fertilizer nitrogen is used per person and year, whereas in China the corresponding value is 22 kg.

In theory, recirculation of nutrients from waste of the exported products seems to be ecologically sound. In practice, there are a number of problems. First, sewage sludge mainly consists of water, which either must be removed (requires energy) or transported, which is expensive and impractical. Second, sewage sludge contains harmful bacteria, human parasites, etc. and has to undergo hygienic treatment. Third, and most severe, is the problem with contaminants, such as heavy metals and organic toxins. Therefore recirculation of sewage sludge and garbage incineration ash is strictly regulated in most countries. In this perspective, replacement of nutrients using newly produced fertilizer can be a better solution from an environmental viewpoint.

Naturally, animal manure produced on the farm should be and is recycled to soil as much as possible. Compared with fertilizers, manure has the advantage of containing organic matter, which improves soil structure. On the other hand, manure contains mostly water (expensive storage and transportation, heavy machinery needed for spreading), and it will lose nitrogen through ammonia emission, both at storage and spreading.

Crops, Varieties, and Cropping Systems

The vegetation found in an agroecosystem is usually divided into crop and weeds, where weeds are unwanted trespassers, which traditionally have been regarded only as negatives. More recently, this view has been modified, and weeds, particularly weedy border zones can be accepted to some extent, as biodiversity enhancers and refuges, for example, for beetles.

The crops used today are products of many years (in some cases millennia) of plant breeding, and properties selected for are usually productivity, product quality, pest resistance, etc. This directed selection, in recent years augmented by direct manipulation of DNA, is one of the main differences between agro- and natural ecosystems. Crop species and varieties are being redistributed all over the world; maize, a staple food in Africa, comes from Central America, common West European and North American cereals such as wheat come from the Middle East, etc. This breeding and distribution of improved crops, together with improved

cultivation/fertilization techniques probably is the main reason for the global success of the human species (three billion in 1960, probably nine billion in 2050). For example, world grain production was 631 Mt in 1950, and in 2000 it had increased to 1840 Mt.

Herbicides, Pesticides, and Fungicides

To reach the goal of high production of crops of good quality, weeds (unwanted plants), pests (unwanted animals), as well as fungal, bacterial, and viral diseases must be kept in check. A monoculture crop is vulnerable to attacks, since one (or a pair) of the pests that enter a field will have a high concentration of food with no transport stretches in between. Potential predators may be absent, since they may need a litter layer on the ground for reproduction, which does not exist in the field, etc. Repeated monocultures may build up specialized pests, such as plant parasitic nematodes. Crop rotations (switching crops from year to year according to a predetermined pattern) can successfully deal with many pests and diseases, and careful soil cultivation can reduce weed problems. Intercropping (growing two or more crops together, such as barley/clover) may also help.

However, most fields will benefit from occasional chemical (or biological) pesticide/herbicide treatment. These types of agrochemicals have a somewhat dubious reputation among laymen and perhaps also ecologists (DDT, Agent Orange, mercury, etc.). Three things should be kept in mind, though. First, the substances and formulations used today are thoroughly tested before approval, and their side effects and the fate of their decomposition products are well known. Second, chemical warfare is common in natural systems – all successful plant species present today have at least some chemical defense against microorganisms and pests. Third, which alternatives do we have? A failed crop in a well-fertilized field will lead to high risks for nutrient losses to the environment. A failed crop in poorer conditions may lead to starvation for the farmer and her family.

Alternative methods, such as increased cultivation, hand weeding, or biological pest reduction by introduction of predators all have their advantages and disadvantages, but there is no `silver bullet' available. In summary, an integrated approach with a combination of methods is the solution, and modern agriculture has moved and is moving in this direction. Of course, for commercial reasons it can be profitable to cultivate, for example, `organic' crops (without fertilizer or pesticides) to obtain a higher price, but from an ecological or environmental viewpoint this approach is not necessarily better.

Agriculture can thus be classified according to the use of agrochemicals, for example, biodynamic, organic, integrated, and industrialized farming. Biodynamic farming forbids the use of conventional agrochemicals and replaces them with exotic homemade concoctions, and organic farming *a priori* forbids conventional agrochemicals. None of these farming systems is firmly based on scientific evidence; instead they are based on a green view of nature that leads to the banning of certain chemicals.

Integrated and industrial farming can also be called `conventional', where economic, legal, and environmental constraints limit the end goal, maximum productivity, and profitability. The main difference between the latter two is that integrated is more environmentally concerned (reduced pesticide use, use of biological pest reduction methods, etc.), and industrialized is more leaning to maximum production with whatever means available, with a minimum of environmental concerns. It should be noted that `conventional' and particularly `industrialized' are somewhat derogatory terms, mainly used by those negative to these approaches.

Migration

Natural ecosystems, for example, East African savannas, can be subjected to major migrations of large herbivores that annually move long distances, following the seasonal changes in rainfall and consequential grass growth. Most natural ecosystems are less subjected to migrations, but, for example, in Northern forests at least migratory birds occur seasonally.

In agroecosystems, migration is usually kept to a minimum. Measures are taken to keep large or small grazers out from the cropped field. In some regions, wild grazers are exterminated (or close to extinction – Western European agricultural regions) and in other regions crop fields are guarded or fenced. However, migration is a component in animal husbandry; cattle is often shifted between pastures, which are given time to recover. Nomadic herding of cattle (Sami people, Masai) is similar to the savanna migrations mentioned above; the cattle and herdsmen follow the annual cycles in grazing opportunities.

Biodiversity

In a cereal monoculture, plant biodiversity is extremely low – if weed control is successful there may be only one species present, a highly specialized and genetically homogeneous wheat variety. This is not common in natural ecosystems, although it can occur in extreme environments. As mentioned above, this means that a pest can have a field day if it can reproduce in the field (or migrate into the field at a large scale).

However, agricultural monocultures still are common and continue to produce good yields. There are several reasons for this. First, there is no simple relation between biodiversity, productivity, or ecosystem stability. A plant monoculture that is well adapted, grows under good conditions, and has a reasonable resistance to pests and diseases can survive and produce well. This is exactly what a highly productive agricultural field is – a well-adapted monoculture. The crop variety has been selected for high production under a number of years with different weather (and on different soils) in a region. A variety that would demand intensive treatment with herbicides, pesticides, and fungicides will not be economical and will be rejected.

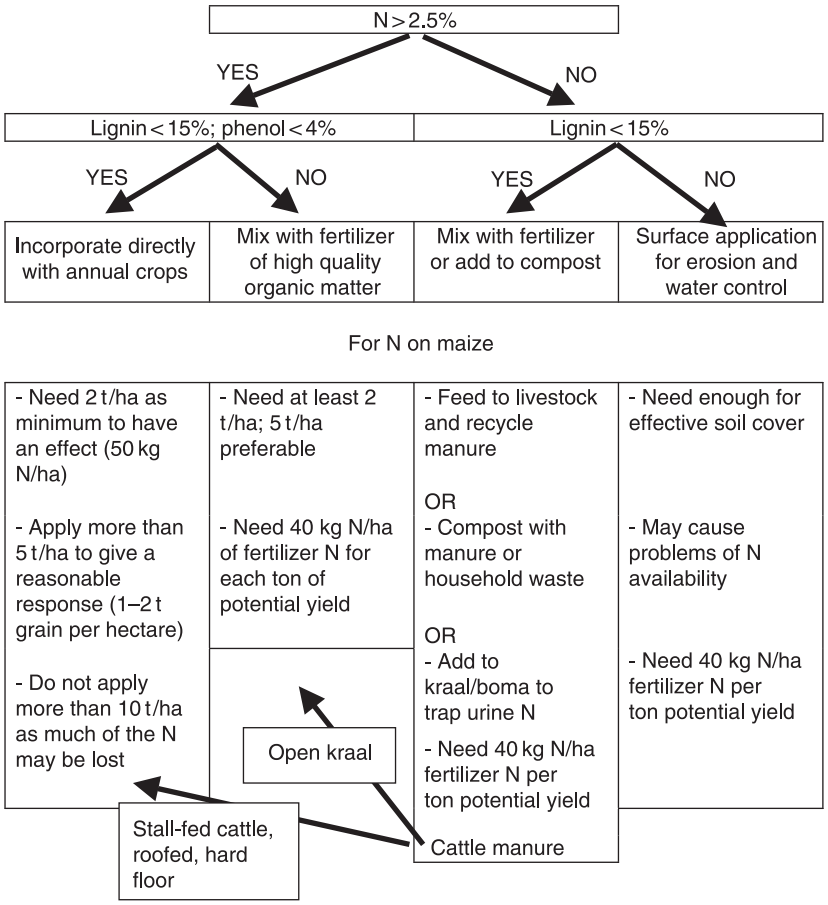


Fig. 2 Example of farmer's decisions regarding N management for a maize crop in sub-Saharan Africa, using a decision support system for organic N management depending on resource quality, expressed as N, lignin, and soluble polyphenol content. General decision matrix (top), more detailed for N economy in a maize cropping system (bottom). Modified from Vanlauwe B, Sanginga N, Giller K, and Merckx R (2004) Management of nitrogen fertilizer in maize-based systems in subhumid areas of sub-Saharan Africa. In: Mosier AR, Syers JK, and Freney JR (eds.) *Agriculture and the Nitrogen Cycle*. 124p. SCOPE 65. Washington Island Press.

Second, the low plant diversity reduces animal diversity in the stand, but perhaps less than one would expect. In a cereal monoculture stand, there can be hundreds of species of insects, mites, springtails, snails, slugs, etc. In the soil under a monoculture the biodiversity is almost always extremely high, though usually lower than in natural systems. Thousands, perhaps millions of bacterial species, tens to hundreds of species of earthworms, enchytraeids, soil insects, springtails, mites, spiders, millipedes, flagellates, amoebae, blue-green algae, etc. can be found. There are no consistent indications that soil functions such as organic matter decomposition is hampered by a low biodiversity under monocultures – a given plant residue will decompose at the same rate under a monoculture as under mixed plants, if soil temperature and moisture are the same.

Third, the last line of defense is the crop protection measures that the farmer takes. For example, in several countries there is a sophisticated monitoring and prediction system for aphid outbreaks. Aphids suck the sap from the crop leaves, but they are also vectors for crop diseases. Therefore their hibernating stages are enumerated, weather is monitored, and if the conditions are `right' the farmers are recommended to spray the fields with an insecticide (or a more specific aphicide) with dose x at date y. In less technically developed regions, experience and skill is a substitute for the model projections, but the principles are the same. It should also be mentioned that in spite of these defenses, pest insects, pathogens, and weeds still reduce worldwide crop yields considerably, and there is a great potential for improvements.

Other Ecosystem Services

The main ecosystem service from agricultural systems is simply to `feed the world'. This simple fact is easily forgotten in the richer parts of the world. However, even in Europe, which for centuries has been thoroughly under agriculture, there are other ecosystem services that are appreciated. In the forest-dominated northern Europe, agriculture actually contributes to biodiversity and landscape diversity. Without agriculture, the forest would cover all land area – the only open areas at lower altitudes would be the lakes and rivers (and the newly clear-cut forest areas, rapidly covered by shrubs). The European rural landscape in general, that is so refreshing for the city-dweller, is an agricultural product.

In other areas of the world, where the agricultural land is not sufficient to properly feed the population, other ecosystem services become relatively less important. However, if agricultural productivity can be increased, some agricultural land can be returned to savanna, forest, or other natural or seminatural states – which would be another type of service from the agroecosystem.

Since the agroecosystem is managed, and more or less sophisticated machinery and management skills are in place, it can easily be converted according to new demands from the society. If the quality requirements are met, agricultural fields can be used for recycling organic waste and ashes, and even for drawing nutrients out of sewage water. Conversion to energy crops is not too difficult (grasses, sugar beet, willow, sugarcane, etc.). Another demand from society, to sequester carbon in the soil to reduce CO_2 in the atmosphere, has recently received much attention. Increasing soil carbon content usually has beneficial effects for soil structure, water-holding capacity and general fertility, and C sequestration, perhaps even with direct payments per ton C sequestered to the farmer, is a new potential service.

The Intelligent Choices

As mentioned in the introduction, an agroecosystem has a purpose. It is designed to obtain certain goals, and the state of the system at any given point in time is a consequence of an array of intelligent choices by the farmer, complementing the border conditions set up by weather and soils, etc. The following decision matrix (**Fig. 2**) illustrates how decisions made by a maize farmer in sub-Saharan Africa can be supported by basic science knowledge. Note that the chemical analyses are not necessary for every farmer and decision. Instead, typical values for the different organic resources are estimated, and the individual farmer uses the rule of the thumb based on these estimates.

In the upper part of the **Fig. 2**, the general decision matrix is shown. Let us assume that we have leaves from a tree, which we know have a low N content and less than 15% of lignin. Then we should mix the leaves with fertilizer or add to compost. Now, in the lower part of **Fig. 2** we can see that if we look in more detail at the N economy of a maize system, we have other options – maybe add the low N material to the cattle corral (kraal/boma) to trap urine N or feed to livestock to produce higher quality organic inputs. Organic resources belonging to the third column from the left could be fed to livestock and the manure thus produced could belong to the first or the second organic resource class, depending on the management of that manure.

All over the world, farmers make these kinds of choices, based not only on biophysical knowledge and constraints, but also on economic and sociopolitical opportunities and constraints. An agroecosystem is not only controlled by farmers, but also by the society the farmer operates in. Subsidies can make growing products that have no market an intelligent choice for the farmer; lack of money can make fertilization impossible, even if it would be profitable in the long run, or real or imaginary environmental concerns from the society can force a farmer to, for example, abandon fertilizer use, cereal cropping, or pig farming.

Summing up, the agroecosystem, although limited by climatic constraints, is a product of decisions made by generations of farmers, supported by advice from agronomists and extension workers – all within a societal context of values, traditions, and legislation. In fact, the present and future agroecosystems are at least equally dependent on the societal context as on the climate and soil. However, the organisms involved are, as in any ecosystem, products of millions of years of evolution, and crop and animal breeding has only contributed with small, although important changes to the germplasm.

See also: Aquatic Ecology: Abundance Biomass Comparison Method; Eutrophication. Behavioral Ecology: Herbivore-Predator Cycles; The Marginal Value Theorem in a Nutshell; Thermoregulation in Animals: Some Fundamentals of Thermal Biology. Ecosystems: Coral Reefs. Evolutionary Ecology: Metagenomics; Natural Selection; Phylogenomics and Phylogenetics. General Ecology: Plant Ecology; Allopatry. Terrestrial and Landscape Ecology: Thermodynamic Properties of Landscape Cover

Further Reading

AndrÅn, O., Lindberg, T., Paustian, K., Rosswall, T. (Eds.), 1990. Ecological Bulletins 40: Ecology of Arable Land - Organisms, Carbon and Nitrogen Cycling. Copenhagen: Ecological Bulletins.

Brussaard, L., 1994. An appraisal of the Dutch program on soil ecology of arable farming systems (1985–1992). Agriculture, Ecosystems and Environment 51 (1–2), 1–6. and following papers.

Clements, D., Shrestha, A. (Eds.), 2004. New Dimensions in Agroecology. Binghamton: The Hawort Press, Inc, p. 553.

Eijsackers, H., Quispel, A. (Eds.), 1988. Ecological Bulletins 39: Ecological Implications of Contemporary Agriculture. Copenhagen: Ecological Bulletins.

Kirchmann, H., 1994. Biological dynamic farming – an occult form of alternative agriculture? Journal of Agricultural and Environmental Ethics 7, 173–187.

Mosier, A.R., Syers, J.K., Freney, J.R. (Eds.), 2004. Agriculture and the Nitrogen Cycle. SCOPE 65 Washington: Island Press, p. 124.

New, T.R., 2005. Invertebrate Conservation and Agricultural Ecosystems. Cambridge: Cambridge University Press, 354pp.

Newman, E.I., 2000. Applied Ecology and Environmental Management, 2nd edn. Blackwell Science.

Vanlauwe, B., Sanginga, N., Giller, K., Merckx, R., 2004. Management of nitrogen fertilizer in maize-based systems in subhumid areas of sub-Saharan Africa. In: Mosier, A.R., Syers, J.K., Freney, J.R. (Eds.), Agriculture and the Nitrogen Cycle. Washington Island Press, p. 124. SCOPE 65.

Woomer, P.L., Swift, M.J., 1994. The Biological Management of Tropical Soil Fertility. Chichester: Wiley.

Relevant Websites

http://www.cgiar.org— Consultancy Group on International Agricultural Research.
http://www.fao.org— Food and Agriculture Organization of the United Nations.

Alpine Ecosystems and the High-Elevation Treeline☆

C Körner, Botanisches Institut der Universität Basel, Basel, Switzerland

Definitions and Boundaries

Ecosystems above the upper climatic limit of trees are termed 'alpine.' Scientifically, the alpine life zone is an altitudinal belt defined by climatic boundaries (Figure 1), and the term 'alpine' does not refer to the European Alps but to naturally treeless high-elevation biota worldwide (mostly grassland and shrubland). 'Alpine' supposedly roots in the pre-Indo-European word *alpo* for steep slopes, still used today in the Basque language. By contrast, in common language, 'alpine' is often used for places anywhere in mountainous terrain, irrespective of elevation (e.g., alpine village, even alpine cities). If a city were truly alpine it would have to be above the climatic treeline, but no such city does exist worldwide. Hence, a distinction must be made between the scientific, biogeographic meaning of alpine (the issue of this article) and common (often touristic) jargon.

The upper limit of the alpine life zone or alpine belt is reached where flowering plants have their high elevation limit. This is often close to the snow line (the elevation at which snow can persist year-round), but commonly a few scattered flowering plants also grow above the snow line, in favorable, equator-facing, and sheltered places. The uppermost part of the alpine belt, where closed ground cover by vegetation is missing, is often termed 'nival,' referring to sparse vegetation in rock and scree fields. The highest place on Earth where flowering plants have been found is in the Central Himalayas at 6200–6350 m above sea level. The high elevation record for flowering plants in Europe is 4500 m in the Swiss Alps.

Depending on latitude, the climatic treeline and hence the lower limit of the alpine belt can be anywhere between close to sea level in subpolar regions (>70° N, >55° S) and close to 5000 m in subtropical continental climates (trees >3 m at ~4800 m in Bolivia and Tibet). In the cool temperate zone (45–50° N), the alpine belt may start anywhere between 1200 and 3500 m (in the European Alps at 2000 m, the Colorado Rocky Mountains at 3400 m); that is, it is lower under strong oceanic influence and higher in the inner parts of continents. The common natural treeline elevation near the equator is 3600–4000 m. The altitudinal width of the alpine belt above treeline is roughly 1000 m. It covers ~2.6% of the globe's terrestrial area, if Antarctica is disregarded.

Given this convention on the two boundaries of the alpine belt, it is important to note that these boundaries are not sharp lines but are centered across gradients which change from place to place and depend on topography and region. Usually, these boundaries are obvious from the great distance (an airplane), but hard to depict on the ground, hence depend on scale.

The Alpine Treeline

Since, by definition, the alpine belt is naturally treeless, the mechanisms by which trees are restricted from growing beyond a certain elevation are key to any understanding of alpine ecosystems. The so-called treeline marks the upper limit of the life-form 'tree' irrespective of the tree species involved. Generally, species which form treelines are *Pinus*, *Picea*, *Abies*, *Juniperus*, and *Larix* among conifers, and *Betula*, *Alnus*, *Erica*, *Polylepis*, *Sorbus*, *Eucalyptus*, and others among non-coniferous families. Because tree occurrence does not stop abruptly, and trees gradually get smaller and finally become crippled, any definition of 'a line' is a convention. The forest line or timberline represents the edge of the closed upper montane forest (note, 'montane' is the biogeographic term for the next lower belt, not to be confused with 'mountain'), the zone of gradual forest opening near the treeline is often termed treeline parkland, and the uppermost position where tree species can survive as small saplings or shrubs among other low-stature vegetation is called the tree species line, with the 'treeline' holding a middle ground, used for the line connecting the uppermost patches of trees >3 m. The whole transition zone from montane forest to alpine heathland is termed treeline ecotone, across which alpine vegetation gains space yielded by the thinning forest. The altitudinal range of the treeline ecotone may be 20–200 m, often <50 m.

Where moisture is permitting tree growth at these altitudes (a minimum of 250–300 mm of precipitation per year), the position of the natural climatic treeline matches with a mean growing season temperature of 6.4 ± 0.7 °C worldwide. The duration of the growing season may vary from 10 weeks at high latitude to a full year in the tropics and its onset and end are defined by a weekly mean air temperature of 0 °C (corresponding to ~3 °C in 10 cm soil depth, where most roots occur). This isotherm sets the

☆*Change History*: February 2013. C Körner updated the 'Definitions and Boundaries' section and references.

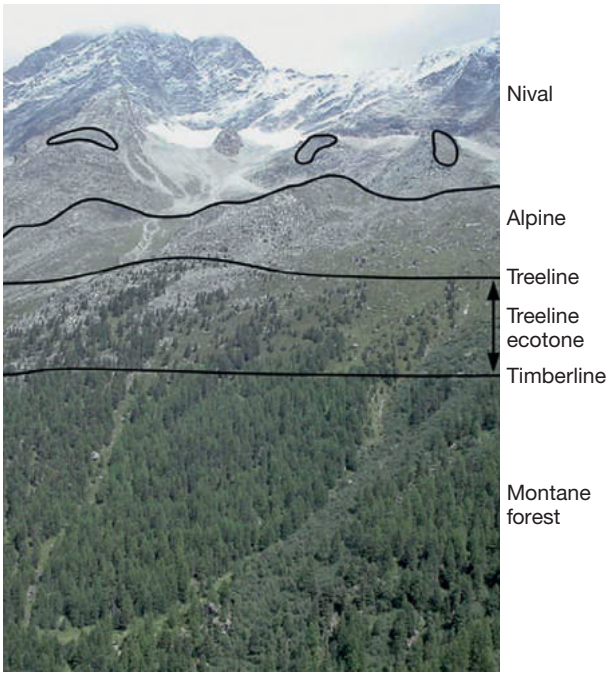


Figure 1 The elevational belts of mountain ecosystems. With increasing altitude, these belts become fragmented and topography (exposure) plays an increasing role. (Example from the Swiss Central Alps with *Pinus cembra* forming the treeline at 2350 m.)

Figure 2 Fire and grazing (both naturally and under human influence) can replace the montane forest, leading to 'alpine-looking' vegetation below the climatic treeline. Here is an example of the Ecuadorian páramos at 3600 m altitude, ~400–500 m below the potential climatic treeline (Páramos El Angel). Giant rosettes of the genus *Espeletia* are the prominent feature of this landscape, with similar vegetation also found in African highlands.

low-elevation climatic threshold for the alpine vegetation, which can be close to 5 °C in dry subtropical mountains and 7.5 °C in cool temperate mountains, a surprisingly narrow range, given the great difference in season length across latitudes.

It is very important not to confuse this climatic (physiological) limit of trees with a multitude of other natural or anthropogenic causes for the local absence of trees such as fire, avalanches, logging and pasturing, loose or missing substrate, waterlogging, or the regional lack of cold-adapted tree species (as is the case, for instance, in Hawaii). In the last case, the treeline observed is a specific tree species line, not representative of the climatic limit of the life-form tree, as can easily be demonstrated by the success of introduced tree species which grow well at much higher elevations in such regions. Open 'alpine-looking' grassland and shrubland may thus occur several hundreds of meters below the climatic treeline; among the most famous of these are the Andean Páramo grasslands with their spectacular giant rosette plants (Figure 2).

Alpine Plants Engineer Their Climatic Environment

Why is there lush alpine vegetation but trees cannot grow? Are alpine plants physiologically superior, able to cope with those low temperatures which otherwise are harming trees? There is good evidence that thermal constraints for growth, that is, building

new tissue, are the same for alpine plants, cold-adapted trees, and winter crops (winter rape and winter wheat), all being halted when tissue temperatures drop below 4–5 °C, and growth is close to zero at 6–7 °C. In contrast, all these species reach 30–50% of maximum rates of photosynthesis at these same temperatures; thus the provision of raw material for growth (sugar) cannot be decisive. Neither are there critical differences in freezing resistance between alpine plants compared to trees. Hence, at tissue level, there is no physiological reason why alpine grasses, herbs, and shrubs should grow at a given low temperature and trees should not.

There are two reasons for success of alpine plant above treeline:

1. By low stature and dense stand structure, alpine plants restrict aerodynamic exchange with the atmosphere, which causes heat to accumulate during periods with solar radiation and permit plants to operate at comparatively warm temperatures, much unlike those experienced by upright, ventilated trees. The life-form 'tree' does not permit any escape from the gradually declining ambient temperatures, whereas alpine plants engineer their microclimate and air-condition their meristems close to the ground so that they can build new tissue at otherwise cold air temperatures above the plant canopy (Figure 3).
2. By developmental flexibility and morphological adaptation, alpine plants are able to make use of short favorable weather conditions, they sprout rapidly, produce only a few, mostly short-lived leaves (~60 days), and have their meristems positioned very close to the ground, in the case of many grasses, sedges, or rosette plants, often 1–2 cm below ground, where the solar-heated soil provides a thermally buffered environment. In contrast, trees operate at longer leaf duration (mostly >120 days; in evergreen treeline conifers, 4–12 years) and leaves take longer to mature, and their aboveground meristems are fully exposed to the cold air temperatures.

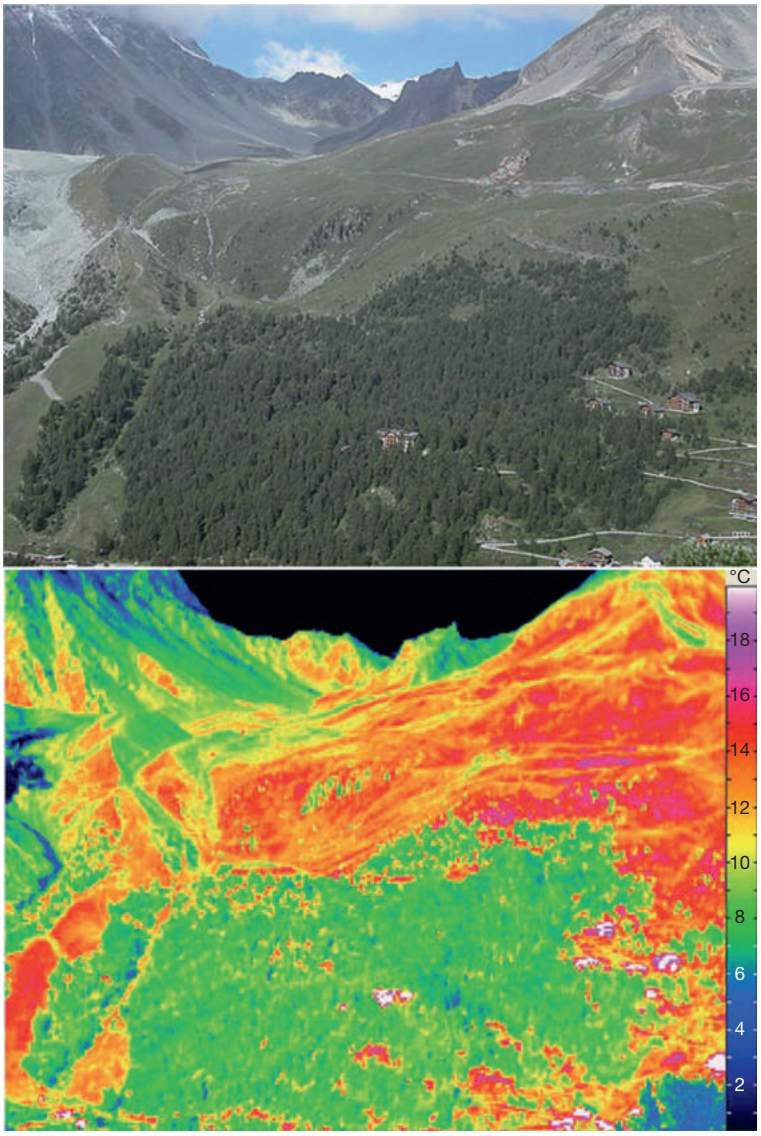


Figure 3 Trees are coupled to air temperature and thus appear 'cool' on this infrared thermograph taken at 10 a.m. on a bright midsummer morning in the Swiss Alps near Arolla. Alpine grassland and shrub heath accumulate heat by decoupling from atmospheric conditions (low stature, dense structures). So the treeline can clearly be depicted as a thermal boundary driven by plant architecture.

The transition from trees to alpine vegetation is thus dictated by plant architecture and not by tissue-specific inferiority of trees compared to alpine plants. This close coupling of trees to atmospheric conditions also explains the surprisingly uniform leveling of treelines across mountain valleys which reminds one of the level of a water reservoir. In contrast, the climate in alpine vegetation varies with compactness and height of the leaf canopy and exposure to the sun. A sun-exposed, sheltered microhabitat at 3000 m of altitude may be warmer than a shaded microhabitat at 1800 m. Altitude per se, or data from a conventional climate station, thus, tells us little about the climate actually experienced by alpine plants. It had long been known that mutual sheltering among alpine plants or leaves/tillers within a plant is very beneficial ('facilitation'), and removing this shelter effect by opening the plant canopy can be disastrous.

Alpine plants are small by 'design' (genetic dwarfs); they are not forced into small stature by the alpine climate directly, though evolution had selected such morphotypes. What seems like a stressful environment is not really stressful for those well adapted. However, there is some additional modulative, direct effect on size by low temperature. Alpine plants that survive in low-altitude rock gardens indeed grow taller than their relatives in the wild. But plants grown in such rock gardens are commonly of montane origin, because most typical alpine plants fade at such high-, low-altitude temperatures, possibly because of overshooting mitochondrial respiration.

Alpine Ecosystem Processes

Almost everything gets slower when it gets cold, but slow production of biomass and slow recycling of dead biomass (litter) go hand in hand, so that the carbon and nutrient cycles remain in balance. Recycling of organic debris is responsible for most of the steady-state nutrient provision and thus controls vigor of growth. When mineral nutrients are added, all alpine vegetation tested had shown immediate growth stimulation, but this holds for most of the world's biota and is not specific to alpine ecosystems. On the other hand, nutrient addition had been shown to make alpine plants more susceptible to stress (softer tissue, reduced winter dormancy) and pathogen impact (e.g., fungal infections) and causes nitrophilous grasses and herbs to overgrow the best-adapted slow-growing alpine specialist species.

It comes as a rather surprising observation that alpine plant productivity – at least in the temperate zone – is only low when expressed as an annual rate of biomass accumulation, but is not low at all when expressed per unit of growing season length. In a 2-month alpine season in the temperate zone alpine belt, the biomass production (above-plus belowground) accumulates to $\sim$400 g dry matter m^{-2} (range 200–600 g m^{-2}). A northern deciduous hardwood forest produces 1200 g m^{-2} in 6 months and a humid tropical forest 2400 g m^{-2} in a 12-month season, all arriving at $\sim$200 g m^{-2} per month. Time constraints of growth are thus the major causes of reduced annual production in closed alpine grass- and shrubland and not physiological limitations in what seems to a human hiker like a rather hostile environment. Acclimation to low temperature, perfect plant architecture, and developmental adjustments can equilibrate these constraints on a unit of time basis. It makes little sense to relate productivity to a 12-month period when 9–10 months show no plant activity because of freezing conditions and/or snow cover.

Similar to carbon and nutrient relations, alpine ecosystem's water relations are largely controlled by seasonality. During the growing season in the humid temperate zone, daily water consumption during bright weather hardly differs across altitude ($\sim$3.5–4 mm evapotranspiration). However, because of the short snow-free season at such latitudes, annual evapotranspiration may be only 250–300 mm compared to 600–700 mm at low altitude; hence, runoff is much higher in alpine altitudes. Given that precipitation often increases with altitude in the temperate zone (a doubling across 2000 m of altitude is not uncommon), annual runoff may be three to five times higher in the alpine belt, with major implications for erosion in steep slopes.

In many tropical and subtropical mountains, moisture availability drops rapidly above the condensation cloud layer at 2000–3000 m altitude, causing the alpine belt to receive very little water, often not more than 200–400 mm per year (e.g., the high Andes, Tenerife, East African volcanoes). The resulting sparse vegetation is often termed alpine semidesert, but because of wide spacing of plants and very little ground cover, those plants which are found in this semiarid alpine landscape were found to be surprisingly well supplied with water even at the end of the dry season (Figure 4). As a rule of thumb, alpine plants are thus better supplied with moisture (even in dry alpine climates) than comparable low-altitude vegetation. True physiologically effective water stress is quite rare in the alpine belt, but moisture shortage in the top soil may restrict nutrient availability periodically, which restricts growth.

Biodiversity in Alpine Ecosystems

For plants and animals to become 'alpine,' they must pass through a selective filter represented by the harsh climatic conditions above treeline. It comes as another surprise that alpine ecosystems are very rich in organismic taxa. It was estimated that the $\sim$2.6% of global land area (outside Antarctica) that can be ascribed to the alpine belt hosts $\sim$4% of all species of flowering plants. In other words, alpine ecosystems are on average similarly rich or even richer in plant species than average low-elevation ecosystems. This is even more surprising if one accounts for the fact that the available land area above treeline shrinks rapidly with elevation (on average, a halving of the area in each successive 170 m belt of altitude). A common explanation for this high species richness is the archipelago nature of high mountains (a fragmentation into climatic 'islands'), the high habitat diversity as it results from gravitational forces (topographic diversity, also termed geodiversity), and the small size of alpine plants, which partly compensates

Figure 4 High-altitude semideserts (near Sajama, Bolivia, 4200 m) are often dominated by sparse tussock grasses, shrubs, and minor herbs in the intertussock space, all together preventing soil erosion, while being used for grazing. The wide spacing mitigates drought stress in an otherwise dry environment.

Figure 5 The four major life-forms of flowering plants in alpine ecosystems: cushion plants (*Azorella compacta*, *Silene exscapa*), herbs (small: *Chrysanthemum alpinum*, tall: *Gentiana punctata*), dwarf shrubs (*Loiseleuria procumbens*, *Salix herbacea*), and tussock-forming graminoids (*Carex curvula*, diverse tall grass tussock).

for the elevational loss of land area. The elevational trends for animal diversity are similar to plants, but some animal taxa decline in diversity with elevation more rapidly (e.g., beetles, earthworms, butterflies) than others (e.g., vertebrates, birds). Often animal diversity peaks at mid-altitudes (close to the treeline ecotone) and then declines.

The four major life-forms of flowering plants in the alpine belt are graminoids (grasses, mostly forming tussocks, sedges, etc.), rosette-forming herbs, dwarf shrubs, and cushion plants (Figure 5). In most parts of the world, bryophytes and lichens (a symbiosis between algae and fungi) contribute an increasing fraction of biodiversity as altitude increases. Each of these life-forms can be subdivided into several subcategories, mostly represented by different forms of clonal growth. Clonal (vegetative) spreading is dominant in all mountains of the world and it secures long-term space occupancy by a 'genet' (a single genetical individual) in a rather unpredictable environment. Because of the topography-driven habitat diversity, rather contrasting morphotypes and physiotypes may be found in close proximity, as for instance succulent (water storing) plants such as alpine cactus or some leaf-succulent Crassulaceae (*Sedum* sp., *Echeveria* sp.) next to wetland or snowbed plants.

Alpine ecosystems are known for their colorful flowers, and it was often thought that this may be a selected-for trait, because it facilitates pollinator visitation. There is also morphological evidence that alpine plants invest relatively more in flowering, given that plant size (and biomass per individual) declines by nearly tenfold from the lowland to the alpine belt, whereas the size of insect-pollinated flowers hardly changes. Furthermore, flower duration increases and so does pollinator visiting duration, and there is no indication that there is a shortage in alpine pollinators. The net outcome is a surprisingly high genetic diversity in what seems like highly fragmented and isolated habitats. Despite the successful reproductive system at the flower-pollinator scale and well-adapted (fast) seed maturation, the real bottleneck is seedling establishment (the risk to survive the first summer and winter), which explains why most alpine plants also propagate clonally.

Overall, mountain biodiversity (the montane belt, the treeline ecotone, and the alpine belt) is a small-scale analog of global biodiversity, because of the compression of large climatic gradients over very short distances. Across a vertical gradient from 1200 to 4200 m in the Tropics one may find a flora and fauna with a preference for climates otherwise only found across several thousand kilometers of latitudinal distance. This is why mountains are ideal places for biodiversity conservation as long as the protected mountain system is large and has migration corridors to prevent biota from becoming trapped in ever-narrowing land area should climatic warming induce altitudinal upward shifts of life zones.

Alpine Ecosystems and Global Change

'Global change' includes changes in atmospheric chemistry (CO_2, CH_4, N_xO_y), the climatic consequences of these changes, and the manifold direct influences of humans on landscapes. All three global change complexes affect alpine biota, either directly or indirectly.

Elevated atmospheric carbon dioxide (CO_2) concentrations affect plant photosynthesis directly, although plants in alpine grassland and glacier fore fields in the Alps were found to be carbon saturated at current ambient CO_2 concentrations (390 ppm). The effect of doubling CO_2 concentrations over four consecutive seasons on net productivity of alpine health was zero. However, not all species responded identically; hence, there is a possibility of gradual shifts in species composition in the long run, with some species getting suppressed and others gaining.

In contrast, even moderate additions of soluble nitrogen fertilizer at rates of those received today by mountain forelands in Central Europe with rains (40–50 kg N $ha^{-1}a^{-1}$) doubled biomass in only 2 years. Even 25 kg $ha^{-1}a^{-1}$ had immediate effects on biomass (+27%), again favoring some species more than others. Atmospheric nitrogen deposition is thus far more important for alpine ecosystems than elevated CO_2. Just for comparison, in agriculture, cereals are fertilized with >100 kg N $ha^{-1}a^{-1}$.

Consequences of climatic change for alpine ecosystems are hard to predict because of the interplay of climatic warming with precipitation. A warmer atmosphere can carry more moisture; hence, increasing precipitation had been predicted for temperate mountain areas. Greater snowpack can shorten the growing season at otherwise higher temperatures. While the temperate zone has seen more late winter snow in recent years, the uppermost reaches of higher plants seem to have profited from climatic warming over the twentieth century. Several authors documented a clear enrichment of summit floras, accelerated in recent decades.

Treeline trees respond to warmer climates by faster growth, but whether and how fast this would cause the treelines of the world to advance upward depends on tree establishment, which is a slow process. Hence, treelines always lagged behind climatic warming during the Holocene by centuries, as evidenced by pollen records. Current trends are largely showing an infilling of gaps in the treeline ecotone, but upward trends await larger-scale confirmation. Eventually any persistent warming will induce upward migration of all biota. By contrast, recent climatic warming has caused the tropical upper montane/alpine climate on Kilimanjaro to become drier, facilitating devastating fires, which depressed the montane forest by several hundred meters with a downslope advance (expansion) of alpine vegetation following.

Land use is still the most important factor for changes in alpine ecosystems. Around the globe, alpine vegetation is used for herding or uncontrolled grazing by lifestock. Much of the treeline ecotone has been converted into pasture land, with both overutilization and erosion (mainly in developing countries) and abandonment of many centuries-old, high-elevation cultural landscapes (mainly industrialized countries) causing problems. The question is not whether there should be pasturing, but how it should be done. Sustainable grazing requires shepherding and observation of traditional practices, which largely prevent soil damage and erosion. Traditional alpine land use has a several thousand years' history and was optimized for maintaining an intact landscape for future generations as opposed to land-hungry newcomers faced with the need of feeding a family today, rather than thinking of sustained livelihood in a given area. All other forms of land use (except mining), as dramatic their negative effects at certain places may look, are less important, because their impact is rather local (e.g., tourism, road projects). Agriculture is by far the most significant factor in terms of affected land area.

Mismanagement of alpine ecosystems has severe consequences (e.g., soil destruction, sediment loading of rivers) not only for the local population but also for people living in large mountain forelands, which depend on the steady supplies of clean water from high-altitude catchments. Almost 50% of mankind consumes mountain resources, largely water and hydroelectric energy; hence, there is an often overlooked teleconnection between alpine ecosystems and highly populated lowlands. Highland poverty is thus affecting the conditions and the economic value of catchments, which goes far beyond the actual agricultural benefits. This insight should lead to better linkages between lowland and highland communities and also include economic benefit sharing with those that perform sustainable land care in alpine ecosystem.

Further Reading

Barthlott W, Lauer W, and Placke A (1996) Global distribution of species diversity in vascular plants: Towards a world map of phytodiversity. *Erdkunde* 50: 317–327.
Billings WD (1988) Alpine vegetation. In: Barbour MG and Billings WD (eds.) *North American terrestrial vegetation*, pp. 392–420. Cambridge: Cambridge University Press.
Bowman WD and Seastedt TR (eds.) (2001) *Structure and function of an alpine ecosystem – Niwot Ridge, Colorado*. Oxford: Oxford University Press.
Chapin FS III and Körner C (eds.) (1995) *Arctic and alpine biodiversity: Patterns, causes and ecosystem consequences*. *Ecological studies* 113. Berlin: Springer.
de Witte L and Stöcklin J (2010) Longevity of clonal plants: Why it matters and how to measure it. *Annals of Botany* 106: 859–870.

de Witte LC, Armbruster GFJ, Gielly L, Taberlet P, and Stöcklin J (2012) AFLP markers reveal high clonal diversity, repeated recruitment and extreme longevity of four arctic-alpine key species. *Molecular Ecology* 21: 1081–1097.
Fabbro T and Körner C (2004) Altitudinal differences in flower traits and reproductive allocation. *Flora* 199: 70–81.
Nagy L, Grabherr G, Körner C, and Thompson DBA (eds.) (2003) *Alpine biodiversity in Europe*. Berlin: Springer.
Hemp A (2005) Climate change-driven forest fires marginalize the impact of ice cap wasting on Kilimanjaro. *Global Change Biology* 11: 1013–1023.
Kalin Arroyo MT, Primack R, and Armesto J (1982) Community studies in pollination ecology in the high temperate Andes of central Chile. Part I: Pollination mechanisms and altitudinal variation. *American Journal of Botany* 69: 82–97.
Körner C (2003) *Alpine plant life*, 2nd ed. Berlin: Springer.
Körner C (2004) Mountain biodiversity, its causes and function. *Ambio* 13: 11–17.
Körner C (2006) Significance of temperature in plant life. In: Morison JIL and Morecroft MD (eds.) *Plant growth and climate change*, pp. 48–69. Oxford: Blackwell.
Körner C (2011) Coldest places on earth with angiosperm plant life. *Alpine Botany* 121: 11–22.
Körner C, Paulsen J, and Spehn EM (2011) A definition of mountains and their bioclimatic belts for global comparisons of biodiversity data. *Alpine Botany* 121: 73–78.
Körner C (2012) *Alpine treelines*. Basel: Springer.
Mark AF, Dickinson KJM, and Hofstede RGM (2000) Alpine vegetation, plant distribution, life forms, and environments in a perhumid New Zealand region: Oceanic and tropical high mountain affinities. *Arctic, Antarctic, and Alpine Research* 32: 240–254.
Messerli B and Ives JD (eds.) (1997) *Mountains of the world: A global priority*. New York: Parthenon.
Meyer E and Thaler K (1995) Animal diversity at high altitudes in the Austrian Central Alps. In: Chapin FS III and Körner C (eds.) *Arctic and alpine biodiversity: Patterns, causes and ecosystem consequences*. *Ecological studies* 113. pp. 97–108. Berlin: Springer.
Miehe G (1989) Vegetation patterns on Mount Everest as influenced by monsoon and föhn. *Vegetatio* 79: 21–32.
Nagy L, Grabherr G, Körner C, and Thompson DBA (2003) *Alpine biodiversity in Europe*. *Ecological studies* 167. Berlin: Springer.
Pauli H, et al. (2012) Recent plant diversity changes on Europe's mountain summits. *Science* 336: 353–355.
Rahbek C (1995) The elevational gradient of species richness: A uniform pattern? *Ecography* 18: 200–205.
Sakai A and Larcher W (1987) *Frost survival of plants. Responses and adaptation to freezing stress*. *Ecological studies* 62. Berlin: Springer.
Spehn EM, Liberman M, and Körner C (2006) *Land use change and mountain biodiversity*. Boca Raton, FL: CRC Press.
Wagner J, Ladinig U, Steinacher G, and Larl I (2012) From the flower bud to the mature seed: Timing and dynamics of flower and seed development in high-mountain plants. In: Lütz C (ed.) *Plants in alpine regions: Cell physiology of adaptation and survival strategies*. Vienna, New York: Springer.
Yoshida T (2006) *Geobotany of the Himalaya*. Tokyo: The Society of Himalayan Botany.

Caves☆

FG Howarth, Bishop Museum, Honolulu, HI, USA

Caves

Caves are defined as natural subterranean voids that are large enough for humans to enter. They occur in many substrata, and cavernous landforms make up a significant portion of the Earth's surface. Limestone caves are the best known. Limestone, calcium carbonate, is mechanically strong yet dissolves in weakly acidic water. Thus over eons great caves can form. Caves form in other soluble rocks, such as dolomite (calcium magnesium carbonate), but they are usually not as extensive as those in limestone. Volcanic eruptions also create caves. The most common are lava tubes that are built by the roofing over and subsequent draining of molten streams of fluid basaltic lava. In addition, cave-like voids form by erosion (e.g., sea caves and talus caves), by earthquakes and by melting water beneath or within glaciers. Depending on their size, shape and interconnectedness, caves develop unique environments that often support distinct ecosystems.

Cave Environments

The physical environment is rigidly constrained by the geological and environmental settings and can be defined with great precision because it is surrounded and buffered by thick layers of rock. Caves can be water-filled or aerial.

Aquatic Environments

Aquatic systems are best developed in limestone caves since water creates these caves. Debris-laden water in voids in insoluble rock will eventually fill caves. A significant exception is found in young basaltic lava that has flowed into the sea. Here, subterranean ecosystems develop in the zone of mixing freshwater and salt water within caves and spaces in the lava. This ecosystem (called anchialine) is fed by food carried by tides and groundwater flow. Frequent volcanism creates new habitat before the older voids fill or erode away. Anchialine systems also occur in limestone and other cavernous substrata in coastal regions. Aquatic cave environments are dark, three-dimensional (3D) mazes, in which food and mates may be difficult to find. In addition, the water can stagnate, locally becoming hypoxic with high concentrations of toxic gases including carbon dioxide and hydrogen sulfide.

Terrestrial Environments

The terrestrial environment in long caves is buffered from climatic events occurring outside. The temperature stays nearly constant, usually near the mean annual surface temperature (MAST); except passages sloping down from an entrance tend to trap cold air and remain a few degrees cooler than MAST. Passages sloping up are often warmer than MAST. The environment is strongly zonal (Fig. 1). Three zones are obvious: an entrance zone where the surface and underground habitats overlap; a twilight zone between the limit of photosynthesis and total darkness; and the dark zone. The dark zone can be further subdivided into three distinct zones: a transition zone where climatic events on the surface still affect the atmosphere, especially relative humidity (RH); a deep zone where the RH remains constant at 100%; and an innermost stagnant air zone where air exchange is too slow to flush the buildup of carbon dioxide and other decomposition gasses. The boundary between each zone is often determined by shape or constrictions in the passage. In many caves, the boundaries are dynamic and change with the seasons.

The subterranean aerial environment is stressful for most organisms. It is a perpetually dark, 3D maze with a water-saturated atmosphere and occasional episodes of toxic gas concentrations. Many of the cues used by surface animals are absent or operate abnormally in caves (e.g., light/dark cycles, wind, scent plumes and sound). Passages can flood during rains, and crevices might drop into pools and water-filled traps. If the habitat is so inhospitable, why and how do surface animals forsake the lighted world and adapt to live there? It is the presence of abundant food resources that provides the impetus for colonization and adaptation.

Food Resources

The main energy source in limestone caves is sinking rivers, which carry-in abundant food not only for aquatic communities but also via flood deposits for terrestrial communities. Rivers are less important in insoluble rock, such as lava, but percolating runoff washes surface debris into caves through crevices. Other major energy sources are brought in by animals that habitually visit or

☆*Change History:* August 2013. FG Howarth updated the abstract, body of the article and further reading sections and added a new Level One section on 'Conservation,' which includes an updated 'Invasive species' section and a new 'Climate change' section.

 doi:10.1016/B978-0-12-409548-9.00725-9

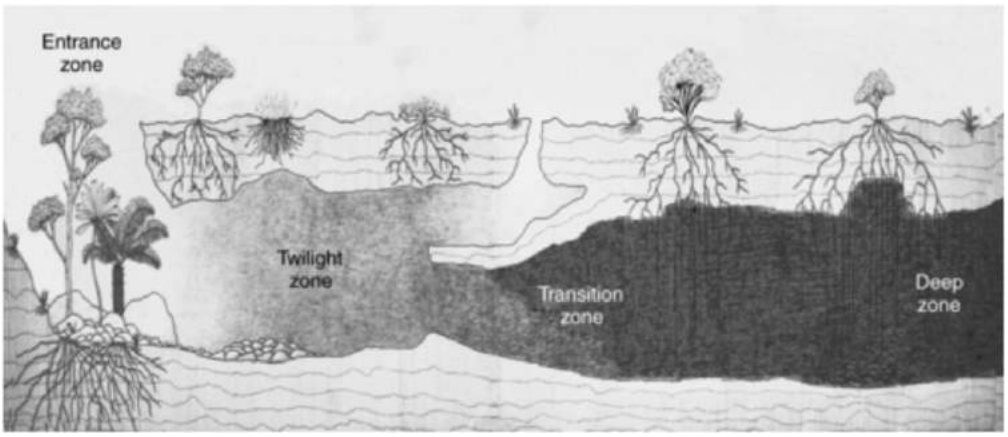


Fig. 1 Schematic profile view of the cave habitat showing the location of principal zones.

roost in caves, plants that are able to send their roots deep underground, chemoautotrophic microorganisms that use minerals (including fossil carbon) and surface animals (i.e., accidentals) that fall or wander into caves and become lost.

Generally in surface habitats, accumulating soil filters water and nutrients and holds these resources near the surface where they are accessible to plant roots and surface-inhabiting organisms. However, in most areas with underlying caves, the soil is thin with areas of exposed bare rock because developing soil is washed or carried into underground voids by water or gravity. Soil formation is limited, and much of the organic matter sinks out of the reach of most surface animals and shallow-rooted plants.

Except for guano deposits, flood deposits, scattered root patches, and other point-source food inputs, the defining feature of cave habitats is the appearance of barren wet rock. Visible food resources in the deep cave are often negligible, and what food deposits there are would be difficult for animals to find in the 3D maze. Food resources in the system of smaller spaces is difficult to sample and quantify, but in theory, some foods may be locally concentrated by water transport, plant roots, or micro point source inputs such as through cracks indirectly or circuitously connecting to the surface. These deposits would be more easily exploited than would widely scattered deposits.

In each biogeographic region, a few members of the surface and soil fauna have invaded cave habitats and adapted to exploit this deep food resource. The colonists usually were pre-adapted; that is, they already possessed useful characteristics resulting from living in damp, dark habitats on the surface.

Cave Communities

Guano Communities

Many animals live in or use caves. Cave-inhabiting vertebrates are relatively well-known. Cave bats, swiftlets (including the edible-nest swiftlet of Southeast Asia) and the oil bird in South America use echolocation to find their way in darkness. Pack rats in North America, along with cave crickets and other arthropods also roost in caves. Large colonies of these cave-nesting animals carry in huge quantities of organic matter with their guano and dead bodies. This rich food resource forms the basis for specialized communities of microorganisms, scavengers, and predators. Arthropods comprise the dominant group of larger animals in this community, and like their vertebrate associates, most species are able to disperse outside caves to establish new colonies.

Deep Cave Communities

In the deeper netherworld, communities of mysterious, obligate cave animals occur. Most are invertebrates, but a few fishes and salamanders have colonized the aquatic realm. Crustaceans (shrimps and their allies) dominate in aquatic ecosystems, and insects, myriapods and spiders and their allies dominate terrestrial systems. Although a few species are specialists on living plant roots or other specific resources, most are generalist predators or scavengers. The relatively high percentage of predators indicates the importance of accidentals as a food resource. However, many presumed predatory species, such as spiders, centipedes, and ground beetles, will also scavenge on dead animals when available. It is not advantageous to have finicky tastes where food is difficult to find. Thus in deep cave communities, the food chain more closely resembles a food web with most species interacting with most of the other species in the community, rather than the energy progressing more or less linearly from plants through plant feeders, scavengers, and omnivores to predators as normally happens in most surface systems.

Adaptations to Cave Life

Animals roosting or living in caves must adapt to cope with the unusual environment. Paramount for the cave-roosting vertebrates is the ability to find their way to and from their roosts at the correct time. Not surprisingly, the birds and bats display uncanny skill in memorizing the complex maze to and from their cave roosts. Pack rats use trails of their urine to navigate in and out of caves. Species using the twilight and transition zones can use the daily meteorological cycle for cues to wake and leave the cave. Those roosting in the deep zone may rely on accurate internal clocks to know when it is beneficial to leave their roost.

Organisms that adapt to live permanently underground must make changes in behavior, physiology, and structure in order to thrive in the stressful environment. They need to find food and mates and successfully reproduce in total darkness. Their hallmark is the loss or reduction of conspicuous structures such as eyes, bodily color, protective armor, and wings. These structures are worthless in total darkness, but they can be lost quickly when selection is relaxed because they are expensive for the body to make and maintain. How such losses could happen quickly is demonstrated by the cave-adapted planthoppers (Cixiidae). The nymphs of surface species feed on plant roots in soil and have reduced eyes and bodily color whereas their adults have big eyes, bold colors, and functional wings. The cave-adapted descendants maintain the nymphal eyes, color, and other structures into adulthood, a phenomenon known as neoteny.

The high relative humidity and occasional episodes of elevated CO_2 concentrations are stressful to most organisms. The blood of insects and other invertebrates will absorb water from saturated atmosphere, and the animals literally will drown unless they have adaptations to excrete the excess water. High levels of CO_2 force animals to breathe more, which increases water absorption. Cave-adapted insects often have modified spiracles to prevent or cope with their air passages filling with water.

Most lava tube arthropods have specialized elongated claws to walk on glassy wet-rock surfaces. Many have elongated legs to step across cracks rather than having to descend and climb the other side. Jumping or falling might land a hapless animal in a pool or water-filled pit or into the clutches of a predator. Small insects are often too heavy or are unable to climb the meniscus at the edge of rock pools and will eventually drown. However, many cave-adapted insects have unique knobs or hairs near the base of each elongated claw and modified behavioral traits that allow them to climb the meniscus and escape. Some of the latter are predators or scavengers, who wait on pools for victims.

Other Cave-Like Habitats

Cavernous rock strata contain abundant additional voids of varying sizes, which may not be passable by humans. These voids are interconnected by a vast system of cracks and solution channels. The smaller capillary-sized spaces are less important biologically because their small size limits the amount of food resources they can hold and transport. Voids larger than about 5 cm can transport large volumes of food as well as serve as habitat for animals. In terms of surface area and extent, these intermediate-size voids are the principal habitat for specialized cave animals. Many aspects of their life history may occur only in these spaces. Some cave species (such as the earwig, *Anisolabis howarthi* (**Fig. 2**), and sheet web spiders, Linyphiidae, in Hawaiian lava tubes) prefer to live in crevices and are only rarely found in caves. In addition, cave-adapted animals have been found living far from caves in cobble deposits beneath rivers, fractured rock strata, talus slopes and buried lava clinker in Japan, Hawai'i, Canary Islands, Australia, North America and Europe. New cave species and ecosystems continue to be discovered in new regions. These discoveries corroborate the view that cave adaptation and the development of cave ecosystems can occur wherever there is suitable underground habitat.

Because these smaller voids are isolated from airflow from the surface, the environment resembles the stagnant air zones of caves. Caves serve as entry points and windows in which humans can observe the fauna living within the voids in the cavernous rock strata. The view is imperfect because so much of the habitat is inaccessible and because the environment is so foreign to human experience.

Fig. 2 The Hawaiian cave earwig, *Anisolabis howarthi* Brindel (family Carcinophoridae). Photo by W. P. Mull.

Fig. 3 The no-eyed big-eyed hunting spider, *Adelocosa anops* Gertsch (family Lycosidae) from caves on the island of Kaua'i. Photo by the author.

Case Study: Hawai'i

Food Web

The main energy sources in Hawaiian lava tube ecosystems are tree roots, which penetrate the lava for several decameters; organic matter, which washes in with percolating rainwater; and accidentals, which are surface and soil animals blundering into the cave. Both living and dead roots are utilized, and this source is probably the most important. Furthermore, both rainwater and accidentals often use the same channels as roots to enter caves, so that root patches often provide food for a wide diversity of cave organisms. The importance of roots in the cave ecosystem makes it desirable to identify the major species. This has become possible only recently by using DNA-sequencing technology. The most important source of roots is supplied by the native pioneer tree on young lava flows: *Metrosideros polymorpha*. Additionally, *Cocculus orbiculatus*, *Dodonaea viscosa*, and *Capparis* are locally important in drier habitats. Several different slimes and oozes occur on wet surfaces and are utilized by scavengers in the cave. They are mostly organic colloids deposited by percolating groundwater, but some may be chemoautotrophic bacteria living on minerals in the lava. Cave-roosting vertebrates do not occur in Hawai'i. Several species of native agrotine moths once roosted in caves in large colonies, but the group has become rare in historic times. The composition of the community that their colonies once supported is unknown. Feeding on living roots are cixiid planthoppers (*Oliarus*). Their nymphs suck xylem sap with piercing mouthparts. The blind flightless adults wander through subterranean voids in search of mates and roots. Potential mates locate each other using species-specific substrate borne songs transmitted along their host roots. Caterpillars of noctuid moths (*Schrankia*) prefer to feed on succulent flushing root tips, but they also occasionally scavenge on rotting plant and animal matter. Tree crickets (*Thaumatogryllus*), terrestrial amphipods (*Spelaeorchestia*), and isopods (*Hawaiioscia* and *Littorophiloscia*) are omnivores but feed extensively on roots. Cave rock crickets (*Caconemobius*) are also omnivorous as well as being opportunistic predators. Feeding on rotting organic material and associated microorganisms are millipedes (*Nannolene*), springtails (*Neanura*, *Sinella*, and *Hawinella*), and phorid flies (*Megaselia*). Terrestrial water treaders (*Cavaticovelia aaa*) suck juices from long-dead arthropods. Feeding in the organic oozes growing on wet cave walls are larvae of crane flies (*Dicranomyia*) and biting midges (*Forcipomyia pholeter*). The blind predators include spiders (*Lycosa howarthi*, *Adelocosa anops* (**Fig. 3**), *Erigone*, *Meioneta*, *Oonops*, and *Theridion*), pseudoscorpions (*Tyrannochthonius*), rock centipedes (*Lithobius*), thread-legged bugs (*Nesidiolestes*), and beetles (*Nesomedon*, *Tachys*, and *Blackburnia*). Most of the cave predators will also scavenge on dead animal material.

Succession

Inhabited Hawaiian lava tubes range in age from 1 month on Hawai'i Island to 2.9 million years on O'ahu Island. On Hawai'i Island colonization and succession of cave ecosystems can be observed. Crickets and spiders colonize the surface of new flows within a month of the flow surface cooling. They hide in caves and crevices by day and emerge at night to feed on windborne debris. *Caconemobius* rock crickets are restricted to living only in this aeolian (wind-supported) ecosystem and disappear with the establishment of plants. The obligate cave species begin to arrive within a year after lava stops flowing in the caves. The predatory wolf spider, *Lycosa howarthi*, arrives first and preys on wayward aeolian arthropods. Other predators and scavenging arthropods – including blind, cave-adapted *Caconemobius* crickets – arrive during the next decade. Under rainforest conditions, plants begin to

invade the surface after a decade, allowing the root feeding cave animals to colonize the caves. *Oliarus* planthoppers arrive about 15 years after the eruption and only 5 years after its host tree, *Metrosideros polymorpha*. The cave-adapted moth, *Schrankia howarthi*, and the underground tree cricket, *Thaumatogryllus cavicola*, arrive later. The cave species colonize new lava tubes from neighboring older flows via underground cracks and voids in the lava. Caves between 500 and 1000 years old are most diverse in cave species. By this time the surface rainforest community is well-developed and productive, while the lava is still young and maximal amount of energy is sinking underground. As soil formation progresses, less water and energy reaches the caves, and the communities slowly starve. In highest rainfall areas, caves support none or only a few species after 10 000 years. Under desert conditions, succession is prolonged for 100 000 years or more. Mesic regimes are intermediate between these two extremes. New lava flows may rejuvenate some buried habitat as well as create new cave habitat. Thus in volcanic caves, succession proceeds upwards with the younger inhabitable caves occurring above barren remnant older caves. The opposite occurs in limestone where solution and down-cutting create younger deeper caves while leaving high and dry remnant passages exposed in cliffs.

Conservation

Caves and their resources are vulnerable to the effects of land-use changes on the surface as well as to changes within the caves. Threats include mining, land-clearing, pollution, water impoundments, recreational use, nonindigenous species invasions and climate change. Two emerging global threats are described in more detail below.

Nonindigenous Species

Several invasive nonindigenous species have invaded cave habitats and are impacting the cave communities. In Hawai'i, the predatory guild is the most troublesome, with some species being implicated in the reduction of vulnerable native species. Among these, the nemertine worm (*Argonemertes dendyi*) and spiders (*Dysdera*, *Nesticella*, and *Eidmanella*) have successfully invaded the stagnant air zone within the smaller spaces. The colonies of cave-roosting moths most likely disappeared from the depredations of the roof rat (*Rattus rattus*) on their roosts and in part from parasites purposefully introduced for biological control of their larvae. Many non-native species (such as *Periplaneta* cockroaches, *Loxosceles* spiders, *Porcellio* isopods, and *Oxychilus* snails) survive well in larger accessible cave passages, where they have some impact, but they appear not to be able to survive in the system of smaller crevices. Alien plants have replaced native species and modified the surface environment over caves thereby directly and indirectly impacting the cave ecosystem below. A few alien trees send roots into caves, creating a dilemma for reserve managers trying to protect both cave and surface habitats since their roots support some generalist native species but not the host-specific planthoppers.

In North America, red imported fire ants (*Solenopsis invicta*) have been implicated in the decline of endangered cave species in Texas. The ants prey on foraging cave crickets, thereby causing critical reduction of food inputs into the caves. Similarly, white-nose syndrome (*Pseudogymnoascus destructans*), a fungal disease of bats recently introduced from Europe, is decimating naïve native bat colonies in eastern North American caves, endangering some bat species and disrupting the food supply in affected caves. With expanding globalization of transportation and commerce, problems caused by nonnative species will become more severe.

Climate Change

Deep caves are buffered environments, and except for areas covered by deep ice sheets, the obligate cave species alive today in temperate regions survived the extreme climatic shifts during the Pleistocene, whereas many of their surface relatives perished. This happenstance suggests that these cave species likely will survive the direct effects of climate change. Even with the predicted increase in extreme droughts and storms, and the expansion or contraction of deserts and forests, many cave species will be able to migrate into and out of deeper cave passages and voids in response to climate change as they do now with changing seasons. A few cave regions are vulnerable; for example, submersion of coastal caves in response to sea level rise and desertification, which would result in less food inputs. In addition, climate change will allow the range expansion of invasive species further enhancing their impacts.

However, the main threats to cave ecosystems resulting from climate change likely will be caused by the responses made by humans as we adapt to climate change. These adaptations include increased water impoundments, mineral and fossil fuel extraction, conversion of marginal land for agriculture, urbanization and other changes in land-use. Cave ecosystems, especially in areas that have not been well-studied, will be out of sight-out of mind and possibly destroyed before the fauna can be documented.

Perspective

The fauna of a large percentage of the world's cave habitats remains unknown to science, and new species continue to be discovered even in well-studied caves. Additional biological surveys are needed to fill gaps in knowledge and improve our understanding of cave ecosystems. Improved methods for sampling the inaccessible smaller voids are needed. The cave environment is a rigorous, high-stress one, which is difficult for humans to access and envision because it is so foreign to human experience. Working in caves can be physically challenging. However, recent innovations in equipment and exploration techniques

allow ecologists to visit the deeper, more rigorous environments. Also, the advent of DNA sequencing technology has been a valuable addition to the cave biologists' tool kit.

In spite of the difficulties of working in the stressful environment, several factors make caves ideal natural laboratories for research in evolutionary and physiological ecology. Since cave habitats are buffered by the surrounding rock, the abiotic factors can be determined with great precision. The number of species in a community is usually manageable and can be studied in total. The remarkable suite of characters shared by cave species worldwide and that evolved independently and in parallel among unrelated species demonstrates the universality of natural selection in forcing adaptions to the similar physical environment. Understanding how this convergent evolution occurs promises to improve our understanding of evolution in general. Questions that are being researched are how organisms adapt to the various environmental stressors; how communities assemble under the influence of resource composition and amount; and how abiotic factors affect ecological processes. For example, a potential overlap between cave and surface ecological studies occurs in some large pit entrances in the tropics. The flora and fauna living in these pits frequently experience CO_2 levels 25–50 times ambient for extended periods of time making them natural laboratories for studying the effects of increased CO_2 levels.

See also: Behavioral Ecology: Herbivore-Predator Cycles. Terrestrial and Landscape Ecology: Landscape Planning

Further Reading

Camacho, A.I. (Ed.), 1992. The natural history of biospeleology. Madrid: Monografias, Museo Nacional de Ciencias Naturales.
Chapman, P., 1993. Caves and cave life. London: Harper Collins Publishers.
Culver, D.C., 1982. Cave life. Cambridge, MA: Harvard University Press.
Culver, D.C., Pipan, T., 2009. The biology of caves and other subterranean habitats (biology of habitats). Oxford, UK: Oxford University Press.
Culver, D.C., White, W.B. (Eds.), 2012. The encyclopedia of caves, 2nd edn. Burlington, MA: Academic Press.
Culver, D.C., Master, L.L., Christman, M.C., Hobbs III, H.H., 2000. Obligate cave fauna of the 48 contiguous United States. Conservation Biology 14, 386–401.
Derkarabetian, S., Steinmann, D.B., Hedin, M., 2010. Repeated and time-correlated morphological convergence in cave-dwelling harvestmen (Opiliones, Laniatores) from montane western North America. Public Library of Science (PLoS ONE) 5 (5), e10388.
Foley, J., Clifford, D., Castle, K., Cryan, P., Ostfeld, R.S., 2011. Investigating and managing the rapid emergence of white-nose syndrome, a novel, fatal, infectious disease of hibernating bats. Conservation Biology 25, 223–231.
Gibert, J., Danielopol, D.L., Stanford, J.A., 1994. Groundwater ecology. San Diego, CA: Academic Press.
Gunn, R.J. (Ed.), 2004. Encyclopedia of caves and karst. New York: Routledge Press.
Howarth, F.G., 1983. Ecology of cave arthropods. Annual Review Entomology 28, 365–389.
Howarth, F.G., 1987. Evolutionary ecology of aeolian and subterranean habitats in Hawaii. Trends Ecolology and Evolution 2, 220–223.
Howarth, F.G., 1993. High-stress subterranean habitats and evolutionary change in cave-inhabiting arthropods. American Naturalist 142, S65–S77.
Howarth, F.G., James, S.A., McDowell, W., Preston, D.J., Yamada, C.T., 2007. Identification of roots in lava tube caves using molecular techniques: implications for conservation of cave faunas. Journal of Insect Conservation 11 (3), 251–261.
Humphries, W.F. (Ed.), 1993. The biogeography of cape range, Western Australia. Perth: Records of the Western Australian Museum. Supplement no 45.
Juberthie, C., Decu, V. (Eds.), 1994–2001. Encyclopaedia Biospeologica 4. Moulis, France: Société de Biospéologie.
Moore, G.W., Sullivan, N., 1997. Speleology: caves and the cave environment, 3rd edn. St. Louis, MO: Cave Books.
Niemiller, M.L., Zigler, K.S., 2013. Patterns of cave biodiversity and endemism in the Appalachians and Interior Plateau of Tennessee, USA. Public Library of Science (PLoS ONE) 8 (5), e64177.
Northup, D.E., Melim, L.A., Spilde, M.N., Hathaway, J.J.M., Garcia, M.G., Moya, M., Stone, F.D., Boston, P.J., Dapkevicius, M.L.N.E., Riquelme, C., 2011. Lava cave microbial communities within mats and secondary mineral deposits: implications for life detection on other planets. Astrobiology 11 (7), 601–618.
Paquin, P., Hedin, M., 2004. The power and perils of 'molecular taxonomy': a case study of eyeless and endangered *Cicurina* (Araneae: Dictynidae) from Texas caves. Molecular Ecology 13, 3239–3255.
Shear, W.A., Taylor, S.J., Wynne, J.J., Krejca, J.K., 2009. Cave millipeds of the United States. VIII. New genera and species of polydesmidan millipeds from caves in the southwestern United States (Diplopoda, Polydesmida, Macrosternodesmidae). Zootaxa 2151, 47–65.
Stone, F.D., Howarth, F.G., Hoch, H., Asche, M., 2012. Root communities in lava tubes. In: Culver, D.C., White, W. (Eds.), Encyclopedia of caves, 2nd edn. Burlington, MA: Elsevier Academic Press, pp. 658–664.
Taylor, S.J., Krejca, J.K., Denight, M.L., 2005. Foraging range and habitat use of Ceuthophilus secretus (Orthoptera: Rhaphidophoridae), a key trogloxene in central Texas cave communities. The American Midland Naturalist 154, 97–114.
Wessel, A., Hoch, H., Asche, M., von Rintelen, T., Stelbrink, B., Heck, V., Stone, F.D., Howarth, F.G., 2013. Founder effects initiated rapid species radiation in Hawaiian cave planthoppers. Proceedings National Academy of Sciences 110 (23), 9391–9396.
Wilkins, H., Culver, D.C., Humphreys, W.F., 2000. In: Subterranean ecosystemsEcosystems of the World 30. Amsterdam: Elsevier Press.

Chaparral

JE Keeley, University of California, Los Angeles, CA, USA

Introduction

Chaparral is the name applied to the evergreen sclerophyllous (hard-leaved) shrub vegetation of southwestern North America, largely concentrated in the coastal zone of California and adjacent Baja California. It is a dense vegetation often retaining many dead spiny branches making it nearly impenetrable (**Fig. 1**). It dominates the foothills of central and southern California but is replaced at higher elevations by forests. On the most arid sites at lower elevations evergreen chaparral is replaced with a lower-stature summer deciduous `soft chaparral' or sage scrub.

Chaparral owes much of its character to the Mediterranean climate of winter rain and summer drought. The severe summer drought, often lasting 6 months or more, inhibits tree growth and enforces the shrub growth form. Intense winter rains coincide with moderate temperatures that allow for rapid plant growth, producing dense shrublands. These factors combine to make this one of the most fire-prone ecosystems in the world. This Mediterranean climate is the result of a subtropical high-pressure cell that forms over the Pacific Ocean. During the summer, this air mass moves northward and blocks water-laden air masses from reaching land, and in winter this high-pressure cell moves toward the equator and allows winter storms to pass onto land. On the Pacific Coast it is wettest in the north, where the effect of the Pacific High is least, and becomes progressively drier to the south, and consequently chaparral dominates more of the landscape in the southern part. Interestingly, these synoptic weather conditions form globally at this same latitude (30–38° north or south) and on the western sides of continents. As a result similar Mediterranean-climate shrublands occur in the Mediterranean Basin of Europe, central Chile, South Africa, and southern Australia.

The Ecological Community

Chaparral is a shrub-dominated vegetation with other growth forms playing minor or temporary successional roles after fire. More than 100 evergreen shrub species occur in chaparral, although sites may have as few as one or more than 20 species, depending on available moisture, slope aspect, and elevation. The most widely distributed shrub is chamise (*Adenostoma fasciculatum*), ranging from Baja to northern California, occurring in either pure chamise chaparral or in mixed stands. It often dominates at low elevations and on xeric south-facing slopes. The short needle-like leaves produce a sparse foliage, and soil litter layers are poorly developed and result in weak soil horizons. Chamise often forms mixed stands of vegetation with a number of species. These include the bright smooth red-barked manzanita (*Arctostaphylos* spp.), the sometimes spiny ceanothus, also known as buckbrush or California lilac (*Ceanothus* spp.). On more mesic north-facing slopes chaparral is commonly dominated by broader-leaved shrubs, including the acorn-producing scrub oak (*Quercus* spp.), the cathartic coffeeberry (*Rhamnus californica*), redberry (*R. crocea*), the rather bitter chaparral cherry (*Prunus ilicifolia*), and chaparral holly (*Heteromeles arbutifolia*), from whence the film capital Hollywood derives its name.

The most common shrub species and the majority of herbaceous species have fire-dependent regeneration, meaning that seeds remain dormant in the soil until stimulated to germinate after fire (see the section titled `Fire' below). These include chamise, manzanita, and ceanothus shrubs, which flower and produce seed most years but seldom produce seedlings without fire. Some ceanothus species are relatively short-lived or are easily shaded out by other shrubs and die after several decades. They, however, persist as a living seed pool in the soil. In addition, a large number of annual species live most of their life as dormant seeds in the soil, perhaps as long as a century or more. Also, many perennial herbs with underground bulbs, known as geophytes, may remain dormant for long periods of time between fires.

All of the other shrub species listed above are not fire dependent and produce seeds that germinate soon after dispersal; however, successful reproduction is relatively uncommon. This is because their seedlings are very sensitive to summer drought and because there are a number of herbivores that live in the chaparral understory and prey on seedlings and other herbaceous vegetation. These include deer mice (*Peromyscus maniculatus*), woodrats (*Neotoma fuscipes*), and brush rabbits (*Sylvilagus bachmani*). Both rodents (mice and rats) are nocturnal; however, evidence of woodrats, or packrats as they are sometimes called, is very evident in many older chaparral stands because of the several foot high nests of twigs they make under the shrub canopy. These animals not only affect community structure by consuming most seedlings and herbaceous species, but also are important vectors for disease and other health threats. For example, deer mice are host to the deadly hanta-virus and woodrats are a critical host for kissing bugs (family Reduviidae) that can cause lethal allergic responses in humans. All animals including reptiles act as hosts for Lyme disease-carrying ticks (*Ixodes pacificus*). The browser of mature shrubs is the black-tailed deer (*Odocoileus hemionus*), although many are attacked by specific gall-forming wasps and aphids. Often scrub oak will have large fruit-like structures produced by gall wasp (family Cynipidae). The adult wasp oviposits on a twig, leaf, or flower and the developing larvae hijack the metabolic activities of the plant cells and force it to produce a highly nutritious spongy parenchymous tissue for the developing wasp larva.

 doi:10.1016/B978-0-444-63768-0.00321-8

Fig. 1 Chaparral shrubland in California. Photo by J. E. Keeley.

These shrubs that reproduce in the absence of fire have successful seedling establishment largely restricted to more mesic plant communities such as adjacent woodlands, or to very old chaparral with deep litter layers that enhance the moisture holding capacity of the soil. When seedlings do establish under the shrub canopy, they typically persist for decades as stunted saplings in the understory. These saplings are heavily browsed by rodents and rabbits and often will produce a swollen woody basal burl that survives browsing and continually sprouts new shoots. If these saplings survive until fire, they are capable of resprouting from their basal burl after fire and exhibit a growth release that enhances their chances of recruiting into the mature canopy during early succession. Thus, in some sense these shrubs may be indirectly fire dependent for completion of their life cycle.

Chaparral has a number of herbaceous or woody (lianas) vines, including manroot (*Marah macrocarpus*) and chaparral honeysuckle (*Lonicera* spp.). These vines overtop the canopy of the shrubs and flower on an annual or near-annual frequency. The former produce fleshy spiny fruits with very large seeds that are highly vulnerable to predation and the latter dry capsules with light seeds that may be wind borne. Both have weak seed dormancy and often establish seedlings in the understory.

Yucca (*Yucca whipplei*) is a fibrous-leaved species that persists as an aboveground rosette of evergreen leaves. It often survives fire because it prefers open rocky sites with very little vegetation to fuel intense fires. Because they are monocotyledonous species they have a central meristem that is protected by the outside leaves, which can withstand severe scorching. This species flowers prolifically after fire and exhibits a remarkable mutualism with the tiny yucca moth (*Tegiticula maculata*). Moth pupae survive in the soil and emerge in the growing season as adults that fly to yucca flowers where they collect pollen. They then instinctively fly to another yucca plant and pollinate the flower, ensuring cross-pollination, and then oviposit an egg in the base of the ovary. This egg soon hatches and the larva feeds on the developing seeds. Yucca moths only reproduce on yucca flowers and yuccas apparently require the pollinator services of this moth for successful seed production, a classic example of symbiosis.

Community Succession

Chaparral succession following some form of disturbance such as fire is somewhat different than in many other ecological communities. Generally all of the species present before fire in chaparral will be present in the first growing season after fire, and thus chaparral has been described as being `auto-successional', meaning it replaces itself. In the absence of disturbance chaparral composition appears to remain somewhat static with relatively few changes in species composition or colonization by new species. In part because of the rather static nature of chaparral, old stands have been described with rather pejorative terms such as `senescent', `senile', `decadent', and `trashy', and considered to be very unproductive with little annual growth. This notion derives largely from wildlife studies done in the mid-twentieth century that concluded, due to the height of shrubs in older stands, there was very little browse production for wildlife. However, if total stand productivity is used as a measure, very old stands of chaparral appear to be very productive and are not justly described as senescent. Also, these older communities appear to retain their resilience to fires and other disturbances, as illustrated by the fact that recovery after fire (see below) in ancient stands (150 years old) recover as well as much younger stands.

Allelopathy

The lack of shrub seedlings and herbaceous plants in the understory of chaparral and related shrublands has led to extensive research on the potential role of allelopathy, which is the chemical suppression by the overstory shrubs of germination (known as enforced dormancy) or growth of understory plants. Often this lack of growth extends to the edge where these shrublands meet grasslands, and forms a distinct bare zone (**Fig. 2**). The importance of allelopathy has long been disputed, with some scientists arguing that animals in the shrub understory are the primary mechanism limiting seedlings and herbaceous species from

Fig. 2 Bare zone between chaparral and grassland. Photo by J. E. Keeley.

establishing. While research has not completely ruled out the possibility of chemical inhibition, it is known that for a large portion of the flora, allelopathy has no role in seed dormancy but rather dormancy is due to innate characteristics that require signals such as heat and smoke to cue germination to postfire environments rich in nutrients and light.

Fire

The marked seasonal change in climate is conducive to massive wildfires, which are spawned by the very dry shrub foliage in the summer and fall and spread by the dense contiguous nature of these shrublands. Fires have likely been an important ecosystem process since the origin of this vegetation in the late Tertiary Period, more than 10 Ma, if not earlier. Until relatively recently the primary source of ignitions was lightning from summer thunderstorms. Fires would largely have been ignited in high interior mountains and coastal areas would have burned less frequently and only when these interior fires were driven by high winds with an offshore flow. In many parts of California such winds occur every autumn and are called Santa Ana winds in southern California and Diablo winds or Mono winds in northern California. When Native Americans colonized California at the end of the Pleistocene Epoch around 12 000 years ago, they too became a source of fires, and as their populations greatly increased over the past few thousand years humans likely surpassed lightning as a source of fire, at least in coastal California. Today humans account for over 95% of all fires along the coast and foothills of California.

Chaparral fires are described as crown fires because the fires are spread through the shrub canopies and usually kill all aboveground foliage. Normally, following a wet winter, high fuel moisture in chaparral shrubs makes them relatively resistant to fire. The amount of dead branches is important to determining fire spread because they respond rapidly to dry weather and combust more readily than living foliage. As a consequence, fires spread readily in older vegetation with a greater accumulation of dead biomass. However, there is a complex interaction between live and dead fuels, wind, humidity, temperature, and topography. In particular, wind accelerates fire spread primarily by heating living fuels and often can result in rapid fire spread in young vegetation with relatively little dead biomass. Fires burning up steep terrain also spread faster for similar reasons.

Community Recovery from Wildfires

Rate of shrub recovery varies with elevation, slope aspect, inclination, degree of coastal influence, and patterns of precipitation. Recovery of shrub biomass is from basal resprouts (**Fig. 3**) and seedling recruitment from a dormant soil-stored seed bank. After a spring or early summer burn, sprouts may arise within a few weeks, whereas after a fall burn, sprout production may be delayed until winter. Regardless of the timing of fire, seed germination is delayed until late winter or early spring and is less common after the first year. Resilience of chaparral to fire disturbance is exemplified by the marked tendency for communities to return rapidly to prefire composition.

Shrub species differ in the extent of postfire regeneration from resprouting versus reproduction from dormant seed banks. Most species of manzanita and ceanothus have no ability to resprout from the base of the dead stem and thus are entirely dependent on seed germination. Such shrubs are termed `obligate-seeders'. A few species of manzanita and ceanothus as well as chamise resprout and reproduce from seeds, and these are referred to as `facultative-seeders'. The majority of shrubs listed above, however, regenerate after fire entirely from resprouts and are `obligate-resprouters'.

In the immediate postfire environment the bulk of plant cover is usually made up of herbaceous species present prior to the fire only as a dormant seed bank or as underground bulbs or corms. This postfire community comprises a rich diversity of herbaceous and weakly woody species, the bulk of which form an ephemeral postfire-successional flora. This `temporary' vegetation is relatively short-lived, and by the fifth year shrubs will have regained dominance of the site and most of the herbaceous species will

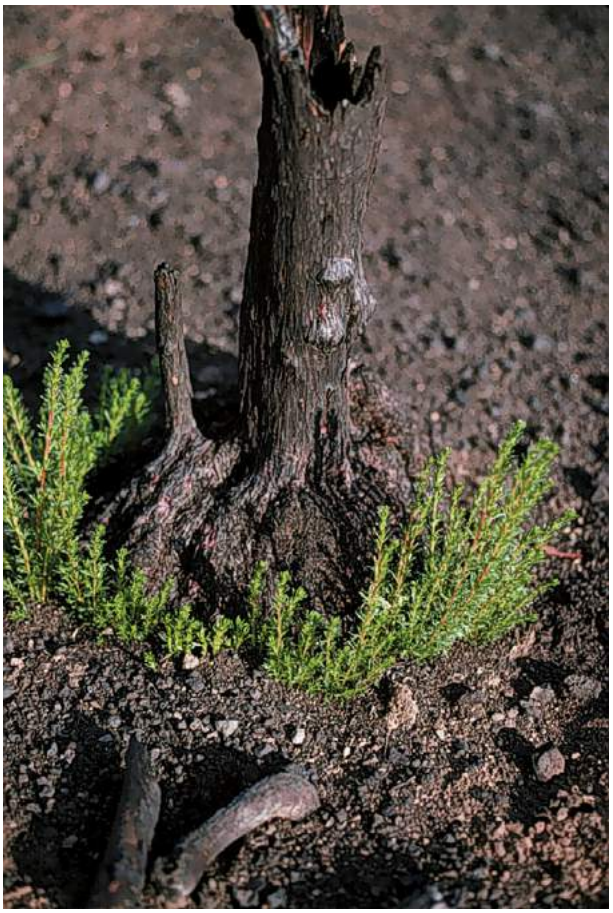


Fig. 3 Postfire resprouts from basal burl of chamise with meter stick. Photo by J. E. Keeley.

return to their dormant state. These postfire endemics arise from dormant seed banks that were generated after the previous fire and typically spend most of their life as dormant seeds. These are termed `postfire endemics' and they retain viable seed banks for more than a century without fire until germination is triggered by heat or smoke of a fire. Postfire endemics are highly restricted to the immediate postfire conditions and if the second year has sufficient precipitation may persist a second year but usually disappear in subsequent years.

Not all of the postfire annuals are so restricted, rather some are quite opportunistic, taking advantage of the open conditions after fire but persisting in other openings in mature chaparral. Such species often produce polymorphic seed pools with both deeply dormant seeds that remain dormant until fire and nondormant seeds capable of establishing in or around mature chaparral. These species fluctuate in relation to annual precipitation patterns, often not appearing at all in dry years.

Herbaceous perennials that live most of their lives as dormant bulbs in the soil commonly comprise a quarter of the postfire species diversity. Nearly all are obligate resprouters, arising from dormant bulbs, corms, or rhizomes and flowering in unison in the first postfire year. Almost none of them produce fire-dependent seeds; however, reproduction is fire dependent because postfire flowering leads to produce nondormant seeds that readily germinate in the second year.

Diversity in chaparral reaches its highest level in the first year or two after fire. It is made up of a large number of relatively minor species and a few very dominant species and is illustrated by dominance–diversity curves (**Fig. 4**). Dominance in chaparral is driven by the fact that a substantial portion of resources are taken by vigorous resprouting shrubs and much less is available for the many annual species regenerating from seed.

Plants are not the only part of the biota that has specialized its life cycle to fire. Smoke beetles (*Melanophila* spp.) are widely distributed in the western US and are attracted by the infrared heat given off by fires. Often while stems are still smoldering they will bore into the scorched wood and lay their eggs.

Seed Germination

Many chaparral species have fire-dependent regeneration, meaning that dormant seeds in the soil require a stimulus from fire for germination. A few species have hard seeds that are cracked by the heat of fire and this stimulates germination. Ceanothus seeds

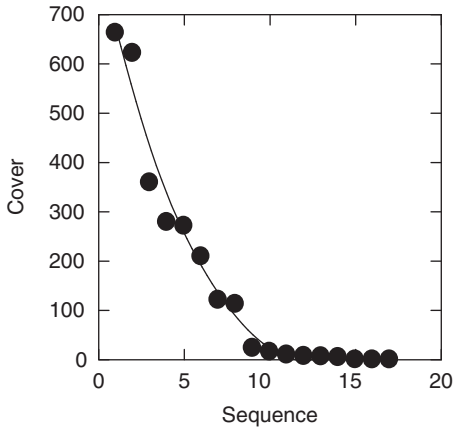


Fig. 4 Dominance–diversity curve based on cover of species in sequence from highest to lowest from postfire chaparral.

are a good example of this germination mode. However, for the majority of species, seeds do not respond to heat but rather to chemicals generated by the burning of plant matter. This can result from exposure to smoke or charred wood. In many of these species seeds will not germinate when placed at room temperature and watered, unless they are first exposed to smoke or charred wood. In natural environments the seeds remain dormant for decades until fire. There is evidence that a variety of chemicals in smoke and charred wood may be responsible for stimulating germination of postfire species, and both inorganic and organic compounds may be involved.

Seeds of many species have a requirement for cold temperatures ($<5\ ^{\circ}$C), which is interpreted as a seasonal cue, but in these chaparral species this requirement is not like the cold stratification requirement of many species from colder climates, where the seeds require a certain duration of cold in order to prevent winter germination. In California species just a short burst of cold often will trigger germination; thus, cold is not a cue that winter is over (as with more northern latitude species) but rather that winter has arrived, which is consistent with the winter germination behavior of these Mediterranean-climate species.

Seed Dispersal

Shrubs can be divided into those with temporal dispersal versus those with strong spatial dispersal. The former are the fire-dependent species that accumulate dormant seed banks, which in essence disperse these shrubs in time, from one fire cycle to the next. Within this group there is limited spatial dispersal. Ceanothus have explosive capsules that shoot seeds a short distance of a meter or two from the parent shrub. Manzanitas drop most of their seeds beneath the parent plant because their small dry fruits are not attractive to birds, although a small number of the seeds are distributed further by coyotes (*Canis latrans*) and bears (*Ursus americans*, historically also included *U. horribilis*). Chamise produces small light fruits that may be carried tens of meters or more by the wind but it appears that most are distributed around the parent shrub.

The postfire endemic annuals also have seeds that are largely dispersed in time rather than in space. Most do not have characteristics suggestive of widespread dispersal. For example, the fire-following whispering bells (*Emmenanthe penduliflora*) derives its name because the flowers and fruits are pendulous and drop seeds directly beneath the parent plant. Postfire endemics in the sunflower family (Asteraceae), a family noted for well-developed dispersal with dandelion-like pappus, commonly have deciduous pappus, which ensures dispersal in time rather than in space.

Shrub species that exhibit fire-free (nonfire-dependent) reproduction have fruits highly attractive to birds and mammals, and the bulk of the seed crop appears to be dispersed by these vectors. Seedling recruitment is sensitive to desiccation and thus it is of some significance that one of the main dispersers of these fruits, the scrub jay (*Aphelocoma californica*), preferentially caches seeds in the shade.

Regional Variation in Fire Regime

California chaparral exhibits regional differences in burning patterns and largely due to regional variation in winds. In much of coastal California autumn winds create severe fire conditions. These occur every year and result in 5–10 days of strong offshore flow with windspeeds of 100 kph or more. These winds result from a high-pressure system in the interior West, and are known as Santa Ana winds in southern California and Diablo or Mono winds in northern California. As these air masses move from the high-pressure cell in the interior to a low-pressure trough off the coast, the air descends and dries adiabatically, resulting in relative humidity below 10%. The fact that these winds occur every year and arrive at the end of an extended drought results in one of the

most severe fire conditions in the world. As a consequence only a small portion of southern California landscape has escaped fire during the last century, and much of the lower-elevation chaparral has burned at an unnaturally high frequency.

In contrast, Santa Ana winds are absent from the southern Sierra Nevada and parts of the central coast, in part to mountain barriers that fail to funnel these winds coastward. This, coupled with lower human population density, has resulted in many fewer fires. As a consequence nearly half of the landscape in the southern Sierra Nevada has not had a fire for well over a century. This condition places these landscapes at the upper end of the historical range of variability. Nonetheless, these older stands of chaparral appear to maintain natural ecosystem processes and exhibit no sign of dying out or replacement by other vegetation types. This is particularly evident, following fires in ancient stands of chaparral from the region, that it exhibits vegetative recovery in cover and diversity indistinguishable from postfire recovery in younger stands.

Future Threats and Management

Degradation and type conversion of native shrublands to alien-dominated grasslands has been noted by numerous investigators, some of whom contend that increased frequency of disturbance is the primary factor that favors non-native annuals over woody native species. In the absence of fire, seeds of non-natives have a low residence time in the soil; thus, the presence of these species on burned sites is more often due to colonization after fire. Typically a repeat fire within the first postfire decade is sufficient to provide an initial foothold for aliens. In addition to outcompeting native plant species, non-native grasses alter the fire regime from a crown-fire regime to a mixture of surface- and crown-fires, where highly combustible grass fuels carry fire between shrub patches. This increases the likelihood of fires and ultimately increases fire frequency. As fire frequency increases there is a threshold beyond which the native shrub cover cannot recover.

Fire management practices potentially conflict with natural resource needs. These landscapes currently experience an unnaturally high frequency of fire; thus, much of it is at risk for alien invasion. When fire managers add to this by using prescription burning and other fuel manipulations, they open up these shrublands and expose them to invasion and potential type conversion to non-native grasslands. In managing these landscapes it might be helpful to consider the fact that the vast majority of alien species in California are opportunistic species that capitalize on disturbance. Adding additional disturbance through prescription burning (or grazing) will only exacerbate the alien problem.

Very little chaparral landscape is protected in parks or wilderness areas. Much of it is in private hands or under federal jurisdiction. Historically, it has largely been managed as rangeland by frequent burning to destroy the chaparral cover, or burned to reduce fuels perceived to be hazardous to more desirable forests or urban environments. Today the expansion of urban development has resulted in large portions of urban communities being juxtaposed with watersheds of potentially dangerous chaparral fuels. Historical studies show that large high-intensity crown fires are a natural part of this ecosystem and there is little reason to believe there will not be more such fires in the future. Fire management has always worked under the philosophy that they can change the vulnerability of communities to wildfires through manipulation of fuels. However, over the past century of such management, every decade has been followed by one of increasing losses to wildfires. Californians need to embrace a different model of how to view fires on these landscapes. Our response needs to be tempered by the realization that these are natural events that cannot be eliminated from the southern California landscape. We can learn much from the science of earthquake or other natural hazard management. No one pretends they can stop earthquakes; rather, they engineer infrastructure to minimize impacts. In the future, living safely with fire is not going to be achieved solely by fire management practices, but will require close integration with urban planning.

See also: Aquatic Ecology: Eutrophication. Terrestrial and Landscape Ecology: Landscape Planning

Further Reading

Arroyo, M.T.K., Zedler, P.H., Fox, M.D. (Eds.), 1995. Ecology and Biogeography of Mediterranean Ecosystems in Chile, California and Australia. New York: Springer.
Christensen, N.L., Muller, C.H., 1975. Effects of fire on factors controlling plant growth in Adenostoma chaparral. Ecological Monographs 45, 29–55.
Halsey, R.W., 2004. Fire, Chaparral, and Survival in Southern California. San Diego, CA: Sunbelt Publications.
Halsey, R.W., 2005. In search of allelopathy: An eco-historical view of the investigation of chemical inhibition in California coastal sage scrub and chamise chaparral. Journal of the Torrey Botanical Society 131, 343–367.
Keeley, J.E., 2000. Chaparral. In: Barbour, M.G., Billings, W.D. (Eds.), North American Terrestrial Vegetation. Cambridge: Cambridge University Press, pp. 203–253.
Keeley, J.E., Fotheringham, C.J., 2003. Impact of past, present, and future fire regimes on North American mediterranean shrublands. In: Veblen, T.T., Baker, W.L., Montenegro, G., Swetnam, T.W. (Eds.), Fire and Climatic Change in Temperate Ecosystems of the Western Americas. New York: Springer, pp. 218–262.
Mooney, H.A. (Ed.), 1977. Convergent Evolution of Chile and California Mediterranean Climate Ecosystems. Stroudsburg, PE: Dowden, Hutchinson and Ross.
Odion, D.C., Davis, F.W., 2000. Fire, soil heating, and the formation of vegetation patterns in chaparral. Ecological Monographs 70, 149–169.
Rundel, P.W., Montenegro, G., Jaksic, F.M. (Eds.), 1998. Landscape Disturbance and Biodiversity in Mediterranean-Type Ecosystems. New York: Springer.
Wells, P.V., 1969. The relation between mode of reproduction and extent of speciation in woody genera of the California chaparral. Evolution 23, 264–267.

Coral Reefs

DE Burkepile and ME Hay, Georgia Institute of Technology, Atlanta, GA, USA

Introduction

Corals are simple, clonal invertebrates that serve as ecosystem engineers, building living structures (reefs) so large that they can be seen from space. These structures, which rival the greatest feats of human engineering, are powered through symbiosis with single-celled algae that are housed within the coral animal. This coral–algal cooperation facilitates a productive ecosystem that can grow in the nutrient-poor `desert' of isolated tropical seas. The rich structural complexity provided by the coral's hard bodies gives shelter to many other species of plants and animals making coral reefs among the Earth's most biologically diverse ecosystems, harboring hundreds of thousands to millions of species worldwide (**Fig. 1**).

Coral reefs also support human societies by providing critical sources of protein, protecting coasts from damaging waves, attracting tourists, and serving as the backbone of the economies for many tropical islands. In addition, coral reefs are crucial in the fight against human diseases, as many of the plants and animals that live on coral reefs produce chemicals that are useful as pharmaceuticals. Reefs have also fascinated naturalists and scientists for centuries. Before publishing his groundbreaking work on natural selection, Charles Darwin published a treatise on reefs in 1842 hypothesizing that coral atolls (rings of reefs in the deep tropical Pacific) were formed around mountain tops as these mountains sank back into the Earth's crust under their own weight. It was more than 100 years before drilling technologies developed to the point where this hypothesis was tested – like with so many other aspects of Darwin's writings, he proved to be correct. For modern ecologists, reefs serve as a model ecosystem for developing basic hypotheses about the ecology and evolution of population structure, of community organization, and about how species diversity evolves and is maintained. In addition, reefs give us a glimpse of the spectacular record of Earth's history because the hard skeletons of corals fossilize to provide a long record of changes in coral distribution and abundance and also record chemical signals of past climatic events, like temperature and sea-level changes. Thus, reefs not only feed and protect humans and other species, but also provide a valuable window into our past, including how our present activities may be changing our environment, and possibly our future.

In this article, we review the major ecological interactions that shape coral reef ecosystems. We pay particular attention to (1) the dynamic relationship between corals and the symbiotic algae living within their tissues, (2) the role of reef herbivores in protecting corals from being overgrown by seaweeds, (3) the numerous ecological processes such as predation, competition, recruitment of juvenile reef organisms, and disturbance that influence the structure of coral reefs, and (4) the dynamic ecological connections between reefs and nearby ecosystems such as seagrass beds and mangrove forests. Finally, we review the current dangers to coral reefs and how these threats undermine the ecological integrity of these diverse ecosystems.

The Coral–Algal Mutualism

Corals are ecosystem engineers in that the growth of their calcium carbonate skeleton creates the biogenic structure on which the entire ecosystem depends. The calcification and growth of reef corals depends on a mutualism between corals and their intracellular, photosynthetic dinoflagellates, *Symbiodinium* spp., also known as zooxanthellae. Both corals and zooxanthellae benefit from this relationship, as corals can receive up to 95% of their carbon from the zooxanthellae's photosynthesis, while the zooxanthellae acquire the nitrogen and other inorganic nutrients in coral excretion products for their growth. Photosynthesis by

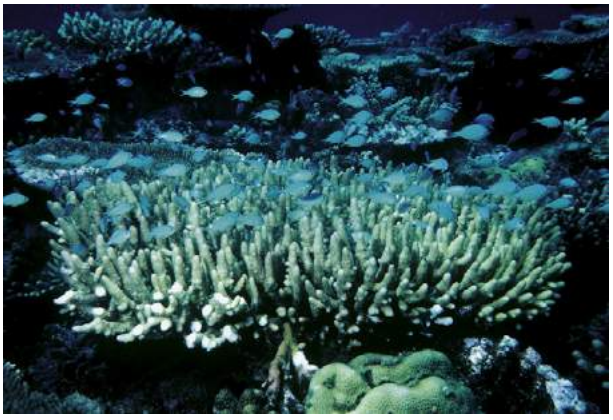


Fig. 1 Coral reefs, like this one in the Indo-Pacific, harbor hundreds of thousands to millions of species worldwide. Photo credit M.E. Hay.

 doi:10.1016/B978-0-444-63768-0.00323-1

zooxanthellae enhances calcification in corals and increases coral growth rates, ultimately leading to reef accretion and the massive reef framework found in many tropical seas. Thus, the physical structure of live and dead corals created by the coral–zooxanthellae mutualism provides heterogeneity and habitat complexity, facilitating the coexistence of diverse plant and animal assemblages.

Although zooxanthellae were initially assumed to represent one species, recent molecular evidence shows that there are at least seven distinct types or clades (referred to as clades A–G). Many corals house multiple clades of zooxanthellae, setting the stage for possible competition among symbionts and for symbiont selectivity by the host. Clades of zooxanthellae differ in their photosynthetic capacity and their tolerance of light, temperature, and other stressors, making them differentially useful to their hosts under changing environmental conditions. When corals are stressed by increasing light levels or temperatures, they often expel their zooxanthellae and become pale in color (called coral bleaching). This process of bleaching may allow corals to take up new clades of zooxanthellae that are better adapted to the new environmental conditions. However, corals that fail to re-acquire zooxanthellae or acquire the wrong clades may ultimately die from the stress, suggesting that a failure of corals to acquire appropriate symbionts can be fatal under changing environmental conditions. Such alterations in the coral–zooxanthellae mutualism may allow corals greater flexibility in adapting to global climate change, which is a major threat to the health of coral reefs and the integrity of the coral–zooxanthellae mutualism.

Ecological Interactions on Coral Reefs

Competition

Competition for limiting resources such as nutrients, space, light, or food is often a strong mechanism limiting the distribution and abundance of species in communities. On many coral reefs, the limiting resource for most benthic organisms is space or light, as most of the reef structure is often occupied (Fig. 2). Consequently, corals have evolved a variety of competitive mechanisms including sweeper tentacles, digestive filaments, and rapid growth rates that allow them to fight neighbors for new space or protect the space they already occupy. Slow growing, massive corals often have the most potent direct competitive mechanisms (i.e., sweeper tentacles and digestive filaments that can sting and directly harm neighboring corals) while branching corals such as many

Fig. 2 Competition for space is often an important ecological force structuring coral reefs as corals and other invertebrates cover most of the benthos on healthy coral reefs. Photo credit M.E. Hay.

Acropora spp. rely on their high growth rates to overtop and shade competitors. Other reef invertebrates such as sponges exude chemical compounds that are toxic to their neighbors, essentially using chemical warfare (termed allelopathy) to gain new space.

Although most early studies of competition on reefs focused on coral–coral competition, more recent studies have examined coral–seaweed competition because reefs are now more commonly overgrown by seaweeds that periodically seem to be killing corals. The conventional wisdom is that seaweeds are competitively superior and can overgrow and kill most corals. Although not all seaweeds are harmful to corals, most studies of coral–algal competition show that direct competition from seaweeds reduces the growth, survivorship, fecundity, and recruitment of many corals. Contact with the calcareous green seaweed *Halimeda opuntia* has even been shown to induce black band disease in some corals. Small, filamentous seaweeds, which are not as directly harmful to corals as are larger, foliose seaweeds, often trap sediments next to coral tissue, and this can smother and kill corals. Thus, even competition with typically innocuous filamentous seaweeds can be harmful on reefs that receive high sediment loads. However, corals are not uniformly susceptible to competition from seaweeds, and competitive outcomes may vary with coral morphology. Foliose corals such as *Agaricia* spp. are more susceptible to seaweed overgrowth than massive corals such as *Montastrea* spp. In addition, seaweeds have disproportionately high negative effects on smaller coral colonies, particularly newly recruited corals, and large stands of seaweed can prevent juvenile corals from recruiting to reefs at all.

Competition on coral reefs is not limited to sessile invertebrates; mobile animals also compete. Because herbivores are abundant on undisturbed coral reefs and standing biomass of seaweeds is generally low in these conditions, competition between herbivores would be expected. When the herbivorous sea urchin *Diadema antillarum* was removed from Caribbean reefs by either purposeful experimental manipulations or by disease outbreak, feeding rates of herbivorous fishes increased, as did the densities of some species, suggesting that fishes and urchins competed for food. *Diadema* also competed intensely with each other for food. However, competition for limiting algal resources generally did not result in a decrease in the size of *Diadema* populations but an increase in the size of their mouthparts (called the Aristotle's lantern) relative to the size of their body. Basically, *Diadema* bodies would shrink in size when food was limiting as a tradeoff between growth and survival.

Herbivory

Because seaweeds can overgrow and kill corals, herbivores are critical for coral reef function because they keep reefs free of seaweeds, thus facilitating the recruitment, growth, and resilience of corals. Fishes and urchins are typically the dominant herbivores on coral reefs with fishes in some reef areas biting the bottom at rates of $>100\,000$ times per m^2 every day. When in sufficient numbers, either fishes alone or sea urchins alone can remove greater than 90% of the daily primary production on reefs. By feeding on seaweeds that are competitively superior to corals, herbivorous fishes both clear the substrate for settling coral larvae and prevent seaweed overgrowth of established corals. In return, the biogenic structure and topographic complexity of reef corals benefit herbivorous reef fishes and urchins by providing food, habitat, and refuges from predation. When herbivores are removed by experimentation, overfishing, or disease, seaweeds replace corals and the biogenic structure of the reef degrades. Both reductions in coral structure and increases in seaweeds are associated with losses of herbivorous reef fishes. Interestingly, large-scale manual removals of seaweeds from reefs have resulted in only temporary increases in herbivorous fish abundance with seaweeds becoming the dominant benthic organism once again after several months. Thus, reductions in seaweeds without recovery of corals may inhibit the recovery of many reef fishes, leading to the continued degradation of coral reefs.

The main herbivorous fishes on coral reefs are generally surgeonfishes (Acanthurdiae) and parrotfishes (Scaridae) with rabbitfishes (Siganidae), chubs (Sparidae), and damselfishes (Pomacentridae) also responsible for considerable herbivory in some locations (**Fig. 3**). Surgeonfishes typically feed on turf algae and foliose seaweeds with some species feeding primarily on detrital material. Parrotfishes have robust jaws with teeth fused into a beak-like formation (hence the name parrotfish), which allows them to feed on tough, calcified seaweeds in addition to algal turfs and foliose seaweeds. Although the important role of herbivores in

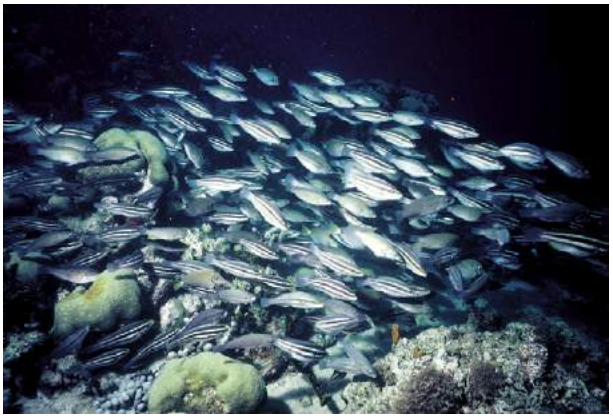


Fig. 3 Herbivores, like this mixed-species school of parrotfishes in the Caribbean, are important to coral reef health because they remove seaweeds that would otherwise overgrow and kill corals. Photo credit M.E. Hay.

influencing reef community is well established, less is known about the importance of individual herbivore species or the role of herbivore diversity in affecting coral reef health. Herbivore diversity should benefit reefs because a more diverse herbivore assemblage should include herbivores with varied attack strategies, which in turn should increase the efficiency of seaweed removal because particular seaweeds are unlikely to be well defended against all types of herbivores. Experimental manipulations of herbivorous fish diversity demonstrate that species-richness is important for reef function because complementary feeding by different herbivorous fishes suppresses upright seaweeds, facilitates crustose corallines and turf algae, reduces coral mortality, and promotes coral growth. Hence not only are herbivores critical for coral reefs, but herbivore species-richness is also essential as a range of feeding strategies and physiologies allows efficient removal of seaweeds and promotes coral health.

Predation

Predation is often a strong top-down force in ecosystems mediating coexistence of lower trophic-level species by preventing competitive exclusion among ecologically similar organisms. In fact, predators often maintain species diversity in ecological communities by preventing expansions of certain prey that would otherwise outcompete competitively inferior organisms and come to dominate the community. If important predators are removed from a food web, the absence of their strong effects can ripple throughout the system, fundamentally altering a variety of predator–prey interactions.

The effects of the largest predators on reefs such as sharks, jacks (Carangidae), and large groupers (Serranidae) are virtually unknown due to the logistical problems of studying such large creatures and the fact that the majority of these species were rare before ecologists began studying reef ecology *in situ* (**Fig. 4**). Although rigorous study of the roles that these fishes play in communities has been limited, a recent model of a Caribbean reef food web suggests that sharks are often the most strongly interacting species in these webs indicating that their removal may have had strong cascading effects on reefs. Further, surveys of lightly fished reefs in the northwestern Hawaiian Islands showed that large apex predators such as sharks and jacks represented $>50\%$ of the total fish biomass as compared to $<3\%$ on heavily fished reefs from the main Hawaiian Islands. These abundant apex predators on lightly fished reefs surely exert a strong top-down force on the community structure of these reefs.

However, the human exploitation of medium-sized predatory fishes has given us the best insight into how predation influences reef communities. On many Pacific coral reefs, outbreaks of the crown-of-thorns starfish, *Acanthaster planci*, cause loss of many square kilometers of coral reefs. These starfish are voracious coral predators that forage in large groups of up to hundreds of

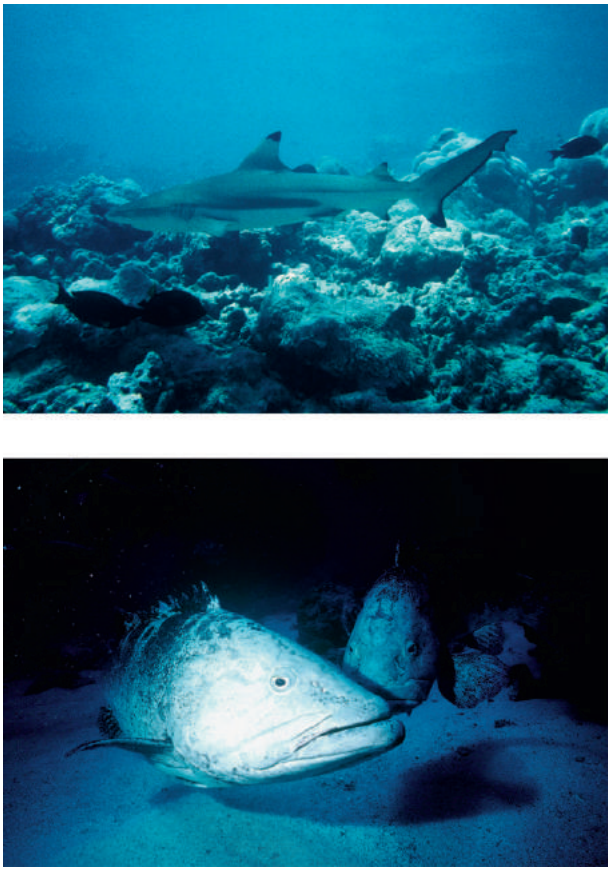


Fig. 4 Apex predators like sharks (top) and groupers (bottom) are now rare on many coral reefs due to overfishing. Photo credit M.E. Hay.

thousands of individuals that can decimate large stands of coral, and their outbreaks have become more frequent since the 1960s when they were first documented. Research in the Fiji islands has shown that outbreaks of *Acanthaster* are correlated to fishing pressure on reefs. *Acanthaster* are 1000 times more dense around islands that have high fishing pressure and low predatory fish abundance than they are on reefs that have light fishing pressure and high predator abundance. High densities of *Acanthaster* decrease cover of reef-building corals and crustose coralline algae while increasing cover of filamentous algae. Thus, the removal of large predators is associated with explosions of *Acanthaster* populations that then have strong cascading effects on the organization of reef communities.

A similar situation exists on many reefs in eastern Africa where intense fishing of triggerfishes and large wrasses allows population explosions of sea urchins such as *Echinometra mathaei*. Reefs unprotected from fishing have six times more urchins than protected reefs, and feeding by these dense urchin populations physically erodes the reef structure once most of the algal biomass has been consumed. This intense grazing decreases coral cover and diversity as well as increases bioerosion rates up to 20-fold compared to reefs that are protected from fishing and have abundant predators. When formerly fished areas were protected from fishing, urchin-eating fishes became more abundant and predation on urchins increased, suggesting that recovery of predatory fish populations should lead to lower urchin populations and a recovery of reef structure over time.

Disturbance

Although biotic interactions (e.g., competition and herbivory) are emphasized as having important consequences for coral reef structure, abiotic disturbances such as hurricanes, temperature fluctuations, sedimentation stress, and sea-level change also produce long-lasting effects on reefs. Coral reefs are one of the hallmark ecosystems strongly influenced by disturbance as the frequency and intensity of hurricanes or disturbance events determines how many species of corals coexist on reefs. If disturbance is very frequent or very intense, then only species that can recolonize disturbed areas quickly or that can withstand intense disturbances will persist. If disturbance is infrequent and mild, then the most competitive species eliminate the less competitive species and come to dominate. However, if disturbance is of an intermediate frequency and intensity, then species with different life-history characteristics (i.e., good colonizers vs. good competitors) can coexist because the disturbance-intolerant species are not displaced frequently and the poor competitors are not outcompeted.

Reefs often recover from acute disturbances such as storms but infrequently recover from chronic disturbances. The coupling of acute natural disturbances with chronic anthropogenic disturbances often leads to precipitous declines in coral reef health. One of the best examples of compounded disturbances driving coral reef decline is from the reefs of Jamaica. Chronic overfishing of herbivorous fishes compounded with two hurricanes and the mass mortality of the herbivorous sea urchin *Diadema antillarum* acted synergistically to force these once coral-dominated reefs into an alternate state of seaweed dominance (**Fig. 5**). In more than two decades, these reefs have shown few signs of recovery. In fact, the episodic effects of natural physical disturbances, coral disease, and coral bleaching along with the constant anthropogenic disturbances of overfishing and pollution have combined to decrease coral cover an average of 80% on reefs throughout the Caribbean over the past few decades. Although disturbance is a natural and integral part of coral reef ecosystems, the compounding of many disturbances over short timescales is often more than reefs can withstand.

Positive Interactions

Ecologists now realize that positive interactions between species can have strong, cascading effects on natural communities and are no less important than negative interactions (i.e., predation or disturbance) in affecting community structure. On reefs, the most obvious positive interaction is the mutualism between corals and their symbiotic algae. Another is the positive feedback between herbivores and corals that maintains a coral-dominated ecosystem. Other crucial positive interactions come from species that are normally thought of as competitors but can mutually benefit each other under the right conditions. Sponges, for example, compete with each other, but can also interact positively. It is more common to find morphologically similar species of sponge growing intermingled in multispecies groups than it is to find a sponge colony growing alone. When growing in these groups, the growth rates of the different species of sponge are often greater than what they would be if these sponges were growing by themselves. This enhancement of growth rates may stem from differences among the species in their susceptibility to predation, pathogens, and physical disturbance. The summed traits of the sponge consortia may enable participants to survive environmental challenges that would be insurmountable for any of them growing alone. Further, sponges are important to the stability and integrity of the reef itself. Sponges actually act as a type of cement that binds the reef together and holds corals in place. When sponges are removed from reefs, storms displace and kill more corals from these reefs than from reefs that have an abundance of sponges.

Net positive interactions may also occur even between consumers and their prey. Herbivorous damselfishes often form mutualisms with some seaweeds on tropical reefs. Through aggressive defense of the algal mats on which they feed, damselfish create patches of species-rich algae on reefs where these algae would normally be grazed to near local extinction by large herbivorous fishes. Although the rapidly growing filamentous algae in the damselfish's territory are its prey, they are also dependent on the territorial behavior of the fish for their persistence at high density. If the territorial fish is removed, its algal lawn is consumed within hours. However, the positive interactions between damselfishes and their algal gardens can be overridden by cooperation among other species of herbivorous fishes that forage in large schools. While schooling would appear to increase the

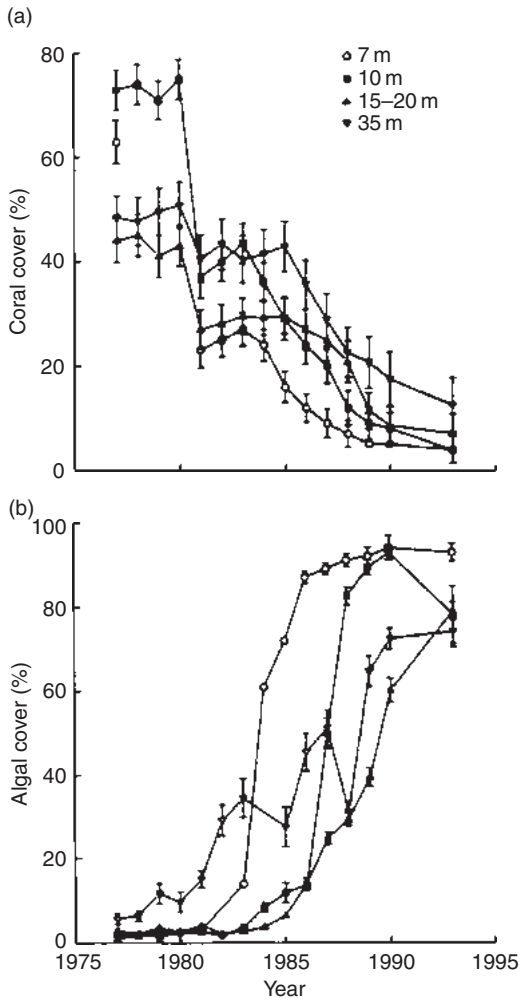


Fig. 5 Degradation of Jamaican coral reefs over two decades. Changes in (a) coral cover and (b) seaweed cover at four depths in Discovery Bay, Jamaica. This decline in corals and increase in seaweeds was the result of synergistic interactions of natural and anthropogenic disturbances including overfishing, hurricanes, disease, and eutrophication. A similar decrease in coral cover occurred throughout Jamaica with coral cover nationwide declining from about 60% to about 4%. From Hughes TP (1994) Catastrophes, phase shifts, and large-scale degradation of a Caribbean coral reef. *Science* 265: 1547–1551.

competition for resources among herbivores, parrotfishes and surgeonfishes often form feeding schools to overrun territories of pugnacious damselfishes. For these large herbivores, increasing school size allows for more bites per individual fish when foraging in and around damselfish territories. The benefits of acting mutually to overwhelm damselfish and gain access to resource-rich habitats must outweigh the potential for competitive interactions between the fishes using these schools. Similarly, piscivorous fishes such as grouper and moray eels often hunt in mixed-species or cooperative foraging groups. For these fishes that forage cooperatively, many mouths may be better than one in terms of overall prey yield to each predator when summed over a lifetime of hunts.

Finally, small wrasses (Labridae) and gobies (Gobiidae) act as cleaner fishes on tropical reefs and remove parasites, mucus, and dead or infected tissue from larger fishes. Reef-based cleaner fish are found at specific cleaning stations, usually situated on prominent portions of the reef. These fish can clean up to 2300 individuals of 132 different species in a day, and some client fish visit cleaners over 100 times a day. If these cleaner fish are removed from reefs, diversity of reef fishes declines, especially for large, transient fishes that may visit reefs specifically to be cleaned. Cleaner fishes can, thus, have a strong effect on parasite loads of their client fishes and on fish-usage patterns across patchy reef environments.

Replenishment of Coral Reefs: The Role of Reproduction and Recruitment in the Ecology of Reefs

Most reef organisms are sessile (corals, sponges, seaweeds) or use only a small portion of a much larger reef habitat (most reef fishes). Thus, colonization of new habitats is achieved through recruitment of juvenile organisms that may drift for long distances

in the plankton before settling onto, and using a small area of reef. Consequently, the production and recruitment of juvenile organisms is a key factor in the ecology of reefs as the replenishment of plant and animal populations is integral to the resilience and recovery of reefs in the face of natural and anthropogenic disturbances.

Coral species differ considerably in their modes of reproduction and in the ability of their larvae to disperse to new reefs. Many corals reproduce both asexually through fragmentation and sexually by the production of gametes. Important reef-building corals such as acroporids are extremely successful at reproducing asexually and are dispersed when storms break apart parent colonies and spread the fragments to new portions of a reef where they can reattach and grow. Sexual reproduction in corals is also variable in that corals are typically either brooders or spawners. Brooders release fertilized larvae into the water column while spawners release sperm and eggs into the water column, where they fertilize and disperse with the ocean currents. These fertilized larvae will eventually exit the plankton and return to reefs as newly recruited juvenile corals. Research from the Great Barrier Reef, Australia has shown that there is large variation in the abundance of coral recruits across both large and small spatial scales. The best predictor of differences in recruitment rates among reefs was the fecundity, not abundance, of adult corals and explained 72% of the variation in recruitment for acroporid corals. Recruitment rates decreased dramatically as the fecundity of adults decreased, but this decrease was not linear; a small decrease in the fecundity of adults resulted in a dramatic decrease in juvenile recruitment (**Fig. 6**). These results suggest that processes such as sedimentation, eutrophication, and competition with seaweeds, all of which reduce the fecundity of adult corals, could dramatically affect the replenishment of coral populations.

Recruitment of juveniles is also important to the replenishment of fish populations and considerable research has gone into determining how recruitment processes affect the assembly of reef fish communities. Most reef fishes, like corals, have planktonic larvae that can disperse wide distances from their point of origin. One of the key questions in the ecology of reef fishes is how the recruitment of juvenile fishes is related to the density of fishes already on the reef (i.e., whether local patterns of recruitment are density dependent or density independent). Despite considerable research on the subject, little consensus has been reached and studies have shown that recruitment rates can be either positively or negatively correlated with adult abundance (positively or negatively density dependent), or show no correlation at all (density independent). These relationships may vary with the species being studied, with location, or with the currents and physical processes prevailing at the time of the test. Continued research is needed to generalize how recruit and adult densities are related and how environmental and biological variables change these relationships.

A key component to the replenishment of populations of coral reef organisms is the extent to which reefs are connected to other reefs (i.e., whether juveniles recruit to reefs from local or distant sources). Coral reefs, and marine ecosystems in general, differ from many terrestrial systems in that juvenile organisms have the potential to ride ocean currents and be dispersed over wide distances potentially connecting geographically distant populations. However, the actual extent that marine populations are connected to each other is still a topic of vigorous debate. This knowledge is crucial to the protection and management of reefs as the connectivity of populations of coral reef organisms will determine whether local populations can be managed with efforts based close to the target population (if the system is fairly closed and recruitment from local populations is frequent) or if management of local populations will necessitate international cooperation (if reefs are fairly open and recruitment is driven by larvae from distant reefs). Thus, solving the question of connectivity among reefs is critical to the preservation of reef health.

Initial models of connectivity for fish populations in the Caribbean suggested that many of the populations were very open and well connected to other populations hundreds of kilometers away. However, these models were based on passive

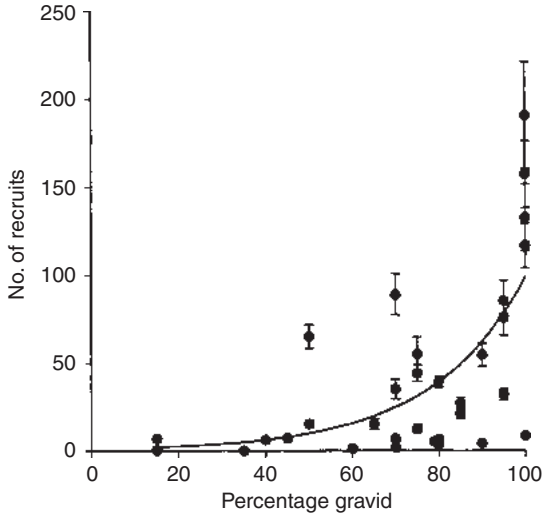


Fig. 6 Relationship between the percentage of coral colonies (*Acropora hyacinthus*) that were gravid and the number of coral recruits. Each point is a separate reef on the Great Barrier Reef, Australia. From Hughes TP, Baird AH, Dinsdale EA, *et al.* (2000) Supply-side ecology works both ways: The link between benthic adults, fecundity, and larval recruits. *Ecology* 81: 2241–2249.

dispersal of fish larvae and simple models of surface currents and did not account for the effect of larval behavior on dispersal or for the effects of fine-scale oceanographic processes. Thus, viewing larvae as passive dispersal agents may overestimate the actual dispersal of larvae and the connections among reefs. Recent models of connectivity in the Caribbean that account for larval behavior suggest that fish populations are less connected than assumed under passive dispersal models and that different regions of the Caribbean are essentially isolated from each other, at least on an ecological timescale. However, there are considerable differences among regions of the Caribbean as to the extent that reefs are connected to distant areas as some regions appear to import a large portion of recruits while other regions are primarily self-seeding. The differences in the relative importance of local and long-distance recruitment of juveniles among regions of the Caribbean underscores the role that careful planning will play in the implementation of marine reserves for protection of coral reefs as understanding how reefs are connected to one another will influence how large reserves should be and where they should be located.

Landscape Ecology of Coral Reefs: Connections of Coral Reefs to Mangrove and Seagrass Systems

Coral reefs are typically found in close proximity to other coastal ecosystems, particularly seagrass beds and mangrove forests. These different ecosystems are often connected to reefs via the movement of animals and nutrients across their boundaries. For example, carnivorous grunts (Haemulidae) forage in seagrass beds at night but school around large coral heads on reefs during the day as a refuge from predation. Coral heads that harbor fish schools receive nutrient supplements from fish excretion, grow up to 23% faster, and have more nitrogen and zooxanthellae per unit area than do corals without resident fishes. Thus, fishes that have no direct trophic link with corals collect nutrients from other ecosystems (seagrass beds) and concentrate these near their host coral. This facilitates coral growth, enhancing the coral's value as a refuge for these fishes and for other reef organisms.

Mangroves and seagrass beds also serve as nursery grounds and provide refuge from predators and an abundance of food for many juvenile fishes that are typically found on coral reefs as adults. Grunts (Haemulidae), snappers (Lutjanidae), barracuda (*Sphyraena barracuda*), and some parrotfishes (Scaridae) are particularly dependent on the presence of nearby mangroves. In Belize, reefs closely associated with mangroves have up to 26 times more biomass of some species of fish than reefs not associated with mangroves. A common species on these reefs, the bluestriped grunt (*Haemulon sciurus*), typically goes through an ontogenetic change in habitat use as it migrates from seagrass beds to mangroves to patch reefs to the forereef as it ages (**Fig. 7**). In areas where mangroves are absent, bluestriped grunts move from seagrass beds directly to patch reefs and are typically smaller than grunts that inhabit patch reefs with nearby mangroves. Thus, mangroves provide important habitat where juvenile grunts feed and increase in size before moving to patch reefs which may subsequently decrease the threat of predation once they move to these reefs. Further, the rainbow parrotfish (*Scarus guacamaia*), the largest herbivorous fish in the Caribbean, is functionally dependent on mangroves for shelter; juveniles of this species live primarily in mangroves, and the species goes locally extinct on reefs when nearby mangroves are removed (**Fig. 7**). Interestingly, density of fishes that have no direct link to mangroves at any stage of their life history can still be influenced by the proximity of mangroves, probably via interactions with mangrove-dependent fishes. Thus, the composition of the fish community on reefs is greatly influenced by the proximity of mangroves, and the rapid removal of mangrove forests from coastlines worldwide will certainly have drastic negative impacts on the ecology coral reefs.

Geographic Distribution of Coral Reefs

Coral reefs exist in tropical areas worldwide (**Fig. 8**). In general, reefs are abundant in areas with shallow coastlines and clear, warm water where riverine discharge of sediments is low. Large coral reefs are rarely found in areas above 29° latitude where ocean temperatures fall below 18 °C for extended periods as this slows coral growth and their capacity to build large reefs; however, zooxanthellate corals can be found in areas with water temperatures as low as 11 °C. In addition, herbivory is often less intense in cooler waters meaning that seaweeds are more abundant in temperate areas and that competition between corals and seaweeds is more intense. The combination of cooler water temperatures and more intense competition with abundant seaweeds likely interact to limit the latitudinal range of large coral reefs. However, when the physical and ecological criteria are met, the results can be phenomenal. For example, the most biologically diverse reefs occur in the tropical Indo-Pacific in the areas around Indonesia and the Philippines and house over 550 species of coral and thousands of species of fish. The Great Barrier Reef off northeastern Australia is the largest reef in the world with more than 2800 individual reefs occupying over 1800 km of the Australian coastline and can be seen from outer space.

Threats to Coral Reefs

Coral reefs are imperiled around the world because of the compounding effects of multiple stressors such as overfishing, pollution, climate change, and change in coastal land use. The decline of reefs is particularly evident in the Caribbean where coral cover has

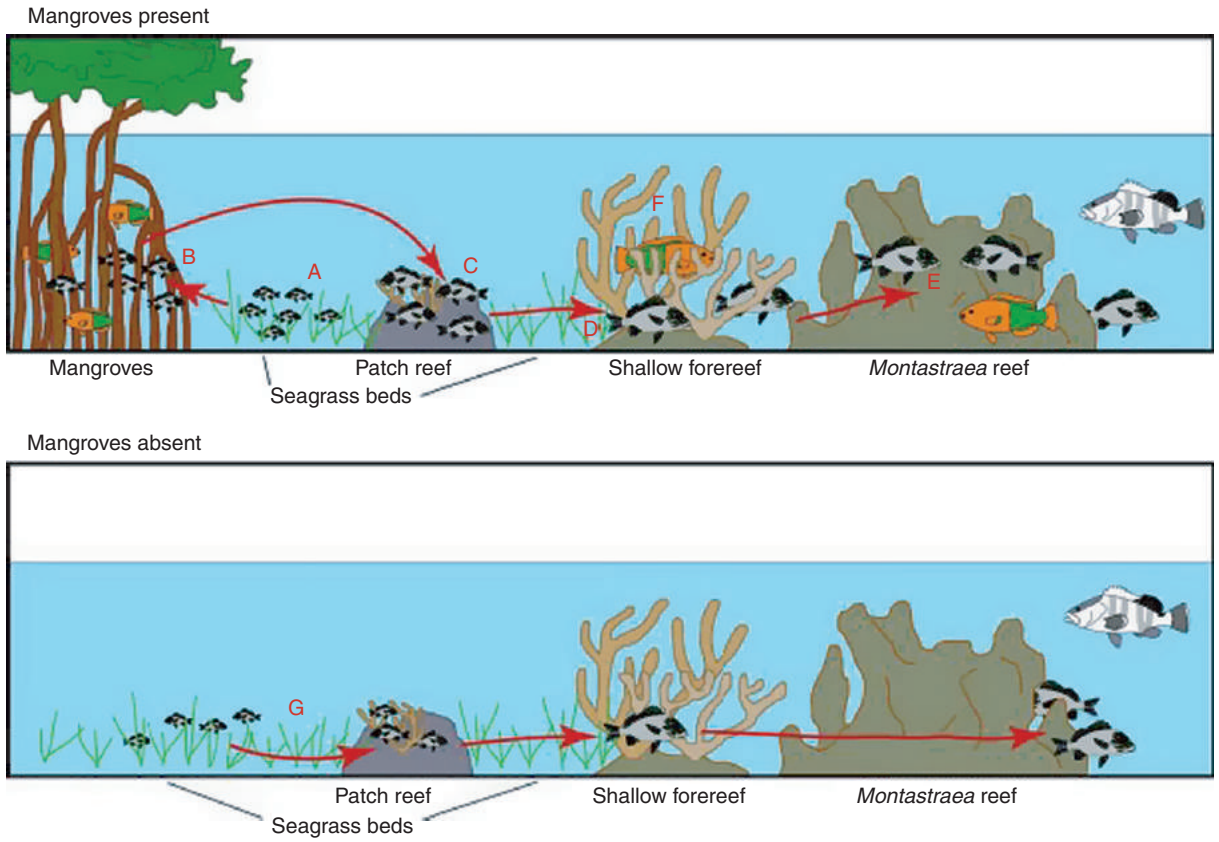


Fig. 7 Schematic illustrating the connection between mangroves and coral reefs. Ecosystem connectivity is stylized for *Haemulon sciurus* (gray and black fish) and *Scarus guacamaia* (orange and green fish) although other parrotfish (scarid), grunt (haemulid), and snapper (lutjanid) species also exhibited similar ontogenetic shifts in habitat use. *H. sciurus* show a substantial shift in size frequency from seagrass (A) to mangroves at approximately 6 cm. On reaching a given size in seagrass beds, juvenile fish then move to mangroves (B) which serve as an intermediate nursery habitat before migrating to patch reefs (C). If mangroves are not present, *H. sciurus* move directly from seagrass to patch reefs, appearing on patch reefs (G) at a smaller size and at lower density (260 ha^{-1} compared to 3925 ha^{-1} in mangrove-rich systems). In the presence of mangroves, the biomass of *H. sciurus* is significantly enhanced on patch reefs, shallow forereefs, and *Montastraea* reefs (C, D, E). *S. guacamaia* (F) has a functional dependency on mangroves and is not seen where mangroves are absent. Illustration describes findings from Mumby PJ, Edwards AJ, Arias-Gonzalez JE, *et al.* (2004) Mangroves enhance the biomass of coral reef fish communities in the Caribbean. *Nature* 427: 533–536. Schematic and description courtesy of Peter J. Mumby.

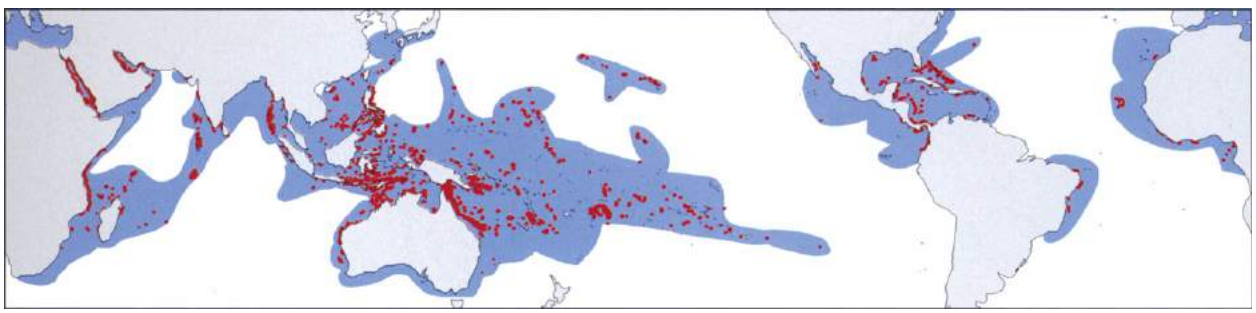


Fig. 8 Worldwide distribution of coral reefs. Coral reefs (red dots) cover roughly 250 000 km^2 of the Earth's surface, but zooxanthellate corals inhabit a far wider range (blue shading). From Veron JEN (2000) *Corals of the World,* vols. 1–3. Townsville, QLD: Australian Institute of Marine Science.

decreased by 80% in recent decades and may drop further as reefs fail to rebound following continued coral bleaching, overfishing, disease outbreaks, and other disturbances. The causes of coral reef decline are many and frequently act synergistically to drive coral reefs to alternate states such as seaweed-dominated reefs or sea urchin barrens (**Fig. 9**). We review the major threats to coral reefs and the role that marine reserves can play in stemming the tide of coral reef decline.

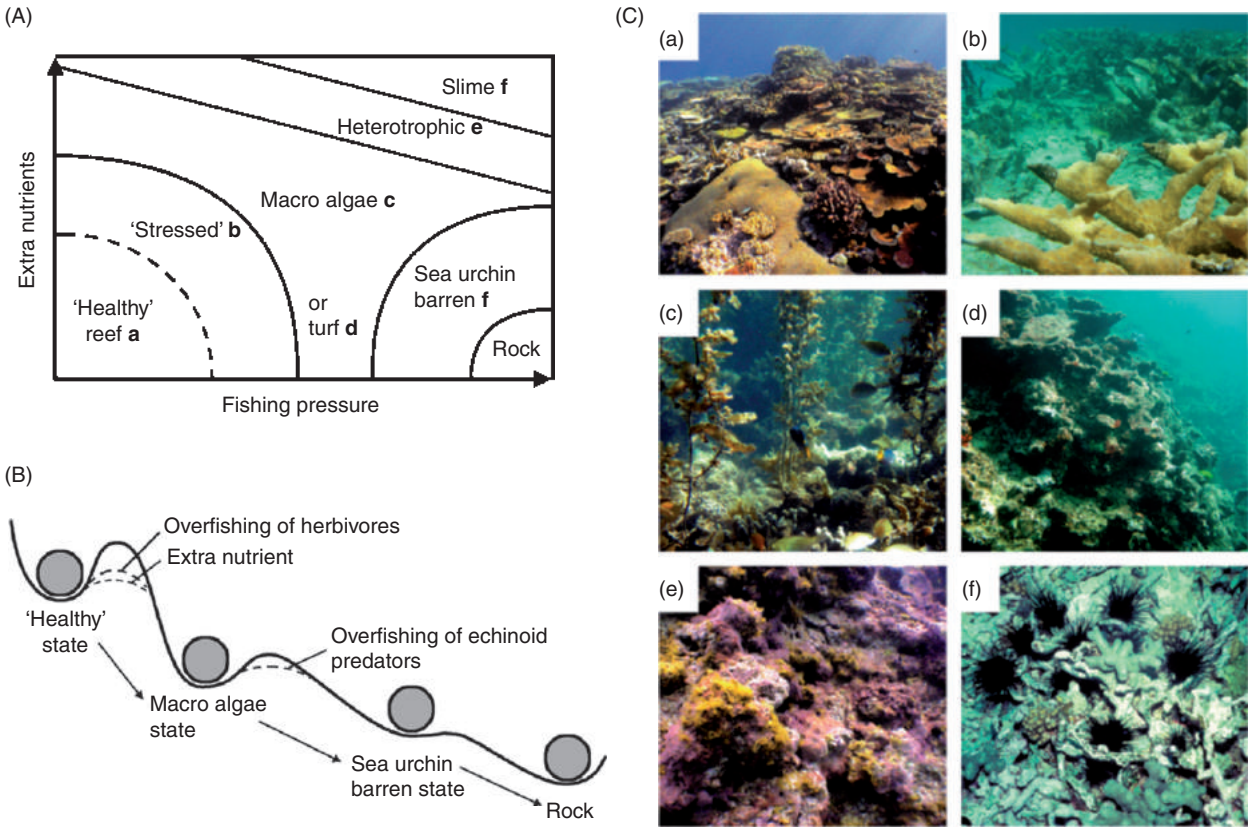


Fig. 9 Alternate states in coral reef ecosystems. (A) A conceptual model showing human-induced transitions between alternate ecosystem states based on empirical evidence of the effects from fishing and excess nutrients. The `stressed' state illustrates loss of resilience and increased vulnerability to phase-shifts. (B) A graphic model depicting transitions between ecosystem states. `Healthy' resilient coral-dominated reefs become progressively more vulnerable owing to fishing pressure, pollution, disease, and coral bleaching. The dotted lines illustrate the loss of resilience that becomes evident when reefs fail to recover from disturbance and slide into less-desirable states. (C) Pictoral representation of the different reef states shown in (A). From Bellwood DR, Hughes TP, Folke C, and Nystrom M (2004) Confronting the coral reef crisis. *Nature* 429: 827–833.

Coral Bleaching

Coral bleaching occurs when corals degrade or expel their dinoflagellate symbionts in response to environmental stressors such as elevated sea surface temperature and increased UV radiation. Although corals can reacquire symbionts and recover in weeks to months, recovered corals may grow slower and have reduced fecundity as compared to previously unbleached corals, giving bleaching-resistant corals an ecological advantage after bleaching events. In severe cases, bleaching may occur on the scale of hundreds to thousands of kilometers and radically alter coral cover and composition with coral mortality from bleaching events approaching 100% in extreme cases. Branching corals such as acroporid and pocilloporid corals are often more susceptible to bleaching and mortality than are massive corals, allowing the slower-growing massive corals to be more persistent on reefs after bouts of strong bleaching. Bleaching events not only decrease live coral cover but also provide large areas for seaweed colonization, and these seaweeds can prevent corals from reestablishing if herbivores are not present in sufficient numbers to suppress seaweed colonization and growth. Additionally, large-scale bleaching and mortality of branching corals can suppress fish populations that are dependent on live coral for shelter and food.

Analyses of coral bleaching on Caribbean reefs over the past two decades suggests that small increases in regional sea surface temperature (0.1 °C) result in large increases in the geographic extent and intensity of coral bleaching events. Given that climate change models suggest an increase in sea surface temperatures of 1–3 °C over the next 50–100 years, coral bleaching events may become an intense, annual stress on coral reefs throughout the Caribbean and even the world. Although corals may adapt and their bleaching thresholds may increase over time as sea surface temperatures rise, the threat of repeated, intense bleaching events over the next several decades is a significant concern. If even the conservative predictions of global climate models are realized, these climate changes could result in the fundamental reorganization of the ecology of coral reefs.

Disease and the Structure of Coral Reef Communities

The impact of diseases on coral reefs has been realized over only the past two decades. Two of the most extensive disease outbreaks have been on reefs in the Caribbean and have fundamentally changed the ecology of Caribbean reefs. In

1983–84 an unknown pathogen swept through the Caribbean and killed approximately 99% of the then abundant sea urchin *Diadema antillarum*. In many areas of the Caribbean, *D. antillarum* had been the dominant herbivore keeping reefs free of most fleshy seaweeds and facilitating recruitment and growth by corals. After the mass mortality, levels of herbivory plummeted and standing crop of seaweeds dramatically increased on many reefs. *D. antillarum* populations are recovering in some areas of the Caribbean, and in these `urchin zones', seaweeds cover 0–20% of the reef as opposed to 30–79% of the reef. Juveniles corals are ten times more abundant in some urchin zones. The potential recovery of this critical herbivore gives hope to Caribbean reefs many of which are still enveloped in a blanket of seaweed.

The other outbreak that altered the structure of Caribbean reefs was the epidemic of white band disease among acroporid corals in the mid to late 1980s. This disease attacked two of the major reef-building corals in the Caribbean *Acorpora palmata* and *A. cervicornis*. These two corals were once so abundant on Caribbean reefs that early coral reef ecologists named characteristic zones on reefs for these dominant corals (i.e., the `palmata zone' and the `cervicornis zone'). These corals that had dominated Caribbean reefs for at least a half million years are now rare to absent on most reefs and have declined so dramatically that they are both being listed as threatened species under the Endangered Species Act in the United States. If the prevalence and severity of coral diseases is linked to pollution and climate change as has been demonstrated for some studies, then a continued increase in the effects of diseases on the ecology of reefs can be expected.

Shifting Baselines, Overfishing, and the Altered Food Webs of Coral Reefs

In many regions of the world, coral reefs are mere remnants of what they were only a few decades ago. These changes to reefs are not adequately appreciated due to the problem of the `shifting baseline syndrome' – reefs that are deemed `normal' today are not what was `normal' only a few decades ago, much less a century or more ago. Each new generation of divers or marine ecologists suffers from reduced expectations of what a healthy coral reef should be. For example, as a graduate student and post-doc, one of us (M.E. Hay) dove on Caribbean reefs dominated by luxuriant stands of elkhorn and staghorn coral (*Acropora palmata* and *A. cerviconis*) the size of football fields and saw reefs abundant with grouper, large herbivorous fishes, and *Diadema* urchins that formed `fields' of gigantic black pincushions on regions of some reefs. In contrast, the younger author here (D.E. Burkepile) has never seen a live stand of elkhorn coral more than a few m^2 and is lucky to see one *Diadema* on most dives. However, both of us dive on reefs that are vastly different from those that the first European colonists would have experienced. Because of this problem of shifting baselines, it is informative for ecologists to explore the history and paleoecology of reefs in order to deduce how reef communities have changed over hundreds, or thousands, of years.

Caribbean reefs were once dominated by sea turtles, crocodiles, manatees, large predatory fishes such as sharks and large groupers, and the now-extinct monk seal. Reefs with such a diversity of charismatic megafauna scarcely exist today anywhere in the world. Centuries of overfishing have made many of these species ecologically extinct and altered the strong trophic interactions that once dominated Caribbean food webs (**Fig. 10**). Including humans into the ecological equation began a process of `fishing down the food web' where large consumers such as sharks and manatees were the primary targets of human harvesting. After the larger animals were depleted, fisheries switched to smaller predators such as groupers and

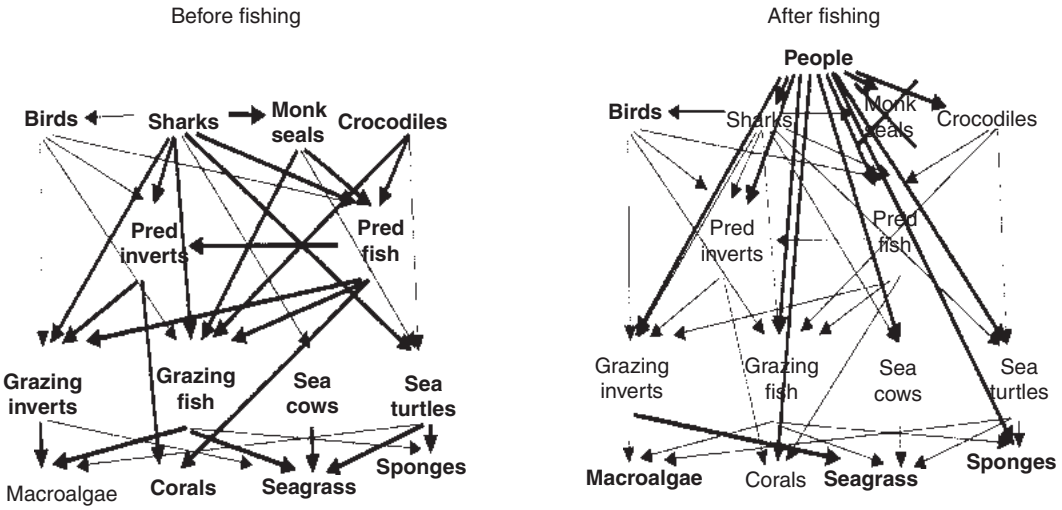


Fig. 10 Simplified coastal food web for coral reefs showing changes in some important top-down interactions due to overfishing; before (left side) and after (right side) fishing. Bold font represents abundant; normal font represents rare; `crossed-out' represents extinct. Thick arrows represent strong interactions; thin arrows represent weak interactions. Modified from Jackson JBC, Kirby MX, Berger WH, *et al.* (2001) Historical overfishing and the recent collapse of coastal ecosystems. *Science* 293: 629–638.

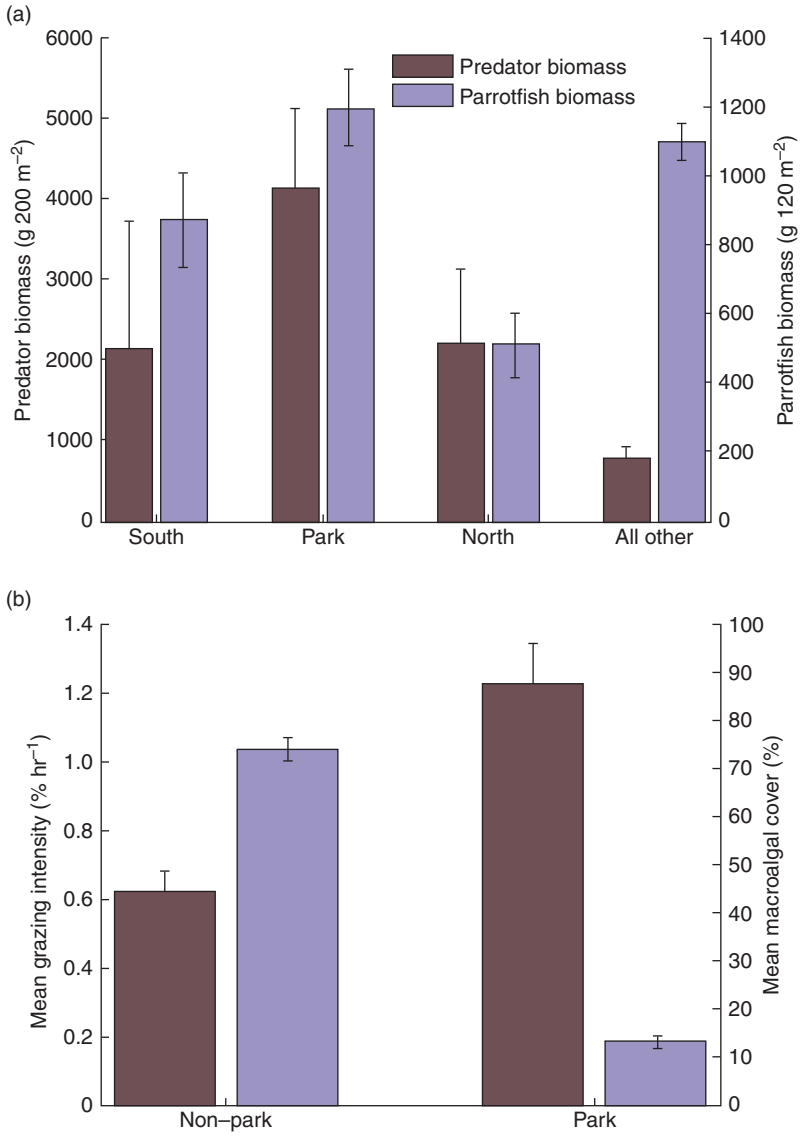


Fig. 11 (a) Patterns of parrotfish biomass and their predators (±SE) within the Exuma Cays (Bahamas) and for all other survey areas combined. `Park' denotes the Exuma Cays Land and Sea Park which is 456 km$^2$ and was established in 1959. `South' and `North' represent reef systems that are near the southern and northern borders of the park. (b) Mean macroalgal cover (gray bars) (±SE) and grazing intensity of parrotfishes (black bars) inside and outside the Exuma Cays Land and Sea Park. Reserve impacts are significant ($P<0.01$) for each variable. From Mumby PJ, Dahlgren CP, Harborne AR, *et al.* (2006) Fishing, trophic cascades, and the process of grazing on coral reefs. *Science* 311: 98–101.

then to herbivores such as parrotfishes. The changes in the connections of these food webs have fundamentally altered the dynamics of these ecosystems and have resulted in cascading effects such as the decline of corals and increase of seaweeds and sponges.

Although the largest megafauna are now largely gone from Caribbean reefs, we have some idea of their historical populations. For example, green turtles were once so abundant that ships' naturalists from the sixteenth to seventeenth centuries remarked that they could navigate to the Cayman Islands via the sounds of turtles swimming and that congregations of turtles seemed so thick as to confound a ship's path. One estimate puts the total pre-Columbian population of green turtles in the Caribbean at greater than 30 million as opposed to the tens of thousands today. Green turtles typically eat seagrasses and seaweeds, but the top-down force that this historical population would have exerted on seagrass production and biomass is unrivaled by any current estimates of herbivory in seagrass beds. Because the biota of coral reefs has changed so dramatically over the past few hundred years, Jeremy Jackson writes that scientists trying to understand the ecological processes that structure coral reefs are ``…trying to understand the ecology of the Serengeti by studying the termites and the locusts while ignoring the elephants and the wildebeest." Basically, the biotic forces that impact coral reefs today are mere shadows of what they once were, and humans have radically changed the ecological and evolutionary trajectories that have influenced coral reef ecosystems for millennia.

Protection and Resurrection of Coral Reefs

One of the saving graces of coral reefs over the next few decades may be the creation and enforcement of marine reserves that protect reefs from overfishing. Overfishing is one of the most devastating threats to reefs, as fishers preferentially remove the large-bodied fishes that are the strongest interactors in these ecosystems, resulting in fundamental changes to the food webs of reefs. The establishment of marine reserves limits or prevents the harvesting of fishes and invertebrates from areas of reef and theoretically allows populations of overharvested species to rebound, reestablishing viable populations of fishes and crucial ecosystem processes on reefs. Recent studies of the efficacy of marine reserves show that reducing fishing pressures on reefs allows increases in the density, biomass, individual size, and diversity of fishes and invertebrates inside the reserves and that these effects occur rapidly and are longlasting. In addition, these reserves not only allow increases in fish density and biomass within the protected areas but also result in the `spillover' of fishes as they migrate from the reserves into unprotected areas. Thus, marine reserves may subsidize fish populations on reefs that are not directly protected from fishing, although the extent to which this spillover effect will actually affect unmanaged reefs is equivocal.

Marine reserves can also restore trophic linkages that enhance the recovery of coral reefs. For some reefs in the Bahamas, long-term protection from fishing (i.e., roughly 50 years of enforcement) has led to increases in the abundance of medium-sized predatory fishes such as the Nassau grouper (*Epinephelus striatus*) (**Fig. 11**). Increases in grouper abundance resulted in increased predation rates on small herbivorous parrotfishes, which would seemingly decrease the rate of herbivory on reefs. However, the protection from fishing also allowed large parrotfishes to recover and actually increased the overall rate of herbivory in the reserve despite increased predation on smaller herbivores (**Fig. 11**). These increased rates of herbivory decreased macroalgal abundance and may increase coral abundance and cover over time if this balance between predation and herbivory can be maintained. Although the benefits of reserves to conservation and fisheries are promising, one of the main challenges to the success of marine reserves is the enforcement of no-harvesting policies once the reserve is established. In many areas, reserves are `paper parks' or parks in name only as there is insufficient money or political will to achieve the enforcement necessary for the reserves to succeed. However, if marine reserves can be implemented and enforced they will be one of the best tools that conservation science currently has to protect, and hopefully resurrect, many coral reefs.

Summary

Fossil evidence shows that corals have dominated many reefs for over 10 000 years. However, the balance between the ecological forces of predation, competition, disturbance, and recruitment that allowed thousands of years of uninterrupted reef growth have now been grossly altered by human activities. Consequently, healthy and dynamic reefs have declined dramatically in the last two decades as a result of overfishing, climate change, pollution, and other anthropogenic insults. Although the ecological future of reefs seems bleak, we hope that creative management of these ecosystems has the potential to protect them for future generations.

See also: Ecological Complexity: Goal Functions and Orientors. Ecological Processes: Biological Nitrogen Fixation. Evolutionary Ecology: r-Strategists/K-Strategists. General Ecology: Temperature Regulation

Further Reading

Bellwood, D.R., Hughes, T.P., Folke, C., Nystrom, M., 2004. Confronting the coral reef crisis. Nature 429, 827–833.
Birkeland, C. (Ed.), 1997. Life and Death of Coral Reefs. New York: Chapman and Hall.
Burkepile, D.E., Hay, M.E., 2006. Herbivore versus nutrient control of marine primary producers: Context-dependent effects. Ecology 87, 3128–3139.
Cowen, R.K., Paris, C.B., Srinivasan, A., 2006. Scaling and connectivity in marine populations. Science 311, 522–527.
Dulvy, N.K., Freckleton, R.P., Polunin, N.V.C., 2004. Coral reef cascade and the indirect effects of predator removal by exploitation. Ecology Letters 7, 410–416.
Gardner, T.A., Cote, I.M., Gill, J.A., Grant, A., Watkinson, A.R., 2003. Long-term region-wide declines in Caribbean corals. Science 301, 958–960.
Halpern, B.S., 2003. The impact of marine reserves: Do reserves work and does reserve size matter? Ecological Applications 13, S117–S137.
Hay, M.E., 1997. The ecology and evolution of seaweed–herbivore interactions on coral reefs. Coral Reefs 16, S67–S76.
Hughes, T.P., 1994. Catastrophes, phase shifts, and large-scale degradation of a Caribbean coral reef. Science 265, 1547–1551.
Hughes, T.P., Baird, A.H., Dinsdale, E.A., *et al.*, 2000. Supply-side ecology works both ways: the link between benthic adults, fecundity, and larval recruits. Ecology 81, 2241–2249.
Jackson, J.B.C., Kirby, M.X., Berger, W.H., *et al.*, 2001. Historical overfishing and the recent collapse of coastal ecosystems. Science 293, 629–638.
Knowlton, N., Rohwer, F., 2003. Multispecies microbial mutualisms on coral reefs: The host as a habitat. American Naturalist 162, S51–S62.
McClanahan, T.R., Mangi, S., 2000. Spillover of exploitable fishes from a marine park and its effect on the adjacent fishery. Ecological Applications 10, 1792–1805.
McCook, L.J., Jompa, J., Diaz-Pulido, G., 2001. Competition between corals and algae on coral reefs: A review of evidence and mechanisms. Coral Reefs 19, 400–417.
Mumby, P.J., Dahlgren, C.P., Harborne, A.R., *et al.*, 2006. Fishing, trophic cascades, and the process of grazing on coral reefs. Science 311, 98–101.
Mumby, P.J., Edwards, A.J., Arias-Gonzalez, J.E., *et al.*, 2004. Mangroves enhance the biomass of coral reef fish communities in the Caribbean. Nature 427, 533–536.
Veron, J.E.N., 2000. Corals of the World, vols. 1–3. Townsville, QLD: Australian Institute of Marine Science.

Desert Streams

TK Harms, RA Sponseller, and NB Grimm, Arizona State University, Tempe, AZ, USA

Distribution and Physical Template

Desert streams occupy arid and semiarid regions defined by low annual precipitation. Semiarid and arid climate zones are found on all continents and include both hot and cold deserts. Although the range of temperatures varies across desert regions, summer temperatures may exceed 40 °C in hot deserts. Annual precipitation ranges from <100 to 300 mm yr^{-1} and combined with high temperature can result in high rates of evapotranspiration. Higher precipitation in the mountains (up to ~ 1000 mm yr^{-1}) can feed streamflow in the low deserts, often supporting perennial flows in large basins. Arid and semiarid regions are characterized by distinct seasons defined by precipitation and/or snowmelt and the amount of precipitation that falls during these seasons shows high interannual variability. This results in extreme seasonal and interannual variation in stream discharge. Indeed, streams in some desert regions flow in response to rain events that occur only once in several years or even less.

Arid and semiarid lands account for over one-third of global lands, making desert streams prominent among aquatic ecosystems. The large geographic area covered by deserts results in a wide variation in temperature and precipitation regimes as well as in geomorphology. Thus, the hydrogeomorphic templates and resulting ecological characteristics of desert streams exhibit a great diversity of patterns. Despite this extensive distribution of desert streams, the vast majority of ecological studies of desert streams have occurred in the southwestern United States, Australia, and Antarctica. Our discussion thus draws from results of studies in these ecosystems. Future studies of desert stream ecosystems in other regions are likely to add new dimensions to the state of our understanding presented here.

Desert stream hydrographs are punctuated by events when discharge may exceed baseflow by several orders of magnitude. Precipitation falling on the catchment rapidly reaches the stream and stream discharge rapidly dissipates following floods. Infiltration of desert soils is minimal and, at the scale of whole basins, much of the water that is not returned to the atmosphere by evapotranspiration reaches streams via overland flow during storms or via infiltration of permeable low-order channel sediments followed by subsurface flow. Resulting flash floods scour the streambed, resulting in downstream export of sediments and aquatic organisms and creating a wide channel. Large floods also deposit alluvial materials in riparian zones and may remove riparian vegetation. These effects vary depending on the scale of the event (see the section titled `Temporal dynamics').

The boundaries of a stream ecosystem in any climate region extend beyond the wetted channel and comprise a stream–riparian corridor (**Fig. 1**). The aquatic ecosystem encompasses surface water as well as the alluvial sediments beneath the streambed where surface and groundwater mix, termed the hyporheic zone. The parafluvial zone is defined by the region of the active channel over which water flows only during floods and in desert streams this region can be much wider than the stream itself. Finally, the riparian zone is the land area surrounding the stream that is significantly influenced by the stream. The availability of water contrasts starkly among these subsystems in desert streams making each subsystem more distinct than in mesic streams. In deserts, the hyporheic zone often contains water and sustains biological activity in the subsurface even in the absence of surface flow. The parafluvial zone contains surface water only during floods and flow quickly recharges the subsurface through coarse sediments or gravel in this zone, leading to short periods of surface flow but sustained subsurface flow. The riparian zone contrasts starkly with desert uplands owing to the presence of shallow groundwater that is accessible to plants.

Desert streams may contain sections of both gaining and losing hydrologic templates. Gaining sections of streams are those in which the water table is sloped toward the stream channel such that groundwater discharges into the surface stream. Losing reaches are characterized by a water table that slopes away from the stream causing surface flow to recharge groundwater. Along losing reaches, the predominate direction of water flow in desert catchments is from the uplands directly into the stream before recharging riparian groundwater; whereas along gaining reaches, water flowing overland into the riparian zone recharges groundwater there before discharging into the surface stream. These contrasting hydrologic templates can have significant effects on nutrient dynamics, water storage, and biota within stream–riparian corridors. Losing reaches, for example, often have no surface flow during dry seasons, whereas gaining reaches are a more permanent source of surface water. Due to permeable sediments, interactions between surface and subsurface water are dynamic. For example, water moves into the hyporheic zone in regions of downwelling and from the hyporheic zone to the surface in regions of upwelling. Water within the hyporheic zone moves through the interstitial spaces of sediments slowing water velocity compared to the surface stream. Such patterns in hydrologic flows have important implications for nutrient cycling and stream productivity.

Temporal Dynamics

Desert streams are highly variable over time, at a range of temporal scales. In addition to the seasonality that typifies many streams, temporal dynamics of desert streams are strongly influenced by disturbances at two extremes (flash flooding and drying) of a

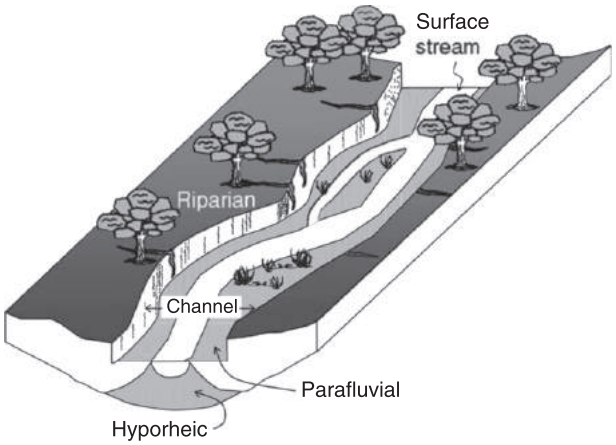


Fig. 1 Schematic drawing of a desert stream–riparian corridor.

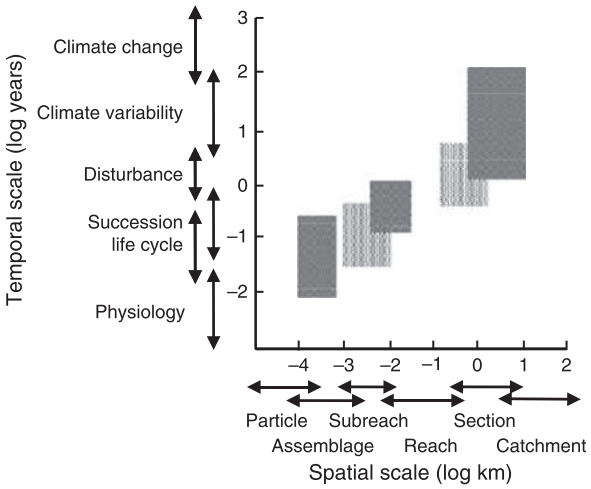


Fig. 2 Temporal scales considered in this section (ordinate) are correlated with spatial scales (abscissa), and each phenomenon discussed is associated with a characteristic range of time and space scales.

hydrologic spectrum. Researchers have largely focused on temporal dynamics at decadal and lower scales; however, decade- to century-scale channel change establishes a geomorphic template upon which these higher-frequency dynamics play out. Our discussion will consider temporal change from low- to high-frequency events (Fig. 2).

The concept of disturbance has various meanings, but in stream ecology disturbance is usually associated with hydrologic extremes that change ecosystem structure and processes. Using terms from disturbance ecology, we can characterize a disturbance regime, which has features such as interannual (or interdecadal) variability, seasonality, and timing, frequency, and magnitude of individual events. Disturbance is intimately connected to succession, which is most simply defined as the change in ecosystem properties on a site following disturbance. Ecosystem components that are affected by a disturbance are those that undergo succession after the disturbance; for example, a flash flood that removes algae and invertebrates but does not affect streamside vegetation initiates succession in the stream but not in the riparian zone.

Successional patterns depict the temporal changes in stream and riparian communities and processes that are superimposed upon a larger temporal scale of variability. For longer-lived riparian vegetation, successional patterns and time frames may be similar to those of terrestrial communities but for stream biota, succession often plays out against a seasonal backdrop. Thus, successional patterns differ between seasons, with faster increases in biomass during warmer months. Successional patterns also vary depending upon the size and nature of the initiating disturbance as well as antecedent conditions (which are themselves influenced by the disturbance regime – timing and clustering of individual events). Whereas disturbances that occur over short timescales may produce predictable recovery sequences, biota recover from longer-term, infrequent disturbances with less regularity. Effects on biota of pronounced interannual variation that characterizes deserts include shifts in community composition of invertebrates and differences in the relative importance of nitrogen-fixing cyanobacteria versus nonfixer algae in the primary producer assemblages.

Flash flooding and drying are the primary disturbances that characterize desert streams. Flood magnitude is usually described by the peak discharge, but other aspects of a flash flood hydrograph – the steepness of its rise and the length of its tail – also determine flood effects. Floods are important geomorphic agents, shaping channel form, as well as disturbances that initiate biotic succession. In deserts, floods connect elements of the landscape, from ridgetops to large rivers to groundwater, which are otherwise isolated and disconnected. Drying is at the opposite hydrologic extreme but is more difficult to treat as a discrete disturbance because it represents a protracted reduction and ultimately loss of stream flow. As drying progresses, there is first a concentration of mobile biota that precedes a concentration of dissolved materials (through evaporation); there may be isolation of sections of a stream and distinctive patterns of surface-water loss; direction of surface–subsurface water exchange may flip; organisms may move into sediments; and, ultimately, surface flow is lost entirely. Drying ends when surface flow resumes, either during a flood or as a gradual increase in discharge.

At the scale of centuries, events that occur only every 50–100 years or so can shape channel form and initiate riparian succession. For example, in the southwestern USA, a period of erosion occurred forming arroyos or gullies and draining the riverine wetlands that were once characteristic of these desert environments. This period left a geomorphic structure that persists today in many southwestern river–riparian ecosystems. Dramatic changes such as these can affect groundwater–surface water interactions and change species composition of the riparian vegetation. Indeed, such large-scale changes have repercussions for many stream characteristics, underscoring the importance of the hydrogeomorphic template in establishing structure and function of stream–riparian ecosystems.

Decadal variability resulting in relatively wet and dry periods in the southwestern USA is related to quasi-cycles of the Pacific Decadal Oscillation and El Niño Southern Oscillation (ENSO). For the southwestern USA, a strong ENSO signal is seen in decadal patterns of winter runoff from the Puerco and Grande Rivers in New Mexico and Sycamore Creek in Arizona. During wetter periods, frequent high-discharge events remove active-channel vegetation, leaving open gravel bars (parafluvial zone). Although these particular characteristics may be unique to desert streams of the southwestern USA, the important point is that larger-scale forcing from global climatic patterns can result in decadal shifts in near-stream riparian vegetation that have profound consequences for stream ecosystem function.

Given the high degree of interannual variability of desert environments, annual averages often carry little information and long-term trends are masked. Years vary not only in the total amount of runoff but also in the temporal distribution and timing of individual events or clusters of events. During the five wettest years of a 30-year period for Sycamore Creek, frequent floods meant the ecosystem was in an early successional state most of the year, whereas stream organisms experienced severe drying conditions over much of the year during the five driest years. Furthermore, years with the same total annual runoff may differ substantially in seasonal distribution of that runoff with consequences for the seasonal patterns of drying. A single, large late-summer flood in 1970 in Sycamore Creek carried the same total volume of water as nine total floods distributed more evenly across the spring season in 1988, with the result that much of the stream was dry during the hottest months in 1970 but was undergoing succession during summer 1988 (**Fig. 3**).

Although seasonality in desert streams may be strongly dependent upon the distribution of disturbance events, other variables in addition to discharge, such as temperature and day length, can influence the biota of desert streams. Deserts are defined only by low precipitation; thus, there is a broad range across the world's deserts in both flow seasonality and annual temperature

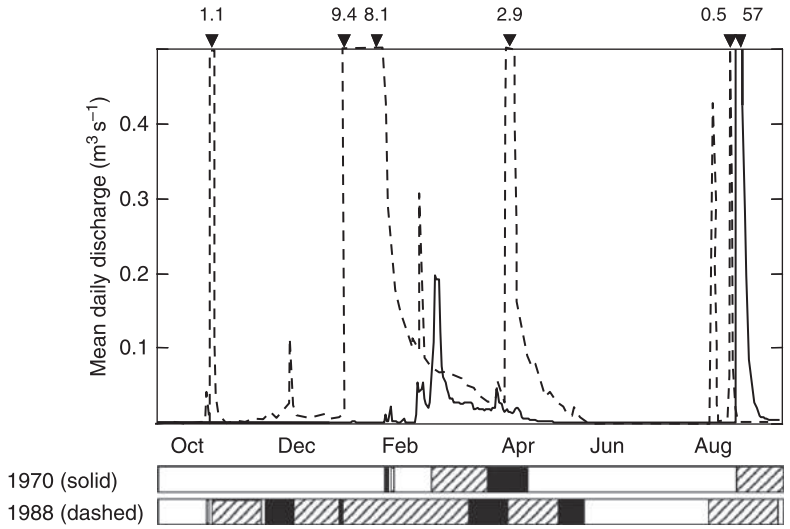


Fig. 3 Contrasting seasonal discharge patterns in 2 years with nearly identical total annual discharge. Bars at bottom show time periods likely to be influenced by postflood succession (hatched), drying (open), or neither (solid). Redrawn from Grimm, N.B., 1993. Implications of climate change for stream communities. In: Kareiva, P., Kingsolver, J., Huey, R. (Eds.), Biotic interactions and Global Change. Sunderland,MA: Sinauer Associates.

distributions. Flow seasonality may vary from highly unpredictable and episodic events to a relatively predictable and sustained increased in discharge associated with a distinct wet season, with consequences for successional patterns. Temperatures of stream water in cool deserts may fluctuate seasonally from near-freezing to 20 °C; in hot deserts, daytime stream temperature can reach >30 °C but may be ameliorated by extensive evaporative cooling.

Temperature variation over the course of a single 24-h period may be nearly as great as seasonal variation. High albedo and low heat capacity of desert land surfaces cause extensive diel fluctuations in air temperatures, leading to wide (though comparatively muted) ranges in stream temperature. Particularly in summer when evapotranspiration rates are very high, streamflow varies measurably over 24 h, causing stranding at stream margins. At points where drying streams sink into the sediments, the end of the stream can migrate up and down the channel by several meters! Desert streams of Antarctica show extreme diel variation in discharge, but by a very different mechanism. Streamflow is generated by solar melting of the vertical walls (ice cliffs) of glaciers. During summer, when the Sun circles around the horizon at a low angle, melting (and streamflow) stops when the cliffs are in shadow.

Biota

Desert stream ecosystems support a diverse assemblage of riparian plants and stream biota. A unifying characteristic of desert stream organisms is the shared evolutionary history in a hydrologically extreme environment. The consequences of this extreme physical template are evident from the variety of adaptations that allow species to thrive in systems prone to flash flooding and prolonged drought. Rather than providing a list of taxonomic names for each group, we place emphasis on life history, behavioral, and morphological adaptations for living in hydrologically variable ecosystems.

Desert stream ecosystems house diverse periphyton communities, which include a variety of filamentous green algae, epilithic, epiphytic, and episammic diatoms (attached to rocks, plants, and sediments, respectively), and nitrogen-fixing cyanobacteria. Both flash flooding and prolonged drought decimate algal biomass in desert streams. Rapid drying is particularly lethal, and algae typically die within hours of exposure to the hot, dry desert environment. Algal species often have physiological adaptations that allow for some resistance to drying, however, and can withstand periods of gradual drying. Such adaptations include the production of extracellular mucilage that increases cellular water retention, and intracellular osmoregulatory solutes that also prevent water loss in drying sediments. In addition to these mechanisms, at the onset of drying, algae may also produce spores, cysts, or zygotes that can reactivate upon rewetting. Benthic algae rapidly recolonize stream sediments following floods, whereas recovery following drought is variable and depends on the degree and modes of drought resistance. In Antarctic desert streams, for example, glacial melt is the primary source of streamflow, and primary producers (cyanobacterial and other microbial assemblages) are activated by higher temperature and renewed flows which may occur seasonally or even on a diel basis. However, these organisms are also able to persist for decades in the absence of liquid water.

A productive and diverse invertebrate fauna characterizes many desert streams, consisting of insect and crustacean taxa residing in both benthic and hyporheic habitats. Life-history characteristics of desert stream invertebrates reflect an evolutionary history in a hydrologically variable ecosystem, and are shaped by both flooding and drying disturbances (**Table 1**). Most stream invertebrate larvae have few mechanisms that confer resistance to either type of hydrologic disturbance. Instead, many species have short developmental times (e.g., 1–3 weeks) that increase the probability of offspring surviving to reproductive maturity in ephemeral environments, and ensure that some aerial adults are available for recolonization following floods or upon rewetting previously dry channels. In addition, organisms with longer life cycles exhibit an array of avoidance behaviors to minimize the effects of flooding and drying disturbance. These include timing reproductive activity to periods of low flood probability, as well as ovipositing eggs in sections of stream that are likely to retain water for longer periods of time (e.g., deep pools, riffles). Finally, air breathing insects (e.g., coleopterans, hemipterans) may exhibit more direct avoidance behaviors, including the use of rainfall as a cue for leaving aquatic habitats before floods.

Relative to mesic counterparts, fish assemblages of arid river systems are species poor, and are composed of taxa that also have specific adaptations to life in hydrologically variable systems. These adaptations include large reproductive efforts, multiple clutches per year, and short developmental times. Such life-history features, along with the ability to migrate long distances during periods of sufficient flow, allow native desert fish to rapidly colonize habitats after disturbances, and result in dramatic temporal fluctuations in population size. In addition, while intense flash floods can decimate fish populations, many desert fish have morphological adaptations that allow for some resistance to high flows. These include depressed skulls, keeled or humped napes, buttressed fins, narrow caudal peduncles, slim bodies, and reduced scales – all of which act to reduce drag and improve swimming ability in turbulent flow.

As desert stream ecosystems contract during drought, fish become isolated in pools where the physical environment can fluctuate dramatically. Although complete water loss is lethal, as streams contract, individuals of many fish species can survive in small pools, as well as beneath logs, stones, and within beds of algae. As a consequence, native desert fish are able to tolerate a broad range of temperatures (7–37 °C); indeed, desert pupfish of western North America can survive in temperatures that exceed 40 °C. Similarly, most desert fish are able to tolerate high salinity and low dissolved oxygen concentration. Others still, like the African lungfish, can burrow into the stream substrate during dry periods and survive for months by breathing atmospheric air with primitive lungs.

Table 1 Invertebrate colonization/recolonization characteristics of desert streams in relation to floods in different physiographic regions

	Mesic	*Hot desert*	*Endorheic cold desert*	*Exorheic cold desert*	*Chapparal*	*Glacial*
Colonization sources[a]	Numerous	Few	Intermediate	Intermediate	Intermediate	Few
Colonization distances[b]	Close	Far	Intermediate	Intermediate	Intermediate/far	Far
Pathways[c]	DD, um, S, O, H	DD, um, S, O	S	dd, um, S, o, h	DD, um, S, O, h	dd, um, d, O, h
Refugia	Abundant	Limited	Limited	Intermediate	Intermediate	Limited
Species diversity	High	Intermediate/high	Low	Intermediate	Intermediate/high	Low
Resilience[d]	High	High	Low	High	High	Unknown
Flood occurrence	Spring/summer	Winter/summer	Winter	Winter/spring/summer	Winter	Spring/summer
Spatial extent of flood	Extensive	Variable	Extensive	Extensive	Extensive	Extensive
Severity of flood	Intermediate	High	High	Intermediate	High	Intermediate

[a]Refers to sources separate from the perturbed stream.

[b]Refers to distance from other unaffected water bodies.

[c]Status at time of spate: DD/dd, downstream drift; um, upstream migration; S/s, survivors; O/o, oviposition; H/h, hyporheic. Upper and lower case indicates major or lesser importance, respectively.

[d]Refers to number of taxa, not individuals, following recovery.

Reproduced from Cushing CE and Gaines WL (1989) Thoughts on recolonization of endorheic cold desert spring-streams. *Journal of the North American Benthological Society* 8: 277–287.

Streamside forests, or riparian zones, stand out as hot spots for aboveground primary productivity in arid landscapes. Arid riparian zones include assemblages of phreatophytic deciduous trees capable of accessing groundwater, as well as shrubs and annual grasses. The overall taxonomic composition of riparian zones is typically in striking contrast to that of the surrounding desert landscape. Deeply rooted riparian trees are well suited to an environment where the water table is temporally variable. Obligate wetland species appear in desert riparian areas with permanent access to shallow groundwater, whereas those found in areas with strong seasonal fluctuations in water table elevation have structures such as tap roots or root architecture that maximizes water capture during precipitation events. Many riparian tree species in arid landscapes actually require over-bank flooding at specific times of the year to induce germination. Riparian vegetation is thought to play an important role in the overall cycling of nutrients in arid landscapes by taking up nutrients present in shallow groundwater and building organic matter pools in riparian soils. Finally, riparian vegetation serves as critical habitat for invertebrate, vertebrate, and avian taxa within arid landscapes.

Energetics

In contrast with streams of temperate and tropical biomes and because of their flood-shaped channel morphology, desert streams generally are not shaded by adjacent riparian vegetation. As a consequence, incidence of photosynthetically active radiation (PAR) reaching desert streams is high, and rates of instream primary production, the process by which energy is captured and organic matter is produced in ecosystems, are among the highest documented for benthic algae. The accrual of algal biomass in turn represents the energetic basis for stream food webs, and is central to the overall ecosystem dynamics of arid streams. For example, the abundance of high-quality benthic algae, together with warm temperatures and selection for rapid growth, result in among the highest rates of secondary production reported for benthic invertebrates. Moreover, owing to high standing stocks and growth rates, invertebrates play an important role in organic matter dynamics and nutrient cycling in desert streams. Indeed, the quantity of organic matter ingested by stream invertebrates can be 2–6 times greater than primary production. Finally, the emergence of desert stream insects represents an important resource for predators in adjacent terrestrial habitats.

At the ecosystem level, high rates of algal productivity in desert streams set them apart from streams of forested regions with respect to the relative rates of production (P) and respiration (R). Specifically, desert streams are often autotrophic ($P > R$). This is in striking contrast to streams of other biomes that receive the bulk of organic matter from outside the stream ecosystem and are often highly heterotrophic ($P << R$). Productivity of desert streams is also influenced by the disturbance regime. Flash floods scour stream channels, decimate existing organisms, and initiate a suite of algal and macroinvertebrate successional processes that correspond to temporal changes in photosynthesis and respiration (**Fig. 4**). Post-flood recovery of heterotrophs is enhanced by availability of organic matter that was stranded or deposited on the stream margins and in the riparian zone during dry periods.

In addition to metabolic changes associated with flash floods, spatial patterns of photosynthesis and respiration and post-flood successional dynamics are further influenced by hydrologic exchange between hyporheic and parafluvial subsystems and the surface stream. Specifically, rates of photosynthesis and the speed of post-flood recovery are greatest where nutrient-rich water from hyporheic and parafluvial sediments enters the surface stream. Conversely, rates of respiration are greatest where oxygen and organic matter from the surface stream enters subsurface and lateral sediments (**Fig. 5**). When both surface and hyporheic processes are taken into account, desert streams may more closely approximate a balanced metabolism ($P = R$), which highlights the connection between the two subsystems.

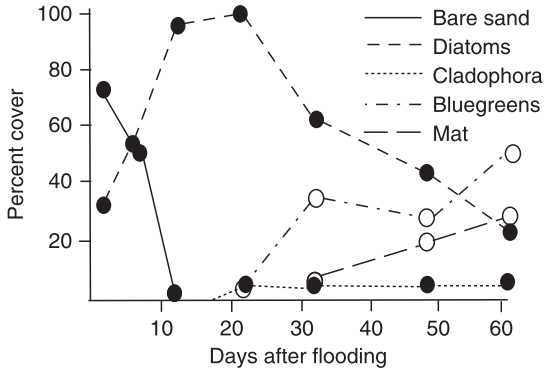


Fig. 4 Recolonization of Sycamore Creek, AZ by primary producers. Redrawn from Fisher, S.G., Gray, L.J., Grimm, N.B., Busch, D.E., 1982. Temporal succession in a desert stream ecosystem following a flash flood. Ecological Monographs 52, 93–110.

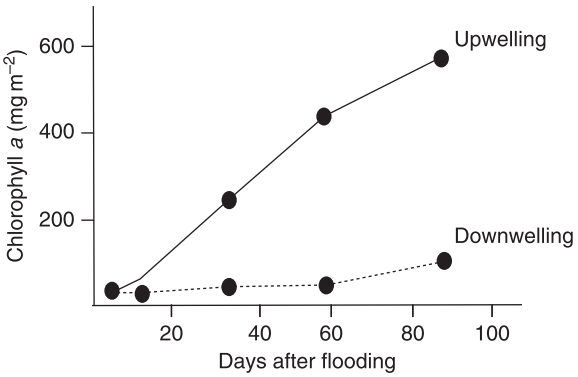


Fig. 5 Comparison of algal colonization in zones of upwelling and downwelling following floods in Sycamore Creek, AZ. Valet, H.M., Fisher, S.G., Grimm, N.P., Camill, P., 1994. Vertical hydrologic exchange and ecological stability of a desert stream ecosystem. Ecology 75 (2), 548–560.

Nutrient Dynamics

Various factors can limit the rate of primary production if demand (requirements of autotrophs) exceeds availability. Limiting factors in streams are typically light and nutrients. Because desert streams often have open canopies and receive abundant light, nutrients are the primary constraint on algal growth. Lack of precipitation in arid and semiarid regions leads to very slow rates of weathering of parent materials, which can lead to phosphorus (P) limitation as rocks are the ultimate source of P in ecosystems. However, in many well-studied arid and semiarid watersheds of the US southwest, volcanic-derived parent materials yield highly dissolved P. Primary production in these desert streams is thus limited by nitrogen (N).

Nutrients enter streams via inputs from upstream, from groundwater and overland flow, in plant materials deposited from the riparian zone, and in the case of N, via fixation of atmospheric N_2 by cyanobacteria. Unidirectional flow of water results in continual input and output of nutrients in dissolved and particulate forms, although inputs of limiting nutrients may be low due to processing that occurred upstream. Nutrient spiraling theory, a set of hypotheses that describe how nutrients move between water column, subsurface, and biotic compartments while being transported downstream, predicts that nutrient uptake should be more efficient under conditions of nutrient limitation. In streams limited by N, for example, inorganic N is rapidly removed from the water column by biota. For hot desert streams, rates of nutrient uptake can be particularly rapid due to high temperatures and light availability, which increase rates of biological reactions. Concurrent with rapid uptake by algae, excretion of inorganic N by invertebrate consumers can represent nearly 30% of the total N delivered to the ecosystem. Over successional time, export of rafting algal mats or stranding of algae on the stream banks during dry periods results in loss of organic N from the stream ecosystem. Budgets of organic N for desert streams thus conform with the successional trajectory of terrestrial ecosystems wherein ecosystems at late successional stages tend to lose nutrients. In contrast to terrestrial ecosystems, however, net primary productivity may continue to be positive during late successional stages of desert streams, resulting in continued uptake of inorganic nutrients by primary producers.

In surface water, nutrient cycling is dominated by uptake of nutrients by algae and benthic biofilms. The dominant pathway of nutrients in the surface is therefore from inorganic to organic forms. Regeneration (mineralization) of inorganic nutrients in the subsurface may in turn resupply dissolved inorganic nutrients. Processes occurring in the hyporheic and parafluvial zones thus contribute strongly to patterns of nutrient availability in desert streams. Water flowing through sediments slows in velocity

allowing for greater interactions between sediment surfaces and materials delivered in water. Microbes inhabiting the interstitial areas of sediments transform nutrients present in these downwelling zones, influencing the spatial distribution of nutrients in the stream channel.

In coarse sediments where dissolved oxygen remains relatively high as water moves through the hyporheic zone, mineralization often dominates N transformations, resulting in a localized increase in streamwater dissolved inorganic N concentrations at locations of upwelling. Increased nutrient availability in zones of upwelling is often associated with hot spots of algal biomass. These patterns are typical of alluvium-dominated reaches where algae are the predominant primary producers. In patches where macrophytes colonize gravel bars and parafluvial zones or in patches of fine sediment deposition, dissolved oxygen concentration in the subsurface is decreased due to root respiration and decomposition of plant-derived organic matter, and hyporheic flows are slowed, all of which lead to hypoxic or anoxic conditions. Hot spots of denitrification are associated with anoxic conditions in the hyporheic zone and water upwelling downstream of such patches is therefore depleted of inorganic N (**Fig. 6**).

Because of these differences in nutrient processing between surface and subsurface flowpaths, streams that undergo drying may exhibit marked spatial variability in nutrient availability. Sections of the stream that dry may continue to harbor subsurface flows and rapid transformation of nutrients for some time after surface flows are depleted. Nutrient inputs and outputs from dry reaches may show strong contrasts in forms or concentration of nutrients. In contrast, reaches characterized by perennial flow tend to show dampened upstream–downstream contrasts due to the homogenizing effects of processes occurring in surface flows.

As with nearly all aquatic ecosystems, the surrounding terrestrial landscape influences nutrient dynamics in desert streams. In deserts, however, hydrologic connectivity between the stream and terrestrial portions of the catchment, including the riparian zone, are variable in time. Deposited nutrients and those stored by plants and microbes may accumulate in the riparian zone and uplands during dry periods. When precipitation or snowmelt events occur, water carries these particulate and dissolved nutrients overland from the uplands to the stream, and between the riparian subsurface and the surface stream. This creates pulses of nutrient transport between desert streams and their watersheds. Pulsed inputs of nutrients result in hot moments of nutrient processing, short time periods with rapid rates of nutrient transformations. Hot moments may account for a significant fraction of annual nutrient processing within riparian zones of desert streams.

Connectivity between terrestrial and aquatic portions of desert stream–riparian corridors may also occur from the stream to the riparian zone. Riparian plants can access water and nutrients from the hyporheic zone as well as from shallow groundwater. Access to these more permanent sources of water and nutrients leads to high productivity in riparian zones relative to desert uplands. Stream biota may transfer nutrients between streams and riparian zones of desert streams. Owing to high rates of primary

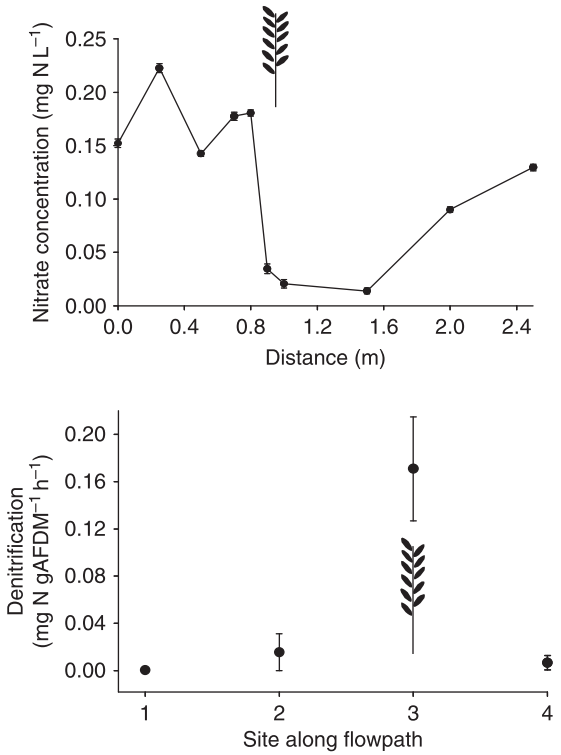


Fig. 6 Changes in concentration of nitrate in water flowing through plant-colonized gravel bars (top). Note precipitous drop in concentration when water encounters a plant patch (denoted by branch figure). Denitrification is a likely mechanism accounting for the drop in concentration of nitrate; *in situ* rates of denitrification increase in plant patches (bottom). Redrawn from Schade, J.D., Fisher, S.G., Grimm, N.B., Seddon, J.A., 2001. The influence of a riparian shrub on nitrogen cycling in a Sonoran desert stream. Ecology 82, 3363–3376.

productivity, insect emergence from desert streams can result in significant exports of nutrients out of the wetted stream. Emerging aquatic insects may thus provide a significant source of nutrients to riparian food webs.

Human Modifications

Desert streams present important challenges to understanding and management of water resources. They exemplify the resource that is most precious to humans inhabiting arid and semiarid regions, yet they are threatened by increasing pressures of human exploitation, agricultural expansion, and urbanization. Direct appropriation of streamflow to support human activities is the most serious threat to desert streams. This takes the form of diversion, interbasin water transfer, and groundwater withdrawal (which reduces baseflow); for example, groundwater withdrawals over the past century have converted the Santa Cruz River in Tucson, Arizona from a perennial to an ephemeral stream. In the Salt River of central Arizona, river diversion into a system of canals that feed agricultural and domestic/industrial demand in Phoenix has left a dry riverbed throughout the metropolitan region. Water extraction, primarily for irrigation, has also resulted in salinization of streams in much of the world and has triggered shifts in the composition of biotic communities.

In their appropriation of water for a variety of uses, people also modify the form and hence the function of desert streams. For example, the creation of canals that are straightened and lined with concrete replaces structurally complex streams with ecosystems that are unlikely to support the ecosystem functions characteristic of unmodified desert streams. Furthermore, impoundment and flow regulation can have profound effects on riparian ecosystems, for example, through the colonization and persistence of exotic plant species that outcompete native species under conditions of lowered water tables and reduced flow variability.

See also: Ecological Complexity: Cybernetics

Further Reading

Boulton, A.J., Peterson, C.G., Grimm, N.B., Fisher, S.G., 1992. Stability of an aquatic macroinvertebrate community in a multi-year hydrologic disturbance regime. Ecology 73, 2192–2207.

Cushing, C.E., Gaines, W.L., 1989. Thoughts on recolonization of endorheic cold desert spring-streams. Journal of the North American Benthological Society 8, 277–287.

Fisher, S.G., Gray, L.J., Grimm, N.B., Busch, D.E., 1982. Temporal succession in a desert stream ecosystem following a flash flood. Ecological Monographs 52, 93–110.

Fisher, S.G., Grimm, N.B., Marti, E., Holmes, R.M., Jones, J.B., 1998. Material spiraling in stream corridors: A telescoping ecosystem model. Ecosystems 1, 19–34.

Fountain, A.G., Lyons, W.B., Burkins, M.B., *et al*., 1999. Physical controls on the Taylor Valley Ecosystem, Antarctica. Bioscience 49, 961–971.

Grimm, N.B., 1993. Implications of climate change for stream communities. In: Kareiva, P., Kingsolver, J., Huey, R. (Eds.), Biotic interactions and Global Change. Sunderland, MA: Sinauer Associates.

Grimm, N.B., Fisher, S.G., 1989. Stability of periphyton and macroinvertebrates to disturbance by flash floods in a desert stream. Journal of the North American Benthological Society 8, 293–307.

Grimm, N.B., Arrowsmith, R.J., Eisinger, C., *et al*., 2004. Effects of urbanization on nutrient biogeochemistry of aridland streams. In: DeFries, R., Asner, G., Houghton, R. (Eds.), Ecosystem Interactions with Land Use Change. Washington, DC: American Geophysical Union, pp. 129–146. Geophysical Monograph Series 153.

Hastings, J.R., Turner, R.M., 1965. The Changing Mile: An Ecological Study of Vegetation Change with Time in the Lower Mile of an Arid and Semi-Arid Region. Tucson: University of Arizona Press.

Holmes, R.M., Jones, J.B., Fisher, S.G., Grimm, N.B., 1996. Denitrification in a nitrogen-limited stream ecosystem. Biogeochemistry 33, 125–146.

McKnight, D.M., Runkel, R.L., Tate, C.M., Duff, J.H., Moorhead, D.L., 2004. Inorganic N and P dynamics of Antarctic glacial meltwater streams as controlled by hyporheic exchange and benthic autotrophic communities. Journal of The North American Benthological Society 23, 171–188.

Minckley, W.L., Melfe, G.K., 1987. Differential selection by flooding in stream-fish communities of the arid American southwest. In: Matthews, W.A., Heins, D.C. (Eds.), Ecology and Evolution of North American Stream Fish Communities. Norman, OK: University of Oklahoma Press, pp. 93–104.

Schade, J.D., Fisher, S.G., Grimm, N.B., Seddon, J.A., 2001. The influence of a riparian shrub on nitrogen cycling in a Sonoran desert stream. Ecology 82, 3363–3376.

Stanley, E.H., Fisher, S.G., Grimm, N.B., 1997. Ecosystem expansion and contraction in streams. Bioscience 47, 427–435.

Stromberg J and Tellman B (eds.) (in press) *Ecology and Conservation of Desert Riparian Ecosystems: The San Pedro River Example*. Tucson: University of Arizona Press.

Valet, H.M., Fisher, S.G., Grimm, N.P., Camill, P., 1994. Vertical hydrologic exchange and ecological stability of a desert stream ecosystem. Ecology 75 (2), 548–560.

Deserts

C Holzapfel, Rutgers University, Newark, NJ, USA

What makes the desert beautiful is that somewhere it hides a well. Antoine de Saint Exupéry

Geography

Definition of Deserts

It is common belief that all deserts are hot and sandy places. While this is generally not true, a common factor of deserts is aridity, the temporal and/or spatial scarceness of water. True deserts can be delineated from other biomes based on their aridity. Of the following groups, only the first two are considered as true deserts here.

Aridity can be divided into four groups:

- extreme arid: less than 60–100 mm mean annual precipitation;
- arid: from 60–100 to 150–250 mm;
- semiarid: from 150–250 to 250–500 mm; and
- nonarid (= mesic): above 500 mm.

Since evaporation depends largely on temperature, bioclimatic aridity cannot be defined solely by the amount of precipitation. Therefore, the higher limits given above refer to areas with high evaporativity in the growing season (e.g., in subtropical areas with rainfall in warm seasons). This is taken into consideration in UNESCO's `World Map of Arid Regions' that defines bioclimatic aridity by P/ET ratios (annual precipitation/mean annual evapotranspiration). P/ET ratios smaller than 0.03 qualify for hyperarid zones (roughly corresponding to the extreme arid zone above) and a ratio of 0.03–0.20 as arid zone (thereby corresponding to the arid zone mentioned above).

Another common way of delineating deserts is based on their vegetation pattern and optional land use. Extreme arid zones typically show contracted vegetation restricted to favorable sites or lack vegetation altogether. Arid zones are characterized by diffuse vegetation. Semiarid zones mostly are characterized by continuous vegetation cover (if edaphic conditions allow for it) and only very locally dry-land farming (without irrigation) is possible. Farming without irrigation becomes a reliable option at larger scale in nonarid zones only.

Based on geographic location and a combination of temperature and geographical causes of aridity, deserts can be separated into five classes:

- *Subtropical deserts*. They are found in the hot dry latitudes between 20° and 30°, both north and south. These deserts lie within the subtropical high pressure belt where the descending part of the Hadley's cell air circulation causes general aridity.
- *Rain shadow deserts*. They are found on the landward side of coastal mountain ranges.
- *Coastal deserts*. Found along coasts bordering very cold ocean currents that typically wring moisture as precipitation from the air before it reaches the land, these deserts are often characterized by fog.
- *Continental interior deserts*. They are found deep within continents and far from major water sources.
- *Polar deserts*. They are found both in the northern and southern cold dry polar regions.

This article focuses on extreme/hyperarid and arid zones and on the first four of geographic desert classes listed above.

Where are Deserts Found?

True deserts are found on all continents except the European subcontinent (see **Fig. 1**). Altogether about 20% of landmass can be classified as desert, making it the largest biome on Earth. **Table 1** gives an overview of the largest deserts. In addition to these major deserts, many smaller, separately named deserts exist; many of these can be classified as local rain shadow deserts. All desert areas of the world border on land semiarid zones. These are either Mediterranean-type climate and vegetation, or dry temperate or tropical grasslands/savannas. The vicinity to these areas is important as many desert organisms are either shared with these transitional biomes or evolved from similar more mesic organisms.

Desert Landforms

According to relief type, two general groups of desert landforms can be distinguished: (1) shield-platform deserts and (2) mountain and basin deserts. The shield-platform deserts are most common in Africa, the Middle East, India, and Australia and are characterized by tablelands and basin lowlands. Mountain and hill slopes in this type of deserts are restricted to ancient mountains or areas with

Fig. 1 Map of world distribution of deserts. Shown are the arid and hyperarid desert region. Borders are somewhat tentative as a clear separation from semidesert scrublands is often not readily possible. Polar deserts are excluded.

Table 1 List of the major desert areas of the world (larger than 50 000 km^2)[a]

Name	*Size (km^2)*	*Type*	*Temperature*	*Countries*
Sahara Desert	8 600 000	Subtropical	Hot	Egypt, Libya, Chad, Mauritania, Morocco, Algeria Tunisia.
Kalahari Desert	260 000	Subtropical	Hot	Botswana, Namibia, South Africa
Namib Desert	135 000	Coastal	Hot	Namibia
Arabian Desert	2 330 000	Subtropical	Hot	Saudi Arabia, Jordan, Iraq, Kuwait, Qatar, United Arab Emirates, Oman, Yemen, Israel
Syrian Desert	260 000	Subtropical	Hot	Syria, Jordan, Iraq
Kavir Desert	260 000	Subtropical	Hot	Iran
Thar Desert	200 000	Subtropical	Hot	India, Pakistan
Gobi Desert	1 300 000	Continental	Cold	Mongolia, China
Taklamakan	270 000	Continental	Cold	China
Karakum Desert	350 000	Continental	Cold	Turkmenistan
Kyzyl Kum	300 000	Continental	Cold	Kazakhstan, Uzbekistan
Great Victoria Desert	647 000	Subtropical	Hot	Australia
Great Sandy Desert	400 000	Subtropical	Hot	Australia
Gibson Desert	155 000	Subtropical	Hot	Australia
Simpson Desert	145 000	Subtropical	Hot	Australia
Great Basin Desert	492 000	Continental	Cold	United States
Chihuahuan Desert	450 000	Subtropical	Hot	Mexico, United States
Sonoran Desert	310 000	Subtropical	Hot	United States, Mexico
Mojave Desert	65 000	Subtropical/rain shadow	Hot (cold)	United States
Atacama Desert	140 000	Coastal	Hot	Chile, Peru
Patagonian and Monte Deserts	673 000	Rain shadow	Cold	Argentina
Antarctic Desert	1 400 000	Polar	Very cold	Antarctica

[a]Various sources.

more recent volcanic activity. The geologically younger mountain and basin deserts (also called mountain and range) are predominant in the Americas and Asia and consist typically of mountain ranges separated by broad alluvially filled valleys. Within the two groups of desert landforms, there are several dominant geomorphological landscape types that are described here briefly.

Desert mountains consist chiefly of sheer rock outcrops and tend to rise abruptly from desert plains. The slopes of these mountains differ according to geological origin of the parent material. Igneous rock mountains tend to be characterized by large debris (boulder fields), while softer sedimentary rocks tend to lack these. Desert mountains (**Fig. 2**) dominate the desert of the USA (38% of desert area), the Sahara (43%), and Arabia (45%).

Piedmont bajada formations (**Fig. 3**) cover roughly a third of the arid Southwest USA but do less so in other desert areas of the world. These formations are built up from alluvial material that tends to accumulate in fans at the mouth of mountain canyons. Individual alluvial fans often coalesce and form large-scale graded slopes called piedmont bajadas (often only `bajadas'). Depending on deposition age and location along the bajada, the fill material is very diverse and differs strongly in alluvial particle size and soil structure, thus creating complex gradients and mosaics of distinct geological landforms. These gradients have been studied extensively in the Sonoran and Mojave Deserts and it had been shown that predominantly, the age and consequent erosion of the alluvial material within these mosaics determine the biological communities that can establish on it.

Desert flats (basins) are another common landscape type (about 20% in the USA and 10–20% in other regions). Often these flats have rather fine-textured soils and with sufficient rainfall vegetation is diffuse and rather evenly spaced across the landscape (**Fig. 4**). In

Fig. 2 Desert mountains: the Cambrian sandstone formations rise almost vertically from the valley floor filled deeply by sands that locally eroded from the mountain fronts. Wadi Ram, Jordan, October 2003. Photograph by C. Holzapfel.

Fig. 3 A piedmont bajada in the Mojave Desert: alluvial fan deposits stemming from a nearby mountain range vary in age and structure (here a mixture of Pleistocene and Holocene deposits). The position along the fan and the composition and structure of the deposits determine hydrology and plant growth (here the common desert shrubs *Larrea tridentata* and *Ambrosia dumosa*). Fremont Valley, California, USA, March 2006. Photograph by C. Holzapfel.

Fig. 4 Desert flats: this large desert basin in the Atacama Desert has very little surface dynamics and fine-textured soil materials are overlain by rocks forming a partial pavement. Among the harshest deserts on Earth, the Atacama receives very little to no rainfall and plant growth is lacking in most years. South of Antofagasta, Chile, October 1994. Photograph by C. Holzapfel.

more arid regions and when rainfall redistribution is patchy due to minor relief differences, distinct banded vegetation patterns can arise. These bands exist mostly in Africa and Australia but are also present in restricted areas in the Middle East and North America. The open areas produce the runoff of rainfall that accumulates in the bands, supporting the growth of vegetation. Another type of flat desert region can be differentiated as hammadas (bedrock fields). These bedrock fields develop *in situ* and depending on the size of rock fragments, can build dense pavements (regs) consisting of densely packed surface stones resting on finer-textured subsoil. This desert landscape type is common in the Sahara and the Middle East and accounts for 40% of the area.

Sand dunes (**Fig. 5**), known as `ergs' in Arabic-speaking countries, are dominant desert landscapes only in extreme arid desert areas (25% of Sahara and Arabian Deserts, less than 1% in arid Southwest USA). Characterized by moving sands, they depend on sources of sand, sufficient wind energy, and favorable accumulation areas. Depending on these factors and prevailing wind direction, different dune types arise. Crescent-shaped dunes (barkhan dunes) form perpendicular to the prevailing wind direction and tend to be highly mobile. Linear dunes form in the direction of the wind and therefore do not move over the desert landscape. This distinction is of biotic importance as the edges of dunes are favorable to plant growth, while the dune crest and upper slopes, due to sand drift and fast erosion, are usually devoid of vegetation. In sandy flats, individual shrubs tend to accumulate sand deposits and eventually form phytogenic hummocks (so-called nebkhas).

Playas (**Fig. 6**) are depressions with very fine-textured, often saline soils. Playas are the beds of former lakes that can be flooded in years of abundant rainfalls. These depressions are known under various names (North Africa and Middle East: chotts, sebkas). Even though individual playas can be large, worldwide they cover only 1% of desert.

Badlands form in areas with clay-rich soils and are typically located at the margins of arid lands, although they are found locally in arid regions as well. Depending on the strength of water forced erosion, badlands are areas with extremely high surface relief, typically forming fantastic `lunar landscapes.'

Dry river beds of ephemeral streams (**Fig. 7**) are of little importance with respect to land cover in deserts (only 1–5% worldwide), but are of immense biological importance. In extreme arid regions, these are the only places with vascular plant growth, and almost all

Fig. 5 Sand dunes: ergs are seas of sands that are constantly on the move. Vast sand deserts are typical for the Sahara (shown here) and Arabian Desert. Douz, Southern Tunisia, March 1986. Photograph by C. Holzapfel.

Fig. 6 Desert playas are often prehistoric lake beds with fine-textured, alkaline soils. Depending on current rainfall and temperature, playas can be flooded and then resemble the former lakes, as this playa on the altiplano of the Andes at an elevation of 4400 m. Plant life on playas is typically sparse but when microorganisms and invertebrates are active, birds such as these Andean flamingos (*Phoenicopterus andinus*) assemble in large numbers. East of San Pedro de Atacama, Chile, November 1994. Photograph by C. Holzapfel.

Fig. 7 Ephemeral stream: in arid areas vegetation concentrates along and in the bed of temporal stream beds. Due to available water in the subsoil that is plant extractable even long after temporal surface flow ceased, most of the primary production and species diversity in extreme deserts is restricted to these habitats. Nahal Zin, Negev Desert, Israel, April 1987. Photograph by C. Holzapfel.

Fig. 8 A flash flood obstructs traffic on a desert road. Depressions and stream beds quickly flood after strong rainfall events due to the high surface runoff in deserts. As in the case here, the precipitation source of the water can be remote and floodwater travels far distances. Due to high stream velocity and carried erosion material, such sudden floods can be disruptive to biotic communities and dangerous to humans. Sedom, Israel, March 1991. Photograph by C. Holzapfel.

animals depend at least at times during their life on the primary production here. This biological importance of ephemeral streams is therefore a foremost feature in deserts, and even though small in size, these landmarks were always distinctly named by human desert dwellers (washes in North America, wadi/*oued* in Arabic-speaking regions, *arroyo seco* in Spanish-speaking regions).

What is Special about Desert Climates?

Life in desert is limited by the scarceness of water. Secondary limiting factors are correlated to the main factor: the dearth of nutrients for producers and food energy for consumers and for both – at least temporally – high heat stress. Precipitation typically is so low that water becomes the controlling factor for biological processes. Precipitation is also highly variable throughout the year and typically occurs in infrequently defined events (discontinuous input). To make things even worse, precipitation varies randomly between years and is therefore not predictable. The coefficient of variation between years in arid areas is typically larger than 30% of the long-term average (and ranges in some extreme deserts to 70%). For comparison, temperate zones and tropical areas typically have coefficient of variation of less than 20%. Individual precipitation events in deserts can be tremendously large (for instance, 394 mm in a single rainstorm in the Peruvian desert that receives a long-term annual precipitation average of 4 mm) and due to surface runoff, large-scale flash flooding can occur (**Fig. 8**). Even though such sudden floods will replenish needed water to desert systems, erosion and direct damage to desert plants are the consequence. Because of the small and temporally highly variable rainfall amounts, deserts have been described by Noy-Meir as ``water-controlled ecosystems with infrequent, discrete, and largely unpredictable water inputs."

Adding to and interacting with the pronounced temporal variation is the high spatial variation of rainfall in deserts. This variation is caused by: (1) oreographic features (e.g., increase with altitude), (2) differences in degree and direction of slopes, and (3) the typically small size of precipitation fronts (often less than1 km in diameter).

Depending on the latitudinal geographical location and the origin of rain fronts, deserts receive either precipitation in the cooler season (cyclonic/frontal rainstorms) or in the warmer season (tropical, convective rainstorms). Some transitional desert regions receive both. The seasonality of rainfall is of great bioclimatological importance as evapotranspiration is larger during the warmer season and rainfall therefore tends to have a smaller biological effect. On the other hand, cold deserts receive precipitation during the cold season mostly in the form of snow and biological activity is then limited both by low temperatures and aridity. Snowmelts in spring create deep-reaching wetting fronts that will hold water available for plant uptake during the warmer growing season. Locally important other water inputs are condensations of atmospheric moisture as dew. These are crucial for plant production in the coastal fog deserts that otherwise do not receive direct precipitations. It is less common inland but can be noticeable in high desert areas as well (for instance in the Negev Desert of Israel). Fog water inputs are directly usable for cryptogamic organisms (e.g., lichens) and many arthropods. Foliar uptake of fog by vascular plants has been demonstrated but its relative importance in the water balance remains controversial. Water vapor tends to move along temperature gradients and can be important in dry soils with strong diurnal radiation. An upward movement of water vapor at night causes formation of dew close to the surface. Such water might sustain germinated plants until they produce roots long enough to reach deeper and wetter soil depths.

Deserts usually experience an extreme diurnal temperature range, with high daylight temperatures (up to 50 °C, the highest temperature recorded in Death Valley was 56.7 °C), and extremely low nighttime temperatures (often dropping below 0 °C). This is caused by very dry air that is transparent to infrared (heat) radiation from both the sun and the ground. Thus during daylight all of the sun's heat reaches the ground. As soon as the sun sets, the desert cools quickly by radiating its heat into space. Clouds reflect ground radiation and desert skies are usually cloudless, thereby increasing the release of heat at night. With intense sun radiation, surface temperatures can be extreme and depending on the color and type of surface can exceed 80 °C.

Desert Soils

The main features of desert soil that affect water and nutrient availability include texture, content of organic matter, pH, and orientation within the landscape. Desert soils show typically little development from parent material and some authors even state that typical developed soils do not exist in deserts. Most desert soils are classified as Aridisols and are differentiated into soils with a clay (argilic) horizon (Argids) and soils without such horizons (Orthids). Other soils, less common in deserts, are mollisols, soils with dark A horizons, and Vertisols, cracking clay soils. Accumulated subsurface horizons with either clays or calcium carbonate (calcic horizons) have clear implication as impediments to water infiltration.

Most desert soils tend to be slightly to highly basic. Such reactivity can negatively affect phosphorous and micronutrient availability as these are generally not in solution at pH $>$ 7.0. Organic matter helps to increase infiltration and via decomposition adds to nutrient availability. It is often distributed unevenly in desert soils (see below).

Soils in deserts have important effects on water inputs as they act as short-term water stores and modify water availability by a number of regulation processes. These regulation processes include direct infiltration and often more importantly runoff and horizontal redistribution of water. Redistribution by runoff tends to be of crucial importance in deserts and contributes to spatially very patchy distribution of water. Relatively impermeable surfaces (e.g., biotic or physical crust in clay-rich soils) create runoff areas that result in catchments that are water rich. Such water redistribution enables patchy plant production even in extreme arid zones, where plant growth would not be possible since evenly distributed sparse rainfalls would not exceed the threshold needed for plant life. Because of sparse plant growth, soil-created redistribution of water is more important than precipitation interception through plant surfaces. However, locally such interception combined with stem flow can create water-rich spots under shrub or tree canopies. In contrast, smaller precipitation events can be locally intercepted and lost by evaporation. This is the reason that soils in the understory of desert shrubs or trees can be either wetter or dryer than the surrounding soil.

Soil texture is of large importance as it affects both infiltration and the movement of wetting fronts. Fine-textured soils that are high in clay and silt fraction tend to impede infiltration, in which wetting fronts move only very slowly, and surface evaporation after rainfalls can be very high. More-coarse-textured soil rich in sand fractions, as for instance sandy loams, is characterized by high infiltration rates and rapid percolation. For this reason, coarse-textured soils are often better for plant growth. As this is in contrast to soils in mesic areas where fine-textured soils are commonly considered to be superior for plant production, this is called the `inverse texture effect'.

Clearly, the orientation and dynamics of soil surfaces within the landscape plays a large role in arid ecosystems. Exposed southern (or northern, depending on the hemisphere) slopes receive high solar radiation and therefore due to higher evapotranspiration, tend to be drier than opposite slopes (Fig. 9). These inclination differences are observable on large-scale landscape level or small-scale microtopography level. An example is the sun-exposed sides of shrub hummocks that are often only raised by a few centimeters, but can be bioclimatically and ecologically very different from the less-exposed side. Slope exposition also plays a role when rainfall directions due to prevailing winds are constant. Rain-exposed slopes can receive up to 80% more water than other slopes.

Biogeography and Biodiversity

General Diversity

The casual observer often assumes that deserts support only low species richness and diversity because of the harsh environmental conditions prevalent in arid areas, but among plants and animals almost all taxa are represented (even aquatic groups like fishes

Fig. 9 Marked phenological and plant composition differences due to slope exposition (southeast facing slope on the left and northwest facing slope on the right). The southeast-facing slope is subject to higher evaporational water losses and receives less direct rainfall compared to the northwest-facing slope. Such abiotic differences result in clear biotic contrast in arid environments, making these systems ecological model cases. Judean Desert, Palestine, December 1989. Photograph by C. Holzapfel.

and amphibians) here, and their species richness may be comparable to that of more mesic environments. Even though detailed comparative data are lacking, it has been argued that the diversity in North American deserts is comparable to some grasslands and even temperate forests. In general, however, evidence based on correlations along climate gradients indicates a decrease of species richness in plants and animals with increasing aridity. Regardless of this, specific taxa can be more species rich in deserts than in bordering less arid systems and regionally show negative relationship of richness with increasing precipitation. Examples for these are reptiles and birds in North America, and ants in Australia. Taxonomic groups that are generally species rich in deserts are rodents, reptiles, some insect groups (e.g., ants and termites), solpugids (camel spiders), and scorpions. In the following, an overview of typical desert taxa is given, and some emphasis is given on the ecological role of these groups in deserts. More specific treatment of ecophysiological adaptations follows in the next section.

Ecological Role and Diversity of Microorganisms

Even though obviously not readily observable, microorganisms inhabit all desert areas and in the extreme arid zones are often the only life forms present. Relatively little is known about the diversity within the lower three kingdoms (Fungi, Protista, Monera) in general and even less is known about the species richness of these groups in deserts. A recent survey that uses `DNA fingerprinting', aiming at resolving bacterial ribosomal DNA, indicated that soils of semiarid sites can harbor higher bacterial richness then mesic sites. Since factors other than water availability are more important (chiefly soil pH) in determining microbial diversity, it can be assumed that true desert can be quite rich as well.

Mycorrhizal fungi seem to be quite important in desert ecosystems, as in more mesic ecosystems. It appears that mycorrhizae of desert plants not only supply the plants with nutrients but also supply moisture during the dry season, at times taking the place of root hairs. Studies conducted in the Chihuahuan Desert indicated that most dominant, perennial species have high arbuscular (AM) fungal infection rates in their coarse roots system, while fine-rooted annual species in comparison show much lower infection rates and are also much less dependent on mycorrhizal associations in general. Worth mentioning are mycorrhizal desert truffles (*Terfezia* and *Tirmania*: *Ascomycetes*), that are host specific to *Helianthemum* species in the arid region of the Middle East and the Mediterranean zones of the Old World. The desert of the American West supports an elusive community of aboveground observable fungi in which the *Gasteromycetes* (puffballs and allies) figure predominately. Another common example is *Podaxis psitillaris* (desert shaggy mane), a species most common in sandy deserts.

Except for their crucial part in mycorrhizal associations, desert microorganisms are noteworthy for their role in three typical desert phenomena: desert crusts, desert varnish, and interstitial communities. Desert crusts are microbiotic communities composed of drought and heat-tolerant algae, cyanobacteria, fungi, lichen, and mosses. These often species-rich communities are held together by sticky polysaccharide secretions and thus form surface crusts. Desiccated crusts are often indiscernible until rainfall or dew moistens the surface and microbial communities become active and green. Under extreme conditions, such crusts can form below the surface. This is possible especially under the protection of semitransparent calcareous or siliceous stones (quartz is a good example) that enables transmission of light up to a depth of 5 cm. The most common life form in crusts (and in some areas also in hot deserts in general) is cyanobacteria. Among their roles in the desert ecosystem are atmospheric fixation of nitrogen and the binding of soil particles. Together with mineral-reducing bacteria, the cyanobacteria are important in soil fertilization and soil formation and thereby have clearly important effects on vascular plants and dependent animal consumers. In hot deserts, cyanobacterial crusts often form smooth surfaces, while in cold deserts, where crust forming interacts with frost heaving, a very rough surface is typical. These different surface types clearly affect vascular plants differently.

Even exposed desert rocks can support life. Clearly the most visible organisms are crustose lichens. However, when conditions become too extreme for growth of lichens, bacteria can still survive on the surface of rocks. Desert varnish, the dark and shiny surface found on sun-exposed, porous stones in hot deserts, is the result of bacterial activity. These bacterial colonies obtain energy from inorganic and organic substances and trap submicroscopic, wind-borne clay particles. These particles accumulate in a thin layer and act as sun protection. Over very long time periods, estimated at thousands of years, these bacterial communities oxidize wind-blown manganese and iron particles and when baked together with clay particles form the dark desert varnish. The color of desert varnish varies depending on the relative proportion of oxidized manganese (dark black) to iron (reddish).

Environmental conditions even more extreme than those that support surface bacterial growth can still allow the formation of interstitial communities. These communities consist mostly of algal species that inhabit the matrix of sedimentary rocks in depth up to 4 mm. These communities can stay dormant for long periods of time and inhabit hot and cold deserts alike (they are known to exist on exposed rocks in Antarctica).

Desert Flora

Even though the geological record indicates that arid conditions existed for a long time (since the Devonian), the current modern desert flora might have originated in the Miocene, expanded in the Pliocene (after restrictions during moist periods in the Cretaceous and Tertiary), and reached its current distribution only during the Pleistocene. Specifically, the deserts in the North American Southwest are relatively young. Overall richness and uniqueness of desert floras reflect size, age, and isolation of desert areas, with larger deserts typically hosting larger numbers of endemic species. Smaller desert regions and edges of larger regions are often characterized by species that evolved in adjacent more mesic areas and partially adapted to arid conditions. A good illustration is the high incidence of Mediterranean plants in desert areas bordering regions with semiarid Mediterranean climates in all parts of the world. In general, desert floras tend to have high affinity to bordering semiarid climate zones, such as Mediterranean climate-type regions and semiarid grassland. Taxonomical studies of many species groups revealed that desert species have evolved (recently) from nondesert species. Biogeographically, strong floristic links exist between old deserts in North Africa, Middle East, and Asia. Floristic similarities among desert regions stretching from North Africa to Central Asia are particularly obvious since no wide barriers of ocean or humid vegetation exist to restrict plant migration; these floristic similarities are present despite strong climatic contrasts ranging from hot environments in North Africa to the much colder, arid Central Asia deserts. Apparent links between the North American Great Basin and Central Asian deserts might be explained by plant migration across the Beringian land bridge. Clear affinities between the deserts in both Americas can be explained by the Panamanian land bridge. In this respect, the distribution of Larrea shrubs is remarkable. The two recognized species – *Larrea divaricata* in South America and *L. tridentata* in North America – are taxonomically and phenotypically very close. It appears that the genus *Larrea* evolved in South America and migrated only tens of thousands of years ago (bird assisted?) to North America where it quickly became the dominant shrub in all warmer desert areas. Corresponding to the isolation of the Australian continent, the flora of the Australian desert is very different from all other deserts of the world.

Dominant plant life forms in deserts reflect water stress conditions typical for deserts (for a treatment of drought adaptations see the following section). While trees are relatively rare and restricted to more-mesic microsites, a wide range of plant life forms can be found that include many short-lived and seasonal active plants (e.g., annual or ephemeral plants and bulbous plants/geophytes). The dominant life forms that visually shape the plant formations are perennial woody plants (mostly shrubs) and fleshy succulent plants (cacti and others). Large succulent species can be dominant in some of the hot desert regions (e.g., the saguaro cactus in the Sonoran Desert). A few plant families are predominant in desert areas. The aster family (Asteraceae) is the most diverse plant family in deserts overall; it is especially numerous in Australia, southern Africa, the Middle East, and North America. Some deserts can be dominated by grass species (Poaceae). Some plant families have their global center of diversity in deserts and most likely evolved here. Notable examples are the chenopods (Chenopodiaceae) that are diverse in arid and semiarid regions of Australia, North America, and from the Sahara to Central Asia. The New World cacti (Cactaceae) are another example of a group of species rich in deserts but relatively sparse in other biomes.

Deserts are home to some of physiognomically extremely unusual plant types. Worth mentioning in this respect are plant characters as the Joshua trees of the Mojave Desert (*Yucca brevifolia*), the famous Welwitchia of the Namib (*Welwitchia mirabilis*), and the boojum tree (**Fig. 10**) of the Sonoran Desert in Baja California (*Fouquieria columnaris*). Exactly why and how deserts host these exceptional plant types is not clearly understood and such `Dr. Seussification' of the desert flora deserves systematic study.

Desert Fauna

The faunas of deserts are often biogeographically more distinct between regions than the desert floras are. Despite this, many similarities exist between the different desert regions. Such phylogenetic similarities typical for the African–Asian deserts are explained by the lack of dispersal barriers, and similarities between North American and Asian regions on one hand and North American and South American regions on the other are likely the result of existing land bridges. The Australia desert fauna, as its desert flora, is very distinct. As mentioned earlier on, almost all animal taxa are present in deserts, but some groups are more diverse than others, with the major deciding factor for this being the general aridity.

Relative to other insect groups, ants and termites are very diverse in deserts. However, their species richness is lower than it is in the Wet Tropics, where these groups originated. These groups reach high population densities and ecological importance is high. With up to 150 species per hectare, the highest species richness for ants is found in Australian deserts. Most desert arthropods are either detrivores (termites, beetles, etc.) or granivores (mostly ants), or are predators feeding on these (scorpions, spiders, etc.). Due to the lack of constant

Fig. 10 Boojum trees (*Fouquieria columnaris*) with associated shrubs, agave, and cacti on a bajada in the Sonoran Desert of Baja California. Cataviña region, Mexico, October 1997. Photograph by C. Holzapfel.

plant production, herbivores are relatively sparse or show pronounced, often dramatic temporal–spatial fluctuations (e.g., mass flights of desert locusts). Species rich substone communities consisting of protozoa, nematodes, mites, and other microarthropods are typical for deserts, creating a microcosm where grazers and predators feed on bacteria, algae, fungi, and detritus.

Fishes live in almost every aquatic habitat on the globe and small, permanent desert water sources are no exceptions. Obviously richness is extremely low, but species often live in very restricted areas and often under extreme conditions. The desert pupfishes (*Cyprinodon* sp.) in the deserts of North America are among the most species-rich groups in deserts. Some species live at temperatures of 45 °C and salt regimes 4 times that of seawater, while some species are restricted to an area as small as 20 m^2 (e.g., the Devil's Hole pupfish in Nevada). These fishes are opportunistic omnivores.

Likewise, desert amphibian communities are depauperate since at least the juvenile stages depend on water. Only a small fraction of the world's amphibians, mainly anurans, are able to occupy deserts.

Reptiles are common and widespread in all deserts and, with the exception of crocodilians and amphisbaenians (worm lizards), all orders are represented in deserts. Relatively few tortoises occur in deserts since they are restricted due to their plant diet. Snakes and lizards are well represented (especially in Australia). The extreme high diversity of reptiles in Australian deserts has been explained by low diversity of mammal and birds which resulted in lower competition for food and lesser predation pressure than in other desert regions. It appears that reptiles as endothermic consumers enjoy an advantage over other ectothermic consumers in the deserts of Australia that are characterized by low-quality plant production.

Even though birds have basic adaptation to cope with dry climates, diversity in deserts worldwide is relatively low and a clear positive relationship between rainfall and bird diversity is typical. Despite this, few desert specialist species developed among the avifauna: sand grouse, lark, parrots, etc.

Likewise, mammals are not very diverse in comparison to other biomes, but some taxa evolved to be true desert groups. Among smaller mammals are the heteromyds in North America, the jirds and gerbils in the African–Asian deserts, and the dayurid marsupials in Australia. Some of the desert mammals are rather large and therefore have advantageous low surface-to-volume ratios (see next section). The `flagships' for this are clearly the camel species (Camelidae) that originated in the Americas in the Miocene and are now naturally found in desert regions of the Old and New Worlds; they are clearly the largest animals in all desert regions. It is of significance that most large herbivorous mammals, including camels, donkeys, goats, sheep, and horses, have been domesticated historically in deserts and semiarid regions and are common as domesticated livestock today. Other large, non-domesticated ungulates such as gazelles, ibexes, and oryxes are generally extinct or at least rare and endangered.

Convergence of Desert Life Forms

Most desert plants and animals initially evolved from ancestors in moister habitats, an evolution that occurred mostly independently on each continent. Despite this phylogenetic divergence, a high degree of similarity of body shape and life form exists among the floras and faunas of different desert regions. Since desert environments are defined by their water limitation and have similar landscapes worldwide, it is not surprising that many organisms show convergent evolution and are morphologically and functionally alike. Similar pressures of natural selection have resulted in similar life forms. In fact, many of these analogous species groups became textbook examples of evolutionary convergence:

- Stem and leaf succulence is found in nonrelated plant taxa: cacti in New World, milkweeds and *Euphorbia* species in the Old World (however, this form is lacking in Australia).
- Bipedal locomotion is found in unrelated small rodent groups: jerboa (family Dipodidae) in the Old World, kangaroo rats (family Heteromyidae) in the New World.

- Bipedal locomotion is shared in a few larger mammals: African springhare (genus *Pedetes*), desert-living kangaroos (Macropodidae) in Australia.
- North American horned lizards (genus *Phrynosoma*) and the Australian thorny devil, the unrelated agamid lizard *Moloch horridus*, share similar grotesque spiny body armors. This has been explained as an adaptive suit that facilitates their need of having a large body due to their specialization on ants. Ants as eusocial insects present a clumped however low digestible source of food (formic acid, chitin). Both lizard groups are in need of a larger digestive system and therefore large bodies that in turn makes them slow moving and in need of protection.

Many adaptations that are discussed in the following sections are typical for all desert regions of the world. A combination of these traits creates the `typical' desert life form that to some extent is similar worldwide.

Ecophysiology and Life Strategies

Strategies for Coping with Drought

All life originated in the sea and all organisms that have left their ancestral home depend on an `inner sea', high internal water content. This phylogenetic inheritage restricts life in many habitats, and obviously deserts are among the harshest in this respect. Even though deserts are not only water limited (they are also low in nutrients and energy resources), adaptations to cope with the spatiotemporal scarcity of water are predominant of most (if not all) true desert organisms.

All desert life forms, animals, plants, and microorganisms alike, employ one or more of three basic strategies to cope with the dearth of water: (1) drought evasion, a strategy of avoiding water stress temporarily in inactive states; (2) drought endurance, a suit of adaptations that reduce actual stress and enable being active during drought; and (3) drought resistance, a suit of adaptations evolved to avoid water stress altogether. Note that water and heat stresses are coupled, thus many of the adaptations mentioned below can be understood as strategies to cope with both.

1. Drought-evading organisms `choose' to pass exceedingly dry periods in dormant stages. Predominant examples are short-lived (ephemeral) plants that survive the dry season or longer periods of drought in the dormant seed stage. Such annual plants are indeed very common in many deserts of the world and compose a large portion of the plant diversity in many areas (up to 80% of species richness). An equivalent for animals can be found in cryptobiosis of invertebrate eggs and larvae. Such aridopassivity can be found in fully developed organisms as well; examples are bulbous geophytes and desert animals that pass dry season belowground inactive (estivation). Choosing of less arid microsites is another way of avoiding drought. In animals, these are typically behavioral space choices (e.g., permanent habitation or temporary use of stress-protected microsites: below shrubs or stones, rock fissures, litter, below tree and shrub canopies, or even soaring in high air). Likewise, many plants are restricted to favorable microsites (e.g., under tree and shrub canopies, runon microsites, algae growing under stones). Some organisms, mostly plants, are able to lose water almost completely and `resurrect' once water becomes available again (poikilohydry: *Selaginella* species, algae, lichens, and moss species).
2. Drought endurance is a main strategy common among the dominant desert organisms worldwide. A suit of ecophysiological, morphological, and behavioral adaptations work together to reduce the most detrimental impacts of water stress.
   Reducing water expenditure. Evergreen desert shrubs are capable of fine-tuned regulation of stomatal movement. Specialized photosynthetic pathways evolved in desert plants that minimize water loss and maximize carboxylation. C_4 and crassulacean acid metabolism (CAM) pathways are adaptations to hot temperatures, compared to the C_3 pathway adapted to colder conditions. Animals of arid regions are able to regulate and restrict water loss by concentrating urine. Birds and reptiles excrete urinary waste as uric acid that can be concentrated and allow reabsorption of water in the urinary tract, a trait not available to mammals. Desert mammals and most other taxa excrete dry feces and reduce the urine flow rate. Water loss through surfaces is reduced in plants through an increase in thick lipid cuticulae, epidermal hair cover, sunken stomata, small surface/volume ratio (leafless plants with photosynthesizing stems – xenomorphic). Animals employ a variety of adaptations that reduce water loss: impermeable integuments (e.g., in arthropods), changes of lipid structure in the epidermis that create diffusion barriers to water vapor (some desert birds), denser hair or feather cover, and small surface-to-volume ratios (common in large mammals).
   Prevention of overheating. High-temperature stress is closely connected to water stress as many of the ways of coping with higher temperatures involve expenditure of water, thereby exacerbating water stress. Examples are transpiration cooling in plants and evaporative cooling in animals (including humans; see below). Desert organisms typically have high heat tolerance and capability to function at high temperatures. The comparatively high-temperature optima and temperature compensation points of photosynthesis in plants and high body temperatures and high lethal temperatures in animals attest that. Among the most thermotolerant species are desert-dwelling ants that forage on extremely hot surfaces. A Saharan desert ant species (*Cataglyphis bicolor*) is noted to hold the record with a critical thermal maximum of $55 \pm 1\ ^{\circ}C$.

Apart from tolerating high temperatures, an array of mechanisms evolved to decrease or dissipate heat loads both in plant and animals. The formation of sheltering boundary layers, employment of insulating structures, and increase of reflection (white color, glossiness) are among these mechanisms. Behavioral space and temporal choices are a contribution to the prevention of overheating. Seeking of sheltered microhabitats and nocturnal activity of many (if not most) desert animals are obvious examples. The nocturnal CO_2 uptake in CAM plants is an interesting analog to this.

3. Drought-resisting organisms employ adaptations that allow them to pass dry periods in an active state without experiencing physiological water stress. The succulence of many typical desert plants worldwide is a form of water storage that enables these plants to use water during dry periods. Examples for taxa that are rich in succulent species are the cacti (Cactaceae) and yuccas (Agavaeae) in the New World and some members of Euphorbiacea and Crassulaceae in the Old World. Succulent plants typically cannot become dormant and therefore require at least periodically predictable precipitation, a requirement that explains the general lack of succulent plants in extreme arid environments where prolonged droughts are common. Most succulent plants have fairly shallow root systems that react very quickly following larger rainfall events. An analog to plant succulence in animals can be found in desert snails that can store large amount of water. The ability of desert mammals (notably the camel) to store large amount of water in the blood is another analogous trait. The accumulation of fat tissue that can be metabolically transformed into water (see below) as a water storage mechanism is somewhat controversial and is more universally understood as being merely an energy source (e.g., fat reserves in desert reptile tails, body of rodents, and the famous camel's hump).

Water Uptake in Deserts

Animals

Vertebrates are able to obtain water from three sources: (1) free water, (2) moisture contained in food, and (3) metabolic water formed during the process of cellular respiration. Some are able to receive water from all three sources, while others are able to exploit only one or two methods. Highly mobile animals tend to be restricted to the use of open water sources that are often sparse and far between. Typical examples are desert birds that fly in regular intervals to the few bodies of water available. To mention are the desert-adapted orders of sand grouse (Pteroclidiformes) and some doves (Columbiformes) that tend to visit standing water in large flocks at dawn and/or dusk. The former are even known to transport water soaked in their specialized belly feathers to their flightless chicks. Many desert animals are able to use available water opportunistically by drinking large quantities in short time. This ability is proverbial in the camel that can take up to 30% of its body weight in a few minutes. Camels and other desert mammals have resistant blood cells that can withstand osmotic imbalance. Animals living in more mesic environments (including humans) would destroy their red blood cell at such high water content in their blood. Much of the free available water has high salinity, and so it is not a surprise that many desert animals show high salt tolerance, for instance by employing salt-excreting glands. Other animals, mostly the ones that are restricted in their mobility (e.g., mammals, reptiles, and insects), rely on water obtained from their food. Carnivorous and insectivorous animals typically receive enough water from their prey. Herbivores do so as well, as long as the moisture content of the consumed plant material is relatively high ($>$ 15% of fresh weight: fresh shoots and leaves, fruits, and berries). The ultimate desert-adapted method however is the extraction of metabolic water. Especially seed-eating (granivorous) animals are able to metabolically oxidize fat, carbohydrate, or protein. Rodents and some groups of desert birds (e.g., larks, Old World and New World sparrows) are able to convert these energy sources into water: 1 g of fat produces 1.1 g of water, 1 g of protein produces 0.4 g of water, and 1 g of carbohydrates produces 0.6 g of water. Schmidt-Nielson has shown that kangaroo rats (genus *Dipodomys*) are able to obtain 90% of their water balance from metabolic water derived from consumed seeds. The remaining 10% is obtained from moisture stored in seeds. The use of already stored body fat as source of water is controversial. It has been argued that metabolizing fat and other storage sources into water requires increased ventilation and therefore increases water loss by transpiration from lung tissue. At the most, no net gain of water will be the result. According to this, the camel's hump might function simply as a fat energy storage facility, one that is situated in one place in order to reduce isolation and allow dissipation of heat.

In areas with high humidity, animals are able to receive water from dew. Such direct uptake as the main source of water is probably restricted to arthropods and some mollusks (snails). There is some evidence that rodents can utilize condensation by water enrichment of stored food (**Fig. 11**).

Fig. 11 Desert sand rat (*Psammomys obesus*). As the scientific name implies, this day-active desert rodent can store large amounts of body fat as reserves during unproductive seasons. Like other desert rodents, it obtains all of its needed water through its plant diet. Negev Desert, Mitzpe Ramon, Israel, May 2003. Photograph by C. Holzapfel.

Plants (and microorganisms)

Plants, with few exceptions, depend on water uptake by their roots from the soil. Due to low soil matrix water potentials and high salinity in arid regions, such soil water is often not readily available. One way for desert plants to overcome this restriction physiologically is to osmoregulate the plant cell water potentials to overcome the low potentials of desert soils, a mechanism that also aids them in extracting water from saline solutions. Indeed, some of the lowest water potentials have been measured in desert shrubs (-8 to -16 MPa (mesic plants rarely go below -2 to -3 MPa)) and salt-tolerant (halophytes) desert perennials (as low as -9 MPa). In general, many desert plants tend be deep rooted and are therefore able to exploit water reserves that tend to be available in the deeper soil layers. Due to the need of desert plants to forage extensively for water, root-to-shoot ratio of desert plants is typically high and rooting depths are larger than in other ecosystems. In extreme cases, as in phreatopytes, rooting depth can exceed 50 m. This was found for mesquite trees (genus *Prosopis*) that are practically independent from local precipitation and are able to maintain very high transpiration rates for prolonged periods. In contrast and as mentioned before, many succulent plants that store water in their tissues tend to be shallowly rooted and are able to intercept even light summer rains that do not cause a deeper recharge of soils and would otherwise be lost to evaporation. Annual plants and most grasses also benefit from being shallowly rooted. In general, many desert plants can react quickly to available water by deploying fast-growing `water roots' from special dormant root meristems. Shallow rooting plants show temporally intensive water exploitation patterns while plants with deeper root systems are characterized by spatially extensive water exploitation patterns.

Some deep-rooted perennial plants exhibit hydraulic redistribution from deeper soils to shallow soils. Water is absorbed from the soil at greater depth during the day and moves via the transpiration stream upward into shallower roots and the aboveground parts of the plant. At night when the air is more humid and plant stomata are closed, plants become often fully hydrated and water may be exuded from the root into the dry shallow soil. This pattern, described as hydraulic lift, may have nutritional benefits for the perennial plant itself, as it enables it to utilize the nutrients from what would have otherwise been dry soil. Released water – on the other hand – might become available for competing plants. Hydraulic lift has been described in almost all of the dominant shrubs of the arid Western US (e.g., *Artemisia tridentata*, *Larrea tridentata*, *Ambrosia dumosa*) and might be prevalent all over the world's arid zones.

Plants of saline habitats, halophytes, must be able to acquire water with high salt concentrations. They need to overcome the high osmotic pressure of saline solutions and need to avoid the potential toxicity of some ions (Na^+, Cl^-). In order to achieve such a high salt tolerance, halophytes employ strategies as osmoregulation, dilution of inner cell salt concentration by succulence, and use specialized salt-excreting glands.

Special water-rich habitats within deserts, for instance, permanent stream sides and springs, attract extrazonal plants that often possess only few aridity adaptations. Found in these oases are wetland plants and some salt-tolerant tree species that can be characterized as `water spenders'. Good examples are palm trees (the date palm *Phoenix dactylifera* and the Californian palm *Washingtonia filifera*) and salt cedars (*Tamarix* species).

Direct uptake of condensed atmospheric water (dew and fog) and water vapor is generally possible only for some specialized poikilohydrous vascular plants, but is of much greater importance for microbiotic organisms such as lichens and cyanobacteria.

Strategies to Cope with Unpredictable Water Resources

A wealth of adaptations arose in desert organisms that allows them to utilize the pronounced spatiotemporal stochasticity of water availability typical to deserts. As detailed before, mobile organisms are able to use spatially patchy water sources that are not available to less-mobile organisms. These sessile organisms often have dormant dispersal units that can reach good microsites where they can establish, reproduce, and eventually send their own diaspores onto other favorable microsites. Such life cycles are typical for short-lived plants (annuals, ephemerals) and some invertebrates. These diaspores typically remain viable for long periods and can `sit and wait' for years with sufficient precipitation. Most annual desert plants form such extensive seed banks. Seeds within such seed banks tend not to germinate equally and even after strong precipitation events, a fraction of the seed will remain dormant. Such fractional dormancy might serve as avoidance of sibling competition as it will reduce densities, but more importantly has been explained as a bet-hedging adaptation in order to cope with rainfall stochasticity. When no supplemental rainfall follows an initial germination triggered by a rainfall event, at least a fraction of the seeds will be available in the following years, thereby ensuring the long-term survival of the population. In addition to dormancy, many desert plants develop some water-sensing adaptation (so called `water clocks') that controls both dispersal and germination. Dry inflorescences of the famous rose of Jericho (*Anastatica hierochuntica*) and other annual plants (e.g., the New World *Chorizanthe rigida*) open up only after abundant rainfall and release only some of their seeds (**Fig. 12**). Many desert plants have morphologically different seeds that differ in dispersal ability and have different germination requirements (amphicarpic plants). In general, a high proportion of desert plants suppress seed dispersal altogether (atelechory). This has been interpreted as an adaptation to remain on the mother site, as it has already been proved to be a favorable location.

Most perennial plants suppress flowering (aridopassive shrubs) or sprouting altogether (e.g., geophytes) in drought years. This is analogous to many desert animals that shift sexual maturity and mating to synchronize with favorable conditions. Similar to plants, sterility is typical for extreme drought years and dispersal and migration (nomadism) are triggered by precipitation regimes. There is some indication that insects and desert shrubs can shift their sex expression with changing rainfall regimes. Especially, monoecious shrubs, plants that have male and female reproductive units on the same individual, can shift their sex ratio with water availability. The male function requires fewer resources from the plant (`cheaper sex'), and is typically the predominant sex in dry years.

Fig. 12 Dry dead plant of the rose of Jericho (*Anastatica hierochuntica* – mustard family Brassicaceae). Seed pods of this annual plant are contained within curled branches forming a ball that opens when moistened and seeds are released only after rainfall events. Dead Sea region, Israel, March 1987. Photograph by C. Holzapfel.

Many desert shrubs tend to break apart into separate shoot sections over time (axial disintegration). This so-called `clonal splitting' is very common for desert shrubs worldwide and has been explained as a risk-spreading adaptation. In time of severe drought, instead of the death of the whole original individual, some segments of the original shrub may survive. The consequence of this growth strategy is often the formation of shrub rings that grow outward and have a dieback zone in the center. Age estimations have been made based on this growth form. Large creosote bush (*Larrea tridentata*) rings in the Mojave Desert, for instance, have been determined to be of an age exceeding 11 000 years (e.g., the famous `King Clone' located by Vasek in 1980).

System Ecology (Ecosystem and Communities)

The leading question in desert ecology is whether aridity alone can explain all aspects of biological systems. If so, desert environments could be understood simply by characterizing the harsh, abiotic environmental factors that prevail in desert systems. Thus, desert systems do not follow the typical ecosystem view and can be described as simplified systems that react to discrete rain events (triggers) by short-term growth production (pulse), interspersed by long-term storage of organic material (seeds, roots, and stems – reserves). This pulse–reserve conceptual desert model is clearly too simplistic; however, it provides an important framework for the description of major ecological components of deserts.

In contrast to this basic view of deserts, two major alternative hypotheses have been developed in regard to the driving factors defining communities and populations in deserts. One hypothesis states that only the primary producers are water limited and all other trophic levels (consumers) are determined by the magnitude of this water-dependent primary production. Another hypothesis postulates that water shortage affects organisms only individually and has no direct effect on higher-order species interactions. According to this view, aridity effects on ecosystems and communities are rather the indirect outcomes of direct physiological and behavioral responses of individual organisms (and their populations) to scarcity of water. Despite the fact that the temporal and spatial lack of water is clearly the driving force behind the individual ecologies of desert species, current research makes it clear that species interactions, including both negative and positive ones, can be strong in deserts. The following sections strive to provide a brief summary of the types of interactions typical to deserts.

Production

Net annual primary production (NPP) is lower in deserts than in most major biomes. However, when taking into account that deserts typically are also characterized by low amounts of permanent plant mass (standing phytomass), relative primary production (the ratio of NPP/standing phytomass) is among the highest worldwide (see **Table 2**). As rainfall fluctuates strongly within and between years, it is no wonder that there is a tremendous spatiotemporal variation in the amount of primary production. However, due to the lack of responsive vegetation structure and typically low levels of soil fertility, deserts are somewhat limited in their biological potential to react to extremely wet years. Semiarid grasslands, rich in very plastic perennial plant structures and therefore exhibiting high potential growth rates, show much larger fluctuations in response to changing water availability (**Fig. 13**). Also water-use efficiency (NPP divided by annual water loss) in deserts is lower than it is in dry grasslands (0.1–0.3 g per 1000 g water in deserts compared to up to 0.7 g in dry grasslands and 1.8 g in forests).

During brief periods when water is available in excess, the typically short supply of nitrogen (and other plant macronutrients) is limiting. Even though nitrogen is limiting in almost all terrestrial ecosystems, deserts are typically more limited due to four reasons: (1) plant growth is triggered by available water faster than nutrients can be replenished by decomposition; (2) desert soils

Table 2 Phytomass and primary production of deserts in comparison to some other major biomes of the world[a]

Plant formation	*Phytomass of mature stands (t ha^{-1})*	*Net annual primary production (t ha^{-1} yr^{-1})*	*Relative primary production*
Tropical forests	60–800	10–50	0.004–0.05
Deciduous forest	370–450	12–20	0.03–0.06
Boreal forest	60–400	2–20	0.03–0.05
Savanna	20–150	2–20	0.1–0.14
Temperate grassland	20–50	1.5–15	0.08–0.3
Tundra	1–30	0.7–4	0.09–0.1
Deserts	1–4.5	0.5–1.5	0.33–0.5

[a]Modified from Evenari, M., Schulze, E.-D., Lange, O., Kappen, L., Buschbom, U., 1976. Plant production in arid and semi-arid areas. In: Lange, O.L., Kappen, L., Schulze, E.-D. (Eds.), Water and Plant Life - Problems and Modern Approaches. Berlin: Springer, pp. 439 -451, and other sources.

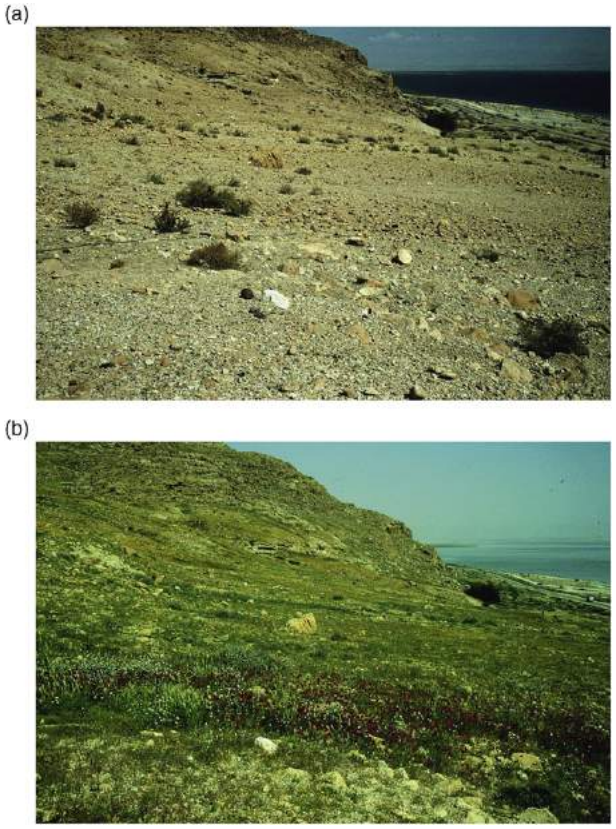


Fig. 13 Extreme, 30-fold differences of plant growth on a rocky desert slope during (a) a dry (precipitation 40 mm, NPP 0.03 t ha^{-1} yr^{-1}) and (b) an extremely wet year (193 mm, 0.87 t ha^{-1} yr^{-1}). Northern Dead Sea, Palestine, March 1991 and March 1992. Photographs by C. Holzapfel.

typically have little nutrient-holding capacities; (3) the nutrient-rich organic matter is located in the upper layers of soils, a layer that is typically too dry for root growth to occur, rendering the nutrients inaccessible; and (4) detritus and other organic material is deposited and accumulated unevenly across the desert surface. Plant debris typically accumulates passively under the canopy of shrubs or is concentrated in nests of animals such as harvester ants and termites. Thus the desert is an `infertile sea' with interspersed islands of fertility.

Resource–Consumer Relationships (Trophic Interactions)

In contrast to some ecosystems, food chains in deserts can be characterized by the importance of the link between producers and consumers via decomposition. Less than in most mesic environments, plant material is typically not directly consumed alive; some estimation puts the amount of energy that moves via decomposition into the food web as above 90% of total primary production. Since food resources are unpredictable, many animals can opportunistically switch from one mode of consumption to another (e.g., many arthropods are either herbivores or decomposers).

Decomposition

Microbial decomposition is often limited by low water availability, resulting in the accumulation of dry plant material and seeds. For that reason, animal detrivores are more important in deserts than in more mesic environments. Examples are darkling beetles, termites, and isopods. Termites are abundant in most of the warmer deserts and are often the dominant decomposers of dead plant material (above- and belowground) that play an extraordinarily important role in nutrient cycling. Since most termites live belowground, they are also important in the formation of soils. A similar phenomenon is displayed by scavenging animals, which are comparatively abundant among the desert fauna. Examples are large mammals (hyenas, coyotes, and jackals) and many birds (Old World and New World vultures, ravens, etc.). Like smaller detrivores in the desert fauna, many of these scavengers can switch to a predatory diet when needed.

Herbivory

Similar to other ecosystems, deserts host a large variety of herbivorous animals that potentially utilize every part of the plants. Some of the drought adaptations of plants, discussed before, also function to deter herbivores. Tough outer layers, spines, and elevated leaf chemicals, all typical for desert plants, can therefore also be understood as mechanisms to protect low and therefore costly primary production. Some plants appear to employ growth forms that make them less conspicuous for herbivores. Remarkable examples are the living stones (*Lithops* species, Aizoaceae) of South Africa that blend with the surrounding rocky desert pavement.

Predation

Abundant detrivorous arthropods are the most important prey source in the desert and provide the base for a relatively large assembly of smaller (e.g., spiders, scorpions) and larger predatory animals (e.g., reptiles, birds). The abundance of long-term stored seeds and fruits in desert systems supports an assembly of a diverse guild of granivores (seed predators). These granivores are recruited from taxonomically much differentiated groups (e.g., ants, birds, rodents), all of them potentially competing for similar food sources. Carnivorous predators can be abundant as well. These predators are mammals, birds, and reptiles (mostly snakes). Because of the relative openness of the desert terrain, prey organisms rely on a number of predator avoidance strategies. Examples are general crypsis (camouflage), `freezing behaviors', and nocturnal activity pattern. Active deterrents are spines (desert hedgehogs, horned lizards), hard shells (desert tortoise), and poisons that can be employed in active predation as well. Strong predator pressure combined with the need for efficient predation in a desert environment poor in prey might be the reason that some of the most poisonous animals we know (e.g., snakes, scorpions, Gila monster) are true desert animals.

Parasitism

Parasitic interactions are often very conspicuous in desert environments. Many desert shrubs show abundant signs of an attack by gall-forming insects. For instance, *Larrea tridentata*, the dominant shrub in all the hot deserts of North America, is attacked by 16 specialized species of gall-forming insects. Parasitic plants, stem and root parasites alike, are common in deserts worldwide. Though detailed studies are lacking, these parasites seem to have the potential of reducing host plant production and performance (Fig. 14).

Nontrophic Species Interactions

Competition among and within species has been recognized as an important force that shaped the communities in all mesic environments and the question whether this is also true for deserts is a natural one, however one that has not been answered univocally. Some researchers conclude that biomass production and densities in desert are typically below a threshold that would

Fig. 14 Heavy infestation of a desert shrub (*Ambrosia dumosa*) by an epiphytic parasite (*Cuscuta* sp.). Parasitic plants can be common in deserts and their effects can add to the abiotic stresses of aridity. Panamint Valley, California, USA, April 1995. Photograph by C. Holzapfel.

necessitate competition for resources. Observing the same density pattern, other researchers state that because such low densities indicate strong resource limitation in desert, strong competition should ensue.

Based on studies of spatial plant community structure, it appears that current competition in deserts is rare; most studies show clumped or neutral patterns – itself a sign of the lack of competition – while only few studies show a clear regular pattern (a sign of past competition). Experimental removal of individual plants in the Mojave Desert, on the other hand, demonstrated interspecific competition among dominant desert shrubs. Spatial studies that assess the size distributions in dependence of distance between desert shrubs typically detect signs of negative association; larger shrubs tend to be spaced farther from each other than smaller ones. Removal experiments with granivorous rodents commonly result in density increase of the remaining species, thereby indicating current competition. The fact that character displacement, the evolution of divergent body features in coexisting species, has been demonstrated for desert rodents is another sign that competition has been of importance at least at one time.

Ecological theory predicts that negative interactions (such as resource competition) decrease in importance with increasing abiotic stress, and positive interactions (such as facilitation) increase. Following this, it should be possible to observe along a mesic to arid gradient a waning of competitive interaction and an increase of facilitative interactions. Indeed, a clear indication of this has been observed in a survey of positive effects among plants that resulted in a proportionally large number of cases from arid regions. In many deserts of the world, one can easily observe the positive association of either young perennials with adult perennials or herbaceous plants with larger perennial plants. Experimentally, it had been shown that the perennials had net positive effect on the smaller sheltered plants. Examples for these so-called `nurse plant effects' are the associations of young succulent plants (often cacti), trees, and shrubs and the prevalent, close association of annual plants with desert shrubs. Typically the larger nurse plant provides canopy shading and increased soil fertility (see above discussion on islands of fertility), and sometimes protection from herbivorous animals to the sheltered plants. In accordance with this prediction, shrub–annual associations tend to be strongly positive in arid sites and less so (or even negative) in less arid sites (**Fig. 15**). As nothing ever in nature is one-sided, these unidirectional facilitative effects are countered by negative effects as the nursed plants can have negative, competitive effects on their benefactor. Competition for water has been shown between annuals and sheltering shrubs and such negative effects are typical once sheltered young succulents outgrow the nurse plant.

Tradeoffs in competitive/facilitative interactions are also found between taxonomically very distant groups. One example is the complex nature of interaction between microbial crusts and vascular plants. For one, these crusts can have very contrasting effects on seed placement. Cold deserts tend to have very rough crust surfaces that facilitate seed deposition and establishment, while the smooth crusts typical to hot deserts decrease such seed entrapment. Because of these differences, no general effect of desert crusts on the performance of vascular plants has been recognized. Nitrogen fixation by cyanobacteria increases nitrogen availability, thereby favoring plant growth; however, the creation of crusts can result in runoff and water redistribution that in turn locally reduces plant performance.

Human Ecology

Origin and History

Humans have lived at the edge of desert and in the desert proper for ever and there are some indications that modern *Homo sapiens* evolved when the world climate turned to be more arid at the end of the Pleistocene. Though lacking many of the physical adaptations of true desert dwellers, we humans might be a desert species after all. One of the adaptations humans bring to live in the desert is a rather high heat tolerance. The combination of upright position that minimizes direct sun exposition during hottest times of the day, the profusion of sweat glands all over the body, and the lack of body hair, together with an energetically

Fig. 15 Clear associations of annual plants with shrubs (here *Ambrosia dumosa*) are common in deserts. Annual plants benefit from nutrient enrichment and shade provided by the shrub canopy and since they usually only provide little benefit to the shrub (e.g., thatch-induced increase in water infiltration and lower soil surface evaporation), they can compete with the shrub for resources. Owens Valley, California, USA, March 1997. Photograph by C. Holzapfel.

conservative way of movements, contributes to our ability to cope with hot deserts. As long as water and salt balances are maintained, humans can perform relatively well under heat stress. This is evidenced by the success of persistence-hunting practices in desert and semideserts, which involves tracking large ungulate prey on foot during midday heat. Such persistence hunting, today only employed by hunter-gatherers in the Kalahari Desert, has been the most successful mode of hunting prior to the domestication of dogs, and uses the relative heat balance advance that a well-hydrated and trained human can have over animal quadrupeds. Recent data show that contemporary hunters run for 2–5 h over distances of 15–35 km at temperatures of 39–42 °C until prey items (mostly antelopes) overheat and can be overcome.

Deserts have been important throughout human history and the first civilizations arose in or close to deserts (Mesopotamia and Egypt). Agriculture practices, often involving irrigation, are sometimes interpreted as cultural ways to deal with the stochasticity of the desert climate. It is interesting to note that the first written law, the codex written by the Babylonian King Hammurabi dating back to 1750 BC, was designed to manage such crucial irrigation systems. It is basically the same set of laws that gave rise to our modern laws. Since ancient history, deserts have been the cradle of great civilizations on one hand and the theater of fierce armed conflict on the other (Fig. 16). One wonders whether the nuclear weapons tests that have been conducted in the deserts of New Mexico and Nevada (among other desert sites worldwide) symbolize that deserts can foster both the beginning and the end of civilization.

Desert Economy

For humans, there are traditionally only three basic ways to sustain themselves in deserts: hunting-gathering, pastoralism, and to some extent agriculture.

Ever since the rise of agriculture in the Neolithic era, foraging as the exclusive mode of production (hunter-gatherers) became limited to areas that were marginal to agriculture or animal husbandry. Naturally deserts are among these zones. Examples of peoples who foraged as hunter-gatherers are the aborigines in Australian deserts (this practice receded since the European discovery of the continent), and the !Kung (bushmen) of the Kalahari, who remain foragers in our times. Recent research on the !Kung people showed that hunting-gathering is a suitable lifestyle that can sustain healthy populations that are even able to spend sufficient leisure time, all this as long as population densities are low. Some Amerindian people employed hunting-gathering in deserts as well. There are some evidences that a later immigration wave of people, the Nadene, linguistically distinct from the first Clovis people, were culturally better adapted to harsh environments and settled first in semiarid grasslands and eventually in deserts (the Navajo and Apache might be the descendents of the Nadene).

Pastoralism is a true desert activity that is also typical for semiarid grasslands. It is obvious that many of the livestock animals that were and are herded by pastoralists originated from arid and semiarid areas and therefore are well adapted to such environments. The ancestors of horses, sheep, and goats evolved in semiarid environments and donkeys and camels in arid environments. People who live as pastoralists in deserts often combine animal husbandry with some scale of horticulture; this combination is called transhumance. In order to use the stochastic desert environment optimally, many pastoralists have to follow rainfall events and are partly or truly nomadic, as is exemplified by the traditional lifestyle of the Bedouin of the Arabian Peninsula (Fig. 17).

The use of agriculture most likely did not evolve in the desert proper, but it has to be mentioned that the first cultured plants, annual grasses, and legumes were domesticated near the edge of the desert in the Middle East (10 000–8000 BC Natufian culture). Independently, in likewise semiarid areas in Mexico (Tehuacan Valley, before 7200 BC), the domestication of Teosinte into corn (*Zea mays*) took place. Deserts harbored in historical times small-scale horticulture near springs and elaborately designed irrigation systems that utilized the effects of runoff and water redistribution. Water-harvesting systems in runoff farms have been found and

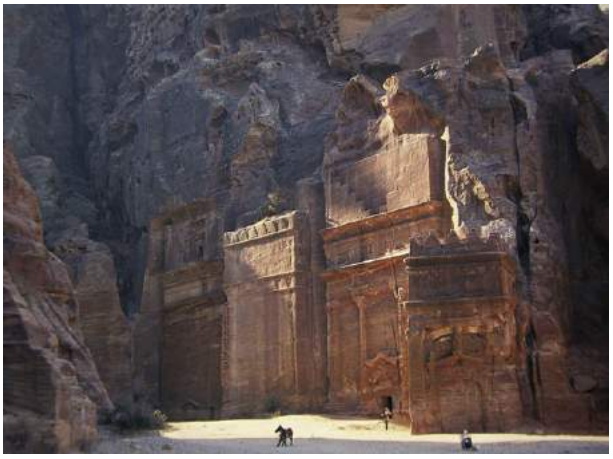


Fig. 16 Many ancient sites thrived near or in deserts. The former Nabatean capital Petra is located in a desert valley surrounded by steep mountains. From here the Nabateans, an Arabic tribe, controlled the trade through the deserts of the Middle East. Petra, Jordan, October 2003. Photograph by C. Holzapfel.

Fig. 17 The nomadic lifestyle is a cultural adaptation of desert-dwelling people to the unpredictability of the desert environment. As still seen here in the Sahara Desert, traditionally camels were essential for transport between grazing areas and arable oases. Douz, Southern Tunisia, March 1986. Photograph by C. Holzapfel.

partly recreated in the Negev Desert (e.g., the Nabatean system in Avdat and Shifta) and in the arid southwest of North America. Large-scale agricultural enterprises depend on permanent water courses. As along the Nile in Egypt and along the Tigris and Euphrates in Mesopotamia, these water sources originated from areas far beyond the desert region. Modern, often large-scale irrigation projects are mostly independent from surface water and use deeper aquifers.

In history, many large cities were established in desert areas (Egypt, Middle East, South America) and there are many cites in deserts in our times (Phoenix, Tucson, Las Vegas). Incidentally, the climate and ecology of urban areas even in the temperate, nonarid zones has many similarities to true deserts (e.g., water limitation due to surface sealing and runoff, high temperatures, etc.).

Human Impact on Deserts

As all ecosystems with low productivity, deserts are fragile to disturbance. Some ecologists go as far as to state that no direct succession occurs at all after disturbance but it is at least obvious that regeneration times after perturbations can be very long. The few long-term studies following disturbance, as for instance the vegetation recovery of ghost towns in the American West, demonstrate these long recovery times that often exceed many decades. It can be generalized that any human impact that changes the soil structure will last very long. Unlike in mesic environments, abandoned agricultural fields in deserts will recover only very slowly (if at all) to natural desert vegetation. Additionally, formerly irrigated fields will have elevated salt concentrations for long periods of time. Soil surface disturbance caused by off-road vehicles inflict severe changes in hydrological characteristics of soils, which might remain permanently. The increase in off-road vehicles in the North American deserts, and increasingly also in the Middle East, is a serious threat to deserts and desert biotas.

Desertification, largely the human-caused extension of the desert, is one the most serious problems facing the globe. Causes of the growth of the desert regions are multifaceted and are a combination of natural long-term variation in the weather, climate destabilization, and human mismanagement due to overpopulation and land-use change. Under the UN Convention to Combat Desertification, desertification is defined as land degradation in arid, semiarid, and dry subhumid areas resulting from various factors, including climatic variations and human activities. The effects of desertification promote poverty among rural people, and by placing stronger pressure on natural resources, such poverty tends to reinforce existing trends toward desertification. It is now clear that in several regions, desert environments are expanding. This process includes general land degradation in arid, semiarid, and also in dry and subhumid areas. Clearly in areas where the vegetation is already under stress from natural or anthropogenic factors, periods of drier-than-average weather may cause degradation of the vegetation. If such pressures are maintained, soil loss and irreversible change in the ecosystem may ensue, so that areas that were formerly savanna or scrubland vegetation are reduced to human-made desert. To counter this process that will increasingly endanger lives and livelihood of millions of people (not to speak of drastic effects on the biodiversity of the planet), synoptic management approaches are needed that combine understanding of the process and investigation into the regional causes of the process, in order to comprehend the effects on the Earth's overall system. It is important to emphasize that desert border areas that undergo desertification will not simply convert into natural deserts. Disturbed and overused semiarid zones are characterized by lower biodiversity than original, natural deserts. Therefore desertification will not simply increase the global area of deserts; it will create large tracts of devastated lands.

Human activity and human-caused climate change will facilitate the migration of ruderal (disturbance-adapted) plant species into locally favorable microsites within the desert. This has been shown for the vegetation along roadsides in the Middle East and in the North American southwest. Even though this might enhance local, small-scale species richness, an overall reduction in regional diversity and a loss of desert-adapted species might follow. Such a strong mixing of former distinct biotic zones has been observed along the edges of deserts in the context of human-caused disturbances and climate change. A wide variety of `extrazonal'

Fig. 18 Many larger desert animals became extinct in the wild due to hunting pressure. These captive Arabian oryx (*Oryx leucoryx*) are part of a breeding effort that led to release operations into formerly occupied desert ranges (Oman, Bahrain, Jordan, Saudi Arabia). Wadi Araba, Israel, May 2003. Photograph by C. Holzapfel.

plants are crossing zonal borderlines, a process that will potentially lead to a marked decrease in large-scale species diversity. This migration by species that are native to the general geographic area but are now spreading into new climatic or biogeographic zones is an overlooked aspect of species invasion.

Due to typically strong abiotic stress, desert areas have been in the past remarkably resistant to invasions by non-native organisms. Notable exceptions have been biological invasions by deliberately introduced organisms in Australian deserts (e.g., rabbits and *Opuntia* species). However, invasion seems to increase rapidly worldwide and many desert areas today show a dramatic increase in the arrival and spread of non-native species. At present, the deserts of the American South West seem to be affected most. Plants originated from the Old World, mostly grasses (e.g., annual *Bromus* species, some perennial grasses), but increasingly members of other plant families also have invaded many desert communities and can have strong impacts on native desert communities. Among the detrimental effects are dramatic changes in fire regimes and direct competition with recruiting shrub seedlings and native annual plants, and even negative effects on adult desert perennials have been demonstrated. The main reason for these trends is due to general land-use changes in desert and desert margins. In the Southwestern US, disturbances due to increasing suburbanization of deserts, besides increases in nutrient depositions, seem to be central agents of these changes.

Endangered Species

Many of the larger vertebrate desert species are threatened and a number of species have been lost to global extinction. The openness of the desert habitat and naturally small population size makes large mammals and birds conspicuous and thus very vulnerable to overhunting. Threatened species include the central Asian wild Bactrian camel (*Camelus bactrianus*), the onager (*Equus hemionus*; a wild ass of southwestern and central Asia), and large antelope species as the addax (*Addax nasomaculatus*) of North Africa and the Arabian oryx (*Oryx leucoryx*; **Fig. 18**). Hunting is also the main reason that larger birds are endangered. Among birds many bustard species are threatened (e.g., the houbara, *Chlamydotis* sp.) or are already extinct (e.g., the Arabian subspecies of the ostrich: *Struthio camelus syriacus*). Large predators have been and continue to be extensively hunted since they are perceived to be a threat to livestock (e.g., desert subspecies of the Old World leopards, *Panthera pardus jarvisi*). International efforts to save many of the larger endangered animals are currently ongoing; many of these efforts involve reintroductions.

Invasive species can have detrimental effect on threatened species as well. An example is the increased fire frequency caused by annual, non-native grasses, which is threatening populations of the desert tortoise (*Gopherus agassizii*) in the deserts of North America.

Desert Research

One of the major attractions of desert ecosystems for scientists lies in their simplicity. Spatial patterns of life are often visible and clear cut and ecologists tend to feel empowered by the sense of ecological understanding. As any desert scholar will have to attest though, this simplicity is only relative. In comparison to more complex systems, deserts seem to invite ecological questions with greater ease than for instance tropical rainforests would. Therefore much of basic ecological knowledge has been founded in desert research and these dry places more often than not were used as simplified models for the green and (forbiddingly) complex world. Thus it is no wonder that the desert has spawned many research efforts, notably among them large, coordinated ecological research enterprises. The permanent research sites established worldwide during the International Biological Program (IBP) are good examples; in the US, many of these continue to be monitored under the Long Term Ecological Research (LTER) program.

See also: **Behavioral Ecology:** Social Behavior and Interactions. **Ecological Complexity:** Emergent Properties. **General Ecology:** Temperature Regulation. **Global Change Ecology:** Climate Change 2: Long-Term Dynamics. **Terrestrial and Landscape Ecology:** Landscape Planning

Further Reading

Belnap, J., Prasse, R., Harper, K.T., 2001. Influence of biological soil crusts on soil environments and vascular plants. In: Belnap, J., Lange, O.L. (Eds.), Biological Soil Crusts: Structure, Function, and Management. Berlin: Springer, pp. 281–300.

Evenari, M., Schulze, E.-D., Lange, O., Kappen, L., Buschbom, U., 1976. Plant production in arid and semi-arid areas. In: Lange, O.L., Kappen, L., Schulze, E.-D. (Eds.), Water and Plant Life – Problems and Modern Approaches. Berlin: Springer, pp. 439–451.

Evenari, M., Shanan, L., Tadmor, N., 1971. The Negev. The Challenge of a Desert. Cambridge, MA: Harvard University Press.

Fonteyn, J., Mahall, B.E., 1981. An experimental analysis of structure in a desert plant community. Journal of Ecology 69, 883–896.

Fowler, N., 1986. The role of competition in plant communities in arid and semiarid regions. Annual Review of Ecology and Systematics 17, 89–110.

McAuliffe, J.R., 1994. Landscape evolution, soil formation, and ecological patterns and processes in Sonoran Desert bajadas. Ecological Monographs 64, 111–148.

Noy-Meir, I., 1973. Desert ecosystems: Environment and producers. Annual Review of Ecology and Systematics 4, 25–41.

Petrov, M.P., 1976. Deserts of the World. New York: Wiley.

Polis, G.A. (Ed.), 1991. The Ecology of Desert Communities. Tucson, AZ: University of Arizona Press.

Rundel, P.W., Gibson, A.C., 1996. Ecological Communities and Processes in a Mojave Desert Ecosystem: Rock Valley, Nevada. Cambridge: Cambridge University Press.

Schmidt-Nielsen, K., 1964. Desert Animals: Physiological Problems of Heat and Water. London: Oxford University Press.

Shmida, A., 1985. Biogeography of the desert flora. In: Evenari, M., Noy-Meir, I., Goodall, D.W. (Eds.), Hot Deserts and Arid Shrublands. Amsterdam: Elsevier, pp. 23–88.

Smith, S.D., Monson, R.K., Anderson, J.E., 1997. Physiological Ecology of North American Desert Plants. Berlin: Springer.

Sowell, J., 2001. Desert Ecology: An Introduction to Life in the Arid Southwest. Salt Lake City, UT: University of Utah Press.

Whitford, W.G., 2002. Ecology of Desert Systems. San Diego: Academic Press.

Dunes

P Moreno-Casasola, Institute of Ecology AC, Xalapa, Mexico

Published by Elsevier B.V.

Introduction

Coastal beaches and dunes have a worldwide distribution. They are common in both temperate and humid tropical areas, in arid climates, and in regions covered by snow during the winter. Beaches and dunes are considered two of the most dynamic systems. They are not permanent structures, but rather huge sand deposits that move and have an episodic supply of sand.

They can be found in deserts as well as on dissipative coasts with a plentiful supply of sediments and where there are strong onshore winds or winds that are parallel to the coastline. Sand dunes are eolian bedforms and beaches are marine geomorphic structures. Dunes form from marine sand delivered to the beach from the near-shore by waves. The exposed sediment is dried by the sun and the wind then transports sand inland to form incipient dunes and foredunes. Tidal range is important in this process since a high range exposes a large intertidal area that often dries out between the tides. These sediments constitute a major source of wind-blown sand given that sand-sized sediments are more easily transported by wind.

Dune size varies considerably. Some of the biggest dunes are found in deserts such as Badain Jaran Desert in the Gobi Desert in China (approximately 500 m), the Sossuvlei Dunes, Namib Desert (380 m), and the Great Sand Dunes National Park Preserve in Colorado, USA (230 m). Along the coast, on the Bassin d'Arcachon, France, is Europe's largest sand dune, the Dune du Pyla, nearly 3 km long, reaching 107 m in height, and moving inland at a rate of 5 m yr^{-1}.

Dune Origin and Formation

Wind is the main agent forming sand dunes. There is an exchange of sediments between beaches and dunes and this is part of a natural process that maintains both morphological stability and ecological diversity. Once exposed, sand is vulnerable to aerodynamic processes.

Wide beaches are formed in the summer and narrower beaches during the winter. Storms erode beaches and transport sand out of the system or to other beaches. Bonding, both by moisture and chemical precipitates, may cause surface adhesions, raising thresholds and reducing erosion. Sometimes salt forms a whitish crust on the sand surface, also bonding sand grains.

Sand grains come in a wide variety of shapes, colors, and densities, depending on their origin and on how long they have been rolling in water currents and wind. Silicate sand and calcium carbonate sand (formed by fractured shells and skeletons) are the more common components of coastal dunes. Sand texture, as well as shape and density, affect transport. Smaller particles are easier to move than larger ones. Sediment size is measured on the Wentworth scale. It is harder for angular grains to become airborne but they may move further once they have. Denser grains are harder to move and often accumulate as lag deposits on the upper beach.

Almost all wind-blown sand travels quite close to the ground, through a mechanism called saltation. Individual grains move in a series of continuous leaps. Once airborne, a grain describes a curve path, and lands hitting the ground at a low angle, but with sufficient force to rebound into the air again. It hits other sand grains that become airborne and do the same thing. In a short time, there is a considerable amount of sand in the air. Under most circumstances, deposition takes place within a short distance although sometimes sand may be transported long distances alongshore where the wind blows parallel to the coast. Deposition is favored by obstacles such as driftwood, clumps of vegetation, boulders, and plastic objects which perturb air flow and create a shelter zone. Small dunes are formed with their tails – called trailing ridges – stretching downwind.

Changes in wind strength and direction cause rapid resedimentation. Often a dune's surface changes by the hour, creating complex stochastic patterns. Over time, these processes create recognizable dune bedforms such as ripples, sand waves, and barchans.

Most coastal dunes form in the presence of vegetation. An important determinant of dune form is the drag imposed by the vegetation on the air flow. Dunes can be classified according to the percentage vegetation cover. At one extreme are dunes that have been stabilized by their vegetation cover (fixed, shore-parallel ridges and parabolic dunes) and at the other are the free wind forms (barchan or sand wave dunes, transverse dunes). Transitional forms are typified by a fragmented topography (hummock dunes).

There is a strong interaction between vegetation and dune form, and there are several patterns of incipient dunes. Plant form modifies sand deposition, forming a leading edge (as in the case of *Ammophila arenaria*), a trailing edge (*Spinifex hirsutus*), or intermittent deposition in clumped vegetation. Perennial grasses such as *Agropyron junceiforme* and *Elymus arenarius* as well as tropical long-branched creepers (*Canavalia rosea* and *Ipomoea pes-caprae*) grow laterally and vertically and are able to raise a dune a meter or two high.

Sand dunes act as a buffer to extreme winds and waves and they also shelter landward communities. They replenish the depleted beach and near-shore during and after storms, and are important in the retention of freshwater tables against salt intrusion. They filter rain water and are also important habitats for plants and animals. People have always appreciated their beauty and recreational value.

R. W. Carter wrote that ``Of all the coastal systems, sand dunes have suffered the greatest degree of human pressure.'' Many have been irreversibly altered by human activities such as tourist developments, golf courses, and urban growth.

Abiotic Factors

Dune ecosystems may be viewed as a series of gradients related to various environmental factors, which operate on different spatial and temporal scales. If we view a profile from the sea landward, we first have the beach (near-shore and back-shore), the embryo or incipient dunes, and the foredune. The first dune ridge (the next inland from the foredune) is normally the highest and forms a continuous sand structure. The second is an older dune ridge, frequently lower because of the reduction in sand supply and the gradual loss of sand. This formation occurs when we have a series of parallel ridges, formed by onshore winds, each ridge lower than the previous. Sand is trapped by vegetation and saltation cannot be initiated beneath the vegetation, unless a blowout forms. Older dune ridges become fragmented when blowouts and parabolic dunes develop. Parabolic dunes are formed when prevailing winds blow at right angles to the dune ridges. Poorly stabilized regions are rapidly eroded but the more vegetated areas on either side remain covered by plants for a longer time. As the bare sand of the central region moves inland, the two horns or tips of the parabola remain attached to the relatively stabilized sand of the trailing ridges. A slack (a dune depression where sand has been blown away until the water table is exposed) may be formed in the middle, between the parabola arms. Parabolic dunes can also be formed in transverse dunes.

Throughout the dune field, there are gradients in salinity, sedimentation, nutrients, flooding, and shelter. Dune vegetation forms a complex spatial mosaic, mainly because of variations in physical gradients which depend on the distance to the sea and topography. Disturbances also result in temporal successions that add another dimension of complexity to the spatial mosaic described.

Sand Movement

Dune movement has only been measured in a few dune systems and most of the published records are based on estimates from maps, the height of sand on fence posts, houses, and trees, etc. The results show that the rate of movement varies considerably among systems, varying from a few centimeters per year to 70 m per month, the latter in New Zealand (personal observation of Patrick Hesp). Dune formation depends on an adequate supply of sand and the wind to transport it. The interaction of wind and vegetation is of primary importance for dune growth. Colonization by plants accelerates dune growth, because surface roughness created by vegetation decreases wind flow and increases sand deposition. Several plants show an inherent capacity to bind sand and are able to develop extensive horizontal and vertical rhizome systems. The growth form and the ecological dynamics of dune plants are important contributors to foredune growth. Rhizomatous growth (as in the grass *Ammophila*) or stoloniferous growth (as in *Ipomoea* or *Spinifex*) can extend the foredune depositional area by 5–15 m in a few months. *Elymus arenaria* (Europe) develops vertical rhizomes 150 cm long and *Ipomoea pes-caprae* (pantropical) can have 25-m-long branches that are buried two or three times along their length. **Fig. 1** shows species that are able to survive and reproduce successfully under high rates of sand mobility in different parts of the world. In each region, sand-tolerating species have evolved, and they play a very important role in

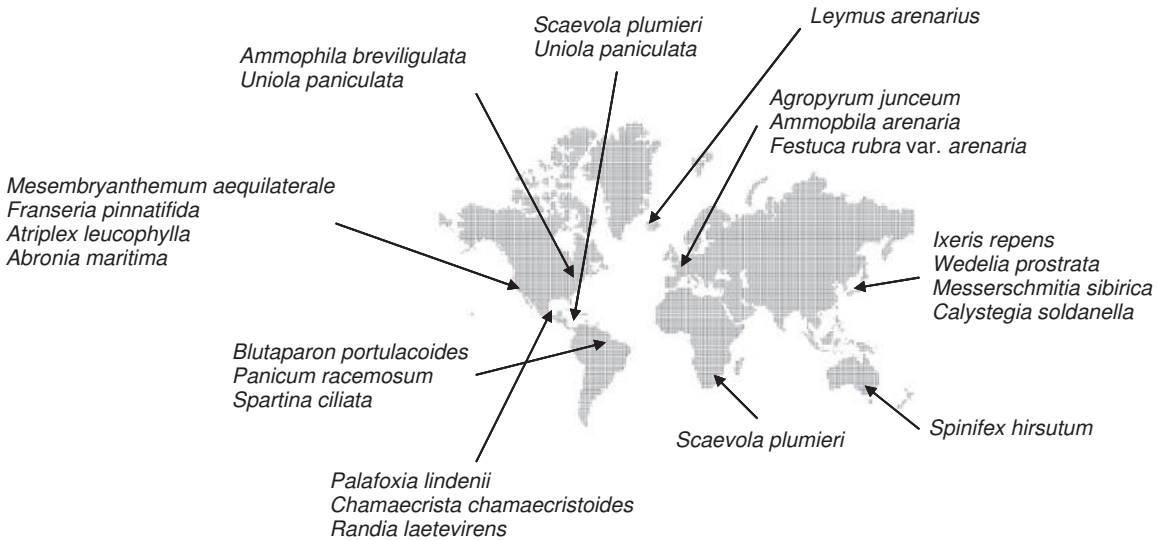


Fig. 1 Species that are able to survive and reproduce successfully under high rates of sand mobility in different parts of the world. Many regions have their own set of species that play important roles in stabilizing sand dunes locally.

dune formation. Sand deposition produces vigorous growth in some of these species; both plant height and plant cover increase, making these species excellent dune fixers. Many hypotheses have been suggested to explain this response of sand dune plants, but there are few studies in which the explanations are based on experimental evidence. Changes in soil temperature, increased space for root development, higher nutrient and moisture availability, a response to darkness, meristem stimulation, and interactions with endomycorrhizae and nematodes are probably factors that play an important role in this response.

Nutrients

There are great differences in the soil properties of young dunes (formed by recently blown sand), and those of more mature dunes in which vegetation has dominated. Newly blown sand from the beach is low in mineral nutrients. Dune soils show marked changes as they age. Pioneer species that initiate dune stabilization are able to live in very poor soils. On fully vegetated dunes, organic matter and nutrients accumulate, and the leaching effects of rainfall decrease. Leaching dissolves carbonate and moves it downward to the water table. With time, the organic matter of nutritionally poor soils of younger dunes increases, and pH decreases. The increase in organic matter content varies among dune systems, depending on the climate and colonizing species. In high-rainfall climates such as Southport (Lancashire, Great Britain), organic matter increases slowly at first but much faster after about 200 years. In Studland, Dorset, the early invasion of *Calluna* is largely responsible for a very rapid increase in organic matter. Primary productivity and the competitive abilities of coastal plants are frequently limited by nutrient availability, with nitrogen deficiency the most severe. As succession advances, plants increase their cover, communities change from grasslands to thickets, and then to tropical or temperate forests, adding nutrients and organic matter to soils. In dune depressions, where water is not a limiting factor for plant establishment, the accumulation of organic matter is faster. Experiments with dune plants have shown that many species are slow-growing and generally show growth responses characteristic of plants from infertile habitats.

Salinity

Soil salinity comes from salt spray and foam blown inland, and the amount of salt usually correlates well with the distance inland or degree of protection from the wind. In some regions with a Mediterranean climate, such as California, soil salinity follows a seasonal progression. Late summer additions by fog and salt spray result in high values at this time of the year. Winter rains leach salt away, salinity decreases, and in early spring reaches its lowest level. Salinity gradients affect species distribution, especially for those plants sensitive to salinity. Germination and growth might be difficult when soil salinity is high. Salts in the soil affect plants by making water less available, and high salinity is considered a physiological drought. Frequently, there are no shared species between the beach and the more sheltered or inland areas of the dunes. Experiments on sea rockets (*Cakile maritima*) and lupines (*Lupinus* spp.) which were sprayed with seawater showed that lupine seedlings were not tolerant of salt spray. The level of salt spray in a Californian beach may be 1 $mg\,cm^{-2}\,d^{-1}$ on a calm day, but is much higher on a windy day. On other beaches and dunes, where onshore winds are not as strong, airborne spray is very low and plants are not subjected to these conditions.

Water

The primary source of water for dune plants is rainfall. Radiation causes considerable diurnal and nocturnal temperature variations. These fluctuations in soil temperature are sufficient to cause the periodical condensation of water vapor in the soil. This produces an increase in water availability from dew that is sufficient to maintain plants in rainless periods. Fog can be another source of water, but in some areas it contains salt. Studies in open dune communities have shown that soil moisture increases to depths of about 60 cm below the dune surface and then tends to fall off. In closed dune communities, where the soil has some organic matter, rainfall is absorbed and held near the surface where it is available to roots. Experiments with *Chamaecrista chamaecristoides* seedlings, a species that thrives in mobile dunes, showed that they had the ability to withstand total lack of watering for more than 80 days. This probably allows them to survive during the dry months of the year in the dunes of the Gulf of Mexico.

In a wet year, there may be widespread flooding in dune depressions. Blowouts are wind hollows or basins of exposed sand within dunes, called slacks or depressions. They frequently arise through the erosion of deflated areas in poorly vegetated dunes. The deflation limit is reached owing to the presence of water, algae, or the accumulation of coarse immovable material. Deposition occurs around the borders of the blowout and vegetation may recolonize the area. The water table falls during the dry season and recovers during the rainy months and the composition of the plant community reflects this groundwater regime. When the soil is completely flooded, the prevailing anaerobic conditions can influence its chemical composition and the concentration of nutrients, affecting plant survival and growth. Flooding can cause the local extinction of some non-wetland species and facilitate the establishment of others.

The frequency and duration of slack inundation are factors that can alter the distribution of vegetation and plant community composition. When flooding takes place occasionally, on very wet years, many of the plants die, and when the water recedes, colonization takes place again. In wet slacks that flood every year, water-loving plants establish and a completely different set of species is found in these areas. Thus community composition will depend on the differential tolerance of plants to the environmental conditions associated with inundation, particularly anoxia. Species are good indicators of the water table depth. In

temperate areas, *Erica tetralix*, *Glyceria maxima*, *Carex nigra*, and *Juncus effusus* are some of the more common species. In Europe, slacks are very important for endemic and rare species. In tropical regions, *Cyperus articulatus*, *Lippia nodiflora*, *Hydrocotyle bonariensis* (Mexico) and *Paspalum maritimum*, *Fimbrisitylis bahiensis*, *Marcetia taxifolia* (Brazil) are frequently found in these depressions. Thickets are also common and in Mexico they are formed by *Pluchea odorata*, *Chrysobalanus icaco*, and *Randia laetevirens*. In Brazil, there is Ericaceae scrub dominated by *Humiria balsmifera*, *Protium icicariba*, and *Leucothoe revoluta*.

Temperature

On open sand dunes, there are considerable diurnal and nocturnal temperature variations. In California, on an August day, when the air temperature was above 15.5 °C 1 m above the ground, the soil surface was at 38 °C and soil 15 cm below the surface was at 19 °C. In a Nevada desert, the soil surface temperature reaches 65.5 °C and in Veracruz, in the coastal dunes in the central Gulf of Mexico, the soil surface also reaches 65 °C. These temperatures are critical for seed germination and seedling establishment. Some species, such as hard-coated legumes, need these temperature oscillations over several weeks to break the hard seed coat. They lie on the soil during the dry season, and the temperature fluctuations break the testa. When the rains come, they are ready to germinate. Vegetation cover reduces these temperature oscillations considerably. There are also temperature differences over short distances because of topography and orientation. In the dunes of temperate regions, there are temperature and vegetation differences depending on dune slope orientation.

Habitats

Coastal dunes are very dynamic systems offering a wide variety of habitats with different physical and biotic conditions, and this allows for the existence of species with very diverse life-history traits. They can be visualized as a permanently changing environment with distinct degrees of stabilization that is closely correlated with topography, the disturbance produced by sand movement, and distance to the sea. Dune habitats can be classified into three types: (1) those where sand movement dominates, sea spray is sometimes important, and nutritionally poor soils prevail (they are formed by the sandy beach, embryo or incipient dunes, foredunes, blowouts, and active dunes); (2) humid and wet slacks or depressions, that is, those habitats which become inundated during the rainy season when the water table rises and they sometimes may even form dune lakes with wetland vegetation; (3) stabilized habitats, which show no sand movement, conditions are less stressful, and there is more organic matter in the soil. Vegetation cover is more continuous – grasslands, thickets, woodlands, and tropical forests.

Fig. 2 shows a beach and dune topographic profile as well as the intensity of some of the abiotic factors mentioned and the areas where they affect the dune system.

Biological Factors

Dune plants are found all over the world, from the frosty regions of Canada and Patagonia, to the tropical areas of the Caribbean, Africa, and the South East, and the dry regions of Australia, Peru, and California. They are subjected to very different climatic conditions, they share few species, and life forms vary. Raunkaier developed an ecologically valuable system of plant classification, based on the position of the vegetative perennating buds or the persistent stem apices in relation to the ground level during the unfavorable season of the year, which can be either the cold winter or the dry summer. There is a strong correlation between the climate of an area and the life forms of the plants present. This system allows comparisons to be made between particular areas or regions. The biological spectrum found in a dune system is an expression of the number of species in each life-form class as a percentage of all the species present. A comparison of the biological spectrum of a dune system in Braunton Burrows (North Devon, Great Britain) and one in La Mancha (Veracruz, Gulf of Mexico) was made. Braunton Burrows is dominated by hemicryptophytic plants (perennating buds are at the surface of the sand) and therophytes (annuals that survive the unfavorable season as seeds); La Mancha is dominated by phanerophytic types (these grow continually, forming stems that often have naked buds projecting into the air, such as in *Hippophae rhamnoides* or *Chamaecrista chamaecristoides*).

Facilitation and Succession

Ecological succession refers to a more-or-less predictable and orderly change in the composition or structure of an ecological community. Facilitation is one of the mechanisms by which succession takes place. It occurs when plant establishment is favored or facilitated by previously established plant communities that ameliorate environmental extremes. Physical factors in dune environments produce a very harsh environment where few plants can survive. Several studies show that facilitation takes place in the early stages of colonization and succession in coastal dunes. As succession proceeds, pioneer species will tend to be replaced by more competitive species, the abiotic environment will become less harsh, and biotic interactions such as competition and predation will be more common.

Dune succession is comprised of a pioneer (also called yellow dunes, associated with the most seaward dunes that are still receiving a significant input of wind-blown sand), intermediate, and mature stages (gray dunes or inland dunes with little or no

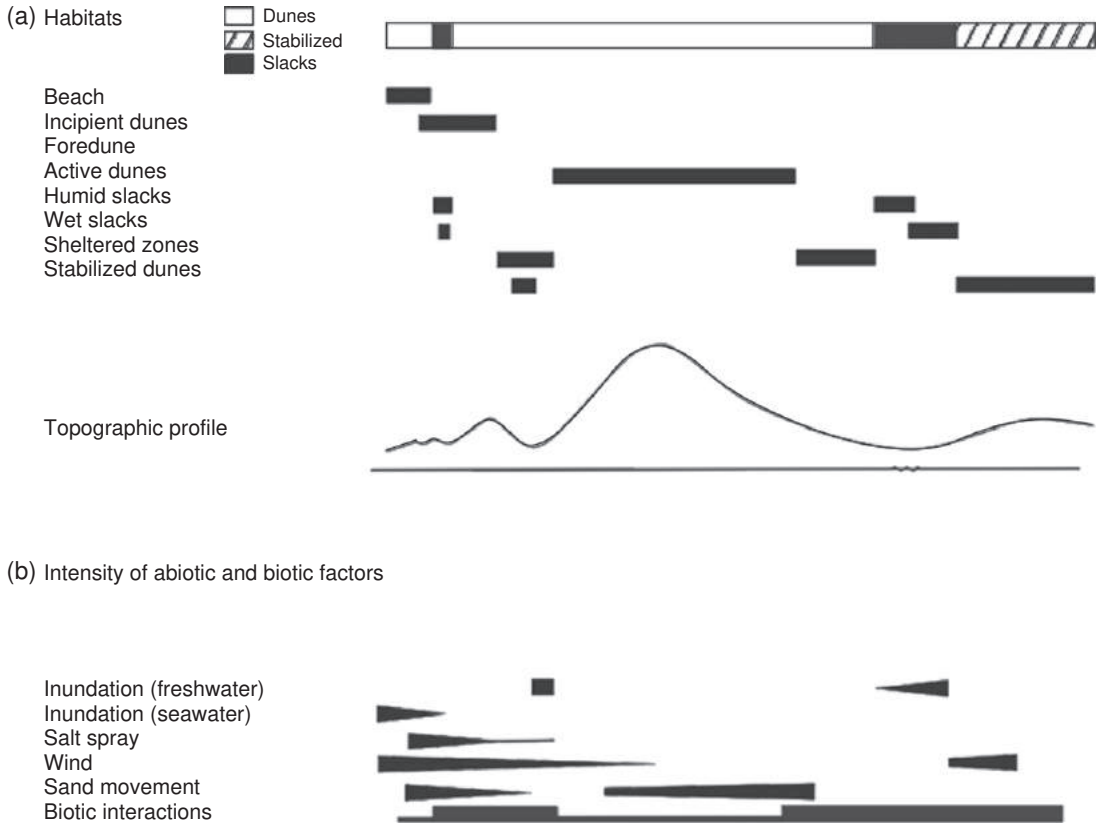


Fig. 2 (a) Beach and dune topographic profile showing each of the habitats. (b) Intensity, indicated by the width of the line, of some of the abiotic factors mentioned along with the areas where they affect the dune system. Reproduced with permission from Moreno-Casasola P and Vázquez G (2006) Las comunidades de las dunas. In: Moreno-Casasola P (ed.) *Entornos veracruzanos: La costa de La Mancha*. Xalapa, Mexico: Instituto de Ecología AC.

sand, a high humus content, and where soil development has occurred). The rate of succession varies with the harshness of the environment. This is related to the abiotic factors mentioned and the vegetation stock. Detailed studies have been undertaken in the Lake Michigan dunes, a salt-free system, in the coastal dunes of Newborough Warren, several sites in Holland, and La Mancha, among others.

Competition, Predation, Disease

Biotic interactions among plants are an important determinant of structure and dynamics. Competition is recognized as one of the most important forces structuring ecological communities. Competition is the interaction of organisms or species such that, for each, the birth or growth rate is depressed and the death rate increased by the presence of the other organisms (or species). Well-known examples of competition between plants growing on coastal dunes are grass and shrub encroachment and the invasion of exotic species.

Grass encroachment occurs when aggressive and competitive grasses spread over dune areas, reducing biodiversity because of the dominance of a few species. Grass encroachment is found in many dune areas, where grasslands become the dominant community type. Among the more aggressive species are *Calamagrotis epigejos*, *Ammophila arenaria*, and *Schizachyrium scoparium*. Shrub encroachment is also common, for example, in the Caribbean (*Coccoloba uvifera*).

Species introduction has been a common practice in dunes, both for dune stabilization and for cattle ranching activities. European marram grass was widely dispersed in other regions that were quite different from its native Europe, mainly to fix sand dunes. Several conifers have also been used for example in Doñana's dune system in southern Spain. African grasses (e.g., *Panicum maximum*) have been brought to America and used to replace local grass species because they have been considered better fodder.

Neither the effects of fauna nor those of grazing animals (especially rabbits) on dunes have received the attention they deserve. The importance of herbivory by rabbits was seen in Great Britain during the outbreak of myxomatosis, a viral disease which infects rabbits. The disappearance of rabbits led to profound changes in the structure of the vegetation, mainly the development of scrub in several dune areas. Rabbits also produce nitrogen and phosphorus enrichment beneath the scrub species under which they find shelter, causing N-fixing root nodules to invade.

Lethal yellowing is a specialized bacterium, an obligate parasite that attacks many species of palms, including the coconut palm (which has become the symbol of tropical beaches). Extensive coconut plantations in the Tropics have been abandoned because of coconut dieback, and shrub encroachment has taken place.

Symbiotic Relations

Symbiotic associations involving nitrogen fixation by microorganisms are frequent in dunes. There are nitrogen-fixing bacteria such as *Rhizobium*, which form a symbiosis with numerous forbs and shrubs in temperate and tropical dune systems. Some of the plants showing nodules are *Ulex europaeous*, *Trifolium* spp., *Lupinus arboreus*, and *Hippophae rahmnoides* in Europe, *Acacia* shrubs in South Africa, and *Chamaecrista chamaecristoides* in Mexico.

In foredunes and mobile dunes, pioneer grasses such as *Ammophila*, *Elytrigia*, and *Uniola* show different degrees of infection by vesicular-arbuscular mycorrhizae (VA). The major benefit to these grasses is probably enhanced phosphorus uptake under conditions of phosphorus limitation. They also help in the aggregation of sand particles. Tropical sand dune plants also frequently show symbiosis with mycorrhizae.

Sand dunes are harsh environments where abiotic factors act as filters that determine species survival. The interactions between abiotic and biotic factors in sand dunes change as dunes mature. They are delicate systems in which plant cover, formed by different vegetation structures and species assemblages, maintains the system in a stabilized condition. Higher diversity is found when there are different habitats. Today, these fragile systems are endangered and the urbanization of the coast is increasing. We must find ways to make our activities and dune conservation compatible.

See also: Behavioral Ecology: Herbivore-Predator Cycles. Ecological Complexity: Cybernetics; Goal Functions and Orientors. Ecosystems: Deserts. General Ecology: Temperature Regulation. Global Change Ecology: Climate Change 2: Long-Term Dynamics; Biosphere: Vernadsky?s Concept

Further Reading

Barbour, M.G., Craig, R.B., Drysdale, F.R., Ghiselin, M.T., 1973. Coastal Ecology: Bodega Head. Los Angeles, CA: University of California Press.
Carter, R.W.G., 1988. Coastal Environments: An Introduction to the Physical, Ecological and Cultural Systems of the Coastlines. New York: Academic Press.
Hesp PA (2000) Coastal sand dunes: Form and function. Massey University Coastal Dune Vegetation Network, New Zealand, Technical Bulletin No. 4, 28pp.
Lortie, C.J., Cushman, J.H., 2007. Effects of a directional abiotic gradient on plant community dynamics and invasion in a coastal dune system. Journal of Ecology 95 (3), 468–481.
Martínez, M.L., Psuty, N.P. (Eds.), 2004. Coastal Dunes: Ecology and Conservation. Berlin: Springer.
Moreno-Casasola, P., Vázquez, G., 2006. Las comunidades de las dunas. In: Moreno-Casasola, P., (Ed.), Entornos veracruzanos: La costa de La Mancha. Xalapa, Mexico: Instituto de Ecología AC.
Olson, J.S., 1956. Rates of succession and soil changes on southern Lake Michigan sand dunes. Botanical Gazette 199, 125–170.
Packham, J.R., Willis, A.J., 1997. Ecology of Dunes, Salt Marshes and Shingle. London: Chapman and Hall.
Pilkey, O.H., Neal, W.J., Riggs, S.R., *et al*., 1998. The North Carolina Shore and Its Barrier Islands: Restless Ribbons of Sand. London: Duke University Press.
Ranwell, D.S., 1972. Ecology of Salt Marshes and Sand Dunes. London: Chapman and Hall.
Rico-Gray, V., 2001. Encyclopedia of Life Sciences: Interspecific Interaction. New York: Macmillan Publishers.
Seeliger, U. (Ed.), 1992. Coastal Plant Communities of Latin America. New York: Academic Press.
Van der Maarel, E., (Ed.), 1993. Dry Coastal Ecosystems, vol. 2A. Amsterdam: Elsevier.
Van der Maarel, E., (Ed.), 1994. Dry Coastal Ecosystems, vol. 2B. Amsterdam: Elsevier.
Van der Maarel, E., (Ed.), 1997. Dry Coastal Ecosystems, vol. 2C. Amsterdam: Elsevier.

Ecosystems☆

Brian D Fath, Towson University, Towson, MD, United States and International Institute for Applied Systems Analysis, Laxenburg, Austria

Introduction

Ecology is a broad and diverse field of study. One basic distinction in ecology is between autecology and synecology, in which the former considers the ecology of individual organisms and populations, mostly concerned with the biological organisms themselves; and the latter, the ecology of relationships among the organisms and populations, which is mostly concerned with communication of material, energy, and information of the entire system of components. In order to study an ecosystem, one must have knowledge of the individual parts; thus, it is dependent on fieldwork and experiments grounded in autecology. However, the focus is much more on how these parts interact, relate to, and influence one another including the physical environmental resources on which life depends. Ecosystem ecology, therefore, is the implementation of synecology. In this manner, the dimensional units used in ecosystem studies are usually the amount of energy or matter moving through the system. This differs from population and community ecology studies in which the dimensional units are typically the number of individuals (**Table 1**).

History of the Ecosystem Concept

The term ecosystem, which is ubiquitous today, both as scientific terminology and in common vernacular, was first used by Arthur Tansley in 1935 in a seminal paper in the journal *Ecology*, entitled "The use and abuse of vegetational concepts and terms." In fact, his reason for coining the term "ecosystem" was in response, as the title says, to a perceived abuse of community concepts by some, such as Clements, Cowles, and Phillips, who interpreted an ecological community as having overt organismal-like properties. The community as organism metaphor bothered Tansley to the extent that he wanted to provide a more scientific footing for the processes and interactions occurring during community development. Tansley describes the ecosystem thusly, "…the fundamental conception is… the whole system, including not only the organism-complex, but also the whole complex of physical factors forming what we call the environment of the biome—the habitat factors in the widest sense." The definition he proposed over 80 years ago sounds fresh today, since it has changed little if at all. The major tenets of this approach are the explicit inclusion of nonliving processes interacting with the biota—in this sense it is more along the Haeckelian lines of ecology than the Darwinian, with an additional emphasis on the system. The latter tied the field closer to the burgeoning disciplines of general system theory and system analysis, and later complex systems theory and socio-ecological metabolism.

While the conceptual underpinning of the ecosystem was now established, the introduction of this term was theoretical, lacking guidance as to how it might be applied as a field of study. There were around this time several whole system energy budgets being developed, particularly for lake ecosystems by North American ecologists such as Juday and Birge in Wisconsin, which were ideal test cases for the ecosystem concept. Building on this work, in 1942, Lindeman's study of Cedar Bog Lake also in Wisconsin was published, providing, for the first time, a clear application of the ecosystem concept. In addition to constructing the food cycle of the aquatic system, he developed a metric—now called the Lindeman efficiency—to assess the efficiency of energy movement from one trophic level to the next based on ecological feeding relations. His conceptual model of Cedar Bog Lake included passive flows to detritus, but these were not included in the trophic enumeration. Since then numerous additional studies have followed this same approach, applying it to many habitats such as terrestrial, aquatic, and urban ecosystems.

Defining an Ecosystem

As stated above, an ecosystem is comprised of the ecological community and its interactions with the nonliving environment. This is often referred to as the interaction of the biotic and abiotic aspects of the ecosphere, however, the term abiotic does a disservice to the overwhelming influence that life has on the planet it inhabits. Many examples of feedback and biotic conditioning of the environment exist such as soil formation and erosion, an oxygen atmosphere and the protective ozone layer, and a balanced carbon cycle. Even simple factors such as temperature, humidity, and soil pH are biologically-mediated leading one to consider that a better term for these features is *conbiotic* rather than abiotic (see entry in this volume). An ecosystem, as a unit of study, must be a bounded system, yet the scale can range from a puddle, to a lake, to a watershed, to a biome. Indeed, ecosystem scale is defined more by the functioning of the system than by any checklist of constituent parts and the scale of analysis should be determined by the problem being addressed. Whereas, individuals perish over time, and even populations cannot survive indefinitely—none can fix their own energy and process

☆*Change History*: October 2017. Fath provided minor edits in the text.

This is an update of B.D. Fath, Ecosystem Ecology, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 1125–1131.

Table 1 Typical dimensional units of study at different ecological scales

Ecological scale	*Dimensions*
Organismal ecology	d*E*/d*t*
Population ecology	d*N*/d*t*
Community ecology	d*N*/d*t*
Ecosystem ecology	d*E*/d*t*

d*E*/d*t*, change in energy over time; d*N*/d*t*, change in number over time.

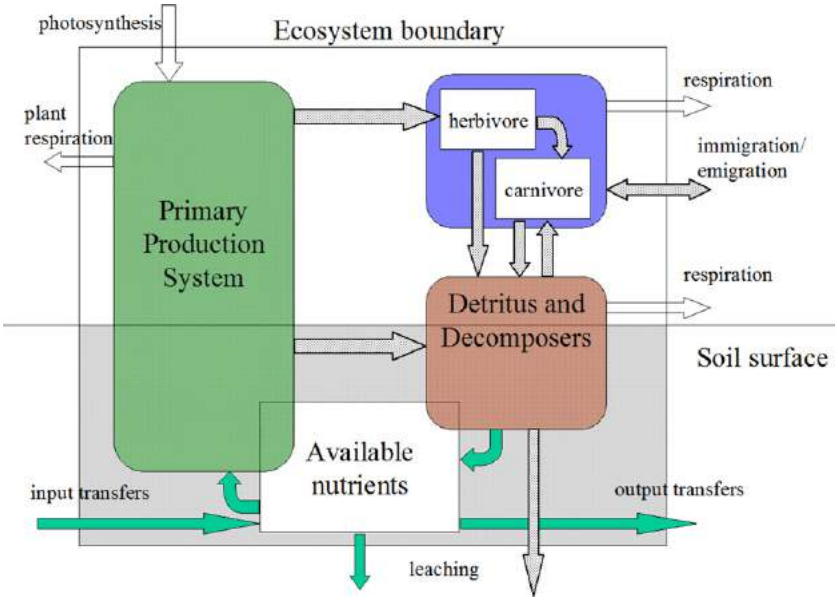


Fig. 1 Conceptual diagram of a simplified ecosystem.

their own wastes—every ecosystem contains the ecological community necessary for sustaining life: primary producers, consumers, and decomposers, and the physical environment for oikos (Fig. 1 shows a simple ecosystem model). It is this feature of ecosystems, that they are the basic unit for sustaining life over the long-term, which provides one of the main reasons for studying them for environmental management and conservation. The two main features of the ecosystem, energy flow and nutrient biogeochemical cycling, comprise the major areas of ecosystem ecology research.

Energy Flow in Ecosystems

The thermodynamic assessment of an ecosystem starts with the recognition that an ecosystem is an open system, in the sense of physics, such that it receives energy and matter input from outside its borders and transfers output back to this environment. Thus, every ecosystem has a system boundary and is embedded in an environment that provides low-entropy energy input and can receive high-entropy energy output. In addition to the external resource source–sink, there is another internal, within system boundary environment with which each organism directly and indirectly interacts. Patten proposed the concept of these two environments, one external and mostly unknowable—other than the input–output interactions, and the second internal and measurable—that is, external to the specific organismal component but within system boundary, as a systems approach to quantify indirect, yet within system interactions. This approach—called environ analysis—relying on the methodologies of input–output analysis, has developed into a powerful analysis tool for understanding complex interactions and dependencies in ecological networks. For now though, let us concern ourselves more generally with what occurs within the ecosystem boundary.

Energy flow in ecosystems begins with the capture of solar radiation by photosynthetic processes in primary producers. (Note, there are also chemoautotrophs that capture energy in the absence of sunlight, but while biologically fascinating, contribute negligible energy flux to the overall global ecological energy balance. Their significance may be in being evolutionarily earlier forms of energy capture.)

$$\text{Energy} + 6CO_2 + 6H_2O \rightarrow C_6H_{12}O_6 + 6O_2 \quad (1)$$

The accumulated organic matter, first as simple sugars then combined with other elements to more complex molecules, represents the gross primary production in the system, some of which is released and used for the primary producers' growth and maintenance through respiration.

$$C_6H_{12}O_6 + 6O_2 \rightarrow 6CO_2 + 6H_2O + \text{Energy} \quad (2)$$

The reminder, or net primary production is available for the rest of the ecosystem consumers including decomposers. Secondary production refers to the energetic availability of the heterotrophic organisms, which accounts for the energy uptake by heterotrophs and the energy used for their maintenance. Overall ecosystem production is supported by the primary producers, whereas ecosystem respiration includes the metabolic activity of all the ecosystem biota (**Table 2**). In this manner, plants provide the essential energetic-basis for all ecological food webs. Since it is often difficult to make direct measurements of ecological production, the change in biomass growth can be used as representative of production.

The captured energy moves through a reticulated network of interactions forming the complex dependency patterns known as food webs. In a simplified food chain, and as first described by Lindeman, the trophic concept is used to assess the distance away from the original energy importation, but in reality the multiple feeding pathways found in ecological food webs make discrete trophic levels a convenient yet inaccurate simplification. Elton observed that one typically finds a decreasing number of organisms as one proceeds up the food chain from primary producers to herbivores, carnivores, and top carnivores—leading him to propose a pyramid of numbers. One can control for the individual variation in body size by considering the biomass at each trophic level rather than the number of individuals—resulting in a pyramid of biomass. The trophic pyramid is a thermodynamically satisfying view of interactions since according to the Second Law energy must be lost during each transformation step; plus energy is used at each level for the maintenance of that level. Under this paradigm, the trophic levels apparently cap out around five or six levels. Fractional trophic levels have been employed to account for organisms feeding at multiple levels, but even these do not usually account for the role of detritus and decomposition, which extend the feeding pathways to higher numbers. However, instead of linking detritus as a source compartment in the ecosystem conceptual model, the standard paradigm is to envision two parallel food webs one with primary producers as the base, and the other with detritus as the base without any input from the rest of the web. If detritus were properly linked as both a source and sink in the ecosystem, then it would be clear that longer energy pathways are possible, if not common. The longer energy flow pathways observed in some studies are not in conflict with the laws of thermodynamics, but they show that ecosystems are more thorough at utilizing the energy within the system, mostly by decomposers, before it is lost as degraded, unavailable energy.

Energy resources flowing through the ecosystem are necessary to maintain all growth and development activities. Organisms follow a clear life history pattern, and while the time scales differ depending on the species, early-stage energy availability is usually used for growth, while later energy surplus is used for maintenance or reproduction. A similar pattern is visible in ecosystem-level growth and development. Net primary production is used to build biomass and physical structure of the ecosystem. The additional structure of photosynthetic material allows for the additional import of solar energy until saturation is reached at about 80% of the available solar radiation. At this point, the overall growth of the ecosystem begins to level off because although gross primary production is high, the overall system supports more and more nonphotosynthetic biomass both in terms of nonphotosynthetic plant material and heterotrophs. When the average gross production is entirely utilized to support and maintain the existing structure, net production is zero and the system has reached a steady state regarding biomass growth. However, the ecosystem continues to develop both in terms of the network organization and in the information capacity. In addition to being a dynamic steady state, it does not persist indefinitely because disturbances afflict the system setting it back to earlier successional stages in which the growth and development processes begins anew, possibly with different results. In this manner, the disturbance acts according to Holling's creative destruction (see entry on Holling Cycle) providing the system the opportunity to develop along a different pathway. Recent work on ecosystem growth and development has focused on the orientation of thermodynamic indicators such as energy throughflow, energy degradation, biomass, work energy capacity, and specific entropy. These orientors provide good system-level indicators of development during succession or restoration of impaired ecosystems.

Biogeochemical Cycles

A useful adage to remember regarding ecosystems is that there are no trashcans in nature. All material is source for some other process. Therefore, another major focus of ecosystem ecology is understanding how the chemical elements necessary for life persist and translocate in pools and fluxes within the ecosphere. The biosphere actively interacts with the three nonliving spheres (hydrosphere, atmosphere, and lithosphere) to provide the available concentration of each for life. This action has a significant impact on the relative distribution of these elements. The simple sugar products of photosynthesis, $C_6H_{12}O_6$, are the base for organic matter so carbon, hydrogen, and oxygen dominate the composition of life, and while oxygen is available in the

Table 2 Ecosystem energetics defined by net and gross production

Net primary production = gross primary production − respiration (autotrophs)
Net secondary production = gross secondary production − respiration (heterotrophs)
Net ecosystem production = gross primary production − ecosystem respiration (autotrophs + heterotrophs)
Net production = biomass (now) − biomass (before)

lithosphere, and hydrogen in the hydrosphere, carbon is actually quite scarce in the environment, making the disproportionate amount of carbon in biomass a hallmark of life. In fact, there are about 20 elements used regularly in living organisms, of which 9 are called the macronutrients as the major constituents of organic matter: hydrogen, oxygen, carbon, nitrogen, calcium, potassium, silicon, magnesium, and phosphorus. Some of these elements are readily available in the abiotic environment, in which case conservation through cycling of the elements is not paramount, however those in scarce supply, such as nitrogen and phosphorus, are reused many times before being released from the system (**Table 3**). These biogeochemical cycles provide the foundation to understand how human modification leads to eutrophication (N and P cycles) and global climate change (C cycle). Therefore, much effort has been made to study and understand these cycles, particularly the carbon, nitrogen, and phosphorus cycles, details of which are addressed elsewhere in this encyclopedia.

Ecosystem Studies

The ecosystem perspective achieved footing in the ecological academic community since it was central to Gene Odum's seminal textbook *Fundamentals of Ecology* first published in 1953. An early implementation of this approach at the institutional scale was attempted was in the International Biological Program (IBP), which was run from 1964 to 1974. The program had many successes in assessing and surveying the earth's ecosystems, but faced the difficulty of compelling a top–down, holistic research paradigm on individual scientific endeavors. As a result of this conflict, the program did not deliver as much as had been hoped, but set the stage for the next generation of ecosystem-scale research. One feature of the IBP that did continue was the use of computer simulation modeling as a tool to understand the complex ecological interrelations. The journal, *Ecological Modeling and Systems Ecology*, was started in 1975 and continues as an active outlet for mathematical and computer-based ecosystem research.

Subsequent to the IBP, the U.S. National Science Foundation officially established the Long-term Ecological Research Sites (LTER) in 1980 but research at several of the sites dates much earlier. Currently, there are 30 such sites ranging from the Coweeta Hydrological Lab in North Carolina, Hubbard Brook Ecosystem Study in New Hampshire, Sevilleta National Wildlife Refuge in New Mexico, and the Baltimore Ecosystem Study (lternet.edu/lter-sites). These projects rely on a vast team of scientists to study the many interactions at this spatial scale. Still, the difficulty lies in putting together all the pieces into an integrated whole picture of the ecosystem.

Smaller-scale, individual-led ecological research is commonly conducted using microcosm and mesocosm experiments. A mesocosm experiment uses designed equipment or enclosures in which environmental factors can be controlled and manipulated to approximate natural conditions. The prevalence of this approach created a wealth of small-scale experimentation but at the expense of larger observational studies, which sparked a fierce debate in the 1990s between the "field" versus "bottle" approach. Indeed, the usefulness of microcosm experiments for ecosystem ecology was brought into question, but the resolution has been that a multiplicity of approaches is useful to address ecological questions.

Biomes

Specific ecosystem characteristics are variable across the globe depending on the location and conditions. Tansley discussed in detail the factors that go into forming a climax community such as edaphic or physiographic. A simple formulation considers the regions' climatograph, a combination of temperature and precipitation and from those two variables gives a good indication of the terrestrial ecosystem (**Table 4**). These climatic conditions determine whether the regions are tropical or temperate, trees or grasses, providing a rich display of ecosystems across the globe.

Human Influence on Ecosystems

Humans have greatly altered and impacted the global biosphere. We recognize now the importance of maintaining functioning ecosystem services both out of our own necessity and for the obligation we have to the ecosphere. In 2000, the United Nations

Table 3 Percentage atomic composition of the biosphere, hydrosphere, atmosphere, and lithosphere for first 10 elements

Biosphere		*Hydrosphere*		*Atmosphere*		*Lithosphere*	
H	49.8	H	65.4	N	78.3	O	62.5
O	24.9	O	33.0	O	21.0	Si	21.22
C	24.9	Cl	0.33	Ar	0.93	Al	6.47
N	0.073	Na	0.28	C	0.04	H	2.92
Ca	0.046	Mg	0.03	Ne	0.002	Na	2.64
K	0.033	S	0.02			Ca	1.94
Si	0.031	Ca	0.006			Fe	1.92
Mg	0.030	K	0.006			Mg	1.84
P	0.017	C	0.002			K	1.42

Table 4 Basic characteristics of major terrestrial biomes

Biome	*Precipitation (cm)*	*Temperature (°C)*
Tundra	<25	−12
Taiga	35–100	10
Temperate deciduous forest	75–150	0–30
Temperate rain forest	200–400	9–12
Tropical rain forest	200–600	25
Desert	<25	35
Grassland	25–75	9–25

Table 5 A few of the tends identified in the Millennium Ecosystem Assessment

50% of all the synthetic nitrogen fertilizer ever used has been used since 1985	20% of the world's coral reefs were lost and 20% degraded in the last several decades
60% of the increase in the atmospheric concentration of CO_2 since 1750 has taken place since 1959	35% of mangrove area has been lost in the last several decades
Approximately 60% of the ecosystem services evaluated are being degraded or used unsustainably	Withdrawals from rivers and lakes doubled since 1960

Table 6 Ecosystem approach principles of the convention on biological diversity

The following 12 principles are complementary and interlinked

Principle

1 The objectives of land, water, and living resource management are a matter of societal choices
2 Management should be decentralized to the lowest appropriate level
3 Ecosystem managers should consider the effects (actual or potential) of their activities on adjacent and other ecosystems
4 Recognizing potential gains from management, there is usually a need to understand and manage the ecosystem in an economic context. Any such ecosystem-management program should:
   (a) Reduce those market distortions that adversely affect biological diversity
   (b) Align incentives to promote biodiversity conservation and sustainable use
   (c) Internalize costs and benefits in the given ecosystem to the extent feasible
5 Conservation of ecosystem structure and functioning, in order to maintain ecosystem services, should be a priority target of the ecosystem approach
6 Ecosystem must be managed within the limits of their functioning
7 The ecosystem approach should be undertaken at the appropriate spatial and temporal scales
8 Recognizing the varying temporal scales and lag-effects that characterize ecosystem processes, objectives for ecosystem management should be set for the long term
9 Management must recognize the change is inevitable
10 The ecosystem approach should seek the appropriate balance between, and integration of, conservation and use of biological diversity
11 The ecosystem approach should consider all forms of relevant information, including scientific and indigenous and local knowledge, innovations, and practices
12 The ecosystem approach should involve all relevant sectors of society and scientific disciplines

Secretary General called for a global ecological assessment, which was recently published as the Millennium Ecosystem Assessment (MEA) (www.maweb.org/en/index.aspx). The report compiled by over 1350 experts from 95 countries found that humans have changed ecosystems more rapidly and extensively over the last 50 years than in any comparable period of time in human history, resulting in a substantial and largely irreversible loss in the diversity of life on Earth (other highlights from the report are presented in **Table 5**). The MEA operated within a framework that identified four primary ecosystem services needed by humans: supporting (nutrient cycling, primary production, soil formation, etc.), provisioning (food, water, timber, fuel, etc.), regulating (climate, flood, disease, etc.), and cultural (aesthetic, spiritual, educational, recreational, etc.). All have shown signs of stress and human pressures during the past century. One positive trend was the increase in food production (crops, livestock, and aquaculture), but this occurred with a concomitant loss of wild fisheries and food capture, along with a substantial increase in the resource inputs required to maintain the high agricultural production. While these observed changes to ecosystems have contributed to substantial net gain in human well-being and economic development, they have come at an increasing cost to the ecosystem health. The loss of this natural capital is typically not properly reflected in economic accounts.

Since the ecosystem provides the necessary functions for life, environmental management principles being devised and implemented today use the ecosystem concept as foundation. In particular, there have been several high profile international efforts such as with the Convention on Biological Diversity (CBD), a treaty initiated in 1992 and signed by 150 government leaders with the express aim to protect and promote biological diversity and sustainable development. The *Ecosystem Approach* adopted within this convention uses scientific methodologies regarding ecological interactions among organisms, their environment, and human activity to promote conservation, sustainability, and equity for managing natural resources. The approach deals with the complex socio-ecological-economic systems by promoting integrated assessment and adaptive management (see entry on Panarchy). The Ecosystem Approach of the CBD is outlined below in 12 principles (**Table 6**). Note particularly principles 5–8 that deal with ecosystem functioning, and taken in the context of the other principles assert how this ecological functioning provides opportunities and constraints for economic and social well-being. Better understanding these issues, such as ecosystem services, scale, time-lags, and dynamics are paramount research areas today.

At a national scale, in the United States, the Endangered Species Act of 1973 was critical legislation that puts into place a recovery plan for any species listed as endangered or threatened. What is impressive is that the legislation rests firmly on an ecosystem approach when it states: "The purposes of this Act are to provide a means whereby the ecosystems upon which endangered species and threatened species depend may be conserved…" This holistic approach was already evident in the Organic Act of 1916, which codified the process of designating National Parks (the first, Yellowstone, was established in 1872) with the "purpose is to conserve the scenery and the natural and historic objects and the wild life therein and to provide for the enjoyment of the same in such manner and by such means as will leave them unimpaired for the enjoyment of future generations." The US Wilderness Act of 1964 provided additional protection to natural places stating they are areas "where the earth and its community of life are untrammeled by man, where man himself is a visitor who does not remain". It is encouraging to see scientific understanding and perspective referenced in legislation. Unfortunately, all of these protections are under threat from politics and businesses that undervalue ecosystems and their services.

Summary

Ecosystems are a unit of organization that include the interactions of the ecological community with its nonliving environment, primarily in terms of the energy flow and nutrient cycling. Research in ecosystem ecology has given us a much better understanding of the processes and functions necessary to sustain life. The work in the natural sciences has outpaced the ability of the social institutions to adapt and implement this knowledge. However, there is reason to be optimistic because the recent focus on the ecosystem approach in major international efforts recognizes that humans, with their cultural diversity, are an integral component of ecosystems.

See also: Conservation Ecology: Protected Area. Ecological Data Analysis and Modelling: Conceptual Diagrams and Flow Diagrams. Evolutionary Ecology: Red Queen Dynamics; Metacommunities. General Ecology: Keystone Species and Keystoneness; Seed Dispersal; Temperature Regulation. Terrestrial and Landscape Ecology: Ecological Engineering: Design Principles

Further Reading

Chapin III, F.S., Matson, P.A., Vitousek, P.M., 2010. Principles of terrestrial ecosystem ecology, 2nd edn New York: Springer, p. 529.
Fath, B.D., Jørgensen, S.E., Patten, B.C., Straškraba, M., 2004. Ecosystem growth and development. Biosystems 77, 213–228.
Golley, F.B., 1993. A history of the ecosystem concept in ecology. New Haven, CT: Yale University Press.
Jørgensen, S.E., Fath, B.D., Bastianoni, S., Marques, J.C., Müller, F., Nielsen, S.N., Patten, B.C., Tiezzi, E., Ulanowicz, R.E., 2007. A new ecology: Systems perspective. Amsterdam: Elsevier, p. 275.
Likens, G.E., Borman, F.H., Johnson, N.M., Fisher, D.W., Pierce, R.S., 1970. Effects of forest cutting and herbicide treatment on nutrient budgets in the Hubbard Brook watershed-ecosystem. Ecological Monographs 20, 23–47.
Lindeman, R.L., 1942. The trophic-dynamic aspect of ecology. Ecology 23, 399–418.
Odum, H.T., 1957. Trophic structure and productivity of Silver Springs, Florida. Ecological Monographs 27, 55–112.
Odum, E.P., 1969. The strategy of ecosystem development. Science 164, 262–270.
Patten, B.C., 1978. Systems approach to the concept of environment. The Ohio Journal of Science 78, 206–222.
Raffaelli, D.G., Frid, C.L.J., 2010. Ecosystem ecology: A new synthesis. Cambridge: Cambridge University Press, p. 173.
Tansley, A.G., 1935. The use and abuse of vegetational concepts and terms. Ecology 16, 284–307.
Weigert, R.G., Owen, D.F., 1971. Trophic structure, available resources and population density in terrestrial versus aquatic ecosystems. Journal of Theoretical Biology 30, 69–81.

The Boreal Forest Ecosystem☆

Donald L DeAngelis, University of Miami, Coral Gables, FL, United States

Introduction

The boreal biome is the largest of all terrestrial biomes, amounting to roughly 15×10^6 km^2, with estimated storage of about 195 billion (195×10^9) metric tons of carbon (C) in aboveground living pools, which is about one-third of the total terrestrial carbon. Approximately three times that amount is stored in soil. The boreal forest biome is also referred to as the "taiga" (Russian) for "swamp forest." Geographically, the boreal forest is located between latitudes 45° and 70° N, and virtually all of it in Canada, Alaska, and Siberia, with portions in European Russia and Fenno-Scandia. The boreal forest is bordered on the north by treeless tundra and on the south by mixed forest. The boreal forest is termed a "biome" by ecologists, a term that refers to a biogeographic unit that is distinguished from other biomes by the structure of its vegetation and dominant plant species. A biome is the largest scale at which ecologists classify vegetation. All parts of a biome tend to be within as the same climatic conditions, but because local conditions differ, a biome may encompass many specific ecosystems (e.g., peatlands, river floodplains, uplands) and plant communities. Despite this diversity within a biome, in referring to the boreal forest we will here use the terms "biome" and "ecosystem type" interchangeably.

Climate and Soils

The climate of the boreal forest is continental and, importantly for the growing season, there tend to be between 30 and 150 days of temperatures above 10°C. Temperature lows can fall below −25°C. Average annual precipitation is 38–50 cm, with the lowest amounts in the northern boreal forest, and greater frequency of precipitation during the summer season. Water is seldom limiting because of the generally flat topography and low rate of evaporation.

Permafrost can occur in the northern parts of this zone, the southern limit coinciding roughly with a mean air temperature of −1°C and snow depth of about 40 cm. The zone of permafrost generally starts at depths ranging from 1.5 to 3 m in the areas of the boreal forest where it occurs. Its occurrence limits soil processes to an upper active layer and impedes water drainage, leading to waterlogged soils. The soil decomposition rate in the taiga is slow, which leads to the accumulation of peat.

Several soil types characterize the boreal forest. The soils of a major part of the boreal forest, lying under a dense coniferous canopy, are heavily podzolized where the soil is permeable. These soils consist largely of spodosols. Intense acid leaching forms a light ash-colored eluvial soil horizon leached of most base-forming cations such as calcium. Thus taiga soils tend to be nutrient poor. Gelisols are common in the north, where permafrost occurs. These are young soils with little profile development. Histosols, which are high in organic matter, form in nonpermafrost wetlands, where decomposition is slowed by hypoxic conditions. These are often referred to as peatlands.

Biodiversity

Tree species richness is far smaller than that in the temperate forests to the south, where more than 100 species are typically observed in 2.5° × 2.5° quadrats in eastern United States. Species richness clearly declines from south to north in the taiga. Whereas 40 or more tree species can be found in the southern taiga in Canada, this declines to 10 or so species near the tundra boundary. Animal species also show strong gradients. Reptile and amphibian species are almost nonexistent above 55°. Mammal species richness declines from close to 40 species to about 20 going northward in the boreal forest biome in North America, while bird species decline from about 130 to less than 100.

Forest Structure and Species

Because many hardwood trees are both sensitive to low winter temperatures and require a long and warm summer, the true boreal forest begins where the few remaining hardwoods become a minor part of the forest. Four coniferous genera dominate a major

☆*Change History:* April 2017. Donald DeAngelis updated sections Introduction, Climate and Soils, Climatic variation and effects on vegetation, Landscape-scale vegetation differences, Herbivorous insect outbreaks, Difference in food webs due to climatic differences, Predator-prey cycling, Fire and its effects, Conservation and global issues Summary.

This is an update of D.L. DeAngelis, Boreal Forest, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 493–495.

part of the taiga; *Picea* (spruce), *Abies* (fir), *Pinus* (pine), and *Larix* (larch). The hardwoods, which largely occur in dwarf form, include *Alnus* (alders), *Populus* (poplars), *Betula* (birches), and *Salix* (willows). The hardwoods tend to be early successional species following disturbances such as fires or erosion/deposition processes on riverbanks, which are eventually shaded out by slower-growing spruces and firs. Much of the main boreal forest is dominated by a few spruce species. These form a dense canopy in the central and southern taiga, with a ground cover of dwarf shrubs, such as cranberries and bilberries, and mosses and lichens. In northern Siberia, huge areas are covered almost solely by larch, and the canopy is much less dense. Pine species, which can withstand a range of harsh conditions, grow in light, sandy soils and other dry areas. As the boreal forest-tundra boundary is approached, conifers thin out to a woodland with lichen and moss dominating the ground. Trees become more and more stunted.

The standing stock of biomass of the boreal forest ranges is estimated at 200 (range 60–400) metric tons per hectare ($t\ ha^{-1}$). This compares with an estimate of $350\ t\ ha^{-1}$ for the temperate deciduous forest and $10\ t\ ha^{-1}$ for the tundra ecosystems. The boreal forest differs from the temperate forest in having a much higher percentage of its total biomass as photosynthetic foliage (7% vs. 1%). It differs from the tundra in having a lower percentage of root biomass (22% vs. 75%).

Climatic Variation and Effects on Vegetation

There is a great deal of climatic variation within the boreal biome. Winter temperatures in the boreal forest of western North America are about 15–20°C colder than in northwestern Europe (Fennoscandia) with half as much precipitation. Across Russia, the climate changes from relatively mild and wet in the Baltic region to extreme continental climate in northeastern Siberia, with mean annual temperatures of -10°C and low precipitation.

The differences in climate between areas of forest at the continental scale bring about differences in vegetation. Typical areas of boreal forest of upland western North America are dominated by white spruce (*Picea glauca*), with lesser amounts of aspens and poplars (*Populus* spp.) in the canopy. In northwestern Europe, Scots pine (*Pinus sylvestris*), Norway spruce (*Picea abies*), and birches (*Betulus* spp.) make up most of the overstory layer. But the main differences are in the understory structure. In North America, there is a relatively tall shrub layer (0.6–2 m high) of willow (Salix spp.) and birches, whereas in northwestern Europe the tall shrub layer is absent. Instead there is a layer of dwarf shrubs (0–0.5 m), mostly ericaceous, such as bilberry (*Vaccinium* spp.). The difference has been attributed to greater tolerance to low winter temperatures of willow and birch shrubs, while smaller shrubs such as bilberry are poor at surviving cold temperatures without a deep snowpack for insulation. The extreme cold and dry conditions of eastern Siberia result in the deciduous needle-leaf conifer, larch (*Larix* spp.) being the most common component of the forest. The deciduous trait allows larches to avoid needle dessication from the extreme winter temperatures.

Landscape-Scale Vegetation Differences

The taiga of interior Alaska is a particular example showing the variability of vegetation communities at the landscape scale. Well-drained south-facing slopes are dominated by communities of white spruce and hardwoods such as birches and aspens, which are adapted to warmer conditions. Along a transect down gentle north-facing slopes, closed black spruce (*Picea mariana*) forest gives way to more open black spruce-*Sphagnum* muskeg with sedge tussocks dominating on areas of stream valley underlain with permafrost. It is also possible for alternative stable states of community to occur in some areas on the same site, depending on the history of that area. This is true of the white spruce and hardwood type community and black spruce community. These communities, both of which can potentially occupy the same sites in the taiga zone, are characterized by different nutrient cycling efficiencies. Black spruce ecosystems have mechanisms that conserve nutrients. Because the litter of black spruce is low in nutrients, it is slow to decompose; thus thick layers of organic matter accumulate on the forest floor under black spruce. This insulates the soil and thus reduces summer soil temperature. Rates of mineralization become even slower on these stands because of the low summer soil temperatures. If a fire disturbance occurs, it may burn off enough of the litter, so that summer soil temperatures are warmer and nutrient cycling is higher. Under these circumstances, white spruce, which needs greater nutrient availability, may be able to invade and dominate. However, any disturbance that reduces nutrient availability to the white spruce may allow some black spruce trees to invade. The black spruces will contribute more low quality (nutrient poor) litter to the ground, causing the insulation factor to increase again, and reducing soil summer temperatures. Eventually, the black spruces may dominate again.

Animals

Animal life in the boreal forest is far less diverse than in most temperate zone ecosystems. A number of bird species are adapted to being residents of the taiga. Grouses such as the capercaillie of the Old World, are adapted to year-round life in the taiga, as are some owls, woodpeckers, tits, nuthatches, crossbills, and crows. Small mammal herbivores of the boreal forest include the squirrels, chipmunks, voles, and snowshoe hares. These provide food for a small number of predators species, including the red fox (*Vulpes vulpes*) and members of the weasel family. The moose (*Alces alces*) (called elk in the Old World) has a wide geographic distribution in the taiga. They are prey for wolves (*Canis lupus*) and occasionally the brown bear (*Ursus arctos*).

One component of the taiga fauna, conspicuous for its frequent devastating effects on thousands of hectares of forest, is that of phytophagous insects. Populations of these insects, which include pine sawflies (*Neodiprion* spp.), spruce budworms (*Choristoneura fumiferana* (Clemens)), bark beetles, and many others that attack conifers, are capable of escaping natural enemies and building up to huge population densities. The large monospecific stands of the boreal forest may be especially vulnerable.

Herbivorous Insect Outbreaks

The boreal forest is known for the occurrence of severe outbreaks of herbivorous outbreaks insects like the spruce budworm. Patterns of outbreaks may be cyclic or irregular (irruptive). Environmental factors can affect outbreaks, and outbreaks of some species of insect have increased in recent times; perhaps due to logging practices and fire suppression, leading to greater densities of tree species preferred by particular insects; for example, there were nine outbreaks of eastern spruce budworm during the nineteenth century, whereas during the first 80 years of the 20th century there were 21, and they were more widespread.

Weather is an important factor in herbivorous insect outbreaks. An example is the spruce budworm, which feeds preferentially on foliage of white spruce and balsam fir (*Abies balsamea*). The caterpillars of this moth do not feed immediately after emergence from the egg, but are dispersed by wind and hibernate over winter. When they emerge in spring they start to feed on *old* fir needles; that is, they are senescence feeders, feeding on needles that are breaking down and releasing nutrients, that is, amino acids, in high concentrations during nutrient translocation. Outbreaks tend to occur when there are large numbers of old trees, where it is likely that many of the needles are old and not very vigorous, so they tend to break down quickly. But the presence of a large stand of old trees is not sufficient in itself to lead to a spruce budworm outbreak. The period of 1948 to 1958 in eastern Canada was a time of summer droughts interspersed with wetter than normal winters. This combination of drying of roots in the summer and waterlogging them in the winter caused dieback of crowns, with rapid aging of leaves and release of high concentrations of amino acids being translocated out of the dying leaves. The extra amino acid in the diets of the spruce budworm caterpillars was enough to increase caterpillar survival, which led to population explosions that predators and parasitoids could not control. DDT was sprayed to stop the outbreaks, but that just made things worse by saving the old stands trees and letting them get older and even more susceptible to outbreaks.

Insect herbivores are estimated to cause more timber losses than fires in the boreal zone. The spruce budworm is the most destructive boreal forest insect which caused as average annual loss during the period of 1982–1987 of 27.3×10^6 ha of forest in eastern North America. Among other defoliator insect pest of the boreal forest, the forest tent caterpillar (*Malacosoma disstria*) caused 2.4×10^6 ha, and the jack pine budworm (*Choristoneura pinus*) a 2.2×10^6 ha annual loss during the same period. Insect damage causes forests to become more susceptible to fire. However, the high numbers of insects during the warm months is a main explanation for the large numbers of birds that migrate from the south to breed in the taiga, especially large numbers of species of warblers and thrushes.

Difference in Food Webs Due to Climatic Differences

The differences in vegetation between western North America and northwestern Europe also lead to differences in the food webs. The tall shrubs of the former provide winter forage for snowshoe hares (*Lepus americanus*) in North American, while the voles (e.g., bank vole (*Myodes and Microtus* spp.)), which are key rodents in Fennoscandia, are better adapted to foraging on the shorter dwarf shrubs that are beneath the snow surface in winter. The snowshoe hare is the prey of relatively specialist predators like the lynx (*Lynx canadensis*), coyote (*Canis latrans*), and great horned owl (*Bubo virginianis*), while mustelids such as the weasel (*Mustela nivalis*) and stoat (*Mustela* ermine) are adapted to feeding on the voles and thus being key predators in northwestern Europe boreal forest. In addition, the snow characteristics differ in the two regions, with deeper and softer snow in western North America, which could have favored the Canadian lynx (*Lynx Canadensis*) as a specialist predator on the snowshoe hare, in contrast with the wetter snow, with hard surface, in northern Europe. The generalist red fox (*Vulpes vulpes*) is better suited to the latter hard-packed snow conditions, and so is an important generalist in northwestern Europe but not western North America. The moose (*Alces alces*)-wolf (*Canis lupus*) predator–prey relationship is important in both systems.

Predator–Prey Cycling

The interaction between a specialist predator and its prey can lead to population cycles, and this occurs between both the snowshoe hare and the Canadian lynx and voles and mustelids. The population cycle of the snowshoe hare is the best studied and most famous. Hare populations exhibit 9–11 year fluctuations in abundance that can be explained in terms of interactions with the lynx, although not all aspects of the cycling are understood. Fire may be an instigator of the cycle, as fires are followed a few decades later by very high levels of edible deciduous vegetation.

Ecosystem Dynamics

In keeping with its position between much warmer climate of the temperate zone and colder climate of the tundra, the boreal forest's indices of production are intermediate between those two ecosystem types. Annual net primary production in the boreal

forest has been estimated at 7.5 (range 4–20) metric tons per hectare ($t\,ha^{-1}\,year^{-1}$). This compares with 11.5 $t\,ha^{-1}\,year^{-1}$ for temperate forest and 1.5 $t\,ha^{-1}\,year^{-1}$ for tundra ecosystems. Mean boreal forest litterfall is estimated to be 7.5 $t\,ha^{-1}\,year^{-1}$, compared with 11.5 and 1.5 $t\,ha^{-1}\,year^{-1}$ for the temperature forest and tundra. Because low temperatures slow decomposition, the rate of litterfall decay in the boreal forest, 0.21 $year^{-1}$, is also intermediate between 0.77 and 0.03 for the temperate forest and tundra. This means that it takes roughly $3 \times (1/0.21) = 14$ years for 95% of a pulse of litter to decompose.

Fire and Its Effects

Fire is an inherent factor in the ecosystem dynamics of the boreal forest. Lightning-caused fires occur on a given area at intervals of 20–100 years in drier areas to 200+ years in wetter areas such as floodplains. Because nutrients tend to be tied up in slowly decomposing organic matter, fire may be important maintaining tree growth by releasing pulses of nutrients periodically. Many taiga plant species have adaptations to fires, such as serotinous cones and early sexual maturity of some conifers, and resprouting capacity of hardwood trees and many herbs and shrubs. Fires also reset the successional cycle, allowing shade intolerant species like birch and aspen to invade.

Fires destroy the highly flammable late successional evergreen forests that are dominated by the spruces (white spruce and black spruce), creating the habitat required by early successional deciduous trees and shrubs. The dominant early successional deciduous species include the trees quaking aspen (*Populus tremuloides*) and Alaska paper birch (*Betula neoalaskana*) and an assemblage of willows (*Salix* spp.). As post-fire succession proceeds, the vegetation becomes progressively dominated by evergreens such as spruces, and ericaceous shrubs (e.g., *Ledum* spp.), as well as green alder (*Alnus viridis* subsp. *fruticosa*). Fires create a mosaic of patches of deciduous and evergreen trees and shrubs created by fire and the subsequent post-fire succession.

Fire is important to the boreal forest. In boreal North America, fire is essential to the existence of most browsing mammal populations, and especially snowshoe hare populations. This is because fire creates most of the habitat mosaic required by these herbivores. Fire destroys late successional evergreen forests dominated by the slowly growing spruces *Picea mariana* and *P. glauca*, which are effectively defended against browsing by resins that contain toxic lipid-soluble secondary metabolites such as the monoterpene camphor that deters feeding by snowshoe hares. However, although spruce is a comparatively poor winter-food for boreal browsers, dense spruce thickets provide the protective cover that these herbivores use to evade their predators. The recently burned patches within an unburned spruce forest matrix are colonized by the more rapidly growing and comparatively poorly defended early successional deciduous species such as the willows (*Salix* spp.), quaking aspen (*Populus tremuloides*) and birches that in are the preferred winter-foods of most North American boreal browsing mammals. Thus, in boreal North America most of the optimal habitat for browsing mammals occurs at the edge of burns where browsers have ready access to both good predator escape cover (spruce forest) and good food (fast growing deciduous species in recently burned patches). For this reason in boreal North America the abundance of browsing mammals is often greatest in a landscape that contains patches of poorly defended early successional deciduous woody species within a late successional spruce matrix. And where the abundance of these browsing mammals is greatest, the intensity of their selective browsing on the recruitment of rapidly growing deciduous tree species such as *B. neoalaskana* and *B. papyrifera* is generally most intense.

Conservation and Global Issues

The boreal forest represents the single largest pool of living biomass on the terrestrial surface. It contains more than 30% of the total terrestrial pool, and it is therefore critically important in global carbon dynamics. Much of the carbon is stored in the ground layer. Currently, the taiga is thought to act as a net sink of carbon, with an estimated 0.54 billion metric tons stored per year. However, global climate change, in the form of higher temperatures, may cause significant changes in the carbon dynamics by increasing decomposition rates faster than photosynthetic rates. Fire frequencies may also increase with temperature, as precipitation is not expected to rise, which will further increase the release of carbon stored in the ground layer. According to some studies, the boreal forest will be a net contributor to CO_2 in the atmosphere under the projected climate changes.

Climate-induced changes in the boreal forest would also have an impact of migrant birds that use the region for reproduction. Changes in tree species composition may challenge the capacity of birds to adapt, as has already the increasing fragmentation of the forest due to clear-cutting in many areas within the biome.

See also: Aquatic Ecology: Deep-Sea Ecology. Behavioral Ecology: Mating Systems. Ecological Data Analysis and Modelling: Modeling Dispersal Processes for Ecological Systems. Ecosystems: Riparian Wetlands. Evolutionary Ecology: Association. General Ecology: Biomass ; Carrying Capacity; Temperature Regulation. Human Ecology and Sustainability: Political Ecology

Further Reading

Bonan, G.B., Shugart, H.H., 1989. Environmental factors and ecological processes in boreal forests. Annual Review of Ecology and Systematics 20, 1–28.

Danell, K., Lundberg, P., Niemalä, P., 1996. Species richness in mammalian herbivores: Patterns in the boreal zone. Ecography 19, 404–409.
Henry, J.D., 2003. Canada's boreal forest. Washington, DC: Smithsonian.
Hunter Jr., M.L., 1992. Paleoecology, landscape ecology, and conservation of neotropical migrant passerines in boreal forests. In: Hagan III, J.M., Johnston, D.W. (Eds.), Ecology and conservation of neotropical migrant landbirds. Washington, DC: Smithsonian.
Knystautus, A., 1987. The natural history of the USSR. New York: McGraw-Hill.
Krebs, C.J., Boutin, S., Boonstra, R., 2001. Ecosystem dynamics of the boreal Forest: The Kluane project. New York: Oxford University Press.
Larsen, J.A., 1980. The boreal ecosystem. New York: Academic Press.
McCullough, D.G., Werner, R.A., Neumann, D., 1998. Fire and insects in northern and boreal forest ecosystems of North America. Annual Review of Entomology 43, 107–127.
Oechel, W. C., and Lawrence, W. T. (1985). Taiga. In: Chabot, B. F., and Mooney, H. A. (eds.). Physiological ecology of North American plant communities, pp. 66–94. New York: Chapman and Hall.
Van Cleve, K., C. T. Dyrness, L. A. Viereck, J. Fox, F. S. Chapin III, and W. Oechel. 1983. Taiga ecosystems in interior Alaska. Bioscience 33:39–44.

Estuaries

RF Dame, Charleston, SC, USA

Introduction

Estuarine ecosystems are among the most complex and complicated systems in the biosphere. Because they are at the interface of terrestrial, freshwater, and marine systems, estuaries are subject to massive fluxes of materials and energy. Further, as a large percentage of the human population lives in close proximity to estuarine and coastal environments, anthropogenic impacts and stress are major driving factors in determining the health and functional status of estuarine ecosystems. In this section, the structure and function of estuarine ecosystems are examined.

Definitions of Estuarine Ecosystems

To set the stage for any discussion of estuarine ecosystems, a clear working definition is needed. One of the simplest and most utilized definitions of an estuarine ecosystem is:

> the zone where freshwater from land runoff mixes with seawater.

Another common definition is:

> An estuary is a semi-enclosed coastal body of water that has free connection with the sea where seawater is diluted by freshwater derived from land drainage.

The preceding definitions focus on the geomorphological and hydrological aspects of estuaries with no mention of the abiotic or physical driving sources of energy, that is, tides and solar insolation. Nor are any biotic components or processes utilized. Thus, the following definition is proposed:

> An estuarine ecosystem is a system composed of relatively heterogeneous biologically diverse subsystems, i.e., water column, mud and sand flats, bivalve reefs and beds, and seagrass meadows as well as salt marshes. These subsystems are connected by mobile animals and tidal water flows that are embedded in the geomorphological structure of creeks as well as channels, and together form one of the most productive natural systems in the biosphere.

Recent quantitative studies indicate that estuaries are ecoclines that are composed of gradients containing relatively heterogeneous subsystems that are environmentally more stable than ecotones (Fig. 1). Ecoclines are boundaries with more gradual, progressive change between freshwater and the sea. In this view, the organisms in the estuary are either from freshwater or from marine environments; there are no brackish water species. Each estuarine system will respond to at least a freshwater and a marine gradient as well as have its own particular combination of biological and physical components and processes. Thus, every estuarine ecosystem is unique.

Geomorphological Types of Estuaries

Bar-Built and Lagoonal

Bar-built or lagoonal estuaries form in the areas behind sandy barrier islands and usually drain relatively small watersheds. The exchange of water between the estuary and the sea occurs through tidal inlets. Astronomical tides and winds are the major forces controlling water circulation and water height. The areas behind barrier islands are generally subject to less wave action and this promotes the development of extensive wetlands. Bar-built estuaries are generally smaller than other estuarine types, suggesting that they have a higher surface area to volume ratio and, therefore, play a greater role in ecological processes than was previously thought. Well-studied examples of bar-built estuaries are found along the temperate and subtropical coasts of eastern North America, Europe, Asia, and the southern and eastern shores of Australia.

Riverine Estuaries

There are two fundamentally different riverine systems (Fig. 2). First, are those that arise in the piedmont, have extensive watersheds, receive substantial freshwater discharge, but only a small portion of their watershed is covered by wetlands. Chesapeake Bay and San Francisco Bay in North America as well as the Eastern Scheldt in Northern Europe are well-studied examples of

 doi:10.1016/B978-0-444-63768-0.00329-2

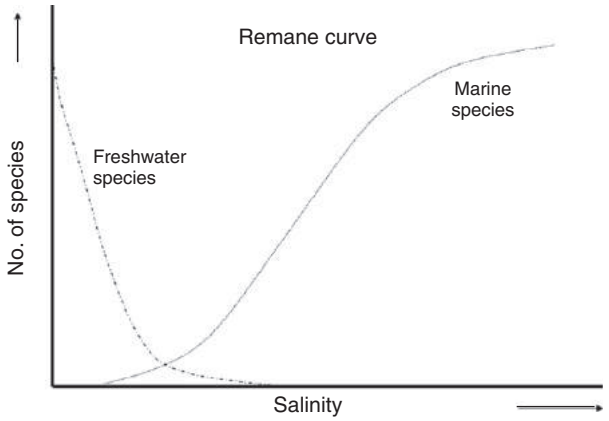


Fig. 1 A generalized Remane curve with number of species plotted versus estuarine salinity gradient.

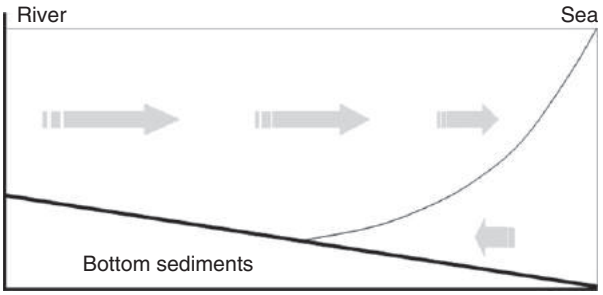


Fig. 2 Generalized material flux patterns in a Riverine estuary.

this type of riverine system. A second type of riverine system known as coastal plain estuaries are characterized by a much gentler slope with proportionally more wetlands than piedmont estuaries. Generally, these systems are less studied, smaller and have a lower, more sluggish flows.

Estuarine Ecosystems and Maturity

In an attempt to place a more ecosystems oriented emphasis on estuaries, the `geohydrologic continuum theory of Marsh-Estuarine ecosystem development' was proposed as a scheme for categorizing estuarine ecosystems. In this theory, the tidal channels of the estuarine ecosystem represent a physical or geohydrologic model of how the ecosystem adapts until there is a change of state. Mature portions of the system are at the ocean–estuary interface, mid-aged components are intermediate within the longitudinal distribution of the system, and young or immature areas are at the land–estuary interface. Mature systems export particulate and dissolved materials, mid-aged areas import particulate and export dissolved materials, and immature systems import both particulate and dissolved materials. Some estuarine ecosystems may have all three types while others may have only one or two.

Estuaries as Complex Systems

While it is generally acknowledged that ecosystems are complex systems, it is appropriate to describe estuarine ecosystems in the context of the complex systems approach. Complexity as used here can be defined by (1) the nonlinear relationships between components; (2) the internal structure created by the connectivity between the subcomponents; (3) the persistence of the internal structure as a form of system memory; (4) the emergence or the capacity of a complex system to be greater than the sum of its parts; (5) the reality that complex systems constantly change and evolve in response to self-organization and dissipation; and (6) behaviors that often lead to multiple alternative states. Thus, estuarine ecosystems are open nonequilibrium systems that exchange matter and energy as well as information with terrestrial and marine ecosystems as well as internal subsystems. These exchanges not only connect various components, but are the essential elements of feedback loops that generate nonlinear behavior and the emergence of structures and behaviors whose sum is greater than the whole. These systems do exhibit alternate states, for example, Chesapeake Bay appears to have a benthic state dominated by oysters and a water column state dominated by plankton.

Major Estuarine Subsystems or Habitats

The landscape approach to estuarine ecosystems focuses on subsystems or habitats as major components within estuaries. Because organisms respond to the amount of change in the physical (abiotic) environment, their reaction to their environment results in subsystems or habitats composed of specific groups of species that are adapted to that particular set of abiotic factors. In estuarine ecosystems, the major abiotic factors are salinity, water velocity, intertidal exposure, and depth.

Water Column

Water is the primary medium for the transport of matter and information in estuarine ecosystems. Freshwater enters the estuary either as precipitation or as an accumulation driven by gravity down-slope through streams and rivers to the estuary. Salt water enters the estuary from the sea via tidal forcing. The gradient of increasing salt concentration from freshwater to marine divides the estuary into zones of salt stress and subsequently into different pelagic subsystems (Fig. 2).

Phytoplankton are small chlorophytic eukaryotes that drift as single cells or chains of cells in estuarine currents. Diatoms and dinoflagellates are the dominant groups while species composition of a specific system is usually determined by salinity, nutrients, and light. They are a major component of the estuarine water column and provide food for many suspension-feeding animals. Planktonic primary production is seasonal and varies from distinct peaks in the arctic to spring and autumn blooms in temperate systems and almost no peaks in tropical estuaries. Average annual planktonic primary production in estuaries is about 200–300 gC m^{-2} yr^{-1} and is mainly a function of light, nutrient availability, and herbivore grazing.

There are two major categories of zooplankton: holoplankton that in most estuaries are dominated by calanoid copepods which spend their entire life in the planktonic state and the diverse meroplankton that only spend their larval state in the plankton. Most estuarine zooplankton are believed to be herbivores and play a major role in connecting carnivores to phytoplankton. They are also thought to be major sources of inorganic nutrients that are available to phytoplankton.

The microbial loop in estuaries is composed of micro- and nano-planktonic bacteria, protozoans, and flagellates. Initially, the microbial loop was thought to play a major role in recycling nutrients with dissolved organic matter (DOM) a major product. However, the recent finding that a sizable proportion of DOM is made up of viruses has forced a major change in the microbial loop model (Fig. 3). The current paradigm of the microbial–viral loop envisions the viruses (10^{10} l^{-1}) as 10 times more abundant than bacteria (10^{9} l^{-1}) and controllers of bacterial diversity and abundance. The viruses are small (20–200 nm), ubiquitous particles that use the process of cell lysis to attack and kill bacteria. As a result, more bacterial biomass is shunted into DOM and away from the macroplankton and suspension-feeding macrobenthos. The much more rapid viral recycling of nutrients also has the potential to generate more stability in the system.

Large mobile animals, birds, terrestrial and aquatic mammals, and fish, shrimps and crabs, are common residents as well as transients in estuarine systems. These animals transform and translocate materials both within the estuary and between the estuary and other systems. The nekton organisms, in particular, use the tidally forced water column as a pathway between deeper channels and intertidal habitats where they seek refuge, feed, and develop.

Marshes and Mangroves

Emergent vascular plant-dominated intertidal wetlands are major subsystems in most estuaries. The two most common habitats are geographically zoned latitudinally with marshes dominating the temperate zone and mangroves the frost free subtropical and tropical zones. Both are found in low-energy wave-protected, sedimentary, high-salinity, and intertidal environments near the

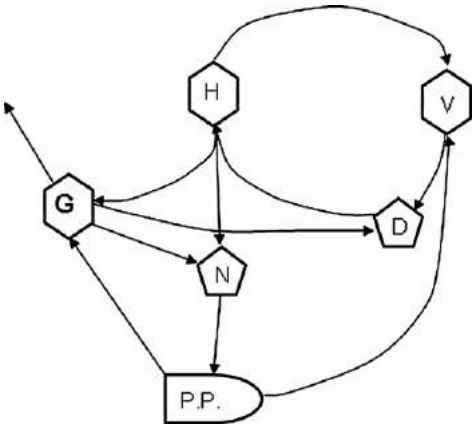


Fig. 3 A simple microbial–viral loop food web for an estuarine system. D, dissolved organic matter; G, grazers; H, heterotrophs; N, nutrients; P. P., primary producers; V, viruses.

Table 1 Primary production in estuaries

Primary producer	*Annual primary production(g C m^{-2})*
Macrophytes	
Spartina	400–2480
Rhizophora	696–2100
Microphytobenthos	50–200
Epiphytes	12–260
Phytoplankton	25–150

mouth of the estuary. While wetlands in the high-salinity portion of estuaries are low in species diversity (almost monocultures) of vascular plants, diversity is much higher in the freshwater reaches.

Salt marshes reach their greatest extent and productivity along the Gulf and southeast Atlantic coast of North America where the cord grass *Spartina alterniflora* dominates. This high production is the result of near ideal conditions of temperature, salinity, light, sediment texture, nutrients, and tidal range. Marsh grasses produce large quantities of both above- and belowground biomass that accumulates in the surrounding sediments (Table 1). The stems and leaves of the grasses also provide a structural base for an epiphytic community that further increases production. Decomposition processes in the organically rich sediments generate a strongly anaerobic reducing environment making the salt marsh a major center for nutrient cycling. The nutrient uptake mechanisms of vascular plants are poisoned by the reducing environment; however, air passages in the roots, rhizomes, and stems of these grasses aerate the surrounding sediments so that nutrient uptake can be maintained. The vertical stems and leaves of *Spartina* also serve as a passive filter that slows water flow, can remove via deposition suspended sediments from the water column, and allows many marshes to maintain their elevation with respect to rising sea level. This same environment provides food and refuge for many economically important nekton.

Mangroves are intertidal, tropical, and subtropical woody vascular plants that fill a niche similar to that of *Spartina*. In the high-salinity portions of the estuary, the red mangrove, *Rhizophora*, dominates. Red mangroves have prop roots that lift the plants above the reducing environment of the surrounding sediments. There is a gradient from high production in riverine swamps to low production in high-salinity scrub areas. On a global scale of increasing light with decreasing latitude, the closer a system to the equator, the higher the mangrove productivity. Nutrients have also been implicated as a major limiting factor on mangrove productivity. There is evidence that mangrove production is enhanced by flushing action of storms. In addition to being a nursery for many fish, shrimps, and crabs, the structural mass of mangroves may form a protective buffer to the impacts of storm surges and tsunamis on coastal and estuarine systems.

Seagrasses

Seagrasses are submerged vascular plants that are found in aerobic, clear-water, high-salinity systems with moderate water flow. Cold water systems are dominated by eel grass, *Zostera*, and in the tropics turtle grass, *Thalassia*, is the major group. These grasses are not found in estuaries with high suspended sediment loads, that is, Georgia and South Carolina where there is insufficient light penetration to support their growth. They are also limited to the upper 20 m of water because water pressure compresses their vascular tissues. Maximum seagrass production can approach 15–20 gC m^{-2} d^{-1}. The high productivity of the seagrass is almost equaled by the productivity of the epiphytes on their leaves; however, the sediment trapping abilities of seagrasses give them an advantage over phytoplankton and epiphytes in nutrient limiting conditions. The structure of the seagrasses provides feeding habitat for many mobile animals as well as deposit feeding and suspension-feeding benthos.

Invertebrate Reefs and Beds

Suspension-feeding benthic animals are common in most estuaries because of the high availability of suspended phytoplankton. A number of bivalves and a few worms can aggregate in very dense, high biomass beds or reefs. These structures are found both intertidally and subtidally in high to moderate salinities. The eastern oyster, *Crassostrea virginica*, in its intertidal form builds some of the most extensive aclonal reefs known. Intertidal beds of *Crassostrea* and *Mytilus* can have biomass densities exceeding 1000 gdb m^{-2}. Depending on the estuary, suspension feeders such as oysters and mussels have been shown to control phytoplankton populations in some systems and influence nutrient cycling by short-circuiting planktonic food webs and reducing the recycle time for essential nutrients. There is evidence that the presence of a significant bivalve suspension-feeder component in estuarine ecosystems enhances system stability.

Mud and Sand Flats

Mud and sand flats are common to the intertidal zone of most estuaries. The major biotic components of tidal flats are bacteria, microbenthic algae, small crustaceans, and burrowing deposit feeders. As in the water column, the microbial–viral loop is thought to play a major role in the decomposition of organic matter in tidal flat sediments. In some estuaries, the microbial–viral loop

utilizing a variety of electron acceptors may represent a significant sink for matter and energy. Thus, the prevailing processes on these flats can potentially redirect the fluxes of matter and energy away from macrofaunal food webs to those dominated by microbial processes. The occurrence of tidal flats was originally attributed to the hydrodynamics and sediment sources in tidal creeks; however, with the application of complexity theory to ecological systems, these flats are also being described as alternative states of salt marshes and bivalve beds.

Material Fluxes

Water Fluxes and Residence Times

Interest in the exchange of nonliving materials and organisms between estuarine ecosystems and the sea was initiated by the first quantitative metabolic studies on the high productivity of marsh dominated estuaries. These studies were first synthesized in simple energy budgets that were found to explain less than 50% of the productivity of estuarine ecosystems. Investigators speculated that the unaccounted-for energy must be exported from the estuarine ecosystem by tidal currents. This idea led to the `outwelling hypothesis' that states that estuarine ecosystems produce much more organic material than can be utilized or stored by the system and that the excess is exported to the coastal ocean where it supports near coastal ocean productivity. While the energy budget or mass balance approach is a cheaper and quicker method of determining the direction of material fluxes, in recent years the direct measurement of material fluxes is favored because this approach provides statistically meaningful results.

Another aspect to the fluxes of materials in estuarine systems is the time the water mass remains in the system or residence time (also known as flushing time or turnover time). Residence time can provide essential information to resource managers on the retention and dispersal of toxins, the incubation of invasive species, and the carrying capacity of a system for benthic suspension feeders (Fig. 4). Recent studies on the physics and geomorphology of water in estuarine tidal channels suggest that the residence time of water may vary greatly from place to place within some estuaries. Such variations have been used to explain growth variations in bivalves in different locations within the same estuary. Traditional estimates of an estuarine system's residence time can be computed from measurements of system volume, tidal prism, and water input to the system. The advent of fast computers and numerical models, however, now allows for much more modeling of these systems with the potential for more sophisticated spatial and temporal management strategies.

In riverine systems, river flow is the main physical cause of material and organismic transport from estuaries to the sea. Each of these systems are a unique and changing feature on the present landscape because rising sea level is drowning their basins and sediments are gradually filling their channels. For example in Chesapeake Bay, 35% of the particulate nitrogen and most of the phosphorus is buried in the sediments of the bay. Of the nitrogen in the bay water column, 31% was exported to the sea and 8.9% was removed from the system as commercial fish harvest. In general, the nutrients transported and exported by riverine estuaries are thought to be a significant source for generating new organic production in the coastal ocean. As many of these systems have dams or have them proposed, managers must take into account the direct and indirect effects of these structures on recreational and coastal fisheries.

In bar-built estuaries, tides are usually the major source of energy for the transport of materials into and out of the estuary. If the fastest currents are on the flooding tides, then the system tends to import suspended particulate material. In contrast, if ebbing tides have the fastest currents, then the system usually exports suspended particulate materials. The Wadden Sea of Northern Europe is a flood-dominated system and North Inlet in South Carolina is an ebb-dominated system.

In shallow, high-insolation, low-precipitation, warm systems, evaporation can dictate the direction of transport. This is the case in some small tropical systems where water loss due to evaporation is replaced by the influx of water and nutrients from the adjacent sea.

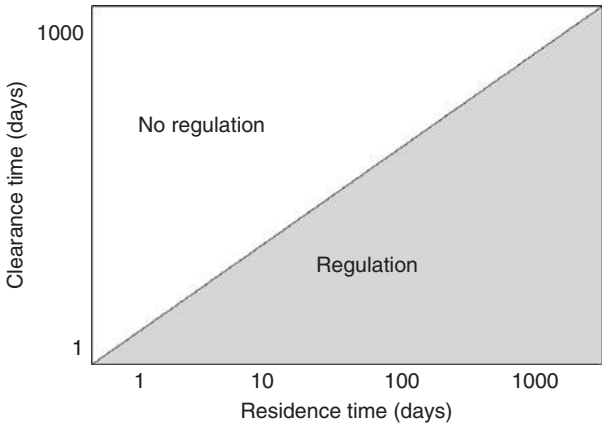


Fig. 4 A plot of water volume residence time versus bivalve clearance time showing areas of potential control by suspersion-feeders.

Organismic Transport

In addition to inanimate materials, the larval and adult stages of many organisms are exchanged between the estuary and the sea. Some organisms may be passively carried by estuarine currents while others may actively swim or take advantage of the direction of tidal flows to move across the estuary–ocean interface.

Primary producers, including phytoplankton and resuspended benthic microalge, depend on passive transport between estuaries and the sea. Most flux studies show that these organisms have a net seasonal or annual transport into the estuary from the coastal ocean. This import has been explained by passive filtration by estuarine wetlands and by active filtration by suspension-feeding animals within the estuary. Protozoans, bacteria, and viruses are also found in the estuarine water column and while they most certainly are passively transported by estuarine currents, the direction of their net flux is yet to be determined.

The exchange of invertebrate larvae between estuaries and the coastal ocean has been explained by two competing schools of thought, the passive and active hypotheses. In the passive hypothesis, the horizontal movements of larvae are mainly a function of current direction and velocity. The active transport school contends that invertebrate larvae swim both vertically and horizontally to take advantage of tidal currents. In one group that includes oysters, the early stage larvae stay high in the water column with later stages sinking to lower depths. This strategy allows downstream movement of early larvae with some exiting the estuary to the sea, while older larvae are entrained in inflowing bottom currents and effectively retained in the estuary. A second approach is used by larvae that migrate vertically in the water column in synchrony with tidal cycles. This strategy allows larvae to maximize upstream transport and retention. A third group has larvae that are immediately transported to the coastal ocean where they stay for weeks before returning into the estuary using wind and tidal currents. A final group uses the coastal ocean during their adult and larval life. In this case, the postlarvae enter the estuary maintaining their position by swimming against tidal currents.

Nekton organisms (fish, crabs, and shrimps) are mobile links between the various subsystems of estuarine ecosystems as well as links between the estuary and the sea. These animals feed and accumulate biomass while in the estuary and then move back to the coastal ocean, thus exporting biomass and inorganic wastes.

Global Climate

While seasonal and latitudinal climatic effects on coastal and estuarine systems have long been documented, the impacts of global climate change (warming or cooling) on estuarine systems have only recently been quantified. Major storms, El Nino-Southern oscillation (ENSO) events, seismic sea waves, or tsunamis and sea level rise (SLR) are global effects that can significantly influence water and material fluxes in estuaries.

Hurricanes and major storms generally influence estuaries through storm surges and short-term increases in precipitation. These enormous pulsed fluxes of water can change the geomorphology of estuaries and their watersheds, massively resuspend sediments, and flush materials off the landscape and into the estuary. Tsunamis can be even larger than storm surges and can have similar impacts to even greater areas of the coastal ocean and estuaries. However, extensive marsh and mangrove wetlands common to estuaries can buffer these pulses of water and reduce the damage they can cause to the coastal landscape.

ENSO events only affect some estuaries. The effect is usually a drought or higher than average precipitation. For example in some South Carolina estuaries, ENSO-induced precipitation and upland runoff can depress salinity up to 75% for as much as 3 months.

SLR is an example of global change on both seasonal and annual time scales that directly influences estuarine systems. Seasonal changes in sea level are the result of air pressure changes at the water's surface and the expansion or contraction of water mass due to heating and cooling. In estuarine systems, these changes are reflected in the depth of the system, but more importantly in the area and time of exposure or submergence in the intertidal zone. SLR will gradually force the transgression of estuaries upslope along the coastal plain. Eventually, SLR will compete with human development for the coastal landscape.

Estuarine Ecosystem Resilience and Restoration

Estuarine ecosystems and subsystems can and do exhibit alternate or multiple states of existence. The ability of an ecosystem to absorb disturbance and resist a change in state is termed ecological resilience, as opposed to engineering resilience, which is the time it takes a system to return to its original state. In the last decades of the twentieth century, ecologists observed that ecosystems were not static entities, but appeared to change in response to external and internal forces. In the Chesapeake Bay estuary, for example, some of the factors causing a state change were over-fishing, increased suspended sediment load, eutrophication, species invasion, and disease. The bay's responses to these forces were slow at first, but with the steady increase in the human population in the bay watershed and with its adherent development, the signs of a state change were dramatically evident. The oyster reefs, a major benthic subsystem or habitat that had dominated the bay for centuries, began to decline rapidly or crash. The benthic-dominated food web was replaced by a planktonic food web. Management efforts to restore the initial oyster-dominated system did not work, probably because they had a single species focus and because ecosystems are strongly nonlinear which means the

path to restoration is different from that leading to the initial change of state and many more components of the ecosystem are involved in addition to the oysters.

See also: Ecological Complexity: Cybernetics; Systems Ecology: Ecological Network Analysis

Further Reading

Alongi, D.L., 1998. Coastal Ecosystem Processes. Boca Raton, FL: CRC Press.
Attrill, M.J., Rundlle, S.D., 2002. Ecotone or ecocline: Ecological boundaries in estuaries. Estuarine, Coastal and Shelf Science 55, 929–936.
Dame, R.F., Childers, D., Koepfler, E., 1992. A geohydrologic continuum theory for the spatial and temporal evolution of marsh-estuarine ecosystems. Netherlands Journal of Sea Research 30, 63–72.
Dame, R.F., Chrzanowski, T., Bildstein, K., *et al*., 1986. The outwelling hypothesis and North Inlet, South Carolina. Marine Ecology Progress Series 33, 217–229.
Dame, R.F., Prins, T.C., 1998. Bivalve carrying capacity in coastal ecosystems. Aquatic Ecology 31, 409–421.
Day, J., Hall, C., Kemp, W., Yanez-Arabcibia, A., 1989. Estuarine Ecology. New York: Wiley.
Gunderson, L.H., Pritchard, L., 2002. Resilience and the Behavior of Large-Scale Systems. Washington, DC: Island Press.
Lotze, H.K., Lenihan, H., Bourque, B., *et al*., 2006. Depletion, degradation, and recovery potential of estuaries and coastal seas. Science 312, 1806–1809.
Mann, K.H., 2000. Ecology of Coastal Waters, 2nd edn. Oxford: Blackwell Science.

Floodplains

BG Lockaby, Auburn University, Auburn, AL, USA
WH Conner, Baruch Institute of Coastal Ecology and Forest Science, Georgetown, SC, USA
J Mitchell, Auburn University, Auburn, AL, USA

Introduction

Globally, floodplains may be of greater value to society than any other ecosystem type. This is because of the critical role that interactions between floodplains and associated streams play in maintaining supplies of clean water. While that role is conceptually simple, the processes which define interactions (i.e., floodplain functions) in aquatic–terrestrial ecotones are exceedingly complex. Consequently, it is necessary to develop some understanding of the ecological mechanisms behind those interactions in order to fully appreciate the importance of floodplain ecosystems. To that end, the goal of this article is to provide a first-iteration overview of floodplain form and function (Fig. 1).

A key concept is that floodplains and associated streams are both causes and reflections of the other's characteristics and functioning. As an example, the climate and geomorphology of a landscape will define the hydrology and initial chemistry of streams. Stream characteristics determine the hydrology, soil characteristics, flora and fauna, and biogeochemistry of the floodplain. In turn, biogeochemical feedback from floodplains to streams helps define the environment seen by aquatic flora and fauna. Thus, a strong interdependency exists between aquatic and terrestrial components of riparian ecotones.

It is critical to understand that land clearing and development, construction of dams and impoundments, pollutant export, and other human activities constitute major influences on streams and floodplains. In some cases, these will override the original hydrology, biogeochemistry, and ecology. As an example, the original hydrology of a riparian system could be dramatically altered by the construction of bermed roadways that cross streams without adequate provision for through flow. Since hydrology is the primary driver of all floodplain functions, corresponding changes in net primary productivity (NPP), species composition of animal and plant communities, and biogeochemistry could be expected to follow.

Geomorphic Origins

Streams in steep topography tend to undergo continual downcutting and, consequently, act as sources of fine and coarse material with little to no opportunity for deposition. Sediment loads are easily carried downstream because the high gradient of the channel imparts sufficient energy for water to retain particles. In many cases, as streams emerge from steeper terrain and move into flatter areas such as coastal plains, the gradient of the channel decreases and flows may spread and lose energy. This promotes the occurrence of overbank flow and creation of deposition surfaces or sediment sinks. However, a floodplain may shift between being a sediment sink or source depending on hydrologic changes induced by climate, anthropogenic activities, or other influences. Downcutting also occurs as older floodplains are abandoned by streams and become terraces which resemble stair steps in cross-section (Fig. 2).

Sedimentation occurs as particles settle during sheetflow and is highly variable both temporally and spatially. Sediment deposition or alluviation makes possible the high soil fertility that is generally associated with many floodplains although there

Fig. 1 Panoramic view of the Timpisque River and the Palo Verde Marsh in Costa Rica.

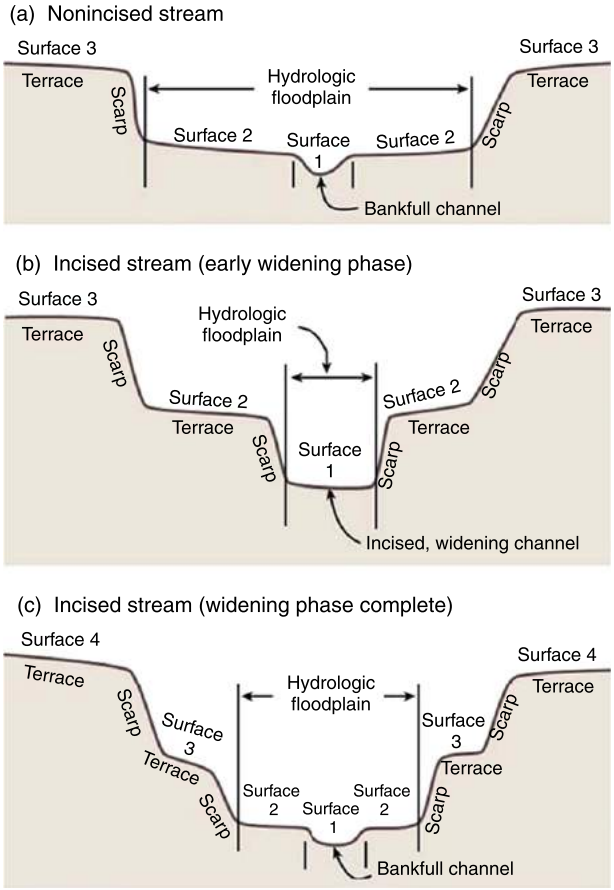


Fig. 2 Terraces in (a) nonincised and (b and c) incised streams. In Stream Corridor Restoration: Princples, Policies, and Practices (10/98). Interagency Stream Restoration Working Group (15 federal agencies) (FISRWG).

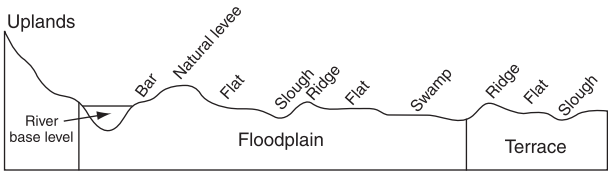


Fig. 3 Cross-sectional view of a floodplain topographic positions. Adapted from Hodges 1998.

are notable exceptions. Rates of sediment accumulation vary markedly among floodplains and, in the southeastern United States for example, range from 1 to 6 mm yr^{-1}.

Deposition and scouring may often occur simultaneously on different portions of floodplains and at different times in individual locations. Consequently, the scale at which sedimentation is assessed is very important in gaining an accurate assessment of net changes. The result of the spatial irregularities is a pattern of swales and berms that generally runs parallel to the stream course. The convex and concave microrelief may represent elevation differentials of only a few centimeters. Nonetheless, those minor differences have major importance in defining soil environments for vegetation and in influencing the extent of contact between floodwaters and the floodplain surface. In many cases, the microtopography of major floodplains is somewhat predictable (Fig. 3) and, similarly, drives spatial patterns of species composition and NPP of vegetation communities. However, changes in microrelief may be much less apparent on some floodplains due to either prolonged or infrequent flooding.

Hydrology

Hydrology is the foremost determinant of vegetation species occurrence, NPP, biogeochemistry, floral and faunal habitat, and all floodplain functions and traits. Consequently, any insights into the nature of floodplain ecosystems and the basis of their societal

value are predicated upon an understanding of hydrology. The `flood-pulse' concept provides one framework within which to develop this understanding.

In this concept, the river and floodplain are considered as a single system and the `rhythm of the pulse' (i.e., the hydroperiod) is the controlling mechanism which regulates exchange of energy and material between the river and floodplain. An influx of sediment and nutrients and export of organic carbon from the floodplain will occur at intervals dependent on the pulse rhythm. Examples of common rhythms include single, long duration and multiple, short duration which might be stereotypic of high-order river floodplains and low-order headwater streams, respectively.

In general, flood frequency and duration may decrease and increase, respectively, as stream order rises. When headwaters originate in mountainous terrain, narrow V-shaped valleys form and hydroperiods may be characterized as flashy (i.e., frequent flooding, with sharp rises and drops associated with stage levels). Hydroperiods reflect the integration of rainfall patterns, water storage capacities, and many other factors across the associated catchments. Consequently, stage level rises and falls are slower due to the `buffering' that is provided by high storage capacities and the greater variability of other factors. Conversely, small catchments have much less storage capacity and, consequently, streams respond rapidly to precipitation events. As a result, floodplains of large rivers can stay flooded for significant portions of a year while low-order floodplains may be inundated frequently but for much shorter periods.

Interchange of water between floodplains and rivers is very complex and involves mutualistic influences. The nuances of those interactions form the basis of the role of floodplains as ecotones and regulators of energy and nutrient exchange. At low stage levels, water within swales and depressions may have originated with the river, precipitation, an upwelling of groundwater, or some combination. From a biogeochemical standpoint, the origin is significant in terms of the degree of spatial and temporal contact with the floodplain. At low stage levels, there is less opportunity for river water to contact the floodplain and, consequently, biogeochemical and dissolved organic carbon exchanges are minimal. As stage levels rise, the potential for the floodplain to influence the biogeochemistry of sheetflow increases as well. However, at some point, increasing floodwater volumes and higher velocities reduce contact with the floodplain. This is because a decreasing proportion of the sheetflow volume is in contact with the floodplain as volumes increase. Similarly, temporal contact is reduced as sheetflow velocities rise.

There is also significant interaction between the river and floodplain in terms of groundwater. Channel waters often generate a head pressure which declines with distance from the stream bank. Groundwater transmittance will decline as hydraulic conductivity of alluvium decreases (e.g., clays have reduced conductance compared to sands). In humid regions, groundwater near the channel moves under pressure and will contact and mix with water that has seeped into the alluvium from adjacent uplands. As a result, groundwater mixing can be quite active during periods of low evapotranspiration.

Biogeochemistry

Once considered purely as nutrient sinks, floodplains are now known to play multiple roles from a geochemical perspective. Based on the type of floodplain, associated vegetation, and the degree and nature of disturbance, floodplains may also serve as sources or transformation zones for nutrients. The widely held perception of floodplains as fertility hot spots belies the complexity associated with input–output budgets as well as the biogeochemical processes within the floodplain ecosystem. In particular, the impact of hydroperiod on biogeochemical processes sets floodplain biogeochemistry apart from that of non-wetland ecosystems. Periodic flooding makes possible nutrient exchange across the aquatic–terrestrial ecotone and controls the nature of decomposition, nutrient uptake and release by vegetation, and many other processes. As an example, the process of denitrification or the anaerobic conversion of nitrate to gaseous forms of nitrogen is very important on floodplains. In addition, the interaction of hydrology and biogeochemistry necessitates the development of unique approaches to the study of nutrient cycling in these ecosystems.

As previously mentioned, floodplains may serve as sinks, sources, or transformation zones for geochemical inputs of nutrients derived from inflow, precipitation, nitrogen fixation, and soil weathering. Multiple roles may proceed simultaneously on the same floodplain if spatial heterogeneity in hydrology, vegetation, disturbance, and nutrient influx so dictate. The use of a geochemical budget allows net inputs to be compared to net outputs and is based on the perspective of the ecosystem as an integrated system.

In general, the factors that promote nutrient sink activity on floodplains include (1) presence of aggrading vegetation; (2) wide carbon: nutrient ratios in living vegetation and detritus; (3) topographic positions conducive to somewhat frequent, short duration, and low-energy flooding; (4) basin geomorphology that promotes significant sediment loads in streams (e.g., redwater, brownwater, or whitewater based on the color of suspended clay); (5) high occurrence of nitrogen-fixers; and (6) until nutrient saturation is approached, association with a river subjected to high anthropogenic nutrient loadings.

Alternatively, rivers draining low gradient basins with sandy soils are often referred to as blackwater systems because their waters are stained with organic substances (Fig. 4). These tend to carry low sediment loads and, consequently, alluviation (i.e., sink activity) is less pronounced. Also, floodplains occupied by mature vegetation communities may act as transformers of nutrients (e.g., inorganic inputs of nitrogen converted to organic outputs) rather than a sink or source. The latter is a key facet of the `kidney function' of these systems and has great significance for maintenance of water quality. Sink activity, such as the filtration and accumulation of sediments (and associated nutrients) from sheetflow also plays a major role in cleansing water (Fig. 5). Finally, floodplains that have been altered in some way by disturbance may function as nutrient sources. The longevity of the source

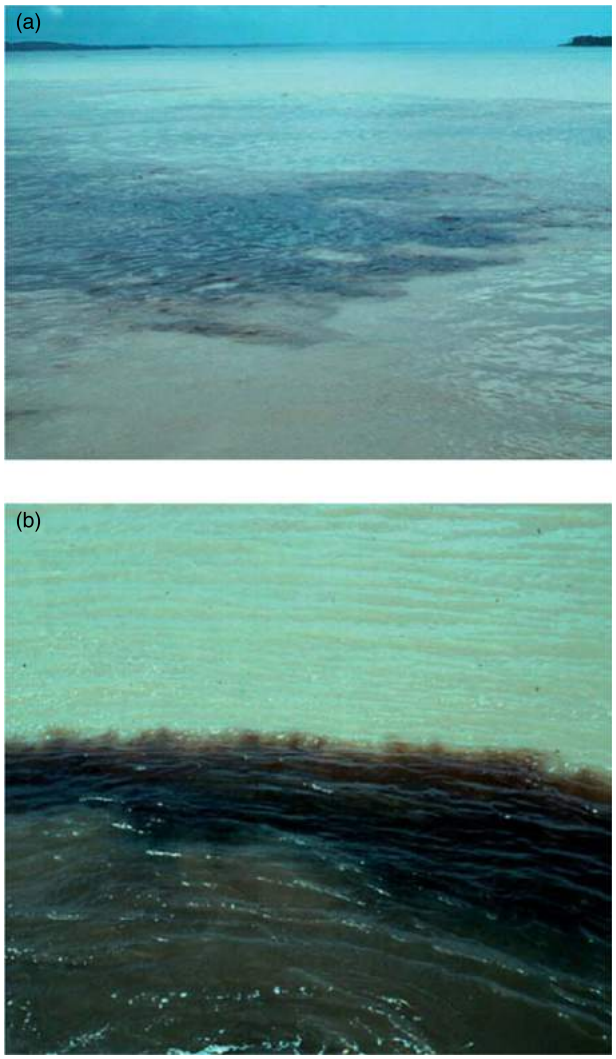


Fig. 4 Amazon River: (a) a broad and (b) a close-up view. The formation of the Amazon River at the `o encontros das aquas' or mixing of the Rio Negro and Rio Solimoes near Manaus, Brazil. The blackwater Rio Negro is contrasted with the sediment-laden Rio Solimoes.

Fig. 5 Flint River – sediment accumulation on the Flint River floodplain near Ft. Valley, GA during floodwater drawdown.

activity could be short-term (e.g., a well-planned forest harvest followed by rapid forest regeneration) or long-term (e.g., conversion to agricultural or urban uses, impoundments, or climate change).

Similarly, all biogeochemical processes within floodplain ecosystems reflect the overriding influence of hydrology. As an example, the timing of litterfall is heavily affected by hydroperiod because different vegetation communities occur under different hydrologic regimes. In the southeastern United States, forest species associated with *Nyssa* may grow under wetter conditions than communities dominated by some species of *Quercus*. On wetter sites, *Nyssa* foliage tends to senesce earlier in the autumn than other floodplain tree species and, consequently, the senesced foliage is exposed to a different microenvironment than litter that falls later in the year. As a result, nutrient release and immobilization sequences are likely to differ among sites.

Mass loss and nutrient dynamics during decomposition are a function of both litter quality and the decomposition microenvironment. Litter quality (the biochemical composition of detritus) is defined by the conditions under which a plant is growing as well as genetics and has been shown to be closely linked to variation in hydroperiod. Also, the frequency and duration of flooding play a dominate role in determining biomass and composition of microbial populations. Key determinants of shifts between nutrient mineralization and immobilization include hydroperiod and nutrient inflow. In the southeastern United States, mass loss rates of foliar litter (with litter quality held constant) are maximized by moderate durations of flooding followed by several months of noninundation.

In general, rates of litter mass loss in forested floodplains exceed those of uplands. Globally, decay constants for temperate floodplain forests average approximately 1.00 while the mean for all temperate deciduous forests is less than 0.80. This differential is partly due to the greater availability of soil moisture (better habitat for microbial populations) during parts of the year. However, mass loss, as measured by disappearance of confined litter, includes both mechanical disintegration as well as metabolic conversion of organic carbon and, consequently, periodic inundation offers greater opportunities for disintegration and export.

The general perception that floodplains are very fertile has led to misconceptions regarding the degree to which insufficient nutrient availability may constrain floodplain NPP. In many cases, it is true that floodplain soils are more fertile than upland counterparts. However, vegetation species found in many floodplains often have higher annual nutrient requirements compared to species adapted to uplands. Consequently, forest vegetation on many floodplains is likely to be nitrogen deficient and, in some cases such as blackwater systems, deficient in phosphorus and base cations as well. An example would be the nutrient-demanding *Populus deltoides* Batr. plantations that grow in extraordinarily fertile soils of the Southern Mississippi Alluvial Valley, USA. In spite of fertile soils and high aboveground NPP (20–25 t ha^{-1} yr^{-1}), those systems would increase in NPP if supplied with additional nitrogen.

The degree to which a floodplain ecosystem is deficient or nondeficient for particular nutrients is critical in regard to that system's potential to act as a nutrient sink. As previously mentioned, the kidney function is enhanced if floodplain vegetation can assimilate incoming nutrients from sources such as polluted water or atmospheric inputs. Once a deficiency is eliminated, it is still possible for floodplain vegetation to assimilate particular nutrients such as nitrogen through luxury consumption. However, a level may be reached after which the vegetation's capacity to retain nutrients is saturated. The latter condition reflects a high degree of biotic stress and is a serious threat to floodplain vegetation associated with eutrophic streams.

Vegetation Community Structure and Composition

Vegetation communities in floodplain systems have developed over hundreds of years as a function of soil type, topography, and hydrology. The type of vegetation growing on a particular floodplain will be dominated by trees or shrubs adapted to the environmental conditions of that floodplain. Hydroperiod is the most important local environmental condition determining composition, and the species found respond to elevation differences relative to the river's flooding regime. Typical floodplain forests begin at the natural levee where coarse-grained deposits result in quickly draining soils and continue as surface elevations decrease away from the river and become more poorly drained.

Structural characteristics of floodplain forests vary depending upon location (**Table 1**). Stem density and basal area are generally greater in the southeastern United States and the humid tropics than in arid areas, but in arid areas basal area can still exceed 50 m^2 ha^{-1}. Basal areas in floodplain forests tend to be as high as or higher than that of upland forests. Almost without exception, the number of tree species increases as flooding decreases. The greatest number of tree species occurs in wet, tropical floodplains such as the Amazon. The understory of floodplain forests is generally lower in density and species numbers, probably due to reduced light levels and the extended flooding conditions.

Adaptations of Floodplain Vegetation

Due to the alternating wet–dry environment experienced by trees growing on floodplains, they have developed a variety of physiological and morphological adaptations that allow survival during flooding. Initially, stimulation of alcohol dehydrogenase (ADH), enzyme activity may provide a temporary means to support essential metabolic functions. The anaerobic pathway is less efficient than the aerobic pathway (39 moles ATP per mole hexose vs. 3 moles ATP per mole hexose), but provides an energy resource while anatomical changes are occurring.

Table 1 Mean structural and aboveground productivity characteristics of floodplain forests

Area	No. of species	Density (no. ha^{-1})	Basal area (m^2 ha^{-1})	Biomass (t ha^{-1})	Aboveground NPP		
					Leaf	Wood (t ha yr^{-1})	Total[a]
Southeastern USA	13	1242	45.0	302	5.36	7.78	13.26
Northeastern USA	10	970	26.1	150			
North Central USA	5	546	29.5				
Western USA	5	310	27.5				
Central USA	12	405	33.5	290	4.20	2.50	8.70
Europe		1237	26.5	314	3.48	17.88	
Central America	10	726	49.9	118	11.61		
Caribbean	27	3359	42.4	224	15.55		
South America	89	687	33.0	413			
Africa	26						
Southeast Asia					9.15		
Australia	12	493		260			

[a]Total NPP does not always equal leaf plus wood as some sources only report total.

The seeds of floodplain tree species require oxygen for germination, and even those species that can grow in permanently to nearly permanent flooded conditions (e.g., *Taxodium* and *Nyssa*) require moist, but not flooded, soil for germination and establishment. Occasional drawdowns are necessary for the survival of tree species. Rapid stem elongation, such as been observed with *Nyssa aquatica*, allows the seedling to get its crown above the water surface of subsequent floods. The dispersal and survival of many wetland tree seeds is dependent upon hydrologic conditions. *Taxodium* and *Nyssa* seeds are produced in the fall and winter between the periods of lowest and highest streamflows, giving the seeds the widest possible range of hydrologic conditions. Overall, seed production of many wetland species seems to be linked to the timing and magnitude of hydrologic events.

Stem hypertrophy, commonly called butt swell or buttressing, is characterized by an increase in diameter of the basal portion of the stem and is common in *Taxodium*, *Fraxinus*, *Nyssa*, and *Pinus* species. Basal swelling can extend from just above the ground level to several meters depending upon the depth and duration of flooding. Swelling generally occurs along that portion of the trunk that is flooded seasonally. Increased air space in the swollen portion of the stem allows increased movement of gases within the plant. Ethylene production has been documented to play a regulatory role in altering growth and stem anatomy of woody plants, and has been found to be higher in flooded *Fraxinus* stems with well-defined hypertrophy than those without stem hypertrophy. Lenticel hypertrophy has long been associated with flooding and acts to increase internal gas transport from the stem to the roots. Duration of flooding does not appear to affect the number of lenticels formed but does affect the size. The formation of hypertrophied lenticels under anoxic conditions also appears to be induced by ethylene. Other commonly observed features in flooded environments include buttress roots and knees. Buttress roots appear as fluted projections at the base of mature trees and extend for several feet from the trunk outward and down into the soil. Because of the shallow nature of root systems in saturated or flooded soils, these buttress roots are thought to provide additional support to the tree. Knees are common in *Taxodium* spp. in the southeastern United States. Their function has not been confirmed, although there is some speculation that they also serve in stability of the tree. In Australia, *Melaleuca* trees on floodplain sites have modified bark structures such as papery bark with internal longitudinal air passages that allow them to tolerate flooded conditions.

Productivity

Riverine floodplains are typically characterized by high productivity. Productivity is enhanced in many floodplain areas by the continued import and retention of nutrient-rich sediments from headwater regions and lateral sources, increased water supply (especially in arid regions), and more oxygenated root zones as a result of flowing waters. The flood pulse advantage has long been recognized, with ancient Egyptians setting taxes based on the extent of the annual flood.

Primary productivity of unaltered, seasonally flooded ecosystems is generally higher than that of floodplain forests that are permanently flooded or those with stagnant waters. Despite the theoretical basis for increased floodplain productivity due to pulsing, it has been difficult to confirm. More recent studies tend to point toward the idea that seasonal flooding can be both a subsidy and a stress. In the southeastern United States, aboveground NPP was similar for upland hardwood, bottomland hardwood, and *Taxodium–Nyssa* forests. The reason for this may be that for some sites, subsidies and stresses occur simultaneously and cancel one another. As a result, flood intensity and duration affect soil moisture, available nutrients, anaerobiosis, and even length of growing season in a complex and nonlinear `push–pull' arrangement. When hydrology is altered rapidly, aboveground productivity is less than in natural forest communities with nearly continuous flooding (**Fig. 6**).

Aboveground biomass in floodplain forests ranges between 100 and 300 t ha^{-1}, although there is one report of a forest in Florida where biomass exceeds 600 t ha^{-1}. Leaves account for only 1–10% of the total aboveground biomass. Belowground biomass has been sampled rarely and varies greatly, but reported values tend to be somewhat lower than the 20% of

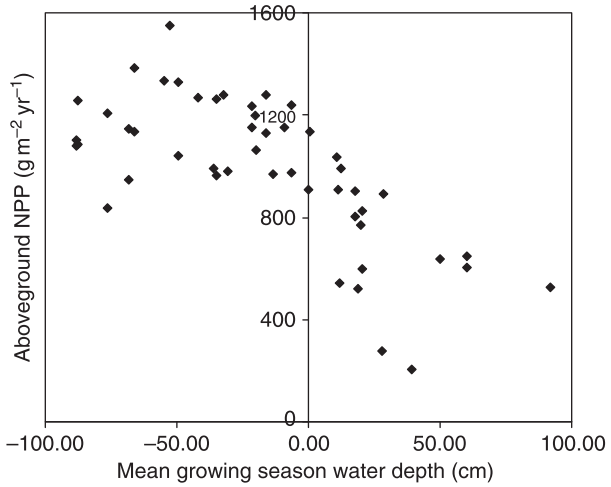


Fig. 6 Relationship between aboveground net primary productivity (NPP) of floodplain forests of the southeastern United States and mean water depth during the growing season.

total biomass often cited for upland species. Total aboveground biomass production (leaves plus stem wood) ranges from 668 to 2136 g m^{-2} yr^{-1}, with leaves accounting for approximately 47% of the production. Although it has been reported that there are no latitudinal patterns in NPP, litterfall production of *Taxodium* forests in the United States shows a curvilinear relationship with latitude with a maximum occurring at about 31.9° N. In northern Australia, litterfall in *Melaleuca* forests has been reported to be 2–3 times greater than that in forests in the southern part of the continent. Changes to natural hydrologic regimes decrease litter production by half. As a result of the high productivity, generally associated with floodplain forests, carbon sequestration is particularly important there.

Anthropogenic Impacts

Rivers and associated floodplains have been vitally linked to civilization throughout history for food production. In order to make farming easier and more productive, rivers have been diverted and floodplains have been deforested and drained or leveed to provide fertile land. A major consequence of the widespread use of floodplains and adjacent uplands for agriculture has been the generation of large sediment loads in associated streams and rivers. As a result, much sediment has been deposited on streambeds and floodplains with negative consequences for aquatic habitat and floodplain vegetation. More recently, impoundments have become commonplace for energy production and water storage and levees continue to be built to provide space for development as well as for farming. Globally, it is estimated that, at a minimum, 75% of total floodplain area has been lost.

Floodplain function is dependent on connectivity between the river and its riparian area. Unfortunately, many anthropogenic impacts eliminate or reduce that connectivity so that key functions such as water filtration are much reduced at the landscape level. Similarly, alterations in hydroperiod caused by human activity often drive changes in composition and productivity of vegetation communities as those species adapted to the former conditions decline and are replaced by others.

Additional impacts include fragmentation of riparian vegetation communities and stimulation of invasive non-native plant invasion. Fragmentation often results in reduced habitat quality while successful invasion by non-native species may cause major alterations in community composition, structure, and function. While ecological restoration of floodplains has attracted widespread interest, economic constraints have primarily limited restoration applications to localized areas. However, notable exceptions include restoration of the portions of the Pantanal River Basin in South America and the Kissimmee River Corridor in Florida, USA.

More recently, urbanization has led to significant and growing impacts on floodplains in many parts of the world. As catchments become developed, the concomitant rise in impervious surface drives major increases in runoff volume and velocity. As a result, rising limbs during flood events become much steeper, a condition that is often associated with higher in-channel velocities. Higher flow velocity increases the rate of channel incision resulting in a lowered groundwater table and reduced connectivity between the stream and floodplain. In addition, urbanization stimulates loadings of nutrients (particularly, nitrogen) and causes a considerable degree of water pollution in general.

Further anthropogenic impacts include channelization of river systems. Channelization has benefited farming and waterborne transportation by reducing flooding and removing obstacles to barge and other water traffic. However, water quality has suffered in many instances since there is again less opportunity for river waters to contact floodplain surfaces and undergo pollutant reduction.

Africa

The African continent has approximately 99 large wetlands, of which 43 are floodplain systems. Some of the larger floodplain systems include the Zaire Swamps (200 000 km^2), the Inner Niger Delta of Mali (320 000 km^2 when flooded), the Sudd of the Upper Nile (16 500 km^2 of permanent swamp and 15 000 km^2 of seasonal floodplain), and the Okavango (14 000 km^2 of permanent swamp and 14 000 km^2 of seasonal floodplain). These floodplain systems are in dynamic equilibrium with the constant flux of pulsing events occurring within them at different spatial and temporal scales. Goods and services resulting from pulsing events include floodplain recession agriculture, fish production, wildlife habitat, livestock grazing, ecotourism, and biodiversity, as well as natural products and medicine.

In semiarid and arid regions of Africa, floodplains are often the only source of year-round water. As in other floodplains around the world, vegetation distribution is strongly related to flooding frequency and duration and microtopography. Dense evergreen tree growth occurs on higher well-drained areas like levees and termite mounds, while grasslands tend to dominate lower, more frequently flooded areas. Typical grasses found growing in these frequently flooded areas (called swamp) include *Phragmites*, *Typha*, and *Polygonum*. Tree and bush genera in less frequently flooded areas include *Hyphanene*, *Borassus*, *Acacia*, *Ficus*, and *Kigelia*.

Floodplain areas are centers of high diversity of animal and plant life. These floodplain areas are of profound importance for fish production and probably serve as spawning and recruitment areas. Interannual fluctuations in fish production have been correlated with the flooding regime. Numerous bird species (over 400 in some floodplains) can be found in these areas, including bee-eaters, jacanas, malachite kingfishers, grey herons, egrets, African fish eagles, and Zaire peacocks. The birds share the floodplains with antelope (sitatunga, waterbuck, puku, and lechwe), hippopotamus, zebra, and buffalo; vegetation ranges from water lilies and papyrus to floodplain forests with minor topographical variations playing an important role in distribution of forest and grassland. Climatic variations are also important, with forest only occurring near rivers in drier areas while in wetter areas forests can extend for a considerable distance away from the river. River meanders tend to be cut off during flooding periods, adding diversity to the floodplain topography.

Unfortunately, very few studies of the ecology of many of the African floodplain systems have been carried out. The most studied floodplain system in Africa is the Okavango Delta. Annual floods travel uninterrupted down the Okavango River and inundate the Okavango Delta from April to September. River water is characterized by moderate levels of nutrients, but when it enters the floodplain it becomes strongly enriched by nutrients via leaching from soil, detritus, and feces. Organic carbon enrichment comes from leaching of floodplain leaf litter and soil, although dissolved organic carbon release from leaf litter is over 2 orders of magnitude greater than for leached soils. This nutrient enrichment has a major impact on aquatic productivity in the delta and illustrates the strong links between terrestrial and aquatic ecosystems.

African floodplains face a different set of challenges as opposed to those in developed countries. In Africa, floodplains generally occur in semiarid areas to arid regions, and flooding is the driving force behind the high productivity of these areas. From as early as the 900s, people have inhabited these areas, and pastoral and agricultural economies are dependent upon the continued presence of the floodplains. Continued pressures from agricultural practices within the floodplains themselves and population growth that demands the transfer of water to alleviate shortages outside of the floodplain need to be addressed to ensure survival of these important ecosystems.

Asia

In northern Asia, there are extensive productive wetlands along the floodplains of rivers. In western Siberia, the river Ob extends over 50 000 km^2 and supports what is called the largest waterfowl breeding and moulting area in Euroasia. The Ob Valley is a labyrinth of intricately arranged channels and floodplain lakes. As in other seasonal floodplains, the region is a land of fluctuating water levels, with seasonal and annual fluctuations in river discharge and flooding patterns. This area avoided any serious human impacts for centuries, but oil and gas exploration has resulted in significant pollution and transformation of the landscape.

The Indus River has long been the lifeblood of arid Pakistan. In earlier times, people used the river's water to cultivate the floodplain, but during the last 100 years, the river has been dammed and diverted into one of the largest and most complex irrigation systems in the world. In the absence of a drainage system to remove irrigation water, evaporation leaves salt in the soil. As a result of this salinization of the soil, combined with waterlogging, over 400 km^2 of irrigated land is lost each year.

Many of the large river systems in South Asia display considerable annual variation in discharge and, during the rainy season, may flood very large expanses of land (e.g., approximately 200 km on each side of the Ganges). In some cases, entire deltaic areas may be inundated. The prolonged, monomodal flooding promotes extensive spatial and temporal contact with floodplains and, consequently, dominates the socioeconomics of the large human populations near those systems. Agricultural activities often cause significant sediment export from upper reaches of many rivers and, as a result, delta tributaries may become clogged. Due to the subsequent reduced flow, salinity can increase in soils and alter species composition in the delta forests.

About 80% of Bangladesh (115 000 km^2) is formed by floodplains of the Ganges, Brahmaputra, and Meghna rivers. In major floods, 57% of the country can be flooded. Availability of water during the dry season makes it possible to grow three crops a year in some areas. Deposition of waterborne sediments keeps the soils fertile and algal growth enriches the soil by fixing nitrogen. As in many parts of the world, forest vegetation of South Asian floodplains strongly reflects variations in hydroperiod and soil. In the Ganges and Brahmaputra River Valleys, within areas with heavy clay soils where flooding occurs for most of the year or

permanently, forest vegetation may be only 5–10 m in height and occur in conjunction with numerous vines. However, combinations of similar flooding regimes and lighter, fertile soils may increase canopy heights by 10 m or more. Many of these riverine forests exhibit a prevalence of evergreen or semievergreen species, although at higher altitudes alders may dominate. Some lowlands, in particular many river deltas such as those of the Ganges–Brahmaputra and Irrawady, are occupied by mangrove forests.

In China, 95% of the population is concentrated in the eastern half of the country, mainly in the vast alluvial plains of the major rivers, the Yellow and Yangtze Rivers primarily. High population densities coupled with high growth rates, rapid urbanization, and industrialization play a major role in most Asian countries. Water resources in this region are under increasing pressure as the demand for domestic supplies, agricultural use, and hydroelectric power increases. Past water resource and agricultural management practices have resulted in rapid loss and degradation of natural wetlands throughout the region. The regulation of rivers and streams through embankments and dams has eliminated floodplains and reduced groundwater recharge. Changing hydrological regimes have increased flooding during the rainy season and reduced availability during dry periods. Water resource management has often resulted in numerous man-made wetlands such as reservoirs and paddy fields that have very different functions and values than natural wetlands, and are in no way a substitute for natural wetlands, particularly floodplain wetlands. In short naturally occurring floodplains in these regions are threatened by numerous human activities, including mining, aquaculture, unsustainable forestry or fisheries practices, and conversion of forests to urban or agricultural land.

Australia

Australia is distinctive in that there are few permanent wetlands due to high evaporation rates and low rainfall. Most wetlands on the continent are intermittent and seasonal. Common features of floodplains are waterholes and lagoons called billabongs that retain water seasonally or permanently, providing important habitat for many animals at different times of the year. Floodplain wetlands tend to be sites of extraordinary biological diversity of waterbirds, native fish, invertebrate species, aquatic plants, and microbes. Key drivers of this biodiversity are the lateral connectivity to the river of the floodplain wetland and the unpredictable flows that create wide ranges of temporally and spatially different aquatic ecosystems.

Humid coastal areas are drained by short, perennial streams, while much of the streamflow in the rest of the country is intermittent or nonexistent because of low and unreliable rainfall, high evaporation, and flat topography. Even under these conditions, forested wetlands can be found throughout Australia, but they can only be classed as true forests in the wettest localities. The largest area of floodplain forested wetland (over 60 000 ha) occurs on the Murray River. Floodplain forests are generally composed of *Melaleuca* or *Eucalyptus* species, but they cannot survive very long periods (>5 months) of flooding. If flooding exceeds several weeks during the growing season, forest canopy cover declines to between 10% and 70%, creating open woodlands.

Tropical floodplain wetlands are found across northern Australia, covering an estimated 98 700 km^2. Vegetation of these wetlands has been mapped at various scales, but there are few specific or long-term analyses of the distribution or successional changes of the plants. The Ord River floodplain in northern Australia encompasses approximately 102 000 ha and is a large system of river, tidal mudflat and floodplain wetlands that supports extensive stands of mangroves, large numbers of waterbirds, and significant numbers of saltwater crocodiles. In southeastern Australia, the Murray–Darling river system drains 14% of the continent and contains the greatest amount of floodplain wetlands on the continent.

In recent years, floodplain areas have undergone considerable change because of animal (buffalo, pigs, cane toads) and plant (mimosa, salvinia, paragrass) invasions, changes in fire regimes, water resource management, and saline intrusion. Dams and the cumulative impact of diversions and upstream river management have turned many floodplains into terrestrial ecosystems. The effect of this change in flooding has not been well studied and data exist only for a fraction of the area affected. Floodplain loss will continue until there is a better understanding of the long-term ecological effects of dams and diversions.

Europe

Floodplains in Europe have been influenced by humans for thousands of years. Civilizations often were established near rivers and frequently utilized floodplain resources for food (agriculture or hunting), power (wood or water mills), and shelter. As communities grew there was an increased need to control the flooding that naturally occurred in the floodplains with the use of dams, dikes, and ditching. These structures altered the hydrology which in turn has altered the forest composition in these areas. Furthermore, channel straightening has caused major hydrologic changes resulting from faster flow and increased groundwater depth. In some of the Danube watersheds, there has been an 80% decrease in first-order streams from 1780 to 1980.

In many areas, depth to groundwater has increased due to the `drying' of the floodplains and this has driven shifts in the composition of vegetation communities. In particular, species such as *Quercus robur*, *Fraxinus* spp., and *Ulmus* spp. are becoming rarer due to the altered hydrology. Forestry practices have induced a further shift from natural systems to faster growing *Populus* clones in many of the floodplains across central Europe. However, reestablishment of the more traditional forest composition of uneven-aged oak, ash, and maple mixes has been achieved in some areas as recently as the past 50 years. Large portions of the forests remain monocultures of even-aged *Acer* or *Fraxinus*.

The Danube Delta represents one of the largest wetlands in Europe and is undergoing eutrophication as a result of increasing nutrient inputs, decreased riparian vegetation, and loss of the filtration function. One major difference between European floodplains and others worldwide is that increased flow and flooding often occurs in the spring as a result of snowmelt in high altitudes.

North America

In the dry climate of the Western United States, water is a limited resource not only for the wildlife but also for the human inhabitants. Although wetland areas comprise a very small portion of total land area (i.e., less than 2%), over 80% of wildlife is dependent on their presence. Rainfall in this region varies from less than 15 cm yr^{-1} in the desert regions to greater than 140 cm in the mountains. In the mountainous regions, rainfall and snowmelt are greater than losses and, therefore, wetlands rarely dry out. However, evapotranspiration in the basin areas is 3–4 times greater than precipitation and, consequently, soil salinization is a stress to which vegetation must adapt. In the driest regions soil salinization prevents vegetative establishment. Ephemeral drains are prevalent in the intermountain west with snow melt and high rain contributing to their flow.

At higher elevations in the United States, where soils are semipermanently inundated or saturated, associations of *Populus*, *Salix*, and *Acer* are found. Floodplain areas flooded or saturated 1–2 months during the growing season are comprised of a wide array of hardwood trees. Common species in the United States include *Fraxinus* spp., *Tilia* spp., *Ulmus* spp., *Liquidambar* spp., *Celtis* spp., *Acer* spp., *Plantanus* spp., and some *Quercus* spp. At the highest elevations, flooding occurs for less than a week to about a month during the growing season. Typical tree species include a variety of *Quercus* spp. and *Carya* spp., with some *Pinus* spp.

Floodplains of the southeastern United States occur within three physiographic regions: (1) coastal plains, (2) piedmont, or (3) Appalachian Mountains. Rainfall is sufficiently prevalent during all seasons except for brief periods of drought. Successional patterns of southern forested floodplains are often dictated by hurricanes, tornados, catastrophic ice storms, and extended drought. Soils are typically acidic, with the exception of near neutral pH soils across much of the Southern Mississippi Alluvial Floodplain and the Selma Chalk geologic region of Alabama and Mississippi. In many floodplains, as one moves in a direction perpendicular to the river, soil textures range from coarse sands near stream channels, fine sands in natural levees, to loams and clays in backwater areas. This separation pattern is a result of particle size and sheet flow velocity.

The lowest elevation, nearly always flooded sites on floodplains in the southeastern United States are occupied by *Taxodium–Nyssa* swamps. In other parts of the world, it appears there are no similar tree species that can survive permanent or long periods of inundation. As long as the floodplain channel remains stable and flooding frequency remains constant, these species should dominate the stands indefinitely.

South America

Much of our current knowledge about forested floodplains has been derived from extensive studies performed in the sub-basins of the Amazon River. In particular, our understanding of floodplain biogeochemistry, NPP, vegetation dynamics, geomorphology, and faunal relationships has been greatly influenced by Amazonian research.

In comparison to river basins in other parts of the world, the water balance of Amazonia lowlands is roughly evenly divided between evapotranspiration and runoff. This contrasts with systems in Asia where runoff dominates due to generally steeper terrain and many African systems where broad floodplains and high potential evapotranspiration result in low runoff. Floodplain forests in South America are typically composed of a small number of fast-growing, early-successional species capable of surviving periodic floods and large amounts of sediment deposition (e.g., *Salix* and *Inga* spp.).

The `flood pulse' concept was originally conceptualized in relation to the Amazon and similar floodplains and can be applied worldwide. The major river floodplains of South America such as the Amazon, Orinoco, and the Parana display singular, river-borne flood pulses of large amplitude and duration. In contrast, inundation on floodplains situated within large depressions such as the Pantanal is generally rainfed (as opposed to overbank flow from rivers), and also displays a singular periodicity but with lower amplitude. Finally, multiple flood pulses that are less predictable in terms of occurrence and amplitude are characteristic of floodplains associated with smaller order streams.

Some of the classic research that defined global variation among floodplains took place in Amazonia and was associated with contrasts between blackwater versus brownwater or whitewater rivers. Similar types of floodplain systems occur in many parts of the world. The color of the river waters is reflective of the geomorphology of particular systems and is a strong indicator of floodplain biogeochemistry, vegetation dynamics, and NPP. Whitewater rivers in the Amazon Basin derive their color from white clay sediments that originate in the Andes. The suspended clays contain higher levels of nutrients (particularly base cations) which, when deposited, often create fertile floodplains labeled varzea.

In contrast, blackwaters are stained by fulvic acids and other organic compounds and are more acidic than whitewater counterparts (pH <5.0 vs. >6.0 for blackwater and whitewater, respectively). Due to the low sediment loads, floodplains associated with blackwater streams are often nutrient poor and are referred to as igapo. Consequently, forest litterfall production on varzea floodplains is often considerably higher than that of the igapo (approximately 10 vs. 5 t ha^{-1} yr^{-1} for the respective system types). Also, the standing crop of fine roots is much higher in igapo soils compared to varzea, a reflection of greater

belowground allocation of biomass as would be expected in resource-poor soils. Such adaptations increase the likelihood of nutrient capture from decomposing igapo litter. The distinctions in hydrology and biogeochemistry between the igapo and varzea also drive major differences in vegetation species occurrence, root, shoot, and reproductive phenology, and community structure.

Distinctions between floodplains types are also important in regard to animal populations. This is particularly true for fish which depend on interactions with inundated floodplains for resource acquisition, reproductive habitats, and other factors. As an example, the lower NPP on igapo floodplains may translate to lower food resources for fish. While the amount of plant detritus exported from varzea floodplains is higher, phytoplankton production also depends on settling of the clay sediments so that sufficient light can penetrate the waters. Although more difficult to document in riverine systems, fish catches are generally much lower in igapo lakes compared to varzea counterparts.

As is the case in much of the world, South American floodplain ecosystems are under pressure from an array of human activities. As an example, the lower reaches of the Parana' River have undergone changes in hydrology due to construction of dams and upstream expansion of agriculture. The altered hydrology, along with increased concentrations of sediment and other contaminants have resulted in heavy impacts to fish populations and concomitant economic declines in local fishing communities. Although there is a growing voice for conservation and protection of natural resources, it is unclear to what extent anthropogenic impacts may be curtailed.

Summary

As pathways between aquatic and terrestrial ecosystems, floodplains perform a myriad of functions that are critical to humanity and all other components of the biosphere. Because of the vital need of all organisms for clean water, the kidney or filtration function is the most important attribute of healthy floodplain systems.

The filtration function entails sediment and nutrient deposition and, consequently, has long made floodplains very attractive for exploitation as agricultural sites. It is ironic that the very function that makes floodplains so important attracts major disturbances which, in turn, result in destruction of the kidney function in those systems. Globally, that destruction is reflected in the magnitude of floodplain loss (i.e., 75%).

While the primary cause of floodplain destruction is shifting from agriculture to urban development, it would be unrealistic to expect that the general magnitude of anthropogenic pressures on these systems will abate. Consequently, an answer to the critical question of whether adequate supplies of clean water exist will become increasingly uncertain. In order to provide a positive answer and, subsequently, protect human health and well-being, it is vital that we more clearly understand how these ecotones operate so that functional floodplains can be maintained and integrated into evolving landscapes.

See also: Global Change Ecology: Microbial Cycles

Further Reading

Brinson, M.M., 1990. Riverine forests. In: Lugo, A.E., Brinson, M.M., Brown, S.L. (Eds.), Forested Wetlands, Vol. 15: Ecosystems of the World. Amsterdam: Elsevier Science Publishers, pp. 87–141.
Cavalcanti, G.G., Lockaby, B.G., 2005. Effects of sediment deposition on fine root dynamics in riparian forests. Soil Science Society of America Journal 69, 729–737.
Groffman, P.M., Bain, D.J., Band, L.E., *et al.*, 2003. Down by the riverside: Urban riparian ecology. Frontiers in Ecology and the Environment 6, 315–321.
Hupp, C.R., 2000. Hydrology, geomorphology and vegetation of coastal plain rivers in the south-eastern USA. Hydrological Processes 14, 2991–3010.
Junk, W.J., 1997. Ecological Studies 126: The Central Amazon Floodplain: Ecology of a Pulsing System. Berlin: Springer.
Lewis Jr., W.M., Hamilton, S.K., Lasi, M.A., Rodriguez, M., Saunders III, J.F., 2000. Ecological determinism on the Orinoco floodplain. Bioscience 50, 681–692.
McClain, M.E., Victoria, R.L., Richey, J.E., 2001. The Biogeochemistry of the Amazon Basin. New York, NY: Oxford University Press.
Megonigal, J.P., Conner, W.H., Kroeger, S., Sharitz, R.R., 1997. Aboveground production in southeastern floodplain forests: A test of the subsidy-stress hypothesis. Ecology 78, 370–384.
Messina, M.G., Conner, W.H., 1998. Southern Forested Wetlands: Ecology and Management. Boca Raton, FL: CRC Press.
Mitsch, W.J., Gosselink, J.G., 2000. Wetlands, 3rd edn. New York, NY: Wiley.
Naiman, R.J., Decamps, H., 1997. The ecology of interfaces: Riparian zones. Annual Review of Ecology Systematics 28, 621–658.
National Academy of Science, 2002. Riparian Areas. Functions and Strategies for Management. Washington, DC: National Academy Press.
Paul, M.J., Meyer, J.L., 2001. Streams in the urban landscape. Annual Review of Ecology and Systematics 32, 333–365.
van Splunder, I., Coops, H., Voesenek, L.A.C.J., Blom, C.W.P.M., 1995. Establishment of alluvial forest species in floodplains: The role of dispersal timing, germination characteristics and water level fluctuations. Acta Botanica Neerlandica 44, 269–278.

Forest Plantations

D Zhang, Auburn University, Auburn, AL, USA
J Stanturf, Center for Forest Disturbance Science, Athens, GA, USA

Between the extremes of afforestation and unaided natural regeneration of natural forests, there is a range of forest conditions in which human intervention occurs. Previously, forest plantations were defined as those forest stands established by planting and/or seeding in the process of afforestation or reforestation. Within plantations, there is a gradient in conditions. At one extreme is the traditional forest plantation concept of a single introduced or indigenous species, planted at uniform density and managed as a single age class (the so-called monoculture). At the other extreme is the planted or seeded mixture of native species, managed for nonconsumptive uses such as biodiversity enhancement. To further complicate matters, many forests established as plantations come to be regarded as secondary or seminatural forests and no longer are classed as plantations. For example, European forests have long traditions of human intervention in site preparation, tree establishment, silviculture, and protection; yet they are not always defined as forest plantations.

Further refinement of the plantation concept is necessary in order to encompass the full range of actual conditions. A useful typology is based on purpose, stand structure, and composition of plantations. Thus, an industrial plantation is established to provide marketable products, which can include timber, biomass feedstock, food, or other products such as rubber. Industrial plantations usually are regularly spaced with even age classes. Home and farm plantations are managed forests but at a smaller scale than industrial plantations, producing fuelwood, fodder, orchard, and garden products but still with regular spacing and even age classes. A wide range of agroforestry systems exist, distinguishable as a complex of treed areas within a dominantly agricultural matrix. Environmental plantations are established to stabilize or improve degraded areas (commonly due to soil erosion, salinization, or dune movement) or to capture amenity values. Environmental plantations differ from industrial plantations by virtue of their purpose; they may still be characterized as regularly spaced with even age classes. Efforts to restore forest ecosystems are increasing and often utilize the technology of plantation establishment, at least initially.

Recently, FAO defined `planted forests' as forests in which trees have been established through planting or seeding by human intervention. This definition is broader than plantations and includes some seminatural forests that are established through assisted natural regeneration, planting or seeding (as many planted forests in Europe that resembled natural forests of the same species mix) and all forest plantations which are established through planting or seeding. Planted forests of native species are classified as forest plantations if characterized by few species, straight, regularly spaced rows, and/or even-aged stands. Forest plantations may be established for different purposes and were divided by FAO into two classes: protective forest plantations which are typically unavailable for wood supply (or at least having wood production as a secondary objective only) and often consist of a mix of species managed on long rotations or under continuous cover; and productive plantation forests which are primarily for timber production purposes.

Fig. 1 shows that, in 2005, some 36% of global forests (about 4 billion ha, covering 30% of total global land area) are natural forests, 53% are modified natural forests, 7% are seminatural forests, and the remaining 4% are forest plantations. Of these forest plantations, productive forest plantations account for 78% and protective forest plantations account for 22%. While natural forests and modified natural forests declined between 1990 and 2005, seminatural forests and forest plantations increased (Fig. 2).

This article provides an overview and economic explanation of global forest plantation development. It also presents factors influencing global forest plantation development and lists the usefulness of forest plantations, including their roles in the conservation of natural forests. Finally, it summarizes the impact of forest plantations on biodiversity and other ecological functions.

An Overview and Economic Explanation of Global Forest Plantation Development

Currently, there are about 109 million ha of productive forest plantations in the world. Productive forest plantations represented 1.9% of global forest area in 1990, 2.4% in 2000, and 2.8% in 2005. The Asia region accounted for 41%; Europe 20%; North and Central America 16%; South America and Africa 10% each; and Oceania 3%.

Forest plantations have been increasing at an increased rate. The area of forest plantations increased about 14 million ha between 2000 and 2005 or about 2.8 million ha per year, 87% of which are in the productive class. The area of productive forest plantations increased by 2.0 million ha per year during 1990–2000 and by 2.5 million ha per year during 2000–05, an increase of 23% compared with the 1990–2000 period. All regions in the world showed an increase in plantation area, with the highest plantation rates found in Asia, particularly in China. The ten countries with the greatest area of productive forest plantations accounted for 79.5 million ha or 73% of the total global area of productive forest plantations (Fig. 3). China, the United States, and the Russian Federation together accounted for more than half of the world's productive plantations.

 doi:10.1016/B978-0-444-63768-0.00331-0

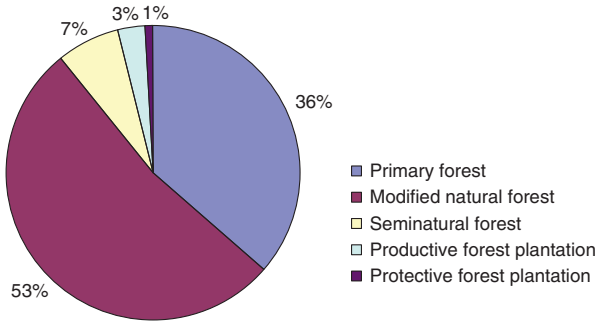


Fig. 1 Global forest characteristics 2005. Modified from FAO (2005) Global forest resources assessment 2005. *FAO Forestry Paper 147*. Rome, Italy.

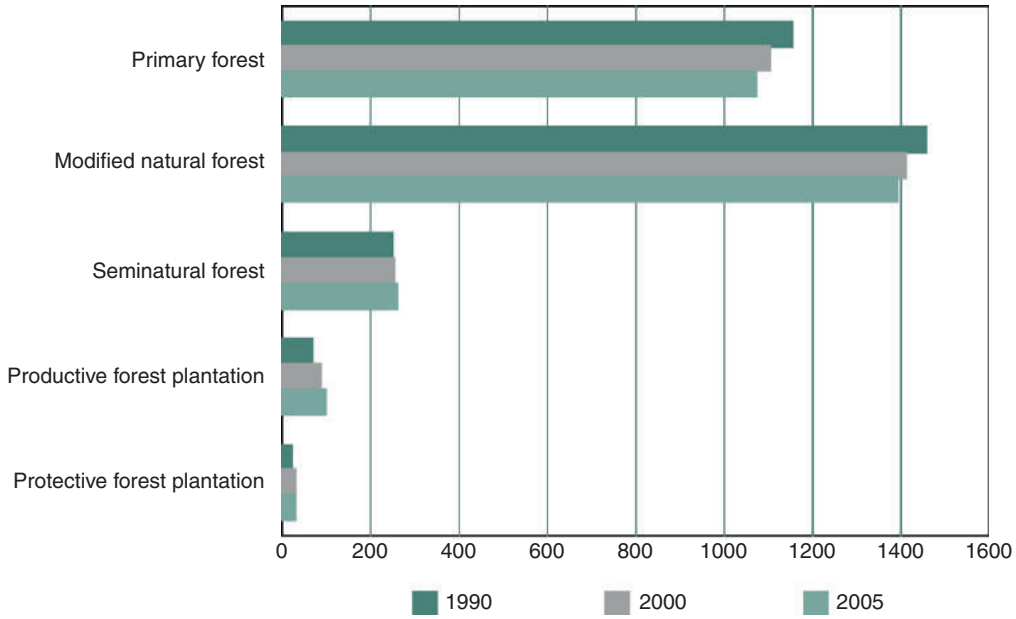


Fig. 2 Global trends in forest characteristics 1990–2005 (million ha). Modified from FAO (2005) global forest resources assessment 2005. *FAO Forestry Paper 147*. Rome, Italy.

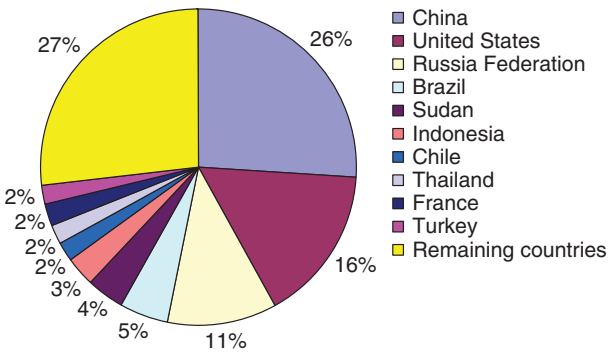


Fig. 3 Ten countries with largest area of productive forest plantations in 2005. Modified from FAO (2005) Global forest resources assessment 2005. *FAO Forestry Paper 147*. Rome, Italy.

Forest plantations, productive or protective, develop in response to a relative scarcity of timber and other goods and services associated with forests. In the early part of modern human history, population was sparse, forests were abundant, and survival, economic development, and territorial control were the primary concerns of governments and society. As forest resources declined, assuring an adequate timber supply gradually caught the attention of rulers and planners and became state policy. Often, the very

first policy implemented would be to regulate timber harvesting schedule and intensity. Society also responded by moving to frontiers farther and farther away from population centers, which in economic terms is called a shift in the extensive margin of timber production. In a nutshell, the production and consumption of forest products were all from natural forests in the early part of human history, and forest plantations were not needed.

When the increase in timber consumption caught up with the ability of a country or a region to produce timber in naturally regenerated forests, citizens and governments would become interested in tree planting. While tree planting occurred at least several thousands of years ago in the Middle East, China, and Europe, and nearly 200 years ago in the Americas, the areas planted with trees through afforestation (planting land that was formerly in a nonforest cover) and reforestation (planting land on which a former forest had been harvested) were relatively insignificant in size before AD 1800. It was only after the industrial revolution that timber consumption increased drastically, due to increasing human population and industrial use of wood – initially as charcoal, then lumber, other solid wood products including mine props and railroad ties, and pulp and paper, and finally for conservation uses – that large-scale forest plantations started to emerge in Europe, North America, Asia, and other regions in the last century, especially in the last few decades.

Thus, forest plantations develop primarily in response to economic necessity. Timber depletion drives the transition of human consumption of natural forests to artificial forests. Early in the development of North America, for example, timber prices were low, and forest lands were more valuable for other uses, especially the production of food. So trees were removed, forest lands were converted to other use, and timber inventory declined. As the standing inventory declines, timber becomes increasingly scarce and timber prices start to rise. As the prices continue to rise for timber in natural forests, the purposeful husbandry of planted forests becomes economically attractive, and productive forest plantations begin to emerge.

Further, timber depletion affects the supply and demand balance for environmental services from natural forests, whether or not these services go through formal markets. Related to this balance is the fact that the demand for most environmental services such as clean water, clean air, and esthetics, which are often produced from or protected by forests, is highly correlated with personal income. As personal income increases, society demands more environmental services from forests, as well as more wood commodities. When natural forests are depleted to the extent that they cannot adequately provide these services, protective forest plantations emerge. In some developing countries, subsistence farming requires forests to protect farming and grass land from potential flooding, dust storms, soil erosion, and desertification, and trees are thus planted for protective purposes whether or not their personal incomes actually grow over time.

Factors Influencing Forest Plantation Development

As mentioned earlier, rising timber prices, caused by timber scarcity, lead to forest plantation development. Thus, timber prices are the primary factor that influences forest plantation development. Holding everything else constant, whenever a country or a region experiences a long period of rising timber prices, forest plantations would develop quickly.

Tree planting also requires land, labor, and capital. The cost of these production factors thus influence forest plantation development. Further, high timber prices, high land costs, and high labor costs force innovation in tree-growing technologies in conventional silvicultural treatments and biotechnologies. A recent report shows that the growth rate of pine plantations in Alabama, a southern state in the US increased about 25% in a decade (from 8.20% in the period from 1982 to 1990, to 10.17% in the period of 1990–2000). This increase in growth rate is attributed to advancement in tree-growing technologies as well as an increase in management intensity.

Government policies influence forest plantation developments as well. Taxes on land and forest-related income, cash subsidies to plant trees, regulations on land use and labor, and free education and extension services to forest farmers all have an impact, positively or negatively, on tree planting. In general, the primary motivation for the private sector to plant trees is to generate financial (or other) benefits from their investment. In some cases, government policies (positive or pervasive incentives in taxes, subsidies) provide or take away a significant proportion of the financial benefits from forest plantation development. Where governments own land, they could conduct afforestation and reforestation activities directly, for purely financial reasons or for social and environmental benefits or both.

The US South is perhaps an important region in timber supply as it produces some 18% of the world's industrial round wood with just 2% of the world's forestlands and 2% of the world's forest inventory. Some 90% of forest lands in the southern US are owned by nonindustrial private and industrial owners, and timber markets are competitive. A study of tree planting showed that tree planting by both forest industry and nonindustrial private landowners was positively related to the availability (measured as previous-year harvest) and the price of land. Planting by forest industry and nonindustrial private landowners was responsive to market signals, positively to softwood pulpwood prices and negatively to planting costs and interest rates. Finally, government subsidy programs, which increase the total plantation area, might have substitution effects on nonindustrial private tree planting. The federal income tax break for reforestation expenses promoted reforestation in the southern US.

Since forests often have a long production cycle, perhaps the most important government policy in promoting forest plantation development is to provide long-term and secure property rights (private property or land tenure) to private landowners or forest farmers. Many theoretical and empirical studies substantiate that long-term and secure property rights promote tree planting activities in both developed and developing countries. For example, in British Columbia, Canada tree planting was done more

often and more promptly following harvest when forest property rights were secure. In Ghana, reforestation was significantly influenced by the form of forest tenure, and more intensive resource management was fostered by more secure forms of tenure.

Forest Plantations and Conservation of Natural Forests

Plantation forests can provide most goods and services that are provided by natural forests. These include timber, nontimber forest products, protection of clean water and clean air, soil erosion control, biodiversity, esthetics, carbon sequestration, and climate control. Nonetheless, as the value of environmental services from natural forests is higher than that from forest plantations, the demand for conservation of natural forests is stronger. It is possible that a division of land, with some land specialized in timber production and other land in providing environmental services, would produce more forest-related goods and services to society. Because forest plantations grow much faster than natural forests, forest plantations are seen as an increasingly important source of timber supply. Should more forest plantations be developed, more natural forests might be saved.

In 1995, natural forests contributed some 78% of global industrial timber supply, and the remaining was from forest plantations. With growing concerns about the status and loss of natural forests, the rapid expansion of protected areas, and large areas of forest unavailable for wood supply, plantations are increasingly expected to serve as a source of timber. The general trend of the sector is for timber supply to shift from natural forests to plantations.

A simple simulation of global timber supply and demand, allowing forest plantations and their productivity to extend at the current rate, has shown that logging on natural forests could fall by half, from about 1.3 billion m^3 in 2000 to about 600 million m^3 in 2025. Thus, forest plantations will have an increasingly significant role in substituting products from natural forests, even if they cannot replace harvests from natural forests for a long period of time.

One side impact of forest plantation development is that the supply of large quantities of low-cost timber could perhaps undermine the value of natural forest stands, leading to more rapid destruction, especially where legal frameworks and law enforcement are inadequate. Therefore from a global perspective, the transition from natural forests as the primary source of timber supply to forest plantations will take a long time. Nonetheless, the transition has been completed in some countries such as New Zealand and Chile.

Direct Ecological Effects of Forest Plantation

Forest plantations have direct ecological effects in addition to the positive impact of reducing pressure on natural forests. Generalizations are difficult, however, in part because plantation management regimes are diverse and the appropriate comparison is not always to unmanaged natural forests. In worst-case scenarios, natural forests or savannas on fragile soils are converted to plantations of exotic species that lower groundwater tables, decrease biodiversity, and develop extreme nutrient deficiencies in successive rotations. While this scenario overstates the impact of plantations, their generally monoculture nature and intensive management raises concerns about the effect of plantations on biodiversity, water, long-term productivity and nutrient cycling, and susceptibility to insects and diseases.

Biodiversity illustrates the complicated ecological impact of forest plantations; although biodiversity encompasses genetic, species, structural, and functional diversity, much of the focus in discussions about diversity has been at the genetic, species, and local ecosystem levels. As has occurred in agriculture, the introduction of genetically improved exotic or native species in forestry increases productivity and carbon-fixation efficiency. In some regions, this introduction has also increased interspecies diversity at landscape and regional scales. In France, compared with 70 natural forest tree species, 30 introduced species that are commonly used in forest plantations have helped increase the interspecies genetic diversity of forests at the local level. In Europe, at least, there is no doubt that the introduction of new tree species has increased the species richness of forests. Nevertheless, exotic species, even those long naturalized species such as Douglas-fir (*Pseudotsuga menziesii*) are unacceptable in nature conservation schemes.

Exotic species can have negative impacts on native species and communities. For example, fast-growing species can replace native forest species because of their natural invasive potential, as have been observed with *Eucalyptus* in northwestern Spain and Portugal. As the introduction of exotic species has potential risks, confirmation of long-term adaptation to local environmental conditions and pest resistance is necessarily the first step for the use of exotic species in extensive plantation programs.

Plantations tend to be even-aged and managed on relatively short rotations; thus, simple stand structures are common. When repeated across a landscape, large areas of similar species and low structural complexity result in a loss of habitat for taxa that require the kind of conditions provided by naturally regenerated stands or old forests. It has been reported that the bird fauna of single-species plantation forests is less diverse than that of natural and seminatural forests. In other cases, however, bird species diversity in plantation forests is comparable with that in naturally generated stands. For example, cottonwood (*Populus deltoides*) plantations in the Mississippi River Valley in the southern United States are intensively managed (rotation lengths of 10–15 years), reaching crown closure in 2 years. In comparison to natural stands, bird species diversity and abundances are similar for all guilds except cavity nesters.

Where avian diversity is decreased in managed forests generally, loss of structure following harvest is usually the cause. In plantations, simplified structure may be exacerbated further by use of exotic species or by monoculture. Because plantations are harvested at or near economic optima, rather than at biological maturity, plantations seldom develop much beyond the stem

exclusion stage of stand development and do not re-establish characteristics of old forests or complex stand structures such as snags and coarse woody debris. Strategies to compensate for the simplifying tendencies of plantations and integrate biodiversity considerations include complex plantations composed of multiple species, varying planting spacing, thinning to variable densities, and retaining uncut patches and snags after harvest. Such biological legacies should benefit invertebrates such as saproxylic beetles as well as fungi, small mammals, and birds.

Silvicultural and site management practices of site preparation, competing vegetation control, and fertilization may reduce understory and groundcover vegetation diversity, although the effects of previous land use such as agriculture may play a larger role. For example, in southern United States industrial pine plantations, understory diversity was correlated with previous land use; lower diversity of native forest species occurred in plantations established on former farmland and higher diversity in plantations on cutover forest land.

Some species can benefit from forest plantations. For example, clear-cutting and short rotations favor the occurrence of ruderal plant species over some long-lived climax species. Forest plantations accommodate edge-specialist bird species and generalist forest species such as deer. Some rare and threatened species have been found to occupy forest plantations, especially when they lost most of their habitat to agricultural and urbanized land uses. For example in the UK, the native red squirrel is out-competed in native woodlands by the gray squirrel introduced from North America but the red squirrel thrives in conifer plantations, which are poor habitat for the gray squirrel.

Spatial considerations play a role in maintaining biodiversity at the landscape scale. Landscape diversity can meet the habitat needs of wildlife and be achieved by varying the size and shape of plantations and incorporating adjacency constraints into harvest scheduling models (i.e., a plantation adjacent to a recently harvested or young stand cannot be harvested until the adjacent stand reaches a certain age or crown height). Retaining areas of naturally regenerated forest, riparian buffers, or open habitat creates a landscape mosaic that combined with prescribed burning in fire-affected ecosystems, adds to landscape diversity. Landscape connectivity that provides dispersal corridors for mobile species is fostered by careful placement of forest roads and firebreaks.

Concerns about plantations and water are as varied as the issues surrounding biodiversity but generally relate to water use, water quality, or alteration of natural drainage. Species of *Eucalyptus* planted outside their native Australia have attracted the most negative attention for their putative excessive water use, especially in Africa and India but *Populus* species have similarly been accused in China of lowering local water tables and adding to drought. Species such as *Eucalyptus camaldulensis*, *E. tereticornis*, and *E. robusta* (and hybrids of these and other eucalypts) are drought tolerant and able to transpire even under considerable moisture stress. On balance they probably do not use more water than adjacent natural forests but certainly use more of the available water than grasslands or agricultural crops. There is little evidence that they can abstract groundwater; however, there is no recharge below the root zone. In the Wheatbelt of Western Australia, removal of the deep-rooted native vegetation including eucalypts and conversion to cereal crops has caused water tables to rise with subsequent salinization of soils and surface water bodies. Plantations of oil mallee crops (*E. polybractea*, *E. kochii* subsp. *plenissima*, and *E. horistes*) are planted to restore natural hydrology and counteract salinization.

Negative effects of plantations on water quality and aquatic resources are more due to intensive management than to use of exotic species. Intensive mechanical site preparation, especially on sloping sites, can result in sediment movement into streams. Chemical herbicides are used to control competing vegetation at various stages in the plantation growth cycle, but usually for site preparation in place of mechanical treatments or early in the life of the stand to release crop species from competitors. Less intense site preparation, formulations of herbicides that are not toxic to insects or other aquatic organisms and break down in soil, careful placement of chemicals to avoid direct application to water bodies, and designation of riparian buffers all have contributed to protection of water quality.

Harvesting practices, especially placement and construction of harvest roads and layout of skidding trails, potentially can degrade water quality. In developed nations, forest practices such as site preparation, harvesting, use of herbicides, and even choice of species may be regulated to some extent. In the United States, best management practices (BMPs) to address nonpoint source pollution and protect water quality have been codified by state agencies and landowners follow them voluntarily. Research shows generally high rates of compliance. Certification schemes substitute the coercive power of the marketplace for that of government; the various certification bodies differ in how they regard plantations, especially with regard to the use of herbicides, exotic species, or genetically modified trees.

Use of inorganic fertilizers to overcome fertility deficiencies, promote rapid growth, and sustain biomass accumulation generally has been found to have little impact on aquatic systems unless fertilizers are applied directly to streams, lakes, rivers, or adjacent riparian zones. Greater attention has focused on nutrient removals in harvests and the potential for intensive management to reduce site fertility and cause a fall-off in productivity of subsequent rotations. Claims of later-rotation productivity declines have been hard to substantiate, however, as general improvements in seed and seedling quality, genetic makeup, site preparation and competition control, and more careful harvesting that conserves site fertility have raised, rather than lowered yields. Nevertheless, there exist documented cases of lowered fertility caused by export of nutrients in the harvested wood. These localized cases have been caused by low initial fertility, often of phosphorus, potassium, or micronutrient deficiencies inherent in the soil parent material that are easily overcome by application of inorganic fertilizers.

In the most intensive management of pine plantations for pulpwood in the southern United States, some companies routinely apply complete nutrient mixes containing all macro- and micronutrients as a precaution, despite lack of demonstrated deficiency of most nutrients except phosphorus and a responsiveness to added nitrogen. A stand may be fertilized with nitrogen up to five times in a 25-year rotation, sometimes in combination with phosphorus. These stands occur mostly on relatively infertile Ultisols

and Spodosols developed on old marine sediments. On better soils (Alfisols, Entisols, and Vertisols), cottonwood plantations managed on 10-year rotations receive only an initial application of nitrogen at planting to promote rapid height growth to better compete with herbaceous competitors. Management of site nutrients in intensive plantations is critical to high yields as well as to protect long-term productivity and may require attention to retaining soil organic matter, especially on sandy soils. Factors to consider include inherent soil fertility (nutrient stocks as well as transformations and fluxes), plant demand and utilization efficiency, and nutrients export in products removed as well as leakages.

It is common wisdom that monoculture plantations are more susceptible than natural forests to insect and disease attacks, yet there is little evidence this is generally true. On the one hand, single-species stands occur naturally and some of these natural vegetation types are the product of periodic, catastrophic disturbances such as pine bark beetles or spruce budworm. On the other hand, one explanation for the often greater productivity of exotic tree species than attained in their native habitat is the lack of yield-reducing insects and diseases. But diversity in the abstract is not a guarantor of lessened risk; diverse, multiple-species stands themselves are not immune to devastating attack by introduced pests, a situation likely to increase in frequency as a result of globalization of trade in timber products.

Often the practices associated with intensive management are the causes of insect and disease problems. For example, the desire to maximize wood production may set the level of tolerable damage from native pests lower than the stable equilibrium levels for the pest; attempts to control the pest at lower levels may cause unstable population growth cycles. The potential risks of plantations stem from their uniformity: the same or a few species, planted closely together, on the same site, over large areas. Pests and pathogens adapted to the dominant species may build up quickly due to food supply and abundant sites for breeding or infection. Proximity of the branches and stems in closely spaced stands may favor buildup of species with low dispersal rates or small effective spread distances. Conversely, the same uniformity of plantations that contributed to the risks of insects and diseases also confers some advantages. Species can be chosen that have resistance to diseases, for example, the greater resistance of loblolly pine (*Pinus taeda*) compared to slash pine (*P. elliottii*) to *Cronartium* rust was one reason loblolly was favored by forest industry in the US South. The shorter rotation length of plantations relative to naturally regenerated stands means trees are fallen before they become overmature and become infected. The compact shape and uniform conditions in plantations facilitate detection and treatment of economically important pests and pathogens.

Plantations may negatively impact adjacent communities – because of invasive natural regeneration of planted trees in adjacent habitat or alteration of local and regional hydrologic cycles and poor management practices may damage aquatic systems. Plantations are certainly simpler and more uniform than naturally regenerated stands or native grasslands, and may support a less diverse flora and fauna. Nevertheless, plantations can contribute to biodiversity conservation at the landscape level by adding structural complexity to otherwise simple grasslands or agricultural landscapes and by fostering the dispersal of forest-dwelling species across these areas.

Further, comparisons of plantations to unmanaged native forests or even naturally regenerated secondary forests are not necessarily the most appropriate comparisons to make. Although the conversion of old-growth forests, native grasslands, or some other natural ecosystem to forest plantations rarely will be desirable from a biodiversity point of view, in that forest plantations often replace other land uses including degraded lands and abandoned agricultural areas. Objective assessments of the potential or actual impacts of forest plantations on biological diversity at different temporal and spatial scales require appropriate reference points. Forest plantations can have either positive or negative impacts on biodiversity at the tree, stand, or landscape level depending on the ecological context in which they found. Impacts on water quantity and quality can be minimized if sustainable practices are followed; similarly with soil resources and long-term site productivity. Both complex plantations for wood production and environmental plantations can beneficially impact local and regional environments.

Lastly, managing forest plantations to produce goods such as timber while at the same time enhancing ecological services such as biodiversity involves tradeoffs; this can be made only with a clear understanding of the ecological context of plantations in the broader landscape. Tradeoffs also require agreement among stakeholders on the desired balance of goods and ecological services from plantations. Thus, there is no single or simple answer to the question of whether forest plantations are `good' or `bad' for the environment.

See also: Ecological Data Analysis and Modelling: Modeling Dispersal Processes for Ecological Systems. Ecological Processes: Nitrification. Ecosystems: Riparian Wetlands. Evolutionary Ecology: Colonization; Ecological Niche. General Ecology: Biomass

Further Reading

Binkley, C.S., 2003. Forestry in the long sweep of history. In: Teeter, L.D., Cashore, B.W., Zhang, D. (Eds.), Forest Policy for Private Forestry: Global and Regional Challenges. Wallingford: CABI Publishing, pp. 1–8.

Brown C (2000) The global outlook for future wood supply from forest plantations. *FAO Working Paper GFPOS/WP/03*. Rome, Italy.

Carnus, J.-M., Parrotta, J., Brockerhoff, E., *et al*., 2006. Planted Forests and Biodiversity. Journal of Forestry 104 (2), 65–77.

Clawson, M., 1979. Forests in the long sweep of history. Science 204, 1168–1174.

Evans, J., Turnbull, J.W., 2004. Plantation Forestry in the Tropics: The Role, Silviculture and Use of Planted Forests for Industrial, Social, Environmental and Agroforestry Purposes, 3rd edn. Oxford: Oxford University Press.

FAO (2001) Global forest resources assessment 2000. *FAO Forestry Paper 140*. Rome, Italy.

FAO (2005) Global forest resources assessment 2005. *FAO Forestry Paper 147*. Rome, Italy.
Harris, T.G., Baldwin, S., Hopkins, A.J., 2004. The south's position in a global forest economy. Forest Landowner 63 (4), 9–11.
Hartsell, A.J., Brown, M.J., 2002. Forest statistics for Alabama, 2000. In: Resource Bulletin SRS-67. Ashville, NC: USDA Forest Service Southern Research Station, p. 76.
Li, Y., Zhang, D., 2007. Tree planting in the US South: A panel data analysis. Southern Journal of Applied Forestry 31 (4), 192–198.
Royer, J.P., Moulton, R.J., 1987. Reforestation incentives: Tax incentives and cost sharing in the South. Journal of Forestry 85 (8), 45–47.
Stanturf, J.A., 2005. What is forest restoration? In: Stanturf, J.A., Madsen, P. (Eds.), Restoration of Boreal and Temperate Forests. Boca Raton, FL: CRC Press, pp. 3–11.
Stanturf, J.A., Kellison, R.C., Broerman, F.S., Jones, S.B., 2003. Pine productivity: Where are we and how did we get here? Journal of Forestry 101 (3), 26–31.
Zhang, D., 2001. Why so much forestland in China would not grow trees? Management World 3, 120–125. (in Chinese).
Zhang, D., Flick, W., 2001. Sticks, carrots, and reforestation investment. Land Economics 77 (3), 443–456.
Zhang, D., Oweridu, E., 2007. Land tenure, market and the establishment of forest plantations in Ghana. Forest Policy and Economics 9, 602–610.
Zhang, D., Pearse, P.H., 1997. The Influence of the form of tenure on reforestation in British Columbia. Forest Ecology and Management 98, 239–250.

Freshwater Lakes

Sven E Jørgensen, Copenhagen University, Copenhagen, Denmark

Introduction

Freshwater lakes and reservoirs are basins filled with freshwater. Only 2.53% of the global water is freshwater; 1.76% of the global water is stored in ice caps, glaciers, and permafrost, while all fresh groundwater makes up 0.76% of the global water. It leaves 0.01% only for the surface freshwater, of which 70% or 0.007% of the global water is stored in the freshwater lakes. As surface water is easily accessible water, the storage of water in lakes and reservoirs becomes very important for the water supply and represents a large proportion of the world's readily accessible water (see Figs. 1 and **2**). Lake water is not only used for human consumption. Other water uses include industrial applications and processes and transportation and generation of hydropower.

The World's Freshwater Lakes

Table 1 gives an overview of 12 important freshwater lakes, including the deepest lake, the lake with the largest surface area, and the lake with the biggest volume. The lakes are not equally distributed in the world. About 10% of the total land is occupied by lakes in Scandinavia, while lakes occupy less than 1% of the land area in Argentina and China.

Importance of Lakes

The lake and reservoir water uses are becoming more intensive and multipurpose, particularly for lakes in heavily populated areas and intensively utilized regions. We can distinguish nine functions of lakes and reservoirs:

1. drinking water supply,
2. irrigation,
3. flood control,
4. aquatic production and fishery,
5. fire and ice ponds,
6. transportation,
7. hydropower,
8. conservation of biodiversity, and
9. recreation.

The multipurpose and extensive use of lakes and reservoirs can often lead to abuse and conflicts. There are numerous examples of such conflicts which are often rooted in inappropriate and insufficient water management.

Water Quality Problems of Lakes and Reservoirs

Nine problems associated with the extensive use of lakes and reservoirs can be identified.

Eutrophication

This is the most pervasive water quality problem on a global scale, being a primary cause of lake deterioration. Eutrophication (nutrient enrichment) represents the natural aging process of many lakes in which they gradually become filled with sediments and organic materials over a typically geologic timescale. Human activities in a drainage basin can, however, dramatically accelerate this process. Its primary cause is the excessive inflow of nutrients (mainly phosphorus, sometimes nitrogen, sometimes both) to a water body from municipal wastewater treatment plants and industries, as well as drainage or runoff from urban areas and agricultural fields. Most lakes in densely inhabited regions of the world suffer from eutrophication, both in industrialized and developing countries. The impacts of the eutrophication process include heavy blooms of phytoplankton in a water body. These blooms will inevitably result in (1) reduced water transparency; (2) decreased oxygen concentration in the water column, particularly in the bottom layer (hypolimnion), which can cause fish kills and the remobilization or resuspension of heavy metals and nutrients into the water column; and (3) significant declines in the biodiversity of the lakes, including the disappearance of

Fig. 1 Lake Baikal, the deepest lake in the world. The volume of Lake Baikal corresponds to almost 20% of all global surface freshwater.

Fig. 2 Crater Lake, Oregon State, the lake famous throughout the world for its clarity. The Secchi disk transparency is 42 m.

sensitive aquatic species. In shallow lakes, eutrophication can also cause an enormous increase in the growth of submerged and emergent rooted aquatic plants, as well as floating plants. This can lead to dramatic changes in the ecosystem structure.

If the sources of nutrients are removed or reduced significantly, the eutrophication problems can be fully controlled (see **Figs. 3** and **4**). Lake Constance, also known as Bodensee, gives very illustrative examples. After the Second World War, the phosphorus concentration in the lake was about 0.01 mg l^{-1} and the lake was oligotrophic. In the year 1980, the lake was mesotrophic to eutrophic and the phosphorus concentration was about 0.08 mg l^{-1}. Due to a massive reduction in the discharge of phosphorus from all sources, wastewater, agricultural drainage water, and septic tanks, it has been possible to reduce the phosphorus concentration to about 0.013 mg l^{-1} today. Lake Biwa, Japan, is illustrative of a partial solution of the problem (**Fig. 5**). The discharge of phosphorus from wastewater has been significantly reduced since the 1970s, but due to almost no reduction in the phosphorus coming from agricultural drainage water, it has only been possible to stabilize the eutrophication level at a phosphorus concentration about 0.035 mg l^{-1}. If on the other hand, the phosphorus in wastewater would not have been reduced, the eutrophication level would have increased.

Table 1 Major freshwater lakes

Lake	Volume (km^3)	Area (km^2)	Max. depth (m)
Lake Baikal	22 995	31 500	1 741
Lake Tanganyika	18 140	32 000	1 471
Lake Superior	12 100	82 100	170
Lake Malawi	6140	22 490	706
Lake Michigan	4920	57 750	110
Lake Huron	3540	59 500	92
Lake Victoria	2700	62 940	80
Lake Titicaca	903	8 559	283
Lake Erie	484	25 700	64
Lake Constance	48.5	571	254
Lake Biwa	27.5	674	104
Lake Maggiore	37.5	213	370

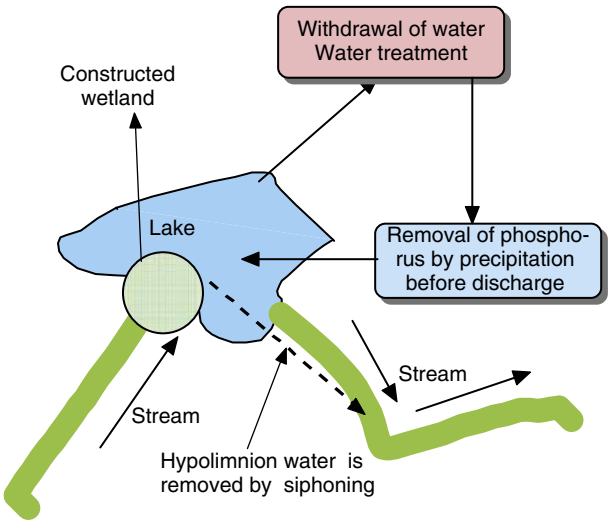


Fig. 3 Abatement of eutrophication requires often the use of several methods at the same time, as shown here: removal of phosphorus from wastewater, construction of wetland to remove phosphorus from the inflowing tributary, and removal of hypolimnic (bottom) water by siphoning.

Fig. 4 Lake Bled, where restoration by siphoning hypolimnic water has been applied.

Acidification

This process of lake deterioration is caused mainly by acid precipitation and deposition. The nitrogen and sulfur compounds that cause this problem are emitted by industrial activities and by the consumption of fossil fuels, and fall to the land surface. The water in a lake can become acidic over time if its drainage basin does not contain the appropriate soil and geologic characteristics to neutralize the acidic water prior to its inflow into the lake. The primary consequence of acidification of lake water is the significant

Fig. 5 Lake Biwa in Japan is a very important recreational area for the population. A museum has been erected to present for the population all aspects of the lake: the culture, the limnology, the geology, and the history.

reduction of species diversity, the extinction of fish populations, and the disruption of lake ecosystem equilibrium. Other causes of lake acidification also exist, including water discharges from mining activities and the direct discharge of industrial waste effluents containing acidic components. Natural sources of acidifying substances include volcanic activities and natural emissions of gases. This problem, because of the geological characteristics, has been a major problem in Scandinavia (except the most southern Scandinavia) and the northeastern United States.

Toxic Contamination

This problem can have direct and dramatic impacts on both human and ecosystem health. Toxic substances originate not only from industrial activities and mining, but also as a result of intensive agriculture practices. Identification of the number of lakes and reservoirs exhibiting toxic contamination will doubtlessly increase in future years as we obtain more information on their concentrations in the environment, particularly in developing countries. Major impacts of toxic substances include the disappearance of sensitive species, as well as their accumulation in lake sediments and biota. The latter can directly and indirectly impact human health. Because the number of risk assessments applied to already-existing chemicals is currently extremely low ($\sim$500), a complete solution of this problem will take many years.

Water-Level Changes

Significant changes in water levels, particularly dropping levels, can be caused by

1. excessive withdrawal of water from lakes and/or their inflowing or out flowing rivers, and
2. the diversion of the inflowing water.

The consequences of water-level changes include: decrease in lake volume and/or surface area; unstable shoreline area communities; changes in lake ecosystem structure; reduced fish spawning areas; increased water retention time (decreased flushing rate), which can accelerate other negative lake processes (e.g., eutrophication, retention of toxic substances); and increased salt concentration, leading to reduced water quality for human uses.

Lake Aral is probably the most illustrative example of this problem. Due to uncontrolled use of the inflowing river water for irrigation, the water level in the lake was reduced by almost 20 m. The lake was divided in two lakes, Large and Small Aral, with together less than half of the original lake area and with a salinity 10 times what it was 40 years ago.

Salinization

This process is an increase in the concentration of salts (all ions, not just sodium and chloride) in lake water, caused by such factors as (1) decreased in-lake water levels; (2) overuse of water in the drainage basin (e.g., cooling water, irrigation); and (3) global climate change. The effects of water salinization include (1) dramatic changes in lake biological structure; (2) lower fish production; and (3) reduced biodiversity. Human utilization of lake water with a high salt concentration also can become very problematic. This problem, however, can at least be partly addressed with the implementation of appropriate environmental management and agricultural practices in a lake's drainage basin.

Siltation

Accelerated soil erosion, resulting from such activities as the overuse or misuse of arable land, mining and/or deforestation in a drainage basin, can lead to the excessive loading of suspended solids (sediment) to lakes. The consequences of these increased loads include the rapid accumulation of sediment within the lake basin, and the increased turbidity (decreased transparency) of the water in the lake. The immediate impacts can be a significant reduction in the number of living organisms in a lake, decreased biodiversity, and reduced fisheries.

Introduction of Exotic Species

The intentional introduction of exotic (nonresident) species has become an almost common practice in some fisheries to increase the production of commercially important species. The introduction of Nile perch into Lake Victoria is a primary example. However, the intentional or unintentional (or sometimes illegal) introduction of exotic species can cause very serious problems in a given lake. The accidental introduction of zebra mussels in Lake Erie and water hyacinths in several lakes of China provides a dramatic example of this phenomenon. The introduction of exotic species can provoke very dramatic changes in the ecosystem structure not only at the biological community level, but also in a lake's chemical–physical environment. The major negative consequences of exotic species include the (1) disappearance of native species; (2) alteration of trophic equilibrium; (3) significant reduction in species diversity; and (4) reduction of water transparency and changes in algae bloom patterns, via chemical–physical feedback processes in a lake.

Overfishing

Unsustainable fishing practices, sometimes combined with other problems, can lead to the collapse of fisheries. It seems to be an increasing problem for many African lakes, particularly Lake Victoria.

Pathogenic Contamination

This problem is caused by discharge of untreated sewage or runoff from livestock farms, a problem in both developing and developed countries. A *Cryptosporidium* outbreak in Lake Michigan (Milwaukee) sickened 400 000 people in the early 1990s. It is likely that many deaths in developing countries are due to consumption of dirty lake water.

See also: Ecological Complexity: Ecological Data Archiving and Sharing. Ecological Processes: Decomposition and Mineralization; Volatilization. General Ecology: Temperature Regulation

Further Reading

ILEC (2005) Managing Lakes and Their Basins for Sustainable Use: A Report for Lake Basin Managers and Stakeholders, 146pp. Kusatsu, Japan: International Lake Environment Committee Foundation. http://www.ilec.or.jp/eg/lbmi/reports/LBMI_Main_Report_22February2006.pdf (accessed October 2007).

ILEC and UNEP (2003) World Lake Vision: A Call to Action, 37pp. Kusatsu, Japan: World Lake Vision Committee. http://www.ilec.or.jp/eg/wlv/complete/wlv_c_english.PDF (accessed October 2007).

Jørgensen, S.E., de Bernard, R., Ballatore, T.J., Muhandiki, V.S., 2003. Lake Watch 2003. The Changing State of the World's Lakes. Kusatsu, Japan: ILEC, 73pp.

Jørgensen, S.E., Löffler, H., Rast and Straškraba, M., (2005). Lake and Reservoir Mangement, 502pp. Amsterdam: Elsevier.

O'Sullivan, P.E., Reynolds, C.S., (2004, 2005). The Lakes Handbook, vols. 1 and 2, 700pp. and 560pp. Blackwell Publishing.

Freshwater Marshes

P Keddy, Southeastern Louisiana University, Hammond, LA, USA

Wetlands are produced by flooding, and as a consequence, have distinctive soils, microorganisms, plants, and animals. The soils are usually anoxic or hypoxic, as water contains less oxygen than air, and any oxygen that is dissolved in the water is rapidly consumed by soil microorganisms. Vast numbers of microorganisms, particularly bacteria, thrive under the wet and hypoxic conditions found in marsh soils. These microbes transform elements including nitrogen, phosphorus, and sulfur among different chemical states. Therefore, wetlands are closely connected to major biogeochemical cycles. The plants in wetlands often have hollow stems to permit movement of atmospheric oxygen downward into their rhizomes and roots. Many species of animals are adapted to living in shallow water, and in habitats that frequently flood. Some of these are small invertebrates (e.g., plankton, shrimp, and clams), while others are larger and more conspicuous (fish, salamanders, frogs, turtles, snakes, alligators, birds, and mammals).

Six Types of Wetlands

There are six major types of wetlands: swamp, marsh, fen, bog, wet meadow, and shallow water (aquatic). These six types are produced by different combinations of flooding, soil nutrients, and climate. A seventh group, saline wetlands, which includes salt marshes and mangroves, is often treated as a distinct wetland type. Saline wetlands occur mostly along coastlines, but also occasionally in noncoastal areas where evaporation exceeds rainfall, such as in arid western North America, northern Africa, or central Eurasia.

Swamps and marshes have mineral soils with sand, silt, or clay. Swamps are dominated by trees or shrubs, whereas marshes are dominated by herbaceous plants such as cattails and reeds (Fig. 1). Such wetlands tend to occur along the margins of rivers (Fig. 2) or lakes, and often receive fresh layers of sediment during annual spring flooding. Marshes are among the world's most biologically productive ecosystems. As a consequence, they are very important for producing wildlife, and for producing human food in the form of shrimp, fish, and waterfowl.

Fens and bogs have organic soils (peat), formed from the accumulation of partially decayed plants. Most peatlands occur at high latitudes in landscapes that were glaciated during the last ice ages. In fens, the layer of peat is relatively thin, allowing the longer roots of the plants to reach the mineral soil beneath. In bogs, plants are entirely rooted in the peat. As peat becomes deeper (the natural trend as fens become bogs), plants become increasingly dependent upon nutrients dissolved in rainwater, eventually producing an `ombrotrophic' bog. The large amounts of organic carbon stored in peatlands help reduce global warming.

Wet meadows occur where land is flooded in some seasons and moist in others, such as along the shores of rivers or lakes. Wet meadows often have high plant diversity, including carnivorous plants and orchids. Examples of wet meadows include wet prairies, slacks between sand dunes, and wet pine savannas. Pine savannas may have up to 40 species of plants in a single square meter, and hundreds of species in a hundred hectares.

Aquatic wetlands are covered in water, usually with plants rooted in the sediment but possessing leaves that extend into the atmosphere. Grasses, sedges, and reeds emerge from shallow water, whereas water lilies and pondweeds with floating leaves occur

Fig. 1 Marshes occur in flooded areas, such as this depression flooded by beavers in Ontario, Canada. As the photo illustrates, marshes form at the interface of land and water. Courtesy of Paul Keddy.

 doi:10.1016/B978-0-444-63768-0.00338-3

in deeper water. Aquatic wetlands provide important habitat for breeding fish and migratory waterfowl. Animals can create aquatic wetlands: beavers build dams to flood stream valleys, and alligators dig small ponds in marshes or wet meadows.

The Distribution of Marshes

Wetlands can occur wherever water affects the soil. Not only are there therefore many kinds of wetlands, but their size and shape is very variable. Wetlands can include small pools in deserts and seepage areas on mountainsides, or they can be long but narrow strips on shorelines of large lakes (Fig. 3), or they may cover vast river floodplains (Fig. 4) and expanses of northern plains. The two largest wetlands in the world (both $>750\,000\ km^2$) are the West Siberian lowland and the Amazon River basin. The West Siberian Lowland consists largely of fens and bogs, but marshes occur along rivers, particularly in the more southern regions (Fig. 5). The Amazon is a tropical lowland with freshwater swamps and marshes containing more kinds of trees and fish than any other region of the world.

Water as the Critical Factor

Water is a critical factor in all marshes. The duration of flooding is the most important factor determining the kind of wetland that occurs. Water can arrive as short pulses of flooding by rivers, as rainfall, or as slow and steady seepage. Each mode of arrival produces different kinds of wetlands. In order to better understand marshes, let us consider four examples of wetlands with very different flooding regimes.

Floodplains. Wetlands along rivers are often flooded by annual pulses of water. These pulses may deposit thick layers of sediment or dissolved nutrients that stimulate plant growth. In floodplains, animal life cycles are often precisely determined by the timing of the flood. Fish may depend upon feeding and breeding in the shallow warm pools left by retreating floodwaters. Birds

Fig. 2 Extensive bulrush (*Schoenoplectus* spp.) marshes along the Ottawa River in central Canada. The stalks of purple flowers indicate the invasion of this marsh by purple loosestrife (*Lythrum salicaria*), a native of Eurasia. Courtesy of Paul Keddy.

Fig. 3 Sedges, grasses, and forbs compose this marsh on the leeward side of a narrow peninsula projecting into one of the Great Lakes (Lake Michigan), Michigan, USA. Courtesy of Cathy Keddy.

Fig. 4 Extensive marshes of bulltongue (*Sagittaria lancifolia*) and American bulrush (*Schoenoplectus americanus*) now occur in coastal Louisiana, USA, where logging destroyed baldcypress forest. Courtesy of Paul Keddy.

Fig. 5 The largest wetland in the world occurs in the Western Siberian Lowland. Although much of this is peatland, marshes occur along the watercourses, particularly in the southern areas. Courtesy of M. Teliatnikov.

may time their nesting to be able to feed their young on the fish and amphibians left behind by receding water. Marshes are often intermixed with swamps, depending upon the duration of flooding (**Fig. 6**). Early human civilizations developed in this type of habitat, along the Nile, Indus, Euphrates and Hwang Ho, where the annual flooding provided fertilized soil and free irrigation.

Peat bogs. Some peat bogs receive water only as rainfall. As a consequence, the water moves slowly, if at all, and contains very few nutrients. Hence, these types of wetlands often are dominated by slow growing mosses and evergreen plants. Most such wetlands occur in the far north in glaciated landscapes. Humans have developed a number of uses for the peat – in Ireland, the peat is cut into blocks and used for fuel. In Canada, the peat is harvested and bagged for sale to gardeners. In Russia, peat is used to fuel electrical plants. Marshes may form on the edges of bogs where nutrients accumulate from runoff, or along river courses where nutrients are more available.

Seepage wetlands. In gently sloping landscapes water can seep slowly through the soil. In northern glaciated landscapes, such seepage can produce fens, which have distinctive species of mosses and plants, and may develop in distinctive parallel ridges. In more southern landscapes, seepage can produce pitcher plant savannas or wet prairies. Often these seepage areas are rather small (only a few hectares in extent) but are locally important because of the rare plants and animals they support. Seepage areas can be larger, and when the water flow is sufficiently abundant, shallow water can move across a landscape in a phenomenon known as sheet flow. The vast Everglades, with its distinctive animals, is a product of sheet flow of water from Lake Okeechobee in south central Florida southward to the ocean.

Temporary wetlands. In many parts of the world, small temporary (or ephemeral) pools form after heavy rain or when snow melts. These pools can go by a variety of local names including vernal pools, woodland ponds, playas or potholes. The aquatic life in these pools is forced to adopt a life cycle that is closely tied to the water levels. Many species of frogs and salamanders breed in such pools, and the young must metamorphose before the pond dries up. Wetland plants may produce large numbers of seeds that remain dormant until rain refills the pond.

Since water has such a critical effect on wetlands, where water levels change, plant and animal communities will change as well. A typical shoreline marsh will often show distinct bands of vegetation (`zonation'), with each kind of plant occupying a narrow

Fig. 6 Southern marshes on the coastal plain of North America may be dominated by a single grass, maidencane (*Panicum hemitomon*). This marsh occupies an opening within a baldcypress swamp, Louisiana, USA. Courtesy of Cathy Keddy.

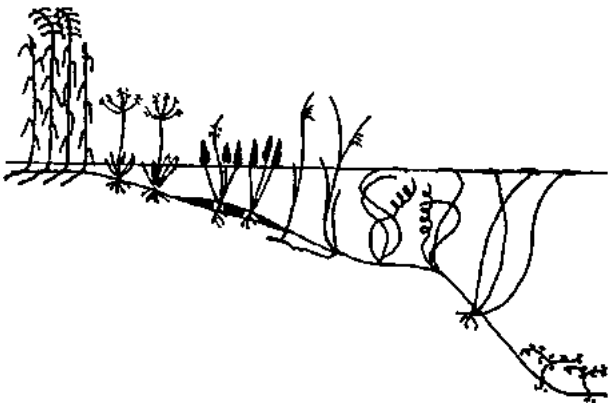


Fig. 7 Different marsh plants tolerate different water levels. Hence, as the water level changes from shallow water (left; seasonally flooded) to deeper water (right; permanently flooded), the plants appear to occur in different zones. Courtesy of Rochelle Lawson.

range of water depths (Fig. 7). Most kinds of animals, including frogs and birds, also have their own set of preferred water depths. Wading birds (egrets, ibis, herons) may feed in different depths of water depending upon the length of their legs. Ducks, geese, and swans can feed at different water depths depending upon the length of their necks. Some water birds (Northern Shoveler, flamingos) strain microorganism from shallow water, while others (cormorants, loons) dive to feed further below the surface. Some ducks prefer wetlands that are densely vegetated, while others prefer more open water. Hence, even small changes in the duration of flooding or depth of water can produce very different plant and animal communities.

Many marsh plants adapt to flooding by producing hollow shoots, which allow oxygen to be transmitted to the rooting zone. The tissue that allows the flow of oxygen is known as aerenchyma. Not only can oxygen move by diffusion, but there are a number of methods in which oxygen moves more rapidly through large clones of plants, entering at one shoot and leaving at another. Consequently, plants can play an important role in oxidizing the soil around their rhizomes, allowing distinctive microbial communities to form. Some marsh plants also have floating leaves (e.g., water lilies) or even float entirely on the surface (e.g., duckweeds). The largest floating leaves in the world (Fig. 8) are those of the Amazon water lily (*Victoria amazonica*). The gargantuan leaves can be 2 m in diameter with an elevated lip around the circumference. There are two gaps in the lip to allow water to drain, and large spines to protect the underwater sections of the foliage.

Other Environmental Factors Affecting Marshes

Nutrients

The main nutrients that affect the growth of marsh plants, and plants in general, are nitrogen and phosphorus. As described above, flood pulses that carry sediment down river courses can produce particularly fertile and productive marshes. Floodplains can therefore be thought of as one natural extreme along a gradient of nutrient supply. At the other end of the gradient lie peat bogs, which depend partly or entirely upon rainfall, and which therefore receive few nutrients. Sphagnum moss is well adapted to

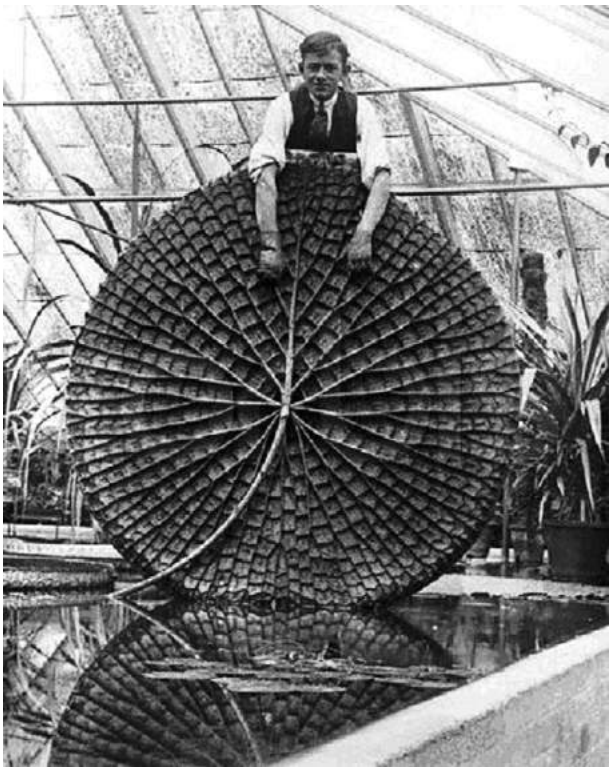


Fig. 8 The Amazon water lily has the largest floating leaves of any wetland plant. Note the prominent ribs on the underside of the leaf. Courtesy of Corbis.

peatlands, and often comprises a large portion of the peat. In between the natural extremes of river floodplains (high nutrients) and peat bogs (low nutrients), one can arrange most other types of wetlands. The type of plants, and their rate of growth, will depend where along this gradient they occur, but most marshes generally occur in more fertile conditions.

While nutrients enhance productivity, paradoxically they can often reduce the diversity of plants and animals. Often, the high productivity is channeled into a few dominant species. One finds large numbers of common species, while the rarer species disappear. Humans often increase nutrient levels in watersheds and wetlands, thereby changing the species present and reducing their diversity. Carnivorous plants are known for tolerating low nutrient levels, because they can obtain added nutrients from their prey. Common examples include pitcher plants (*Sarracenia* spp.), bladderworts (*Utricularia* spp.), and butterworts (*Pinguicula* spp.). Cattails (*Typha* spp.) and certain grasses (*Phalaris arundinacea*) are particularly well known for rapid growth and an ability to dominate marshes at higher nutrient levels.

Disturbance

A disturbance can be narrowly defined as any factor that removes biomass from a plant. In marshes, sources of disturbance may include waves in lakes, fire, grazing, or (in the north) scouring by winter ice. One of the principal effects of disturbances is the creation of gaps in the vegetation, allowing new kinds of plants to establish from buried seeds. Most marshes have large densities of buried seeds, often more than 1000 seeds m^{-2}. After disturbance, marsh plants can also re-emerge from buried rhizomes. Hence, cycles of disturbance play an important role in creating marshes.

Although the presence of fire in wetlands may seem paradoxical, fire can often occur during periods of drought. Northern peatlands, cattail marshes on lakeshores, wet prairies, and seepage areas in savannas can burn under the appropriate conditions. In northern peatlands, a fire can remove thousands of years of peat accumulation in a few days, even uncovering boulders and rock ridges that were buried beneath the peat. In marshes, fire can selectively remove shrubs and small trees, preventing the marsh from turning into a swamp. In the Everglades, burning can create depressions that then cause marshes to revert to aquatic conditions.

Animals that feed upon plants often cause only small and local effects. Think of a moose grazing on water lilies, a muskrat feeding on grasses, or a hippopotamus feeding on water hyacinth. Often the small patch of removed foliage is quickly replaced by new growth. But when herbivores become overly abundant, they can destroy the marsh vegetation entirely. In northern North America along Hudson Bay, Canada geese (*Branta canadensis*) are now so abundant that they remove all vegetation from expanses of coastal marsh. In southern North America, along the Gulf of Mexico, an introduced mammal, nutria (*Myocastor coypus*), similarly can strip marsh vegetation to coastal mudflats. To some extent, disturbance by herbivores is a natural phenomenon, one

Fig. 9 Marshes provide essential habitat for many kinds of wading birds including flamingos, Jurong Bird Park, Singapore. Courtesy of Corbis.

Fig. 10 Alligators are one of the many species that benefit from protected marshes such as the Everglades, Loxahatchee National Wildlife Refuge, Florida, USA. Courtesy of Paul Keddy.

that has occurred cyclically throughout history. However, in the above two examples, one suspects humans may be the ultimate cause of the large-scale overgrazing (see the next section).

Periodic droughts may at times function like a natural disturbance by killing adult plants, and allowing new species to re-establish from buried seeds. Vernal pools and prairie potholes both have plant and animal species that are adapted to this kind of cyclical disturbance.

Plant and Animal Diversity in Wetlands

Wetlands are important for protecting biological diversity. Their high productivity provides abundant food, and the water provides an important added resource. Hence, wetlands often have large populations of animals and wading birds. The Camargue in Southern Europe, for example, is considered to be the European equivalent of the Everglades. Both have species of wading birds such as storks and flamingos (Fig. 9). Large numbers of other kinds of species including fish, frogs, salamanders, turtles, alligators (Fig. 10), crocodiles, and mammals require flooded conditions for all or at least part of the year. If the wetlands are drained, all of the species dependent upon them will disappear.

All wetlands, however, do not support the same species. Often, as already noted in the section entitled `Disturbance', small differences in water level or nutrient supplies will produce distinctive types of wetlands. Hence, wetlands that are variable in water levels and fertility will frequently support more kinds of species than wetlands that are uniform. Along the Amazon River floodplain, for example, different kinds of swamp, marsh, grassland, and aquatic communities form in response to different flooding regimes, and each has its own complement of animal species. In the Great Lakes, different flood durations similarly produce different types of wetlands, from aquatic situations in deeper water, to marshes and wet meadows in shallower water. Some types of frogs, such as bullfrogs, require deeper water, while others, such as gray tree frogs, require shrubs.

Human Impacts

Humans have had, and continue to have, serious impacts upon wetlands in general, and marshes in particular. Some human impacts include draining, damming, eutrophication, and alteration of food webs. Let us consider these in turn.

One of the most obvious ways in which humans affect wetlands is by draining them. When the wetlands are drained, the soil becomes oxidized, and terrestrial plants and animals replace the wetland plants and animals. Often, drainage is followed by conversion to agriculture or human settlement, entirely removing the marshes that once existed. Vast areas of farmland in Europe, Asia, and North America were once marshes and have now been converted to crops for human consumption. Many countries now have laws to protect wetlands from further development, although the degree of protection provided, and the degree of enforcement, varies from one region of the world to another. Wetlands are also often included in protected areas such as national parks and ecological reserves.

Construction of dams can also have severe negative effects upon wetlands. The dams may be built for flood control, irrigation, or generating electricity. The wetland behind the dam may be destroyed by the prolonged flooding, whereas the wetlands downstream are disrupted by the lack of normal flood pulses. A single dam can therefore affect a vast area of wetlands. The degree of damage depends upon the pattern of water level fluctuations in the reservoir behind the dam, but in general large areas of marsh are lost both upstream and downstream from the dam. Sediment that would have expanded and fertilized wetlands during periodic floods becomes trapped behind the dam. Most of the world's large rivers have now been significantly affected by dams. To protect wetlands, it is necessary to identify rivers that are still relatively natural and to prevent further dams from being constructed. In other cases, it is possible to remove dams and allow natural processes to resume. An artificial levee can be considered a special type of dam that is built parallel to a river to prevent it from flooding into adjoining lands. Levees harm marshes by preventing the annual flooding, and by allowing cropland and cities to move into floodplains.

Humans can also affect wetlands by changing the nutrients in the water. Sewage from cities provides a specific `point source' of nutrients, particularly nitrogen and phosphorus, that enter water courses then spread into wetlands. Activities such as agriculture and forestry provide `diffuse sources' of nutrients, where runoff from large areas carries dissolved nutrients, and nutrients attached to clay particles, into the water and into adjoining marshes. The added nutrients can stimulate plant growth, which may seem to be beneficial – but it often leads to significant changes in the biota. Rarer plants and animals that are adapted to low fertility are replaced by more common plants and animals that exploit fertile conditions. Rapid growth of algae, followed by decay, can eliminate oxygen from lakes, causing fish kills. Protecting the quality of marshes therefore requires two sets of actions. First, it is necessary to control the obvious point sources of pollution by building sewage treatment plants. Second, it is necessary to use entire landscapes with care, with the broad objective of reducing nutrients in runoff. This can involve carefully timing the fertilization of crops, maintaining areas of natural vegetation along watercourses, fencing cattle away from stream valleys, minimizing construction of new logging roads, and avoiding construction on steep hill sides.

Herbivores are common in wetlands, and a natural part of energy flow from plants to carnivores. Common examples of large herbivores include moose, geese, muskrats, and hippopotamuses. Humans can disrupt wetlands by disrupting the natural balance between herbivores and plants. Herbivores can increase to destructive levels in several ways. When humans introduce new species of herbivores, rates of damage to plants may increase greatly – for example, nutria introduced from South America are causing significant damage to coastal wetlands in Louisiana. When humans reduce predation on herbivores, they may also increase to higher than natural levels. Killing alligators may damage wetlands by allowing herbivores such as nutria to reach high population densities; similarly, the loss of natural predators may be one of the reasons that Canada geese have multiplied to levels where they can destroy wetlands around Hudson Bay. There is also evidence that when humans harvest blue crabs, snails that the crabs normally eat begin to multiply and damage coastal marshes. These types of effects are difficult to study, since the effects may be indirect and take place over the long term.

Road networks are a final cause of damage to wetlands. The obvious effects of roads include the filling of wetlands, and the blocking of lateral flow of water into or out of wetlands. But there are many other effects. When amphibians migrate across roads to breeding sites, vast numbers can be killed by cars. In northern climates, the road salt put on roads as a de-icer can flow into adjoining wetlands. Snakes may be attracted to the warm asphalt and killed by passing cars. Invasive plant species can arrive along newly constructed ditches. Overall, roads change a landscape by accelerating logging, agriculture, hunting, and urban development. As a consequence, the quality of the marshes in a landscape is linked to two factors: the abundance of roads (a negative effect) and the abundance of forest (a positive effect). Although it may not be obvious, halting road construction (or removing unwanted roads) and protecting forests (or replanting new areas of forest) may have important consequences for all the marshes in a landscape.

Wetland Restoration

Humans have caused much damage to wetlands over the past thousand years, and the effects have increased as human populations and technological power have grown. We have seen some examples of damage in the preceding section. In response to such past abuses, humans have also begun consciously re-creating wetlands. There are a growing number of efforts to create new wetlands and enhance existing wetlands. Along both the Rhine River and the Mississippi River, some levees have been breached, allowing floodwater to return and marshes to recover. Depressions left by mining, or deliberately constructed for wetlands, can be flooded to recreate small marshes in highly developed landscapes. Construction of dams and roads has been more carefully regulated.

Table 1 The world's largest wetlands (areas rounded to the nearest 1000 km^2)

Rank	Continent	Wetland	Description	Area (km^2)	Source
1	Eurasia	West Siberian Lowland	Bogs, mires, fens	2 745 000	Solomeshch, chapter 2
2	South America	Amazon River basin	Savanna and forested floodplain	1 738 000	Junk and Piedade, chapter 3
3	North America	Hudson Bay Lowland	Bogs, fens, swamps, marshes	374 000	Abraham and Keddy, chapter 4
4	Africa	Congo River basin	Swamps, riverine forest, wet prairie	189 000	Campbell, chapter 5
5	North America	Mackenzie River basin	Bogs, fens, swamps, marshes	166 000	Vitt *et al.*, chapter 6
6	South America	Pantanal	Savannas, grasslands, riverine forest	138 000	Alho, chapter 7
7	North America	Mississippi River basin	Bottomland hardwood forest,swamps, marshes	108 000	Shaffer *et al.*, chapter 8
8	Africa	Lake Chad basin	Grass and shrub savanna, shrubsteppe, marshes	106 000	Lemoalle, chapter 9
9	Africa	River Nile basin	Swamps, marshes	92 000	Springuel and Ali, chapter 10
10	North America	Prairie potholes	Marshes, meadows	63 000	van der Valk, chapter 11
11	South America	Magellanic moorland	Peatlands	44 000	Arroyo *et al.*, chapter 12

Modified from Fraser LH and Keddy PA (eds.) (2005) *The World's Largest Wetlands: Ecology and Conservation*. Cambridge: Cambridge University Press.

The future of marshes will likely depend upon two human activities: our success at protecting existing marshes from damage and our success at restoring marshes that have already been damaged. The list of the world's largest wetlands in **Table 1** provides an important set of targets for global conservation.

Summary

Marshes are produced by flooding, and, as a consequence, have distinctive soils, microorganisms, plants, and animals. The soils are usually anoxic or hypoxic, allowing vast numbers of microorganisms, particularly bacteria, to transform elements including nitrogen, phosphorus, and sulfur among different chemical states. Marsh plants often have hollow stems to permit movement of atmospheric oxygen downward into their rhizomes and roots. Marshes are some of the most biologically productive habitats in the world, and therefore support large numbers of animals, from shrimps and fish through to birds and mammals. Marshes are one of six types of wetlands, the others being swamp, fen, bog, wet meadow, and shallow water. Humans can affect marshes by changing water levels with drainage ditches, canals, dams, or levees. Other human impacts can arise from pollution by added nutrients, overharvesting of selected species, or building road networks in landscapes.

See also: Aquatic Ecology: Eutrophication. Behavioral Ecology: Herbivore-Predator Cycles. Global Change Ecology: Energy Flows in the Biosphere. Human Ecology and Sustainability: Ecological Footprint

Further Reading

Fraser, L.H., Keddy, P.A. (Eds.), 2005. The World's Largest Wetlands: Ecology and Conservation. Cambridge: Cambridge University Press.
Keddy, P.A., 2000. Wetland Ecology. Cambridge: Cambridge University Press.
Middleton, B.A. (Ed.), 2002. Flood Pulsing in Wetlands: Restoring the Natural Hydrological Balance. New York: Wiley.
Mitsch, W.J., Gosselink, J.G., 2000. Wetlands, 3rd edn. New York: Wiley.
Patten, B.C. (Ed.), 1990. Wetlands and Shallow Continental Water Bodies, Vol. 1: Natural and Human Relationships. The Hague: SPB Academic Publishing.
Whigham, D.F., Dykyjova, D., Hejnyt, S. (Eds.), 1992. Wetlands of the World 1. Dordrecht: Kluwer Academic Publishers.

Greenhouses, Microcosms, and Mesocosms

WH Adey, Smithsonian Institution, Washington, DC, USA
PC Kangas, University of Maryland, College Park, MD, USA

Introduction

An ecosystem is an assemblage of organisms living together and interacting with each other and their environment. An element of biodiversity and biogeochemical time stability is implied, although dimensions are optional, ranging from the biosphere subset, the biome, to perhaps a field or pond. Ecosystems with their complex food webs and biotic physical/chemical relationships are self-organizing due to the genetic information existing in the genome of each species. Even when spatially well bounded, ecosystems are not closed. At the very least, they are subject to energy input and energy and materials exchange with adjacent ecosystems. Often, ecosystems demonstrate biotic exchange with adjacent ecosystems that can be complex and include reproductive and seasonal phases.

The development of an ecosystem in a greenhouse implies that the ecosystem is solar driven, and thus no deeper than the photic zone of the ocean, although the basic principles discussed could generally apply to deep ocean ecosystems. Greenhouse placement generally requires spatial limitation, and scaling of model to analog is necessary for many physical and biotic factors. In some cases, it is intuitive, and in others must be empirically demonstrated by trial and error in comparison with analog function. All of the following case studies demonstrate aspects of this necessary scaling exercise. Normal, biogeochemical, and biotic exchange with adjacent ecosystems must also be simulated.

The reasons for placing or developing ecosystems within greenhouses for research or educational purposes have varied enormously and have ranged from the strongly funded, multiscientist research endeavors, to the classroom aquarium or terrarium. Even the naming of the research field of endeavor has varied widely: to name a few, synthetic ecology, ecological engineering, controlled ecology, closed systems ecology, ecosystem modeling, etc. The systems themselves are called living systems models, microcosms, mesocosms, macrocosms, ecotaria, living machines, closed ecological life support systems (CELSS), etc.

In this article, we limit our discussions to those serious research efforts in which significant effort is expended to match biodiversity, food web and symbiotic relationships, as well as biogeochemical function to analog wild ecosystems. By definition, such systems cannot be closed; however, the known interchanges, biotic and biogeochemical, with adjacent ecosystems must be known, studied in the wild, and be simulated so that the essential functional characteristic of the analog ecosystem can be maintained.

Hundreds, perhaps thousands, of microcosm studies of liter or few liter dimensions of a very limited biodiversity have been undertaken to elucidate component ecosystem function, often related to toxic compound effect. Rarely could these studies be regarded as the modeling of an ecosystem. At the other extreme, perhaps the most complex ecosystem modeling effort ever undertaken was the Biosphere II project in Arizona during the 1980s and 1990s. Biosphere II was an ecologically well-conceived collection of interacting terrestrial marine and freshwater ecosystems. However, it was intentionally operated as a closed system because of its planned space station future. Several decades ago, it was widely regarded among ecologists that even though greenhouse enclosure provided a critical element of control over variables, the difficulties inherent in enclosure and operation were too great to allow ecosystem modeling. As the examples we provide below show, this judgment was only minimally correct and perhaps no more severe than the breakup of wild ecosystems by development or farming expansion. Human expansion and perturbation has severely altered many of the ecosystems on Earth, and has altered all ecosystems in at least minor ways. The entire biosphere has been in effect placed in a poorly operated greenhouse, with the atmosphere serving as its upper `glass' roof. There can be no valid argument against greenhouse enclosure of ecosystems for research and education. It is simply one end of a complex spectrum of interacting biota and biogeochemistry that we seek to understand. Indeed, in many ways, such model ecosystems may be `purer' than their wild counterparts.

Physical/Chemical Control Parameters

The Enclosure

The shape of an ecosystem relative to its controlling physical and energy parameters can be crucial. In the case of aquatic systems, the relative thickness of the water mass and its relationship to the bottom establish the basic character of an ecosystem. A large body of water would be dominated by true plankters, normally living most of their lives suspended in mid and surface waters, with little benthic (or bottom) influence, whereas the shallow stream or narrow lagoon of a few meters in depth is benthic dominated. Light enters only through the air–water interface of a water ecosystem, and the shape of the containing body of water relative to depth, as well as water turbidity, determines the photosynthetic versus heterotrophic character of the ecosystem. The direction of current flow and wave action through an aquatic system relative to the position and orientation of its communities is

 doi:10.1016/B978-0-444-63768-0.00343-7

Fig. 1 Florida Everglades mesocosm during construction. The butyl rubber-lined concrete block walls were used to constrain the entire system as well as to physically separate the salinity subcomponents thereby creating a salinity gradient. The plastic box at the lower right contains the tide controller, which determines the tide level in the estuary (center tank and the four smaller units behind).

critical to simulate in any model. The direction, frequency, and strength of wind relative to forest or field size can also be critical to systems function, as can be the physical dimension and density of such ecosystems.

The all-glass or acrylic aquarium box, ranging from about 40 l (10 gallons) to 1000 l (250 gallons), is a standard piece of equipment in terrestrial and aquatic modeling, and by drilling holes to attach pipes and linking all-glass tanks in complex arrays, many aspects of wild ecosystems can be modeled with reasonable accuracy.

The construction of molded fiberglass tanks or poured concrete or concrete block tanks, sealed with a wide variety of newer sealants, has considerable advantages for larger systems.

Ideally, the ecosystem envelope would be like that of the boundary of the mathematical modeler, a theoretical boundary allowing the controlling of exchange but not having any inherent characteristics. Walls, whatever their nature, unless rather esoteric measures are used to prevent organisms and organic molecules from using their surfaces, or blocking wind or current, are intrusions into the model ecosystem that may or may not be acceptable. For a small model of a planktonic system, the presence of uncleaned walls may prevent the system from being plankton dominated. To some degree, walls also interact with the water and atmosphere of the ecosystem they contain. For most purposes, glass and many plastics are ideal in this respect.

Greenhouse walls and roofs can block ultraviolet light and, for most ecosystem models, a component of artificial light is probably essential to achieve both the intensity and spectral veracity of natural light. Reinforced cement block or concrete can be valuable construction materials for large systems; however, concrete interacts with both water and atmosphere, being one of the limiters of veracity in of Biosphere II (as we describe below), and must be sealed with epoxy or other, carefully considered resins (Fig. 1).

Many chemical elements and compounds used in construction are toxic. Some of these are only mildly poisonous and are often required by organisms as elements in small quantities and only become toxic in excess. Others are always toxic and only concentration determines effect. Glass, acrylics, epoxies, polyesters, polypropylenes, polyethylenes, nylons, Teflon, and silicones, among others, are structural materials commonly used in model/greenhouse construction. When properly cured these materials are generally inert, nonbiodegradable, and nontoxic. Many metals and organic additives easily find their way into construction processes and must be avoided or sealed off.

Physical/Chemical Environment

Many of the physical/chemical parameters of ecosystem, such as temperature, salinity, pH, hardness, and oxygen, are more or less obvious and generally accepted as crucial. Others such as light, wind, tides, currents, and wave action, have often been neglected or at least minimally considered in their effects on ecosystem models.

Light

Whole communities or parts of ecosystems, where plants are major components, typically capture a maximum of 6% of the incident light energy in photosynthesis. Nevertheless, full light is often required to achieve that peak transfer of energy. Also much higher capture efficiencies are possible when forcing energy such as wind and wave are present. In many cases, if greenhouse roofs cannot be opened, artificial lighting will have to be introduced to achieve the correct spectrum and intensity to drive the primary production characteristics of an analog ecosystem.

Water Supply/Water Environment

Whether a terrestrial or aquatic ecosystem is planned, the supply and internal transfer of water is critical. Air- and water-handling systems need to be carefully designed to prevent water contamination. Since water sequestration and loss is more or less inevitable,

Fig. 2 Red mangrove community in the Florida Everglades mesocosm from the engineering control pad. The large fan in the upper center provides wind for the mangrove communities. The box in the lower right is one of a bank of five algal turf scrubbers (ATSs) that control water quality (nutrients, pH, O_2) in the coastal system.

the water quality of both initial water and later top-ups must be carefully controlled. Rarely would tap water be acceptable. Water is the universal solvent, whether in liquid or gaseous form, and often `sequesters' gases. Most ecosystems in greenhouses require the dedicated monitoring and control of atmospheric and water quality. Managed aquatic plant systems, such as algal turf scrubbers (ATSs), have been successfully used to manage water quality of adjacent ecosystems interaction, as we describe in some of the examples. Such systems can also control atmospheric quality (Fig. 2).

Water and Air Movement

In virtually all water ecosystems, the water flows, and in most shallow water systems it oscillates (surges) as well. In models, this flow and surge are developed, at least initially by pumps. However, standard impellor pumps destroy or damage many plankters, particularly larger zooplankton and swimming invertebrate larvae. Several approaches are available to solve this problem, including using slow-moving piston pumps, membrane pumps, and Archimedes screws (Fig. 3). All of these devices can work well, though relative performance is not fully quantified.

In terrestrial environments, fans for wind and air handlers for heat and air conditioning, as well as the cooling or heating surfaces employed have the same effect on flying insects and birds. On the other hand, in the wild, ultraviolet light, wind, and rain have critical controlling effects on many plant predators. These factors cannot be omitted.

Biotic Parameters

Ecosystem Structuring Elements

Some communities of organisms are structured by physical elements – a sandy beach, or rock, for example. However, in most terrestrial environments and in many shallow aquatic environments, plants and algae are the structuring elements. They not only provide the food and water and atmospheric chemistry but also greatly increase surface for attachment and cover. In general, plants also provide a spatial heterogeneity (spatial surface) that does not exist in the physical world. Particularly in the marine environment, where calcification is enhanced, many animals join plants to provide a community structure that consists of reef or shell framework. This framework is calcium carbonate (or other organic solid such as chiton) instead of (or along with) plant cellulose. In constructing any living ecosystem, it is essential that these structuring elements be first developed as `colonial' stages soon after the physical environment is formed.

Ecosystem Subunits

In the construction of greenhouse ecosystems, subunit installation can be utilized. However, it would be impossible to individually extract and emplace the tens to hundreds of species amounting to hundreds to millions of individuals that occur in these subunits. Installation of sub-blocks of wild ecosystem includes the microspecies and keeps their relationships intact. For example, soil blocks, or in the marine or aquatic habitat, mud or rock blocks, can be introduced into the preexisting physical/chemical elements of the model ecosystem.

Repeated efforts must be taken to install rock, soil, mud, or `planktonic blocks'. These injections should be periodically carried out during system stocking; at completion of development, they should be followed by several final injections. The process of cutting out, or otherwise extracting, an ecological block or ecosystem subunit and transporting it to the waiting model can be stressful to the community of organisms within the block. Even in the model, the block meets conditions that at least initially

Fig. 3 Engineering/control pad in the Florida Everglades mesocosm. The green diagonal tube in the center is an Archimedes screw that lifts water from the coastal tank (far right) for distribution to the estuary (back right), the ATS (left-foreground), and the wave generator (out of view to lower right).

Fig. 4 Salt marsh community 1 year after establishment of the Florida Everglades mesocosm. Young white mangrove tree to left. After 5 years of self-organization the system followed a succession to a white mangrove/buttonwood swamp.

consist of the raw physical/chemical environment unameliorated by the effects of a functioning community of organisms. The first block injections are likely to lose species. However, with each addition, the diversity of reproductively successful species increases.

All ecological communities are patchy. An island, coral reef, a large salt marsh, a field, even a forest, all differ from place to place. Chance factors of organism settlement, negative and positive interaction between species, the local effect of environment, and real differences of environment (wave exposure, current, etc.) all lead to patchiness within a community. The model itself, no matter how accurate, is a patch, or several patches, that the modeler hopes represents a `mean' of most wild patches.

After the structuring elements are established, and the entire pool of available species from the type community given a chance at immigration into the model, the model will self-organize. In the form of the genetic codes of its constituent species, the ecosystem carries a tremendous quantity of information with regard to its structure and function. Particularly since we know and understand only a small part of this information, we should be loath to subvert ecosystem self-organization (**Fig. 4**).

Care in adhering to wild density levels will help prevent overstocking, overgrazing, and overpredation until the model is better understood. Single members of species guilds can be selected to perform a function, and thus reproductive density is achieved without exceeding ecosystem density requirements. In general, the larger the population of any single species, the more likely it is that breeding success will be achieved.

Ecosystem Interchange

However, one arbitrarily draws the boundary, no ecosystem occurs in total isolation. In many cases when such boundaries are arbitrary, major survival effects must be provided by adjacent ecosystems. For example, coral reefs and most shallow benthic communities are greatly dependent on the effects of adjacent open bodies of water for food, oxygen, and wave and current `drive'. Typically, filters, the core element of aquarium science, have been devised to fill the need for the larger, less animal-dense body of adjacent open water that has been `filtered' by the settling or loss of organic particles to deep water. However, such filters to a large extent usurp the role normally

provided by plants in most of the communities that are modeled. Unfortunately, in so doing they do not add oxygen as the plants do, and they raise nutrient levels. Both bacterial and foam fractionation methods remove organic particulates and swimming plankters, including reproductive stages that should be part of ecosystem function. Managed aquatic plant systems, such as ATSs, have been successfully used to manage water quality of adjacent ecosystems interaction, as we describe in some of the examples.

Although terrestrial systems, in general, may be less difficult in this regard, simulation of biotic interchange may be crucial. For example, birds and mammals often change ecosystems seasonally and even diurnally and the effects may be critical. Many insects are seasonal, some for very short periods, and often cross ecosystem boundaries. In some cases, it may be possible to provide these interactions through a human manager; however, a refugium, or alternate ecosystem may be necessary to achieve veracity.

The Operational Imperative

Successful enclosed ecosystem operation requires the monitoring of a large number of physical and chemical factors. To a large extent, this can be automated with electronic sensors, and the data can be logged and the system computer controlled. Some chemical parameters require wet chemistry, though a once-a-week analysis is usually sufficient in a well-run system. Like any piece of complex laboratory equipment (a scanning electron microscope, for example) a dedicated and highly trained technician is needed to manage the monitoring equipment, though in a well-tuned system, considerable time can be available for other duties.

An operational feature that is rarely discussed, and in practice is mostly anecdotal is that of population instability. A mesocosm, in effect, is a few-square-meter patch of a larger ecosystem. In the wild, ecosystem patches of a few square meters can be subject to considerable short-term variability, though stability is achieved to some extent by the smoothing effect of the larger ecosystem that may be measured in square kilometers.

Microcosms and mesocosms require an ecologist, fully acquainted with `normal' community structure of the `wild-type' system. Effectively, that ecologist/operator performs as the highest, and most omnivorous, predator. In the cases of algal or insect `explosions' the operator's function is obvious, a once-a-week cropping or `grazing' (i.e., hand harvest) until the explosion tendency subsides. In other cases, the short-term introduction of predator to carry out the limited cropping or grazing role can be quite successful. These `managed predators' can be kept in a refugium unit where they are readily available for such service.

Case Study: Coral Reef Microcosm

The Caribbean coral reef ecosystem model shown in Fig. 5 received natural sunlight from one side, south-facing at 37.5° N latitude; the metabolic unit had six 160 W VHF flourescent lamps (to match tropical intensity), step-cycled to bring mid-day peak intensity to approximately 800 uE m^{-2} s^{-1} and total incoming light to 220 Langleys/day (Fig. 6). The ATS, lighted at night, had three 100 W metal halide lamps. The discussion presented represents data accumulated throughout the 9th year of 10 years of operation.

The physical and chemical components of the microcosm were measured in the metabolic unit and closely match those of the St. Croix analog (Table 1). The pH of the microcosm ranges from 7.96 ± 0.01 ($n = 62$) in the morning to 8.29 ± 0.10 ($n = 39$) in the late afternoon. Because of linked interacting photosynthesis and calcification in the ecosystem, calcium concentrations and

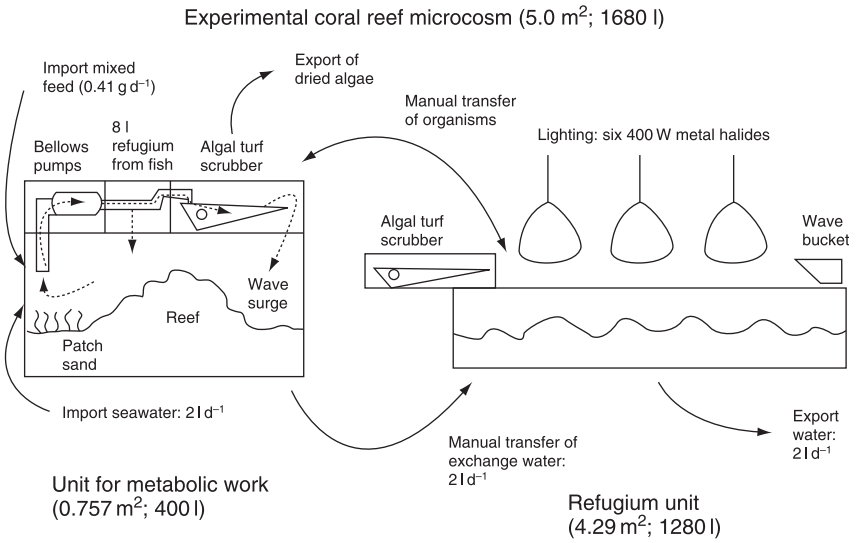


Fig. 5 Diagram of coral reef microcosm with its refugium.

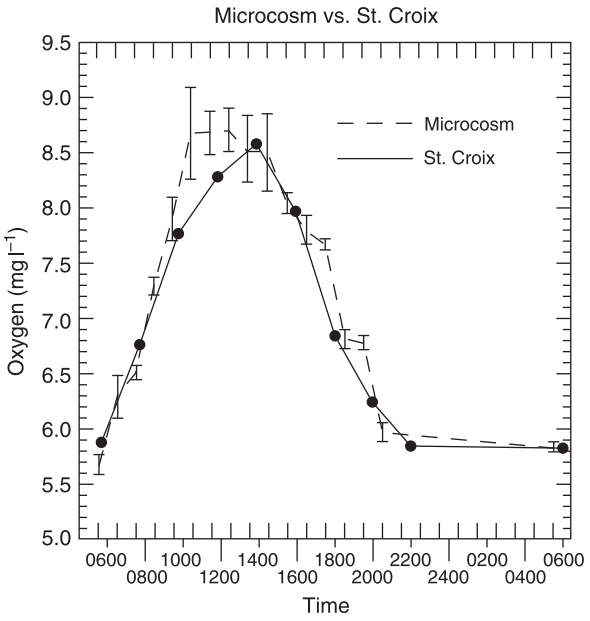


Fig. 6 Comparison of mean daily oxygen concentration in the coral reef microcosm in comparison with that over the wild analog reef (1-year means).

Table 1 Comparison of physical/chemical parameters between coral reef microcosm and the wild analog reef

	Microcosm	*St. Croix Reefs (fore-reef)*[a]
Temperature (°C) (am–pm)	26.5 ± 0.03 (n = 365)–27.4 ± 0.02 (n = 362)	24.0–28.5
Salinities (ppt)	35.8 ± 0.02 (n = 365)	35.5[b]
pH (am–pm)	7.96 ± 0.01 (n = 62)–8.29 ± 0.02 (n = 39)	8.05–8.35[c]
Oxygen concentration (mg l^{-1}) (am–pm)	5.7 ± 0.1 (n = 14)–8.7 ± 0.2 (n = 11)	5.8–8.5
GPP (g O_2 m^{-2} d^{-1}); (mmol O_2 m^{-2} d^{-1})	14.2 ± 1.0 (n = 4); 444 ± 3 (n = 4)	15.7; 491
Daytime NPP (g O_2 m^{-2} day^{-1}); (mmol O_2 m^{-2} d^{-1})	7.3 ± 0.3 (n = 4); 228 ± 9 (n = 4)	8.9; 278
Respiration (g O_2 m^{-2} h^{-1}); (mmol O_2 m^{-2} h^{-1})	0.49 ± 0.04 (n = 4); 15.3 ± 1.3 (n = 4)	0.67; 20.9
$N - NO_2^- + NO_3^2$(μmol)	0.56 ± 0.07 (n = 6)	0.28
Calcium (mg l^{-1}); (mmol l^{-1})	491 ± 6 (n = 33); 12.3 ± 0.2 (n = 33)	417.2[d]; 10.4
Alkalinity (meq l^{-1})	2.88 ± 0.04 (n = 59)	2.47[b]
Light[e] (Langleys d^{-1})	220	430 (surface); 220 (5 m deep infore-reef)

[a]The St. Croix data is from Adey and Steneck (1985).
[b]Tropical Atlantic means from Millero and Sohn (1992); no data available for St. Croix.
[c]Values from Enewetak and Moorea (Odum and Odum, 1955; Gattuso *et al.*, 1997).
[d]Tropical Atlantic means from Sverdrup *et al.* (1942); no data available for St. Croix.
[e]The light levels of the system were measured with a pyranograph. All of the physical and chemical components of the microcosm are compared to the fore-reef of St. Croix since light levels are equivalent (Kirk, 1983; Adey and Steneck, 1985). For references, see Small A and Adey W (2001) Reef corals, zooxanthellae and free-living algae: A microcosm study that demonstrates synergy between calcification and primary production. *Ecological Engineering* 16: 443–457.

alkalinity continually fall during the day and are stable or rise slightly at night. Calcium was added each morning as a solution of aragonite dissolved in HCl at approximately 24 000 mg l^{-1}. To keep microcosm concentrations above 420 mg l^{-1}, after a full day of calcification, the mean concentration of calcium in the system was maintained at 491 ± 6 mg l^{-1}. Bicarbonate, was added as either $NaHCO_3$ or $KHCO_3$ dissolved in distilled water. The mean alkalinity was 2.88 meq l^{-1} (n = 59), in order to maintain levels above 2.40 meq l^{-1}. Water quality (nutrients, oxygen, and pH) of the system was controlled by algal turf scrubbing.

The mean oxygen concentration of the microcosm as shown in **Fig. 6** is very close to that of the analog St. Croix reef. Net primary productivity (NPP) and respiration (R) were calculated based on the rate of oxygen increase and decrease, respectively, across the point of saturation (6.5 mg 1^{-1} O_2), to avoid atmospheric fluxes. This gave a mean gross primary productivity (GPP) of 14.2 ± 1.0 gO_2 m^{-2} d^{-1}, as compared to the mean GPP for the analog fore-reef at 15.7 gO_2 m^{-2} d^{-1}. The difference between the microcosm and reefs *in situ* can be accounted for by the difference in spatial heterogeneity; topographic relief on the St. Croix fore-reef typically ranges from 1 to 2 m, while in the microcosm only 10–30 cm is possible.

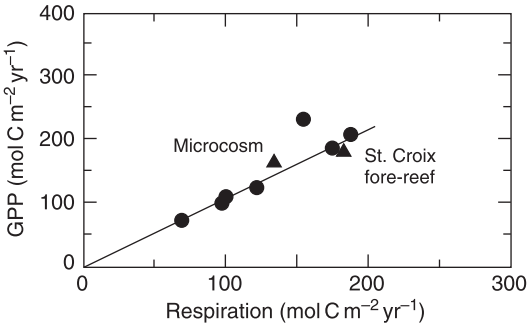


Fig. 7 GPP as a function of respiration in the coral reef microcosm and its wild analog reef in comparison with selected worldwide reefs.

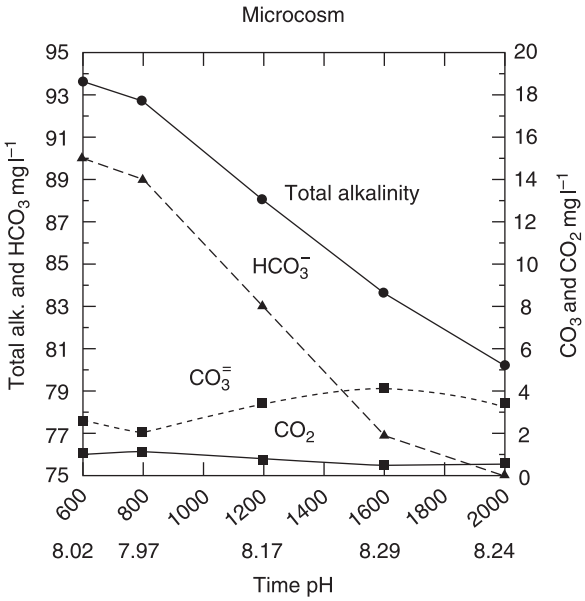


Fig. 8 Mean daytime carbonate cycle in coral reef microcosm calculated by namograph from pH and total alkalinity data.

In Fig. 7, GPP versus R for the microcosm and its analog are plotted, showing that both are well within the range of typical wild reefs. Even though primary productivity of the microcosm is very close to the wild analog, the fact that respiration is somewhat lower probably relates to the proportionally lower spatial heterogeneity in the microcosm.

Whole ecosystem calcification in the coral reef model, at 4.0 ± 0.2 kg $CaCO_3$ m^{-2} yr^{-1}, is related to its primary components (stony coral 17.6%, *Halimeda* 7.4%, *Tridacna* 9.0%, algal turf, coralline and foraminifera 29.4%, and miscellaneous invertebrates 36%). Through analysis of the microcosm's daily carbonate system, it is demonstrated that bicarbonate ion (Fig. 8), not carbonate ion, is the principal component of total alkalinity reduction in the water column.

This coral reef microcosm contained 534 identified species within 27 phyla (Table 2), with an estimated 30% unaccounted for due to lack of taxonomic specialists. Because of the length of time that this model system was closed to biotic interchange, virtually all of the biotic composition of the system (over 95%) had to be maintained by reproduction. Based on standard species/area relationships ($S = kA^z$, where S = species richness and A = area), the predicted pan tropic coral reef biodiversity calculated from the model biodiversity (at three million species) exceeds that of recent estimates for wild coral reefs.

Case Study: Florida Everglades Mesocosm

This greenhouse-scale mesocosm is a 98 500 l butyl-lined, concrete block tank divided into seven connected sections of varying salinity (Fig. 9). Each section contains water, algae, animals, sediments, and wetland–coastal plants representative of habitats along a transect from the full salinity Gulf of Mexico through the estuarine Ten-Thousand Islands and into the freshwater Florida Everglades (Fig. 10).

As in the wild analog, the Gulf Shore and estuary are part of the same dynamic water mass. Here, the estuarine salinity gradient is created by pump-driven tidal inflow interacting through open weir constrictions and against downstream freshwater flow. Tank

Table 2 Families of organisms, with numbers of species and genera found in the Coral reef microcosm after 10 years of operation, 7 years in closure

Plants, algae, and cyanobacteria
Division Cyanophota
Chroococcaceae 6/5
Pleurocapsaceae 4/2
UID Family 4/4
Oscillatoriaceae 8/6
Rivulariaceae 4/1
Scytonemataceae 1/1
Phylum Rhodophyta
Goniotrichaceae 2/2
Acrochaetiaceae 2/2
Gelidiaceae 1/1
Wurdemanniaceae 1/1
Peysonneliaceae 3/1
Corallinaceae 11/8
Hypneaceae 1/*1*
Rhodymeniaceae 3/2
Champiaceae 1/1
Ceramiaceae 3/3
Delesseriaceae 1/1
Rhodomelaceae 7/6
Phylum Chromophycota
Cryptomonadaceae 2/2
Hemidiscaceae 1/1
Diatomaceae 6/4
Naviculaceae 9/4
Cymbellaceae 3/1
Entomoneidaceae 1/1
Nitzchiaceae 6/4
Epithemiaceae 3/1
Mastogloiaceae 1/1
Achnanthaceae 9/3
Gymnodiniaceae 6/4or5
Gonyaulacaceae 1/1
Prorocentraceae 2/1
Zooxanthellaceae 1/1
Ectocarpaceae 2/2
Phylum Chlorophycota
Ulvaceae 1/1
Cladophoraceae 4/2
Valoniaceae 2/2
Derbesiaceae 3/1
Caulerpaceae 3/1
Codiaceae 6/2
Colochaetaceae 1/1
Phylum Magnoliophyta
Hydrocharitaceae 1/1
Kingdom Protista
Phylum Percolozoa
Vahlkampfiidae 2/1
UID Family 2/2
Stephanopogonidae 2/1
Phylum Euglenozoa
UID Family 4/3
Bondonidae 7/1
Phylum Choanozoa
Codosigidae 2/2
Salpingoecidae 1/1

(*Continued*)

Table 2 Continued

Phylum Rhizopoda
Acanthamoebidae 1/1
Hartmannellidae 1/1
Hyalodiscidae 1/1
Mayorellidae 2/2
Reticulosidae 2/2
Saccamoebidae 1/1
Thecamoebidae 1/1
Trichosphaeridae 1/1
Vampyrellidae 1/1
Allogromiidae 1/1
Ammodiscidae 1/1
Astrorhizidae 1/1
Ataxophragmiidae 1/1
Bolivinitidae 3/1
Cibicidiidae 1/1
Cymbaloporidae 1/1
Discorbidae 5/2
Homotremidae 1/1
Peneroplidae 1/1
Miliolidae 10/2
Planorbulinidae 2/2
Siphonidae 1/1
Soritidae 4/4
Textulariidae 1/1
Phylum Ciliophora
Kentrophoridae 1/1
Blepharismidae 2/2
Condylostomatidae 1/1
Folliculinidae 4/3
Peritromidae 2/1
Protocruziidae 2/1
Aspidiscidae 7/1
Chaetospiridae 1/1
Discocephalidae 1/1
Euplotidae 11/3
Keronidae 7/2
Oxytrichidae 1/1
Psilotrichidae 1/1
Ptycocyclidae 2/1
Spirofilidae 1/1
Strombidiidae 1/1
Uronychiidae 2/1
Urostylidae 4/2
Cinetochilidae 1/1
Cyclidiidae 3/1
Pleuronematidae 3/1
Uronematidae 1/1
Vaginicolidae 1/1
Vorticellidae 2/1
Parameciidae 1/1
Colepidae 2/1
Metacystidae 3/2
Prorodontidae 1/1
Amphileptidae 3/3
Enchelyidae 1/1
Lacrymariidae 4/1
Phylum Heliozoa
Actinophyridae 2/1
Phylum Placozoa
Family UID 5
Phylum Porifera
Plakinidae 2/1

Table 2 Continued

Geodiidae 5/2
Pachastrellidae 1/1
Tetillidae 1/1
Suberitidae 1/1
Spirastrellidae 2/2
Clionidae 4/2
Tethyidae 2/1
Chonrdrosiidae 1/1
Axinellidae 1/1
Agelasidae 1/1
Haliclonidae 4/1
Oceanapiidae 1/1
Mycalidae 1/1
Dexmoxyidae 1/1
Halichondridae 2/1
Clathrinidae 1/1
Leucettidae 1/1
UIDFamily 2/?
Eumetazoa
Phylum Cnidaria
UID Family 3/?
Eudendridae 1/1
Olindiae 1/1
Plexauridae 1/1
Anthothelidae 1/1
Briareidae 1/1
Alcyoniidae 2/2
Actiniidae 3/2
Aiptasiidae 1/1
Stichodactylidae 1/1
Actinodiscidae 4/3
Corallimorphidae 3/2
Acroporidae 2/2
Caryophylliidae 1/1
Faviidae 3/2
Mussidae 1/1
Poritidae 3/1
Zoanthidae 3/2
Cerianthidae 1/1
Phylum Platyhelminthes
UID Family 1/1
Anaperidae 3/2
Nemertodermatidae 1/1
Kalyptorychidae 1/1
Phylum Nemertea
UID Family 2/2
Micruridae 1/1
Lineidae 1/1
Phylum Gastrotricha
Chaetonotidae 3/1
Phylum Rotifera
UID Family 2/?
Phylum Tardigrada
Batillipedidae 1/1
Phylum Nemata
Draconematidae 3/1
Phylum Mollusca
Acanthochitonidae 1/1
Fissurellidae 2/2
Acmaeidae 1/1
Trochidae 1/1
Turbinidae 1/1
Phasianellidae 1/1
Neritidae 1/1

(*Continued*)

Table 2 Continued

Rissoidae 1/1
Rissoellidae 1/1
Vitrinellidae 1/1
Vermetidae 1/1
Phyramidellidae 1/1
Fasciolariidae 2/2
Olividae 1/1
Marginellidae 1/1
Mitridae 1/1
Bullidae 1/1
UID Family 4/?
Mytilidae 2/1
Arcidae 2/1
Glycymerididae 1/1
Isognomonidae 1/1
Limidae 1/1
Pectinidae 1/1
Chamidae 1/1
Lucinidae 2/2
Carditidae 1/1
Tridacnidae 2/1
Tellinidae 1/1
Phylum Annelida
Syllidae 3/2
Amphinomidae 1/1
Eunicidae 3/1
Lumbrineridae 1/1
Dorvilleidae 1/1
Orbiniidae 1/1
Spionidae 1/1
Chaetopteridae 1/1
Paraonidae 1/1
Cirratulidae 4/3
Ctenodrilidae 4/3
Capitellidae 3/3
Muldanidae 1/1
Oweniidae 1/1
Terebellidae 2/1
Sabellidae 14/4
Serpulidae 6/6
Spirorbidae 2/2
Dinophilidae 1/1
Phylum Sipuncula
Golfingiidae 1/1
Phascolosomatidae 3/2
Phascolionidae 1/1
Aspidosiphonidae 3/2
Phylum Arthropoda
Halacaridae 1/1
UID Family 2/?
Cyprididae 2/2
Bairdiiaae 1/1
Paradoxostomatidae 1/1
Pseudocyclopidae 1/1
Ridgewayiidae 2/1
Ambunguipedidae 1/1
Argestidae 1/1
Diosaccidae 1/1
Harpacticidae 1/1
Louriniidae 1/1
Thalestridae 1/1
Tisbidae 1/1
Mysidae 1/1
Apseudidae 2/1

Table 2 Continued

Paratanaidae 1/1
Tanaidae 1/1
Paranthuridae 1/1
Sphaeromatidae 1/1
Stenetriidae 1/1
Juniridae 1/1
Lysianassidae 1/1
Gammaridae 4/4
Leucothoidae 1/1
Anamixidae 1/1
Corophiidae 1/1
Amphithoidae 2/2
Alpheidae 2/2
Hippolytidae 2/1
Nephropidae 1/1
Diogenidae 1/1
Xanthidae 2/?
Phylum Echinodermata
Ophiocomidae 1/1
Ophiactidae 1/1
Cidaroidae 1/1
Toxopneustidae 1/1
Holothuriidae 1/1
Chirotidae 1/1
Phylum Chordata
Ascidiacea UID Fam..1/1
Grammidae 1/1
Chaetodontidae 1/1
Pomacentridae 5/4
Acanthuridae 1/1

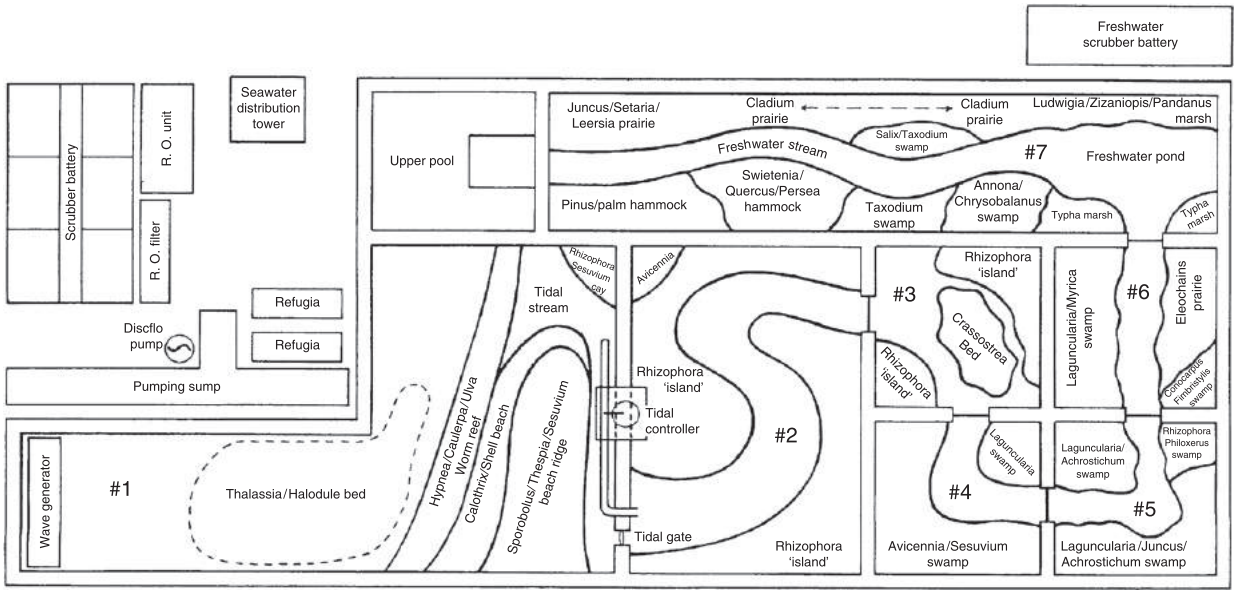


Fig. 9 Plan view of the Florida Everglades mesocosm and its critical engineering components.

#1, the Gulf Shore, acts as a tidal reservoir for the estuary, thereby saving the need for a blank reservoir (Fig. 11). The primary pump is an 800 lpm, Discflo™ unit that utilizes a rotating disk, rather than plankton-destructive impellers. The freshwater is derived from rain and from reverse osmosis extraction from the Gulf Shore (the equivalent of Gulf evaporation and resulting

Fig. 10 Florida Everglades mesocosm approximately 4 years after construction showing salt marsh, black mangrove, and red mangrove communities (from front to background at left) and lower freshwater stream at right. At this point the greenhouse roof is providing a significant constraint to community succession by limiting vertical growth of mangrove and hammock trees.

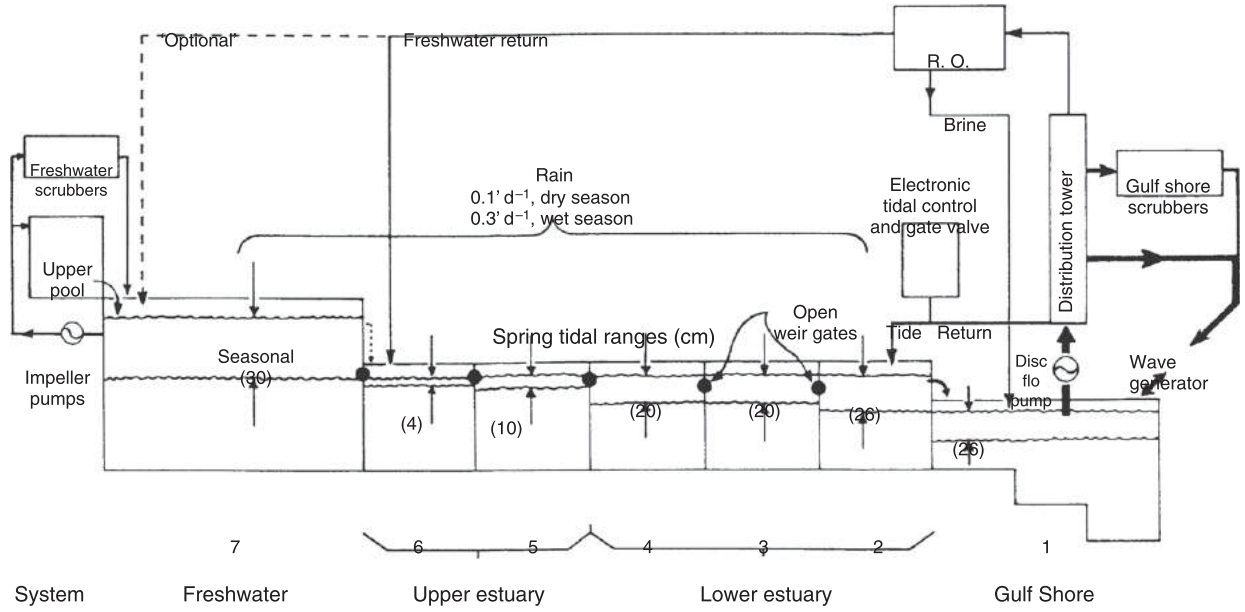


Fig. 11 Vertical/longitudinal section through Florida Everglades mesocosm showing water management system and tide levels.

rainfall in the wild). All aquatic organisms, including adult invertebrates, can move from the estuary to the Gulf Shore. All organisms that can survive Discflo™ pumping (including small fish) can return to the estuary via tidal inflow. The freshwater system, at times, flows directly into the uppermost estuary and technically all organisms can enter the estuary from freshwater.

The initial stocking of the mesocosm was completed in mid-1988, and small collections continued to be injected through 1990. During this period, partial censuses for key organisms were undertaken, and, where required, additional stocking was carried out. From late 1990 to late 1994, the system was operated as a biotically closed system, with minor human interaction, functioning as an omnivorous predator.

Major physical/chemical parameters are shown in **Table 3**. Dissolved nitrogen was monitored as nitrite plus nitrate in each of the community units (tanks); these were typically at levels of 5–8 μM ($NO_2 + NO_3$) through the middle of the estuary, and at 3–5 μM ($NO_2 + NO_3$), in the Gulf (#1) system. Levels average a few μM higher in winter than in summer. Nutrient flow-through is achieved by algal export, in the ATS banks. When levels drop below 1–2 μM in the Gulf (#1) system (typically in summer), the dried scrubber algae are redistributed to the system.

After 4 years of biotically closed operation, the Florida Everglades mesocosm was censused for organisms. The abundance of the principal higher plants, algae, invertebrates, and fish are shown in **Figs. 12–14**. A total of 369 species (not including bacteria, fungi and the minor `worm' phyla) was tallied. Excluding algae, protists, and small invertebrates, which could not be censused during introduction, it can be estimated that approximately 20–40% of the introduced species survived through the 4 years of

Table 3 Physical/chemical parameters of the Florida Everglades mesocosm

Parameter	Tank #1	Tank #2	Tank #3	Tank #4	Tank #5	Tank #6	Tank #7
Temperature °C							
Spring	23.4	23.2	22.5	22.6	22.0	21.3	23.2
Summer	25.7	25.5	25.4	25.7	25.6	25.1	25.1
Fall	22.2	22.3	21.8	22.0	21.7	21.3	22.1
Winter	21.0	20.9	19.9	19.6	19.0	18.4	21.9
Salinity, ppt	31.6	31.2	30.5	28.7	19.7	0.7	0.1
$[NO_2 + NO_3]$ μM							
Tap H_2O as top up[a]	7.2	7.6	8.2	6.3	5.4	6.6	6.7
Milli RO as top up[b]	1.4	1.7	2.3	1.8	0.9	1.7	1.4
Tidal range cm/0.5 day	13–26	13–26	13–20	11–20	6–10	0–4	0
Hydroperiod cm yr^{-1}	0	0	0	0	0	0	30.5

[a]Nutrient levels in system as $(NO_2 + NO_3)$ μM when `Tap H_2O as top up'.
[b]`Milli RO as top up' refers to mean system values when reverse osmosis water from Milli RO™ is used as evaporative replacement.

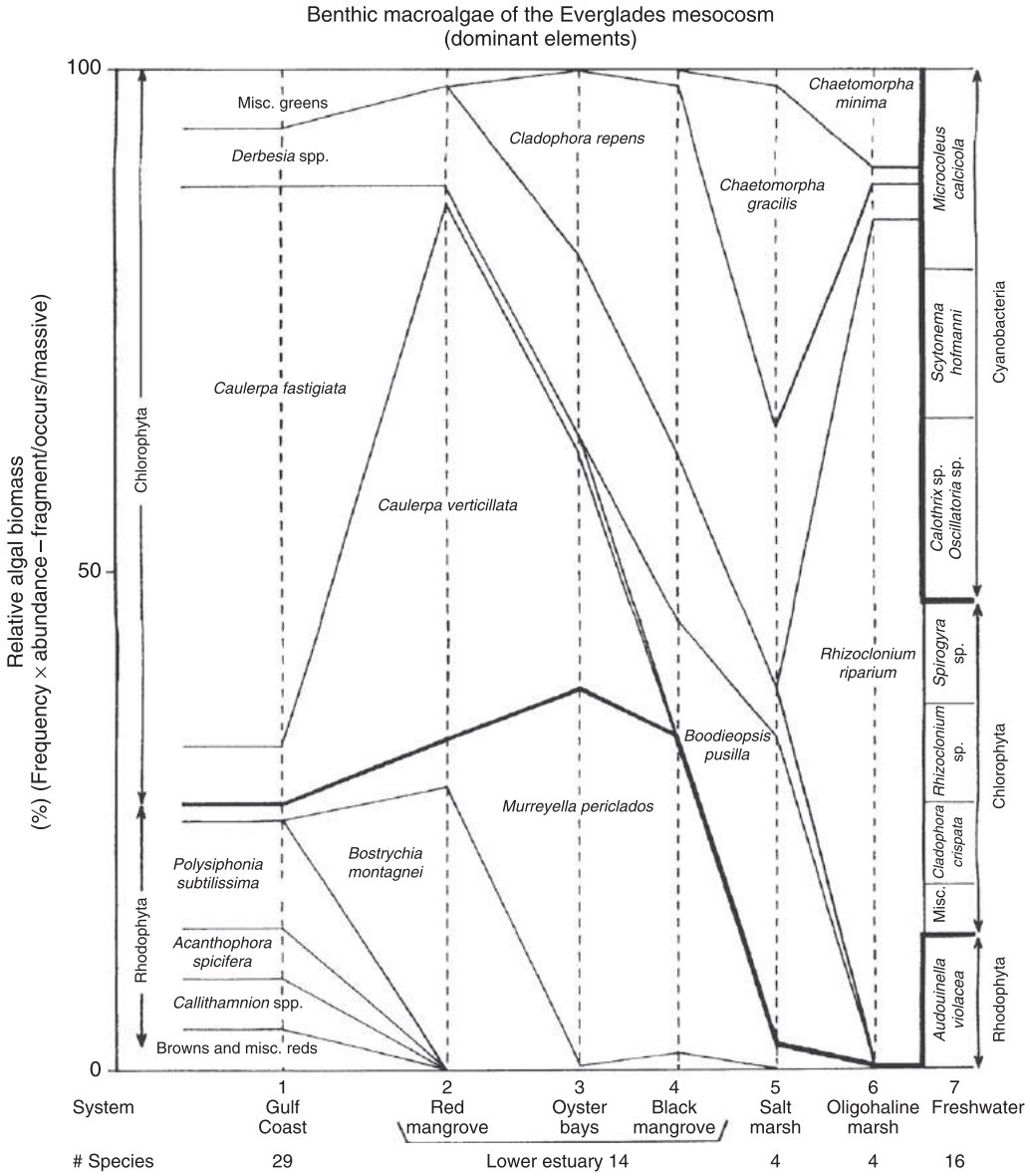


Fig. 12 Relative biomass of dominant benthic algae in Florida Everglades mesocosm.

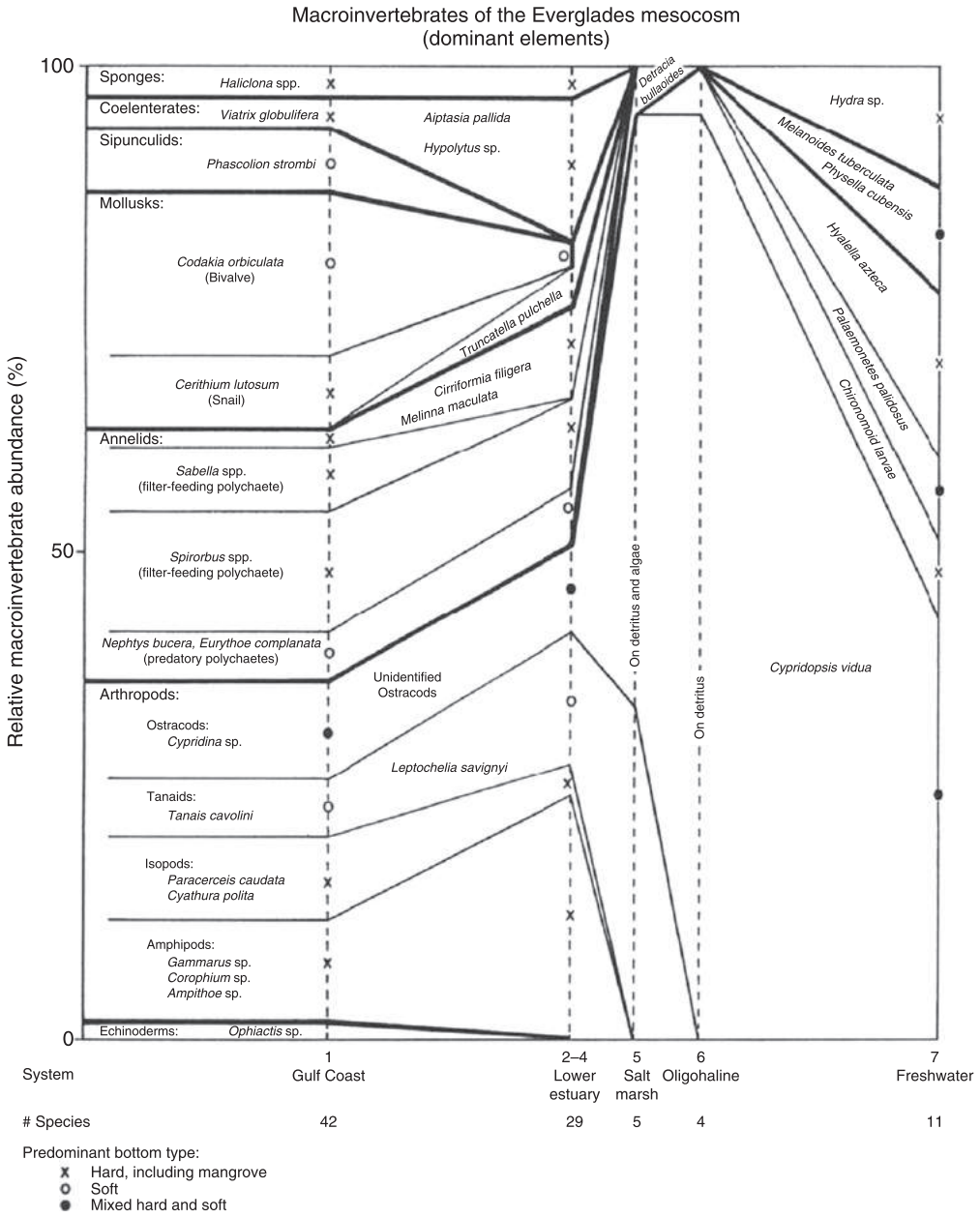


Fig. 13 Relative invertebrate abundance in the Florida Everglades mesocosm.

biotic closure. In most cases these were the dominating species in the analog ecosystems. At the time of termination of the system as a carefully monitored mesocosm, only 15–30% of the originally introduced species were reproductively maintaining populations. However, in most cases, as Figs. 12–14 show, these were the species that provided primary structure and metabolism in the analog ecosystem.

Case Study: Biosphere 2

Biosphere 2, located near Tucson, AZ, USA, is the largest greenhouse system ever built with nearly three acres (1.2 ha) of enclosed space. It is unique in surpassing any other greenhouse ecosystem in size, complexity, and duration of operation. The system was originally intended as a model of the Earth's biosphere (e.g., biosphere 1) with several tropical and subtropical ecosystems, an agricultural area, wastewater treatment wetlands, and a human habitat, along with a factory-sized machinery area for maintaining physical–chemical conditions. It was built to develop bioregenerative technology for future space travel, to educate the public about biosphere-scale issues and for basic ecological research. Atmospheric

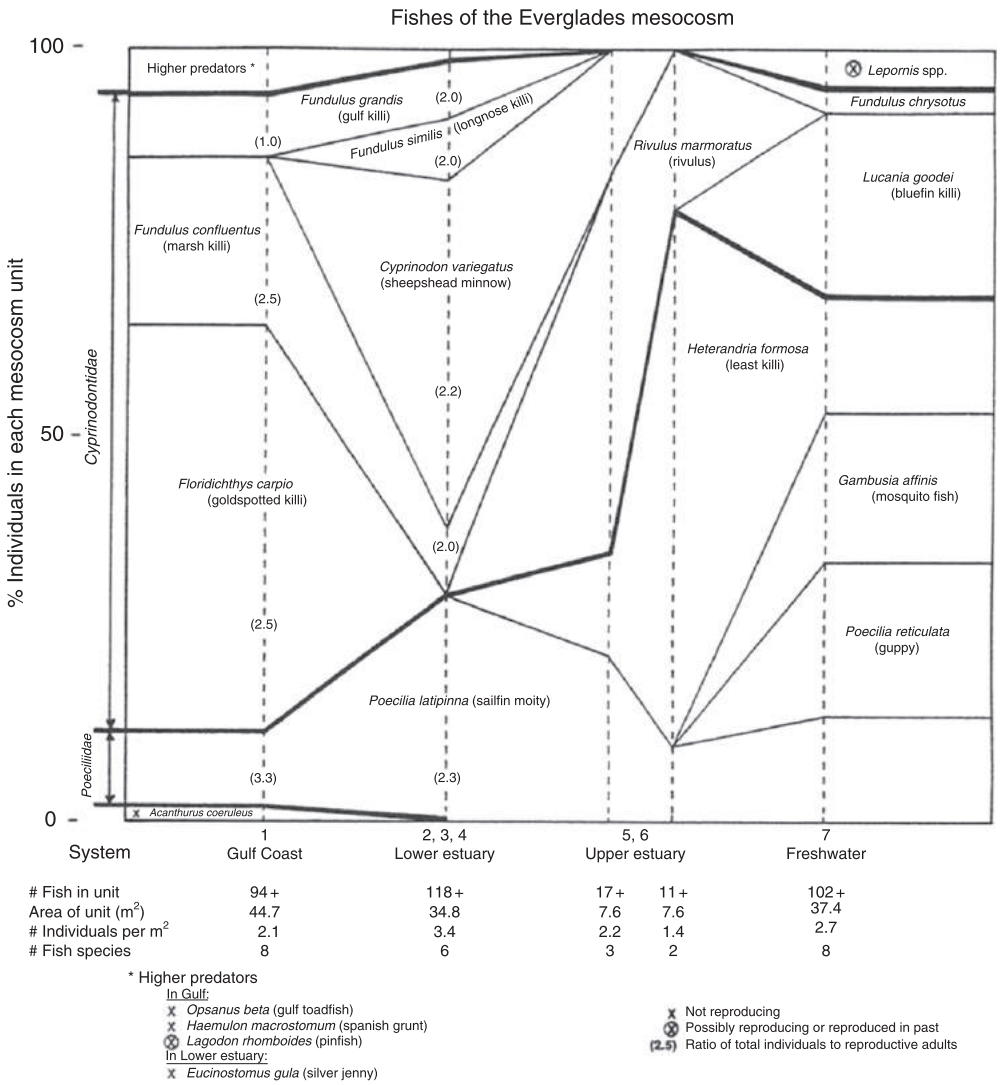


Fig. 14 Distribution of fish (% of total) in Florida Everglades mesocosm.

closure of gas cycles was part of the system design, which was tested with a prototype module of 11 000 ft^3 (312 m^3) from 1988 to 1990.

A number of ecologists were consulted for the creation of the greenhouse's ecosystems which included plots of rainforest, desert, savanna, mangrove estuary, and ocean with coral reef. Thousands of species were added to the greenhouse intentionally and unintentionally (i.e., in ecosystem sub-blocks, as described above), from existing tropical systems as distant as Venezuela and from the local Arizona desert. After construction the ecosystems self-organized and many of the added species went extinct within the system as expected. Success, in terms of replication of the analog ecosystems in nature, has varied among the different model ecosystems, but most have developed and sustained a significant degree of ecological integrity.

Two experiments were conducted in Biosphere II during which humans were enclosed inside the system: the first for 2 years (1991–93) and the second for 6 months (1994). These experiments tested concepts of sustainability at a very basic level since the humans had to rely on the overall greenhouse system for life support function. However, changes in the gas cycles within the greenhouse caused the human experiments to be modified and ultimately terminated. During the first human experiment oxygen concentration in the atmosphere decreased dramatically because high rates of soil respiration released more carbon dioxide than was taken up in photosynthesis; some of the carbon dioxide was absorbed as carbonates in the concrete of the greenhouse foundation. Oxygen had to be pumped into the system to maintain the humans so that the 2-year test could be completed. During the second human experiment, buildup of noxious concentrations of nitrous oxide in the atmosphere from microbial metabolism caused the experiment to be shut down ahead of the planned schedule.

At least two of the basic principles of ecosystem modeling discussed in the introduction were violated in this system. Incoming light was greatly reduced, due to the glass and significant support structure, resulting in insufficient photosynthesis and primary productive to balance respiration. This could have been offset by introducing a subset of highly efficient photosynthesis (such as

provided by an ATS), using artificial lighting; indeed, some ATS systems were used, but only as a minor element of control on the ocean system. Monitored exchange with the external environment could also have been employed. Also, the concrete as an atmospheric reactant should have been sealed with a nonreactive material, such as glass or plastic.

Much controversy developed during these human experiments. Colombia University took over management of the system from 1996 to 2003. During this time period the research program changed from human enclosure experiments to work on global climate change.

See also: Ecosystems: Lagoons; Mangrove Wetlands; Tropical Seasonal Forest. General Ecology: Edaphic Factor

Further Reading

Adey, W., Finn, M., Kangas, P., *et al*., 1996. A Florida Everglades mesocosm – model veracity after four years of self-organization. Ecological Engineering 6, 171–224.
Adey, W., Loveland, K., 2007. Dynamic Aquaria: Building and Restoring Living Ecosytems, 3rd edn. San Diego: Elsevier/Academic Press, 505pp.
Kangas, P., 2004. Ecological Engineering: Principles and Practice. Boca Raton: Lewis Publishers, CRC Press, 452pp.
Körner, C., Arnone III, J., 1992. Responses to elevated carbon dioxide in artificial tropical ecosystems. Science 257, 1672–1675.
Marino, B.D.V., Odum, H.T., 1999. Biosphere 2: Research Past and Present. 358pp Amsterdam: Elsevier, (Also Special Issue of *Ecological Engineering* 13: 3–14).
Osmund, B., Aranyev, G., Berry, J., *et al*., 2004. Changing the way we think about global change research: Scaling up in experimental ecosystem science. Global Change Biology 10, 393–407.
Petersen, J., Kemp, W.M., Bartleson, R., *et al*., 2003. Multi-scale experiments in coastal ecology: Improving realism and advancing theory. Bioscience 53, 1181–1197.
Small, A., Adey, W., 2001. Reef corals, zooxanthellae and free-living algae: A microcosm study that demonstrates synergy between calcification and primary production. Ecological Engineering 16, 443–457.
Walter, A., Carmen Lambrecht, S., 2004. Biosphere 2, center as a unique tool for environmental studies. Journal of Environmental Monitoring 6, 267–277.

Lagoons

G Harris, University of Tasmania, Hobart, TAS, Australia

Background

Coastal lagoons are estuarine basins where freshwater inflows are trapped behind coastal dune systems, sand spits, or barrier islands which impede exchange with the ocean. They are most frequent in regions where freshwater inflows to the coast are small or seasonal, so that exchange with the ocean may not occur for months or years at a time. Many occupy shallow drowned valleys formed when the sea level was lower during the last ice age and subsequently flooded by postglacial sea level rise. The tidal range is usually small. Accordingly, coastal lagoons are frequently found in warm temperate, dry subtropical, or Mediterranean regions along moderately sheltered coasts. Lagoons are infrequent in wetter temperate and tropical regions where freshwater inflows are sufficient to scour out river mouths and keep them open. Here estuaries are dominated by salt marshes in temperate and mangroves in tropical climes. A particularly good example is the series of coastal habitats on the southern and eastern coastline of Australia which change from open temperate estuaries and salt marshes in the wetter southern regions of Tasmania, through a series of coastal lagoons of varying sizes and ecologies along the south and east coasts, to open subtropical and tropical estuaries, reefs, and mangroves in the warmer and wetter north. A similar, although inverted, sequence can be seen running south along the east coasts of Canada, and the northeastern, central, and southeastern coasts of the USA. The resulting lagoons have varying water residence times, depending on volume, climate, freshwater inflow volumes, and the tidal prism.

Some lagoons are predominantly freshwater or brackish, while others are predominantly marine; so the dominant organisms in coastal lagoons reflect the balance of freshwater and marine influences. All are influenced by the local biogeography. Thus, the dominant species in Northern Hemisphere lagoons are quite different from those in their Southern Hemisphere equivalents. Different coastal regions of the globe differ in their biodiversity; for example, the endemic biodiversity of seagrasses is very high in Australian waters. Nevertheless, two points are worthy of note. First, there is great functional similarity between systems despite differing in the actual species involved. Second, human activity is quickly moving species around the world so that there are large numbers of what might be called `feral' introduced species in coastal waters close to ports and large cities.

Coastal lagoons are ecologically diverse and provide habitats for many birds, fish, and plants. The interactions between the species in estuaries and coastal lagoons produce valuable ecosystem services. Indeed, the value of ecosystem services calculated for such systems by Costanza *et al*. was the highest of any ecosystem studied. Lagoons are also esthetically pleasing and desirable places to live, providing harbors, fertile catchments, and ocean access for cities and towns; thus, they have long been the sites of rapid urban and industrial development. Habitat change and other threats to lagoons now compromise these valuable services. All around the world they are threatened by land-use change in their catchments, urbanization, agriculture, fisheries, transport, tourism, climate change, and sea level rise. Coastal waters and lagoons are therefore definitive examples of the problems of multiple use management. Rapid population growth in coastal areas is common in many western countries (particularly the common `sea-change' phenomenon, in which there is a trend toward rapid population growth along coasts), so the threats and challenges are increasing rapidly. Climate change and sea level rise are also becoming issues to be dealt with. In tropical and subtropical regions there is both evidence of rapid coastal habitat loss and population growth as well as an increased frequency of severe hurricanes. Modified systems impacted by severe hurricanes and tsunamis appear to be more fragile in the face of extreme events and certainly do not degrade gracefully.

Research and the management of coastal systems require a synthesis of social, economic, and ecological disciplines. Around the world there are a number of major research and management programs which aim to apply ecosystem knowledge to the effective management of coastal resources. Current examples include work in Chesapeake Bay and the Comprehensive Everglades Restoration Plan in the USA. In Italy the lagoon of Venice is a classic example. In Australia major programs have been undertaken in coastal embayments and lagoons in Adelaide (Gulf of St. Vincent), Brisbane (Moreton Bay), and Melbourne (Port Phillip Bay). (For details on these programs and useful links, see www.chesapeakebay.net, www.evergladesplan.org, and www.healthywaterways.org.) Land-use change (both urbanization and agriculture) in catchments, together with the use of coastal lagoons for transport and tourism, has led to a combination of changes in physical structures (both dredging and construction of seawalls and other barriers), altered hydrology and tidal exchanges, increased nutrient loads, and inputs of toxicants. The resulting symptoms of environmental degradation include algal blooms (which may be toxic), loss of biodiversity, and ecological integrity (including the loss of seagrasses and other important functional groups), anoxia in bottom waters, loss of important biogeochemical functions (denitrification efficiency), and the disturbances caused by introduced, `feral' species from ships and ballast water.

Inputs – Catchment Loads

Land-use change in catchments changes the hydrology of rivers and streams and increases nutrient loads to lagoons. Rivers draining clear catchments, or those with extensive urbanization, show `flashier' flow patterns with water levels rising and falling

quickly after rainfall. The hydrological balance and water residence times of the lagoons are altered as a result. While nutrient loads are generally proportional to catchment area (Fig. 1), loads from cleared agricultural or urban catchments are higher than those from forested catchments, the nutrient loads being proportional to the amount of cleared land or the human population in the catchment. Carbon, nitrogen, and phosphorus loads all increase; C loads from wastewaters may lead to biochemical oxygen demands (BODs) and anoxia, while increased N and P loads stimulate algal blooms and the growth of epiphytes in seagrasses. A further problem is the fact that forested catchments tend to export organic forms of N and P (which are less biologically active in receiving waters), whereas cleared and developed catchments tend to export biologically available inorganic forms of N and P. Thus, both nutrient loads and the availability of those loads increase when catchments are cleared and developed.

N is in many cases (particularly in warmer coastal waters) the key limiting element in lagoons because of high denitrification efficiencies in sediments and long water residence times in summer. In temperate waters N and P may be co-limiting or the limitation may vary seasonally and on an event basis. Overall the climate regime, geomorphology, and biogeochemistry of coastal lagoons seem to lead to extensive N limitation and denitrification is an important process which determines many ecological outcomes. The effect of land-use change on N loads is therefore a key area of concern. A considerable amount of work has been done on the export of N from catchments around the world. Catchments tend to retain on average about 25% of the N applied to them and export about 75%. There are both latitudinal and seasonal factors which affect this figure. Catchment exports on the eastern coast of North America show an effect of latitude, with warmer, southern catchments with perennial vegetation exporting about 10% of applied N and more northerly catchments with seasonal vegetation growth exporting as much as 40% of applied N, particularly in winter. P exports tend to come primarily from sewage and other wastewater discharges, and also from erosion and agricultural runoff. Catchment loads show evidence of self-organized pattern and process in catchments – nutrient loads and stoichiometries change over time at all scales and the distribution of inflowing nutrients may be fractal.

Fates and Effects – Physics and Mixing

Water movement and mixing are driven by the effects of wind and tide on coastal lagoons. The basic hydrodynamics of coastal systems are well represented by physics-based simulation models of various kinds. A number of two- and three-dimensional (2D and 3D) models now exist (both research tools and commercially available products) which can adequately represent wind-induced wave patterns and currents, tidal exchanges and circulation, and changes in surface elevations due to tides and winds. (For an introduction to a variety of models, see www.estuary-guide.net/toolbox or www.smig.usgs.gov; models by Delft Hydraulics at www.wldelft.nl and DHI www.dhigroup.com/. Input data required are basic meteorological data: wind speed and direction, plus solar insolation, and a detailed knowledge of the morphometry and bathymetry of the lagoon in question. Based on the conservation of mass and momentum and various turbulence closure schemes, it is possible to adequately model and predict both velocity fields and turbulent diffusion in the water column. Calibration and validation data are obtained from *in situ* current meters and pressure sensors. Bottom stress, sediment resuspension, and wave-induced erosion can also be represented. It is thus possible to model the effects of various climate and engineering scenarios, everything from sea level rise to construction projects of various kinds. These models are widely used to develop environmental impact statements (EIS) for major projects and to manage major dredging projects around the world. Only some of these models are capable of long-term predictions of water balance and

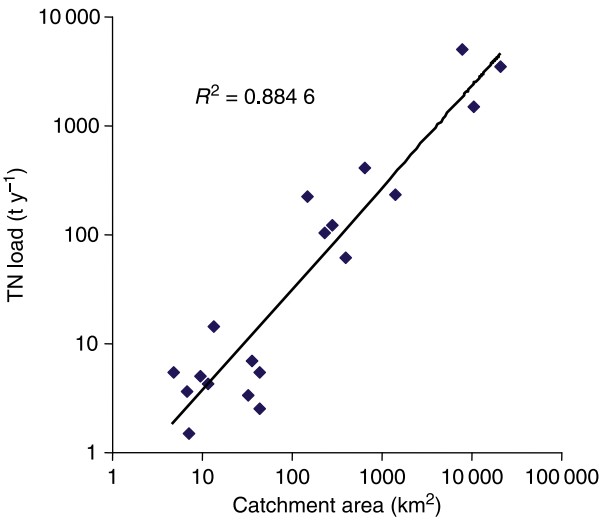


Fig. 1 The empirical relationship between catchment area (km^2) and the total nitrogen load (tonnes per year) to their associated coastal lagoons. Data from the catchments of 19 coastal lagoons on the east coast of Australia. Details of data sources are given in Harris, G.P., 1999. Comparison of the biogeochemistry of lakes and estuaries: Ecosystem processes, functional groups, hysteresis effects and interactions between macro- and microbiology. Marine and Freshwater Research 50, 791–811.

of water residence times. Such predictions require careful analysis of long-term meteorological records and good predictive models of inflows and evaporation. Nevertheless such models also exist.

Fates and Effects – Ecological Impacts and Prediction

Given the nature of the threats, the value of ecosystem services delivered, and the importance of ecosystem management, there have been many studies of ecosystems in coastal lagoons. As noted above some of the ecosystem studies have been in the form of major multidisciplinary programs. The knowledge obtained has then frequently been encapsulated in various kinds of predictive ecological models which attempt to provide answers to `what if' questions from environmental managers and engineers seeking to implement catchment works or reductions in wastewater discharges. The ecological models are driven by the hydrodynamic models described above – the physical setting provides the basic context for the ecological response. In many cases the knowledge has also been built into a variety of EIS and risk assessments which attempt to judge the possible detrimental effects of land-use change, port construction, harbor dredging, and other engineering developments in urban and industrial areas.

Empirical Knowledge and Models

Despite the pandemonium of interactions between species in coastal marine systems (or perhaps because of it), there are some high-level empirical relationships which can be used for diagnosis and management. Much as Vollenwieder discovered in lakes there are some predictable high-level properties of coastal marine systems. For example, the total algal biomass (as chlorophyll *a*) responds to N loads just as lakes respond to P loads. This is further evidence of the importance of N as a limiting element in marine systems, and for the key of P as the limiting element in freshwater systems. The differing biogeochemistry of marine and freshwater ecosystems is explicable on the basis of the evolutionary history and geochemistry of the two systems. The existence of a relationship between N and algal biomass is evidence for a kind of `envelope dynamics' of these diverse systems. N does not limit growth rates of the plankton so much as the overall biomass. As a result of high growth rates, grazing, and rapid nutrient regeneration in surface waters, the total community biomass reaches an upper limit set by the overall rate of supply of N. This is a form of `extremal principle' of these pelagic ecosystems which indicates that with sufficient biodiversity then an upper limit to maximum nutrient use efficiencies can be reached. A similar model of high-level ecosystem properties has been developed in which some fundamental physiological properties of phytoplankton (the slope of the P vs. I curve at low light and the maximum photosynthetic rate) are used to develop a production model based on biomass, photosynthetic properties, and incident light. This amounts to saying that even in shallow coastal systems it is possible to get some reasonable empirical predictions of the physiological (photosynthetic parameters and nutrient uptake efficiencies) and ecosystem responses to some driving forces (nutrient loads and incident light).

A second form of empirical determinant of system function is set by the stoichiometry and biogeochemistry of these systems. The characteristic elemental ratios in the key organisms (algae, grazers, bacteria, macrophytes) and the ratios of elemental turnover set limits on the overall system performance. The predominant element ration in pelagic marine organisms is the Redfield ratio (106C:15N:1P). This aspect of the biogeochemistry of coastal lagoons has been used in a global comparison of the biogeochemistry of these systems by the IGBP LOICZ program. Knowing the loading rates of major nutrients, the concentrations of nutrients in the water column, and the rates of tidal exchange allows simple mass balance models of C, N, and P to be constructed. The salt and water budget of these systems can be used to obtain bulk hydrological fluxes. Making stoichiometric assumptions via the Redfield ratio about fluxes of C, N, and P (as well as oxygen) in the plankton and across the sediment interface allows estimates to be made of the overall autotrophic–heterotrophic balance of the system as well as nitrogen fixation and denitrification rates (essentially by estimating the `missing N' based on the C, N, and P stoichiometry). These techniques have made it possible to do global comparisons of the biogeochemistry of lagoons around the world and to examine the effects of inflows, tidal exchanges, and latitude or climate. This has been a major contribution to the knowledge of the ways in which major elements are processed and transported from the land to the ocean through the coastal zone.

The overall impression is that pristine lagoons (loaded by largely organic forms of C, N, and P) are mostly net heterotrophic and strong sinks for N through denitrification. More eutrophic systems with higher N and P loads (and more of those in inorganic forms) tend to be net autotrophic and, if dominated by cyanobacterial blooms, net N fixing systems. Decomposition of these blooms may be sufficiently rapid to cause anoxia in bottom waters and lead to the cessation of denitrification and the export of N (as ammonia) on the falling tide. Warm temperate and subtropical lagoons – with low hydrological and nutrient loads – seem to have higher denitrification efficiencies than temperate systems. They are often heterotrophic and strongly N limited systems. An extreme is Port Phillip Bay in Melbourne which has low freshwater inflows, high evaporation, a long water residence time (*c.* 1 year), high denitrification efficiency (60–80%) and is so N limited that it imports N from the coastal ocean on the rising tide. Temperate lagoons and estuaries have higher freshwater and nutrient inflows, are more eutrophic (autotrophic), and are exporters of N. Temperate systems are therefore more likely to show occasional P limitation. Overall, the cycling of the major elements is driven by the stoichiometry of the major functional groups of organisms.

Thus in biodiverse ecosystems it is possible to obtain some high-level state predictors from a knowledge of key drivers and the basic physiology and stoichiometry of the dominant organisms. The predictions so produced are not perfect but they do capture a large fraction of the behavior of these systems. At this level these models can be used for the management of nutrient loads to coastal lagoons.

Detailed Simulation Models of Ecosystems, Functional Groups, and Major Species

Many of the questions that are asked of ecologists studying coastal systems are of a more detailed nature and relate to loss or recovery of major species, functions, or functional groups – ecosystem services and assets if you like. Examples would be dominant algal groups, seagrasses, macroalgae, denitrification rates, benthic biodiversity, fish recruitment, etc. At this level a large number of dynamical ecological simulation models of shallow marine systems have been constructed. There is much more uncertainty in the ecological models than there is in the physical models. Much of the required ecological detail is unknown, key parameters can be ill-defined, the data are usually sparse in space and time, and the computational resources are not adequate to the task of a complete simulation of the entire system. Ecological models are therefore abstractions which attempt to represent the major ecological features and functions of the greatest relevance to the task at hand. Nevertheless, 30 years of research in lagoons and coastal systems around the world have uncovered a number of major functional groups and ecosystem services which, when coupled together in models, give some guide as to the overall ecological responses.

The generic models of coastal systems use two basic functional components. A nutrient, phytoplankton, zooplankton (NPZ) model for the water column, and a benthic model incorporating the necessary functional groups – macroalgae, zoo- and phyto-benthos, seagrasses – with the groups chosen to represent the particular system of interest. All functional groups are represented by their basic physiologies and stoichiometries and the interconnections (grazing, trophic closure, decomposition, and denitrification rates) are represented by established relationships. The NPZ models adequately predict the average chlorophyll of lagoons and, when coupled with 3D physical models, can give predictions of the spatial distribution of algal biomass in response to climate and catchment drivers. For reasons which will become clear below, these models only predict average biomass levels and cannot predict all the dynamics of the various trophic levels. The coupling between the plankton and the benthos in lagoons is nonlinear and results in some strongly nonlinear responses of the overall system to changes in nutrient loads. Basically, there is competition between the plankton and the benthos for light and nutrients which can drive switches in system state. Thus, lagoons, much like shallow lakes, may show state switches between clear, seagrass-dominated states and turbid, plankton-dominated states.

The major driver of the state switches is the high denitrification efficiencies exhibited by the diverse phyto- and zoobenthos in lagoons with strong marine influences. As long as there is sufficient oxygen in bottom waters, diverse zoobenthos burrow and churn over the sediments causing extensive bioturbation and 3D structure in the sediments. Clams, prawns, polychaete worms, crabs, and other invertebrates set up a complex system of burrows and ventilate the sediments through feeding currents and respiratory activity. Given sufficient light at the sediment surface the phytobenthos (particularly diatoms, the microphytobenthos, MPB) photosynthesize rapidly and set up strong gradient of oxygen in the top few millimeters of the sediment. These gradients, together with the strong 3D microstructure of the sediments set up by the zoobenthos, favor the co-occurrence of adjacent oxic and anoxic microzones which are required for efficient denitrification. N taken up by the plankton sinks is actively denitrified by the sediment system. In marine systems the abundance of sulfate in seawater ensures that P is not strongly sequestered by the sediments. Thus, the basis of the LOICZ models lies in the efficiency of denitrification of N in sediments and the more or less conservative behavior of P in these systems. These ecosystem services are supported by the high biodiversity of the coastal marine benthos.

In lagoons with higher nutrient loads, the entire ecosystem may switch to an alternative state. Increased N loads stimulate the growth of plankton in the water column and shade off the MPB. The increased planktonic production sinks to the bottom depleting oxygen and reducing the diversity of zoobenthos, restricting the community structure to those species resistant to low oxygen concentrations. Active decomposition in anaerobic sediments together with reduced bioturbation leads to the cessation of denitrification and the release of ammonia from the sediments. So instead of actively denitrifying and eliminating the N load, the system becomes internally fertilized and algal production rises further. This is analogous to the internal fertilization of eutrophic lakes through the release of P from anoxic sediments. In both cases the switch is caused by a change in redox conditions and the change in performance of suites of microbial populations. Once switched to a more eutrophic state (algal bloom dominated), these lagoons do not easily revert to their clear and macrophyte dominated state. Loads must be strongly reduced to get them to switch back – something which may not be possible if the catchment has been modified by urban or agricultural development. There is thus evidence for strong hysteresis in the response of these ecosystems to various impacts.

The overall biodiversity and nutrient cycling performance of coastal lagoons therefore depends on the relative influences of marine and freshwaters, the differing biodiversity of marine and freshwater ecosystems, the relative C, N, and P loads to the plankton and the benthos, and on seasonality, latitude, and climate drivers. Nevertheless, at least the broad features of their behavior can be explained and predicted on the basis of sediment geochemistry, and the stoichiometry and physiology of the major functional groups in these ecosystems. Empirical work on a number of lagoons up the east coast of Australia allowed Scanes *et al.* to effectively determine the response of `titrating' these systems with nutrients. As the N load to the lagoons was increased, seagrasses were lost and algal blooms were stimulated. Even at a crude level of visual assessments it was possible to rank these systems in order of loading and to show that the pattern of response was entirely similar to that predicted by the models (**Fig. 2**). Thus, despite difference in biogeochemistry and biodiversity, shallow lakes and coastal lagoons have broadly similar response to increased nutrient loads and other forms of human impact. Even broad indicators of system state reveal consistent patterns of change.

So oligotrophic lagoons with a Mediterranean climate (warm temperatures in summer and long water residence times) and strong marine influences can be strong sinks for N, whereas cooler, temperate lagoons and estuaries with larger freshwater inflows and higher productivity may export N and be frequently P limited. As the LOICZ program intended, we have managed a broad understanding of the ways in which the coastal zone influences the transport of major elements from land to ocean.

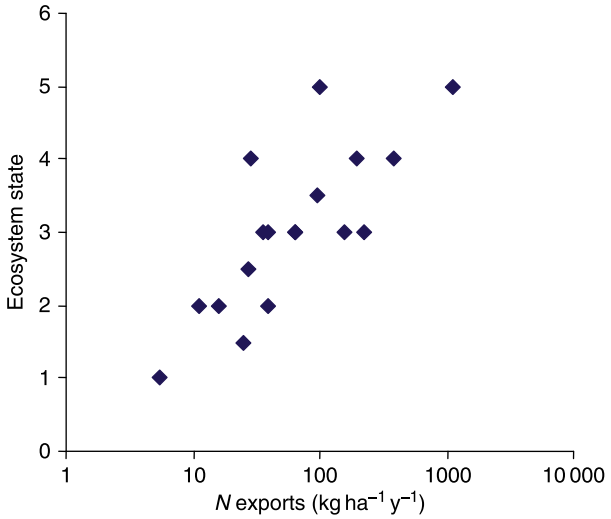


Fig. 2 The empirical `ecosystem titration' relationship between catchment N exports and the resulting ecosystem state in 17 coastal lagoons on the east coast of Australia. Ecosystem state is defined as 1, pristine; 3, showing marked seagrass loss and the growth of macrophytic algae; 5–6, dominated by nuisance algal blooms (some of which may be toxic). Data from personal observations and reworked from Scanes, P., Coade, G., Large, D., Roach, T., 1998. Developing criteria for acceptable loads of nutrients from catchments. In: Proceedings of the Coastal Nutrients Workshop, Sydney (October 1997). Artarmon, Sydney: Australian Water and Wastewater Association, pp. 89–99. Artarmon, Sydney: Australian Water and Wastewater Association.

Nonequilibrium Dynamics

If more detailed descriptions and predictions are required (e.g., the diversity and abundance of individual species and other specific ecosystem services and assets), then the predictive ability is less. One of the reasons for this is the fact, alluded to above, that these are nonequilibrium systems which respond to individual events (storms and engineering works) over long time periods. The elimination and invasion of species may take decades and the responses of freshwater lagoons, for example, to salt incursions may also take decades. A particularly good example is Lake Wellington in the Gippsland Lakes system in Victoria, Australia. The entire system is slowly responding to the ingress of salt made possible by the opening of the lagoon system mouth (Lakes Entrance) in 1883. Lake Wellington, the lake farthest inland, remained fresh until after the 1967 drought when a combination of high N and P loads from agriculture, the extraction of water from the inflowing La Trobe River for power station cooling and irrigation, and the incursion of salt killed all the freshwater macrophytes in the Lake. In a few years the lake switched from its previous clear and macrophyte dominated state to being turbid and dominated by toxic algal blooms. It does not appear to be possible to switch it back.

The response of these lagoon systems to climate and other perturbations is nonlinear and complex because of the interactions between the major functional groups and because the timescales of response of the major groups differ strongly. Phytoplankton may respond to changes in loads and water residence times in a matter of days, whereas seagrasses take decades or longer to recover. By perturbing a simple coupled plankton-benthos model with storm events and `spiked' N loads, Webster and Harris showed that the threshold load for the elimination of seagrasses could be altered considerably depending on the characteristics of the input loads. So the response of the system was a function of the overall load and the frequency and magnitude of events. Climate change and catchment development both alter the overall C, N, and P load to lagoons as well as the characteristics of that load, so that ecological responses by lagoons are highly complex and change over time depending on a variety of modifications and management actions. Consequently, lagoons are always responding to the last storm or intervention and the abundance of key species drifts to and fro over time as the entire plankton–sediment system responds.

The picture is made more complex by the evidence for strong trophic cascades in marine as well as freshwater systems. Coastal ecosystems are frequently over-fished; larger predators and grazers are removed by human hand. Removal of the `charismatic megafauna' of coastal systems, together with beds of shellfish and other edible species, has changed the ecology of many lagoons and estuaries. Coastal ecosystems around the world have also been strongly modified by the removal of natural physical structures (mangroves and reefs) which confer resilience in the face of extreme events. We have removed both larger fish and benthic filter feeders from many systems compromising function and the ability to respond to changes in catchment loads. Overall there has been a consistent simplification of both physical and ecosystem structures (removal of reefs and macrobiota, simplification of food chains, etc.) and a trend toward more eutrophic (nutrient rich) and simplified systems dominated by microbiota, especially algae and bacteria. We know less about the response of ecosystems to changes in the `top down' trophic structure than we do about the responses to `bottom up' catchment drivers; nevertheless, there is good evidence for similar nonlinearities and state switches in response. A nonequilibrium view of coastal lagoons changes the way we look at them. Overall there is a need to pay attention to

the `precariousness' of these systems and manage them adaptively for resilience and response to natural and anthropogenic impacts. Despite being over-fished and highly modified, there is still a need for the ecosystem services they produce.

Emerging Concepts – Multifractal Distributions of Species and Biomass

The underlying complexity of interactions and species distributions is displayed when detailed (high-frequency) observations are made of the spatial and temporal distributions of biomass and species. There is now much evidence to show that the underlying distribution of the plankton and the MPB are fractal or multifractal. Similarly, high-frequency observations in catchments show similar multifractal and even paradoxical properties of hydrological and nutrient loads. So underlying all the generalizations discussed above lies a pattern of behavior which gives strong evidence of self-generated complexity which arises from the pandemonium of interactions between species and functional groups. Indeed, we can probably argue that the kinds of general, system level, responses described above would not occur if it were not for the underlying complexity. While making high-level statements about ecosystem behavior possible, these small-scale, multifractal properties (and the possibilities created by emergence) cause problems when we wish to make predictions at the meso-scale level of dominant species and functional groups. Because of the work that has been done across the levels of organization, coastal lagoons are very good examples of a new kind of ecology – an ecology of resilience and change, rather than an ecology and equilibrium and stasis.

One fundamental problem that these new insights reveal is that most of the data we presently use for the analysis of coastal lagoons are collected too infrequently to be useful for anything other than the analysis of broad trends. Data collected weekly or less frequently are strongly aliased and cannot reveal the true scales of pattern and process. It is just possible to analyze daily data for new insights and processes but high-frequency data – collected at scales of hours and minutes – reveal a wealth of new information. Aliased data combined with frequentist statistical techniques that `control error' actually remove information from multifractally distributed data and raise the possibility of serious type I and II errors in ecological interpretations. Most importantly, there is information contained in the time series of multivariate data that can be collected from coastal systems. Most analyses of ecological data from ecological systems use univariate data and because of the infrequent data collection schedules – including gaps and irregular time intervals – time series analyses are not possible.

We are just beginning to find new technologies and techniques to study the high-frequency multivariate behavior of these systems using moorings and other *in situ* instruments. New electrode technologies make on-line access to data possible and throw up new possibilities for new kinds of observations of system state. We are beginning to realize that in addition to the `top down' causation of climate and trophic interactions, there is also a `bottom up' driver of complexity and the strong possibility of the emergence of high-level properties from the interactions between individuals. New forms of statistical analyses display information in time series of complex and emergent systems. This emerging understanding of complexity and emergent properties changes the ways in which we should approach EIS and risk assessments. We now know that interactions and self-generated complexity, together with hysteresis effects at the system level, can cause surprising things to happen as a result of anthropogenic change. Coastal lagoons are now classic examples of this. That means that risk assessments and EIS cannot look at impacts and changes in isolation; somehow we must develop integrated risk assessment tools that examine the interactive and synergistic effects of human impacts on coastal ecosystems. A further level of complexity is contained in the similar complex and emergent properties of the interactions between agents in the coupled environmental and socioeconomic (ESE) system in which all coastal lagoons are set. Multiple use management decisions are set in a complex web of ESE interactions across scales. Decisions made about industrial and engineering developments for financial capital reasons influence both social capital and ecological (natural capital) outcomes. Feedbacks ensure that this is also a highly nonlinear set of interactions. What we do know is that the prevalent practices of coastal management and exploitation are not resilient in the face of extreme events and that they do not degrade `gracefully' when impacted by hurricanes and tsunamis. New management practices will be required.

See also: Aquatic Ecology: Equilibrium Concept in Phytoplankton Communities. Evolutionary Ecology: Red Queen Dynamics

Further Reading

Adger, W.N., Hughes, T.P., Folke, C., Carpenter, S.R., Rockström, J., 2005. Socio-ecological resilience to coastal disasters. Science 309, 1036–1039.
Aksnes, D.L., 1995. Ecological modelling in coastal waters: Towards predictive physical–chemical–biological simulation models. Ophelia 41, 5–35.
Berelson, W.M., Townsend, T., Heggie, D., *et al*., 1999. Modelling bio-irrigation rates in the sediments of Port Phillip Bay. Marine and Freshwater Research 50, 573–579.
Brawley, J.W., Brush, M.J., Kremer, J.N., Nixon, S.W., 2003. Potential applications of an empirical phytoplankton production model to shallow water ecosystems. Ecological Modelling 160, 55–61.
Costanza, R., d'Arge, R., de Groot, R., *et al*., 1998. The value of ecosystem services: Putting the issues in perspective. Ecological Economics 25, 67–72.
Fasham, M.J.R., Ducklow, H.W., Mckelvie, S.M., 1990. A nitrogen-based model of plankton dynamics in the oceanic mixed layer. Journal of Marine Research 48, 591–639.
Flynn, K.J., 2001. A mechanistic model for describing dynamic multi-nutrient, light, temperature interactions in phytoplankton. Journal of Plankton Research 23, 977–997.
Gordon, D.C., Boudreau, P.R., Mann, K.H., *et al*., 1996. LOICZ biogeochemical modelling guidelines. *LOICZ Reports and Studies*, *No. 5.* Texel: LOICZ.
Griffiths, S.P., 2001. Factors influencing fish composition in an Australian intermittently open estuary. Is stability salinity-dependent? Estuarine, Coastal and Shelf Science 52, 739–751.

Harris, G.P., 1999. Comparison of the biogeochemistry of lakes and estuaries: Ecosystem processes, functional groups, hysteresis effects and interactions between macro- and microbiology. Marine and Freshwater Research 50, 791–811.
Harris, G.P., 2001. The biogeochemistry of nitrogen and phosphorus in Australian catchments, rivers and estuaries: Effects of land use and flow regulation and comparisons with global patterns. Marine and Freshwater Research 52, 139–149.
Harris, G.P., 2006. Seeking Sustainability in a World of Complexity. Cambridge: Cambridge University Press.
Harris, G.P., Heathwaite, A.L., 2005. Inadmissible evidence: Knowledge and prediction in land and waterscapes. Journal of Hydrology 304, 3–19.
Hinga, K.R., Jeon, H., Lewis, N.F., 1995. Marine eutrophication review. Part 1: Quantifying the effects of nitrogen enrichment on phytoplankton in coastal ecosystems. Part 2: Bibliography with abstracts. *NOAA Coastal Ocean program, Decision Analysis Series, No 4.* Silver Spring, MD: US Dept of Commerce, NOAA Coastal Ocean Office.
Howarth, R.W., 1998. An assessment of human influences on fluxes of nitrogen from the terrestrial landscape to the estuaries and continental shelves of the North Atlantic Ocean. Nutrient Cycling in Agroecosystems 52, 213–223.
Howarth, R.W., Billen, G., Swaney, D., *et al.*, 1996. Regional nitrogen budgets and the riverine N and P fluxes for the drainages to the North Atlantic Ocean – Natural and human influences. Biogeochemistry 35, 75–139.
Lotze, H.K., Lenihan, H.S., Bourque, B.J., *et al.*, 2006. Depletion, degradation and recovery potential of estuaries and coastal seas. Science 312, 1806–1809.
McComb, A.J., 1995. Eutrophic Shallow Estuaries and Lagoons. Boca Raton: CRC Press.
Mitra, A., 2006. A multi-nutrient model for the description of stoichiometric modulation of predation in micro- and mesozooplankton. Journal of Plankton Research 28, 597–611.
Moll, A., Radach, G., 2003. Review of three-dimensional ecological modelling related to the North Sea shelf system. Part 1: Models and their results. Progress in Oceanography 57, 175–217.
Murray, A.G., Parslow, J.S., 1999. Modelling of nutrient impacts in Port Phillip Bay – A semi-enclosed marine Australian ecosystem. Marine and Freshwater Research 50, 597–611.
Nicholson, G.J., Longmore, A.R., 1999. Causes of observed temporal variability of nutrient fluxes from a southern Australian marine embayment. Marine and Freshwater Research 50, 581–588.
Occhipinti-Ambrogi, A., Savini, D., 2003. Biological invasions as a component of global change in stressed marine ecosystems. Marine Pollution Bulletin 46, 542–551.
Pollard, D.A., 1994. A comparison of fish assemblages and fisheries in intermittently open and permanently open coastal lagoons on the south coast of New South Wales, south-eastern Australia. Estuaries 17, 631–646.
Roy, P.S., Williams, R.J., Jones, A.R., *et al.*, 2001. Structure and function of south-east Australian estuaries. Estuarine, Coastal and Shelf Science 53, 351–384.
Scanes, P., Coade, G., Large, D., Roach, T., 1998. Developing criteria for acceptable loads of nutrients from catchments. In: Proceedings of the Coastal Nutrients Workshop, Sydney (October 1997). Artarmon, Sydney: Australian Water and Wastewater Association, pp. 89–99.
Scheffer, M., 1998. Shallow Lakes. London: Chapman and Hall.
Scheffer, M., Carpenter, S., de Young, B., 2005. Cascading effects of overfishing marine systems. Trends in Ecology and Evolution 20, 579–581.
Seitzinger, S.P., 1987. Nitrogen biogeochemistry in an unpolluted estuary: The importance of benthic denitrification. Marine Ecology – Progress Series 41, 177–186.
Seitzinger, S.P., 1988. Denitrification in freshwater and coastal marine systems: Ecological and geochemical significance. Limnology and Oceanography 33, 702–724.
Seuront, L., Gentilhomme, V., Lagadeuc, Y., 2002. Small-scale nutrient patches in tidally mixed coastal waters. Marine Ecology-Progress Series 232, 29–44.
Seuront, L., Spilmont, N., 2002. Self-organized criticality in intertidal microphytobenthos patterns. Physica A 313, 513–539.
Smith SV and Crossland CJ (1999) Australasian estuarine systems: Carbon, nitrogen and phosphorus fluxes. *LOICZ Reports and Studies, No. 12.* Texel: LOICZ.
Sterner, R.W., Elser, J.J., 2002. Ecological Stoichiometry: The Biology of Elements from Molecules to the Biosphere. Princeton, NJ: Princeton University Press.
Vollenweider, R.A., 1968. In: Scientific fundamentals of the eutrophication of lakes and flowing waters, with particular reference to nitrogen and phosphorus as factors in eutrophication. Paris: OECD, p. 182. *Technical Report DAS/SCI/68.27.*
Walker, D.I., Prince, R.I.T., 1987. Distribution and biogeography of seagrass species on the northwest coast of Australia. Aquatic Botany 29, 19–32.
Walker, S.J., 1999. Coupled hydrodynamic and transport models of Port Phillip Bay, a semi-enclosed bay in south-eastern Australia. Marine and Freshwater Research 50, 469–481.
Webster, I., Harris, G.P., 2004. Anthropogenic impacts on the ecosystems of coastal lagoons: Modelling fundamental biogeochemical processes and management implications. Marine and Freshwater Research 55, 67–78.

Relevant Websites

http://www.chesapeakebay.net - Chesapeake Bay Programme.
http://www.dhigroup.com - DHI.
http://www.evergladesplan.org - Everglades.
http://www.healthywaterways.org - Healthy Waterways.
http://www.estuary-guide.net - Toolbox, The Estuary Guide.
http://www.wldelft.nl - wl delft hydraulics.

Mangrove Wetlands

RR Twilley, Louisiana State University, Baton Rouge, LA, USA

Introduction

Mangroves refer to a unique group of forested wetlands that dominate the intertidal zone of tropical and subtropical coastal landscapes generally between 25° N and 25° S latitude. These tropical forests grow along continental margins between land and the sea across the entire salinity spectrum from nearly freshwater (oligohaline) to marine (euhaline) conditions. The coastal forests also inhabit nearly every type of coastal geomorphic formation from riverine deltas to oceanic reefs – another example of the tremendous `biodiversity' of mangrove ecosystems. Mangroves are trees considered as a group of halophytes with species from 12 genera in eight different families. A total of 36 species has been described from the Indo-West-Pacific area, but fewer than ten species are found in the new world tropics. The term mangroves may best define a specific type of tree, whereas mangrove wetlands refers to whole-plant associations with other community assemblages in the intertidal zone, similar to the term `mangal' introduced by Macnae to refer to swamp ecosystems. In addition, the habitats of tropical estuaries consist of a variety of primary producers and secondary consumers distributed in bays and lagoons that have the intertidal zone dominated by mangrove wetlands. These may be referred to as mangrove-dominated estuaries.

There are numerous reviews and books that describe the ecology and management of mangroves around the world, including references describing techniques to study the ecology of mangrove wetlands.

Ecogeomorphology of Mangroves

The environmental settings of mangroves are a complex behavior of regional climate, tides, river discharge, wind, and oceanographic currents (Fig. 1). There are about 240×10^3 km^2 of mangroves that dominate tropical continental margins from river deltas, lagoons, and estuarine settings to islands in oceanic formations (noncontinental). The landform characteristics of a coastal region together with geophysical processes control the basic patterns in forest structure and growth. These coastal geomorphic settings can be found in a variety of life zones that depend on regional climate and oceanographic processes. Hydroperiod of mangroves resulting from gradients in microtopography and tidal hydrology (Fig. 1) can influence the zonation of mangroves from shoreline to more inland locations forming ecological types of mangrove wetlands. Lugo and Snedaker identified ecological types of mangroves based on topographic location and patterns of inundation at local scales (riverine, fringe, basin, and dwarf; Fig. 1) that Woodroffe summarized into basically three geomorphic types (riverine, fringe, and inland). A combination of ecological types of mangroves can occur within any one of the geomorphic settings occurring at a hierarchy of spatial scales that can be used to classify mangrove wetlands.

Various combinations of geophysical processes and geomorphologic landscapes produce gradients of regulators, resources, and hydroperiod that control mangrove growth (Fig. 2). Regulator gradients include salinity, sulfide, pH, and redox that are non-resource variables that influence mangrove growth. Resource gradients include nutrients, light, space, and other variables that are consumed and contribute to mangrove productivity. The third gradient, hydroperiod, is one of the critical characteristics of wetland landscapes that controls wetland productivity. The interactions of these three gradients have been proposed as a constraint envelope for defining the structure and productivity of mangrove wetlands based on the relative degree of stress conditions (Fig. 2). At low levels of stress for all three environmental gradients (such as low salinity, high nutrients, and intermediate flooding), mangrove wetlands reach their maximum levels of biomass and net ecosystem productivity.

Soil nutrients are not uniformly distributed within mangrove ecosystems, resulting in multiple patterns of nutrient limitation. Along a microtidal gradient in carbonate reef islands, trees were generally N-limited in the fringe zone and P-limited in the interior or scrub zone. Fertilization studies demonstrated that not all ecological processes respond similarly to or are limited by the same nutrient. It is also apparent that mangrove forests growing in other ecogeomorphic settings are also prone to P-limitation associated with different geophysical processes. One of the most critical regulator gradients (Fig. 2) controlling mangrove establishment, seedling survival, growth, height, and zonation is salinity, depending on their ability to balance water and salt. Interspecific differential response of mangrove propagules to salinity occurs at salinities from 45 to 60 g kg^{-1}. The δ^{13}C and δ^{15}N signatures of mangrove leaf tissue can indicate stress conditions such as drought, limited nutrients, and hypersalinity across a variety of environmental settings.

Biodiversity

Mangrove ecosystems support a variety of marine and estuarine food webs involving an extraordinarily large number of animal species and complex heterotrophic microorganism food web. In the New World tropics, extensive surveys of the composition and ecology of mangrove nekton have found 26–114 species of fish. In addition to the marine and estuarine food webs and associated

 doi:10.1016/B978-0-444-63768-0.00346-2

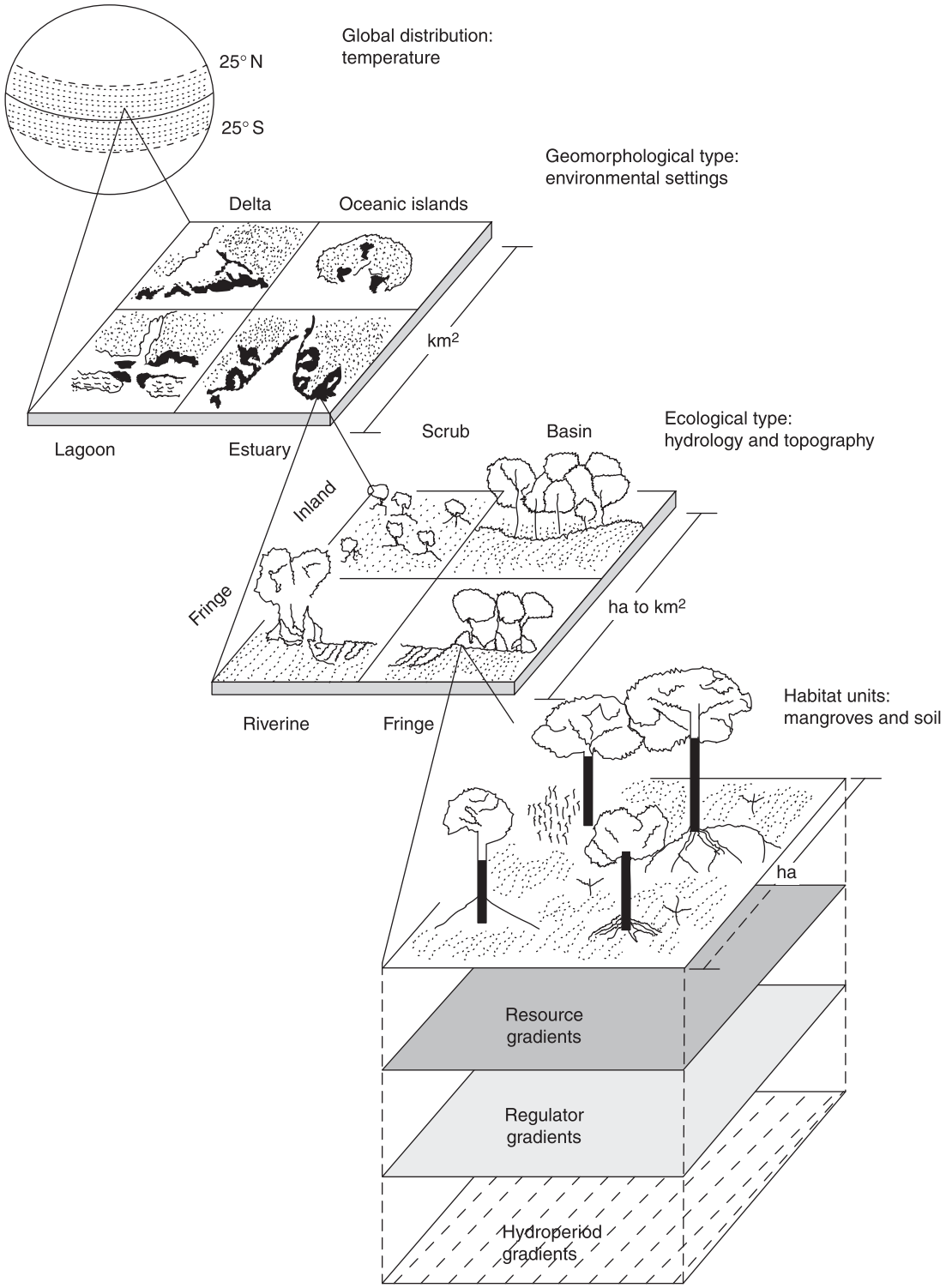


Fig. 1 Hierarchical classification system to describe patterns of mangrove structure and function based on global, geomorphic (regional), and ecological (local) factors that control the concentration of nutrient resources and regulators in soil along gradients from fringe to more interior locations from shore. Modified from Twilley RR, Gottfried RR, Rivera-Monroy VH, Armijos MM, and Bodero A (1998) An approach and preliminary model of integrating ecological and economic constraints of environmental quality in the Guayas River estuary, Ecuador. *Environmental Science and Policy* 1: 271–288 and Twilley RR and Rivera-Monroy VH (2005) Developing performance measures of mangrove wetlands using simulation models of hydrology, nutrient biogeochemistry, and community dynamics. *Journal of Coastal Research* 40: 79–93.

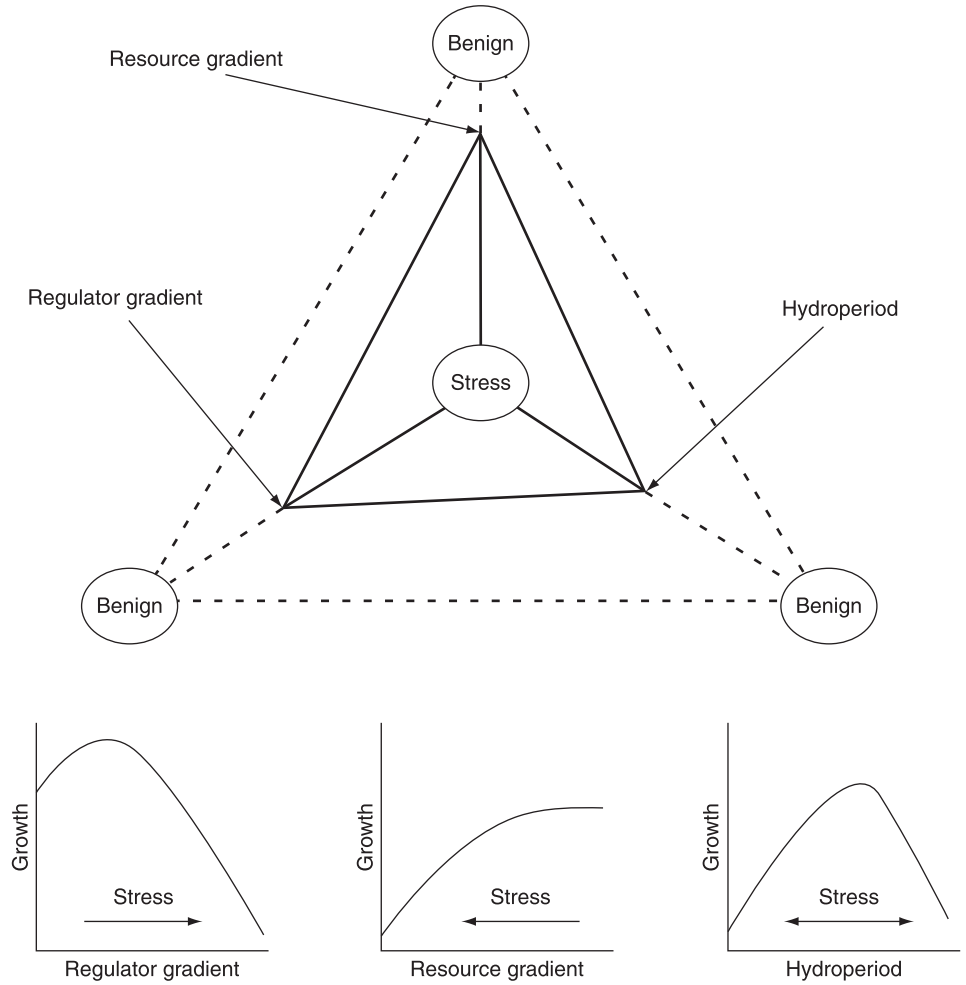


Fig. 2 Interaction of three factors controlling the productivity of coastal wetlands, including regulator gradients, resource gradients, and hydroperiod. The bottom panel defines stress conditions associated with how gradients in each factor control growth of wetland vegetation. From Twilley RR and Rivera-Monroy VH (2005) Developing performance measures of mangrove wetlands using simulation models of hydrology, nutrient biogeochemistry, and community dynamics. *Journal of Coastal Research* 40: 79–93.

species, there are a relatively large number and variety of animals that range from terrestrial insects to birds that live in and/or feed directly on mangrove vegetation. These include sessile organisms (such as oysters and tunicates), arboreal feeders (such as foliovores and frugivores), and ground-level seed predators. Sponges, tunicates, and a variety of other forms of epibionts on prop roots of mangroves are highly diverse, especially along mangrove shorelines with little terrigenous input. Over 200 species of insects have been documented in mangroves in the Florida Keys, similar to the richness of insects and faunal biota observed in other parts of the Caribbean. One of the most published links between mangrove biodiversity and ecosystem function may be the presence of crabs in mangrove wetlands. Crabs can influence forest structure, litter dynamics, and nutrient cycling of mangrove wetlands, suggesting that they are a keystone guild in these forested ecosystems.

Ecosystem Processes

Succession

Succession in mangroves has often been equated with zonation, wherein `pioneer species' would be found in the fringe zones, and zones of vegetation more landward would `recapitulate' the successional sequence toward terrestrial communities. Zonation in mangrove communities has variously been accounted for by a number of biological factors, including salinity tolerance of individual species, seedling dispersal patterns resulting from different sizes of mangrove propagules, differential consumption by grapsid crabs and other consumers, and interspecific competition. Snedaker proposed the establishment of stable monospecific zones wherein each species is best adapted to flourish due to the interaction of physiological tolerances of species with environmental conditions. Geological surveys of the intertidal zone of Tabasco, Mexico, demonstrated that the zonation and structure

of mangrove wetlands are responsive to eustatic changes in sea level, and that mangrove zones can be viewed as steady-state zones migrating toward or away from the sea, depending on its level. Thus, both monospecific and mixed vegetation zones of mangrove wetlands represent steady-state adjustments rather than successional stages. Many models of mangrove succession are based on how gap dynamics influence spatial patches of community dynamics across the landscape.

Productivity and Litter Dynamics

Tree height and aboveground biomass of mangrove wetlands throughout the tropics decrease at higher latitudes, indicating the constraint of climate on forest development in the subtropical climates. In addition, mangrove biomass can vary dramatically within any given latitude, an indication that local effects of regulators, resources, or hydroperiod may significantly limit the potential for forest development at all latitudes. The primary productivity of mangroves is most often evaluated by measuring the rate of litter fall, as recorded for other forested wetlands. Regional rates in litter production in mangroves are a function of water turnover within the forest, and rank among the ecological types is as follows: riverine > fringe > basin > scrub.

The dynamics of mangrove litter, including productivity, decomposition, and export, can determine the coupling of mangroves to the secondary productivity and biogeochemistry of coastal ecosystems. Patterns of leaf-litter turnover have been proposed to vary among ecological types of mangroves with greater litter export in sites with increasing tidal inundation (riverine > fringe > basin). However, several studies in the Old World tropics in higher-energy coastal environments of Australia and Malaysia have emphasized the influence of crabs on the fate of mangrove leaf litter, rather than geophysical processes. In these coastal environments, crabs consume 28–79% of the annual leaf fall. A similar biological factor was observed in the neotropics where the crab *Ucides occidentalis*s in the Guayas River estuary (Ecuador) processed leaf litter at similar rates observed in Old World tropics. Differences in litter turnover rates among mangrove wetlands are a combination of species-specific degradation rates, hydrology (tidal frequency), soil fertility, and biological factors such as crabs.

Nutrient Biogeochemistry

The nutrient biogeochemistry of mangrove wetlands as either a nutrient source or sink depends on the process of material exchange at the interface between mangrove wetlands and the estuary, which is largely controlled by tides (tidal exchange, TE, in Fig. 3). Nutrient exchanges may occur either with coastal waters (TE) or with the atmosphere (atmosphere exchange, AE), depending on whether the nutrient has a gas phase or not (Fig. 3). Substantial amounts of carbon and nitrogen can exchange with the atmosphere, resulting in very complex mechanisms both at the interface with coastal waters and with the atmosphere that influence the mass balance of these nutrients. In addition, there are internal processes, including root uptake (UT), retranslocation (RT) in the canopy, litter fall (LF), regeneration (RG), immobilization (IM), and sedimentation (SD) (Fig. 3). The balance of these nutrient flows will determine the exchanges across the wetland boundary.

There are very few comprehensive budgets of carbon, nitrogen, or phosphorus for mangrove ecosystems. Mangrove sediments have a high potential in the removal of N from surface waters, yet estimates of denitrification have a large range from a low of 0.53 μmol m^{-2} h^{-1}, to 9.7–261 μmol N m^{-2} h^{-1} in mangrove forests receiving effluents from sewage treatment plants. Small amendments of $^{15}NO_3$ followed by direct measures of $^{15}N_2$ production have shown that denitrification accounts for <10% of the applied isotope suggesting that NO_3 is accumulated in the litter via immobilization on the forest floor rather than a sink to the atmosphere. The other nutrient sink in mangrove wetlands is the burial of nitrogen and phosphorus associated with sedimentation. A survey of sedimentation and nutrient accumulation among five sites in south Florida and Mexico indicates patterns associated with the ecological types of mangroves, with rates of about 5.5 g m^{-2} yr^{-1}. This rate is higher than nitrogen loss via denitrification, indicating the significance of burial as nitrogen sink in mangrove ecosystems. Intrasystem nutrient recycling mechanisms in the canopy may be a site of nitrogen conservation in mangroves and, together with leaf longevity, could influence the nitrogen demand of these ecosystems. The significance of this ecological process to the nutrient budget of different mangrove wetlands has not been determined.

Surveys of nitrogen exchange demonstrate some of the principles of determining the function of mangrove wetlands as a nutrient sink. The largest nitrogen flux of nitrogen from sites in Mexico and Australia is export of particulate nitrogen, consistent with organic carbon representing the largest flux from most mangroves (Fig. 3). Compared to other flux studies of mangroves, there seems to be a pattern of net inorganic fluxes into the wetlands and corresponding flux of organic nutrients out. The best summary may be that mangrove wetlands transform the tidal import of inorganic nutrients into organic nutrients that are then exported to coastal waters. Carbon export from mangrove ecosystems ranges from 1.86 to 401 g Cm^{-2} yr^{-1}, with an average rate of about 210 g Cm^{-2} yr^{-1}. Carbon export from mangrove wetlands is nearly double the rate of average carbon export from salt marshes, which may be associated with the more buoyant mangrove leaf litter, higher precipitation in tropical wetlands, and greater tidal amplitude in mangrove systems studied.

Mangrove Food Webs

The function of mangrove wetlands as a source of habitat and food to estuarine-dependent fisheries is one of the most celebrated values of forested wetlands. There are several excellent reviews that describe the secondary productivity of tropical mangrove

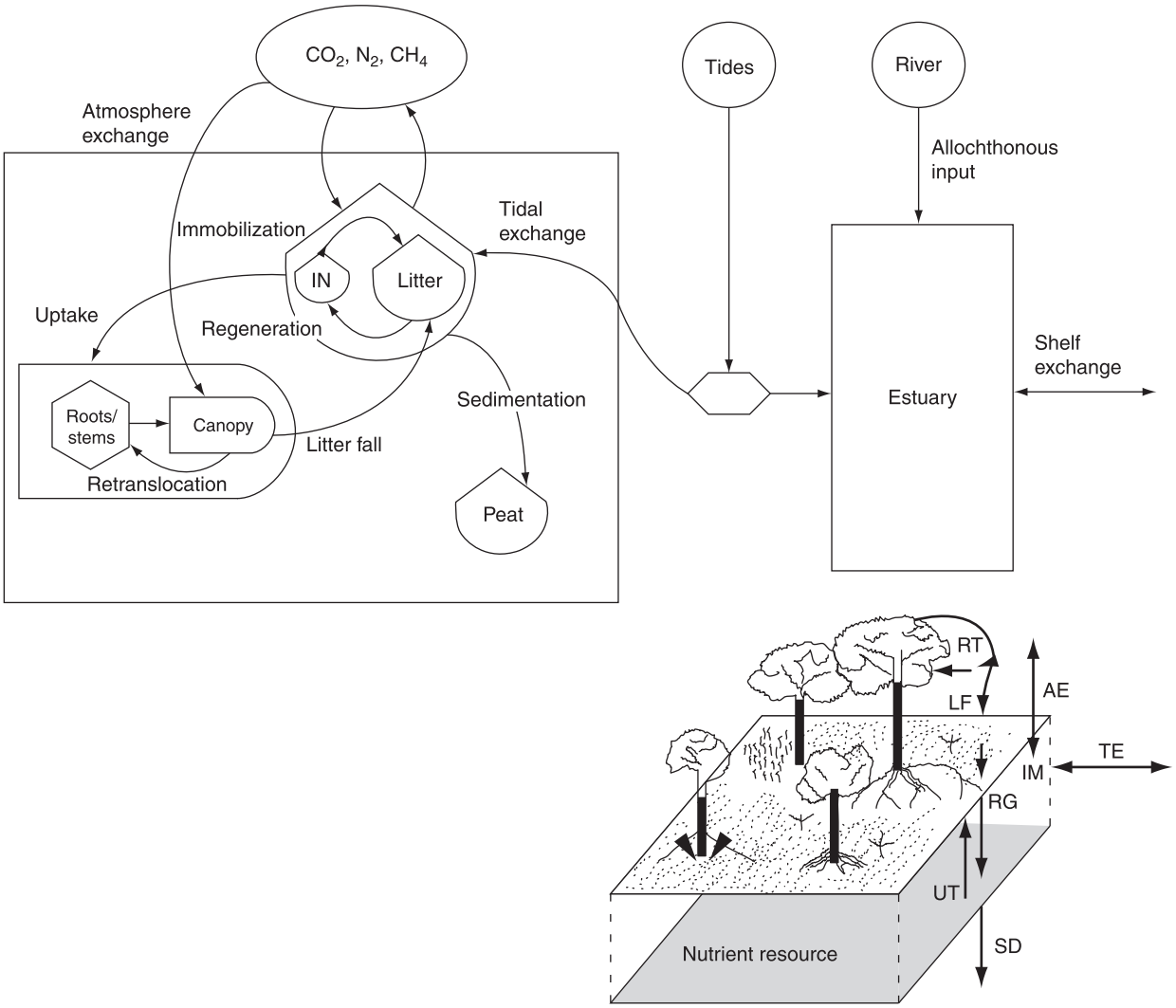


Fig. 3 Upper panel: Schematic of the various fluxes of organic matter and nutrients in a mangrove ecosystem, including exchange with the estuary (IN = inorganic nutrients). Lower panel: A diagram of a mangrove wetland with soil nutrient resources describing the various processes associated with intrasystem cycling and exchange. From Twilley RR (1997) Mangrove wetlands. In: Messina M and Connor W (eds.) *Southern Forested Wetlands: Ecology and Management*, pp. 445-473. Boca Raton, FL: CRC Press.

ecosystems. The original `outwelling hypothesis' of mangroves has been revised from the original paradigms based on comparisons among different mangrove estuaries using natural isotope abundance to trace mangrove organic matter through estuarine food chains. There are seasonal and spatial differences in the amount of mangrove detritus that can be measured in shrimp and fish that inhabit mangrove estuaries. If the distance from the source of mangrove detritus increases, the proportion of carbon in the tissue of shrimp from mangrove detritus decreases as the signal of carbon phytoplankton increases. The seasonal timing of mangrove export of detritus relative to the migration of estuarine-dependent fisheries may also dilute the contribution of mangrove detritus from the food webs among diverse sites. The migratory nature of many of the nekton communities and the seasonal pulsing of both organic detritus input and *in situ* productivity result in very complex linkages of mangroves with estuarine-dependent fisheries. In addition, mangrove detritus low in nitrogen relative to carbon may be modified by the microbial community and then utilized by higher trophic levels, masking the direct utilization of this organic matter as an energy source.

Impacts of Environmental Change

Mangroves are arguably an excellent indicator of how ecosystems will respond to the manifold impacts of global environmental change and land-use disturbance. Given present patterns, the combined effects of climate and land-use change will be noticeably evident in reduced goods and services of mangroves to human systems throughout the tropics in the twenty-first century. For example, accelerated rates in sea-level rise have been speculated as the most critical environmental change affecting the continued

existence of mangrove ecosystems. Numerous processes contribute to vertical accretion of mangroves at a rate that balances the increase in regional sea-level rise. Critical rates in sea-level rise have been estimated above which there is a projected collapse of mangrove ecosystems. While some speculation suggests that mangroves cannot sustain existence at sea-level rise >1.2–2.3 mm yr^{-1}, there is evidence that mangroves located in particular environmental settings existed through periods of accelerated sea level rise. Mangroves in Australia can keep pace with changes in sea-level rise with rates ranging from 0.2 to 6 mm yr^{-1} in the south Alligator tidal river. Also, mangrove forests in many estuaries in northern Australia tolerated sea-level rise of 8–10 mm yr^{-1} in the early Holocene. Many of these mangroves receive terrigenous sediments and exist in macrotidal environments, with critical rates that are much different than for mangroves in microtidal and carbonate environments. In addition, mangrove areas can be sustained along the coastline by migrating inland under conditions of increased sea-level rise. But this inland migration will depend on whether suitable inshore landscapes are available. The most significant recent restriction to mangrove colonization is human land use of available landscapes.

Mangroves in many coastal regions such as Gulf of Mexico and Caribbean are distributed in latitudes where the frequency of hurricanes and cyclones is high, resulting in strong effect on mangrove forest structure and community dynamics. Several patterns have been observed in Florida, Puerto Rico, Mauritius, and British Honduras. Species attributes and availability of propagules are important factors along with the severity of storm and sediment disturbance in projecting recovery patterns. Frequent storm disturbance tends to favor species capable of constant or timely flowering, abundant seedling or sprouting, fast growth in open conditions, and early reproductive maturity. Woody debris resulting from these disturbances have an important role in biogeochemical properties of disturbed mangrove forests. Although mangrove trees show these `traits', it is important to consider the cumulative impact of human activities on these ecosystems in conjunction with the complex natural cycle of regeneration and growth of mangrove forests. Cyclonic disturbance in areas with higher rates of sea-level rise has been demonstrated to cause sediment collapse (drop in surface elevation) that reduces the ability of mangroves to recolonize disturbed areas. Yet this potential impact may vary across ecogeomorphic types of mangroves.

River (and surface runoff) diversions that deprive tropical coastal deltas of freshwater and silt result in losses of mangrove species diversity and organic production, and alter the terrestrial and aquatic food webs that mangrove ecosystems support. Freshwater diversion of the Indus River to agriculture in Sind Province over the last several hundred years has reduced the once species-rich Indus River delta to a sparse community dominated by *Avicennia marina*. It is also responsible for causing significant erosion of the seafront due to sediment starvation and the silting-in of the abandoned spill rivers. A similar phenomenon has been observed in southwestern Bangladesh following natural changes in river channels of the Ganges and the construction of the Farakka barrage that reduced the dry season flow of freshwater into the mangrove-dominated western Sundarbans. Freshwater starvation, both natural and human-induced, has had negative impacts on the biodiversity of mangroves in the Ganges River delta as well along the dry coastal life zone of Colombia (the Cíenaga Grande de Santa Marta lagoon).

Deforestation of mangrove wetlands is associated with many uses of coastal environments, including urban, agriculture, and aquaculture reclamation, as well as the use of forest timber for furniture, energy, chip wood, and construction materials. Two reclamation activities that have contributed to examples of massive mangrove deforestation are agriculture and aquaculture enterprises. Agriculture impacts on mangroves are most noted in West Africa and parts of Indonesia. Many of the large agricultural uses are found in humid coastal areas or deltas where freshwater is abundant and intertidal lands are seasonally available for crop production. Mariculture use of the tropical intertidal zones, in the construction and operation of shrimp ponds, has become one of the most significant environmental changes of mangrove wetlands and water quality of tropical estuaries in the last several decades.

Oil spills represent contaminants to mangroves that can alter the succession, productivity, and nutrient cycling of these coastal forested wetlands. These impacts have been well documented in ecological studies in Puerto Rico, Panama, and Gulf of Mexico. An oil slick in a mangrove wetland will cause a certain mortality of trees depending on the concentration of hydrocarbons and species of trees, as well as the edaphic stress levels already existing at the site. Thus, those mangroves in dry coastal environments may be more vulnerable to oil spills than those in more humid environments.

Management and Restoration

Mangroves produce a variety of forest products, support the productivity of economically important estuarine- dependent fisheries, and modify the water quality in warm-temperate and tropical estuarine ecosystems. These goods and services lead to increased human utilization of mangrove resources that vary throughout the tropics depending on economic and cultural constraints (**Fig. 4**). Economic constraints are usually in the form of available capital to fund land-use changes in coastal regions, as well as river-basin development. Cultural constraints are complex and determine the degree of environmental management and natural resource utilization. However, the sustainable utilization of coastal resources, to a large degree, is controlled by these two social conditions of a region. Human use and value of mangrove wetlands are therefore a combination of both the ecological properties of these coastal ecosystems together with patterns of social exploitation. Therefore, any best management plan designed to provide for the sustainable utilization of mangrove wetlands has to consider both the ecological and social constraints of the region. Humans are part of all ecosystems, and management of natural resources is a combination of policies that seek to regulate the actions of societies within limitations that are imposed by the environment. Recent emphasis has been placed on comprehensive ecosystem restoration programs that represent changes in management of landscapes to reduce impacts on natural processes that enhance system recovery.

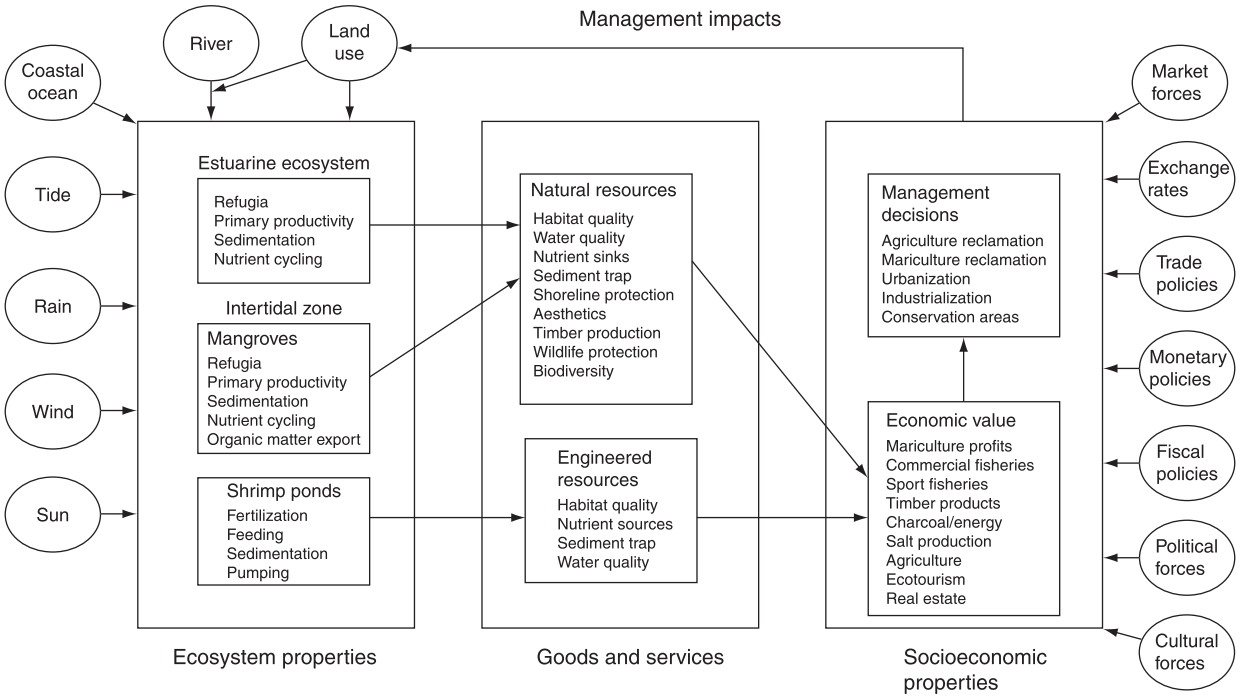


Fig. 4 Conceptual framework constraints of environmental setting and human activities on ecosystem properties, ecological functions, and uses of mangrove ecosystems that determine management decisions in coastal environments. From Twilley RR, Gottfried RR, Rivera-Monroy VH, Armijos MM, and Bodero A (1998) An approach and preliminary model of integrating ecological and economic constraints of environmental quality in the Guayas River estuary, Ecuador. *Environmental Science and Policy* 1: 271–288.

There have been several reviews of mangrove restoration, which collectively have alluded to the concept that since these forested wetlands are adapted to stressed environments, they are relatively amenable to restoration efforts. The success of mangrove restoration is the establishment of the proper environmental settings that control the characteristic structure and function of mangrove wetlands. The goal of ecological restoration is to return a degraded mangrove site back to either the natural condition (restoration) or to some other new condition (rehabilitation). The rates of change in the ecological characteristics of mangrove wetlands between natural, degraded, and some rehabilitated condition will depend on the type of environmental impact, the magnitude of the impact, and the ecogeomorphic type of mangrove wetland that is impacted. The success of any mangrove restoration project depends on the establishment of proper site conditions (geophysical processes and geomorphic features) along with ecological processes of the site such as the availability of propagules and the recruitment of these individuals to sapling stage of development. Some of the key parameters of a restoration project include the elevation of the landscape to provide the proper hydrology of the site, recognizing the significance of natural processes to sustaining the restored condition, and proper planting techniques to enhance recruitment. Several models of different properties of mangroves have been developed during the last decade to help facilitate planning and design of mangrove restoration projects and improve our management of these critical features of coastal landscape.

See also: **Aquatic Ecology:** Equilibrium Concept in Phytoplankton Communities. **Ecological Complexity:** Goal Functions and Orientors. **Ecological Data Analysis and Modelling:** Modeling Dispersal Processes for Ecological Systems. **Ecosystems:** Riparian Wetlands. **Evolutionary Ecology:** Association. **General Ecology:** Carrying Capacity. **Global Change Ecology:** Energy Flows in the Biosphere

Further Reading

Alexander, T.R., 1967. Effect of Hurricane Betsy on the southeastern Everglades. Quarterly Journal of the Florida Academy of Sciences 30, 10–24.
Allen, J.A., Ewel, K.C., Keeland, B.D., Tara, T., Smith, T.J., 2000. Downed wood in Micronesian mangrove forests. Wetlands 20, 169–176.
Alongi, D.M., Christoffersen, P., Tirendi, F., Robertson, A.I., 1992. The influence of freshwater and material export on sedimentary facies and benthic processes within the Fly Delta and adjacent Gulf of Papua (Papua New Guinea). Continental Shelf Research 12, 287–326.
Bacon, P.R., 1990. The ecology and management of swamp forests in the Guianas and Caribbean region. In: Lugo, A.E., Brinson, M., Brown, S. (Eds.), Ecosystems of the World 15: Forested Wetlands. Amsterdam: Elsevier Press, pp. 213–250.
Bacon, P.R., 1994. Template for evaluation of impacts of sea level rise on Caribbean coastal wetlands. Ecological Engineering 3, 171–186.
Baldwin, A., Egnotovich, M., Ford, M., Platt, W., 2001. Regeneration in fringe mangrove forests damaged by Hurricane Andrew. Plant Ecology 157, 151–164.
Ball, M.C., 1980. Patterns of secondary succession in a mangrove forest of southern Florida. Oecologia 44, 226–235.

Ball, M.C., 1988. Ecophysiology of mangroves. Trees 2, 129–142.
Berger, U., Hildenbrandt, H., 2000. A new approach to spatially explicit modelling of forest dynamics: Spacing, ageing and neighborhood competition of mangrove trees. Ecological Modelling 132, 287–302.
Blasco, F., 1984. Climatic factors and the biology of mangrove plants. In: Snedaker, S.C., Snedaker, J.G. (Eds.), The Mangrove Ecosystem: Research Methods. Paris: UNESCO, pp. 18–35.
Botero, L., 1990. Massive mangrove mortality on the Caribbean coast of Colombia. Vida Silvestre Neotropical 2, 77–78.
Boto, K.G., Saffingna, P., Clough, B., 1985. Role of nitrate in nitrogen nutrition of the mangrove *Avicennia Marina*. Marine Ecology Progress Series 21, 259–265.
Boto, K.G., Wellington, J.T., 1988. Seasonal variations in concentrations and fluxes of dissolved organic and inorganic materials in a tropical, tidally-dominated, mangrove waterway. Marine Ecology Progress Series 50, 151–160.
Brown, S., Lugo, A.E., 1994. Rehabilitation of tropical lands: A key to sustaining development. Restoration Ecology 2, 97–111.
Camilleri, J.C., 1992. Leaf-litter processing by invertebrates in a mangrove forest in Queensland. Marine Biology 114, 139–145.
Carlton, J.M., 1974. Land-building and stabilization by mangroves. Environmental Conservation 1, 285.
Chapman, V.J., 1976. Mangrove Vegetation. Vaduz, Germany: J. Cramer.
Chen, R., Twilley, R.R., 1998. A gap dynamic model of mangrove forest development along gradients of soil salinity and nutrient resources. Journal of Ecology 86, 1–12.
Chen, R., Twilley, R.R., 1998. A simulation model of organic matter and nutrient accumulation in mangrove wetland soils. Biogeochemistry 44, 93–118.
Chen, R., Twilley, R.R., 1999. Patterns of mangrove forest structure and soil nutrient dynamics along the Shark River Estuary, Florida. Estuaries 22, 955–970.
Cintrón, G., 1990. Restoration of mangrove systems. In Symposium on Habitat Restoration. Washington, DC: National Oceanic and Atmospheric Administration.
Cintrón, G., Lugo, A.E., Martinez, R., Cintrón, B.B., and Encarnacion, L., 1981. Impact of oil in the tropical marine environment, pp. 18–27. Technical Publication. Division of Marine Resources, Department of Natural Resources of Puerto Rico.
Corredor, J.E., Morell, M.J., 1994. Nitrate depuration of secondary sewage effluents in mangrove sediments. Estuaries 17, 295–300.
Craighead, F.C., Gilbert, V.C., 1962. The effects of Hurricane Donna on the vegetation of southern Florida. Quarterly Journal of the Florida Academy of Sciences 25, 1–28.
Davis, J.H., 1940.*Carnegie Institution, Publication No. 517* In: The ecology and geologic role of mangroves in Florida. Washington, DC: Carneigie Institution, pp. 303–412.
Davis, S., Childers, D.L., Day, J.W.J., Rudnick, D., Sklar, F., 2001. Wetland-water column exchanges of carbon, nitrogen, and phosphorus in a southern everglades dwarf mangrove. Estuaries 24, 610–622.
Davis, S., Childers, D.L., Day, J.W., Rudnick, D., Sklar, F., 2003. Factors affecting the concentration and flux of materials in two southern everglades mangrove wetlands. Marine Ecology Progress Series 253, 85–96.
Duke, N.C., 2001. Gap creation and regenerative process driving diversity and structure of mangrove ecosystems. Wetlands Ecology and Management 9, 257–269.
Duke, N.C., Pinzon, Z., 1993. Mangrove forests. In: Keller, B.D., Jackson, J.B.C. (Eds.), Long-Term Assessment of the Oil Spill at Bahia las Minas, Panama, Synthesis Report, Volumn II, Technical Report. New Orleans, LA: US Dept of the Interior, Minerals Management Service, Gulf of Mexico OCS Regional Office, pp. 447–553.
Ellison, J.C., 1993. Mangrove retreat with rising sea-level Bermuda. Estuarine, Coastal and Shelf Science 37, 75–87.
Ellison, A.M., 2000. Mangrove restoration: Do we know enough? Restoration Ecology 8, 219–229.
Ellison, A.M., Farnsworth, E.J., 1992. The ecology of Belizean mangrove-root fouling communities: Patterns of epibiont distribution and abundance, and effects on root growth. Hydrobiologia 20, 1–12.
Ellison, J.C., Stoddart, D.R., 1991. Mangrove ecosystem collapse during predicted sea-level rise: Holocene analogues and implications. Journal of Coastal Research 7, 151–165.
Ewe, S.M.L., Gaiser, E.E., Childers, D.L., *et al*., 2006. Spatial and temporal patterns of aboveground net primary productivity (ANPP) in the Florida Coastal Everglades. Hydrobiologia 569, 459–474.
Ewel, K.C., Ong, J.E., Twilley, R., 1998. Different kinds of mangrove swamps provide different goods and services. Global Ecology and Biogeography Letters 7, 83–94.
Ewel, K.C., Zheng, S.F., Pinzon, Z.S., Bourgeois, J.A., 1998. Environmental effects of canopy gap formation in high-rainfall mangrove forests. Biotropica 30, 510–518.
Farnsworth, E.J., Ellison, A.M., 1991. Patterns of herbivory in Belizean mangrove swamps. Biotropica 23, 555–567.
Farnsworth, E.J., Ellison, A.M., 1993. Dynamics of herbivory in Belizean mangal. Journal of Tropical Ecology 9, 435–453.
Farquhar, G.D., Ball, M.C., von Caemmerer, S., Roksandic, Z., 1982. Effect of salinity and humidity on $\delta^{13}C$ values of halophytes – evidence for diffusional isotope fractionation determined by the ratios of intercellular/atmospheric CO_2 under different environmental conditions. Oecologia (Berlin) 52, 121–137.
Fell, J.W., Master, I.M., 1973. Fungi associated with the degradation of mangrove (*Rhizophora mangle* L.) leaves in south Florida. In: Stevenson, L.H., Colwell, R.R. (Eds.), Estuarine Microbial Ecology. Columbia, SC: University of South Carolina Press, pp. 455–465.
Feller, I.C., 1993. *Effects of Nutrient Enrichment on Growth and Herbivory of Dwarf Red mangrove*. PhD Dissertation, Georgetown University.
Feller, I.C., 1995. Effects of nutrient enrichment on growth and herbivory of dwarf red mangrove (*Rhizophora mangle*). Ecological Monographs 65, 477–505.
Feller, I.C., McKee, K.L., 1999. Small gap creation in Belizean mangrove forests by a wood-boring insect. Biotropica 31, 607–617.
Feller, I.C., Whigham, D.F., McKee, K.L., Lovelock, C.E., 2003. Nitrogen limitation of growth and nutrient dynamics in a disturbed mangrove forest, Indian River Lagoon, Florida. Oecologia 134, 405–414.
Feller, I.C., Whigham, D.F., O'Neill, J.P., McKee, K.L., 1999. Effects of nutrient enrichment on within-stand cycling in a mangrove forest. Ecology 80, 2193–2205.
Field, C.D., 1996. Restoration of mangrove ecosystems. In International Society for Mangrove Ecosystems. Hong Kong: South China Printing.
Fry, B., Bern, A.L., Ross, M.S., Meeder, J.F., 2000. $\delta^{15}N$ studies of nitrogen use by the red mangrove, *Rhizophora mangle* L., in south Florida. Estuarine, Coastal and Shelf Science 50, 723–735.
Fry, B., Smith III, T.J., 2002. Stable isotope studies of red mangroves and filter feeders from the Shark River estuary, Florida. Bulletin of Marine Sciences 70, 871–890.
Garrity, S.D., Levings, S.C., Burns, K.A., 1994. The Galeta oil spill. I. Long-term effects on the physical structure of the mangrove fringe. Estuarine, Coastal and Shelf Science 38, 327–348.
Getter, C.D., Scott, G.I., Michel, J., 1981. The effects of oil spills on mangrove forests: A comparison of five oil spill sites in the Gulf of Mexico and the Caribbean Sea. In: Proceedings of the 1981 Oil Spill Conference. Washington, DC: API/EPA/USCG, pp. 65–111.
Gilmore Jr., R.G., Snedaker, S.C., 1993. Mangrove forests. In: Martin, W.H., Boyce, S.G., Echternacht, A.C. (Eds.), Biodiversity of the Southeastern United States/Lowland Terrestrial Communities. New York: Wiley, pp. 165–198.
Glynn, P.W., Almodovar, L.R., Gonzalez, J.G., 1964. Effects of hurricane Edith on marine life in La Parguera, Pueto Rico. Caribbean Journal of Science 4, 335–345.
Gosselink, J.G., Turner, R.E., 1978. The role of hydrology in freshwater wetland ecosystems. In: Good, D.F.W.R.E., Simpson, R.L. (Eds.), Freshwater Wetlands: Ecological Processes and Management Potential. New York: Academic Press, pp. 633–678.
Hedgpeth, J.W., 1957. Classification of marine environments. Geological Society of America, Memoir 67 (1), 17–28.
Huston, M.A., 1994. Biological Diversity. Cambridge: Cambridge University Press.
Iizumi, H., 1986. Soil nutrient dynamics. In: Cragg, S., Polunin, N. (Eds.), Workshop on Mangrove Ecosystem Dynamics. New Delhi: UNDP/UNESCO Regional Project, p. 171. (RAS/79/002).
Jones, D.A., 1984. Crabs of the mangal ecosystem. In: Por, F.D., Dor, I. (Eds.), Hydrobiology of the Mangal. The Hague: Dr. W. Junk Publishers, pp. 89–109.
Koch, M.S., Snedaker, S.C., 1997. Factors influencing *Rhizophora mangle* L. seedlings development into the sapling stage across resource and stress gradients in subtropical Florida. Biotropica 29, 427–439.
Krauss, K.W., Allen, J.A., Cahoon, D.R., 2003. Differential rates of vertical accretion and elevation change among aerial root types in Micronesian mangrove forests. Estuarine Coastal and Shelf Science 56, 251–259.

Krauss, K.W., Doyle, T.W., Twilley, R.R., Smith, T.J., Whelan, K.R.T., Sullivan, J.K., 2005. Woody debris in the mangrove forests of south Florida. Biotropica 37, 9–15.
Kristensen, E., Andersen, F.Ø., Kofoed, L.H., 1988. Preliminary assessment of benthic community metabolism in a Southeast Asian mangrove swamp. Marine Ecology Progress Series 48, 137–145.
Lee, S.Y., 1989. Litter production and turnover of the mangrove *Kandelia candel* (L.) Druce in a Hong Kong tidal shrimp pond. Estuarine, Coastal and Shelf Science 29, 75–87.
Leh, C.M.U., Sasekumar, A., 1985. The food of sesarmid crabs in Malaysian mangrove forests. Malay Naturalist Journal 39, 135–145.
Lewis, R.R., 1982. Mangrove forests. In: Lewis, R.R. (Ed.), Creation and Restoration of Coastal Plant Communities. Boca Raton, FL: CRC Press, pp. 153–171.
Lewis, R.R., 1990. Creation and restoration of coastal plain wetlands in Florida. In: Kusler, J.A., Kentula, M.E. (Eds.), Wetland Creation and Restoration. Washington, DC: Island Press, pp. 73–101.
Lewis, R.R., 1990. Creation and restoration of coastal wetlands in Puerto Rico and the US Virgin Islands. In: Kusler, J.A., Kentula, M.E. (Eds.), Wetland Creation and Restoration. Washington, DC: Island Press, pp. 103–123.
Lin, G., Sternberg, L.S.L., 1992. Differences in morphology, carbon isotope ratios, and photosynthesis between scrub and fringe mangroves in Florida, USA. Aquatic Botany 42, 303–313.
Lin, G., Sternberg, L.S.L., 1992. Effect of growth form, salinity, nutrient and sulfide on photosynthesis, carbon isotope discrimination and growth of red mangrove (*Rhizophora mangle* L.). Australian Journal of Plant Physiology 19, 509–517.
Lovelock, C.E., Feller, I.C., Mckee, K.L., Engelbrecht, B.M.J., Ball, M.C., 2004. The effect of nutrient enrichment on growth, photosynthesis and hydraulic conductance of dwarf mangroves in Panama. Functional Ecology 18, 25–33.
Lugo, A.E., 1980. Mangrove ecosystems: Successional or steady state? Biotropica 12, 65–72.
Lugo, A.E., 1998. Mangrove forests: A tough system to invade but an easy one to rehabilitate. Marine Pollution Bulletin 37, 427–430.
Lugo, A.E., Snedaker, S.C., 1974. The ecology of mangroves. Annual Review of Ecology and Systematics 5, 39–64.
Lynch, J.C., Meriwether, J.R., McKee, B.A., Vera-Herrera, F., Twilley, R.R., 1989. Recent accretion in mangrove ecosystems based on ^{137}Cs and ^{210}Pb. Estuaries 12, 284–299.
Macnae, W., 1968. A general account of the fauna and flora of mangrove swamps and forests in the Indo-West-Pacific region. Advances in Marine Biology 6, 73–270.
Malley, D.F., 1978. Degradation of mangrove leaf litter by the tropical sesarmid crab *Chiromanthes onychophorum*. Marine Biology 49, 377–386.
McKee, K.L., 1993. Soil physicochemical patterns and mangrove species distribution – Reciprocal effects? Journal of Ecology 81, 477–487.
McKee, K.L., Feller, I.C., Popp, M., Wanek, W., 2002. Mangrove isotopic (δ^{15}N and δ^{13}C) fractionation across a nitrogen vs. phosphorus limitation gradient. Ecology 83, 1065–1075.
Medina, E., Francisco, M., 1997. Osmolality and δ^{13}C of leaf tissues of mangrove species from environments of contrasting rainfall and salinity. Estuarine, Coastal and Shelf Science 45, 337–344.
Naidoo, G., 1985. Effects of waterlogging and salinity on plant–water relations and on the accumulation of solutes in three mangrove species. Aquatic Botany 22, 133–143.
Nedwell, D.B., 1975. Inorganic nitrogen metabolism in a eutrophicated tropical mangrove estuary. Water Research 9, 221–231.
Nixon, S.W., 1980. Between coastal marshes and coastal waters – A review of twenty years of speculation and research on the role of salt marshes in estuarine productivity and water chemistry. In: Hamilton, P., MacDonald, K.B. (Eds.), Estuarine and Wetland Processes with Emphasis on Modeling. New York: Plenum Press, pp. 437–525.
Odum, W.E., Heald, E.J., 1972. Trophic analysis of an estuarine mangrove community. Bulletin Marine Science 22, 671–738.
Odum, W.E., McIvor, C.C., 1990. Mangroves. In: Myers, R.L., Ewel, J.J. (Eds.), Ecosystems of Florida. Orlando, FL: University of Central Florida Press, pp. 517–548.
Odum, W.E., McIvor, C.C., Smith, T.J., 1982. The Ecology of the Mangroves of South Florida: A Community Profile. FWS/OBS-81/24. Washington, DC: US Fish and Wildlife Service, Office of Biological Resources.
Parkinson, R.W., DeLaune, R.D., White, J.R., 1994. Holocene sea-level rise and the fate of mangrove forests within the wider Caribbean region. Journal of Coastal Research 10, 1077–1086.
Pinzon, Z.S., Ewel, K.C., Putz, F.E., 2003. Gap formation and forest regeneration in a Micronesian mangrove forest. Journal of Tropical Ecology 19, 143–153.
Ponnamperuma, F.N., 1984. Mangrove swamps in south and Southeast Asia as potential rice lands. In: Soepadmo, E., Rao, A.N., McIntosh, D.J. (Eds.), Proceedings Asian Mangrove Symposium. Kuala Lumpur: University of Malaya, pp. 672–683.
Pool, D.J., Lugo, A.E., Snedaker, S.C., 1975. Litter production in mangrove forests of southern Florida and Puerto Rico. In: Walsh, G., Snedaker, S., Teas, H. (Eds.), Proceedings of the International Symposium on the Biology and Management of Mangroves. Gainesville, FL: Institute of Food and Agricultural Sciences. University of Florida, pp. 213–237.
Rabinowitz, D., 1978. Early growth of mangrove seedlings in Panama, and an hypothesis concerning the relationship of dispersal and zonation. Journal of Biogeography 5, 113–133.
Rivera-Monroy, V.H., Day, J.W., Twilley, R.R., Vera-Herrera, F., Coronado-Molina, C., 1995. Flux of nitrogen and sediment in a fringe mangrove forest in Terminos Lagoon, Mexico. Estuarine, Coastal and Shelf Science 40, 139–160.
Rivera-Monroy, V.H., Twilley, R.R., 1996. The relative role of denitrification and immobilization in the fate of inorganic nitrogen in mangrove sediments. Limnology and Oceanography 41, 284–296.
Rivera-Monroy, V.H., Twilley, R.R., Boustany, R.G., Day, J.W., Vera-Herrera, F., Ramirez, Md.C., 1995. Direct denitrification in mangrove sediments in Terminos Lagoon, Mexico. Marine Ecology Progress Series 97, 97–109.
Rivera-Monroy, V.H., Twilley, R.R., Bone, D., *et al.*, 2004. A conceptual framework to develop long-term ecological research and management objectives in the wider Caribbean region. Bioscience 54, 843–856.
Robertson, A.I., 1986. Leaf-burying crabs: Their influence on energy flow and export from mixed mangrove forests (*Rhizophora* spp.) in northeastern Australia. Journal of Experimental Marine Biology and Ecology 102, 237–248.
Robertson, A.I., Alongi, D.M., 1992. Tropical Mangrove Ecosystems, vol. 41. Washington, DC: American Geophysical Union.
Robertson, A.I., Alongi, D.M., Boto, K.G., 1992. Food chains and carbon fluxes. In: Robertson, A.I., Alongi, D.M. (Eds.), Tropical Mangrove Ecosystems. Washington, DC: American Geophysical Union, pp. 293–326.
Robertson, A.I., Blaber, S.J.M., 1992. Plankton, epibenthos and fish communities. In: Robertson, A.I., Alongi, D.M. (Eds.), Tropical Mangrove Ecosystems. Washington, DC: American Geophysical Union, pp. 173–224.
Robertson, A.I., Daniel, P.A., 1989. The influence of crabs on litter processing in high intertidal mangrove forests in tropical Australia. Oecologia 78, 191–198.
Robertson, A.I., Duke, N.C., 1990. Mangrove fish-communities in tropical Queensland, Australia: Spatial and temporal patterns in densities, biomass and community structure. Marine Biology 104, 369–379.
Rodelli, M.R., Gearing, J.N., Gearing, P.J., Marshall, N., Sasekumar, A., 1984. Stable isotope ratio as a tracer of mangrove carbon in Malaysian ecosystems. Oecologia 61, 326–333.
Rojas-Galaviz, J.L., Yáñez-Arancibia, A., Day Jr., J.W., Vera-Herrera, F.R., 1992. Estuarine primary producers: Laguna de Terminos-a study case. In: Seeliger, U. (Ed.), Coastal Plant Communities of Latin America. San Diego, CA: Academic Press, pp. 141–154.
Romero, L.M., Smith, T.J., Fourqurean, J.W., 2005. Changes in mass and nutrient content of wood during decomposition in a south Florida mangrove forest. Journal of Ecology 93, 618–631.
Ross, M.S., Meeder, J.F., Sah, J.P., Ruiz, L.P., Telesnicki, G.J., 2000. The Southeast saline Everglades revisited: 50 Years of coastal vegetation change. Journal of Vegetation Science 11, 101–112.
Roth, L.C., 1992. Hurricanes and mangrove regeneration: Effects of Hurricane Juan, October 1988, on the vegetation of Isla del Venado, Bluefields, Nicaragua. Biotropica 24, 375–384.

Rützler, K., Feller, C., 1988. Mangrove swamp communities. Oceanus 30, 16–24.
Rützler, K., Feller, C., 1996. Carribbean mangrove swamps. Scientific American 274, 94–99.
Saenger, P., Hegerl, E.J., and Davie, J.D.S., 1983. *Global Status of Mangrove Ecosystems. Commission on Ecology Paper No. 3*, pp. 83. International Union for the Conservation of Nature (IUCN).
Saenger, P., Snedaker, S.C., 1993. Pantropical trends in mangrove above-ground biomass and annual litterfall. Oecologia 96, 293–299.
Sauer, J.D., 1962. Effects of recent tropical cyclones on the coastal vegetation of Mauritius. Journal of Ecology 50, 275–290.
Scholander, P.F., Hammel, H.T., Hemmingsen, E., Garay, W., 1962. Salt balance in mangroves. Plant Physiology 37, 722–729.
Sherman, R.E., Fahey, T.J., Battles, J.J., 2000. Small-scale disturbance and regeneration dynamics in a neotropical mangrove forest. Journal of Ecology 88, 165–178.
Simberloff, D.S., Wilson, E.O., 1969. Experimental zoogeography of islands: The colonization of empty islands. Ecology 50, 278–289.
Smith III, T.J., 1987. Seed predation in relation to tree dominance and distribution in mangrove forests. Ecology 68, 266–273.
Smith III, T.J., 1992. Forest structure. In: Robertson, A.I., Alongi, D.M. (Eds.), Tropical Mangrove Ecosystems. Washington, DC: American Geophysical Union, pp. 101–136.
Smith, T.J., Boto, K.G., Frusher, S.D., Giddins, R.L., 1991. Keystone species and mangrove forest dynamics: The influence of burrowing by crabs on soil nutrient status and forest productivity. Estuarine Coastal and Shelf Science 33, 419–432.
Smith III, T.J., Robblee, M.B., Wanless, H.R., Doyle, T.W., 1994. Mangroves, hurricanes, and lightning strikes. BioScience 44, 256–262.
Snedaker, S., 1982. Mangrove species zonation: Why? In: Sen, D.N., Rajpurohit, K.S. (Eds.), Tasks for Vegetation Science, vol. 2. The Hague: Junk, pp. 111–125.
Snedaker, S.C., 1986. Traditional uses of South American mangrove resources and the socio-economic effect of ecosystem changes. In: Kunstadter, P., Bird, E.C.F., Sabhasri, S. (Eds.), Proceedings, Workshop on Man in the Mangroves. Tokyo: United Nations University, pp. 104–112.
Snedaker, S.C., 1989. Overview of ecology of mangroves and information needs for Florida Bay. Bulletin of Marine Science 44, 341–347.
Snedaker, S.C., Meeder, J.F., Ross, M.S., Ford, R.G., 1994. Discussion of Ellison, JC and Stoddart, DR 1991. Mangrove ecosystem collapse during predicted sea-level rise: Holocene analogues and implications. Journal of Coastal Research 7, 151–165. *Journal of Coastal Research* 10: 497–498.
Snedaker, S.C., Snedaker, J.G., 1984. The Mangrove Ecosystem: Research Methods. London: UNESCO.
Sousa, W.P., Quek, S.P., Mitchell, B.J., 2003. Regeneration of Rhizophora mangle in a Caribbean mangrove forest: Interacting effects of canopy disturbance and a stem-boring beetle. Oecologia 137, 436–445.
Sutherland, J.P., 1980. Dynamics of the epibenthic community on roots of the mangrove Rhizophora mangle, at Bahia de Buche, Venezuela. Marine Biology 58, 75–84.
Teas, H.J., 1981. Restoration of mangrove ecosystems. In: Carey, R.C., Markovits, P.S., Kirkwood, J.B. (Eds.), Proceedings of Workshop on Coastal Ecosystems of the Southeastern United States. Reno, NV: US Fish and Wildlife Service, Office of Biological Services, FWS/OBS-80/59, pp. 95–103.
Thayer, G.W., Colby, D.R., Hettler Jr., W.F., 1987. Utilization of the red mangrove prop root habitat by fishes in south Florida. Marine Ecology Progress Series 35, 25–38.
Thom, B., 1967. Mangrove ecology and deltaic morphology: Tabasco, Mexico. Journal of Ecology 55, 301–343.
Thom, B.G., 1982. Mangrove ecology – A geomorphological perspective. In: Clough, B.F. (Ed.), Mangrove Ecosystems in Australia. Canberra: Australian National University Press, pp. 3–17.
Thom, B.G., 1984. Coastal landforms and geomorphic processes. In: Sneadeker, S.C., Sneadeker, J.G. (Eds.), The Mangrove Ecosystem: Research Methods. Paris: UNESCO, pp. 3–17.
Tilman, D., 1982. Resource Competition. Princeton, NJ: Princeton University Press.
Tomlinson, P.B., 1995. The Botany of Mangroves. New York: Cambridge University Press.
Twilley, R.R., 1988. Coupling of mangroves to the productivity of estuarine and coastal waters. In: Jansson, B.O. (Ed.), Coastal-Offshore Ecosystems: Interactions. Berlin: Springer, pp. 155–180.
Twilley, R.R., 1995. Properties of mangroves ecosystems and their relation to the energy signature of coastal environments. In: Hall, C.A.S. (Ed.), Maximum Power. Denver, CO: Colorado Press, pp. 43–62.
Twilley, R.R., 1997. Mangrove wetlands. In: Messina, M., Connor, W. (Eds.), Southern Forested Wetlands: Ecology and Management. Boca Raton, FL: CRC Press, pp. 445–473.
Twilley, R.R., Cárdenas, W., Rivera-Monroy, V.H., *et al*., 2000. Ecology of the Gulf of Guayaquil and the Guayas River Estuary. In: Seeliger, U., Kjerve, B.J. (Eds.), Coastal Marine Ecosystems of Latin America. New York: Springer, pp. 245–263.
Twilley, R.R., Chen, R.H., 1998. A water budget and hydrology model of a basin mangrove forest in Rookery Bay, Florida. Marine and Freshwater Research 49, 309–323.
Twilley, R.R., Chen, R.H., Hargis, T., 1992. Carbon sinks in mangroves and their implications to carbon budget of tropical coastal ecosystems. Water, Air and Soil Pollution 64, 265–288.
Twilley, R.R., Gottfried, R.R., Rivera-Monroy, V.H., Armijos, M.M., Bodero, A., 1998. An approach and preliminary model of integrating ecological and economic constraints of environmental quality in the Guayas River estuary, Ecuador. Environmental Science and Policy 1, 271–288.
Twilley, R.R., Lugo, A.E., Patterson-Zucca, C., 1986. Production, standing crop, and decomposition of litter in basin mangrove forests in southwest Florida. Ecology 67, 670–683.
Twilley, R.R., Pozo, M., Garcia, V.H., Rivera-Monroy, V.H., Zambrano, R., Bodero, A., 1997. Litter dynamics in riverine mangrove forests in the Guayas River estuary, Ecuador. Oecologia 111, 109–122.
Twilley, R.R., Rivera-Monroy, V.H., 2005. Developing performance measures of mangrove wetlands using simulation models of hydrology, nutrient biogeochemistry, and community dynamics. Journal of Coastal Research 40, 79–93.
Twilley, R.R., Rivera-Monroy, V.H., Chen, R., Botero, L., 1998. Adapting and ecological mangrove model to simulate trajectories in restoration ecology. Marine Pollution Bulletin 37, 404–419.
Twilley, R.R., Snedaker, S.C., Yañez-Arancibia, A., Medina, E., 1996. Biodiversity and ecosystem processes in tropical estuaries: Perspectives from mangrove ecosystems. In: Mooney, H., Cushman, H., Medina, E. (Eds.), Biodiversity and Ecosystem Functions: A Global Perspective. New York: Wiley, pp. 327–370.
Vermeer, D.E., 1963. Effects of Hurricane Hattie, 1961, on the cays of British Honduras. Zeitschrift fur Geomorphologie 7, 332–354.
Wadsworth, F.H., 1959. Growth and regeneration of white mangrove in Puerto Rico. Caribbean Forester 20, 59–69.
Waisel, Y., 1972. In: Biology of Halophytes. New York: Academic Press, p. 395.
Walsh, G.E., 1974. Mangroves: A review. In: Reimold, R., Queen, W. (Eds.), Ecology of Halophytes. New York: Academic Press, pp. 51–174.
Wanless, H.R., Parkinson, R.W., Tedesco, L.P., 1994. Sea level control on stability of Everglades wetlands. In: Davis, S., Ogden, J. (Eds.), Everglades: The Ecosystem and Its Restoration. Delray Beach, FL: St. Lucie Press, pp. 199–223.
Watson, J., 1928. Mangrove Forests of the Malay Peninsula. Singapore: Fraser & Neave.
Woodroffe, C.D., 1990. The impact of sea-level rise on mangrove shoreline. Progress in Physical Geography 14, 483–520.
Woodroffe, C., 1992. Mangrove sediments and geomorphology. In: Robertson, A.I., Alongi, D.M. (Eds.), Tropical Mangrove Ecosystems. Washington, DC: American Geophysical Union, pp. 7–42.
Woodroffe, C.D., Chappell, J., Thom, B.G., Wallensky, E., 1986. Geomorphological dynamics and evolution of the South Alligator tidal river and plains. In ANU, North Australia Research Unit Monograph 3. Darwin: North Australian Research Unit.
Yañez-Arancibia, A., 1985. Fish Community Ecology in Estuaries and Coastal Lagoons: Towards an Ecosystem Integration. Mexico City: UNAM Press.
Yañez-Arancibia, A., Day Jr., J.W., 1982. Ecological characterization of Terminos Lagoon, a tropical lagoon-estuarine system in the Southern Gulf of Mexico. Oceanologica Acta SP, 431–440.

Yañez-Arancibia, A., Day Jr., J.W., 1988. Ecology of Coastal Ecosystems in the Southern Gulf of Mexico: The Terminos Lagoon Region. Mexico City: Universidad Nacional Autonoma de Mexico, Ciudad Universitaria, Mexico.

Yáñez-Arancibia, A., Lara-Domínguez, A.L., Day, J.W., 1993. Interactions between mangrove and seagrass habitats mediated by estuarine nekton assemblages: Coupling of primary and secondary production. Hydrobiologia 264, 1–12.

Yáñez-Arancibia, A., Lara-Domínguez, A.L., Rojas-Galaviz, J.L., *et al*., 1988. Seasonal biomass and diversity of estuarine fishes coupled with tropical habitat heterogeneity (southern Gulf of Mexico). Journal of Fish Biology 33 (supplement A), 191–200.

Peatlands☆

DH Vitt, Southern Illinois University, Carbondale, IL, USA

Introduction

Peatlands, or mires as they are sometimes called, are characterized by often deep accumulations of incompletely decomposed organic material, or peat. Peat accumulates when carbon that is sequestered in plant biomass through the process of photosynthesis exceeds the long-term loss of this carbon to the atmosphere via decomposition plus losses of carbon dissolved in water removed from the peatland through hydrological flow. Globally, peatlands contain about 30% of the world's terrestrial soil carbon, while covering only about 3–4% of the Earth's surface, and as such their carbon storage is considerably greater than their land surface area might indicate. Peatlands, in general, are relatively species poor when compared to upland communities in the same geographic region. However, due to the specialized environmental conditions often associated with peatlands, plants, and animals found only in these ecosystems are sometimes present. Peatlands are especially known for the presence of carnivorous plants such as *Sarracenia* and *Drosera* and for the occurrence of a large number of species of peat mosses (the genus *Sphagnum*).

Occurrence

Globally, peatlands occupy about 4 million km^2, with the boreal and subarctic peatland area estimated to be approximately 3 460 000 km^2, or about 87% of the world's peatlands. Six countries have greater than 50 000 km^2 of peatland and these account for 93% of the world's peatlands – five of these countries are predominantly boreal. Russia contains 1.42 million km^2, Canada 1.235 million km^2, the US 625 000 km^2, Finland 96 000 km^2, and Sweden 70 000 km^2; Indonesia has an estimated 270 000 km^2 as well. Although peat-forming plant communities occur in most of the world's nine zonobiomes, they are most prevalent in zonobiome VIII (cold temperate), or more commonly termed the boreal forest or taiga (Figure 1). The world's largest peatland complex is located in western Siberia (especially noteworthy is the Great Vasyugan Mire located between the Ob and Irtysh Rivers at about 58° N and 75° W). Two other large peatland complexes are the Hudson Bay Lowland in eastern Canada and the Mackenzie River Basin in northwestern Canada. Although peatlands have long been associated with cool, oceanic climatic regimes such as those in Britain and Ireland and indeed peatlands are common in these areas, in fact peatlands are most abundant in areas where the regional climate is continental with short cool summers and long cold winters, where the vegetation is coniferous and evergreen, and the upland soils are podzolic.

Environmental Limiting Factors

The initiation, development, and succession of peatland ecosystems are influenced by a number of regional, external factors. Especially important are hydrological and landscape position, climate, and substrate chemistry. These regional allogenic factors

☆*Change History:* June 2013. DH Vitt introduced small edits in the text of the article including sections Environmental Limiting Factors and Nutrient Sequestration (Oligotrophication), and changed Figure 3.

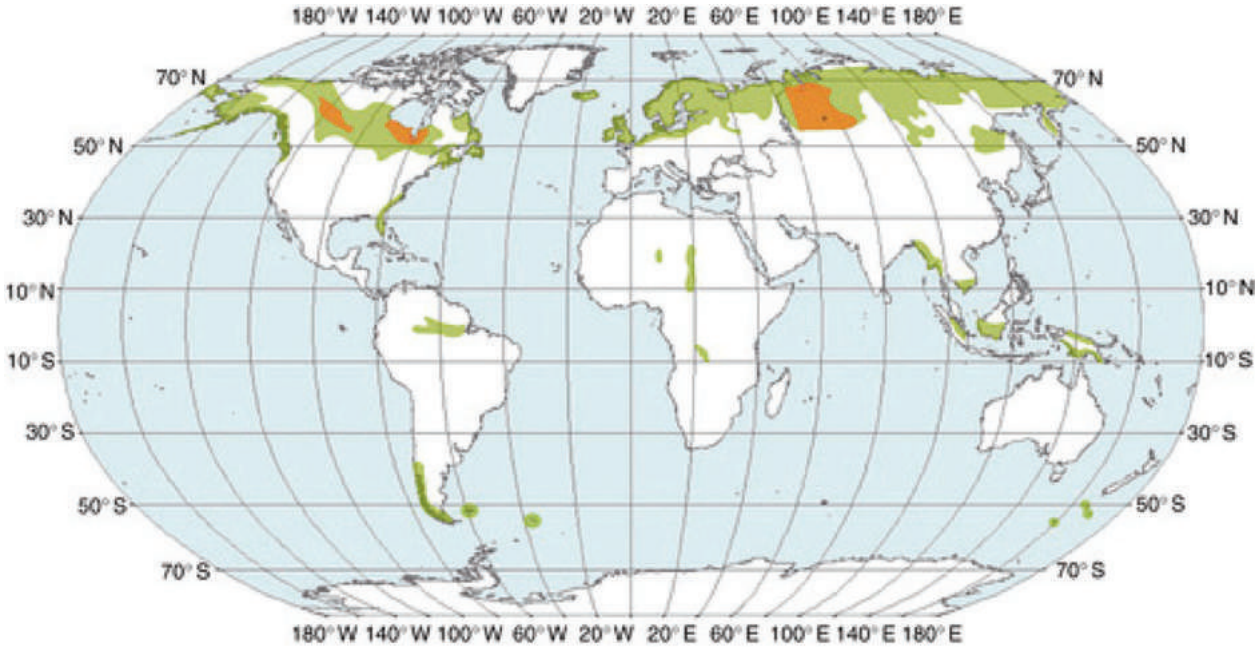


Figure 1 Estimated global distribution of peatlands. Areas colored in light green are those having >10% peat cover. The orange areas in North America and Siberia are the world's largest peatland complexes. The dot in western Siberia is the location for the Vasyugan peatland.

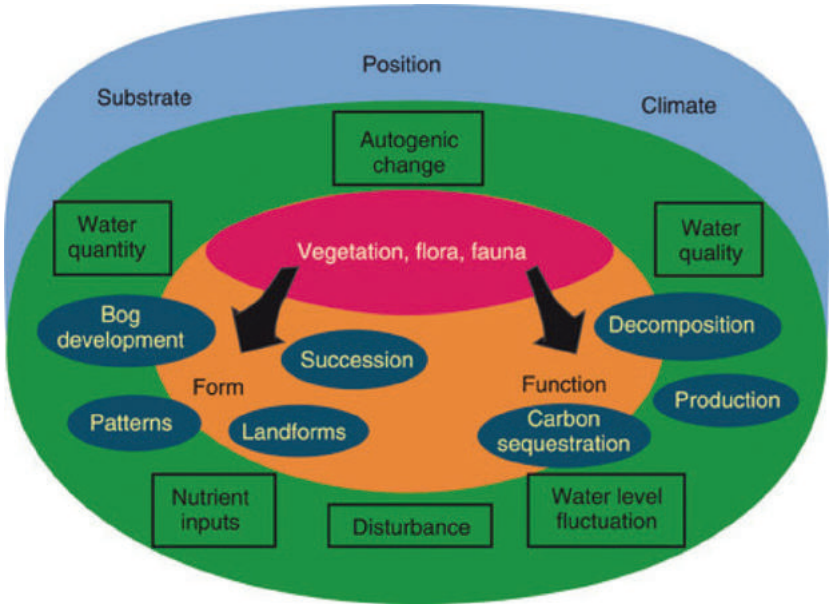


Figure 2 Substrate, position, and climate are regional factors that influence six local factors that are shown in boxes in the diagram. These local drivers direct both the form and function of bogs and fens. Adapted from Vitt DH (2006) Peatlands: Canada's past and future carbon legacy. In: Bhatti J, Lal R, Price M, and Apps MJ (eds.) *Climate Change and Carbon in Managed Forests*, pp. 201–216. Boca Raton, FL: CRC Press.

determine a number of site-specific factors that influence individual peatland sites. These local factors include rate of water flow, quantity of nutrient inputs, the overall chemistry of the water in contact with the peatland, and the amount of water level fluctuation. Additionally, there are a number of internal, or autogenic, processes that help regulate peatland form and function (Figure 2). These allogenic and autogenic factors operate in an ever-changing world of disturbance that includes natural disturbances, especially wildfire, as well as anthropogenic disturbances such as mining, forestry, and agriculture.

Peatland form and function are dependent on the process of peat accumulation and the pattern of loss or gain of carbon from habitats. Peat accumulation is dependent on the input of organic matter produced by photosynthesis. This organic matter is first accumulated in the upper, aerobic (or acrotelm) peat column wherein relatively rapid rates of decomposition occur. The rate at which this partially decomposed organic matter is deposited into the water-saturated, anaerobic peat column (the catotelm), wherein the rate of decomposition is extremely slow, largely determines the amount of carbon that will accumulate at a given site. Thus, the amount of carbon, and hence the quantity of peat, that is deposited at a peatland site is dependent on photosynthesis, aerobic decomposition within the acrotelm, and subsequent anaerobic processes in the catotelm, including methanogenesis and sulfate reduction. Recent examination of many peat cores across the northern high latitudes provides evidence that the carbon accumulated over the past 1000 years is related to contemporary growing season length and photosynthetically active radiation. This analysis suggests that net primary production is more important than decomposition in long-term carbon accumulation.

Peatland Types

Peat-forming wetlands are in general ecosystems that have accumulated sufficient organic matter over time to have a well-developed layer of peat. In many soil classifications, this is defined as soils having greater than 30% organic matter that forms deposits greater than 30–40 cm in depth. Non-peat-forming wetlands such as marshes (wetlands without trees) and swamps (wetlands dominated by a tree layer) mostly have less than 30–40 cm of accumulated organic material and over time have not been able to sustain continued accumulation of a carbon-rich peat deposit. Numerous classifications have been proposed that distinguish between various peatland types. For example, peatlands have been classified based on the source of water that has the primary influence on the peatland. Thus, peatlands that are influenced by water that has been in contact with soil or lake waters are termed geogenous and are divided into three types. Peatlands may be topogenous (influenced by stagnant water, mostly soil water, but also nonflowing water bodies as well), limnogenous (influenced by flood water from water courses resulting in lateral flow away from the direction of stream flow), or soligenous (influenced by flowing water, especially sheet flow on gentle slopes, including seepages and springs). Contrasted to these geogenous types of peatlands, others may be ombrogenous (influenced only by rain water and snow).

Peatlands are extremely variable in vegetation structure; they may be forested (closed canopy), wooded (open canopy), shrub dominated, or sedge dominated. Ground layers may be moss dominated, lichen dominated, or bare. Finally, peatlands vary as to where they occur on the landscape: in association with streams, lakes, springs, and seeps or isolated at higher elevations in the watershed. Peatlands often occur on the landscape as 'complex peatlands', wherein several distinctive peatland types occur together (**Figure 3**). Finally, and perhaps most universally utilized, is a classification that combines aspects of hydrology, vegetation, and chemistry into a functional classification of peat-forming wetlands. In general, this view of peatlands would consider hydrology as fundamental to peatland function and recognize two peatland types – fens and bogs.

Fens are peatlands that develop under the influence of geogenous (or minerogenous waters) waters that influence the peatland after being in contact with surrounding mineral, or upland, substrates). Waters contacting individual peatlands have variable amounts of dissolved minerals (especially base cations (Na^+, K^+, Ca^{2+}, Mg^{2+}) and associated anions (HCO_3^-, SO^{2-}, Cl^-), and may also vary in the amount of nutrients (N and P) as well as the number of hydrogen ions. Further complicating this minerotrophy is variation in the flow of water, including amount of flow and as well as source of the water (surface, ground, lake, or stream). Peatlands receiving water only from the atmosphere via precipitation are hydrologically isolated from the surrounding landscape. These ombrogenous peatlands, or bogs, are ombrotrophic ecosystems receiving nutrients and minerals only from atmospherically deposited sources.

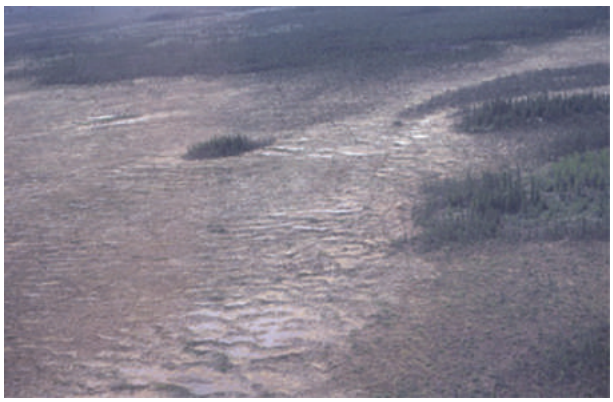


Figure 3 Peatland complex in northern Alberta, Canada. Patterned fen in left foreground, bog island with localized permafrost (large trees) and melted internal lawns to left, and curved treed bog island to right background. Small tree-covered, oval island in center is upland.

In summary and from a hydrological perspective, in fens water flows into and through the peatland after it has been in contact with surrounding materials, whereas in bogs water is deposited directly on the peatland surface and then flows through and out of the bog directly onto the surrounding landscape. Thus, fens are always lower in elevation than the surrounding landscape, while bogs are slightly raised about the connecting upland areas.

The recognition that hydrology is the prime factor for dividing peatlands into fens and bogs dates back to the early 1800 s. However, in the 1940s, Einar DuReitz recognized that vegetation composition and floristic indicators could be used to further characterize bogs and fens. Somewhat later, Hugo Sjörs associated these floristic indicators with variation in pH and electrical conductivity (as a surrogate for total ionic content of the water). The results of these early field studies in Sweden provided an overarching view of how hydrology, water chemistry, and flora are associated, and more recent studies delineate how these combined attributes together form a functional classification of northern peatlands that provides an ecosystem perspective.

Bogs

Bogs are functionally ombrotrophic. At least in the Northern Hemisphere, they have ground layers dominated by the bryophyte genus *Sphagnum* (**Figure 4**). Sedges (*Carex* spp.) are absent or nearly so. The shrub layer is well developed and trees may or may not be present. Nearly, all of the vascular plants have associations with mycorrhizal fungi. Microrelief of raised mounds (hummocks) and depressions (hollows) is generally well developed. The peat column consists of a deep anaerobic layer (the catotelm), wherein decompositional processes are extremely slow and a surficial layer of 1–10 dm of the peat column that occupies an aerobic zone (the acrotelm). The acrotelm extends upward from the anaerobic catotelm and is mostly made up of living and dead components of *Sphagnum* plants, wherein vascular plant roots and fallen vascular plant aboveground litter occur. Well-developed acrotelms are unique to ombrotrophic bogs and provide opportunities to study atmospheric deposition and ecosystem response to such deposition.

Bogs are acidic ecosystems that have pH's of around 3.5–4.5. Base cations are limited owing to the ombrogenous source of water and to the cation exchange abilities of *Sphagnum* (see below). Bicarbonate is lacking in bogs and carbon is dissolved in the water column only as CO_2. The lack of geogenous waters limits nutrient inputs to those derived only from atmospheric deposition, and thus nitrogen and phosphorus are in short supply.

Bogs appear to be limited in distribution to areas where precipitation exceeds potential evapotranspiration. In many oceanic regions of the Northern Hemisphere (especially Britain, Ireland, Fennoscandia, and coastal eastern Canada), bogs form large treeless expanses. In Europe, the Ericaceous shrub, *Calluna vulgaris*, forms a characteristic component of these treeless landscapes. Many of these oceanic bogs are patterned, with a series of pools of waters separated by raised linear ridges. This sometimes spectacular pool/ridge topography forms either concentric or eccentric patterns (**Figure 5**), with water flowing from the highest raised center of the bog to the lower surrounding edges. Runoff from the surrounding upland (and from the raised bog itself) is concentrated at the margins of these raised bogs and due to increased nutrients, decomposition processes are greater and peat accumulation somewhat less. Thus, the central, open, raised 'mire expanse' part of a bog is surrounded by a wetter, often shaded lagg, or moat, and this 'mire margin' zone may be dominated by plants indicative of fens. Some oceanic bogs have a rather flat mire expanse, with occasional pools of water. Whereas the mire expanse surface of these raised bogs is flat, the dome of water contained within the bog peat is convex and thus the driest part of the bog is at the edges just before contact with the fen lagg. This marginal, relatively dry upslope to the mire expanse is usually treed and is termed the 'rand'.

In continental areas, bogs have a very different appearance (**Figure 6**). These continental bogs have a conspicuous tree layer and abundant shrubs (mostly *Ledum* spp. or *Chamaedaphne calyculata*) while pools of water are not present. In North America, the endemic tree species, *Picea mariana*, dominates these continental bogs, while in Russia bogs have scattered individuals of *Pinus sylvestris*. Farther north in the subarctic and northern boreal zones, peat soils contain permafrost. When entire bog landforms are frozen, the bog becomes drier and dominated by lichens (especially species of the reindeer lichen, *Cladina*). Unfrozen or melted

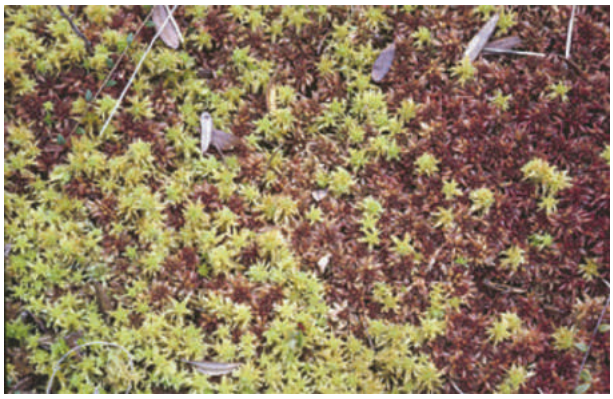


Figure 4 Mixed lawn of the peatmosses: *Sphagnum angustifolium*, mostly to the left, and *S. magellanicum* (red), mostly to right.

Figure 5 An oceanic eccentric bog. Maine, USA. Highest elevation of bog is to center left, with elongate axis sloping to distant right. Photo is courtesy of Ronald B. Davis.

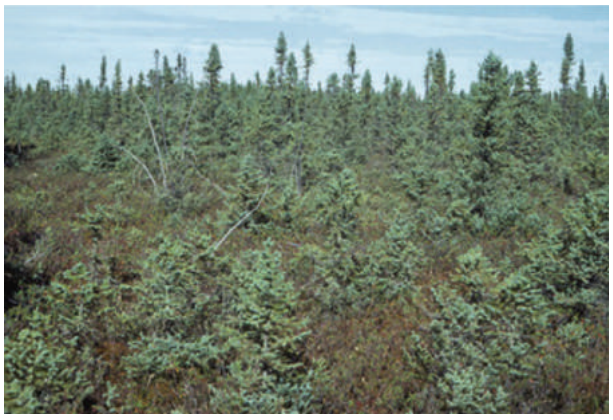


Figure 6 A continental ombrotrophic bog from western Canada. Tree species is *Picea mariana* (black spruce).

areas contained within these peat plateaus are easily recognized features termed collapse scars (Figure 7). Peat plateaus form extensive landscapes across the subarctic zone of both North America and Siberia. Farther south in the boreal zone, bog landforms may contain only scattered pockets of permafrost (frost mounds), that over the past several decades have been actively melting. Recent melting of the raised frost mounds results in collapse of the mound and active revegetation by fen vegetation to form wet, internal lawns with associated dead and leaning trees (Figure 8).

Fens

Fens are peatlands that are minerotrophic that when compared to bogs have higher amounts of base cations and associated anions. All fens have an abundance of *Carex* and *Eriophorum* spp. and water levels at or near the surface of the peat (thus acrotelms are poorly developed). Unlike bogs that are characterized by high microrelief of hummocks and hollows, fens feature a more level topography of extensive carpets and lawns dominated by species of mosses (Figure 9). Depending on the characteristics of the surrounding water, fens can by divided into three types.

Poor fens

These *Sphagnum*-dominated peatlands are associated with acidic waters (pH 4.5–5.5) that contain the least amount of base cations and little or no bicarbonate alkalinity.

Rich fens

True mosses dominate the ground layer of rich fens, especially a series of species that are red-brown in color and often termed 'brown mosses'. Examples of important genera would be *Drepanocladus*, *Hamatocaulis*, *Warnstorfia*, *Meesia*, *Campylium*, *Calliergon*, and *Scorpidium*. Waters have pH varying from 5.5 to more than 8.0 and base cations are relatively abundant, especially calcium. Alkalinity varies from very little to extremely high amounts of bicarbonate. Rich fens occur as two types centered on the chemistry of the pore waters. 'Moderate-rich fens' have pH values between 5.5 and 7.0, with little alkalinity. Both brown mosses and some mesotrophic species of *Sphagnum* (e.g., *S. obtusum*, *S. subsecundum*, *S. teres*, and *S. warnstorfii*) dominate the ground layer. 'Extreme-

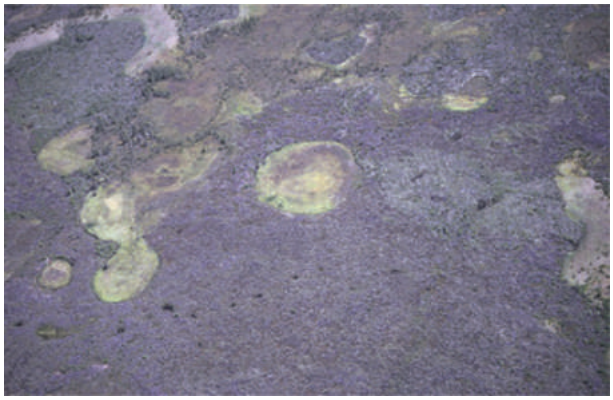


Figure 7 Extensive peat plateaus with permafrost (whitish areas dominated by the reindeer lichens in the genus *Cladina*), with isolated collapse scars (without permafrost – greenish circular to oblong areas), and with lush growth of *Sphagnum* species and sedges.

Figure 8 Bog dominated by *Picea mariana* in background, with dead snags in foreground, indicating recent permafrost collapse and the formation of an internal lawn and dominated by carpet and lawn species of *Sphagnum*.

rich fens' are bicarbonate-rich peatlands, often with deposits of marl (precipitated $CaCO_3$) and pH ranging from around neutral to over 8.0. Species of *Scorpidium*, *Campylium*, and *Hamatocaulis* dominate the ground layer.

Whereas water quality (= chemistry) is the main factor controlling fen type and flora, water quantity (= flow) controls vegetation structure and surface topography. Fens, whether poor or rich, are vegetationally extremely variable, ranging from sites having abundant trees (dominated by *Larix laricina* in North America), to sites dominated by shrubs (mostly *Betula*, *Alnus*, and *Salix*), to sites having only sedges and mosses. Topographically, fens may be homogeneous and dominated by lawns and carpets. However, as water flowing through the fen increases, the surface vegetation develops a reticulation of wet pools and carpets separated by slightly raised ridges. Further increase in flow of water directs the patterns into linear pools (some filled with floating vegetation = carpets), sometimes termed flarks, alternating with linear ridges (termed strings; **Figure 10**). These pool/string complexes are oriented perpendicular to water flow, with smaller pools always upstream from the larger ones. Especially prevalent in Scandinavia and Russia, these patterned fens and associated bog islands form extensive peatlands termed aapamires.

Important Processes in Peatlands

Acidification

Sphagnum species have cell walls rich in uronic acids that in aqueous solution readily exchange a hydrogen ion for a base cation. The base cations that are in solution in bogs and poor fens are received by the peatland from atmospheric deposition or inflowing water and are always associated with an inorganic anion (HCO_3^-, SO_4^{2-}, Cl^-). When the base cation is exchanged for the organically produced H^+, acidity of the peatland waters is produced. This acidity thus originated through the exchange of an inorganic base cation for an H^+ produced by *Sphagnum* growth – hence this is termed inorganic acidity. Inorganic acidity relies on the presence of base cations and can only produce acidity when base cations are present in the pore water to exchange. Inorganic acidity is an extremely powerful process when abundant base cations are present such as in rich fens transitional to poor fens and in poor fens. In bogs, with limited supplies of base cations due to their ombrogenous water supply, inorganic acidity is less important.

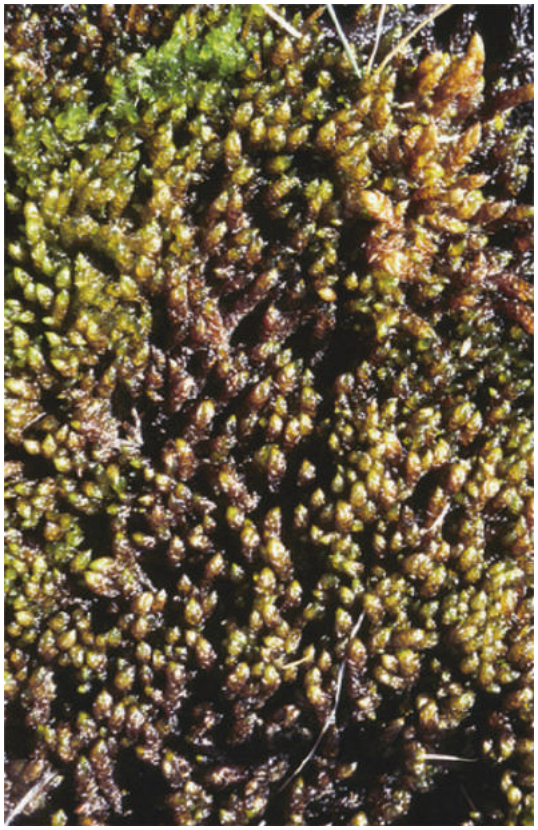


Figure 9 A carpet of the brown moss, *Scorpidium scorpioides*, a characteristic species of rich fens.

Figure 10 A patterned fen in western Canada characterized by elongate pools (flarks) separated by raised ridges (strings), oriented perpendicular to water flow.

Organic material produced by plants is decomposed and carbon mineralized through bacterial and fungal respiration. Under aerobic conditions, bacteria break down long cellulose chains and in doing so eventually produce short-chained molecules that are small enough to be dissolved in the pore waters. This dissolved organic carbon (DOC) may be lost to the peatlands via runoff or may remain suspended in the pore waters for some length of time. These decompositional processes produce acidity through dissociation of humic acids, acidity that is completely produced via organic processes; hence, peatland acidity produced via decompositional processes, and extremely important in ombrotrophic bogs, is termed organic acidity.

Rich fens, with pH above 7.0, also accumulate deep deposits of peat and are well buffered by large inputs of bicarbonate alkalinity. With continued inputs of bicarbonate, rich fens may remain stable for millennia, dominated by brown mosses that have less capacity for inorganic acidification, but strong tolerance for the alkaline peatland waters. However, as rich fens accumulate peat to depths of several meters, there is the possibility that the active surface layer will become more isolated from the bicarbonate inputs and alkalinity may decrease to the point that some tolerant species of *Sphagnum* may invade. If *Sphagnum* species establish,

then cation exchange proceeds, acidity increases while alkalinity decreases, and rich fen plant species are replaced by poor fen species tolerating acidic conditions. This acidification of rich fens has been documented in the paleorecord wherein the change from rich fen to poor fen vegetation takes place extremely rapidly, perhaps in the order of 100–300 years. As a result, these transitional rich fen–poor fen communities are short-lived on the landscape and among the most rare of peatland types.

Water Retention

The surface of a peatland lies on a column of water contained within the peat column. The peatland surface consists of a nearly complete cover of mosses (either peat mosses (*Sphagnum*) or true mosses (brown mosses)) that are continually pushed upward by the accumulating peat. This upward growth is limited only by the abilities of the peat and living moss layer to maintain a continuous water column that allows the living moss layer to grow. The vascular plants that grow in this water-soaked peat column produce roots that are largely contained in the small upper aerobic part of the peat. The mosses, however, alive and growing only from their uppermost stem apices, must maintain contact with the water column; thus, wicking and retaining of water above the saturated water column is paramount for maintenance of the moss layer. Peatland mosses have special modifications that help in this regard. Although some brown mosses have adaptations for water retention, such as the development of a tomentum of rhizoids along the stems, numerous branches along the stem that provide small spaces for capillarity, and leaves that have enlarged bases that retain water, it is in species of *Sphagnum* where water retention (up to 20 times dry plant weight) is greatly enhanced through a number of morphological modifications. *Sphagnum* has unistratose (one-celled thick) leaves consisting of alternating, large, dead, hyaline cells and small, partially enclosed, living, green cells. The walls of the hyaline cells are perforated with pores and are strengthened by the presence of cross-fibrils. Stems and branches are often encased in an outer layer of one or more rows of dead, enlarged cells. All of these hyaline cells have lost their living cell contents very early on in development and as a result the ratio of carbon to nitrogen is high. In addition to the features that allow the plants to hold water internally, the entire *Sphagnum* plant is a series of tiny spaces that serve as reservoirs for capillarity. The branches are surrounded by numerous, overlapping, very concave branch leaves (one-cell thick). The branches are attached to the stem in fascicles of three to five branches, half of which hang along the stem and half extend outward at more or less 90°. The fascicles of branches originate at the stem apex, and slowly develop while still close together at the apex of the stem. This group of maturing branches, the capitulum, along with the top 1–5 cm of mature stem and associated branches form a dense canopy. In total, this canopy (**Figure 11**) consists of numerous small spaces of different sizes and, along with the dead hyaline cells of the leaves and branches, provides the mechanism for wicking and retention of capillary water far above the actual water table, which in turn provides the framework for the aerobic peat column that is so characteristic of bogs.

Nutrient Sequestration (Oligotrophification)

Peat forms due to slow decompositional processes that allow organic materials to be deposited as peat. As organic material is deposited, it contains within its carbon matrix nutrients, especially nitrogen and phosphorus, which were originally incorporated in the cell structure of the living plants, especially those of *Sphagnum* and brown mosses. Relatively rapid decomposition in the acrotelm mineralizes only a portion of the total nutrients tied up in the plant material, making these available for further plant growth as well as fungal and bacterial processing. However, upon entry to the catotelm, almost all decompositional activity stops and the nutrients become tied to organic materials in unavailable forms. Thus, rather than being recycled and remaining available

Figure 11 A longitudinal view of the canopy of *Sphagnum*; each stem is terminated by a capitulum of young branches. The branches along the stem are covered with numerous overlapping leaves and organized into fascicles that have branches that hang down along the stem as well as branches that spread outward from the stem allowing the individual stems to be evenly spaced from one another.

for new plant growth, nitrogen and phosphorus become part of long-term unavailable nutrient pools. The lack of ability to utilize this unavailable pool of nutrients causes peatlands over time to become more oligotrophic at their surface yet also having large amounts of stored nitrogen and phosphorus. For example, *Sphagnum* peat is generally about 1% nitrogen; however, almost all of this catotelmic nitrogen is unavailable for plant and microorganism use while in place in the peat deposit. When exposed to the atmosphere (e.g., as a garden amendment), the carbon is oxidized to CO_2 and the nitrogen is mineralized to NO_3^- and NH_4^+ and available for plant uptake. Although the actual percent of nitrogen, and other nutrients, may not be as high as that in inorganic soils, the total amount in the soil within any one square meter surface area of the peatland is greater in peat soils due to the depth of the peat present. This oligotrophification, and consequently nutrient storage, is autogenetically enhanced through the buildup of the peat column, placing the peat surface farther from the source of minerotrophic inputs. The long-term result of oligotrophication is the regional storage of large pools of carbon and nitrogen, with stored nitrogen estimated at 9–16% of the global terrestrial soil pool. In ombrotrophic bogs, these high amounts of stored nitrogen are seemingly paradoxical to nitrogen inputs only from the atmosphere, mostly in areas where pristine bogs receive some of the lowest atmospheric inputs globally (about 1 kg N ha^{-1} yr^{-1}). Across Canada, these ombrotrophic peatlands contain approximately 8–40 times more nitrogen in their upper peat columns than can be explained by atmospheric inputs alone. Recently, this missing pool of nitrogen has been accounted for by high rates of N-fixation from prokaryotes associated with *Sphagnum* canopies.

Methane Production

Methane is a highly potent greenhouse gas that originates from both natural and anthropogenic origins. On a weight basis, methane is 21 times more efficient at trapping heat and warming the planet than carbon dioxide. Methane emissions from wetlands account for more than 75% of the global emissions from all natural sources. Methane is a highly reduced compound produced as the end product of anaerobic decomposition by a group of microorganisms called methanogens, which phylogenetically belong to *Archaea*. These strict anaerobes can utilize only a limited variety of substrates with H_2–CO_2 and acetate being the most important too. The H_2–CO_2 dependent methanogenesis is considered the dominant pathway of methane production in boreal peatlands. However, acetate-dependent methanogenesis sometimes dominates in fens. In rich fens, higher nutrient availability promotes the growth of vascular plants (primarily sedges). Roots of these vascular plants penetrate deep into the peat column and therefore transport potential carbon-rich substrates, such as acetate, into the anaerobic layer. Rapid decomposition of organic matter also provides abundant substrates for methanogens. Poor fens, with lower vascular plant cover than that of rich fens, generally have lower potentials for CH_4, and a higher portion of the produced CH_4 comes from H_2–CO_2. Similar to poor fens, *Sphagnum*-dominated bogs also have a higher proportion of CH_4 produced from H_2–CO_2, and it may be that the dominance of mosses (without roots) and mycorrhizal vascular plants (without deep carbon-rich roots), along with the reduced abundance of sedges with well-developed deep roots, prohibit movement of labile carbon substrates to the anaerobic peat layer. Low decomposition rates in acidic bogs also limit the amount of acetate that can be produced during peat decomposition, which in turn limits the acetoclastic pathway. Methanogen diversity in bogs is very low and the composition of the methanogen community in bogs also differs greatly from that characteristic of fens. In general, higher CH_4 production is found in peatlands with higher vascular plant cover, and higher water tables are found in rich fens.

Sulfate Reduction

In peatlands, sulfur occurs in several different redox states (S valences ranging from +6 in SO_4^{2-} to −2 in hydrogen sulfide (H_2S), S-containing amino acids, and other compounds), and conversions between these states are the direct result of microbially mediated transformations. In bogs, the sole sulfur input is via atmospheric deposition, while in fens atmospheric deposition can be augmented by surface and/or groundwater inputs, which may contain sulfur derived from weathering of minerals in rock and soil. Regardless of the sulfur source, when sulfur enters a peatland, there are a variety of pathways through which it can cycle. In the aerobic zone, sulfate can be adsorbed onto soil particles, or assimilated by both plants and microbes. In the anaerobic zone, sulfate can also be adsorbed onto soil particles, assimilated by plants or microbes, or reduced by sulfate-reducing bacteria through the process of dissimilatory sulfate reduction. Dissimilatory sulfate reduction is a chemoheterotrophic process whereby bacteria in at least 19 different genera oxidize organic matter to meet their energy requirements using sulfate as the terminal electron acceptor. Thus, this process is one way in which carbon is lost from the catotelm. If the sulfate is reduced by sulfate-reducing bacteria, the end product (S^{2-}) can have several different fates. In the catotelm, where $S^{2-}$is formed, it can react with hydrogen, to produce H_2S gas, which can diffuse upwardly into or through the acrotelm where it can be either oxidized to sulfate, or lost to the atmosphere. Alternatively, H_2S can react by nucleophilic attack with organic matter to form organic or C-bonded sulfur (CBS). If Fe is present, S^{2-}can react with Fe to form FeS and FeS_2 (pyrite), which is referred to as reduced inorganic sulfur (RIS). The RIS pool tends to be unstable in peat and can be reoxidized aerobically with oxygen if the water table falls, or anaerobically probably using Fe_3^+ as an anaerobic electron acceptor. If Hg is present, and combines with $S^{2-}$to form neutrally charged HgS, then Hg sulfide is capable of passive diffusion across cell membranes of bacteria that methylate Hg. Alternatively, bacteria can transfer the methoxy groups of naturally occurring compounds, such as syringic acid, to S^{2-}, and form methyl sulfide (MeSH) or dimethyl sulfide (DMS), although the exact mechanisms by which this occurs are still unknown.

Initiation and Development of Peatlands

Peatlands initiate in one of four ways. The first, the most common, appears to through paludification (or swamping), wherein peat forms on previously drier, vegetated habitats on inorganic soils and in the absence of a body of water, generally due to regional water table rise and associated climatic moderation. Additionally, local site factors also have strong influences on paludification. Second, peat may form directly on fresh, moist, nonvegetated mineral soils. This primary peat formation occurs directly after glacial retreat or on former inundated land that has risen due to isostatic rebound. Third, shallow bodies of water may gradually be filled in by vegetation that develops floating and grounded mats – thus terrestrializing the former aquatic habitat. Both lake chemistry and morphometry as well as species of plants in the local area influence the rates and vegetative succession. Fourth, peat may form and be deposited on shallow basins once occupied by extinct Early Holocene lakes. These former lake basins, lined with vegetated impervious lake clays, provide hydrologically suitable sites for subsequent peat development.

Across the boreal zone, peatland initiation appears to be extremely sensitive to climatic controls. For example, in oceanic areas, peatlands often initiated soon after glacial retreat some 10 000–12 000 years ago. Many of these oceanic peatlands began as bogs and have maintained bog vegetation throughout their entire development. In more continental conditions, most peatlands were largely initiated through paludification. In areas where the bedrock is acidic, most of these early peatlands were poor fens, whereas in areas where soils are base rich and alkaline, rich fens dominated the early stages. Like oceanic peatlands, subcontinental peatlands initiated soon after glacial retreat; however, throughout most of the large expanses of boreal Canada and Siberia, peatland initiation was delayed until after the Early Holocene dry period, initiating 6000–7000 years ago. Many of these peatlands initiated as rich or poor fens and have remained as fens for their entire existence, whereas others have undergone succession and today are truly ombrotrophic bogs. A recent study in western Canada correlated peak times of peatland initiation to Holocene climatic events that are evident in US Midwest lakes, North Atlantic cold cycles, and differing rates of peat accumulation in the one rich fen studied in western Canada.

Peatlands as Carbon Sinks

Peat is about 51% carbon and peatlands hold about 270–370 Pg (petagram) of carbon or about one-third of the world's soil carbon. For example in Alberta (Canada), where peatlands cover about 21% of the provincial landscape, the carbon in peatlands amounts to 13.5 Pg compared to 0.8 Pg in agricultural soils, 2.3 Pg in lake sediments, and 2.7 Pg in the province's forests. Estimates for apparent long-term carbon accumulation in oceanic, boreal, and subarctic peatlands range from around 19 to 25 g C m^{-2} yr^{-2}. However, disturbances can have a dramatic effect on carbon accumulation. Wildfire, peat extraction, dams and associated flooding, mining, oil and gas extraction, and other disturbances all reduce the potential for peatlands to sequester carbon, while only permafrost melting of frost mounds in boreal peatlands has been documented to have a positive effect on carbon sequestration. One recent study has suggested that effects from disturbance in Canada's western boreal region have reduced the regional carbon flux (amount of carbon sequestered in the regional peatlands) from about 8940 Gg (gigagram) C yr^{-1} under undisturbed conditions to 1319 Gg carbon sequestered per year under the present disturbance regime, yet only 13% of the peatlands have been affected by recent disturbance. These data suggest that although for the long-term peatlands in the boreal forest region have been a carbon sink and have been removing carbon from the atmosphere, at the present time, due to disturbance, this capacity is greatly diminished. Furthermore, when disturbance is examined in more detail, it is wildfire that is the single greatest contributor to loss of carbon sequestration, both from a direct loss as a result of the fire itself as well as from a loss of carbon accumulation due to post-fire recovery losses. If wildfire greatly increases as is predicted by climate change models, then the effectiveness of peatlands to sequester carbon may be greatly reduced and it has been proposed that an increase of only 17% in the area burned annually could convert these peatlands to a regional net source of carbon to the atmosphere. If boreal peatlands become a source for atmospheric carbon, then the carbon contained within the current boreal peatland pool, in total, is approximately two-thirds of all the carbon in the atmosphere.

Further Reading

Bauerochse A and Haßmann H (eds.) (2003) *Peatlands: Archaeological sites–archives of nature–nature conservation-wise use*, Proceedings of the Peatland Conference 2002 in Hanover, Germany, Hanover: Verlag Marie Leidorf GmbH.

Davis RB and Anderson DS (1991) *The eccentric bogs of Maine: A rare wetland type in the united states*. Orono: Maine Agricultural Experiment Station, Technical Bulletin 146.

Feehan J (1996) *The bogs of Ireland: an introduction to the natural, cultural and industrial heritage of Irish peatlands*. Dublin: Dublin Environmental Institute.

Fraser LH and Kelly PA (eds.) (2005) *The World's largest wetlands: Their ecology and conservation*. Cambridge: Cambridge University Press.

Gore AJP (1983) *Ecosystems of the world. Mires – swamp, bog, fen and moor*. Amsterdam: Elsevier Scientific, 2 vols.

Joosten H and Clarke D (2002) *Wise use of mires and peatlands – background and principles including a framework for decision-making*. Jyväskylä, Finland: International Mire Conservation Group and International Peat Society. http://www.mirewiseuse.com.

Larsen JA (1982) *The ecology of the northern lowland bogs and conifer forests*. New York: Academic Press.

Moore PD (ed.) (1984) *European mires*. New York: Academic Press.

Moore PD and Bellamy DJ (1974) *Peatlands*. London: Elek Scientific.

National Wetlands Working Group (1988) *Wetlands of Canada. Ecological land classification series, No. 24*. Ottawa: Sustainable Development Branch, Environment Canada, and Montreal: Polyscience Publications.

Parkyn L, Stoneman RE, and Ingram HAP (1997) *Conserving peatlands*. NewYork: CAB International.
Vitt DH (2000) Peatlands: ecosystems dominated by bryophytes. In: Shaw AJ and Goffinet B (eds.) *Bryophyte biology*, pp. 312–343. Cambridge: Cambridge University Press.
Vitt DH (2006) Peatlands: Canada's past and future carbon legacy. In: Bhatti J, Lal R, Price M, and Apps MJ (eds.) *Climate change and carbon in managed forests*, pp. 201–216. Boca Raton, FL: CRC Press.
Wieder RK and Vitt DH (eds.) (2006) *Boreal peatland ecosystems*. Berlin, Heidelburg, New York: Springer.
Wright HE Jr., Coffin BA, and Aaseng NE (1992) *The patterned peatlands of Minnesota*. Minneapolis: University of Minnesota Press.

Polar Terrestrial Ecology

TV Callaghan, Royal Swedish Academy of Sciences Abisko Scientific Research Station, Abisko, Sweden

The polar regions are situated at latitudes beyond which the Earth's angle to the Sun is shallow and the input of thermal radiation is low. During the winter period, the Sun is below the horizon and there are prolonged periods of darkness. The resulting low-temperature regimes dominate ecological processes, either directly by affecting plant growth, microbial activity, animal behavior, organism reproduction and survival, or indirectly by controlling the length of the snow- and ice-free periods in which most primary production and dependent biological activity occurs, the availability of water in liquid form, and the expansion and contraction, and other active layer properties in generally primitive soils underlain by permafrost. Feedback mechanisms from polar regions and their ecological systems to the climate system affect local, regional, and global climate. The balance between greenhouse gas emissions from decomposition, particularly soil microbial respiration, and photosynthesis has resulted in a large net accumulation of carbon in arctic soils while ice and snow that cover low, tundra vegetation reflect incoming radiation. Both mechanisms lead to cooling. In contrast, global ocean circulation leads to the redistribution of the Earth heat by cooling the tropics and warming the high latitudes.

Both the Arctic and the Antarctic are characterized by vast wilderness areas that are generally young, as most land areas, with some extensive exceptions in the Arctic, were glaciated in the Pleistocene. Polar regions host some of the Earth's most extreme environments and organisms such as snow algae, lichens that inhabit the crevices within crystalline rocks, and the simple communities of soil fauna in the dry valleys of Antarctica.

Polar environments vary between the Arctic and the Antarctic and also within each region (Figs. 1 and 2).

The Arctic is dominated by a polar ocean surrounded by continental land masses and islands, whereas the Antarctic is dominated by a polar, largely ice-covered, land mass surrounded by oceans. Terrestrial ecosystems are extensive (7.5 million km^2) and varied and stretch from the closed canopy northern boreal forests in the south, through the latitudinal treeline ecotone and tundra wetlands to the polar desert in the north. Along this latitudinal gradient, mean July temperature varies from about 12 °C in the south to 2 °C in the north, total annual precipitation varies from about 250 to 75 mm (mainly as snow), and net primary production varies from about 1000 to 1 g m^{-2} yr^{-1}. Approximately 6000 animal and 5800 plant species inhabit arctic lands (3% and 5% of global biodiversity, respectively). Biodiversity decreases geometrically along this gradient. Although plant biodiversity is low in comparison with many biomes, it is surprisingly high per square meter because of the small scale of plants, and over 6000 species of animals and plants have been cataloged in and around Svalbard at about 79° N. There is also large environmental variation associated with the climatic effects of northern ocean currents: in arctic Norway, Sweden, and Finland, forests grow north of the Arctic Circle (66.7° N) because of the warming effect of the northward-flowing Gulf Stream, whereas polar bears and tundra

Fig. 1 Ecosystems of the Arctic.

Fig. 2 Coastal ecosystem, sub-Antarctic South Georgia.

vegetation are found at about 51° N in eastern Canada because of the cooling influence of southward-flowing cold ocean currents. In the Arctic, indigenous and other arctic peoples have been part of the ecosystem for millennia.

The large land masses of the Arctic have great connectivity with land masses further south: great rivers flow from low latitudes to the Arctic Ocean, and mammals and hundreds of millions of birds migrate between the summer breeding grounds in the Arctic and overwintering areas in boreal or temperate regions. Food chains in the Arctic are more complex than those in the Antarctic and at the top of the chain are mammalian carnivores such as the polar bears, wolves, and arctic foxes. Population cycles characteristic of arctic animals together with relatively few species in each trophic level can result in ecological instability and ecological cascades: increasing numbers of snow geese in arctic Canada have denuded vegetation resulting in habitat hypersalinity.

The Antarctic land mass covers some 12.4 million km^2 but less than 1% is seasonally ice free. In the Antarctic, the major environmental variation is associated with the relatively moist and `warm' maritime climate of the west coast of the Antarctic Peninsula (temperatures are between 0 and 2 °C for 2–4 months in summer) contrasted with the cold, dry polar desert climate of the continental land mass. Consequently, most biological activity and most species are found on the west coast of the Antarctic Peninsula. Vegetation is dominated by relatively simple plant communities of lichens, mosses, and liverworts that support simple soil invertebrate communities. Only two species of higher plant and higher insects occur. Terrestrial mammals are absent and this short trophic structure, together with the isolation of the land mass, has enabled the establishment of a highly specialized, and commonly endemic fauna of ground-nesting birds (e.g., penguins) and seals that depend on the coastal land areas for breeding and moulting, and the sea, for food. Nutrients for plant growth in these areas are mainly derived from the sea and are deposited on land by wind or birds. In contrast, over much of the tundra, low nutrient availability to plants limits primary production. There are no indigenous peoples in the Antarctic and human activities there have been restricted to the past 200 years.

Human activity has, until recently, influenced both Arctic and Antarctic ecosystems less than most biomes on Earth. However, the polar amplification of global climate change together with the inherent sensitivity of polar ecological systems to invasion by species from warmer latitudes has resulted in the vulnerability of polar ecosystems which are now under threat of rapid change.

The Future: Polar Regions and Climate Change

The polar regions are undergoing rapid climate change. There is a general amplification of global warming in the Arctic: surface air temperatures have warmed at approximately twice the global rate, although there are local variations. The average warming north of 60° N has been 1–2 °C since a temperature minimum in the 1960s and 1970s with the largest increase (*c.* 1 °C per decade) in winter and spring. Continental arctic land masses together with the Antarctic Peninsula are the most rapidly warming areas of the globe. Precipitation in the Arctic shows trends of a small increase over the past century (about 1% per decade), but the trends vary greatly from place to place and measurements are very uncertain. There are reductions in Arctic sea ice, river and lake ice in much of the sub-Arctic, and Arctic glaciers. Reduction in Arctic sea ice has occurred at a rate of 8.9% per decade for September relative to the 1979 values and there was an un-predicted extreme reduction in 2007. Permafrost has warmed. Although changes in the active layer depth have no general trend, in some sub-Arctic locations, discontinuous permafrost is rapidly disappearing and changes in permafrost are driving changes in hydrology and ecosystems. In Arctic Russia, ponds are drying in the continuous permafrost zone and waterlogging is occurring where there is discontinuous permafrost.

In Antarctica, temperature trends show considerable spatial variability: the Antarctic Peninsula shows significant warming over the last 50 years, whereas cooling has occurred around the Amundsen-Scott Station at the South Pole and in the Dry Valleys. Consequently, there is no continent-wide polar amplification of global change in Antarctica.

Current polar warming is leading to changes in species' ranges and abundance and a northward and upward extension of the sub-Arctic treelines. Forest is projected to displace considerable areas of tundra in some places. Species tend to relocate, as they have in the past, rather than adapt to new climate regimes. However, this process is likely to lead to the loss of some species: polar

bears and other ice-dependent organisms are particularly at threat. In other areas, where rates of species relocation are slower than climate change, the incidence of pests, disease, and fire is likely to increase. Changes in vegetation, particularly a transition from grasses to shrubs, have been reported in the North American Arctic, and satellite imagery has indicated an increase in the `normalized difference vegetation index' (a measure of photosynthetically active biomass) over much of the Arctic. This index has increased by an average of about 10% for all tundra regions of North America, probably because of a longer growing season. However, such increases in productivity and changes in plant functional types have been shown experimentally to displace mosses and lichens that are now major components of Arctic vegetation.

In Antarctica, warming has caused major regional changes in terrestrial and marine ecosystems. The abundances of krill, Adelie, and Emperor penguins and Weddell seals have declined but the abundances of the only two native higher plants has increased. On continental Antarctica, climate change is affecting the vegetation composed of algae, lichens, and mosses. Introductions of alien species, facilitated by increased warming and increased human activity, are particular threats to southern ecosystems. Recent studies on sub-Antarctic islands have shown increases in the abundance of alien species and negative impacts on the local biota. In contrast, cooling has caused clear local impacts in the Dry Valleys where a 6–9% reduction in lake primary production and a 10% per year decline in soil invertebrates has occurred.

The responses of polar environments to climatic warming include feedbacks to the global climate system and other global impacts. Increased runoff from arctic rivers could affect the thermohaline circulation that redistributes the Earth's heat, thereby causing cooling in the North Atlantic and further warming in the tropics. Reductions in sea ice extent and snow cover together with a shift in vegetation from tundra to shrubs or forests are likely to reduce albedo (reflectivity of the surface) and lead to further warming despite the increased uptake of carbon dioxide by a more productive vegetation. Thawing permafrost is likely to release methane, a particularly powerful greenhouse gas, and evidence of this is already available from various arctic areas.

Not all impacts of climate warming in polar regions are disadvantageous to society: the reduction of sea ice in the Arctic is likely to lead to increased marine access to resources and new fisheries and reduced length of sea routes, while warming on land will probably lead to increased productivity and increased potential for forestry and agriculture.

See also: General Ecology: Ecological Efficiency

Further Reading

Anisimor, O.A., Vaughan, D.G., Callaghan, T.V., *et al.*, 2007. Polar regions (Arctic and Antarctic). In: Parry, M.L., Canziani, O.F., Palutikof, J.P., Hanson, C.E., Van der Linder, P.J. (Eds.), Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press, pp. 655–685.

Callaghan, T.V., Björn, L.O., Chapin III, F.S., *et al.*, 2005. Tundra and polar desert ecosystems. In: ACIA. Arctic Climate Impacts Assessment. Cambridge: Cambridge University Press, pp. 243–352.

Chapin III, F.S., Berman, M., Callaghan, T.V., *et al.*, 2005. Polar ecosystems. In: Hassan, R., Scholes, R., Ash, N. (Eds.), Ecosystems and Human Well-Being: Current State and Trends, vol. 1. Washington, DC: Island Press, pp. 719–743.

Convey, P., 2001. Antarctic ecosystems. In: Levin, S.A. (Ed.), Encyclopaedia of Biodiversity, vol. 1. San Diego: Academic Press, pp. 171–184.

Nutall, M., Callaghan, T.V., 2000. In: The Arctic: Environment, People, Policy. Reading: Harwood Academic Publishers, p. 647.

Richter-Menge, J., Overland, J., Hanna, E., *et al.*, 2007. State of the Arctic Report.

Walther, G.R., Post, E., Convey, P., *et al.*, 2002. Ecological responses to recent climate change. Nature 416 (6879), 389–395.

Riparian Wetlands

KM Wantzen, University of Konstanz, Konstanz, Germany
WJ Junk, Max Planck Institute for Limnology, Plön, Germany

Introduction

The riparian zone of running water systems is a site of intensive ecological interactions between the aquatic and the terrestrial parts of the stream valley. Wetlands that occur in this zone exchange water with the aquifer and with the main channel during flood events (Fig. 1). Riparian wetlands are buffer zones for the water budget of the landscape: they take up excess water from flood events and release it gradually afterwards.

Modern ecological theory recognizes the important role riparian wetlands play for biodiversity and for the energy and matter budgets along the whole range of river courses. The carbon and nutrient budgets are influenced by dissolved and particulate substances from the bordering terrestrial ecosystems, by the autochthonous production from the wetland plants, and by allochthonous organic matter delivered by the floodwater. The proportions between these sources are defined by the hydrological patterns, landscape morphology, and climatic conditions.

The crossover between humid and dry conditions creates habitats for organisms coming from either aquatic or terrestrial ecosystems, and for those biota that are specialized on wetland conditions. As the transversal dimension of streamside wetlands is generally small, their overall importance for landscape ecology, biogeochemistry, and biodiversity is often overlooked. However, the total size of these wetlands can be considerable in areas with dense stream networks. Moreover, the corridor-shaped extension of riparian wetlands makes them perfect pathways for the gene flow between remote populations of aquatic and terrestrial biota. Many ecological services are uniquely provided by riparian wetlands, including erosion control, filtering of nutrients and pesticides from adjacent cropland, mitigation of floods, and recreation, which increases their conservation value in a socioeconomic context.

There is a large array of environmental conditions that vary between the different types of riparian wetlands, especially climatic region and prevailing vegetation type, and landscape morphology and hydrologic patterns. This article deals with the different types of riparian wetlands, their deterministic environmental conditions, prevailing ecological processes, typical biota, and aspects of conservation.

Definitions and Concepts

There are many definitions of riparian wetlands. A hydrological definition defines riparian wetlands as

> lowland terrestrial ecotones which derive their high water tables and alluvial soils from drainage and erosion of adjacent uplands on the one side or from periodic flooding from aquatic ecosystems on the other (McCormick, 1979)

A functional definition states that

> riparian areas are three-dimensional ecotones of interaction that include terrestrial and aquatic ecosystems, that extend down to the groundwater, up above the canopy, outward across the floodplain, up the near-slopes that drain to the water, laterally into the terrestrial ecosystem, and along the water course at a variable with (Ilhard *et al.*, 2000).

Both definitions point to the ecotonal character of riparian wetlands between water bodies on one side and the upland on the other. Riparian wetlands can be, at the smallest scale, the immediate water's edge where some aquatic plants and animals form a distinct community, and pass to periodically flooded areas of a few tens of meters width. At medium scale they form bands of vegetation, and at the largest scale they form extended floodplains of tens of kilometers width along large rivers. In this case, complexity of the riparian wetlands increases so much that many scientists give them the status of specific ecosystems.

There are several concepts that deal with different aspects of stream and river ecology but two of them are of specific interest to rivers and riparian zones. The `river continuum concept' (RCC) of Vannote *et al.*, describes the longitudinal processes in the river channel and the impact of the riparian vegetation on the physical and chemical conditions and as carbon source to the aquatic communities in the channel. The `flood pulse concept' (FPC) of Junk *et al.* stresses the lateral interaction between the floodplain and the river channel and describes the specific physical, chemical, and biological processes and plant and animal communities inside the floodplain. The predictions of the RCC fit well for rivers with narrow riparian zones but with increasing lateral extent and complexity of the riparian zone the FPC becomes more important. Here, we restrict our discussion to riparian wetlands along streams and low-order rivers. Since lateral extent of the riparian zone along low-order rivers can vary considerably in different parts of the same river or between different rivers of the same river order, the applicability of the concepts may also vary.

 doi:10.1016/B978-0-444-63768-0.00352-8

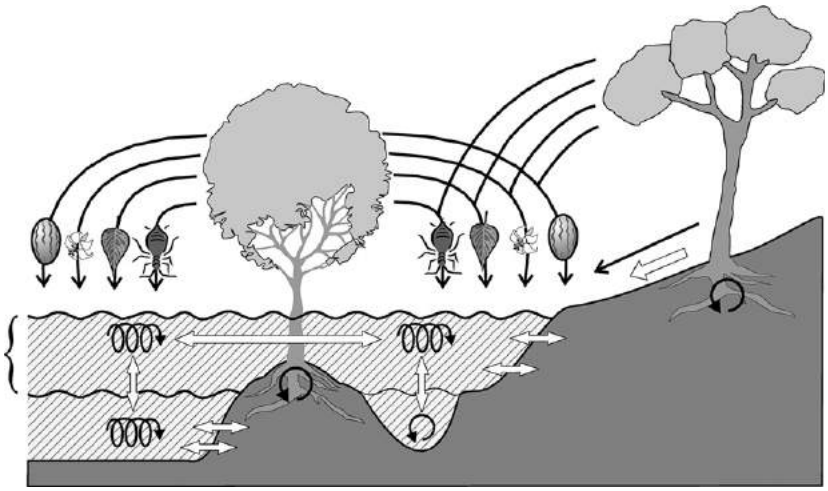


Fig. 1 Inputs, turnover, and exchange of organic matter in the stream channel (left) and a riparian wetland water body (center) at low and high water levels. Black arrows indicate organic matter inputs, white arrows indicate water exchange pathways, spirals indicate nutrient spiralling or downriver transport, and circular arrows indicate sites of organic matter turnover *in situ*. Curly brace indicates water-level fluctuations during flood events. Modified from Wantzen, K.M., Yule, C., Tockner, K., Junk, W.J., 2006. Riparian wetlands. In: Dudgeon, D. (Ed.), Tropical Stream Ecology. Amsterdam: Elsevier, pp. 199–217.

Environmental Conditions Determining Riparian Wetlands

Riparian habitats are integral parts of a larger landscape and therefore influenced by factors operating at various special and temporal scales. The physical setting that determines rivers and streams basically defines the riparian wetlands; however, some environmental features have specific importance on the wetlands that will be dealt with in the following.

Spatial and Temporal Scales

At the regional scale, geomorphology, climate, and vegetation affect channel morphology, sediment input, stream hydrology, and nutrient inputs. At the local scale, land use and related alteration to stream habitats, but also the activity of bioengineers such as beavers, can be of significant influence. At short timescales, individual heavy rainfall events affect the riparian systems; at an annual basis climate-induced changes in light, temperature, and precipitation trigger important cyclic biological events, such as autochthonous primary and secondary production, litterfall, decomposition, and spawning and hatching of animals. On multiannual timescales, extreme flood and drought events, debris-torrents, landslides, heavy storms or fire can have dramatic effects on the riparian zone and its biota.

Climatic Region

Climate controls the availability of the water in the wetlands and the activity period of the organisms. If the flooding and activity periods match, the floodborne resources can be used by the adapted floodplain biota (e.g., during summer floods). On the other hand, winter floods are generally less deleterious for little-flood-adapted tree species.

In the boreal and temperate regions, freezing and drought in winter and snowmelt floods in spring are predictable drivers of the interplay between surface water and groundwater in riparian wetland hydrology. Ice jams may cause stochastic flood events in winter. Normally, stream runoff is reduced during winter, and groundwater-fed riparian wetlands discharge into the stream channel as long as possible. In wetlands with organic sediments, this water is often loaded with large amounts of dissolved organic carbon. In shallow streams that freeze completely during winter, riparian wetlands may serve as refuges for the aquatic fauna, for example, for amphibians and turtles. Spring snowmelt events generally provoke prolonged flood events that exceed the duration of rain-driven floods. These long floods can connect the riparian wetland water bodies to the stream, so that organic matter and biota become exchanged. At the same time, there is often an infiltration (downwelling) of surface water into the riparian groundwater body.

In seasonal wet-and-dry climates (both Mediterranean and tropical savanna climates) water supply by rainfall is limited to a period of several months during which very strong rainstorms may occur. These events, albeit short, are of great importance for the release of dissolved substances and for the exchange of organic substances and biota between wetland and main water course. Moreover, energy-rich organic matter (e.g., fruits) may become flushed from the terrestrial parts of the catchment into riparian wetlands. On the other hand, flash-floods can cause scouring and erosion of fine sediments (including organic matter). During the dry season, groundwater levels are lower and may cause a seasonal drought in the riparian wetlands. In these periods, the aquatic biota either estivate or migrate into the permanent water bodies, and large parts of the stocked organic matter become mineralized.

However, even in strongly seasonal zones, like the Brazilian Cerrado, groundwater supply may be large enough to support permanent deposition of undecomposed organic matter.

The distribution of water-conductive (coarse) and impermeable substrate (bedrock and loam) of the valley bottom influences the thickness of the stagnant water body in the riparian zone and thus the extension of organic matter layers. Permanently humid conditions are found in many riparian wetlands of the boreal zone and in the humid tropics. These permanent riparian wetlands can accumulate large amounts of organic carbon. In tropical Southeast Asia (Malayan Peninsula and parts of Borneo), a special case of riparian wetland occurs, the peat swamps. These swamps develop when mangrove forests proceed seawards, and the hinterland soils lose their salt content. Here, large amounts of organic matter from the trees become deposited and the streams flow within these accumulations.

Valley Size, Morphology, and Connectivity

The common textbook pattern of steep valleys in the upper sections of the streams and open, shallow floodplains in the lower river sections holds true only for very few cases in nature. Rather, we find these two valley types interspersed in an alternating pattern like `beads on a string'. Shallow areas are more likely to bear extended riparian wetlands; however, if groundwater levels are high enough, even steep valleys may be covered with wetlands. The morphology of riparian wetlands can be described by the entrenchment ratio (i.e., the ratio of valley width at 50-years flood level to stream width at bankfull level) or by the belt width ratio, that is the distance between opposing meander bends over a stream section to stream width at bankfull level. Fifty-years flood often intersect the terrace slope.

Riparian wetlands of different catchments may be linked with each other through swamp areas (e.g., in old eroded landscapes of the Brazilian and Guyana Shields in South America) so that biogeographical barriers can be overcome by aquatic biota even without a permanent connection between the water courses. The term connectivity describes the degree by which a floodplain water body is linked to the main channel. Riparian wetlands may also be connected to the stream, either in a direct connection by a short channel, or indirectly by a longer channel which may be intercepted by a pond. In some cases, these channels can be cryptic/hidden when they are formed by macropores in the organic soils. Alluvial riparian wetlands may be connected to the stream via the hyporheic interstitial zone provided that the sediments are coarse enough to conduct water. Wetlands without any of these pathways exchange water, biota, and organic matter with the main channel during overbank flow of the stream. Purely aquatic organisms depend on the existence of connection channels to migrate between wetland and main water body. For example, amphibia are especially sensitive to fish predation, so that the highest biodiversity of amphibia is found at riparian wetland habitats with the lowest accessibility for fish.

Hydrology and Substrate Type

The slope of the landscape and the rock characteristics of the catchment define the physical habitat characteristics of the stream–wetland system. Riparian wetlands provide habitats with different hydraulic and substrate conditions than the stream channel. Although flooding in streams is generally shorter, less predictable and `spikier' than in large rivers, there is a large number of exchange processes between the main channel and the riparian zone during these flood events. Major flood events, albeit rare, act as `reset mechanism' in the floodplain that rejuvenates the sediment structure and the successional stage of the vegetation. Between these rare events, riparian wetlands act as sinks for fine particles and organic sediments that were washed out of the stream channel, the terrestrial zone of the catchment, or derive from an autochthonous biomass production.

Vegetation

Vegetation bordering to and growing within riparian wetlands fulfils many functions: it delivers both substrate for colonization and food resources for aquatic animals, it strips nutrients from the incoming water, and it provides raw material for the organic soils. It retards nutrient loss, filters nutrient input from the upland, reduces runoff by evapotranspiration, and buffers water-level fluctuations. Shading by tree canopies reduces light conditions for algal and macrophyte primary production and it equilibrates soil temperatures. Therefore, riparian wetlands differ completely according to their vegetation cover.

Unvegetated riparian wetlands occur at sites where establishment of higher plants is hampered by strong sediment movement (e.g., high-gradient and braided rivers), low temperatures (high elevation and polar zones), rocky surfaces, or periodical drought (desert rivers). The lack of shading and nutrient competition by higher plants favors growth of algae on the inorganic sediments, and productivity may be high, at least periodically.

High altitudes and/or elevated groundwater levels may preclude tree growth but allow the development of grass or herbal vegetation on riparian wetlands. Hillside swamp springs (helokrenes) can coalesce and form extensive marshes far above the flood level of the stream channel, so that the distinction between `riparian' and `common' wetland is difficult.

The tree species of forested riparian wetland are adapted to periodical or permanent waterlogging of the soils. They contribute an important input of organic carbon to the stream system. Large tree logs shape habitat structure by controlling flow and routing of water and sediment between stream channel and wetland. Tree roots increase sediment stability, sequester nutrients, and form habitats.

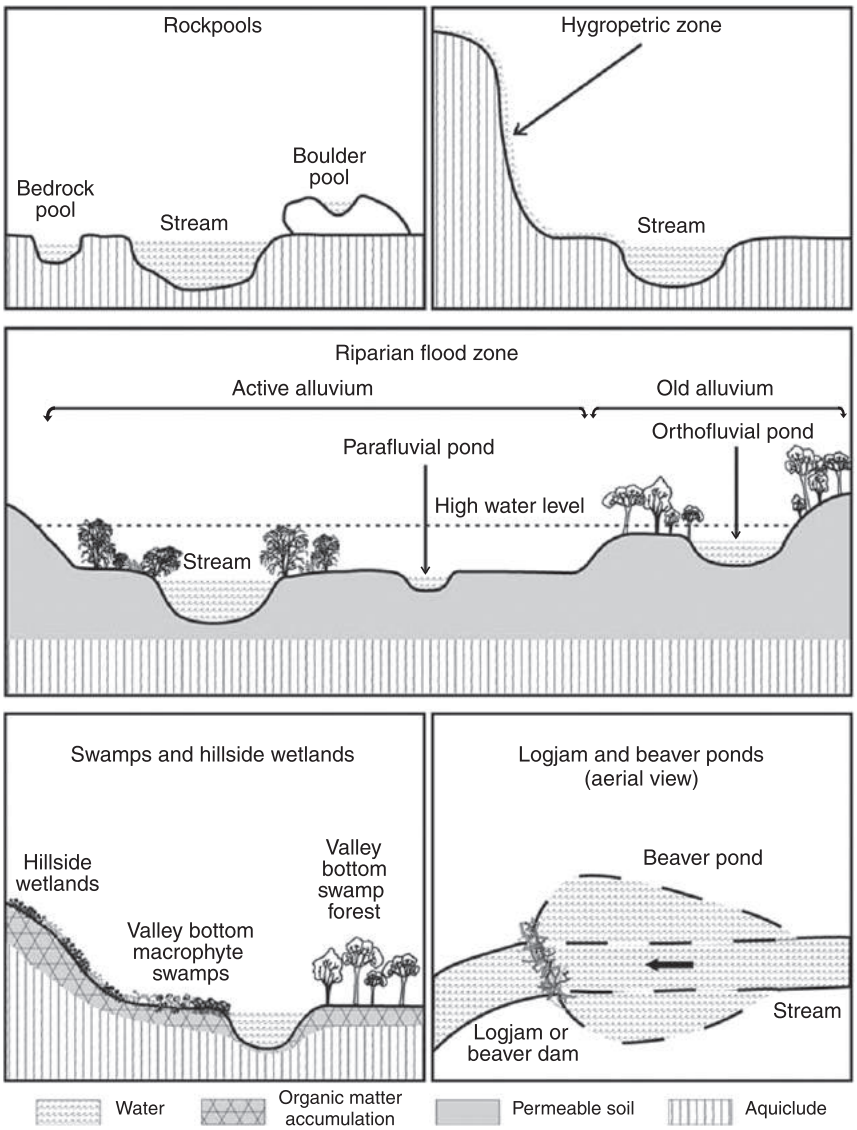


Fig. 2 Types of riparian wetlands.

Types of Riparian Wetlands

Riparian wetlands are very variable in size and environmental characteristics. In the following, we list the most common types according to their hydrological and substrate characteristics (Figs. 2 and **3**).

Hygropetric Zone

At sites where groundwater outflows run over rocky surfaces, hygropetric zones develop. In the thin water film, there is a vivid algal production and a diverse, however, less-studied fauna of invertebrates (mostly aquatic moths, chironomids, and other dipterans). Biota of the hygropetric zone need to be adapted to harsh environmental conditions such as periodical freezing and drying of the surfaces.

Rockpools

Many streams run through bedrock or large boulders which have slots that fill with flood or rain water. Biota colonizing these pools have to be adapted to relatively short filling periods, high water temperatures, and solar radiation. High algal production and low predator pressure (at least at the beginning of the filling period) attract many invertebrate grazers.

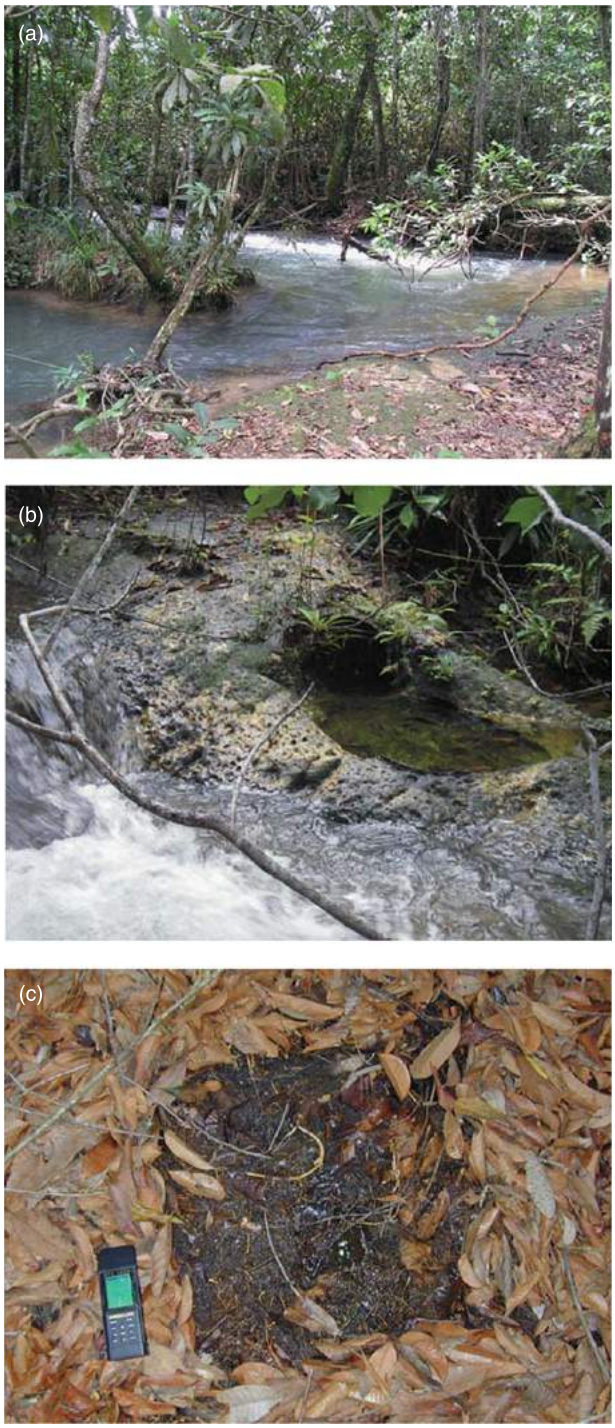


Fig. 3 Photographs of riparian wetlands (Tenente Amaral Stream, Mato Grosso, Brazil): (a) Stream channel with hygropetric zone (foreground) and floodplain forest (background), (b) Rockpool carved into the sandstone bedrock, (c) moist organic soil colonized by many aquatic invertebrate taxa. Leaf litter was removed. All photographs by K. M. Wantzen.

Parafluvial and Orthofluvial Ponds

In alluvial stream floodplains, permanent or temporary ponds develop from riverine dynamics either within the active channel (parafluvial pond) or in the riparian zone (orthofluvial pond). They are fed by both surface water and groundwater. In coarse-grained sediments, these ponds are connected to the main channel by the hyporheic interstitial zone, that is, an ecotone between groundwater and surface water that extends below and at either side of the stream channel. In fine-grained sediments (including organic soils), the contribution of groundwater is much more important, and these ponds are often brownish from dissolved

organic matter (humic acids and yellow substances). Para- and orthofluvial ponds contribute disproportionately to total species richness along riparian corridors.

Riparian Flood Zones

Even if no basin-like structures are present, flooding events create wetted zones on either side of the stream, independent of sediment type. Extension and permanence of the wetted zone depends on the valley shape, the porosity of the sediments, and eventual backflooding from tributary streams. In temporarily flooded forests with thick organic layers and in stranded debris dams, the moisture conditions may be long enough to bridge the gap between two flood events, so that many aquatic biota such as chironomids and other midges can complete their larval development in these semiaquatic habitats.

Riparian Valley Swamps

Swamps occur on soils that are waterlogged for most of the year. The lack of oxygen in the sediments allows accumulation of organic matter and selects for tree or herb species that have specific adaptations to these conditions, for example, pressure ventilation in the roots. The vegetation consists of either macrophytes or trees. Due to the shading and oxygen consumption during decomposition of organic matter, some of these riparian wetlands are hostile environments for aquatic metazoa that depend on dissolved oxygen. Some trees such as the Australian gum (*Melaleuca* sp.) shed bark which release secondary compounds that influence biota.

Hillside Wetlands

In areas where the aquiclude extends laterally from the stream, the riparian swamps can merge into hillside wetlands far above the flood level. Given that waterloggedness is permanently provided, these ecosystems tend to develop black organic soil layers from undecomposed plant material. The anoxic conditions in these soils favor denitrification and nitrogen may become a limiting factor for plant growth. Carnivorous plants (Droseraceae, Lentibulariacea, Sarraceniaceae) that replenish their nitrogen budget with animal protein are commonly found in these habitats. At sites where drainage is better, woody plants invade these natural meadows. The soft texture of the soils and their position in hill slope gradients makes these ecosystems highly vulnerable to gully erosion.

Logjam Ponds and Beaver Ponds

Falling riparian trees are stochastic events which may have dramatic consequences for the hydraulics of a stream system. Many tree species are soft-wooded, and tree dynamics are generally high in riparian wetlands. A fallen log blocks the current and creates a dam that accumulates fine particles. These natural reservoirs often extend far into the riparian zone.

Dams built by beaver (*Castor* sp.) can significantly alter the hydrological and biogeochemical characteristics of entire headwater drainage networks in Northern America and Eurasia. Fur trade led to the regional extinction of beavers. Few decades after reintroduction of beavers on a peninsula in Minnesota, they converted a large part of the area into wetlands, which led to a manifold increase in the soil nutrient concentrations. The activity of beavers considerably enhances the biodiversity of wetland-depending species. Beavers increase regional habitat heterogeneity because they regularly abandon impounded areas when the food supply is exhausted and colonize new ones, thereby creating a shifting mosaic of patches in variable stages of plant succession.

Typical Biota and Biodiversity in Riparian Wetlands

The importance of riparian wetland habitats for the conservation of biodiversity is well documented for several watersheds. Riparian areas generally have more water available to plants and animals than adjacent uplands. This is of specific importance in regions with a pronounced dry season, where lack of water affects plant growth. Abundance and richness of plant and animal species tend to be greater than in adjacent uplands because they share characteristics with the adjacent upland and aquatic ecosystems and harbor a set of specific riparian species. Because of their richness and their spatial distribution, the relative contribution of riparian ecosystems to total compositional diversity far exceeds the proportion of the landscape they occupy.

Apart from beavers, several other biota act as `ecological engineers' that create and modify riparian wetlands. African hippopotamus deepen pools and form trails that increase the ponding of the water. Several crocodilians maintain open water channels. Digging mammals, freshwater crabs, and insects like mole crickets increase the pore space in riparian soils and enhance the water exchange between wetland and stream channel. Similar macropores develop from fouling tree roots. Plants also strongly modify the habitat characteristics in riparian wetlands, either actively, by influencing soil, moisture, and light conditions or, passively, by changing the hydraulic conditions through tree fall or organic debris dams.

Typical wetland species are adapted to the amphibious characteristics of the habitats. They are either permanent wetland dwellers that cope with aquatic and dry conditions or they temporarily colonize the wetlands during either the dry or the wet phase. There are many animal species that permanently colonize riparian wetlands, especially anurans, snakes, turtles, racoons, otters, and many smaller mammals, like muskrats, voles, and shrews. Aquatic insects have developed special adaptations to survive periodical droughts, for example, by having short larval periods or drought resistance. Many birds profit by the rich food offered from the aquatic habitats like dippers, kingfishers, jacamars, warblers, and rails. Periodical colonizers from terrestrial ecosystems are bats, elks, moose, and several carnivorous mammals and birds. Many aquatic species like fish and aquatic invertebrates periodically colonize riparian wetlands. Riparian wetland biota belong to the most threatened species as they suffer from both the impacts on the terrestrial and aquatic systems, and many riparian species are threatened with extinction. The effects of extinction of a species are especially high if it is an ecological engineer or a keystone species, for example, a top predator. Extinction of wolves in the Yellowstone National Park in the US led to overbrowsing of broad-leaved riparian trees by increased elk populations.

Ecological Services of Riparian Wetlands

Riparian wetlands are intrinsically linked to both the stream and the surrounding terrestrial ecosystems of the catchment. In many places of the world, however, riparian zones have remained the only remnants of both wetland and woody habitats available for wildlife. They are surrounded by intensively used areas for either agriculture or urban colonization. The performance of riparian wetlands to provide ecological services becomes reduced by the same degree as these bordering ecosystems become degraded. However, even in degraded landscapes, the beneficial effects of the riparian wetland ecosystems are astonishingly high. For humans, healthy riparian wetlands are vital as filters and nutrient attenuators to protect water quality for drinking, fisheries, and recreation.

Nutrient Buffering

Riparian wetlands are natural traps for fine sediments and for organic matter, but they may vary from a nutrient sink to a nutrient source at different times of a year depending on high or low water levels. Particle-bound nutrients, such as orthophosphate ions, become deposited in the riparian wetlands during spates and may accumulate there. This may increase the amount of phosphate that becomes released during the following flood event. Therefore, technical plans for phosphorus retention in artificial wetlands in agricultural landscapes include a hydraulic design which hampers the release of particles from the wetland, for example, by providing continuous, and sufficiently broad wetland buffer strips along the streams.

For the removal of nitrogen inputs from floodwater and from lateral groundwater inputs, riparian wetlands are very efficient. Generally it can be taken for granted that the slower the water flow (both ground and surface water) the higher is the nitrate uptake rate; however, the precise flow pathways in the sediments have to be considered. In anoxic soils, reduction and denitrification processes transform inorganic nitrogen forms into nitrogen gas which is then released into the atmosphere. Once the nitrate has been completely reduced, sulphate is also reduced in the anoxic sediments. Nitrogen also becomes immobilized by bacterial growth and/or condensation of cleaved phenolics during the aerobic decay of organic matter. Aquatic macrophytes and trees growing in the riparian wetlands are very efficient in nitrogen stripping by incorporating mineral nitrogen forms into their biomass. They can represent the most important nitrogen sinks in riparian systems. Some riparian wetland plants (e.g., alder, *Alnus* sp., and several leguminous trees) have symbiotic bacteria associated to their roots that can fix atmospheric nitrogen when this nutrient is scarce in the soils. Thus, not all riparian wetlands exclusively remove nitrogen.

Carbon Cycle

Like other wetlands, riparian wetlands are important players in the carbon cycle of the watershed. They accumulate large amounts of coarse particulate organic matter (CPOM) and they release dissolved organic matter into the stream and gaseous carbon-compounds into the atmosphere (**Fig. 1**).

In the boreal zone, the spring snowmelt runoff contributes to more than half of the annual total organic carbon (TOC) export. The larger the riparian wetland zone, the bigger the amount of exported TOC. On the other hand, riparian wetlands receive large amounts of dissolved carbon from litter leachates from the surrounding forests, especially during the leaf-fall period. These leachates can be an important source for phosphorous and other nutrients, as well as for labile carbon compounds. These substances enhance heterotrophic microbial (bacterial and fungal) activity.

Spring snowmelt also carries large amounts of fine particulate organic matter (POM). Riparian wetlands often provide surface structures that act like a comb to accumulate these particles (e.g., macrophytes), and enhance the production of detritivores. Additional POM is produced by riparian trees. The general trend for litter production to increase with decreasing latitude (valid in forests) is overlain by species-specific productivity and physiological constraints due to the waterloggedness in riparian wetlands. Here, the litter production is generally higher in periodically flooded, than in permanently flooded, wetlands. Depending on the oxygen content of the soils, the chemical composition of the leaves, and the activity of detritivores, more or less dense layers of `leaf peat' can accumulate in the sediments. This organic matter stock can be increased by undecomposed tree logs and bark. A

reduction of the water level in the riparian wetlands leads to an increased mineralization of the carbon stocks and enhances the release of carbon dioxide.

Hydrological Buffering and Local Climate

Riparian wetlands have an equilibrating effect on hydrological budgets. Riparian vegetation dissipates the kinetic energy of surface flows during spates. Riparian wetlands store stormwater and release it gradually to the stream channel or to the aquifer between rainstorm events. Moreover, they are important recharge areas for aquifers. Several current restoration programmes try to increase this recharge function of riparian wetlands in order to stabilize the groundwater stocks for drinking water purposes.

Riparian wetland trees and macrophytes contribute considerably to evapotranspiration and to local and regional climate conditions. The rate of vapor release depends on the plant functional group which needs to be considered for basin-scale water budgets.

Corridor Function for Migrating Species

Riverine wetlands represent a web of ecological corridors and stepstones. In intense agricultural areas they can be considered as `green veins' that maintain contact and gene flow between isolated forested patches. Providing shadow, balanced air temperatures and moisture, shelter, resting places, food and water supply, they cover the requirements of a great deal of amphibian, reptile, bird, and mammal species. These not only use the longitudinal connection but also migrate laterally and thus reach the next corridor aside. Moreover, long-range migrating birds use the green corridors of riparian zones in general as landmarks for migration. Networks of riparian corridors also facilitate the movement of non-native species. In some US riparian zones, their richness was about one-third greater in riparian zones than on uplands and the mean number and the cover of non-native plant species were more than 50% greater than in uplands.

Refugia and Feeding Ground for Riverine Biota

During flood, drought, and freezing events, but also during pollution accidents in the stream channel, connected riparian wetland habitats represent refugia for riverine animals. In extreme cases, residual populations from the wetlands may contribute to the recolonization of defaunated stream reaches. Riparian wetlands also act as traps and storage sites for seeds both from the upstream and from the uphill areas. The seed banks contain propagules from plants that represent a large range of moisture tolerances, life spans, and growth forms. These seeds may also become mobilized and transported during spate events.

Riparian wetlands offer a large variety of food sources. Connected wetland water bodies `comb out' fine organic particles including drifting algae from the stream water, they receive aerial and lateral inputs of the vegetation, and they have a proper primary productivity which profits by the increased nutrient input and storage from the surroundings. Many riverine fish and invertebrate species are known to migrate actively into the riparian wetlands in order to profit by the terrestrial resources that are available during flood periods. In analogy to the `floodpulse advantage' of fish in large river floodplains, stream biota that temporarily colonize riparian wetlands have better growth conditions than those that remain permanently in the stream channel. For example, the macroinvertebrate community of riparian sedge-meadows in Maine (USA) is dominated by detritivorous mayfly larvae (over 80% of the invertebrate biomass) during a 2-month period in spring. The larvae use the stream channel as a refuge and use the riparian wetland as feeding ground where they perform over 80% of their growth.

Reciprocal Subsidies between Aquatic and Terrestrial Ecosystems

Many aquatic species profit by the terrestrial production and vice versa. Apart from leaf litter, large quantities of fruits, flowers, seeds, as well as insects and feces fall from the tree canopies into the streams where they represent important energy and nutrient sources for the biota. In Amazonian low-order rainforest streams, terrestrial invertebrates make up a major portion of the gut content of most fish species. Fruits and seeds are preferred food items for larger fish species that colonize medium- and high-order rivers. Riparian wetlands increase the area of this active exchange zone, and they retain these energy-rich resources for a longer period than a stream bank alone would do.

Aquatic organisms also contribute to the terrestrial food webs. For example, bats are known to forage on the secondary production of emerging insects in riparian wetlands, and the shoreline harbors a large number of terrestrial predators, such as spiders, tiger beetles, and riparian lizards. Experimental interruption of these linkages (e.g., by covering whole streams with greenhouses) has shown that the alteration of riparian habitats may reduce the energy transfer between the channel and the riparian zone.

Recreation

The sound of the nearby stream, the equilibrated climate, and the occurrence of attractive animal and plant species render riparian wetlands highly attractive for recreation purposes such as hiking, bird-watching, or meditation. These can be combined with `in-

channel' recreation activities such as canoeing, rafting, or fishing, and represent an economically valuable ecosystem service, that should be considered in management and conservation plans.

Conservation

Water is becoming scarce in many areas worldwide. Water mining reduces water levels, but high and stable groundwater tables are a prerequisite for the existence of riparian wetlands. In addition to direct water withdrawal, predictions about climatic changes include other threats. Increased stochasticity of the runoff patterns and reduced snowmelt floods are severe threats to the existence of riparian wetlands. The riparian zones of streams and rivers have been sought after by humans since early days. High productivity, reliable water supply, and climatic stability make these ecosystems suitable for a range of human-use types, such as wood extraction, hunting, aquaculture, and agriculture. In areas of intensive agriculture, riparian zones including their wetlands have shrunk to narrow strips or have completely vanished. On the other hand, the ecosystem services are good socioeconomical arguments to restore and enlarge riparian wetlands.

For conservation planning, it is very important to bear in mind that riparian wetlands are very diverse and have typical regional characteristics. Secondly, the whole riparian zone is very dynamic. Many tree species are relatively short-lived and well adapted to changes in the floodplain morphology or in the hydrology of the wetland. The existence of variable hydrological patterns is a prerequisite for the coexistence of annually varying plant and animal communities. Often, large-scale projects restore riparian zones including wetlands according to a single pattern that does not consider these dynamic changes in habitat and species diversity. If large flood events are precluded by dam constructions in the upstream region, the natural habitat dynamics are blocked and the vegetation will develop towards a late-successional stage without pioneer vegetation, and with a reduced range of moisture tolerance. Several studies could prove that once the hydrological fluctuations become reduced by water-level regulation, exotic species can invade river valleys more efficiently.

While many animal species depend exclusively on the specific habitat conditions of wetlands, most riparian amphibians and reptiles migrate into the drier zones of the aquatic–terrestrial ecotones for a part of their life cycle. This makes them vulnerable to increased mortality in the neighboring ecosystems, especially if these have been converted into agricultural or urban use. Therefore, a buffer zone considering the home range of these species is needed to fully protect these species.

See also: Ecological Processes: Nitrification. Evolutionary Ecology: Ecological Niche. Global Change Ecology: Energy Flows in the Biosphere. Human Ecology and Sustainability: Ecological Footprint

Further Reading

Ilhardt, B.L., Verry, E.S., Palik, B.J., 2000. Defining riparian areas. In: Verry, E.S., Hornbeck, J.W., Dolloff, C.A. (Eds.), Riparian Management in Forests of the Continental Eastern United States. Boca Raton, London, New York, Washington, DC: Lewis Publishers, pp. 23–42.

Junk, W.J., Wantzen, K.M., 2004. The flood pulse concept: New aspects, approaches, and applications – An update. In: Welcomme, R.L., Petr, T. (Eds.), Proceedings of the Second International Symposium on the Management of Large Rivers for Fisheries, vol. 2. Bangkok: FAO Regional Office for Asia and the Pacific, pp. 117–149.

Lachavanne, J.-B., Juge, R. (Eds.), 1997. Man and the Biosphere Series, Vol. 18: Biodiversity in Land–Inland Water Ecotones. Paris: UNESCO and The Parthenon Publishing Group.

McCormick JF (1979) A summary of the national riparian symposium. In: U.S. Department of Agriculture, Forest Service (ed.) *General Technical Report WO-12 Strategies for Protection and Management of Floodplain Wetlands and Other Riparian Ecosystems*, pp. 362–363pp. Washington, DC: US Department of Agriculture, Forest Service.

Mitsch, W.J., Gosselink, J.G., 2000. Wetlands, 3rd edn. New York: Chichester, Weinheim, Brisbane, Singapore Toronto: Wiley.

Naiman, R.J., DÅcamps, H., McClain, M.E., 2005. Riparia – Ecology, Conservation, and Management of Streamside Communities. Amsterdam: Elsevier.

Peterjohn, W.T., Correll, D.L., 1984. Nutrient dynamics in an agricultural watershed: Observations on the role of a riparian watershed. Ecology 65, 1466–1475.

Verry, E.S., Hornbeck, J.W., Dolloff, C.A. (Eds.), 2000. Riparian Management in Forests of the Continental Eastern United States. Boca Raton, London, New York, Washington, DC: Lewis Publishers.

Wantzen, K.M., Yule, C., Tockner, K., Junk, W.J., 2006. Riparian wetlands. In: Dudgeon, D. (Ed.), Tropical Stream Ecology. Amsterdam: Elsevier, pp. 199–217.

Rivers and Streams: Ecosystem Dynamics and Integrating Paradigms☆

Kenneth W Cummins, Michigan State University, East Lansing, MI, United States
Margaret A Wilzbach, Humboldt State University, Arcata, CA, United States

Nomenclature

Calories (cal), kilocalories (kcal, 1000 cal)
The heat equivalent of a given mass

Glossary

Allochthonous Organic inputs to lotic ecosystems that originate from plants and animals of terrestrial or marine origin.

Autochthonous Organic inputs that are generated within a lotic ecosystem through photosynthesis of autotrophic members of the aquatic community.

Autotroph An organism that synthesizes organic matter from inorganic materials using sunlight or chemical energy. Important autotrophs in lotic ecosystems include vascular plants, mosses and liverworts, periphyton and phytoplankton, some bacteria and protists. Autotrophic ecosystems are those in which trophic pathways are predominantly fueled by in-stream (autochthonous) primary production.

Ecotone A transition zone between two adjacent ecological communities.

Flocculation The process in which colloids come out of suspension and aggregate or precipitate to form particulate matter.

Heterotroph An organism that cannot fix carbon from inorganic sources; its energy derives from consumption of living or dead organic matter. Important heterotrophs in streams and rivers are microbes, protozoans and micrometazoans, macroinvertebrates, and vertebrates. Heterotrophic ecosystems are those in which energy transfers are dominated by community respiration (of microbes plus aquatic plants plus macroinvertebrates).

Hyporheos The groundwater interface in streams where a mixture of surface water and groundwater can be found, located both beneath the active stream channel and within the riparian zone of most streams and rivers.

Invertebrate functional feeding group Collection of macroinvertebrate taxa with similar morphological and behavioral mechanisms for procuring food.

Nutrient spiraling The interdependent processes in lotic ecosystems of nutrient cycling and downstream transport.

Paradigm A widely accepted conceptual model or theory.

Introduction

Research scientists, watershed managers, and conservationists alike agree that following an ecosystem perspective is the most productive way to examine streams and rivers. The integration of physical–chemical with biological processes, which is the study of ecosystems, has largely replaced single physical factor or single-species approaches to management and rehabilitation of running waters. In the discussion that follows, fluxes of energy and matter into, through, and out of lotic ecosystems are the basic processes embraced by integrating paradigms (conceptual models) that continue to underlie inquiry into the structure and function of streams and rivers.

Energy Flux

Energy Sources

Streams and rivers are driven almost entirely by two alternate energy sources: (1) sunlight that fuels the in-stream growth of aquatic plants (primary production), and (2) plant litter from stream-side (riparian) vegetation. The relationship between these two energy drivers is essentially inverse. The heavier the riparian cover over the stream/river channel, the greater the plant litter inputs and the greater the limitation of light reaching the water and therefore in-stream algae and vascular plant growth. In contrast to non-filamentous algae, very few stream/river consumers utilize macrophytes, filamentous algae, and rooted vascular plants. Rather, macrophytes enter the energy transfer to consumers as detritus after they die. Stream and river systems in which the

☆*Change History*: October 2017. KW Cummins and MA Wilzbach introduced small edits in the text of the article and in figure captions; added a sixth functional feeding group category in the section Feeding Roles and Food Webs; added new references and deleted others; added a table (Table 2) to compare use of biomass versus numeric data on functional feeding group ratios as surrogates for lotic ecosystem attributes; and added a Glossary to define technical terms.

This is an update of K.W. Cummins and M.A. Wilzbach, Rivers and Streams: Ecosystem Dynamics and Integrating Paradigms, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 3084–3095.

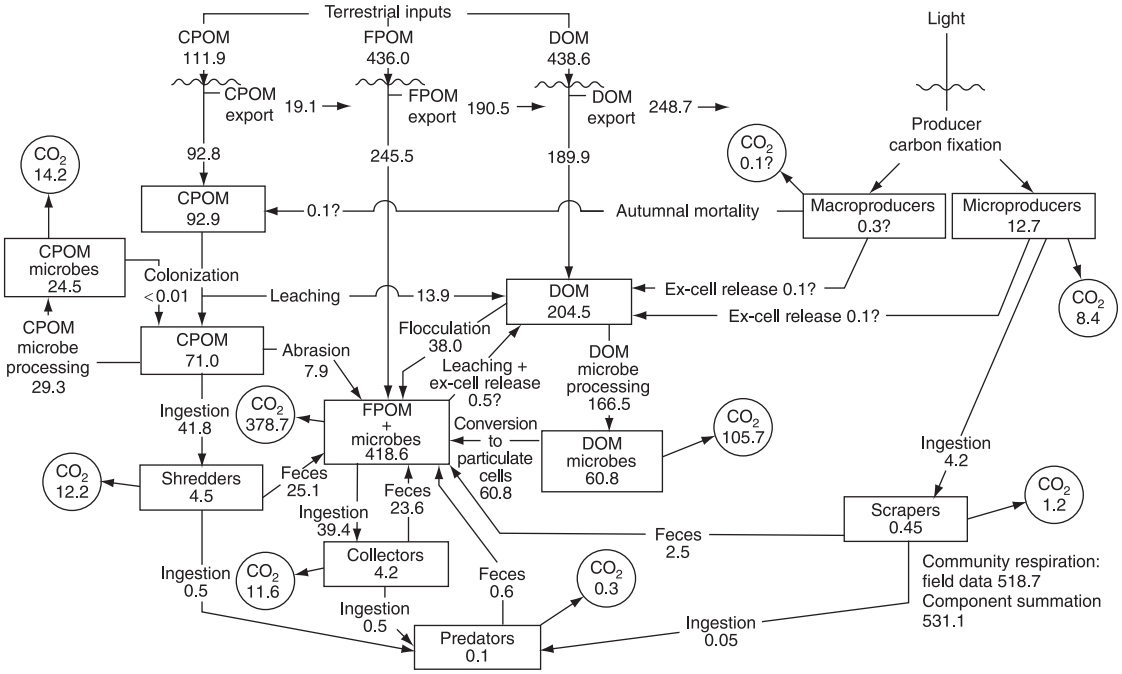


Fig. 1 Example of an energy budget for a small woodland stream ecosystem (Augusta Creek watershed, Michigan, USA). All values are in grams ash-free dry mass m^{-2} $year^{-1}$. *Squares* represent pools of organic matter in various states; *arrows* represent transfers and *circles* represent respiratory consumption of organic matter. From Saunders GW, *et al.* (1980) In: LeCren ED and McConnell RH (eds.) The functioning of freshwater ecosystems. Great Britain: Cambridge University Press

majority of the energy transfer is from in-stream plant growth to consumers are termed autotrophic. Those systems dominated by the detrital pathways of energy transfer are heterotrophic. As discussed in the "river continuum concept" (RCC), the relative importance of these two energy sources changes with stream size. Smaller streams in forested catchments are usually dominated by litter energy sources and wider, mid-sized stream segments are dominated by plant growth in the water. Larger rivers are dependent upon organic matter (OM) delivered from the upstream tributary network.

A model of energy flux, that is, the transfer of energy between trophic levels of plants and animals, was produced by Lindeman in the early 1940s and various forms of this model have more or less been the basis for the investigation of energy flux in running waters ever since. These studies most frequently take the form of energy budgets; an accounting of the energy in and the energy out of a given ecosystem (Fig. 1) or biological population (Fig. 2) or community within the system. OM budgets are useful in identifying the sources, magnitude, and fates of energy and provide insight into internal dynamics of lotic ecosystems. At the system level, inputs include autotrophic production plus energy originating from the surrounding terrestrial environment (allochthonous) that is brought in by various physical vectors. Outputs include community respiration and losses by downstream transport. Energy retained within a stream reach over a given time interval is referred to as storage. Comparisons of energy flux among and between trophic levels commonly express biomass as caloric equivalents. Animal ingestion, egestion, and growth (increase in mass) are all measured as biomass. Respiration (metabolism) is readily converted to calories consumed using an oxy-calorific equivalent. Tables are available that provide conversions of mass to calories for freshwater organisms.

Feeding Roles and Food Webs

Feeding studies of benthic macroinvertebrates have shown that, based on food ingested, most taxa are omnivorous. For example, invertebrates that chew riparian-derived leaf litter in streams, termed "shredders," ingest not only the leaf tissue and associated microbiota (e.g., fungi, bacteria, protozoans, and micro-arthropods), but also diatoms and other algae that may be attached to the leaf surface, as well as very small macroinvertebrates (e.g., first-instar midge larvae). For this reason, trophic-level analysis does not lend itself well to simple trophic categorization of stream macroinvertebrates.

An alternate classification technique, originally described by Cummins in the early 1970s, involves the functional analysis of stream/river invertebrate feeding. The method is based on the combined morphological and behavioral mechanisms of food acquisition used by the invertebrates and five fundamental categories of their food found in running waters (Fig. 3). There is a direct correspondence between the availability of categories of nutritional resources and the relative abundance of invertebrate populations that are adapted to efficiently harvest a given food resource. In 1974, five invertebrate functional feeding groups (FFG) were proposed. These include shredders, filtering collectors, gathering collectors, scrapers, and predators. These groups partition five food resource categories in running waters that are defined on the basis of particle size and type: (1) coarse particulate organic

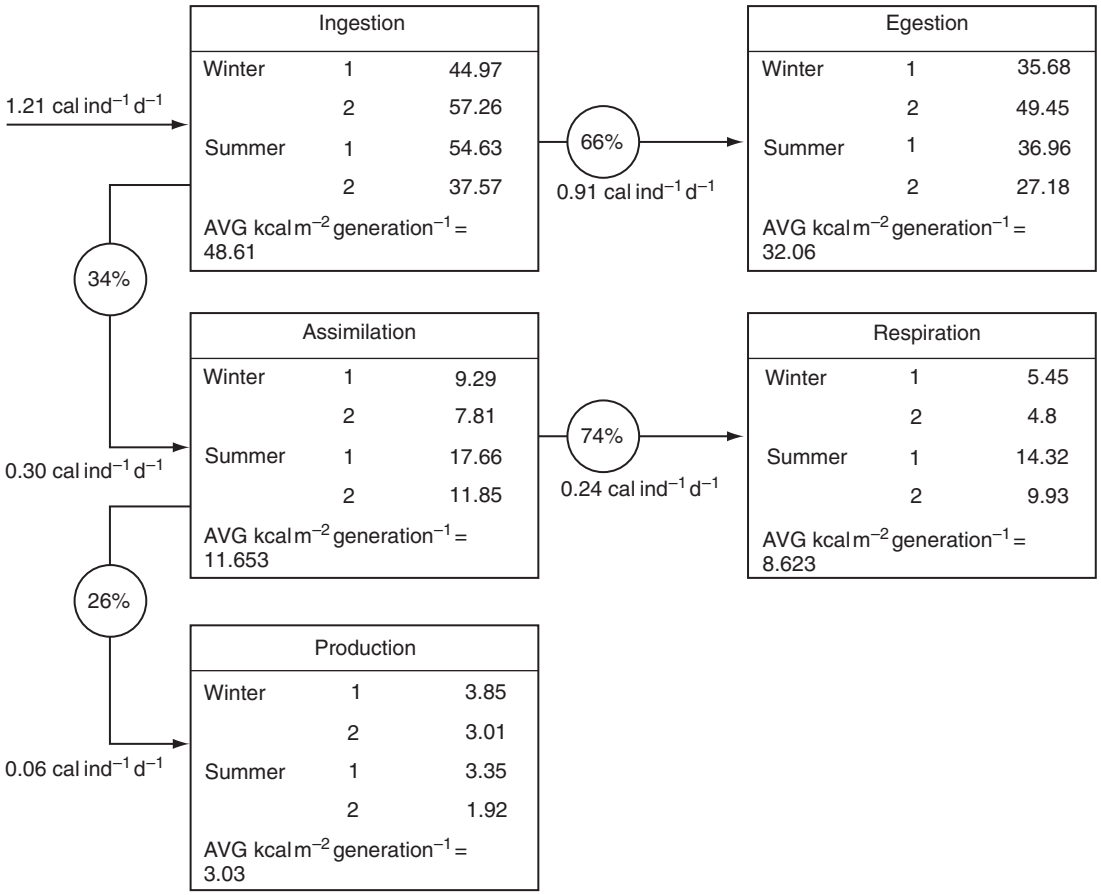


Fig. 2 Example of an energy budget constructed over 2 years of study for a population of a stream invertebrate (*Glossosoma nigrior*, Trichoptera) from Augusta Creek, Michigan, USA. The budget is based on independent measurements of ingestion, production, and respiration. Modified from Cummins, K.W., (1975). Macroinvertebrates. In: Whitton BA (ed.) River Ecology. Berkeley: University of California Press.

matter (CPOM), which is primarily riparian plant litter that has been conditioned, that is, microbially colonized, within the stream; (2) fine particulate organic matter (FPOM), which are particles generally smaller than 1 mm in diameter that are largely derived from the biological and physical breakdown of CPOM and whose surfaces are colonized and metabolized by bacteria; (3) periphyton, that is, tightly accreted algae and associated organic material; (4) live vascular aquatic plants; and (5) prey, that is, invertebrate species or larval/nymphal stages small enough to be captured and consumed by invertebrate predators. Recently, two other categories were added: herbivore shredders and herbivore piercers. Herbivore shredders are macroinvertebrates that are adapted to consume live aquatic vascular plant tissue. Herbivore piercers, which are limited to some species within the caddisfly family Hydroptilidae, pierce individual algal cells to suction their cell contents.

As the relative availability of the basic food resources changes, there is a concomitant change in the corresponding ratios of the FFGs of freshwater invertebrates adapted to specific resource categories.

Obligate and facultative members occur within each FFG. These can be different species or different stages in the growth period of the life cycle of a given species. For example, it is likely that most aquatic insects, including predators, are facultative gathering collectors as first-instars newly hatched from the egg. It is with obligate forms that linkages between invertebrates with their food resource categories are most reliable. The distinction between obligate and facultative status is best described by the efficiency with which a given invertebrate converts the resource acquired to growth; that is, obligate forms are more efficient consumers of a given food resource, such as conditioned leaf litter, than are facultative forms. For example, shredders feeding on litter consume the fungal-rich leaf matrix, whereas scrapers only abrade the much less microbially colonized, and therefore less nutritious, leaf cuticle. The high efficiency of obligate forms feeding on a particular resource category is in contrast with the wider array of food types consumed by facultative forms, but with lower efficiency. Facultative forms are more flexible, and often exhibit a greater affinity for one of the five food resource categories. The same morpho-behavioral mechanisms can result in the ingestion of a wide range of food items, the intake of which constitutes herbivory (consumption of living algal and vascular plants), detritivory (consumption of dead OM), or carnivory (consumption of live animal prey).

Although intake of food types changes from season to season, habitat to habitat, and with growth stage, limitations in food acquisition mechanisms have been shaped over evolutionary time and these are relatively more fixed than the food items ingested.

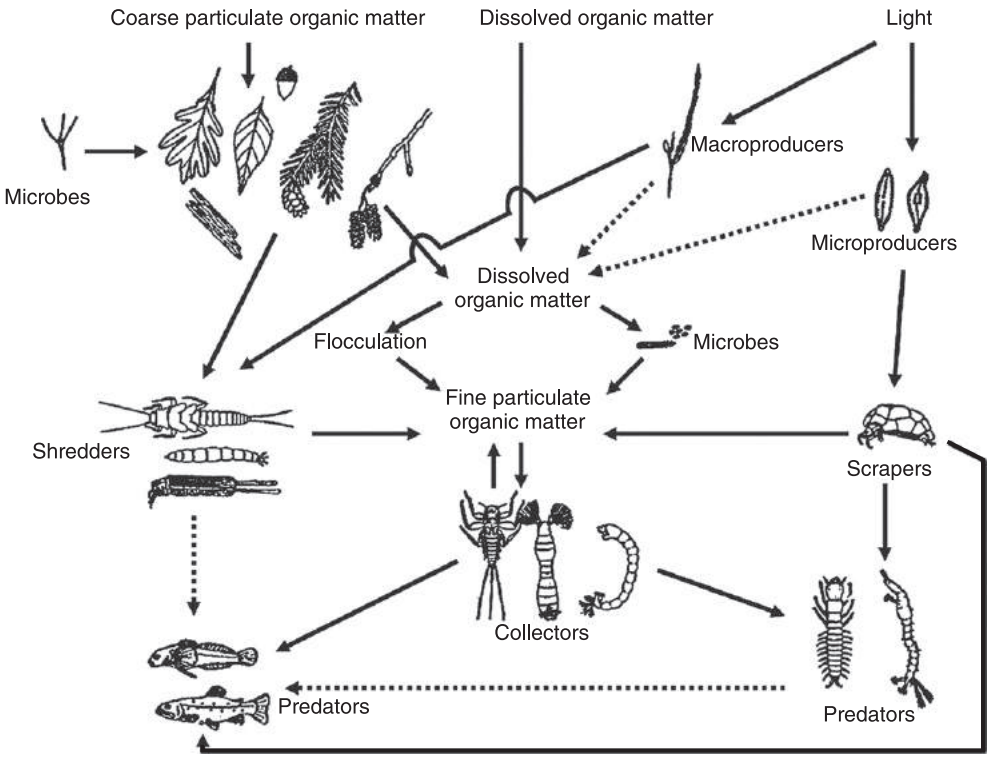


Fig. 3 Conceptual model of invertebrate functional feeding groups and their food resources in a small, forested stream ecosystem. *Dashed lines* indicate additional or alternate pathways. Modified from Cummins K.W., (1974). Structure and function of stream ecosystems. Bioscience 24: 631–641.

Morphological structures that enable aquatic insects to harvest a given food resource category exhibit significant similarities across diverse taxa. This convergent or parallel evolution lies at the heart of the FFG classification method. For example, larvae of the 26 North American caddisfly (Trichoptera) families are spread among the six major FFGs. The less highly evolved mayflies (Ephemeroptera) and stoneflies (Plecoptera) are adapted to acquire fewer food resource categories (**Table 1**).

An advantage of the FFG procedure is that it does not require detailed taxonomic separations of the invertebrates. Broad, easily distinguished characteristics allow FFG classification, preferably in the field with live specimens. Separations usually involve systematic distinctions at the level of family or higher, and cut across taxonomic lines. As an example, two groups of case-bearing larval caddisflies (Trichoptera) are sufficient to separate FFG categories at better than 90% efficiency. All families, or genera within families, of Trichoptera that construct mineral cases are FFG scrapers; those that construct organic cases are FFG shredders.

Given the coupling of FFGs with food resource categories, ratios of the different FFG groups can serve as surrogates for ecosystem attributes. For example, the ratio of the functional groups linked to in-stream primary production (scrapers plus those shredders that may harvest live plant tissue) to those groups dependent upon the CPOM and FPOM heterotrophic food resources (shredders of detrital material plus gathering and filtering collectors) provides an index of the ratio of autotrophy to heterotrophy at the lotic ecosystem level. When measured directly, an ecosystem ratio of autotrophy/heterotrophy >1 indicates an autotrophic system. A surrogate FFG ratio of >0.75 has been experimentally measured in such autotrophic stream/river systems. The interpretation of FFG ratios as surrogates for lotic ecosystem attributes, such as autotrophy versus heterotrophy, can differ somewhat whether the data are expressed numerically (the usual case) or gravimetrically (**Table 2**). The conversion of numbers to biomass equivalents can readily be accomplished using the relationship:

$$Y = aX^b$$

where Y = dry mass in mg, X = body length in mm, a = Y intercept, and b = slope of Y on X (Benke *et al.*, 1999).

Flux of Matter

Nutrient Cycles and Spiraling

The limnological study of standing waters has always been dominated by a conceptual model of closed ecosystems, in which nutrients recycle seasonally, totally within the system. The unidirectional flow of running waters necessitated modifying this view of closed cycles in lakes to an open-cycle model; that is, the open nutrient cycles in streams and rivers follow a spiraling pattern in

Table 1 Total number of families and number of families assigned to each functional feeding group of three major orders of benthic aquatic insects found in North America

Order	*Number of families by dominant functional feeding group*							
	Total number families	*Shredder-detritivores*	*Shredder-herbivores*	*Scrapers*	*Filtering collectors*	*Gathering collectors*	*Predators*	*Filamentous algal piercers*
Ephemeroptera	23			2	5	12	4	
Plecoptera	9	6					3	
Trichoptera	26	7	2	8	4	3	1	1

Assignment of functional group based on largest number of species in the family listed as in that functional group.

which nutrients generated (or delivered) at one point along a stream or river complete the recycling to their initial state at a displaced location downstream (Fig. 4). Total spiral length represents the sum of the distance traveled by an element as an inorganic solute or organic particle until its uptake by the biota, plus the distance traveled within the biota until its release back into the water column. If nutrients such as nitrogen or phosphorous are cycled rapidly, the spirals are "tight," that is, the downstream completion of the cycle is short. If cycling is slow, the closing of the loop is displaced a longer distance downstream and the spirals are more open. The tighter the loops of nutrient spirals, the more retentive (conservative) is the stream or river reach.

Transport and Storage of OM

The transport and storage of OM in running-water ecosystems involves complex interactions between (1) the state of the OM, (2) the source of the OM, and (3) the physical, chemical, and biological retention potential for any given reach of stream or river.

State of the OM

OM can be partitioned into three broad categories: (1) dissolved (DOM, size range $<0.45\,\mu m$); (2) fine particles (FPOM, size range $>0.45\,\mu m$ to 1 mm); and (3) coarse particles (CPOM, size range >1 mm). Although FPOM particles are colonized primarily on the surface by bacteria, CPOM is colonized by fungi, bacteria, and microzoans that penetrate the matrix of the plant material. Aquatic hyphomycete fungi usually penetrate the CPOM leaf and needle litter first. Bacteria and microzoans follow the fungal hyphal tracks into the matrix of the CPOM. The OM in solution (DOM) includes a full range of molecules from simple very labile ones such as sugars and amino acids to complex recalcitrant ones such as phenolic compounds.

Sources of the OM

A major source of OM in streams (orders 0–5, where stream order 0 represents headwater tributaries that are intermittent or ephemeral, and the smallest permanently flowing stream is first order) is the riparian zone. This border of stream-side vegetation produces litter (e.g., leaves, needles, bud and flower scales, seeds and fruits, small wood and bark) most of these enter streams on a seasonal schedule depending upon the relative proportions of deciduous and evergreen species. Other sources of OM are solutions and particles from bank erosion, DOM leachates from riparian litter, exudates, and leachates from periphytic algae and vascular aquatic plants together with their physical fragmentation and moribund tissues.

Physical, Chemical, and Biological Retention Potential

The retention of DOM involves physical flocculation of the OM in solution with divalent cations, such at Ca^{2+}, and biological uptake by resident bacteria and fungi. Chemical reactions between the smaller molecular weight organic compounds may precede the physical complexing with cations. The rate and extent of biological uptake of DOM depends upon factors such as the lability or recalcitrance of the compounds, density and composition of the microbial flora, and water temperature. These mechanisms that convert DOM to FPOM, flocculation and microbial uptake, are quite important ecosystem processes. The conversion of DOM in solution to particles significantly increases the retention of the OM. The difference in the efficiency of retention of OM between soft, stained water streams and hard, clear water streams accounts in part for the greater productivity of the latter. The POM that results from the conversion of DOM is more likely to remain in a given reach of stream or river and enter into trophic pathways.

Retention of POM depends upon channel geomorphology. Large wood debris (LWD), branches and exposed bank roots, coarse sediments, backwaters, side channels, and settling pools are all important retention features. For any given reach of stream or river, a major source of OM is transport from upstream. In addition, OM is retained when bankfull-flow is exceeded and material is deposited on the upper banks or on the floodplain. OM is returned to the channel when water levels recede. Whether these

Table 2 Comparison of functional feeding group (FFG) ratios determined by numbers versus biomass

Functional group	*Taxa*	*Number*	*Biomass (mg)*	*Calculated ratios*	*Numbers ratio*	*Biomass ratio*
SC (scrapers)	Heptageniidae (5 genera)	135	12.659	P/R (gross primary production/community respiration): SC/ DSH + GC + FC Threshold = >0.75	167/(34 + 147 + 21) = 0.83	17.255/(2.877 + 23.864 + 6.624) = 0.52
	Ephemerellidae (*Drunella*)	2	0.965			
	Glossosomatidae (*Glossosoma*)	4	0.700		Autotrophic	Heterotrophic
	Uenoidae (*Neothrema*)	14	2.882			
	Gastropoda (*Juga*)	12	0.049			
SC totals	9 Taxa	167	17.255			
DSH (detrital shredders)	Lepidostomatidae (*Lepidostoma*)	6	0.833	CPOM/FPOM (primarily riparian leaf litter/organic fragments and feces): DSH/GC + FC Fall-winter threshold = >0.50 Spring–summer threshold ≥ 0.25	34/(147 + 21) = 0.20	2.877/(23.864 + 6.624) = 0.09
	Sericostomatidae (*Gumaga*)	3	0.165			
	Capniidae	11	0.234		Shredders not dominant	Shredders not dominant
	Leuctridae	9	0.956			
	Peltoperlidae	5	0.689			
DSH totals	5 Taxa	34	2.877			
GC (gathering Collectors)	Baetidae	51	1.591	TFPOM/BFPOM (suspended load/sediment storage): FC/GC Threshold = >0.50	21/47 = 0.45	6.624/23.864 = 0.28
	Leptophlebiidae	25	14.465			
	Ephemerellidae (3 genera)	11	3.760		Moderate supply of FPOM in transport	Moderate but smaller supply of FPOM in transport
	Elmidae larvae (*Cleptelmis*)	15	3.262			
	Chironomini	39	0.614			
	Collembola	6	0.172			
GC totals	8 Taxa	147	23.864			
FC (filtering collectors)	Hydropsychidae	5	2.118	Substrate stability (coarse sediments, large wood, or rooted aquatic vascular plants): SC + FC/ DSH + GC Threshold = >0.50	(167 + 21)/(34 + 147) = 1.04	(17.255 + 6.624)/(2.877 + 23.864) = 0.89
	Philopotamidae	14	4.228			
	Simuliidae	2	0.278			
FC totals	3 Taxa	21	6.624		Stable substrates abundant	Stable substrates abundant
12.749 P (predators)	Perlidae	11	10.031	Top down/bottom up invertebrate community structure (predator vs. food regulation of invertebrate populations): P/SC + DHS + GC + FC Threshold = 0.10–0.20	62/(167 + 34 + 147 + 21 + 62) = 0.14	12.749/(17.255 + 2.877 + 23.864 + 6.624 + 12.749) = 0.20
	Chloroperlidae	41	2.567			
	Tipulidae (*Eriocera, Hexatoma*)	8	0.057			
	Tanypodinae	2	0.074			
	6 Taxa	62	12.749		Normal predator levels indicating bottom up regulation	Normal predator levels indicating bottom up regulation
Total all samples	31 Taxa	431	63.369			

Winter data from Prairie Creek, northern California. *CPOM* = coarse particulate organic matter; *FPOM* = fine particulate organic matter; *transport TFPOM* = suspended in stream water column; *benthic BFPOM* = entrained on or in stream bottom sediments. Proposed ratio thresholds are based on FFG data collected simultaneously with measurements of stream ecosystem attributes in North American systems. Herbivorous shredders are not present in this system.

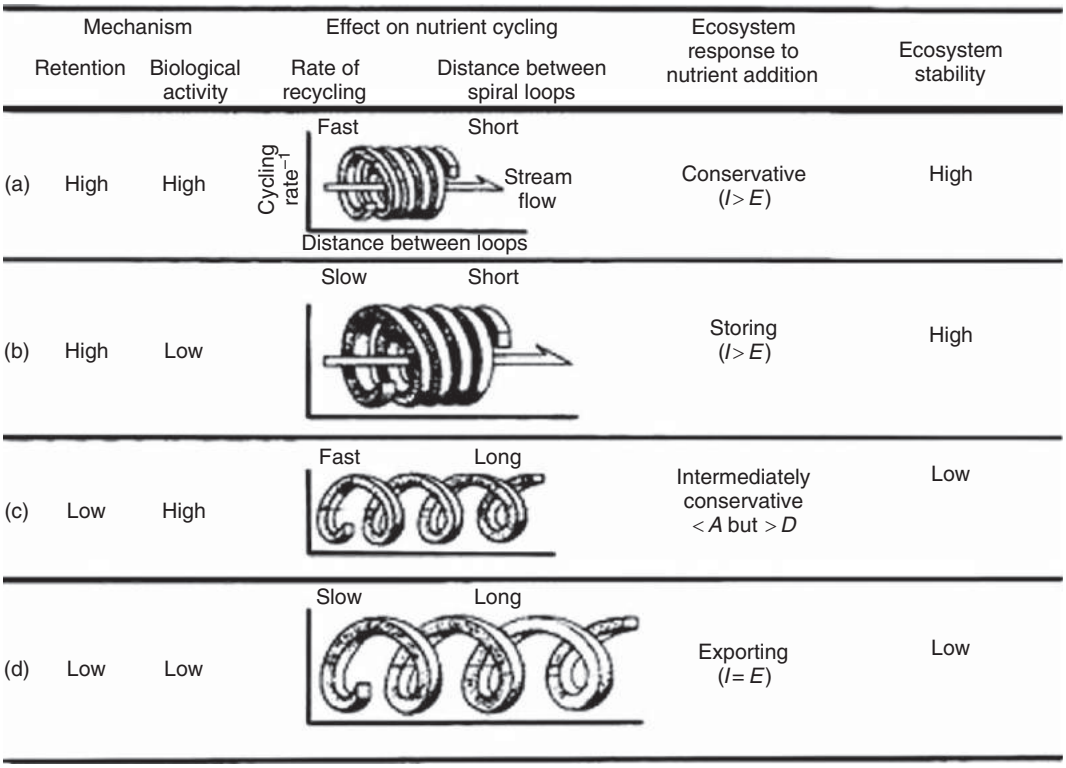


Fig. 4 Nutrient spiraling depicted as the effects of different interactions between the distance of downstream movement (velocity × time) and measures of biological activity such as metabolism by benthic microbes. Modified from Minshall, G.W., Petersen, R.C., Cummins, K.W., *et al.* (1983). Interbiome comparison of stream ecosystem dynamics. Ecological Monographs 53: 1–25.

off-channel areas serve as sources or sinks for OM over an annual cycle depends upon the configuration of the upper banks and floodplains and the patterns of the flood flows. The general fertility of floodplains suggests that they are largely sinks.

Integrative Paradigms in Lotic Ecology

Paradigms, or conceptual models, have continued to be developed, modified, and integrated since the 1980s. The RCC, arguably the most encompassing of these, has guided a large portion of the research on lotic ecosystems in the interim. However, a number of other models have served to elucidate specific components of running-water structure and function or have proposed alternative broad integrating principles.

The RCC

The major goal of the architects of the RCC was to examine the patterns of biological adaptation that overlay the physical setting (template) of stream/river channels in a watershed. The RCC views entire fluvial systems, from headwaters to their mouths, as continuously integrated series of physical gradients together with the linked adjustments in the associated biota. The RCC was founded on many antecedent studies and additional correlates have been incorporated into the general paradigm. Subsequent views and critiques of portions of the RCC also have had significant impact on the present form of the RCC as a general model of lotic ecosystem structure and function. The physical RCC template, and biological communities adapted to it, are viewed as changing in a predictable fashion from stream headwaters to river mouth (**Figs. 5–7**). Major generalizations of the RCC involve seasonal spatial variations in OM supply (e.g., algal/detrital biomass), structure of the invertebrate community, and resource partitioning along drainage networks (**Fig. 8**).

The RCC predicts that recognizable patterns in the structure of biological communities and the input, utilization, and storage of OM will be observable along the continuum. Light limitation inhibits primary production in the headwaters (orders 1–3) due to shading of the channel by riparian vegetation and in the larger rivers (orders greater than 7 or 8) because of light attenuation through the turbid water column that is typical of the lower portions of stream/river networks. The cumulative effect of drainage networks tends to increase nutrient levels in the downstream direction. Overall periphyton and rooted aquatic vascular plant biomass, and insect and fish diversity are all maximized in the mid-sized running waters (orders 4–6). High biotic diversity is

Fig. 5 A headwater stream in the Cascade Mountains of Oregon, USA at winter base flow. The coniferous riparian zone provides partial canopy closure and supplies large woody debris to the channel. Photo by Wilzbach.

Fig. 6 The Firehole River, a mid-sized stream flowing through woodland meadow in Yellowstone National Park, USA. Reduced riparian shading enables abundant growth in in-stream algae. Photo by Wilzbach.

supported in rivers in this size range both because of the variety of habitats and food resources for consumer organisms, and because of overlapping ranges of organisms with evolutionary terrestrial origins (such as the insects) that are dominant in the headwaters with those of marine origins (e.g., annelids and mollusks) that are more prevalent downstream (Fig. 9).

The RCC has been widely utilized as an organizing principle and has been the subject of many studies resulting in various tests of the concept. As would be expected, a degree of unpredictability in the physical template leads to correspondingly less predictability in the overlay of biological communities. This lack of predictability is often a function of the spatial and temporal scales

Fig. 7 The Smith River in coastal northern California, USA. This high order river with a canyon-controlled channel is dependent upon organic matter delivered from the upstream tributary network. Photo by Wilzbach.

of reference employed and it can also be induced by human interference. For example, systems may appear more or less variable over time spans of less than a decade, the time period of observation, but long-term variability, at the scale of centuries or greater, is usually obscured by short-term variations.

Other Paradigms

At least eight paradigms, other than the RCC, continue to guide the development of running-water ecosystem theory (**Table 3**). These include serial discontinuity, hierarchical scales, riparian zone influences, flood pulse, hyporheic dynamics, hydraulic stream ecology, patch dynamics, and network dynamics.

Serial discontinuity

Interruptions in the longitudinal continuum, as proposed by the RCC, are caused by engineered impoundments which serve to reset the general patterns of biotic organization. Above a dam, the system exhibits characteristics of a higher order than the impounded stream. Below a dam, the regulated flows often completely alter the seasonal hydrological patterns of the receiving channel. For example, the normal pattern imposed by the discontinuity resulting from a dam is to decrease the flows during natural high-water periods (reservoir storage phase) and release the water during natural low-flow periods (reservoir release phase). The storage phase retains water to prevent flooding and to provide a later water supply during dry periods. During the release phase, water is delivered for irrigation, drinking, recreation, and, in some cases, to improve fish habitat. In some basins, interruptions in the longitudinal profile of river networks occur in the form of natural lakes or impoundments. Although these may change the sequencing of stream orders, they do not change the annual hydrograph.

Hierarchical scales

The hierarchical scales paradigm addresses a weakness of the RCC. The relative significance of the factors driving the physical, chemical, and biological components of running waters changes with scale. The data on which the RCC is based were all collected at the reach scale and during only several seasons. The hierarchical approach recognizes that ecosystem processes operating at the reach scale and over short time periods do not adequately represent the patterns viewed over greater spatial and temporal scales. Therefore, descriptions of stream/river ecosystem structure and function must be placed in the appropriate context of space and time.

Ecotones

Several paradigms focus on the ecotones that bridge between the stream/river channel and its surroundings and underpinnings. These include the riparian border primarily along small streams, the aquatic terrestrial interface of large rivers with floodplains, and the subsurface region of the sediments beneath running waters (the hyporheic zone).

Riparian zone influences

The riparian zone paradigm attempts to integrate the physical processes that shape the valley floor of streams and rivers with the coupled succession of terrestrial plant communities in the riparian zone along the channel and the role they play in the formation of stream habitat and the production of nutritional resources for organisms that reside in running waters. The ecotone between wetted channel and terrestrial bank vegetation that constitutes the riparian zone is a critical coupling, especially in small streams. The confined or unconstrained nature of the channel system, for example, exert a major influence on the nature of the riparian vegetation that develops along the banks. Definition of the lateral boundaries of the riparian zone, or buffer strip width in the

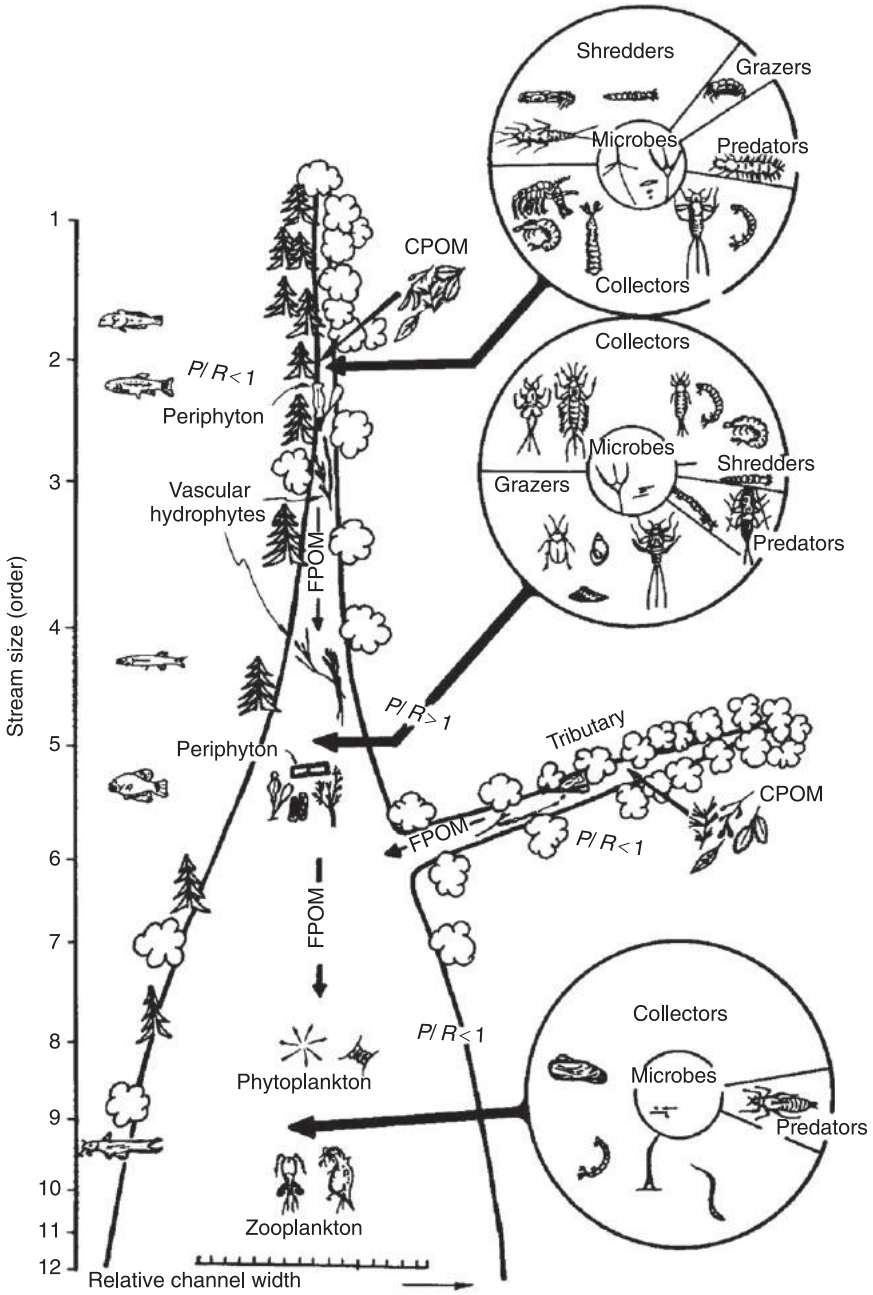


Fig. 8 The "river continuum concept" (RCC). A proposed relationship between stream size (order) and the progressive shift in structural and functional attributes of lotic biotic communities. The heterotrophic headwaters and the large rivers are both characterized by an autotrophic index, or *P*/*R* (ratio of gross primary production to total community respiration) of less than 1 (*P*/*R* = <1). The largely unshaded mid-sized rivers are generally classified as autotrophic with a *P*/*R* = >1. The invertebrate communities of the headwaters are dominated by shredders and collectors, the mid-sized rivers by grazers (= scrapers) and collectors. The large rivers are dominated by FPOM-feeding collectors. Fish community structure grades from invertivores in the headwaters to invertivores and piscivores in the mid-sized rivers to planktivores and bottom-feeding detritivores and invertivores in the largest rivers. From Vannote, R.L., Minshall, G.W., Cummins, K.W., Sedell, J.R., Cushing, C.E., 1980. The river continuum concept. Canadian Journal of Fisheries and Aquatic Sciences 37, 130–137.

parlance of timber managers, is a continuing debate. From the perspective of the stream/river ecosystem per se, the functional roles of the riparian corridor encompass differing areas along the stream bank (**Fig. 10**). Shading of the channel, which along with nutrient levels regulate in-stream primary production, which in turn depends upon the height and foliage density of the vegetation, steepness of the side slopes, and aspect (compass direction) of the channel. The width of the riparian zone that yields litter inputs and large woody debris to the channel can also vary with height and species composition of the stream-side vegetation. Seasonal timing of litter drop and its introduction into the stream produces patterns around which the life cycles of many steam invertebrates have become adapted. This coupling between riparian litter inputs and stream invertebrates is most direct for

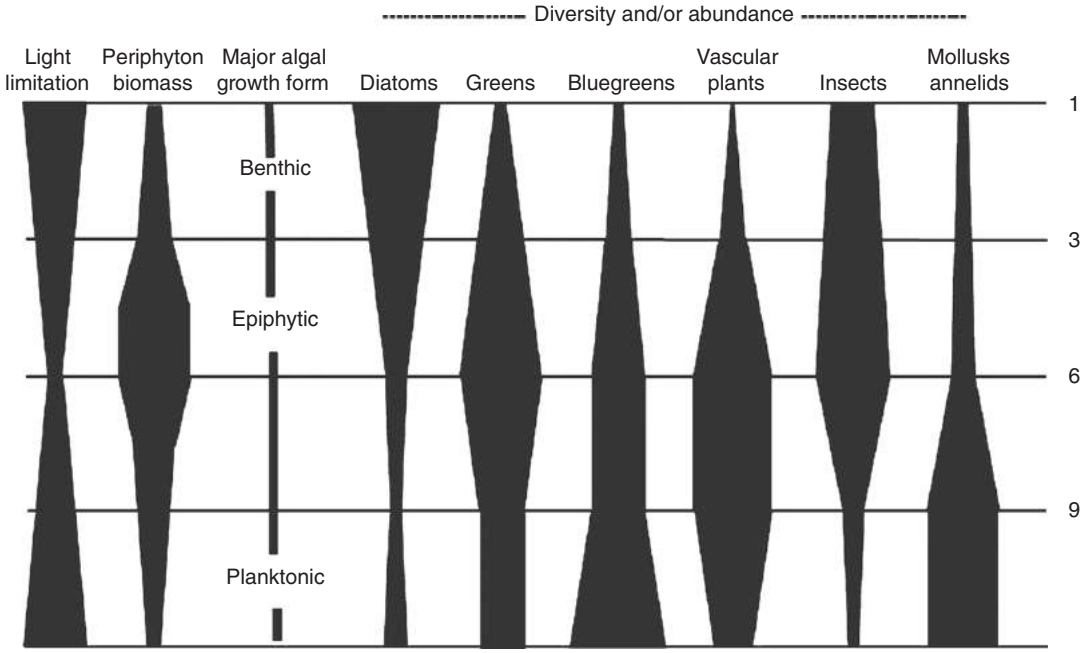


Fig. 9 Patterns in categories of biotic diversity, from small streams to large rivers, compared on a relative scale for each parameter, as predicted by the "river continuum concept." Numbers at the right are general stream/river order ranges. Modified from Cummins, K.W., (1997). Stream ecosystem paradigms. In: CNR – Instituto de Ricerca Sulle Acque. Prospettive di recerca in ecologia delle acque. Roma, Italia.

invertebrate shredders that feed on conditioned litter (Fig. 11). The timing of the inputs of litter to streams varies among

Table 3 Comparison of the most appropriate scales of application for eight commonly used paradigms (conceptual models) for running water analysis

Basin or reach scale	*Stream orders or m reach length*	*Scale designation*	*RCC*	*HS*	*RZI*	*FPC*	*HD*	*HSE*	*PD*	*ND*
Basin	Macro	0/1 Order to estuary	a	a	b	a	c	c	c	b
	Meso	0/1 Order to order 6	a	a	b	a	c	c	c	b
	Micro	0/1 Order to order 2–5	a	a	a, e	d	c	c	c, a	b
Reach	Macro	>1000 m	d	a	b	d	a	c	a	c
	Meso	100–1000 m	d	a	a	c	a	a	a	a
	Micro	<10 m	d	a	a	c	a	a	a	d

Headings: *RCC*, river continuum concept; *HS*, hierarchical scales; *RZI*, riparian zone influences; *FPC*, flood pulse concept; *HD*, hyporheic dynamics; *HSE*, Hydraulic stream ecology; *PD*, patch dynamics; *ND*, network dynamics. *Table entries*: *a*, most direct influence on stream biota and ecosystem processes; *b*, if channels are braided, ranking moves down in scale to a lower order; *c*, beyond the scale to detect specific (local) differences; *d*, influence too local to detect general, large-scale patterns; *e*, may be of less direct importance in naturally deranged (lake-interrupted) or beaver-influenced stream drainages.

ecoregions and with the species composition of the riparian vegetation. Most deciduous species of riparian trees and shrubs are colonized by aquatic hyphomycete fungi and fed on by shredders much more rapidly than evergreen species such as conifers and rhododendrons and some deciduous species such as oaks (Fig. 12). Roots of riparian vegetation stabilize banks at the edge of the channel and influence the chemistry of subsurface flow into the channel. The width of the riparian zone that encompasses these root functions also varies. Thus, a complete definition of the riparian zone that encompasses all these functions would need to be of sufficient width to accommodate all of them. Just as zones of influence of these riparian functions vary, so do the zones of associated management. For example, if a goal is to manage for the long-term input of large woody debris to the channel from the riparian zone to provide habitat structure for fish and invertebrates, all trees tall enough to reach the channel when they fall and large enough to provide habitat structure should be left in place. However, to accommodate variations in channel and bank morphology and the composition of the riparian vegetation, this management would need to be implemented at the reach scale.

Flood pulse concept

The flood pulse concept addresses the ecotones between rivers and their floodplains. Unlike the lateral riparian influence on stream ecosystem processes where the impact is largely from the landscape to the stream, the flood pulse concept emphasizes the

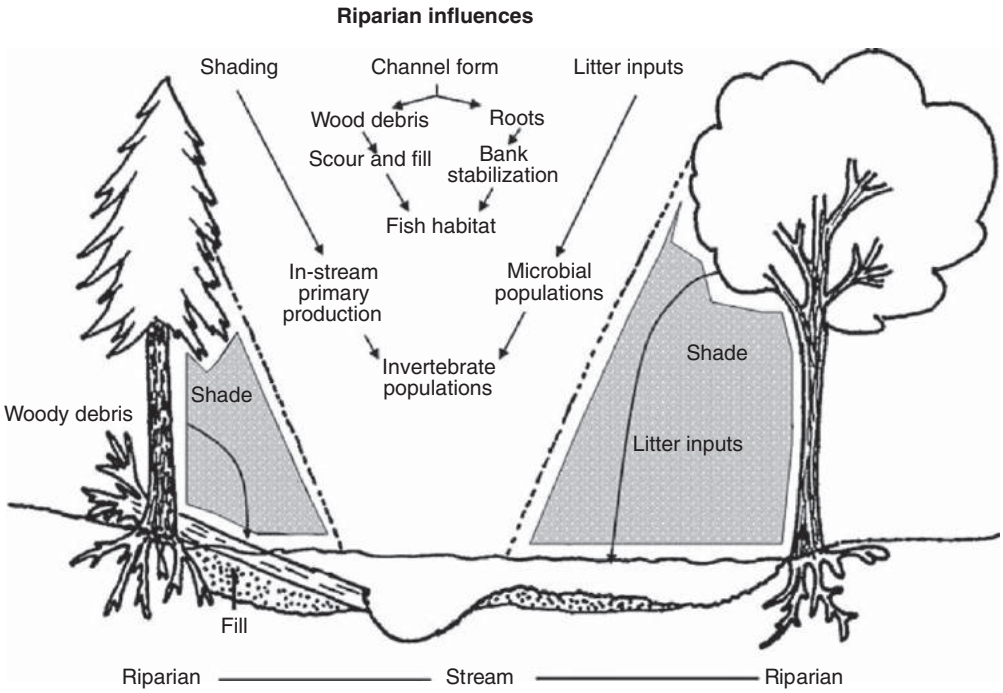


Fig. 10 Influences of the riparian zone on streams. Redrawn from Cummins, K.W., (1988). The study of stream ecosystems: A function view. In: Pomeroy, L.R., Alberts, J.J., (eds). New York: Springer.

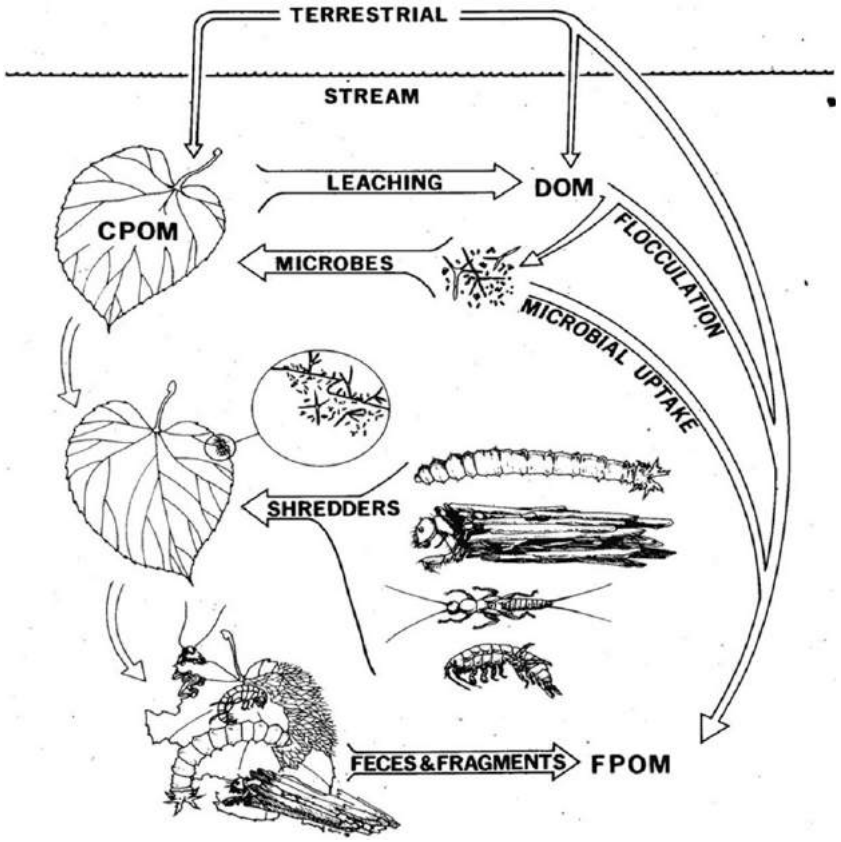


Fig. 11 Conceptual model of the sequence of leaf litter entrainment in a stream through use by shredders and the production of FPOM in the form of shredder feces and fragmentation of the litter. From Cummins, K.W., Klug, M.J., (1980). Feeding ecology of stream invertebrates. Annual Review of Ecology and Systematics 10: 147–172.

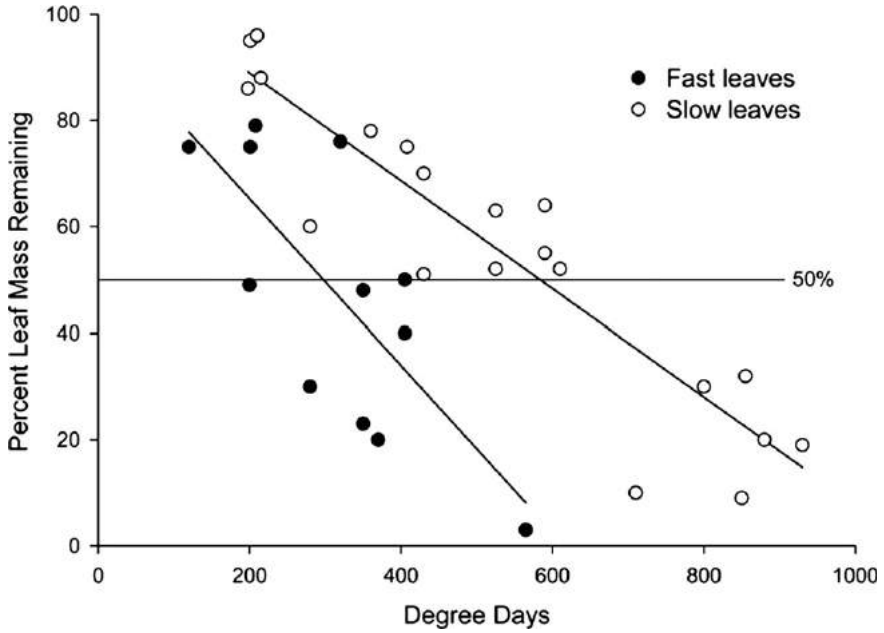


Fig. 12 Linear regressions of processing (turnover) rate per degree day against degree days for a fast leaf type (basswood) and a slow leaf type (oak). Modified from Cummins, K.W., Wilzbach, M.A., Gates, D.M., Perry, J.B., Taliaferro, W.B., (1980). Shredders and riparian vegetation. BioScience 39: 24–30.

reciprocal exchange between the major river channel and its floodplain. A consequence of this distinction is that the overwhelming bulk of riverine animal biomass derives directly or indirectly from production on the floodplain and not from downstream transport of OM produced higher in the watershed. Although the importance of this aquatic/terrestrial transition zone, the floodplain ecotones, is widely acknowledged, there are few hard data to indicate whether over annual cycles, or longer periods, the primary movement of nutrients and biomass is onto or off the floodplain, or in balance. The general perception of "fertile floodplains" suggests that the periodically inundated floodplains are sinks relative to the river channel. However, the high productivity of adult fish in many floodplain rivers and the concentration of reproductive activity on the floodplain supports the notion that floodplains are sources and the river is a sink, gradually exporting to the sea. At any rate, the seasonal pulsing of river discharge, the flood pulse, is the major force controlling existence, productivity, and interactions of biota in river–floodplain ecosystems.

For any given storm or series of storms, the movement of material and organisms on to the floodplain follows the rising limbs of the hydrographs, and the return to the river channel follows the falling limbs. Unfortunately, application of the flood pulse concept is restricted because of the wide-scale engineering modifications that have isolated rivers from their floodplains. Natural exceptions to the flood pulse concept are rivers flowing through deeply incised canyons.

Hyporheic dynamics

The hyporheos is the subsurface region beneath and adjacent to stream/river channels that exchanges water with the surface water. This surface water–ground water ecotone is spatially and temporally very dynamic. The conceptual framework of the hyporheic dynamics paradigm has resulted in the incorporation of channel-aquifer dynamics into the general model of the RCC. The hyporheic zone occurs, at least to some extent, beneath and lateral to the active channel from the headwaters to river mouths, except in bedrock channels. Water, solutes, inorganic and organic complexes, and uniquely adapted biota move through interstitial pathways into and out of the sediments. These flow paths are determined by the bedform of the channel. Where the bedform is convex, there is groundwater recharge from the channel into the sediments. These locations where oxygenated water is driven into the sediments are often spawning sites of salmonid fishes. Concave bedforms are sites of groundwater upwelling into the channel. Invertebrates found in the hyporheos include the small early instars of a wide range of taxa and larger forms at times of extremes inflow, either high or low. In addition, there are some microbenthic forms that are specifically adapted to a groundwater existence. The presence of hyporheic invertebrates is determined largely by siltation and the availability of oxygen. If interstitial sediment spaces are filled with fine sediments and/or conditions are anaerobic, the fauna will be excluded.

Hydraulic stream ecology

The hydraulic stream ecology model emphasizes that the local responses of stream organisms to flow conditions can serve as an organizing principle for running waters. Lotic animals are not well-adapted to hydraulic stress and can sustain exposure to such stress for only short periods of time. Common patterns of diurnal and seasonal drift of stream invertebrates along vertical

gradients of sediment and current velocity are a manifestation of this stress response. The model identifies mean water velocity and depth as more critical than characteristics of substrates in determining the distributions of stream animals. The model has yet to adequately incorporate the extensive data that implicates the diurnal light cycle as the major control parameter of invertebrate stream drift.

Patch dynamics

The patch dynamics concept emphasizes the patchy distribution of riverine habitats in space and time, and argues that an ever-shifting mosaic of patches enables a greater number of species to co-occur than would be the case under greater environmental constancy. Environmental conditions are predictable in aggregate, but not within a particular patch, and these aggregate conditions confer some regularity in species composition. In this model, particular patch types can be found at any point along the general longitudinal gradient proposed by the RCC. However, there are clear examples of invertebrate "patches" that at least change in abundance from headwaters along the continuum to large rivers. For example, small headwater streams (orders 1–3) are generally better shaded than higher orders (>3) and sustain less suitable algal periphyton to support scrapers. Further, the dominance of the CPOM–detrital shredder linkage correlates with stream width and the close availability of riparian tree and/or shrub litter, and this generally matches with stream orders 1–3. The extension of the shading of periphyton growth and the riparian CPOM–shredder linkage to larger rivers can occur along braided channels, but these "patches" will always be more abundant in the headwaters than in mid-sized or larger rivers.

Network dynamics

The network dynamics hypothesis, which combines the hierarchical scales and patch dynamics models, is based on the observation that there are abrupt changes that occur at the confluences of tributaries with the receiving channel. Changes in water and sediment flux at these locations result in changes in the morphology of the receiving channel and its floodplain. In this view, the branching nature of river channel network, together with infrequent natural disturbances, such as fire, storms, and floods, are the formative elements of the spatial and temporal organization of the non-uniform distribution of riverine habitats. Further, the tributary junctions are proposed hot spots of biological activity. Some data show increased fish diversity and abundance at these junctions, but the influence on other components of the biota has yet to be investigated. The "network dynamics hypothesis" does not address "patches" represented by braided channels.

Whether hydraulic characteristics, tributary junctions, or other patch phenomena represent local conditions that need to be integrated along river continua to account for whole-profile trends that are clearly apparent, or whether such phenomena are localized specific modifiers that differentially affect stream orders along profiles has yet to be demonstrated clearly.

Conservation and Human Alterations of Streams and Rivers

A great challenge for stream and river ecology in the 21st century will be the restoration of degraded running-water ecosystems while preserving those systems that still remain in good condition. Restoration will dominate in more developed regions where modifications of running waters and their watersheds have been more extensive. In less-developed regions, preservation of many running waters may still be possible, but the distinction between pristine and degraded systems is disappearing rapidly. The historical scientific databases for running waters are generally poor, with largely anecdotal or very incomplete information available. The lotic ecosystem paradigms described above can serve as tools for evaluating present conditions of running waters, surmising their likely antecedent condition, and developing targets and strategies for restoration. Because the majority of degraded streams and rivers have changed beyond our ability to return them to their historical state, it is more logical to use the term rehabilitation. Often the actions will take the form of returning certain organisms or processes to a condition that addresses societal objectives.

In the context of preserving and rehabilitating streams and rivers, it will be important to enlist the best scientific understanding of the structure and function of running-water ecosystems. For example, regulations governing the protection and width of riparian buffer strips, designed to protect stream organisms (usually fish), vary among political boundaries or regulatory authorities, and are wider in some areas, narrower in others. However, managers and environmentalists should not limit their view of riparian buffers as only a matter of vegetative composition and buffer width with the sole aim of providing shading to reduce water temperatures, a source of large woody debris, or stream bank stabilization. This view of riparian buffers ignores the often completely different in-stream trophic role played by the coupled riparian ecosystem. The buffer width required to produce shade, litter, large wood, nutrients, and bank stabilization are often quite different. Thus, the management and rehabilitation of a given reach of running water requires an integrated approach that acknowledges all the riparian functions and places the actions within the context of the larger watershed.

See also: Behavioral Ecology: Herbivore-Predator Cycles. Ecological Complexity: Goal Functions and Orientors; Self-Organization. Global Change Ecology: Microbial Cycles

Further Reading

Allan, J.D., Castillo, M.M., 2007. Stream ecology structure and function. The Netherlands: Springer.

Benda, L., Poff, N.L., Miller, D., *et al.*, 2004. The network dynamics hypothesis: how channel networks structure riverine habitats. Bioscience 54, 413–427.

Benke, A.C., Huryn, A.D., Smock, A., Wallace, J.B., 1999. Length-mass relationships for freshwater macroinvertebrates in North America with particular reference to the southeastern United States. Journal of the North American Benthological Society 18, 308–343.

Cummins, K.W., Klug, M.J., 1979. Feeding ecology of stream invertebrates. Annual Review of Ecology and Systematics 10, 147–172.

Cummins, K.W., Wilzbach, M.A., Gates, D.M., Perry, J.B., Taliaferro, W.B., 1989. Shredders and riparian vegetation. Bioscience 39, 24–30.

Frissell, C.A., Liss, W.J., Warren, C.E., Hurley, M.D., 1986. A hierarchical framework for stream classification: viewing streams in a watershed context. Environmental Management 10, 199–214.

Gregory, S.V., Swanson, F.J., McKee, W.A., Cummins, K.W., 1991. An ecosystem perspective of riparian zones. Bioscience 41 (8), 540–551.

Hauer, F.R., Lamberti, G.A. (Eds.), 2017. Ecosystem structure, *Vol. 1.* Methods in stream ecology,Ecosystem structure, vol. 1. Amsterdam: Elsevier Publishing Company.

Junk, W.J., Bayley, P.B., Sparks, R.E., 1989. The flood pulse concept in river-floodplain systems. Canadian Journal of Fisheries and Aquatic Sciences 106, 110–127. Special Publication.

Merritt, R.W., Cummins, K.W., Berg, M.B. (Eds.), 2008. An introduction to the aquatic insects of North America, 4th edn Dubuque, IA: Kendall/Hunt Publishing Company.

Merritt, R.W., Cummins, K.W., Berg, M.B., 2017. Trophic relationships of macroinvertebrates. In: Hauer, F.R., Lamberti, G.A. (Eds.), Methods in stream ecology,Ecosystem structure, vol. 1. Amsterdam: Elsevier Publishing Company.Chapter 20.

Minshall, G.W., Petersen, R.C., Cummins, K.W., 1983. Interbiome comparison of stream ecosystem dynamics. Ecological Monographs 53, 1–25.

Saunders, G.W., *et al.*, 1980. In: LeCren, E.D., McConnell, R.H. (Eds.), The functioning of freshwater ecosystems. Great Britain: Cambridge University Press.

Stanford, J.A., Ward, J.V., 1993. An ecosystem perspective of alluvial rivers: connectivity and the hyporheic corridor. Journal of the North American Benthological Society 12, 48–60.

Statzner, B., Higler, B., 1986. Stream hydraulics as a major determinant of benthic invertebrate zonation patterns. Freshwater Biology 16, 127–139.

Tank, J.L., Rosi-Marshall, E.J., Griffiths, N.A., Entrekin, S.A., Stephen, L., 2010. Review of allochthonous organic matter dynamics and metabolism in streams. Journal of the North American Benthological Society 29, 118–146.

Townsend, C.R., 1986. The patch dynamics concept of stream community ecology. Journal of the North American Benthological Society 8, 36–50.

Vannote, R.L., Minshall, G.W., Cummins, K.W., Sedell, J.R., Cushing, C.E., 1980. The river continuum concept. Canadian Journal of Fisheries and Aquatic Sciences 37, 130–137.

Ward, J.V., Stanford, J.A., 1983a. The serial discontinuity concept of river ecosystems. In: Fontaine, T.D., Bartell, S.M. (Eds.), Dynamics of lotic ecosystems. Ann Arbor, MI: Ann Arbor Science Publications, pp. 29–42.

Rivers and Streams: Physical Setting and Adapted Biota☆

Margaret A Wilzbach, Humboldt State University, Arcata, CA, United States
Kenneth W Cummins, Michigan State University, East Lansing, MI, United States

Glossary

Alluvial Deposition of mineral or organic particles by a river. An alluvial river is one in which the channel bed and banks are composed of transported sediment.
Benthic Of, relating to, or occurring at the bottom of a body of water. *Benthos* refers to the community of organisms that live on or in the bottom sediments of a river or other aquatic environment.
Community respiration Metabolism of organic matter by both plants and heterotrophs (bacteria, fungi, and animals).
Detritivore An organism that feeds on and breaks down dead plant or animal matter.
Floodplain Low-lying areas of land bordering a river subject to flooding and formed largely from deposited river sediments.
Geomorphology The study of the physical features of the surface of the earth and the processes which create them.
Hydrology The science that encompasses the occurrence, distribution, movement and properties of the waters of the earth and their relationship with the environment within each phase of the hydrologic cycle.
Invertivore An organism which feeds on invertebrates.
Lotic Of, related to, or living in moving waters.
Point bar Accumulation of deposited sediments along the inner bank of a stream meander.
Primary production Rate of formation of organic matter from inorganic carbon by photosynthesizing organisms in the presence of light.
Riparian zone The land area adjacent to a stream channel influenced by stream-derived moisture.
Water table Upper surface of groundwater below which soil is saturated with water that fills all voids and interstices.

Introduction

Streams and rivers are enormously important ecologically, economically, recreationally, and esthetically. This importance far outweighs their proportional significance on the landscape. Running waters constitute less than 1/1000th of the land surface and of freshwater resources of Earth and contribute only 1/5,000th of annual global freshwater budgets. Streams and rivers are significant agents of erosion and serve a range of human needs, including transportation, waste disposal, recreation, and water

☆*Change History*: February 2018. MA Wilzbach and KW Cummins added an Abstract and a Glossary to define technical terms; added new references and deleted others; introduced small edits in the text of the article and in figure captions; added information on the hyporheic zone, changing patterns in runoff in response to climate change, and impact of altered hydrologic regimes on species diversity and imperilment (in Hydrologic Processes): added discussion of channel units other than pools and riffles and reach types other than pool-riffle, the time frame over which different components of river morphology respond to changes in discharge and sediment supply (in Channel Morphology); added information on forces propelling and resisting current and defined large woody debris (in Physical Factors); and added information on the extent of species imperilment of mollusks (in Adapted Biota).

for drinking, irrigation, hydropower, cooling, and cleaning. At the same time, flooding of streams and rivers pose potential natural hazards to human populations. Irrespective of their direct consequences for people, streams and rivers are rich, complex ecosystems that are diagnostic of the integrity of the watersheds through which they course. Unfortunately, species dependent on flowing waters are in global decline, exhibiting extinction rates four to five times greater than those for terrestrial species.

Streams are generally conceived of as smaller, shallower, and narrower bodies of flowing water than rivers. However, the difference between streams and rivers lacks clear distinction in the literature of the last 100 years. For the purposes of this article, streams refer to channels in drainage networks up to order 5 and rivers as orders 6 and above (see definition of stream order under the section titled "Channel Morphology"). In this article, we discuss the history of stream ecology, followed by a treatment of the physical and chemical setting and biological features of major groups of lotic organisms. In a companion article, we cover ecosystem dynamics and integrating paradigms in stream and river ecology.

History of the Discipline of Stream and River Ecology

The formal published beginning of the study of flowing waters (lotic ecology) dates to the early 20th century in Europe, where initial work focused on the distribution, abundance, and taxonomic composition of lotic organisms. In North America, ecological studies of streams began shortly after. In the 1930s, fishery biology dominated North American stream ecology. Stream and river studies worldwide remained descriptive through the 1950s and this period marked the beginning of a focus by stream ecologists on human impacts. Descriptive studies detailed the taxonomic composition and density of the benthic invertebrate fauna found in reaches of streams and rivers variously affected by human impacts. Beginning in the 1960s and 1970s, research interest shifted to synthetic views of flowing-water ecosystems, with research addressing energy flow and organic matter budgets in small catchments. In 1970, H.B.N. Hynes, pioneer of modern stream ecology, published his landmark book, *The Ecology of Running Waters*, which summarized concepts and literature to that point. With the 1980s came the realization that running-water dynamics demanded an integrated spatial and temporal perspective, and that whole catchments were the basic units of stream/river ecology. For example, holistic organic budget analyses of running-water ecosystems cannot be constructed unless both spatial and temporal scales are applied.

The hallmark of lotic research during the 1980s and 1990s was its interdisciplinary nature. Interactions among stream biologists, hydrologists, geomorphologists, microbiologists, and terrestrial plant ecologists focused attention on physical processes and greater spatial and temporal scales. This perspective of stream ecosystems continues to direct the science in the 21st century, aided immensely by the incorporation of geographic information systems (GIS) analysis and incorporation of concepts from landscape ecology. Emergent areas of study within stream ecology include urban stream ecology, bioassessment, genetic barcoding, food web studies, nutrient cycling, and restoration.

Despite recognition among stream ecologists that the watershed or catchment represents the fundamental unit for the study of streams and rivers, measurements of lotic ecosystem structure and function continue to be commonly made at the reach or microscale level. Impetus has been strong to extend the scope of understanding to the watershed mesoscale and beyond, because ecosystem processes exhibit effects of differing importance at different spatial and temporal scales and these processes interact across scales. Growing concern about impacts of climate change on lotic ecosystems has provided additional motivation to analyze entire basins or all the basins in continental regions. Integration of data-rich studies at the reach level to entire watersheds and the coarse resolution of regional basin analysis relying on satellite imagery remain a challenge for lotic ecologists in the 21st century. The River Continuum Concept and other stream/river conceptual models, coupled with modeling and field experiments conducted at an expanded spatial and temporal scale, continue to aid in the integration of knowledge about lotic ecosystems.

The Physical and Chemical Setting

Stream and river biota evolved in response to and in concert with the physical and chemical setting. Traditionally the domain of hydrologists, geomorphologists, and chemists, the study of processes driving the physical and chemical templates have been embraced by stream ecologists for interpreting patterns in organismic distributions and lotic ecosystem structure and function. From a purely physical perspective, the primary function of rivers is to transfer runoff and move weathering products away from the Earth's land surfaces for delivery to the oceans or terminal water bodies. Despite tremendous variability in the morphology and behavior of rivers, each river reflects the interaction between geomorphic and hydrologic processes. We summarize these processes and their effect on river morphology, and discuss major physical and chemical drivers of river ecosystem dynamics and organismic biology.

Hydrologic Processes

The total amount of the Earth's water does not change. Water is continuously recycled among various storage compartments within the biosphere in a process referred to as the hydrologic cycle (Fig. 1). The cycle involves solar-driven evaporation of moisture from land and evapotranspiration from terrestrial vegetation, cloud formation, and precipitation.

Annual global precipitation averages about 100 cm; the majority of this evaporates and relatively little falls directly into streams. The remainder that falls on land surfaces either infiltrates into the soil or becomes surface runoff. Relative contributions of different

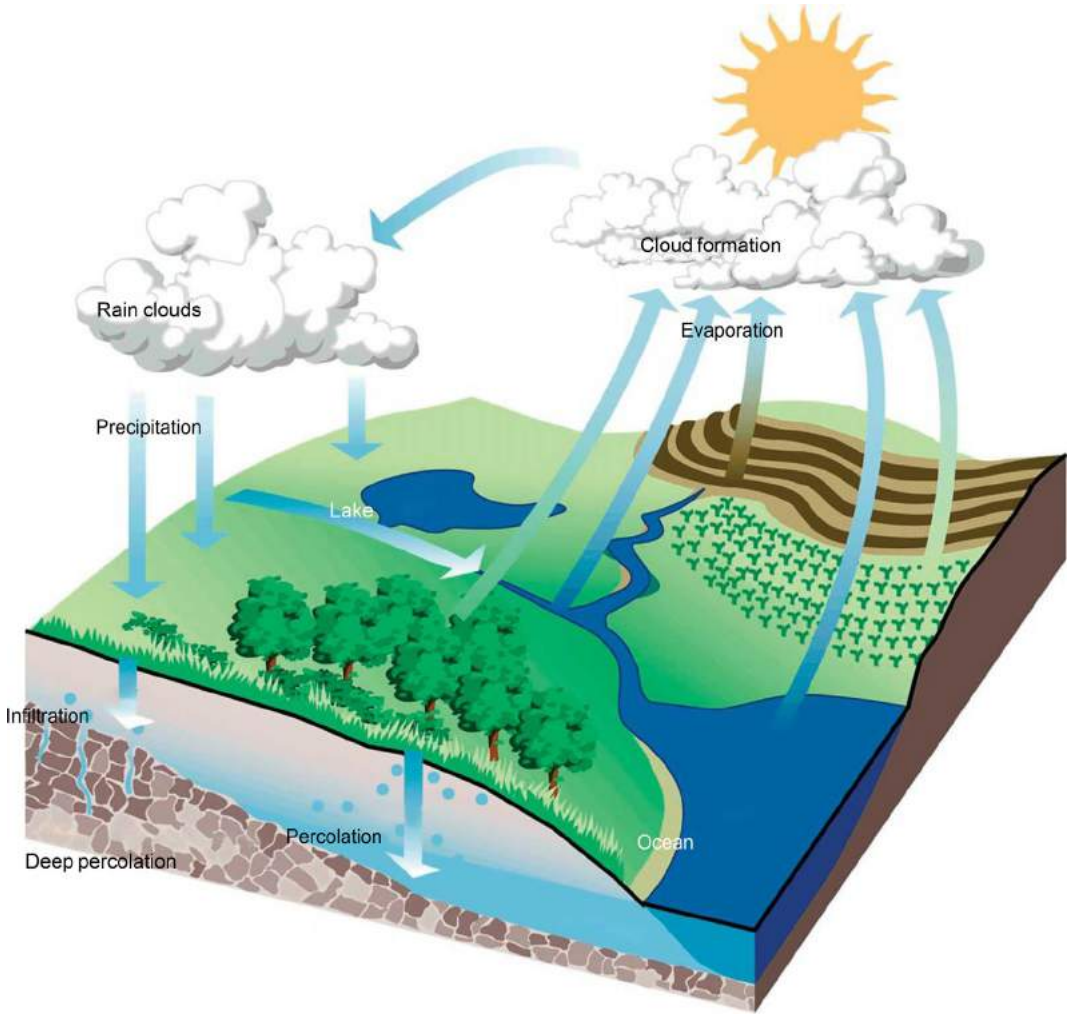


Fig. 1 The hydrologic cycle. From Stream Corridor Restoration: Principles, Processes, and Practices, 10/98, by the Federal Interagency Stream Restoration Working Group (FISRWG).

pathways by which water enters streams and rivers varies with climate, geology, watershed physiography, soils, vegetation, and land use.

Water that infiltrates becomes groundwater, which makes up the largest supply of unfrozen freshwater. Groundwater discharges gradually to stream channels through springs or direct seepage when a channel intersects the groundwater table. The hyporheic zone is the region beneath and alongside a streambed where shallow groundwater mixes with surface water. This zone plays an important role in lotic nutrient dynamics and benthic ecology.

Baseflow describes the proportion of total streamflow contributed from groundwater, and it sustains streamflow during periods of little or no precipitation. Streams can be distinguished on the basis of the balance and timing of stormflow versus baseflow that they receive. Ephemeral streams carry water only during and immediately after a precipitation event, and never intersect the water table. Intermittent streams flow only during certain times of year from springs or from runoff from precipitation, and may flow above or below the water table (Fig. 2). Perennial streams flow continuously during wet and dry periods, and receive both stormflow and groundwater flow. The duration, timing, frequency, magnitude, and predictability of flow greatly affect the composition and life-history attributes of stream communities. A study of 900 rivers published in the American Meteorological Society's Journal of Climate in 2009 concluded that river flows into the oceans have decreased significantly over the last 50 years as a result of changing patterns of precipitation and rising temperatures, and predicted a continuing trend of reduced runoff and greater water scarcity.

Discharge, the most fundamental of hydrological measurements, describes the volume of water passing a channel cross-section per unit time and is computed as the product of the cross-sectional area and the average velocity of water in the cross-section. Any increase in discharge must result in an increase in width, depth, or velocity of a stream channel, or some combination of these. Discharge usually increases in a downstream direction through tributary inputs and groundwater addition, accompanied by

Fig. 2 An intermittent stream at 3.4 km elevation in the Andes Mountains in Chile, bordered by riparian vegetation of herbs and grasses. Intermittent streams are often important in exporting invertebrates and organic detritus to downstream fish-bearing reaches. Photo credit: KW Cummins.

increases in channel width, depth, and velocity. An estimated 35,000 km^3 of water is discharged annually by rivers to the world's oceans, with the Amazon River alone accounting for nearly 15% of the total.

Hydrographs depict changes in discharge over time. The hydrograph of an individual storm event typically displays in succession a steep rising limb from direct runoff, a peak, and a gradually falling recession limb as the stream returns to prestorm conditions (Fig. 3). Variability in the shapes of hydrographs among streams reflects differences in the climatic, geomorphic, and geologic attributes of their watersheds, and differences in the distribution of runoff sources (e.g., subsurface vs. surface, snowmelt vs. rainfall-dominated).

Discharge records of sufficient duration allow prediction of the magnitude, frequency, and periodicity of flood events for a given river and year. Recurrence interval (T, in years) for an individual flood is estimated as

$$T = (n + 1)/m$$

where n is the number of years of record, and m is ranked magnitude of the flood over the period of record, with the largest event scored as $m = 1$. The reciprocal of T is the exceedance probability, which describes the statistical likelihood that a certain discharge will be equaled or exceeded in any given year. Thus, a 1-in-100-year flood has a probability of 1% of occurring in any given year. Recurrence interval information provides important context for interpretation of studies conducted on lotic organisms. Widespread alteration of the timing and magnitude of the natural flow regime of river systems throughout the world has resulted in reduced biotic diversity and loss of sensitive aquatic species.

Geomorphic Processes

Discharge and sediment supply represent the physical energy and matter that move through river systems, and the form and profile of rivers change over time to accommodate the energy and matter delivered to them. Sediments supplied to rivers originate primarily from the physical and chemical weathering of bedrock and soils, transported to rivers largely by direct runoff of rainfall

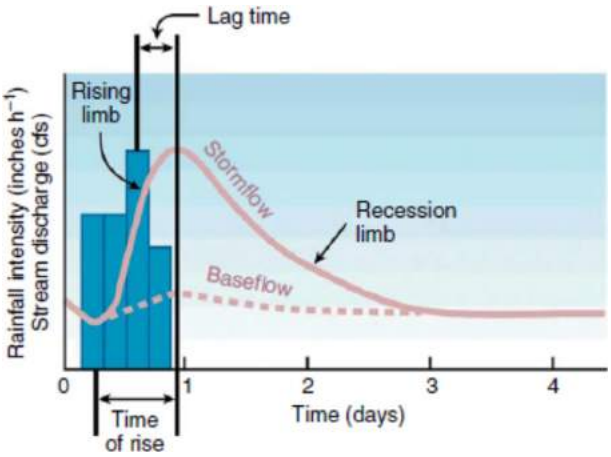


Fig. 3 Stream hydrograph from a rainstorm event. From Stream Corridor Restoration: Principles, Processes, and Practices, 10/98, by the Federal Interagency Stream Restoration Working Group (FISRWG).

and, especially in areas of unstable terrain, sometimes via mass wasting processes. Erosion of alluvial sediments stored within the channel, its banks, and floodplain also contribute to sediment supply.

Within a channel, sediment movement is initiated as a function of drag and lift forces exerted on deposited particles. The greater the shear stress exerted on the streambed, the greater the sediment particle size that can be entrained into the water column. Stream competence and stream capacity refer to the largest particle size moved by a given set of flows and the total amount of sediment that can be transported, respectively. Coarse sediment moves along the stream/river bottom as bedload, and fine sediment moves downstream in the water column as suspended load. As suspended particles absorb and scatter light, high concentrations of suspended sediment reduce light penetration needed for aquatic primary production. Sediment deposition occurs when water velocity falls below the settling velocity for a given particle size, with the largest particle sizes deposited first. Whereas sediments may be temporarily deposited within mid-channel or point bars, longer-term storage occurs on floodplains and elevated alluvial terraces.

Channel Morphology

Within a reach, channel cross-sections reflect the interaction between bed and bank materials and flow, and vary from symmetrical in riffles to asymmetrical in pools as flow meanders. Bankfull discharge occurs every 1.5–2 years on average in unregulated systems. The bankfull cross-sectional area of rivers (i.e., the product of surface width and mean depth) is correlated with streamflow and drainage area.

Channel pattern is described by its sinuosity (amount of curvature) and thread (multiple channel braiding). Sinuosity index is measured as channel length along the thalweg (a line drawn through the deepest points of successive cross-sections along the length of the channel), divided by valley length. Meandering channels have a sinuosity index value exceeding 1.5. Erosion of the channel bank carves the river bends, with the fastest current at the outside of the bend where the bank erodes. The greater the curve, the faster the water flows around the bend, deflecting to the other bank and forming the next curve. This pattern repeats downstream, creating regular swings in the river with a meander wavelength approximately 11 times the channel width.

Riffles are topographic high spots along the channel composed of the coarsest bedload sediments transported by the river, and with a water surface slope that is steeper than the mean stream gradient at low flow (Fig. 4). Riffles are typically spaced every five to seven channel widths in reaches that develop pool-riffle sequences. Pools are topographic depressions with fine sediments and reduced velocity. Reach types other than pool-riffle (e.g., cascades, step-pools, plane-bed) and channel geomorphic units other than pools and riffles (e.g., cascades, runs) have also been identified in stream classification schemes (e.g., Rosgen, 1996).

Change in discharge and sediment supply induce adjustments in river morphology on a time scale that increases with the length scale of morphological attribute. Channel width and depth dimensions adjust on a time step of years or less, and meander wavelength and reach gradient adjusts over hundreds of years or longer. Adjustments in the longitudinal profile of a river to changes in discharge and sediment supply occur over thousands of years. River profiles are generally concave, with declining gradients from headwaters to river mouths. The concavity reflects the adjustment between climate and tectonic setting (land relief and base level) and geology, which control sediment supply and resistance to erosion. Base level describes the limit to which a river cannot erode its channel. For streams emptying into the ocean, this is sea level.

Within a drainage basin, stream channels and their networks grow in size and complexity in a downstream direction as described by stream order (Fig. 5). A first-order stream lacks permanently flowing upstream tributaries and order number increases only where two streams of equal order join. Employing this system, the Mississippi and the Nile Rivers at their mouths are order 10. There are usually three to four times as many streams of order $n-1$ as of order n, each of which is roughly half as long, and drains a little more than one-fifth of the land area. In the United States, nearly half of the approximately 5,200,000 km of total river length are first

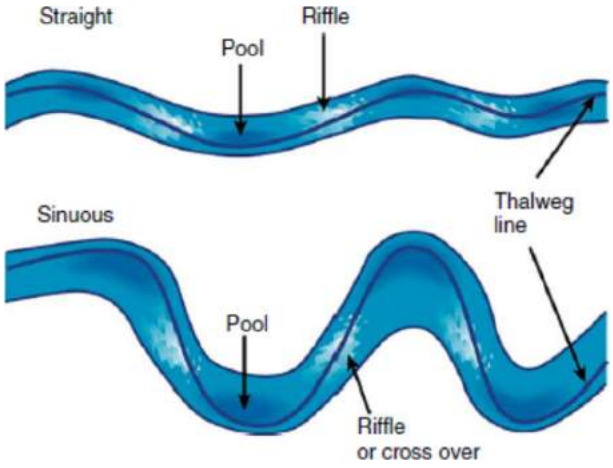


Fig. 4 Riffle and pool sequences in straight and sinuous streams. From Stream Corridor Restoration: Principles, Processes, and Practices, 10/98, by the Federal Interagency Stream Restoration Working Group (FISRWG).

order. As described in the River Continuum Concept, many patterns in biological composition and energy inputs have been found to be correlated with stream order.

Drainage basins, or watersheds, are the total area of land draining water, sediment, and dissolved materials to a common outlet. Watersheds occur at multiple scales, ranging from the largest river basins to first-order watersheds of only a few hectares. Larger watersheds are comprised of smaller watersheds and stream segments in a nested hierarchy. The size and shape of the watershed, and the pattern of the drainage network within the watersheds, exerts a strong influence on the fluxes of energy, matter, and organisms in river systems. Because some energy, matter, and organisms move across and through landscapes independently of drainage basins, a complete perspective of stream ecology requires consideration of landscape ecology.

Physical Factors

Current

Current, the central defining physical feature of running-water systems, refers to the downstream flow of water, propelled by gravity. The forward momentum of water is resisted by friction from the channel bed and banks and viscosity of the water (i.e., resistance to mixing). Velocity, or speed, of the current (m s^{-1}) shapes substrate composition, affects delivery of dissolved oxygen, nutrients, and food and removal of waste materials, and influences the physical forces exerted on organisms in the streambed or water column. Velocity rarely exceeds 3 m s^{-1} in running waters.

Flow in running waters is complex and highly variable in space and time. At a given velocity, flow may be laminar, moving in parallel layers which slide past each other at differing speeds with little mixing, or turbulent, where flow is chaotic and vertically mixed. The dimensionless Reynolds number (*Re*), the ratio of inertial to viscous forces, predicts the occurrence of laminar versus turbulent flow. High inertia promotes turbulence. Viscosity is the resistance of water to deformation, due to coherence of molecules. At $Re < 500$, flow is laminar; at $Re > 2000$ flow is turbulent; intermediate values have transitional flow. Although laminar flow is rare in running waters, microenvironments may contain laminar flow, even within turbulent, high-flow settings. *Re* can be estimated for individual organisms as well as for bulk flow. *Re* increased with the length of an organism. Organisms with high *Re*, such as trout, often exhibit a streamlined body shape to minimize turbulent drag.

In a channel cross-section, a vertical velocity gradient decreases exponentially with depth. Highest velocities are at the water surface where friction is least, and zero at the deepest point of the bottom where friction is the greatest. Mean current velocity is commonly found at 60% of the depth from the surface to the bottom. A boundary layer extends from the streambed to a depth where velocity is no longer reduced by friction and a thin viscous sublayer of laminar flow exists at its base.

Microorganisms and small benthic macroinvertebrates may experience shelter from fluid forces within the narrow sublayer very close to the streambed. However, the sublayer is so thin that most stream organisms must contend with complex, turbulent flow. A variety of morphological and behavioral adaptations in organisms reduce drag and lift forces and risks of downstream displacement. Attributes of macroinvertebrates, for example, can include small size, dorso-ventral flattening to reduce exposure to current, streamlining to reduce current drag, the development of silk, claws, hooks, suckers, and friction pads as holdfasts, and directed movement away from high-velocity areas.

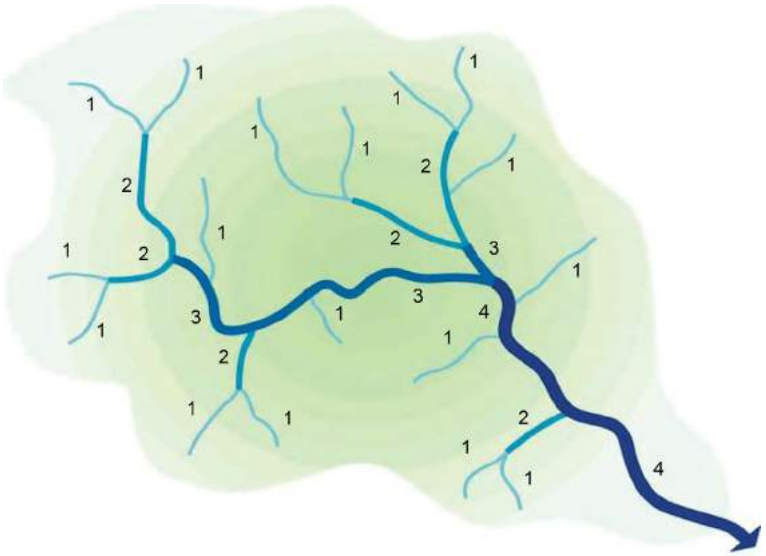


Fig. 5 Ordering of stream segments within a drainage network. From Stream Corridor Restoration: Principles, Processes, and Practices, 10/98, by the Federal Interagency Stream Restoration Working Group (FISRWG).

Substrate

For riverine organisms, substrate provides food or a surface where food accumulates, a refuge from flow and predators, a location for carrying out activities such as resting, reproduction, and movement, and material for construction of cases and tubes. Algal growth, invertebrate growth and development, and fish egg incubation largely occur on or within the substrate. Substrate includes both inorganic and organic materials, often in a heterogeneous mixture. Mineral composition of the substrate is determined by parent geology, modified by the current. Organic materials include aquatic plants and terrestrial inputs from the surrounding catchment ranging from minute fragments and leaves to fallen trees (Fig. 6).

Particle sizes of mineral substrates are commonly categorized using the Wentworth scale, in which each size category represents a doubling of the diameter of the next smaller size category (Table 1). Organic substrate particles <1 mm in diameter and >0.45 μm are classified as fine particulate organic matter or PFOM, and often function as food rather than substrate for benthic organisms. Organic materials >1 mm in diameter are referred to as coarse particulate organic matter or CPOM. CPOM serves as substrate or food for litter-feeding invertebrates (Fig. 7). Large woody debris (LWD) is at the upper end of the size range of CPOM, and is used more often as substrate than invertebrate food. Although different definitions are in use, a common definition specifies large woody debris (LWD) as any piece of wood that is greater than, or equal to 10 cm in diameter over a length of 1.5 m or more. Other than particle size, substrate attributes including shape, surface texture, sorting, and stability also influence benthic community structure, but these are less easily quantified. Diversity, biomass, and abundance of benthic invertebrates generally increases with mineral particle size from sand to pebbles and cobble, but decreases for large boulders and bedrock.

Evaluation of the ecological role of substrate is difficult because of its heterogeneity and covariance with velocity and oxygen supply. Mean particle size generally decreases downstream, but this trend may be locally interrupted at tributary junctions. Spatial heterogeneity in substrate at multiple scales is also apparent both vertically and horizontally within the streambed. Substrate embeddedness describes the degree to which larger sediments, such as cobbles, are surrounded or covered by fine sand and silt. Significant embeddedness reduces streambed surface area and organic matter storage, the flow of oxygen and nutrients to incubating fish eggs and aquatic invertebrates, and entrance to and movement within the streambed.

Temperature

Temperature affects all life processes, including those in running waters. Decomposition, primary production, community respiration, and nutrient cycling are temperature dependent. Most stream organisms are ectothermic, and their metabolism, growth rates, life cycles, and overall productivities are temperature dependent. Annual temperature changes often serve as environmental cues for and/or regulate life-history events of invertebrates and fishes, especially emergence or spawning. The temperature regime sets limits to where species can live, and many species are adapted to specific thermal regimes. Increasing water temperature decreases dissolved oxygen solubility at the same time that it increases metabolic demand. Thus preferences of coldwater organisms such as salmonid fishes may have as much to do with temperature effects on oxygen availability as with effects of temperature per se.

Stream temperature is the net result of heat exchange via (1) net solar radiation, modified by cloud cover, day length, sun angle, vegetation, and topographic shading; (2) evaporation and convection; (3) conduction, or heat exchange with the streambed; and (4) advection from upstream water inputs, including groundwater and tributaries. On a diel basis, stream temperature varies less than air temperature because of the high specific heat of water. The greatest daily fluxes occur in summer in temperate regions, with a minimum before dawn. Groundwater inputs, which usually enter the channel at a temperature within 1°C of mean annual air temperature, can greatly affect the flux, as can canopy cover. At the catchment level, daily temperature flux increases with distance from the headwaters, with a maximum in mid-order segments. Thermal stratification is rare except in large rivers and at tributary junctions. Seasonal variations in temperature mirror trends in mean monthly air temperature, with the timing of the summer maximum often lagging the timing of maximum solar radiation. Year-to-year variation in monthly mean temperatures is low, typically less than 2°C. The annual temperature range of temperate streams is generally 0–25°C, and 0–40°C in intermittent desert

Fig. 6 Small headwater stream in old-growth Douglas-fir forest in Oregon (United States), showing large woody debris spanning the channel. This spanner log forms a retention structure for organic detritus and sediment as well as refugia and habitat when the channel is inundated by high flows. Photo credit: MA Wilzbach.

Table 1 Size categories of inorganic substrates in streams and rivers, based on the Wentworth grain size scale

Size category	*Particle diameter (range in mm)*
Boulder	>256
Cobble	
Large	128–256
Small	64–128
Pebble	
Large	32–64
Small	16–32
Gravel	
Coarse	8–16
Medium	4–8
Fine	2–4
Sand	
Very coarse	1–2
Coarse	0.5–1
Medium	0.25–0.5
Fine	0.125–0.25
Very fine	0.063–0.125
Silt	<0.063

Fig. 7 Accumulation of leaf litter in a second-order stream in Oregon (United States) flowing through a second-growth forest with a red alder riparian zone. The litter that is retained at the leading edge of the cobbles provides the major food resource for stream invertebrate shredders and habitat for other invertebrates. Photo credit: KW Cummins.

streams. The lower Amazon River is always within one or two degrees of 29°C. Extremes in temperature occur in hot springs, which can exceed 80°C, and in subarctic and arctic streams that may completely freeze in winter. Surface freezing is usually prevented by snow and ice bridging, but underwater ice may form on streambeds as anchor ice or in the water column as slush or frazil.

Stream ecologists often evaluate temperature effects on stream organisms and ecosystem processes based on degree-day accumulation rather than temperature maxima or minima. Degree-days, which are calculated by summing daily mean temperatures above 0°C, can differ among streams with similar maximum or minimum average daily temperatures. Such differences can affect number of annual generations of aquatic insects and production potential of aquatic organisms.

Water chemistry

Constituents of river water include dissolved gases, dissolved inorganic ions and compounds, particulate inorganic material, particulate organic material, and dissolved organic ions and compounds. Primary dissolved gases are oxygen, carbon dioxide, and nitrogen. Dissolved inorganic ions and compounds include major and minor ion groups and trace elements, which occur in minute quantities, such as copper, zinc, iron, and aluminum. Minor ions such as nitrogen and phosphorus are nutrients essential to plant and animal growth. Major ion groups include cations of calcium, magnesium, sodium, and potassium and the anions bicarbonate, sulfate, and chloride.

Concentrations of dissolved gases and major ions affect pH, which measures hydrogen ion activity. The pH affects many cellular processes of organisms and determines the solubility and biological availability of nutrients and heavy metals. Hardness measures calcium and magnesium concentrations, and is associated with but not identical to alkalinity, which measures the ability of streamwater to absorb hydrogen ions, thus buffering changes in pH. Alkalinity is primarily due to bicarbonate and carbonate ions.

Total dissolved solids, the sum of the concentrations of major cations and anions, is often estimated as specific conductance. Hardness, alkalinity, and ionic concentrations are often positively correlated with stream productivity and taxonomic richness. Particulate inorganic and organic materials together make up the suspended load in lotic systems, and contribute to turbidity.

Carbon dioxide and oxygen are the most biologically important dissolved gases. Diffusion from the atmosphere maintains concentrations of both oxygen (O_2) and carbon dioxide (CO_2) in streams at close to equilibrium. However, CO_2 is more soluble in water than is O_2, which is 30 times less available in water than air. Groundwater and sites of organic matter decomposition are low in O_2 and enriched in CO_2. Photosynthesis and respiration can alter diel concentrations of oxygen and carbon dioxide in productive systems, with O_2 elevated and CO_2 reduced during the day, and the reverse occurring at night. Where production is high relative to diffusion, diel changes in O_2 concentration can be used to estimate photosynthesis and respiration. Because current and turbulence continually renew O_2 supply, O_2 concentrations are usually problematic for stream organisms only in sites severely contaminated with organic pollutants, or in conditions of high temperatures, drought, and dense populations of aquatic plants. Low-oxygen concentrations are better tolerated by stream animals at faster than slower current speeds.

Natural spatial variability in lotic chemistry largely reflects differences in the types of rocks available for weathering and the amount, chemical composition, and distribution of precipitation. For example, total dissolved solids are approximately twice as great in rivers draining sedimentary terrain compared to those draining terrain of igneous and metamorphic rock. Most rivers contain 0.01%–0.02% dissolved minerals, about 1/20–1/40th the salt concentration of the oceans, with an average concentration of 100 mg L^{-1}. Generally $\geq$50% of this is bicarbonate and 10%–30% is chloride and sulfate. River water contains more dissolved solids than does rainwater, because of evaporation, weathering, and anthropogenic inputs. Rainwater, although nearly pure, contains dissolved minerals from dust particles and droplets of ocean spray.

Rainwater is also naturally acidic due to atmospheric carbon dioxide dissolving in water droplets, forming a weak carbonic acid (H_2CO_3). In catchments with erosion-resistant rock, little buffering capacity, or where decaying plant matter is abundant, stream-water can be acidic even in the absence of pollution. Water percolating through the soil enters the stream enriched with CO_2 from plant and microbial respiration, and forms carbonic acid. The carbonic acid dissolves the calcium carbonate in rocks, producing calcium bicarbonate, which is soluble in water and the source of carbon atoms for aquatic photosynthesis. Dissolution of calcium carbonate increases the amount of stream calcium and bicarbonate ions and the latter dissociates to carbonate ions. At equilibrium, bicarbonate and carbonate ions dissociate, forming hydroxyl ions and resulting in weak alkaline waters, with a pH >7. At equilibrium, water resists changes in pH because the addition of hydrogen ions is neutralized by the hydroxyl ions formed by dissociation of bicarbonate and carbonate, and added hydroxyl ions react with bicarbonate to form carbonate and water. Thus the buffering capacity of a stream is largely determined by its calcium bicarbonate content. The pH of most natural running waters ranges between 6.5 and 8.5, with values below 5 or above 9 being harmful to most stream organisms. Industrially derived sulfuric and nitric acids have seriously lowered pH in surface waters of large areas of Europe and North America, resulting in reduced species diversity and density.

The Adapted Biota

Many taxonomic groups inhabit running waters. Key biological attributes, life histories, and distribution patterns of major groupings of organisms that play a central role in energy flux within trophic webs or that are of significant human interest—namely algae, macrophytes, benthic macroinvertebrates, and fishes—are summarized below.

Algae

Algae are the most important primary producers in running-water ecosystems and are drivers of biogeochemical cycling. Because of their sessile nature and short life cycles, algal assemblages are often used to evaluate stream ecosystem health and are particularly precise indicators of ecological change from nutrient contamination and agricultural land use. Algae lack vascular tissue, bear chlorophyll *a* and lack multicellular gametangia. Algal evolution has radiated to include several diverse kingdoms. For example, blue-green algae are classified as bacteria, and dinoflagellate algae as protozoans. Algal taxonomy is based on pigmentation, the chemistry and structure of internal storage products and cell walls, and the number and type of flagella. Divisions of algae common in streams include the Bacillariophyta (diatoms), Chlorophyta (green algae), Cyanophyta or Cyanobacteria (blue-green algae), Chrysophyta (yellow-green algae), and Rhodophyta (red algae). Of these, the diatoms, green algae, and cyanobacteria are most prevalent. Assemblages of benthic algae attached to submerged substrates are dominant members of the periphyton (also referred to as biofilm or aufwuchs), which represents a complex assemblage of algae, detritus, bacteria, fungi, and meiofauna held together in a polysaccharide matrix. Free-floating algae within the water column, the phytoplankton, occur chiefly in slowly moving lowland rivers as sloughed benthic cells or as exports from connected standing waters.

Diatoms are extremely abundant in freshwater, and typically contribute the majority of species within the periphyton. Generally microscopic, diatoms are brownish-colored single-celled algae constructed of two overlapping siliceous cell walls, or valves, fit together like the halves of a petri dish to form the frustule. Valves are connected to each other by one or more "girdle" bands. Diatoms are largely classified to genus on the basis of unique cell ornamentation, which may include pores (punctae), lines (striae), or ribs (costae). The symmetry of these decorations defines two groups: radially symmetrical centric diatoms and bilaterally symmetrical pennate diatoms. Diatoms may occur individually, in chains, or in colonies, and those with a divided cell wall

(raphe) are able to move. In temperate streams, diatoms commonly exhibit two blooms: in spring prior to shading by deciduous canopies as water temperatures rise and nutrients are plentiful; and in fall following leaf abscission, when nutrients released from decaying green algae and deciduous litter are available. Diatoms constitute a high-quality, rapid-turnover food resource for many stream invertebrates. Representative diatoms common in stream periphyton are shown in Fig. 8.

Green algae can be unicellular or multicellular, occur in a variety of habitats, and are taxonomically distinguished by the number and arrangement of flagella, their method of cell division, and their habitat. Green algal forms may be microscopic or readily visible without a microscope as a thallus (undifferentiated body tissue) or as colonies of filaments. Filamentous forms may be branched or unbranched. Green algae provide attachment sites for diatoms, and are a source of FPOM and photosynthetic oxygen, but are not a major source of food for most herbivorous invertebrates.

Blue-green algae, or cyanobacteria, are prokaryotic organisms of ancient lineage which contain the photosynthetic pigment phycocyanin, used to capture light for photosynthesis. They occur in a variety of habitats and are one of very few groups of organisms that can convert inert atmospheric nitrogen into an organic form. Blue-green algae may be filamentous or nonfilamentous, and only filamentous forms with heterocysts are capable of nitrogen fixation in aerobic settings. Several of the heterocyst-containing filamentous taxa (e.g., *Anabaena*, *Aphanizomenon*, and *Microcystis*) can form dense blooms and produce toxins in warm, nutrient-rich waters. Nitrogen-fixing *Nostoc*, common in small streams, forms a unique commensal association with the chironomid midge *Cricotopus*.

Macrophytes

Macrophytes include vascular flowering plants, mosses and liverworts, some encrusting lichens, and a few large algal forms such as the Charales and the filamentous green alga *Cladophora*. Light and current are among the most important factors limiting the occurrence of macrophytes in running waters. Macrophytes can be categorized into those that are attached to the substrate, those that are rooted into the substrate, and free-floating plants. Attached plants include the mosses and liverworts, certain lichens, and some flowering plants of the tropics. The attached macrophytes are predominantly found in cool, headwater streams. Mosses are unusual in their requirement for free CO_2, rather than bicarbonate, as their carbon source. In shaded, turbulent streams, their contribution to primary production may override that of periphyton. Mosses also support very high densities of macroinvertebrates. Rooted plants include submerged (e.g., Hydrocharitaceae, Ceratophyllaceae, and Halorgidaceae) and emergent (e.g., Potamogetonaceae, Ranunculaceae, and Cruciferae) forms. They usually require slow currents, moderate depth, low turbidity, and fine sediments for rooting. Rooted plants are most common in mid-sized rivers and along the margins of larger rivers where they reduce current velocity, increase sedimentation, and provide substrate for epiphytic microflora. They also provide valuable habitat for some animals. Tough, flexible stems and leaves, attachment by adventitious roots, rhizomes or stolons, and vegetative reproduction are among the adaptations of macrophytes to running water. Free-floating plants (e.g., Lemnaceae and Pontederiaceae) are of minor importance in running waters at temperate latitudes as they depend largely on standing-water conditions. They may accumulate significant biomass in subtropical and tropical settings. Macrophytes in lotic ecosystems contribute to energy flow predominantly through decomposer food chains, as few macroinvertebrates feed on the living plants.

Benthic Macroinvertebrates

Three invertebrate phyla are common in running waters: Annelida (segmented worms), Mollusca (snails, clams, and mussels), and Arthropoda (crustaceans and insects). Both annelids and molluscs are of marine evolutionary origin. While often abundant in fine and coarse sediments in rivers, aquatic annelids are challenging to identify, and information about their distribution in river systems is poorly understood. Molluscs are most abundant and diverse in larger rivers. Arthropoda dominate the headwaters, but are abundant all along drainage networks. Representative taxa are illustrated in Fig. 9.

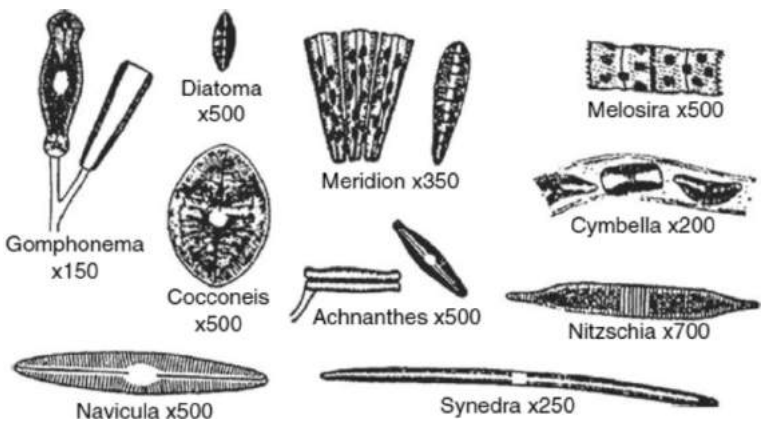


Fig. 8 Representative diatoms common in stream periphyton. From Hynes H. B. N. (1970) *The ecology of running waters.* Liverpool: Liverpool University Press.

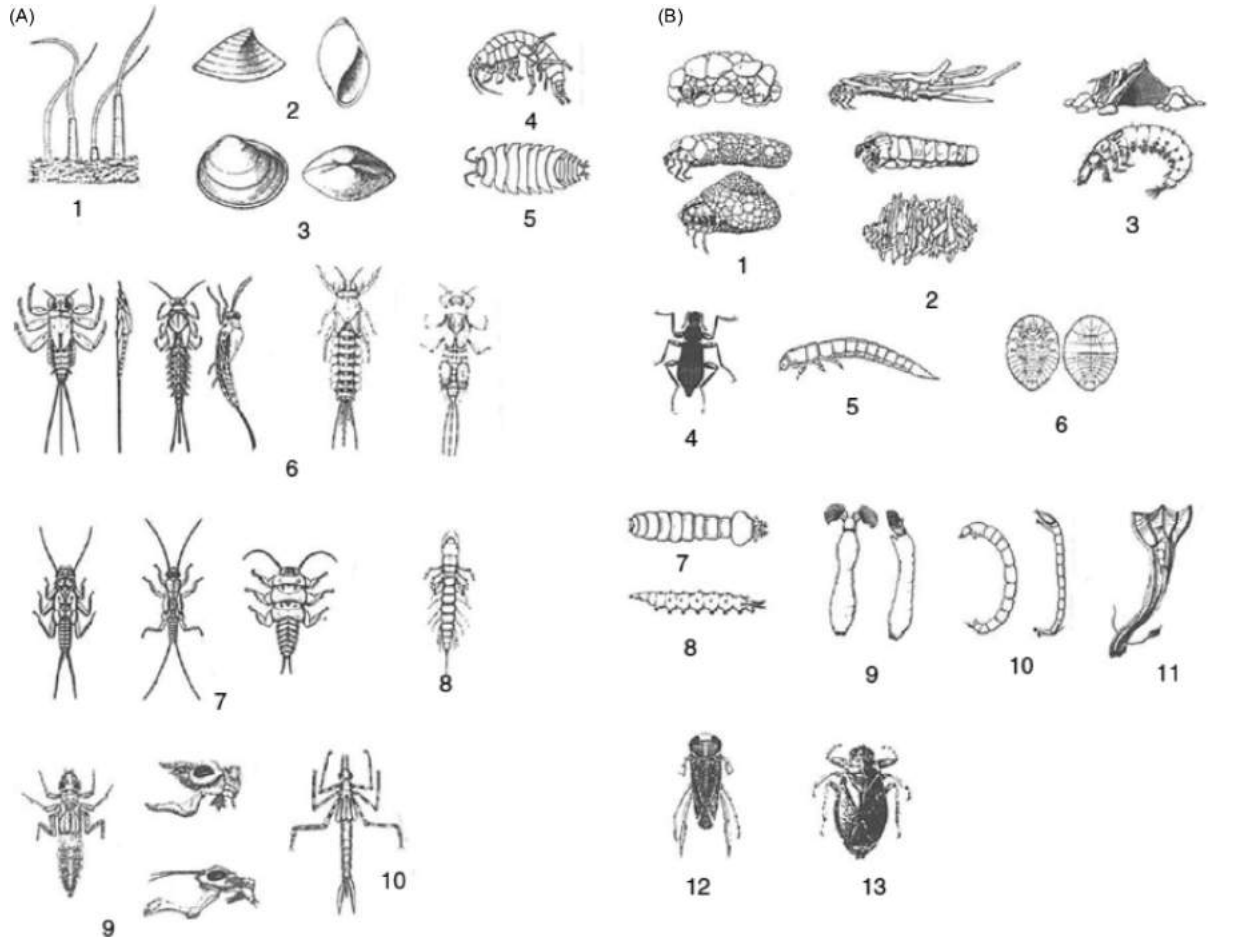


Fig. 9 (A) Examples of lotic benthic invertebrates. 1, Annelida, Oligochaeta (Tubificidae); 2, Mollusca, Gastropoda (*left*, Ancylidae; *right*, Physidae); 3, Mollusca, Bivalvia (Spaeridae: *left*, lateral view*; right*, dorsal view); 4, Crustacea, Amphipoda; 5, Crustacea, Isopoda; 6, Insecta, Ephemeroptera; 7, Insecta, Plecoptera; 8, Insecta, Megaloptera (Sialidae); 9, Insecta, Odonata, Anisoptera (*left*, nymph; *right upper*, lateral view of head with extended labium; 10, Insecta, Odonata, Zygoptera (*right*, nymph; *left lower*, lateral view of head with extended labium). (B) Examples of lotic benthic invertebrates. 1, Insecta, Tricoptera (mineral case bearers); 2, Insecta, Trichoptera (organic case bearers); 3, Insecta, Trichoptera (net spinner, fixed retreat above); 4, Insecta, Coleoptera (Elmidae adult); 5, Insecta, Coleoptera (Elmidae larvae); 6, Insecta, Coleoptera, Psphenidae larvae (*left*, ventral; *right*, dorsal); 7, Insecta, Diptera, Tipulidae; 8, Insecta, Diptera, Athericidae; 9, Insecta, Diptera, Simuliidae (*left*, dorsal; *right*, lateral view); 10, Insecta, Diptera, Chironomidae (*left*, Chironominae; *right*, Tanypodinae); 11, Insecta, Diptera, Chironomidae (filtering tube of *Rheotanytarsusi*); 12, Insecta, Hemiptera, Corixidae; 13, Insecta, Hemiptera, Belastomatidae.

The Oligochaeta are the most abundant and diverse of the freshwater annelids, and are notable for their ability to inhabit low-oxygen environments. Oligochaetes inhabit the sediments, some in tubes, and are almost all gathering-collector detritivores. Their bodies are segmented with two pairs of stout, lateral chetae on each segment. Leeches (Hirudinea), a minor annelid group occurring in small streams to mid-sized rivers, feed as gathering collectors, predators, or ectoparasites.

A calcium requirement for shell formation limits the distribution of gastropod (limpets and snails) and bivalve (clams and mussels) molluscs in streams and rivers. Limpets, such as *Ferrissia* (Ancylidae), frequent small, fast-flowing streams where their hydrodynamic shape and sucker formed by the mantle allow them to move over rocks in the current and scrape loose attached algal food with a rasping radula. Snails, such as *Physa*, are abundant scrapers in river macrophyte beds that use radulae to rasp vascular plant surfaces, removing periphyton and epidermal plant tissue. Clams and mussels are filtering collectors that burrow in the sediments with their incurrent and excurrent siphons exposed, pumping water in to extract dissolved oxygen and FPOM, and out to eliminate wastes. Bivalve mollusks are sensitive to water quality and have been used worldwide as indicators of lotic ecosystem health. Mollusks are one of the most imperiled groups of animals worldwide: of 693 recorded extinctions of animal species since 1500, 42% have been mollusks, almost entirely nonmarine forms. In the United States and Canada, 202 of 300 freshwater unionoid clams are presumed extinct or critically imperiled.

Common Crustacea of running waters include Amphipoda (scuds), Isopoda (aquatic pill bugs), benthic Copepoda (Harpacticoida), and Decapoda (crayfish and freshwater shrimps). Most isopods and amphipods (except *Hyallela*) are detrital shredders feeding on stream-conditioned riparian litter in headwater streams. Although decapod shrimps and crayfish have species found in all sizes of running waters, the former tend to be more abundant in streams, the latter in mid-sized rivers. Decapods are scavengers and facultative shredders of plant litter. These crustaceans have always been of interest because of their large size,

commercial food and bait value, and importance as food for large game fish. The minute harpacticoid copepods are poorly known, but are often found in small streams to large rivers where they feed on fine particulate organic matter.

Aquatic insects (Arthropoda) are the most conspicuous and best-studied invertebrates of running waters. They can be subdivided into the more primitive hemimetabolous orders, in which immature nymphs gradually metamorphose into mature winged adults, and the more evolved holometabolus orders that have larval and pupal stages. Insect growth is accomplished by the nymphs or larvae over weeks to years, while the adults feed little and are short lived (a day to weeks). Thirteen orders of aquatic or semiaquatic (occurring at lotic margins) taxa have been distinguished. The hemimetabolous mayflies (Ephemeroptera), stoneflies (Plecoptera), dragonflies and damselflies (Odonata), and the holometabolous caddisflies (Trichoptera), and dobsonflies and alderflies (Megaloptera) all have aquatic larvae. These signature taxa are represented in almost all unpolluted lotic ecosystems. Mayflies, which are the only insects that molt as winged subadults (subimagos) to sexually mature adults (imagos), are commonly imitated in the sport of flyfishing. All of the odonate and about half of the plecopteran nymphs are predaceous. The dragonflies and damselflies occur in small streams to large rivers, with many species associated with aquatic vascular plants. Nonpredaceous stonefly nymphs feed by shredding conditioned riparian litter. Some of the predaceous Megaloptera are among the largest of the lotic aquatic insects, and they are typical of slow-flowing areas and often associated with submerged woody debris.

Caddisflies are a large aquatic order in which a majority of species construct portable cases made of plant pieces (characteristic of shredders) or mineral particles (characteristic of scrapers) held together with silk extruded from glands in the head. All the cases are lined with silk into which hooks on the hind prolegs are hooked to maintain the larvae in the case. Larvae circulate water through the case by undulating the abdomen to irrigate the gills and integument and to facilitate respiration. Five families of Trichoptera larvae, and all families in the pupal stage, construct nonportable, fixed retreats of organic and mineral material. Most larvae of the five families spin silk nets with which they filter out FPOM food from flowing water. Species of the family Rhyacophilidae are free-ranging without cases and almost exclusively predaceous.

The holometabolous Coleoptera (beetles), Diptera (true flies), Lepidoptera (aquatic moths), and Hymenoptera (aquatic wasps) constitute the largest insect orders and have some aquatic or semiaquatic representatives, as do the spongeflies of the Neuroptera. The beetles are the only aquatic insects with representatives in which both the larvae and adults live in the water. One family of Diptera, the midges (Chironomidae), is usually more abundant and diverse in running waters than all other aquatic insects combined. Chironomid species are found in all lotic habitats and all functional feeding group categories. Difficulty in identifying Chironomid larvae has hampered their use in ecological studies. Very few aquatic moths occur in running waters. Hymenoptera, a large terrestrial order containing many social species, has some parasitic forms in which the females enter the water to oviposit in the immatures of aquatic and semiaquatic orders. The larvae of spongeflies inhabit freshwater sponges where they feed directly on sponge tissue or prey upon other animals.

The hemimetabolous Hemiptera (true bugs), Orthoptera (grasshoppers, etc.), and Collembola (springtails) have aquatic or semiaquatic species. All widely distributed hemipterans are active predators, occupying a full range of slow water and marginal habitats where they capture prey and imbibe their body fluids using piercing mouth parts. Orthoptera and Collembola of running waters are all semiaquatic and function as detritus-gathering collectors. We further describe functional feeding roles in Rivers and Streams: Ecosystem Dynamics and Integrating Paradigms.

Fishes

Fishes, the principal group of vertebrates found in running waters, are of great human interest because of their commercial and recreational value. Approximately 46% (about 15,000 species) of the estimated total number of the world's fishes live in freshwater. Almost all freshwater fish taxa have representatives that occur in running waters, with varying degrees of river dependency and saltwater tolerance. Groups with little or no tolerance for saltwater (e.g., Cyprinidae, Centrarchidae, and Characidae) are classified as primary freshwater fishes; these have dispersed through freshwater routes or evolved in place from distant marine ancestors. Secondary freshwater fishes (e.g., Cichlidae and Poeciliidae) have some tolerance to seawater but are usually restricted to freshwater. Diadromous fishes migrate between freshwater and saltwater. Anadromous fishes, including many salmonids, lampreys, shad, and sturgeon, spend most of their lives in the sea and migrate to freshwater to reproduce. American and European eels are catadromous fishes, which spend most of their lives in freshwater and migrate to the sea to reproduce. Catadromy appears to be more prevalent in the tropics, with anadromy more common at higher latitudes.

Researchers have commonly observed longitudinal gradients of fish assemblages in river systems, resulting in numerous attempts to classify stream zones by the dominant fish species or assemblage found. Because fish faunas vary considerably among geographic and climatic regions, zonation schemes usually have only regional application. Longitudinal gradients arise as the result of species addition and/or replacement, and reflect adaptations to type and volume of habitat and available food along the river continuum. Fishes occupying headwater reaches, typified by salmonids and sculpins, have high metabolic rates and demands for oxygen. Salmonids are active, streamlined fishes with strong powers of locomotion, which maintain position in swift water to feed upon drifting invertebrates. Sculpins, with depressed heads and large pectoral fins, hold close to the streambed and forage for invertebrates among stones on the bottom. Upstream fishes are usually solitary in habit and may exhibit territoriality associated with breeding, food, or spatial resources. Their distribution extends downstream where oxygen and temperatures are suitable, to join deeper-bodied fishes tolerant of warmer temperatures and reduced oxygen. Species richness is usually greatest in the mid-order segments, in association with overall habitat heterogeneity and increased pool development. In high-order reaches, fish assemblages include larger, deep-bodied fishes such as suckers and catfishes feeding on bottom deposits, planktivorous fishes, and fishes that feed on large molluscs.

Further Reading

Allan JD and Castillo MM (2007) *Stream ecology structure and function*. Netherlands: Springer.
Dudgeon D, Arthington AH, Gessner MO, et al. (2006) Freshwater biodiversity: Importance, threats, status and conservation challenges. *Biological Reviews* 81: 163–182.
Giller PS and Malmqvist B (1998) *The biology of streams and rivers*. Oxford: Oxford University Press.
Hauer FR and Lamberti GA (eds.) (2017) *Methods in stream ecology: Volume 1. Ecosystem structure*. Amsterdam: Elsevier Publishing Company.
Hynes HBN (1970) *The ecology of running waters*. Liverpool: Liverpool University Press.
Knighton D (1998) *Fluvial forms and processes: A new perspective*. London: Arnold Publishers.
Leopold LB (1994) *A view of the river*. Cambridge, MA: Harvard University Press.
Melles SJ, Jones NE, and Schmidt B (2012) Review of theoretical developments in stream ecology and their influence on stream classification and conservation planning. *Freshwater Biology* 57: 415–434.
Merritt RW, Cummins KW, and Berg MB (eds.) (2008) *An introduction to the aquatic insects of North America*, 4th edn. Dubuque, IA: Kendall/Hunt Publishing Company.
Rosgen D (1996) *Applied river morphology*. Pagosa Springs, CO: Wildland Hydrology.
Stevenson RJ, Bothwell ML, and Lowe RL (1996) *Algal ecology: Freshwater benthic ecosystems*. London: Academic Press.
Thorpe J and Rogers DC (eds.) (2014) *Thorp and Covich's freshwater invertebrates: Ecology and general biology*, 4th edn. London: Academic Press.

Rocky Intertidal Zone

PS Petraitis and JAD Fisher, University of Pennsylvania, Philadelphia, PA, USA
S Dudgeon, California State University, Northridge, CA, USA

Introduction

The British ecologist A. J. Southward described the intertidal zone as ``the region of the shore between the highest level washed by the waves and the lowest level uncovered by the tide,'' and thus communities on rocky intertidal shores are primarily defined by the tides and the presence of hard surfaces. The types of organisms, the number of species, and the distribution and abundance of individual species found in a particular rocky intertidal community also depend on the physical aspects of the shore, the supply of resources, food and larvae from overlying water, the biological interactions among the species present, and the regional pool of species. Although rocky intertidal shores cover only a small fraction of the Earth's surface, they contain a large diversity of organisms – ranging from highly productive microalgae to transient vertebrate predators (Fig. 1).

Physical Aspects of the Shore

Tides

Tides are caused by the gravitational effects of the Moon and Sun, which ideally produce a cycle of two high tides and two low tides per day. However, the amplitude and frequency of the tides are altered by the phases of the Moon, the Earth's orbit and declination, latitude, and the configurations of the shoreline and the seafloor. The tidal range tends to be smaller toward the equator and can vary from several meters in high latitudes to less than tens of centimeters near the equator. Configuration of the coast and the ocean basin can cause harmonic resonances and create tides that vary dramatically in amplitude and frequency. In extreme cases, the reinforcing and canceling effects can produce a single high and low tide per day or almost no change over the course of a day.

The timing of low tides can have a profound effect by exposing organisms to extreme conditions. For example, the lowest tides in the Gulf of Maine, USA tend to occur near dusk or dawn, and so organisms are rarely exposed to mid-day sun in the summer but are often exposed to below freezing temperatures on winter mornings. In contrast, the lowest summer tides in southeastern Australia occur mid-day and expose organisms to extraordinarily high temperatures.

Characteristics of the Shore

Any firm stable surface in the intertidal zone has the potential to support the organisms that commonly occur in rocky intertidal communities, and at low tide, intertidal habitats can range from dry rock to filled tide pools. Rock surfaces can vary from very hard to relatively soft rock such as from granite to sandstone and can range from smooth platforms to irregular fields of stone cobbles and boulders. Topography, inclination, color, and texture of the rock affect rate of drying and surface temperature, which can limit the distribution and abundance of species. Man-made surfaces such as rock jetties and wooden pier pilings and biogenic surfaces such as mangrove roots can also support communities that are indistinguishable from the communities found on nearby rocky shores.

Tide pools can be very different than the surrounding shore because of thermal variability, changes in salinity from evaporation and runoff, and changes in pH, nutrients, and oxygen levels caused by algae. Pools often support residents such as sea urchins, snails, and fish that would otherwise be restricted to subtidal areas.

The amount of wave surge affects the types of organisms found on the shore and their distribution. Wave surge and breaking waves tend to expand the extent of the intertidal zone and distribution of species by continually wetting the shore and allowing species to extend farther up the shore. Wave surge can also cause mobile animals to seek refuge and can limit the distribution of slow moving species, and the force of breaking waves can damage and sweep away organisms. Sand and debris such as logs swept up by the waves can scour organisms off the surface. In areas of low wave surge, sedimentation of sand and silt may bury organisms or clog gills and other filter-feeding structures.

Attached Organisms

Unlike terrestrial habitats, which depend largely on local plant material to support resident animal populations, rocky intertidal assemblages are supported not only by algal primary production but also by secondary production from suspension feeders, such as barnacles and mussels, which link the ocean's productivity to the shore.

Fig. 1 Closeup of predatory snails, mussels, barnacles, and brown algae in Maine, USA. Photo by P. S. Petraitis.

Algae

The term `algae' refers to an extraordinarily diverse and heterogeneous group comprising about seven major lineages, or roughly 41% of the kingdom-level branches in the Eukarya domain. Most lineages consist of unicellular microalgae, but the multicellular macroalgae that dominate many rocky shores worldwide occur in only three groups (Rhodophyta, Chlorophyta, and Phaeophyta) (Fig. 2).

Microalgae are ubiquitous and although inconspicuous, they are important members of rocky intertidal communities. For example, diatoms are the primary food source of many grazing gastropods and form biofilms, which facilitate settlement of invertebrate larvae and stabilize meiofaunal assemblages.

Benthic macroalgae (i.e., seaweeds) dominate many rocky shores, especially the low- and mid-intertidal zones of temperate regions, and many exhibit morphologies adaptive for life on wave swept shores. The idealized body plan of a seaweed consists of a holdfast, a stipe, and one or more blades. The holdfast usually attaches the alga either by thin encrusting layers of cells tightly appressed to the rock surface or by a massive, thick proliferation of tissue that often produce mucilaginous `glues' to adhere the tissue to the rock. The stipes are analogous to plant stems and display remarkable material properties that enable seaweeds to withstand the tremendous hydrodynamic forces imposed by breaking waves. The blade is the principal structure for the exchange of gases and nutrients, and the capture of light for photosynthesis. Blades also contain reproductive tissue, either within a vegetative blade, or in sporophylls (i.e., special blades for reproduction). Some larger brown seaweeds, such as fucoids and kelps, have gas-filled floats called pneumatocysts that buoy the blade so that it remains closer to the surface where light intensity is greater.

The diversity and complexity of the life cycles of most seaweeds contributes to their great abundance on rocky shores. The life cycle of most seaweeds consists of an alternation of separate gametophyte and sporophyte generations. The two generations can either look the same (i.e., isomorphic) or different (heteromorphic). In some species, the heteromorphic generations are so different that they were originally described as different species. Heteromorphic life histories are hypothesized to represent an adaptation to grazing pressure, and heteromorphic generations clearly show tradeoffs with respect to competitive ability, resistance to disturbance and longevity associated with upright foliose and flat encrusting morphologies.

Sessile Invertebrates

Adults of many invertebrate species are attached permanently to the rock or other organisms (epibiota). These include members of the phyla Porifera (sponges), Cnidaria (hydroids and sea anemones), Annelida (tube-building polychates), Arthropoda (barnacles), Mollusca (mussels and clams), Bryozoa (moss animals), and Chordata (tunicates). Suspension feeding – either by pumping water through a sieve structure or trapping particles carried on induced or external currents – is a common feature of sessile animals and serves to transfer inputs of energy and nutrients produced in the water column into the intertidal zone via the ingestion of plankton. Additionally, by feeding on locally derived detritus, suspension feeders capture some of the nutrients that are produced by neighboring inhabitants.

Sessile intertidal animals are often physically or chemically defended against predation and display plastic phenotypes in response to changing environmental conditions because they are fixed in place and cannot move to avoid predators. For example, the presence of the predatory gastropod *Acanthina angelica* induces change in the shell shape of its barnacle prey *Chthamalus anisopoma*, and the barnacle forms a curved shell making it more difficult for the predator to attack.

Fig. 2 Extensive brown algal beds in Maine, USA. Photo by P. S. Petraitis.

Mobile Organisms

Mobile invertebrates and vertebrates that are found on rocky intertidal shores are typically divided into two categories based on the amount of time spent between tidemarks. Resident species remain in the intertidal zone throughout most of their life and face a large range of local physical conditions that they mitigate by a variety of behavioral and physiological adaptations. Many residents find shelter during low tides, either between rocks, under algae, or in tide pools, while other species attach to exposed rock surfaces just ahead of the incoming tide. Transient species are those that spend only a small part of their life cycles in the intertidal zone (e.g., as juveniles) or are those that enter and leave the intertidal zone during low or high tide.

Invertebrates

Large, mobile invertebrate consumers are ecologically the most intensively studied guild on rocky shores and include species from Turbellaria (flatworms), Crustacea (e.g., crabs, shrimp, amphipods, and isopods), Annelida (e.g., polychaetes), Gastropoda (e.g., snails, nudibranchs, and chitons), and Echinodermata (sea urchins, brittle stars, and sea stars). Herbivores range from grazers of diatom films to browsers of macroalgae, and predators exploit a variety of methods (crushing, stinging, drilling, and partial consumption) to overcome the defenses of their prey.

Small mobile metazoans (roughly 0.1–1 mm and collectively termed meiofauna) thrive on and among the algae, animals, and the trapped sediments on rocky shores. Meiofauna include consumers from many invertebrate phyla, that – due to their small sizes, extremely high abundances, and high turnover rates – are an important guild of consumers whose effects have largely been neglected in comparison to studies of larger invertebrates.

Vertebrates

Vertebrates tend to be transient species that use the intertidal zone to feed or hide and include fish and marine mammals that enter at high tide and birds and terrestrial mammals that enter at low tide (Fig. 3). For instance, marine iguanas (*Amblyrhynchus cristatus*) of the Galápagos Islands, Ecuador forage extensively on intertidal algae on lava reefs during low tides. The major exceptions are resident intertidal fishes, which are often cryptic and less than 10 cm in length. Resident and transient fishes include hundreds of species from dozens of families, though members of the families Blenniidae, Gobiidae, and Labridae are the most common.

Birds and mammals, characterized by high endothermic metabolic rates and large body sizes, have significant impacts on intertidal communities even at low densities. Birds include locally nesting and migratory species and can remove millions of invertebrates during a season. In addition, birds in some communities provide major inputs of nutrients via guano and prey remains. More than two dozen terrestrial mammals, mostly carnivores, rodents, and artiodactyls, have been reported as consumers or scavengers of rocky intertidal organisms on every continent except Antarctica. Most recorded prey species are mollusks, crabs, or fish. Probably one of the most unusual cases is a population of feral rabbits on a small island off the coast of South Africa that forage on seaweeds in the intertidal zone. Given the mobility of vertebrates, their impact on rocky intertidal shores has been difficult to assess and intertidal activity is often discovered by finding exclusively intertidal animals or algae in the gut contents of otherwise pelagic or terrestrial species.

Little is known about the effects of harvesting by humans in the rocky intertidal zone. Results from a few large-scale studies in Australia, Chile, and South Africa, however, have demonstrated that harvesting has had significant effects on intertidal assemblages.

Fig. 3 Rocky shore in Central California, USA with elephant seals on the beach. Photo by S. Dudgeon.

Zonation

Patterns

Rocky intertidal shores often display a vertical zonation of fauna and flora associated with the strong environmental gradient produced by the rise and fall of the tides. For example, most moderately exposed rocky shores of the northern hemisphere have kelps at the littoral sublittoral interface, followed by rhodophyte algae dominating the low intertidal zone, by fucoid algae, mussels, and barnacles dominating the mid-intertidal zone, and by cyanobacteria, lichens, and a variety of small tufted, encrusting, or filamentous ephemeral seaweeds occurring in the high intertidal zone. While species from many phyla may be found together, often a single species or group is so common; vertical zones are named according to the dominant group (e.g. the intertidal balanoid zone named after barnacles in the family Balanidae).

Combinations of various physical factors acting upon different inhabitants in intertidal zones that vary in their exposure to waves can lead to complex patterns of distribution and abundance along shorelines in a particular region. Nevertheless, some general patterns are evident at a regional scale. Geographically, vertical zonation patterns are most pronounced on temperate rocky shores where species diversity is high and tidal amplitudes tend to be greatest. On rocky shores in the tropics, biotic zones are compressed into narrow vertical bands because of small tidal amplitudes. In polar regions, annual ice scour and low species diversity tend to obscure any conspicuous vertical zonation.

Causes

It is often stated that the upper limits of organisms are set by physical factors, whereas the lower limits are set by biological interactions but there are many exceptions to this rule. The specific causes of the zonation seen on most rocky shorelines vary with geographic location, but zonation results primarily from behavior of larvae and adults, tolerance to physiological stress, the effects of consumers, and the interplay between production and the presence of neighbors.

Adult movements and larval behavior during settlement from the plankton onto rocky shores have major effects on the distribution of animals. For example, studies of barnacles have shown that vertical zonation of larvae in the water column contributes to corresponding vertical zonations of both larval settlement and adults on the shore, a pattern previously ascribed solely to interspecific competition. For seaweeds, behavior is a relatively unimportant cause of their zonation since adult seaweeds are sessile and settling spores are mostly passively transported.

Marine organisms living higher on the shore are faced with more frequent and extreme physiological challenges than their lower shore counterparts, and the upper limits of intertidal distributions for most species are set by cellular dehydration. Dehydration can occur either from freezing during winter or simply desiccation associated with long emersion times. High temperatures and wind, which accelerate the rate of water loss from tissues, exacerbate the effects of desiccation.

Primary and secondary production by sessile organisms can be limited at higher tidal elevations because nutrients and other resources can be acquired only when immersed. Respiration rates of seaweeds and invertebrates are temperature dependent and thus can be greater when an organism is exposed at low tide. For seaweeds, prolonged exposure to dehydration also reduces photosynthesis.

The reduced productivity associated with increased exposure at higher tidal elevations modifies intra- and interspecific interactions. For instance, competition between seaweeds, which may be intense lower on the shore, is reduced at higher tidal elevations and enables coexistence. Competition among intertidal seaweeds is hierarchical with lower shore species dominating those of the higher shore. Thus, fucoid species of the mid intertidal zone are outcompeted for space in the low zone by foliose red seaweeds that pre-empt space with an encrusting perennial holdfast. There is also a competitive hierarchy among mid intertidal

zone fucoids with those typically occurring lower on the shore competitively dominant to those higher up. This is most apparent on European rocky shores where the diversity of intertidal fucoids is greatest.

Grazing rates tend to be greater lower on the shore, although there are cases of herbivory by insects setting the upper limits of ephemeral green algae. Grazing by sea urchins at the interface with the sub-littoral zone can limit the lower distributions of macroalgae, but there is little evidence for grazing on perennial seaweeds setting the lower limits of those taxa within the intertidal zone. Grazing of perennial seaweeds is most intense at the sporeling stage soon after settlement. Grazing by gastropods and small crustaceans certainly contributes to losses of biomass of established individuals, but does not affect distributions within the intertidal zone. In contrast, the grazing of established ephemeral species both on emergent rock and tidepools is intense during spring and summer in many regions eventually eliminating those algae from their respective habitats. There are also many examples of consumers using seaweeds as habitat as well as food.

Rocky Intertidal Shores as an Important System in Development of Ecology

The rocky intertidal zone has been a stronghold for ecological research, and the success of intertidal experiments stems in part from the fact that intertidal assemblages are often comprised of the few species that are able to survive the environmental variation associated with the cycling of tides. In addition, many resident intertidal species are small, common, and slow moving or fixed in one place. Thus rocky intertidal shores historically appeared as simple, well-defined habitats in which easily observed and manipulated local interactions control the dynamics of the assemblages. Such initial appearances, however, have been deceiving, and variation in recruitment of offspring from the plankton, a characteristic of many marine species, has stimulated an increased appreciation of the role of oceanographic conditions.

Descriptive Studies: Research Prior to 1960

Descriptions of rocky shores and speculation about the causes of vertical zonation go back more than 195 years. Before the 1960s, ecologists had published descriptions of intertidal areas from more than a dozen large geographical regions that spanned much of the globe and included both sides of the North Pacific and North Atlantic; Greenland; the West Indies; South and Central America; the coasts of Africa; the Mediterranean; the Black Sea; Indian Ocean Islands; Singapore; Pacific Islands, Australia, and Tazmania. These early accounts of the rocky intertidal remain a potentially valuable source for comparison to contemporary patterns of species distributions due to local species extinctions and introductions.

The Rise of Experimental Studies: 1960–80

Direct experimental manipulation of intertidal organisms accelerated in the 1960s with the groundbreaking work of J. H. Connell and R. T. Paine. Connell manipulated the presence of two species of barnacles in Scotland by selectively removing individuals from small tiles fashioned from the sandstone rock from the shore. He showed that the lower limit of the high intertidal species *Chthamalus stellatus* was set by competition with the mid zone species *Balanus* (now *Semibalanus*) *balanoides* and that the upper limit of *S. balanoides* was set by physical factors. Paine removed the predatory seastar *Pisaster ochraceus* from an area of the intertidal shore in Washington and showed that *Pisaster* was responsible for controlling mussels, which are successful competitors for space and dominate the intertidal shore in the absence of *Pisaster*. These early investigations provided a framework for the rapid growth of experimental studies that characterized the field in recent decades (Fig. 4).

Fig. 4 Grindstone Neck in Maine, USA with Mount Desert Island in the background. This site was used by Menge and Lubchenco in their groundbreaking work in the 1970s. Photo by P. S. Petraitis.

In general, the observation and experimental manipulations of mobile consumers and their prey has often revealed predation by mobile consumers as an important factor that contributes to the structure of rocky intertidal assemblages. Consumers have been repeatedly shown to be prey species- and prey size-selective, while algal grazing consumers can inadvertently remove newly settled animals and algae as well as their intended prey.

Supply-Side Ecology and External Drivers: 1980–2005

Marine ecologists have known for a long time that success of many intertidal species depend on the supply of propagules (larvae, zygotes, and spores) from the plankton, but it was not until the 1980s that experiments were executed to assess how the supply of propagules influenced the patterns of distribution and abundance of adults in benthic assemblages.

Propagule supply and early post-settlement mortality markedly influence both the strength of interactions among established individuals and overall patterns of distribution and abundance on rocky shores. Abundance of established individuals is often directly proportional to the density of settlement and consequently, and strength of adult interactions depends on variation of settlement. In contrast, if settlement is high enough to consistently saturate the system, then local populations tend to be driven by strong interactions among adults regardless of settlement variation. In some cases, heavy early postsettlement mortality can lead to low densities of adults despite an abundance of settlers, and this has been shown for several seaweeds and many invertebrate species. The causes of variation in propagule supply can be classified into two broad categories – oceanographic transport or regional offshore production. Although invertebrate larvae and some macroalgal spores are motile, their movements are most directly important at small spatial scales near the substrate just prior to settlement. By and large, propagules of benthic species are transported at the mercy of currents and other oceanic transport phenomena. For instance, coastal upwelling results in a net offshore transport of propagules and leads to a reduction in settlement along a shoreline. This commonly occurs with invertebrate species that have long residence times in the plankton. In contrast, seaweeds, which have very short planktonic stages, often dominate intertidal sites within regions characterized by seasonal or permanent upwelling (Fig. 5).

Regional offshore production influences the supply of larvae to a coastal habitat in two ways. First, phytoplankton production in nearby waters offshore affects the abundance of planktotrophic larvae that feed for several weeks in the plankton potentially leading to greater larval supply in areas with greater phytoplankton production. Second and in opposition, increased production in offshore can generate increased resources and habitat for the associated pelagic community that preys upon larvae and thus leads to a reduced larval supply.

Unresolved Problems and Future Directions

Marine ecologists have been remarkably successful in advancing our knowledge of how strong local interactions affect the composition of communities, yet it is not yet clear how the results of small-scale experiments can be scaled up into broad scale generalizations. This is one of the major challenges of rocky intertidal ecology since practical, everyday concerns of management, commercial harvesting, biodiversity, and restoration demand answers on the scale of square kilometers of habitat, not square meters of experimental site. One current approach has been to use teams of researchers undertake identical small-scale experiments over a broad geographical region (e.g., EuroRock in Great Britain and Europe) or over similar oceanographic conditions (e.g., the ongoing studies of rocky shore in upwelling systems on the Pacific Rim by PISCO). Another approach has been the integration of `real time' physical, chemical, biological data from in situ and remote sensors (e.g., satellites that can reveal near shore temperature and primary productivity) with experimental studies on community dynamics.

Fig. 5 The intertidal zone near Antofagasta in northern Chile, a region with upwelling and abundant seaweeds. Photo by P. S. Petraitis.

Neither approach solves the difficulties of working with large mobile consumers such as mammals, whose importance is under appreciated because of the difficulties inherent with studying mammals. Even the rat (*Rattus norvegicus*) – the most widely recorded introduced intertidal mammal with the broadest documented intertidal diet – likely remains underreported as a rocky intertidal consumer from many coastal locations where it is known to be established. It is likely that rocky intertidal organisms supply terrestrial consumers significant amounts of energy, yet there are few data on intertidal–terrestrial linkages and how intertidal shores serve as important subsidies for terrestrial habitats.

It is also unclear if detailed information from one area can be informative about another area. For example, rocky intertidal shores on both sides of the Atlantic Ocean look surprisingly alike with not only the same species of plants and animals present but also similarities in their abundances and distributions. The similarity is so striking that a good marine ecologist, knowing little more than the direction of the prevailing swells, can list the 20 most common species on any 100 m stretch of shoreline. The average beachcomber could not tell if he or she were in Brittany, Ireland, Nova Scotia or Maine. The causes of this similarity are not well understood. Rocky shores in Europe and North America may look similar because of strong biological interactions maintain species in balance or because of historical accident, and these opposing views are endpoints on a continuum but represent one of the major intellectual debates in ecology today.

Finally ecosystems are not static, and rocky intertidal systems, which lie at a land–sea boundary, will be doubly affected by climate change as both oceanic conditions such as storm frequency and surge extent, and terrestrial conditions, such as air temperatures, are altered. Such changes could affect local communities by altering the disturbance dynamics and changing the geographic limits of intertidal species.

Further Reading

Connell, J.H., 1961. The influence of interspecific competition and other factors on the distribution of the barnacle *Chthamalus stellatus*. Ecology 42, 710–723.
Denny, M.W., 1988. Biology and Mechanics of the Wave-Swept Environment. Princeton, NJ: Princeton University Press.
Graham, L.E., Wilcox, L.W., 2000. Algae. Upper Saddle River, NJ: Prentice-Hall.
Horn, M.H., Martin, K.L.M., Chotkowski, M.A. (Eds.), 1999. Intertidal Fishes: Life in Two Worlds. San Diego, CA: Academic Press.
Koehl, M.A.R., Rosenfeld, A.W., 2006. Wave-Swept Shore: The Rigors of Life on a Rocky Coast. Berkeley, CA: University of California Press.
Levinton, J.S., 2001. Marine Biology. New York: Oxford University Press.
Lewis, J.R., 1964. The Ecology of Rocky Shores. London: English Universities Press.
Little, C., Kitching, J.A., 1996. The Biology of Rocky Shores. New York: Oxford University Press.
Moore, P.G., Seed, R. (Eds.), 1986. The Ecology of Rocky Coasts. New York: Columbia University Press.
Ricketts, E.F., Calvin, J., Hedgpeth, J.W., 1992. Between Pacific Tides, 5th edn. Stanford, CA: Stanford University Press, revised by Phillips DW.
Southward, A.J., 1958. The zonation of plants and animals on rocky sea shores. Biological Reviews of the Cambridge Philosophical Society 33, 137–177.
Stephenson, T.A., Stephenson, A., 1972. Life between Tidemarks on Rocky Shores. San Fransisco, CA: W. H. Freeman.
Underwood, A.J., 1979. The ecology of intertidal gastropods. Advances in Marine Biology 16, 111–210.
Underwood, A.J., Chapman, M.G. (Eds.), 1996. Coastal Marine Ecology of Temperate Australia. Sydney: University of New South Wales Press.
Underwood, A.J., Keough, M.J., 2001. Supply side ecology: The nature and consequences of variations in recruitment of intertidal organisms. In: Bertness, M.D., Gaines, S.D., Hay, M.E. (Eds.), Marine Community Ecology. Sunderland, MA: Sinauer Associates, pp. 183–200.

Salt Marshes

JB Zedler, CL Bonin, DJ Larkin, and A Varty, University of Wisconsin, Madison, WI, USA

Physiography

Salt marshes are saline (typically at or above seawater, $>34\ g\ l^{-1}$) ecosystems with characteristic geomorphology (sedimentary environments, fine soil texture, and relatively flat topography), herbaceous vegetation, and diverse invertebrates and birds. They occur along shores in estuaries, lagoons, forelands (open areas), and barrier islands in marine environments, and in shallow inland sinks where salts accumulate. They are not found where waves, currents, or streamflow create strong erosive forces. Salt (which stresses most species) severely limits the pool of plant species that can colonize saline sediments, and wetness typically confines the vegetation to herbaceous species, although some species are long-lived `subshrubs'. Given a near-surface water table, most shrubs and trees cannot establish their extensive root systems.

Plants of tidal marshes are usually able to colonize sediment above mean high water during neap tides (MHWN = average higher high-tide level during lower-amplitude neap tides, which alternate with the broader-amplitude spring tides). Sediment stabilization by halophytes initiates salt marsh formation. Plants not only slow water flow and allow sediments to settle out, but also their roots help hold sediments in place. Gradual accretion around plant shoots can further elevate the shoreline, allowing development of a marsh plain and transition to upland. This process can reverse, with tides eroding accumulated sediments. When sedimentation is outweighed by erosion, salt marshes retreat.

The overriding physiochemical influence is salt, which comes from marine waters, from exposed or uplifted marine sediments, or from evaporation of low-salinity water in arid-region sinks. Salt marshes along coasts typically have tidal influence (**Fig. 1**), although many nontidal lagoons have saline shores that support salt marsh vegetation. Salt marshes in inland settings occur in shallow sinks (e.g., around the Great Salt Lake, Utah, USA). The salts that contribute to salinity are primarily those of four cations (sodium, potassium, magnesium, calcium) and three anions (carbonates, sulfates, and chlorides); the relative proportions differ widely among soils of inland salt marshes, but sodium chloride is the predominant salt of seawater.

Tidal regimes differ around the globe, but most tidal marshes experience two daily high tides of slightly different magnitude, while some have the same high and low tides from day to day. Levels alternate weekly as neap and spring amplitudes, with the amplitudes readily predicted given gravitational forces between the Earth, the Moon, and the Sun (astronomic tides). Forces vary in relation to global position and coastal morphology; in southern California, mean astronomic tidal range is 3 m, while in the Bay of Fundy it is 16 m. The influence of seasonal low- and high-pressure systems on water-level oscillations (atmospheric tides) also vary greatly. For example, in Western Australia's Swan River Estuary, atmospheric tides outweigh astronomic tides. In the Gulf of Mexico, astronomic tides are minimal because of limited seawater connection with the Atlantic Ocean. Water levels within the Gulf vary only a few centimeters except during storms and seiches.

In tidal systems, marsh vegetation generally ranges from MHWN to the highest astronomic tide. Depending on tidal amplitude and the slope of the shore, salt marshes can be very narrow or kilometers wide. Strong wave action limits the lower salt marsh boundary, but a sheltered area can extend the lower boundary below MHWN.

Animal diversity is high, especially among the benthic and epibenthic invertebrates and the arthropods in the soil or plant canopies. Species that complete their life cycles within salt marshes either tolerate changing salinity and inundation regimes or avoid them by moving elsewhere or reducing contact. Globally, salt marshes are known to support large populations of migratory birds in addition to resident birds, insects, spiders, snails, crabs, and fin and shellfish. Indeed, foraging is the most visible activity in salt marshes.

Extent

Salt marsh area is not well inventoried. The global extent of pan, brackish, and saline wetlands is approximately 435 000 km^2, or 0.3% of the total surface area and 5% of total wetland area. In USA, the 48 conterminous states have about 1.7 Mha of salt marshes, out of a total of 42 Mha of wetlands.

While broadly distributed, salt marshes are most common in temperate and higher latitudes where the temperature of the warmest month is $>0\ ^{\circ}C$. Closer to the equator, where the mean temperatures of the coldest months are $>20\ ^{\circ}C$, salt marshes are generally replaced by mangroves. Salt marshes sometimes occur inland of mangroves or instead of mangroves where woody plants have been removed.

Habitat Diversity

Habitats within the salt marsh vary with elevation, microtopography, and proximity to land or deeper water. In southern California, the high marsh, marsh plain, and cordgrass (*Spartina foliosa*) habitat tend to follow elevation contours, although

 doi:10.1016/B978-0-444-63768-0.00357-7

Fig. 1 A tidal marsh in San Quintin Bay seen from the air. Image by the Pacific Estuarine Research Lab.

cordgrass is often restricted to low elevations adjacent to bay and channel margins. Other habitats are related to minor variations in topography, which impound fresh or tidal water. For example, back-levee depressions, tidal pools, and salt pans occur where drainage is somewhat impaired. Salt marshes along the Atlantic Coast of USA are very extensive, with *S. alterniflora* creating a monotype except for a narrow transition at the inland boundary where succulent halophytes or salt pans are found.

Tidal creeks provide diverse habitats for plants and animals. Banks are often full of crab burrows, and creek bottoms harbor burrowing invertebrates and fishes. They also serve as conduits for fish, fish larve, phyto- and zooplankton, plant propagules, sediments, and dissolved materials, which move between the salt marsh and subtidal channels.

Adjacent habitats can include small, unvegetated salt pans that dry and develop a salt crust, especially during neap tides. Salt pans occur where salt concentrations exceed tolerance of halophytes. During heavy rains or high tides, water fills the pan, creating temporary habitat for aquatic algae and animals and permanent habitat for the species that survive the dry spells *in situ* as resting stages. More extensive salt pans are sometimes called salt flats. Other nearby habitats usually include mudflats (where inundation levels exceed tolerance of halophytes), brackish marsh (where salinities are low enough for brackish plants to outcompete halophytes), sandy or cobble beaches (where wave force excludes herbaceous vegetation), sand dunes (where soils are too coarse and dry for salt marsh plants), and river channels (where freshwater enters the estuary and is not sufficiently saline).

Salt Marsh Plants

Salt-tolerant plants (halophytes) include herbaceous forbs, graminoids, and dwarf or subshrubs. Many of the forbs are succulent (e.g., *Sarcocornia* and *Salicornia* spp.). Graminoids often dominate Arctic salt marshes, while subshrubs dominate salt marshes in Mediterranean and subtropical climates. Many salt marshes support monotypic stands of cordgrass (*Spartina* spp.) (**Table 1**).

Floristic diversity of salt marshes is low because few species are adapted to saline soil. Members of the family Chenopodiaceae comprise a large proportion of the flora (e.g., species of *Arthrocnemum*, *Atriplex*, *Chenopodium*, *Salicornia*, *Sarcocornia*, and *Suaeda*). In contrast to the flowering plants, salt marsh algae are diverse in both species and functional groups (green macroalgae, cyanobacteria, diatoms, and flagellates).

NaCl is a dual stressor, as it challenges osmotic regulation and sodium is toxic to enzyme systems. Salt marsh halophytes cope with salt by excluding entry into roots, sequestering salts intracellularly (leading to succulence), and excreting salt via glands, usually on leaf surfaces. One succulent, *Batis maritima,* continually drops its older salt-laden leaves, which are then washed away by the tide. I. Mendelssohn has attributed moisture uptake from seawater to the ability of some species to synthesize prolines.

Prolonged inundation reduces the supply of oxygen to soils, causing anoxia and stressing vascular plants. In addition, abundant sulfate in seawater is reduced to sulfide in salt marsh soil, with high sulfide concentrations, which are toxic to roots.

Salt marsh vascular plants withstand brief inundation but do not tolerate prolonged submergence, as occurs when a lagoon mouth closes to tidal flushing and water levels rise after rainfall. Salt marshes in lagoons thus experience irregular episodes of dieback and regeneration in relation to ocean inlet condition.

Regular inundation benefits halophytes by importing nutrients and washing away salts. Salts that accumulate on the soil surface during daytime low tides and salts excreted by halophytes are removed by tidal efflux. Thus, soil salinities are relatively stable where tidal inundation and drainage occur frequently. Inland salt marshes, however, experience infrequent reductions in salinity during rainfall, and soils can become extremely hypersaline (e.g., $>10\%$ salt). In between irregular inundation events, halophytes and resident animals endure hypersaline drought.

Table 1 Representative species of global salt marshes based on a summary by Paul Adam

Arctic Puccinellia phryganodes dominates the lower elevations
Boreal Triglochin maritima and *Salicornia europea* are widespread. Brackish conditions have extensive cover of *Carex* spp.
Temperate
Europe: *Puccinellia maritima* dominated lower elevations historically (but *Spartina anglica* often replaces it). *Juncus maritimus* dominates the upper marsh; *Atriplex portulacoides* is widespread
USA:
Atlantic Coast: *Spartina alterniflora* is extensive across seaward marsh plain; *S. patens* occurs more inland
Gulf of Mexico: *Spartina alterniflora* and *Juncus roemerianus* dominate large areas
Pacific Northwest: *Distichlis spicata* in more saline areas, *Carex lyngbei* in less saline areas
California: *Spartina foliosa* along bays, *Sarcocornia pacifica* inland
Japan: *Zoysia sinica* dominates the mid-marsh
Australasia: *Sarcocornia quinqueflora* dominates the lower marsh, *Juncus kraussii* the upper marsh
South Africa: *Sarcocornia* spp. are abundant in the lower marsh, *Juncus kraussii* in the upper marsh. *Spartina maritima* is sometimes present
Dry coasts vegetation tends toward subshrubs, such as *Sarcocornia*, *Suaeda*, *Limoniastrum*, and *Frankenia* species
Tropical Sporobolus virginicus and *Paspalum vaginatum* form extensive grasslands. *Batis maritima*, *Sesuvium portulacastrum*, and *Cressa cretica* are also found

Salt Marsh Animals

The salt marsh fauna includes a broad taxonomic spectrum of invertebrates, fishes, birds, and mammals, but few amphibians and reptiles. Resident fauna are adapted to the land–sea interface, while transient users benefit from the foraging, nursery, and reproductive support functions.

Salt marsh animals cope with inundation regimes that differ seasonally, monthly, daily, and hourly. Vertebrates accomplish this largely through mobility. For example, fishes exploit marsh surface foraging opportunities during high tides and then retreat to subtidal waters. Birds time their use to take advantage of either low or high tide. Residents, such as the light-footed clapper rail (*Rallus longirostris levipes*), nest during the minimum tidal amplitude. Migrants, such as curlews, move upslope at a high tide and feed during low tide during their seasonal visits. Many invertebrates move away from adverse conditions. Some beetles climb tall plants to escape rising tides. A springtail, *Anurida maritime*, has a circatidal rhythm of 12.4 h that enables it to emerge for feeding shortly after tides ebb and retreat underground prior to the next inundation. For less-mobile fauna, physiological adaptations are essential. Gastropods avoid desiccation during low tides by sealing their shells. Some arthropods avert drowning by trapping air bubbles in their epidermal hairs during high tides.

Another challenge is fluctuating salinities, which salt marsh residents handle with exceptional osmoregulatory ability. The southern California intertidal crab species *Hemigrapsus oregonensis* and *Pachygrapsus crassipes* are able to hypo- and hyperosmoregulate when exposed to salt concentrations ranging from 50% to 150% of seawater (brackish to hypersaline). Tidal marsh fishes also have wide salinity tolerances. Cyprinodontiform tidal marsh fishes can tolerate salinities as high as 80–90 ppt. One species, *Fundulus majalis*, hatched at salinities up to 72–73 ppt. Lower salinity limits for mussels can be as low as $3\ g\ l^{-1}$ and they can tolerate high salinities as well, with mussels able to tolerate losing up to 38% of their water content. Even birds have adaptations for dealing with salt water and saline foods; for example, the Savannah sparrow *(Passerculus sandwichensis beldingi)* has specialized glands that excrete salt through the nares.

Because salt marshes have continuously changing hydrology, small differences in elevation and topography (e.g., shallow, low order tidal creeks) influence foraging activities of fishes and birds by regulating inundation and exposure times, enhancing marsh access for fishes, and increasing edge habitat. Ephemeral pools of just centimeters in depth provide valuable bird habitat, enhance macroinvertebrate abundance and diversity, and support reproductive, nursery, and feeding support functions for fishes.

Ecology

Salt marshes are well studied relative to their limited global area. Knowledge of salt marsh ecology is strongest for vegetation, soil processes, and food webs. Conservation is an emerging issue, given threats of sea-level rise in concert with global warming.

Vegetation and Soils

In Europe, salt marsh ecology developed around floristics and phytosociology. In USA, research on the Atlantic and Gulf Coasts characterized salt marsh ecosystem functioning, especially productivity, microbial activities, outwelling of organic matter, food webs, and support of commercial fisheries, while on the Pacific Coast, studies concern the impacts of invasive species of *Spartina* and effects of extreme events on vegetation dynamics. In Canada, effects of geese damaging vegetation are a research focus. Studies of USA's inland salt marshes have contributed knowledge of waterfowl support functions and halophyte salt tolerance. In South Africa's small estuaries, *Spartina* productivity and shifts of vegetation in response to altered freshwater inflows have been explored.

In Asia, widespread plantings of *S. alterniflora* have been undertaken in order to extend coastal land area, provide forage, and produce grass for human use. In general, salt marshes of Asia, Central America, and South America are poorly known.

Salt marshes develop primarily on fine sediments, but salt marsh plants can grow on sand and sometimes gravel. Older salt marshes have peaty soils, especially in cooler latitudes where decomposition is slow.

Both roots and burrowing invertebrates affect soil structure by creating macropores in soil. Invertebrates also cause bioturbation, a process whereby sediments are re-suspended and potentially eroded away. This activity can be countered by algae and other microorganisms, which form biofilms on the soil surface. Biofilms cement soil particles and reduce erosion; they also add organic matter, and those that contain cyanobacteria fix nitrogen.

Salt marsh soils are often anoxic just below the surface due to high organic matter content and abundant moisture for microorganisms. This is especially so in lower intertidal areas and in impounded marshes. Tidal marsh soils are typically high in sulfur, which forms sulfides that blacken the soil, emit a distinctive rotten-egg smell, and stress many plants. Across intertidal elevation ranges, soil microorganisms, sulfides, and inundation regimes reduce species richness where inundation is most prolonged, often to a single, tolerant species.

Food Webs

Studies of salt marshes have made important advances in food web theory. Early papers focused on primary productivity measurements and attempts to explain differences in rates within and among salt marshes. The energy-subsidy model described *S. alterniflora*'s high productivity at low elevations as a function of increased rates of nutrient delivery and waste removal, due to frequent tidal inundation. It also explained how salt marshes with decreasing tidal energy across Long Island, New York, had a corresponding decrease in *S. alterniflora* productivity. R. E. Turner added the role of climate by relating higher productivity of *S. alterniflora* to warmer latitudes.

In the 1960s, E. Odum's interest in energy flow led several investigators at the University of Georgia to quantify productivity, consumption, and decomposition of various components of Sapelo Island salt marshes. J. Teal's energy-flow diagram depicted Georgia's *S. alterniflora* marsh as exporting organic matter. Although estimated by subtraction rather than measurement, detrital export became a textbook example of how ecosystems channel and dissipate energy.

Later, advances were made in exploring the quantity and fate of detritus derived from salt marsh primary producers. J. Teal's suggestion that substantial organic matter is transported to estuarine waters supported E. Odum's `outwelling hypothesis', that estuarine-derived foods drive coastal food webs and benefit commercial fisheries. A number of ecosystem-scale tests of outwelling ensued, and although outwelling did not prove to be universal, the research demonstrated connectivity between riverine, salt marsh, and open-water ecosystems. Also, the copious detrital organic matter provided by salt marshes was shown to be high in nutritional value once detrital particles were enriched by microorganisms, but microalgae were also shown to be an important food source. Even though their standing crop is low, high turnover rates lead to high primary productivity. In salt marshes with ample light penetration through the vascular plant canopy, microalgae can be as productive as macrophytes, and some species (notably cyanobacteria) are much richer in proteins and lipids. Algae also hold much of the labile nitrogen in salt marshes, widely thought to be the limiting factor for growth of invertebrate grazers.

Food webs are driven by both `bottom-up' or `top-down' processes. Evidence for bottom-up control of trophic interactions comes from experimental addition of nitrogen. Nitrogen has been shown to limit algae, vascular plants, grazers, and predatory invertebrates in nearly every salt marsh field experiment. Recently, however, P. V. Sundareshwar and colleagues showed that phosphorus can limit microbial communities in coastal salt marshes.

Despite widespread evidence for bottom-up effects, there is expanded recognition of the top-down role of consumers in regulating salt marsh food webs. Populations of lesser snow geese have increased due to agricultural grains that are left in the fields after harvest, and large flocks now cause large-scale destruction of vegetation in Arctic salt marshes due to rampant herbivory. In Atlantic salt marshes of southern USA, snail herbivory accompanies drought-induced die-back of *S. alterniflora*.

Ecosystem Services

Several ecosystem services provided by salt marshes are appreciated by society, and some protective measures are in place. The regular rise and fall of water in salt marshes, either daily with tides or seasonally with rainfall, enhances at least six valued functions:

Denitrification improves water quality. The sediments of tidal marshes are well suited to denitrification, which occurs most rapidly at oxic–anoxic interfaces. The first step, nitrification, occurs near soil–water or root–soil interfaces or along pores where oxygen enters the soil at low tide. The second step requires anoxic conditions and proceeds rapidly where moisture is sufficient for bacteria to respire and remove oxygen. In this step, nitrate is reduced to nitrogen gas in a series of microbially mediated steps. The rise and fall of tide waters ensures that oxic and anoxic conditions coexist.

Carbon sequestration slows greenhouse warming. The high net primary productivity of salt marshes creates high potential for carbon storage and the anoxic soils slow decomposition, so carbon can accumulate as peat. Large standing crops of roots, rhizomes, and litter are fractionated by a diversity of invertebrates and microorganisms and incorporated into soil. Rates are potentially highest at

cooler latitudes, where decomposition is slowed by low temperatures. Sea-level rise is also a key factor; as coastal water levels become deeper, decomposition slows. Sedimentation also buries organic matter, making it less likely to decompose. With sea level rising a millimeter or more per year, on average, salt marsh vegetation can build new rooting zones above dead roots and rhizomes of past decades. Along the USA Gulf of Mexico, the ability of salt marshes to keep up with rising sea level is attributed to root and rhizome accumulation, not just sedimentation. If decomposition proceeds anaerobically to states that produce methane, however, not only is carbon storage reversed, but carbon is also released in a form that contributes more to global warming than carbon dioxide.

Fin- and shellfish have commercial value. Tidal marshes are valued for their nursery function, meaning that the young of many fishes, crabs, and shrimp make use of estuarine waters as `rearing grounds'. In the USA, it is estimated that some 60% of commercial species spend at least part of their life cycle in estuaries. Several attributes of salt marshes contribute to the food-web-support function, including high productivity of both algae and vascular plants, detritus production and export to shallow water-feeding areas, refuge from deepwater predators, plant canopy cover as a refuge from predatory birds, warmer temperatures that can accelerate growth, and potential to escape disease-causing organisms and parasites that might have narrower salinity tolerance.

Forage is used to feed livestock. In Europe and Asia, graziers move cattle, horses, sheep, or goats onto the marsh plain during low tides. It is common to see ponies tethered to stakes in *Puccinellia*-dominated salt marshes of UK. The temporary availability (between tides) allows recovery between use and, potentially, high-quality forage and salt for livestock.

Recreational opportunities and esthetics are appreciated by people who live near or visit coastal areas. By virtue of their low-growing vegetation and locations between open water and urban areas, salt marshes attract both wildlife and people. The combination provides high value for birdwatchers, hikers, joggers, and artists. Where there is flat topography above and near the salt marsh, the needs of elderly and disabled visitors can be accommodated along with hikers, school children, and those seeking a refuge from city life. Of particular interest is the ever-changing view, as tides rise and fall along marine coasts, and as water levels change with season in inland systems. Visitor centers have been constructed near many urban salt marshes. Ecotourism then adds economic value to the local municipality as well as the larger region.

Shorelines are anchored by salt marsh vegetation. Recent damages from hurricanes and tsunamis have called attention to the protection that wetland vegetation provides to coastal lands, and especially high-cost real estate. Water flow is slowed by stems and leaves of salt marsh plants, and their roots and rhizomes bind inflowing sediments. Mucilage produced by biofilms (algae, fungi, and bacteria) can then cement particles until new plant growth anchors the substrate. The stems of vascular plants are often coated with biofilms, particularly those of tuft-forming cyanobacteria, such that the total surface area available for sediment-trapping and anchoring is greatly enhanced. Floating mats of green macroalgae *(Ulva, Enteromorpha)* also collect sediments and, when they move to the wrack line and join other debris, add to accretion at the upper marsh plain boundary.

Challenges for Salt Marsh Conservation

Habitat Loss

Estuaries, where rivers meet the sea, are not only suitable for salt marsh development but also ideal places for human habitation. The ocean–river connection is a navigational link, flat land is easy to build upon, the river provides drinking water, the salt marsh and coastal fisheries provide food, outgoing tides facilitate wastewater disposal, and seawater provides an essential preservative and universal seasoning, NaCl. Thus, many cities, such as Venice, Boston, Amsterdam, London, Buenos Aires, Washington, DC, and Los Angeles, were built on or rapidly grew to displace salt marsh ecosystems. Major ports within smaller natural bays, such as San Diego, have displaced nearly all the natural salt marsh, while others, such as San Francisco, sustain large salt marshes despite extensive conversion.

The process of converting salt marsh into nontidal land was historically called reclamation. The practice of building embankments to exclude tidal flows eliminated thousands of hectares of European salt marshes. In the Netherlands, embankments reclaimed substantial land as polders for agriculture. In USA, reclamation reduced salt marsh area by 25% between 1932 and 1954. While the trend is to halt or reverse this practice, estuaries are being dammed in Korea to create tillable fields from mudflats. In Vietnam, Mexico, and other coastal nations, salt marshes are yielding to fish and shrimp impoundments. In such cases, people who use mudflats for fishing and crabbing are displaced by farmers.

Although salt marshes are highly valued, they are increasingly threatened by human population growth. It is estimated that 75% of the global population will live within 60 km of the coast. Thus, coastal ecosystems are particularly at risk.

Eutrophication

Salt marshes are enriched when phosphorus and/or nitrogen flow into waters that ultimately flood the salt marsh. Agricultural fertilizers applied to fields throughout coastal watersheds move downslope into waters that flow toward salt marshes. Because many salt marshes are nitrogen-limited, the effect is to increase the productivity of both algae and vascular plants. Increased nitrogen loading stimulates algal growth, especially of green macroalgae, which form large mats that can smother vascular plants and benthic invertebrates. Indirect degradation occurs when microbial decomposition increases oxygen demand, causing soil hypoxia or anoxia and sulfide toxicity.

I. Valiela's long-term eutrophication experiment in a New England salt marsh indicates that nitrogen addition shifts *S. alterniflora* to *S. patens* and increases competition for light. Such altered competitive relationships are likely widespread, especially where considerable nitrogen is deposited from the air (e.g., from dairy operations in the Netherlands).

Sediment Supply

Both reduced and enhanced sediment supplies can threaten the persistence of salt marsh ecosystems. Sediment supplies are reduced when water is removed from rivers for irrigation, human consumption, and industrial use, or when overbank flooding is prevented by engineering works. Reduced sediment supply from the Mississippi River is one factor contributing to salt marsh loss in Louisiana.

Excessive sediments flow into salt marshes where the catchment has lost vegetative cover as a result of logging, farming, or development. Inflows also occur where mining operations discharge materials directly to streams. Wastes from California's gold rush are still making their way to San Francisco Bay. At a much smaller estuary, the marsh plain of Tijuana Estuary in southern California has elevated 25–35 cm since 1963 due to erosion from rapidly urbanizing canyons in nearby Tijuana, Mexico. The impacts have been losses of microtopographic variation and local species richness.

Global Change

Increases in global mean temperature will have substantial impacts on the world's salt marshes. Sea levels rise when high-elevation glaciers and polar ice caps melt and when seawater warms and expands. The impacts of more rapidly rising sea level depend on rates of sedimentation and uplift. If sediment accretion is equal to sea-level rise, the salt marsh remains in place, but when sea-level rise exceeds sediment accretion, the salt marsh moves inland – unless bluffs or development limits salt marsh migration. As sea level rises relative to the land, salt marsh communities will experience increased inundation, such that plant and animal species should shift upslope. However, not all species will be able to disperse or migrate as rapidly as tidal conditions change. In a few cases, for example, Scandinavia, the coast is still rebounding from the pressure of former glaciers, and land is rising faster than sea level. Salt marsh is then lost at the upper end and slowly gained near the water.

Globally, mean sea level has risen 10–25 cm during the last century. Current models predict an additional 5.6–30 cm rise in sea level by 2040. In areas of rapid shifts in sedimentation or high erosion due to wind and waves, salt marshes are destabilized and threatened with compositional changes and/or loss of marsh area. Salt marshes are also threatened by subsidence; if the land settles faster than sediment or roots and rhizomes can accumulate, vegetated areas convert to open water. USA's largest area of salt marsh loss is along the Louisiana coastal plain, where subsidence, decreased sedimentation, canal dredging, levee construction, and other human disturbances eliminate more than 4300 ha yr^{-1}.

Coastal watersheds that experience increased storminess as a result of climate change will discharge water, sediments, nutrients, and contaminants more erratically than at present, with resulting impacts on salt marshes downstream.

Soil salinity might also rise with higher temperatures, increased evaporation, and increased evapotranspiration. With more rainfall and freshwater flooding, however, soil salinity might decrease. The net effect of warming on salt marsh soil salinity is difficult to predict. Increased storminess could translate into more or stronger dune washover events during high tides, and stronger ocean swells would transport seawater further inland. The toxic effect of salt on upland vegetation, coupled with persistent salt in the soil, would favor halophytes over glycophytes in an increasingly broader wetland–upland transition areas (**Fig. 2**). This prediction is most likely for areas of low annual rainfall, such as Mediterranean-type climates.

Climate change is likely to affect species differently, potentially altering competitive relationships. Photosynthesis, transpiration, nutrient cycling, phenology, and decomposition are influenced by temperature. Salt marshes with a mixture of C3 and C4

Fig. 2 Saltmarsh vegetation from the upland–wetland interface (foreground) to San Quintin Bay, Baja California Peninsula, Mexico. Photo by J. Zedler.

plants might shift toward C4 plants as mean temperature climbs; however, elevated CO_2 might favor C3 species. In subtropical regions, a warming trend and sea-level rise would likely allow mangroves to move northward and displace salt marshes.

Impacts of climate change to plants and animals are difficult to estimate. European ecologists, however, have detailed information on bird use of salt marshes and can predict shifts in invertebrate foods and shorebirds given various scenarios of sea-level rise.

Invasive Species

Plant and animal species are inadvertently moved around the globe when ships take on ballast water in one port and discharge it in another; seeds of alien plants and either live animals or dormant stages are then available to colonize salt marshes. When the USA resumed trade with China, new invaders gained access to San Francisco Bay. Fred Nichols traced the arrival of a small clam, *Potamocorbula amurensis*, to 1876. Now it coats some benthos with thousands of clams/m^2.

Other alien species have been intentionally introduced. In the 1950s, the US Army Corps of Engineers experimentally introduced *S. alterniflora* onto several dredge spoil islands to stabilize the material and provide wildlife habitat. A region-wide invasion of the Pacific Northwestern USA followed several decades of `benign' behavior. Today, the species is dominant along the lower edge of salt marsh shorelines, where it displaces oysters and eliminates shorebird-feeding habitat.

Once a species has taken up residence, it might hybridize with native species and become more aggressive, either as the hybrid or subsequent genetic variants. Such is the case for *S. alterniflora*, which has been widely planted in Europe, China, Great Britain Australia, and New Zealand. In Great Britain, it hybridized with the native *S. maritima* to form *S. townsendii*, which then underwent chromosomal doubling to form *S. anglica*. *S. anglica* can grow at lower elevations than native species and vigorously colonizes mudflats. Dense clones of *S. anglica* reduce habitat for wading birds and displace native salt marsh plants.

Non-native strains of *Phragmites australis* were introduced to the USA 200 years ago, and they have since spread throughout much of North America. Today, the alien strain dominates the less-saline portions of salt marshes in the northeastern USA, where it displaces native plant species, alters soil conditions, and decreases waterfowl use. Disturbances such as ditching or dredging open salt marsh canopies and allow invasion of *P. australis*, while eutrophication, altered hydrologic regimes, and increased sedimentation favor its spread.

Invasive plant species have been linked to reduced diversity, shifts in trophic structure, habitat alteration, and changes in nutrient cycling. Invasive alien animals are equally problematic. In San Francisco Bay wetlands, alien mudsnails outcompete native ones and the Australasian isopod, *Sphaeroma quoyanum*, burrows into and destabilizes creek banks of tidal marshes, causing erosion. Marsh edge losses exceeding 100 cm yr^{-1} have been reported in heavily infested areas. Another invader, the green crab, *Carcinus maenus*, has altered the food web of Bodega Bay, California, by reducing densities of a native crab, two native mussels, and other invertebrates. As the green crab moves north, it will likely reduce food availability for shorebirds.

In the southeastern USA, fur farmers introduced nutria (*Myocastor coypus*) from South America in the 1930s. These rodents feed on roots of salt marsh plants. When fur clothing went out of style, nutria populations expanded and began converting large areas of marsh to mudflat and open water.

Chemical Contamination

Chemical contaminants accumulate in salt marshes that receive surface-water runoff and/or direct discharges of waste materials. Among the most toxic are halogenated hydrocarbons, which include many insecticides, herbicides, and industrial chemicals. When accumulated in the tissues of salt marsh animals a wide range of disorders can result, for example, immunosuppression, reproductive abnormalities, and cancer.

Petroleum hydrocarbons pollute harbors and remnant salt marshes following oil spills, urban runoff, and influxes of industrial effluent and municipal waste. Once they move into anoxic sediments, they can persist for decades, reducing primary production, altering benthic food webs, and accumulating in bird tissues. Polycyclic aromatic hydrocarbons have additional carcinogenic and mutagenic potential for aquatic organisms.

Heavy metals are also toxic to aquatic organisms and can impair feeding, respiration, physiological and neurological function, and reproduction, as well as promote tissue degeneration and increase rates of genetic mutation. Mercury is especially problematic because it is methylated in the anoxic soils of salt marshes and is then able to bioaccumulate in food chains.

Salt marsh plants in urban areas take up, accumulate, and release heavy metals. Judith Weis and others have found lowered benthic diversity and impaired fish behavior in contaminated sites. Fish are slower to catch prey and less able to avoid predators where heavy metals contaminate their habitat.

Research Value

Tidal marshes include an impressive array of environmental conditions within about a meter of elevational range. The compressed environmental gradient invites studies of species $\times$ abiotic factors, and over time, their contributions proceeded from community ecology to ecosystem science and, finally, integration of the two.

Community Ecology

The limited number of vascular plant species has made salt marshes very suitable for both descriptive and manipulative studies. Early researchers attributed plant species distributions to their physiological tolerance for the abiotic environment, without regard to species interactions. J. A. Silander and J. Antonovics used perturbation-response methods to determine that biotic forces also affected species distributions. Others effectively used reciprocal transplanting to examine the relative importance of abiotic conditions and interspecific competition to species distributions. For example, S. Pennings and R. Callaway revealed interspecific interactions among southern California halophytes, and S. Hacker and M. Bertness reported interspecific interactions among New England halophytes. Manipulative transplantation has shown that species distributions respond to abiotic conditions, facilitation, and competition.

The wide latitudinal range of salt marshes allowed study of community structure and function in relation to sea-level variations, for example, James Morris documented and modeled interannual variations in salinity and its effect on *S. alterniflora* growth. Such studies led to predictions of changes in response to global climate change.

Ecosystem Functioning

The monotypic nature of USA Atlantic Coast salt marshes aided early studies of vascular plant productivity and considerable literature developed around the rates of productivity and alternative methods of calculating gross and net productivity – work that transferred to grasslands and other nonwoody vegetation. Nitrogen dynamics were a later focus. The first marine system to have a nitrogen budget was Great Sippewisset Marsh in Massachusetts. The budget quantified nitrogen inputs from groundwater, precipitation, nitrogen fixation, and tidal flow, and nitrogen outputs from tidal exchange, denitrification, and buried sediments.

Integrating Structure and Function

A long controversy over the causes of height variation in *Spartina* spp. has involved USA researchers on both the Atlantic and Pacific Coasts and has linked plant and ecosystem ecology. The most convincing evidence for a genetic (`nature') component is that of D. Seliskar and J. Gallagher, who grew genotypes from Massachusetts, Georgia, and Delaware for 11 years in a common garden and documented persistent phenotypic differences. A series of papers on soil biogeochemistry explained the role of `nurture'. Nitrogen was shown to be a key limiting factor for *S. alterniflora* plant growth because nitrate is quickly reduced to ammonia by bacteria in poorly drained areas away from creeks, where soils have lower soil redox potential. Sulfate-reducing bacteria were also implicated, because they reduce sulfate to sulfide, which impairs the growth of sensitive plant species. Increased soil redox potential and greater pore water turnover in creek-side habitat contributes to taller height forms of *S. alterniflora*. Thus, both genetics and environment influence height forms of *S. alterniflora,* an outcome of both community and ecosystem research.

Restoration

With recognition of lost ecosystem services, interest in restoring salt marshes is growing in Europe and the USA. One way that the British are combating rising sea levels is via `managed retreat', which involves breaching of embankments to restore tidal flushing to lands that were once salt marshes. In the Netherlands, tidal influence is being reinstated to various polders along the southwestern coast to restore natural processes and diverse estuarine biota to former polders.

Fig. 3 Tidal marsh vegetation is typically dominated by salt-tolerant grasses and succulent forbs, easily distinguished in this restored marsh at Tijuana Estuary, near San Diego, California. Photo by J. Zedler.

Some of the earliest salt marsh restoration in USA has been accomplished as mitigation for damages to other sites as required by federal regulatory agencies. In North Carolina, *S. alterniflora* was being replanted in the 1970s, and the practice has expanded widely to mitigate damages due to development.

Some of the most innovative research on wetland restoration has been accomplished in salt marshes by replicating variables in restoration sites; for example, D. Seliskar and J. Gallagher showed that genotypic variation in *S. alterniflora* has implications for nearly every component of the food web (in Delaware), T. Minello and R. Zimmerman showed that channels in replanted salt marshes enhanced fish support (Galveston Bay, Texas), I. Mendelssohn and N. Kuhn showed that dredge spoil addition accelerated *S. alterniflora* recovery in subsiding wetlands (Louisiana), Cornu showed that topographic variation across a tidal floodplain affected salmon use (Oregon), and J. Callaway, G. Sullivan, J. Zedler, and others showed that planting diverse assemblages and incising tidal creeks jumpstarted ecosystem functioning in salt marsh restoration sites (Tijuana Estuary, California) (**Fig. 3**). In Spain, restoration of tidal ponds is being accomplished in replicate excavations that test the effect of size and depth on use by salt marsh animals (Doñana Marshlands).

In conclusion, salt marshes perform highly valued ecosystem services that are lost when habitats are developed and/or degraded. Further innovations will likely take place in both the practice and science of restoration, because salt marshes are highly amenable to experimentation.

See also: Behavioral Ecology: Herbivore-Predator Cycles. Ecological Complexity: Goal Functions and Orientors; Panarchy. Ecological Processes: Nitrification. Human Ecology and Sustainability: Ecological Footprint

Further Reading

Adam, P., 1990. Saltmarsh Ecology. Cambridge, UK: Cambridge University Press.
Adam, P., 2002. Saltmarshes in a time of change. Environmental Conservation 29, 39–61.
Allen, J.R.L., Pye, K., 1992. Saltmarshes: Morphodynamics, Conservation and Engineering Significance. Cambridge, UK: Cambridge University Press.
Chapman, V.J., 1960. Salt Marshes and Salt Deserts of the World. Plant Science Monographs. London: Leonard Hill [Books] Limited.
Daiber, F.C., 1982. Animals of the Tidal Marsh. New York, NY: Van Nostrand Reinhold Co.
Long, S.P., Mason, C.F., 1983. Saltmarsh Ecology. Glasgow: Blackie & Sons Ltd.
Pennings, S.C., Bertness, M.D., 2000. Salt marsh communities. In: Bertness, M.D., Gaines, S.D., Hay, M.E. (Eds.), Marine Community Ecology. Sunderland, MD: Sinauer Associates Inc., pp. 289–316.
Pomeroy, L.R., Weigert, R.G., 1981. The Ecology of a Salt Marsh. New York: Springer.
Reimold, R.J., Queen, W.H. (Eds.), 1974. The Ecology of Halophytes. New York, NY: Academic Press Incorporated.
Seliskar, D.M., Gallagher, J.L., Burdick, D.M., Mutz, L.A., 2002. The regulation of ecosystem functions by ecotypic variation in the dominant plant: A *Spartina alterniflora* salt-marsh case study. Journal of Ecology 90, 1–11.
Threlkeld, S., (ed.). *Estuaries and Coasts: Journal of the Estuarine Research Foundation.* Lawrence, KS: Estuarine Research Federation.
Weinstein, M.P., Kreeger, D.A. (Eds.), 2000. Concepts and Controversies in Tidal Marsh Ecology. Boston, MA: Kluwer Academic Publishers.
Zedler, J.B. (Ed.), 2001. Handbook for Restoring Tidal Wetlands. New York, NY: CRC Press.
Zedler, J.B., Adam, P., 2002. Saltmarshes. In: Perrow, M.R., Davy, A.J. (Eds.), Handbook of Ecological Restoration, vol. 2. Ress, Cambridge, UK: Cambridge University Press, pp. 238–266.

Ecosystems: Savanna☆

Lindsay B Hutley, Charles Darwin University, Darwin, NT, Australia
Samantha A Setterfield, The University of Western Australia, Perth, WA, Australia

Introduction

This article examines the ecological features of one of the most important tropical ecosystems, the savannas. Savannas feature the coexistence of both trees and herbaceous plants and are distinct from grasslands (absence of woody plants) and closed forests (tree dominant). Savanna ecosystems occur in over 20 countries, largely in the seasonal tropics. Much of the world's livestock occurs in savanna, underlining their social and economic importance. Approximately 20% of the world's land surface is covered with savanna vegetation, which produces almost 30% of global net primary production. With both tree and herbaceous life-forms, savanna biodiversity is high, often higher than associated dry deciduous forests. Globally, the tenure of savanna lands incorporates pastoral, private use, indigenous, and national parks, with the disparate management aims of grazing, agriculture, mining, tourism, subsistence livelihoods, and conservation. Given their size, savannas affect global carbon, nutrient, and water cycles and with their frequent fires, significantly influence atmospheric chemistry. Savanna ecosystems have existed for approximately 8 million years, although paradoxically, many ecologists regard savannas as an ecologically unstable mixture of trees and grasses. Savanna boundaries are dynamic in space and time and their occurrence and structure are determined by a combination of environmental factors, such as available water, nutrients, the frequency of disturbances (e.g., fire and herbivory) and stochastic weather events. This range of factors results in significant structural variation and providing an overarching and strict definition of what constitutes a savanna has been problematic. This article will provide a commonly used definition, describe savanna distribution and examine factors that influence their structure and function. Understanding the determinants of savanna functioning, resilience and stability are vital ingredients for improved management. Savannas are under increasing development pressure and threats to their long-term sustainability are examined.

Definition and Occurrence

Savanna ecosystems predominantly occur in the seasonal tropics and are a unique mix of coexisting trees, shrubs, and grasses (Fig. 1). Debate surrounds the use and definition of the term savanna, reflecting the range of tree–grass ratios found in these ecosystems. Savanna ecosystems feature a range of structures, from near treeless grasslands to woody dominant open-forest/woodlands of up to 80% woody cover. A widely used, generic definition describes a savanna ecosystem as one consisting of a continuous or near continuous grass dominated understorey, with a discontinuous woody overstorey. Woody components can be a mix of trees and shrubs of evergreen or deciduous phenology, broad or needle leafed. The grass-dominated understory can consist of a mix of species with either annual or perennial habit (often >1 m in height). Ecosystems that fit this definition have ambiguously been termed woodlands, rangelands, grasslands, wooded grasslands, shrublands, open-forests, or parklands.

Savanna formations occur on all continents of the world (Fig. 2), with the largest extent found in the wet–dry tropical regions of Africa, South America, and Australia. Smaller areas occur in Asia, including Sri Lanka, Thailand, Vietnam, and Papua New Guinea.

☆*Change History*: April 2018. Lindsay B. Hutley, Samantha A Setterfield updated (1) minor refinements to text, expression; (2) replacement of Fig. 5 which depicts the non-linear relationship of rainfall with savanna tree basal area for Africa. A more recent study of Lehmann et al 2014 published in the journal Science has now been included as Fig 5. In this new Figure, the same relationship is explored but now for savanna across the 3 major savanna regions of Africa, South America and north Australia. This provides a more comprehensive assessment of drivers of savanna productivity; (3) revised section on Threats to savanna ecosystems that now includes more of a focus on climate change impacts and concepts of differential tree and C4 grass response to CO_2 fertilization, impacts on flammability. Also mentioned is the increase in savanna clearing and land use change, particularly in South America (e.g., the cerrado); and (4) two changes to the recommended reading list.

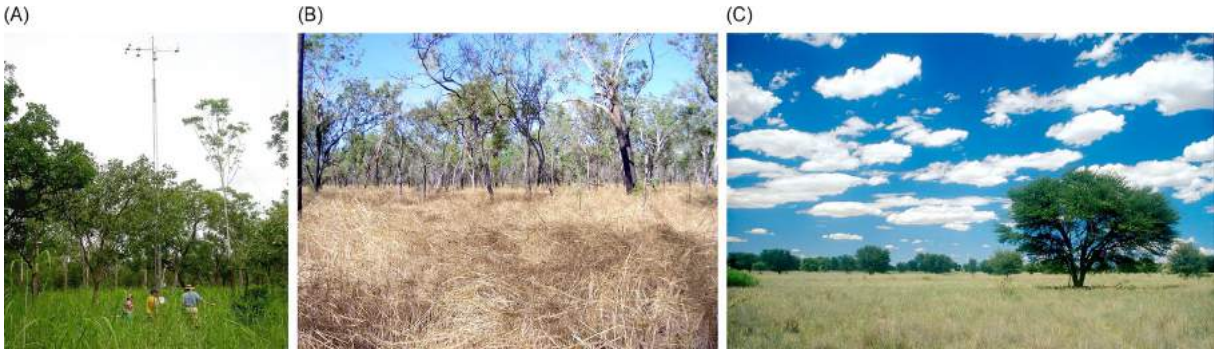


Fig. 1 Savanna ecosystems of the world, featuring the coexistence of a discontinuous woody overstorey with a continuous herbaceous understorey. Plates A and B are of a north Australian savanna site that receives approximately 1100 mm rainfall and is dominated by evergreen trees (*Eucalyptus* sp.) and tall C4 tropical grasses (*Sarga* spp.). Canopy fullness and grass growth are significantly differently in the wet (A) and dry (B) seasons. Tower-mounted instrumentation in Plate A is monitoring ecosystem productivity and water use over wet and dry seasons. Plate C, African savanna of the Kalahari Gemsbok National Park, Botswana. Photo courtesy of Joerg Tews.

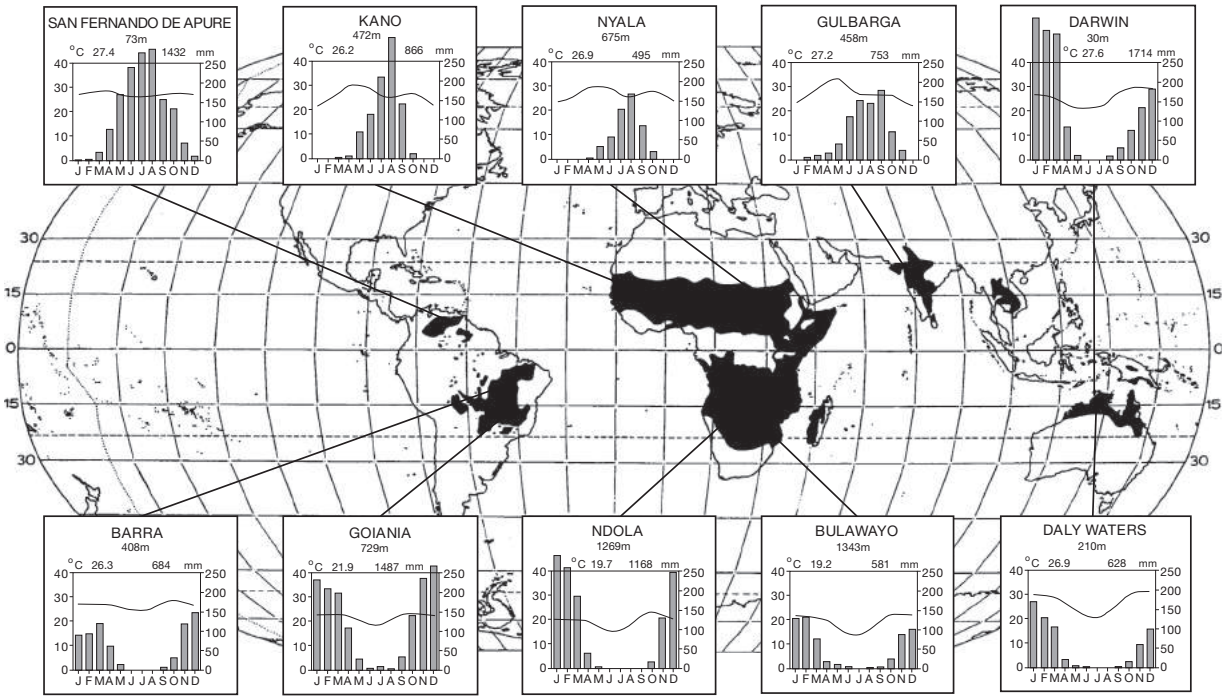


Fig. 2 The distribution of the world's savannas. Temperature and monthly rainfall data for a range of savannas are also given, with highly seasonal rainfall clearly evident.

Savanna also occurs in India, although these tree and grass systems tend to be derived from dry deciduous forest and subhumid deciduous forest due to land use changes and population pressure. Tropical savanna occupies an area of approximately 27.6 Mkm^2. Tree–grass mixtures also occur in temperate regions, in North America (Florida, TX), Mediterranean Europe, and Russia, although these temperate savannas are far smaller in extent at approximately 5 Mkm^2. In total, the savanna biome occupies 1/5th of the global land area and supports a large and growing human population.

The existence of a dry season is a defining feature of savannas; rainfall is seasonal and ranges from 300 to 2000 mm, with a dry season lasting between 2 and 9 months of the year. There can be a single, extended dry season or several shorter dry periods. Interannual variation of rainfall is typically high, as is the commencement and cessation of the wet season and growing season length, making cropping in savanna lands difficult. Indeed, historical rainfall plays an important role in determining the vegetation structure of a savanna. Seasonally available moisture dramatically influences plant productivity, which in turn determines the timing of available resources for savanna animals.

Given their wide biogeographic range, savannas occur on a number of soils types, typically oxisols, ultisols, entisols, and alfisols (using US Soil Taxonomy). In general, these soils are ancient and highly weathered, low in organic matter and cation exchange capacity (CEC). Oxisols occur in tropical savanna regions of South America and central and eastern African savanna and consist of

highly weathered, transported, and deposited material occurring on fluvial terraces. Extensive weathering of primary minerals has occurred and they are dominated by clay minerals such as kaolinite and gibbsite which have low CEC. Also present in the soil are acidic Fe and Al sesquioxides, which limits nutrient availability, especially phosphorus. Savanna soils tend to be sands to sandy loams, deep and well drained but with low soil moisture holding capacity. Entisols that occur in Australian savanna also feature the occurrence of ferruginous gravels, further reducing water and nutrient holding capacity. Bioturbation by earthworms, termites, and to a lesser extent ants are critical in the cycling nutrients through the poor soil systems. Termites essentially act as primary consumers and in savannas that lack a significant herbivore biomass (e.g., Australian and some South American savannas), they have an ecological function similar to that of herbivorous mammals.

Savannas of Australia, Africa, and South America

Tropical savanna is the predominant vegetation type across the northern quarter of Australia where rainfall is above 500 mm per year, an area of ~2 million km^2 (Figs. 1A, B and 2). These savannas are open-woodlands and open-forests, with tree cover declining as rainfall decreases with distance from the northern coast. The overstorey flora is typically dominated by *Eucalyptus*, *Corymbia* species. *Melaleuca and Eucalyptus* assemblages occur in the drier regions of this biome where annual rainfall <1000 mm year^{-1}. The ground layer is dominated by annual and perennial grasses from the *Sorghum*, *Heteropogon*, and *Schizachyrium* genera. A variety of other tall grasses (>1 m height) dominate the ground layer of the monsoonal savannas, which extend from Western Australia to the Cape York Peninsula in Queensland. *Heteropogon contortus* (black speargrass) dominates the tropical savanna understory in eastern Queensland, with *Themeda triandra*, *Aristida*, *Bothriochloa*, and *Chrysopogon bladhii* becoming more dominant as rainfall declines. *Acacia*-dominated savanna communities include extensive areas of brigalow (*A. harpophylla*), lancewood (*A. shirleyi*), and gidgee (*A. cambagei* and *A. georginae*).

The neotropical savannas of South America cover more than 2 million km^2. The Brazilian *cerrado* and the Colombian and Venezuelan *llanos* are a continuous formation, interrupted by narrow gallery forests. The *cerradão* includes a range of vegetation formations from the pure or almost pure grassland of *camp limpo*, to open-woodland with scattered tree cover of *campo cerrado*. These savanna can grade into denser woodland or open-forests, the *cerradão*, where tree cover is greater than 50%. The dominant grasses are *Andropogon*, *Aristida*, *Paspalum*, and *Trachypogon*. The Orinoco *llanos* of are grasslands or with scattered trees typically less than 8 m tall. Common trees include *Byrsonima* spp. *Curatella americana*, *Bowdichia virgioides*, and grasses include *Trachypogon* and *Andropogon*. Hyperseasonally flooded savannas and esteros (savanna wetland) occur in Brazil and Bolivia. Other savanna types, such as savanna parkland and mixed woodland occur through tropical America.

The African savannas occur across a range of soil types within a rainfall range of 200–1800 mm. One of the most extensive savanna areas is the *miombo* which covers about 2.7 million km^2 across central and southern Africa. The *miombo* is characterized by a discontinuous canopy of 10–12 m tall deciduous species of *Brachystegia*, *Isoberlinia*, and *Julbernardia*, with an herbaceous layer of tall grasses including mainly *Andropogon* species. In southern Africa, fine-leaved savannas, dominated by *Acacia* species, occur over fertile soils in low-lying, semiarid (250–650 mm/year) areas. Broadleaved savannas, including *Burkea africana*, *Combretum* spp., *Brachystegia*, occur on weathered, infertile soils. The northern *Sudanian* savannas have scattered deciduous trees, typically *Isoberlinia doka*, over xerophytic grasslands. These are bordered on the north by the drier *Sahelian* savannas and on the south by the wetter *Guinea*-type savannas. The arid and semiarid east African savannas are grasslands (*Aristida* spp., *Brachiara* spp.) with scattered shrubs or trees (including *Acacia*, *Grewia*, *Commiphora*), for example, the Serengeti.

Savannas occur throughout Asia, although many of these are derived from human disturbance. Savanna is fairly extensive in the Indian subcontinent although tree clearing has increased their extent, and many areas have been converted to agriculture. The most significant and widespread savanna type in Southeast Asia is the dry dipterocarp forest, which occurs in Vietnam, Laos, Cambodia, Thailand, Burma, and a small area in India. The region receives 1000–1500 mm rain per year and are dominated by the deciduous *Dipterocarpus* species, which can grow to 20 m, over a dense grass and herb layer including *Imperata cylindrical*, *Apluda mutica*, and *Apluda mutica* (pygmy bamboo).

Adaptive Traits of Savanna Vegetation

Savanna plants display a suite of traits to cope with seasonal drought, low water and nutrient availability, and the impacts of regular fire and herbivory. Adaptive traits which aid in the survival of fire of woody plants include thick insulating bark, high wood moisture content, elevated and well-separated crowns, and significant resprouting capacity. Resprouting can occur via lignotubers and from other underground and stem basal tissues following the death of aerial stems. This enables recovery with minimal developmental costs. Vegetative reproduction from roots, rhizomes, or stolons is dominant in much of the savanna biome. Adaptations to low nutrient availability include root mycorrhizal associations, particularly of ectomychorrhizae. Savanna trees can rapidly translocate sequestered nutrients from the leaves to other tissues (e.g., bark) prior to leaf fall. Woody savanna plants often have thorns that restrict grazing, as well as chemical features, such as tannins making leaves less palatable. Savanna grasses also display morphological features, such as serrated edges and chemical features, including tannins and silica bodies to restrict grazing.

The herbaceous grass layer is dominated by grasses with the C4 photosynthetic pathway. This pathway enables high photosynthetic rates at high temperatures and irradiance and low water availability. Most savanna trees and shrubs have the C3 photosynthetic pathway that has a higher efficiency under low light when compared to the C4 pathway; a characteristic which facilitates

recruitment and establishment under shaded tree canopies. The growth of savanna plants tends to occur mostly during the wet season with senescence or dormancy in the fire prone dry season, a trait that facilitates persistence in unfavorable conditions. Annual herbaceous species persist via a soil seed bank whereas above-ground parts of perennial herbaceous species die during dry periods, with dormant, regenerative buds protected within below-ground rhizomes or by cataphylls. Some annual grass species use hygroscopically active awns and pointed calluses on their seed that enables penetration into the surface soil protecting them from fire. Perennial herbaceous species require wet season rains to produce their first green shoots as carbohydrate storage from the previous wet season is limited. Rainfall stimulates germination of annual herbaceous and grass species and the early wet season is a period of rapid growth. Most herbaceous species flower in the wet season, although in contrast, many woody species flower in the dry season.

Woody species have evolved physiological and morphological mechanisms to either tolerate (evergreen habit) or avoid (deciduous habit) prolonged periods of water stress. Deep-rooting woody plants (usually evergreen) can access water resources throughout the year and provides them with their full photosynthetic capacity when favorable conditions occur. Deciduous species rehydrate stems prior to onset of wet season rains, which is then followed by leaf expansion to maximize photosynthetic activity during the wet season. Deciduousness and evergreeness represent extremes of physiological adaptations to survive the seasonal savanna climate. Evergreen species invest more resources in longer lived leaves, whereas deciduous species tend to support shorter lived leaves with high leaf photosynthetic capacity. Deciduous species need to acquire enough nutrient and photosynthate to ensure persistence and reproduction during the wet season, whereas evergreen species tend to have slower growth rates but persist throughout the seasonal cycle. Evergreeness also allows opportunistic acquisition of resources when soil nutrients are severely limiting and the cost of producing new leaves to respond to change in soil moisture is prohibitive.

Although this section has described broad seasonal growth patterns, it is important to note that the world's savanna plants include a high diversity of species and life-forms, with many distinct phenological patterns. All periods of the climatic cycle is favorable to certain vegetative or flowering phenophases in at least one group of species.

Environmental Factors Determining Savanna Structure

The adaptive traits described above enable individual plant survival in seasonally variable climates, but what environmental factors operate at a landscape or regional scale that determine savanna structure? Evidence suggests that four key environmental factors are responsible; (1) plant available moisture (PAM), (2) plant available nutrients (PAN), (3) fire regime (frequency, severity), and (4) herbivory. Herbivores include vertebrate and invertebrates and consist of both browsers consuming woody biomass and grazers consuming grasses and herbs. The overarching determinants of savanna physiognomy (relative abundance of the tree and grass layer) are climate and soil type (PAM and PAN), which determines the potential growth and survival of trees and grasses at a given site. Growth potential is moderated by disturbance agents, fire, herbivory, and stochastic events (such as cyclones). These factors act in concert to influence both competitive interactions and facilitation of tree and grass growth and determine savanna structure, floristics, and productivity (Fig. 3). The interaction of these factors is poorly understood and their variation in space and time make experimental testing and isolation of any single determinant difficult. Spatial heterogeneity of vegetation due local site histories (determined by antecedent rainfall, fire history, and herbivore numbers) and an inability to quantify these factors exacerbates this difficulty.

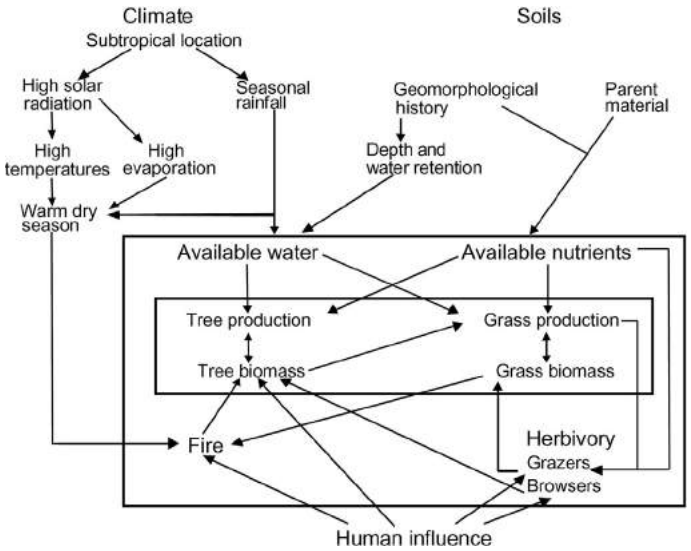


Fig. 3 Interactions between the environmental determinants of savanna structure. The relative tree and grass biomass and productivity is determined by available water, nutrient, and disturbance regime (fire and herbivory). These determinants are in turn characterized by climate and soil type for any given location. Reproduced with permission from House et al. (2003). *Journal of Biogeography* **30**, 1763–1777 (Blackwell Publishing Ltd., Oxford).

Available Moisture and Nutrients

Savannas tend occur in warm climates with an annual drought, with soils typically low in nutrient capital and poor water holding capacity. Trees and grass interactions are dominated by competition for water and nutrients, rather than light or growing space. At a broad, continental scale, PAM is the most significant of the four ecological determinants, with increasing rainfall correlated with increased tree cover and in general, a decreased grass biomass. PAM can be quantified via a range of parameters, from simple measures such as annual rainfall or via water balance parameters (rainfall as a fraction of potential or actual evapotranspiration) or soil characteristics (water release characteristics, soil storage capacity). At fine spatial scales, soil physiochemical properties (PAN) have a more significant influence and the interaction with PAM is often termed the PAM/PAN plane. Nutrient availability is largely a function of soil moisture and dry season nutrient uptake and nitrogen mineralization in particular, is limited by low levels of PAM. Significant plant growth is only possible during periods of high PAM that releases available nutrient via mineralization. Soils of semiarid savanna can have a higher intrinsic fertility when compared to highly, leached soils of mesic sites, but this nutrient capital is only available for uptake during moist periods. Savanna vegetation receiving similar rainfall can exhibit contrasting structure and floristics, simply due to fine-scale changes in soil type. A good example of this interaction comes from the long-term savanna research site of Nylsvley in South Africa; here, nutrient-poor, broad-leafed savanna, dominated by *Burkea africana*, surround patches of nutrient-rich soil that support a very different savanna type, a fine-leafed savanna dominated by *Acacia tortilis*. Both savanna types experience the same climate, but differences in soil parent material result in higher levels of soil available N and P in the fine-leafed patches. Productivity of the fine-leafed savanna is approximately double that of the broad-leafed system and attracts a larger grazing and browsing fauna. Similarly in South American savanna, soil acidity and aluminum levels significantly affect structure and floristics independent of rainfall.

Fire

Fire is an important landscape-scale determinant that impacts all of the world's savanna. Fire is an inevitable consequence of the annual cycle of profuse herbaceous production during the wet season followed by curing of this material in the dry season, when climatic conditions are ideal for burning. Savanna fires are virtually all surface fires, consuming the highly flammable herbaceous layer. Crown fires rarely occur as the fuel loads and dominated by fine fuels (grass and leaf litter) and the foliage of savanna trees and shrubs tend to be of low flammability. Human ignitions largely control fire behavior and extent in savanna, with some fires started by dry lightning strikes. Savanna fires spread rapidly through the surface fuels and high soil temperatures do not persist for longer than a few seconds to minutes. While these fires have a significant impact on above-ground plant parts, there is limited impact on savanna seed banks or above- and below-ground regenerative plant parts.

Fire has a major role in restricting tree establishment and growth, as evident from long-term fire exclusion plots (>25 years) in southern African and north Australian savanna (Fig. 4), which have resulted in a woody thickening. Frequent fire events can reduce tree seedling establishment and the ability of saplings to escape the flame zone via height growth. This limitation on tree establishment enables grass persistence and growth, maintaining the fuel load. The aerial stems of small seedlings and suckers are often killed during fire but the individuals are able to resprout from lignotubers or from other underground and stem basal tissues. Seedlings less than 6 months old have been observed to resprout in some species (e.g., *Eucalyptus miniata*) and frequent fire in the savannas will kill or maintain tree seedlings as a suppressed woody sprout layer until there is a sufficient fire-free period for them to "escape" the fire damage zone. Species can survive for at least 40 years as suppressed sprouts, during which time they develop significant lignotubers which aid in rapid growth during fire-free periods.

Fig. 4 Impacts of overgrazing and fire on savanna structure. Plate A is an overgrazed native grass paddock in semiarid savanna in north Australia (Kidman Springs Station, Victoria Rivers District, NT) at the end of the wet season of 1973. This site would be subjected to wind and water erosion, resulting in further decline in health and productivity of such sites. Exclusion of grazing and fire (Plate B) has resulted in a complete recovery of structure and function, with return of trees and grasses stabilizing soil surfaces, increased water capture, and a recovery in nutrient availability and cycling. Photos courtesy of John Ludwig, CSIRO.

The timing of fires in relation to reproductive phenology can constrain or promote plant reproduction. Studies on the woody species in the Brazilian *cerrado* and mesic Australian savannas have indicated that frequent fire can reduce seed production and sexual recruitment and could cause a shift in species composition, favoring vegetatively reproducing species. However, fire is also important for the sexual regeneration of some species, as burning induces flowering and fruit dehiscence in many *cerrado* species and facilitates pollination in others. Most perennial grass species are generally unaffected by burning and regenerate from basal leaf sheaths protected underground. Some perennial (e.g., *Trachypogon plumosus*) and annual (e.g., *Andropogon brevifolius*) grass species decrease in abundance after a long-term absence of fire.

Prior to human occupation and use of fire in savannas, lightning would have been the dominant source of ignition and it is likely that extensive but infrequent fires would have occurred. In Australia, humans have intentionally used fire for at least 45,000 years and in Africa for potentially 1 million years or more. Large proportions of savanna regions are burnt each year for a variety of reasons; land clearing, livestock management, property protection, conservation management, and cultural purposes. In African savannas, fires burn between 25% and 50% of the arid "Sudan Zone" and 60%–80% of the humid "Guinea Zone" each year. In the high rainfall (>1000 mm annual rainfall), ~50% of *Eucalyptus* dominated open-forest savanna is burnt in northern Australia. With the progression of the dry season, fire intensity increases due to fuel accumulation from curing litterfall and grass senescence, resulting in an increased combustibility of fuels. This and more severe fire weather (i.e., higher temperatures, wind speeds, and lower humidity) result in an increased landscape flammability. Early dry season fires (when fuel accumulation is low and curing incomplete) tend to be low intensity, patchy, and limited in extent whereas fires later in the season are of higher intensity and produce more extensive and homogeneous burning. Impacts on vegetation depend on fire intensity, distribution, and timing (fire regime) in relation to the vegetative and phenological cycles. Determining direct effects of fire on savannas is often difficult due to confounding effects of herbivory. Nevertheless, long-term burning experiments have shown that the higher intensity, late dry-season fires are the most damaging to woody species.

Herbivory

Two common images of savannas are herbivory by large, native ungulates, particularly in Africa and the widespread grazing by domestic herds, particularly cattle. A more neglected group of savanna herbivores are the invertebrates, particularly grasshoppers, caterpillars, ants, and termites. Mammal herbivores are typically categorized as grazers, browsers, or mixed feeders, who can vary their diet depending on food availability. Mammal and insect herbivores impact on savanna structure and function via consumption of biomass, seed predation, trampling of understory, and the pushing over and killing of trees and shrubs. The importance of herbivory as a determinant varies between savanna regions, and appears to largely reflect the abundance of large herbivores present. Large herbivore diversity and abundance are much higher in Africa than in Australia, Asia, or South America. More than 40 large wild herbivore species have been described in African savanna. In contrast, only six species of megapod marsupial have been considered as large herbivorous mammals in the Australian savannas, and only three species of ungulates are regarded as native South American savanna inhabitants. Domestic animals, particularly cattle, buffalos, sheep, and goats, are now the dominant, large herbivores in most savannas.

Large herbivores can lead to changes in species composition, woody vegetation density, and soil structure. For example, grazing pressure in Africa and Australia has led to a decrease in palatable, perennial, grazing-sensitive tussock grasses, and an increase in less palatable perennial and annual grass and forb species. Changes to the soil surface can occur, including loss of crusts (important in nutrient cycling), development of scalds, compaction, increased runoff, soil erosion, and nutrient loss. In parts of Africa, woody vegetation density has sometimes been reduced by large herbivores, for example uprooting of trees by elephants when browsing. Browsers such as giraffes can reduce woody seedling and sapling growth thereby keeping them within a fire-sensitive height for decades. By contrast, in many of the world's savannas the density of woody vegetation has increased at the expense of herbaceous vegetation; one of the major causes has been high rates of herbivory. A decrease in grass biomass following grazing leads to a reduction fuel and thus fire frequency and intensity, enhancing the survival of saplings and adult tress. Fire also affects herbivory as herbivores may favor postfire vegetation regrowth. Clearly, fire and herbivory have an interactive effect on savanna structure and function.

While less spectacular than large browsers and grazers, insects are often the dominant group of herbivores in savannas, especially on infertile soils supporting low mammal biomass. There is a paucity of data describing their abundance or role in these ecosystems. In a broad-leaved, low fertility savanna of southern Africa, a grasshopper biomass of 0.73 kg ha^{-1} can consume almost 100 kg ha^{-1} of plant material and damage an additional 36 kg ha^{-1}. This represents a loss of 16% of above-ground grass production. Grasshoppers and caterpillars can account for up to half the grass herbivory, although the rate and proportion varies substantially between years. Fertile, fine-leaved savannas are able to support a larger mammal biomass and the proportion of herbivory resulting from insect consumption is lower when compared to infertile African sites. The impact of insect herbivores on physiognomy has not been established but they are clearly important herbivores in savannas through their impact on productivity and ecosystem properties.

Conceptual Models of Tree and Grass Coexistence

Interactions between the coexisting life-forms in savanna communities are complex and over the last 40 years, a range of conceptual or theoretical models has been proposed to explain tree and grass mixtures. Contrasting models have all been supported by empirical evidence for particular sites, but no single model has emerged that provides a generic mechanism explaining coexistence. Models can be classified into several categories. Competition-based models feature spatial and temporal separation of resource usage by trees and grasses that minimizes competition and enables the persistence of both life-forms. Alternatively, demographic-based models have been described, where mixtures are maintained by disturbance, resulting in bottlenecks in tree recruitment and/or limitations to tree growth and grasses can persist. Table 1 provides a summary of these models. Root–niche separation models suggest there is a spatial separation of tree and grass root systems, with grasses exploiting upper soil horizons and trees developing deeper root systems. Trees rely on excess moisture (and nutrient) draining from surface horizons to deeper soil layers. Phenological separation models invoke differences in the timing of growth between trees and grasses. Leaf canopy development and growth in many savanna trees occurs prior to the onset of the wet season, often before grasses have germinated or initiated leaf development. As a result, trees can have exclusive access to resources at the beginning of the growing season, with grasses more competitive during the growing season proper. Given their deeper root systems, tree growth persists longer into the dry season, providing an additional period of resource acquisition at a time when grasses may be senescing. This spatial and temporal separation of resource usage is thought to minimize competition, enabling coexistence. Other competition models suggest that trees density becomes self-limiting at a threshold of PAM and PAN and are thus unable to completely exclude grasses. These models assume high rainfall years favor tree growth and recruitment, with poor years favoring grasses and high interannual variability of rainfall maintaining a relatively stable equilibrium of trees and grasses over time.

Alternatively, savannas can be viewed as metastable ecosystems (narrow range of stabile states) with a dynamic structure over time. Demographic-based models suggest that determinants of tree demographics and recruitment processes ultimately set the tree–grass ratio (Table 1). Fire, herbivory, and climatic variability are fundamental drivers of tree recruitment and growth, with high levels of disturbance resulting in demographic bottlenecks that constrain recruitment and/or growth of woody components and grass persistence results. At high rainfall sites, in the absence of disturbance, the ecosystem tends toward forest. High levels of disturbance, particularly fire, can push the ecosystem toward a more open canopy or grassland; this ecosystem trajectory is more likely at low rainfall sites.

There is observational and experimental data to support all of the above models and it is highly likely that savanna structure and function results from the interaction of all processes. In many savannas, root distribution is spatially separated with mature trees exploiting deeper soil horizons as the competitive root–niche separation model predicts. Root partitioning favors tree growth in semiarid systems where rainfall occurs during periods when grass growth is dormant; rainfall can drain to deep layers supporting tree components. By contrast, in semiarid savanna where rainfall and growing seasons coincide, investment in deep root systems could result in tree water stress, as rainfall events tend to be sporadic and small in nature, with little deep drainage. In this case, surface roots are more effective at exploiting moisture and mineralized nutrients following these discrete events. In these savannas, tree and grass competition for water and nutrients would be intense. In mesic savanna sites, root competition between both trees and grass roots in upper soil layers is apparent, contrary to predictions of niche-separation models. Mesic savannas of north Australia (rainfall >1000 mm) are dominated by evergreen *Eucalyptus* tree species and during the wet season these trees compete with high growth-rate C4 grasses for water and nutrients in upper soil layers (0–30 cm). However, by the late dry season, tree root

Table 1 Conceptual models explaining the coexistence of trees and grasses in savanna ecosystems in equilibrium (tree:grass ratio relatively stable at a given site), nonequilibrium (tree:grass ratio variable), or disequilibrium (disturbance agents essential for the maintenance of tree:grass coexistence)

Competition-based mechanisms of coexistence	*Demographic-based mechanisms of coexistence*
Spatial and temporal niche separation of resource usage enables both life-forms to coexist	Climatic variation and disturbance impacts on tree demography Extremes of climate and disturbance influence tree germination and/or establishment and/or transition to mature size classes enabling coexistence
Root–niche separation	Tree and grasses exploit deep and shallow soil horizons
Phenological separation	At low rainfall sites, tree establishment and growth occurs only in above average rainfall periods
Temporal differences in leaf expansion and growth, trees have exclusive access to resources at beginning and end of growing season, grasses competitive during growing season	At high rainfall sites, high fuel production maintains frequent fire to limit tree dominance
Balanced competition	Trees are the superior competitor but become self-limiting for a given rainfall and unable to exclude grasses
Competition–colonization	Rainfall variability results in a tradeoff between tree and grass competition and colonization potential. Higher than mean rainfall favors tree growth, lower than mean favors grasses
Primary determinants PAM, PAN	Primary determinants PAM variability, PAN, fire regime, herbivory
Secondary determinants	Fire regime, herbivory

activity has shifted to subsoil layers (up to 5 m depth) and herbaceous species have either senesced or are physiologically dormant. These root dynamics suggest grasses are essentially drought avoiders but are able to compete with trees during the wet season. This system serves as an example of where both root–niche and phenological separation is occurring.

Tree to tree competition is also significant, as suggested by the strong relationship observed in most savanna regions between annual rainfall and indices of tree abundance, be it tree cover or tree basal area (area occupied by tree stems, Fig. 5). As PAM decreases, tree abundance declines, although there is significant variability at any given rainfall driven by differences in soil type, fire regime, herbivory, and/or land use. Competition models fail to consider impacts of savanna determinants on different demographics of a population, such as recruitment, seedling establishment, and tree sapling growth. Root–niche or phenological separation models largely consider impacts acting on mature individuals, whereas demographic models include impacts of climate variability and disturbance on critical life-history stages (e.g., seedling establishment and accession to fire-tolerant size classes). Demographic models assume that savanna tree dynamics are central to savanna ecosystem functioning and that savanna trees are the superior competitors under most conditions; grass persistence only occurs when determinants act to limit tree abundance. It is clear that competition, both within and between savanna life-forms, occurs and that tree abundance is moderated by climate

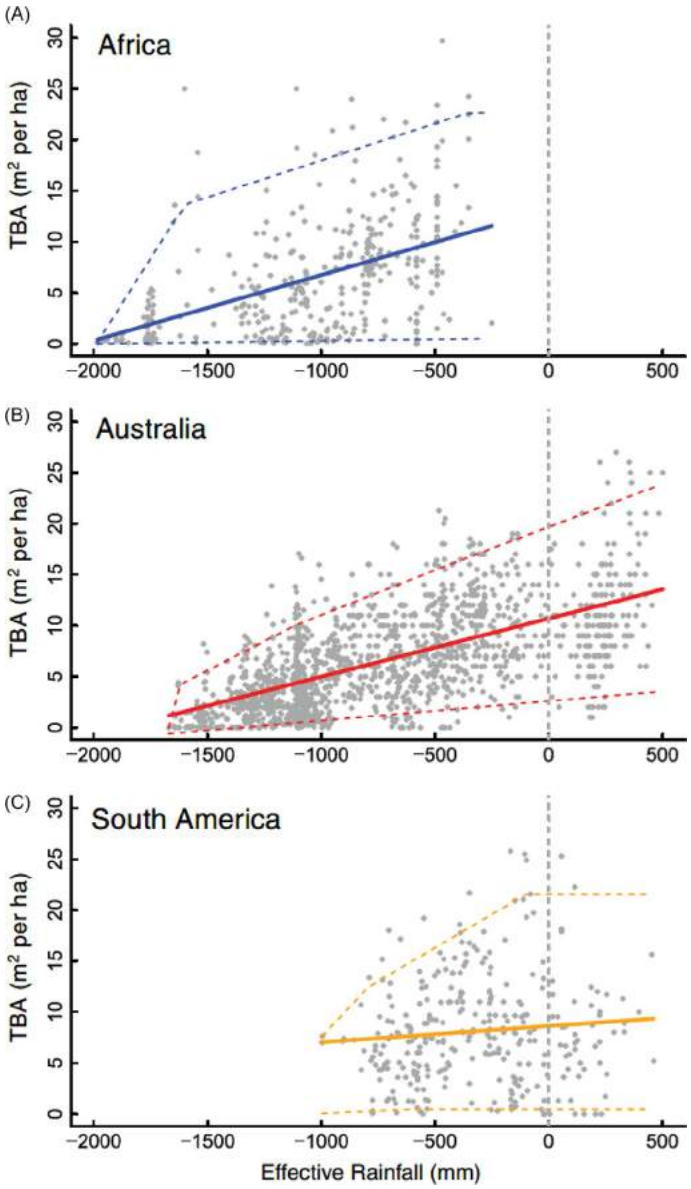


Fig. 5 Relationship between effective rainfall (rainfall–potential evapotranspiration) and tree basal area for African, Australian, and South American savannas. Rainfall sets a maximal climate-determined woody cover with other factors such as available nutrient, fire regime, and herbivory also determining woody cover at any given site. From Lehmann et al. (2014). *Science* **343**, 548–552 (American Association for the Advancement of Science), with permission.

variability and disturbance. A more comprehensive model would integrate both competition and demographic theories to yield a model in which competitive effects are considered for each life-history stage.

The complexity inherent in these models is evident when savanna structure is correlated with any of the environmental determinants. Fig. 5 describes the relationship between tree basal area and effective rainfall, which is calculated as rainfall–potential evapotranspiration, and is a useful surrogate for PAM. Tree basal area data are shown for African, Australian, and South American savanna with a large range of basal are possible at any given effective rainfall, especially for African sites. Rainfall sets an upper limit on tree basal area, with points below this limit determined PAM as well as PAN (soil type) and fire regime playing a significant role. At semiarid savanna sites (<650 mm rainfall), it is likely that rainfall limits tree cover and canopy closure, permitting grass coexistence. At rainfalls >650 mm, tree canopy closure may be possible, with disturbance limiting tree dominance. For African and Australian savanna, the relationship (slope) is similar, with basal area increasing with effective rainfall. This relationship is far weaker for the wetter South American savannas, suggesting other factors such as soil characteristics are playing a more significant role in determining productivity.

Savanna Biomass and Productivity

Global net primary production (NPP), the net production of plant biomass is approximately 67.6 Gt C $year^{-1}$ of which almost 30% occurs in savanna ecosystems (19.9 Gt $year^{-1}$). This production occurs on 18% of the global land surface, demonstrating that savannas are relatively productive ecosystems but subjected to frequent disturbances, such as fire and herbivory that limits biomass accumulation. Mean savanna NPP has been estimated at 7.2 t C ha^{-1} $year^{-1}$ (Table 2), lower than typical values for the other major tropical ecosystem, rainforest, which ranges from 10 to 15 t C ha^{-1}. Savanna NPP and biomass varies by an order of magnitude (Table 2), as would be expected given their geographic range and structural variation. The relative production of trees versus grasses is also highly variable, but in general, NPP of the C4 grass layer is two to three times that of woody NPP. Biomass stored in above- and below-ground pools gives the root:shoot ratio and global biomass data gives a global mean of approximately 2 (Table 2). This reflects the investment in root systems and below-ground storage organs, such as lignotubers, to maintain uptake of moisture and nutrient from sandy, nutrient poor savanna soils and to survive annual drought and disturbance.

Savanna photosynthesis and growth is highly seasonal and interannual variability high. Mesic savanna may receive annual rainfall associated with rainforest ecosystems, yet productivity is significantly lower, due largely to annual drought, poor soils, and impacts of disturbance. Long-term (as opposed to annual) estimates of savanna productivity need to include loss of biomass due to fire and herbivory. Including fire and herbivory impacts in productivity estimates gives the carbon sequestration rate, which represents the net gain (sink) or loss of carbon from the ecosystem to the atmosphere. While wet season productivity can be very high in savannas, much of a wet-season's herbaceous productivity can be lost via fire or grazing. Woody biomass tends to be a less dynamic, longer-term carbon storage pool than the herbaceous components of savanna. Savanna fire results in a significant release of greenhouse gases, including CO_2, CO, methane, nonmethane hydrocarbons, nitrous oxide, particulate matter and aerosols, equivalent to 0.5–4.2 Gt C $year^{-1}$. Fire reduces net savanna sequestration rate by about 50% and protection of savannas from fire and grazing results in an increase in woody biomass which can result in a long term increase in stored soil carbon. Savanna sink strength in mesic Orinoco savannas in South America (~1500 mm annual rainfall) has been measured at 1 t C ha^{-1} $year^{-1}$, with this sink maintained over a 25 year period in plots with fire and grazing excluded. Similarly, the carbon sink strength of north Australian, *Eucalyptus* dominated savannas receiving approximately the same rainfall has also been estimated at approximately 0.5–1 t C ha^{-1} $year^{-1}$, with this sink measured at sites burnt but not grazed. This carbon is likely being stored in woody biomass and soil organic carbon pools, with a small fraction being stored as black carbon (charcoal), a resilient carbon pool. Savanna soil carbon storage is by far the largest terrestrial pool (Table 2) and soil carbon represents a longer-term carbon storage when compared

Table 2 Savanna biomass, soil carbon stocks, and productivity

Parameter	*Mean (sd)*	*Range*
Biomass and soil stocks (t C ha^{-1})		
Aboveground biomass	10.6 (9.0)	1.8–34
Belowground biomass	19.5 (14.9)	4.9–52
Total biomass	33.0 (22.9)	9.4–84
Root:shoot ratio	2.1 (2.0)	0.6–7.6
Soil organic carbon	174.2 (126.0)	18–373
Savanna area (Mkm^{-2})	27.6	
Total carbon pool (Gt C)	326	
Productivity (t C ha^{-1} $year^{-1}$)		
NPP	7.2 (5.1)	1.4–22.8
NEP	0.14	

Data from Grace, J., San, J.J., Meir, P., Miranda, H.S., Montes, R.A. (2006). Productivity and carbon fluxes of tropical savannas. *Journal of Biogeography* **33**, 387–400 (Blackwell Publishing Ltd., Oxford).

to the more dynamic vegetation components. Burning also influences nutrient dynamics via losses due to volatilization (vaporization) of lighter elements such as nitrogen and sulfur. At a global scale, savannas and tropical seasonally dry forests represent a significant source of N_2O to the atmosphere (4.4 Tg N_2O year^{-1}). Shifts to a more frequent fire regime may result in a net loss of nitrogen, significant as savanna are in general nitrogen poor. Many grass species are able to recover quickly after fire, with regrowth attractive to grazing animals, due to the relatively high nutrient content of the foliage.

Threats to Savanna Ecosystems

Savannas are the location of human evolution, and humans are an integral component of these ecosystems. Humans have influenced the determinants of savannas for thousands of years via modification to nutrient availability from fire and clearing for agriculture. Human cultures have used fire as a vegetation management tool and introduced animal husbandry systems, changing grazing, and browsing pressures, modifying tree–grass competitive balances (e.g., Fig. 4). Savannas now face a range of threats, many operating synergistically. A contemporary impact is now being experienced via climate change and its influence on rainfall distribution, temperature increases, climate conditions conducive to fire and increased atmospheric CO_2 concentration. Impacts of climate change will be complex. Elevated atmospheric CO_2 is likely to favor woody growth over grasses given the increased carbon allocation to roots and lignotubers and the higher water and light use efficiencies of the C3 photosynthetic pathway as utilized by woody components when compared to the response of C4 photosynthesis of grasses. Tree:grass balance will be shifted as tree saplings may grow to fire-tolerant sizes faster, limiting the impact of fires that maintain grasses in savanna. However, gains in productivity may be offset by enhanced flammability due to shifts in fire weather from higher temperatures, evaporation rates, and fuel curing. This impact will likely vary in different countries and regions due to the variation in dominance of different phenological strategies. An extended growing season have greater impact on deciduous species (with longer leaf retention) than it would on evergreen species.

Human usage of the savanna biome is increasing, which can lead to degradation of vegetation and soil resources, resulting in nutrient losses and shifts in water balance and availability. Conversion to agricultural production is resulting in the clearing of large areas of the world' savannas. For example, ~50% of the *cerradão* and *llanos* has now been cleared or altered for agricultural crops such as coffee, soybeans, rice, corn, and beans. In addition, an increasing threat is from afforestation, such as growing exotic eucalypt and acacia plantations in Brazil and Nigeria, palm oil in Colombia, and fruit plantations in China. Plantations have replaced large areas of cerrado with few remaining areas of natural vegetation in the eastern cerrado region. Alterations in grazing pressure and fire suppression in managed savannas has also resulted in woody dominance, which ultimately reduces grazing production, severely impacting communities relying on cattle-derived incomes and reducing local biodiversity. This thickening or woody encroachment is being observed in areas subjected to extensive grazing activities in both African and Australian savannas and may be enhanced by increasing atmospheric CO_2 concentration.

Clearing for alternative land uses can result in exotic species invasions, a problem for much of the world's savannas. African savanna, especially in South Africa, are being invaded by woody species, often *Acacia* or *Eucalyptus* species from Australia, introduced for fuel wood or timber production. Reduced herbivory of these species results in high growth rates and water use. The development of thickets reduces deep drainage, groundwater recharge, and streamflow, consequently affecting water supplies. In an attempt to increase the grazing potential of north Australian and South American savanna, fast growing African grasses such as *Andropogon gayanus*, *Melinis minutiflora*, *Panicum maximum*, and *Urochola* spp. have been introduced. They are often more productive than native species resulting in more flammable fuel loads. At infested sites in north Australia, resultant fire intensity is five times that observed from native grass savanna causing major declines in tree cover. This in turn will result in a demographic bottleneck, long-term loss in tree cover and the instigation of a grass–fire cycle.

All of the above examples involve human impacts acting on one or more of the determinants of savanna structure and function. Clearly increased knowledge of their interactions will provide improved understanding of savanna processes and enable better management in a rapidly changing world. Savannas may be ideal ecosystems for agro-forestry applications, rather than traditional cropping systems. Small shifts in fire regime may dramatically increase productivity, thus savanna systems could be used for carbon sequestration and greenhouse gas mitigation schemes, providing alternative livelihoods and aiding in the maintenance of biodiversity.

Further Reading

Andersen AN, Cook GD, and Williams RJ (2003) *Fire in tropical savannas: The Kapalga experiment*. New York: Springer.
Baruch Z (2005) Vegetation–environment relationships and classification of the seasonal savannas in Venezuela. *Flora* 200: 49–64.
Baudena M, Dekker SC, van Bodegom PM, Cuesta B, Higgins SI, Lehsten V, Reick CH, Rietkerk M, Scheiter S, Yin Z, Zavala MA, and Brovkin V (2015) Forests, savannas, and grasslands: Bridging the knowledge gap between ecology and dynamic global vegetation models. *Biogeosciences* 12: 1833–1848.
Furley PA (1999) The nature and diversity of neotropical savanna vegetation with particular reference to the Brazilian cerrados. *Global Ecology and Biogeography* 8: 223–241.
Grace J, San JJ, Meir P, Miranda HS, and Montes RA (2006) Productivity and carbon fluxes of tropical savannas. *Journal of Biogeography* 33: 387–400.
Higgins SI, Bond WJ, and Trollope WSW (2000) Fire, resprouting and variability: A recipe for grass–tree coexistence in savanna. *Journal of Ecology* 88: 213–229.
van Langevelde F, van de Vijver CADM, Kumar L, van de Koppel J, de Ridder N, van Andel J, Skidmore AK, Hearne JW, Stroosnijder L, Bond WJ, Prins HHT, and Rietkerk M (2003) Effects of fire and herbivory on the stability of savanna ecosystems. *Ecology* 84: 337–350.

Mistry J (2000) *World savanna: Ecology and human use.* Harlow: Prentice Hall.
Ratnam J, Bond WJ, Fensham RJ, Hoffmann WA, Archibald S, Lehmann CER, Anderson MT, Higgins SI, and Sankaran M (2011) When is a 'forest' a savanna, and why does it matter? *Global Ecology and Biogeography* 20: 653–660.
Rossiter NA, Setterfield SA, Douglas MM, and Hutley LB (2003) Testing the grass–fire cycle: Exotic grass invasion in the tropical savannas of northern Australia. *Diversity and Distributions* 9: 169–176.
Sankaran M, Hanan NP, Scholes RJ, Ratnam J, Augustine DJ, Cade BS, Gignoux J, Higgins SI, Le Roux X, Ludwig F, Ardo J, Banyikwa F, Bronn A, Bucini G, Caylor KK, Coughenour MB, Diouf A, Ekaya W, Feral CJ, February EC, Frost PGH, Hiernaux P, Hrabar H, Metzger KL, Prins HHT, Ringrose S, Sea W, Tews J, Worden J, and Zambatis N (2005) Determinants of woody cover in African savanna. *Nature* 438: 846–849.
Scholes RJ and Archer SR (1997) Tree and grass interactions in savanna. *Annual Review of Ecology and Systematics* 28: 517–544.
Scholes RJ and Walker BH (eds.) (1993) *An African savanna: Synthesis of the Nylsvley study.* Cambridge: Cambridge University Press.
Solbrig OT and Young MD (eds.) (1993) *The world's savannas: Economic driving forces, ecological constraints, and policy options for sustainable land use.* New York: Parthenon Publishing Group.
du Toit JT, Rogers KH, and Bigg HC (eds.) (2003) *The Kruger experience: Ecology and management of savanna heterogeneity.* Washington: Island Press.

Steppes and Prairies

JM Briggs, Arizona State University, Tempe, AZ, USA
AK Knapp, Colorado State University, Fort Collins, CO, USA
SL Collins, University of New Mexico, Albuquerque, NM, USA

Steppes and prairies (grasslands) are ecosystems that are dominated by grasses and to help understand grasslands, it is important to know something about grass morphology and growth forms. The remarkable ability of grasses to thrive in so many ecological settings and their resilience to disturbance is largely attributable to their growth form. Grasses are characterized by streamlined reduction and simplicity with tillers being the key adaptive structural element of the plant (Fig. 1). Tillers originate from growing parts (meristems) typically just near, at, or below the surface of the soil. The meristems that produce tillers are generally well protected by their location near or beneath the soil surface. It is the location of the meristem that explains much of the resilience of grasses and thus grasslands to disturbance.

Grass leaves are narrow and generally well-supplied with fibrous supporting tissue that has thick-walled cells. These features, along with a capacity to fold or roll the leaves along the vertical plane, permit the plant to endure periods of water stress without collapse. Another feature of grass leaves is the presence of siliceous deposits and silicified cells (phytoliths). Although silica is present in many plant families, phytoliths are characteristic of grasses. Phytoliths often have distinctive forms within taxonomic groups and since they persist in soil profiles for a very long time, they can be used by paleobotanists to determine shifts in dominance from one grass form to another. Silica also makes grass forage very abrasive and it is now generally accepted that the evolution of abrasion-resistant teeth present in many modern grazing animals was an evolutionary response to tooth-wearing effects of a diet high in grass. This also suggests that the grasses and their megaherbivore grazers are highly coevolved. But recent discovery of grass phytoliths in Late Cretaceous dinosaur coprolites in India suggest that grasses were already substantially differentiated and that abrasive phytoliths were present in many grasses before the explosion of grazers in the Oligocene and Miocene time periods.

Grasses show a very large variation in the way tillers are aggregated as they expand from their origin, but two general forms of grasses are recognized: bunch-forming (caespitose) and sod-forming (rhizomatous). This description captures the major features of the dominant grass species but there are some species and groups that deviate from this general pattern. The most obvious include the woody bamboos (some of which can reach tree size and for the most part are restricted to forest habitats in the tropics and subtropics).

In addition to growth form, grasses can also be roughly divided into two categories based upon their photosynthetic pathways: cool season (C_3) and warm season (C_4). C_4 photosynthesis is a variation on the typical C_3 pathway and is thought to have an advantage in high-light and -temperature environments typical of many grassland regions worldwide. Throughout the world today, tropical, subtropical, arid, semiarid, and mesic grasslands are typically dominated by C_4 grasses while in cooler high-elevation or northern climates, C_3 grasses are more common.

Grasslands

As mentioned above, ecosystems in which grasses and grass-like plants (including sedges and rushes and collectively known as graminoids) dominate the vegetation are termed grasslands. In its narrow sense, `grassland' may be defined as ground covered by vegetation dominated by grasses, with little or no tree cover. UNESCO defines grassland as ``land covered with herbaceous plants with less than 10 percent tree and shrub cover'' and wooded grassland as 10–40% tree and shrub cover. Grassland ecosystems are notable for two characteristics: they have properties that readily allow for agricultural exploitation through the management of domesticated plants or herbivores, and a climate that is quite variable both spatially and temporally. They are found in regions where drought is fairly common but where precipitation is sufficient for their growth. In addition, they can also dominate wetlands in both freshwater and coastal regions. They also occur in sites where more predictable rainfall occurs and soils are shallow or poorly drained, or in areas with topography too steep for woody plants. To put it simply, grasslands usually occupy that area between wetter areas dominated by woody plants and arid desert vegetation.

Grassland biomes occur on every continent except Antarctica. It is estimated that grasslands once covered as much as 25–40% of the Earth's land surface although much of the original extent of native grassland has been plowed and converted to other grass production (corn and wheat) or other row crops such as soybeans. Indeed, grasslands are important from both agronomic and ecological perspectives. Grasslands are the basis of an extensive livestock production industry in North America and elsewhere. In addition, grasslands sequester and retain large amounts of soil carbon and thus, they are an important component of the global carbon cycle.

Indeed, because grasslands store a significant amount of carbon in their soils and they contain relatively high biodiversity, then now play a prominent role in the discussion about biofuel production. Biofuels may offer a mechanism to generate energy that releases less carbon into the atmosphere. Some energy producers recommend intensive agricultural production of corn, or other grasses such as switchgrass or elephant grass for biofuel production. However, agricultural practices have significant energy costs that may reduce the value of these fuel sources. A recent study has suggested, however, that diverse prairie communities on

 doi:10.1016/B978-0-444-63768-0.00359-0

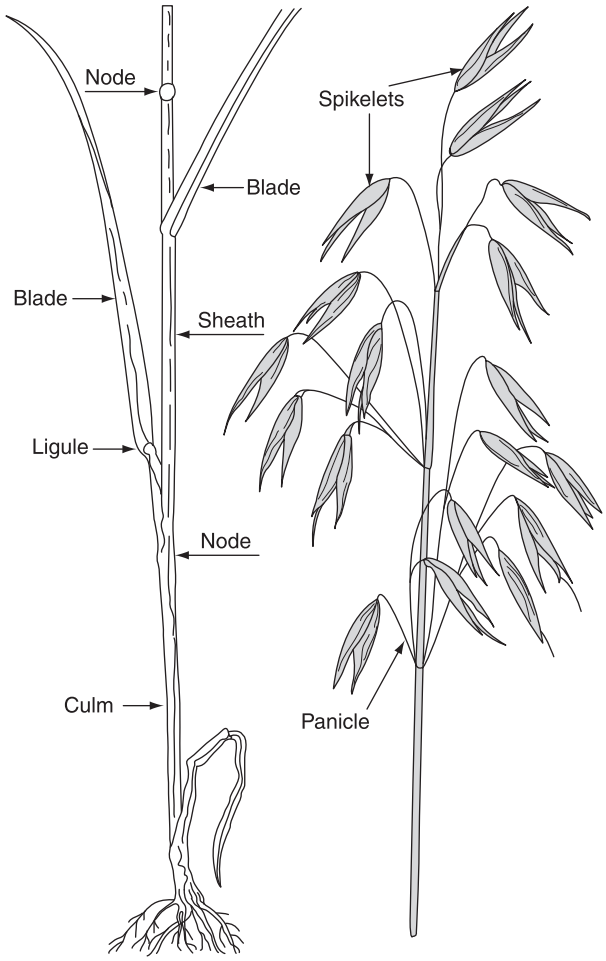


Fig. 1 Common oat, *Avena sativa*, × ½. From Hubbard (1984).

marginal lands are potentially `carbon negative' because they provide significant biomass for fuel and store carbon belowground. Much additional research is needed to assess the sustainability of grasslands for biofuel production, but the prospects are certainly tantalizing to energy producers and conservationists alike.

Grassland Types

It is estimated that prior to the European settlement of North America, the largest continuous grasslands in the United States stretched across the Great Plains from the Rocky Mountains and deserts of the Southwestern states to the Mississippi river. Other extensive grasslands are, or were, found in Europe, South America, Asia, and Africa (**Fig. 2**). Grasslands can be broadly categorized as temperate or tropical. Temperate grasslands have cold winters and warm to hot summers and often have deep fertile soils. Surprisingly, plant growth in temperate grasslands is often nutrient limited because much of the soil nitrogen is stored in forms unavailable for plant uptake. These nutrients, however, are made available to plants when plowing disrupts the structure of the soil. The combination of high soil fertility and relatively gentle topography made grasslands ideal candidates for conversion to crop production and thus have led to the demise of much of the grasslands across the world.

Grasslands in the Midwestern United States that receive the most rainfall (75–90 cm) are the most productive and are termed tallgrass prairies. Historically, these were most abundant in Iowa, Illinois, Minnesota, Missouri, and Kansas. The driest grasslands (25–35 cm of rainfall) and least productive are termed shortgrass prairie or steppe. These grasslands are common in Texas, Colorado, Wyoming, and New Mexico. Grasslands intermediate between these extremes are termed mid- or mixed grass prairies. In tallgrass prairie, the grasses may grow to 3 m tall in wet years. In shortgrass prairie, grasses seldom grow beyond 25 cm in height. In all temperate grasslands, production of root biomass belowground exceeds foliage production aboveground. Worldwide, other names for temperate grasslands include steppes throughout most of Europe and Asia, veld in Africa, *puszta* in Hungary, and the pampas in South America.

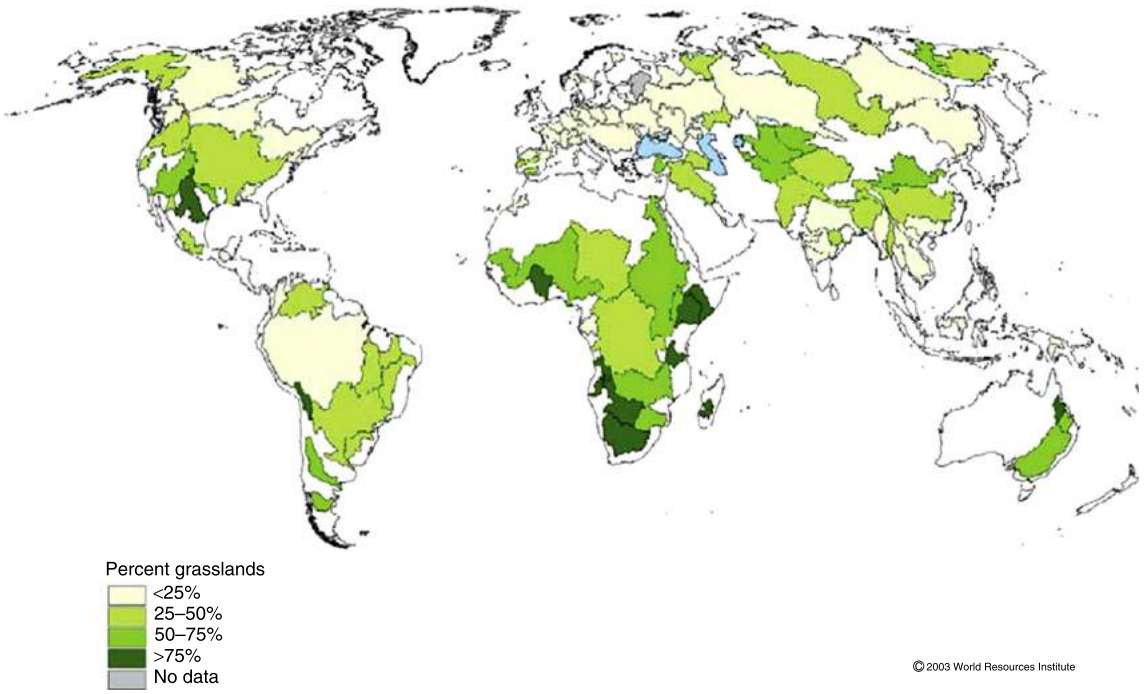


Fig. 2 Map of the grasslands of the world. World Resources Institute – PAGE, 2000. Sources: GLCCD, 1998. Loveland, T.R., Reed, B.C., Brown, J.F., et al., 1998. Development of a Global Land Cover Characteristics Database and IGBP DISCover from 1 km AVHRR Data. International Journal of Remote Sensing 21 (6–7), 1303–1330. Available online at http://edcaac.usgs.gov/glcc/glcc.html. Global Land Cover Characteristics Database, Version 1. Olson, J.S., 1994. In: Global Ecosystem Framework – Definitions. Sioux Falls, SD: USGS EDC, p. 39.

Tropical grasslands are warm throughout the year but have pronounced wet and dry seasons. Tropical grassland soils are often less fertile than temperate grassland soils, perhaps due to the high amount of rainfall (50–130 cm) that occurs during the wet season and washes (or leaches) nutrients out of the soil. Most tropical grasslands have a greater density of woody shrubs and trees than temperate grasslands. Some tropical grasslands can be more productive than temperate grasslands. However, other tropical grasslands grow on soils that are quite infertile or these grasslands are periodically stressed by seasonal flooding. As a result, their productivity is reduced and may be similar to that of temperate grasslands. As noted for temperate grasslands, root production belowground far exceeds foliage production in all tropical grasslands. Other names for tropical grasslands include velds in Africa, and the compos and llanos in South America.

Although temperate and tropical grasslands encompass the most extensive grass dominated ecosystems, grasses are present in most types of vegetation and regions of the world. Where grasses are locally dominant they may form desert grassland, Mediterranean grassland, subalpine and alpine grasslands (sometimes referred to as meadows or parks), and even coastal grassland. Most grasslands are dominated by perennial (long-lived) plants, but there are some annual grasslands in which the dominant species must reestablish each year by seed. Intensively managed, human-planted, and maintained grasslands (e.g., pastures, lawns) occur worldwide as well.

The Grassland Environment

Grassland climates can be described as wet or dry, hot or cold (typically in the same season), but on average are intermediate between the climates of deserts and forests. The climate of grasslands is best described as one of extremes. Average temperatures and yearly amounts of rainfall may not be much different from desert or forested areas, but dry periods during which the plants suffer from water stress occur in most years in both temperate and tropical grasslands. An excellent example of this comes from North America, where in the area around Washington, DC (dominated by eastern deciduous forest), the annual precipitation is ~102 cm whereas at Lawrence, KS (dominated historically by tallgrass prairie), the annual precipitation is ~100 cm. But the way the rainfall is distributed is notably different. At Lawrence, KS, over 60% of the rainfall occurs in the growing season (April–September), whereas at Washington, DC, the precipitation is uniformly distributed throughout the year. The open nature of grasslands is accompanied by the presence of sustained high wind speeds. Windy conditions increase the evaporation of water from grasslands and this increases water stress in the plants and animals. Another factor that increases water stress is the high input

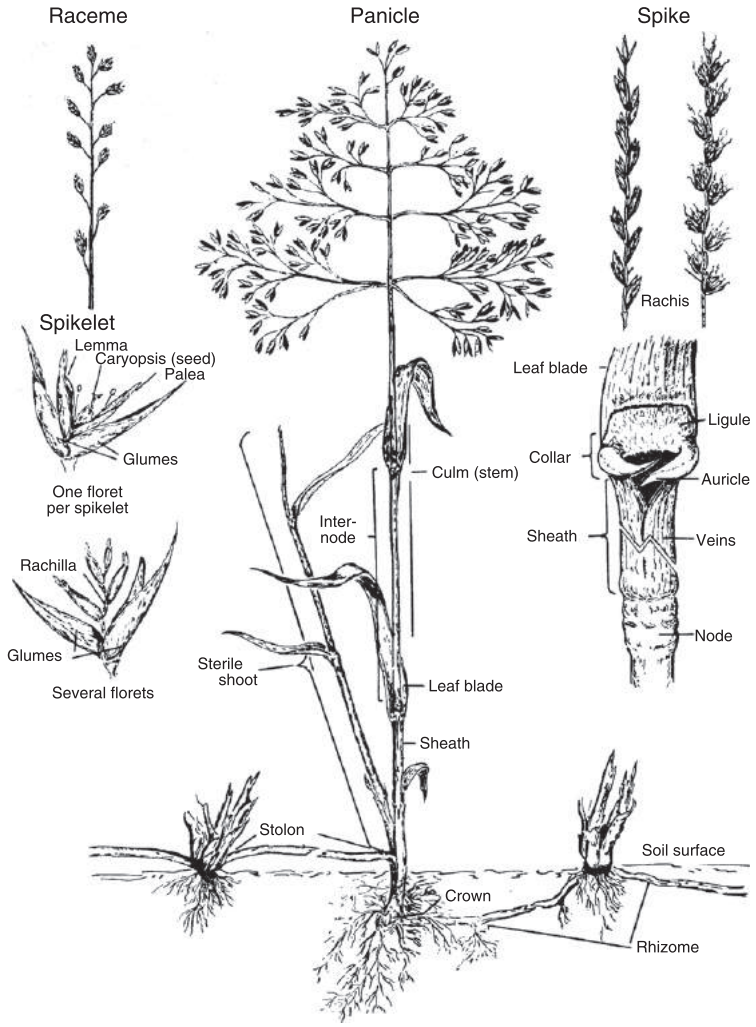


Fig. 3 Structure and architecture of the grass plant. From Ohlenbusch *et al.* (1983).

of solar radiation in these open ecosystems. This leads to the convective uplift of moist air and results in intense summer thunderstorms. Rain falling in these intense storms may not be effectively captured by the soil and the subsequent runoff of this water into streams reduces the moisture available to grassland plants and animals. In addition to periods of water stress within the growing season, consecutive years of extreme drought are more common in grassland than in adjacent forested areas. Such droughts may kill even mature trees, but the grasses and other grassland plants have extensive root systems and belowground buds that help them survive and grow after drought periods (Fig. 3).

Fire in Grasslands

It is generally recognized that climate, fire, and grazing are three primary factors that are responsible for the origin, maintenance, and structure of the most extensive natural grasslands. These factors are not always independent (i.e., grazing reduces standing crop biomass which can be viewed simply as a fuel for fire, and biomass is also highly dependent upon the amount of precipitation). Historically, fires were a frequent occurrence in most large grasslands. Most grasslands are not harmed by fire, many benefit from fire, and some depend on fire for their existence. When grasses are dormant, the moisture content of the senesced foliage is low and this fine-textured fuel ignites easily and burns rapidly. The characteristic high wind speeds and lack of natural fire breaks in grasslands allow fire to cover large areas quickly. Because fire moves rapidly and much of the fuel is above the ground, temperatures peak rapidly and soil heating into the range that is biological damaging ($>60\ ^{\circ}C$) occurs for only a short period of time and only at the surface or maybe a few centimeters into the soil. Thus, the important parts of the grasses (roots and buds) have excellent protection against even the most intense grass fires. Fires have been documented to be started by lightning and set intentionally by humans in both tropical and temperate grasslands. Fires are most common in grasslands with high levels of plant productivity, such as tallgrass prairies, and in these grasslands fire is important for keeping trees and adjacent forests from

Fig. 4 Photograph of a spring fire at the Konza Prairie Biological Field Station. The fire in the background is occurring ~2 weeks after the area in the foreground was burned. Photograph by Alan K. Knapp.

encroaching into grasslands. Many tree species are killed by fire, or if they are not killed, they are damaged severely because their active growing points are aboveground. Grassland plants survive and even thrive after fire because their buds are belowground where they are protected from lethal temperatures (**Fig. 4**).

The response of grassland species to fire mostly depends upon the production potential of the grassland. In the more highly productive grasslands (e.g., tallgrass prairie), fire in the dormant season (usually right before the growing season) results in an increase in growth of the grasses and thus greater plant production or total biomass. This occurs because the buildup of dead biomass (detritus) from previous years inhibits growth; fire removes this layer. However, in drier grasslands, or even in years in productive grasslands when the precipitation is low, the burning of this dead plant material may cause the soil to become excessively dry due to high evaporation losses. As a result, plants become water-stressed and growth is reduced after fire, thus resulting in lower productivity. It is only with long-term data that the true impact of fires on grasslands can be determined (**Fig. 5**).

So what are the mechanism(s) behind the increase in production in mesic grasslands after a fire? One of the most common misconceptions is that fire in grasslands increases productivity by increasing (releasing) the amount of nitrogen (N), a key limiting nutrient in terrestrial ecosystems. Actually, soil N decreases with burning. However, as mentioned above, the primary mechanism by which fire increases production in tallgrass prairie is through the removal of the accumulation of detritus produced in previous years. Standing dead biomass has been reported to accumulate to levels of up to 1000 g m^{-2} in tallgrass prairie and a steady state is achieved *c.* 3–5 years after a fire. The specific effects of this blanket of dead biomass on production are numerous and manifest on individual through the ecosystem levels. This detritus may accumulate to >30 cm deep, and this nonphotosynthetic biomass shades the soil surface and emerging shoots. This reduction in light available to shoots in sites without fire occurs for up to 2 months and because soil moisture is usually high in the spring, loss of energy at this time is especially critical for primary production. In concert with reductions in light available to the grasses, the early spring temperature environment is much different between burned and unburned sites, with burned sites having a higher temperature favoring the dominant C_4 grasses. All of these factors result in less production in unburned tallgrass compared to annually burned prairie (**Fig. 5**). Other evidence that fire does not increase N availability in mesic grasslands comes from N fertilization experiments. Within tallgrass prairie, in annually burned sites, N fertilizer had a strong impact on production, but in sites that have not been burned for several years, additional N did not enhance production and sites with intermediate fire histories had intermediate responses to N fertilization. The results of many studies suggest that one generality regarding grasses and fire is that grasses tolerate fire extremely well and in most cases reach their maximum production in the immediate post-fire years. One qualification to this statement is that the beneficial effect of fire is not uniform across all precipitation gradients. In addition, the growth form type of the dominant grass is also very important. Highly productive grasslands on the high end of precipitation gradients show moderate to high positive response to burning whereas more arid grasslands and some bunchgrass grasslands show reduced productivity in the first few years after fire.

Most grasslands have an active growing season as well as a dormant season. Although fire can occur year-round in many grasslands, fire is most likely to occur during the dormant season and it is most rare in the middle of the growing season during normal (non drought) years. Given the fact that so many aspects of a grassland change during the yearly cycle, it seems fair to expect that a fire in different seasons would have dramatically different impacts. However, in spite of the many studies that have examined the impact of fires at different times of the year, there does not seem to be a general consensus on fire seasonality. Rather, it is probably best to say that grasslands seem somewhat sensitive to `season of burn'. In one long-term study, it was found that the dominant grass in the tallgrass prairie (*Andropogon gerardii*) increased with burning in autumn, winter, or spring (dormant season), whereas burning in summer (growing season) resulted in an increase in many of the subdominant grasses with a reduction in *A. gerardii*.

Research indicates that community structure and ecosystem functioning in grasslands are impacted strongly by fire frequency. Plant species composition, in particular, differs dramatically between annually burned and less frequently burned sites in mesic grasslands. In tallgrass prairie, annually burned sites are dominated strongly by C_4 perennial grasses. Although C_4 grasses retain

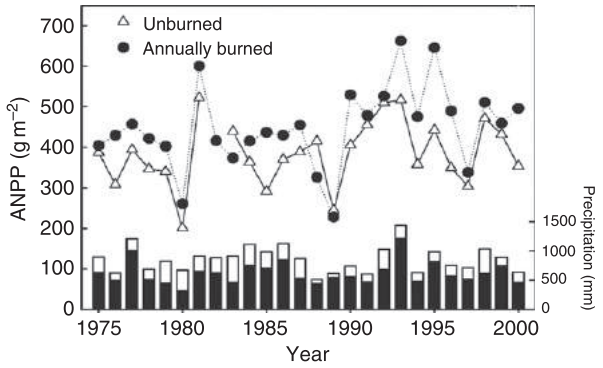


Fig. 5 Long-term record (26 years) of aboveground net primary production (ANPP) at Konza Prairie Biological Field Station from unburned sites (clear triangles) and annually burned sites (solid circles). The growing season precipitation (April–September; solid bars) and annual precipitation (clear bars) is also shown.

dominance at infrequently burned sites, C_3 grasses, forbs, and woody species are considerably more abundant resulting in greater diversity and heterogeneity in unburned prairie. In fact, the flora on annually burned sites is a nested subset of that found on less frequently burned areas. Thus, the differences reflect shifts in dominance between frequently and infrequently burned sites, rather than difference in composition *per se*. Again as with response of production to fire, there appears to be a gradient of response in community structure to grassland fires. In more northern prairies of North America, burning has not been shown to strongly affect community structure. However, these northern grasslands are dominated by C_3 grasses, which tend to decrease with burning, unlike the C_4 grasses that dominate prairies in warmer climates. Thus, the role of competition and fire in structuring grassland plant communities may increase along a latitudinal gradient throughout the Great Plains.

At a mesic grassland (Konza Prairie Biological Station), a clear picture of fire effects on plant community structure has emerged from the long-term (>20 years) empirical and experimental research done at the site. In the absence of large herbivores, the system is strongly driven by bottom-up forces associated with light, soil resource availability, and differential ability to compete under low-resource conditions. Although light availability increases with burning, the abundance of other critical limiting resources, N and water, declines as fire frequency increases. This is especially true in upland areas (with shallow soils) where production is likely limited by water. These changes in resource availability favor the growth and dominance of a small number of perennial C_4 grasses and forbs. As dominance by these competitive species increases, general declines in plant species diversity and community heterogeneity occur.

Impact of Fire on Consumers

Direct effects

Most grassland animals are not harmed by fire, particularly if fires occur during the dormant season. Those animals living belowground are well protected, and most grassland birds and mammals are mobile enough to avoid direct contact with fire. For example, there were few differences in the kinds and abundances of ground-dwelling beetles in frequently and infrequently burned Kansas tallgrass prairie. Insects that live in and on the stems and leaves of the plants are the ones that are most affected by fire. Fire has been shown to reduce directly the abundance of caterpillars which means fewer butterflies, which are important pollinators, in frequently burned prairies. Fortunately, most natural fires are patchy in that many unburned areas remain throughout a larger burned area. These patches serve as refugia for many insect populations. Given that these animals have short generation times these refugia often allow insect populations to recover quickly following a fire.

Indirect effects

Given the distinct effects of fire frequency on plant community structure and dynamics within and among burning treatments, it seems plausible that consumers that depend on the primary producers for food and habitat structure will be indirectly affected because fire alters food availability and habitat structure. Given that fire usually homogenizes grassland plant communities, one would predict that this would hold true for consumers. However, there does not appear to be tight linkages between changes in vegetation composition and structure animal populations. Indeed, work in an Oklahoma prairie shows that more grassland birds occur in areas with patchy burns than in areas that are uniformly burned or not burned. Much more work on how fire affects habitat heterogeneity and grassland consumers communities is needed.

Grazing in Grasslands

Grazing is a form of herbivory in which most of the leaves or other plant parts (small roots and root hairs) are consumed by herbivores. Grazing, both above- and belowground, is an important process in all grasslands. The long association of grazers and

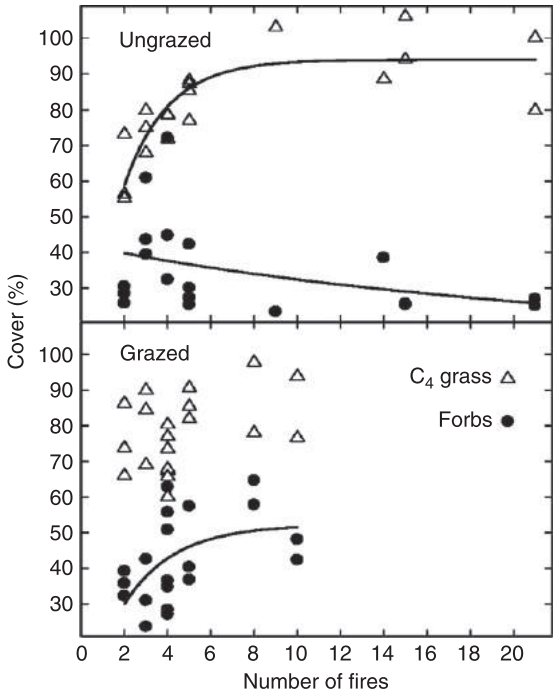


Fig. 6 Aboveground biomass removal by large ungulates modulates plant community responses to fire in mesic grasslands. In ungrazed prairie (top), cover of dominant C_4 grasses increased with increasing fire frequency, while cover of forbs decreased, resulting in a loss of diversity. However, in prairie grazed by bison (bottom), the cover of forbs was positively correlated with fire frequency and the cover of grasses was unaffected, resulting in high diversity in spite of frequent fires. From Collins, S.L., Knapp, A.K., Briggs, J.M., Blair, J.M., Steinauer, E.M., 1998. Modulation of diversity by grazing and mowing in native tallgrass prairie. Science 280 (5364), 745–747.

grasslands has prompted the hypothesis that grasses and their megaherbivore grazers are a highly coevolved system, but, as mentioned above, there is some more recent evidence that this might not be the case. However, there is no disagreement that large grazers have been a factor in grassland ecology since their origin. The herbivory actions of many other smaller organisms including small mammals and insects may be equally important. There is no doubt that the impact of native grazers in grasslands can be extensive and work on the East African Serengeti plains estimated that 15% to >90% of the annual aboveground net primary productivity can be consumed by ungulates. However, data from small mammal exclosures suggest that small mammals can also impact grasslands as when small mammals were excluded from plots in Kenya; biomass was 40–50% higher than in adjacent plots where small mammals occurred.

Due to the ability of grasses to cope with high rates of herbivory, many former natural grasslands are now being managed for the production of domestic livestock, primarily cattle in North and South America and Africa, as well as sheep in Europe, New Zealand, and other parts of the world. Grasslands present a vast and readily exploited resource for domestic grazers. However, like many resources, grasslands can be overexploited (discussed in more detail below).

Grazing systems can be roughly divided into two main types – commercial and traditional – with the traditional type often mainly aimed at subsistence. Commercial grazing of natural grasslands is very often at a large scale and commonly involves a single species, usually beef cattle or sheep for wool production. Some of the largest areas of extensive commercial grazing developed in the nineteenth century on land which had not previously been heavily grazed by ruminants; these grazing industries were mainly developed in the Americas and Australia, and to a much less degree in southern and eastern Africa. Traditional livestock production systems vary according to climate and the overall farming systems of the area. They also use a wider range of livestock, including buffaloes, asses, goats, yaks, and camels. In traditional farming systems, livestock are often mainly kept for subsistence and savings, and are frequently multipurpose, providing meat, milk, and manure as fuel.

Grazing aboveground by large herbivores alters grasslands in several ways. Grazers remove fuel and may lessen the frequency and intensity of fires. Most large grazers such as cattle or bison primarily consume the grasses; thus the less abundant forb species (broad-leafed, herbaceous plants) may increase in abundance and new species may invade the space that is made available. Thus, fire reduces heterogeneity in mesic grassland (a few species dominate) while grazers increase heterogeneity regardless of fire frequency. In other words, grazing decouples the impact of fire in productive grasslands (Fig. 6). As a result; grazing increases plant species diversity in mesic grasslands. In xeric grasslands, on the other hand, grazing may lower species diversity, particularly by altering the availability of suitable microsites for forb species. These effects are strongly dependent on grazing intensity. Overgrazing may rapidly degrade grasslands to systems dominated by weedy and non-native plant species.

Grazers may also accelerate the conversion of plant nutrients from forms that are unavailable for plant uptake to forms that can be readily used. Essential plant nutrients, such as nitrogen, are bound for long periods of time in unavailable (organic) forms in

plant foliage, stems, and roots. These plant parts are slowly decomposed by microbes and the nutrients they contain are only gradually released in available (inorganic) forms. This decomposition process may take more than a year or two. Grazers consume these plant parts and excrete a portion of the nutrients they contain in plant-available forms. This happens very quickly compared to the slow decomposition process, and nutrients are excreted in high concentrations in small patches. Thus, grazers may increase the availability of potentially limiting nutrients to plants as well as alter the spatial distribution of these resources.

Some grasses and grassland plants can compensate for aboveground tissue lost to grazers by growing faster after grazing has occurred. Thus, even though 50% of the grass foliage may be consumed by bison or wildebeest, when compared to ungrazed plants at the end of the season, the grazed grasses may be only slightly smaller, the same size, or even larger than ungrazed plants. This latter phenomenon, called `overcompensation' is controversial, yet the ability of grasses to compensate partially or fully for foliage lost to grazers is well established. Compensation occurs for several reasons, including an increase in light available to growing shoots in grazed areas, greater nutrient availability to regrowing plants, and increased soil water availability. The latter occurs after grazing because the large root system of the grasses is able to supply abundant water to a relatively small amount of regrowing leaf tissue.

As with fire, the impact of grazing on grasslands depends upon where in the precipitation gradient the grassland occurs (usually more mesic grasslands can recover more quickly than arid grasslands) as well as the growth form – cespitose (bunch-forming grasses) versus rhizomatous grasses. But another key factor is the evolutionary history of the grassland. In general, grasslands with a long evolutionary history of grazers, as in Africa, are very resilient to grazing whereas grasslands with a short evolutionary history such as desert grasslands in North America can easily be damaged by even light grazing.

Threats to Grasslands and Restoration of Grasslands

Grassland environments are key agricultural areas worldwide. In North America and elsewhere, grasslands are considered to be endangered ecosystems. For example, in US Great Plains up to 99% of native grassland ecosystems in some states have been plowed and converted to agricultural use or lost due to urbanization. Similar but less dramatic losses of mixed and shortgrass prairies have occurred in other areas. While the loss of native grasslands due to agricultural conversion is still occurring in some places, dramatic increases in woody shrub and tree species threatens many remaining tracts of grasslands. Indeed, across the world, the last remaining native grasslands are being threatened by an increase in the abundance of native woody species from expansion of woody plant cover originating from both within the ecosystem and from adjacent ecosystems. Increased cover and abundance of woody species in grasslands and savannas have been observed worldwide with well-known examples from Australia, Africa, and South America. In North America, this phenomenon has been documented in mesic tallgrass prairies of the eastern Great Plains, subtropical grasslands and savannas of Texas, desert grasslands of the Southwest, and the upper Great Basin. Purported drivers of the increase in woody plant abundance are numerous and include changes in climate, atmospheric CO_2 concentration, nitrogen deposition, grazing pressure, and changes in disturbance regimes such as the frequency and intensity of fire. Although the drivers vary, the consequences for grassland ecosystems are strikingly consistent. In many areas, the expansion of woody species increases net primary production and carbon storage, but reduces biodiversity. The full impact of shrub encroachment on grassland environments remains to be seen.

Another threat to native grasslands is the increase of non-native grass species. For example, in California, it is estimated that an area of approximately 7 000 000 ha (about 25% of the area of California) has been converted to grassland dominated by non-native annuals primarily of Mediterranean origin. Conversion to non-native annual vegetation was so fast, so extensive, and so complete that the original extent and species composition of native perennial grasslands is unknown. In addition, across the western US, invasive exotic grasses are now dominant in many areas and these species have a significant impact on natural disturbance regimes. For example, the propensity for annual grasses to carry and survive fires is now a major element in the arid and semiarid areas in western North America. In the Mojave and Sonoran deserts of the American Southwest, in particular, fires are now much more common than they were historically, which may reduce the abundance of many native cactus and shrub species in these areas. This annual-grass-fire syndrome is also present in native grasslands of Australia and managers there and in North America are using growing season fire to try to reduce the number of annual plants that set seed and thus reduce the populations of exotics, usually with very mixed results.

Conservation and Restoration

Because grasslands have tremendous economic value as grazing lands and also serve as critical habitat for many plant and animal species, efforts to conserve the remaining grasslands and restore grasslands on agricultural land are underway in many states and around the world. The most obvious conservation practice is the protection and management of existing grasslands. This includes both private and public lands. Probably the largest private holder of grasslands in the world is The Nature Conservancy. The Nature Conservancy is a global organization that works in all 50 states in the United States of America, and in 27 countries, including Canada, Mexico, Australia, and countries throughout the Asia-Pacific region, the Caribbean, and the Latin America.

However, as mentioned numerous times, the factors that led to the establishment of grasslands and, in particular, the organic-rich soils derived from the dominant biota have facilitated the agricultural exploitation of grasslands. Consequently, many

grasslands that were historically persistent have been converted to cropland. Thus, restoration of grasslands is also a very important conservation practice. Grassland restoration is the process of recreating grassland (including plant and animal communities, and ecosystem processes) where one existed but now is gone. Grassland restoration can include planting a new grassland where one had been broken and farmed, or it can include improving a degraded grassland (e.g., one that was never plowed but lost many plant and animal species due to prior land management practices). Restoration practices of existing grasslands may include reintroducing fires into grasslands following extended periods of fire suppression. On areas that have been moderately to heavily grazed (but not completely overgrazed), reducing the intensity of grazing may be required. In addition, mowing is also a cost-effective method of restoring grasslands. Mowing can be effective on sites that have been invaded by brush and forest, but the grasses are still present.

In areas where the grasses are completely absent (agriculture fields) or in a very degraded state, reseeding of grasses is usually necessary. There are proven techniques, complete with specialized equipment (seed drills) for restoration of grasslands, and, for the most part, it is fairly easy to get the dominant grasses established in an area. Indeed, some of the earliest examples of restoration ecology come from efforts to restore native tallgrass prairie in North America. As a result, the market for restoration of grasslands (at least in North America) has developed to the point that obtaining enough grass seed (sometimes even local native seed) is not a problem. A bigger challenge, however, in restored grasslands is increasing establishment of the nongrass species which are so critical for biodiversity. Seeds may be more difficult to obtain (especially for rarer plants), and then getting the forbs to survive and reproduce in many grassland restoration projects has been challenging. Further research is needed regarding what management techniques are important to their establishment and growth in these restored areas.

In addition to the prairie flora that is at risk, grassland animals (particularly birds and butterflies) suffer when grassland quality declines. In North America, grassland birds were historically found in vast numbers across the prairies of the western Great Plains. Today, the birds of these and other grasslands around the world have shown steeper, more consistent, and more geographically widespread declines than any other group. These losses are a direct result of the declining quantity and quality of habitat due to human activities like conversion of native prairie to agriculture, urban development, and suppression of naturally occurring fire.

See also: Aquatic Ecology: Eutrophication. Behavioral Ecology: Herbivore-Predator Cycles. Ecological Data Analysis and Modelling: Modeling Dispersal Processes for Ecological Systems

Further Reading

Borchert, J.R., 1950. The climate of the central North American grassland. Annals of the Association of American Geographers 40, 1–39.
Briggs, J.M., Knapp, A.K., Blair, J.M., *et al.*, 2005. An ecosystem in transition: Woody plant expansion into mesic grassland. BioScience 55, 243–254.
Collins, S.L., Knapp, A.K., Briggs, J.M., Blair, J.M., Steinauer, E.M., 1998. Modulation of diversity by grazing and mowing in native tallgrass prairie. Science 280 (5364), 745–747.
Collins, S.L., Wallace, L.L., 1990. Fire in North American Tallgrass Prairies. Norman, OK: University of Oklahoma Press.
Frank, D.A., Inouye, R.S., 1994. Temporal variation in actual evapotranspiration of terrestrial ecosystems: Patterns and ecological implications. Journal of Biogeography 21, 401–411.
French, N. (Ed.), 1979. Perspectives in Grassland Ecology. Results and Applications of the United States International Biosphere Programme Grassland Biome Study. New York: Springer.
Knapp, A.K., Blair, J.M., Briggs, J.M., *et al.*, 1999. The keystone role of bison in North American tallgrass prairie. BioScience 49, 39–50.
Knapp, A.K., Briggs, J.M., Hartnett, D.C., Collins, S.L., 1998. In: Grassland Dynamics: Long-Term Ecological Research in Tallgrass Prairie. New York: Oxford University Press, p. 364.
Loveland, T.R., Reed, B.C., Brown, J.F., *et al.*, 1998. Development of a Global Land Cover Characteristics Database and IGBP DISCover from 1 km AVHRR Data. International Journal of Remote Sensing 21 (6–7), 1303–1330.
McNaughton, S.J., 1985. Ecology of a grazing ecosystem: The serengeti. Ecological Monographs 55, 259–294.
Milchunas, D.G., Sala, O.E., Lauenroth, W.K., 1988. A generalized model of the effects of grazing by large herbivores on grassland community structure. American Naturalist 132, 87–106.
Oesterheld, M., Loreti, J., Semmartin, M., Paruelo, J.M., 1999. Grazing, fire, and climate effects on primary productivity of grasslands and savannas. In: Walker, L.R. (Ed.), Ecosystems of the World. Amsterdam: Elsevier, pp. 287–306.
Olson, J.S., 1994. In: Global Ecosystem Framework – Definitions. Sioux Falls, SD: USGS EDC, p. 39.
Prasad, V., Strömberg, C.A.E., Alimohammadian, H., Sahni, A., 2005. Dinosaur coprolites and the early evolution of grasses and grazers. Science 310, 1177–1190.
Sala, O.E., Parton, W.J., Joyce, L.A., Lauenroth, W.K., 1988. Primary production of the central grassland region of the United States. Ecology 69, 40–45.
Samson, F., Knopf, F., 1994. Prairie conservation in North America. BioScience 44, 418–421.
Weaver, J.E., 1954. North American Prairie. Lincoln, NE: Johnsen Publishing Company.

Swamps

C Trettin, USDA, Forest Service, Charleston, SC, USA

Introduction

Swamp is a general term that is defined as ``spongy land, low ground filled with water, soft wet ground" (Webster, 1983), hence its association with a wide variety of terrestrial ecosystems. Typically, a swamp is considered a forested wetland. A wetland is a type of terrestrial ecosystem that has a hydrologic regime where the soil is saturated near the surface during the growing season; the soil has hydric properties, expressing characteristics of anaerobic conditions; and the dominant vegetation is hydrophytic with adaptations for living in the wet soils. In the case of a swamp, the forest species are adapted to the wet soil conditions. Without geographic context, there is little functional information conveyed by the term swamp other than the prevalence of wetland conditions and dense forest vegetation (**Fig. 1**).

The following discussion is designed to convey the common hydrologic settings, soil conditions, and vegetative communities that occur within the common usage of the term swamp. References focusing on swamp forests should be consulted for specific geographic regions.

General Properties of a Swamp

Hydrology

The hydrologic setting controls the form and function of the wetland because of the dependence on excess water to mediate biological and geochemical reactions. There are four general settings that may be used to characterize the swamp hydrology (**Fig. 2**). The riverine or floodplain setting is the most commonly associated hydrogeomorphic setting for swamps. It is characterized by periodic flooding from the river or stream, and it may also receive runoff from adjoining uplands. The periodicity, and flood depth and duration are the key factors that affect the type of forest communities present in the swamp. Depressional wetlands occur where there are surface depressions which receive water from the surrounding uplands, directly from precipitation, and in certain instances, they may also intersect a shallow water table. Lacustrine and estuarine fringe wetlands receive their water primarily from an open-water body; runoff from adjoining uplands and precipitation also contribute to the water balance. The common hydrologic attribute of swamps in each of those settings is the presence of water above the soil surface, but the period of inundation varies widely. While it is common for swamps to have flooded conditions for periods ranging from days to months on an annual basis, it is not uncommon for there to be multiyear intervals between flood events. The factors that affect the flooding regime include timing and amount of precipitation, groundwater level, land use in the watershed contributing to the swamp, and evapotranspiration.

Soils

Swamp soils cover the full range of texture classes and degrees of organic matter accumulation (**Fig. 3**). The wet mineral soils are characteristic of riverine and depressional settings. The histic mineral soils have a moderately thick accumulation of surface organic matter (<40 cm) reflecting prolonged periods of saturation and little scouring action if located in a floodplain, hence they may be found in any of the four hydrologic settings. The histosols or peat soils have a thick layer (>40 cm) of organic matter accumulation, representing the long periods of saturation on an annual basis. These soils typically occur in depressional settings and are not common in floodplains due to the periodic scouring that occurs during flood events.

Fig. 1 Bottomland hardwood swamp, characteristic of floodplains in the Southeastern United States.

 doi:10.1016/B978-0-444-63768-0.00360-7

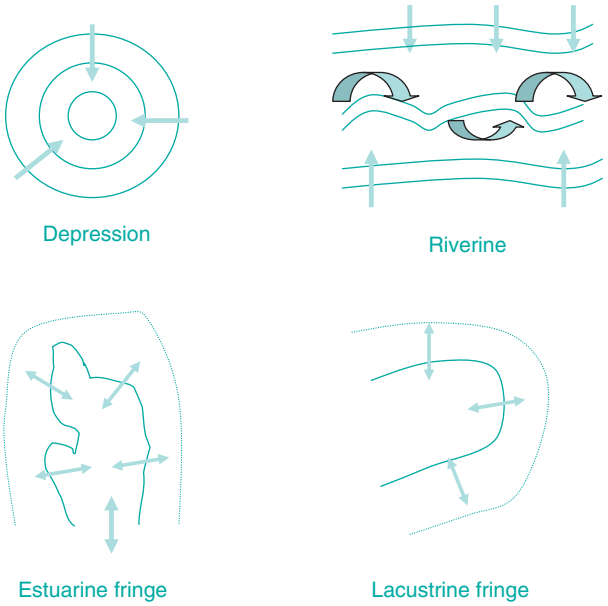


Fig. 2 Influence of geomorphic position on hydrology. The arrows show the dominant direction of water flow for the four dominant types of geomorphic positions that are characteristic for swamps. After Vasander H (1996) Peatlands in Finland, 64pp. Helsinki: Finnish Peatland Society.

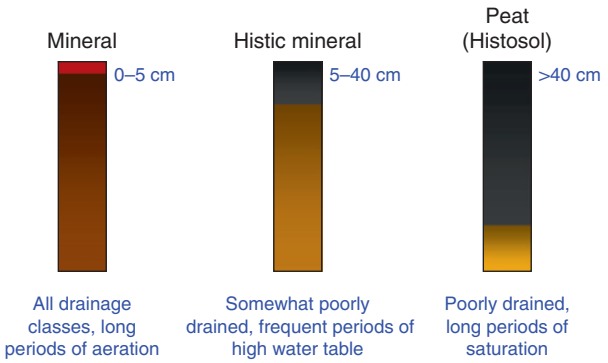


Fig. 3 Types of soils common in swamp forests. The three categories reflect the amount of organic matter that has accumulated on the soil surface, which is in turn controlled by the soil drainage and hydrology. After Trettin CC, Jurgensen MF, Gale MR, and McLaughlin JA (1995) Soil carbon in northern forested wetlands: Impacts of silvicultural practices. In: McFee WW and Kelly JM (eds.) *Carbon Forms and Functions in Forest Soils*, pp. 437–461. Madison, WI: Soil Science Society of America.

Vegetation

The term swamp generally implies a forested wetland. However, due to the wide range of physical settings (see the previous two sections) and geographic locations ranging from the boreal to the tropical climatic zones on each continent, there are no consistent characteristics or attributes beyond the occurrence of hydrophytic vegetation. Accordingly, swamps may be dominated by either conifers or angiosperms, but a common situation would be a mixture of species and communities reflecting relatively minor differences in a microsite. For example, while a floodplain forest may be broadly characterized as a bottomland hardwood swamp, it contains a mosaic of vegetative communities which reflect small differences in hydrology and soils.

Ecological Functions

The ecological functions of swamps are significant, because of their prevalence and the wide range of conditions that they occupy. The following overview highlights some of the major ecosystem functions that are provided from swamp wetlands; specifics for a particular type of swamp are available from the regional references.

Hydrology

Hydrologic functions that are mediated by swamp wetlands depend on the hydrogeomorphic setting. Riverine swamps provide temporary storage for floodwaters, thereby reducing the peak flow to downstream areas. This function is physically based, with little interaction with the type of forest vegetation. However, changes in land use, especially conversion to agriculture, in the floodplain, may reduce the water storage potential, resulting in enhanced downstream conveyance of flow. The flood storage function also serves to sustain stream flow, as the waters slowly drain from the area. Swamps occurring in a depressional setting may be a source of groundwater recharge, where accumulated surface water slowly infiltrates through the subsurface sediments. In estuarine and lacustrine settings, swamps occurring at the land–water margin are important for the stability of the shoreline.

Water quality

The effects of a swamp on water quality depend on the hydrogeomorphic setting. The riverine swamp affects water quality in two primary ways – by physical and biogeochemical reactions. Sediment removal is an important function of the riverine swamps; this is a process where sediment in the floodwaters settles out onto the floodplain surface. The deposited sediment provides nutrients to the swamp vegetation and it represents the removal of a contaminant from the floodwater. Floodplains with dense understory vegetation can be more effective than open forest settings in filtering sediment from the floodwaters.

The floodplain and riparian zone swamps may also remove chemical constituents from the water, particularly nitrogen and phosphorus. As a result of the anaerobic soil conditions, nitrate nitrogen, which is a common pollutant in surface and shallow-subsurface runoff, can be converted to nitrogen gas, thereby removing it from the water. The removal of phosphorus compounds typically involves reactions associated with the sediments.

Habitat

Swamps are important for the diversity of habitat conditions that they provide. At the large scale, swamps comprise part of the mosaic of land types, yielding wet, vegetative conditions among uplands. At smaller scales, within a swamp, there are a multitude of habitat conditions that are largely dictated by elevation relative to the mean high water level.

Terrestrial

The terrestrial habitats provided by swamps are diverse due to variations in vegetative composition and structure, which are largely regulated by the hydrologic conditions of the site. The habitat also changes through the development of the forest. In early successional stages, the vegetation is typically a dense combination of shrubs and trees; then, as the trees gain dominance, the shrub layers die back yielding a less dense understory. Correspondingly, the habitat conditions for amphibians, birds, reptiles, and mammals change as the stand evolves. The swamp forests are particularly important habitat for birds, especially migratory song birds.

Aquatic

Swamps also provide important aquatic habitat for fish, birds, and amphibians. Organic matter produced in the swamp is an important energy source for aquatic organisms, including those living in water bodies within the swamp and also larger receiving bodies such as lakes, rivers, and oceans. In floodplains, the floating debris and logs provide physical structures that are an important component of the aquatic habitat.

Restoration

In many areas, swamps have been converted into agricultural use, through the use of drainage systems and clearing of the forest vegetation. The merits of restoring the converted wetlands back to swamp forests include the reestablishment of flood water storage, in the case of floodplains, and the development of wildlife habitat. The restoration of swamp forests is complicated by the myriad of soil and hydrologic conditions that one may encounter, and the effects of past management practices which necessitate the restoration may also exacerbate the situation. However, with proper consideration of the hydrologic setting and matching species to the soil and water regimes, functional restoration is feasible. The typical sequence of restoring swamp forests is to reestablish the wetland hydrology by blocking drainage ditches, and planting appropriate tree and understory species.

Ecosystem Services and Values

Swamps provide both direct and indirect values to society. Direct values include raw materials, such as timber and food stocks. Indirect values include floodwater storage, water supply, water quality, recreation, esthetics, wildlife diversity, and biodiversity. The valuation will depend on inherent characteristics of the resource that are largely constrained by the biogeographic zone and location within a watershed, societal norms, and economic conditions.

See also: Ecological Processes: Decomposition and Mineralization. Global Change Ecology: Energy Flows in the Biosphere

Further Reading

Barton, C., Nelson, E.A., Kolka, R.K., *et al*., 2000. Restoration of a severely impacted riparian wetland system – The Pen Branch Project. Ecological Engineering 15, S3–S15.

Burke, M.K., Lockaby, B.G., Conner, W.H., 1999. Aboveground production and nutrient circulation along a flooding gradient in a South Carolina Coastal Plain forest. Canadian Journal of Forest Research 29, 1402–1418.

Conner, W.H., Buford, M.J., 1998. Southern deepwater swamps. In: Messina, M.G., Conner, H. (Eds.), Southern Forested Wetlands Ecology and Management. Boca Raton, FL: CRC Press, pp. 261–287.

Conner, W.H., Hill, H.L., Whitehead, E.M., *et al*., 2001.*General Technical Report SRS-43* In: Forested wetlands of the Southern United States: A bibliography. Asheville, NC: US Department of Agriculture, Forest Service, Southern Research Station, p. 133.

Conner, R.N., Jones, S.D., Gretchen, D., 1994. Snag condition and woodpecker foraging ecology in a bottomland hardwood forest. Wilson Bulletin 106 (2), 242–257.

Conner, W.H., McLeod, K., 2000. Restoration methods for deepwater swamps. In: Holland, M.M., Warren, M.L., Stanturf, J.A. (Eds.), Proceedings of a Conference on Sustainability of Wetlands and Water Resources, 23–25 May. Oxford, MS: US Department of Agriculture, Forest Service, Southern Research Station.

de Groot R, Stuip M, Finlayson M, and Davidson N (2006) Valuing wetlands: Guidance for valuing the benefits derived from wetland ecosystem services. *Ramsar Technical Report No. 3, CBD Technical Series No. 27*, Convention on Biological Diversity. Gland, Switzerland: Ramsar Convention Secretariat. http://www.cbd.int/doc/publications/cbd-ts-27.pdf (accessed November 2007).

Messina, M.G., Conner, W.H. (Eds.), 1998. Southern Forested Wetlands Ecology and Management. Boca Raton, FL: CRC Press, p. 347.

Mitch, W.J., Gosselink, J.G., 2000. Wetlands. New York: Wiley, p. 920.

National Wetlands Working Group (NWWG), 1988. Wetlands of Canada. In: Ecological Land Classification Series, No. 24. Ottawa: Sustainable Development Branch, Environment Canada, p. 452.

Stanturf, J.A., Gardiner, E.S., Outcalt, K., Conner, W.H., Guldin, J.M., 2004.*General Technical Report SRS–75* In: Restoration of southern ecosystems. Asheville, NC: US Department of Agriculture, Forest Service, Southern Research Station, pp. 123–211.

Trettin, C.C., Jurgensen, M.F., Gale, M.R., McLaughlin, J.A., 1995. Soil carbon in northern forested wetlands: Impacts of silvicultural practices. In: McFee, W.W., Kelly, J.M. (Eds.), Carbon Forms and Functions in Forest Soils. Madison, WI: Soil Science Society of America, pp. 437–461.

Vasander, H., 1996. In: Peatlands in Finland. Helsinki: Finnish Peatland Society, p. 64.

Webster, N., 1983. Unabridged Dictionary, 2nd edn. Cleveland, OH: Dorset and Baber.

Relevant Websites

http://www.aswm.org
Association of State Wetland Managers.

http://www.ncl.ac.uk
Mangrove Swamps WWW Sites, Newcastle University.

http://www.ramsar.org
Ramsar Convention on Wetlands.

http://www.sws.org
Society of Wetland Scientists.

http://www.epa.gov
Wetlands at US Environmental Protection Agency.

http://www.wetlands.org
Wetlands International.

http://www.panda.org
World Wildlife Fund.

Temperate Forest

WS Currie and KM Bergen, University of Michigan, Ann Arbor, MI, USA

Introduction

The temperate forest biome is characterized by a distinct seasonality that includes a long growing season together with a cold winter season in which much of the vegetation may be dormant. The strong seasonality drives physiological events to occur at regular annual intervals for plant species. These include bud break, flowering, and foliar and shoot extension. As the growing season ends, marked by dropping temperatures and shortening photoperiod (day length), trees and shrubs undergo seasonal physiological changes that include the senescence and abscission of foliage (although in evergreen species some foliage is also retained) and the setting of buds for the next growing season. Because of the cold winters, the dominant woody vegetation is characterized by freeze-hardy species. During the winter season, the air temperature drops below freezing and soils are frozen or cold and wet, impeding decomposition of plant litter and promoting the accumulation of an organic layer on the soil surface.

The temperate forest is distributed over portions of five regions of the globe: North America, South America, Europe, Asia, and Australia–New Zealand (**Fig. 1**). Within this biome, distinct biogeographic units are recognized, particularly the mixed-deciduous temperate forest (the largest in terms of area), the mixed-evergreen temperate forest (sometimes called subtropical evergreen), and the temperate rainforest. Major taxa include pines (*Pinus* spp.), maples (*Acer* spp.), beeches (*Fagus* spp., *Nothofagus* spp.), and oaks (*Quercus* spp.) in the mixed-deciduous and mixed-evergreen temperate forests; spruces (*Picea* spp.), Douglas-fir (*Pseudotsuga menziesii*), and redwoods (*Sequoia sempervirens*, *Sequoiadendron giganteum*) in the Northern Hemisphere temperate rainforests; and southern beeches (*Nothofagus* spp.) and eucalyptus (*Eucalyptus* spp.) in Southern Hemisphere temperate rainforests.

Within a continent, forests in the temperate biome grade into subdivisions based on latitude, elevation, and large-scale patterns of precipitation. In North America, for example, the predominant natural vegetation in the eastern United States and the southern reaches of eastern Canada is mixed-deciduous temperate forest. This forest grades to the south through broad-leaved-coniferous mixtures to the mixed-evergreen forest along the Atlantic coastal plain (**Fig. 2**). Temperate rainforests in North America are found in the coastal Pacific Northwest where marine climates together with orographic lifting produce high rainfall. In South America temperate forests are found in Chile and parts of Patagonia. In Europe, within the temperate forest biome the mixed-deciduous forests dominate in the western continent, Great Britain, southern Eastern Europe, and southern European Russia. In Near-East Asia, the temperate forest occurs in Turkey and Iran and a narrow band is found in Central Asia as a transition between the boreal forest to the north and steppe to the south. The temperate forests in East Asia occur predominantly in northern and central China, but also over most of Japan, Korea, and of the southern tip of Siberia. Temperate forests, including rainforests, are also found in parts of New Zealand, the southeast coast of Australia, and Tasmania.

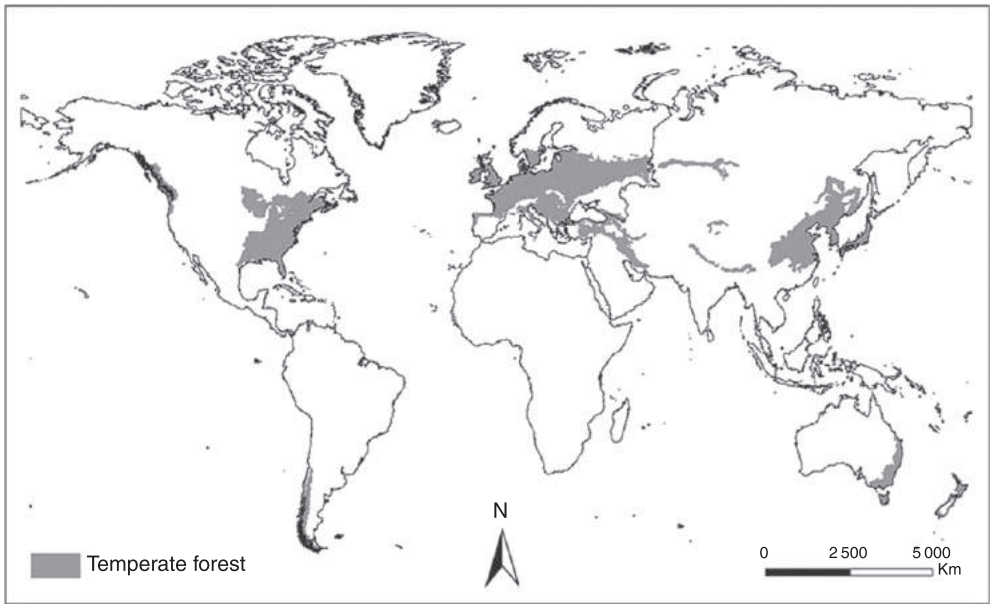


Fig. 1 The distribution of temperate forests of the world. Map data: Olson, D.M., Dinerstein, E., Wikramanayake, E.D., et al., 2001. Terrestrial ecoregions of the world: A new map of life on earth. BioScience 51, 933–938. Map prepared by the Environmental Spatial Analysis Laboratory, University of Michigan, USA, 2006.

Fig. 2 The edge of a mixed coniferous–deciduous forest in southeastern Maine, USA. Photo by W. S. Currie.

Temperate forests are distinguished from boreal forests by having a 4–6-month (140–200 days) frost-free growing season with on average at least 4 months at 10 °C or above and mean annual temperatures from 5 °C to 20 °C. At higher latitudes, the temperate forest transitions to the boreal forest, a biome of evergreen cold-tolerant forests with much shorter growing seasons. The latter are also found in middle latitudes as montane forests at high elevations and are often closer, floristically and functionally, to boreal than to temperate forests.

The occurrence of frost (at 0 °C or colder) differentiates the extratropical (including temperate) from tropical regions. Moisture also distinguishes temperate regions from drier forested regions, such as chaparral and wetter forested regions such as tropical rainforests. In temperate regions, precipitation exceeds potential evaporation and water is available at approximately 50–200 cm yr^{-1}. Precipitation in most temperate regions is fairly evenly distributed throughout the year in contrast to the tropics where there are typically pronounced wet and dry seasons.

Physiography, Climate, and the Temperate Forest Biome

Climatic and Physiographic Controls on the Distribution of Temperate Forests

The geographic distributions of the different vegetation biomes of the world are dependent on the physical environment and climate in the form of light, temperature, and moisture. In middle latitudes (30°–60° N and S), these controls result in a temperate forest biome within each hemisphere that is discontinuous, separated by the oceans and the tropics, and by moisture and physiographic barriers. The present-day distribution of temperate forests derives not only from present climatic controls but also from paleoclimates and past connections among the continents. Climates during the Pleistocene (*c.* 1.8 million to 10 000 years ago) set the stage for the present-day distribution. During glacial maxima, ice sheets covered large parts of Europe and North and South America as well as isolated areas in East Asia. In North and South America, plants migrated to unglaciated refugia and re-migrated, as glaciers receded, to their present-day distributions. Evidence suggests that many genera of forest trees that remain in North America and unglaciated East Asia were extirpated from Europe because the east–west running Alpine range blocked migrations to refugia during Pleistocene glaciations. Similarly important were continental connections between North America (the Nearctic), East Asia, and Europe (the Palearctic) at different points in geologic history. As a result, floristic differences are relatively small across the Holarctic, which spans from the west coast of North America to the east coast of Asia and includes the majority of the world's temperate forests.

Temperate forests occur across a wide range of local physiographic landforms, from rocky slopes to rolling plains and river floodplains, although generally under non-extreme physiographic conditions. Trees that occupy slopes or well-drained substrates with low organic matter such as sandy outwash plains (e.g., pines and some oaks) are adapted for drier (xeric) sites low in nutrients. Trees adapted for moderate (mesic) sites are found on plains, glacial moraines, or low hills with greater stocks of soil organic matter. Nutrient- and moisture-demanding broad-leaved species, for example, maples and beeches, thrive in mesic landscapes. Trees occupying river floodplains, wetlands, or bogs have environments that can be very moist to wet (wet-mesic to hydric). These soils are relatively rich in organic matter but trees in these landscapes must be adapted to withstand flooding, including long periods with wet, anoxic soils with low nutrient availability.

Climatic and Physiographic Subdivisions

Given the great geographic extent of temperate forests it is not surprising that regional differences are observed. Systematic classifications of ecoregions and climates describe subdivisions within the biome (**Table 1**). The extensive temperate mixed-deciduous forest occurs primarily in Bailey's warm continental division (210), hot continental division (220), and marine division (240); these are Köppen–Trewartha classes *Dcb*, *Dca*, *Do*, and *Cf*. The warm continental division has snowy cold winters, while the

Table 1 Temperate forest biome types and corresponding geographic regions, Bailey ecoregions, and Köppen–Trewartha climate classes

Temperate forest type	*Geographic region*	*Bailey ecoregion*	*Köppen–Trewartha*[a] *climate class*
Temperate mixed-deciduous forest	• Eastern North America • Asia • Europe • South America • Australia/New Zealand	Humid temperate domain (200) • Warm continental division (210) • Hot continental division (220) • Marine division (240)	*Dcb*: Temperate continental, cool summer *Dca*: Temperate continental, warm summer *Do*: Temperate oceanic *Cf*: Humid subtropical
Temperate mixed-evergreen forest	• Southeast North America • Asia • South America • Australia/New Zealand	Humid temperate domain (200) • Subtropical division (230)	*Cf*: Humid subtropical
Temperate rainforest	• Northwestern North America • South America • Southeast Australia/New Zealand	Humid temperate domain (200) • Marine division (240)	*Cf*: Humid subtropical *Do*: Temperate oceanic

[a]*Dc*: Temperate continental: 4–7 months above 10 °C, coldest month below 0 °C; *Cf*: Humid subtropical: 8 months 10 °C, coldest month below 18 °C, no dry season; *Do*: Temperate oceanic: 4–7 months above 10 °C, coldest month above 0 °C.

hot continental division has warmer, wetter summers and milder winters. In the marine division (240) winters are mild, summers relatively cool, and precipitation occurs most of the year.

The temperate mixed-evergreen forests occur primarily in Bailey's temperate and rainy subtropical division (230) which is most analogous to the Köppen–Trewartha mid- and lower-latitude *Cf* (humid subtropical) class (**Table 1**). These climates have no dry season, with even the driest months having at least 30 mm of rain, and have hot summers with the average temperature of warmest month greater than 22 °C.

Temperate rainforest conditions largely occur where ocean moisture is abundant and prevented from moving inland by mountain ranges. These conditions occur in particular continental placements within Bailey's marine division (240) and Köppen–Trewartha *Do* class in higher latitudes and within Bailey's subtropical division (230) and Köppen–Trewartha *Do* and *Cf* classes in lower latitudes (**Table 1**).

Disturbance and Forest Structure

Major disturbances occur naturally in temperate forests, although particular locations vary in the types, frequencies, and severities of disturbance. Major natural disturbances include fires, windthrow during severe storms, ice storms, flooding, disease, and irruptions of defoliating or wood-boring insects. The array of natural disturbances that occur at a particular location constitutes its disturbance regime, a strong force in shaping forest structure and composition. Smaller-scale disturbances also shape forests over long time periods in the absence of a major disturbance. These include the production of forest gaps from the mortality of one to a few large trees. In some cases, idiosyncratic combinations of processes may produce repeated disturbance. An example is `fir waves' that occur only in Japan and the northeastern US. In these waves of mortality that pass through the forest repeatedly, a fungal pathogen weakens the roots in mature trees while wind gusts cause the weakened roots to break as they rub against sharp gravel in the rocky soil. Because of the repetitive nature of natural disturbances and the long lifetimes of temperate forest trees, trees are often adapted (through what is termed `vital attributes') to withstand particular disturbances or to regenerate following disturbance. Some examples are trees that re-sprout from stumps following fire or from branches following windthrow, cones that require fire to open, and seeds that germinate best on exposed soil.

Human activities have substantially altered the disturbance regimes in many temperate forests. The large-scale harvesting of trees for timber, whether cutting selected sizes or species of trees or cutting all of the trees in a stand, are relatively new forms of disturbance that now affect forest structure and community composition throughout much of the temperate biome. Human activities also cause large-scale chronic disturbances, including polluted rainfall (e.g., acid rain) that causes soil acidification and nitrogen enrichment over large regions of the US, Western Europe, and increasingly in eastern Asia. Still another category of human-induced disturbance is in the introduction of invasive species. In the eastern US, the introduction of a fungal pathogen in the early twentieth century caused the chestnut blight, essentially eradicating one of the dominant trees (the American chestnut, *Castanea dentata*) from a large region.

Structural Layers of Vegetation

Disturbances in temperate forests vary not only in their type and frequency but also in their intensity or severity, the latter gauged by the percentage of vegetation mortality. A major disturbance that causes widespread or near total mortality of trees in a forest stand, followed by the development of a new (secondary) forest stand, is known as a stand-initiating event. Following such an

event, but mediated by the occurrence and severities of subsequent disturbances, the vertical structure of a forest stand tends to grow more complex over time. More favorable site conditions such as organic-rich, fertile soils and ample moisture also promote structural complexity. With full development, the vertical structure includes a canopy overstory, understory, a shrub layer, and an herbaceous layer. In achieving such development the forest passes through several stages. These include a stand initiation stage in which seedlings and saplings dominate and new species may continue to arrive; a stem exclusion stage in which the canopy closes, shading out shorter individuals; an understory re-initiation stage in which shade-tolerant species grow as seedlings and saplings; and finally an old-growth or steady-state stage. In the old-growth stage, the overstory typically includes both canopy dominants and subdominants (the latter with crowns only partially in sunlight) together with understory and shrub layers made up of mature, shade-tolerant individuals. Old-growth stands can be identified through a few key characteristics, including a distribution of age and size classes of trees, the absence of saw-cut stumps, and the presence of decaying logs the size of overstory trees.

The understory in a structurally complex temperate forest stand comprises trees and shrubs that spend their entire life cycle there as well as young or suppressed individuals of potential canopy-dominant species. Understory-tolerant species are those that can survive in, or even require, the shade of a forest canopy (e.g., sugar maple, *Acer saccharum*). In old-growth stands or those not recently disturbed it is common to see shade-tolerant species in both the understory and overstory because the overstory trees are those that regenerated in the shade of the canopy. Some temperate forests have a dense layer of understory shrubs, for example *Kalmia* spp., *Rhododendron* spp., and *Vaccinium* spp. (blueberry). The herbaceous layer of a temperate forest commonly contains mosses, lichens, vines, and forbs. Many shrubs and herbs are adapted to low-light environments or grow before canopy leaf extension in the spring or after overstory leaf abscission in fall; in summer only about 10% of full sunlight reaches the herbaceous layer, but this figure can rise to 70% in deciduous stands in winter. Shrubs and herbs that require more light grow in well-lighted gaps or extend their crowns into openings. Vines grow into forest canopies to access light and may be plentiful following a disturbance that kills canopy trees but leaves the dead trees standing.

Soils and Woody Debris

Soils provide a physical rooting medium, the capacity to store and release water, and the capacity to store and release nutrients for growing trees. The soils of the temperate forest regions occur in five orders of the system of soil taxonomy, namely Spodosols, Alfisols, Ultisols, Entisols, and Inceptisols. They range from somewhat infertile (Spodosols) to quite fertile (Alfisols). Spodosols are characterized by a heavily leached surface mineral horizon and a deeper accumulation of Al and Fe-rich organomineral complexes. Spodosols form under coniferous or mixed forests in relatively cool regions with substantial hydrologic leaching, particularly at the northern borders of the biome in the Northern Hemisphere. Further toward the subtropical in cooler areas of eastern North America, Europe, and parts of Asia and Australia, Alfisols form, characterized by organic-rich mineral soil horizons throughout the soil profile, moderate leaching and high fertility. Ultisols, the oldest and most highly weathered soils in temperate zones, are located in the unglaciated and warmer portions of the biome, including southern North America, Asia, Australia, and New Zealand. Because of their advanced age and weathering, these can be deep soils with relatively poor fertility. Inceptisols and Entisols, the youngest soils characterized by less weathering and poor horizon development, are widely distributed in temperate forests. In particular, these form in areas where glaciers left behind new parent material either as till or outwash.

A characteristic that distinguishes temperate from tropical forest soils is the much larger stores of soil organic matter typically present in temperate soils. In temperate regions, litter in various stages of decomposition from fresh litter to humified matter often accumulates atop the mineral soil, forming the forest floor. This organic layer is key in retaining water, retaining and releasing nutrients, and providing animal habitat. It varies in thickness from a few centimeter to tens of centimeters, depending on the age of the stand, the soil pH, the inherent decomposability of the species of litter, the amount of rainfall, and the presence or absence of earthworms.

An additional important category of organic detritus found in many temperate forests is coarse woody debris. This includes standing dead trees and downed, decomposing logs. Rotting woody debris provides a rooting medium, a habitat for soil fauna, a substrate for the saprotrophic flow of energy to the food web, and a means for returning nutrient elements to soils, as well as important structural material for forest streams. Logs undergo a wide range of decay rates, from relatively rapid (a few years) where logs are small and wetting–drying cycles are rapid, to very slow (lasting to a century) where logs are large and the environment is wet and cool. In harvested or managed forests, coarse woody debris may be absent because logs are removed for timber. In unmanaged temperate forests, the long time periods needed for large logs to be produced and decomposed produces a U-shaped curve in the mass of woody debris over time (**Fig. 3**). After a stand-initiating disturbance, woody debris from the previous stand accumulates rapidly and then decays slowly. A lag time of several decades typically exists before woody debris from the new stand begins to accumulate. If the new stand remains even-aged, a second peak may occur as the new stand passes through the stem exclusion stage of development and widespread mortality occurs in smaller trees that compete unsuccessfully for light after the canopy has closed.

Ecological Communities and Succession

Vegetation Communities

Temperate forest vegetation communities span the range from single-species stands to mixed-species stands as well as the range from even-aged to all-aged stands. Which type of community is present at any point in space and time depends on the site

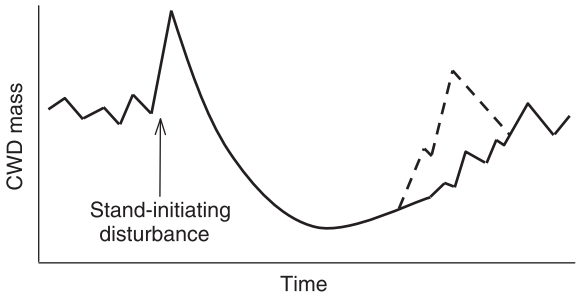


Fig. 3 Dynamics in the mass of coarse woody debris (CWD) before, during, and after a major stand-initiating disturbance in a temperate forest. The solid line represents a U-shaped curve in CWD mass over time. The dashed line represents a secondary peak that may occur if the newly initiated stand remains even-aged and undergoes a self-thinning stage.

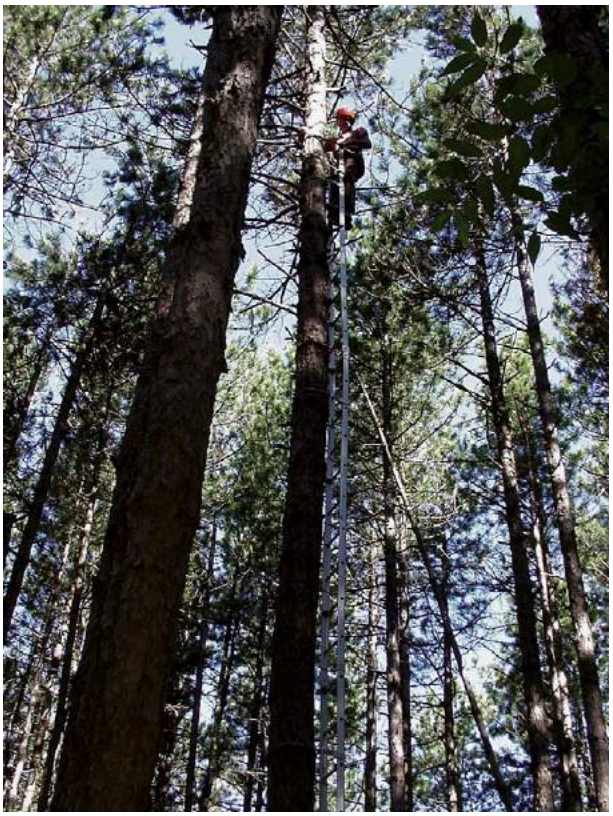


Fig. 4 The canopy of an even-aged red pine plantation, aged about 75 years, in Massachusetts, USA. Photo by W. S. Currie.

physiography, soil, and climate together with its disturbance history. Species such as pines, eucalyptus, cottonwoods (*Populus* spp.), and others may form natural single-species, even-aged stands (Fig. 4). Pioneer species such as aspens (*Populus* spp.) and some pines may initially form even-aged monocultures which eventually diversify in composition and vertical structure as growth, self-thinning, or succession proceeds. Long-lived hardwoods and other conifers also form stands where, increasingly with forest age, great diversity exists in tree ages and sizes. An example of the latter are hemlock-northern hardwood forests of the Great Lakes region of the United States. If horizontal structure and heterogeneity are taken into account, small patch mosaics of even-aged forests of varying ages form larger landscapes of mixed-aged stands, known at the landscape scale as a shifting-mosaic steady state.

It is easy to observe apparent associations of forest trees that occur at certain scales, for example old-growth hemlock–sugar maple (*Tsuga canadensis–Acer saccharum*) stands that form at the scale of square kilometers, or the oak–hickory (*Quercus–Carya*) associations that form more loosely in secondary forests over hundreds of thousands of square kilometers. However, a long-standing debate concerns whether forest communities represent organized associations or simply continuously varying associations as tree species respond individualistically to environmental gradients.

Temperate forest tree species form apparent associations with one another and with the abiotic environment not only across space but also over time at a particular location. A key organizing principle in understanding such temporal associations is the concept of succession, or the replacement of one dominant species or set of dominant species with another, over time, on a particular soil. Primary succession refers to the replacement of species over time occurring in the first forest stand to grow on a newly exposed soil, for example, following the retreat of a glacier. Secondary succession refers to species replacement over time following a major disturbance such as massive windthrow, mortality, or forest harvest. Early-successional species, termed pioneers, are those that are able to fix nitrogen from the atmosphere (see the section titled Nutrient cycling) or those that grow rapidly under high-light conditions but cannot tolerate shade. Late-successional forest species are typically those that can tolerate low-light or low-nutrient conditions as understory trees, while continuing to grow over long time periods, eventually reaching the overstory. Forest ecologists have long sought general principles of succession – for example, the identification of a deterministic sequence leading to a particular stable endpoint or `climax' vegetation community in a particular climate and physiographic landform. Current understanding, however, emphasizes that while certain successional mechanisms exist, the particular sequences and possible endpoints of succession at a particular location are typically numerous, ultimately depending on a complex interplay among competition, species arrival, regeneration, disturbance regimes, and species' modification of the environment.

Temperate Forest Fauna

Faunal biodiversity in temperate forests is not as great as that in tropical forests, but is greater than in boreal forests. Because temperate forests are highly seasonal in their climate and cycles of vegetation physiology and production, faunal life cycles, ecology, and populations are often tied strongly to the seasons. Animal habitats within temperate forests are numerous and heterogeneous, including soils, the forest floor, woody debris, woody stems, and the layers of vegetation canopies. Although some animals depend on particular tree species, many are more dependent on certain aspects of forest structure.

In the temperate forest, the greatest concentration of fauna is on and just below the forest floor, in the litter, humus, and soil. Animals not only inhabit these strata, but through their activities drive soil carbon and nutrient cycling. Also within these strata are gradients of moisture, temperature, gases, and organic matter. Soil microhabitats are pore spaces, water film on soil particles, plant remains, the rhizosphere, and tunnels and burrows. Together, soil fauna and saprophytic flora contribute to the decomposition of organic matter. While most decomposition and nutrient release take place in the warm, humid summers, coinciding with the growing season of the vegetation, microorganisms and invertebrates can remain active below the insulating winter snowpack. Some animals occupy the litter in summer and move to the mineral soil in winter.

Because of the moist soil conditions, many temperate forest floors are home to reptiles (turtles and lizards) and amphibians (toads, frogs, newts, and salamanders). In the mixed-deciduous temperate forests there are over 230 species of reptiles and amphibians. These animals live on the forest floor close to streams, depressions, or lakes where there is available moisture. Lizards are found in moist woods and also in disturbed areas. Turtles live in or near bodies of water and toads and frogs are widespread, needing only shallow water. Temperate forest streams and rivers can support abundant fish populations, particularly under less-disturbed conditions and in coastal temperate rainforests.

Mammal populations in temperate forests tend to be comprised of scattered individuals or groups, and their habitat ranges from the forest floor to the canopy layers. Examples of small mammals are squirrels, rabbits, mice, chipmunks, skunks, and bats. Very large mammals are the exception and in temperate forests may include bear, mountain lions, deer, and other ungulates such as moose and elk. These mammals depend on the herb and shrub layers of the forest in addition to the litter and woody debris for food and habitat. Edge areas form transition zone habitats; for example, deer and other large animals usually live near the edges of forest openings with the trees providing shelter while edible ground vegetation is available in the openings all year.

Trunks are also habitats for spiders, beetles, and slugs. Birds are especially versatile across habitat structures; they are found on the forest floor and in several of the vegetation layers depending on nesting and foraging preferences. Types of birds that breed in mixed-deciduous forests include bark foragers (woodpeckers, flickers), canopy gleaners and pursuers (chickadee, vireo, flycatchers), ground species (thrushes, ovenbirds), and warblers. Deciduous forest are also breeding habitat for larger avian species including turkeys, vultures, owls, and hawks. In addition, moths, butterflies, and other flying insects feed and reproduce in the canopy, the understory, and the forest floor.

Water and Energy Flow, Nutrient Cycling, and Carbon Balance

Water, Evapotranspiration, and Energy

Water enters temperate forests as rainfall, snowfall, fog, and the direct condensation of water vapor onto plant or soil surfaces. Some water, amounting to less than 10% of rainfall under most conditions, is lost immediately to the atmosphere through evaporation. Depending on the season, water drips from the forest canopy to enter soils or accumulates as snow until a mid-winter thaw or spring snowmelt. Entering the soil, water is stored, taken up by plant roots, or moves to groundwater or surface water. Water taken up by plants moves upward through the xylem and exits as water vapor through leaf stomates in the process of

transpiration. Typically, through the combined processes of evaporation and transpiration, less than half of the annual precipitation is passed directly back to the atmosphere as water vapor. Somewhat more than half of the annual precipitation passes through the rooting zone of the soil to enter groundwater or surface water such as streams and lakes.

Evapotranspiration, or evaporation and transpiration taken together, makes a large contribution to the ecosystem energy budget and to the regulation of temperature. In the conversion of liquid water to gas, evapotranspiration carries away large amounts of heat as latent heat. This cooling effect combines with other terms in the energy budget of a forest canopy to regulate the temperature of leaves and of the forest as a whole. Other major terms in the energy budget include the absorption or reflection of short-wave (sunlight) and long-wave radiation (from sunlight and from the atmosphere), the emission of long-wave radiation, and the gain or loss of sensible heat from the atmosphere. On a typical summer day the vegetation canopy absorbs energy in short-wave radiation from the Sun and dissipates the energy as sensible and latent heat to the atmosphere, heating the troposphere from below. On warm days with strong sunlight, the ability of forest tree canopies to dissipate heat allows the trees to maintain leaf temperatures closer to the photosynthetic optimum while also minimizing plant respiration. The opening and closing of stomates, governing transpiration, is under plant physiological control and is an important aspect of plant adaptation to life in a particular environment. During prolonged periods of drought, when trees are less able to use water to cool the canopy and maintain leaf turgor, foliar wilting and tissue damage can occur. Some temperate forest trees can be unexpectedly drought-deciduous, dropping their foliage during a late summer drought.

The photosynthetic conversion of light energy to stored chemical energy is a minor term in the physical energy budget of a forest, amounting to no more than 2% of the energy in sunlight. At the same time, this energy conversion represents the largest term in the ecological energy budget of a forest. The energy stored in photosynthate drives the life processes of all of the plants and animals in the ecosystem. A large portion of this energy is consumed by the vegetation itself through plant respiration, supplying energy for growth, metabolism, and reproduction. Another large flow of energy enters the food web through herbivory; herbivores eat seeds, fruits, and living plant tissues. The consumption of living leaves by insects, while normally minor, can grow during insect irruptions to encompass virtually the entire forest canopy over large areas. Similarly, the consumption of living leaves by forest ungulates including deer and moose are typically small energy fluxes at the ecosystem scale (although the browsing of seedlings and saplings can have a strong impact on forest regeneration and the future composition of the vegetation community). The chief means of energy flow to the faunal food web is through the saprotrophic pathway. Fungi and bacteria (often called soil flora) decompose dead and senesced plant material including leaves, roots, and woody debris. The soil flora is grazed upon by soil microfauna, which are in turn preyed upon by other fauna including arthropods, amphibians, and birds.

Nutrient Cycling and Carbon Balance

To achieve the high levels of productivity typical of temperate forests, trees require ample and reliable supplies of nutrient elements. Those required in the largest supplies include N, P, K, Ca, Mg, S, and Mn. Trees acquire most of their nutrients through root uptake from soils, which store nutrients in soil solution, on the surfaces of mineral grains, on the surfaces of organic matter, and in decomposing organic matter itself. A forest ecosystem receives inputs of nutrients from the atmosphere and from mineral weathering, experiences losses of nutrients via leaching (the water-driven movement of elements out of the rooting zone, ultimately to streams), and cycles nutrients internally (**Fig. 5**). A key internal cycle is the plant–soil cycle in which an element such as calcium (Ca) is taken up by plant roots, used nutritionally by the tree, returned to the soil in foliar litterfall, and returned to the pool of soil-available nutrients during decomposition of the litter. Temperate forests are characterized by the fact that, for most nutrient elements required for plant growth, the internal cycling is greater than the ecosystem inputs and losses of these elements.

While most of the required nutrient elements can be released through mineral weathering, a notable exception is nitrogen (N). Temperate forests rely on inputs of N from the atmosphere. Combined with the fact that trees have a high demand for N, and with the fact that N is strongly retained in unavailable forms in soil organic matter, this makes N the most limiting nutrient for plant growth in most temperate forests. Trees have a high demand for N because photosynthesis and plant metabolism require enzymes, which are made of N-rich amino acids. Amino acids are also one of the primary needs of herbivores that consume plant tissues, including defoliating insects, deer that browse saplings, and beavers that girdle trees by eating the cambium around the tree base. Given the high demand for N by forest trees, it is somewhat ironic that trees in temperate forests are surrounded by two large, potential sources of N that are limited in availability because of the chemical form of the N. The first is N_2 gas, which is the primary constituent of the atmosphere. Most forest trees cannot access gaseous N_2, although a few exceptions include red alder (*Alnus rubra*) and black locust (*Robinia pseudoacacia*) in the USA, which access atmospheric N_2 through the process of N fixation (**Fig. 5**). In this process, symbiotic bacteria living in root nodules fix the N_2 into plant-available forms. The second large pool of poorly available N occurs in humus and soil organic matter, made up of partially decayed and humified plant and microbial detritus. Typically, large accumulations of N are bound in this material in large, polyfunctional macromolecules that form during litter decomposition. Temperate forests are characterized by the combined facts that (1) cold, wet winters impede microbial decomposition and allow these pools of organic matter to accumulate, and (2) warm, humid summers promote decomposition by fungi, causing these soil organic matter pools to turn over and release nutrients at slow but continuing rates. Nutrient release during decomposition is termed mineralization because N is converted from organic to the inorganic forms of nitrate (NO_3) and ammonium (NH_4) which are easily taken up and used by plants (**Fig. 5**).

Carbon is the primary elemental constituent of both forest vegetation and the organic matter in forest soils. Carbon (C) is not considered a nutrient element *per se* because a C atom passes through a forest once, in a single direction, closely linked to the flow of energy; unlike nutrients, carbon does not cycle between plants and soils repeatedly. Forests are highly open systems with respect

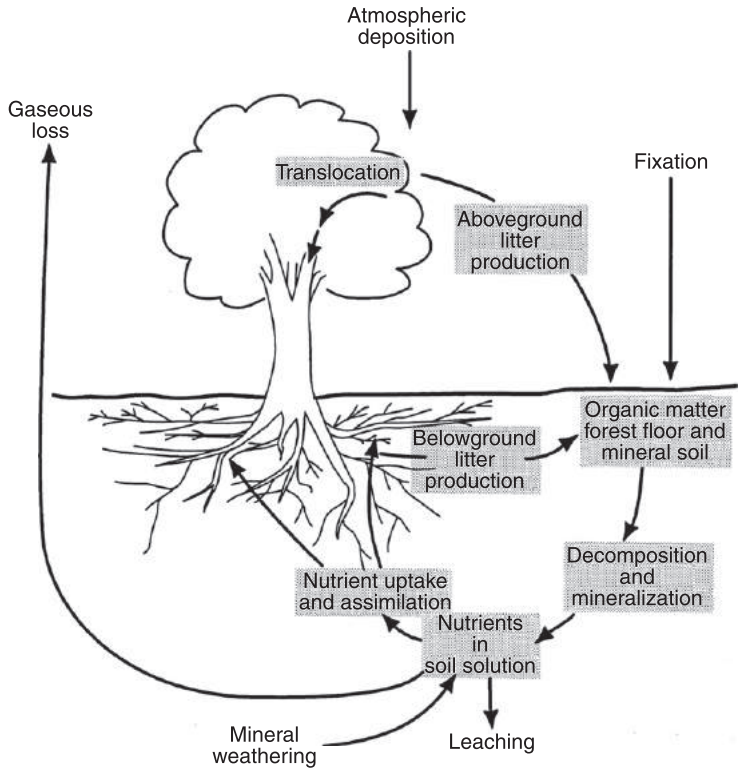


Fig. 5 Schematic diagram of generalized nutrient cycling in temperate forests. Shaded terms represent nutrient cycling fluxes within the system, while unshaded terms represent ecosystem inputs or losses. From Barnes, B.V., Zak, D.R., Denton, S.R., Spurr, S.H., 1998. Forest Ecology. New York: Wiley.

to carbon, exchanging large quantities of CO_2 with the atmosphere. The carbon balance of a temperate forest arises from the interplay among processes controlling forest sources and sinks of atmospheric CO_2. Photosynthesis, or primary production, converts atmospheric CO_2 to reduced organic compounds, storing energy and C in the forest. Autotrophic respiration, the conversion of organic compounds to CO_2 by plants, provides energy for plant metabolism. Heterotrophic respiration, the conversion of organics to CO_2 by herbivores, microorganisms, soil fauna, and other animals in the food web, releases energy for animal life processes. Fire, the rapid oxidation of organics, also releases CO_2 to the atmosphere. Depending on the balance among these processes, temperate forests can either store or release large quantities of carbon. The primary storage pools include growing trees (particularly the woody stems), the forest floor, standing and downed woody debris (Fig. 6), and soil organic matter. The transfer of carbon among these pools is linked to forest disturbance and stand dynamics including aggradation and succession. The flows of carbon into and out of the ecosystem are closely coupled to the availability of water, flows of energy, and the cycling of nutrients.

Temperate Forest Land Cover

Historical Land Cover and Land-Cover Change

Temperate forests in all regions of the globe have been significantly altered by human activities for thousands of years. Their moderate climates, fertile soils, and vegetation productivity have been favorable to human settlement and clearing for agriculture, as well as direct use of trees themselves for lumber and fuels. Agricultural and settlement activities have included development of urban areas, widespread grain and other crop (e.g., corn, vegetables) cultivation, livestock grazing, gathering of mulch, and alteration of natural water drainage. Under these historical pressures, it is estimated that only 1–2% of the original temperate forest remains as never-harvested remnants scattered around the globe. The vast majority of temperate forest land cover is in secondary forest responding to human harvest or other human-induced disturbance.

The longest histories of substantial forest clearing have been in Asia and Europe. In China clearing for agriculture probably began some 5000 years ago, where the Chinese civilization is believed to have begun around the Huang He (Yellow River). The primary sociopolitical factor contributing to deforestation of China over the centuries has probably been the focus on an agriculture-based economy. At present, there is negligible large-scale reforestation in temperate China and significant soil erosion problems hampering reforestation.

Fig. 6 An old-growth sugar maple–birch–hemlock forest showing a large piece of downed woody debris. This forest is located in northern Michigan, USA. Photo by W. S. Currie.

Forest clearing for agriculture in Europe began over 5000 years ago starting in present-day Turkey and Greece and moving northwest through Middle Europe to Northern Europe. Forests of Britain were substantially cleared for agriculture and grazing. Woodlands regained some area in the Middle Ages; however, even remaining European temperate forests were degraded, being used for fuelwood, woodland pasture, and later for charcoal. Coppice practices promoted species that re-sprouted more quickly than beech – including maples and oaks, and this activity altered the natural floristic composition. Tall trees in Britain and Western Europe were removed for shipbuilding. Manorial estates provided some of the few refuges for natural forests. Reforestation in recent centuries in Europe began subsequent to reduction in the use of woodlands for pasture and fuel; reforestation has also occurred through the introduction of planted managed forests and scientific forestry. However, spruce, pine, and larch have been widely planted on areas previously occupied by once deciduous temperate forests.

North American indigenous populations cleared or burned small areas for some agriculture, but land-cover change in North American temperate forests began at large scales in the late sixteenth century with the European settlement. Eastern North American was rapidly cleared as the population moved westward in the nineteenth century. By the start of the twentieth century only a small amount of the original North American temperate forest remained. When the richer soils of the topographically level Midwest and Great Plains were found to be more productive for agriculture than those of eastern North America, eastern farms were abandoned and natural forests began to re-grow. At present, secondary forests are regrowing in the eastern and central United States.

In the Near East the temperate forest occurs in a narrow belt including in Turkey and Iran. This area probably served as a plant refugium during the Ice Ages and the floristic composition is more diverse than that in Europe. Some forests have been exploited for coppice, timber, or grazing and others transformed into agriculture and fruit-tree plantations. Beech forests are the most significant of the present-day broad-leaved forests in the region. In the small area of temperate deciduous forest in South America, forests have been moderately altered since the arrival of the Spaniards in the sixteenth century; the further south one goes the more recently the vegetation has been undisturbed and wooded areas remain. Australia first saw introduction of European agricultural practices only approximately 150 years ago.

Present-Day Land Cover and Rates of Change

The global temperate forest continues to be changed by a combination of long-term effects of historical land-cover change and by present-day change agents. Present-day drivers of land-cover change in temperate forests include accelerated population growth, continued industrialization, and changes in agricultural practices. These are expressed on the landscape as continued clearing for settlement and agriculture in some regions, abandonment of agriculture and reforestation in other regions, and widespread alteration in landscape spatial structure and biodiversity.

While rates of tropical deforestation increased between 50 and 90% in the 1980s, the area of temperate forests has remained constant or increased in the last 50 years in the form of new second-growth forests. In some areas, in eastern North America and parts of Northern Europe, farming is less economically viable than in other parts of the temperate region, leading to reforestation in these areas. Preservation in the form of parks has expanded by active conservation efforts worldwide. Managed forestry has maintained existing temperate forest lands by re-planting after harvest, and sustainable forestry practices are receiving increasing attention.

While the temperate forests may have stabilized or increased in terms of total area, most regions continue to experience other alterations manifested in the landscape spatial patterns and forest biodiversity. Today the temperate forest biome is a mosaic of settlements, patches of forest, and agriculture. Large expanses of unbroken forests from past centuries have been replaced by

considerable landscape-scale heterogeneity and fragmentation. Temperate forest communities have changed compositionally, as disturbance regimes have shifted from natural to a combination of natural and human-caused, producing different patterns of regeneration and succession. While some recently established nature preserves have a natural forest structure, reduced biodiversity characterizes many temperate managed and secondary forests. Considerable present-day challenges lie in understanding and addressing the impacts of land-use change and other aspects of global environmental change in the temperate biome on forest biodiversity and forest ecology.

See also: **Ecological Complexity:** Goal Functions and Orientors. **Ecological Data Analysis and Modelling:** Modeling Dispersal Processes for Ecological Systems. **Ecological Processes:** Nitrification. **Ecosystems:** Riparian Wetlands. **Evolutionary Ecology:** Colonization. **General Ecology:** Biomass; Carrying Capacity

Further Reading

Bailey, R.G., 1998. Ecoregions: The Ecosystem Geography of the Oceans and Continents. New York: Springer.
Barbour, M.G., Billings, W.D., 2000. North American Terrestrial Vegetation. Cambridge, UK: Cambridge University Press.
Barnes, B.V., Zak, D.R., Denton, S.R., Spurr, S.H., 1998. Forest Ecology. New York: Wiley.
Currie, W.S., Yanai, R.D., Piatek, K.B., Prescott, C.E., Goodale, C.L., 2003. Processes affecting carbon storage in the forest floor and in downed woody debris. In: Kimble, J.M., Heath, L.S., Birdsey, R.A., Lal, R. (Eds.), The Potential for U.S. Forests to Sequester Carbon and Mitigate the Greenhouse Effect. Boca Raton, FL: Lewis Publishers, pp. 135–157.
Frelich, L.E., 2002. Forest Dynamics and Disturbance Regimes. Studies from Temperate Evergreen-Deciduous Forests. Cambridge Studies in Ecology. Cambridge, UK: Cambridge University Press.
Lajtha, K., 2000. Ecosystem nutrient balance and dynamics. In: Sala, O., Jackson, R.B., Mooney, H., Howarth, R.W. (Eds.), Methods in Ecosystem Science. New York: Springer, pp. 249–264.
Olson, D.M., Dinerstein, E., Wikramanayake, E.D., *et al.*, 2001. Terrestrial ecoregions of the world: A new map of life on earth. BioScience 51, 933–938.
Rohrig, E., Ulrich, B., 1991. Temperate Deciduous Forests. Amsterdam: Elsevier.

Temporary Waters

EA Colburn, Harvard University, Petersham, MA, USA

Overview

What Are Temporary Waters, and Why Are They of Interest Ecologically?

Temporary waters are shallow lakes, ponds, pools, rivers, streams, seeps, wetlands, depressions, and microhabitats that contain water for a limited period of time and are otherwise dry. They occur across the globe, on all continents and oceanic islands, at all latitudes, and in all biomes, wherever water can collect long enough for allow aquatic life to develop.

Numerous and widespread, many temporary waters are small and easily studied. Their communities are diverse, with much among-site variation (i.e., high β diversity), and differ from those in permanent waters, contributing to regional (γ) biodiversity. Endemic species are often present. Organisms survive through species-specific behavioral, physiological, and life-history adaptations. Community composition and structure change in response to environmental variations. Temporary waters are highly productive and their food webs are relatively simple. For all of these reasons, temporary waters lend themselves to surveys and experimental manipulations designed to test hypotheses about biological adaptation, population regulation, evolutionary processes, community composition and structure, and ecosystem functioning.

In many parts of the world, most temporary waters have been lost. The conservation and restoration of vulnerable temporary waters is a major thrust of applied ecology. Also important are applications of ecological understanding to the control of disease vectors, especially pathogen-transmitting mosquitoes, from temporary water habitats.

What Is Covered in This Article?

This article is divided into two sections. The first introduces temporary waters – definitions, important variables, types, geographic distributions, and terminology. The second section examines the ecology of temporary waters, with an overview of the biota and their adaptations, and summaries of some key questions in organismal and community ecology, ecosystem ecology, and applied ecology.

Introducing Temporary Waters

Definition

In temporary waters, aquatic habitat is present for noncontinuous lengths of time, in contrast to permanent water bodies, which are always flooded except under unusual conditions such as extreme droughts. This discontinuity in the availability of water is the defining characteristic of temporary waters.

For this article, temporary waters include temporary inland salt waters, whose chemistry and biota are allied to fresh waters and not to marine ones, but they do not include coastal areas flooded by ocean tides. Also excluded from this discussion are subterranean waters.

Important Variables

Apart from periodic drying, there are no hard and fast rules about the characteristics of temporary waters. Classification may be useful, provided it contributes to understanding. The important considerations governing how to classify temporary waters in a given situation should be: what is the purpose of classification, and what are the desired outcomes in terms of distinguishing different types of temporary water bodies? Researchers have developed many approaches to classifying temporary waters using the descriptive variables listed below.

Geography

Regional location (e.g., Ontario, Malay Archipelago), latitude (e.g., tropical, Arctic), or climate (e.g., humid, arid) may contribute to similarities among temporary waters.

Biome

Temporary waters occur in all terrestrial biomes, even the wettest. Regardless of their location globally, habitats within a particular biome, with similar hydrologic characteristics on similar substrates, often are much alike.

Table 1 Major types of temporary waters found throughout the world

Rockpools or rock pools – Accumulations of rainwater or floodwater in depressions on exposed bedrock or boulders
Rainpools or rain pools – Accumulations of rainwater on any substrate
Seasonal woodland pools – Fill annually, usually as a result of winter or spring rains, and from melting snow in northern areas, and dry later in the year
Grassland pools – Temporary ponds in grassland environments
Marsh pools – Temporary ponds that occur within larger grass, sedge, or rush-dominated wetlands and remain flooded after most of the wetland has drawn down
Swamp pools – Depressions within larger wooded wetlands that remain flooded after the surface of the swamp has dried
Floodplains – Land areas that are inundated seasonally by high waters spilling over the banks of rivers and streams
Floodplain pools – Low areas in floodplains that remain flooded after floodwaters have withdrawn and left most of the floodplain dry
Springs, seeps, and spring seeps – Sources of water derived from groundwater or from subsurface flow reaching the land surface after heavy rains. Springs are expressions of the groundwater table and tend to be relatively permanent; seeps may be more transitory. Both vary in output with rainfall over the source area, and both may provide seasonal or continuous sources of water. Flow from springs and seeps may extend from the source as marshes, pools, or streams that may contain water during cool or wet seasons and become dry during periods of high temperature and/or low precipitation
Intermittent headwater streams – The smallest tributaries at the head of stream systems, often seasonal in their flow, containing water during the wet and/or cooler months and becoming dry during the hot/dry months
Arid-land rivers, intermittent rivers, or ephemeral rivers – Flowing waters that occur in regions where the groundwater table is far below the surface and where annual potential evapotranspiration is greater than precipitation. They typically flow only during the rainy season, when runoff travels over the land and is carried downstream; some only carry storm runoff, but others may have extended flow maintained by seasonal groundwater discharges. During the dry season, there may be water below the surface and in isolated pools within the channel, and there may be brief spates of flow following cloudbursts
Dry lakes or playas – Shallow water bodies in arid regions, especially in closed basins, where water collects from large areas. Due to the arid conditions, the water usually evaporates rapidly. A long history of flooding and drying leads to accumulations of salts in these basins, and dry lakes are typicaly saline. Many dry lakes occupy basins that contained large freshwater lakes earlier in geological history. Deposits of salts and sediments left behind when the lakes dried may be tens or hundreds of meters deep beneath the lake beds, and they contribute to saline conditions in the playa
Sinkholes or sink holes – Depressions created in calcareous bedrock by the gradual dissolution of the rock by water. They range in diameter and depth from meters to kilometers. Sinkholes that contain water are fed by groundwater, precipitation, and/or streamflow and include both permanent and temporary waters
Snowmelt pools, icemelt pools, and meltwater pools – Formed by the seasonal melting of ice and snow in the Arctic and Antarctic, along the margins of icefields and mountain glaciers, and in areas that receive snowfall
Meltwater streams – Flowing waters that develop seasonally as glaciers, icefields, and winter snows melt; they often flow during the day and stop flowing at night as low temperatures inhibit melting
Plant-associated microhabitats or natural containers (phytotelmata) – Microhabitats formed where plants produce small depressions in which water can collect (see **Table 2**)
Artificial containers – Any human-made concavity where water can collect, including gutters, birdbaths, tires, empty cans, tractor ruts, canoes, split and discarded coconuts, and other water-holding depressions

Water body type

Temporary waters may be lotic (flowing) or lentic (still). There are several major categories, and many unique regional names (**Tables 1**–**3**). Some categories overlap; for example, pools formed after thunderstorms on exposed rocks on coastal Scandinavian islands are both rainpools and rockpools.

Substrate

Substrate (e.g., rock, organic debris, sand, clay, limestone, mud, basalt, wood) influences hydrology, water chemistry, and temperature and is an important habitat variable in its own right (e.g., for seed germination or shelter for burrowing animals).

Size

Some classifications distinguish microhabitats, mesohabitats, and macrohabitats.

Hydrology

Hydrologic variables are the most important factors influencing aquatic life in temporary waters.

Water sources

Water sources include groundwater, runoff, precipitation, snowmelt, streamflow, and floodwater.

Flood timing

Flood timing encompasses both season and predictability. Vernal, estival, autumnal, and hibernal (or brumal) refer respectively to filling in spring, summer, fall, or winter. Intermittent systems flood predictably at annual (seasonal) to multiyear intervals. Waters that flood unpredictably are ephemeral if they fill several times a year, and episodic if they fill just once or twice a decade.

Table 2 Examples of phytotelmata and other natural containers that provide temporary aquatic habitats for mosquito larvae and other organisms

Ant nests
Insect-bored bamboo, bamboo stumps
Fungal cap concavities
Log holes
Buttress-root slits
Eggshells
Flower bracts
Fruits
Horns
Leaf axils
Fallen leaves
Nuts
Modified leaves of pitcher plants and analogs
Pods
Reeds
Rockholes, potholes
Mollusk shells
Skulls and other skeletal remains
Stumps and trunk cavities
Treeholes

Derived from Index in Laird, M., 1988. The Biology of Larval Mosquito Habitats. Boston: Academic Press.

Seasonality and predictability of flooding influence the biota. Predictable filling of Mediterranean vernal pools by rainfall during the winter growing season facilitates plant growth and has contributed to the development of an endemic flora. When the pools dry, high summer temperatures prevent the establishment of terrestrial vegetation.

Flood duration, or hydroperiod

Across most categories of temporary waters, there is a continuum of flood duration: days, weeks, months, or years. Ephemeral waters are flooded for hours, days, or weeks. Intermittent refers to flood durations of several months. Semipermanent or near-permanent waters dry only occasionally, during major droughts. Within a water body, the hydroperiod varies across filling cycles, depending on weather, with some waters being more stable than others (**Fig. 1**).

Typically, with increasing hydroperiod, the potential aquatic community becomes richer, and the adaptations of the flora and fauna become less extreme. Waters with shorter hydroperiods have fewer total species but more that are unique to temporary habitats.

Chemistry

Important chemical characteristics include salinity (fresh, $<3\ g\ l^{-1}$ salts; brackish, 3–35 $g\ l^{-1}$; saline, $\geq 35\ g\ l^{-1}$), major ions (e.g., sulfate- vs. chloride-dominated desert waters), color (e.g., clear vs. stained dark with organic acids), pH, and dissolved oxygen.

Distribution of Temporary Waters

Most types of temporary waters occur widely across the world's biomes, from the poles to the equator. Their numbers and varieties vary with annual precipitation, temperatures, and local geology and geography. They are most common in arid or cold areas where liquid water is unable to persist for long periods of time.

Tropical rainforests

Tropical rainforests, although well watered, contain many temporary waters. Cavities in bromeliads and other epiphytes retain rainwater (**Fig. 2**) where decaying organic materials support microorganisms, insects, and amphibians. Rainpools on the forest floor fill, dry within days, and support distinct communities. Lowlands of great tropical river systems, including the Amazon and the Paraná–Paraguay, are inundated during the rainy season, and the retreating floodwaters create a mosaic of ponds that retain water for varying time periods.

Boreal and temperate forests

Temporary waters in deciduous and coniferous forests include rainpools, rockpools, and treeholes. Intermittent headwater streams dry in summer when forest trees are transpiring (**Fig. 3**). Floodplain pools fill in spring or after major storms. Seasonal woodland pools, commonly called vernal pools, fill from groundwater, snowmelt, and spring rainfall and dry in summer (**Fig. 4**). Many are

Table 3 Some terms used to describe temporary waters around the world

Avens	France: depressions hollowed out in limestone
Baias	South America: temporary lakes
Billabongs	Australia: pools that are left behind in floodplains as large, seasonal rivers recede after flooding
Bogs	Worldwide: freshwater peatlands with acidic water chemistry; usually with limited connection to other surface waters, often fed exclusively by rainfall
Buffalo wallows	North America: created by buffalo (*Bison bison*) rubbing their bodies on the ground, these shallow excavations on the prairies fill seasonally with water
California vernal pools	Western North America: seasonally flooded pools in Mediterranean scrub of western North America, especially California, and characterized especially by rich plant communities with large numbers of endemic species
Carolina bays	North America: round or oval depressions of uncertain origin in the coastal plain of the Southeastern United States, often supporting endemic plant communities and temporary-pond fauna
Corixos	South America: temporary-water bodies in floodplains, especially in the Pantanal region
Dambos	Southern Africa: shallow, treeless, seasonally inundated wetlands at heads of drainage networks
Dismals	North America: swamps or marshes in the Mid-Atlantic region of Virginia, Delaware, and the Carolinas
Doline	Western Balkan states/Dinaric Alps: depressions and sinkholes in limestone
Fens	Worldwide: freshwater peatlands with alkaline water chemistry
Gator holes	North America: excavations made by alligators (*Alligator mississippiensis*) in the Florida Everglades; they remain flooded when waters recede and serve as refugia for aquatic animals during droughts
Gnammas	Western Australia: temporary waters formed on granitic outcrops
Heaths	Great Britain: freshwater peatlands with acidic water chemistry
Mires	Northern Europe: freshwater peatlands with acidic water chemistry
Kettles, kettle holes	Worldwide, in areas affected by continental glaciation in the past: largely circular depressions formed by the melting of blocks of ice calved off of retreating continental glaciers and buried in morainal debris
Moors	Great Britain: freshwater peatlands with acidic water chemistry located on hilltops
Mosses	Scotland: raised bogs, i.e., freshwater peatlands with acidic water chemistry located on hilltops or above the groundwater table
Muskegs	North America: freshwater peatlands with acidic water chemistry (Algonquin)
Oshanas	Namibia and Angola: linearly linked shallow pans that are filled by floodwater and precipitation
Pakahi	New Zealand: shallow, groundwater-flooded areas with acid soils, inappropriate for cultivation (Maori)
Pans, panes, pannes	Worldwide: shallow temporary waters that flood periodically from rainfall in arid regions; also refers to temporary pools that form in salt marshes from monthly flooding by spring tides
Phytotelmata	Worldwide: a technical term describing temporary waters associated with plants, in axils of leaves or branches, modified pitchers and similar structures, nuts
Plunge pools	Worldwide: deep holes that form in bedrock at the base of waterfalls through the action of water over time, and that retain water for a period of time after the stream has dried up
Pocosins	North America: upland-coastal floodplains or groundwater-flooded seasonal wetlands in the South Atlantic United States
Potholes, pot holes	Worldwide: rockpools in or along streambanks and streambeds, created by the action of water and rock scouring out round depressions into boulders or bedrock. Potholes may be a few centimeters to more than a meter in diameter
Prairie potholes	North America: in the Great Plains, largely circular depressions formed as blocks of ice left by departing continental glaciers were covered by morainal debris and then melted
Ramblas	Spain: temporary streams that usually flow only after rainstorms
Sabkhas, seabkhas	Arabian Gulf: saline lakes
Salinas	South America: saline lakes
Sinkholes	Worldwide: depressions in limestone, formed by the solution of surface rock or by the collapse of underground caverns or caves collapse where the subsurface has been dissolved by gradual solution in water. Sinkholes may be dry on the bottom, intermittently flooded, or contain water continuously
Sinking creeks	North America: flowing streams that disappear from the surface into one of the many cracks or sinkholes in limestone regions, or into the ground in arid areas
Sloughs	Worldwide: the term has a variety of meanings. In Great Britain it refers to muddy and shallow waters. In North America it is used to refer to prairie potholes, temporary ponds, oxbow wetlands, permanent ponds, deepwater areas in the Everglades, brackish marshes on the west coast, seasonally flowing depressions in forests, freshwater wetlands in the Great Plains. Some use the term to refer to areas where water is not stagnant but rather, flowing slowly; others specifically define sloughs as areas with stagnant water
Swallow holes	Great Britain: sinkholes in limestone, especially deep holes through which water funnels underground
Takyrs	Turkmenistan: pans in the desert
Tenajas	North America: rockpools, usually in temporary stream channels, that remain flooded for several months after the stream dries; some develop plant communities similar to those in vernal pools
Turloughs	Ireland: temporary waters formed in limestone, filled primarily by groundwater although may sometimes fill from precipitation; usually fill in fall and dry in spring or early summer
Vasante	South America: temporary streambeds connecting lakes in the Pantanal region during the rainy season
Vernal pools	North America: temporary woodland pools that fill in spring and dry in summer; applied more broadly to all seasonal woodland pools that reach maximum depth and volume in spring. Worldwide, the term applies to any temporary pools that fill in spring. The term `California vernal pools' is used to represent a class of Mediterranean biome temporary ponds characterized primarily by their endemic plant communities

(*Continued*)

Table 3 Continued

Vleis	Southern Africa: seasonally inundated wetlands in southern Africa, typically flooded by rivers at high water
Whale wallows	Eastern North America: seasonal woodland ponds along the Delaware coast in the United States

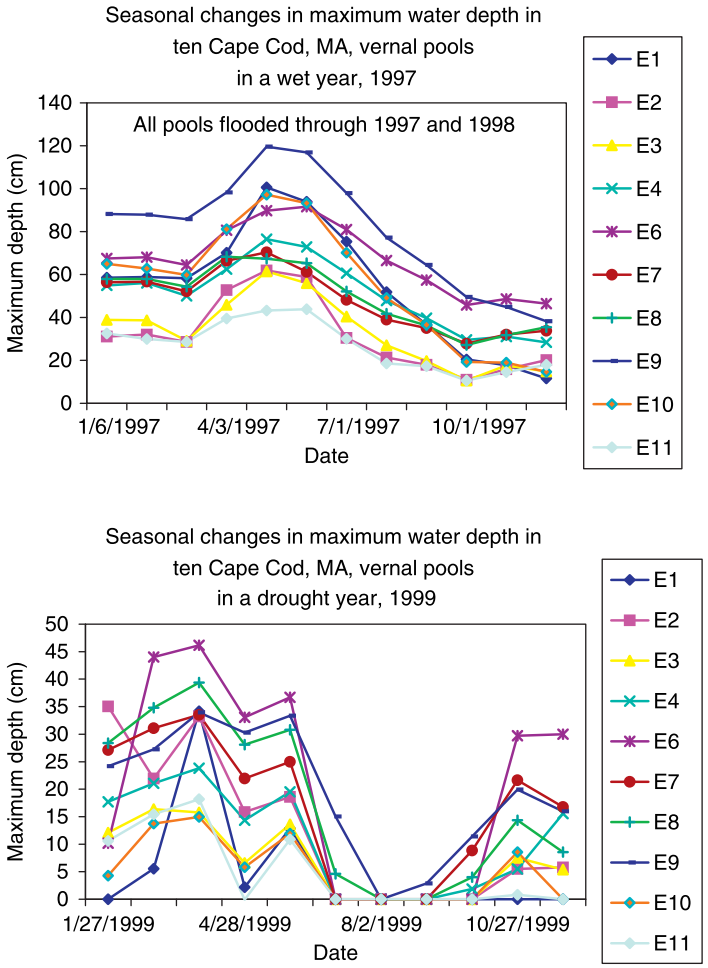


Fig. 1 Water depths differ within and between years in ten temporary ponds clustered together on Cape Cod, MA, USA.

Fig. 2 Bromeliads and other plants serve as natural containers for rainwater and provide microhabitats for microorganisms, mosquito larvae, and some tropical amphibians.

Fig. 3 Intermittent headwater streams drain up to 80% of the landscape in temperate forests and support distinctive communities of aquatic invertebrates and stream salamanders.

Fig. 4 Temporary woodland ponds, or vernal pools, are common in temperate and boreal forests. Water levels vary significantly over time. In this pool, normal high water reaches the base of encircling maples, and in wet years it is more than a meter deep.

important breeding habitats for amphibians, crustaceans, and aquatic insects. Carolina Bays in the Southeastern United States, and other previously unglaciated systems, support endemic plants.

Tundra and icefields

Where temperatures are cold and the growing season is short, temporary waters appear when the summer sun melts glaciers, ice, and snow. For a few months, Antarctic rockpools, high-mountain ponds, and myriad shallow water bodies perched over permafrost in Arctic tundra teem with bacteria, protozoans, planktonic crustaceans, and insect larvae. This broth of aquatic life provides food for nesting birds which flock in hundreds of thousands to high latitudes to raise their young.

Fig. 5 A salt crust left by evaporating water overlies dry lakes, or playas, in many desert basins.

Mediterranean scrub

The Mediterranean scrub biome occurs along the Mediterranean Sea; from Baja California to eastern Washington; and in parts of Chile, southern Africa, and Australia. Rains during the winter–spring growing season collect above impervious substrates, forming water bodies known as vernal pools, vleis, pans, Mediterranean temporary pools, and gnammas. They support endemic floras, including *Isoëtes* spp.; endemic faunas, including fairy shrimp and other crustaceans; and cosmopolitan temporary-pool plants and animals. On other substrates, temporary pools are less predictable and lack endemics. Most rivers in this biome flow only during the wet season, although isolated pools retain water for part of the dry season. Treeholes and other natural containers provide microhabitats after rains.

Deserts

In deserts, extreme aridity, high temperatures, salinity, and isolation of waters are especially stressful for aquatic life. Brief rainstorms create ephemeral pools on rocks and other surfaces. Extended rains collect water from large areas to fill closed basins, forming shallow, usually saline lakes that leave extensive deposits of encrusting salts upon drying (Fig. 5). Many rivers and streams flow seasonally, especially in wet winters, or flash-flood unpredictably after storms, leaving behind pools of varying permanence (Fig. 6). Permanent springs overflow during winter, creating seasonal streams, marshes, and thickets. Salts accumulate in the soil along the edges of desert waters, and temporary water bodies are generally brackish or saline.

Grasslands

Temporary waters in grasslands include pools, marshes, floodplains, and seasonal rivers and streams. Rich assemblages of plants, invertebrates, and amphibians occur in these waters and are critical for bird populations in the prairie pothole region of North America (Fig. 7); the Eurasian steppes; the Indus, Ganges, Assam, Sylhet, and lower Mekong river plains in Asia; the southern African veldt; and the Pampas, Campos, and Pantanal regions of South America.

The Ecology of Temporary Waters

The Biota

All major groups of freshwater organisms occur in temporary waters. Many families and genera are found in similar habitats throughout the world, and there are some cosmopolitan species.

Hundreds of species of prokaryotes, including photobacteria and bacterial decomposers, proctotists, including green algae and diatoms, and protozoans, including ciliates, flagellates, and sarcodines, have been identified from temporary waters. Plants include mosses and liverworts, ferns, grasses, sedges, rushes, spike rushes, and other taxa typical of local wetlands. Microinvertebrates include rotifers, tardigrades, and gastrotrichs. Arthropods, especially water mites, crustaceans, and insects, dominate the macroinvertebrates. Many microcrustacean species, especially ostracodes and copepods, swim in the water, feed in the sediments, or cling to surfaces. Branchiopod crustaceans, particularly Notostraca (tadpole shrimps) (Fig. 8), Anostraca (fairy shrimps), Conchostraca (clam shrimps), and some Anomola (daphnias and other water fleas), are largely restricted to temporary waters. All aquatic insect orders include temporary-water species, with the largest number in the Diptera, or true flies. Water mites selectively feed on insects and crustaceans. Platyhelminthes (flatworms and flukes), Annelida (segmented worms and leeches), Nematoda (roundworms), Nematomorpha (gordian worms), and Mollusca (snails and bivalves) are also represented. Annual tropical killifishes (Cyprinodontiformes) and African and South American lungfishes (Lepidosireniformes) survive periodic drying, and fully aquatic fishes move seasonally into floodplains and intermittent headwater streams to feed and breed. Most anuran amphibian species, including true frogs, treefrogs, and toads, and some salamanders, preferentially breed in temporary habitats. For many of the

Fig. 6 (a) Seasonal rivers in arid regions flow seasonally. (b) When flow ceases, the pools that remain persist for varying lengths of time.

Fig. 7 Prairie potholes dot the landscape of the upper Great Plains in North America, provide habitat for aquatic life, and support breeding waterfowl. Reproduced from Sloan, C.E., 1972. USGS Professional Paper 585-C Ground-Water Hydrology of Prairie Potholes in North Dakota. Reston, VA: US Geological Survey.

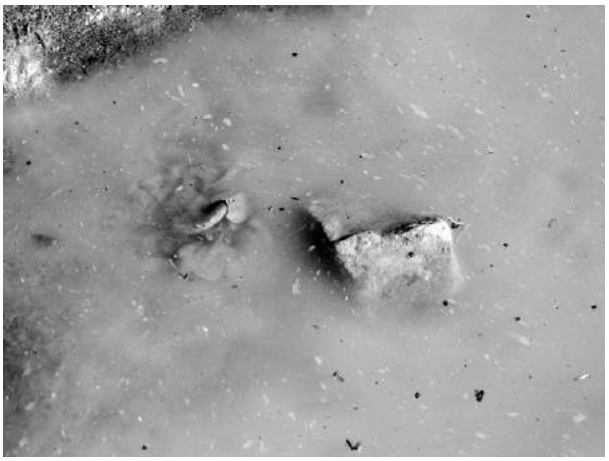


Fig. 8 Notostracan crustaceans, known commonly as tadpole shrimp (left of center), are temporary-water specialists. Diapausing eggs lie for months, years, or decades in sediments of desert playas, rockpools, and woodland ponds. Hatching upon flooding, the animals are voracious predators and scavengers and their presence restricts the distributions of other temporary pool animals.

world's birds, temporary waters are critical food sources during breeding or migration. Reptiles and mammals feed and hydrate in these seasonal waters.

Autecology: Organisms and Populations

Temporary waters lend themselves to thousands of ecological questions about adaptations, population regulation, and evolutionary pathways linking sibling species and cosmopolitan taxa.

Adaptations to drying

Inhabitants of temporary waters are distinguished by their ability to survive periodic drying. Adaptations include diapause, quiescence, and active avoidance. Rapid responses to flooding, fast growth, and flexibility in initiating the drying response maximize organisms' habitat use.

Diapause

Diapause involves suspended development. Hormonally controlled, and initiated and terminated by specific environmental cues, diapause is the most common and most effective drought-survival mechanism. It can allow survival over years – even decades – of continuous drying.

The rapid appearance of living organisms when water fills formerly dry puddles, containers, and floodplains is not, as formerly believed, spontaneous generation, or life miraculously developed from nothing. Instead, much of the life in newly flooded areas emerges from cysts, spores, seeds, or eggs diapausing on the dry substrate.

Found from bacteria to fishes in temporary waters, diapause is common in organisms with limited dispersal. Typically, the organism is replaced by a small, highly desiccation-resistant structure that awaits rehydration in the sediment. The substrate reservoir of diapausing microbes, plants, and animals is termed a seed bank, egg bank, or propagule bank. Diapause also occurs in larval and adult stages. Reproductive diapause is seen in some insects, and certain flatworms and annelids enter diapause after encysting in mucus.

Other dormancy

Other responses to drying involve decreased activity and lowered oxygen consumption. Some bdelloid rotifers, tardigrades, and nematodes survive complete dehydration to revive when flooded. Perennial plants may lose their leaves and die back to subsurface roots, tubers, or rhizomes when water levels recede. African and South American lungfish on drying floodplains encase in mud, breathe air, and more than halve their metabolism. Many mollusks burrow into sediments and estivate. Some insect pupae become dormant, delaying adult emergence. Dormancy is generally less effective than diapause for surviving extended drying or unpredictable flooding.

Avoidance

Anatomical, behavioral, or physiological adaptations can help organisms avoid drying. Some plants extend long roots deep into groundwater (**Fig. 9**). Crayfish excavate burrows that remain flooded after surface drying. Animals in intermittent streams move downward into areas of high moisture or subsurface flow. Some insects and fish migrate between permanent and temporary waters. Amphibians and some insects have aquatic larvae and terrestrial adults.

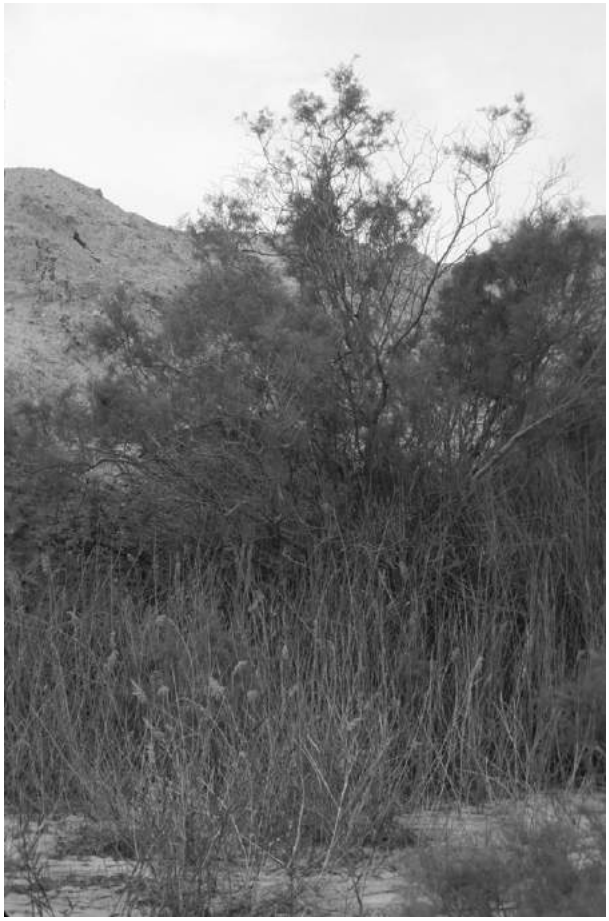


Fig. 9 Along the banks of ephemeral rivers and seasonal waterbodies in arid regions, deep-rooted trees and shrubs such as these tamarisks (*Tamarix* spp.) tap the groundwater and can influence drying of the waters at the surface. A gradient of increasing salinity tolerance is seen in plants radiating outward from the water source.

Physiological ecology

Life in temporary waters may require biochemical modifications and major physiological adaptations. Many endemic plants from temporary waters use C_4 or crassulacean acid metabolism (CAM) photosynthesis, biochemical pathways that use water more efficiently than the C_3 photosynthesis of most plants. Species along salinity gradients show increasing osmoregulatory specializations.

Temporary waters have large local thermal gradients and over time may be subfreezing or above 40 °C. Biological processes vary with temperature, typically doubling with each 10 °C increase. Most species grow within narrow temperature ranges, and thermal cues regulate many life-cycle events. Enzymes need to function over temperature ranges found in temporary water bodies during organisms' life cycles. For example, inhabitants of Antarctic rockpools are active in cold water, and they diapause or secrete antifreeze substances to avoid consequences of subfreezing temperatures; their physiology differs markedly from relatives in temperate or desert pools.

Most freshwater plants and animals cannot regulate internal ionic concentrations in salt water. Inhabitants of many desert waters have impermeable body surfaces, salt-exporting cells, modified life histories, and well-developed drought-resisting adaptations (**Fig. 10**). Many are active in winter, when salinities and temperatures are low. Energetic costs of osmotic and ionic regulation must be compensated for by benefits, such as abundant food or reduced predation. Distributions in desert waters reflect species' physiological tolerances, with no plants and few highly adapted animals in hypersaline pools, greater diversity at low salinities, and low richness in highly unpredictable, ephemeral, freshwater rainpools.

Populations

Bet hedging

Desiccation-resistant seeds of annual plants from temporary waters germinate when chemicals in the seed coat are washed away. Seeds from the same parent have different levels of resistance, ensuring that not all germinate at once, and that some remain in the sediment seed bank. Similarly, crustaceans and some insects form an egg bank comparable to the seed bank of plants; some of the eggs hatch upon flooding, others hatch another time. African and South American annual killifish deposit diapausing eggs into the

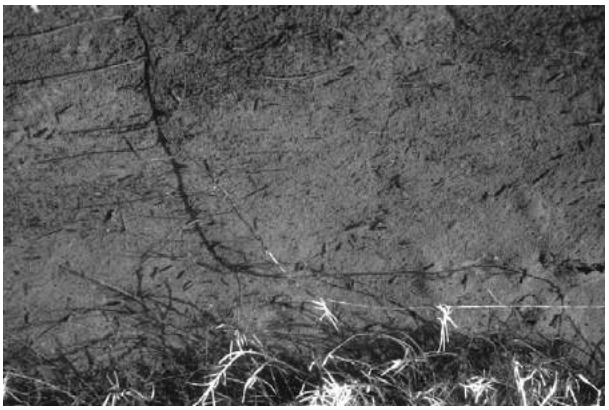


Fig. 10 The stick-like cases of salt-tolerant caddisfly larvae (Insecta: Trichoptera: Limnephilidae: *Limnephilus assimilis*) litter the bottom of Salt Creek in Death Valley, CA, in winter. The ability to regulate hemolymph osmotic and ionic concentrations in brackish waters, rapid growth, adult reproductive diapause during the hot summer months, and the presence of fewer predators than in low-salinity waters contribute to the species' persistence in this temporary desert stream.

egg bank in the sediment of floodplain pools, and these eggs, too, hatch differentially upon flooding. This strategy of spreading risk is predicted from game theory and is termed bet hedging. The new field of resurrection ecology uses egg and seed banks to establish new communities in restoration projects, and obtains insights into evolutionary processes by growing individuals from samples collected a century or more ago and comparing them with modern individuals.

Life-history strategies

Numerous studies address short-term controls on population growth and survival in temporary waters. What are appropriate responses to flooding and drying, when they occur at different times from one year to the next, or to salinity and temperature? Theories of *r*- and *K*-selection predict that some species produce many, small progeny, raising the odds that some will survive. Others produce fewer, but larger progeny with more reserves to support them over adverse conditions. Examples tending toward both extremes can be found in temporary waters. Environmental conditions and evolutionary history shape species' responses, and many cues stimulate the initiation and termination of life cycles (**Tables 4** and **5**).

Short, irregular hydroperiods should favor *r*-selected life histories including rapid hatching/germination upon flooding, fast growth, and timely entry of many propagules into a drought-resistant state. Under longer, predictable hydroperiods, *k*-selection should produce slower growth, larger sizes, and longer life spans. In rockpools worldwide, with hydroperiods from hours to weeks, algae, insects, and crustaceans in the most ephemeral pools complete development in less than 24 h. Life spans are longer in longer-duration pools, and in those with predictable flood regimes. Temperate *Eubranchipus* fairy shrimp and some aedine mosquitoes have one generation per year; they hatch, grow, mature, mate, deposit diapausing eggs, and die. In some fingernail clams, only young individuals resist drying, and they enter obligatory diapause as soon as they are born. Similar patterns are seen in many species.

Life-history tradeoffs, such as the ability to grow while water remains, potentially allow production of more offspring, or development to a larger size, which may enhance survival and reproductive fitness. A longer developmental period may also mean higher intraspecific densities and competition as habitat shrinks, and it increases the risk of being stranded if drying occurs rapidly. *Haematococcus pluvialis*, a photosynthetic flagellate related to *Volvox* from rockpools worldwide, is typical – diapausing spores develop rapidly when flooded, the organisms grow and reproduce, and upon pool drying, the motile cells form aplanospores that withstand drying and high temperatures. The diapausing spores form in less than a day, at any stage in the life cycle, providing *Haematococcus* with flexibility in the face of variable habitat duration. Widely different taxa grow rapidly to a minimum size threshold, after which they can reproduce and grow through multiple generations (e.g., *Daphnia* spp., snails, some bivalves), or become larger (e.g., amphibians, insects), as long as water is present, or until other cues initiate diapause, dormancy, or transformation.

Complex life histories

Many species' life cycles are complex. Post-hatching populations of the cosmopolitan water flea *Daphnia pulex* are all-female and reproduce parthenogenetically while conditions are favorable. When drying threatens, males are produced, and fertilized eggs develop into diapausing, drought-resisting ephippia that lie in the substrate until the next flooding event and temperature cues stimulate hatching. Similar alternating generations occur in some rotifers.

Diving beetles (*Agabus* spp.) have a 2-year life cycle. They hatch from eggs in temporary pools and, when mature, fly to permanent waters to overwinter, returning to pools to breed the following spring. The eggs they leave hatch the following year, before the next wave of adults arrives. Some water-mite larvae that parasitize *Agabus* and other migratory insects are transported by

Table 4 Some cues stimulating the termination of diapause in temporary waters

Hydration
Hydration plus temperature
Hydration plus chemical cues
Hydration plus chemical and thermal cues
Hydration after drying (a minimum period of drying may be required)
Hydration plus chemical cues after drying
Hydration plus chemical and thermal cues after drying
Hydration after drying and low temperatures or freezing (a minimum period of drying and exposure to cold temperatures may be required)
Hydration plus chemical cues after drying and low temperatures or freezing
Hydration plus chemical and thermal cues after drying and low temperatures or freezing
Photoperiod in combination with one or more of the above

Table 5 Some cues initiating life stages adapted to drying in temporary waters[a]

Developmental stage (obligate diapause/dormancy/transformation once development reaches a critical threshold, regardless of habitat favorability)
Developmental stage plus other cues (diapause/dormancy/transformation initiated facultatively after development reaches a critical threshold, only after habitat becomes unfavorable)
Water temperature
Photoperiod
Chemical cues (pH, dissolved oxygen, chemical signals from predators or competitors, salinity, nutrients, other)
Drawdown-associated cues (chemical concentrations, crowding, depth)

[a]Note that if drying occurs before necessary developmental thresholds have been reached, cues cannot initiate drought-resistant stages, and organisms may die without completing their life cycles.

their hosts from temporary waters in fall and back in spring; they then pass through two predatory life stages before laying eggs that hatch into new parasitic larvae.

Dispersal, population maintenance, and evolutionary ecology

How nonmotile organisms disperse has long fascinated biologists and has implications for community composition and stability. Mechanisms include transport by wind; bird feet, feathers, and digestive tracts; water; humans; and insects. Many temporary-water populations are units of metapopulations; they undergo periodic local extinctions and are recolonized from other waters or provide colonizers for other sites. Genetic analysis and modeling help determine the extent of genetic mixing needed to maintain populations or allow divergence.

Endemic species, especially in crustaceans and some plant taxa, are widespread in temporary waters. New species of copepods, anostracans, and other crustaceans are still being identified from all over the world and provide exciting opportunities to understand evolutionary processes.

Community Ecology

Community studies include questions about local and regional biodiversity; community composition and structure in relation to environmental and biological variables and disturbances; patterns of colonization and extinction; predator–prey, host–parasite, and competitive interactions between species; and food webs.

Comparable temporary waters differ in their biota. Distributions of species, and thus community composition, shift along gradients of size, hydroperiod, predictability, and salinity, with richness increasing with decreasing stress. Community composition may change between years, and it can also vary seasonally, with a succession of new hatches and migrants entering waters over time. The presence of potential community members as unhatched propagules in the sediment complicates assessments of community composition and structure.

Community theory

The theory of island biogeography postulates that species richness in isolated habitats is regulated by local extinction and colonization and should vary with habitat size and proximity to potential sources of colonizers. The intermediate disturbance hypothesis predicts high richness in communities subject to a moderate degree of disturbance or stress; according to this model, high stress leads to mortality in all but fast-growing individuals, and under low stress, inter- and intraspecific interactions such as competition and predation determine community structure. Other models look at resource and habitat partitioning/niche diversification, temporal offsets in life histories, and other mechanisms controlling community composition and structure. Studies of amphibians, plants, invertebrates, and algae in temperate woodland pools, Mediterranean temporary pools, Negev and Namibian desert pools, Scandinavian rockpools, Arctic snowmelt pools, and other areas show complex relationships between

community composition and habitat variables such as size, hydroperiod, frequency of flooding, hydrologic predictability, distance from other waters, and salinity. The data suggest that community richness is related to both degrees of disturbance and the predictability of disturbance. Isolation is also important, with greater richness in waters that are connected to larger bodies (e.g., in floodplains) but also fewer taxa specifically adapted to temporary habitats. Species pools in individual water bodies are poor in comparison to the regional set of species (**Table 6**), and experimental assemblages comprised of larger subsets of available species function differently than the smaller natural communities.

Interspecific interactions

Food web manipulations allow examination of relationships among species and show interesting relationships. For instance, algae grow better when grazed by tadpoles than alone. Some potential competitors avoid conflict by preferentially choosing waters with different hydrologic or other characteristics when the other species are present. The survival outcomes for some species of amphibians and insects when they co-occur with competitors depend on which species becomes established first.

Certain aquatic insects, crustaceans, and vertebrates can survive in pools with long flood durations, but they are typically found only at the more ephemeral end of the flooding continuum. They are excluded from the longer hydroperiod pools by predators such as amphibian larvae, tadpole shrimp, and water bugs. The ovipositing females of some species explicitly avoid pools with predators. For example, vulnerable species of mosquitoes avoid laying eggs in pools containing predatory backswimmers, whereas predation-resistant midge larvae do not; American toads (*Bufo americanus*) avoid temporary pools with omnivorous wood frog tadpoles (*Rana sylvatica*).

Ecosystem Ecology

There are many questions about temporary waters as ecosystems. How do tiny, intermittently flooded water bodies produce huge numbers of insects, amphibians, and other organisms? How do nutrients, carbon, and energy flow within temporary waters, and between them and adjacent terrestrial landscapes?

Some temporary waters are among the most productive ecosystems known. In some temporary habitats, photosynthesis by microscopic producers is the base of the food web. For many, from microhabitats in plant leaves to large woodland and floodplain pools, decomposing detritus is the primary energy source. Much remains to be learned about the sources and fluxes of energy and nutrients in temporary-water ecosystems.

Applied Ecology

Vector Control

Mosquito-borne diseases including encephalitis, yellow fever, West Nile fever, and, especially, malaria affect millions of people and are a major focus of world health agencies. Most mosquitoes breed in temporary waters. Their populations have expanded following human alterations of natural habitats, the creation of flooded areas by equipment and land-use change, and the dispersal of water-retaining containers. The effective long-term control of disease vectors requires understanding of the ecology of the pest animals and of their habitats.

Table 6 Regional species pools (β diversity) are greater than local species pools (α diversity), as illustrated by numbers of non-dipteran macroinvertebrates found in early spring from nine adjacent temporary pools on Cape Cod, Massachusetts, USA

Water body	Number of taxa
Pool 1	34
Pool 2	22
Pool 3	24
Pool 4	37
Pool 5	12
Pool 6	38
Pool 7	48
Pool 8	28
Pool 9	22
Total species	89

Modified from **fig. 2** in Colburn, E.A., 2004. Vernal Pools: Natural History and Conservation. Blacksburg, VA: McDonald and Woodward.

Ecological Engineering and Conservation

Bird and amphibian populations and unique aquatic species depend on temporary waters, and the overall contributions of these systems to biodiversity are still being explored. Losses of these habitats are severe (e.g., loss estimates for California vernal pools exceed 90%), and remaining sites face draining, filling, excavation, pollution, water abstraction, invasive species, and climate change. In many regions, seasonal rivers and pools are important water sources; elsewhere, temporary waters provide the only arable areas. Many have been dammed, or converted for rice culture and other crops. Hydraulics, hydrology, surface–groundwater interactions, and biology affect management of these systems for human use and conservation, habitat restoration, and habitat creation.

See also: Aquatic Ecology: Deep-Sea Ecology. Ecological Processes: Volatilization. General Ecology: Demography

Further Reading

Batzer, D.P., Rader, R.B., Wissinger, S.A. (Eds.), 1999. Invertebrates in Freshwater Wetlands of North America: Ecology and Management. New York: Wiley.
Belk, D.A., Cole, G.A., 1975. Adaptational biology of desert temporary-pond inhabitants. In: Hadley, N.F. (Ed.), Environmental Physiology of Desert Organisms. Stroudsburg, PA: Dowden, Hutchinson and Ross, Inc., pp. 207–226.
Caceres, C.E., 1997. Dormancy in invertebrates. Invertebrate Biology 116 (4), 371–383.
Calhoun, A.J.K., DeMaynadier, P. (Eds.), 2007. Science and Conservation of Vernal Pools in Northeastern North America. New York: CRC Press.
Colburn, E.A., 2004. Vernal Pools: Natural History and Conservation. Blacksburg, VA: McDonald and Woodward.
Eriksen, C., Belk, D., 1999. Fairy Shrimps of California's Pools, Puddles, and Playas. Eureka, CA: Mad River Press.
Fryer, G., 1996. Diapause, a potent force in the evolution of fresh-water crustaceans. Hydrobiologia 320, 1–14.
Hartland-Rowe, R., 1972. The limnology of temporary waters and the ecology of Euphyllopoda. In: Clark, R.B., Wooton, E.F. (Eds.), Essays in Hydrobiology. Exeter, UK: University of Exeter, pp. 15–31.
Laird, M., 1988. The Biology of Larval Mosquito Habitats. Boston: Academic Press.
Simovich, M., Hathaway, S., 1997. Diversified bet-hedging as a reproductive strategy of some ephemeral pool anostracans (Branchiopoda). Journal of Crustacean Biology 16 (3), 448–452.
Sloan, C.E., 1972. *USGS Professional Paper 585-C* Ground-Water Hydrology of Prairie Potholes in North Dakota. Reston, VA: US Geological Survey.
Wiggins, G.B., Mackay, R.J., Smith, I.M., 1980. Evolutionary and ecological strategies of animals in annual temporary pools. Archiv für Hydrobiologie (*Supplement*) 38, 97–206.
Williams, D.D., 1987. The Ecology of Temporary Waters. Portland, OR: Timber Press.
Williams, D.D., 2006. The Biology of Temporary Waters. London: Oxford University Press.
Witham, C.W., Bauder, E.T., Belk, D., Ferren Jr., W.R., Ornduff, R. (Eds.), 1998. Ecology, Conservation, and Management of Vernal Pool Ecosystems – Proceedings from a 1996 Conference. Sacramento, CA: California Native Plant Society.
Zedler, P.H., 1987. *The Ecology of Southern California Vernal Pools. Biological Report 85(7.11).* Washington, DC: US Fish and Wildlife Service.

Tropical Ecology☆

Harald Beck, Towson University, Towson, MD, United States

Geography of the Tropics

The tropics include all geographic regions of the Earth that extend from the equator towards the northern hemisphere up to the Tropic of Cancer (23°30′ latitude), and in the southern hemisphere down to the Tropic of Capricorn (23°30′ latitude, Fig. 1). Tropical regions cover only about 7% of the Earth's biosphere but harbor more than 50% of the world's species. Different types of forests dominate the plant community within tropical latitudes; around 58% of rainforests occurs in the Neotropics, which encompasses southern Mexico, Central America, and most of South America. Some 32% of the world's rainforests are located in Brazil, the remaining 42% occur in the Paleotropics, a region including Africa, Madagascar, Southeast Asia, New Guinea, and parts of Australia.

Tropical Climates

Most people imagine the tropics as steamy lush evergreen forests with high humidity and hot temperature throughout the year. However, a wide range of climates occur within tropical latitudes, ranging from snow peaked mountains (i.e., Andes in South American, Mount Kilimanjaro in Africa) to deserts (i.e., central Australia, Kalahari Desert in Africa). However, anthropogenic climate changes have been altering tropical climates, foremost perception patterns and temperature, which has negative effects on the terrestrial and aquatic biotas and adverse socio-economic consequences for humans.

Temperature

Tropical regions receive perpendicular solar radiation at noon almost year-round, thus the mean annual temperature is higher and seasonal changes are less pronounced than in areas at higher latitudes. The intensive solar radiation also increases evapotranspiration (see below).

Typically, low-altitude tropical areas have an annual mean temperature of 26°C (range 23–35°C). However, cloud cover, particularly during the rainy season, can reduce sunlight by absorbing photosynthetically active radiation. A fundamental question is the degree to which sunlight is a limiting factor for plants. Graham and colleagues increased light availability by installing high-intensive lamps above the canopy of a common tree species in Central America. They found that branch growth, number of flower buds, and fruits production increased significantly compared to control trees. Because the researchers quantified additional variables, they suggested that light rather than water or temperature was the main limiting factor for that tree species.

Precipitation

The amount and onset of precipitation is primarily determined by the Intertropical Convergence Zone, an equatorial low pressure system that follows the sun's zenith. Because the evaporative power of the equatorial sun is at its maximum, areas within tropical latitudes receive the highest annual rainfall, except in some locations (i.e., within Africa, Australia). Similar to temperature, rainfall is tied to geography and can range from 0 mm (Chilenian Atacama desert) to over 11,900 mm (Hawaii) annually. Apart from some ever-wet regions (i.e., Sundaland in southeastern Asia, New Guinea), most tropical regions have a predictable annual seasonality with one or two rainy seasons (monsoon) alternating with a dry season. During the dry season, which may range from 1 to 6 months, water lost due to evapotranspiration is greater than the amount of rainfall.

To what extent precipitation is a limiting factor for plants is an essential question and a challenging one to answer experimentally. Researchers have carried out a remarkable large-scale rainfall exclusion experiment in the Amazon over several years. Results have shown that most trees had reduced transpiration and photosynthesis rates, resulting in lower leaf production, reduced trunk growth and stunted saplings growth. Large canopy trees that were fully exposed to sunlight had an increase in annual mortality, from 1% (the typically background mortality) to 9%.

Other factors that contribute to the variations in tropical climates include cold ocean currents (i.e. Humboldt Current), warm ocean currents (i.e. El Niño Southern Oscillation), distance from oceans, and prevailing wind conditions (i.e. trade winds).

☆*Change History:* November 2017. H. Beck made updates to the text and the references section.

This is an update of H. Beck, Tropical Ecology, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 3616–3624.

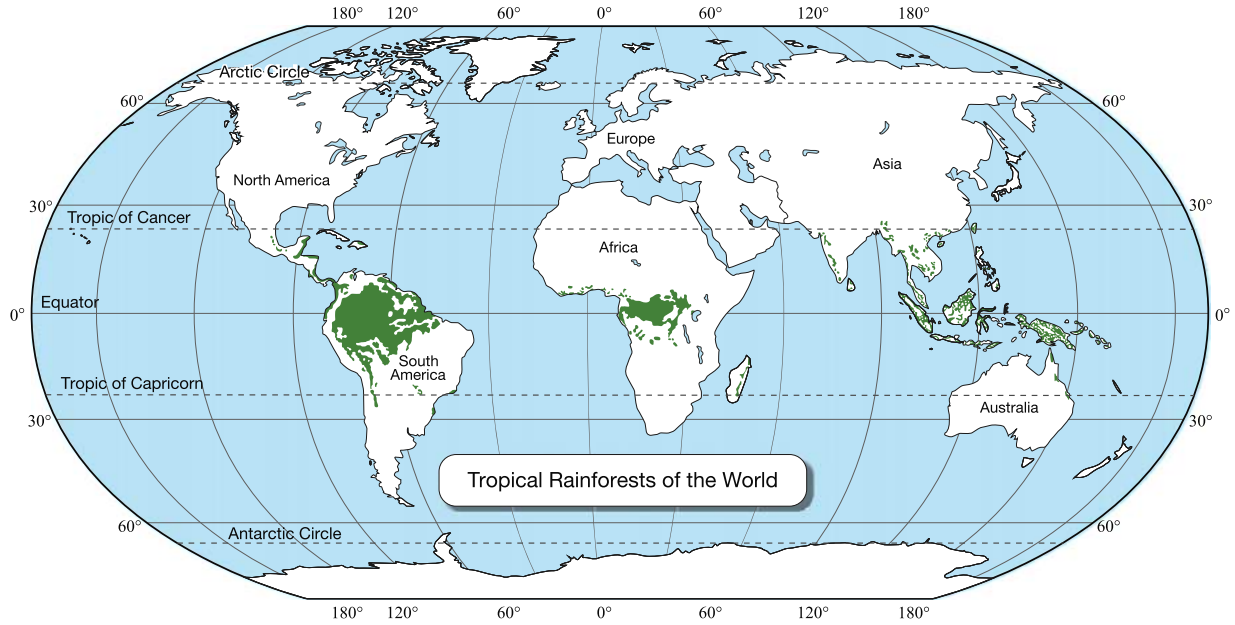


Fig. 1 Approximate distribution of tropical rainforests of the World. The tropical latitudes are centered at the equator and extend northwards up to the Tropic of Cancer (23°30′ latitude) and southwards to the Tropic of Capricorn (23°30′ latitude). Over half of the original rainforests have been destroyed by human activities and currently only around 6% remain; most of these are in different stages of degradation. Map created by Dr. Paporn Thebpanya.

El Niño Southern Oscillation

Only recently, scientists have begun to understand the direct and indirect effects of El Niño events. El Niño events occur every 2–8 years with varying intensity, resulting in interannual climate variation on a global scale. El Niño episodes can lead to below-average rainfall and above average-temperature in some areas (i.e. in Indonesia, New Guinea, West Africa, and Amazonia), whereas in other areas it can lead to abnormally high rainfall, sometimes resulting in floods (i.e. in South America). Studies found that El Niño events can affect plant and animal species across the tropics. Above-average solar radiation during El Niño events in some areas in Central and South America have resulted in drier and sunnier climates, favoring higher fruit production. However, during the subsequent milder dry season the fruit production was unusually low, leading to famine and high mortality of numerous frugivorous and granivorous species. Conversely, in Southeast Asia, after El Niño event, mass flowering and fruiting of many species occurred, triggering migration of numerous animal species, and increased reproduction.

Future studies monitoring the climate, Earth surface temperature, fruit production, and animal densities across numerous geographic locations are needed to better understand the ecological impacts of this phenomenon. Anthropogenic climate changes also affect the frequency, locality, and intensity of El Niño events. Researcher now call those events "extreme El Niños" which disrupted global weather patterns and negatively impact aquatic and terrestrial ecosystems, and socio-economic of humans living in the impacted areas (i.e. drought, flooding, and wildfire).

Seasonality Drives Many Ecological Processes

Scientists found that annual seasonality directly and indirectly affects the ecology of most organisms. For example, many tree species flower during the dry season to optimize cross-pollination by insects. The vast majority of plants that rely on wind for seed dispersal fruit during the dry season. In areas with a prolonged dry season, many tree species are deciduous and produce new leaves at the onset of the subsequent rainy season. Seasonality also can affect recycling pathways of organic matter, nutrient availability and energy flow.

Many studies further demonstrated that plants and animals that experience pronounced seasonality evolved unique adaptation. For instance, most plants that rely on animals as seed dispersers, fruit at the onset or during the rainy season, whereas many mammals (i.e. rodents) cache seeds for later consumption. Another strategy to cope with seasonal food shortage is migration. One of the most spectacular animal migrations can be observed in east Africa, where millions of herbivores including wildebeest, zebras, antelopes, and gazelles migrate to greener pastures. Migration also occurs in other tropical regions such as in Sumatra and Malaysia, where the bearded pigs migrate to track mast fruiting dipterocarp trees.

Long-term satellite climate monitoring, improved satellite animal tracking (GPS) technology, and GIS software will certainly provide new exciting details on the influence of seasonality on migration routes at large and small spatial scale.

Biogeography of Tropical Organisms

In the 19th century, while working in the Malayan Archipelago, the British naturalist Alfred Wallace was one of the first to observe and describe an underlying pattern in the distribution of species. Bird species with Oriental origin occur west of the border line, whereas bird species with Australian origin occur east of it. In recognition for his pioneering work, this line is called the Wallace Line. This fundamental question as to why species are distributed the way they are remains to this day an area of interesting research. To address it, scientists may use a combination of geological (i.e. plate tectonics, volcanism), historical (i.e. glaciation, dispersal), geographical (i.e. altitude, stream routes), molecular techniques (i.e. genetic), and climatic (rainfall, wind direction) factors. For example, taxa with large geographic distributions (i.e. ants, ferns) might have originated before certain continental plates separated. A recent molecular study found that Old World driver ants and New World army ants had a common ancestor before the southern supercontinent Gondwana (over 105 mya.) broke apart. Since then, they evolved into thousands of different species on both continents. One would, however, not expect to find army ants on tropical islands (i.e. Galapagos), unless they were able to disperse over water. In fact, many plant and animal taxa dispersed from mainland to distant islands. For example, with over 3900 km to the nearest mainland, the Hawaii archipelago is the most isolated area in the world. Nevertheless over 23,680 species including 8427 insects (5462 endemic), 294 birds (63 endemic), and 44 mammals (2 endemic) occur on the Hawaiian Islands. Dispersal to isolated islands has been repeatedly possible for numerous volant (flying) and non-volant species. The latter ones required driftwood for their long journey. On the other hand, many typical tropical plant (i.e. Annonaceae) or animal taxa (i.e. Ursidae) are absent, demonstrating that long-distance dispersal is limited to taxa with certain intrinsic characteristics that allow for long journeys.

Successful colonizing species may evolve into new species, a process called adaptive radiation to occupy not-yet-filled ecological niches. For instance, ancestral finch species that arrived on Hawaii evolved different bill shapes to explore new resources, including flowers, resulting in over 40 Hawaiian honeycreepers. Other examples of adaptive radiation include Galapagos finches, marsupials in Australia and the Neotropics, and lemurs in Madagascar. Adaptive radiation can lead to high levels of endemism, and increases topical gamma diversity (the regional diversity of all habitats).

Active volcanoes can form land bridges between separated areas, thereby allowing species exchange. The most dramatic example of faunal exchange occurred (some 3 mya ago) after volcanic activity lifted the Isthmus of Panama out of the sea, connecting North and South America. Many animal species migrated in both directions, a phenomenon known as The Great American Interchange.

Prior to the land bridge formation, South America was isolated for almost 80 million years, since breaking and drifting away from Africa. Therefore, more archaic mammals including anteaters, armadillos, marsupials, sloths, and now extinct species such as the giant ground sloth and the saber-toothed marsupials dominated its fauna. North America, on the other hand, had been repeatedly connected to Eurasia and harbored more modern mammal species including bears, camels, cats, dogs, elephants, horses, peccaries, rodents, and tapirs. Most of the original North American species underwent explosive adaptive radiation and today comprise over 50% of South America's mammal species. The South American species, however, were less successful; only one armadillo, opossum, and porcupine species survived in North America.

New fossil records and a better understanding of phylogenetics, paleoclimatology and paleogenetics, among other disciplines, will further improve our understanding of past and extant species distribution.

Tropical Species Richness: Anyone's Guess

How many species are out there? This is one of the most fundamental questions in biology, yet we do not know the answer. Until the early 1980s, biologists estimated that around 2 million species occur worldwide. In 1982, however, Terry Erwin fumigated the canopies of tropical trees with insecticides. After a downpour of invertebrates, mostly unknown species, he estimated that in the topics alone there might be as many as 30 million species. More recent estimates of global species richness suggest between 1.8 and 8.7 million species ($\pm$1.3 million), of which only 1.7 million are known to science. Every year new species are found either in museum collections or in the field, including the spectacular discoveries in 2005 of three new primate species from India, Africa, and South America. In 2016 three new mouse lemur species have been discovered in Madagascar.

Most groups of organisms exhibit a tendency for increases in species richness or biodiversity from the poles towards the tropical equatorial region. This phenomenon often referred to as the Latitudinal Gradient in Species Diversity, is one of the most widely recognized patterns in biogeography. Scientists have argued for over a century about its underlining mechanism. In 1808, the German naturalist Alexander von Humboldt was the first to suggest that energy (sun radiation) is the mechanism driving this relationship. Since then over 100 hypotheses have been proposed to explain increased biodiversity in the tropics but we still lack a satisfactory answer. Most hypotheses focused on: historical events, energy availability, productivity (or both combined), species-area relationship, stability, disturbance, spatial heterogeneity, patchiness, habitat complexity, evolutionary rate, and direct interactions (i.e. predation, competition, mutualisms).

A synopsis of some of the main hypotheses and empirical evidences are discussed in the following section.

- *Historical events*: Continental glaciation during the late Pleistocene in northern latitudes may have accelerated the extinction of many species, thus preventing species to reach higher diversity. Given sufficient time (i.e. millions of years), species will reach equilibrium and the latitudinal gradient in species diversity might vanish.

- *Energy availability and productivity*: Because the tropics receive more solar energy and rainfall, there should be an increase in net primary productivity compared to the capricious seasonality of higher latitudes.
- *Intermediate disturbances hypothesis*: Species richness should be the highest in communities with intermediate levels (i.e. temporal and spatial) of disturbance (i.e. fires, hurricanes, and treefalls), because no single species can attain dominance no equilibrium is reached. At low disturbance levels, however, competitively dominant species would exclude subordinate species. Whereas at high levels of disturbance, selection would favor only few fast-growing species.
- *Evolutionary rate hypothesis* (or climate-speciation hypothesis): Higher ambient temperature in tropical regions may result in higher mutation rates and shorter generation times, which may result in higher speciation rates compared to higher latitudes. Thus, tropical organisms would evolve at a faster rate than temperate organisms.

Joseph Connell used treefall gaps to support his intermediate disturbance hypothesis. Treefalls are the most frequently occurring disturbance in tropical forests. These light gaps in the canopy allow more sunlight to reach the ground and can trigger the germination of many heliophilic (pioneer) species. Studies have found that gaps increased plant and animal species richness. For instance, researchers found higher species richness of insectivorous birds, and a distinct gap community of lizards, frugivorous birds, insectivorous bats, and small mammals. Other research indicated higher small mammal species richness in gaps than in the undisturbed understory.

To test whether the rate of speciation is faster for tropical organisms than their temperate counterparts, John Wiens and his collaborator combined three previously independent ideas and proposed the Tropical Conservatism Hypothesis. Later, John Wiens et al. used Neotropical treefrogs as model organisms to test this hypothesis. Their results supported all predictions and the authors argued convincingly that the tropical environment was not responsible for an increased speciation rate, but temperate regions were colonized more recently, thus when given sufficient time, more species will evolve in temperate regions.

Contradictory results were found in another recent study. Shane Wright and her colleagues tested the Climate—Speciation Hypothesis (whether warmer tropical climate leads to higher metabolic and mutation rates, resulting in higher speciation rates). This hypothesis is very similar to that one John Wiens et al. tested. To date, the Wright et al. study is one of the most comprehensive ones because they compared the rate of plant evolution across a wide geographic distribution (including Borneo, New Guinea, Australia and South America) with closely related temperate plant species (including North America, Australia, Eurasia and New Zealand). They found that tropical plant species had more than twice the rate of molecular evolution (nucleotide substitution) compared to temperate plants.

The mixed results from these studies demonstrate that understanding tropical diversity remains a complex and challenging endeavor. Considering geologically diverse settings, the evolutionary, and biogeographic history of taxa, it seems more likely that multiple factors rather than a single "holy grail" hypothesis will explain the underlying mechanisms responsible for high species diversity found in the tropics.

Tree Plots: A Wealth of Knowledge for Plant Ecology

Tree plots have been a powerful approach to quantify and compare structural differences between forests across tropical areas. Plots range between 0.1 and 50 ha in which all trees of a given diameter and height are labeled, identified, and various measurements are taken annually. Depending on the specific question, scale is important because the larger a plot, the more likely it will include individuals of rare species. However, because of the high tree density and species richness of tropical forests, it may take years and many dedicated people to establish a single 50 ha plot. For example, the initial census of a 50 ha plot in Malaysia rendered over 335,000 individuals from 814 species. Therefore, few 50 ha tree plots are established (i.e. in Borneo, Ecuador, India, Panama, Malaysia, Sri Lanka, and Thailand). But many smaller plots have been set up, for example, over 450 have been established in the Amazon. Combining data from smaller plots within a region can be a powerful approach. For example, ter Steege et al. pooled data from 275 tree plots (ranging from 0.4 to 4 ha) scattered throughout the entire Amazon Basin. They tested the Rainfall—Density Hypothesis, which predicts a positive relationship between rainfall and species diversity. The authors found that the length of the dry season negatively correlated with tree density and maximum α-diversity (diversity within a particular area). This is one of the few studies that confirmed, over a very large scale, that rainfall is a major factor for local tree diversity.

The 50 ha tree plots have been extremely valuable to test other major ecological hypotheses, i.e. testing the spatial distribution of tree species. Results indicate that the vast majority of tree species are spatially clumped (in the neighborhood of a given species there is a higher than average density of conspecifics), rather than randomly distributed, and rare species are even more clumped than common species.

Other results relate to recruitment pattern of juvenile plants, impact of disturbance (i.e. gaps, drought), density dependent effects, community ecology, management and conservation related question, genetic structure among individuals, and seed rain. Furthermore, thanks to long-term data records, scientists can test the effects of El Niño events, and anthropogenic climate changes (see below).

Tree plot studies have stimulated a wealth for hypotheses and have increased our understanding of spatial dynamics and plant diversity. Despite this, many funding agencies are not keen on supporting long-term projects. More projects like this across different tropical ecosystems are needed to address fundamental ecological questions.

Interactions and Interdependencies of Tropics Species

To illustrate the manifold and complex interactions and interdependencies of tropical species, a synthesis of numerous studies focusing on one plant and one animal genus is provided in the following section.

The Ficus Genus: Master of Many Trades

Ficus, the genus to which fig trees belong, has over 1000 species that occur throughout the pantropics. Figs have a wider variety of growth forms than any other tropical plant genus including shrubs, woody lianas, hemiepiphytes, epiphytes, and trees. Some figs (strangler trees) start their lifecycle as epiphytes on other trees and eventually become majestic free standing trees. Soon after a bird or monkey deposit a sticky fig seed on a large tree, it starts germinating (similar to mistletoes). The young plant grows and sends aerial roots downward. Once the roots reach, the soil, they engage in mutualistic interactions with mycorrhizal fungi. The fig provides carbohydrates while the fungi facilitates the uptake of water, minerals (i.e. phosphorus), and other nutrients. The plant continues to grow, eventually overtaking the host's canopy. Meanwhile, the fig roots expand, forming a tight network that starts to constrict the host trunk. Eventually, the host tree dies and slowly decomposes within the woody network of the now free standing fig tree (Fig. 2).

Compared to the smooth trunk of most trees, the trunk of strangler figs started as network of roots and therefore contains many crevices and holes. The higher structural heterogeneity of strangler trunks provides a variety of microhabitats such as den, nest, and foraging habitats for invertebrates (i.e. ants, bees, spiders) and vertebrates (geckos, lizards, rodents, marsupials, birds). Species that create habitat for other species are called ecosystem engineers (sensu Jones) or niche constructors (sensu Odling, Smee, Laland and Feldman). Thus strangler figs are considered autogenic engineers because they modify the environment by modifying itself.

The pollination of fig flowers is a textbook example of coevolution. In general, each fig species has its own highly specialized wasp species that pollinates its flowers. Hundreds of little flowers are enclosed within a small (0.5–0.6 cm) fruitlike globular structure called synconium. Inside the sealed synconium are also eggs laid by fig wasps before the die. After development, male wasps hatch first and inseminate the unhatched females. Later these pre-born impregnated females hatch, picking up pollen as they chew exit holes through the synconium. The freed wasps visit other flowering figs, chew entrance holes into the synconium, pollinate its flowers, lay their eggs, and die. There are also parasitic fig wasp species that utilize the synconium and consume fig tissues without providing any pollination service.

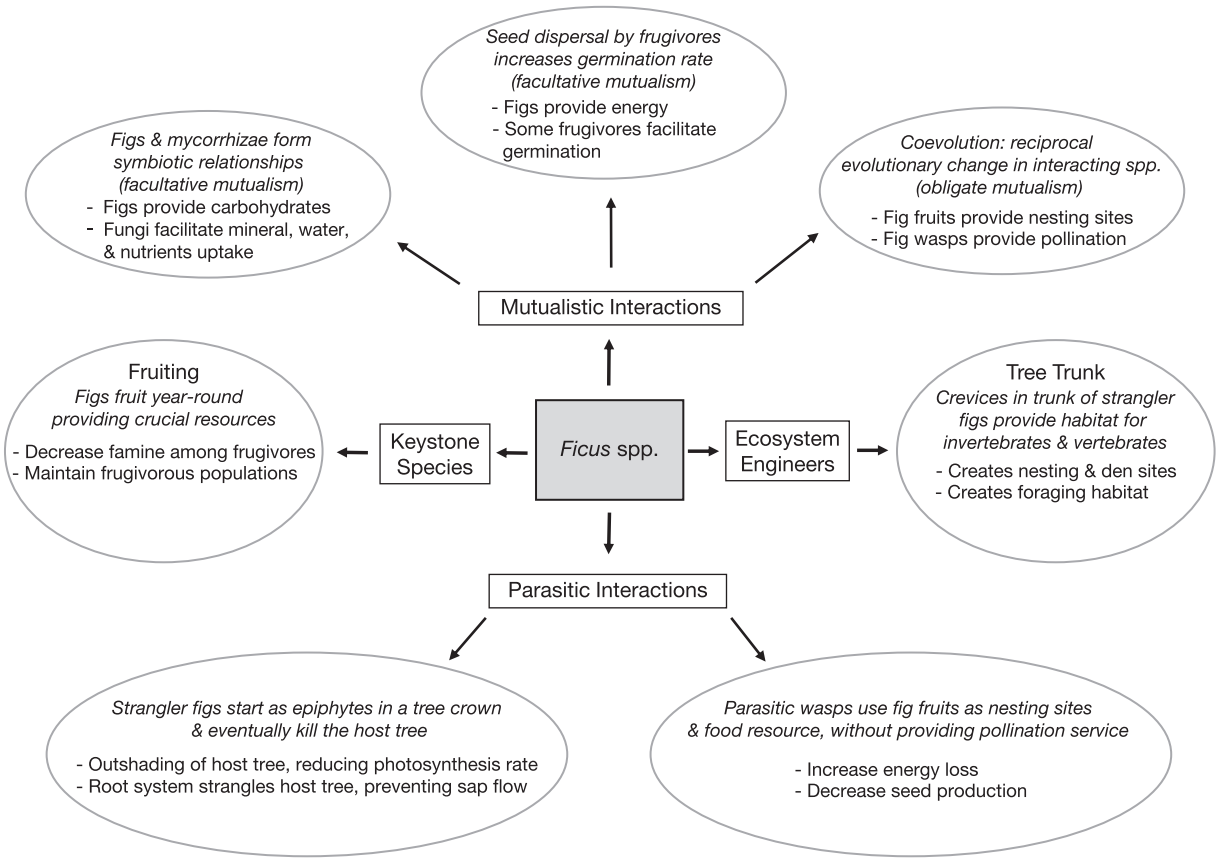


Fig. 2 Summary of some interactions and resulting interdependencies of Pantropical figs with other species and their ecological ramifications.

Fruiting fig trees are magnets for a myriad of species such as pigeons, hornbills, toucans, parrots, macaws, bats, flying foxes, and monkeys. In fact is has been demonstrated that these seed dispersers have evolutionary shaped the fruit trait such as size, color, and odor. This is another great example highlighting how tropical mutualisms can generate and maintain biodiversity. As animals move through the fruit-loaded canopy, they create a fruit rain and terrestrial species like duikers, peccaries, and rodents can thrive on them. Most of the tiny seeds (0.5–3 mm) survive digestion and are dispersed at different scales, depending on retention time and movement pattern of the animal species. Some seeds will end up in crowns of other trees, presenting an opportunity for strangler figs (see above), while others will be deposited on the forest floor.

Because fig species fruit asynchronously year round, including the dry season when the vast majority of other plants do not fruit, figs are critical for many animal species survival; it is for this reason that figs are called Keystone Species for frugivorous vertebrates. Studies showed that fig fruits can constitute over 50% of chimpanzees' diet (Africa), up to 70% for some primate species (Peru), and almost the entire diet of some Neotropical bat species. It is safe to say that without fig fruits during the dry season, the density of many vertebrate species would dramatically decline, if not crash (**Fig. 2**).

Neotropical Peccaries (*Tayassuidae*)

Three (potentially four) peccary species (a pig-like mammal) occur in the Neotropics. All of them are gregarious species, for example, white-lipped peccaries can occur in groups of several hundred individuals and represent the largest terrestrial biomass (230 kg km^{-2}) for mammals in Neotropical forests. Peccaries are omnivores and utilize seeds, roots, fungi, invertebrates, and vertebrates. They consume fruits from over 207 species and destroy the seeds of over 79% of those species. Peccaries are primarily seed predators and only small seeds such as those of figs, escape their mastication and digestive system and are dispersed (endozoochory) over long-distances. Trees that drop their fruits underneath the canopy attract herds of peccaries, which ferociously bulldoze through the soil and leaf litter and trample juvenile plants while searching for fruits. Peccaries prefer seeds infested with nutritional insect larvae such as bruchid beetles, thereby they may also control insect populations in a top-down fashion, and indirectly enhance future seed survival. Some seeds are too hard to be cracked by peccaries; in those cases peccaries chew off the fruit pulp and expectorate (spit out) the seeds in close vicinity of the parent tree. Some of those seeds are accidentally trampled deep into the soil and are thereby protected from insect predation. This shot-distance dispersal can lead to clumped distribution of plants i.e. Mauritia palm swamps (**Fig. 3**).

Numerous plants have hooks on their seed coat that allow them to attach to the hair of animals and fall off later. This dispersal mode is called epizoochory. Some of those seeds have been found in the fur of peccaries, an indication that they may facilitate the dispersal of epizoochorous species.

Studies have shown that because peccaries destroy a large number of seeds and seedlings of many plant species, they play a fundamental role in regulating recruitment, demography, and the spatial distribution of plants; thereby reducing competitive exclusion among plants and promoting plant species diversity.

Considering the high biomass of peccaries particularly of white-lipped and their consumption of such a large diversity of fruits, they can (out)-compete many other frugivorous species (including the collared peccaries), thus affecting their population dynamics.

Peccaries can be considered ecosystem engineers, because their rooting and bulldozing behavior leads to the removal of leaf-litter and soil. Leaf-litter can act as physical or chemical barrier to the establishment of litter-gap dependent species. Thus peccaries create new habitats which may permit the establishment of litter-gap dependent species.

Peccaries also create and maintain wallows. Research indicates that most of these wallows contain water year-round, including the dry season when most other terrestrial water bodies dry up. Studies found that wallows are critical breeding habitats for several amphibian species which go locally extinct shortly after peccaries are extirpated.

Because of habitat fragmentation and hunting, peccary populations, particularly white-lipped and Chacoan peccaries, are continuously declining, and they are one of the most endangered mammal species throughout the Neotropics. Aside from a few isolated white-lipped peccary populations this species is extirpated throughout Central America. Considering their manifold interactions with other species, local extinction of peccaries may result in changes in the distribution, community composition, and species diversity of plant and animal species.

Anthropogenic Impacts on Tropical Ecosystems

Both terrestrial and aquatic tropical ecosystems provide critical ecological services that humans, no matter where they live on this planet, depend upon. For example, tropical forests affect the global carbon cycle and via carbon sequestration (photosynthesis) store up to 50% of all terrestrial carbon. They produce over 40% of the world's oxygen, influence the global weather, and provide fresh water, food, wood, rubber, and other chemicals for billions of people. Currently, over 41% of the World's human population live within the tropics and estimates suggest that by 2050 their population will increase to over 50%. Human overpopulation and their activities have devastating effects on tropical ecosystems. For instance, habitat destruction, deforestation, fragmentation, forest fires, mining, infrastructure constructions, and selective logging have already reduced global rainforests to one half of their

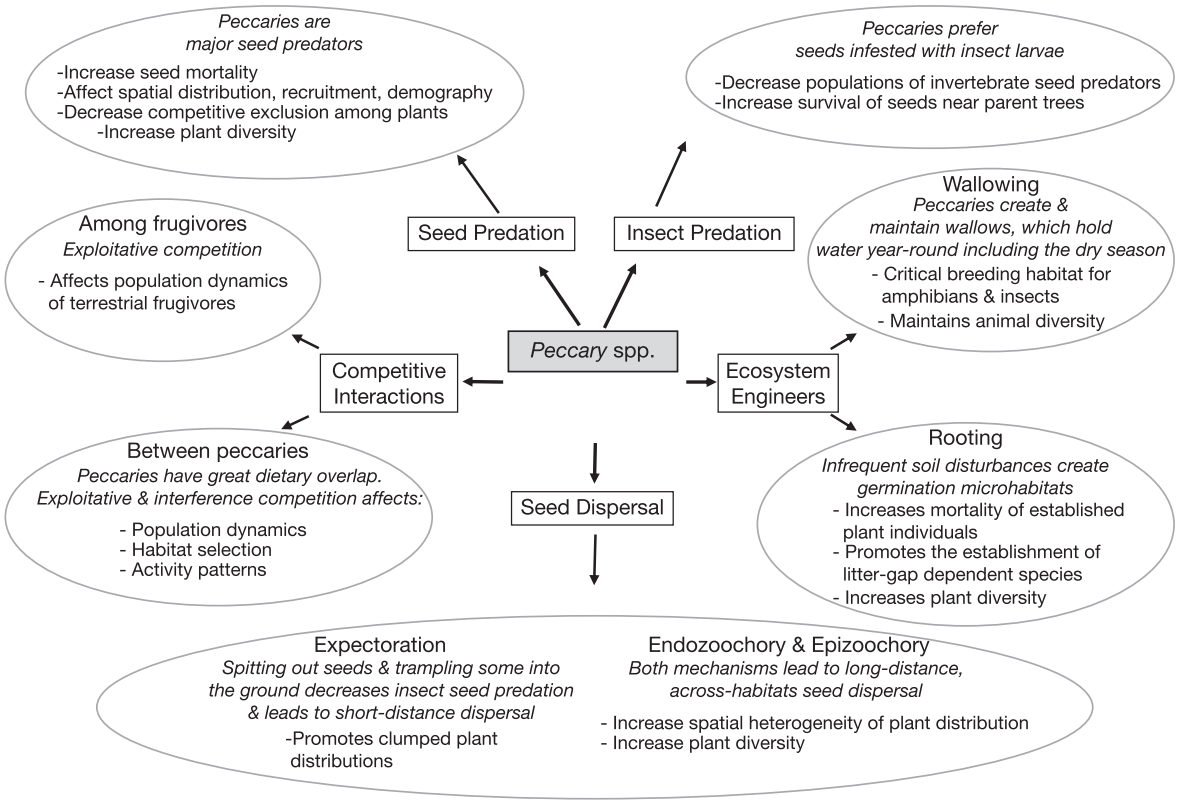


Fig. 3 Summary of some interactions and interdependencies of Neotropical peccaries with other species and their ecological ramifications. Adapted from Beck, H. (2005). Seed predation and dispersal by peccaries throughout the Neotropics and its consequences: A review and synthesis. In Forget, P-M., Lambert, J. E., Hulme, J. E. and Vander Wall, S. B. (eds.) *Seed fate: predation, dispersal and seedling establishment*, pp. 77–115. Wallingfort: CABI Publishing, with permission from CABI Publishing, UK.

original size. If the current rate of deforestation continues, then the world's rainforests and many of their species will vanish within 40–100 years.

Legal and illegal hunting and wildlife trade (problems that also occur in many established national parks), brought many species to the brink of extinction, including gorillas, chimpanzees, rhinoceros, tigers, macaws, parrots, amphibians, mahogany, cacti, and orchids, to name a few. Defaunation driven by unsustainable overhunting has led to a "Bushmeat Crisis" occurring in many tropical regions. The original densities of most large mammals in tropical forests have been reduced to around 10%. Many species are locally extinct, described as The Empty Forest (sensu Redford). Because many of these animal species have evolved non-redundant mutualistic interactions with plants (**Figs. 2** and **3**), or may also affect other animal species (i.e., via competition, predator or prey interactions), almost every extinction event will result in trophic cascading effects and negative consequences for remaining species and their ecosystem.

Anthropogenic induced global climate change might become the biggest threats to tropical ecosystems and its inhabitants. Climate change has and will continue, with increased magnitude, to alter global weather patterns such as the frequency, passage, and intensity of El Niños, Hurricanes, Typhoons, and seasonal rainfall patters. Higher temperatures and several mega-droughts (in 2005 and 2010) across the Amazon caused unprecedented mortality among drought-sensitive tree species. In fact, the biomass loss was so large that the Amazon, rather than being a carbon sink, became a carbon source, by releasing greenhouse gases and contributed to global warming. Other consequences of global climate change include alterations to forest structure, dynamics, species composition, and loss of biodiversity. These effects will further cascade and affect species that had mutualistic interactions with these drought-sensitive plants.

The complex abiotic factors, biotic interactions, and interdependencies of tropical species that are interrupted by human activities will result in an extraordinary ecological meltdown, species extinction, and the destruction of most of the tropical ecosystems.

Meanwhile how about humans, their food security, and economy? Billions of humans may be starving, dying, or become environmental migrants with consequences that will also profoundly affect countries outside of the tropics. As I am writing this chapter, millions of migrants from North Africa and the Middle East shocked the neighboring regions and the European Union to its core. Imagine what will happen if billions have to move! National and international political and social actions are needed to truly address our self-inflicted threats: global climate change, overpopulation, poverty, and our obsession with a constant increase in GDP and economic growth. We must reform our socio-economical thinking towards the development of renewable energy

forms and real sustainability. We need comprehensive science-based agreements, true commitment to an ecologically-balanced world from individual governments and their citizens, and stronger support and efforts by watchdog and conservation organizations may be able to slow down this ecological crisis.

Conclusions

Tropical ecology advanced impressively since the last century. A mosaic of individual studies has slowly revealed the larger picture. New insights into fundamental questions such as the impact of climate, the distribution of species, and the interactions among species have helped improve management and conservation of our natural resources and provided future research directions. The field of tropical biology and, in particular, conservation has recruited more enthusiastic local students from tropical regions than ever before. Their voice can now be heard loud and clear in their own countries but they must become even louder, if we really want to avoid further loss of species and most tropical ecosystems.

See also: Ecological Data Analysis and Modelling: Statistical Inference. Ecosystems: Riparian Wetlands. Evolutionary Ecology: Colonization; Association. General Ecology: Biomass

Further Reading

Alonso, A., Sukumar, R., 2016. Tropical conservation, perspectives on local and global priorities. New York: Oxford University Press.
Altrichter, M., Taber, A., Beck, H., Reyna-Hurtado, R., Lizarraga, L., Keuroghlian, K., Sanderson, E.W., 2016. Range-wide declines for a key Neotropical ecosystem architect, the near threatened white-lipped peccary Tayassu pecari. Oryx 46, 87–98.
Beck, H., 2005. Seed predation and dispersal by peccaries throughout the Neotropics and its consequences: A review and synthesis. In: Forget, P.-M., Lambert, J.E., Hulme, J. E., Vander Wall, S.B. (Eds.), Seed fate: Predation, dispersal and seedling establishment. CABI Publishing: Wallingfort, pp. 77–115.
Beck, H., 2006. A review of peccary-palm interactions and their ecological ramifications across the Neotropics. Journal of Mammalogy 87, 519–530.
Beck, H., 2008. Linking Amazon forest dynamics with mammalian diversity. Germany: Vdm Verlag Dr. Müller.
Beck, H., Thebpanya, P., Filiaggi, M., 2010. Do Neotropical peccary species (Tayassuidae) function as ecosystem engineers for anurans? Journal of Tropical Ecology 26, 407–414.
Beck, H., Snodgrass, J., Thebpanya, P., 2013. Long-term exclosure of large terrestrial vertebrates: Implications of defaunation for seedling demographics in the Amazon rainforest. Biological Conservation 163, 115–121.
Bush, M.B., Flenley, J.R., Gosling, W.D. (Eds.), 2011. Tropical rainforest responses to climate change, 2nd edn Chichester: Springer.
Cai, W., Borlace, S., Lengaigne, M., Rensch, P., Collins, M., Vecchi, G., Axel, T., Santoso, A., McPhaden, M.J., Wu, L., England, M., Guilyardi, E., Jin, F.-F., 2014. Increasing frequency of extreme El Niño events due to greenhouse warming. Nature Climate Change 4, 111–116.
Connell, J.H., 1978. Diversity in tropical rain forests and coral reefs. Science 199, 1302–1310.
Costello, M.J., Wilson, S., Houlding, B., 2011. Predicting total global species richness using rates of species description and estimates of taxonomic effort. Systematic Biology 61, 871–883.
Hotaling, S., Foley, M.E., Lawrence, N.M., Bocanegra, J., Blanco, M.B., Rasoloarison, R., Kappeler, P.M., Barrett, M.A., Yoder, A.D., Weisrock, D.W., 2016. Species discovery and validation in a cryptic radiation of endangered primates: Coalescent-based species delimitation in Madagascar's mouse lemurs. Molecular Ecology. 2016. doi:10.1111/mec.13604.
Jones, C.G., Lawton, J.H., Shachak, M., 1994. Organisms as ecosystem engineers. Oikos 69, 373–386.
Lewis, S.L., Brando, P.M., Phillips, O.L., van der Heijden, G.M.F., Nepstad, D., 2011. The 2010 Amazon drought. Science 331, 554.
Lomáscolo, S.B., Levey, D.J., Kimball, R.T., Bolker, B.M., Alborn, H.T., 2010. Dispersers shape fruit diversity in Ficus (Moracea). Proceedings of the National Academy of Sciences 17, 14668–14672.
Mora, C., Tittensor, D.P., Adl, S., Simpson, A.G.B., Worm, B., 2011. How many species are there on Earth and in the ocean? PLoS Biology 9. e1001127doi:10.1371/journal.pbio.1001127.
Pacheco, L.F., Altrichter, M., Beck, H., Buchori, D., Owusu, E.H., 2016. Conservation as the new paradigm for development. In: Aguirre, A.A., Sukumar, R. (Eds.), Tropical conservation: Perspectives on local and globa priorities. New York: Oxford University Press, pp. 390–402.
Phillips, O.L., Lewis, S.L., Baker, T.R., Chao, K.-J., Higuchi, N., 2008. The changing Amazon forest. Philosophical Transaction of the Royal Society B 363, 1819–1827.
Phillips, O.L., *et al*., 2009. Drought sensitivity of the Amazon rainforest. Science 323, 1344–1347.
Phillips, O.L., *et al*., 2010. Drought-mortality relationships for tropical forests. New Phytologist 187, 631–646.
Redford, K.H., 1992. The empty forest. Bioscience 42, 412–422.
Taber, A., Altrichter, M., Beck, H., Gongora, J., 2011. The Tayassuidae. In: Wilson, D.E., Mittermeier, R.A. (Eds.), Handbook of the mammals of the world: hoofed mammals,vol. 2. Barcelona, Spain: Lynx Edicions, pp. 292–307.
Taber, A., Beck, H., Gonzalez, S., Altrichter, M., Duarta, J.M.B., Reyna-Hurtado, R., 2016. Why Neotropical forest ungulates matter. In: Aguirre, A.A., Sukumar, R. (Eds.), Tropical conservation: Perspectives on local and global priorities. Oxford: University Press, pp. 255–261.
Williams, J.W., Jackson, S.T., Kutzbach, J.E., 2007. Projected distributions of novel and disappearing climates by 2100 AD. Proceedings of the National Academy of Sciences of the United States of America 104, 5738–5742.
Wright, S.D., Keeling, D.J., Gillman, L.N., 2006. The road from Santa Rosalia: A faster tempo of evolution in tropical climates. Proceedings of the National Academy of Sciences 103, 7718–7722.

Tropical Rainforest

RB Waide, University of New Mexico, Albuquerque, NM, USA

Introduction

Of all of the Earth's ecosystems, tropical rainforests exist on the extremes of temperature, rainfall, biodiversity, and structural complexity. Tropical rainforests exist only where high year-round temperatures are found in conjunction with moderate to high rainfall, which limits both their latitudinal and elevational distribution. Compared to other ecosystems, tropical rainforests have high numbers of plant and animal species and often show great specificity in their biological relationships. This high taxonomic diversity contributes to high functional diversity, which results in a complex forest structure comprising many different life forms and sizes of plants. This diversity and structural complexity makes tropical rainforests one of the most interesting and complex ecosystems on Earth, and, as such, tropical rainforests have captivated the imaginations of scientists and the public alike.

Definitions

The common use of the term `tropical rainforest' varies among regions of the globe depending on the ecological context. There is general agreement that tropical rainforests are tall, dense evergreen forests existing in wet and warm places, but since there is a degree of subjectivity in these terms, the name `tropical rainforest' is applied to forests on different continents that may be quite different structurally. Moreover, since the characteristics that define tropical rainforest grade with latitude and elevation, the boundary between tropical rainforest and other forest types is by necessity arbitrary.

The classical definition of tropical rainforest focuses on features of the vegetation: evergreen, hygrophilus, tall, and rich in lianas and epiphytes. Additional characteristics of tropical rainforest include the dominance of woody plants (**Fig. 1**), principally trees; high species richness; sparse undergrowth; relatively slender trunks compared to trees of temperate forests; straight boles without branches except near the top; buttresses (**Fig. 2**); large, dark green leaves with entire margins; the occurrence of flowers on the trunk or branches; and inconspicuous green or white flowers.

Alternative definitions of tropical rainforest focus on characteristics of the forest community and its environment, including the proportion of deciduous trees in the canopy, the elevation of the forest, and the length and severity of the dry season. Some local classifications of forest type incorporate floristic information, but such classifications require detailed knowledge of the flora as well as trained experts to implement them.

Several schemes attempt to classify vegetation types based on climatic conditions including temperature, rainfall, length of dry periods, and evapotranspiration. Such classification systems avoid the subjectivity inherent in definitions that depend on relative terms, but by necessity are oversimplifications of the factors controlling the distribution of tropical rainforests. The Köppen classification uses the average annual precipitation, average monthly precipitation, and average monthly temperature to divide the globe into six major climate regions and their subregions. Under this system, the tropical rain climate has no month whose mean temperature is less than 18 °C and the mean rainfall of the driest month is >60 mm. In the Holdridge classification, rainforests are defined as areas where the ratio of potential evapotranspiration to rainfall is low, with tropical lowland rainforests occurring where the mean annual temperature exceeds 24 °C. Shifts between forests classes are determined by changes in rainfall and temperature related to elevation and latitude. These classification systems work well in areas where plant formations are strongly controlled by climate, such as Central America and northern South America, but are less useful where edaphic factors or other environmental factors are major controlling factors, as in the lower Amazonian region in Brazil.

The rest of this article focuses on lowland, evergreen tropical forests occurring in hot, wet conditions. Tropical forests at higher elevations or in areas with a pronounced dry season are covered elsewhere.

Distribution

Tropical rainforests exist wherever conditions are appropriate, but are mostly confined to a broad belt around the equator. The latitudinal distribution of tropical rainforests is limited by the distribution of freezing temperatures, which tropical plants are unable to withstand. A circumglobal belt of dry conditions also limits the distribution of tropical rainforest and, except in rare cases, prevents a continuous transition between tropical and temperate forests. Some gaps occur in the equatorial band of tropical rainforests, such as in eastern Africa, where prevailing conditions are too dry for tropical rainforest to develop. Moreover, in some areas on the eastern margins of continents, conditions suitable for tropical rainforest exist outside of the tropics. In all areas suitable for the development of tropical rainforest, human actions limit the present distribution of forests.

Fig. 1 Tall, relatively slender trees with straight boles are characteristic of tropical forest on Barro Colorado Island, Panama. Credit: Nicholas V. K. Brokaw.

Fig. 2 Broad buttresses are found on many rainforest trees, such as this specimen from Providencia, Antioquia, Colombia. Credit: Robert B. Waide.

Large areas of tropical rainforest exist on continents and large islands that straddle the equator. Roughly half of the tropical rainforests on the planet occur in three areas in tropical America. The largest of these forest areas (somewhat over 3 million km^2) occupies the drainage basins of the Amazon and Orinoco Rivers in northern and central Brazil and surrounding countries. A narrow strip of tropical rainforest runs along the Atlantic coast of Brazil from 7° to 28° S (from Recife nearly to São Paulo), but less than 5% of this forest remains in its original condition. A third block of forest occupies southern Mexico, Central America, and the area of northern South America west of the Andes. Many Caribbean islands also have small areas of tropical rainforest.

In Africa, another large block of tropical rainforest occupies the basin of the Congo River in the Democratic Republic of the Congo, the Republic of the Congo, Gabon, and Cameroon. Previously, part of this forest extended into Nigeria. Belts of tropical rainforest also extended along the coast of West Africa and the eastern part of the island of Madagascar, but little remains of these forests but isolated patches.

A third large area of tropical rainforest existed on the Malay Peninsula and the islands of Borneo, Sumatra, and Java. Sulawesi, the Philippines, and many of the smaller islands in Indonesia also have substantial areas of rainforest, but the condition of the remaining forest varies widely from island to island. Rainforests also occupied parts of mainland Southeast Asia where rainfall was sufficiently high. Isolated patches of tropical rainforest occur in the area of the Western Ghats in India and on the island of Sri Lanka. Most of the island of New Guinea supports tropical rainforest, and there is also a small area of rainforest in NE Australia. Patches of tropical rainforest also occur on some of the Pacific Islands (Solomons, New Hebrides, Fiji, Samoa, New Caledonia).

Climate and Soils

Tropical rainforests are found under a surprisingly wide range of climatic conditions. Annual rainfall is generally high in tropical rainforest compared to other ecosystems, but can range from 1700 to 10 000 mm. Many rainforests experience 1–4 dry months a year, when the rainfall is less than the water lost through evaporation and transpiration. These annual dry periods exert a strong effect on the phenology of biotic processes such as flowering and fruiting. In some tropical rainforests, rainfall is uniformly high throughout the year, and no annual dry periods exist. In these forests, dry periods may occur at multiyear intervals and trigger strongly synchronized biotic responses including mass flowering, increased animal reproduction, and migration. In some parts of the world, these multiyear cycles are related to periodic El Niño–Southern Oscillation (ENSO) events. While strong ENSO events result in more severe dry periods in many tropical rainforests, the strongest biological consequences seem to occur in forests that do not normally experience a dry season, especially in areas of Indonesia and Malaysia.

Mean annual temperatures in tropical rainforests generally fall in the range between 24 and 28 °C near the equator, but a consistent characteristic is the absence of a cool season. In general, diurnal temperature differences (6–10 °C) exceed monthly differences. The amount of solar radiation is higher in the topics than in temperate zones, but tropical rainforests generally have lower available solar radiation than drier tropical forests because of the greater amounts of water vapor and increased cloudiness in more humid climates. As a result, plant growth in closed-canopy tropical rainforests is often light limited.

The environments of tropical rainforests are characterized by high relative humidity during the daytime and generally saturated conditions at night. However, because much of the rainfall in tropical rainforests occurs in intense events, even months with high rainfall can have periods of a few days when little or no rain falls, saturation deficits increase, and plants wilt. The dry periods can be exacerbated by winds; evaporation rates are higher in the trade wind zone than in equatorial forests where average wind speeds are less.

Tropical cyclones can have severe effects on tropical rainforests. In general, areas within 10° latitude of the equator are not subject to tropical cyclones, but tropical rainforests in the Caribbean, Madagascar, northeastern Australia, many oceanic islands, and parts of Central America and Southeast Asia are affected by these storms. The strongest tropical cyclones can have severe but, in most cases, temporary effects on forest structure and composition. Tree mortality can be high as a result of one of these storms but the forest recovers quickly through regeneration, new growth, and refoliation of damaged trees. In those areas of tropical rainforest subject to recurrent tropical cyclones, forest structure and the biological traits of forest species may be affected by the frequency and intensity of storms.

The soils underlying tropical rainforests can have important effects on plant distribution and primary productivity. The complex interactions between soil characteristics (e.g., soil texture, age, drainage characteristics, nutrients) and the considerable topographic and geographic variation in these characteristics make it difficult to determine the importance of specific soil properties. Most areas supporting tropical rainforests have very old soils that are highly leached and weathered and as a result acidic and infertile. Such soils have low levels of the nutrients necessary for plant growth and high levels of toxic aluminum and thus are unsuitable for most forms of permanent agriculture. However, these soils can sustain high-diversity, high-biomass tropical rainforests because plants of these forests recycle nutrients efficiently. Some tropical rainforests occur on relatively fertile volcanic or floodplain soils and can sustain permanent agriculture.

In areas of the Amazon with low local relief, soil properties can have a strong effect on plant communities and therefore overall biodiversity. Small changes in topography and the depth of sand overlying the clay subsoil can cause large changes in the plant community.

Forest Structure

Tropical rainforests support a more diverse set of organisms than other kinds of forests. The number of different life forms, or synusiae, is greater in tropical forests that in temperate forests. A synusia is a group of organisms whose members are ecologically equivalent. When applied to plants, the term reflects an aggregation of species with similar life form and function. Autotrophic plants (e.g., those that photosynthesize) include those that do not need mechanical support (i.e., trees, shrubs, and herbs) and those that do (i.e., climber, epiphytes, and hemi-epiphytes; **Fig. 3**). Heterotrophic plants include saprophytes and parasites.

The structure of a tropical rainforest arises from each synusia's methods for obtaining resources for survival and growth: water, nutrients, and sunlight. In some forests, photosynthetic, self-supporting plants seem to form distinct strata depending on their size.

Fig. 3 Bromeliads and other epiphytes cover branches of trees in La Planada Reserve, Colombia. Reproduced by permission of Art Wolfe/Photo Researchers, Inc.

Such stratification is by no means a uniform characteristic of tropical rainforests. Photosynthetic plants that are not self-supporting use other plants as a platform for growth. Climbers (lianes) have roots in the ground but use other plants to support their elongated stems. Epiphytes depend on their host plants for support only, although a specialized group of this synusia (the mistletoes) obtains both support and water and dissolved substances from the support tree. Hemi-epiphytes initially live as epiphytes on supporting plants but eventually send roots down to the ground. Saprophytic and parasitic plants obtain required energy and nutrients from other living or dead plants, and therefore do not require light for growth or reproduction.

Background mortality of individual trees from natural causes is a major cause of spatial heterogeneity in tropical rainforests. Gaps caused by dead or fallen trees change the structure and the environmental characteristics of the forest. However, tropical rainforests are also dynamic ecosystems subject to a large number of natural disturbances including storms, lightning strikes, landslides, and the effects of animals, all of which can produce gaps in the forest canopy.

The canopy is an important structural element of tropical rainforests because the height and degree of closure of the canopy plays an important role in determining conditions in the understory (**Fig. 4**). Moreover, the lack of easy access to forest canopies means that their importance as a source of biodiversity and an influence on ecosystem processes has probably been underestimated. Forest canopies have important roles in the regulation of nutrient cycling and in the storage of carbon. Large pools of nutrients exist in live and dead components of the canopy, and decomposition of organic matter in the canopy influences access to these nutrient pools. The forest canopy serves to filter air- and waterborne nutrients and to provide a site for nitrogen fixation. Canopy-dwelling organisms are efficient at acquiring and storing nutrients, thus providing a buffer for pulsed nutrient releases. Forest canopies are rich in species of plants and animals that are independent of the forest floor. Moreover, canopy trees and their epiphytes provide important sources of food for birds, mammals, and insects that occupy other strata.

Biodiversity

Understanding of the biodiversity of tropical rainforests is still being refined. New species of all taxonomic groups are found every year, and knowledge of the diversity of some taxa, especially insects, is rudimentary. Tropical rainforests are extremely rich in species of all taxa compared to other terrestrial ecosystems. For example, the tropical rainforests of the world have an estimated 175 000 species of plants, which constitutes about two-thirds of the global total. Considerable variation in diversity occurs among tropical rainforests around the world, with the largest number of tree species (>250 species per hectare) occurring in Amazonia and Malaysia, followed by the islands of New Guinea and Madagascar, and then Africa. The largest areas of tropical rainforest (Neotropics, Africa) have the greatest number of primate species. Similar comparisons for other taxa are difficult because of the lack of data.

Conservation Issues

Because of their global significance with regard to carbon storage and the maintenance of biodiversity, conservation of tropical rainforests is an important and hotly debated topic. Solution of conservation issues is made more difficult because most areas of tropical rainforest occur in countries that are trying to increase the standard of living of their people. Partly because of this controversy, adequate data to judge the loss of tropical forests are difficult to come by. However, it is clear that tropical forests, including tropical rainforests, are disappearing at an increasing rate. The percent of the original forest habitat that has been lost exceeds 90% for some countries (Ghana, Bangladesh, Philippines). Estimates suggest that very little tropical rainforest will remain by the year 2050. The ultimate causes of forest loss include increasing populations in countries with tropical rainforest, extreme poverty, and the lack of effective government protection for forests. Proximate causes of forest loss include logging, clearcutting for agriculture, loss of ecosystem integrity because of

Fig. 4 Canopy of lowland tropical rainforest of La Selva Biological Station, Costa Rica. Photographed from a light plane flying 200 feet above the canopy. Reproduced by permission of Gregory G. Dimijian, MD/Photo Researchers, Inc.

Fig. 5 Rainforest has been cleared for timber and agriculture in this subsistence farm in Providencia, Antioquia, Colombia. Credit: Robert B. Waide.

forest fragmentation, and hunting (Fig. 5). Hunting of large animals may have insidious effects on forest structure, as the populations of prey species may explode when released from predation. Increased populations of small mammals, for example, may have severe effects on other organisms, leading to the breakdown of whole ecosystems over time.

Because the issues facing tropical rainforests vary considerably from one place to another, generic conservation solutions are not practical. However, the major elements of a conservation strategy for tropical rainforests will include the creation of reserves to protect biodiversity, the regulation of exploitative use of tropical rainforest products, the engagement of traditional societies, the development of sustainable use strategies that will address the issue of poverty, and an increased effort by developed countries to form partnerships with developing countries.

See also: Ecological Data Analysis and Modelling: Statistical Inference. Ecosystems: Riparian Wetlands. Evolutionary Ecology: Association. General Ecology: Biomass. Global Change Ecology: Phosphorus Cycle

Further Reading

Denslow, J.S., Padoch, C. (Eds.), 1988. People of the Tropical Rain Forest. Berkeley, CA: University of California Press.
Gentry, A.H. (Ed.), 1990. Four Neotropical Rainforests. New Haven, CT: Yale University Press.
Golley, F.B. (Ed.), 1989. Tropical Rain Forest Ecosystems. New York, NY: Elsevier.
Primack, R., Corlett, R., 2005. Tropical Rain Forests: An Ecological and Biogeographical Comparison. Oxford: Blackwell Science.
Richards, P.W., 1996. The Tropical Rain Forest: An Ecological Study. Cambridge: Cambridge University Press.
Sutton, S.L., Whitmore, T.C., Chadwick, A.C. (Eds.), 1983. Tropical Rain Forest: Ecology and Management. Oxford, UK: Blackwell Scientific Publications.
Terborgh, J., 1992. Diversity and the Tropical Rain Forest. New York: Scientific American Library.
Whitmore, T.C., 1998. An Introduction to Tropical Rain Forests. New York, NY: Oxford University Press.

Tropical Seasonal Forest☆

Egbert G Leigh Jr., Smithsonian Tropical Research Institute, Panama, Republic of Panama

Glossary

Eddy flux towers Towers with three-dimensional anemometers above the forest canopy, adjacent to devices for sampling the air's CO_2 content. Vertical movement of CO_2 (which is downward when the forest is photosynthesizing) is calculated from the covariance between the vertical velocity of the air and its CO_2 content which these instruments measure. These data allow calculation of gross productivity.

Evapotranspiration, actual The water evaporated from a forest or transpired by its leaves (released through the leaves' *stomata* when they are absorbing CO_2 from the atmosphere).

Evapotranspiration, potential The water that would evaporate from a forest or be transpired from its leaves if rainfall kept soil moisture content reasonably high without waterlogging it.

Gross productivity The total rate at which the forest's photosynthesis transforms carbon dioxide and water into carbohydrate, usually measured as the amount of carbon in the carbohydrates thus produced.

Phenology The seasonal timing of events such as leaf flush, leaf fall, flowering and fruiting, either in individual trees or in the forest as a whole.

Savanna Tropical tall grassland, that burns during the dry season, with scattered fire-resistant trees.

Species turnover The change in species composition encountered in passing from one place to another. Zones where species turnover is concentrated are sometimes called "biogeographical knots."

Stomata Pores in leaves which, when opened to let in CO_2, causes transpiration, the release of water vapor from the leaf, which pulls replacement water by capillary traction up from roots through *xylem vessels* into these leaves.

Trade-off A circumstance where selection to enhance one structure or ability necessarily diminishes another, as selection for earlier reproduction necessarily decreases longevity.

Xylem vessels Tubes <0.5 mm in diameter through which transpiration pulls water from the roots up to the leaves.

Definitions and Distinctions

Tropical humid seasonal forest is defined here as lowland tropical forest (mean annual temperature $\geq 20^{\circ}$C, mean temperature of the coolest month $\geq 18^{\circ}$C) which averages ≥ 1700 mm year^{-1} of rainfall and ≥ 100 mm month^{-1} for at least 7 months of the year, with a single severe dry season averaging <200 mm for the year's driest quarter and <100 mm month^{-1} for at least 3 consecutive months (there may be a second, milder dry season). Its wetter counterpart, tropical rainforest (everwet forest), which is almost entirely evergreen, averages ≥ 1700 mm year^{-1} of rainfall and ≥ 100 mm nearly every month (no 2 consecutive months average <100 mm). Its drier counterpart, tropical dry seasonal forest, where many trees lose their leaves in the dry season, averages 650–1700 mm year^{-1} of rain and ≤ 100 mm in the year's driest quarter and ≥ 5 consecutive months averaging <100 mm (**Table 1**). On less fertile, more acid soils, the same climate supports savanna, tall fire-prone grassland with scattered trees (Pennington and Lavin 2016, p. 25). There are also severely seasonal evergreen forests, some averaging ≥ 2500 mm year^{-1} of rainfall, with a single, long dry season of six or more consecutive months averaging <100 mm apiece.

This article's definition of seasonal forest is arbitrary. We will soon discuss a forest, Khao Chong in Thailand, with only one dry month, averaging 50 mm (**Table 1**), which behaves like a seasonal, not an everwet, forest. Second, many forest characteristics vary continuously with the length of the dry season, without "natural divisions". Third, rapidly draining soils that hold little water support tree species typical of drier climates (Pyke *et al.* 2001, p. 563), whereas soils better able to hold water, which release it slowly enough that plants can still extract some water at the height of dry season, support tree species typical of wetter climates (Kursar *et al.*, 2005). Finally, a criterion based on the ratio of a forest's actual to its potential, evapotranspiration must be a better way of predicting a forest's degree of limitation by water shortage than the length of the dry season or the rainfall during the year's driest quarter. Data on potential and actual evapotranspiration, however, are far less readily available.

Distribution and Evolution of Different Forest Types

Most rainforest occurs near the equator, where the midday sun is within 18° of zenith all year round (normally, the nearer the midday sun to zenith, the higher the rainfall 6 weeks later). Rainforest also occurs at higher latitudes near the base of east-coast

☆*Change History*: October 2017. Egbert Leigh expended and greatly revised this article in January 2017.

This is an update of E.G. Leigh Jr., Tropical Seasonal Forest, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 3629–3632.

 doi:10.1016/B978-0-12-409548-9.10921-2

Table 1 Sample rainfall régimes for different forest types. Average rainfall for each month, the year (total), and the driest quarter (P_3)

Forest type	*Everwet (rain)*	*Mildly seasonal forest*	*Seasonal forest*	*Dry forest*	*Dry forest*
Site	Pasoh, Malaysia	Khao Chong, Thailand	BCI, Panama[a]	HKK, Thailand[b]	Kirindy, Madagascar
Latitude	2°59′N	7°34′N	9°9′N	15°38′N	20°3′S
January	94 mm	140 mm	71 mm	6 mm	254 mm
February	109 mm	50 mm	32 mm	30 mm	188 mm
March	153 mm	168 mm	23 mm	39 mm	108 mm
April	167 mm	148 mm	106 mm	82 mm	22 mm
May	162 mm	251 mm	245 mm	226 mm	5 mm
June	125 mm	262 mm	275 mm	120 mm	0 mm
July	115 mm	238 mm	237 mm	123 mm	1 mm
August	120 mm	282 mm	322 mm	155 mm	2 mm
September	162 mm	302 mm	309 mm	278 mm	5 mm
October	189 mm	340 mm	364 mm	360 mm	23 mm
November	224 mm	346 mm	360 mm	47 mm	42 mm
December	168 mm	264 mm	202 mm	10 mm	149 mm
Total	1788 mm	2787 mm	2551 mm	1476 mm	799 mm
Driest quarter	356 mm	358 mm	131 mm	46 mm	3 mm

[a]BCI stands for Barro Colorado Island.
[b]HKK stands for Huai Kha Khaeng.
Data for Pasoh, BCI and HKK are from Leigh (2004); data for Khao Chong, from Bunjavejchewin S. and Davies S. J. (unpublished, used with permission), and data for Kirindy are from Sorg and Rohner (1996).

slopes facing trade winds, as in Puerto Rico's Luquillo Mountains (18°N), La Selva, Costa Rica (10°N), Las Tuxtlas, Mexico (18°N), and most of the east coasts of Madagascar and the Philippines. Humid seasonal forests flank equatorial forests to north and south. Nearly all of the African forest block, including Makokou (Gabon), Korup (Cameroun), and the Congo's Ituri forest, as well as the forests of Ghana, the Ivory Coast, and Liberia, are humid seasonal forests. Large blocks of southern and southwestern Amazonia are humid seasonal forest, including Cocha Cashu in Peru's Parque Nacional Manú (12°S). Humid seasonal forest also extends northward from roughly 7°N, near the Thai–Malaysian border, into the Isthmus of Kra, where Khao Chong is located, and formerly covered most of lowland Java. The best-studied humid seasonal forest, Panama's Barro Colorado Island (**Fig. 1**), is at 9°N, but in Central America, humid seasonal forest extends to Belize (18°N). The Amazonian humid forest block is separated from dry forest by a large expanse of savanna on its southeast, and a smaller one on its northwest, flank (Simon and Pennington, 2012). Likewise, in Africa, dry seasonal forest (forêt dense sèche) and woodland (forêt claire) are separated from humid seasonal forest by bands of humid savanna whose tall grass supports really hot dry-season fires (Menault *et al.*, 1995). Dry seasonal forest occurs at higher latitudes than humid seasonal forest, as does Thailand's Huai Kha Khaeng (16°N), and in rainshadow, as on Madagascar's west coast (the location of Kirindy Forest, 20°S), and on the west coast of Central and North America from Costa Rica's Parque Nacional Santa Rosa (11°N) to Chamela, Mexico (19°N) and beyond. Severely seasonal evergreen forest occurs at 13–15°N in the Western Ghats of southwest India (Pascal, 1988).

Savanna only evolved $\leq$8 million years ago when productive fire-prone tall grassland became widespread in South America, Africa, and Asia, and many tree clades from both humid and dry forests colonized these new grasslands (Simon and Pennington, 2012; Maurin *et al.*, 2014). On the other hand, flowering rainforest existed all through the Cenozoic, and Neotropical dry forest evolved >30 million years ago (Simon and Pennington, 2012, p. 712). Curiously, Neotropical dry forest is rarely invaded by rainforest or savanna clades, and dry forest clades are far less likely to colonize other biomes than rainforest or savanna clades (Pennington and Hughes, 2014; Pennington and Lavin, 2016).

What Difference Does Even a Short Dry Season Make?

This difference is clearest in Southeast Asia, so we start with evidence from that region, as summarized by Bunyavejchewin et al. (2011) and Ashton (2014).

Southeast Asia

On the Isthmus of Kra between the Malay Peninsula and Thailand, rainfall during the year's driest quarter declines from >300 mm at 5°N to 112 mm at 9°N. On this isthmus, 200 plant genera reach their southern limit where seasonal forest ends, at the Kangar-Pattani line near 7°N, whereas 375 genera reach their northern limit at this line, the northern limit of rainforest. Of the 157 species of the tree family Dipterocarpaceae south of this line, only 27 occur north of it, along with 19 other dipterocarp species which only occur north of this line (Bunyavejchewin et al., 2011).

Fig. 1 View of old-growth seasonal forest on Barro Colorado Island, Panama, facing southward, downslope from the island's central plateau. From left to right, the canopy trees are *Prioria copaifera* (Leguminosae), an epiphyte-laden *Anacardium excelsum* (Anacardiaceae), and *Quararibea asterolepis* (Malvaceae). Drawing by Daniel Glanz, used by permission from the artist.

Baltzer et al. (2009) identified trade-offs driving this "biogeographic knot." In the everwet forest of Pasoh Reserve at 3°N in the Malay Peninsula, trees of species restricted to everwet forest grow faster and survive better than those that also live in the more seasonal Khao Chong forest, north of the Kangar-Pattani line in Thailand. Despite the slight differences in seasonality, species also living at Khao Chong were more drought-tolerant. Their narrower, thicker-walled vessels are less likely to fail under water stress, but they supply their leaves with water more slowly when conditions are favorable, thus reducing photosynthesis. Pasoh trees of species present in Khao Chong also have lower photosynthetic capacity per unit leaf area, and absorb CO_2 from the air less readily (have lower stomatal conductance) than everwet specialists.

Moreover, seasonal forests (even Khao Chong!) differ in phenology from their everwet counterparts (Bunyavejchewin et al., 2011). Everwet forests have little seasonal variation in leaf flush. Many of their tree species fruit simultaneously every few years in episodes of "gregarious fruiting," often associated with El Niño. In seasonal forests, phenology conforms far more nearly to the annual cycle of the seasons, perhaps because it is best for seeds to germinate early in the rainy season, to have time to sink deep roots before the next dry season.

Finally, everwet forests have higher, often far higher, tree diversity than their seasonal counterparts (**Table 2**).

The Neotropics

Although all forest near the Panama Canal is seasonal, there is a gradient of increasing seasonality from the Caribbean to the Pacific side where, over 80 km, annual rainfall declines from 3200 to 1800 mm and rainfall during the year's driest quarter declines from 168 to 75 mm. Condit et al. (2004) showed that species turnover between the two coasts was nearly complete. The 10 most common tree species on a 5 ha plot near the canal's Caribbean end included 1191 of its 2492 trees $\geq$10 cm in trunk diameter, but only one of these species was represented by trees this large (just four such trees) on a 4 ha plot near the Canal's Pacific end. Similarly, the 10 most common species on the Pacific-side plot included 690 of its 1083 trees $\geq$10 cm trunk diameter, but none of these species was represented by a tree this large on the Caribbean-side plot.

Trade-offs driving this "quasi-knot"

The trade-offs driving this "quasi-knot" differ somewhat from those in Southeast Asia because in Panama, as in Ghana (Richards, 1996), soils of wetter forest are less fertile. Engelbrecht et al. (2007) defined the drought sensitivity of a tree species's seedlings as the proportion of seedlings surviving 22 weeks without watering divided by the proportion of comparable but well-watered

Table 2 The relation of tree diversity to dry season severity

Site	*Latitude*	*No. of dry months*	*P*	P_3	P_1	*N*	*S*
Kirindy, Madagascar	20.0°S	8	711	14	0	788	45
Mudumalai, India	11.6°N	4	1250	66	7	245	20
Huai Kha Khaeng, Thailand	15.6°N	6	1476	46	6	438	65
Khao Chong, Thailand	7.6°N	1	2787	358	50	495	195
Pasoh, Malaysia	3.0°N	1	1788	346	94	531	206
Lambir Hills, Malaysia	4.2°N	0	2664	498	153	637	247
Santa Rosa, Costa Rica	11.0°N	5	1668	8	1	354	56
Parque Metropolitana, Panama	9.0°N	4	1832	75	15	318	36
Barro Colorado Island, Panama	9.2°N	3	2551	131	23	429	91
San Lorenzo, Panama	9.3°N	3	3203	168	41	569	87
Santa Rita Ridge, Panama	~9.5°N	2	3586	231	57	479	172
Bajo Calima, Colombia	3.9°N	0	7000	900	200	664	252
Manú, Peru	11.9°S	3	2028	185	55	586	174
Yasuni, Amazonia, Ecuador	0.7°N	0	3081	594	174	701	257

Number *N* of trees ≥10 cm dbh, number *S* of species among them, average annual rainfall *P*, average rainfall P_3 for the year's driest quarter and average rainfall P_1 for the year's driest month, at selected sites.

Data for Kirindy, Madagascar are from Abraham et al. (1996); for Mudumalai, from Sukumar et al. (2004); for Huai Kha Khaeng, from Bunyavejchewin et al. (2004); for Khao Chong, Thailand, from S. Bunyavejchewin and S. J. Davies (unpublished, used with permission); for Pasoh, from Manokaran et al. (2004), for Lambir Hills, from Lee et al. (2004); tree data for Santa Rosa, Costa Rica, are from Burnham (1997), climate data are from the Instituto Meteorological Nacional, Costa Rica, Departamiento de Informacion; tree data for Parque Metropolitana, San Lorenzo and Santa Rita Ridge are from Santiago et al. (2004), rainfall data for Parque Metropolitana and San Lorenzo, Panama are unpublished data from Steven Paton, Smithsonian Tropical Research Institute, used with permission; rainfall data for Santa Rita Ridge are from the Autoridad de Canal de Panama, Hydrologic Resources Section; data for BCI from Leigh et al. (2004); data for Bajo Calima, Colombia, from Faber-Langendoen and Gentry (1991); tree data for Manú, Peru are averages from 13.9 ha from Pitman et al. (2002); rainfall data for Manú were read off from the page facing p. 1 of Gentry (1990); data for Yasuní are from Valencia et al. (2004).

seedlings surviving during these 22 weeks. The more sensitive to drought a species's seedlings, the higher its number of trees ≥1 cm dbh on the Caribbean-side 5-ha plot per tree ≥1 cm dbh on the Pacific-side 4-ha plot ($r^2 = 0.44$, $N = 23$ species).

In another study near the Panama Canal, Santiago et al. (2004, 2005) and Santiago and Mulkey (2005) showed that the species turnover reflected a trade-off between tolerating drought and tolerating poor soil. On the poor soils of wetter sites, leaves are tougher, longer-lived and have less nitrogen per gram than their dry forest counterparts. In wetter forest litterfall, richer in lignin, is more resistant to nutrient cycling, which keeps the soil poor. Finally, Spear et al. (2015) showed that a trade-off between drought tolerance and pest resistance contributed to this species turnover. Moreover, plants forced to grow more slowly by some harsh circumstance such as poor soil must invest more in anti-herbivore defense (Coley et al., 1985).

Phenology and seasonality

From Colombia to tropical Mexico, leaf flush, flowering and fruiting are less seasonal in more nearly everwet forests (Hilty, 1980). Although, as Wright and Calderón (2006) showed, Neotropical trees flower and fruit more abundantly during El Niños, at least in seasonal forest, and El Niños favor gregarious fruiting in Malaysia, Neotropical trees of everwet forest do not fruit gregariously.

On Barro Colorado Island, Panama, and Amazonian Peru's Parque Manu, synchronous seasonal shortage in production of fruit and new leaves largely control populations of vertebrate herbivores (none of which weighs over 500 kg), although big cats may be required to stabilize this control over the long term (Terborgh, 1983, 1990; Leigh, 1999). In paleotropical seasonal forests, seasonal shortages of fruit and new leaves do not coincide, which may explain why paleotropical forests support a higher biomass of primates than their neotropical counterparts (van Schaik et al., 1993).

Indeed, on Barro Colorado Island, the onset of rainy season is like the coming of spring (Foster, 1982). Although some seasonal forest plants, including big trees with colorful flowers, depend on dry season rains to trigger synchronous flowering, many more plants, especially those with small flowers pollinated by small, easily desiccated insects, flower after the rains begin, when small insects become abundant. Many trees bear fruit and/or flush leaves at or soon before the rains. Most canopy trees in seasonal forest can reach water even at the height of the dry season. The onset of the rainy season is also a season of abundance for vertebrate herbivores, the most favorable season for bearing and rearing young.

Seasonality, pest pressure, and tree diversity

In the Neotropics, as in the Paleotropics, tree diversity is higher in forests receiving more rain during the year's driest quarter (**Table 2**). Unlike dry season, poor soil does not diminish tree diversity (Duque et al., 2017) unless the poverty is extreme (Leigh, 1999). Work in Neotropical forests has shown that pest pressure is the primary factor maintaining tropical tree diversity. Kursar et al. (2009) showed that, on Barro Colorado Island, coexisting, recently diverged, species of the tree genus *Inga* differ primarily, often only, in anti-pest defenses, as if the best way to coexist with a related species was to avoid its pests.

On a 50-ha plot at Barro Colorado, all woody plants ≥1 cm trunk diameter of free-standing species were mapped, marked, identified and their trunk diameter measured every 5 years, to learn how tropical tree species coexist. Later, Comita et al. (2010)

marked, identified, and measured the height of all tree seedlings on a 1 × 1 m plot at the center of each of the plot's 5 × 5 m quadrats every year for several years. These data showed that a seedling grew more slowly, and was more likely to die, the more plants of its species were nearby. Moreover, the more a seedling's prospects suffered from an additional nearby plant of its species, the rarer that species, as if the sensitivity of a species's seedlings to near neighbors of its species governed that species's relative abundance. Mangan et al. (2010) found evidence that soil-borne pathogens were what kept each tree species too rare to crowd out the others. In six tree species, a seedling grew more slowly, and died faster, when grown in soil from under trees of its species than when grown in soil from under trees of other species. Moreover, seedlings of rarer species suffered more from being grown in soil from under trees of their own species. Soil pathogens may play the primary role in limiting populations of different tree species.

The role of pests in maintaining the diversity of tropical trees was demonstrated in humid seasonal forest, but pest pressure is probably more intense in more nearly everwet forests, causing their higher tree diversity. First, dry season depresses pest populations, so flushing leaves before the rains begin reduces herbivory (Aide, 1992). In the tropical dry forest of south India, with 1200 mm year^{-1} of rain and a 5-month dry season, trees flush edible new leaves several weeks before the rains begin, thereby reducing losses to insect herbivores (Murali and Sukumar, 1993). By the time rains bring out the insects, these leaves are tougher and harder to eat. Everwet forests, however, offer no opportunity for this ploy. As we have seen, in less seasonal forests mature leaves are tougher and less nutritious, as if deterring pests were a more urgent priority in less seasonal forest. Finally, Alwyn Gentry showed that in forests with weaker dry seasons a higher proportion of tree species attracted mammals to disperse their seeds as if, in less seasonal forests, it was more urgent to disperse seeds further from their parent and its pests (Leigh, 1999, p. 192).

Mildly and Strongly Seasonal Humid Forest

Humid seasonal tropical forest can be separated into mildly seasonal forest, limited more by light than water, and strongly seasonal, primarily water-limited forest. Using various remote-sensing indices, Guan et al. (2015) inferred that in most tropical forest with >2000 mm rain year^{-1}, monthly gross production (total photosynthesis) is at least as high during the dry as the rainy season. This is untrue for the wettest (2671 mm year^{-1}) of the sites compared in **Fig. 1** of Wu et al. (2016), judging by eddy flux measurements of gross production. On the other hand, at all three sites there with >100 mm rain in the year's driest quarter, gross production drops in the latter part of the rainy season, and rises again all through dry season, unhindered by lack of water. Here, leaf flush, and probably flowering, increases in the dry season (van Schaik et al., 1993). In sum, the remote sensing criteria of Guan et al. (2015) probably suffice to distinguish mildly from strongly seasonal forest. If so, most of South America within 5° of the equator, and all Malaysia, Sumatra and Borneo, are everwet or mildly seasonal (Guan et al., 2015).

Strongly seasonal humid forest is primarily water-limited. They cannot benefit from increased dry season light, especially later in the dry season, so total monthly photosynthesis will be lower in the dry season, especially later in that season, than during the rainy season. Guan et al. (2015) infer from their remote sensing indices that most forests with <2000 mm rain year^{-1}, including those on the southern margins of Amazonia, nearly all African forest, and most Thai, and Burmese forest, is strongly seasonal. Eddy flux data from Jaru Reserve, Brazil, with 2040 mm rain year^{-1} but <70 mm during its driest quarter, and from Thai dry forests with 5-month dry seasons, averaging <10 mm during the year's driest 2 months—dry deciduous forest with 1680 mm rain year^{-1} and dry evergreen forest with 1240 mm year^{-1}—gross production declines or reaches its nadir during the dry season (Wu et al., 2016, Fig. 1; Huete et al., 2008, Figs. 2–4). In strongly seasonal forest, shallow-rooted trees should flower and flush leaves when the rains begin, deep-rooted trees in dry season when light is most available (van Schaik et al., 1993).

Of course, whether a forest is mildly or strongly seasonal depends not only on annual rainfall or its seasonality, but on the depth and water-holding capacity of the soil and the rooting depth of its trees. Amazonian forest near Paragominas, Brazil (3°S) averages 1803 mm year^{-1}, 260 mm during its 6-month dry season, and <100 mm during its driest quarter, yet it is not only evergreen, but mildly seasonal. During a 4-year period, evapotranspiration, which is proportional to gross production, averaged 694 mm during the wet, and 821 mm during the dry, half-year. This period, which included the 1992 El Niño year with only 1100 mm, averaged 1550 mm year^{-1}, insufficient to sustain such transpiration indefinitely. This transpiration rate is sustainable with 1800 mm rain year^{-1}, for the tree's deep roots (up to 18 m) allow them to draw water from the top 10 m of soil, which the following rainy season's excess suffices to replenish (Jipp et al., 2008).

The Distinction Between Humid and Dry Seasonal Forest

No clear-cut Kangar-Pattani line separates humid from dry seasonal forest along a gradual climatic gradient. Moreover, dry forests on different continents differ markedly enough to render a universal definition difficult to attain.

Neither the Americas nor Madagascar have long been settled by human beings, and their dry forest trees are not adapted to surviving fire, although fire-prone savanna appeared in South America >7 million years ago (Simon and Pennington, 2012). Indeed, fires mostly stop at the edge of these forests. In these regions, dry forests are characterized by a higher proportion of deciduous trees. This proportion however, is a continuous variate (**Table 3**): distinguishing humid from dry seasonal forests is an arbitrary business.

Trees in these dry seasonal forests differ greatly in how long they are leafless (Pichon et al., 2015) and in the ways they cope with water stress—whether by strong xylem vessels and deep roots, or by storing water in their trunks, shedding their leaves in the

Table 3 Seasonality and deciduousness

Site	*P*	N_{dry}	P_{dry}	P_3	f_d (%)	f_{bd}
Barro Colorado, Panama	2551	3	131	131	12	No data
Santa Rosa, Guanacaste, Costa Rica	1668	5	42	8	~70	No data
Huai Kha Khaeng, Thailand	1476	6	214	46	17	0%
Badrala, Madagascar	1425	7	75	0	80	17%
Kirindy, Madagascar	799	8	100	3	96	7%
Chamela, Jalisco, Mexico	700	8	150	0	96	No data

Annual rainfall *P*(mm), number of consecutive dry months with <100 mm rain N_{dry}, total dry season rainfall P_{dry}, rainfall during the year's driest quarter P_3, fraction of tree species at least briefly deciduous f_d, and fraction of brevideciduous trees (leafless for ≤ 10 days) f_{bd}.

Data for Kirindy are from Sorg and Rohner (1996); data for Badrala from Pichon *et al.* (2015); rainfall data for Barro Colorado and Huai Kha Khaeng are from Leigh (2004); rainfall data for Santa Rosa, Guanacaste are from the Instituto Meteorological Nacional, Costa Rica, Departamiento de Informacion; data on deciduousness at Huai Kha Khaeng are from Williams *et al.* (2008); data on deciduousness in other forests and climate data for Chamela (Jalisco) are from Reich (1995).

Table 4 Impact of fire on the density of small stems

Number of stems ha^{-1}	*Rainforest*	*Thailand dry forest*	*India dry forest*	*Miombo*	*Humid forest*
≥ 1 cm trunk diameter	6707	1450	363	<888	8112
≥ 5 cm trunk diameter	1375	738	272	530	1301
≥ 10 cm trunk diameter	531	438	252	266	438
≥ 30 cm trunk diameter	76	128	110	38	77

Density of stems of various sizes in everwet Malaysian rainforest, fire-prone Thai and south Indian dry forest, and miombo woodland and fire-free humid seasonal forest in the Democratic Republic of the Congo (DRC).

Data for rainforest are for Pasoh, Thailand dry forest for Huai Kha Khaeng, and humid forest for Ituri (Edoro) are from Losos *et al.* (2004); data for miombo are those for Kaspa in Malaisse (1978); data for Indian dry forest are from Sukumar *et al.* (2005).

dry season, and or photosynthesizing with CO_2 respired by the bark without releasing water vapor to the atmosphere (Santiago et al., 2016). The same strategy may be deployed in many climates. Magnificent tall baobabs, *Adansonia* spp. (Malvaceae, Bombacoideae) grow under ~800 mm rain year^{-1} on Madagascar's west coast. They store water in their trunks which they use only to flush new leaves, in response to increasing photoperiod, before the rains come. A look-alike on Barro Colorado Island, Panama, with ~2600 mm rain year^{-1}, the cuipo *Cavanillesia platanifolia* (Bombacoideae), does likewise, as does this island's *Pseudobombax septenatum* (Bombacoideae), which also has vertical green stomate-less stripes on its trunk for photosynthesizing CO_2 respired by the bark (Ávila-Lovera and Ezcurra, 2016). Nitrogen-fixing bacteria enable dry forest trees of the families Papilionaceae and Mimosaceae to both increase water use efficiency and produce more effective anti-herbivore compounds than competitors which lack such bacteria (Pellegrini et al., 2016).

South Asian dry forests, where human beings have lived for a million years, are distinguished by the frequency of forest floor fires (>1 per decade) and their trees' tolerance of these fires (Bunyavejchewin et al., 2011). Fires reduce the density of smaller stems in these forests well below those in humid seasonal forest (**Table 4**).

Many of these forests are largely evergreen, because their poor soil does not provide enough nutrients to renew leaves every year. Yet evergreen and deciduous forests seem equally productive: Huete et al. (2008) found that in Thailand, a mixed deciduous forest on good soil with 1650 mm rain year^{-1} and a dry evergreen forest on poor soil with 1240 mm rain year^{-1}, both with 5-month dry seasons and <20 mm during the year's driest 2 months, have nearly equal gross production, ~ 27.5 tons (Mg) carbon ha^{-1} year^{-1} compared to 34 in everwet forest 60 km N of Manaus (Wu et al., 2016, Fig. 1).

The dry forest at Mudumalai in south India suffers several fires per decade, thanks to its grassy understory. These fires reduce the density of smaller stems well below the levels of dry forest in Thailand (**Table 4**). Sukumar et al. (2005) found that the Mudumalai region averages two elephants km^{-2}: they sometimes inflict great mortality on stems ~5 cm in diameter. As reproduction is so difficult, canopy trees are adapted to live long: mortality of trees over 30 cm trunk diameter is 0.6% year^{-1} at Mudumalai (Sukumar et al., 2004, Table 33.7), compared to 2% year^{-1} on Barro Colorado (Leigh et al., 2004, Table 24.7, 1985–95) and 1.9% year^{-1} at Thailand's Huai Kha Khaeng (Bunyavejchewin et al. 2004, Table 27.7). Like Thai forests with herbaceous or grassy understory, Mudumalai supports high mammal biomass (10 tons (Mg) km^{-2} compared to 1.4 tons km^{-2} at Peru's Manu, Leigh, 1999, p. 164). These mammals support tigers, leopards, and dholes (wild dogs) (Sukumar et al., 2004). Indeed, Mudumalai's ecosystem is one of the most intact known, extraordinarily so for an area so long frequented by human beings.

Dry forests in Africa also burn, on the average, nearly ten times more often, about once per 12 years, than equally dry South American forests (Pellegrini et al., 2016). In Africa, Menault et al. (1995) distinguished two types of dry forest, fire-sensitive "true dry forest," forêt dense sèche, and fire-tolerant woodlands, forêt claire. Both types normally have deciduous canopies. True dry forest also has a layer of understory trees, and casts enough shade that the herb layer is too sparse to carry a fire. The continuous canopy of woodlands, even when

fully leafed, lets through enough light to support a dense herb layer that can carry a fire when dry. Near Lumumbashi, Democratic Republic of the Congo, Malaisse (1978) compared forêt claire, there called miombo, with forêt dense sèche, there called muhulu. Miombo differs far more than muhulu from humid seasonal forest. Miombo lets 27% of the incoming light reach the herb layer, compared to muhulu's 2.3%; its average diurnal temperature range is 17°C, compared to muhulu's 10°C. Although dry forest at Mudumalai. South India, has a thick fire-prone grassy understory, it has three times the miombo's density of trees over 30 cm trunk diameter (**Table 4**). Yet tree diversity is higher in miombo than muhulu. Miombo presumably evolved millions of years ago, before human beings started burning vegetation.

The Importance of Humid Seasonal Forest

Humid seasonal forests have served as essential guides to understanding more diverse everwet forests. In seasonal forest, tree diversity is high enough to challenge explanation, but not high enough to overwhelm the mammologist who needs to know what his study animals are eating or the ethologist who needs to learn what plants her fascinatingly social bees are pollinating, and from what plants—and why these?—they are robbing nectar.

At Barro Colorado Island, Panama, work by independent researchers pursuing questions of their own design developed understanding of its ecosystem's organizing processes. There, biologists could discover how seasonal shortages of fruit and new leaves help regulate populations of vertebrate herbivores, what climate fluctuations cause the occasional famine, and what role birds and bats play in limiting populations of herbivorous insects (Leigh, 1999; Wright et al., 1999; Kalka et al., 2008). Field and laboratory techniques were developed there for analyzing the trade-off between growing fast in bright light and surviving in shade (Leigh, 1999, p. 188; Wright et al., 2003). Studies there on selected pest-limited plant populations (Leigh, 1999, p. 192) prompted Stephen Hubbell and Robin Foster to establish a 50-ha plot where all free-standing woody plants at least 1 cm in trunk diameter were censused. The first census, completed in 1983 took 3 years, and would have taken much longer, had the flora been less well known. This census found 305 species among 235,000 "trees" at least 1 cm stem diameter (Leigh et al., 2004). The plot was recensused every 5 years from 1985 on to track growth, mortality and recruitment. Combining plot data with seedling data from small subplots allowed Comita et al. (2010) to show that a tree species's relative abundance on the plot was greatly influenced by its sensitivity to near neighbors of the same species. Barro Colorado's 50-ha plot inspired similar plots around the world, in everwet, humid seasonal, and dry seasonal forests (Losos and Leigh, 2004): they provided many of this article's data. Plots in continental everwet forest reveal their formidable diversity. A 25-ha plot 80 km N of Manaus, Brazil, has 1302 species among 150,000 "trees" at least 1 cm stem diameter, and 930 species among 13,345 trees at least 10 cm trunk diameter (Duque et al., 2017). Even so, thanks to the unparalleled dialogue between Barro Colorado's 50-ha plot directors and independent biologists with varied interest and expertise, this plot remains a primary source of ideas on new data to collect and new questions to pose these data.

Another, somewhat less seasonal, much more diverse humid forest that has attracted effective question-driven research by different independent investigators is the Cocha Cashu Research Station in southeastern Amazonian Peru's Parque Nacional Manú. As at Barro Colorado, research at Manú is devoted to studying processes that organize its ecological community. Manú's community is more complete than Barro Colorado's in its variety of habitats and its assemblage of larger predators such as jaguars, pumas and harpy eagles. Work there has demonstrated the roles of predation and habitat diversity in governing mammal populations (Terborgh, 1983, 1990). Question-oriented work has also revealed the role of migration in organizing the savanna ecosystem of the Serengeti (McNaughton, 1985). Studies focused on a particular site are particularly fruitful for they often set each other in enlightening perspective.

Study of whole tropical forest communities, including animals, is essential for a proper understanding of their trees. Animal pollinators allow tree species to maintain high genetic diversity even when pests keep them rare. Animals dispersing seeds of rare species allow some of the resulting seedlings to escape their pests until they have grown big enough to resist them. These animals allow plants to shift resources from anti-pest defense to fast growth, thus promoting tropical forest productivity (Leigh, 1999), as does the intense predation on herbivorous insects by birds and bats. Steinberg et al. (1995) graphically illustrated how profoundly predators can shape the properties of plant communities. In the Aleutians, sea otters historically ate enough sea urchins—principal consumers of kelp—to keep shallow-water kelps nearly urchin-free. These kelps can thus severely reduce their anti-herbivore defense, fueling fast growth, whereas their New Zealand counterparts must invest heavily in urchin-deterring toxins because they have no sea otters to protect them. Steinberg *et al.*'s findings are a clear and urgent lesson for tropical forest ecologists: ignoring animals is not the way to understand tropical forest.

See also: Ecosystems: Riparian Wetlands. Evolutionary Ecology: Colonization; Association. General Ecology: Biomass. Global Change Ecology: Phosphorus Cycle

References

Abraham, J.-P., Benja, R., Randrianasolo, M., *et al.*, 1996. Tree diversity on small plots in Madagascar: A preliminary review. Revue d'Ecologie (La Terre et la Vie) 51, 93–117.
Aide, T.M., 1992. Dry season leaf production: An escape from herbivory. Biotropica 24, 532–537.
Ashton, P., 2014. On the forests of tropical Asia. Kew: Kew Publications.

Ávila-Lovera, A., Ezcurra, E., 2016. Stem-succulent trees from the old and new world tropics. In: Goldstein, G., Santiago, L.S. (Eds.), Tropical tree physiology. Switzerland: Springer, pp. 45–65.
Baltzer, J.L., Grégoire, D.M., Bunyavejchewin, S., Moor, N.S.M., Davies, S.J., 2009. Coordination of foliar and wood anatomical traits contributes to tropical tree distributions and productivity along the Malay-Thai peninsula. American Journal of Botany 96, 2214–2223.
Bunyavejchewin, S., Baker, P.J., LaFrankie, J.V., Ashton, P.S., 2004. Huai Kha Khaeng forest dynamics plot, Thailand. In: Losos, E.C., Leigh Jr, E.G. (Eds.), Tropical forest diversity and dynamism. Chicago, IL: University of Chicago Press, pp. 482–491.
Bunyavejchewin, S., Baker, P.J., Davies, S.J., 2011. Tropical seasonally dry forests in continental Southest Asia: Structure, composition and dynamics. In: McShea, W.J., Davies, S.J., Bhumpakphan, N. (Eds.), The ecology and conservation of dry forests in Asia. Washington DC: Smithsonian Institution Scholarly Press, pp. 9–35.
Burnham, R.J., 1997. Stand characteristics and leaf litter composition of a dry forest hectare in Santa Rosa National Park, Costa Rica. Biotropica 29, 384–395.
Coley, P.D., Bryant, J.P., Chapin, F.S. III, 1985. Resource availability and plant anti-herbivore defense. Science 230, 895–899.
Comita, L.S., Muller-Landau, H.C., Aguilar, S., Hubbell, S.P., 2010. Asymmetric density dependence shapes species abundances in a tropical tree community. Science 329, 330–332.
Condit, R., Aguilar, S., Hernandez, A., *et al*., 2004. Tropical forest dynamics across a rainfall gradient and the impact of an El Niño dry season. Journal of Tropical Ecology 20, 51–72.
Duque, A., Muller-Landau, H.C., Valencia, R., *et al*., 2017. Insights into regional patterns of Amazon forest structure, diversity and dominance from three large terra-firme forest dynamics plots. Biodiversity and Conservation 26, 669–686.
Engelbrecht, B.M.J., Comita, L.S., Condit, R., *et al*., 2007. Drought sensitivity shapes species distribution patterns in tropical forests. Nature 447, 80–82.
Faber-Langendoen, D., Gentry, A.H., 1991. The structure and diversity of rain forests at Bajo Calima, Chocó region, Western Colombia. Biotropica 23, 2–11.
Foster, R.B., 1982. The seasonal rhythm of fruitfall on Barro Colorado Island. In: Leigh Jr., E.G., Rand, A.S., Windsor, D.M. (Eds.), The ecology of a tropical forest: Seasonal rhythms and long-term changes. Washington, DC: Smithsonian Institution Press, pp. 151–172.
Gentry, A.H. (Ed.), 1990. Four Neotropical rainforests. New Haven, CT: Yale University Press.
Guan, K., Pan, M., Li, H., *et al*., 2015. Photosynthetic seasonality of global tropical forests constrained by hydroclimate. Nature Geoscience 8, 284–289.
Hilty, S.L., 1980. Flowering and fruiting periodicity in a premontane forest in Pacific Colombia. Biotropica 12, 292–306.
Huete, A.R., Restrepo-Coupe, N., Ratana, P., *et al*., 2008. Multiple site tower flux and remote sensing comparisons of tropical forest dynamics in monsoon Asia. Agricultural and Forest Meteorology 148, 748–760.
Jipp, P.H., Nepstad, D.C., Cassel, D.K., de Carvalho, C.R., 2008. Deep soil moisture storage and transpiration in forests and pastures of seasonally-dry Amazonia. Climatic Change 39, 395–412.
Kalka, M.B., Smith, A.R., Kalko, E.K.V., 2008. Bats limit arthropods and herbivory in a tropical forest. Science 320, 71.
Kursar, T.A., Engelbrecht, B.M.J., Tyree, M.T., 2005. A comparison of methods for determining soil water availability in two sites in Panama with similar rainfall but distinct tree communities. Journal of Tropical Ecology 21, 297–305.
Kursar, T.A., Dexter, K.G., Lokvam, J., *et al*., 2009. The evolution of antiherbivore defenses and their contribution to species coexistence in the tropical tree genus *Inga*. Proceedings of the National Academy of Sciences of the United States of America 106, 18073–18078.
Lee, H.S., Tan, S., Davies, S.J., *et al*., 2004. Lambir forest dynamics plot, Sarawak, Malaysia. In: Losos, E.C., Leigh Jr, E.G. (Eds.), Tropical forest diversity and dynamism. Chicago, IL: University of Chicago Press, pp. 527–539.
Leigh Jr., E.G., 1999. Tropical forest ecology. New York: Oxford University Press.
Leigh Jr., E.G., 2004. How wet are the wet tropics? In: Losos, E.C., Leigh Jr., E.G. (Eds.), Tropical forest diversity and dynamism. Chicago, IL: Chicago University Press, pp. 43–55.
Leigh Jr., E.G., Lao, S.L., Condit, R., *et al*., 2004. Barro Colorado Island forest dynamics plot, Panama. In: Losos, E.C., Leigh Jr., E.G. (Eds.), Tropical forest diversity and dynamism. Chicago, IL: University of Chicago Press, pp. 451–463.
Losos EC and the CTFS Working Group (2004) The structure of tropical forests. In: Losos EC and Leigh EG Jr. (eds.) *Tropical forest diversity and dynamism*, pp. 69–78. Chicago, IL: University of Chicago Press.
Losos, E.C., Leigh Jr., E.G. (Eds.), 2004. Tropical forest diversity and dynamism. Findings from a large-scale plot network. Chicago, IL: University of Chicago Press.
Malaisse, F., 1978. The miombo ecosystem. In: UNESCO/UNEP/FAO (Ed.), Tropical forest ecosystems. Paris: UNESCO, pp. 589–606.
Mangan, S.A., Schnitzer, S.A., Herre, E.A., *et al*., 2010. Negative plant-soil feedback predicts tree species relative abundance in a tropical forest. Nature 466, 752–755.
Manokaran, N., Quah, E.S., Ashton, P.S., *et al*., 2004. Pasoh forest dynamics plot, peninsular Malaysia. In: Losos, E.C., Leigh Jr., E.G. (Eds.), Tropical forest diversity and dynamism. Chicago, IL: University of Chicago Press, pp. 585–598.
Maurin, O., Davies, T.J., Burrows, J.E., *et al*., 2014. Savanna fire and the origins of the "underground forests" of Africa. New Phytologist 204, 201–214.
McNaughton, S.J., 1985. Ecology of a grazing ecosystem: The Serengeti. Ecological Monographs 55, 259–294.
Menault, J.C., LePage, M., Abbadie, L., 1995. Savannas, woodlands and dry forests in Africa. In: Bullock, S.H., Mooney, H.A., Medina, E. (Eds.), Seasonally dry tropical forests. Cambridge: Cambridge University Press, pp. 64–92.
Murali, K.S., Sukumar, R., 1993. Leaf flushing phenology and herbivory in a tropical dry deciduous forest, southern India. Oecologia 94, 114–119.
Pascal, J.P., 1988. Wet evergreen forests of the western Ghats of India. Pondicherry, India: Institut Francais de Pondichéry.
Pellegrini, A.F.A., Staver, A.C., Hedin, L.O., Charles-Dominique, T., Tourgée, A., 2016. Aridity, not fire, favors nitrogen-fixing plants across tropical forest and savanna biomes. Ecology 97, 2177–2183.
Pennington, R.T., Hughes, C.E., 2014. The remarkable congruence of new and old world savanna origins. New Phytologist 204, 4–6.
Pennington, R.T., Lavin, M., 2016. The contrasting nature of woody plant species in different neotropical forest biomes reflects differences in ecological stability. New Phytologist 210, 25–37.
Pichon, C., Hladik, A., Hladik, C.M., *et al*., 2015. Leaf phenological patterns of trees, shrubs and lianas in a dry semi-deciduous forest of north-western Madagascar: Functional types and adaptive significance. Revue d'Ecologie (la Terre et la Vie) 70, 197–212.
Pitman, N.C.A., Terborgh, J.W., Silman, M.R., *et al*., 2002. A comparison of tree species diversity in upper Amazonian forests. Ecology 83, 3210–3224.
Pyke, C.R., Condit, R., Aguilar, S., Lao, S., 2001. Floristic composition across a climatic gradient in a neotropical lowland forest. Journal of Vegetation Science 12, 553–566.
Reich, P.B., 1995. Phenology of tropical forests: Patterns, causes and consequences. Canadian Journal of Botany 73, 164–174.
Richards, P.W., 1996. The tropical rain forest, 2nd edn. Cambridge: Cambridge University Press.
Santiago, L.S., Mulkey, S.S., 2005. Leaf productivity along a precipitation gradient in lowland Panama: Patterns from leaf to ecosystem. Trees 19, 349–356.
Santiago, L.S., Kitajima, K., Wright, S.J., Mulkey, S.S., 2004. Coordinated changes in photosynthesis, water relations and leaf nutritional traits of canopy trees along a precipitation gradient in lowland tropical forest. Oecologia 139, 495–502.
Santiago, L.S., Schuur, E.A.G., Silvera, K., 2005. Nutrient cycling and plant-soil feedbacks along a precipitation gradient in lowland Panama. Journal of Tropical Ecology 21, 461–470.
Santiago, L.S., Bonal, D., De Guzman, M.E., Ávila-Lovera, E., 2016. Drought survival strategies of tropical trees. In: Goldstein, G., Santiago, L.S. (Eds.), Tropical Treee Physiology. Switzerland: Springer, pp. 243–258.
van Schaik, C.P., Terborgh, J.W., Wright, S.J., 1993. The phenology of tropical forests: Adaptive significance and consequences for primary consumers. Annual Review of Ecology and Systematics 24, 353–377.

Simon, M.F., Pennington, T., 2012. Evidence for adaptation to fire regimes in the tropical savannas of the Brazilian cerrado. International Journal of Plant Sciences 173, 711–723.
Sorg, J.P., Rohner, U., 1996. Climate and tree phenology of the dry deciduous forest of the Kirindy forest. Primate Report 46, 57–80.
Spear, E.R., Coley, P.D., Kursar, T.A., 2015. Do pathogens limit the distributions of tropical trees across a rainfall gradient? Journal of Ecology 103, 165–174.
Steinberg, P.D., Estes, J.A., Winter, F.C., 1995. Evolutionary consequences of food chain length in kelp forest communities. Proceedings of the National Academy of Sciences of the United States of America 92, 8145–8148.
Sukumar, R., Suresh, H.S., Dattaraja, H.S., John, R., Joshi, N.V., 2004. Mudumalai forest dynamics plot, India. In: Losos, E.C., Leigh Jr., E.G. (Eds.), Tropical forest diversity and dynamism. Chicago, IL: University of Chicago Press, pp. 551–563.
Sukumar, R., Suresh, H.S., Dattaraja, H.S., Srinidhi, S., Nath, C., 2005. The dynamics of a tropical dry forest in India: Climate, fire, elephants and the evolution of life-history strategies. In: Burslem, D.F.R.P., Pinard, M.A., Hartley, S.I. (Eds.), Biotic interactions in the tropics. Cambridge: Cambridge University Press, pp. 510–529.
Terborgh, J., 1983. Five new world primates: A study in comparative ecology. Princeton, NJ: Princeton University Press.
Terborgh, J., 1990. An overview of research at Cocha Cashu Biological Station. In: Gentry, A.H. (Ed.), Four Neotropical rainforests. New Haven, CT: Yale University Press, pp. 48–59.
Valencia, R., Condit, R., Foster, R.B., *et al.*, 2004. Yasuní forest dynamics plot, Ecuador. In: Losos, E.C., Leigh Jr., E.G. (Eds.), Tropical forest diversity and dynamism. Chicago, IL: University of Chicago Press, pp. 609–620.
Williams, L.J., Bunyavejchewin, S., Baker, P.J., 2008. Deciduousness in a seasonal tropical forest in western Thailand: Interannual and interspecific variation in timing, duration and environmental cues. Oecologia 155, 571–582.
Wright, S.J., Calderón, O., 2006. Seasonal, El Niño, and longer term changes in flower and seed production in a moist tropical forest. Ecology Letters 9, 35–44.
Wright, S.J., Carrasco, C., Calderón, O., Paton, S., 1999. The El Niño Southern Oscillation, variable fruit production, and famine in a tropical forest. Ecology 80, 1632–1647.
Wright, S.J., Muller-Landau, H.C., Condit, R., Hubbell, S.P., 2003. Gap-dependent recruitment, realized vital rates, and size distributions of tropical trees. Ecology 84, 3174–3185.
Wu, J., Albert, L.P., Lopes, A.P., *et al.*, 2016. Leaf development and demography explain photosynthetic seasonality in Amazon evergreen forests. Science 351, 972–976.

Tundra☆

R Harmsen and P Grogan, Queen's University, Kingston, ON, Canada

Introduction

Tundra ecosystems are widely distributed over all continents. Tundra is characterized by climatic stress consisting of low temperatures, strong winds, low precipitation, frost action, and long periods of shortage of liquid water caused by freezing and/or drought. These stresses combine to create what is called the periglacial environment, which is defined by repeated effects of freezing and thawing on soils and water bodies. There are many different kinds of tundra. The two main categories are arctic and alpine tundra. Each of these can be divided into subcategories or can be seen as gradients from a richly vegetated tundra with tall shrubbery adjacent to the 'tree line' through categories with less vegetation to barren areas with only a minimum of vegetation adjacent to icefields and permanently frozen polar or alpine areas. Included in the tundra biome are tundra ponds, lakes, streams, marshes, and other wetlands.

Since tundra is found at the cold limit of life forms on Earth, climatic changes of the past have had major effects on tundra ecosystems and the plant and animal species of these systems. With each Pleistocene ice age, big areas of arctic tundra were eradicated, while others shifted southward and areas of forest or prairie became tundra, only to be reversed during the subsequent interglacial warm periods. Similar changes would have occurred in mountainous regions. These major changes resulted in the local extinction of species and in the disruption of coevolved, interactive plant and animal assemblages. These changes in tundra communities persist today, resulting in low species diversity and relatively simple food webs compared to lower latitude ecosystems. During the current interglacial, many areas that were pushed down by the weight of the ice were first flooded by the rise in sea level, but have subsequently, in part at least, rebounded and developed into tundra. Some parts of Beringia (eastern Siberia, lowland Alaska, and the Yukon) were not glaciated, and remained tundra during the last glacial period. Species of plants and animals that now form arctic tundra communities survived the ice ages either south of the glaciated land area or in unglaciated refuges such as Beringia.

The Periglacial Environment

Periglacial conditions are the result of current or geologically recent frost and ice formations on a landscape. Glaciers affect landscapes in major ways, which can have long-lasting effects on geomorphology, drainage systems, and soils. But even temperature regimes that cause frequent freeze–thaw cycles – for example, annually in the high arctic and daily on high tropical mountains –have not only direct effects on plants and animals, but also indirect effects on soils and water, which results in specific types of erosion and the formation of characteristic landscapes. Furthermore, the presence of permafrost under tundra ecosystems is of critical importance, in that it forms an impenetrable floor, preventing biological penetration and vertical movement of water and nutrients.

Freeze–thaw cycles cause expansion and contraction of soils and water, while the gradual freezing of wet soils will also cause a nonrandom redistribution of water into ice lenses and ice wedges. These processes can result in frost heave and long-term vertical and horizontal movements of soils, debris, and even large rocks, creating typical landscape features (such as polygons), frost mounds (such as palsas and pingos), slope solifluction, and others. These characteristic land forms in turn affect the vegetation and all other life forms.

Permafrost is defined as ground that has been continuously frozen for at least 2 years, and it is very common in the arctic tundra, but less frequently found in alpine tundra sites, as alpine landscapes are more diverse and the summers are warmer. It is not found in any but the highest tropical mountains. During spring, the thawing of the soil starts from the surface down, gradually releasing the vegetation from the grip of the frozen soil. There is usually an overlap between snow melt and the thawing of the soil, especially in undulating landscapes. All melt water must run off, accumulate in low areas, or evaporate, as no vertical movement of water is possible due to the impervious permafrost. This can cause erosion, affecting plants and small animals. During the summer, thawing of the soil above the permafrost continues till autumn, when the surface may already have started to refreeze. During later autumn and early winter, the frost will penetrate deeper into the soil from the surface, as it also comes up from the main body of the permafrost. This process can cause considerable expansion and result in frost heave and can cause much damage to root systems and animal burrows. In many areas, nutrient supply from tundra soils is low because soil decomposition is limited by the cool temperatures and high moisture contents resulting from the impenetrable permafrost below.

Some lakes (e.g., kettle lakes and moraine lakes) have their origins in major ice formations dating from the latest ice age, while others are recent formations. Tundra lakes and ponds are severely affected by annual freezing, especially those lakes that freeze each winter right to the bottom and beyond to the permafrost. Freezing of lake ice causes expansion and results in the shoreline with its vegetation being elevated above the surrounding lowlands. Lake sediments are often high in inorganic matter from spring runoff, but low in organic matter due to low productivity, reflecting low nutrient levels. Frost action and wind effects on ice tend to disturb lake sediments in shallow lakes.

☆*Change History:* February 2013. R Harmsen updated all sections and the Reference List.

Landscape and Species Diversity

Many parts of the arctic tundra are flat, especially in areas adjacent to the sea. These areas are often covered with ponds and shallow lakes, separated by marshes and connected by meandering streams and rivers. These areas can accumulate peat and develop into fens. Along the seashore, these habitats tend to merge into salt marshes, brackish lagoons, and beach ridges. On higher ground, with hills and rock outcrops, the landscape diversity is much greater, especially because north and south facing slopes have very different microclimates, and hence, very different biological communities. Here one can also find deep lakes and fast-flowing rivers. Both erosion and the underlying rock type also add to the ecosystem diversity. In mountainous areas, the low altitude arctic tundra merges at higher altitudes into a more alpine version.

The diversity of alpine tundra worldwide is enormous, as it is found on all continents and in many climatic zones. Snow accumulation during winter, combined with slope, wind, and summer climate affects the length of the growing season of alpine tundra ecosystems. Tropical alpine tundra occurs only at very high altitudes, with unique climates varying from desert to some of the wettest conditions on Earth (Fig. 1). It should also be noted that many of the alpine tundra zones are isolated from other such zones by hundreds or even thousands of kilometers, so their flora and fauna have undergone independent evolution. Especially, geologically old high mountains contain many endemic species derived from local forest or savannah species. For instance, the Southern Alps of New Zealand have over 600 species of alpine plants, very few of which are found elsewhere on Earth.

Roughly 5% of Earth's surface is covered by arctic vegetation and 3% by alpine vegetation. The alpine tundra worldwide, as well as per hectare for most alpine systems, has a much higher biodiversity than the arctic lowland tundra. Species richness declines with altitude on mountains and with latitude in the arctic, and is also dependent on local climatic conditions, nutrient availability, etc.

Vegetation and Succession

Whether one climbs a mountain and crosses the timberline, or travels northward in the arctic and crosses the tree line, one enters the low tundra, which is characterized by shrubs. A combination of low temperatures, shallow soils, and strong winds prevents tree growth, but a tight shrub cover manages to thrive under such conditions. On each mountainous area on Earth, shrub tundras can be found, which are superficially quite similar to other isolated alpine shrub tundra communities; even many of the individual species have a remarkably similar appearance. However, mostly unrelated species form such shrub communities in different parts

Fig. 1 Mount Kenya, tropical Africa. High tropical alpine tundra. In the foreground, a boulder moraine with lichens, mosses, scattered tussock grasses, and a few rosettes of the large *Seneciodendron keniensis*. In the middle ground, a sparse stand of the yet larger *Seneciodendron keniodendron*. The genus *Seneciodendron* is endemic to the east-central African mountains. In the background, the Tyndall Glacier. Photo by W. C. Mahaney.

Fig. 2 Hudson Bay Lowlands, Northern Manitoba, Canada, 60°N. Low arctic willow (*Salix* spp.) and graminoid tundra. Note the radio collar on the polar bear.

of the world. For instance, most of the species of the shrub vegetation on East Africa's Mount Kenya, New Guinea's Mount Wilhelm, and Pico Mucuñuque of the Venezuelan Andes belong to different families. This is a good example of convergent evolution acting on divergent taxa, causing similar adaptation to a specific environment in distantly related taxonomic groups. The shrub zone in the Canadian arctic has a more impoverished vegetation than the shrub zones on tropical mountains. It is dominated by several species of willow and birch, and a smattering of other species (Fig. 2). Again, the arctic tundra in Greenland, Scandinavia, or Siberia also looks very similar, but in this case the species are all close relatives or even the same circumpolar species on the different continents. Another difference is that on tropical mountains there are a lot of shrub species that are not found below the tree line, whereas in the arctic many of the shrub species are also found south of the tree line. These differences are the result of the different effects of the ice ages, which on mountains merely caused the vertical movement of more or less entire plant communities up and down alpine valleys and slopes, while in the arctic, changes in the climate can cause north–south displacements of the conditions suitable for shrub tundra over hundreds of kilometers.

The more typical graminoid, forb, and moss tundra found higher up the mountains and further north in the arctic is adapted to extreme cold, long periods of temperatures permanently below freezing (and permanent darkness in the arctic), and strong winds. It is the strong winds blowing ice crystals that abrade any vegetation above snow level, combined with desiccation that makes tree and tall shrub growth impossible in high arctic and alpine tundra. Especially in arctic deserts, where snow cover is low, vegetation remains low, very close to the ground (Fig. 3). For instance, on Banks Island at 70°N, arctic willow (*Salix arctica*) grows horizontally along the ground, forming matted areas of intertwining branches that form catkins and leaves in summer. One such willow can live and grow for decades. All grasses, sedges, and forbs die in autumn and survive the winter as below-ground root masses, or as ground-hugging rosettes.

One advantage of being a plant in a dense, close-to-the-ground plant community is that on cool, sunny summer days, radiant heat from the 24 h solar radiation is trapped within the air between the plants, keeping temperatures high enough for growth and maturation of seeds. There are very few annual plants in the high arctic tundra, because the season is not long enough for them to germinate, grow, and reproduce. A few very small species, such as *Koenigia islandica* and *Montia lamprosperma*, maintain an annual life strategy. Uniquely, a few species of semiparasitic members of the Scrophulariaceae, such as *Euphrasia arctica*, do so as well. These species have a distinct early season advantage, being able to grow very rapidly by gaining nutrients and photosynthate from neighboring perennials.

The frequent disturbances due to the freeze–thaw cycles often lead to local eradication of vegetation. This creates openings for reinvasion and subsequent succession. One of the most interesting examples of this is the result of solifluction of soil clumps on south-facing slopes in the arctic. Soil clumps with vegetation surrounded by clefts get heated by the sun on the downslope side, causing them to thaw out and slump downwards, burying the lowest vegetation, while at the same time exposing a small strip of upslope bare soil (Fig. 4). It takes up to 30 years for the clump to make one entire downhill rotation. On each clump, one can see a successional sequence of plant maturity, species composition, and diversity, as the oldest community gets buried and an opening appears at the top end for reinvasion. Succession on a larger scale occurs after slope collapses, frost mounting, stream erosion, mud deposits after flooding, etc.

Ecosystem Structure and Function

Very few species remain active within arctic tundra ecosystems during the winter. Only mammals such as the muskox (*Ovibos moschatus*), the reindeer (*Rangifer tarandus*), the arctic hare (*Lepus arcticus*), lemmings, and the wolf (*Canis lupus arctos*) remain fully active. A few birds, for example, the raven (*Corvus corax*) and the rock ptarmigan (*Lagopus mutus*), manage as well. During the autumn and early winter months, soil microbial metabolic activity continues down to at least $-12\,^{\circ}C$. The vast majority of

Fig. 3 Banks Island, Northwest Territories, Canada 70°N. Upland high arctic tundra, also described as arctic desert. The vegetation is dominated by mountain avens (*Dryas integrifolia*), various species of arctic vetch (*Oxytropis* spp. and *Astragalus* spp.), and scattered clumps of small graminoids. In the background is the Thomsen River valley with sedge meadow tundra and tundra ponds.

Fig. 4 Banks Island, Northwest Territories, Canada 70°N. Two types of high alpine tundra. In the foreground, a wet graminoid tundra fed by snowmelt water. On the opposite slope, a sparsely vegetated dry tundra showing solifluction. The muskoxen feed primarily on the graminoid slope, but will venture onto drier tundra types to feed on high nitrogen species such as arctic vetch.

organisms that spend the winter on the tundra do so in some form of dormancy. Alpine tundra, being much more diverse, and much of it having periods of daylight throughout the year, varies greatly in the degree of winter activity of the fauna. The brief summer on the tundra is enormously productive and provides food for a wide variety of organisms. The vegetation starts to bloom and grow as soon as the snow starts to melt. At that time of year the sun hardly sets, if at all, and temperatures rise quickly. Dormant overwintering insect larvae start to feed and eggs eclose to add innumerable larvae in the snow melt ponds, in the soil, and on the new vegetation. The ecosystem seems to burst into active life. High availability of edible vegetation, exploding insect, bird and rodent populations, and young birds lasts till just before the freeze-up in autumn (**Fig. 4**).

Many bird species migrate annually from more southerly wintering sites to the tundra to breed, taking advantage of, and adding to, the burst of summer productivity. Some of these species arrive in extremely large numbers. Most of these birds are insectivorous or feed on pond crustaceans; some such as loons and grebes are piscivorous, falcons and hawks are predators, and geese are herbivorous. The colonially nesting geese, especially, can have major destructive effects on the vegetation, which in turn can affect many other species.

In some tundra ecosystems, some small mammals, especially two species of lemmings, show extreme oscillations in population density, making them keystone species in the tundra ecosystem. For instance, on Banks Island in northern Canada, both the collared lemming (*Dicrostonyx torquatus*) and the brown lemming (*Lemmus sibiricus*) undergo sharp population oscillations within a 3–5-year period. At peak populations the lemmings are all over the place, whereas the year after it is hard to find a single lemming. During the outbreak phase, several predatory birds, including snowy owls (*Nyctea scandiaca*), rough-legged hawks (*Buteo lagopus*), and jaegers (*Stercorarius* spp.), migrate long distances and concentrate in the regions with high lemming populations. They lay large clutches and raise many young, only to disperse to other areas when the lemming population collapses (**Fig. 5**). Mammalian predators are not as able to respond by migration. Arctic foxes (*Alopex lagopus*) and ermines (*Mustela erminea*) are the main mammalian predators; they also take advantage of lemming outbreaks with large litters. However, this leaves relatively dense populations of these predators after the collapse of the

Fig. 5 Nest of snowy owl (*Nyctea scandiaca*) with six eggs and one hatchling. Snowy owls start incubating as soon as their first egg is laid so that the young are hatched sequentially. Note the seven dead lemmings surrounding the nest, intended as food for the hatchlings. Later that summer, the lemming population crashed. Only the two eldest hatchlings survived to fledge, the others being eaten by the older ones.

lemming population. This has a major feed-forward effect in that the half-starved predators exert a strong negative effect on other less favored prey species, mostly birds, from small passerines to ducklings and even goslings. Only after the predator population has collapsed can the lemming population start to grow again.

The ultimate cause of the collapse of the lemming population is not the predation pressure, but the exhaustion of quality vegetation and a delay in nutrient cycling. However, once the lemming population has collapsed, the subsequently declining predator population can drive the lemming population further down to its minimum. The vegetation, litter layer, and soils are strongly affected by the lemming cycles. This is shown by the enormous difference between the tundra in northern Canada and central Greenland, as in Greenland there are no lemmings, much more accumulated litter, differences in relative abundance of plant species, and far fewer predators. Exclosure experiments in Canadian tundra have produced similar results.

Special Adaptations to Tundra Conditions

Many species have evolved special adaptations to the rigorous, but often predictable, conditions of the tundra. This article presents four cases of such adaptations as examples of this phenomenon: the muskox, two species of arctic bumblebee, an alpine lobelia, and two congeneric alpine beetle species.

The Muskox of Banks Island in Canada's Northwest Territories

The muskox (*Ovibos moschatus*) is a surviving species of the Pleistocene megafauna; it survived the ice age both in Beringia and south of the ice sheet in what is now southern Canada and the northern United States. It has a very long adaptive history in arctic conditions, which shows in a number of very effective adaptations to extreme cold. Besides the obvious anatomical features such as the extremely effective insulating wool under the shaggy guard hair and the front hooves that are perfectly shaped to scratch the hard arctic snow to expose vegetation, this animal has a set of integrated physiological and behavioral traits making up a unique reproductive strategy. A muskox cow responds to her nutritional condition in autumn by not going into a reproductively active phase when in poor condition, and only doing so early in the rutting season when she is in excellent condition. This means that cows in poor condition, which would not have been able to survive the winter and produce a calf the next spring, will live and have another chance at reproduction the next year. The cows that do get pregnant, when faced with a bad winter will either abort their fetus or abandon the calf after birth. Since most calves are born well before the snowmelt and reappearance of new fodder, the cows have to be in good shape to not only carry the calf to birth, but also lactate for several weeks. However, only calves born early in the year have a good chance of gaining enough weight and reserve fat to survive their first winter.

Integrated with this strategy are some significant traits. At birth, the calf weight over cow weight ratio is one of the lowest among ungulates, making abortion or abandonment a relatively minor cost for the cow, which can then cut lactation. Once the calf is born and the cow is lactating, she licks the calf when it urinates and swallows the urine. The urea of the urine is rebuilt into protein by the cow's gut flora and will eventually be available for milk production. This is important because storage of protein over the winter is difficult, and late winter forage is scarce and low in protein. As soon as new forage is available during snowmelt, the cows graze selectively on high-protein vegetation, such as willow catkins and sprouting rosettes of arctic vetch (*Oxytropis* spp.). In far northern parts of their range, muskox cows live long lives, but reproduce only every second or third year and still lose some of their calves.

Two Species of Bumble Bee from the Canadian Arctic

The first author has a personal recollection of working in early July on the tundra on northern Banks Island when in the middle of a snow squall a bumblebee flew by. This seemingly incongruous event is explained by the fact that the common large bumblebee (*Bombus polaris*) has an unusually well insulated thorax, which allows it to keep its flight muscles at $\sim 30\,^{\circ}C$ even when the ambient temperature drops to the freezing point. What is even more special about this species is that the queen also keeps her abdomen near 30 °C, which presumably allows the eggs to develop faster. However, early in the season the queen also warms her eggs and larvae in the nest by inserting her abdomen into the middle of the nest and producing heat by vibrating her flight muscles and circulating the heat to her abdomen. After she has overwintered, the queen builds the nest, often in an abandoned lemming burrow, using bits of dead vegetation and muskox wool. There she raises one brood of workers before switching to start raising reproductive bees for the next year. The other species of bumblebee (*B. hyperboreus*) found on Banks Island is an obligatory brood parasite of *B. polaris*. The queens of this species lay eggs only for reproductives, and lay them in the nest of their host species. This strategy is obviously adapted to the very short summer season in the high arctic tundra, but it also depends on the presence of *B. polaris*. The ratio of the densities of the two species is stabilized by frequency-dependent selection.

Flightless Beetles of the Genus *Parasystatus* on Mount Kenya

In the tussock grass alpine tundra of Mount Kenya between 3200 and 4000 m, there are six described species and at least one undescribed species of the genus *Parasystatus*. These large beetles must be adapted to the diurnal extremes of the climate, which has been described as summer each day and winter each night. Two of these species, *P. elongates* and one undescribed species, have been studied in some detail as to their adaptation to the nightly frost of that zone. *P. elongates* spends its entire larval and pupal development inside a tussock of the grass *Festuca abyssinica*, where it is not affected by the nightly frost. As an adult beetle, it is active by day, shielded from the intense solar radiation by inflated elytra and a shiny, reflective outer cuticle. At night, the beetle hides under vegetation to avoid the worst of the frost; it has an ineffectively high supercooling point (Cooling a liquid to below its freezing point without phase transition; here pertaining to the avoidance of ice formation due to the presence of antifreeze substances and/or the absence of crystallization nuclei.), but an effective freeze tolerance. The other species of the same genus is active well into the night, and protects itself with a much lower supercooling point, but is freezing sensitive. These two different physiological adaptations to nightly frost within one genus indicate that the two species have independently invaded the alpine tundra, rather than having arisen through speciation in the alpine zone. Their being flightless – a typical adaptation to mountain top ecosystems – also rules out invasion from another mountain.

The Giant Lobelia and Its Insect Commensals on Kilimanjaro

Between 3000 and 4000 m on the slopes of Mount Kilimanjaro, the giant lobelia (*Lobelia deckenii*) also has to face the stress of nightly frost, which can be severe due to parts of the Kilimanjaro alpine tundra being relatively dry. The plant has evolved into a ball-shaped rosette consisting of a fleshy center surrounded by concave spiky leaves, which are arranged in such a manner as to trap rainwater. A single plant can contain, trapped in its rosette, a compartmentalized mass of several liters of water. This volume is large enough to prevent it from freezing right to the middle in any one night. Indeed, the center of the plant where the growing tip is located maintains a very even temperature throughout the diurnal cycle. Not surprisingly, this water mass of the lobelia plants with its relatively even temperature has become the breeding environment for a few species of insects with aquatic larvae, the most abundant of which is a chironomid midge. The water in the lobelias also contains microorganisms, which feed on decomposing debris and are in turn food for the insect larvae.

Global Warming and Other Anthropogenic Effects

Extensive research in the arctic and alpine regions including ice core analysis, paleolimnology, palynology, and geomorphology has provided a detailed picture of the climatic history of these regions. This allows us to conclude that, as well as the major changes at the end of the last ice age, frequent climate oscillations have subsequently occurred that caused major changes in tundra ecosystems. Furthermore, there have been times when tundra types existed that are no longer extant. The species complexes that now exist consist of species that have been sufficiently flexible and/or dispersible to have survived the climatic and landscape oscillations of the past. However, this does not necessarily bode well for the future of tundra ecosystems and species, as anthropogenic changes are certain to be increasingly imposed on Earth. Already, the most likely reason for the extinction of most of the Pleistocene arctic megafauna is a combination of climate change and human hunting. The disappearance of the large herbivores at this time caused a major switch in plant dominance on tundra ecosystems from graminoids to mosses, with concomitant changes in long-term soil and peat formation. We must expect similar major changes in the coming century, associated with at least some extinctions. Climate change will be severe and direct human effects will also increase. Already, several species are declining due to pollution and overhunting. Some of the most at-risk tundras (and associated endemic tundra species) will be isolated alpine tundra systems on relatively low mountains, where climatic warming will cause the entire system to be replaced by forest.

Climate change in the arctic is already causing major effects on the tundra ecology. The mean annual air temperatures in the arctic have risen by 2 °C since the 1960s. The sea-ice cover on the Arctic Ocean and the water ways adjacent to islands and continental shores have been declining at a rate of 18% per decade since 1979. By 2012, it was at only 50% of the 1979–2000 average. The two measures, sea-ice cover and air temperature are strongly interdependent, as ice-free water absorbs more solar radiation than ice, causing warming, which in turn melts more ice. Open water also evaporates faster than ice, resulting in more precipitation over adjacent land. Warmer temperatures are increasing the growth of vegetation – especially of shrubs, leading to a 'greening' of the arctic. This in turn has a major effect on the community structure and species composition of the various tundra ecosystems.

The cultural practices of native northern peoples are being affected by earlier breakup and melting of ice in spring, which affects their ability to travel to hunting grounds. In *Inuktituk* (the Inuit native language), the word for 'July' means 'the time of the lake ice breaking,' but the ice now often breaks in June. There is little doubt that the people and the ecology of the arctic are already being significantly affected by the climate warming caused by the anthropogenic fossil fuel emissions that almost entirely derive from the activities of people further south.

See also: General Ecology: Temperature Regulation

Further Reading

A.C.I.A., 2005. Arctic climate impact assessment. Scientific Report. New York: Cambridge University Press.
Chapin, F.S., Jefferies, R.L., Reynolds, J.F., Shaver, G.R., Svoboda, J. (Eds.), 1992. Arctic ecosystems in a changing climate. San Diego: Academic Press.
Chapin, F.S., Körner, C., 1995. Arctic and alpine biodiversity: Patterns, causes and ecosystem consequences. Berlin: Springer.
Coe, M.J., 1967. The ecology of the alpine zone of Mount Kenya. The Hague: Junk.
Craeford, R.M.M. (Ed.), 1997. Disturbance and recovery in Arctic lands. Dordrecht: Kluwer Academic.
French, H.M., Williams, P., 2007. The periglacial environment. Toronto: Wiley.
Goulson, D., 2003. Bumblebees: Their behavior and ecology. Oxford: Oxford University Press.
I.P.C.C., 2007. Climate change: Impacts, adaptation and vulnerability. New York: Cambridge University Press.
Jones, H.G., Pomeroy, J.W., Walker, D.A., Hoham, R.W., 2001. Snow ecology: An interdisciplinary examination of snow-covered ecosystems. Cambridge: Cambridge University Press.
Laws, R.M., 1984. Antarctic ecology. London: Academic Press.
Mahaney, W.C., 1989. Quaternary and environmental research on east African mountains. Rotterdam: Balkema.
Pielou, E.C., 1995. A naturalist's guide to the Arctic. Chicago: University of Chicago Press.
Pienitz, R., Douglas, M.S.V., Smol, J. (Eds.), 2004. Long-term environmental change in Arctic and Antarctic lakes. Dordrecht: Springer.
Rosswall, T., Heal, O.W. (Eds.), 1975. Structure and function of tundra ecosystems,Ecological bulletin, vol. 20. p. 450.
Serreze, M.C., 2010. Understanding recent climate change. Conservation Biology 24, 10–17.
Wielgolaski, F.E. (Ed.), 1997. Ecosystems of the world 3: Polar and alpine tundra. Amsterdam: Elsevier.

Upwelling Ecosystems

TR Anderson and MI Lucas, National Oceanography Centre, Southampton, UK

Introduction

Throughout the world's oceans, phytoplankton community structure and rates of primary production are determined by the interplay between available light and nutrient supply (NO_3^-, Si, PO_4^{2-}, dissolved Fe) as well as by grazing. Winds blowing over the ocean create a surface mixed layer, the depth of which is of great importance for production by phytoplankton. If mixing is vigorous, as is often the case at high latitudes, then nutrients are plentiful but plankton circulating within a mixed layer that may be hundreds of meters deep are exposed to low average light intensities. In contrast, mixing is inhibited in warm stratified waters such as those of the vast subtropical gyres that cover 40% of the surface ocean, in which case light is plentiful and limitation is instead by nutrients. The unique physical circulation of upwelling systems leads to conditions that, to varying degrees, provide both light and nutrients together in quantities that considerably exceed rate-limiting requirements for sustaining maximal growth rates of phytoplankton. As a result, upwelling ecosystems are among the most productive in the ocean.

Upwelling Circulation

The Coriolis effect, whereby the Earth's rotation causes moving bodies at its surface to be deflected, means that wind-driven ocean currents turn right in the Northern Hemisphere, and left in the Southern Hemisphere. The result is horizontal flow at the ocean surface in the so-called Ekman layer, typically tens of meters deep. Upwelling occurs in areas where this flow diverges, the Ekman flow or divergence, so that water displaced at the surface must be replaced by deeper water from beneath. Depending on the nature of this divergence, two major types of upwelling systems can be distinguished.

First, coastal upwelling systems occur where the Ekman layer is directed offshore resulting in flow divergence near the coast. Such systems tend to occur on the eastern boundary of ocean basins, major examples being the Canary, Benguela, Humboldt (Peru), and California Current systems (**Fig. 1**). Offshore Ekman flow in eastern boundary current (EBC) systems is driven by local equatorward winds associated with the pressure gradient between the quasi-stationary atmospheric high-pressure systems over the subtropical oceans relative to adjacent continental low-pressure atmospheric systems. Seasonal north–south progressions of these high-pressure systems (poleward in spring, summer) cause increased upwelling and nutrient supply that, along with increased day length and light, drive latitudinal shifts in phytoplankton biomass and productivity. The other major coastal upwelling system is the Somali Current, driven by seasonal monsoon winds of the Arabian Sea. Coastal upwelling is often enhanced by topographical features such as capes or canyons where local upwelling cells form.

Second, upwelling occurs in the open ocean, being most marked where easterly trade winds give rise to Ekman divergence north and south of the equator. The resulting area of equatorial upwelling in the Pacific is vast, extending westwards from the coast of South America to beyond the international date line. A smaller belt of upwelling occurs in the equatorial Atlantic. In the Southern Ocean, the Antarctic Circumpolar Current contains another zonal upwelling region, most vigorous between 50° and 60°S, driven by northerly Ekman flow that is generated by the strongest prevailing westerly winds in the 40–50° latitudes.

General Characteristics

Nutrients are present in high (e.g., NO_3^-, 35; Si, 30–60; PO_4^{2-}, 1–2 $\mu mol\ l^{-1}$) concentrations in subsurface waters of the global oceans. Upwelling brings them to the surface, fertilizing the resident phytoplankton assemblage. Stratification of the surface mixed layer between the strongest upwelling pulses provides favorable light conditions for algae to grow and take up the nutrients at their disposal. Resulting rates of primary production are often among the highest seen in marine systems. Coastal upwelling systems, for example, occupy just 0.5% of the ocean surface area, yet contribute to 2% of global marine primary production. Supporting an abundance of higher trophic orders such as fish, birds, seals, and whales, they also contain some of the world's major fin-fisheries.

Intermittence is a key feature of upwelling systems. Upwelling intensity is seasonally episodic in systems such as the Canary Current, Benguela, and Somali systems, whereas in others such as the Humboldt and Southern Ocean, upwelling is semi-continuous all year round. In all systems, wind strength varies on shorter timescales of days to weeks, leading to periods of strong and weak upwelling, or times when upwelling ceases altogether. Organisms must be able to tolerate these changes in upwelling intensity and the resulting impact on nutrient supply and spatiotemporal variation in food resources, as well as variations that may occur from year to year and on longer timescales. In addition, they face the prospect of either themselves, or their reproductive products, being swept away in the Ekman layer toward less-favorable habitats. A key feature of these organisms, including phytoplankton, zooplankton, and fish, is that their life histories and behavior are specifically geared toward maintaining populations in the regions of the upwelling centers.

 doi:10.1016/B978-0-444-63768-0.00363-2

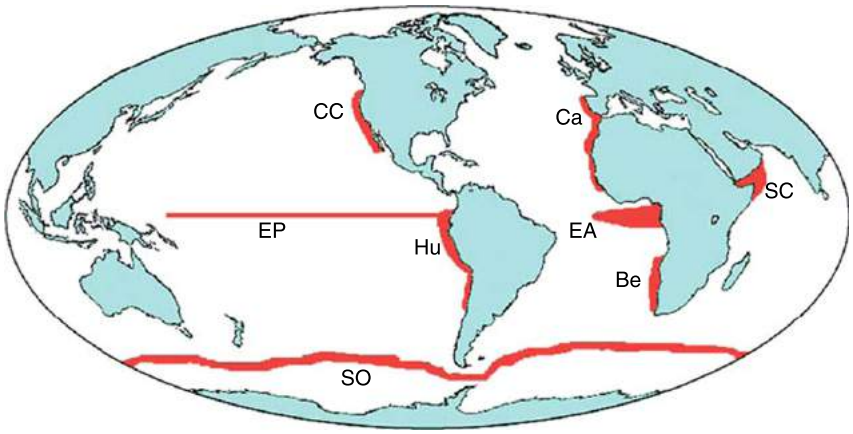


Fig. 1 Global map of major upwelling systems. Be, Benguela; Ca, Canary; CC, California Current; EA, Equatorial Atlantic; EP, Equatorial Pacific; Hu, Humboldt; SC, Somali Current; SO, Southern Ocean.

Understanding the ecosystem structure and functioning of upwelling systems, and in particular how they are influenced by climate variability, is increasingly recognized as essential to the management of sustainable fishery resources.

Primary Production and Lower Trophic Levels

Primary Production

The EBC systems provide a highly favorable combination of light and nutrient supply for primary production by virtue of a strongly shoaling pycnocline (the density gradient that signifies the base of the mixed layer) toward the coast and a relatively shallow (<500 m) shelf environment. A near-surface pycnocline has the dual effect of facilitating the injection of nutrients into surface waters and maintaining phytoplankton in a shallow (usually <50 m) well-lit euphotic layer environment. Shallow shelf sediments augment the nutrient concentration of deeper upwelling source waters. As the four major EBC systems lie predominantly in mid-latitudes (40° N/S to 10° N/S), insolation rates are seasonally high, providing both the light and necessary surface warming and stratification to optimally drive photosynthetic carbon fixation supported by a nutrient-replete environment. Taken together, they have a combined productivity estimated to be ~ 1 Gt C yr^{-1}. Chlorophyll concentrations typically exceed 2 mg m^{-3} but can reach up to 50 mg m^{-3} locally where intense dinoflagellate blooms are present. Highest production rates are consistently found in the Humboldt system (2–6 g C m^{-2} d^{-1}) due to a higher average irradiance and a less-fluctuating nutrient environment associated with relatively consistent upwelling.

A key feature of primary production in upwelling ecosystems is that it is fuelled by large amounts of NO_3^- that is `new' (i.e., allochthonous) to the euphotic zone. Phytoplankton production based on nitrate uptake is therefore termed `new production'. Nitrate arises almost entirely from remineralization of organic matter below the pycnocline, notably from dead phytoplankton and other material that had earlier `rained' down from surface waters. In contrast, `regenerated production' is based on nitrogen (NH_4^+, urea, and dissolved organic nitrogen) excreted by organisms within (i.e., autochthonous) the euphotic zone. The relative importance of new production is often stated by expressing it as a fraction of total phytoplankton production (i.e., the sum of new and regenerated production), this fraction being known as the *f*-ratio. Values for the *f*-ratio are usually high (~ 0.5–0.7) in most EBC upwelling ecosystems, although in seasonally pulsed upwelling systems such as the Southern Benguela, the yearly averaged *f*-ratio is lower (~ 0.3). High rates of new production make available carbon and nutrients for transfer to higher trophic levels and are the fundamental reason why EBC systems can sustain productive fisheries. They also provide the potential for large downward fluxes of sinking particles from the euphotic zone in the event that phytoplankton are inefficiently grazed. In contrast, systems based on regenerated production gradually run downhill unless there are new inputs of nitrogen because nutrients are never recycled with 100% efficiency.

Phytoplankton Community Structure

A characteristic succession is seen in the composition of phytoplankton communities of coastal upwelling ecosystems. This succession is driven by changes in the nutrient and light environment, linked closely with upwelling frequency and the three-dimensional (3D) circulation of water as it flows away from the upwelling centers. Large individual (20–200 μm), colonial, and chain-forming diatoms of up to 500 μm in length proliferate as newly upwelled water arrives and stabilizes in the sunlit surface layer. Analogous to weeds, these algae grow quickly because of intrinsically fast growth rates and an ability to take up nutrients rapidly, provided that concentrations remain sufficiently high. Cell division rates of 2–4 d^{-1} quickly result in population growth

that outstrips zooplankton herbivory, leading to extensive diatom-dominated blooms. The accumulating chlorophyll biomass can exceed 6 mg m^{-3} in just a few days, high enough to be easily visible in ocean color satellite imagery (Fig. 2).

Most diatom species within EBC systems are adapted to avoid lateral dispersal in surface currents away from the upwelling centers. Many species trigger a resting stage in their life cycle in response to diminishing nutrient availability. Spores are formed that sink rapidly and become entrained into deeper shelf-edge waters and surface sediments. Sinking of vegetative cells, or chain-formation that increases sinking rates, provide alternative strategies to counteract dispersal. Spores and physiologically inactive diatoms remain within the sediment–water interface layer and await the next upwelling event that will entrain them back into near-shore and nutrient-rich sunlit surface waters, so initiating another bloom event. The construction of silica tests and spiny frustules from dissolved silicate, along with their large size and ability to form chains and colonies, offer initial protection from herbivorous zooplankton such as copepods. The efficacy of grazers is further weakened in pulsed upwelling systems that do not settle into steady state. As initial exponential growth rates of diatoms are much faster (hours, days) than those of herbivorous copepod consumers (weeks), episodic and short-term upwelling events (days) produce a `mismatch' between phytoplankton and zooplankton, the former breaking free from top-down grazer control that might otherwise prevent blooms from occurring. Nevertheless, the grazing that does occur permits carbon to be efficiently transferred to pelagic fish in a short two-step food chain (diatoms $\Rightarrow$ mesozooplankton $\Rightarrow$ pelagic fish).

A shift in phytoplankton community structure occurs as nutrients become depleted in the well-stratified surface waters downstream of coastal upwelling centers. Diatoms give way to smaller cells such as nanoplanktonic phytoflagellates (2–20 μm) as well as other smaller picoautotrophs (<2 μm) that do not require Si and are better able to scavenge nutrients at low concentrations because of their high surface area to volume ratio. As nutrients in the surface layers are scarce, these small phytoplankton primarily occupy the thermocline where nutrients diffuse slowly from below. A deep chlorophyll maximum (DCM) develops that involves a delicate tradeoff between maximizing nutrient availability, but having sufficient light in a near light-limited environment. Microzooplankton grazers keep the numbers and biomass of these small phytoplankton in check (<0.5–1 μg chl *a* l^{-1}). Within this `microbial loop', particulate organic nitrogen (PON) is efficiently recycled via microzooplankton and bacteria into NH_4^+ and urea to support further phytoplankton growth. As nitrogen remineralization is usually balanced by the rapid uptake of such regenerated nitrogen by phytoplankton, concentrations of these nutrients remain low (<0.5–1 μmol l^{-1}). Paradoxically therefore, grazing pressure is essential to support further algal growth. Carbon is only inefficiently transferred along

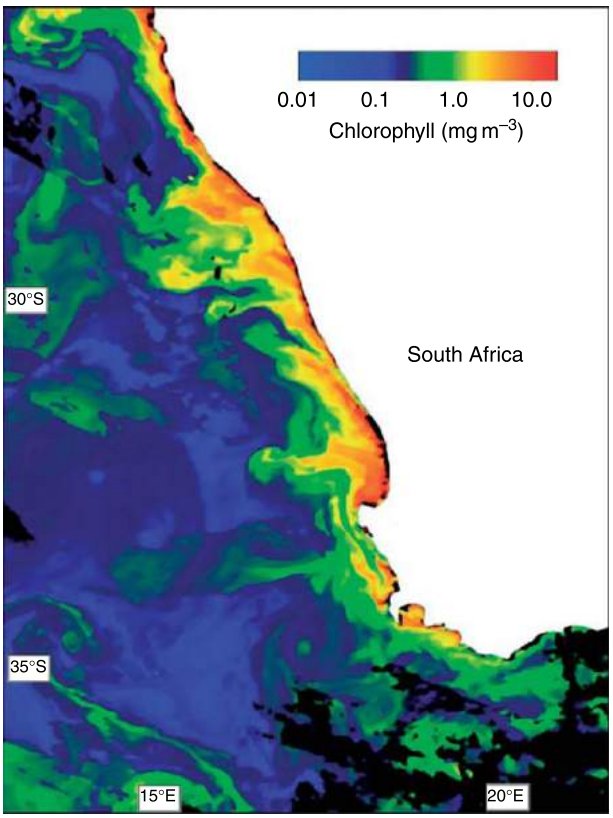


Fig. 2 Satellite image of chlorophyll biomass in the Southern Benguela region. Note high chlorophyll concentrations inshore where upwelling of nutrient rich water is strongest, and also offshore filaments in the chlorophyll signal which weakens offshore as nutrients become exhausted. Courtesy Stewart Bernard, Univ. Cape Town.

an extended food chain (pico-, nanoplankton ⇒ microzooplankton ⇒ mesozooplankton ⇒ fish) because, at each step, something approaching 90% of the transferred carbon is lost through respiration.

Oxygen Depletion

Dead and decaying material is a feature of all ecosystems, those of upwelling areas being no exception. Detrital particles are produced in abundance, either as senescent phytoplankton or as zooplankton fecal material. A `rain' of sinking particulate organic material (`marine snow') is exported from the euphotic zone and decomposed either in mid-water by heterotrophic bacteria, or by both benthic organisms and bacteria on the seafloor. Large quantities of oxygen are consumed, creating an oxygen minimum zone (OMZ) in the sediments and the overlying water. The resulting oxygen concentrations of <0.5 ml l^{-1} are inhospitable to many animals, pelagic and benthic alike.

Coastal upwelling systems are particularly susceptible to hypoxic (<0.5 ml l^{-1}) or anoxic (near-zero O_2) events because of the high rates of diatom-dominated phytoplankton productivity that quite suddenly become nutrient-limited and therefore senescent. The oxycline may often extend to near the surface (<50 m). Most zooplankton that can actively migrate often do so in order to maintain their position in the oxygenated waters near the surface. Others, such as *Eucalanus inermis* in the Humboldt system, are able to withstand low oxygen concentrations, and indeed are known to congregate in the OMZ, perhaps exploiting it as a refuge from predation. In similar fashion, juvenile hake (*Merlucius capensis* and *M. paradoxus*) in the northern Benguela system off Namibia are known to exploit the OMZ as a refuge from their cannabilistic parents!

Oxygen consumption in upwelling areas leads to extensive seafloor habitats subject to permanent hypoxia. Diversity is low, but those animals that tolerate low-oxygen conditions are abundant. Calcareous foraminiferans, nematodes, and annelids utilize the influx of organic material from above, but rely on anaerobic metabolism to do so. Chemoautotrophic bacteria use NO_3^- and sulfur as terminal electron acceptors instead of O_2, first stripping the anoxic water column of NO_3^- (denitrification) before specific anaerobic sulfur and sulfate-reducing bacteria release sulfurous and foul-smelling H_2S into the water column creating so-called `black-tides' and `sulfur eruptions'. However, it appears that some nitrogen losses previously ascribed to denitrification should instead be linked to the process of ammonium oxidation `annamox', first described in Dutch sewerage works.

Offshore hypoxia can have profound effects on the near-shore intertidal environment. When gentle upwelling begins, low-oxygen water is driven into the near-shore and intertidal zone, killing all before it except those most resistant to hypoxia. In the Benguela system, crayfish `walkouts' by animals fleeing from low-O_2 water can leave thousands of tons of crayfish stranded on the beaches in sheltered embayments.

Zooplankton

Great variety is seen in the zooplankton of upwelling ecosystems. Smallest are the microzooplankton, including ciliates, heterotrophic dinoflagellates, and flagellates (typical size 2–5 μm) that efficiently graze the smallest phytoplankton. Reproducing by cell division, their high growth rates (e.g., 1.0 d^{-1}) are similar to those of their algal prey such that their grazing is sufficient to prevent small phytoplankton cells from blooming. Microzooplankton do not fit the general paradigm that organisms eat prey items significantly smaller than themselves. Instead, they have evolved various specialized feeding mechanisms including direct engulfment, tube feeding in which a feeding tube, the peduncle, pierces prey which then has its insides sucked out, and pallium feeding, where a feeding veil envelops and digests prey *in situ*. Prey items as large as, or larger than, the microzooplankter's own body size can be consumed using these adaptations. It has been suggested, for example, that heterotrophic dinoflagellates are able to compete with copepods for diatom prey, although the extent to which this competition operates in marine ecosystems is as yet poorly known.

It is the larger mesozooplankton ($\sim$0.2–2 mm), notably copepods and to a lesser extent euphausiids, that are the major grazers of diatoms and which form the main trophic link with fish and other higher trophic levels. Phytoplankton are captured by filter-feeding or by selective particle capture (raptorial feeding) based on size and/or palatability. *Calanus* is the dominant copepod genus in upwelling ecosystems. Although not reproducing as fast as microzooplankton, it may achieve as many as ten generations per year, each with its own life cycle of eggs, nauplii, copepodites, and adults. When food conditions are favorable, fecundity is high and egg production is rapid. Upon hatching, the planktivorous juvenile stages are swept along in the Ekman layer and will starve if they do not encounter suitably dense patches of appropriately sized food particles. For adults, survival is enhanced by the storage of energy reserves in the form of lipids. Given their size, copepods are unable to maintain their position within upwelling systems by swimming against lateral advection that is often offshore. This problem is overcome by diel vertical migration. Offshore surface Ekman flow is balanced by deeper shoreward flow onto the shelf. By migrating into this deeper layer by day, copepods utilize the natural circulation pattern to maintain their inshore position where food resources are richest.

Euphausiids are also a significant component of the zooplankton community in upwelling ecosystems, for example, *Euphasia lucens* in the Benguela system. They are much larger (1–2 cm) than copepods and have a longer life span of about a year. This longevity, along with an omnivorous diet, means that euphausiids are better able to cope with the fluctuating food conditions of upwelling ecosystems than are copepods. Nevertheless, physical transport away from upwelling centers remains a problem, and these animals also employ diel vertical migration into the subsurface countercurrent to maintain their position in the flow field. Their larger size makes euphausiids a key prey item for larger zooplankton consumers, including baleen whales that are often temporary residents of upwelling systems.

Open-Ocean Upwelling Systems

The general principles and characteristics that govern productivity in EBC regions apply also to the major open-ocean upwelling systems (Equatorial Pacific, Equatorial Atlantic, and Southern Ocean). There are nevertheless key differences, notably that upwelling strength tends to be lower and there is no influence of the seabed (e.g., in supplying iron) on euphotic zone processes, now that it is 3000–4000 m beneath the ocean surface.

The Equatorial Pacific is a vast upwelling system, as well as being a good example of a so-called high-nutrient low-chlorophyll (HNLC) ecosystem. Phytoplankton biomass is generally low and relatively constant ($\sim$0.2–0.4 mg chl *a* m^{-3}) which, along with low productivity of $\sim$0.1–0.5 g C m^{-2} d^{-1}, occurs despite the presence of sufficient macronutrients (NO_3^-, Si, PO_4^{2-}) and light. Iron, however, is in short supply. This micronutrient is needed by phytoplankton to harvest light using their photosynthetic machinery (photosystems I and II), as well as by the enzymes nitrate and nitrite reductase to reduce NO_3^- within cells to NH_4^+. Without a sedimentary source, aeolian supply is the primary source of Fe to the open ocean. However, most aeolian dust supply is from the Saharan desert, far-distant from the Equatorial Pacific. The resulting shortage of iron impacts most severely on large cells, notably diatoms, because of their inability to compete with smaller phytoplankton at low nutrient concentrations. In the western basin, phytoplankton biomass is dominated by small solitary picoplanktonic cells (0.2–2 μm) within a DCM comprising prochlorophytes, *Synechococcus*, and small eukaryotes. These cells utilize what little iron supply there is from the waters upwelled from below, starving the surface ocean of this element. Diatoms are more abundant ($\sim$6%) to the east of $\sim$140 °W where deep nutrient-rich upwelling outcrops at the surface but, nevertheless, the overall biomass is still picoplankton dominated. Grazing by microzooplankton keeps phytoplankton stocks in check, but small natural enhancements of iron that occur in the Equatorial Pacific in response to the passage of tropical instability waves promote transient increases in primary production.

The upwelling region at the Antarctic Polar Front is another HNLC system with low Fe concentrations. Chlorophyll biomass is typically $\sim$0.5 mg m^{-3} in the austral summer with a productivity of $\sim$0.5–1 g C m^{-2} d^{-1}. Unlike the Equatorial Pacific, however, winter gales drive deep mixing that entrains nutrients, including Fe, into surface waters. This is sufficient to initiate short diatom-dominated blooms in the early spring (September, October) as the light environment improves. Iron limitation throughout the rest of the year opens the way for a more typical HNLC community of pico- and nanoplanktonic phytoflagellates that are microzooplankton controlled. Populations of the prymnesiophyte *Phaeocystis antarctica* may also develop, an organism that exists both as solitary cells and mucilaginous colonies, and which is the main producer of volatile organic sulfur (dimethyl sulphide, DMS) in the region.

The Equatorial Atlantic is unique among open-ocean upwelling systems both in terms of its hydrography and because of its close proximity to aeolian dust sources from the Sahara meaning that productivity is far less Fe-limited than in other open ocean systems. Between June and January, water flowing from the Amazon basin floods eastwards (the North Equatorial Counter Current) across the northern margin of the equatorial upwelling region. Stripped of nutrients as it crosses the Amazonian shelf, this fresher, buoyant water forms a layer $\sim$40 m deep that caps the nutrient-rich water below and also limits light penetration. A DCM forms at the juncture of these two water types, the low light intensities being particularly well exploited by the cyanobacterium *Prochlorococcus*. The nutrient-depleted waters above are home to a separate community in which nitrogen fixers such as *Trichodesmium* utilize atmospheric nitrogen as a nutrient source. The aeolian flux of dust plays an important role since nitrogen fixers have a particularly high requirement for Fe.

Between February and May, the Amazonian outflow diverts northwards toward the Caribbean. Phytoplankton now find themselves $\sim$40 m closer to the surface in a higher light environment and productivity increases but, as the rate of upwelling is weak, the upward flux of nutrients is insufficient to support diatom blooms except in the eastern basin near the African coast.

In addition to the major upwelling systems described above, the upper ocean contains numerous mesoscale eddies - whirling current systems analogous to weather systems in the atmosphere but only about a tenth of the size (tens instead of hundreds of kilometers across). Produced through the conversion of potential energy to kinetic energy as part of the ocean's annual energy cycle, both cyclonic and anticyclonic eddies (depending on the vertical structure of the water column) can result in the localized doming of isopycnals (constant density surfaces) and upwelling of nutrient-rich waters into the euphotic zone as they form. Eddies themselves then decay as they release their potential energy over periods of weeks to months, vertical motions both upward and downward occurring on their periphery during this time. Primary production is generally stimulated through nutrient enrichment. As in other upwelling systems, regions of higher nutrients and shallower mixed layer depth associated with eddies tend to promote the growth of larger phytoplankton cells such as diatoms, their concentrations typically being higher within eddies than in surrounding waters. Ubiquitous in nature, eddies provide a significant vertical transport mechanism for nutrients throughout much of the world's oceans.

Fish and Higher Trophic Levels

The EBC upwelling ecosystems of the world support major commercial fisheries based on the shoals of sardine, anchovy, and mackerel that thrive on the abundance of phytoplankton and zooplankton food. In the Humboldt system alone, for example, catches have been around 12 million tons in peak years, although this decreases by >50% during unfavorable conditions. Indeed, stocks of different fish species have been highly variable over the years, suggesting a remarkable responsiveness in ecosystem structure to changing conditions. Understanding the links between fish, lower and higher trophic levels, and environment is essential to ensuring the sustainable management of these important fish resources.

Small Pelagic Fish

The food chain of upwelling systems embraces phytoplankton and zooplankton at its base, linking to small pelagic fish which are in turn consumed by higher predators such as piscivorous fish, birds, and seals (Fig. 3). A curious aspect of this trophic network is that there are many species at low (phytoplankton, zooplankton) and at high trophic levels, but only a few species of small pelagic fish in between. Indeed, the fish biomass of coastal upwelling systems is typically dominated by either a single species of sardine (*Sardinops*) or a single species of anchovy (*Engraulis*) at any one time.

Although food resources are generally favorable, the 3D circulation makes upwelling systems a hazardous environment for fish. Losses of eggs and juvenile stages may occur due to offshore transport or because of starvation when being carried by currents from the spawning to nursery areas. Spawning grounds are therefore often strategically positioned in quieter areas surrounding the upwelling centers such as downstream of capes in sheltered embayments. The result is a complex network of spawning grounds, transport pathways, and migration patterns, a typical example being the Benguela system (Fig. 4). Anchovy spawn on the Western Agulhas Bank in spring and summer (with a maximum in November), while sardine have a longer spawning season in the same area, peaking in both October and March. Once fertilized, eggs and larvae drift northwards in the Benguela `Jet'. Larvae feed selectively on small particles and juvenile fish recruitment occurs at several locations north of St. Helena Bay on the West Coast. There, anchovy recruits feed primarily on larger zooplankton (copepods) as they slowly migrate southwards, returning as 1-year-old adults to the Agulhas Bank to spawn in the following austral spring/summer.

The survival of small pelagic fish is thus determined to a large degree by direct physical factors such as circulation patterns and the intensity and duration of upwelling that simultaneously control egg and larval survival, recruitment success, and food supply. Population control is therefore neither exclusively `bottom-up' via primary producers nor `top-down' by higher predators. Instead it is from the `waist', both up and down, the so-called `wasp-waist' hypothesis. Small pelagic fish provide higher tropic levels such as birds and seals with food, while at the same time keeping phytoplankton and zooplankton numbers in check. As a result, ecosystem functioning as a whole may be remarkably sensitive to fluctuations in pelagic fish numbers. Direct environmental forcing or commercial fishery exploitation of `wasp-waist' populations may cause disruption to these ecosystems by undermining the stability of the entire food web.

Bottom-up and top-down controls of small pelagic fish populations are nevertheless by no means unimportant. Sardines and anchovy, for example, have different feeding strategies. Sardines are mostly indiscriminate filter feeders on phytoplankton and smaller zooplankton, including small cyclopoid copepods, whereas anchovy use biting behavior to selectively ingest individual particles such as larger ($\sim$2 mm) copepods and euphausiids. Strong upwelling should therefore favor anchovy by promoting the diatom growth that supports larger zooplankton. In contrast, the more nutrient-depleted waters present during periods of weaker upwelling favor smaller phytoplankton and consequently smaller zooplankton that are preferred by filter-feeding sardines.

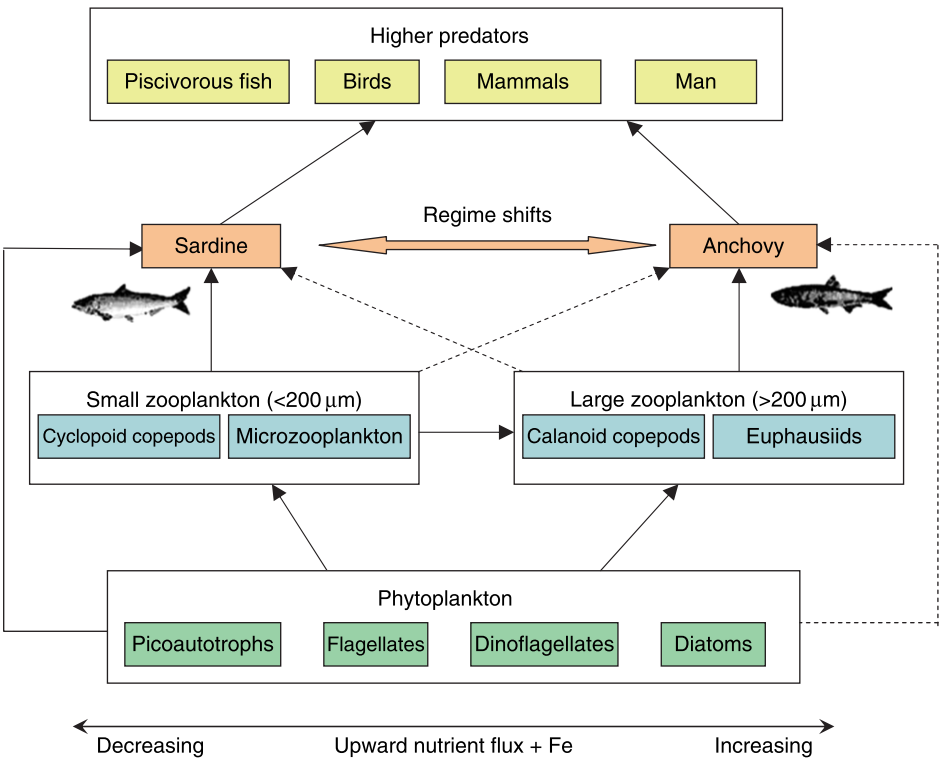


Fig. 3 Idealized flow diagram for an upwelling ecosystem food web. Dashed arrows indicate weak flows.

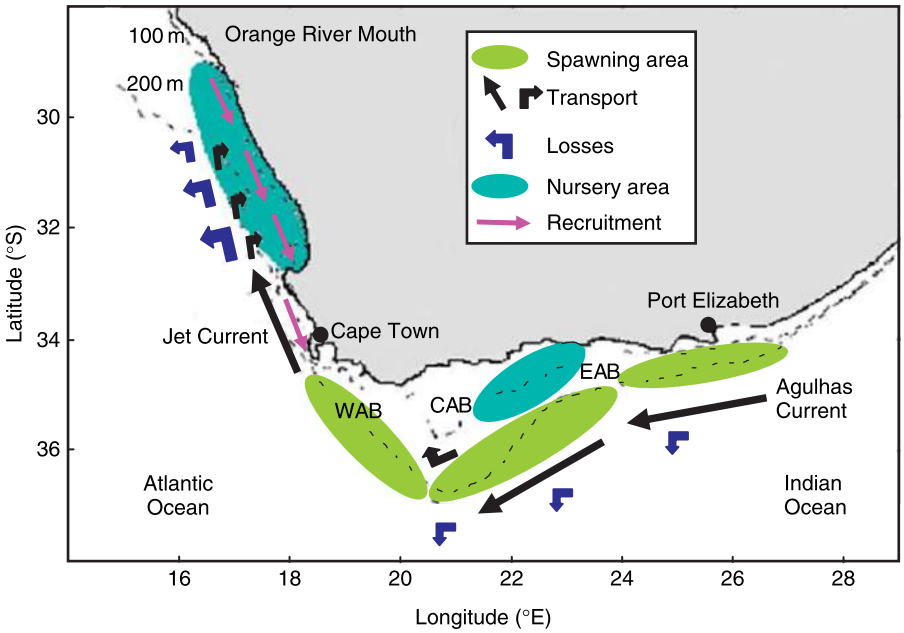


Fig. 4 Map of the Southern Benguela off South Africa showing the locations of small pelagic fish spawning and nursery grounds and transport and loss processes that impact on eggs and larvae. WAB, CAB, and EAB indicate the Western, Central, and Eastern Agulhas Banks, respectively. Redrawn from Lehodey, P., Alheit, J., Barange, M., et al., 2006. Climate variability, fish and fisheries. Journal of Climate 19, 5009–5030.

Variability in small pelagic fish populations has consequences for their predators. Evidence from the Benguela region shows that during periods of pelagic fish abundance populations of piscivorous fish (e.g., snoek, hake), seals and birds (gannets, cormorants) generally increase and, in doing so, begin to exert a stronger top-down control on the small pelagic fish. This in turn relaxes anchovy predation pressure on copepods and so mesozooplankton numbers recover, in turn exerting a higher grazing pressure on phytoplankton. Not only does top-down predator control on small pelagics equal or exceed that by commercial fishermen, it can substantially shape community structure right down to the level of primary producers.

Fish Production

High primary production fuelled by new nutrients undoubtedly contributes to the prodigious fish production of coastal upwelling systems. Back in 1969, John Ryther proposed that high fish yield should be expected where phytoplankton cells are large, or exist as colonies or chains, thereby leading to only one or two trophic links from primary producers to fish. The greater number of trophic links stemming from smaller phytoplankton cells should lead instead to greater respiration losses and recycling of organic matter.

Fish catches vary markedly between EBC systems, typical values being 0.05%, 0.09%, and 0.16% of primary production for the Canary, Benguela, and Humboldt Current systems respectively. Much of this variation may be due to differences in upwelling intensity and frequency that impact on lower trophic levels and fish recruitment, although intensity of fishing may also play a part. The strongly seasonal and pulsed nature of upwelling in the Canary and Benguela systems, for example, leads to temporal mismatches between primary producers and copepods, depressing their fecundity and therefore population size, so reducing the food supply for fish. In contrast the Humboldt system experiences less variation in the intensity of upwelling, leading to tighter coupling between phytoplankton and copepods, greater zooplankton production, and ultimately higher fish production. The low fish catch per unit primary production of the Canary system relative to that of the Benguela is due to its narrow continental shelf (20 km; the Benguela's is ~85 km), such that proportionately more primary production may be advected offshore away from the main zones of fish production.

The impact of environment on the reproductive success of fish in upwelling ecosystems can be thought of in terms of a fundamental triad of processes: enrichment (nutrient supply for primary production), concentration processes (convergence, water-column stability), and retention (within favorable habitat). Acting at the base of the food chain, nutrient enrichment via upwelled water fertilizes primary producers, although excessive wind may deepen the surface mixed layer, leading to light limitation. Upwelling also stimulates small-scale turbulence, increasing encounter rates between both zooplankton and fish larvae and their prey. On a larger scale, convergence promotes food particle aggregation, but divergent upwelling flow will tend to dissipate particles offshore. Based on these pros and cons, an optimal level of upwelling intensity can be defined, the `optimal environmental window' (OEW), that maximizes fish yield (Fig. 5). When on the left side of the OEW (too little wind) upwelling is weak and primary production, and hence also food for fish, is restricted by insufficient nutrient supply. On the other hand too

much upwelling (right side of the OEW) leads to dispersal of organisms away from upwelling centers and provokes light limitation in phytoplankton as they are mixed deeper into the water column because stratification is not established.

Higher Trophic Levels

The abundance of zooplankton and small pelagic fish in coastal upwelling systems provides food for a range of higher trophic levels including piscivorous fish, seabirds, pinnipeds, and cetaceans. Predatory fish such as horse mackerel and deep-water hake are themselves important fishery resources, the latter being caught by mid- and deep-water trawling. Another economically viable product of upwelling systems, particularly in the Humboldt and Benguela, is the production of bird droppings, guano, which is prized as a fertilizer because of its high N and P content. The so-called `guano birds' such as the guanay cormorant, Peruvian booby, Chilean pelican, Cape cormorant, and Cape gannet are part of resident seabird populations that breed along the coast and on adjacent islands feeding on small pelagic fish such as anchovies and sardines. Both the Benguela and Humboldt systems also support populations of small-sized penguins. The African Penguin (*Spheniscus demersus*) extends from central Namibia to Algoa Bay on the south coast of South Africa. Its population has dwindled from more than one million in 1900 to about 200 000 now, feeding mainly on a diet of pelagic schooling fish (anchovy, sardine, redeye). Reasons for the declining population are ecologically complex, but include competition with commercial fisheries for food, habitat degradation because of removal of guano from islands that they burrow into for nesting, pollution (oiling), and predation by seals. The Humboldt Penguin (*Spheniscus humboldti*) breeds mainly from 5 to 33 °S along the Peruvian and Chilean coast, with another small colony at 42 °S. Like the African Penguin, it also feeds on small pelagic fish, and its population has declined to around 30 000 for similar reasons.

Because of the episodic nature of upwelling, higher predators must endure large seasonal or interannual fluctuations in prey availability, for example, in response to El Niño events (see below). Resident seabird and pinniped populations are particularly susceptible. In catastrophic instances, starvation may cause adult seabirds to die, although more often food scarcity affects breeding success by decreasing the proportion of adults breeding and the growth rate of the hatchlings. In similar fashion, there is often a high incidence of seal pup mortality during food shortages, the adult females being unable to provide enough milk for their survival. Seabirds and pinnipeds are able to employ various strategies to compensate when food resources are in short supply including increasing the time and/or distance spent foraging for offspring, delaying reproduction until such time that food becomes available, or ceasing breeding efforts altogether. Others may target alternative food resources, such as squid, or migrate to other regions within the system where food resources are more plentiful.

Many migratory species are attracted to the high productivity of upwelling systems. Blue whales, for example, feed on dense swarms of euphausiids that exist in the California Current system. Many birds that nest elsewhere also benefit from the abundance of food in upwelling areas. The California Current system, for example, is visited by sooty shearwaters, which breed off South America, and red-necked phalaropes which nest in the Arctic. Arctic terns migrate to the Southern Hemisphere in austral winter, feeding in the Benguela and Southern Ocean upwelling systems.

The catch of pelagic anchovy and sardine by the predators described above is generally considered to surpass that by commercial purse-seine fishermen, even in the Humboldt which is heavily exploited. Seals in particular are unpopular competitors, not least because they cause damage to nets and generally interfere with fishing operations. Nevertheless, the potential consequences of overfishing in upwelling ecosystems should not be underestimated. Over the last few decades the fishing industry has progressively concentrated on species at relatively low trophic levels, with emphasis on small pelagic species such as sardine and anchovy with decreased catches of predatory fish such as hake and horse mackerel. An apparent consequence of this `fishing down marine food webs' has been a decline in pelagic fish and a proliferation of jellyfish that occupy the vacant niche, the two utilizing the same food resources. Jellyfish biomass in the northern Benguela off Namibia, for example, is now thought to exceed that of commercially

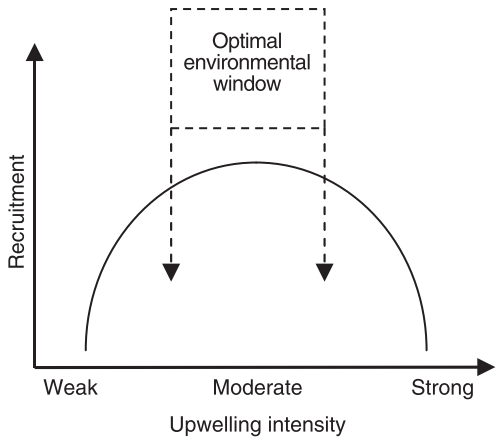


Fig. 5 Optimal environmental window for fish recruitment. Reproduced from Cury, P., Roy, C., 1989. Optimal environmental window and pelagic fish recruitment success in upwelling areas. Canadian Journal of Fisheries and Aquatic Sciences 46, 670–680.

important fish stocks. Once established, this regime shift may be difficult to reverse because jellyfish are predatory on fish eggs and juveniles.

Climatic Forcing

Changes have occurred in the dominant fish species of upwelling ecosystems, from year to year, over decades and indeed centuries. Worldwide, populations of anchovy and sardine have exhibited `flip-flops' in which one species is replaced by the other. Fishing is one factor that may influence these changes. Comparison of the fish catches of different upwelling systems, however, reveals a remarkable synchronicity in their behavior (Fig. 6) suggesting a climate linkage via global `teleconnections'. The implication is that fish populations are driven primarily by natural climate variability and its influence on ecosystem structure and recruitment success. Short-term events such as El Niño cause calamitous declines in fish stocks that lead to hardship for wildlife such as bird and seal populations, and of course fishermen. Superimposed on this short-term variability are longer-term trends that occur in response to factors such as climate change. Understanding these variations and their causes is crucial to the maintenance of sustainable fisheries in upwelling areas.

El Niño

The El Niño Southern Oscillation (ENSO) is the most important example of a relatively short-term impact of climatic forcing on upwelling ecosystems, with a typical periodicity of 3–5 years. In normal (La Niña) years, easterly trade winds blow across the surface of the equatorial Pacific from Peru/Chile to Indonesia creating the general divergent open ocean upwelling that occurs

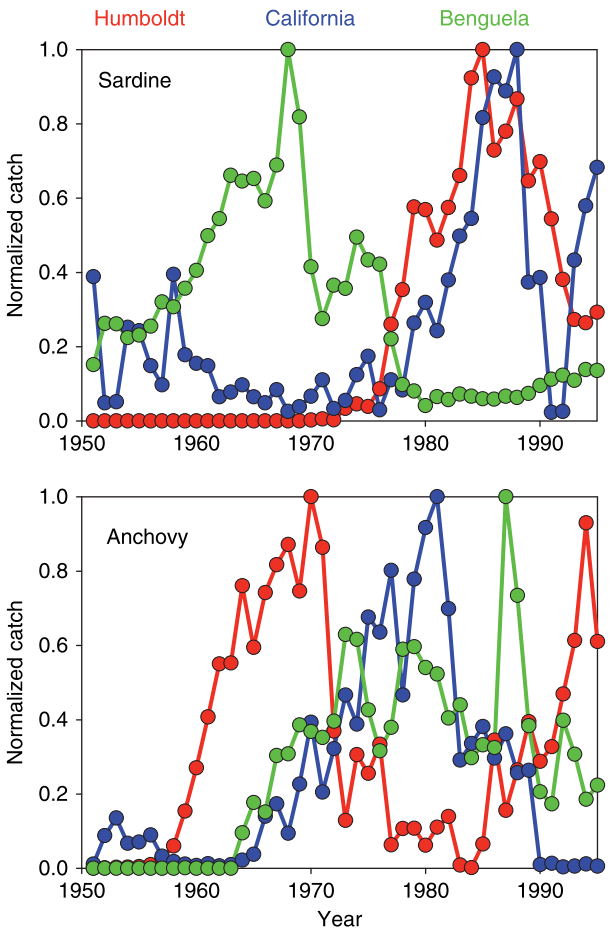


Fig. 6 Comparison of sardine and anchovy catches in the Humboldt, California, and Benguela systems. Data are normalized to maximum catch (million tons): sardine: 5.62, Humboldt; 0.29, California; 1.51, Benguela; anchovy: 12.9, Humboldt; 0.32, California; 0.97, Benguela. Reproduced from Schwartzlose, R.A., Alheit, J., Bakun, A., et al., 1999. Worldwide large-scale fluctuations of sardine and anchovy populations. South African Journal of Marine Science 21, 289–347.

across the eastern half of the equatorial Pacific. This process sets up a surface temperature gradient of $<20\ ^{\circ}C$ in the east to $>30\ ^{\circ}C$ in the west, resulting in a shallow thermocline ($\sim$20 m) in the east, but a much deeper one ($\sim$80 m) in the west. At its eastern end, along-shore winds off Peru and Chile drive the coastal upwelling of the Humboldt Current System. El Niño occurs when the easterly trade winds lose intensity, allowing warm water from Indonesia and eastern Australia to flood eastwards across the Pacific, `capping' the deeper nutrient-rich waters that lie below (Fig. 7). The coastal winds that drive upwelling in the Humboldt System possess insufficient energy to erode and mix this stratified surface layer.

Upwelling continues during El Niño but the water arriving at the surface is depleted in nutrients with drastic consequences for marine life. Diatom blooms are suppressed with a dramatic shift to a community structure dominated by the small cells of the microbial loop. Fish stocks collapse in response to the low availability of food, the starvation of adults and/or larvae leading to recruitment failure. High mortality rates are also seen in top predators. Major ENSO events occurred in 1972 and 1976, as well as in later years, economic disaster following in their wake. For example, revenue losses to Chile and Peru resulting from decimated fish stocks were about $8 billion for the 1997–98 ENSO event.

Further afield, the effects of ENSO events are felt throughout the Southern Hemisphere, and indeed the globe. The warm waters of El Niño in turn affect atmospheric circulation, the resulting teleconnections instigating changes in other upwelling systems. For example, so-called `Benguela Niño' events occur about 6 months after the onset of activities in the Pacific. During these events, the Angola-Benguela front moves southwards by several hundred kilometres, bringing low-oxygen warm water into the Namibian upwelling region that results in a southward displacement of pelagic fish stocks.

Long-Term Climate Variability

Having high fecundity, small pelagic fish are able to recover from events such as El Niño within a year or two. Yet the observed anchovy–sardine flip-flops persist over many years suggesting that climatic factors operating on longer timescales play an important role in structuring coastal upwelling ecosystems and fish stocks (Fig. 6).

Anchovy were the dominant fish species in the Humboldt Current System until the mid-1970s. Catches peaked at about 12.9 million tons in 1970, but were followed by a severe decline that may have been precipitated by the major El Niño of 1972. Recovery of the anchovy stock did not occur until the mid-1980s, sardine being dominant during the interim period. The variability of fish catches appears to have followed cycles of around 55–65 years over the last century. Analysis of atmospheric circulation patterns (e.g., the `atmospheric circulation index', ACI) reveals that the dominant direction of air masses has also changed on similar timescales. Fish scales preserved in anoxic and undisturbed shelf sediments off California and off Namibia reveal 50–70 year cycles of anchovy and sardine abundance that are linked to changes in sea surface temperature over the last 1600 years. In the Humboldt system, regime shifts appear to correlate with lasting periods of warm or cold temperature anomalies related to the approach and retreat of warm subtropical water toward the coasts of Peru and Chile. Sardine are favored during periods of warm water intrusion (1970–85) whereas the anchovy fishery prospers during periods when temperatures remain relatively cool (1950–70, 1985 to present). Resolving the underlying, probably basin-scale, physical processes that lead to such patterns, along with teleconnections linking different upwelling systems, remains a priority for scientific investigation.

Changes in local atmospheric pressure gradients with global warming might be expected to increase upwelling frequency and intensity, with accompanying changes in the structure and function of ecosystems. Nutrient concentrations have for example

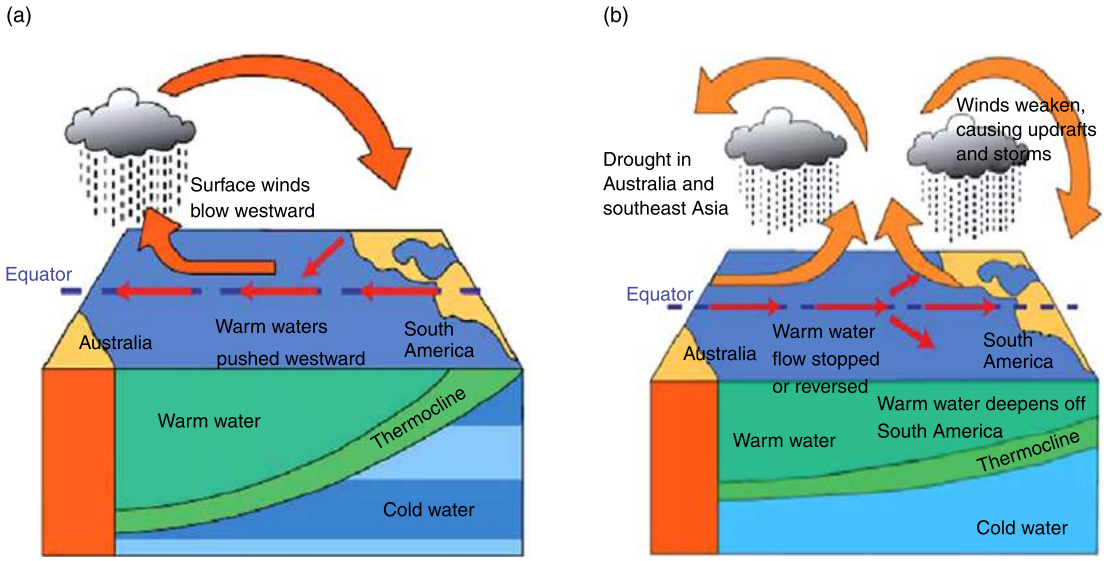


Fig. 7 Atmospheric and ocean circulation patterns associated with La Niña (a) and El Niño (b). Reproduce from SEPM Photo CD-5, *Oceanography Series* (edited by Peter A. Scholle), with permission from Society for Sedimentary Geology (SEPM).

increased in the Benguela region over recent decades, suggesting an increase in upwelling. There has at the same time been a shift from sardine prior to the mid-1960s, to anchovy in subsequent years. The recent dominance of anchovy has impacted on the zooplankton population by selective predation pressure on larger zooplankton, so that smaller cyclopoid copepods become dominant.

Variability in the trophic structure of upwelling ecosystems, be it short-term regime shifts or longer-term trends, occurs as a consequence of a range of processes operating via both environment and man's direct intervention by fishing. Understanding these interactions in order to predict the response of upwelling ecosystems to climatic forcing and fishing strategies involves unravelling a multiplicity of factors that affect primary production, zooplankton and fish recruitment, a challenging task for the scientific community.

See also: Ecological Complexity: Goal Functions and Orientors. Global Change Ecology: Biogeocoenosis as an Elementary Unit of Biogeochemical Work in the Biosphere

Further Reading

Alheit, P., Niquen, M., 2004. Regime shifts in the Humboldt Current ecosystem. Progress in Oceanography 60, 201–222.
Bakun, A., 1990. Global climate change and intensification of coastal ocean upwelling. Science 247, 198–201.
Barange, M., Harris, R. (Eds.), 2003. Marine Ecosystems and Global Change. IGBP Science no. 5., p. 32. Stockholm IGBP.
Croll, D.A., Marinovic, B., Benson, S., *et al*., 2004. From wind to whales: Trophic links in a coastal upwelling system. Marine Ecology Progress Series 289, 117–130.
Cury, P., Roy, C., 1989. Optimal environmental window and pelagic fish recruitment success in upwelling areas. Canadian Journal of Fisheries and Aquatic Sciences 46, 670–680.
Cury, P., Bakun, A., Crawford, R.J.M., *et al*., 2000. Small pelagics in upwelling systems: Patterns of interaction and structural changes in `wasp-waist' ecosystems. ICES Journal of Marine Science 57, 603–618.
Cury, P., Shannon, L., 2004. Regime shifts in upwelling ecosystems: Observed changes and possible mechanisms in the northern and southern Benguela. Progress in Oceanography 60, 223–243.
Hare, C.E., DiTullio, G.R., Trick, C.G., *et al*., 2005. Phytoplankton community structure changes following simulated upwelled iron inputs in the Peru upwelling region. Aquatic Microbial Ecology 38, 269–282.
Lehodey, P., Alheit, J., Barange, M., *et al*., 2006. Climate variability, fish and fisheries. Journal of Climate 19, 5009–5030.
Lynam, C.P., Gibbons, M.J., Axelsen, B.E., *et al*., 2006. Jellyfish overtake fish in a heavily fished ecosystem. Current Biology 16, R492–R493.
Mann, K.H., Lazier, J.R.N., 2006. Dynamics of Marine Ecosystems. Biological–Physical Interactions in the Ocean. Oxford, UK: Blackwell.
Moloney, C.L., Jarre, A., Arancibia, H., *et al*., 2005. Comparing the Benguela and Humboldt marine upwelling ecosystems with indicators derived from inter-calibrated models. ICES Journal of Marine Science 62, 493–502.
Murray, J.W., Barber, R.T., Roman, M.R., Bacon, M.P., Feely, R.A., 1994. Physical and biological controls on carbon cycling in the Equatorial Pacific. Science 266, 58–65.
Payne, A.I.L., Brink, K.H., Mann, K.H., Hilborn, R., 1992. Benguela trophic functioning. South African Journal of Marine Science 12. 1–1108.
Peterson, W., 1998. Life cycle strategies of copepods in coastal upwelling zones. Journal of Marine Science 15, 313–326.
Ryther, J.H., 1969. Photosynthesis and fish production in the sea. Science 166, 72–76.
Schwartzlose, R.A., Alheit, J., Bakun, A., *et al*., 1999. Worldwide large-scale fluctuations of sardine and anchovy populations. South African Journal of Marine Science 21, 289–347.
Summerhayes CP, Emeis K-C, Angel MV, Smith RL, and Zeitschel B (eds) Upwelling in the Ocean: *Modern Processes and Ancient Records*, 422pp. New York: Wiley.
Van der Lingen, C.D., Shannon, L.J., Cury, P., *et al*., 2006. Resource and ecosystem variability, including regime shifts, in the Benguela Current System. In: Shannon, V., Hempel, G., Malanotte-Rizzoli, P., Moloney, C., Woods, J. (Eds.), *Benguela: Predicting a Large Marine Ecosystem*, Large Marine Ecosystem Series, vol. 14. Amsterdam: Elsevier, pp. 147–184.

Adaptation☆

David J Booth and Peter Biro, University of Technology, Sydney, NSW, Australia

Introduction

A biological adaptation is a structure, physiological process, or behavioral trait of an organism that has altered incrementally, mainly through natural selection in response to increases in reproductive success in the organisms that showed those traits most strongly. Four major types of adaptations affect population dynamics, and in turn are shaped by population dynamics. Structural adaptations are special body parts of an organism that help it to survive in its natural habitat, for example, its skin color, shape, and body covering. Behavioral adaptations are ways in which a particular organism behaves to survive in its natural habitat. Physiological adaptations are systems present in an organism that allow it to perform certain biochemical reactions optimally, while life-history adaptations are parameters affecting growth and reproduction such as age at sexual maturity, reproductive investment, body size, and longevity.

Here, we will not distinguish adaptations (features produced by natural selection for their current function), from exaptations (features that perform functions, which were produced by natural selection, but not for their current use). We will emphasize the fact that adaptations are both a cause and a consequence of population dynamics. "Most behavioral and life-history theory assumes populations with stable dynamics, yet considerable empirical work shows that natural selection constantly changes population structure, resulting in disequilibrium, and hence unstable population dynamics." Adaptations can be relatively fixed (genetically based and slow to evolve), or highly plastic (changing from moment to moment, or evolving quickly), or both.

Processes of Evolutionary Adaptation

Adaptation (changes within a species over time) and speciation (in which new species arise) are two main processes that explain the observed diversity of species. For these to occur through evolution (natural selection) an organism must be viable at all stages of its development which places constraints on the rate and form of development, behavior and structure of organisms. Since adaption are to the local environment, a species may differ considerably across its distributional range, or across any climate boundaries.

Types of Adaptations

Structural Adaptations

Adaptations of feeding, antipredator, and reproductive morphology can directly affect population demography, and if they vary in space or time, can consequently drive variation in population demography. Alternative feeding morphologies in response to increased competition may increase population size but with reduced growth. Considerable variation in morphology associated with resource use is a classic example of local adaptation to the environment, and can therefore lead to variable population dynamics. Alternatively, structural adaptations may reduce variation in population dynamics. For instance, inducible defenses (spines in plankton, body shape in frogs, etc.) that evolve as adaptations to predation intensity, may stabilize dynamics by reducing mortality rates and therefore the likelihood of density dependence. Classic examples of structural plasticity reducing variation in predation rate include the evolution of head spination in cladocerans such as *Daphnia* species only in lakes which have high predator densities. Predators in turn may vary their morphology in response to prey type or abundance. Populations of freshwater roach fish such as Crucian carp (*Carassius carassius*) may exhibit local adaptation to zooplankton prey size and structure by altering the morphology of their gill rakers which are used to sieve plankton. Two species of Darwins' finches in the Galapagos have been

☆*Change History:* December 2017. D.J. Booth made minor changes to the text and references.

shown to vary in body size and bill shape over the last 30 years in response to changes in food supply. Tadpoles of some frog species develop deeper tails in ponds with predators, and there are many other examples of "inducible defenses."

Life-History Adaptations

A whole suite of different life-history characters can vary and adapt according to fluctuations in an animal's own abundance, feeding opportunities, predation, and environmental conditions. In turn, variation in life-history characters can affect spatial and temporal variation in the population abundance. Often these characters have a genetic basis and can evolve (sometimes rapidly), or they can be quite plastic within the individual, or even both. The most frequently considered and important life-history characters include growth rate, age at maturity, reproductive investment, reproductive strategy, body size, and longevity. Although they can be considered in isolation, they are often tightly correlated with one another. For example, greater growth is often associated with delayed age at maturity, greater reproductive investment, and larger body size.

Effects of changing population density on life history through competition and predation are perhaps most important and widely studied. Generally speaking, increases in population density result in decreases in age at maturity, reduced energy investment toward reproduction, smaller body size, and adoption of alternative reproductive strategies. Increased competition for food among conspecifics is a common underlying cause, though risk of cannibalism/predation can also indirectly cause such effects by confining and concentrating vulnerable individuals to refuges where feeding opportunities become limited. If so, then competition and predation are not independent factors affecting behavior and life history as so much of the literature would imply. Such "density-dependent" reductions in age at maturity and reproductive investment frequently result in smaller numbers and sizes of offspring, resulting in reduced survival and density-dependent mortality which tend to stabilize population fluctuations. Such delayed mortality is often termed "delayed density dependence" and is generally viewed as an important mechanism for stabilizing population dynamics. As suggested above, competition and predation can interact to affect changes in population dynamics through short-term effects on reproductive investment and juvenile survival in prey populations. For instance, density-dependent limitation of food resources reduces condition in young hares in Northern Canada and, in combination with predation, is thought to initiate the decline phase of their fluctuating population dynamics.

Adoption of alternative reproductive tactics is another life-history adaptation to strong reproductive competition. Examples include adoption of sneaking versus territorial/aggressive mating tactics in several fish species, for example, the bluegill sunfish (*Lepomis macrochirus*) whereby small and early maturing males become "sneakers" that dart-in and ejaculate over the eggs of a copulating pair of large individuals. This alternative "sneaking" life history (early maturing, small body size) is thought to increase fitness in fish and mammal populations where reproductive success is normally size based. Such alternative reproductive tactics serve to compensate, at least in part, for reduced reproductive success in early maturing individuals. Some of the best-known life-history responses to variation in feeding and predation conditions come from guppy populations in Trinidad, where greater predation rates result in compensating increases in reproductive investment (more eggs), earlier maturity, rapid juvenile growth, decreased coloration, and smaller population sizes than populations from streams with few or no predators. These changes in life history should serve to mitigate effects of predation mortality and prevent extinction, though its effect on population dynamics is unclear. Similar in its effect to that of natural predators, commercial fishing can alter life-history parameters. For instance, in Atlantic cod (*Gadus morhua*) and other species, "longevity overfishing" (preferential removal of the oldest, largest, and most fecund individuals) may cause a shift in phenotype to smaller body size, fewer eggs, early maturity, and consequent lower population productivity and fishery value. These changes make it difficult for the population to rebound from low numbers, and therefore increase the risk of local extinction due to the low numbers of adult fish remaining in combination with fewer young being produced which are small and less fecund than larger ones.

Variation in climatic conditions can also directly select for life-history traits that increase probability of survival in harsh environs, stabilize dynamics, and allow persistence. For instance, fish have been shown to evolve more rapid growth rates with increasing latitude and mammals evolve larger body size with increasing latitude. More rapid juvenile growth rates in fish increase the probability of surviving their first winter by accumulating body reserves and fat for the longer winters at northern latitudes. However, more rapid growth often requires greater feeding effort, increasing exposure to predators, resulting in a significant mortality cost for rapid growth rates. This may result in lower initial recruitment, but may be offset by the benefits of greater overwinter survival. Larger absolute body size in mammals at northern latitudes is thought to decrease the ratio of exposed body surface area to body mass, thus decreasing mass-specific metabolic expenditures required to keep warm during long northern winters (following Bergmann's rule). Rapid growth to accumulate resources and larger body size to minimize heat loss are both mechanisms that minimize environmentally related mortality and allow persistence and stability in these extreme environments.

Sex-ratio adjustment can also buffer a population against fluctuations. Birds and wasps in particular may use selective infanticide to maximize survival, while turtles and lizards have temperature-dependent sex determination. In the Australian water dragon lizard, for example, females are produced at hot and cool temperatures, and males at intermediate temperatures. Females take control of sex determination through changes in their nest site selection latitudinally and altitudinally and adjust sex ratios.

It should be noted that not all life-history adaptations will act to stabilize population dynamics, and there is current debate in this area. Adaptive food choice by consumers has been suggested as a major factor in population and community stability. However, a review of reproductive traits suggested that, while adaptive timing of reproduction could lead to stability, adaptations in reproductive investment and allocation of reproductive investment to offspring should destabilize population dynamics. Whether adaptive life-history strategies stabilize or destabilize population dynamics depends on the combination of optimized traits under some tradeoff.

There is surprisingly little direct evidence that intraspecific genetic variation can influence population growth and life history. However, one recent study of the Glanville fritillary butterfly (*Melitaea cinxia*) in Finland showed that variants of one gene (*Pgi*) influence population growth in a complex and habitat-dependent manner. The *Pgi* gene has several alleles: one of the homozygotes and one of the heterozygotes are common, and are linked to a higher flight metabolic rate and to be more fecund than the other heterozygotes. In small meadows, growth was highest when the two former genotypes predominated, but in larger meadows, the latter genotype was favored, possibly because butterflies with it mature later but also die later, allowing them to exploit a larger habitat more thoroughly.

Local adaptations to specific environments are well documented, and have important effects on overall population dynamics. For instance, pathogen–host populations, such as those associated with the rabies virus, illustrate the interaction between life-history dynamics, spatial spread, and evolutionary changes in infectious diseases. This is of critical importance for understanding extant epidemiological patterns and is prerequisite to constructing a predictive theory of disease emergence.

Behavioral Adaptations

In response to fluctuations in the abundance of competitors, food, and predators, individuals adapt behaviorally through changes in territoriality, reproductive behavior, foraging activity, and habitat use. By doing so, individuals can mitigate some of the negative effects of increased competition, decreases in food abundance, and predation. Density-dependent behavioral variation is most likely to directly affect population dynamics, primarily by promoting stability through reductions in survival at high density which dampen oscillations. In contrast with life-history characters, behavior can be extremely plastic and change from moment to moment with varying conditions. However, behavior also has a genetic basis allowing longer-term evolution of less plastic variation in behavior. Below we discuss some important behavioral adaptations to dynamic changes in competition and predation, emphasizing the importance of highly plastic behavior as well as less flexible behavioral adaptations.

For instance, territoriality has been shown to be a direct mechanism for density-dependent losses of young animals from natal habitats and linked to density-dependent rates of recruitment and adult population size. Given a minimum territory size, increases in conspecific density must therefore result in the aggressive exclusion of individuals into available habitats which are likely to be less favorable for growth and protection from predators. However, increases in food abundance can moderate these competitive effects. When territorial defense involves significant energy expenditure, increases in food abundance can result in decreases in territory size and therefore permit an increase in local population density. Thus, territoriality is flexible and variation in local density and food abundance interact to affect changes in territorial behavior that in turn affects local density. As individuals grow and energetic demands increase, territory size should also correspondingly increase and result in delayed density-dependent emigration of individuals and local "population regulation."

Animals that are not territorial or under circumstances which do not favor territorial monopolization of resources (e.g., unpredictable resources) frequently respond to increases in density with increased rates of foraging activity and space use. When individuals deplete food resources with increases in density, then mobile animals must increase activity rates and/or use of space to search out and find new sources of food. Increases in activity with density have been shown to compensate (at least in part) for low local food abundance to maintain growth rates (a benefit), but greater activity rates also increase encounter rates and visibility to predators (a cost). Density-dependent rates of foraging activity by prey, and corresponding activity-dependent vulnerability to predation has been demonstrated for taxa ranging from aquatic insects to large ungulates and appears to be a common mechanism for delayed density dependence and potentially stabilizing effects on population dynamics. For instance, increases in density and/or decreases in food abundance results in increases in foraging activity in tadpoles and fish and greater spacing of individuals in groups of shorebirds, resulting in greater vulnerability to predators and elevated mortality rates. Increased competition due to high density, in combination with two or more very distinct sources of food, can also combine to select for divergent foraging behavior (activity and habitat use) among individuals within a population. In this scenario, a population may diverge into distinct pools of active and sedentary individuals that specialize on different food items located in different habitats. In marine environments, most species produce propagules that disperse large distances in open ocean. For marine invertebrates and fishes, arrival at adult habitat (settlement) is characterized by larvae choosing to settle with resident conspecifics. This choice affects rate of conspecific aggression, group and individual survival, growth, and reproductive success.

Although behavior is usually thought of as being plastic and highly adaptable, recent work has begun to highlight the fact that behavioral tendencies exist (termed "behavioral syndromes"), analogous to human personality. It is thought that behavior that is not completely flexible evolves in response to unpredictable fluctuations in biotic and environmental conditions. Because fluctuations in conspecific and predator abundance are often unpredictable and change over long time intervals, it is unreasonable to expect that an animal can integrate current and future conditions to make optimal choices that maximize fitness. Rather, an inflexible component to behavior with a genetic basis can allow natural selection to optimize (over the long term) behaviors which allow persistence and fitness maximization in the face of fluctuations in competition, predation, and environmental conditions. Indeed, many animals ranging from insects, frogs, fish, mammals, and primates all display individual behavioral tendencies, such as consistent tendencies to be active and aggressive across many situations and contexts. Thus, individual behavior has a tendency or trajectory that is genetically determined, but also possess flexibility to current conditions. Maintenance of variation in behavioral tendencies then may be a mechanism for ensuring persistence over the long term, rather than short-term fitness maximization. For instance, animal populations on predator-free islands have evolved to become "tame," because antipredator behavior no longer represents a long-term persistence mechanism and reduces feeding rates in the short term. Similarly, evolution of different dispersal propensity in response to increased competition and reductions in food and/or predators can have effects on populations and their

growth (range expansion). For example, cane toads are an exotic pest species that appear to be evolving a greater propensity to be active dispersers at the front of their invasive distribution in northern Australia. Those at the "front" appear to be more active, move in a directional manner, and have longer legs than those at the point of introduction a few decades earlier.

Physiological Adaptations

Clearly, physiology of organisms limits scope for their populations to occupy habitats, grow, disperse, etc. At a gross level, physiology delimits occupation of aqueous versus aerial habitats, but more subtly, individual variation in physiological responses to the abiotic (e.g., rainfall, water quality) and biotic environment (e.g., scope to escape predators) can bear directly on population distribution, growth rate, and fecundity.

Ambient temperature is a key factor determining geographical distribution and energy budget, especially of ectothermic animals. Climate change will, therefore, cause changes in the population dynamics of organisms. In trout, for instance, adaptation leads to optimal growth when temperatures are within normal range, but can result in smaller population size when temperatures deviate substantially from the norm. This occurs primarily because metabolism increases exponentially with temperature, and because consumption is also reduced, thus reducing growth rates. More generally, when temperatures are greater than or less than optimal, individual "scope for growth" is reduced, leading to often negative effects of slower growth and smaller body size. Another example is where timing of breeding in birds does not adjust to increases in temperature due to climate warming leading to poor chick survival and reduced population size.

Energy allocation to growth and reproduction may also closely respond to local environments, with anticipated harsh conditions (e.g., freezing lakes, food shortages, long-distance migration) often preceded by buildup of storage lipids in many taxa. This may remove energy from growth and reproduction in the short term but improve lifetime success. A species that occupies a large latitudinal range, for instance, may vary its "energy budget" in response to temperatures. For example, northern Atlantic forms of silverside fishes allocate more energy to lipids than southern populations, and animals may channel energy to growth (when small and vulnerable as juveniles) and later allocate energy to fat reserves for periods of resource shortage and to reproduction.

Evolutionary Adaptation Under Human-Caused Climate Change

Evolutionary adaptation can be rapid and potentially therefore help species counter stressful conditions or take ecological opportunities arising from climate change. This means that within a community, there may be climate-change winners and losers, depending for instance on a species' adaptive capacity near its physiological limits. For example, corals exist at water temperatures often at their physiological upper thermal limits, and different rates of adaptation of co-occurring species may predict which corals will survive, if any, in the face of unprecedented global ocean warming.

Conclusion

Natural selection works through responses of individuals within populations to different environments. Therefore, population dynamics is intimately linked to natural selection, and speciation may have its foundations in within-species polymorphisms for resources. Adaptations may serve to either enhance population stability or lead to increased variation in demographic parameters, and are key to species performance across a wide range of environments, such as across latitudes. A new suite of methods, such as structurally dynamic modeling, are required which take into account the consequences of adaptations at the population level (such as plankton size shifts) to model responses of ecosystems to changed impacts. Evolutionary processes such as adaptation need to be incorporated into management programs, for instance those designed to minimize biodiversity loss under rapid climate change.

Further Reading

Billerbeck JM, Schultz ET, and Conover DO (2000) Adaptive variation in energy acquisition and allocation among latitudinal populations of the Atlantic silverside. *Oecologia* 122: 210–219.

Biro PA, Abrahams MV, Post JR, and Parkinson EA (2006) Behavioural trade-offs between growth and mortality explain evolution of submaximal growth rates. *Journal of Animal Ecology* 75: 1165–1171.

Conover DO and Schultz ET (1997) Natural selection and the evolution of growth rate in the early life history: What are the trade-offs? In: Chambers RC and Trippel EA (eds.) *Early life history and recruitment in fish populations*, pp. 305–332. London: Chapman and Hall.

Gould SJ and Lewontin RC (1979) The spandrels of san Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London. Series B, Biological Sciences* 205: 581–598.

Hoffmann AA and Sgro CM (2011) Climate change and evolutionary adaptation. *Nature* 470: 479–485.

Lima SL (1998) Stress and decision-making under the risk of predation: Recent developments from behavioral, reproductive, and ecological perspectives. *Advances in the Study of Behavior* 27: 215–290.

Logan CA, Dunne JP, Eakin CM, and Donner SD (2014) Incorporating adaptive responses into future projections of coral bleaching. *Global Change Biology* 20: 125–139.

Logan CA, Dunne JP, Eakin CM, and Donner SD (2014) Incorporating adaptive responses into future projections of coral bleaching. *Global Change Biology* 20: 125–139.

Sih A, Bell AM, Johnson JC, and Ziemba RE (2004) Behavioral syndromes: An integrative overview. *The Quarterly Review of Biology* 79: 241–277.

Skúlason S and Smith TB (1995) Resource polymorphisms in vertebrates. *Trends in Ecology and Evolution* 10: 366–370.

Allee Effects

John M Drake, University of Georgia, Athens, GA, United States
Luděk Berec, Institute of Entomology, České Budějovice, Czech Republic
Andrew M Kramer, University of South Florida, Tampa, FL, United States

Glossary

Allee effect Positive association between individual fitness and population size or density.
Allee threshold Critical population size below which average lifetime reproductive output is less than one.
Anthropogenic rarity Species rarity caused by human activity.
Component Allee effect Positive association between a vital rate and population size or density.
Demographic Allee effect Positive association between per capita population growth rate and population size or density.
Evolutionary suicide An adaptive process leading to extinction.
Inverse density dependence Positive association between per capita population growth rate and population density; also called "positive density dependence."
Strong Allee effect An Allee effect that is sufficiently large to cause an unstable equilibrium at some small, non-zero population size; this critical population size is called the "Allee threshold."
Weak Allee effect An Allee effect that is too small to cause an unstable equilibrium.

Theory of Allee Effects

Summary

Allee effects are an individual and population-level phenomenon in which the expression of reproductive output or other measures of fitness increases with density over some interval of small population sizes. Because of this positive association, Allee effects are sometimes called *inverse density dependence* or *positive density dependence*, in contrast to the negative density dependence caused by intra-specific competition. Population-level phenomena due to this density-dependent process include accelerating invasions, the appearance of critical phenomena, and evolutionary suicide.

Population Dynamics

The dynamics of populations subject to Allee effects are typically studied by translating average individual fitness into a per capita intrinsic rate of population increase. Denoting the population size or density by n, we suppose that the vital rates of the population may be represented in terms of a per capita birth rate, $b(n)$, and a per capita death rate, $d(n)$, each of which is expressed in terms of individuals per time. The balance of vital rates gives the per capita rate of increase $g(n) = b(n) - d(n)$. A *component Allee effect* may be expressed in either term, that is, a component Allee effect occurs if $b(n)$ is an increasing function of n or if $d(n)$ is a decreasing function of n. The existence of a component Allee effect does not guarantee the existence of a *demographic Allee effect*, defined to occur when the per capita growth rate $g(n)$ is an increasing function of n (over some interval), because density may affect both reproduction and mortality simultaneously. Examples are shown in **Fig. 1**. Scenario 1 shows a component Allee effect in the death rate (*green line* decreasing) resulting in a demographic Allee effect (*blue line* increasing). Similarly, Scenario 2 shows a component Allee effect in the birth rate (*black line* increasing) resulting in a demographic Allee effect (*blue line* increasing). In contrast, Scenario 3 shows a component Allee effect in the death rate (*green line* decreasing) but no demographic Allee effect (*blue line* decreasing) due to a countervailing effect in the birth rate (*black line* decreasing faster than *green line*). Likewise, Scenario 4 shows a component Allee effect in the birth rate (*black line* increasing) but no demographic Allee effect (*blue line* decreasing) due to a countervailing effect in the death rate (*green line* increasing faster than *black line*).

To obtain the rate of growth or decline for the population as a whole, the per capita population growth rate is multiplied by the population size n. It is conventional to write this growth rate as a differential equation, that is,

$$dn/dt = ng(n) = n(b(n) - d(n)).$$

For instance, Dennis (2002) considers a model where mate encounter limits reproduction at small population sizes. In this model, the birth rate is given by

$$b(n) = \lambda n/(\theta + n),$$

where λ is a maximum per capita growth rate and $n/(\theta + n)$ reflects the suppression of reproduction due to mate limitation. In this

 doi:10.1016/B978-0-12-409548-9.10587-1

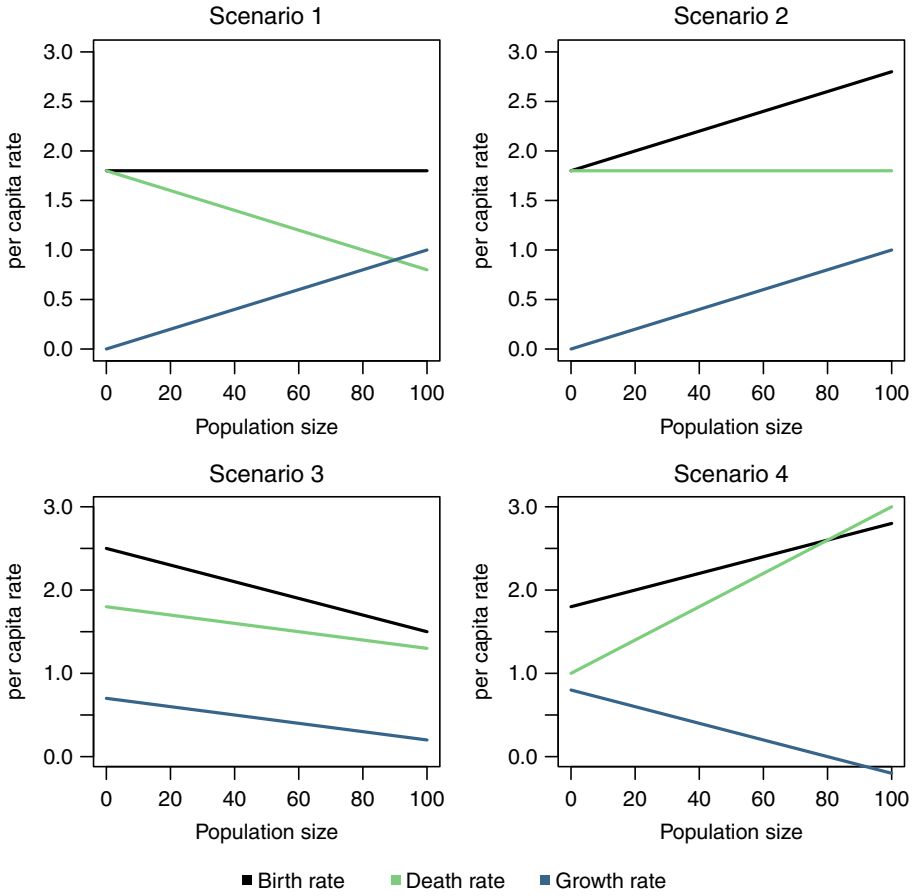


Fig. 1 Component and demographic Allee effects. Scenario 1: component Allee effect in death rate (*green line* decreasing) resulting in a demographic Allee effect (*blue line* increasing). Scenario 2: component Allee effect in birth rate (*black line* increasing) resulting in a demographic Allee effect (*blue line* increasing). Scenario 3: component Allee effect in death rate (*green line* decreasing) but no demographic Allee effect (*blue line* decreasing) due to a countervailing effect in the birth rate (*black line* decreasing faster than *green line*). Scenario 4: component Allee effect in birth rate (*black line* increasing) but no demographic Allee effect (*blue line* decreasing) due to a countervailing effect in the death rate (*green line* increasing faster than *black line*).

model, θ is a half saturation constant corresponding to the population size at which the per capita birth rate is half what it would be if matings were not depressing population growth (**Fig. 2**).

If we take the death rate to be a constant, μ, this gives rise to the mate limitation model of population dynamics

$$dn/dt = ng(n) = n(\lambda n/(\theta + n) - \mu n).$$

If $\lambda > \mu$, this model has a positive, unstable equilibrium, called the *Allee threshold*, at $n^* = \theta\mu/(\lambda - \mu)$ as shown in **Fig. 3**. A population with such an unstable equilibrium is called a *strong Allee effect*. In contrast, a model with an acceleration in population size, but where the unstable equilibrium is at extinction ($n^* = 0$) is called a *weak Allee effect*.

This model does not exhibit any negative density dependence, even at large population sizes. Dennis (2002) has shown how a logistic model for population regulation can be modified to include an Allee effect. The resulting model is

$$dn/dt = ng(n) = n(r - rn/k - \lambda\theta/(\theta + n)),$$

where r is the intrinsic rate of increase and k is the carrying capacity. This model has the shape in **Fig. 4**. A cubic model that gives a similar shape is the function.

$$dn/dt = ng(n) = n(r(n/a - 1)(1 - n/k)),$$

where the new parameter a governs the location of the Allee threshold.

This theory has been extended to a variety of more complicated models, including models of demographic stochasticity and extinction (Dennis, 2002), spatial models (Lewis and Kareiva, 1993; Kanarek and Webb, 2010), age-structured models (Cushing, 1994), discrete-time models (Veit and Lewis, 1996; Allen *et al.*, 2005), and models structured by two sexes that represent mate limitation directly (Engen *et al.*, 2003). These models make a range of theoretically testable predictions. For instance, a strong Allee effect gives rise to a sigmoidal relationship between initial population size and the chance of extinction (Dennis, 2002; Kaul *et al.*, 2016). A strong Allee effect coupled to spatial diffusion gives rise to a critical patch size that must be occupied by an incipient

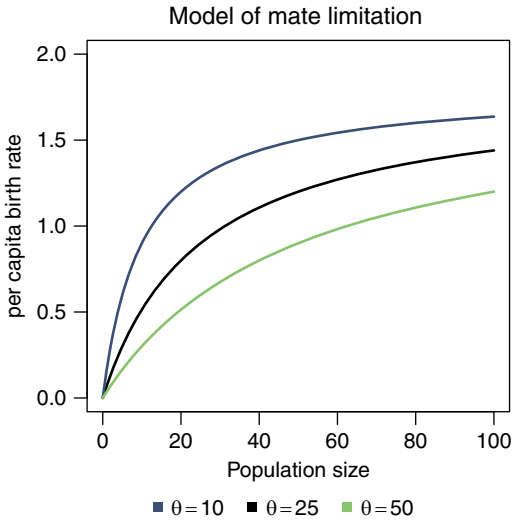


Fig. 2 Mate-limited Allee effects. The mating function $n/(\theta + n)$ causes the per capita birth rate to increase with population size until it reaches an asymptote.

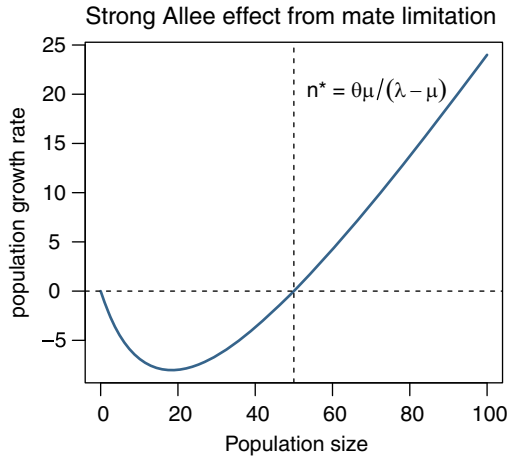


Fig. 3 Strong Allee effect. A strong Allee effect occurs when there is a non-zero population size below which the population growth rate is negative and above which it is positive. This Allee threshold, here shown at $n^* = 50$, is a critical population size below which the population cannot persist.

population to invade and an accelerating invasion wave (Lewis and Kareiva, 1993; Vercken *et al.*, 2011). And, in populations that reproduce through mating of two sexes, fluctuations in the sex ratio can give rise to an additional source of demographic noise, increasing the risk of extinction (Engen *et al.*, 2003). Thus, the theory of Allee effects gives rise to predictions that may be empirically tested in field populations. Nevertheless, at this time, the theory of Allee effects is much more advanced than the empirical study of Allee effects in nature.

Evidence for Allee Effects

Overview

Following two reviews in the late 1990s (Courchamp *et al.*, 1999; Stephens and Sutherland, 1999), however, there was a rapid increase in studies of Allee effects in natural populations, as well a retrospective examination of previous studies that had been overlooked (Courchamp *et al.*, 2008; Kramer *et al.*, 2009). The majority of the new work was driven by applications in species conservation, invasive species management, biological control, pest outbreak and harvest (Kramer *et al.*, 2009). By the end of the 2000s, component or demographic Allee effects had been detected in plants, terrestrial arthropods, aquatic invertebrates, fish, birds, and mammals, totaling >60 species (Kramer *et al.*, 2009). Each taxonomic group is discussed in more detail below.

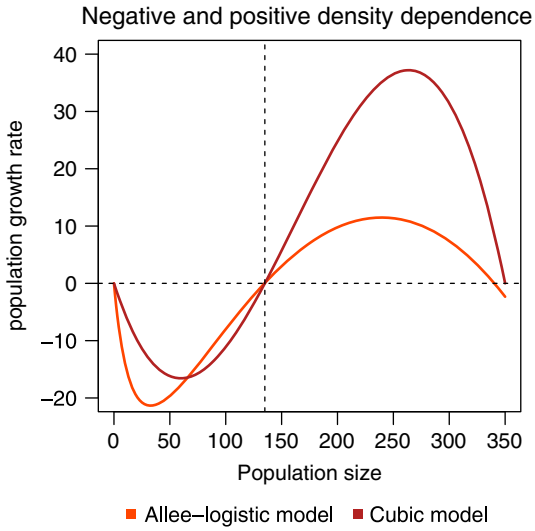


Fig. 4 Models with both positive and negative density dependence. Models with both positive density dependence at small population sizes (Allee effects) and negative density dependence (at large population sizes) are needed for understanding the long-run dynamics of populations subject to Allee effects, for instance to calculate extinction probabilities or the critical patch size of an invading population.

The diverse group of affected species and numerous mechanisms documented to cause Allee effects show these effects to be common in nature and with potential for widespread influence on population dynamics (Courchamp *et al.*, 2008). Most cited evidence was based on researchers detecting component Allee effects, primarily from mate limitation, predator satiation, and mechanisms related to social cooperation (Gascoigne and Lipcius, 2004; Gascoigne *et al.*, 2009; Kramer *et al.*, 2009). Whether these mechanisms commonly cause positive density dependence in natural populations is still actively debated. Kramer *et al.* (2009) found that only 23 studies showed any evidence for a demographic Allee effect, but also that most studies were not designed to detect demographic Allee effects. Convincingly identifying critical thresholds due to a demographic Allee effect was even rarer, occurring in only 7 of the 23 studies. Several meta-analyses have approached the question of the prevalence of demographic Allee effects by assembling data on multiple populations and applying a consistent method to detect positive density dependence in these large datasets. Gregory *et al.* (2010) fit several population models with and without Allee effects to 1198 time series of population abundance. The models with Allee effects were rarely the top-ranked model (1.1%) and had a combined support of 12% (Gregory *et al.*, 2010). This limited evidence for Allee effects was similar to low detection rates in previously published meta-analyses (Myers *et al.*, 1995; Sibly *et al.*, 2007). More recent reviews of insect (Fauvergue, 2013) and fish (Keith and Hutchings, 2012) populations also found only limited evidence of demographic Allee effects.

Causal explanations for these findings include (i) that demographic Allee effects do not occur regularly in natural populations due to fitness tradeoffs, and (ii) that Allee effects operate only at densities well below those exhibited by most populations. Alternatively, these results may also be a consequence of the difficulty of detecting positive density dependence in the presence of short time series, climatic variability and measurement error (Gregory *et al.*, 2010), that is that the failure to detect Allee effects is a measurement problem. Given the inherent difficulties in detecting and censusing sparse populations, it is not surprising that most meta-analyses have used purpose-built or specially adapted statistical methods to detect Allee effects with the goal of achieving higher statistical power. These include least squares model fits to population models (Gregory *et al.*, 2010) and Bayesian analysis of variance applied to the standardized number of recruits (Keith and Hutchings, 2012). The most heavily analyzed datasets have been marine fisheries datasets (reviewed in Hutchings, 2015), which have reported divergent findings depending on method used. Confirming the presence of a critical threshold for positive population growth is even more difficult, but has been accomplished for individual species or populations by quantifying the relationship between population growth rate or per capita growth rate and population density (Kramer *et al.*, 2009). An alternative method for detecting demographic Allee effects depends instead on comparing probability of establishment and extinction to population density (Hopper and Roush, 1993). Rigorous application of this approach relies on the theoretically conjectured differences in pattern between extinction risk due to demographic stochasticity and risk due to positive density dependence (Dennis, 2002). This method can be particularly powerful for experimental tests of Allee effects using replicate populations (Kaul *et al.*, 2016).

Evidence in Vertebrates

Component Allee effects have been detected in mammals, birds, and fish (Courchamp *et al.*, 2008; Kramer *et al.*, 2009). Examples in amphibians or reptiles are presently undocumented. Mechanisms confirmed to induce Allee effects include mate limitation, cooperative defense, cooperative breeding, cooperative feeding, predator satiation and dispersal (Courchamp *et al.*, 2008, Kramer *et al.*, 2009). Cooperative and predator-dependent methods are the most common, whereas mate limitation has been studied (and

detected) less often for vertebrates, perhaps not surprising since most vertebrates are highly mobile and often exhibit advanced social organization and signal-based communications among individuals. Multiple studies on vertebrates have provided clear evidence for the importance of Allee effects to managed populations and those of conservation concern (e.g., Angulo *et al.*, 2007), including the detection of critical thresholds for population persistence (e.g., Wittmer *et al.*, 2010).

The influence of Allee effects on the dynamics of fisheries has been of particular interest, resulting in competing conclusions for and against the relative importance of positive density dependence compared with compensatory dynamics (reviewed in Keith and Hutchings, 2012, Hutchings, 2015). However, because Allee effects appear to be a possible explanation for multiple cases where overfished populations have failed to recover following the removal of fishing pressure, continued work on fisheries data has led to innovative methods for detecting Allee effects. Outcomes of this research include recent attempts to estimate approximate rules for what level of population depletion leads to impaired recovery, with estimates of 10% of maximum population size for marine fishes, 15% of maximum population size for terrestrial mammals and 19% of maximum population size for marine mammals (Hutchings, 2015).

Evidence in Invertebrates

Studies detecting Allee effects in invertebrates are similar in number to those for vertebrates (Kramer *et al.*, 2009). These detections include terrestrial arthropods, aquatic arthropods, and aquatic molluscs subject to mate limitation, dispersal, cooperative feeding, cooperative defense, and predator satiation (Courchamp *et al.*, 2008; Kramer *et al.*, 2009). Mate limitation is the dominant causal mechanism in terrestrial and aquatic invertebrates. While this may be due partly to study bias, it strongly suggests that invertebrates experience a relatively greater challenge to mate-finding than vertebrates, and that variation in traits facilitating mate encounter is likely to be an important aspect of extinction and invasion risk in these species. One well-studied example of Allee effects in wild populations is the invasion of Gypsy moth (*Lymantria dispar*) in North America. The speed of gypsy moth spread and population dynamics where programs to monitor and slow that spread have been deployed provide evidence that (i) Allee effects may be expressed in landscape-level features of population biology (Vercken *et al.*, 2011), (ii) that there exist critical population thresholds for population establishment, and (iii) that these thresholds are spatially and temporally variable (Tobin *et al.*, 2007, Fauvergue, 2013).

Evidence in Plants

Allee effects have not been as widely considered in studies of plant populations, but have been shown to be important in populations of conservation concern (van Tussenbroek *et al.*, 2016) and invading species (Taylor *et al.*, 2004). Pollen limitation (a kind of mate limitation) is nearly universally implicated in these studies. The constraint of pollen limitation has often been hypothesized as the primary driver in the evolution of selfing in plants (see section on "Evolutionary Ecology of Allee effects"). Most of these studies document only the component Allee effects, but Groom (1998) found experimental support for reduced persistence in isolated patches due to pollen limitation.

Evidence in Microorganisms

There has been very little research on Allee effects in microorganisms. Two possible reasons are (i) the predominance of asexual reproduction removes the mechanism of mate limitation from consideration, and (ii) the numerically "large" population sizes of microbial populations begs questions about the relevance of theory exclusively concerned with "small" populations. Recently, Allee effects have been engineered in *Escherichia coli* (Smith *et al.*, 2014), and detected in experimental population of *Vibrio fischeri* (Kaul *et al.*, 2016). In both cases, the circumstances under which Allee effects were detected were highly contrived experimental settings. In another example, the social bacterium *Myxococcus xanthus* exhibits density-dependent spore production in laboratory populations (Kuspa *et al.*, 1992), resulting in heightened extinction risk at low densities and when the population is dominated by genotypes that do not contribute to fruiting body formation (Fiegna and Velicer, 2003). Further research is required to determine when analogous conditions occur in nature. We anticipate that the number of examples where positive density dependence and component Allee effects in microorganisms are demonstrated will increase as techniques for studying these populations improve.

Evolutionary Ecology of Allee Effects

Density-Dependent Selection

The presence of an Allee effect in any population is due to the fitness consequences of sparsity. For this reason, many species are presumed to have acquired adaptations expressed at small population sizes to mitigate these fitness consequences (Courchamp *et al.*, 2008; Berec *et al.*, 2018). For instance, mate-finding adaptations include the ability to move faster or more efficiently when rare, the evolution of signaling systems such as sex or aggregation pheromones, or shifts in mating system such as the facultative appearance of hermaphrodites or clonally reproducing individuals in sparse populations (Gascoigne *et al.*, 2009). Even once acquired, such adaptations may not cause the complete disappearance of demographic Allee effects, however. For instance, gypsy

moth *Lymantria dispar* (Tobin *et al.*, 2011) and the large copepod *Hesperodiaptomus shoshone* (Kramer *et al.*, 2008) both exhibit demographic Allee effects due to reduced mate encounter when rare, despite the evolution of sophisticated communication systems for mate-finding involving female-produced chemical signals that increase the encounter rate with males. Determining the etiology of traits that affect the expression of Allee effects is complicated, however, by the existence of unknown historic environments in which evolution occurred. For instance, the pheromone systems used by these species for signaling may be an evolutionary adaptation for alleviating the mate-finding Allee effect. Alternatively, it could be an adaptation acquired to signal mate quality at high densities but co-opted to facilitate mate encounter at low densities. In general, a challenge for evolutionary biology is to understand which adaptations have evolved to counteract the effects of low density and which to provide an advantage when many competitors are available.

It is further possible that dense populations of individuals may acquire traits that cease to be advantageous if those populations are reduced in size. This kind of density-dependent selection has been observed in the gamete morphology and reproductive performance of three congeneric sea urchins (Levitan, 2002). The species that commonly lives at the lowest density, *Strongylocentrotus droebachiensis*, has evolved relatively larger eggs and slow, long-lived sperm, gametes that perform best under sperm limitation. In contrast, the urchin *S. purpuratus* that lives at the highest density evolved smaller eggs and fast, short-lived sperm, gametes that perform best under sperm competition. The urchin *S. franciscanus*, which is typically found at intermediate densities, has gametes with intermediate traits. More generally, Berec *et al.* (2018) found that density-dependent selection in mate-finding traits resulted in higher Allee thresholds in populations kept at higher densities under a variety of realistic scenarios in a simulation model. In interpreting these results, it is important to keep in mind that most documented Allee effects in nature occur in species that are *anthropogenically rare* because they have been reduced in abundance by recent human activity (Courchamp *et al.*, 2008; Gascoigne *et al.*, 2009; Kramer *et al.*, 2009). Density-dependent selection thus offers an explanation for the relatively high frequency of demographic Allee effects in anthropogenically rare species.

Trait Evolution

Given the difficulty of experimentally studying demographic Allee effects and the time scales over which major innovations in mating system, gamete structure, or social behaviors are expected to evolve, it is not surprising that there is a great deal more theory than empirical evidence about the relation of Allee effects to biological evolution. The theory can be roughly divided into two nonexclusive categories: (i) theories about how Allee effects respond to the evolution of traits that shape them, and (ii) theories about how various traits evolve in populations not exposed versus exposed to Allee effects. Models of both kinds provide important insights into how populations may persist at low densities, and whether they may escape from chronic rarity.

One leading line of work suggests that when there is sufficiently large phenotypic variation in a trait that affects the strength of a demographic Allee effect, populations below the Allee threshold may escape extinction through an *evolutionary rescue*, a process of frequency dependent selection in which traits are acquired that prevent a declining species from going extinct (Gomulkiewicz and Holt, 1995). Under such circumstances, even though the population declines initially, genetic adaptation causes the Allee threshold to decline at a faster rate, eventually falling below the realized density in the population, in turn preventing population extinction (Kanarek and Webb, 2010; Cushing and Hudson, 2012; Cushing, 2015). This process may increase the chance that a population subject to an Allee effect invades a novel habitat, but its real effect is likely to be dominated by additional immigrations that sustain the initially small invading population (Kanarek *et al.*, 2015).

Another line of theory suggests that evolution also affects the strength of Allee effects in populations expressing strong Allee effects (i.e., populations on the path to extinction). Since Allee effects are intimately related to individual fitness, any mutation in a trait shaping a component Allee effect may be affected. Outcomes of such evolution may vary greatly. For example, rare plants capable of selfing may face two opposing selection pressures: inbreeding (when selfing dominates) and pollen limitation (when selfing is rare). When inbreeding depression is weak, small or low-density populations of outcrossing plants may avoid a component Allee effect due to pollen limitation by evolving selfing, while evolution may lead plants to become completely non-selfing under strong inbreeding depression despite the requirement for pollen-ovule encounter to achieve fertilization (Cheptou, 2004). Inbreeding depression is a fitness cost of selfing and may itself induce a component Allee effect if selection is weak and deleterious mutations are not purged sufficiently fast (Courchamp *et al.*, 2008; Luque *et al.*, 2016). Such antagonistic Allee effects can also arise in animals, for instance when a mate-finding Allee effect is combined with a predation-driven Allee effect, or if Allee effects due to foraging facilitation and anti-predator defense behavior affect viability of a species (Berec *et al.*, 2007). One or the other Allee effect then dominates, depending on ecological and evolutionary context.

Selection for selfing may also lead to *evolutionary suicide* (Cheptou, 2004), however, when an individual fitness increase due to trait change is accompanied by a decrease in the population size or density such that at some point the population collapses and goes extinct (Webb, 2003; Rankin and López-Sepulcre, 2005). Allee effects are a common ingredient of evolutionary suicide: the population goes extinct after it falls below the Allee threshold (which may itself increase in size as a result of the trait evolution; Parvinen, 2007; Courchamp *et al.*, 2008). The social bacterium *Myxococcus xanthus* appears to represent one of a few observed cases of evolutionary suicide, in this case triggered by a strong demographic Allee effect (Fiegna and Velicer, 2003).

A relatively new line of theoretical inquiry investigates traits of mate search rate. When search rate evolves and shapes a mate-finding Allee effects, mating systems, fitness trade-offs and the searching sex are all predicted to have a crucial role in determining

the Allee threshold (Berec *et al.*, 2018). Specifically, costs imposed by mate search may be crucial to determining how density influences the evolution of Allee effects. In a simulation study, Berec *et al.* (2018) found that when viability declines in fast searchers, the attained trait value commonly minimizes the Allee threshold irrespective of population density. However, density-dependent selection arose when a search-fecundity trade-off was imposed, leading to mate search rates that result in lower Allee thresholds in populations kept at lower densities. Further, the final outcome was sensitive to mating system. When there were no costs to mate search, runaway selection on mate search rate gave rise to evolutionary suicide with males as the searching sex, but not with females as the searching sex (Berec *et al.*, 2018).

A number of theoretical studies consider demographic Allee effects as a selection pressure on evolution of various traits, and explore whether populations subject to Allee effects evolve different characteristics as compared with those that are not subject to it. For instance, demographic Allee effects were shown to greatly facilitate local adaptation to sink environments (Holt *et al.*, 2004). They were also found to modify propensity of individuals to disperse, such as to reduce the distance that individuals move (Travis and Dytham, 2002) or to reduce sex-bias in dispersal (Meier *et al.*, 2011). However, evolution may also act to speed up the rate at which populations subject to a mate-finding Allee effect spread, in spite of diluting the population and making mate finding more difficult (Shaw and Kokko, 2015). This and other somewhat counter-intuitive evolutionary outcomes affected by and/or shaping Allee effects (some of which are mentioned above) underscore the need for theory to illuminate the evolutionarily counter-intuitive properties of mate-limitation and other population phenomena induced by sparsity as well as the need for empirical studies.

Conclusion

Allee effects are one of the classic phenomena of population ecology. The existence of a threshold size for the viability of populations subject to a strong Allee effect gives rise to a wide range of consequences for the establishment and extinction of biological populations, the speed of spread, and the expected persistence over long times. Despite the rich body of dynamical theory that elaborates on these predictions, the evidence for Allee effects in nature is relatively weak. Possible reasons for this include that population processes that give rise to Allee effects such as mate limitation are genuinely rare, that Allee effects are common but difficult to detect, and that populations that would otherwise be subject to Allee effects quickly evolve adaptations that prevent the expression of low fitness at small population sizes. Current research is seeking to illuminate which, if any, of these explanations is correct.

See also: Behavioral Ecology: Age Structure and Population Dynamics. Conservation Ecology: Source–Sink Landscape. Ecological Complexity: Population Dynamics: Stability. Evolutionary Ecology: Fitness; Association; Adaptation. General Ecology: Abundance; Demography; Cooperation; Growth Models

References

Allen, L.J.S., Fagan, J.F., Högnäs, G., Fagerholm, H., 2005. Population extinction in discrete-time stochastic population models with an Allee effect. Journal of Difference Equations and Applications 11, 273–293.
Angulo, E., Roemer, G.W., Berec, L., Gascoigne, J., Courchamp, F., 2007. Double Allee effects and extinction in the island fox. Conservation Biology 21, 1082–1091.
Berec, L., Kramer, A.M., Bernhauerova, V., Drake, J.M., 2018. Density-dependent selection on mate search and evolution of Allee effects. Journal of Animal Ecology. 87 (1), 24–35.
Berec, L., Angulo, E., Courchamp, F., 2007. Multiple Allee effects and population management. Trends in Ecology and Evolution 22, 185–191.
Cheptou, P.-O., 2004. Allee effect and self-fertilization in hermaphrodites: Reproductive assurance in demographically stable populations. Evolution 58, 2613–2621.
Courchamp, F., Berec, L., Gascoigne, J., 2008. Allee effects in ecology and conservation. Oxford University Press.
Courchamp, F., Clutton-Brock, T., Grenfell, B., 1999. Inverse density dependence and the Allee effect. Trends in Ecology and Evolution 14, 405–410.
Cushing, J.M., 1994. Oscillations in age-structured population models with an Allee effect. Journal of Computational and Applied Mathematics 52, 71–80.
Cushing, J.M., 2015. The evolutionary dynamics of a population model with a strong Allee effect. Mathematical Biosciences and Engineering 12, 643–660.
Cushing, J.M., Hudson, J.T., 2012. Evolutionary dynamics and strong Allee effects. Journal of Biological Dynamics 6, 941–958.
Dennis, B., 2002. Allee effects in stochastic populations. Oikos 96, 389–401.
Engen, S., Lande, R., Sæther, B.-E., 2003. Demographic stochasticity and Allee effects in populations with two sexes. Ecology 84, 2378–2386.
Fauvergue, X., 2013. A review of mate-finding Allee effects in insects: From individual behavior to population management. Entomologia Experimentalis et Applicata 146, 79–92.
Fiegna, F., Velicer, G.J., 2003. Competitive fates of bacterial social parasites: Persistence and self-induced extinction of *Myxococcus xanthus* cheaters. Proceedings of the Royal Society B 270, 1527–1534.
Gascoigne, J., Berec, L., Gregory, S., Courchamp, F., 2009. Dangerously few liaisons: A review of mate-finding Allee effects. Population Ecology 51, 355–372.
Gascoigne, J.C., Lipcius, R.N., 2004. Allee effects driven by predation. Journal of Applied Ecology 41, 801–810.
Gomulkiewicz, R., Holt, R.D., 1995. When does evolution by natural selection prevent extinction? Evolution 49, 201–207.
Gregory, S.D., Bradshaw, C.J.A., Brook, B.W., Courchamp, F., 2010. Limited evidence for the demographic Allee effect from numerous species across taxa. Ecology 91, 2151–2161.
Groom, M.J., 1998. Allee effects limit population viability of an annual plant. The American Naturalist 151, 487–496.
Holt, R.D., Knight, T.M., Barfield, M., 2004. Allee effects, immigration and the evolution of species' niches. American Naturalist 163, 253–262.
Hopper, K.R., Roush, R.T., 1993. Mate finding, dispersal, number released, and the success of biological control introductions. Ecological Entomology 18, 321–331.

Hutchings, J.A., 2015. Thresholds for impaired species recovery. Proceedings of the Royal Society B 282.20150654
Kanarek, A.R., Webb, C.T., Barfield, M., Holt, R.D., 2015. Overcoming Allee effects through evolutionary, genetic, and demographic rescue. Journal of Biological Dynamics 9, 15–33.
Kanarek, A.R., Webb, C.T., 2010. Allee effects, adaptive evolution, and invasion success. Evolutionary Applications 3, 122–135.
Kaul, R.B., Kramer, A.M., Dobbs, F.C., Drake, J.M., 2016. Experimental demonstration of an Allee effect in microbial populations. Biology Letters 12.20160070
Keith, D.M., Hutchings, J.A., 2012. Population dynamics of marine fishes at low abundance. Canadian Journal of Fisheries and Aquatic Sciences 69, 1150–1163.
Kramer, A.M., Sarnelle, O., Knapp, R.A., 2008. Allee effect limits colonization success of sexually reproducing zooplankton. Ecology 89, 2760–2769.
Kramer, A.M., Dennis, B., Liebhold, A.M., Drake, J.M., 2009. The evidence for Allee effects. Population Ecology 51, 341–354.
Kuspa, A., Plamann, L., Kaiser, D., 1992. A-signalling and the cell density requirement for *Myxococcus xanthus* development. Journal of Bacteriology 174, 7360–7369.
Lewis, M.A., Kareiva, P., 1993. Allee dynamics and the spread of invading organisms. Theoretical Population Biology 43, 141–158.
Levitan, D.R., 2002. Density-dependent selection on gamete traits in three congeneric sea urchins. Ecology 83, 464–479.
Luque, G.M., Vayssade, C., Facon, B., *et al.*, 2016. The genetic Allee effect: A unified framework for the genetics and demography of small populations. Ecosphere 7.e01413
Meier, C.M., Starrfelt, J., Kokko, H., 2011. Mate limitation causes sexes to coevolve towards more similar dispersal kernels. Oikos 120, 1459–1468.
Myers, R.A., Barrowman, N.J., Hutchings, J.A., Rosenberg, A.A., 1995. Population dynamics of exploited fish stocks at low population levels. Science 269, 1106–1108.
Parvinen, K., 2007. Evolutionary suicide in a discrete-time metapopulation model. Evolutionary Ecology Research 9, 619–633.
Rankin, D.J., López-Sepulcre, A., 2005. Can adaptation lead to extinction? Oikos 111, 616–619.
Shaw, A.K., Kokko, H., 2015. Dispersal evolution in the presence of Allee effects can speed up or slow down invasions. American Naturalist 185, 631–639.
Sibly, R.M., Barker, D., Hone, J., Pagel, M., 2007. On the stability of populations of mammals, birds, fish and insects. Ecology Letters 10, 970–976.
Smith, R., Tan, C., Srimani, J.K., *et al.*, 2014. Programmed Allee effect in bacteria causes a tradeoff between population spread and survival. Proceedings of the National Academy of Sciences 111, 1969–1974.
Stephens, P.A., Sutherland, W.J., 1999. Consequences of the Allee effect for behaviour, ecology and conservation. Trends in Ecology and Evolution 14, 401–405.
Taylor, C.M., Davis, H.G., Civille, J.C., Grevstad, F.S., Hastings, A., 2004. Consequences of an Allee effect in the invasion of a pacific estuary by *Spartina alterniflora*. Ecology 85, 3254–3266.
Tobin, P.C., Whitmire, S.L., Johnson, D.M., *et al.*, 2007. Invasion speed is affected by geographical variation in the strength of Allee effects. Ecology Letters 10, 36–43.
Tobin, P.C., Berec, L., Liebhold, A.M., 2011. Exploiting Allee effects for managing biological invasions. Ecology Letters 14, 615–624.
Travis, J.M.J., Dytham, C., 2002. Dispersal evolution during invasions. Evolutionary Ecology Research 4, 1119–1129.
van Tussenbroek, B.I., Soissons, L.M., Bouma, T.J., *et al.*, 2016. Pollen limitation may be a common Allee effect in marine hydrophilous plants: Implications for decline and recovery in seagrasses. Oecologia 182, 595–609.
Veit, R.R., Lewis, M.A., 1996. Dispersal, population growth, and the Allee effect: Dynamics of the house finch invasion of eastern North America. American Naturalist 148, 255–274.
Vercken, E., Kramer, A.M., Tobin, P.C., Drake, J.M., 2011. Critical patch size generated by Allee effect in gypsy moth (*Lymantria dispar* L.). Ecology Letters 14, 179–186.
Webb, C., 2003. A complete classification of Darwinian extinction in ecological interactions. American Naturalist 161, 181–205.
Wittmer, H.U., Ahrens, R.N.M., McLellan, B.N., 2010. Viability of mountain caribou in British Columbia, Canada: Effects of habitat change and population density. Biological Conservation 143, 86–93.

Further Reading

Krkošek, M., Connors, B.M., Lewis, M.A., Poulin, R., 2012. Allee effects may slow the spread of parasites in a coastal marine ecosystem. The American Naturalist 179, 401–412.
Tobin, P.C., Onufrieva, K.S., Thorpe, K.W., 2013. The relationship between male moth density and female mating success in invading populations of *Lymantria dispar*. Entomologia Experimentalis et Applicata 146, 103–111.

Association☆

Christine Angelini, University of Florida, Gainesville, FL, United States

Glossary

Association Term used for a wide range of situations where different species occur together.
Community All types of organisms found together at a given place and time.
Facilitation Interaction among species in which one improves environmental conditions or enhances resource availability to support the other.
Mutualism Interaction among different organisms in which both partners benefit.
Parasitism An interaction among different organisms in which one partner benefits at the expense of the other.
Symbiosis An interaction between two different organisms living in close proximity to one another that typically benefits both organisms.

Introduction

The most general definition of associate is "to keep company with," and the term association has accordingly been used in ecology for a wide range of situations in which different species occur together. In the early 20th century, the term was first used to refer to a group of plant species that occurs predictably together under a given set of environmental conditions, for example, the oak-hickory forest of the southeastern United States. The plant association was the forerunner of the modern concept of the ecological community, which includes all types of organisms found together at a given place and time. Ecological associations vary in both the intimacy of interaction between the species and in the types of benefits accruing to them (Fig. 1). In general, associations between species can be understood in terms of the timescales over which species interact as well the dynamic balance between fitness costs and benefits to the parties involved. Casual and intimate associations, as well as parasitic and mutualistic associations, are points along complementary continuums. Importantly, changes in environmental conditions or evolutionary changes in the interacting species can cause a given association to shift along each continuum to a new position.

Types of Associations

Intimacy: Casual to Obligate Associations

Associations between species vary from casual and fleeting to mutually obligatory and coevolved (Fig. 1). At the casual end of this spectrum, associations may arise with little or no interaction, simply because species have similar environmental requirements and tolerances, a view championed in the 1920s by Henry Gleason in his individualistic concept of plant associations. A variety of essentially fortuitous casual associations among species can enhance the fitness of one or both parties. For example, among both terrestrial plants and seaweeds, toxic or otherwise herbivore-resistant species can provide small-scale refuges for the germination and growth of other, more palatable species that would otherwise be eaten. Although both herbivore-resistant and edible species have equal or greater fitness when living alone in appropriate environments, such associational defenses broaden the range of conditions under which the edible species can thrive, often increasing diversity in the immediate vicinity of the well-defended plant.

Among animals, casual associations form when one species enhances food availability or reduces risk of predation or parasitism to benefit the other species. For example, birds foraging in mixed-species flocks can facilitate higher food intake by individual birds as the activities of one species flush out prey or open up tough foods (nuts, vertebrate carcasses) that are then available to others, and can provide predator protection as wary, alert species provide warning calls to alert other birds of the presence of predators. A more specialized association occurs between oxpecker birds and African ungulates: by removing ticks from their hosts, oxpeckers get a steady source of food, while the host gets relief from parasitism. A similar relationship holds among cottonmouth snakes and roosting birds where the presence of snakes protects birds and their chicks from rodent nest predators, like raccoons and rats, and the birds drop scraps of fish and other food to feed the snakes. In these specialized cases, the fitness of both partners increases as a result of this casual association.

More persistent and regular interactions among species can lead to the evolution of obligate associations, often involving adaptations in one or both of the interacting species. At the far end of this gradient of intimacy is symbiosis ("living together")

☆*Change History*: January 2018. Christine Angelini updated the abstract, main body, and bibliography.

This is an update of J.E. Duffy, Association, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 269–272.

 doi:10.1016/B978-0-12-409548-9.11110-8

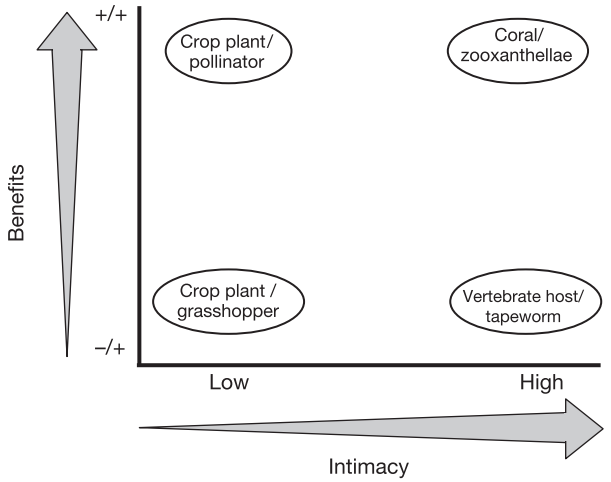


Fig. 1 Schematic illustration of the two axes of association, and the positions of some common associations in this space. The "benefits" axis ranges from parasitism (positive fitness consequences for parasite and negative for host) to mutualism (benefits to both parties). The "intimacy" axis ranges from casual and fleeting to intimate and obligatory for at least one of the associates.

between species that live in close association for most or all of their lives. The most extreme case of symbiosis is the eukaryotic cell itself, which is the product of an ancient association among formerly free-living prokaryotes that have become fully integrated physiologically and genetically and now reproduce as a single organism. Another conspicuous example of an intimate association, noted by Darwin, is the coevolution of tubular flowers (i.e., those that are shaped like trumpets and are typically nectar-rich) with hummingbirds and long-tongued insects. Similarly, reef-building corals have evolved to become dependent on their symbiotic algae; while the corals provide protection and habitat for the algae, the algae provide sugars derived from photosynthesis for their coral hosts. When bleaching events force corals to expel their algae, the corals can die of starvation if not rapidly recolonized their algae symbionts.

Benefits: Parasitism to Mutualism

A second gradient of association involves the symmetry of benefits to the interacting species (Fig. 1). Associations range from parasitism, in which one party benefits at the expense of the other through mutualism, in which both parties benefit. Between these extremes lies commensalism, in which one party benefits while the other is unaffected.

Parasites include microbes, intestinal worms, and invertebrates as well as plant- and animal-feeding insects, which are estimated to make up more than half of all animal species. The most familiar parasites associate closely and permanently with their hosts, and are strongly adapted to maintain their host-dependent life cycles. For instance, castrator barnacles are adapted to shed their hard-shells when they make contact with a crab host and squeeze through chinks in the crab's claw joints before forming tendrils, or roots, that wrap around and leach nutrients from its host's organs. Once the barnacle reaches sufficient size, it alters its host's hormone levels, rendering its host infertile while it produces an egg mast that soon releases larvae into the water column. Thus, all stages of the castrator barnacle's life cycle have evolved to enable this parasite to maintain its obligate association with its host. A similar relationship forms between mistletoe, which are a diverse group of obligate hemiparasites, and the trees they utilize as hosts where the mistletoe root into and then draw water and nutrients from their tree host. Thus, while the mistletoe fitness benefits, the host tree fitness suffers, from this parasitic association. But other parasites have a less intimate association with their hosts. Among the more bizarre such parasites are certain cichlid fish of the African Great Lakes that feed solely on the scales or eyes of other fish, which they obtain by surreptitiously attacking and sucking from the living victim.

Many associations in nature are commensal, benefiting one party with little or no impact on the other. Commensalism commonly involves a larger host species and a much smaller guest species that exploits the host's organic products or structure. Examples include many animals that live on plants or corals, as well as certain microbes that associate with the human gut or skin with no appreciable effect on the host's fitness. Epiphytes, for instance, are a highly diverse group of plants that live opportunistically on the surface of other plants, but do not extract water or nutrients from them. Spatially dominant organisms whose physical structure provides habitat and ameliorates environmental stress, such as trees, kelps, or corals, are referred to as foundation species and support many other plant and animal species that may have little to no reciprocal effect on the foundation species.

Mutualistic associations, like parasitic ones, take a variety of forms, from casual to obligate, and provide a variety of benefits including food, protection, and dispersal. The foraging mutualisms among mixed-species flocks and birds and snakes mentioned above represent the casual end, while the symbiotic associations between plants and their mycorrhizal fungi, certain insect groups and their mycetomes, corals and their algae (zooxanthellae), giant tube-worms of hydrothermal vents and their chemosynthetic bacteria, and the complex microbial ecosystems housed within the gut of ruminant ungulates fall out on the obligate end of the continuum. Among the most familiar mutualisms and those that are perhaps most important for sustaining human societies are those between flowering plants and the insects and other animals that pollinate them. Although not symbiotic due to the mobility

of many pollinators, many plant–pollinator associations are highly specialized and coevolved, and have profound effects on terrestrial ecosystem structure and functioning.

Mutualisms based on protection from enemies are also common, especially in tropical ecosystems where predation pressure is generally stronger than in temperature or artic systems. In coral reefs, for example, certain crabs and shrimp live only within the protective, stony branches of their coral hosts where they feed on the coral's secretions. In return, the crustaceans attack predatory starfish that attempt to eat their coral hosts. Thus, both mutualist partners provide predator protection to the other. A similar mutually beneficial relationship arises between ants and Acacia trees in African savannas browsed by large grazers, including giraffes and elephants. While the ants aggressively attack these grazers to protect their Acacia hosts, the hosts provide housing and carbohydrates to the ants, which create domatia within the tree. In areas where large grazers have been removed due to poaching or experimental exclusions, this mutualism can fall apart however, indicating that cooperation in such partnerships can be highly sensitive to shifts in environmental conditions.

Evolution of Associations

Selection for Association

Theory predicts that mutualistic associations should be most favored under conditions where risk of mortality from abiotic stress or predation is high, or in nutrient-poor environments where the symbiosis provides the host with nutritional benefits. Under such conditions association can provide escape from mortality or starvation that outweighs the costs imposed by competition for resources with the associate. These predictions are generally consistent with evidence from the distribution of mutualistic and commensal associations. For example, predation pressure is generally higher, and nutrient availability is generally lower, in tropical than in temperate marine waters. Accordingly, symbiosis with photosynthetic algae is much more common in tropical corals than in temperate species, and protective commensalisms between corals and sedentary invertebrates is much more common among tropical than temperate shrimp. Protective mutualisms between plants and ants, like those described among Acacia and ants in savannas, also occur primarily in the tropics. In temperate latitudes, mutualistic associations also arise, typically within more physically stressful areas. On cobble beaches characteristic of the shorelines of northeastern North America, for instance, the dominant plant *Spartina alterniflora* stabilizes and shades the cobbles, allowing numerous other plant species to colonize and thrive where they could otherwise not establish due to the heat and mobility of the cobbles.

The Coevolutionary Arms Race

When associations produce fitness benefits and/or costs to the organisms involved, their interactions will generate natural selection on one another, potentially producing evolutionary changes in one or both species. Because parasites reduce host fitness, they impose selection on the host to defend itself. In turn, parasites experience selection to overcome host defenses, which then generates selection for more effective host defenses and so on. This reciprocal selection is termed a coevolutionary arms race. Arms races also occur between predators and their prey as occurs between shell-crushing predatory crabs and snails where the crabs have evolved increasingly powerful and complex claws in response to the snails evolving increasingly thick or spiked shells.

Animals and plants have evolved a wide range of biochemical, behavioral, and life history adaptations that reduce the impacts of enemies. That such defensive features are in fact evolutionary responses to enemy pressure is supported by their geographic distribution. For example, waterfleas (*Daphnia*) from lakes with predatory fish show genetically determined predator avoidance behavior that is not present in conspecifics from fishless lakes. Many seaweeds from the tropics, where herbivore pressure is generally intense, are better defended against grazing than their relatives from more temperate areas. Conversely, selection on parasites to circumvent host defenses is illustrated by parasitic flukes that infest snails in New Zealand lakes. The flukes are better able to infect snails from their own than from other lakes, suggesting that they have adapted to the traits of the local hosts with which they have experience. Such local adaptation between hosts and parasites is common.

Specialization

Specialized associations are surprisingly common among animal species, particularly among parasites that must complete development on a single individual host. In such associations, the host constitutes the entire environment of the parasite. Detailed studies of tropical plant-feeding insect species have demonstrated that more than half are extreme specialists, feeding on only one or a few closely related species of plants. Surveys of coral-reef shrimps symbiotic with sponges yield a similar pattern. Moreover, phylogenetic analyses reveal that association with plants has strongly enhanced the evolutionary diversification of insects. Hence, the tendency of organisms to form specialized interactions with other species has been a major generator of Earth's biological diversity.

Ecology of Associations

Context Dependency

The nature of the relationship between associated species is dynamic in both ecological and evolutionary time, and subject to shift with changing conditions. The fluidity of transition between parasitism, commensalism, and mutualism at evolutionary time scales is hinted at by fungi in the family Clavicipitaceae, several of which grow within the tissues of grasses. Some species are parasites, producing diseases such as ergot in rye. Other species are commensal, with little appreciable impact on the grass host, and others still are mutualistic, producing toxic alkaloids that protect their grass hosts from grazing. Similarly, among yucca moths, most species form obligate pollination mutualisms with their yucca hosts, but several species have become parasitic; laying their eggs in yucca fruits they have not pollinated and thereby exacting a heavy cost on the host. Both the Clavicipitaceae fungi and yucca moth examples highlight that over evolutionary time scales associations can diverge in ways that can significantly affect the distribution and fitness of both associated species.

On shorter time scales and at smaller, landscape-scales, the relationship between associated species can also change in response to natural and anthropogenic shifts in environmental conditions. Global warming in particular is now recognized as an important mediator of the strength and prevalence of associations among microbial parasites and their hosts. An alarming example of this trend is the prolonged, and widespread, outbreak of a chytrid fungus, parasitic in amphibians, in Central America since the late 1980s. The epidemic, which appears responsible for the extinction of over 70 species of frogs and toads in the region, has recently been linked to improved conditions for the fungus resulting from warming temperatures. Additionally, during increasingly frequent and severe droughts, the relationship between ribbed mussels that form dense aggregations in salt marshes and cordgrass that dominants marsh landscapes shifts from a facilitative mutualism where the cordgrass benefits from, but does not depend on the mussels, to an obligate mutualism. This is because the mussel enhance water storage and reduce soil salinity stress to protect the vegetation from otherwise lethal soil conditions during the drought events.

Importantly, the nature of associations between species is also dynamically changing as a result of sifts in species' distributions due local or total species extinctions or introductions. For example, overfishing of predatory fish and crabs in New England salt marshes has opened the door for European green crabs to invade these marshes, a generalist predator which is unexpectedly benefiting marsh cordgrass by suppressing herbivore densities. Rebounding sea otter populations in coastal California is similarly reestablishing positive interactions between the sea otters and the dominant seagrass in the region as a result of the sea otters strongly suppressing previously abundant herbivore populations. Thus, associations should be considered dynamic and sensitive to changes in environmental conditions at ecological and evolutionary time scales.

Community and Ecosystem Consequences

In general, mutualistic associations expand the range of conditions beyond those in which either species could live alone, whereas parasitism tends to restrict the range of conditions under which the host can live. Consequently, associations have strongly influenced the distribution and abundance of organisms, and indeed the basic structure of ecosystems. Mutualistic associations have allowed associated species to dominate their environments, and to colonize new environments that would be unsuitable for either species alone. As a result, mutualistic and parasitic associations profoundly influence community structure, or the composition of species in an ecosystem, and the ways in which these species then control nutrient, water, and energy flows.

Perhaps most dramatically, the evolutionary diversification and dominance of terrestrial ecosystems by flowering plants has been aided by several such associations. Insects and other animals pollinate flowers much more efficiently than does the wind, and the specificity of their behavior has fostered reproductive isolation and diversification of plant species. Once pollinated, many plants produce fruits or nuts whose movement through the environment is commonly aided by a diverse collection of mammal and invertebrate dispersers that play a secondary role in mediating patterns in plant distribution. Concurrently, the roots of most terrestrial plants are intimately associated with mutualistic fungi (mycorrhizae) that can enhance their uptake of nutrients and water from the soil. One of the most ecologically important mutualisms between plants and microbes involves the fixation of organic nitrogen by bacteria (rhizobia) in the roots of legumes. By introducing new usable nitrogen into soils, this association is important in facilitating plant succession, mediating community organization, and providing a source of limiting nitrogen to grazing livestock and other animals. Reef-building corals similarly thrive in extremely nutrient-poor ocean waters, and form the foundation of the most diverse marine ecosystems on Earth, as a result of the tight recycling of nutrients by their endosymbiotic algae.

Mutualisms have also promoted dominance of certain mobile consumers. Leafcutter ants are the major herbivores of the New World tropics because their elaborate farming of fungi on harvested foliage allows them to use virtually any plant species as nourishment, in stark contrast to most herbivorous insects, which use one or a few closely related plant species as food. In lower termites, symbiotic bacteria and protists within the gut produce cellulases that allow them to feed on one of the poorest food sources in the world: wood. Similarly, the mutualistic gut flora of ruminant grazers such as wildebeest, bison, and many domestic livestock contribute to their high densities despite relatively poor nutritional quality of their forage. These densities in turn cause large grazing vertebrates to play a key role in determining the structure and functioning of their ecosystems.

Human welfare is critically dependent on associations with certain beneficial organisms, and is threatened by its association with others. Over the last few thousand years, various crop plants, livestock, and household animals have been domesticated as humans selectively bred them for characteristics we consider desirable. The food security and physical power provided by our association with these domestic organisms has been central to making humans the most abundant and powerful species in Earth's history. Remarkably, similar agricultural associations have also evolved in several groups of insects, helping them dominate their environments.

See also: Aquatic Ecology: Microbial Communities. Ecological Processes: Succession and Colonization. Evolutionary Ecology: Evolution of Parasitism. Evolutionary Ecology: Coevolution; Red Queen Dynamics; Microbiomes and Holobionts. General Ecology: Community; Pollination; Cooperation; Dominance

Further Reading

Angelini, C., Griffin, J.N., Derksen-Hooijberg, M., Lamers, L.P.M., Smolders, A., van der Heide, T., Silliman, B.R., 2016. A keystone mutualism underpins salt marsh resilience to drought. Nature Communications. 12473.
Bruno, J.F., Stachowicz, J.J., Bertness, M.D., 2003. Inclusion of facilitation into ecological theory. Trends in Ecology and Evolution 18, 119–125.
Dickman, C.R., 1992. Commensal and mutualistic interactions among terrestrial vertebrates. Trends in Ecology and Evolution 7, 194–197.
Mueller, U.G., Gerardo, N.M., Aanen, D.K., Six, D.L., Schultz, T.R., 2005. The evolution of agriculture in insects. Annual Review of Ecology Evolution and Systematics 36, 563–595.
Price, P.W., 1980. Evolutionary biology of parasites. Princeton, NJ: Princeton University Press.
Strong, D.R., Lawton, J.H., Southwood, R., 1984. Insects on plants community patterns and mechanisms. Cambridge, MA: Harvard University Press.
Thompson, J.N., 1994. The Coevolutionary process. Chicago, IL: University of Chicago Press.

Relevant Websites

Coevolution—http://evolution.berkeley.edu/evolibrary/article/evo_33.
Consequences of parasitism—https://www.nature.com/scitable/knowledge/library/ecological-consequences-of-parasitism-13255694.
Plant—microbe associations in the rhizosphere—https://www.nature.com/scitable/knowledge/library/the-rhizosphere-roots-soil-and-67500617.
Plant—pollinator mutualisms—https://www.nature.com/scitable/knowledge/library/mighty-mutualisms-the-nature-of-plant-pollinator-13235427.

Body Size, Energetics, and Evolution

FA Smith, University of New Mexico, Albuquerque, NM, USA

Introduction

Living things vary tremendously in their characteristic body size. The range of mass spans more than 21 orders of magnitude, from the smallest known organisms (mycoplasm) at $\sim 10^{-13}$ g to the largest (blue whale) at 10^{8} g (180–200 t). Even insects, a group typically thought of as relatively diminutive, have body sizes ranging over more than three orders of magnitude (**Table 1**). This includes feather-winged beetles, which readily could move through the eye of a needle, and the Goliath beetle from Africa, which is bigger than a typical dinner plate. The body mass span is extended considerably if one considers extinct insect species such as Meganeura with wingspans of ~ 70 cm. For centuries, biologists have been interested in what underlies this incredible diversity; why organisms evolve a certain size; and what the ecological and evolutionary consequences and tradeoffs are of each.

Body Size Matters

How animals interact with their environment is strongly influenced by their body size: the relative importance of forces like the surface tension of water, and the influence of gravity and temperature differ greatly depending on the absolute size of organisms. A cat cannot walk on water or up a wall, but some insects or lizards can. Likewise, both tree trunks and the limbs of terrestrial animals must be strong enough to provide support against the force of gravity but not so large that they are crushed under their own weight, or interfere with efficient functioning (transport of water and nutrients in the case of a tree, locomotion in the case of the animal). For a quadrupedal animal, ~ 140 t is the estimated maximum mass before the width of the limbs would have to be so large as to support the weight of the animal without interfering the movement; this value is close to that estimated for the sauropod *Argentinosaursus*, the largest known terrestrial animal (~ 100 t). Aquatic organisms face different constraints. Although the influence of gravity is ameliorated, allowing the evolution of much larger size (~ 200 t in the case of the blue whale), water has ~ 24 times the heat conductance of air (0.58 vs. 0.024 $\mathrm{W\,m^{-1}\,K^{-1}}$). In practical terms, this means that endotherms (animals maintaining a constant body temperature) must expend considerably more energy maintaining homeostasis. This likely limits the minimum body size; the smallest truly aquatic endotherms are ~ 100 kg, a mass that reflects selection on the ability of neonates to successfully thermoregulate. Because neonate mass is tightly correlated with maternal mass, an adult mass less than 100 kg results in offspring too small to successfully thermoregulate in water.

Biological Scaling

Not only do the structure of organisms and their dimensions change in a regular way with size, but many fundamental physiological, ecological, and evolutionary factors also scale in predictable ways. If the relationship is linear or geometric with body mass with a slope of ~ 1, the scaling is termed isometric (iso = same, metric = measure); the gut capacity of animals is an example of a trait that scales isometrically with body mass. Many relationships scale nonlinearly with body mass, with slopes less than or greater than 1; these are known as allometric relationships (allos = different, metron = measure; **Fig. 1**), a term coined by Julian Huxley and Georges Teissier in 1936. Allometric relationships were first noted in the late 1890s by Eugene Dubois and Louis Lapicque working independently on the relationship between brain and body mass, and have been extensively explored in paleontology, physiological ecology, and other disciplines.

Among traits that scale allometrically are many fundamental physiological processes such as metabolic rate, fecundity, home range, and cost of locomotion. Allometric or isometric relationships with body size are often formulated as power functions:

$$Y = aM^{b}$$

where Y is the variable of interest, M is body size, b is the slope of the relationship (representing how the variable of interest changes with differences in body size), and represents a taxon-specific constant, sometimes referred to as the normalization or proportionality constant (the intercept at unity body mass when $M = 1$). Power laws are often logarithmically transformed such that

$$\log \mathrm{Y} = \log a + b\log M$$

because the exponent becomes the slope of a straight line, facilitating computations and comparisons. The intercept often varies in a regular way among groups; marsupials, for example, have a metabolic rate 30% lower than other mammals, which is reflected in the value of their normalization constant. These body size relationships allow comparisons within and among species at different taxonomic levels and also allow reasonably accurate predictions of many biological rates and times. Often what appear to be

Table 1 Examples of the range of body size seen in various taxa

Taxa	*Smallest*		*Largest*		*Orders of magnitude*
Mammal	*Suncus etruscus* (pygmy shrew)	~1.8 g	*Balaenoptera musculus* (blue whale)	~180 t	8 (mass)
Bird	*Mellisuga helenae* (Bee hummingbird)	6.2 cm, ~1.8 g	*Struthio camelus* (North African ostrich)	~2.75 m, ~156 kg	2 (length)
Tree	*Salix herbacea* (dwarf willow)	1–6 cm tall	*Sequoiadendron giganteum* (giant sequoia)	83.8 mtall	3 (length)
Fish	*Paedocypris progenetica*	~7.9 mm long	*Rhincodon typus* (whale shark)	12.6 m long	4 (length)
Frog	*Eleutherodactylus iberia*	~9.8 mm	*Conraua goliath* (Goliath frog)	~32 cm, 3.3 kg	~1–2 (length)
Spider	*Patu marplesi* (Samoan moss spider)	0.3 mm	*Theraphosa blondi* (Goliath bird-eating spider)	28 cm170 g	4 (length)
Insect	*Nanosella fungi* (Feather-winged beetle)	0.25 mm	*Goliathus goliatus* (Goliath beetle)	>110 mm	3 (length)

For roughly cylindrical organisms, mass scales as the cube of length, so a difference of one order of magnitude in length equates to a three-order difference in mass.

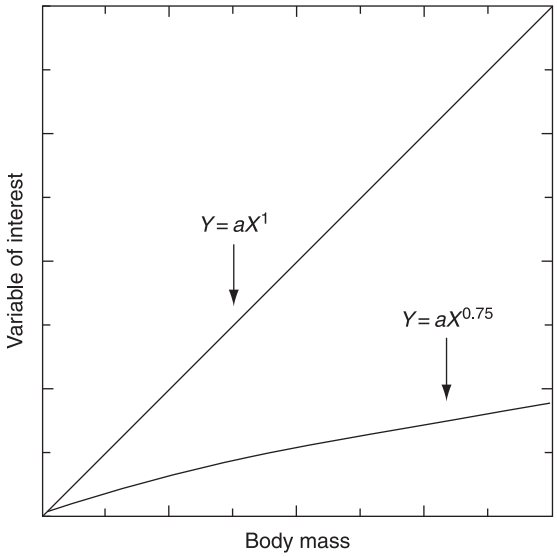


Fig. 1 Difference between isometric and allometric scaling in arithmetic space. Such relationships are typically logarithmically transformed to obtain a linear slope for ease in computation and take the form $\log Y = \log a + b \log X$.

significant differences among organisms are a simple consequence of scaling effects. True deviations from predicted values can provide important insights into evolutionary history and adaptation. Considerable research has gone into formulating and comparing allometric relationships for a whole variety of traits and taxa.

Body Size, Energetics, and Food Acquisition

All organisms require energy for the essential activities of survival, reproduction, and growth. Consequently, knowledge of energetics is central to an understanding of the selective forces that shape an organism's physiology, natural history, and evolution. The environment imposes intense selective pressures on organisms over both short and long time intervals. The occupation of novel environments, or abrupt environmental alterations, for example, can radically alter the pattern of energetic allocation between the essential activities of survival, reproduction, and growth.

The rate at which energy (E) is acquired, transformed, and used is known as the metabolic rate (MR); it drives the rate of all biological activities of and within the organism. Biologists often measure metabolic rate in calories (cal; defined as the energy required to heat 1 g water by 1 °C) or joules (1 cal = 4.1840 J). More fundamentally, the metabolic rates of organisms reflect their energetic demands or footprint on the environment. Strikingly, the metabolic rate of mammals scales consistently with body size with an exponent of 3/4 (**Fig. 2**); this is often referred to as the 'mouse to elephant curve'. The 3/4 scaling relationship between metabolic rate and body size was first proposed by Max Kleiber in 1932 and has been the object of intense debate and study. Much of the controversy centered on whether the relationship was related to surface area, which would result in an exponent of 2/3, or whether it reflected other constraints. Several comprehensive studies have firmly demonstrated a 3/4 scaling exponent and extend

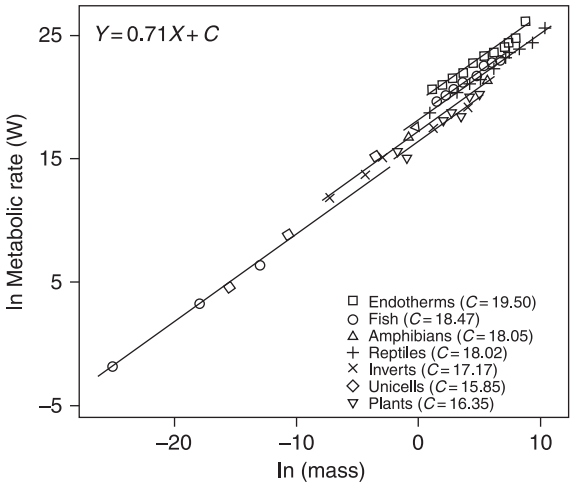


Fig. 2 The logarithmic relationship between temperature-corrected metabolic rate (in watts) and body mass (in grams) for various taxa, ranging from unicellular organisms, invertebrates, and different groups of vertebrates, to plants. The overall slope provides an estimate of the allometric exponent; the intercepts are the normalization or proportionality constants for each group (see text). Differences in the intercepts reflect taxon-specific biology. The observed slope of 0.71 ± 2 is close to the predicted value of 3/4. Drawn with permission from Brown JH, Gillooly JF, Allen AP, Savage VM, and West GB (2004) Toward a metabolic theory of ecology. *Ecology* 85: 1771–1789.

the relationship to organisms as diverse as microbes, invertebrates, to the largest mammals and trees. Over 90% of the variation in metabolic rate across species can be explained by body mass, with the residual variation reflecting unique evolutionary or biological adaptations specific to particular groups. Although still somewhat controversial, recent studies convincingly demonstrate that the underlying constraints are a result of the design properties of the vascular system. Specifically, the mechanism involves limitations on rates of uptake of resources across surfaces and rates of distribution across fractal-like branching networks within organisms.

The ecological and evolutionary consequences of allometric scaling of metabolism are profound. In practical terms it means that each gram of an animal the size of a mouse or shrew uses 20 times more energy than a gram of elephant or giraffe. Thus, food acquisition, processing, and passage rates are typically much more rapid for small animals (**Fig. 3**), and the type of digestive strategy that can be utilized is heavily influenced by body size. True herbivory – that is, the ability to obtain energy from plant structural materials as opposed to relying largely on the easily digestible cell contents – is largely controlled by residence time in the gut or fermentation chamber. If passage rates are rapid as they are for small animals, insufficient time may elapse for the microbes to ferment plant materials and consequently limited energy can be obtained from this digestive strategy. Some small herbivores have evolved specialized adaptations to get around these constraints. These include a highly convoluted cecum and microvilli (increasing surface area), or shunts for selectively retaining materials to effectively increase residence time. While such adaptations may allow more efficient use of plant structural materials, the consequence is that small herbivores are often energetically limited, which in turn influences other essential activities such as reproduction. Most small animals are much more selective and forage on higher quality resources.

The Influence of Temperature

Temperature plays a crucial role in energetics. The total energy required by an animal is not only a function of size, but is also dependent on whether it maintains a constant body temperature, that is, whether it is endo- (endo = inside) or ectothermic (ecto = outside, thermic = to heat). Endotherms maintain their body temperature within a narrow range by means of heat generated by their metabolism. Maintaining homeothermy (a constant body temperature) consumes ~90% of the energy intake of the animal, but allows activity largely independent of environmental temperatures. Only birds and mammals have adopted this evolutionary pathway – mammals typically maintain core temperatures of ~37–40 °C, while birds maintain slightly higher core temperatures of ~39–43 °C. Ectotherms, such as reptiles, fish, and other taxa, do not utilize metabolic energy to maintain a constant body temperature. Consequently, their absolute energy requirements are considerably less. However, ectotherms are generally incapable of intense activity over sustained periods of unfavorable environmental temperatures. Moreover, the metabolic rate of ectotherms is influenced by ambient environmental temperatures. Some ectotherms behaviorally thermoregulate by shifting among different microclimates to maintain a more consistent core temperature. The thermal inertia resulting from huge body masses achieved by sauropods in the Mesozoic probably meant that they were, for all practical purposes, homoeothermic. The implications of endo- versus ectothermy go beyond differing metabolic requirements. In general, ectotherms grow slower and mature at a larger body size in colder environments. Fecundity is related to adult body size, with larger individuals having larger clutches.

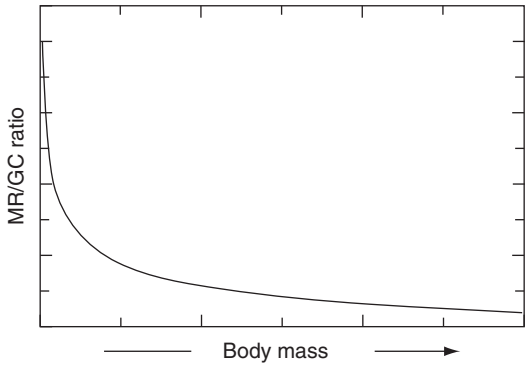


Fig. 3 While metabolic rate scales allometrically with mass to the 3/4 power, gut capacity is isometric with body mass. Thus, the ratio of metabolic rate to gut capacity scales negatively with mass to the 1/4 (e.g., $M^{0.75}/M^{1} = M^{-0.25}$). This fundamental constraint mandates higher-quality food and/or high passage rates for small animals to meet their higher per gram metabolic requirements. Food that takes longer to process (i.e., plant structural materials such as leaves) becomes progressively more difficult to handle and digest. Consequently, the smallest vertebrates are insectivores, subsisting on a ubiquitous and high-energy food source that is relatively easy to process.

It is well known that rate of biological activity rises exponentially with temperature. Mechanistically, this is because the increase in the kinetic energy of molecules results in substrates colliding with active sites more frequently. Physiologists express the relationship between metabolic rate and temperature as the Q_{10}, the rate increase for each 10^0 rise in temperature. A Q_{10} of 2, for example, means that the metabolic rate doubles for each 10^0 rise.

Recently, investigators have incorporated the important influence of temperature on metabolic rate (MR) by adding the Boltzmann factor to the allometric equation relating it to body size. In this formulation, the relationship is stated:

$$\mathrm{MR} = M^{3/4} e^{-E/kT}$$

where E represents the activation energy, k is the Boltzmann's constant, and T is the absolute temperature in degrees kelvin. Similarly, the rates of many other fundamental life-history traits demonstrate temperature dependence. Thus, the addition of the Boltzmann factor into allometric relationships allows comparisons across the entire range of living organisms, regardless of thermal regulation regime, life history, or size (**Fig. 2**). The robust relationships found suggest that the combined effects of body size, temperature, and resource supply constrain metabolic and other fundamental biological rates for all taxa.

Factors Influencing Body Size over Space and Evolutionary Time

Regular patterns of morphological variation with abiotic factors (especially temperature) have been observed repeatedly over time and space. These 'ecogeographic' gradients often involve body size, although coloration or dimensions can be influenced as well. The existence of ecogeographic patterns across space demonstrates the ability of species to adapt to fluctuating abiotic conditions, as well as underscoring the strong selection imposed on organisms by their environment.

Bergmann's Rule

Bergmann's rule is the principle that within a broadly distributed genus, species of larger size are found in colder environments, and species of smaller size are found in warmer areas. Although originally formulated in terms of species within a genus, it is often recast in terms of populations within a species. The rule (named after the German physiologist, Carl Bergmann) appears to be valid for the majority (62–83%) of vertebrates (**Fig. 4**). This includes endotherms (e.g., birds and mammals) as well as numerous species of ectotherms (bacteria, protests, plants, insects, marine organisms, and turtles). The most notable exceptions include lizards and snakes. There may be a tendency for larger-bodied animals to conform to the rule more closely than smaller bodied animals, perhaps reflecting a reduced ability to avoid stressful environments by burrowing or other means. In addition to being a general pattern across space, Bergmann's rule has been observed in populations over historical and evolutionary time when exposed to varying thermal regimes (**Fig. 5**).

Environmental temperature directly influences the energetic relationships and physiology of animals. Consequently, Bergmann's rule is often interpreted as a direct response to temperature. As organisms increase in body size, surface area increases more slowly than does volume (surface area $\propto$ length^2, vs. volume $\propto$ length^3), such that SA scales as $\sim V^{2/3}$. Because heat loss is proportional to surface area, this means that larger animals lose less heat per unit mass than smaller animals and are at an advantage under cold environmental conditions. Conversely, smaller animals have a greater surface to volume ratio and are more capable of dissipating heat under thermally stressful conditions. Numerous other mechanisms have also been postulated to

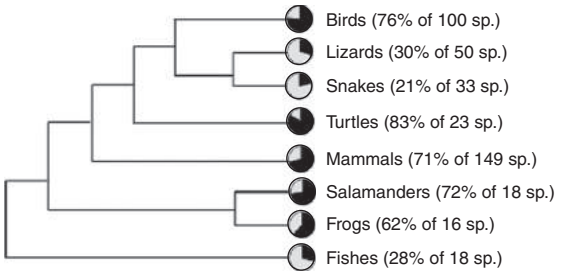


Fig. 4 Percent of various vertebrate groups that conform to Bergmann's rule. Although endothermic vertebrates demonstrate a strong body size-cline with temperature, ectotherms are much more variable in their adherence to the rule. Drawn with permission from Millien V, Lyons SK, Olson L, Smith FA, Wilson AB, and Yom-Tov Y (2006) Ecotypic variation in the context of global climate change: Revisiting the rules. *Ecology Letters* 9: 853–869.

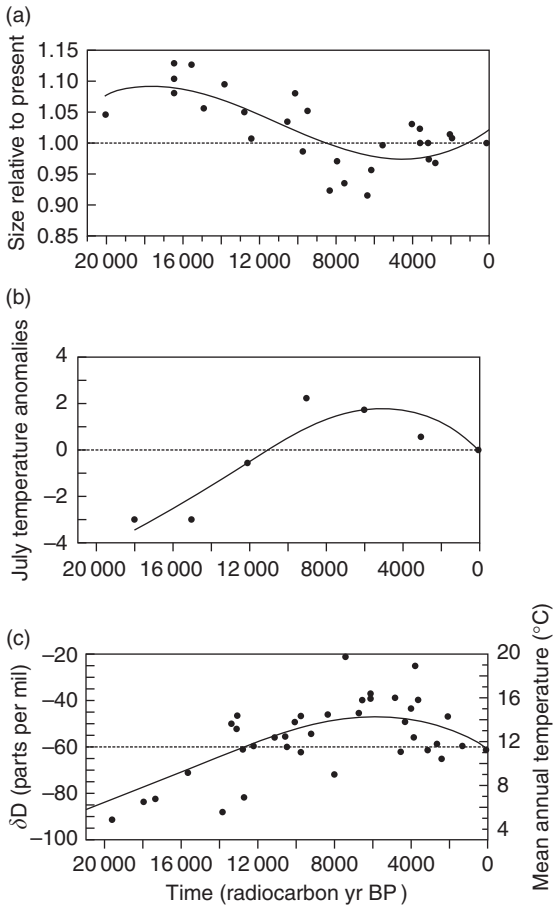


Fig. 5 Third-order regression equations fitted to data for woodrat body size and environmental temperature over the Late Quaternary illustrating adherence to Bergmann's rule over evolutionary time. (a) Mean body size of woodrats from various locations across the southwestern United States plotted as a function of radiocarbon date and expressed relative to the size of the animals at the same sites today. (b) Simulated July temperature anomalies from the National Center for Atmospheric Research Climate Circulation Model (NCAR-CCMO) expressed as deviations from modern temperature. (c) Mean annual temperature as estimated from deuterium isotope ratios measured in fossil leaves. All regression equations were constrained to yield contemporary values (as indicated by the dotted line in each panel) at 0 years BP. Drawn with permission from Smith FA, Betancourt JL, and Brown JH (1995) Evolution of body size in the woodrat over the past 25 000 years of climate change. *Science* 270: 2012–2014.

explain the size-cline, including productivity gradients, selection on life-history characteristics, development rates, and other factors related to thermal characteristics of the environment. Although the validity of the rule is generally accepted, no general consensus has yet been reached about the underlying mechanism(s) generating the gradient in size.

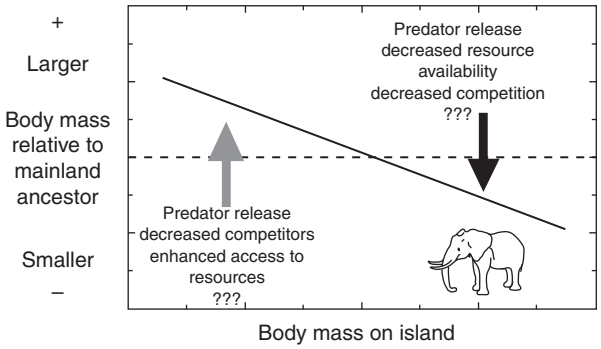


Fig. 6 Schematic of body size changes in insular habitats leading to dwarfing or gigantism. Some of the selective forces postulated to be important factors in insular habitats are indicated; these may or not apply (and depending on the taxa may select for or against larger body size). These are not meant to be inclusive list. The magnitude of evolutionary shifts is known to be dependent on the size of the island and on the degree of isolation.

Island Rule

The island rule, or Foster's rule, is the general principle that larger vertebrates isolated on islands tend to become smaller, and small ones tend to become larger. It was originally proposed by J. Bristol Foster in 1964 in an analysis of islands of the coast of western North America and Europe. Over the past few decades, considerable research has gone into establishing the fundamental patterns and proposing a variety of underlying casual mechanisms. The island rule has been amply demonstrated for many groups of mammals and reptiles, although whether mammalian carnivores conform to this pattern is still unclear. The fossil record is replete with examples of both gigantism and dwarfism on islands: during the Pleistocene the Mediterranean islands contained numerous pygmy species of elephants, rhinos, deer, as well as giant rabbits, shrews, and mice. Pygmy mammoths occupied the Channel Islands off the California coast and giant hutias ranged on islands in West Indies, with one species reaching masses of 50–200 kg. Giant birds related to the pigeon were found on Fiji and the Mascarenes, and giant tortoises ranged on numerous islands in the Indian Ocean. Many of these large animals have subsequently gone extinct, in large part due to human exploitation.

Classical explanations for the island rule focus on energetics. Island communities are typically relatively depauperate compared to mainlands. The reduced species diversity often results in few or fewer competitors or predators. With the removal of constraints imposed by the diverse mainland suite of predators and competitors, organisms evolve to a size that increases the net energy that they can obtain in the insular habitat. For large animals, where large size no longer ameliorates predation pressure, this may result in dwarfing of body size. For small animals gigantism may be favored because of increased ecological or physiological access to food resources (Fig. 6). The degree of dwarfing or gigantism may be related to the reduction in overall species diversity as well as resource availability. Note that the distinctive selective pressures on islands have also resulted in other major evolutionary transformations, such as the tendency for a loss of flight in birds, and reductions in the dispersal ability of plants.

Cope's Rule

Cope's rule is an avowed tendency for lineages of organisms to become larger over evolutionary or geologic time. Named for Edward Drinker Cope, a prominent paleontologist who first proposed it in 1896, it remains unclear whether it is a valid phenomenon or an artifact of sampling or investigator biases. Studies of several vertebrate groups such as mammals and dinosaurs appear to have generally upheld the tendency for descendent lineages to be larger than ancestral ones. Other studies, including those of marine organisms have not. Moreover, no clear underlying causal mechanism has yet been accepted to explain why an evolutionary trend toward larger body size should be favored over geologic time. Larger body size can increase the fitness of an organism if it results in a net increase in the ability to acquire resources and energy from the environment. Several traits such as foraging ability, predator avoidance, reproductive success, and thermal efficiency may be positively related to larger size. Although the fitness of the individual may increase with size, there are considerable disadvantages to large body size at the population, species, or clade level. Significant among these is that longer generation times and lowered population densities concomitant with large size might render organisms more susceptible to environmental perturbations and/or extinction events.

Patterns at Higher Taxonomic Levels

Because body size so strongly constrains the energetic demands of organisms and their interaction with the environment, it is not surprising that it also influences space and home range requirements, population densities, and other important population and community level characteristics. The statistical study of broad, consistent patterns between organisms and their environment is called macroecology, a term coined by James Brown and Brian Maurer in 1989. For example, the population density of mammals scales negatively with body mass, with about the same exponent, as does the positive scaling of metabolic rate. The coupling of these relationships results in a well-known macroecological pattern dubbed the 'energetic equivalence rule', which suggests that

the energy used by a local population of a species in a community is largely independent of its body size. Similar allometric relationships have been observed between abundance and body size for other groups including algae, terrestrial plants, and marine phytoplankton. Evidence from recent research suggests that there may well be general mechanisms that link the metabolism of individuals to higher-order emergent properties of communities and ecosystems. Thus, the study of body size is likely to continue to be a fruitful area of research in years to come.

See also: Behavioral Ecology: Thermoregulation in Animals: Some Fundamentals of Thermal Biology. Ecological Data Analysis and Modelling: Forest Models. Ecological Processes: Allometric Theory: Extrapolations From Individuals to Ecosystems. Evolutionary Ecology: Clines. General Ecology: Ecophysiology; Metabolic Theories in Ecology: The Dynamic Energy Budget Theory and the Metabolic Theory of Ecology; Ecological Efficiency

Further Reading

Alroy, J., 1998. Cope's rule and the dynamics of body mass evolution in North American fossil mammals. Science 280, 731–734.
Brown, J.H., 1995. Macroecology. Chicago: University of Chicago Press.
Brown, J.H., Gillooly, J.F., Allen, A.P., Savage, V.M., West, G.B., 2004. Toward a metabolic theory of ecology. Ecology 85, 1771–1789.
Calder, W.A., 1984. Size, Function and Life History. Mineola: Dover Publications.
Cope, E.D., 1896. The Primary Factors of Organic Evolution. Chicago: Open Court Publishing.
Damuth, J., 1981. Population density and body size in mammals. Nature 290, 699–700.
Enquist, B.J., Brown, J.H., West, G.B., 1998. Allometric scaling of plant energetics and population density. Nature 395, 163–165.
Foster, J.B., 1964. Evolution of mammals on islands. Nature 202, 234.
Haldane, J.B.S., 1927. Possible Worlds and Other Papers. London: Chatto and Windus.
Kleiber, M., 1961. The Fire of Life. New York: Wiley.
Peters, R.H., 1983. The Ecological Implications of Body Size. Cambridge: Cambridge University Press.
Millien, V., Lyons, S.K., Olson, L., Smith, F.A., Wilson, A.B., Yom-Tov, Y., 2006. Ecotypic variation in the context of global climate change: Revisiting the rules. Ecology Letters 9, 853–869.
Smith, F.A., Betancourt, J.L., Brown, J.H., 1995. Evolution of body size in the woodrat over the past 25 000 years of climate change. Science 270, 2012–2014.
Thompson D'Arcy, W., 1942. On Growth and Form. New York: Dover Publications.

Clines

EE Sotka, College of Charleston, Charleston, SC, USA

A cline is a gradient of a phenotypic or genetic character within a single species. The geographic distances across which characters shift can range from meters to thousands of kilometers. Clines are especially frequent within geographically widespread species. There is strong evidence that natural selection plays a central role in maintaining clines, in part because much of the spatial variation in a given trait reflects shifts in the biotic and abiotic environments. Clines are known as `taxonomist's nightmares and evolutionist's delights' because their evolution informs several contentious issues in ecology and evolution, including the degree and nature of natural selection, the process of dispersal and gene flow, historical demography, and speciation. Below, we briefly describe a few examples of clinal variation, outline the theoretical frameworks that underlie modern analyses of genetic clines, and describe the role of clines in understanding parapatric speciation.

Cline Examples

Though clines were formally defined as recently as 1938 by Julian Huxley, the gradual change of characters within species has been observed by naturalists for centuries. Consequently, the number of published examples of clinal variation is staggering and include clines in morphology, physiology, behavior, and genetic loci. Some morphological clines are so common as to be cited as a `rule' of nature. The oldest and most contentious of these clines is Bergmann's rule, which posits that body size increases with latitude. The pattern is widespread within and among species of mammals, birds, and some insect groups (e.g., Drosophilids; **Fig. 1**). It appears likely that selection must act on the latitudinal cline in body size because the pattern has evolved within multiple lineages of organisms and on several continents. Further, some of these clines develop over extremely short periods of time. For example, the invasive populations of the fruit fly *Drosophila suboscura* in North America has evolved a latitudinal cline similar to that seen in native Europe in less than 20 years (**Fig. 1**). As with many published clinal patterns, however, the selective mechanisms underlying Bergmann's rule are unclear. Air temperature sharply declines with latitude and represents the most obvious environmental factor operating on body size, but the exact manner in which temperature drives body size evolution remains unsolved. Other common morphological clines include Allen's rule (populations of homeotherms in colder climates have shorter appendages) and Gloger's rule (populations in more arid environments are paler in color). These and other clines are often discordant. For example, within 100 years of invading North America, house sparrows (*Passer domesticus*) evolved a north–south cline in body size (Bergmann's rule) and east–west cline in coloration (Gloger's rule).

Physiological traits also reveal clinal variation. For example, subpopulations of the vascular plant *Anthoxanthum odoratum* can grow on soil containing extremely high concentrations of zinc from mine waste. These plants are highly localized to an area less than 500 m across. When plants are tested in the laboratory, physiological tolerance for zinc of plants is positively related to the level of zinc in the soil. This pattern indicates selection for zinc-tolerance on contaminated soil and a cost to maintaining zinc-tolerance on normal soils. Other organisms reveal clinal variation in the physiological responses to temperature, salinity, and day length among other environmental factors. The best examples of clines in behavioral traits occur within hybrid zones between differentiated populations and species, where mating behavior for each of the parental types changes over space.

Clines at genetic loci (i.e., allozymes, microsatellites, protein-coding loci) are commonly under direct selection or tightly coupled to loci under selection. For example, the LdH of the killifish *Fundulus heteroclitus* shifts with latitude along the east coast of the United States. One allozyme is more efficient in warmer temperatures, while another is more efficient at cooler temperatures. This cline is also seen in humans across European and Middle Eastern populations. However, some clines may occur at loci that are neither under direct nor indirect selection. In theory, these neutral clines must be the consequence of secondary contact (i.e., historically separated populations are reconnected) that are not at equilibrium. At equilibrium, neutral clines should have collapsed because of the ongoing introgression of alleles due to gene flow.

Much of our understanding of single-species clines comes from hybrid zones between two different species because the theory of clines between and within species is broadly similar. One of the first applications of cline theory was by Szymura and Barton, who studied the hybrid zone between two *Bombina* toads in Poland (**Fig. 2**). There is striking concordance between morphological and allozyme clines across the same geographic space in this system. This pattern suggests that the species evolved in allopatry and their alleles are currently introgressing into the cline. The high level of linkage disequilibrium – the nonrandom gametic association of alleles between two or more loci – also suggests that strong selection helps to maintain these clines in the face of ongoing gene flow.

The Evolution of Clines at Equilibrium

There are at least five forces that affect the form of a cline: the genetics of the character, genetic drift, population density, the magnitude, direction and type of selection, and the magnitude and direction of gene flow. Arguably, the two most influential of

 doi:10.1016/B978-0-444-63768-0.00469-8

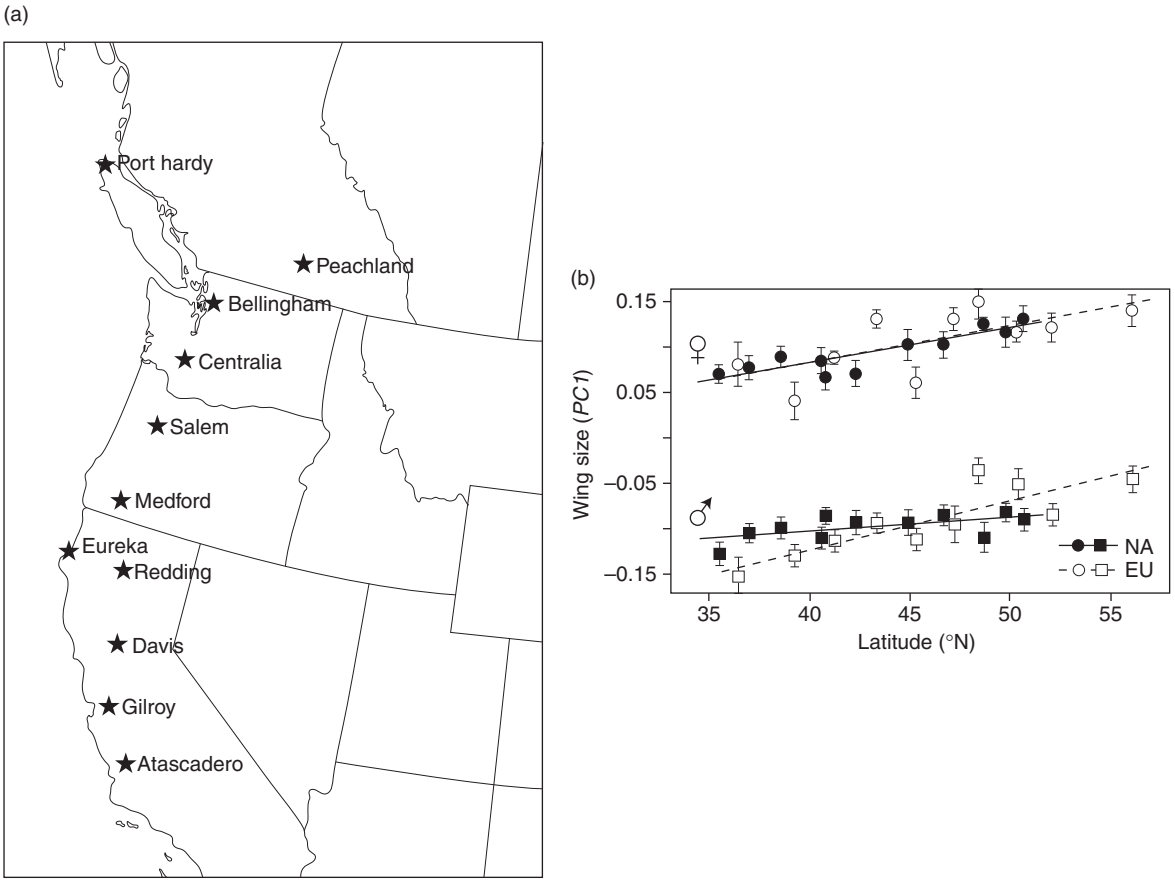


Fig. 1 *Drosophila* wing size cline in North America and Europe, as a proxy for insect body size. In less than 20 years, a wing size cline for North American females (see map) has evolved that is statistically indistinguishable from that in Europe (upper lines). Males, however, are different (lower lines). Whereas in Europe, the slope for males and females are similar, in North America, the slope of the male cline is very shallow. The error bars indicate ± 1 standard error. Figures taken from Gilchrist GW, Huey RB, and Serra L (2001) Rapid evolution of wing size clines in *Drosophila subobscura*. *Genetica* 112–113: 273–286.

these forces are selection and gene flow. At equilibrium, the width of a cline largely represents a balance between the diversifying effects of selection and the homogenizing effects of dispersal (**Fig. 3**). When selection (s) represents the difference in fitness between genotypes at the center of the cline, and σ is the standard deviation of the distance from parents to offspring along a linear gradient (which is broadly proportional to the width of the dispersal cloud around parents and is linearly related to the distance that an offspring moves on average), then the width of the cline at equilibrium is proportional to $\sigma s^{-1/2}$. Thus, narrow clines in highly dispersive organisms will be maintained only when there are high levels of selection, while narrow clines in poorly dispersed organisms can be maintained by weaker selection. The analytical power of clines is that if the magnitude of selection were known, this could be used along with the width of a cline to estimate dispersal. Conversely, if the magnitude of dispersal were estimated, this could be used to infer selection. The balance between selection and dispersal at the center of the cline is

$$w^2 = \frac{K\sigma^2}{s_e} \qquad [1]$$

where w is the cline width, s_e is the `effective' selection coefficient, and $K$ is a multiplier that depends on the type of selection. Strict application of this equation requires assumptions be met that might be rare in natural settings, including Gaussian dispersal, weak selection, and genetic equilibrium. The `effective' selection coefficient acting on the clinal locus includes direct selection on the locus that displays clinal variation in addition to the cumulative levels of indirect selection on linked loci. The multiplier K varies from about 3 in the case of exogenous selection across an ecotone to 4 in the case of heterozygote disadvantage at the center of a cline. Frequency-dependent selection against rare genotypes can increase K to 8–12, because frequency-dependent selection is effectively weaker than heterozygote disadvantage. Because cline width is proportional to the square root of K, different types of selection give cline widths that are similar to within about a factor of 2. In general, however, there are substantial deviations from the theoretical expectation when any type of selection is strong ($s > 0.2$).

Estimation of the clinal width is relatively straightforward. Traditionally, cline width has been defined as the geographic distance between populations that contain 10% or 20% and 80% or 90% of the parental gene frequencies, but in the theory of eqn

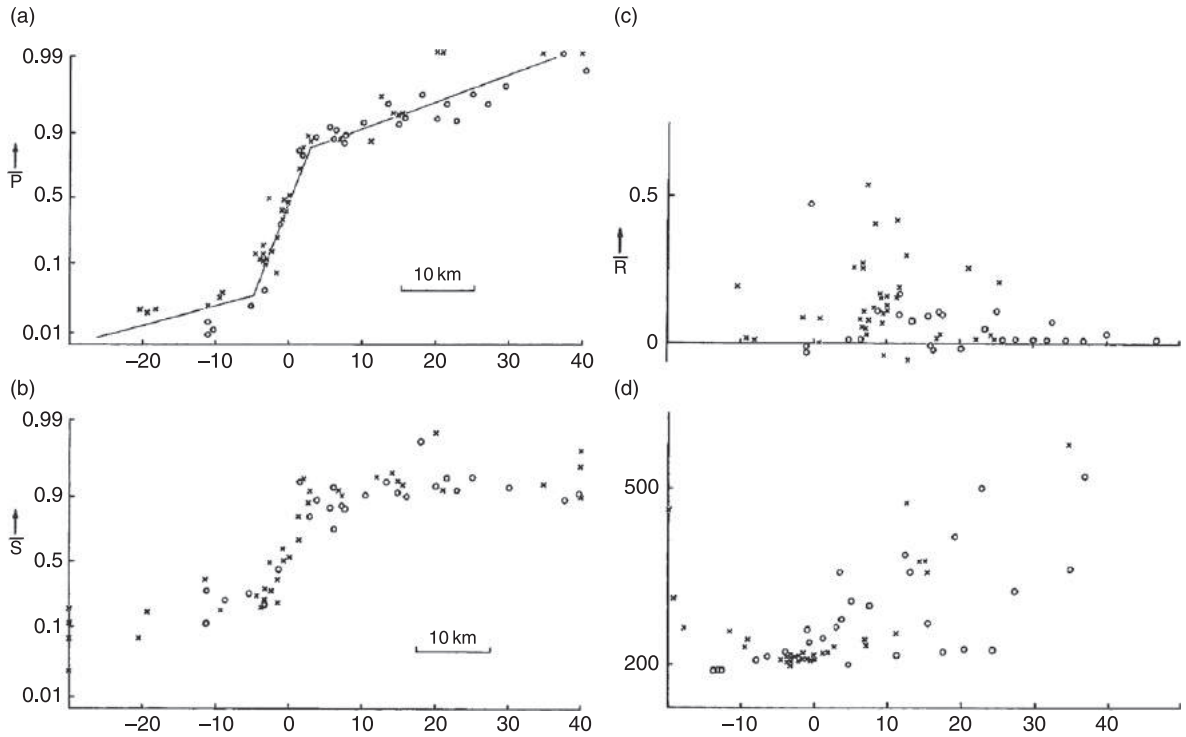


Fig. 2 The hybrid zone between two *Bombina* toad species in Poland across approximately 10 km. (a) Frequencies of *B. variegata* allozymes averaged across all loci. (b) Frequencies of seven morphological characters. (c) Standardized linkage disequilibrium, *R*, averaged across all pairs of loci. There is concordance of morphological and allozyme characters and the highest values of linkage disequilibrium within the hybrid zone. From Szymura JM and Barton NH (1991) The genetic structure of the hybrid zone between the Fire-bellied toads, *Bombina bombina* and *B. Variegata*: comparisons between transects and between loci. *Evolution* 42: 237–261.

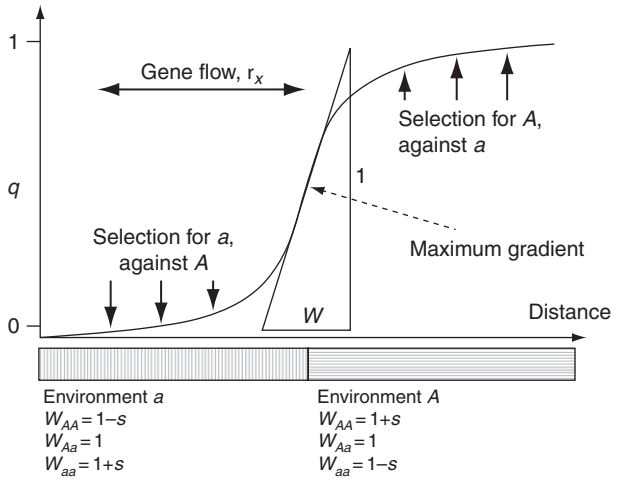


Fig. 3 Selection in continuous populations. Selection, *s*, in a continuous population may favor allele *a* on the left of the diagram, and allele *A* on the right. At equilibrium, the gene frequencies will form a sigmoid cline over the boundary between gene flow and selection (σ_x/s). The case of different environments is shown here; however, similar clines are formed in the case of intrinsic or frequency-dependent selection, for example, in contact zones between races differing in an underdominant chromosomal rearrangement, or in warning color pattern. From Mallet J (2001) Gene flow. In: Woiwood IP, Reynolds DR, and Thomas CD (eds.) *Insect Movement: Mechanisms and Consequences,* pp. 337–360. New York: CABI Publishing.

[1], cline width is the inverse of the maximum slope of the cline. An estimate of width may assume that allele frequencies vary between 0.0 and 1.0 along the cline, or alternatively, if the populations are not fixed on either side, cline width $w = \Delta p/\text{slope}$, where Δp is the change in gene frequencies among parental populations at the ends of the cline and slope is the slope at the center of the cline.

Using clines to estimate selection

The consequence of this equilibrial relationship between cline width, selection, and average dispersal distance is that evolutionary ecologists can utilize clinal theory to disentangle the relative strengths of selection and dispersal. For example, the strength of selection acting on loci within a cline of known width can be inferred using formula [1] when empirical estimates of dispersal are generated from direct censuses. However, selection estimates are probably far lower than the actual levels. This is largely because direct methods of measuring dispersal regularly underestimate the frequency of long-distance dispersal.

Selection operates either because hybrids of the parental lines are generally less fit, or alternatively, parents or hybrids may be less fit in non-native environments. Those different modes of selection have been called endogenous and exogenous selection, respectively, and they lead to similar consequences when clines occur between differentially adapted species or populations. Most commonly, clines are the result of some mix of both modes of selection. The scenario by which hybrid zones are maintained from selection favoring hybrids within a narrow zone of intermediate habitat (termed bounded hybrid superiority) is thought to be extremely rare, but when present, does not apply to the theoretical expectations.

The distinction between endogenous and exogenous selection is crucial for understanding the potential mobility of the hybrid zone. For example, if endogenous selection operates (selection against hybrids), then the transition zone tends to shift toward any region that had low population densities. Alternatively, hybrid zones maintained by exogenous selection tend to remain stationary at a particular place on an ecological gradient, or shift in geographic position when the environment changes.

Using clines to estimate dispersal

In a similar manner, evolutionary ecologists have begun to utilize clinal theory to estimate dispersal distances of organisms. This approach may become particularly useful to marine researchers, because most marine invertebrates produce hundreds to thousands of microscopic offspring per parent and as a consequence, there are few empirical estimates of larval dispersal of marine organisms to within several orders of magnitude. However, inferring mean dispersal distance requires a precise estimate of the selection coefficient (s_e), which is itself a logistically difficult task. Field-based experiments can sometimes detect rather strong selection coefficients, but weaker levels of selection are more difficult to measure. Laboratory-based experiments can be more sensitive, but in many cases, their results may not be generalizable to more natural conditions.

In response to the difficulty of estimating selection coefficients directly, researchers have instead focused on estimating linkage disequilibrium (LD) because LD correlates with the magnitude of selection. Positive values of LD along a cline generally reflect an excess of parental gametic haplotypes and a reduction of hybrid gametic haplotypes within the cline (see **Fig. 2** for example). LD is generated when either endogenous or exogenous selection acts within the clines and is weakened by recombination. The net effect is that hybrids are less readily generated or maintained, parental alleles do not readily recombine, and a greater than expected number of parental gametes or haplotypes are encountered within a hybrid zone. Consequently, the higher the rate of migration across the clines of a given width, the larger the number of parental genotypes found within the clines, and the higher the degree of LD (i.e., selection). Because LD is generated by selection after migration in each generation, it is largely equivalent to `effective' selection when $s_e < 0.10$. When greater levels of selection maintain the clines, LD is not strictly equivalent to s_e, but rather approximates selection within an order of magnitude. It is not strictly generated by epistatic interactions between loci.

For clines at equilibrium, the balance between selection and dispersal can be represented at the center of a cline by

$$w^2 \sim \frac{\sigma^2}{D_{AB} r} \qquad [2]$$

where r is the rate of recombination and D is the maximum level of LD between loci (A and B). The basic formulation of D is the two-locus deviation from random expectation, or

$$D_{AB} = p_{AB} - p_A p_B \qquad [3]$$

where p_A is the frequency of allele A at locus 1, p_B is the frequency of allele B at locus 2, and p_{AB} is the frequency of AB. If one assumes the loci are unlinked, then the rate of recombination is $r = 0.5$. Because changes in allele frequencies (p) affect the maximum potential genetic disequilibrium, eqn [2] can be replaced with one that uses the correlation coefficient between loci, R_{AB}:

$$R_{AB} = \frac{D_{AB}}{\sqrt{p_A p_B (1 - p_A)(1 - p_B)}} \qquad [4]$$

so that at the center of a perfect cline when $p_A = p_B = 0.5$,

$$w^2 \sim \frac{4\sigma^2}{R_{AB} r} \qquad [5]$$

The clinal theory outlined here is explained in much greater detail by its architects (see the section titled `Further reading' for more details), and includes a host of factors that complicate its applicability to particular data sets. For example, the equations largely assume an evolutionarily static balance between selection and dispersal. If the cline is new because a new population is invading a region after human introduction, the relationships between selection, width, and dispersal will be very different. For most clines, however, stabilization of clines occurs very rapidly. It has been estimated, for example, that if $s = 0.1$, then clines stabilize within on the order of 10 generations.

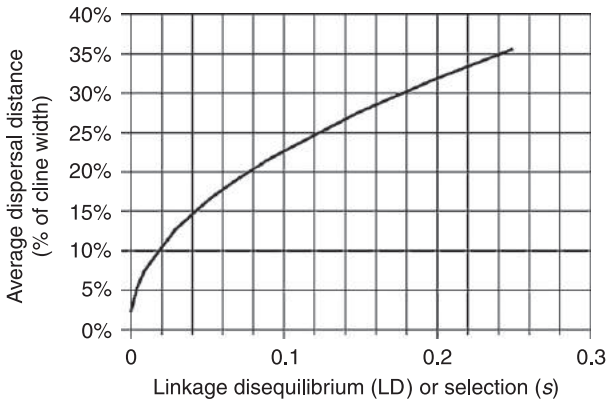


Fig. 4 The correlation between linkage disequilibrium (LD), selection (s), and dispersal distance within a stable cline. The curve is based on formulas [1] and [2]. Dispersal is given as the proportion of the clinal width (e.g., 20% of a 100 km clinal width is 20 km). Under reasonable levels of selection, the dispersal distance is a fraction of the clinal width. From Sotka EE and Palumbi SR (2006) The use of genetic clines to estimate dispersal distances of marine larvae. *Ecology* 87: 1094–1103.

A second complication is that clinal theory is directly relevant for loci under direct selection, such as some allozymes or morphological and physiological traits. However, many hybrid zones are detected using molecular markers (e.g., microsatellites) that are less likely to be under direct selection. If these neutral genes are not physically linked with loci under direct selection, then recombination quickly breaks up linkage disequilibria between loci, and the introgression of the neutral alleles across the cline will occur unimpeded. Thus, genetic differences among populations may be extremely strong immediately after secondary contact (i.e., after historically separated populations reconnect) and begin to weaken after hundreds or thousands of generations of gene flow. As a consequence, such neutral clines often look like a `staircase' of several steps of allele frequency across the hybrid zone (neutral markers are rarely fixed on both sides of the hybrid zone) and the introgression will eventually homogenize allele frequencies. The net effect of the flattening of the cline is a rather weak but artificial increase in estimated clinal width, and any given empirical estimate of selection in a snapshot of time will infer a rate of dispersal that is somewhat lower than the actual dispersal. In fact, a neutral cline will flatten at a rate proportional to a predictable product of dispersal and time; the width of a cline of neutral genes t generations after two differentiated populations come together is expected to be about $2.51\sigma\ (t)^{-1/2}$, assuming equal population sizes. On the other hand, clines at neutral genes can be stable if these genes are physically linked to genes under selection. Linkage of neutral markers to many genes under selection results in an effective selection (s_e) that helps sculpt neutral gene clines in a manner analogous to the action of non-neutral clines.

Third, these relationships are largely based on an underlying diffusion approximation and may be violated by rare long distance dispersal. Fourth, shifts in population densities, physical barriers to dispersal, and asymmetric gene flow can slightly alter the relationships between cline width, dispersal, and selection values.

The consequence of these and other complications is that estimates of dispersal based on selection or LD will be robust and of the right order, but may not be precise. Further, the dispersal estimates reflect the distances travelled within the clinal region only.

Still, in the absence of detailed information on linkage or selection, it is possible to produce a first approximation of larval dispersal distance using clinal theory. Even under high levels of selection (e.g., $s \sim 0.25$) and at equilibrium, the average geographic distance dispersed by offspring – as measured by the variance in distance between parent and offspring or neighborhood size – is less than about a third of the cline width ($s < 0.35\ w$; **Fig. 4**). Although this is a crude approximation; it suggests that in general, populations on the endpoints of clines do not typically disperse offspring across the entire cline width in one generation. Instead, typical propagules may requires 3–5 generations to traverse the cline. Only if selection were very large (e.g., hybrids were infertile) relative to rates of recombination would a cline be maintained by a dispersal distance that was as long as the cline was wide. In cases where selection was measurable but ecologically moderate ($s \sim 0.1$), then $s \sim 0.11\ w$. In other words, for selection that ranges from moderate to strong, clines are generally several times wider than average dispersal distance. For clines subject to weak selection, the subsequent clines can be an order of magnitude wider than dispersal, or more.

Clines and Parapatric Speciation

It is a general rule of thumb of evolutionary biology that it is far easier to describe the maintenance of genetic variation than to describe its generation. This is particularly true for understanding the evolution of clinal variation. On the one hand, clines may represent the consequence of a recent collision of formally separated populations that evolved in allopatry. These situations, termed secondary introgression are thought to dominate systems with steep clines (**Fig. 1b**) as when two species hybridize within a tension zone.

On the other hand, it is known that clines can be generated by spatially variable selection acting on a set of genetically identical subpopulations. One example of this comes from the *Drosophila* invader into North America (see **Fig. 1a**), in which a cline was

generated within two decades of the invasion. These situations reflect primary contact clines. The main problem for evolutionary ecologists is that it can be difficult to know whether a cline is the consequence of primary contact or secondary introgression. Recent authors indicate that cline theory can be utilized in either case, but these theoretical expectations provide little insight into the generation of the cline itself.

The mechanism of clinal generation is important to speciation, or the generation of species. Traditionally, natural historians believed that speciation across clines was commonplace, but today, there is far more skepticism largely because the evidence for such speciation is weak and elusive. The generation of `good' species within a cline is called parapatric speciation and will proceed through the following steps. First, the environment generates strong local genetic differences among subpopulations in multiple traits and is reflected in a series of a steep clines. If such differences persisted, then hybrids of subpopulations from the endpoints of the cline will become less fit, a situation called a tension zone. The exact mechanism by which hybrids are less fertile or viable can be intrinsic (e.g., genetic incompatibilities) or extrinsic (e.g., less competitive in the environment). As a consequence of the fitness cost of hybridization, there will be strong selection for prezygotic traits that minimize cross-breeding and thereby lead to the generation of `good' species, a process referred to as reinforcement.

However, to date, several of the predictions of parapatric speciation are poorly supported. There is growing evidence from molecular and other lines of evidence that most hybrid zones are not primary contact zones, but are instead the consequence of secondary introgression of formerly allopatric populations. Further, prezygotic isolation does not appear to be the primary force maintaining most hybrid zones.

See also: Aquatic Ecology: Microbial Communities. Behavioral Ecology: Biological Rhythms; Habitat Selection and Habitat Suitability Preferences. Conservation Ecology: Connectivity and Ecological Networks. Ecological Data Analysis and Modelling: Species Distribution Modeling. Evolutionary Ecology: Isolation; Body Size, Energetics, and Evolution; Natural Selection. General Ecology: Migration and Movement

Further Reading

Arnold, M., 1997. Natural Hybridization and Evolution. New York: Oxford University Press.
Barton, N.H., 1982. The structure of the hybrid zone in *Uroderma bilobatum* (Chiroptera: Phyllostomatidae). Evolution 36, 863–866.
Barton, N.H., Gale, K.S., 1993. Genetic analysis of hybrid zones. In: Harrison, R. (Ed.), Hybrid Zones and the Evolutionary Process. New York: Oxford University Press, pp. 13–45.
Barton, N.H., Hewitt, G.M., 1985. Analysis of hybrid zones. Annual Review in Ecology and Systematics 16, 113–148.
Endler, J.A., 1977. Geographic Variation, Speciation, and Clines. Princeton: Princeton University Press.
Gilchrist, G.W., Huey, R.B., Serra, L., 2001. Rapid evolution of wing size clines in *Drosophila subobscura*. Genetica 112–113, 273–286.
Hare, M.P., Guenther, C., Fagan, W.F., 2006. Nonrandom larval dispersal can steepen marine clines. Evolution 59, 2509–2517.
Huxley, J.S., 1938. Clines: An auxiliary taxonomic principle. Nature 142, 219–220.
Mallet, J., 2001. Gene flow. In: Woiwood, I.P., Reynolds, D.R., Thomas, C.D. (Eds.), Insect Movement: Mechanisms and Consequences. New York: CABI Publishing, pp. 337–360.
Nagylaki, T., 1978. Clines with asymmetric migration. Genetics 88, 813–827.
Slatkin, M., 1973. Gene flow and selection in a cline. Genetics 75, 733–756.
Sotka, E.E., Palumbi, S.R., 2006. The use of genetic clines to estimate dispersal distances of marine larvae. Ecology 87, 1094–1103.
Szymura, J.M., Barton, N.H., 1991. The genetic structure of the hybrid zone between the Fire-bellied toads, *Bombina bombina* and *B. Variegata*: comparisons between transects and between loci. Evolution 42, 237–261.
Woodruff, D.S., 1978. Mechanisms of speciation. Science 199, 1329–1330.

Coevolution

RB Langerhans, Harvard University, Cambridge, MA, USA

Introduction

What Is Coevolution?

All organismal populations experience multiple selective pressures deriving from varied aspects of their environment. In addition to abiotic features (e.g., climate), this 'environment' is often comprised of many other organisms. Thus, most populations evolve in response to interactions with other species. While the abiotic components of the environment cannot evolve in response to organisms, the biotic components can – and this phenomenon has played an integral role in the evolution of phenotypic diversity. Coevolution is reciprocal evolutionary change between interacting species driven by natural selection. That is, each player in a coevolutionary relationship evolves adaptations in response to its interaction with the other player(s). Although this general concept has been around since Darwin, the term 'coevolution' was coined by Paul Ehrlich and Peter Raven in a classic article in 1964, "Butterflies and plants: A study in coevolution." Since then, the field of research examining coevolution has blossomed into a large-scale research program.

The Broad Importance of Coevolution

Coevolution is undisputed as one of the most important processes shaping biodiversity. The importance of coevolution goes far beyond the classic examples, such as predator–prey coevolutionary arms races, figs and fig wasps, yuccas and yucca moths, ants and acacias, and fungal farming by several taxa. Coevolution's influence spans all subdisciplines within ecology and evolutionary biology. Indeed, a large extent of the historical and ongoing patterns of phenotypic evolution and species diversification is the product of coevolution. Coevolution can stem from numerous types of species interactions that are commonplace on this planet, such as interspecific competition for resources, predator–prey interactions, host–parasite interactions, plant–herbivore interactions, and flower–pollinator interactions. Even the eukaryotic cell originated from a symbiotic relationship where one of the species evolved into the organelles we now call mitochondria. A similar scenario is responsible for the formation of chloroplasts, and thus the origin of plants. Most vertebrate and invertebrate species rely heavily on coevolved symbionts residing within their digestive system or other special organs to allow proper digestion and growth. Coral reefs, and the communities they support, depend largely on coevolved symbioses between corals and zooanthellae, as well as interactions with other corals and algae-feeding fish. The symbiotic organisms, lichens, are critically important during primary succession in terrestrial ecosystems. Even the colonization of land by plants was facilitated by mutualistic interactions with mycorrhizal fungi. Coevolution's influence is so far reaching that the history of life on Earth would be unrecognizable in the absence of it.

Empirical Evidence for Coevolution

Despite the widespread importance of coevolutionary interactions, empirical demonstration of coevolution is a difficult task. This derives from the inherent difficulties of demonstrating adaptation – much less reciprocal adaptation – that has plagued evolutionary biology for many decades. Nevertheless, a great deal of supportive evidence has been gathered for coevolution, coming from an assortment of tests for an array of hypotheses stemming from coevolutionary theory. As is discussed below, coevolution comes in many forms, and thus there are different manners in which researchers approach the question of coevolutionary association. As a general approach, researchers often examine phenotypic, ecological, and genetic evidence to test the hypothesis that organisms are evolving (or have evolved) in response to one another.

The Basics of Coevolution

Types of Coevolution

A few different categories of coevolution are often discussed by scientists in ecology and evolutionary biology: pairwise coevolution, diffuse coevolution, and gene-for-gene coevolution. Pairwise coevolution (or 'specific' coevolution) describes tight coevolutionary relationships between two species. Diffuse coevolution (or 'guild' coevolution) refers to reciprocal evolutionary responses between suites of species. This type of coevolution emphasizes that most species experience a complex suite of selective pressures derived from numerous other species, and their evolutionary responses change the selective environment for other species. Gene-for-gene coevolution (or 'matching gene' coevolution) describes the specific case where coevolution involves gene-for-gene correspondence among species, such as when hosts and parasites have complementary genes for resistance and virulence.

 doi:10.1016/B978-0-444-63768-0.00471-6

Symbiosis and the Nature of Coevolutionary Interactions

Some of the classic, and most obvious, examples of coevolutionary interactions involve two species that live in continuing, intimate associations, termed symbiosis. With such tight ecological associations, strong and repeated coevolutionary responses are easy to envision. Symbiosis is commonly used to refer to all relationships between different species, and thus an association need not be extremely close to qualify as symbiotic. There are five major types of interspecific relationships: antagonism, parasitism, amensalism, commensalism, and mutualism (**Table 1**). Antagonism describes the scenario where both members of the interaction are harmed, such as interspecific competition. In parasitic coevolution, one member benefits (parasite), while the other is harmed (host). Parasitism is an extremely common form of coevolution. Parasitic interactions do not only include associations where one organism lives in or on a second organism, but also encompass some other very common interactions such as herbivory and predation. Sometimes the benefiting member kills the harmed member (e.g., parasitoid–host, predator–prey), whereas other times the harmed member is merely injured (e.g., parasite–host, herbivore–plant). Amensalism is where one member is harmed, while the other member is neither positively nor negatively affected. A common example of amensalism is the production of a chemical compound by one member as part of its normal metabolism which is detrimental to another organism (e.g., allelopathy in plants, toxic skin secretions in animals). Commensalism describes a symbiotic relationship where one member benefits, while the other member is neither helped nor harmed. It is possible for some interactions to be parasitic under some circumstances (e.g., low host nutrition), but commensal during others (e.g., high host nutrition). Mutualism is a coevolutionary relationship where both members benefit. For instance, clownfish gain protection from sea anemones, and anemones gain food from clownfish. Many relationships may change in nature over time, space, or ecological context, and the distribution of these types of outcomes for a given interaction (e.g., percentage of outcomes that are antagonistic vs. mutualistic) can be important in determining the coevolutionary responses that will be elicited.

The Red Queen Hypothesis

The Red Queen hypothesis was first proposed by Leigh Van Valen in 1973, and is a coevolutionary hypothesis describing how reciprocal evolutionary effects among species can lead to some particularly interesting outcomes. While Van Valen specifically addressed macroevolutionary extinction probabilities, the hypothesis has since become much more general, providing an evolutionary explanation for numerous characters (e.g., sex, mating systems, pathogen virulence, maintenance of genetic diversity), and coevolutionary arms races in general. The conceptual basis of the Red Queen hypothesis is that species (or populations) must continually evolve new adaptations in response to evolutionary changes in other organisms to avoid extinction. The term is derived from Lewis Carroll's *Through the Looking Glass*, where the Red Queen informs Alice that "here, you see, it takes all the running you can do to keep in the same place." Thus, with organisms, it may require multitudes of evolutionary adjustments just to keep from going extinct.

The Red Queen hypothesis serves as a primary explanation for the evolution of sexual reproduction. As parasites (or other selective agents) become specialized on common host genotypes, frequency-dependent selection favors sexual reproduction (i.e., recombination) in host populations (which produces novel genotypes, increasing the rate of adaptation). The Red Queen hypothesis also describes how coevolution can produce extinction probabilities that are relatively constant over millions of years, which is consistent with much of the fossil record. Thus, extinction resistance of lineages does not improve over time, but rather remains fairly constant because the probability of evolutionary change in one species leading to extinction in another species should be independent of species age (they are constantly evolving with their changing environments, not constantly improving with respect to a static background environment).

The Geographic Mosaic Theory of Coevolution

Owing to relatively recent developments in ecological, evolutionary, genetic, mathematical, and phylogenetic studies, coevolution is viewed today as an ongoing, highly dynamic process where populations interact across geographical landscapes. This contrasts with the historical view dating back to Darwin where coevolution was largely visualized as a slow, directional molding of species' traits through long periods of evolutionary time. Recent and ongoing research in coevolution is revealing that ecological and

Table 1 The five types of relationships between species. Effects of interaction are negative (−), positive (+), or indifferent (0)

Type of interaction	*Effects of interaction on*	
	Species 1	*Species 2*
Antagonism	−	−
Parasitism	−	+
Amensalism	−	0
Commensalism	+	0
Mutualism	+	+

evolutionary timescales are often one and the same, where effects of selection can be very rapid (strong intragenerational shifts), and evolutionary responses and counter-responses can be observable in only a few generations. This conceptualization of coevolution places a strong emphasis on the geographic context of coevolution and the continual reshaping of species' traits across geographic landscapes.

John Thompson has championed this new framework for studying coevolution, the geographic mosaic theory of coevolution (Fig. 1). This theory posits that coevolutionary interactions have three components driving evolutionary change:

- *Geographic selection mosaics*. Natural selection arising from interspecific interactions varies among populations.
- *Coevolutionary hot spots*. Interactions are subject to reciprocal selection only within some local communities (coevolutionary hot spots), embedded within a broader matrix of communities where selection is nonreciprocal or where only one of the participants occurs (coevolutionary cold spots).
- *Trait remixing*. Spatial distributions of potentially coevolving genes and traits are continually being altered due to new mutations, gene flow, genetic drift, and extinction of local populations.

Thus, with this theory, populations are placed within a context of geographic selection mosaics, providing a more complicated, but more realistic view, than previous perspectives. A couple of examples from nature help illustrate how this framework facilitates the understanding of how coevolution shapes species traits and interactions across landscapes: (1) garter snakes and newts, and (2) conifers and crossbills. These examples demonstrate how coevolutionary hot spots can result in geographic patterns in coevolved traits. First, *Taricha granulosa* newts and *Thamnophis sirtalis* garter snakes inhabit western North America, and show strong evidence of coevolving traits across this region. The newts possess a potent neurotoxin, which paralyzes and often kills a predator that has ingested a newt. However, the garter snake has evolved varying amounts of resistance to the neurotoxin (this resistance is physiologically costly). Newt toxicity levels and snake resistance levels are tightly matched across the geographic landscape (i.e., where newts are more toxic, snakes have greater resistance). In addition, snake populations, where newts do not co-occur, exhibit very low levels of resistance. Research has revealed two coevolutionary hot spots, and a number of intermediate and cold spots. Second, both red crossbills (*Loxia* spp.) and red squirrels (*Tamiasciurus hudsonicus*) prey upon pine cones. Where squirrels are the primary seed predator, conifers have evolved heavier pine cones with fewer seeds and thinner scales, which defends against squirrels. Where crossbills are the major seed predator, pine cones are lighter with more seeds and thicker scales, which defends against crossbills. Crossbills in turn have evolved counter adaptations to consume the seeds, exhibiting deeper, less curved bills where pine cones have thick scales compared to areas where pine cones have thin scales. Thus, trees are evolving in response to crossbills and squirrels in different populations, and crossbills are evolving in response to these evolutionary changes in the trees.

Coevolution Drives Diversification

Cospeciation and Phylogenies

Farenholz's Rule (originally proposed in 1913) posits that parasites and their hosts speciate in synchrony. This process, the joint speciation of two or more lineages that are ecologically associated (coevolving), has since been termed cospeciation (or parallel cladogenesis). While most research to date has examined parasite–host interactions, other coevolutionary relationships may also exhibit cospeciation; however, we will focus on parasites and hosts for this discussion.

If the process of cospeciation were the only one operating, then phylogenetic trees of parasites and their hosts should be topologically identical (i.e., exact mirror images of each other; Fig. 2). However, virtually all such phylogenies are not perfectly concordant. This implies that other processes must also be at work, such as host switching, speciating independently of their host, members going extinct, failure to colonize all descendants of a speciating host lineage, or failure to speciate when its host does. Further, even when concordant, it is possible that one of the groups (often the parasite) has colonized the other (the host) – host

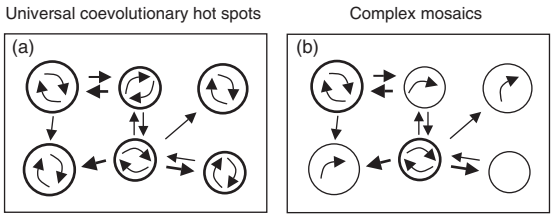


Fig. 1 Hypothetical illustration of the geographic mosaic of coevolution. The example depicts interactions between two species within local communities (arrows within circles). Interaction arrows represent different types of selection acting on different species. Arrows between communities reflect the magnitude of gene flow. Thick circles represent coevolutionary hot spots, while thin circles are cold spots. (a) Coevolution occurs in all communities, although the interaction coevolves in different ways among communities. (b) Coevolutionary hot spots occur within a matrix of cold spots. Adapted from *The Geographic Mosaic of Coevolution* by John Thompson (see Further Reading).

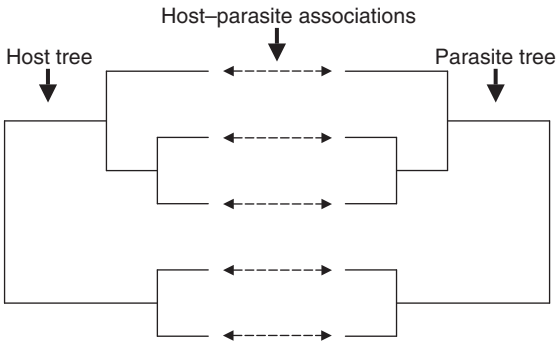


Fig. 2 Hypothetical phylogenies for host and parasite clades illustrating evidence for cospeciation.

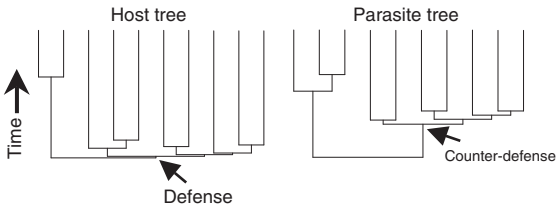


Fig. 3 Hypothetical illustration of escape-and-radiate coevolution. In this example, a host lineage has undergone an adaptive radiation after evolving an effective defense against the parasite clade. During this time, interaction with the parasites does not occur. Then, the parasite lineage evolved a counter-defense and also radiated. Because the parasite clade did not cospeciate or colonize each new host in a systematic manner, the parasite–host associations are not anticipated to exhibit one-for-one matching as in Fig. 2.

shifts might correspond to the host phylogeny because closely related hosts are more liable to colonization by closely related parasites. Comparisons of species' phylogenies can produce insight (with limitations) into the coevolutionary history of interacting organisms. Thus, as phylogenetic data, and more sophisticated tree-building methods, have become widely available, phylogenies have become very useful in the study of coevolution. A number of empirical studies now, at least partially, support the general notion of cospeciation occurring with some regularity in some parasite–host interactions.

Coevolution, Divergence, and Adaptive Radiation

Coevolutionary interactions can lead to phenotypic divergence among populations, speciation, and adaptive radiation. In addition to the process of cospeciation described earlier, diversifying coevolution across landscapes, which has now been demonstrated in several empirical systems, can (and apparently does) contribute to the formation of new species, as well as phenotypic differentiation within species. One hypothesis of coevolutionary diversification has its roots in the original article that first coined the term coevolution (Ehrlich and Raven article mentioned above), and proposes that reciprocal diversification among members of a coevolutionary association (often parasite–host) results from reciprocal adaptation, geographic differentiation, speciation, and periods of noninteraction in the diversifying lineages. This process is called escape-and-radiate coevolution (Fig. 3). In this process, one member evolves a defense (or some other innovation greatly reducing impact of interaction with other member), which enables a radiation due to the expansion of ecological opportunity. During the radiation, interaction among the members is minimal or nonexistent. Then, the other member evolves a counter-defense to overcome the innovation and radiates as well, producing reciprocal radiation events among the members. It is believed that diffuse coevolution among plants and herbivores may often follow such a coevolutionary diversification process.

Coevolutionary interactions may also often produce character displacement (exaggerated phenotypic divergence in sympatry) within local hot spots. This coevolutionary displacement is typically embedded within the broader geographic mosaic of coevolution among species, but may also result in fixation or speciation. A number of potential outcomes may result from coevolutionary displacement: character displacement among competitors in coevolutionary hot spots, displacement via apparent competition in hot spots, replicated community structure in hot spots, and trait overdispersion in competitive networks. Such displacement has been demonstrated in numerous systems in nature.

See also: Ecosystems: Coral Reefs. Evolutionary Ecology: Natural Selection. Eco-Evolutionary Dynamics. Red Queen Dynamics. Evolutionary Ecology. Microbiomes and Holobionts. Eco-Immunology: Past, Present, and Future. Association. Association. General Ecology: Pollination. Biodiversity

Further Reading

Combes, C., 2001. Parasitism: The Ecology and Evolution of Intimate Interactions. Chicago: University of Chicago Press.
Dawkins, R., Krebs, J.R., 1979. Arms races between and within species. Proceedings of the Royal Society of London B, Biological Sciences 205, 489–511.
Ehrlich, P.R., Raven, P.H., 1964. Butterflies and plants: A study in coevolution. Evolution 18, 586–608.
Farrell, B.D., Mitter, C., 1998. The timing of insect/plant diversification: Might *Tetraopes* (Coleoptera: Cerambycidae) and *Asclepias* (Asclepiadaceae) have co-evolved? Biological Journal of the Linnean Society 63, 553–577.
Norton, D.A., Carpenter, M.A., 1998. Mistletoes as parasites: Host specificity and speciation. Trends in Ecology and Evolution 13, 101–105.
Page, R.D.M. (Ed.), 2003. Tangled Trees: Phylogeny, Cospeciation, and Coevolution. Chicago: University of Chicago Press.
Paracer, S., Ahmadjian, V., 2000. Symbiosis: An Introduction to Biological Associations. Oxford: Oxford University Press.
Pellmyr, O., Herrera, C.M. (Eds.), 2002. Plant–Animal Interactions: An Evolutionary Approach. Malden, MA: Blackwell Science.
Silliman, B.R., Newell, S.Y., 2003. Fungal farming in a snail. Proceedings of the National Academy of Sciences of the United States of America 26, 15643–15648.
Thompson, J.N., 1999. The evolution of species interactions. Science 284, 2116–2118.
Thompson, J.N., 2005. The Geographic Mosaic of Coevolution. Chicago: University of Chicago Press.
Van Valen, L., 1973. A new evolutionary law. Evolutionary Theory 1, 1–30.
Waser, N.M., Ollerton, J. (Eds.), 2006. Plant–Pollinator Interactions: From Specialization to Generalization. Chicago: University of Chicago Press.

Coexistence☆

David J Booth and Bradley R Murray, University of Technology, Sydney, NSW, Australia

Introduction

One of the most perplexing issues in ecology is the coexistence of large numbers of species within communities across a range of spatial scales. The rainforests of South America and southeast Asia, for instance, are home to a high diversity of plant species, where over 200 species can be found in 1 ha of forest. On coral reefs, species richness is high at both large spatial scales (e.g., over 1500 species of fish and 400 corals on the Great Barrier Reef) and at the scale of small patches of reefs (e.g., over 200 species of fish on a reef only 50 m wide). The coexistence of large numbers of species is particularly puzzling given that many coexisting species seem to perform the same ecological roles within plant and animal assemblages. If so many species are doing the same thing, how can they all coexist in a world where resources are often limited and there is competition for these resources? To fully appreciate how such large numbers of species can coexist, we need to have a thorough understanding of the dynamics of their populations and, in particular, of how individuals within a species' population interact with each other, and with populations of other species, in the context of ambient environmental conditions. Indeed, the coexistence of individuals within a population is intriguing, given that interactions within species can in fact be stronger than interactions among the populations of different species. The ecologist George E. Hutchinson posed the *Paradox of the Plankton* in 1961, questioning how so many species of marine plankton could coexist in a relatively homogenous medium with limited food, seemingly in violation of Gause's rules of competitive exclusion. The research that Hutchinson stimulated over the next 45 years has given us a broad understanding of how organisms can coexist.

The Link Between Population Dynamics and Coexistence

The study of population dynamics involves investigating changes in features of populations such as numbers of individuals, individual weights and ages, and the biotic and abiotic conditions influencing those changes. The coexistence of the populations of different species simply refers to the persistence of two or more species' populations over time in a given area. Fluctuations in the dynamics of species' populations determine where, how, and when they will interact with individuals of the same species and individuals of other species. Those population interactions, and the ability to avoid interactions, are important properties determining the coexistence of large numbers of species. This article demonstrates the critical link between the dynamics of species' populations driving interactions among organisms and mechanisms promoting coexistence among species.

Competition and Coexistence

One of the most fundamental ways in which the populations of species interact is by competing with one another. In an environment with just a single species, the population of this species grows exponentially unless limited by crowding and a short supply of resources. Individuals of the same species have very similar requirements from the environment for growth and reproduction, and once individuals become crowded, individuals must compete with each other for limited resources. In this single-species system, the term "intraspecific competition" refers to the competitive interactions among individuals of the same species, where the growth, reproduction, and survival of some individuals are negatively affected. Grouping species (i.e., those species that aggregate for social or resource reasons) are particularly susceptible to crowding, and so may be expected to illustrate the most traits that enable coexistence. For instance, coral reef damselfishes can occupy small coral heads in groups of over 20 individuals, with a linear size-based dominance hierarchy (the largest fish is aggressively dominant over the next largest, and so forth; see **Fig. 1**). Larger groups suffer slower growth overall, due to increased competition for zooplankton food, and smaller (lower-status) fish within groups grow more slowly. However, survivorship and hence maturity are higher in larger groups due to greater vigilance for predators and predator swamping, so that coexistence is beneficial overall.

The situation becomes more complex when the dynamics of two or more species and their interactions in an environment need to be considered. A competitive interaction between the populations of different species is termed "interspecific competition," where growth, reproduction, and survival of one species is negatively affected either directly (e.g., interference competition) or indirectly (e.g., resource exploitation). For example, asymmetrical competition between two species can occur where one species has an advantage in resource exploitation or is able to interfere with another's ability to exploit resources. A great deal of effort has been invested in elucidating the role of interspecific competition in species coexistence.

☆*Change History:* March 2018, D.J. Booth made changes to the text, figures and references.

This is an update of D.J. Booth and B.R. Murray, Coexistence, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 664–668.

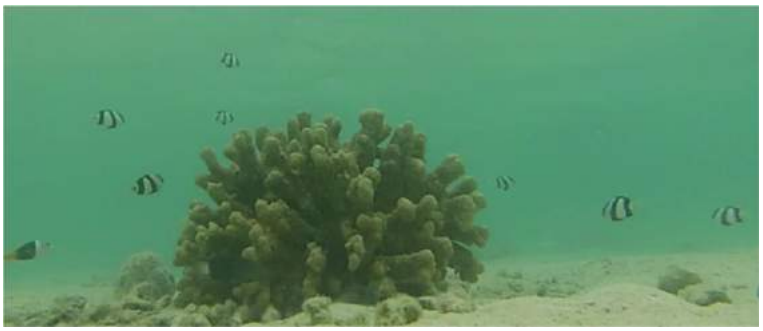


Fig. 1 Social group of the humbug damselfish (*Dascyllus aruanus*). Individuals coexist in dominance hierarchies. Photo by David Booth.

Interestingly, however, a paucity of opportunities for interspecific competition in some communities can result in the coexistence of large numbers of species. This situation may arise when the individuals of species are clumped (aggregated) in their spatial distribution. For example, populations of many shrubby plant species are spatially clumped in open-forest plant communities in southeastern Australia. Plant species that are better competitors for resources would normally out-compete and eliminate inferior competitors, but competitively inferior species are able to coexist with competitively superior species where there is clumping of individuals. This is because individuals of competitively superior species are mostly competing with each other when clumped, and not with populations of inferior competitors, thus reducing the chance of weaker competitors being eliminated.

Competitive networks can promote coexistence in crowded communities by allowing no one species to be competitively dominant in all situations. For example, if species A is dominant over species B (i.e., A > B), B > C, and C > A, a simple three-species network will result. On coral reefs, such networks in corals apparently promote coexistence in the face of overgrowth.

The Importance of Niche Differentiation for Coexistence

Competitive interactions among the populations of two species will lead to the exclusion of one of the species when the realized niche of the superior competitor encompasses the fundamental niche of the inferior competitor. This is known as the competitive exclusion principle. (Note that the fundamental niche of a species describes all possible combinations of resources and conditions under which species' populations can grow, survive, and reproduce; the realized niche describes the more limited set of resources and conditions necessary simply for the persistence of species' populations in the presence of competitors and predators.) The fact that so many species in a vast array of ecological communities are able to coexist means that species must differ in their realized niches. Such niche differentiation is critical to avoid competitive exclusion and allow coexistence.

Niche differentiation refers to differences among species in their physiology, morphology, and behavior, and concurrently in their use of resources and tolerance of conditions. Perhaps the simplest idea is that species partition available resources; that is, there is differential utilization of resources among species. Resources are often separated either spatially and/or temporally. For example, native mammals can often coexist in the same area because they forage and seek protection in different microhabitats provided by variation in native vegetation. In the Australian arid zone, for instance, the spinifex hopping-mouse *Notomys alexis* forages in the open while the sandy inland mouse *Pseudomys hermannsburgensis* is more commonly associated with hummocks of spinifex grass. The two species differ considerably in their morphology, related to their differential use of spatially segregated resources. The hopping-mouse has large hind legs for bipedal motion, which provides for rapid movement and quick changes in direction which are necessary for predator avoidance in the open habitats in which it forages. On the other hand, the sandy inland mouse is much more "mouse"-like in its limb structure, using all four limbs for quadrupedal motion, an efficient means of movement in tangled spinifex clumps.

The Roles of Predator–Prey Interactions and Density Dependence in Coexistence

Individuals in populations and communities of species can coexist provided shared resources are not in short supply. Predation is one process that precludes resource depletion and may thus facilitate coexistence. Predators may act as mediators of coexistence by harvesting all prey types equally, and so preventing resource depletion. Also, predators may switch to the most common of alternative prey types, which again will prevent competitive extinction. Lotka–Volterra predator–prey dynamics can explain coexistence of closely matched predators and prey, as seen in classic cycling of abundances of hares and lynx in Canada.

Tradeoffs and Species Coexistence

A "super-species" is the best at colonizing new patches of habitat, at outcompeting all neighbors and at avoiding predators. Such a species does not exist. The benefits obtained from performing one ecological function well come at the cost of performing a

different function. Such differences among species are referred to as tradeoffs. The population dynamics of different species are influenced by interspecific tradeoffs, which subsequently influence the possibilities for coexistence among species. It is important to note that in this context, tradeoffs represent niche differentiation among species (see above).

One way to consider a tradeoff is as a negative functional interaction between traits. In plants, for example, shrubs and trees must invest much of their resources into support structures to obtain height. This comes at the expense of investment in photosynthetic tissue. In contrast, vines must use host plants for structural support in order to reach any considerable height. The use of host plants for support results in a greater allocation of energy (which would otherwise be invested in stem tissue) toward photosynthetic tissue.

In spatially structured landscapes, the competition–colonization tradeoff is an important mechanism that can generate coexistence. Simply, while some species are good competitors, others are better colonizers. No one species can be good at both: there is a tradeoff, such that species with high fecundity and dispersal are poorer competitors, while species that are good competitors necessarily have lower fecundity and poorer dispersal. Under a fairly strict set of conditions, relating to the dynamics of competing populations, the competition–colonization tradeoff predicts the coexistence of a potentially infinite number of species. These conditions include the idea that there is a strict competitive hierarchy among species (i.e., competitive asymmetry), such that the superior competitor will always win competitive interactions with competitively inferior species. Being a good competitor comes at the cost of being a poorer colonizer, such that inferior competitors that are better colonizers can occupy sites that have not been colonized by the superior competitor.

For plants, the interspecific competition–colonization tradeoff has been linked to the life-history tradeoff relating seed size to seed number. From a finite pool of resources, a plant can make either many small seeds or few large seeds. There is fairly clear empirical evidence that seedlings from large seeds out-compete seedlings from small seeds during early seedling growth. The larger production of (smaller) seeds in other species leads to superior colonization. These differences in seed size and seed output neatly match the requirements of the competition–colonization tradeoff model for coexistence. Theoretical work, matched with neat empirical research, by Mark Rees and his research group reveals the complex nature of the competition–colonization tradeoff in annual plant communities. Importantly, the tradeoff appears to work as the sole mechanism of coexistence in the real world only when competitive asymmetries are extreme.

One form of coexistence (called density-dependent coexistence) requires that life-history strategies differ such that an advantage at one stage of the life cycle implies a disadvantage at another stage in the life cycle. With two species, each gains a relative advantage at some level of total density, and where densities fluctuate, both species may persist indefinitely. Such life-history tradeoffs appear in small-bodied fish assemblages in freshwater lakes, for instance, although density-dependent coexistence has not been demonstrated unequivocally in nature as yet.

Coexistence Through Variable Supply of Young

Several models of species coexistence have been developed through empirical studies of coral reef ecosystems, where species diversity at very small spatial scales can be considerably high. How do small (<100 m^2) patch reefs support over 300 species of fishes and many more invertebrates and algae? First, many reef species exhibit "functional versatility" whereby their use of resources (food, shelter) is flexible enough to allow many apparently overlapping species to coexist. The range of microhabitats present on coral reefs is large, also facilitating coexistence. However, the patchy and variable supply of young onto the reef (with parallels in other marine organisms and terrestrial insects) can also promote coexistence. The "lottery hypothesis" suggests that, should a reef resident be removed (e.g., through mortality) the replacement would be drawn at random from the larval species pool adjacent to the reef at that time. Therefore, even if certain species were competitively displaced from a habitat patch, it could be replenished through the larval pool. Hypothetical successional climaxes of reefs, forests, and many other ecosystems would never be reached, so resulting coexistence would not be stable. One consequence would be enhancement of biodiversity in these situations, through invasion of larvae, seeds, and other propagules or migrants. Such invasion could be facilitated by mortalities or displacement of residents through a disturbance (see the next section).

Coexistence Mediated by Disturbance

Physical disturbance of habitat can also facilitate coexistence of individuals and species. Gap formation in rainforests through storms and fire is a major mechanism promoting coexistence and by which many species are maintained. Cyclones and smaller-scale storms on coral reefs may shift coral assemblages from domination by massive, slow-growing forms to faster-growing branching taxa. Crown-of-thorns seastar invasions can decimate reefs at small scales, leaving a mosaic of habitats in their wake. By preventing reefs reaching successional climaxes, disturbances such as these may enhance coexistence of the species that live there.

Neutral Models of Coexistence

Recently, there has been a surge of interest in ecology for neutral models of coexistence. In neutral models of coexistence, niche differentiation among species is removed from consideration and species are considered to be equivalent. However, fundamental

Table 1 Some models and explanations of processes that promote coexistence among species and individuals within habitats

Model	*Explanation*	*Example*
Lottery	Available space occupied at random from larval/propagule species pool	Species with dispersive larvae/propagules (e.g., coral reef fishes) occupying space at random
Recruitment limitation	As above, but habitat occupation limited by patchy supply of larvae/propagules	Marine benthic organisms esp., at edge of their range
Niche diversification		Many available microhabitats promote coexistence
Different plant parts used by invertebrates as a resource		
Tradeoffs	High performance in one life-history feature at the expense of reduced performance in another	Plants either produce many small seeds (good colonizing ability) or few large seeds (good competitive ability)
Competitive networks	No species is dominant over all others	Overgrowing corals
Predation limitation	High predation (esp. on common prey-switching) promotes coexistence	Predation on lower rocky intertidal marine environments
Abiotic disturbance	Physical factor(s) that create space for colonization/replenishment	Storm gaps in rainforest canopies
Intraspecific aggregation	Species undergo higher intraspecific than interspecific competition	Patchily distributed and aggregated plant populations of different species
Asynchronous life histories	Seasonal resource use asynchronous among species, promoting coexistence (e.g., reproductive cycling, population irruptions)	Cicadas
Neutral theory	No niche diversification among species, probabilities of birth, death, and immigration and interspecific competition determine coexistence	High diversity of plant species in rainforests

properties of population dynamics (e.g., probabilities of birth, death, and immigration) and interspecific competition play an important role in determining coexistence. Many of the patterns of coexistence and diversity, species–area relationships, and the distribution of abundances among species in the real world can be explained through neutral models that do not include niche differentiation among species. Frequency-dependent natural selection can also generate coexistence between species that are apparently neutral (i.e., there is no niche differentiation between the species). In this situation, when species are rare (in low-abundance), they adapt to a high frequency of interspecific competitive interactions, thus providing for their persistence. When species are common (in high-abundance), they adapt to a high frequency of intraspecific competitive interactions, as they are more likely to encounter conspecifics in the habitat in which they are abundant. Under these conditions, the species are said to have evolved "pseudo-neutrality." Nevertheless, the problem is that a community theory that accounts neither for species differences nor for patterns of species distribution across real physical geography must be incomplete.

Conclusion

A wide range of ecological interactions separately or in concert, facilitate coexistence within populations and between species (for a summary see **Table 1**). Coexistence mechanisms can explain both the enormous species richness in some habitats and the dominance of a few taxa in others. Close coexistence of individuals within a species, or very similar species, is possible through a combination of biotic interactions and physical disturbances which may prevent "equilibrium" conditions ever being reached.

See also: Behavioral Ecology: Herbivore-Predator Cycles; Competition; Food Specialization; Age Structure and Population Dynamics. Conservation Ecology: Source–Sink Landscape; Invasive Plant Species. Ecological Complexity: Complex Ecological Networks. Ecological Data Analysis and Modelling: Species Distribution Modeling. Ecological Processes: Succession and Colonization. Evolutionary Ecology: Colonization; Gause?s Competitive Exclusion Principle; Life-History Patterns; Evolutionary Ecology; Ecological Niche; Metacommunities. General Ecology: Community; Habitat; Ecophysiology; The Intermediate Disturbance Hypothesis; Biodiversity. Terrestrial and Landscape Ecology: Spatial Distribution

Further Reading

Allesina, S., Levine, J.M., 2011. A competitive network theory of species diversity. Proceedings of the National Academy of Sciences of the United States of America 108, 5638–5642.

Bell, G., 2001. Neutral macroecology. Science 293, 2413–2418.

Booth, D.J., 1995. Survivorship and growth within social groups of the domino damselfish *Dascyllus albisella*. Ecology 76, 91–106.

Booth, D.J., Brosnan, D.M., 1995. The role of recruitment dynamics in rocky shore and coral reef fish communities. Advances in Ecological Research 26, 309–385.
Buss, L.W., Jackson, J.B.C., 1979. Competitive networks: Nontransitive competitive relationships in cryptic coral reef environments. American Naturalist 113, 223–234.
Chase, J.M., Leibold, M.A., 2003. Ecological niches: Linking classical and contemporary approaches. Chicago: University of Chicago Press.
Dickman, C.R., 2006. Species interactions: Direct effects. In: Attiwill, P., Wilson, B. (Eds.), Ecology: An Australian perspective, 2nd edn South Melbourne: Oxford University Press, pp. 285–302.
Dickman, C.R., 2006. Species interactions: Indirect effects. In: Attiwill, P., Wilson, B. (Eds.), Ecology: An Australian perspective, 2nd edn South Melbourne: Oxford University Press, pp. 303–316.
Dickman, C.R., Murray, B.R., 2006. Species interactions: Complex effects. In: Attiwill, P., Wilson, B. (Eds.), Ecology: An Australian perspective, 2nd edn South Melbourne: Oxford University Press, pp. 313–334.
Falster, D.S., Murray, B.R., Lepschi, B.J., 2001. Linking abundance, occupancy and spatial structure: An empirical test of a neutral model in an open-forest woody plant community in eastern Australia. Journal of Biogeography 28, 317–323.
Hubbell, S.P., 2001. The unified neutral theory of species abundance and diversity. Princeton, NJ: Princeton University Press.
Levine, J.M., Rees, M., 2002. Coexistence and relative abundance in annual plant assemblages: The roles of competition and colonization. American Naturalist 160, 452–467.
Murray, B.R., Dickman, C.R., 1994. Granivory and microhabitat use in Australian desert rodents: Are seeds important? Oecologia 99, 216–225.
Rees, M., 1995. Community structure in sand dune annuals: Is seed weight a key quantity? Journal of Ecology 83, 857–863.
Tilman, D., 1994. Competition and biodiversity in spatially structured habitats. Ecology 75, 2–16.
Turcotte, M.M., Levine, J.M., 2016. Phenotypic plasticity and species coexistence. Trends in Ecology and Evolution 31, 803–813.
Turnbull, L.A., Coomes, D., Hector, A., Rees, M., 2004. Seed mass and the competition/colonization trade-off: Competitive interactions and spatial patterns in a guild of annual plants. Journal of Ecology 92, 97–109.

Colonization

MJ Donahue, Humboldt State University, Arcata, CA, USA
CT Lee, Stanford University, Stanford, CA, USA

Introduction

Colonization is the arrival of individuals to areas of suitable habitat that are currently uninhabited by individuals of the same species. Populations are established (or re-established) in uninhabited areas by the successful colonizers that survive and reproduce. Colonization, or recolonization, is a spatial process central to several fundamental concepts in ecology, including the spatial structure of populations, species coexistence, succession, disturbance and recovery, invasive species, and speciation.

Colonization is one possible consequence of dispersal. Dispersal is the permanent movement of an individual from one location to another (commonly, a seed, larva, or juvenile stage moving from its natal area to the area it will inhabit as an adult). If the individual disperses to an area uninhabited by conspecifics, then dispersal has resulted in colonization. If that colonizer survives to reproduce, then a new population has been established.

Colonization occurs at a range of spatial scales. For sessile organisms in which space is a primary resource (e.g., plants, sessile marine invertebrates), the death of an individual may open space for recolonization by another individual. At this smallest scale, the balance between extinction (individual mortality) and colonization (settlement of another individual in that location) determines the persistence of a population. Small disturbances, such as gopher mounds in grasslands or tear-outs in intertidal mussel beds, create small habitat patches that allow good colonizers to coexist with dominant competitors, enhancing local biodiversity through competition–colonization tradeoff. Larger-scale disturbances, due to fire, logging, or sedimentation–erosion dynamics, may result in a successional mosaic, a landscape in which different areas were disturbed at different times in the past, resulting in a range of successional stages and, therefore, higher regional biodiversity. Colonization is also essential for range expansion and spatial spread. Anthropogenic effects have increased the rates of spread of many species by enhancing colonization. Over a longer timescale, colonization of new niche space may lead to niche expansion or speciation.

We first consider characteristics which enhance species' colonization ability. Then, we use the central ecological concept of colonization–extinction balance to consider the implications of colonization from the smallest scale of an individual to the largest scale of species ranges. In doing so, we consider the implications of colonization for population growth, species coexistence, disturbance, succession, species invasions, and speciation.

Factors Enhancing Colonization Ability

Organisms vary considerably in their ability to colonize new habitat. **Figs. 1a** and **1b** illustrate the variation between taxa of freshwater zooplankton in the time needed to colonize experimental mesocosms, and in the total number of mesocosms colonized over the course of 2 years. In general, whether or not a particular organism colonizes a given new habitat depends on life-history characteristics, behavior, and environmental factors. Classically, the *r*-selected life-history strategy is associated with strong colonization ability through the production of many, low-quality propagules, often adapted to long-distance dispersal via wind or water. This strategy increases the probability that propagules will arrive in a patch because of the vast number of propagules released. However, strategies that increase the probability of establishment, rather than simply the probability of arrival, also enhance colonization success, including habitat selection and competitive ability. Many animals use habitat selection to detect appropriate habitat, thereby increasing the probability of arrival and subsequent establishment. This is particularly true for specialist organisms with highly specific habitat requirements or organisms that utilize short-lived resources (e.g., rotting fruit or hydrothermal vents). Good competitors, especially those with clonal growth, often make good colonizers; few propagules may arrive, but those that do are likely to be successful.

Among herbaceous plants, an annual life-history strategy, rapid growth, and the production of many, small seeds are characteristic of the `ruderal' strategy (*sensu* Grime; see section titled **`Further reading'**). Ruderals rapidly colonize habitat opened by disturbance, allowing them to coexist with stronger competitors that colonize more slowly but ultimately exclude the ruderal species. Ruderals are *r*-strategists that increase colonization ability by increasing the probability of arrival at a new patch. Other life-history characteristics may enhance the probability of establishment after arrival, such as phenotypic plasticity and the ability to reproduce asexually. Phenotypic plasticity allows an organism to use the morphology or behavior that is appropriate to its new environment. The ability to reproduce asexually, whether by clonal growth or selfing, is clearly advantageous to establishment because only a single individual is necessary to produce a viable population. In their experimental mesocosms, however, Caceres and Soluk (2002) found that the colonization ability of freshwater zooplankton was not necessarily predicted well by mode of reproduction (see the section titled **`Further reading'**).

 doi:10.1016/B978-0-444-63768-0.00472-8

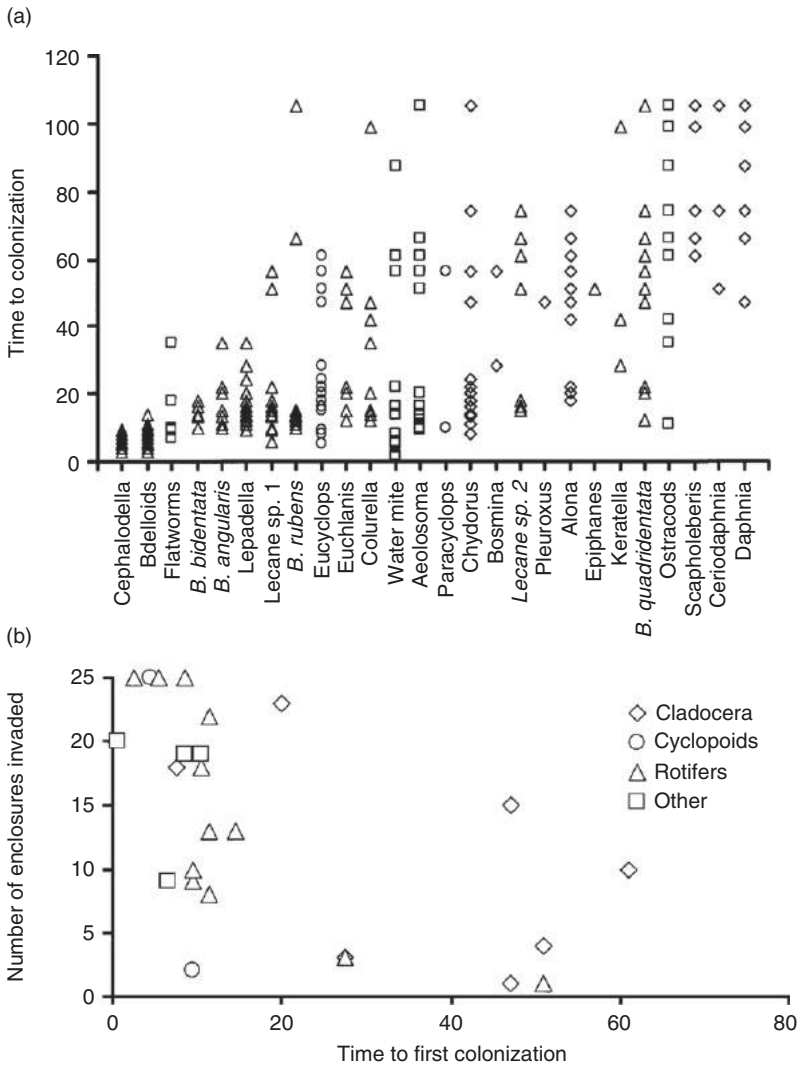


Fig. 1 (a) Time to colonization (weeks) into each mesocosm for each of 26 zooplankton taxa. Taxa are listed in order of average time to colonization. Diamonds indicate cladocera; circles indicate copepods; triangles indicate rotifers; squares indicate all other taxa. Because colonization by individuals is difficult to observe, `colonization' here also includes establishment of reproducing populations large enough to be detected in samples. (b) Negative correlation between the week each taxon was first found in the array, and the number of mesocosms that were invaded over the 2 years. The other category includes annelids, flatworms, water mites, and ostracods. Reprinted from Caceres, C.E., Soluk, D.A., 2002. Blowing in the wind: A field test of overland dispersal and colonization by aquatic invertebrates. Oecologia 131, 402–408.

Among sessile marine invertebrates, there are two contrasting life-history strategies that may enhance colonization ability. Some taxa produce many, small larvae that feed during an extended larval period (weeks to months) in the plankton and are, therefore, transported over long distances. This long larval period results in long-distance transport far from their natal areas, enhancing the probability of arrival in uninhabited areas. Other taxa produce fewer, larger, larvae with enough yolk to survive a shorter (hours to days), nonfeeding planktonic period or, more extreme, larvae that crawl away and have no planktonic transport at all. While these taxa are less likely to be transported far from their natal areas, when they do arrive they are more likely to colonize successfully and establish a new population. This increased colonization success is due to the increased survival of colonizing individuals but also to more rapid local population growth since subsequent reproduction is retained locally. These examples make it clear that there are two strategies for colonization success: produce many propagules to ensure arrival at new habitat or produce well-supplied or highly competitive propagules that are likely to succeed if they arrive.

Competition–Colonization Tradeoff

Ecologists contrast *r*- and *K*-strategy life histories to identify the broad pattern of life-history tradeoffs between short-lived organisms with fast maturation, good dispersal ability, and many, small propagules and long-lived organisms with slow

maturation, strong competitive ability, and few, well-supplied propagules, respectively. Because resources are finite, selection has forced a tradeoff between colonization ability and competitive ability. This tradeoff also produces a mechanism of species coexistence, called competition–colonization tradeoff.

In systems where disturbance creates colonization opportunities, competition–colonization tradeoffs can lead to coexistence between multiple competitors that share a single, limiting resource. This tradeoff is most easily imagined when the limiting resource is space; the dominant species is a better competitor and will overgrow the fugitive species, which is a better colonizer. Disturbance results in mortality of both species and opens space on the landscape. If the fugitive species can colonize and exploit this new space fast enough, then the dominant and fugitive species can coexist.

We can see competition–colonization tradeoff as one case of colonization–extinction balance. Species coexistence is determined by the relative colonization rates of each species, the mortality rate of both species due to disturbance, and the mortality rate of the fugitive species by competitive exclusion. For coexistence to occur, the fugitive species and dominant species must each solve the problem of colonization–extinction balance; the dominant species must have a high-enough colonization rate to avoid extinction by disturbance alone, while the fugitive species must have a higher colonization rate to withstand extinction from both disturbance and competitive exclusion. It acts at the scale of an individual or small group of individuals and determines whether one or both species will persist.

This pattern of coexistence has been best documented between organisms that compete for space. One classic example of competition–colonization tradeoff is the interaction between the annual sea palm, *Postelsia palmaeformis*, and the long-lived mussel, *Mytilus californianus*, in the rocky intertidal of the northeast Pacific. *M. californianus* is the competitive dominant, which will exclude *P. palmaeformis* over time. However, in areas of high wave energy, patches of M. *californianus* are ripped from the rocky substrate and *P. palmaeformis* colonizes these open patches more quickly than *M. californianus*. Transplant experiments demonstrate that *P. palmaeformis* can survive and reproduce in areas of low wave energy; however, in the absence of wave disturbance, it is overgrown by *M. californianus*. Therefore, when wave energy is high enough, *M. californianus* mortality due to disturbance and the faster colonization rate of *P. palmaeformis* allow these two species to coexist. For more on this examples, see section entitled **Further Reading**.

In Laikipia district of Kenya, fire-prone bushland savannah is dominated by a single species of swollen-thorn acacia tree, *Acacia drepanolobium*, which is host to four acacia-ant species that are obligate mutualists with *A. drepanolobium*. The ants nest within the swollen thorns and feed on nectaries of *A. drepanolobium*; the presence of ants deters herbivores from browsing on *A. drepanolobium*. Any one acacia tree hosts a single colony of ants. The four species of ants coexist on a single limiting resource of host trees; the species form a strict competitive hierarchy, in which more dominant species displace subordinate species from neighboring host trees. Unoccupied trees arise due to fire and elephant disturbance, which destroys colonies but not trees, and by small trees growing into a habitable size. A tree may be colonized by a colony from a neighboring tree or by a foundress queen. The two subordinate ant species both have higher colonization rates; the most subordinate species produces many more foundress queens to colonize newly available mature trees while the second most subordinate species has a higher than expected rate of colonizing empty neighboring trees by colony expansion. This tradeoff between competitive dominance and colonization ability is one mechanism of coexistence for these four species sharing a single limiting resource. For more on this examples, see the section titled **`Further reading'**.

Competition–Colonization Tradeoff and the Successional Mosaic

In systems where disturbances act on larger spatial and temporal scales, competition–colonization tradeoffs can maintain regional species diversity. Large-scale disturbances, such as fire, logging, or sedimentation–erosion dynamics, open up tracts of unoccupied habitat. These newly opened areas are colonized by a predictable sequence of species in a process called ecological succession. Newly disturbed areas are first colonized by ruderal pioneer species. These species may be outcompeted or overgrown by larger or longer-lived organisms. Over time, competitively superior species will dominate. When different areas in a landscape are disturbed at different times, the landscape becomes a patchwork of communities in different stages of succession, called a `successional mosaic'. When disturbance is moderate, this successional mosaic is thought to result in high diversity at the regional scale because communities in all successional stages are represented. This is called the `intermediate disturbance hypothesis' and is discussed in detail in another article in this volume.

The intermediate disturbance hypothesis has been proposed as an explanation for patterns of macroinvertebrate diversity in lotic streams in the Taieri River catchment in New Zealand. In this system, the intensity and frequency of disturbance are determined by the flow rate and frequency, respectively, of periodic high-discharge events. Disturbance varies from reach to reach in the catchment due to stream slope, distance from headwaters, and local topography, among other factors. Macroinvertebrate diversity is a unimodal function of disturbance intensity. At high disturbance intensity, some stream macroinvertebrates cannot survive a disturbance or colonize fast enough after a disturbance to persist in high-disturbance reaches. At low disturbance intensity, macroinvertebrates are distributed evenly, indicating competition among the species that remain long after a disturbance. For more on this examples, see section entitled **`Further reading'**.

A critical piece of this `successional mosaic' is the availability of colonists from other populations on the landscape. In streams, these colonists can come from several sources: upstream habitat, flying adults from another reach, or refuge habitat that organisms use to escape the effects of disturbance. If disturbance is too frequent (or to rare) across the entire landscape, then there will be no

areas in the later (or earlier) stages of succession to provide colonists to areas in other stages of succession. At this scale, colonization–extinction balance can influence the diversity of local communities (e.g., stream reaches) with different disturbance regimes and whole regions (e.g., river catchment) through the successional mosaic.

Metapopulation Biology

On larger spatial scales still, extinction of entire populations can provide the requisite unoccupied suitable habitat for colonization to occur. If recolonization from extant local populations occurs quickly enough to balance local extinction, persistence of a species on a regional scale can occur. Such a regional population, called a metapopulation is an important concept for understanding population dynamics in patchy habitats.

A classic metapopulation is a regional population composed of many local populations, each of which may be extant or extinct at any one point in time. This situation occurs when patches of suitable habitat are separated by uninhabitable areas. If habitat patches are essentially equal, and the dynamics of individual populations are fast relative to interpatch dynamics, then the important characteristic of a given patch is whether or not it is occupied. Thus, the primary variable in classic metapopulation theory is the proportion of patches that is occupied (or the proportion of populations that is extant), and the determinants of this proportion are the rate of local population extinction and the colonization rate. Extinction occurs because of demographic stochasticity; no finite population can persist indefinitely. Colonization occurs via movement of individuals from extant populations to empty patches, and all patches are assumed to be equally accessible to all others. The theory, first set out in 1970 by Richard Levins, describes the dynamics of patch occupancy by the equation

$$\mathrm{d}p/\mathrm{d}t = cp(1-p) - ep$$

Here, p is the proportion of occupied patches, e is the extinction rate per occupied patch, and c is the rate of colonization per occupied patch per unoccupied patch. Alternatively, if p is the probability that a given patch is occupied, e is the probability that a given extant population goes extinct, and c is the probability that an individual from a given extant population colonizes a given empty patch. The equilibrium proportion of extant populations is $P = 1 - e/c$. Thus, the colonization rate must exceed the extinction rate to ensure persistence of the regional ensemble, and the magnitude of the excess determines the species' regional abundance. Factors such as the number of potential colonists leaving occupied patches, their dispersal ability, and their ability to establish new populations in unoccupied habitat patches are all included in the overall rate of colonization, c, in this model framework.

The classic metapopulation model involves several assumptions. While few natural populations may strictly meet these assumptions, the model provides a powerful conceptual framework for patchy populations, with important implications for conservation and evolution as well as population biology. It also serves as a starting point for introducing complications such as spatially restricted dispersal, spatial aggregation of and/or correlation between local populations, differential size or quality of habitat patches, etc. Such modifications may allow the model to describe the dynamics of the many natural populations that meet a looser definition of a metapopulation as a regional population comprised of local populations that are interconnected by dispersal.

The metapopulation approach is perhaps most useful for modeling disease dynamics, since many of the classic model's assumptions are best met by populations of microorganisms. From the point of view of the disease, occupied patches are infected hosts and empty ones are susceptible hosts. The disease spreads if colonization via infection of susceptible hosts exceeds extinction via host immune response. The Kermack–McKendrick model, proposed to describe the dynamics of bubonic plague and cholera, is one of the simplest examples of this approach. The model consists of three differential equations describing the dynamics of susceptible hosts (suitable, unoccupied patches, $|S|$), infected hosts (suitable, occupied patches, $|I|$), and recovered hosts who have developed immunity from disease (unsuitable patches, $|R|$):

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta SI$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \beta SI - \gamma I$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I$$

where β is the infection rate per infected host per susceptible host, and γ is the recovery rate per infected host. The infection rate in this epidemic model is thus analogous to the colonization rate in the classic metapopulation model. The major difference in this case is that with extinction of disease populations due to host immune response, habitat patches are also rendered unsuitable. Therefore, the question of most interest is not how many patches are infected at equilibrium, but whether or not the disease ever spreads. In this model, spread occurs if $R_0 > 1$, where

$$R_0 = \frac{\beta S}{\gamma}$$

Thus, although models of disease dynamics and the questions they ask and answer may differ in important ways from basic metapopulation models, the role of colonization and its balance with extinction is clear in both cases. More sophisticated metapopulation models for disease allow host population dynamics and host mobility, as well as any of the refinements to the metapopulation model mentioned above.

Island Biogeography

We turn now to considering the balance between colonization and extinction on the spatial scale of continents, oceanic islands, and the distances between them. In doing so, we move from a single species occupying many patches to many species in a single large patch. The influential theory for this situation is `island biogeography theory', and covers it in detail. In brief, the theory predicts that the number of species on an island depends on a dynamic balance between colonization of the island by species from the mainland and extinction of species already on the island. The mainland serves as a permanent pool of a fixed number of colonist species, and the island is permanently suitable for all of them, but each species colonizes the island only when not currently present there. As a result, as the number of species occupying the island increases, the rate of colonization per unit time decreases. If all species were equal in their colonization ability, then the colonization rate would decrease linearly with the number of species on the island. Since species usually differ in colonization ability, however, as we discussed above, the poorer colonizers take longer to arrive, and the better colonizers are likely already present when they do. Thus, we expect the decrease in the colonization rate to be more gradual at higher island species richness, and the rate should be a convex decreasing function of species richness on the island. The exact value of the colonization rate function at any richness depends on the specifics of a given situation, including the identities of the species in the colonist pool and the isolation of the island from the mainland. The intersection of the colonization function with the extinction function (an increasing function of island richness) determines the richness at which colonization and extinction balance. This is the island's equilibrial or long-term species richness.

Empirical tests of the general theory have upheld its essential points. The classic model, however, does not detail interaction between colonizing species (such as facilitation of colonization by earlier colonists, for example and also excludes any effects on species richness of evolution on either the mainland or the island. Despite these omissions, the model provides a valuable conceptual framework for understanding the role of colonization in determining diversity and community composition over large spatial scales, and also provides a starting point for understanding spatial systems other than oceanic islands. For instance, the theory can apply to habitat islands such as mountaintops or lakes, as long as a permanent `mainland' habitat exists to provide colonists. `Island biogeography theory', where colonization occurs between pairs of very unequal habitat patches, and metapopulation theory, where colonization takes place between many equal patches, are two special cases of colonization–extinction interactions in patchy habitats. These two classical theories are thus useful tools for understanding the dynamics of the many patchy natural systems on regional to continental spatial scales. Metacommunity theory is a contemporary synthesis of these and other conceptual approaches to spatially structured communities.

Availability of New, Suitable Habitat on a Large Scale

Biological invasion is a prominent example of colonization of new areas of suitable habitat. Invasive species often have life-history characteristics that make them good colonizers. In cases where invasive spread occurs most quickly, many secondary colonization events follow the initial introduction, so that rapid spread occurs by the establishment, increase, and eventual coalescence of many small and widely dispersed populations descended from the initial colonizers of the new habitat. Species that exhibit such rapid, patchy spread include cheatgrass, *Bromus tectorum*, in western North America; smooth cordgrass, *Spartina alterniflora*, in Willapa Bay, Washington State; and the Argentine ant, *Linepithema humile*, which disperses poorly on its own but covers long distances when assisted by humans. Because many conspicuous invasive species disperse widely, invaders often fall into the `ruderal' or `fugitive' life-history categories; but many (such as yellow starthistle, *Centaurea solstitialis*, and the Argentine ant, *L. humile*) are very strong competitors, while others (such as saltcedar, *Tamarix ramosissina*) tolerate extreme environments very well. Whether they are good colonizers because they disperse often, survive well during dispersal, and/or find many types of habitat acceptable for vigorous growth (sometimes despite the presence of a prior resident), invasive species increasingly take advantage of new dispersal pathways opened by the global human economy to reach previously unavailable areas of suitable habitat. These issues are of primary importance in attempts to slow, stop, or prevent biological invasions that are potentially devastating to native communities.

Changes in habitat suitability can provide new opportunities for colonization on large spatial scales. One way to be a good colonizer is to enhance the suitability of a new habitat, and some species can facilitate their own spread into new areas by changing the properties of the surrounding ecosystem. For instance, colonization by alien grasses can increase the fuel available to fire and thereby increase fire frequency, area, and intensity. The grasses recover quickly from fire and colonize burned areas, generating a positive feedback loop that favors the grasses' spread. Alternatively, one colonist can open large areas of suitable habitat for another; when the European green crab, *Carcinus maenas*, arrived in San Francisco Bay, its predation upon two species of native clams reduced their densities dramatically and allowed the rapid spread of the eastern gem clam, *Gemma gemma*. Finally, climatic change can also drive changes in habitat suitability, with colonization of the newly suitable areas resulting in geographic shifts in

species' ranges. Northward plant colonization of newly uncovered habitat after withdrawal of glaciers at the end of the last ice age is well documented in the pollen record. In recent decades, changes in intertidal community composition in the northeastern Pacific reveal northward shifts of several southern species into waters that previously were too cold for them. Climate change can also lead to colonization by facilitating dispersal. Low sea levels may have facilitated the spread of humans from Eurasia to the Americas by revealing the Bering land bridge. Understanding interactions between climate change and colonization is an urgent priority for conserving biodiversity and anticipating likely novel community assemblages, given rapid predicted natural and anthropogenic habitat changes in the future.

One of the most dramatic instances of colonization of new habitat occurs not in physical space but in niche space. Organisms can increase the area of suitable habitat by evolving to use new areas. The global influenza pandemic of 1918, recently shown to be derived from avian influenza, is a spectacular example of a virus' evolution to take advantage of an entirely new host, and other deadly human diseases throughout history (including recent times) may have their origin in niche shifts of animal disease. Colonization of a completely new habitat type frequently occurs after colonization in physical space, as evidenced most clearly by adaptive radiation on species-poor islands. The lag between the initial colonization of a new area by an invasive species and the onset of rapid spatial spread is sometimes also attributed to genetic adaptation to the new environment. Genetic founder effects may also play a role in this adaptation process. In any case, niche shifts precipitated by physical colonization can be an important process in allopatric speciation.

See also: Behavioral Ecology: Dispersal–Migration; Competition. Ecological Data Analysis and Modelling: Modeling Dispersal Processes for Ecological Systems; Metapopulation Models. Ecological Processes: Physical Transport Processes in Ecology: Advection, Diffusion, and Dispersion; Succession and Colonization. Evolutionary Ecology: Pioneer Species; Allee Effects; Ecological Niche; r-Strategists/K-Strategists. General Ecology: Succession; Migration and Movement; Seed Dispersal. Terrestrial and Landscape Ecology: Island Biogeography; Spatial Distribution

Further Reading

Amarasekare, P., Possingham, H., 2001. Patch dynamics and metapopulation theory: The case of successional species. Journal of Theoretical Biology 209, 333–344.
Anderson, R.M., May, R.M., 1979. Population biology of infectious diseases: Part I. Nature 280, 361–367.
Caceres, C.E., Soluk, D.A., 2002. Blowing in the wind: A field test of overland dispersal and colonization by aquatic invertebrates. Oecologia 131, 402–408.
D'Antonio, C.M., Vitousek, P.M., 1992. Biological invasions by exotic grasses, the grass fire cycle, and global change. Annual Review of Ecology and Systematics 23, 63–87.
Grime, J.P., 1977. Evidence for existence of three primary strategies in plants and its relevance to ecological and evolutionary theory. American Naturalist 111, 1169–1194.
Grosholz, E.D., 2005. Recent biological invasion may hasten invasional meltdown by accelerating historical introductions. Proceedings of the National Academy of Sciences of the United States of America 102, 1088–1091.
Keeling, M.J., Gilligan, C.A., 2000. Metapopulation dynamics of bubonic plague. Nature 407, 903–906.
Kermack, W.O., McKendrick, A.G., 1927. A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London A 115, 700–721.
Levins, R., 1969. Some demographic and genetic consequences of environmental heterogeneity for biological control. Bulletin of the Entomological Society of America 15, 237–240.
Paine, R.T., 1979. Disaster, catastrophe, and local persistence of the sea palm, *Postelsia palmaeformis*. Science 205, 685–687.
Sagarin, R.D., Barry, J.P., Gilman, S.E., Baxter, C.H., 1999. Climate related changes in an intertidal community over short and long time scales. Ecological Monographs 69, 465–490.
Stanton, M.L., Palmer, T.M., Young, T.P., 2002. Competition–colonization trade-offs in a guild of African acacia-ants. Ecological Monographs 72, 347–363.
Suarez, A.V., Holway, D.A., Case, T.J., 2001. Predicting patterns of spread in biological invasions dominated by jump dispersal: Insights from Argentine ants. Proceedings of the National Academy of Sciences of the United States of America 98, 1095–1100.
Townsend, C.R., Scarsbrook, M.R., Doledee, S., 1997. The intermediate disturbance hypothesis, refugia, and biodiversity in streams. Limnology and Oceanography 42, 938–949.

Dominance and Its Evolution☆

Reinhard Bürger, University of Vienna, Vienna, Austria
Homayoun C Bagheri, University of Zurich, Zurich, Switzerland

Introduction

As Gregor Mendel noted in his seminal paper of 1865, it had been known for some time that hybrids are usually not exactly intermediate between the parental species. Mendel introduced the term "dominant" to refer to characters that are transmitted almost unchanged to the hybrids, and "recessive" to those that become latent. In the meantime, more than a century of genetics has shown that dominance is a ubiquitous phenomenon with most deleterious mutations being recessive. Deleterious mutations are constantly created anew, while selection continually eliminates such mutations from populations. However, mutations can more easily persist and spread through a population, when heterozygote individuals do not (fully) experience the deleterious effects, that is, the mutation is recessive. Dominance is not only ubiquitous, but clearly of great evolutionary importance because it masks the effects of deleterious mutations as long as they are rare and are found predominantly in heterozygous individuals. Mutations are rarely completely dominant or recessive, but usually intermediate, also called codominant. Complete absence of dominance is uncommon, however. The degree of dominance plays an important role in the explanation of several evolutionary phenomena. These include the evolution of sexual reproduction and recombination, the evolution of selfing, Haldane's rule on inviable species hybrids, and the maintenance of genetic variation of metric traits. It is also well established now that dominance is a property of the phenotype and not the gene because genes that affect several traits can be dominant for one trait and recessive for another.

In 1928, R. A. Fisher raised the question why most deleterious mutations are recessive and proposed a model for the evolutionary modification of dominance relations. This was motivated by early experiments by W. Bateson and others which had demonstrated that dominance relations can be modified by changing the genetic background or, later, by selection. Fisher's theory was criticized by Sewall Wright for reasons discussed below. As an alternative, he suggested that the explanation of dominance has a strong physiological component. In agreement with Wright, J. B. S. Haldane held that dominance provides a factor of safety, in that a single dose of the most active allele would, suitably regulated, provide enough of the gene product for normal viability. Fisher's hypothesis led to fierce controversies between proponents of "physiological" and "evolutionary" explanations. In 1981, Kacser and Burns developed a molecular explanation using metabolic control theory which, by many, was taken as evidence that an evolutionary explanation of dominance is obsolete. Still, there are many demonstrated cases of evolution of dominance in natural systems. However, the demonstrated cases of evolution of dominance in natural systems, some of them discussed below, require an explanation. In the meantime, alternative molecular theories, also based on biochemical principles, have been proposed. Some of them provide the potential for evolutionary modification of dominance. Despite its central importance and a huge body of empirical and theoretical work, no generally valid explanation for dominance has been provided until today. In fact, disagreement among scientists about several issues related to its explanation abound.

We shall start with a brief summary of empirical and experimental facts. Then we shall highlight and evaluate the various explanations of dominance (molecular basis of dominance), the theories that have been proposed for the evolution of dominance, as well as some of the main controversies. Finally, we will suggest a possible reconciliation.

Empirical Facts

It has been known for long, and is a generally accepted fact now, that mutations with major (deleterious) fitness effects are almost always recessive. In the simplest case of two alleles, A and a, at a locus, the fitness of the three possible genotypes, AA, Aa, and aa, are usually denoted by 1, $1 - hs$, $1 - s$. Here, s is the selection coefficient and h the dominance coefficient. A mutant is neutral if $s = 0$ and lethal if $s = 1$. If $h = 0$, then A is dominant and a is recessive; if $h = 1$, then a is dominant and A recessive; if $h = {}^1/_2$, one says there is no dominance; if $0 < h < 1$, one speaks of intermediate dominance or codominance; if $h < 0$, then there is overdominance, whereas underdominance prevails if $h > 1$.

Estimates of Dominance

Measuring the coefficient of dominance, h, in particular, for mutations of small effect (s), which constitute the vast majority of all mutations, is difficult. Direct estimates can be obtained from spontaneous mutations in mutation–accumulation experiments of highly inbred lines. Such assays are very laborious. By assuming mutation–selection balance, indirect estimates can be obtained

☆*Change History*: April 2018. B. Wertheim added a single sentence to the first paragraph of the chapter.

This is an update of R. Bürger and H.C. Bagheri, Dominance and Its Evolution, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 945–952.

 doi:10.1016/B978-0-12-409548-9.11201-1

from the structure of segregating populations. These estimates, however, are based on additional assumptions that often cannot be checked and also face statistical challenges. Therefore, they are not very reliable, and the development of methods for measuring dominance coefficients is an active field of research.

All known data demonstrate widespread correlations between dominance and fitness effects of mutations; that is, mutations with a large deleterious effect tend to be more recessive than those with a small effect. This is true for coding genes and, more recently, has also been shown for noncoding genes. Because fitness and dominance effects vary across loci, and the locus at which a mutation occurs can often not be determined, in general, only average coefficients of dominance ($\bar{h}$) and selection ($\bar{s}$) can be inferred. Available data suggest that $\bar{h}$ is about 0.1–0.2 and decays approximately exponentially with increasing s.

Mutations affect fitness through phenotypic traits. In fact, the fitness effect is a consequence of a mutation's effect on one or several traits. If these are closely correlated with fitness, then the fitness effect may be large. Many traits, however, for instance morphological traits, influence fitness only weakly. Dominance effects have been estimated with respect to traits that can be measured on a metric scale. For such metric traits (e.g., body size, oil content in maize, or abdominal bristle number in Drosophila, the latter being extremely well studied genetically), average dominance coefficients appear to be close to 0.5 and depend only weakly on the size of the mutational effect. In particular, no significant skew has been observed.

Compared with intermediate dominance, over- and underdominance are rare. The probably best-known example for overdominance is sickle-cell anemia, which provides immunity against malaria in heterozygous state. Over- and underdominance often seem to be the consequence of frequency-dependent selection or spatially heterogeneous selection. The first occurs if the heterozygotes have an overall advantage, for instance because they are generalists and can use most or all available resources. Underdominance can easily occur by migration–selection balance if each of the two homozygotes is optimally adapted to a different ecological niche and heterozygotes are maladapted because they do not have their own niche.

Modification of Dominance

Numerous experiments have shown that dominance relations can be changed experimentally by substitution of the genetic background or by selection. Therefore, certain genes have the ability to modify the character expression, and such genes seem to be abundant.

Most empirical evidence for evolutionary modification of dominance concerns wing patterns in lepidopteran species, mimicry in butterflies, and pesticide resistance. A classical example is provided by industrial melanism in the peppered moth *Biston betularia*, a lepidopteran species. During the second half of the 19th century, dark forms of this usually light-colored insect became more common in industrial areas and within about 50 years nearly replaced the pale form. One of the reasons why the melanic form gained a selective advantage was that the tree trunks on which these butterflies rested became blackened due to pollution which killed light lichen, and the pale form, originally cryptically colored, became better visible to their predators. Interestingly, the dark form became dominant during this process. Extensive experiments by H. B. D. Kettlewell in the 1950s and 1960s suggest that selection for an unlinked modifier occurred during the spread of the favorable mutant. Later, similar experiments were performed in related species with slightly different outcomes, which may be explainable by linked modifiers that, initially, had a different frequency as in *B. betularia*. Since the ecological setting is complex, other selective forces as well as migration–selection balance may also have played an important role. In all of these cases it seems that an originally recessive or intermediate allele became more or less dominant.

A quite different situation was observed in some cases of mimicry by P. M. Shepard, C. A. Clarke, and others in the 1950s and 1960s. There, frequency-dependent selection maintains overdominance, and one of the homozygous phenotypes is modified to resemble the heteroyzgotes.

Another well-studied class of examples for dominance evolution concerns insecticide-resistance genes. Often the genes causing resistance, and sometimes even those causing dominance, are known. Since pesticide applications are heterogeneous in space, and alleles conferring an advantage in treated areas tend to be disadvantageous in untreated areas, underdominance is often maintained by migration–selection balance.

Explanations of Dominance and Its Evolution

Basic Scenarios for the Evolution of Dominance

Almost all population-genetic explanations for the evolution of dominance are based on scenarios involving selection. Distilled down to its simplest form, the common idea in these models is the notion that, given a set of homozygous and heterozygous genotypic variants, one or more of the less-fit variants evolve to resemble the fitter variant at the phenotypic level. Consider a starting condition in which the phenotypes associated with a pair of allele variants are codominant. The original scenario proposed by Fisher was one in which the fitness landscape associated with the three possible genotypes mimics the qualitative shape of the phenotypic landscape. Hence one of the homozygous variants is fitter than its alternate, and the heterozygote is intermediate in fitness. **Fig. 1**A depicts such a scenario. In the latter case, dominance could evolve if the phenotype of the heterozygote is modified (via selection) to resemble the fitter homozygote. This would lead to dominance of the fitter phenotype, and a form of robustness with respect to mutations, since now the effects of the less-desirable allele are masked in the heterozygotes.

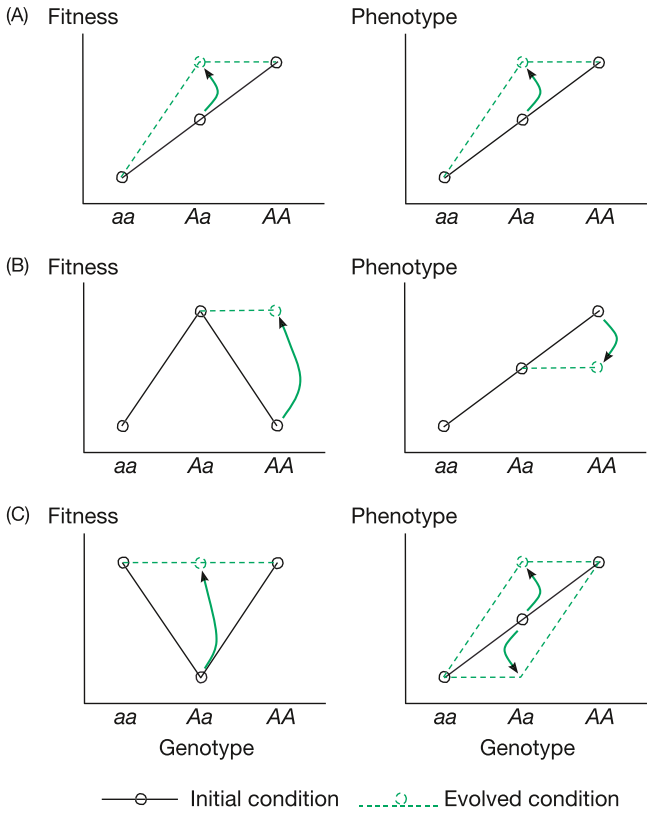


Fig. 1 Idealized representation of some situations under which dominance could evolve. (A) Superior homozygote as initial condition, (B) heterozygote superiority as initial condition, (C) heterozygote inferiority as initial condition.

The noteworthy concept here is that a "phenotypic" landscape that is described as function of the genotype at a single locus is not necessarily fixed. Consequently, fitness in relation to a single-locus genotype is also not necessarily fixed. In this scheme, the question that remains is what are the causes by which the relation between genotype and phenotype can change. One possibility is a change in the environment. A second is a change in the genetic background. The latter case is usually the main protagonist in models of dominance evolution. In the simplest case, changes in the genetic background can be represented as changes in alleles at a "modifier locus,", which in turn modify dominance relations with respect to alleles at a "primary locus." All such modifier models and their more complicated derivatives inherently assume that the relation between genes at separate loci and the phenotype are nonadditive (i.e., nonindependent) and that the latter relation can be modified by gene interactions. Nonadditivity, whereby allele substitutions at one locus can alter the phenotypic effects of substitutions at another locus, is referred to as "epistasis." Models of dominance evolution via changes in the genetic background inherently assume the existence of epistasis.

If mutational effects at one locus can be modified by substitutions at another, then the general scheme presented in Fig. 1A is not the only scenario by which dominance could evolve. Two commonly considered scenarios are presented in Fig. 1B and C. Both rely on balanced polymorphisms as the starting condition. One possible scenario is when the heterozygote is superior in fitness to the homozygotes (Fig. 1B). Dominance would evolve under such a condition if the phenotype of one of the homozygotes were modified to resemble the fitter heterozygote. The reverse starting condition can be also considered, whereupon either homozygote is superior to the heterozygote (Fig. 1C). In the latter case, dominance would evolve if the heterozygote were modified to resemble either one of the homozygotes. The latter starting condition is more complicated from the population-genetic perspective, since the maintenance of a balanced polymorphism involving two fit genotypes depends on frequency-dependent selection.

The situations presented in Fig. 1 deal with the possibility of dominance modification due to selection—provided that a physiological trait of interest can be modified, and that it can be subject to selection. In the next section, we consider some of the physiological grounds for the manifestation of dominance modification. Whether population conditions allow for selection to be an effective force is itself a decisive matter. This will be discussed in a subsequent section.

Mechanistic Explanations for Dominance: The Case of Metabolism

During the early years of the 20th century, one of the more prominent explanations for dominance, advocated by W. Bateson, R. C. Punnett, and others, was the "presence-and-absence" hypothesis. Consider a situation in which a phenotype or trait comes about as a result of the action of a set of gene products. One can conceive of a situation in which each gene i plays a role X_i for the

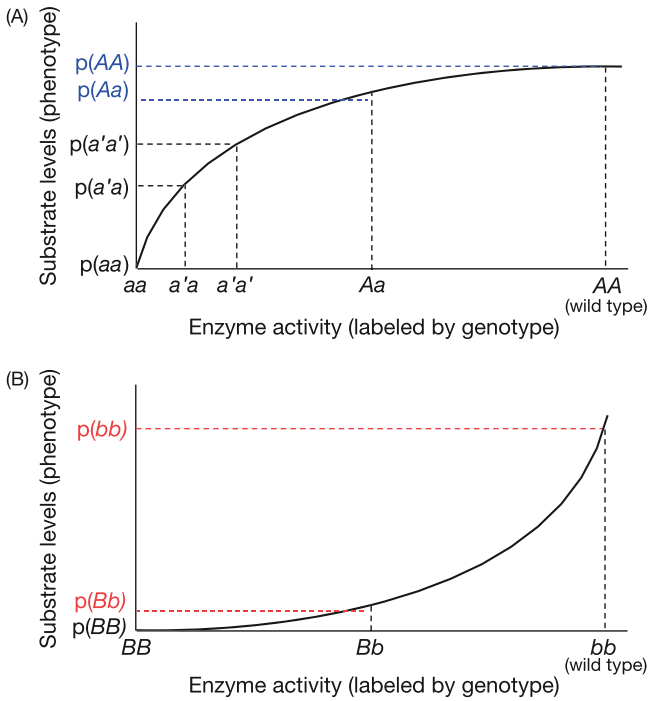


Fig. 2 Physiological model for dominance based on enzyme activity and substrate accumulation. (A) Dominant wild type (Wright's physiological model), (B) recessive wild type (hypothetical model).

formation of a given phenotype *P*. The presence–absence hypothesis assumes that as long as there is one functional copy of gene *i*, then the role X_i is satisfied, and hence *P* remains at wild-type levels. With advances in physiological genetics, and subsequent advances in molecular biology, the presence–absence hypothesis has been abandoned. Nonetheless, we should note that the prime weakness in the presence–absence hypothesis was not necessarily its incomplete representation of the underlying mechanisms; all mechanistic representations of genotype–phenotype relations are at some level incomplete. By the standards of today, the presence–absence hypothesis is more a logical argument than a detailed exposition on mechanism. However, its major failure was that as a phenomenological description, it failed to fit much of the mounting experimental evidence; it was not general enough to encompass many exceptions.

Given present knowledge of the variety of molecular processes occurring in the cell, such as gene regulation, signal transduction, and metabolism, it would now seem unreasonable to use any specific mechanistic model as a general model for dominance modification. Depending on the underlying causes of a given phenotypic trait, the proximal causes for dominance modification and the constraints placed upon it can be very different from one case to the next. Hence, any given molecular explanation may serve as an instantiation of general concepts pertaining to dominance modification, but it is unlikely to serve as the grand model for modification. Nonetheless, if any given modification model is sufficiently well-developed, it can be used to test some general evolutionary questions.

The case study that has been most central to addressing dominance evolution revolves around phenotypes that are directly dependent on the functioning of metabolic enzymes. Shortly after Fisher proposed his hypothesis, both Wright and Haldane argued that the proximal causes for the manifestation of dominance would have to depend on the underlying physiology. At the time, enzymes and their biochemistry were at the forefront of mechanistic explanations for the inner workings of organisms. Accordingly, they also had a central role in Wright and Haldane's explanations. With the assumption that most genes code for enzymes, and the use of a simplified nonlinear model of enzyme pathways, Wright tried to address dominance from a physiological perspective. Based on his model, he hypothesized that the rate of substrate formation throughout a pathway will have a diminishing returns relation to enzyme concentrations (**Fig. 2**A). Under such circumstances, when enzyme concentrations are in the flat region of the curve, then reductions in enzyme concentration will have relatively small effects on substrate levels. Accordingly, let us assume that a wild-type homozygote *AA* has two copies of a gene coding for an enzyme *E*, while the mutant heterozygote *Aa* has one functional copy. Furthermore, let us assume that enzyme concentrations are proportional to copy numbers, and hence that the concentration of functional enzyme *E*, and consequently enzyme activity, is halved in *Aa* genotypes in comparison to *AA*. If the *AA* genotype lies deep in the plateau region of the curve in **Fig. 2**A, then the reduction of enzyme concentrations in *Aa* genotypes to half of *AA* would not have an appreciable effect on phenotype, and the wild-type phenotype associated with *AA* would be considered dominant. In Wright's model, dominance of the wild type results from the convex shape of the relation between enzyme activity and phenotype. On the other hand, a concave curve such as the one in **Fig. 2**B would lead to the wild type being recessive. In the context of **Fig. 2**A, dominance of the wild type would evolve when wild-type enzyme

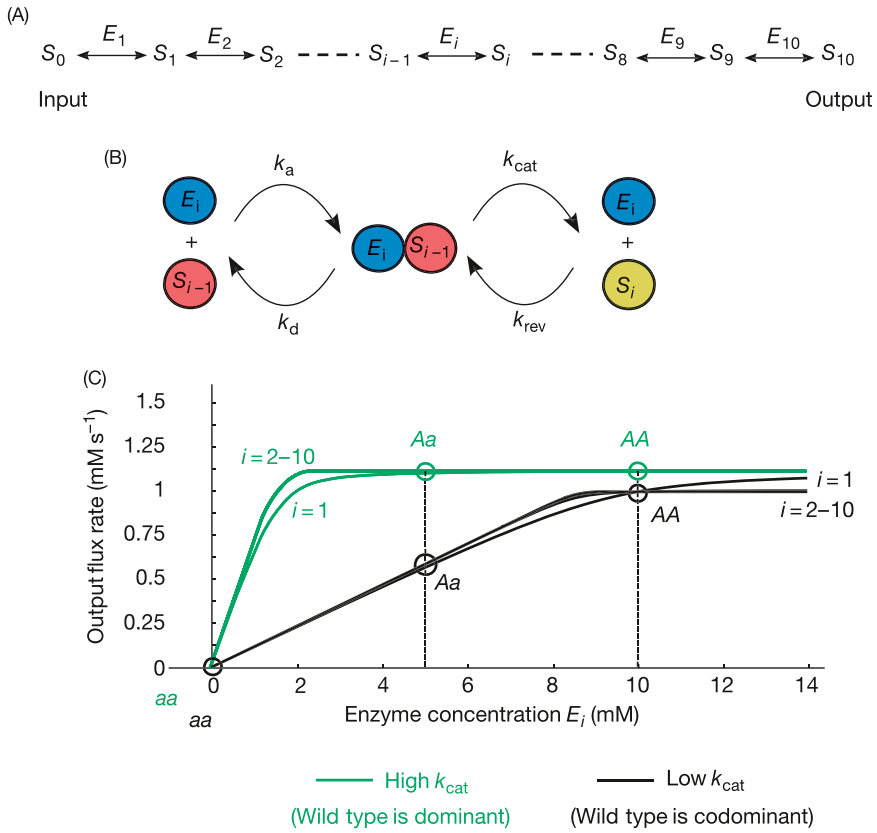


Fig. 3 Model of dominance modification for the flux phenotype of a 10-enzyme pathway. (A) Ten-enzyme sequential pathway, (B) details of each enzyme-catalyzed reaction, (C) flux phenotype of pathway as a function of enzyme concentrations.

activity is pushed to the flat end of the curve. Conversely, as exemplified by the $a'a'$ homozygote in Fig. 2A, a phenotype that is associated with the steeper part of the curve will tend toward codominance with respect to the null *aa* homozygote.

The physiological model discussed above ignores compensatory effects such as gene regulation, feedback, redundancy, and nonsequential pathway topologies. Nonetheless, as a model, it is a plausible starting point for addressing questions related to dominance and its evolutionary origins; in fact, it has been central to research on this topic. An important question that arises within this model, is the reason whereby wild-type enzyme activities should be at the high end of the curve. Haldane's suggestion was that organisms could evolve a "factor of safety" against perturbations. Such perturbations could in principle be either environmental or genetic. Another explanation could be that high enzyme activities are simply required for individual fitness, and because of the convex shape of the curve, dominance evolves as a side effect of selection for high enzyme activity.

As an alternative explanation for dominance, Kacser and Burns suggested that in metabolic systems, one could explain dominance without making recourse to evolutionary explanations. Their argument is based on models that are at first sight similar to Wright's model. However, there is a fundamental difference. In Wright's model, levels of dominance can be modified by changing the enzyme activity associated with the "wild-type" homozygote. The Kacser and Burns model (from here on abbreviated as the KB model) places a crucial constraint on this modification. In the latter case, dominance can be modified at any given locus in the same sense as in Wright's model. However, the KB model also suggests that modifications that result in increased phenotypic sensitivity to any enzyme are compensated by decreased sensitivity to other enzymes. Furthermore, the model also implies that the more enzymes are involved in a pathway, the smaller the average effect of each enzyme on the phenotype. Based on these premises, Kacser and Burns concluded that dominance—which depends on insensitivity of the phenotype to enzyme concentration changes—is a de facto property of metabolic pathways and that evolutionary explanations based on dominance modification are not necessary.

Many of the mechanistic details of Wright's model do not reflect what is known of enzyme kinetics today. However, his model does preserve some of the nonlinear aspects of chemical transformations. On the other hand, the KB model is more akin to modern conceptions of enzyme kinetics. However, the model contains some approximations that eliminate important nonlinearities in enzyme kinetics—namely, enzyme saturation. The consequence is that the scope of the KB model is much more limited than originally intended and, in hindsight, not a good candidate for investigating the evolutionary origin of dominance. For example, Bagheri and Wagner have suggested that if one considers simple models of enzyme kinetics, with the possibility of enzyme saturation, one finds that system robustness or dominance is not a de facto property of enzyme systems. In particular, the compensatory constraints on dominance modification that are postulated in the KB model are eliminated when the system allows

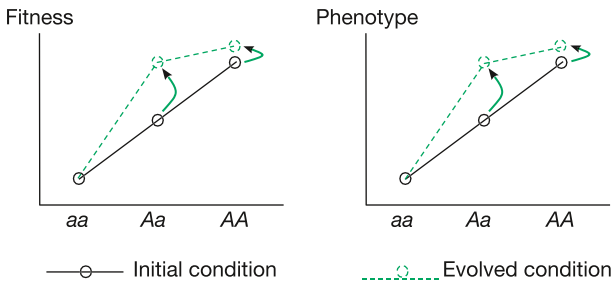


Fig. 4 Example of dominance evolution due to a modifier with multiple effects.

for saturation. In fact, the modulation of enzyme saturation levels allows for dominance modification, and hence, dominance evolution. **Fig. 3** shows an example of a 10-enzyme pathway, which contrary to the KB model is modeled with the possibility of enzyme saturation. The pathway consists of a constant input S_0 and an output substrate S_{10}, whose rate of production (flux, dS_{10}/dt) is the phenotype of interest (**Fig. 3**A). As indicated by **Fig. 3**B, each enzyme-catalyzed reaction can be conceptualized as a process by which a substrate and an enzyme join to form an enzyme–substrate complex, which subsequently dissociates into a product and the unaltered enzyme. Saturation levels can be altered by mutations that either change concentration levels or the dissociation constant k_{cat}. Flux for such a model can be obtained by numerical integration of the resulting system of differential equations. **Fig. 3**C shows the flux function for two realizations of a 10-enzyme pathway. The green lines trace a pathway whose enzymes possess relatively high k_{cat} values. The black lines trace a pathway with low k_{cat} values. Within each case, flux responses to changes in enzymes 2–10 are very similar, and the corresponding curves are almost indistinguishable. Responses to enzyme 1 are slightly more sensitive. What is important to note in this figure is that we are observing two different endpoints of possible evolutionary trajectories. In the low k_{cat} case, the wild-type phenotype is codominant with respect to mutations that halve the concentration of "any" enzyme. In contrast, in the high k_{cat} case, the wild-type phenotype is dominant with respect to mutations that halve the concentration of any enzyme. Hence, one can conceptualize dominance evolution as an evolutionary process whereby k_{cat} levels have been gradually modified from low to high values. High k_{cat} values result in low enzyme saturation levels, and consequently, higher dominance levels for the wild type.

A further consequence of the model in **Fig. 3**C is the important observation that there is a slight difference between the wild-type *AA* phenotypes of the high and low k_{cat} pathways. This means that if one is selecting for the high-flux phenotype, one can select modifiers due to their direct fitness effects rather than their dominance modification effects. Direct effects can be relatively small, but they are not dependent on the frequency of *a* alleles. Meanwhile, modifying effects can be larger, but nonetheless they occur less frequently due to their dependence on the frequency of *a* alleles. Hence, selection for modifiers with multiple effects would circumvent some of the frequency sensitivity problems associated with modifier evolution. A schematic drawing of such a scenario is shown in **Fig. 4**, which corresponds to the enzyme kinetic model in **Fig. 3**. Dominance evolution in such a scenario is more likely to be successful than in cases where modifiers only have a dominance modification effect (e.g., **Fig. 1**A). Interestingly, despite the lack of an explicit enzyme kinetic representation, the mathematical properties of Wright's model are similar to the enzyme kinetic model presented here; namely, Wright's model allows for the possibility of dominance evolution through multiple effects, while the KB model does not.

Evolutionary Explanations for Dominance

As discussed above, dominance relations are not necessarily fixed but can be modified and may even evolve. Fisher's basic argument was the following. At mutation–selection balance, deleterious alleles are kept at low frequency in a population but, nevertheless, reduce its mean fitness. Therefore, genes that protect the population from the effects of recurrent deleterious mutations should be selectively favored and increase in frequency. Since most rare alleles occur in heterozygotes, modification of the heterozygous phenotype would increase mean fitness. Wright introduced a model that formalized Fisher's proposition and has become the archetype for the evolutionary modification of dominance. It assumes a primary locus at which a wild-type allele, *A*, and a deleterious mutant allele, *a*, occur. In addition, it assumes a so-called modifier locus, with alleles *M* and *m*, where *M* modifies the fitness effects of heterozygous genotypes as follows:

$$\begin{array}{cccc} & AA & Aa & aa \\ MM & 1 & 1 & 1-s \\ Mm & 1 & 1-ks & 1-s \\ mm & 1 & 1-hs & 1-s \end{array} \qquad (1)$$

Here, the intensity of selection against *a* is $s > 0$, and the dominance coefficients satisfy $0 \le k \le h \le 1$ and $h \neq 0$. Mutation is assumed to occur from allele *A* to allele *a* at rate *u*, and recombination between the two loci occurs with frequency *r*, with $0 \le r \le {}^1/_2$. Fisher, and in order to refute him, also Wright, assumed that evolution commences at mutation–selection balance between *A* and *a* when a new allele *M* is introduced at the modifier locus at low frequency. At this initial state, with fitnesses given by the last line of (1), the mean fitness of the population is approximately $1 - 2u$. If *M* became fixed, which was not proved by Fisher or Wright but only much later, the

first line in (1) applies, and the mean fitness increased to about $1 - u$. This model is in line with the graphical representation in Fig. 1A. Several more elaborate analyses in the 1960s and 1970s corroborated and extended Wright's analysis. Given that per-locus mutation rates are very small, typically of the order of 10^{-6} to 10^{-5}, this is a very small fitness increase. Whereas Fisher considered this to be sufficient, Wright argued for the contrary because, in his opinion, genetic drift could easily overcome this tiny fitness advantage and probably cause the loss of the modifier. In addition, the slightest disadvantage of the modifier would nullify its favorable effect. However, Fisher adhered to his proposition because he considered population sizes to be very large, of the order of 10^9 or higher. Hence, in contrast to Wright, genetic drift was only of minor relevance to him. Today, we know that so-called effective population sizes (this is what matters for the effects of genetic drift) are much lower than actual population size, and for most higher organisms (even species of Drosophila) less than 10^6. For these and other reasons, Fisher's proposition is now generally considered to be untenable.

Several other hypotheses, usually also formalized in terms of modifier models, and mostly in line with one of the mechanisms depicted in Fig. 1B or C, have been advanced to explain various cases of dominance evolution. If, for instance, as in Fig. 1B, a homozygous phenotype is changed to resemble the advantageous heterozygote, then the potential fitness increase is substantial and there are no theoretical obstacles to such an occurrence. Several cases of mimicry fall into this category. Similarly, if, as in several cases of pesticide resistance, underdominance is maintained by spatially heterogeneous selection, heterozygotes are frequent and their fitness disadvantage can be reduced by modifiers. In general, whenever there is a balanced polymorphism, as in Fig. 1B or C, there are many heterozygotes present and evolution of dominance can readily occur.

An interesting hypothesis, in line with the formal model (1), was considered by Haldane in 1956. He suggested that when a gene was sweeping through a population as a result of natural selection, heterozygotes would be very frequent, and this would provide an opportunity for intense selection for modification of the heterozygote. A case in point seems to be industrial melanism in moths and butterflies. Apparently, selection of the modifier has occurred during the spread of a favorable mutant. This requires that, when the favorable mutant arises, or a rare mutant becomes favorable through a change in the environment, modifiers are already present in the population. During such a process, heterozygotes become very frequent, and a modifier can rise to high frequency and become fixed. It was proved by Bürger and Wagner in the early 1980s that this mechanism indeed works. It also helps to explain several obstacles raised by backcrossing experiments in various cases of industrial melanism. Another interpretation of this process, put forward by G. S. Mani in 1980, maintains that heterozygotes have been kept at high frequency by migration–selection balance, a scenario studied in more detail by S. Otto and D. Bourguet in 1998.

Despite many well-established cases of evolution of dominance and despite several mathematical analyses that have demonstrated that dominance evolution can occur under a number of scenarios (but not under Fisher's original one), evolutionary explanations have been largely dismissed and are ignored by most textbooks.

There are several reasons why evolutionary explanations of dominance are considered to be irrelevant or even wrong. One is that Fisher's original proposition is untenable and many seem to believe that this applies to any evolutionary explanation of dominance without realizing how different they can be. A second and related reason is, as first pointed out by Charlesworth, the strong inverse correlation between the deleteriousness of a mutant and its degree of dominance. Such a correlation cannot, and was not, predicted from Fisher's theory. If, however, evolution occurs as in Haldane's model and the population is finite, the situation changes. A third, very important reason, is that Kacser and Burns' metabolic theory has been widely accepted, and thus an evolutionary explanation of dominance had seemed unnecessary. Finally, Orr showed that in "artificial" diploids of the normally haploid alga Chlamydomonas, wild types are about as often dominant as in diploid species. He claimed that this finding falsifies Fisher's theory. It, however, only shows that dominance can occur without evolution. Moreover, as discussed by Saved and Mayo, a similar observation has already been made in 1947 by D. Lewis.

What is clear by now is that, even if the theory of Kacser and Burns is eventually replaced by a more general or different theory, molecular pathways can often generate dominance. Therefore, a priori, dominance does not necessarily require an evolutionary explanation. However, it should also be noted that molecular pathways can allow for dominance modification, and hence dominance evolution. Furthermore, and this must not be ignored, many instances of evolution of dominance have been demonstrated, noteworthily in well-understood ecological settings, and these require an evolutionary explanation.

We close this discussion by pointing out that evolution of dominance, by whatever mechanism, is a special case of evolution of robustness, and modifier models are conceptually the simplest, and oldest, models to study it. More generally, epistatic interactions among loci can lead to the evolution of robustness and of genetic architecture, for instance, because new mutants can change the effects of current as well as future (mutant) alleles at other loci. In recent times, this has become an active field of research, and due to the complexity of the problems involved, is likely to remain so in the coming years.

See also: Evolutionary Ecology: Genetic Drift; Units of Selection; Natural Selection; Fitness; Hardy–Weinberg Equilibrium; Hardy–Weinberg Equilibrium

Further Reading

Bagheri, H.C., 2006. Unresolved boundaries of evolutionary theory and the question of how inheritance systems evolve: 75 years of debate on the evolution of dominance. Journal of Experimental Zoology 306B, 329–359.
Bagheri, H.C., Wagner, G.P., 2004. Evolution of dominance in metabolic pathways. Genetics 168, 1713–1735.

Bourguet, D., 1999. The evolution of dominance. Heredity 83, 1–4.
Bourguet, D., Genissel, A., Raymond, M., 2000. Insecticide resistance and dominance levels. Journal of Economic Entomology 93, 1588–1595.
Charlesworth, B., 1979. Evidence against Fisher's theory of dominance. Nature 278, 848–849.
Deng, H.-W., Gao, G., Li, J.-L., 2002. Estimation of deleterious genomic mutation parameters in natural populations by accounting for variable mutation effects across loci. Genetics 162, 1487–1500.
Fernández, B., García-Dorado, A., Caballero, A., 2005. The effect of antagonistic pleiotropy on the estimation of the average coefficient of dominance of deleterious mutations. Genetics 171, 2097–2112.
Kacser, H., Burns, J.A., 1981. The molecular basis of dominance. Genetics 97, 639–666.
Mayo, O., Bürger, R., 1997. The evolution of dominance: A theory whose time as passed? Biological Reviews 72, 97–110.
Orr, A.H., 1991. A test of Fisher's theory of dominance. Proceedings of the National Academy of Sciences of the United States of America 88, 11413–11415.
Phadnis, N., Fry, J., 2005. Widespread correlations between dominance and homozygous effects of mutations: Implications for theories of dominance. Genetics 171, 385–392.
Savageau, M.A., 1992. Dominance according to metabolic control analysis: Major achievement or house of cards? Journal of Theoretical Biology 154, 131–136.
Sved, J.A., Mayo, O., 1970. The evolution of dominance. In: Kojima, K. (Ed.), Mathematical topics in quantitative genetics. Berlin: Springer, pp. 289–316.
Wagner, G.P., Bürger, R., 1985. On the evolution of dominance modifiers. Part II: A non-equilibrium approach to the evolution of genetic systems. Journal of Theoretical Biology 113, 475–500.
Wright, S., 1934. Physiological and evolutionary theories of dominance. American Naturalist 68, 25–53.
Zhang, X.-S., Wang, J., Hill, W.G., 2003. Influence of dominance, leptokurtosis and pleiotropy of deleterious mutations on quantitative genetic variation at mutation–selection balance. Genetics 166, 597–610.

Eco-Evolutionary Dynamics

Gabriel Pigeon and Fanie Pelletier, Université de Sherbrooke, Sherbrooke, QC, Canada

Glossary

Animal model A mixed effect model (i.e., a form of linear regression in which the explanatory terms are a mixture of both "fixed" and "random" effects) where one of the random effects of interest is the additive genetic value of individual animals. It is used to partition plastic from genetic part of traits.

Co-evolution Evolution of two or more species having a close ecological relationship, where both species adapt to the evolutionary changes occurring in the other species, thereby affecting each other's evolution.

Contemporary evolution Evolutionary changes observable over less than a few hundred years. The concept is also sometimes referred to as microevolution or rapid evolution or evolution on short time-scale.

Eco-evolutionary feedbacks The reciprocal interactions between the ecology of populations, communities, and ecosystems.

Evolution Change in allele frequencies in a population over time.

Evolvability The ability to respond to selection.

Extended phenotype All the effects a gene has on the outside world that may influence its chances of being replicated. These can include effects on the organism in which the gene resides, the environment, or other organisms.

Functional traits Any morpho-physio-phenological traits which impact fitness indirectly via its effects on performance traits.

Phenotypic plasticity Change in the average phenotype expressed by a genotype in different environments.

Standing genetic variation Allelic variation that is currently segregating within a population, as opposed to alleles that appear by new mutation events.

The classic view is that natural selection determines which phenotypes persist or go extinct, over a long time-scale while population growth, community, and ecosystem processes are mostly driven by ecological factors such as density-dependent competition and stochastic factors on a shorter time-scale. Thus, micro-evolutionary changes are assumed to occur on an evolutionary time-scale and scientists have typically ignored the potential feedback of evolution on ecological processes. However, the realization that selection can be strong and that evolutionary changes can occur over ecological time-scale (also termed contemporary time-scale) has led researchers to ask a very important question: can evolutionary changes feedback on ecological dynamics? Several studies have shown that evolutionary processes can have quantifiable effects on ecological dynamics. Under some circumstances, evolution can occur within a few generations, and it is therefore necessary to consider the possibility that evolutionary changes may feedback on ecological processes and vice versa. Eco-evolutionary dynamics studies explore both the unidirectional effects of ecological changes on evolutionary processes (often referred to as the ECO to EVO links) and evolutionary changes on ecological processes (often referred to as the EVO to ECO link), and, even more importantly, they document the bidirectional eco-evolutionary feedbacks on contemporary time-scale. Considering that the ECO to EVO links have been recognized for decades and deeply investigated in evolutionary ecology, the main contribution of eco-evolutionary dynamics is to bring insights on the less explored EVO to ECO links and to document feedbacks between these processes over a contemporary time-scale.

Natural Selection

All ecological and evolutionary changes ultimately occur as a result of alterations to the birth and death patterns of individuals. As the changes in birth and death are often associated with specific phenotypes (selection) and because those phenotypes generally have a genetic basis, evolutionary responses are expected. Assuming no other changes, it is expected that an evolutionary response should improve reproduction and survival of the next generation. Changes in phenotypes, however, are not always synonymous with evolutionary changes. Indeed, phenotypic plasticity may result in trait changes even in the absence of any evolutionary changes. Furthermore, plasticity itself can evolve and most studies to date have not been able to partition the ecological effects of trait changes into their plastic and evolutionary components. While the distinction between genetic variation and phenotypic plasticity is important to understand phenotypic variation and to identify whether the phenomena of interest is an ECO to ECO link or an EVO to ECO one, the effect of the changes in traits on the ecological processes of interest will be due to changes in the mean phenotype. Because it is the phenotype of individuals that will influence interactions between individuals or the effect of trait changes on the environment, eco-evolutionary dynamics studies often take a phenotypic perspective and assume that part of the trait changes has a heritable component.

 doi:10.1016/B978-0-12-409548-9.10548-2

The main drivers of natural selection are environmental. Those drivers are very diverse and include changes in predation pressure, density, sex ratio, weather, etc. One of the classical examples that have documented natural selection in the wild is the study on the effects of drought on the beak size of Darwin's finches (*Geospiza fortis*). Another classical example comes from the peppered moth (*Biston betularia*), where the incidence of the dark form increased due to industrialization during the 19th century. Lichen cover had decreased due to industrialization, making the dark form more cryptic and less predated on than the normal form. This selective advantage led to a change in frequency of the dark form, a change that was later shown to be genetic. This change in frequency of the dark form compared to the light one is thus one of the clearest and most intuitive examples of evolution under natural selection. A dynamic vision of the impact of ecology on evolution is therefore necessary, and all studies investigating the ecological drivers of natural selection on short time-scale document the ECO to EVO links of eco-evolutionary dynamics.

Contemporary Evolution

As previously mentioned, evolution was traditionally considered to be very slow, occurring over millions of years. For example, Darwin believed that evolution by natural selection could not allow for rapid changes, since it operated only by taking advantage from slight small successive variations. Experiments from animal breeding, however, have taught us that if selection is strong, evolutionary responses can occur over only a few generations. For instance, evolutionary changes in the resistance of certain breeds of cattle (Belmont Adaptaur; *Bos taurus*) to ectoparasite resistance (tick and worm) have been produced in as little as 15 years of selection by breeders.

Another important step in documenting the ECO to EVO links of eco-evolutionary dynamics is to evaluate whether evolutionary responses to natural selection occur and to identify the circumstances under which these responses are likely to occur on a contemporary time-scale. Several empirical studies have indeed revealed that even in a natural context, traits can also evolve over a relatively short time period in response to changing environments. Cases of contemporary evolution were also observed in humans. A study of a pre-industrial human population reported evolutionary changes in the age at first reproduction in less than two centuries. These evolutionary changes in phenotype occurred on similar time-scale as ecological processes. Thus, interactions between evolutionary changes and ecological processes are possible. Beyond that realization, however, a critical question is not whether such reciprocal interactions are possible but whether they are strong enough to matter in our understanding of ecological processes. For example, are these changes large enough to change the fate of a population, alter the composition of a community, or to a larger extent, even slow down the nutrient cycles?

From Evolution to Ecology (The EVO to ECO Link)

As just exemplified in previous sections, the clear and rapid effect of ecological change on evolution of different phenotypes is well established. The consequence of these phenotypes changes on various levels of biological organizations, however, is still under investigation. The phenotype of an individual can shape its interactions with its biotic and abiotic environment: its competitiveness, its ability to acquire resources, its ability to survive, grow and reproduce. By extension, the phenotypes present in a population will determine the possible interactions this population can have with its environment: its growth or decline, its coexistence with competing species or with predators, its geographical range. These changes in the community can further cascade and cause changes in the properties of the ecosystem. In response to ecological changes, evolutionary changes in traits can hence have a profound impact on the ecological landscape. The ecological processes affected can be on several levels of organization, ranging from the population to the ecosystem level (**Fig. 1**). The expectation is that effects should decrease with higher hierarchical levels as the effects of the evolutionary changes are buffered at each level. However, direct links and interactions between levels may also magnify the effects. Research on the impact of evolutionary changes on each of these organizational levels is starting to emerge, with studies at the population and community levels being most abundant.

Population Level Eco-Evolutionary Dynamics

One aspect of evolutionary dynamics is aiming to document the links between changes in the distribution of phenotypic traits generated by selection and population dynamics (**Fig. 1**; yellow box). As previously mentioned, natural selection and population dynamics are linked by birth and death. It is therefore necessary to consider the possibility of an eco-evolutionary feedback for traits linked to demography. When the trait change also causes changes in the fitness landscape, feedback loops can occur. The best example would be in the case of a trait under density-dependent selection. Density-dependent selection may favor the evolution of traits more closely linked to competitive ability at high densities than at low densities. These competitive traits, however, are often costly, yielding increased mortality, or reduced growth and fecundity. Consider a hypothetical example where the level of aggressiveness was selected differently according to density. When density is high, competition for resources increases and so does the optimal level of aggressiveness. In this case, the fitness landscape may resemble **Fig. 2A**, with an optimal fitness with high aggressiveness at high density, but low aggressiveness being favored at low densities. When density is high, directional selection,

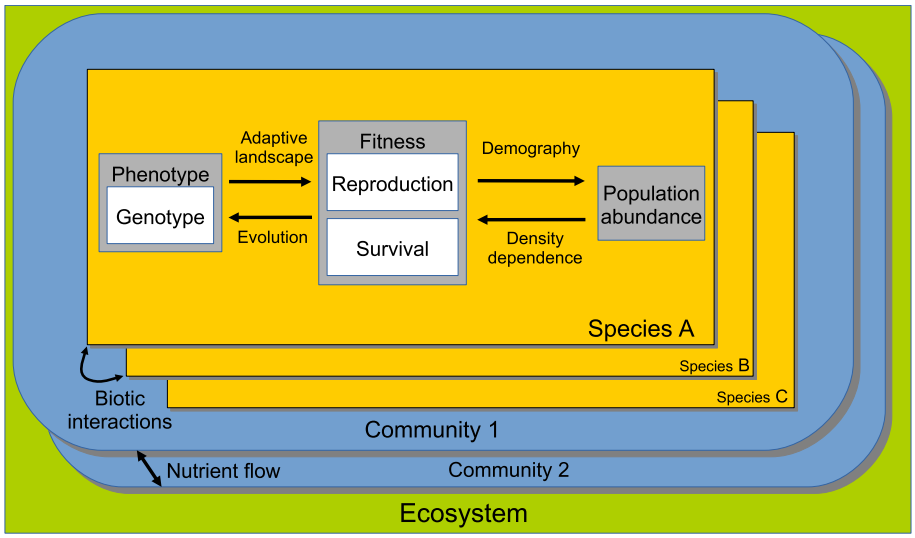


Fig. 1 Conceptual model of eco-evolutionary dynamics, demonstrating the potential effects of changes in genotype on higher levels of ecological process going from population, community to ecosystem levels of organization. *Yellow* boxes represent processes within the species level, where phenotype is based on genotype, which affects fitness through the fitness landscape. Fitness is determined by reproduction and survival which influence population abundance. Population abundance can feedback to genotype via density abundance, which changes the selective pressure, and leads to evolution. Different species interact within a community (*Blue box*). These interactions can also be based on the genotype of species (not shown on the figure). Communities are linked within the ecosystem (*green box*) via subsidies and nutrient flow.

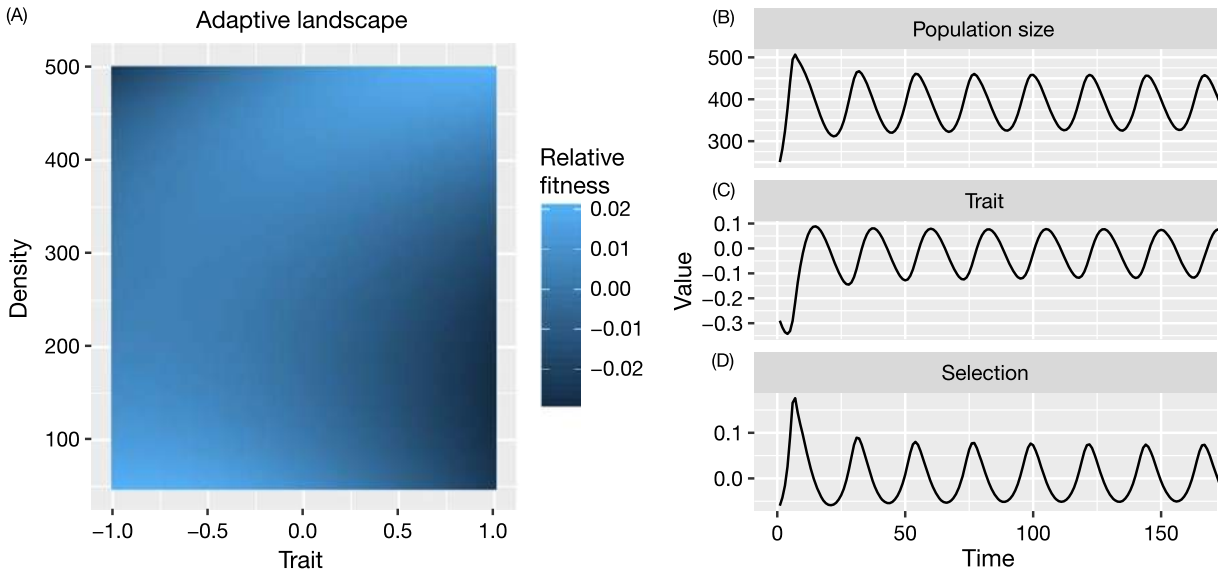


Fig. 2 Simulated example of eco-evolutionary feedback loops at the population level. (A) Fitness landscape where relative fitness of the trait of interest depends on density. The *right panel* shows the resulting cyclic fluctuations in (B) abundance, (C) mean trait value within the population, and (D) selection differential.

combined with phenotypic variability and heritability, may lead to evolutionary changes toward more and more aggressive individuals. If aggressiveness is negatively correlated with population growth rate (e.g., because more aggressive individuals have a lower life expectancy or ability to raise offspring), changes in this trait could lead to a decrease in population growth important enough to lower density. This population level change will, in turn, generate a change in the selective pressures (**Fig. 2D**) and will favor a return to less aggressive phenotypes with lower mortality. If the demographic effects of trait changes are faster than the genotypic changes, which is likely, the evolutionary lag can induce cycles in the population dynamics with the density and mean phenotype alternating between states of high density and aggressiveness and low density and aggressiveness (**Fig. 2B** and **C**). Evidence of such dynamics has been found in nature in the side-blotched lizards (*Uta stansburiana*), where reproductive strategies

whose fitness consequences were density-dependent effects induced cycles alternating between morphs with few competitive offspring and morphs with numerous less competitive offspring (Sinervo *et al.*, 2000).

Additional evidence that traits tightly linked to fitness may lead to EVO to ECO effects come from evolutionary changes in life-history traits. For example, size-selective harvesting, targeting large and older individuals, has led to a reduction in the age at maturity of some heavily fished species. Such evolutionary responses favor r-like life-history strategies which, in turn, can result in increased population growth rate as a larger proportion of the population produces recruits. Similarly, selection favoring higher reproductive output can increase the production of recruits, leading to increased population growth rate. However, evolution of earlier maturity is not a guarantee of an increased population growth rate given that trade-offs between different life-history traits are common. The increase in population growth rate caused by earlier maturation could be counteracted by a reduction in survival. Similarly, larger recruitment could reduce the quality and survival prospects of the offspring produced. All of these considerations need to be taken into account when trying to predict the ecological response to evolutionary changes in life-history traits. Given this complexity, empirical studies are often necessary to obtain realistic and thorough understanding of the complete ecological response to these changes. As populations become adapted, population growth rates may increase until the population approaches carrying capacity, where density-dependence becomes a constraint. Carrying capacity, however, is also partly determined by the ability of individuals to acquire resources, which is also a phenotypic trait under selection. As such, evolutionary changes in traits may also cause changes in the carrying capacity and in density dependence in natural populations. Apart from a few exceptions, however, eco-evolutionary theory has seldom been challenged with empirical data outside the lab.

It is worth mentioning that evolution does not necessarily lead to an increase in population growth rate. Frequency-dependent selection may result in the evolution of a phenotype conferring a lower mean fitness. Additionally, it is important to remember that natural selection acts primarily on traits at the individual level, and not to improve the performance of the population as a whole. For example, in carnivores, sexually selected infanticide will favor male fitness of the perpetrator but will have a negative effect on offspring survival and ultimately reduce population growth. Further, the effects of evolutionary changes on population growth rate may be relatively weak if the population is regulated by other intrinsic or extrinsic factors which are independent of the evolving trait. The real challenge in eco-evolutionary dynamics is therefore not only to show the existence of interactions between trait changes and population parameters, but also to quantify their biological importance at the population level and evaluate whether those effects can cascade at higher level of biological organization such as the community.

Community-Level Eco-Evolutionary Dynamics

Through changes in population growth, eco-evolutionary dynamics will alter the abundance of the evolving species with potential cascading effects on the community. For example, the abundance of predators will undoubtedly influence the abundance of its prey, as often exemplified by the Canada lynx (*Lynx canadensis*) and snowshoe hare cycles (*Lepus americanus*) in most ecology textbook. However, community processes are not solely determined by the abundance of its constituent species. Trait changes in a species could also have direct effect on community-level dynamics because phenotypic traits will affect how species interacts with their abiotic and biotic environments. As a consequence, evolutionary changes in those traits could impact species niche and hence community assemblages. The increased focus on functional traits to study community ecology is a testament of the ecological importance that traits play in structuring communities. Thus, it has been suggested that the impact of evolutionary changes in traits of a species could be as important as those of changes in functional traits in a community assemblage.

Functional traits, in combination with environment filtering, have been found to be key drivers of community assemblage. Evolutionary changes in traits may therefore shift the ecological niche of a species and therefore its distribution, leading to new community compositions as one species range expands or contracts. For example, a plant may adapt to the local climatic conditions at its range margin allowing it to expand its niche toward colder and formerly unsuitable climates. Evolutionary changes in traits can also impact interspecific competition, trophic interactions, mutualism, parasitism, etc. These new biotic interactions can have several ecological consequences. One of the possible community-level outcomes of eco-evolutionary dynamics is a switch to an alternative stable state, which is often in response to a strongly altered selective regime. This can lead to changes in community assemblage. A second possible outcome of changes in biotic interactions is co-evolution. Indeed, co-evolution can be considered a special case of eco-evolutionary dynamic. Selection and evolution of one of the co-evolving species induces changes in the selective pressure on the second co-evolving species which in turn evolves in response to these changes, creating a feedback loop centered on the interaction between the two co-evolving species. Another possible outcome of eco-evolutionary dynamics at the community level is the emergence of cyclic evolutionary dynamics.

A great example of altered cyclic dynamics was shown in an experiment tracking the predator–prey dynamics of rotifer and green algae (**Fig. 3**) in the presence or absence of evolution in the prey. The ability of the prey to evolve (or not) in response to a high predation pressure was controlled by setting up the experiment either using clonal populations (i.e., no genetic diversity) or using populations composed of several genotypes. This experiment showed that the population abundance of algae fluctuates with the abundance of rotifers, the main predator, leading short population cycles in the absence of evolution by the prey (**Fig. 3A**). In prey populations that have the opportunity to evolve, the algae populations develop effective defense mechanisms that allow the populations to increase in size. As those defense mechanisms are costly, they are traded off with competitive abilities and the populations are doomed to crash at high densities. The evolution by the prey increased the duration of cycle and changed the

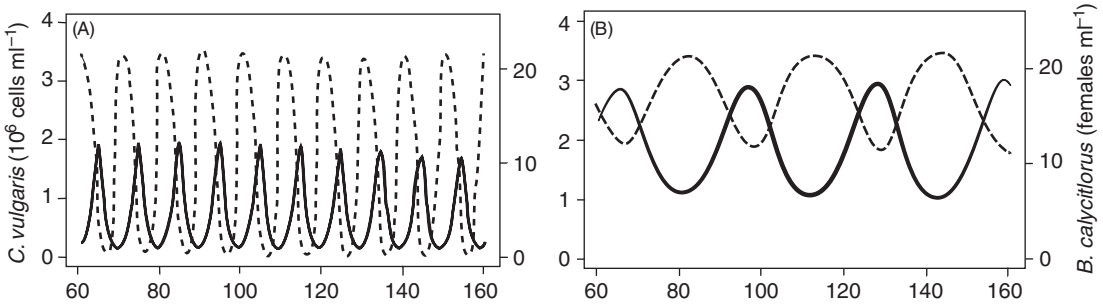


Fig. 3 Cycles predicted by the model for the *Brachionus* predator (*solid line*) and *Chlorella* prey (*dashed line*) in a single-clone system (A) and a multiple-clone system (B). Adapted from Yoshida, T., Jones, L. E., Ellner, S. P., Fussmann, G. F. and Hairston, N. G. (2003). Rapid evolution drives ecological dynamics in a predator-prey system. *Nature* **424**, 303–306.

predator–prey dynamics (Fig. 3B). This experiment provides support for the eco-evolutionary effects on predator–prey interactions.

The impact of evolutionary changes on communities will likely vary depending on which species and traits evolve. For example, the evolution of keystone species, such as an apex predator, is likely to have profound impacts on the community assemblage. For example, a reduction in size for a top predator will affect its ability to consume large prey. Alternative prey (smaller species) can then be exploited shifting the size distribution of the prey community. However, the ability of a species to influence its community probably also increases with its prevalence, suggesting that community effects are not limited to the evolution of top predators or traits with large ecological effects. Eco-evolutionary dynamics may also play a major role in the dynamics of invasive species. Indeed, invasive species with a high potential for adaptation to new environments will be more likely to be successful invaders. Similarly, the arrival of an invasive species will modify the selection pressures on endogenous species, making potential evolutionary changes in the endogenous species with unknown evolutionary consequences. Researchers working on invasive species (e.g., Pimentel) were among the first to underline the potential importance of the EVO to ECO links.

An alternative way to study the EVO to ECO links on community processes that complement the approach used in the studies presented above is the "focal-species composite-response" approach. The community composition, structure, and properties can be considered as an extended phenotype. As such, the focal species is able to evolve and as it does, it can have indirect genetic effects on its community. An example of a study using such an approach was conducted on North American cottonwoods (*Populus* sp.) and their arthropod communities. This study revealed that arthropod communities were highly heritable. Their results suggested that cottonwoods possessed an unknown trait that was heritable and which imposed selection upon arthropods. This trait generated indirect genetic effects on the fitness of the arthropod community members and produced distinct arthropod communities with distinguishable phenotypes potentially affecting other processes at the ecosystem level.

Ecosystem Level Eco-Evolutionary Dynamics

Eco-evolutionary dynamics at the ecosystem level has proven particularly challenging to study because of the logistic difficulties associated with working at large spatial scales while maintaining the resolution needed to measure evolutionary changes. Quantifying ecosystem-level effects of evolutionary changes may be further hindered because many ecosystem processes may require a long time to respond or for the effects to be measurable. Thus, studies focusing on this aspect of eco-evolutionary dynamics have initially relied on contrasting the effect of standing genetic variation on ecosystem processes compared to other ecological variables. One of the best examples of such an approach was conducted on Populus hybrids that differ in their polyphenol production of a genetically based trait. The study revealed that the concentration of condensed tannins was the best predictor of decomposition and nitrogen cycling, suggesting that intraspecific variation in that trait was an important driver of ecosystem functions.

An alternative approach to overcome the logistic difficulties of monitoring ecosystem level eco-evolutionary dynamics has been to use mathematical models. Recently, ecosystem modeling has been used to evaluate the ecological feedback of evolutionary changes in demography and life-history traits in response to size-selective fisheries. These models suggest that by selecting against large and more fecund fish, biomass of heavily harvested fish is greatly reduced. Some key studies have now provided strong evidence that those changes are likely to be, at last partially, genetically based. Given the critical importance of fish body size in the dynamics of consumer-resource aquatic networks, allometric trophic network model has revealed that heavily fished ecosystems are more likely to further decrease fish abundance and increase temporal variability in their food resources.

Theory suggests that the consequences of evolutionary changes in dispersal may also play an important role in ecosystem level eco-evolutionary dynamics. For example, the transition of species from an anadromous to freshwater resident life history can be important because anadromous fishes provide important subsidies of marine-derived nutrients to rivers, lakes, and streams. Additionally, studies on experimental metacommunities have suggested that dispersal rates may influence ecosystem properties, with productivity peaking at intermediate dispersal rates. Thus, evolutionary changes in migration propensity could be strong drivers of ecosystem processes.

Studies in mesocosms suggested that evolution could be a strong driver of ecosystem functions. For example, one study using guppies (*Poecilia reticulata*) and rivulus (*Rivulus hartii*) as model species compared the relative importance of evolution, coevolution, and species invasion on algal dynamics, invertebrate biomass, and decomposition rates. The evolution of guppies in response to alternative predation regimes significantly influenced algal biomass and accrual rates. More importantly, the effects of evolution and coevolution were larger than the effects of species invasion for some ecosystem responses, suggesting that under some circumstances within species changes in color and life-history traits may be as important a driver of ecosystem processes as changes in species composition.

Eco-Evolutionary Feedback

Most studies focus on the ECO to EVO link or the EVO to ECO link independently. Eco-evolutionary feedbacks have seldom been studied and can be pictured as the *Holy Grail* of eco-evolutionary dynamics studies. The occurrence of those two-ways feedbacks over long ecological time-scales are evident but are generally ignored over contemporary time-scales. It has been proposed that two main requirements need to be met in order to expect eco-evolutionary feedbacks: first, there must be a strong effect of the phenotype on the environment (e.g., organisms may structure or construct their environment) and second, the new environment must cause the subsequent evolution of the phenotype of the organisms inhabiting it. Those requirements have been suggested because for feedbacks to occur on an ecological time-scale, the evolving phenotype needs to have enough effects on the environment for new selective pressures to emerge, affecting in return the phenotype of the evolving species. Models of eco-evolutionary dynamics, such as the one presented in "From Evolution to Ecology (The EVO to ECO Link)" section on ecosystem consequences of size selective fisheries, are looking at those two-way interactions but empirical studies documenting both links are still scarce.

Cryptic Eco-Evolutionary Dynamics

More recently some researchers have argued that eco-evolutionary dynamics interactions may be most important when they are the least obvious. This phenomenon has been referred to as cryptic eco-evolutionary dynamics. Cryptic dynamics arise in a system when processes interact in a way that effectively conceals the action of one, or more, of those component processes. Thus, eco-evolutionary feedbacks, as described in previous sections, could occur but may not have measurable consequences on ecological processes because several processes can obscure their effects. First, evolution can be cryptic for example, when several genotypes may be favored in a varying environment because they produce the same outward phenotype. Such counter-gradient variation may be common and result in cryptic eco-evolutionary dynamics. Similarly, evolution of tolerance or resistance may result in no obvious change in the phenotype unless one is careful and controls for the relevant challenging environmental conditions. These types of cryptic eco-evolutionary dynamics may be of great ecological importance because they contribute to ecological stability. Another process leading to ecological stability with the potential for cryptic dynamics is environmental tracking, which would reduce the fluctuation in population size even as the evolving trait changes to accommodate some environmental changes. While the ecological consequences of cryptic eco-evolutionary dynamics are null or quasi-null, their importance becomes apparent only when compared, in practice or in theory, to an equivalent case in the absence of evolution. Eco-evolutionary biologists should therefore keep in mind that the most parsimonious explanation (no effect of evolution) may not always be the best answer if we are more interested in explanatory theory than mere predictions.

Approaches to Studying Eco-Evolutionary Dynamics

To date, a lot of our knowledge on eco-evolutionary dynamics comes from laboratory systems where the importance of interactions between ecology and evolution might have been amplified due to their quantification in controlled environments. Thus, to evaluate the importance EVO to ECO links over a short time-scale in natural environments, researchers have used various approaches to assess under which conditions feedbacks between ecology and evolution are expected and to quantify their importance in the wild.

Modeling

Modeling is probably the approach to the study of eco-evolutionary dynamics where most progresses have been made thanks to the theoretical study of processes such as co-evolution and host–parasite interactions. These theoretical models come in a variety of forms and reviewing them all is outside the scope of this review. Models of adaptive dynamics have also been very informative to understand eco-evolutionary dynamics. In this mathematical framework, an environment generated by genotypes is modeled and mutants are introduced to this fictive initial environment. Mutants with higher fitness are able to invade the environment while mutants with lower fitness are not. The presence of these successful mutants in turn changes the subsequent environment. The

ensuing dynamics can be simulated until a steady state is reached. The elaboration of such models has led to a better understanding of how previously stable coexistence dynamics can be perturbed when evolutionary changes occur much more quickly than ecological changes. Individual-based models, where every individual is modeled specifically from its genotype and phenotype to its reproduction and survival, have also been insightful. Such models have been used, for example, to predict evolutionary consequences of commercial fishing on population dynamics as well as the indirect consequences on aquatic communities.

A drawback of a theoretical modeling approach is that while it provides convincing evidence that eco-evolutionary dynamics can happen and could have significant impacts on ecological processes, it does not provide evidence of their presence in nature or information on their strength and prevalence. Mathematical models do provide insight into the condition under which eco-evolutionary dynamics are more likely to be present. Those models also provide hypotheses and predictions, which can then be verified in nature. Combining modeling to experimental or observational studies is a powerful approach, having the generalizability and precision of modeling while gaining the realism of the real world. The mechanisms that are supposed to be behind the observed effect can be modeled to verify the interpretation of the observed results. Alternatively, an eco-evolutionary process can be modeled and the result of the modeled is compared to the results of a similar experimental setup.

Observational Studies

Most observational studies are based on some sort of comparative approaches contrasting situations with and without evolution. This comparison can be made using two different populations (one that went through an evolutionary response and the other that did not), or by monitoring one single population that showed an evolutionary response in time. The researchers can then measure the effects that those evolutionary changes had on population dynamics as well as community and ecosystem processes. The changes in ecological processes due to evolution can be measured directly by correlating them to variation in genotypes. For example, such an approach was used in our previous example on cottonwood tree and arthropods revealing that the genotype of trees explained around 60% of the variation in arthropod communities (see "Community Level Eco-Evolutionary Dynamics" section). In this way, the community effect of individuals could be considered as an extended phenotype and standard quantitative method could be used.

Mathematical methods, such as the Geber method, strengthen the conclusion that can be made from observational studies by decomposing changes in ecological processes due to evolution from other processes. The Geber method consists in comparing the relative importance of different factors in generating the change in some ecological variables. Using the Geber method, one first measures the temporal changes in various factors, including evolutionary and environmental variables, as well as their effects on an ecological variable of interest. These resulting effects are then used to estimate the relative contribution of the change in each factor to the observed change in the ecological variable of interest allowing the quantification of the influence of trait changes on population growth. This method was used to partition the effects of traits and environmental conditions on population growth in five populations of wild ungulates and revealed that variation in birth mass in those species explained just as much of the variance in population growth rate as variation in weather conditions.

As previously mentioned, a key distinction to make when studying eco-evolutionary dynamics is that of evolutionary changes and phenotypic plasticity. A useful tool to achieve this in nature is the animal model, provided that pedigree information is available. First developed in the animal breeding and statistical genetic fields, the animal model is a form of mixed effects model which uses relatedness information from a pedigree to decompose phenotypic variance into its different genetic and environmental components. Using this information, the animal model allows estimating key parameters such as the heritability of a trait or the genetic correlations between traits. The animal model is now widely used in the field of quantitative genetics. The major advantage of the animal model over more traditional methods for estimating heritability, such as parent-offspring regressions or half-sib designs, is that it makes use of all the information from all types of relationship within the pedigrees. Further, other factors having potential effects on the phenotype can also be easily incorporated into the animal model.

Another framework in the study of eco-evolutionary dynamics, which has received increasing attention, is the integral projection matrix, or IPM. IPMs are an extension of matrix models where the distribution of continuous trait is tracked over time. IPMs are based on four basic functions: survival, development, recruitment, and inheritance. The survival function determines the survival of individuals according to their trait; the development function determines the change in traits such as size among survivors (growth); the recruitment function determines the recruitment from individuals according to their trait; and the inheritance function describes the relationship between the offspring and parent traits. Extensions of the IPMs which track changes in genotypes through time in addition to the trait are being developed. Given that the model tracks the distribution of the trait over time, its resulting survival, recruitment, and population growth, this tool could be well suited to the study of eco-evolutionary dynamic, although careful consideration of the genetic interpretation must be made. For instance, the inheritance function used to parametrize IPMs determines the trait distribution of the offspring at the next model iteration but it is not the heritability of the trait as defined in quantitative genetics.

Experimental Manipulation

Experimental manipulation provides the most convincing demonstration of the effects of eco-evolutionary dynamics. Experimental manipulation allows careful decomposition of the plastic, genetic, and other environmental effects on population and

community dynamics. Experiments with chemostats (i.e., aquatic microcosm system used for the culture of microorganisms in which the chemical composition is kept at a controlled level) have provided very convincing evidence of the importance of eco-evolutionary dynamics. This type of experiment was used to study predator–prey cycles of communities (algae-rotifer) with or without genetic diversity as previously explain in this article ("Community Level Eco-Evolutionary Dynamics" section and **Fig. 3**). In a later experiment using a similar experimental setup, it was shown that rapid evolution of rotifers (lower investment in sexual reproduction) caused an important shift in the nutrient balance of the ecosystem.

The major advantage of experimental manipulations, such as chemostats, and the reason why they have been so successful in demonstrating effects of evolution on ecology is that environmental conditions can be carefully altered to cause rapid and important evolutionary changes. Changes in genotype, phenotype, and their "ecological" consequences can be precisely measured. However, this strength is also their most criticized weakness because such study may amplify the effect size due the controlled environment in which they are conducted. Moreover, whether results obtained in a "bottle" can be extrapolated to natural populations or not is often questioned. The recent realization that natural conditions can also influence genetic variation, and hence evolvability, also cast doubts on the generalizability of laboratory experiments. Steps have been taken to remedy this issue by using mesocosms, which consist of large experimental set-ups, which aim at being closer to natural systems while still permitting replication and experimentation.

Conclusion

The fact that ecology and evolution interact has been known since Darwin. Evolutionary biologists have studied the ECO to EVO links for decades and have shown how natural and artificial changes in ecological conditions can drive evolution. The main novelty of eco-evolutionary dynamics is to emphasize the EVO to ECO links. The potential for evolution to feedback on ecological processes over long time-scales has been recognized for decades, but it is only recently that researchers have realized that interactions between these processes might also be important over short time-scales. Eco-evolutionary studies try to identify the circumstances where the EVO to ECO links may matter and more importantly quantify the importance of evolutionary changes relative to all other factors known to affect the ecology of wild species.

See also: Behavioral Ecology: Herbivore-Predator Cycles; Age Structure and Population Dynamics. Ecological Complexity: Population Dynamics: Stability. Ecological Data Analysis and Modelling: Ecological Models: Individual-Based Models; Grassland Models. Ecological Processes: Predation and Its Effects on Individuals: From Individual to Species. Ecosystems: Ecosystems. Evolutionary Ecology: Coevolution; Natural Selection; Fitness; Red Queen Dynamics; Evolutionary Ecology; Coexistence. General Ecology: Abundance; Community; Demography. Terrestrial and Landscape Ecology: Anthropogenic Landscapes

Reference

Sinervo, B., Svensson, E., Comendant, T., 2000. Density cycles and an offspring quantity and quality game driven by natural selection. Nature 406, 985–988.

Further Reading

Bailey, J.K., Hendry, A.P., Kinnison, M.T., Post, D.M., Palkovacs, E.P., Pelletier, F., Harmon, L.J., Schweitzer, J.A., 2009. From genes to ecosystems: An emerging synthesis of eco-evolutionary dynamics. New Phytologist 184, 746–749.

Hendry, A.P., 2016. Eco-evolutionary dynamics. New Jersey: Princeton University Press.

Kinnison, M.T., Hairston, N.G., Hendry, A.P., 2015. Cryptic eco-evolutionary dynamics. Annals of the New York Academy of Sciences 1360, 120–144.

Kokko, H., López-Sepulcre, A., 2007. The ecogenetic link between demography and evolution: Can we bridge the gap between theory and data? Ecology Letters 10, 773–782.

Kuparinen, A., Boit, A., Valdovinos, F.S., Lassaux, H., Martinez, N.D., 2016. Fishing-induced life-history changes degrade and destabilize fishery ecosystems. Scientific Reports 6, 1–8.

Pelletier, F., Clutton-Brock, T., Pemberton, J., Tuljapurkar, S., Coulson, T., 2007. The evolutionary demography of ecological change: Linking trait variation and population growth. Science 315, 1571–1574.

Post, D.M., Palkovacs, E.P., 2009. Eco-evolutionary feedbacks in community and ecosystem ecology: Interactions between the ecological theater and the evolutionary play. Philosophical Transactions of the Royal Society, B: Biological Sciences 364, 1629–1640.

van Benthem, K.J., Bruijning, M., Bonnet, T., Jongejans, E., Postma, E., Ozgul, A., 2017. Disentangling evolutionary, plastic and demographic processes underlying trait dynamics: A review of four frameworks. Methods in Ecology and Evolution 8 (1), 75–85.

Yoshida, T., Jones, L.E., Ellner, S.P., Fussmann, G.F., Hairston, N.G., 2003. Rapid evolution drives ecological dynamics in a predator-prey system. Nature 424, 303–306.

Eco-Immunology: Past, Present, and Future

Meredith Kernbach, Chloe Ramsay, Jason R Rohr, and Lynn B Martin, University of South Florida, Tampa, FL, United States

A Brief History of Immunology

Immunology and especially ecoimmunology are young fields. Although Thucydides speculated as long ago as 400 BC about immunity after noticing that survivors of the plague became protected from second infections, the first true immunological study was by Ilya Methchnikoff who observed cellular defensive responses in starfish to wood splinters (Newell-McGloughlin and Re, 2006). By the 18th century in Europe (Riedel, 2005), immune processes were exploited in the interest of human health even though the mechanistic bases of protection were unknown. In ensuing years, Robert Koch, Louis Pasteur, and others improved such vaccination and other hygienic measures, although explicitly immunological insight, such as the distinction between humoral and cellular immunity, was not gained until the early 20th century (Tauber, 2003; Kaufmann, 2008).

By the mid 20th century, coincident with the molecular revolution, immunology grew rapidly including discoveries about tumor regulation, allergies, anaphylaxis, and the structure and synthesis of immunoglobulins. As many of these discoveries were made possible through the emergence of cellular and later molecular tools, immunology thrived as an almost exclusively reductionist discipline. Since its inception, the field focused predominantly on how host cells and tissues engaged various "dangerous" (Matzinger, 1994) and nonself (Burnet, 1961) threats, but recently, the emphasis has been made that such directives have shortcomings that might have affected the accumulation of knowledge (Tauber, 2017). For instance, important discoveries in other fields, such as the ability to classically condition immune responses (Ader and Cohen, 1982) and predict the size of disease epidemics using mathematics (Anderson and May, 1985), developed separately. Indeed, the field of epidemiology, which eventually instigated the discipline now recognized as disease ecology, has only recently and partly begun to merge with immunology (Brock et al., 2014). Some immunology, namely vaccinology, is well integrated with epidemiology, but mostly intellectual pursuits about infectious disease biology have developed and operated independently of immunology (Tauber, 2017). The most striking distinction among infectious disease-related disciplines is how they deal with variation. Almost all immunologists have tried to control or eliminate it, going to extremes of breeding hosts in sterile conditions, standardizing genetic backgrounds, and/or attempting to ensure that temperature, diet, photoperiod, and other factors are identical among study groups. Most other disease biologists, and especially ecoimmunologists, instead strive to explain existing variation among or within hosts (Martin et al., 2010). Ecoimmunologists (and "wild immunologists" sensu Pedersen and Babayan, 2011) study the defenses of thousands of different host species, oftentimes in their natural environments (Plowright et al., 2008). For these immunologists, variation is the topic of interest.

Ecoimmunology Past

Ecoimmunology emerged from the work of behavioral ecologists interested in understanding elaborate sexual traits such as songs, ornaments, and mate-directed behaviors (e.g., courtship dances). Zahavi (1975) argued that these traits acted as handicaps, allowing females to use them as honest signals of male quality. The argument was that only mates of genuine high quality would be able to produce and keep such costly traits and achieve reproductive success too. Around the same time, Hamilton and Zuk (1982) extended the handicap hypothesis, proposing that male ornaments are good indicators of mate quality because the costs of elaborate traits should compromise the ability of males to combat infections (Hamilton and Zuk, 1982). If a male could maintain a large tail, produce an elaborate song, and/or execute a complicated display while infected, he would be a truly good male. Females should choose such males, especially if male traits had a genetic basis, which they often do. These ideas were later extended in the form of the immunocompetence handicap hypothesis (ICHH), which proposed a physiological mechanism for trade-offs between ornaments and immunity (Folstad and Karter, 1992). Observing that high levels of androgen hormones positively affected male ornamentation but negatively affected many immune functions, they proposed the first ecoimmunological

hypothesis. They suggested that the benefits of testosterone for attracting mates were traded off against the adverse effects of testosterone on immunity, providing a mechanistic basis for handicaps. Although this hypothesis has since been tested extensively (Hunt et al., 1997; Enstrom et al., 1997) and in some sense launched the field of ecoimmunology, androgen effects on immune defenses are not as consistent as they originally conveyed (Arredouani et al., 2014; Roberts et al., 2004). Alternate versions of the ICHH have since been proposed and supported in organisms that do not use androgens (Nunn et al., 2009), and these propositions, which are also based on eco-evolutionary principles (i.e., Bateman's hypothesis) explain intra- and interspecific immune variation reasonably well.

The term, ecological immunology, was first introduced in 1996 by Sheldon and Verhulst, and was proposed then as a valuable research area because most hosts remained susceptible to infections in spite of a strong history of selection by parasites. At the time, Sheldon, Verhulst, and others (Lochmiller and Deerenberg, 2000; Schmid-Hempel and Ebert, 2003) parroted a central proposition of evolutionary biology to justify ecoimmunology research: in most environments, resources are limited, so trade-offs should affect whether and how hosts defend themselves against infections. From that simple conjecture, the field developed rapidly (Martin et al., 2011). Growth increased especially rapidly as the costs of immune defense became clear. It was long known that immune defenses required a supply of amino acids and calories (Beisel, 1977), but the eco-evolutionary ramifications of such costs were never before considered. In the livestock industry especially, the costs of immunity were well studied because of the economic consequences of breeding the largest and fastest growing, yet immunologically well-protected, animals. For these reasons, diet effects on immune functions were extensively studied in domesticated species (Klurfeld, 1993; Klasing, 2013). Not too long after Sheldon and Verhulst's call to action, data corroborating the costliness of immune defenses started to emerge from ecologists (Demas et al., 1997; Martin et al., 2003, 2007; Lochmiller and Deerenberg, 2000). The next steps were to probe whether such immune costs were large enough to lead to consequential trade-offs in nature (Norris et al., 1994; Richner et al., 1995).

In the ensuing 20 years, an enormous number of studies provided both correlative and experimental evidence revealing important costs of immunity for many vertebrates and invertebrates (Saino et al., 1997; Festa-Bianchet, 1989; Graham et al., 2010; Moret and Schmid-Hempel, 2000). Just this year, a metaanalysis involving many of these studies revealed significant costs of immunity for almost every taxon ever studied (Brace et al., 2017). Songbirds and a few insects (i.e., bumblebees, *Drosophila* sp.) were the organisms of choice for much of the early work in this field. Several avian brood size experiments concluded that many measures of immune defense decreased with increasing reproductive output (Gustafsson et al., 1994; Norris et al., 1994; Norris and Evans, 2000). Trade-offs between immune function and other costly life processes, such as feather regrowth, were observed in domesticated chickens (Alodan and Mashaly, 1999), wild house sparrows (Martin, 2005), and other species. Even migratory disposition was found to impact immunity (Møller and Erritzøe, 1998), although results were mixed depending on several factors (Owen and Moore, 2006; van Gils et al., 2007). Growth too was found to affect and be affected by immunity (Rivera et al., 1998). Today, the costs of immunity and defenses generally are so ensconced in ecology (Connors and Nickol, 1991; Lochmiller et al., 1993) that we no longer question their relevance. Now, we seek to understand whether and how their effects can percolate through communities (Beldomenico and Begon, 2010; Paull et al., 2012).

Immunocompetence: The Ecoimmunologist's Red Herring

Coincident with the rise of ecoimmunology were frustration and healthy skepticism about the immunological tools of ecologists. What ecologists and organismal biologists wanted to measure was immunocompetence, or "how hosts prevent or control infections." What they used to describe such a complex trait were coarse measures at best (Adamo, 2004). This approach was understandably a frustration to traditional immunologists who spent years developing precise tools to characterize the responses of (mostly) murine immune systems to diverse viral, microbial, and metazoan threats. Most prominent among those first ecological tools was the phytohemagglutin (PHA) skin test. This assay was thought to quantify cell-mediated immunity and by extension serve as a measure of T-cell sensitive parasite control (Duffy and Ball, 2002; Faivre et al., 2003; etc.). In practice, one injected a small amount of PHA (in solution) under the skin and measured resultant swelling using pressure-sensitive calipers. Although this technique revealed several surprising things about the immune systems of wild animals (e.g., seasonal variability in strength (Martin et al., 2004) and nonlinear scaling with body mass (Tella et al., 2002)), rarely were PHA measures related to control of a given infection, so its relevance as a metric of immunocompetence remained unclear. For these reasons, the PHA and related tests (i.e., leukocyte counts, quantification of total immunoglobulins) began to fall out of favor.

At that time, ecoimmunologists started to rely less on easy-to-use tools, instead favoring functional readouts of host protection or tools used in the labs of modern immunologists (Fassbinder-Orth, 2014; Boughton et al., 2011). One assay involving agglutination of foreign red blood cells became quite popular (Deerenberg et al., 1997); another involving the killing of various microbes by blood plasma and serum quickly grew in popularity (Millet et al., 2007; Liebl and Martin, 2009). As ecoimmunologists started to accept that measuring single measures could never capture something as nebulous as immunocompetence (Martin et al., 2006; Viney et al., 2005), progress in the field grew. Not too long after, ecologists would propose new frameworks and concepts that would eventually spill into traditional immunology (Schmid-Hempel and Ebert, 2003; Viney et al., 2005). Foremost among these was the concept of parasite tolerance (Råberg et al., 2007).

Parasite tolerance emphasizes that hosts can cope with infection in a different manner than usually emphasized in immunology. Historically, immunology focused on mechanisms to regulate the numbers of parasites (i.e., resistance), just as a few behavioral ecologists had emphasized avoidance and other behaviors as a way to escape infection. The additional possibility, that hosts could

mitigate parasite damage instead of controlling parasite burden (Medzhitov et al., 2012), had never really been embraced, perhaps because as a predominantly biomedical field the thought of permitting enduring infections was too distasteful to pursue (Acevedo-Whitehouse and Cunningham, 2006; Hill, 2001; Schroder and Bowie, 2005). In the mid-20th century, though, plant biologists demonstrated quite clearly that hosts can cope feeding insects, fungi and other enemies better than trying to prevent exposure to these organisms altogether. Animal biologists eventually co-opted this concept, defining parasite tolerance mathematically as the relationship between host fitness or performance and parasite burden (Roy and Kirchner, 2000; Råberg et al., 2009). Today, research on parasite tolerance abounds and includes studies that attempt to refine its measurement (Louie et al., 2016), resolve its mechanistic basis (Sears et al., 2011; Ayres and Schneider, 2012; Soares et al., 2017; Medzhitov et al., 2012), and elucidate its eco-evolutionary effects at levels of biological organization above individuals (Gervasi et al., 2017; Johnson and Levin, 2013).

Ecoimmunology Present

From Immunocompetence to Host Competence

Ecoimmunology continues to be a popular field, as evidenced by the many reviews (Little et al., 2012; Demas and Nelson, 2011; Demas et al., 2011; Adamo, 2004; Raffel et al., 2008; Martin, 2009; Martin et al., 2006, 2010; Adelman and Martin, 2009; Ezenwa et al., 2016) of its history, purview, and implications. Of the many concepts that currently captivate researchers, one, host competence (i.e., the propensity of an individual to generate infections in another individual) is facilitating integration with other disease research at multiple levels of biological organization (Gervasi et al., 2015, 2017; Barron et al., 2015; Martin et al., 2016; Keesing et al., 2012; Luis et al., 2013). Competence is underlain by many physiological and behavioral characteristics that affect exposure, elimination, and subsequent transmission of a parasite (Ferrari et al., 2004). Variability in host competence within and among individuals and species can contribute to spatiotemporal heterogeneity in disease at the community level (Barron et al., 2015; Gervasi et al., 2015; Martin et al., 2017). To date, there have been two prime ecological areas of study with regards to competence: the dilution effect and superspreading.

The first, the dilution effect, represents the recurrent observation that host diversity often reduces disease risk (Civitello et al., 2015; Ostfeld and Keesing, 2000). Although the scale of study and other factors can eliminate or even reverse the effects of biodiversity on disease risk (Wood et al., 2014), dilution effects are common enough (LoGiudice et al., 2003; Keesing et al., 2010; Ezenwa et al., 2006) that many have been motivated to investigate their mechanistic bases. Prime among them is heterogeneity in host competence. Potential hosts in an environment differ dramatically in their ability to replicate parasites (Schmidt and Ostfeld, 2001). To date, the Lyme disease system has been best-studied with regards to the role of competence in dilution effects. Lyme disease is caused by infections with the bacterium, *Borrelia burgdorferi* (Bb), with is transmitted by ixodid ticks. A few mammalian hosts are much more competent than others and play important roles in the local dynamics of Lyme disease. The white-footed mouse (*Peromyscus leucopus*) is especially competent for Bb, and its immune system and particularly its antimicrobial responses are quite distinct from other mammals (Previtali et al., 2012), even other *Peromyscus* (Martin et al., 2007, 2008). What is also particularly compelling and somewhat concerning about this and similar species is that it is competent for multiple tick-borne, disease-causing microbes (Ostfeld et al., 2014). Intriguingly, variation in competence can be predicted by the life history and behavioral traits of hosts, phenomena that provide opportunities for the mitigation of human disease risk and perhaps even predictions about under- or unstudied species in new ecosystems or novel parasites (Han et al., 2015). Evidence is gathering in other systems, such as vertebrate-trematode, snail-trematode, vertebrate-fungal, and plant-fungal systems, that the most abundant and widespread hosts, with life-history traits that favor reproduction and dispersal, are often the most competent (Johnson et al., 2012; Sears et al., 2015; Venesky et al., 2014; Lively and Dyndahl, 2000; Parker et al., 2015).

A second reason ecologists are now interested in host competence is embodied in the hallmark superspreader, Mary Mallon, better known as Typhoid Mary (Soper, 1939). Typhoid Mary was said to be responsible for the infections of 22 people, which proved fatal for three of those individuals. As an asymptomatic typhoid fever carrier, Mary came into contact with many others while working as a cook. Typhoid Mary is not an aberration it seems. Superspreaders appear to be important for other diseases, such as acute respiratory syndrome, or SARS, which first emerged in late 2002 and rapidly spread via air travel through May 2003. There were nearly 8000 reported cases of SARS (Riley et al., 2003), but transmission events varied quite extensively across patients. Patient A, the first known infected patient, reportedly transmitted the SARS pathogen to 33 other individuals (Shen et al., 2004). Scientists concluded that SARS was likely transmitted to three other individuals, all of whom infected more than eight additional people (Shen et al., 2004).

Although only recently emphasized, superspreaders might be common in many host-parasite systems (Lloyd-Smith et al., 2005; Kilpatrick et al., 2006) including tuberculosis and measles outbreaks (Stein, 2011). Indeed, the 20:80 rule in epidemiology highlights that 20% of infected individuals tend to be responsible for 80% of new infections (Woolhouse et al., 1997). Methods for identifying these keystone individuals are only just being developed (Modlmeier et al., 2014; Paull et al., 2012), and often identification is quite hard to do because hard-to-quantify behavioral differences are what distinguishes hosts. Mechanistically, though, there are several promising avenues to consider as biomarkers, and there is a large need for such work, as competence can be quite labile in some species (Gervasi et al., 2016). In other words, whereas host species vary quite a bit in their competence, individual hosts within species too could be quite different too. Various natural and anthropogenic stressors, resource depletion, habitat quality, and other factors could alter the traits that hosts use to cope with infections (Gervasi et al., 2015; Martin et al., 2010).

For instance, glucocorticoids, which have profound and diverse effects on immune function (Martin, 2009), dramatically altered the ability of avian hosts to serve as competent reservoirs for West Nile virus. Birds with chronically (surgically) elevated stress hormones were 2× more attractive (Gervasi et al., 2016) and enduringly more infectious (Gervasi et al., 2017) to biting vectors than controls. Some of these effects were related to the expression of cytokines (e.g., interferon gamma), molecules expressed by leukocytes and other cells that directly antagonize parasites or coordinate cellular defenses (Martin et al., 2016).

Causes and Consequences of Co-Infection

Another growing area of interest in ecoimmunology and a prime example of the ongoing merger between ecoimmunology and disease ecology (Hawley and Altizer, 2011; Brock et al., 2014) involves co-infections. In the wild, it is probably much more common to be infected by many versus a single pathogen, which makes the common practice in immunology to study one parasite alone naturally unrepresentative. When a host is co-infected, interactions between pathogens can alter the outcome of infections (Corbett et al., 2003; Bromenshenk et al., 2010; Johnson and Hoverman, 2012). Pathogens can interact directly, competing for resources or space (Kuris and Lafferty, 1994), but they can also interact indirectly, through the immune system (Sousa, 1992). The best known and most extreme example of co-infection in hosts and how the immune system can play a role in the outcome of disease entails the human immunodeficiency virus, HIV (Shankar et al., 2014; Moreno et al., 2000; Abu-Raddad et al., 2006). HIV reduces the functionality of the host immune system, leaving the host open to secondary infections. For instance, co-infection with HIV and tuberculosis (TB) accelerates the deterioration of the host immune system and increases the negative effects of both pathogens (Shankar et al., 2014). Patients with leishmaniasis and HIV experienced more relapses in leishmaniasis symptoms after treatment with pentavalent antimony than individuals who were not co-infected (Moreno et al., 2000). Co-infected individuals had significantly lower levels of serum transforming growth factor-beta (TGF-β1; Moreno et al., 2000). The authors postulated that the lower levels of TGF-β1 and the host's inability to mount a leishmaniasis-specific immune response could be the reason for more common relapses (Moreno et al., 2000). Co-infection with HIV and malaria has been shown to increase transmission of both of these pathogens. HIV facilitates malaria transmission and infection with malaria strongly activates CD4 cells, which then facilitates HIV transmission because HIV can replicate rapidly in the CD4 cells (Alemu et al., 2013).

Outcomes of co-infection can be dependent on the host's immune resource allocation and, as described above, the type of immune response necessary to combat the infections in question (Alemu et al., 2013; Graham, 2008; Sousa, 1992). In situations where resources allocated to the host's immune response are limited, mounting a sufficient immune response can be challenging. This is particularly true in cases where the immune system mounts differing responses to the pathogens from different taxa. Given limited resources or a lack of coevolutionary history between host and pathogen (Graham, 2002; Graham et al., 2005), activating and maintaining two distinct immune pathways might be impossible. Subsequently, co-infection could have greater impacts on hosts or even host populations than single infections. For instance, in African buffalo (*Syncerus caffer*), co-infection with bovine tuberculosis (BTB) and gastrointestinal helminths decreased mortality risk in individual buffalo. Individuals infected with BTB alone had higher levels of interferon gamma (IFN-y) than individuals that were coinfected with worms. Altogether, interactions between these two infections led to higher prevalence of helminth infections in the population because of the effect of BTB on host immune systems (Ezenwa and Jolles, 2015).

Co-infection with similar pathogens can also increase disease burden. Co-infection of murine hosts with *B. burgdorferi* and *Ehrlichia bacteria*, the causative agent of Granulocytic Ehrlichiosis, increased pathogen loads and Lyme arthritis severity relative to hosts infected with *B. burgdorferi* alone. Co-infected hosts also had lower levels of interleukin-12 (IL-12), interferon-y (IFN-y), and fewer INF-y receptors on macrophages. These Th1-type cytokines are typically upregulated to combat bacterial infections. Conversely, they saw an increase in interleukin-6 (IL-6) levels with co-infection (Thomas et al., 2001). Increases in IL-6 levels can decrease the effects of Lyme arthritis in C57BL/6 mice (Anguita et al., 1996), which means that despite increased burden, this immune response might decrease symptoms. Indeed, co-infections with similar pathogens need not always increase disease burden. Co-infections with similar pathogens could foster cross immunity in hosts. In these cases, parasites are so similar that host responses to one infection will be protective against subsequent infections. For example, when mice were infected with genetically distinct clones of rodent malaria, the primary infecting clone established. However, if the secondary infecting clone was added after the primary infection had established, the secondary clone did not reach as high a burden (de Roode et al., 2005). Moreover, exposure to like pathogens at similar times can cause a more rapid or larger magnitude systemic response. This scenario could lead to more rapid and effective protection against both pathogens. Co-infections are known alter host immunity (Thomas et al., 2001; Moreno et al., 2000), which in turn, can change the outcome of infection and increase pathogen transmission between hosts (Alemu et al., 2013; Ezenwa and Jolles, 2015). Ecoimmunologists are working to better understand how pathogen interactions are immune-mediated and how co-infections alter transmission dynamics and population disease dynamics.

Ecoimmunology Future

The above text about dilution effects, superspreading, and coinfections highlights the need to understand the effects of anthropogenic change on disease in natural systems. Since the inception of the field, scientists have probed how global changes will affect immune responses of individuals and the impacts of these effects on populations and communities (Martin et al., 2010; Martin and Boruta, 2014). For years now we have known that reproduction, molt, migration, and development all induce context-dependent

costs on an individual, and these costs affect their ability to respond to infections (Brace et al., 2017). Ongoing, escalating global changes are apt to alter or sometimes even exacerbate immune-related trade-offs. For example, amphibian species that reside in human-impacted forest exhibit reduced bactericidal killing ability (Hopkins and DuRant, 2011). Poor habitat quality and resource scarcity also correlate with land development (Klurfeld, 1993; Moret and Schmid-Hempel, 2000; Cropper and Griffiths, 1994), and these factors are known from other fields to affect immune responses (Glick et al., 1981). Habitat destruction is not the sole consequence of human disturbance in several circumstances. There is evidence that chemical toxins affect susceptibility to pathogens (Ross et al., 2000). In a metaanalysis, pesticide exposure at environmentally relevant levels was shown to decrease immune function and increase pathogen load in fish and amphibians (Rohr and McCoy, 2010). Pesticides have been shown to decrease a variety of immune functions in these taxa including decreases in proliferative activity of B and T lymphocytes (Rymuszka et al., 2007), lower levels of melano-macrophages in the liver, eosinophils in the blood (Rohr et al., 2008), and total white blood cell counts (Christin et al., 2004).

In its short lifespan, ecoimmunology has made many important contributions and begun to merge with disease ecology to produce one of the most integrated biological research fields (Brock et al., 2014; Ezenwa et al., 2016). Going forward, an integrated disease ecology promises to have more positive impacts on our understanding of immune variation among and within species, which could be helpful in medicine, mitigation of global climate change effects, and management and prevention of zoonotic and other disease outbreaks (Tauber, 2017).

References

Abu-Raddad LJ, Patnaik P, and Kublin JG (2006) Dual infection with HIV and malaria fuels the spread of both diseases in Sub-Saharan Africa. *Science* 314: 1603–1606.

Acevedo-Whitehouse K and Cunningham A (2006) Is MHC enough for understanding wildlife immunogenetics? *Trends in Ecology & Evolution* 21(8): 433–438.

Adamo SA (2004) How should behavioral ecologists interpret measurements of immunity? *Animal Behavior* 68(6): 1443–1449.

Adelman JS and Martin LB (2009) Vertebrate sickness behaviors: Adaptive and integrated neuroendocrine immune responses. *Integrative and Comparative Biology* 49(3): 202–214.

Ader R and Cohen N (1982) Behaviorally conditioned immunosuppression and murine systemic lupus erythematosus. *Science* 215(4539): 1534–1536.

Alemu A, Shiferaw Y, Addis Z, Mathewos B, and Birhan W (2013) Effect of malaria on HIV/AIDS transmission and progression. *Parasites & Vectors* 6: 18.

Alodan MA and Mashaly MM (1999) Effect of induced molting in laying hens on production and immune parameters. *Poultry Science* 78(2): 171–177.

Anderson R and May R (1985) Helminth infections of humans: Mathematical models. *Population Dynamics, and Control* 24: 1–101.

Anguita J, Persing DH, Rincon M, Barthold SW, and Fikrig E (1996) Effect of anti-interleukin 12 treatment on murine Lyme borreliosis. *Journal of Clinical Investigation* 97: 1028–1034.

Arredouani S, Kissick H, Dunn L, and Sanda M (2014) PTP1B regulates lymphocytes responses androgen deprivation. *Journal for Immunotherapy of Cancer* 2(Suppl 1): P2.

Ayres JS and Schneider DS (2012) Tolerance of infections. *Annual Review of Immunology* 30: 271–279.

Barron D, Gervasi S, Pruitt J, and Martin L (2015) Behavioral competence: How host behaviors can interact to influence parasite transmission risk. *Current Opinion in Behavioral Sciences* 6: 35–40.

Beisel WR (1977) Magnitude of the host nutritional responses to infection. *The American Journal of Clinical Nutrition* 30(8): 1236–1247.

Beldomenico P and Begon M (2010) Disease spread, susceptibility and infection intensity: Vicious circles? *Trends in Ecology & Evolution* 25(1): 21–27.

Boughton RK, Joop G, and Armitage SAO (2011) Outdoor immunology: Methodological considerations for ecologists. *Functional Ecology* 25(1): 81–100.

Brace AJ, Lajeunesse MJ, Ardia DR, Hawley DM, Adelman JS, Buchanan KL, Fair J, Grindstaff J, Matson KD, and Martin LB (2017) Costs of immune responses are related to host body size and lifespan. *Journal of Experimental Zoology Part A* 327: 254–261.

Brock PM, Murdock CC, and Martin LB (2014) The history of Ecoimmunology and its integration with disease ecology. *Integrative and Comparative Biology* 54(3): 353–362.

Bromenshenk JJ, Henderson CB, Wick CH, Stanford MF, Zulich AW, Jabbour RE, Deshpande SV, McCubbin PE, Seccomb RA, Welch PM, Williams T, Firth DR, Skowronski E, Lehmann MM, Bilimoria SL, Gress J, Wanner KW, and Cramer RA (2010) Iridovirus and microsporidian linked to honey bee decline. *PLoS One* 5(10): e13181.

Burnet FM (1961) Immunological recognition of self. *Science* 133: 307–311.

Christin M, Menard L, Gendron A, Ruby S, Cyr R, Marcogliese D, Rollins-Smith L, and Fournier M (2004) Effects of agricultural pesticides on the immune system of *Xenopus laevis* and *Rana pipiens*. *Aquatic Toxicology* 67(1): 33–43.

Civitello DJ, Cohen J, Fatima H, Halstead NT, Liriano J, McMahon TA, Ortega CN, Sauer EL, Sehgal T, Young S, and Rohr JR (2015) Biodiversity inhibits parasites: Broad evidence for the dilution effect. *Proceedings of the National Academy of Sciences of the United States of America* 112(28): 8667–8671.

Connors VA and Nickol BB (1991) Effects of *Plagiorhynchus cylindraceus* (Acanthocephala) on the energy metabolism of adult starlings, sturnus vulgaris. *Parasitology* 103(3): 395–402.

Corbett EL, Watt CJ, Walker N, Maher D, Williams BG, Raviglione MC, and Dye C (2003) The growing burden of tuberculosis: Global trends and interactions with the HIV epidemic. *Archives of Internal Medicine* 163: 1009–1021.

Cropper M and Griffiths C (1994) The interaction of population growth and environmental quality. *The American Economic Review* 84(2): 250–254.

de Roode JC, Helinksi MEH, Anwar MA, and Read AF (2005) Dynamics of multiple infection and within-host competition in genetically diverse malaria infections. *The American Naturalist* 166(5): 531–542.

Deerenberg C, Arpanius V, Daan S, and Bos N (1997) Reproductive effort decreases antibody responsiveness. *Proceedings of the Royal Society B* 264: 1021–1029.

Demas G and Nelson R (2011) *Ecoimmunology*. Oxford: Oxford University Press.

Demas GE, DeVries AC, Nelson RJ, and Nelson RJ (1997) Effects of photoperiod and 2-deoxy-D-glucose-induced metabolic stress on immune function in female deer mice. *The American Journal of Physiology* 272(6 Pt 2): R1762–7.

Demas GE, Adamo SA, and French SS (2011) Neuroendocrine-immune crosstalk in vertebrates and invertebrates: Implications for host defence. *Functional Ecology* 25(1): 29–39.

Duffy DL and Ball GF (2002) Song predicts immunocompetence in male European starlings (*Sturnus vulgaris*). *Proceedings of the Royal Society B* 269: 847–852.

Enstrom DA, Ketterson ED, and Nolan VJ (1997) Testosterone and mate choice in the dark-eyed junco. *Animal Behavior* 54: 1135–1146.

Ezenwa VO and Jolles AE (2015) Opposite effects of anthelminthic treatment on microbial infection at individual versus population scales. *Science* 347: 175–177.

Ezenwa VO, Godsey MS, King RJ, and Guptill SC (2006) Avian diversity and West Nile virus: Testing associations between biodiversity and infectious disease risk. *Proceedings of the Biological Sciences* 273(1582): 109–117.

Ezenwa VO, Archie EA, Craft ME, Hawley DM, Martin LB, Moore J, and White L (2016) Host behaviour-parasite feedback: An essential link between animal behaviour and disease ecology. *Proceedings of the Royal Society B* 283(1828): 20153078.

Faivre B, Grégoire A, Préault M, Cézilly F, and Sorci G (2003) Immune activation rapidly mirrored in a secondary sexual trait. *Science* 300(5616): 103.

Fassbinder-Orth CA (2014) Gene expression quantification methods in ecoimmunology: From qPCR to RNA-Seq. *Integrative and Comparative Biology* 54: 396–406.

Ferrari N, Cattadori IM, Nespereira J, Rizzoli A, and Hudson PJ (2004) The role of host sex in parasite dynamics: Field experiments on the yellow-necked mouse *Apodemus flavicollis*. *Ecology Letters* 7(2): 88–94.
Festa-Bianchet M (1989) Individual differences, parasites, and the costs of reproduction for bighorn ewes (*Ovis canadensis*). *The Journal of Animal Ecology* 58(3): 785.
Folstad I and Karter AJ (1992) Parasites, bright males, and the immunocompetence handicap. *The American Naturalist* 139: 603–622.
Gervasi SS, Civitello DJ, Kilvitis HJ, and Martin LB (2015) The context of host competence: A role for plasticity in host-parasite dynamics. *Trends in Parasitology* 31(9): 419–425.
Gervasi SS, Burkett-Cadena N, Burgan SC, Schrey AW, Hassan HK, Unnasch TR, and Martin LB (2016) Host stress hormones alter vector feeding preferences, success, and productivity. *Proceedings of the Royal Society B: Biological Sciences* 283: 20161278.
Gervasi SS, Stephens PR, Hua J, Searle CL, Xie GY, Urbina J, Olson DH, Bancroft BA, Weis V, Hammond JI, Relyea RA, and Blaustein AR (2017) Linking ecology and epidemiology to understand predictors of multi-host responses to an emerging pathogen, the amphibian Chytrid fungus. *PLoS One* 12(1): e0167882.
Glick B, Day EJ, and Thompson D (1981) Calorie-protein deficiencies and the immune response of the chicken I. Humoral immunity. *Poultry Science* 60(11): 2494–2500.
Graham AL (2002) When T-helper cells don't help: Immunopathology during concomitant infection. *The Quarterly Review of Biology* 77: 409–433.
Graham AL (2008) Ecological rules governing helminth-microparasite coinfection. *Proceedings of the National Academy of Sciences* 105(2): 566–570.
Graham AL, Lamb TJ, Read AF, and Allen JE (2005) Malaria-filaria co-infection in mice makes malarial disease more severe unless filarial infection achieves patency. *Journal of Infectious Diseases* 191: 410–421.
Graham AL, Hayward AD, Watt KA, Pilkington JG, Pemberton JM, and Nussey DH (2010) Fitness correlates of heritable variation in antibody responsiveness in a wild mammal. *Science* 330: 662–665.
Gustafsson L, Nordling D, Andersson MS, Sheldon BC, and Qvarnstrom A (1994) Infectious diseases, reproductive effort and the cost of reproduction in birds. *Philosophical Transactions of the Royal Society B* 346(1317): 323–331.
Hamilton WD and Zuk M (1982) Heritable true fitness and bright birds: A role for parasites? *Science* 218: 384–387. https://doi.org/10.1126/science.7123238.
Han BA, Park AW, Jolles AE, and Altizer S (2015) Infectious diseases transmission and behavioural allometry in wild mammals. *Journal of Animal Ecology* 84(3): 637–646.
Hawley DM and Altizer SM (2011) Disease ecology meets ecological immunology: Understanding the links between organismal immunity and infection dynamics in natural populations. *Functional Ecology* 25(1): 48–60.
Hill A (2001) Immunogenetics and genomics. *The Lancet* 357(9273): 2037–2041.
Hopkins WA and DuRant SE (2011) Innate immunity and stress physiology of eastern hellbenders (*Cryptobranchus alleganiensis*) from two stream reaches with differing habitat quality. *General and Comparative Endocrinology* 174(2): 107–115.
Hunt KE, Hahn TP, and Wingfield JC (1997) Testosterone implants increase song but not aggression in male *Lapland longspurs*. *Animal Behavior* 54: 1177–1192.
Johnson PTJ and Hoverman JT (2012) Parasite diversity and coinfection determine pathogen infection success and host fitness. *Proceedings of the National Academy of Sciences* 109(23): 9006–9011.
Johnson PJT and Levin BR (2013) Pharmacodynamics, population dynamics, and evolution of persistence in *Staphylococcus aureus*. *PLoS Genetics* 9(1): e1003123.
Johnson PTJ, Rohr JR, Hoverman JT, Kellermanns E, Bowerman J, and Lunde KB (2012) Living fast and dying of infection: Host life history drives interspecific variation in infection and disease risk. *Ecology Letters* 15: 235–242.
Kaufmann SHE (2008) Immunology's foundation: The 100-year anniversary of the Nobel Prize to Paul Ehrlich and Elie Metchnikoff. *Nature Immunology* 9(7): 705–712.
Keesing F, Belden LK, Daszak P, Dobson A, Harvell CD, Holt RD, Hudson P, Jolles A, Jones KE, Mitchell CE, Myers SS, Bogich T, and Ostfeld RS (2010) Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature* 468(7324): 647–652.
Keesing F, Michelle HH, Michael T, McHenry DJ, Duerr S, Brunner J, Killilea M, LoGiudice K, Schmidt KA, and Ostfeld RS (2012) Reservoir competence of vertebrate hosts for *Anaplasma phagocytophilum*. *Emerging Infectious Diseases* 18(12): 2013–2016.
Kilpatrick AM, Daszak P, Jones MJ, Marra PP, and Kramer LD (2006) Host heterogeneity dominates West Nile virus transmission. *Proceedings of the Biological Sciences* 273(1599): 2327–2333.
Klasing KC (2013) In: Swayne DE (ed.) *Nutritional diseases, in diseases of poultry*. Chichester: John Wiley & Sons, Ltd. Chapter 29.
Klurfeld DM (1993) Cholesterol as an immunomodulator. In: *Nutrition and Immunology*, pp. 79–89. New York: Plenum Press.
Kuris AM and Lafferty KD (1994) Community structure: Larval trematodes in snail hosts. *Annual Review of Ecology and Systematics* 25: 189–217.
Liebl AL and Martin LB (2009) Simple quantification of antimicrobial capacity of blood using spectrophotometry. *Functional Ecology* 23: 1091–1096.
Little TJ, Allen JE, Babayan SA, Matthews KR, and Colegrave N (2012) Harnessing evolutionary biology to combat infectious disease. *Nature Medicine* 18(2): 217–220.
Lively CM and Dyndahl MF (2000) Parasite adaptation to locally common host genotypes. *Nature* 405: 679–681.
Lloyd-Smith J, Schreiber S, Kopp P, and Getz W (2005) Superspreading and the effect of individual variation on disease emergence. *Nature* 438(7066): 355.
Lochmiller RL and Deerenberg C (2000) Trade-offs in evolutionary immunology: Just what is the cost of immunity? *Oikos* 88(1): 87–98.
Lochmiller RL, Vestey MR, and Boren JC (1993) Relationship between protein nutritional status and Immunocompetence in northern bobwhite chicks. *The Auk* 110(3): 503–510.
LoGiudice K, Ostfeld RS, Schmidt KA, and Keesing F (2003) The ecology of infectious disease: Effects of host diversity and community composition on Lyme disease risk. *Proceedings of the National Academy of Sciences of the United States of America* 100(2): 567–571.
Louie A, Song KH, Hotson A, Thomas Tate A, and Schneider DS (2016) How many parameters does it take to describe disease tolerance? *PLoS Biology* 14(4): e1002435.
Luis AD, Hayman DTS, O'Shea TJ, Cryan PM, Gilbert AT, Pulliam JRC, Mills JN, Timonin ME, Willis CK, Cunningham AA, Fooks AR, Rupprecht CE, Wood JL, and Webb CTA (2013) Comparison of bats and rodents as reservoirs of zoonotic viruses: Are bats special? *Proceedings of the Royal Society B: Biological Sciences* 280(1756): 20122753.
Martin LB (2005) Trade-offs between molt and immune activity in two populations of house sparrows (Passer domesticus). *Canadian Journal of Zoology* 83(6): 780–787.
Martin LB (2009) Stress and immunity in wild vertebrates: Timing is everything. *General and Comparative Endocrinology* 163: 70–76.
Martin LB and Boruta M (2014) The impacts of urbanization on avian disease transmission and emergence. In: Gil D and Brumm H (eds.) *Avian urban ecology*. Oxford: Oxford University Press.
Martin LB, Scheuerlein A, and Wikelski M (2003) Immune activity elevates energy expenditure of house sparrows: A link between direct and indirect costs. *Proceedings of the Royal Society of London B: Biological Sciences* 270: 153–158.
Martin LB, Pless M, Svoboda J, and Wikelski M (2004) Immune activity in temperate and tropical house sparrows: A common-garden experiment. *Ecology* 85: 2323–2331.
Martin LB, Weil ZM, and Nelson RJ (2006) Refining approaches and diversifying directions in ecoimmunology. *Integrative and Comparative Biology* 46(6): 1030–1039.
Martin LB, Navara KJ, Weil ZM, and Nelson RJ (2007) Immunological memory is compromised by food restriction in deer mice, *Peromyscus maniculatus*. *American Journal of Physiology* 292: R316–320.
Martin LB, Johnson EM, Hutch CR, and Nelson RJ (2008) 6-MBOA affects testis size, but not delayed-type hypersensitivity, in white-footed mice (*Peromyscus leucopus*). *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology* 149(2): 181–187.
Martin LB, Hopkins WA, Mydlarz LD, and Rohr JR (2010) The effects of anthropogenic global changes on immune functions and disease resistance. *Annals of the New York Academy of Sciences* 1195(1): 129–148.
Martin LB, Hawley DM, and Ardia DR (2011) An introduction to ecological immunology. *Functional Ecology* 25(1): 1–4.
Martin LB, Burgan SC, Adelman JS, and Gervasi SS (2016) Host competence: An organismal trait to integrate immunology and epidemiology. *Integrative and Comparative Biology* 56(6): 1225–1237.
Martin LB, Kilvitis HJ, Brace AJ, Cooper L, Haussmann MF, Mutati A, Fasanello V, O'Brien S, and Ardia DR (2017) Costs of immunity and their role in the range expansion of the house sparrow in Kenya. *The Journal of Experimental Biology* 220(12): 2228–2235.
Matzinger P (1994) Tolerance, danger, and the extended family. *Annual Review of Immunology* 12(1): 991–1045.
Medzhitov R, Schneider DS, and Soares MP (2012) Disease tolerance as a defense strategy. *Science* 335(6071): 936–941.

Millet S, Bennett J, Lee KA, Hau M, and Klasing KC (2007) Quantifying and comparing constitutive immunity across avian species. *Developmental and Comparative Immunology* 31: 188–201.
Modlmeier AP, Keiser CN, Watters JV, Sih A, and Pruitt JN (2014) The keystone individual concept: An ecological and evolutionary overview. *Animal Behaviour* 89: 53–62.
Møller AP and Erritzøe J (1998) Host immune defence and migration in birds. *Evolutionary Ecology* 12(8): 945–953.
Moreno J, Canavate C, Chamizo C, Laguna F, and Alvar J (2000) HIV—Leishmania infantum co-infection: Humoral and cellular immune responses to the parasite after chemotherapy. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 94(3): 328–332.
Moret Y and Schmid-Hempel P (2000) Survival for immunity: The price of immune system activation for bumblebee workers. *Science* 290(5494): 1166–1168.
Newell-Mcgloughlin M and Re EB (2006) *The evolution of biotechnology from natufians to nanotechnology*. Dordrecht, the Netherlands: Springer.
Norris K and Evans MR (2000) Ecological immunology: Life history trade-offs and immune defense in birds. *Behavioral Ecology* 11(1): 19–26.
Norris K, Anwar M, and Read AF (1994) Reproductive effort influences the prevalence of Haematozoan parasites in great tits. *The Journal of Animal Ecology* 63(3): 601.
Nunn CL, Lindenfors P, Purshall ER, and Rolff J (2009) On sexual dimorphism in immune function. *Philosophical Transactions of the Royal Society of London B* 364: 61–69.
Ostfeld RS and Keesing F (2000) Biodiversity series: The function of biodiversity in the ecology of vector-borne zoonotic diseases. *Canadian Journal of Zoology* 78(12): 2061–2078.
Ostfeld RS, Levi T, Jolles AE, Martin LB, Hosseini PR, and Keesing F (2014) Life history and demographic drivers of reservoir competence for three tick-borne zoonotic pathogens. *PLoS One* 9(9): e107387.
Owen JC and Moore FR (2006) Seasonal differences in immunological condition of three species of thrushes. *The Condor* 108(2): 389–398.
Parker IM, Saunders M, Bontrager M, Weitz AP, Hendricks R, Magarey R, Suiter K, and Gilbert GS (2015) Phylogenetic structure and host abundance drive disease pressure in communities. *Nature* 520: 542–544.
Paull SH, Song S, McClure KM, Sackett LC, Kilpatrick AM, and Johnson PT (2012) From superspreaders to disease hotspots: Linking transmission across hosts and space. *Frontiers in Ecology and the Environment* 10(2): 75–82. https://doi.org/10.1890/110111.
Pedersen A and Babayan S (2011) Wild immunology. *Molecular Ecology* 20(5): 872–880.
Plowright RK, Sokolow SH, Gorman ME, Daszak P, and Foley JE (2008) Causal inference in disease ecology: Investigating ecological drivers of disease emergence. *Frontiers in Ecology and the Environment* 6(8): 420–429.
Previtali A, Hanselmann R, Jolles A, Keesing F, Martin L, and Ostfeld R (2012) Relationship between pace of life and immune responses in wild rodents. *Oikos* 121: 1483–1492.
Råberg L, Sim D, and Read AF (2007) Disentangling genetic variation for resistance and tolerance to infectious diseases in animals. *Science* 318(5851): 812–814.
Råberg L, Graham AL, and Read AF (2009) Decomposing health: Tolerance and resistance to parasites in animals. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 364(1513): 37–49.
Raffel TR, Martin LB, and Rohr JR (2008) Parasites as predators: Unifying natural enemy ecology. *Trends in Ecology & Evolution* 23(11): 610–618.
Richner H, Christe P, and Oppliger A (1995) Paternal investment affects prevalence of malaria. *Evolution* 92: 1192–1194.
Riedel S (2005) Edward Jenner and the history of smallpox and vaccination. *Proceedings (Baylor University Medical Center)* 18(1): 21–25.
Riley S, Fraser C, Donnelly CA, Ghani AC, Abu-Raddad LJ, Hedley AJ, Leung GM, Ho LM, Lam TH, Thach TQ, Chau P, Chan KP, Lo SV, Leung PY, Tsang T, Ho W, Lee KH, Lau EM, Ferguson NM, and Anderson RM (2003) Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of public health interventions. *Science* 300(5627): 1961–1966.
Rivera DL, Olister SM, Liu X, Thompson JH, Zhang XJ, Pennline K, Azuero R, Clark DA, and Miller MJ (1998) Interleukin-10 attenuates experimental fetal growth restriction and demise. *FASEB Journal* 12(2): 189–197.
Roberts ML, Buchanan KL, and Evans MR (2004) Testing the immunocompetence handicap hypothesis: A review of the evidence. *Animal Behavior* 68: 227–239.
Rohr JR and McCoy KA (2010) A qualitative meta-analysis reveals consistent effects of atrazine on freshwater fish and amphibians. *Environmental Health Perspectives* 118(1): 20–32.
Rohr J, Schotthoefer A, Raffel T, Carrick H, Hoverman J, Johnson C, Lieske C, Piwoni MD, Schoff PK, and Beasley VR (2008) Agrochemicals increase trematode infections in a declining amphibian species. *Nature* 455: 1235–1239.
Ross PS, Ellis GM, Ikonomou MG, Barrett-Lennard LG, and Addison RF (2000) High PCB concentrations in free-ranging Pacific killer whales, *Orcinus orca*: Effects of age, sex and dietary preference. *Marine Pollution Bulletin* 40(6): 504–515.
Roy BA and Kirchner JW (2000) Evolution dynamics of pathogen resistance and tolerance. *Evolution* 54(1): 51–63.
Rymuszka A, Sieroslawska A, Bownik A, and Skowronski T (2007) In vitro effects of pure microcystin-LR on the lymphocyte proliferation in rainbow trout (*Oncorhynchus mykiss*). *Fish & Shellfish Immunology* 22(3): 289–292.
Saino N, Calza S, and pape Moller A (1997) Immunocompetence of nestling barn swallows in relation to brood size and parental effort. *The Journal of Animal Ecology* 66(6): 827.
Schmid-Hempel P and Ebert D (2003) On the evolutionary ecology of specific immune defence. *Trends in Ecology & Evolution* 18: 27–32.
Schmidt KA and Ostfeld RS (2001) Biodiversity and the dilution effect in disease ecology. *Ecology* 82(3): 609–619.
Schroder M and Bowie A (2005) TLR3 in antiviral immunity: Key player or bystander? *Trends in Immunology* 26(9): 462–468.
Sears B, Rohr J, Allen J, and Martin L (2011) The economy of inflammation: When is less more? *Trends in Parasitology* 27(9): 382–387.
Sears BF, Snyder PW, and Rohr JR (2015) Host life history and host-parasite syntopy predict behavioural resistance and tolerance of parasites. *Journal of Animal Ecology* 84: 625–636.
Shankar EM, Vignesh R, Ellegard R, Barathan M, Chong YK, Bador MK, Rukumani DV, Sabet NS, Kamarulzaman A, Velu V, and Larsson M (2014) HIV-mycobacterium tuberculosis co-infection: A danger-couple model of disease pathogenesis. *Pathogens and Disease* 70(2): 110–118.
Shen Z, Ning F, Zhou W, He X, Lin C, Chin DP, Zhu Z, and Schuchat A (2004) Superspreading SARS events, Beijing, 2003. *Emerging Infectious Diseases* 10(2): 256–260.
Soares MP, Teixeira L, and Moita LF (2017) Disease tolerance and immunity in host protection against infection. *Nature Reviews Immunology* 17: 83–96.
Soper GA (1939) The curious career of typhoid Mary. *Bulletin of the New York Academy of Medicine* 15(10): 698–712.
Sousa WP (1992) Interspecific interactions among larval trematode parasites of freshwater and marine snails. *American Zoologist* 32: 583–592.
Stein RA (2011) Super-spreaders in infectious diseases. *International Journal of Infectious Diseases* 15(8): e510–e513.
Tauber AI (2003) Metchnikoff and the phagocytosis theory. *Nature Reviews* 4: 897–901.
Tauber AI (2017) *Immunity: The evolution of an idea*. Oxford: Oxford University Press.
Tella JL, Scheuerlein A, and Ricklefs RE (2002) Is cell-mediated immunity related to the evolution of life history strategies of birds? *Proceedings of the Royal Society of London B* 269: 1059–1066.
Thomas V, Anguita J, Barthold SW, and Fikrig E (2001) Coinfection with *Borrelia burgdorferi* and the agent of human Granulocytic Ehrlichiosis alters murine immune responses, pathogen burden, and severity of Lyme arthritis. *Infection and Immunity* 69(5): 3359–3371.
van Gils JA, Munster VJ, Radersma R, Liefhebber D, Fouchier RAM, and Klaassen M (2007) Hampered foraging and migratory performance in swans infected with low-pathogenic avian influenza A virus. *PLoS One* 2(1): e184.
Venesky MD, Liu X, Sauer EL, and Rohr JR (2014) Linking manipulative experiments to field data to test the dilution effect. *Journal of Animal Ecology* 83: 557–565.
Viney ME, Riley EM, and Buchanan KL (2005) Optimal immune responses: Immunocompetence revisited. *Trends in Ecology & Evolution* 20(12): 665–669.
Wood CL, Lafferty KD, DeLeo G, Young HS, Hudson PJ, and Kuris AM (2014) Does biodiversity protect humans against infectious disease? *Ecology* 95(4): 817–832.
Woolhouse MEJ, Dye C, Etard JF, Smith T, Charlwood JD, Garnett GP, Hagan P, Hii JLK, Ndhlovu PD, Quinnell RJ, Watts CH, Chandiwana SK, and Anderson RM (1997) Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. *Proceedings of the National Academy of Sciences* 94(1): 338–342.
Zahavi A (1975) Mate selection—A selection for a handicap. *Journal of Theoretical Biology* 53(1): 205–214.

Further Reading

Caldwell RM, Schafer JF, Compton LE, and Patterson FL (1958) Tolerance to cereal leaf rusts. *Science* 128(3326): 714–715.

Chandler AC, Read CP, and Nicholas HO (1950) Observations on certain phases of nutrition and host-parasite relations of *Hymenolepis diminuta* in white rats. *The Journal of Parasitology* 36(6): 523.

Ezenwa VO (2016) Helminth-microparasite co-infection in wildlife: Lessons from ruminants, rodents, and rabbits. *Parasite Immunology* 38: 527–534.

Franceschi C and Campisi J (2014) Chronic inflammation (Inflammaging) and its potential contribution to age-associated diseases. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 69(Suppl 1): S4–S9.

Freeman BM and Manning ACC (1975) The response of the immature fowl to multiple injections of adrenocorticotrophic hormone. *British Poultry Science* 16(2): 121–129.

Gervasi SS and Foufopoulos J (2007) Costs of plasticity: Responses to desiccation decrease post-metamorphic immune function in a pond-breeding amphibian. *Functional Ecology* 22: 100–108.

Karvonen M, Tuomilehto J, Libman I, and LaPorte R (1993) A review of the recent epidemiological data on the worldwide incidence of type 1 (insulin-dependent) diabetes mellitus. *Diabetologia* 36(10): 883–892.

Koski KG and Scott ME (2001) Gastrointestinal nematodes, nutrition and immunity: Breaking the negative spiral. *Annual Review of Nutrition* 21(1): 297–321.

Lessells MC (1991) The evolution of life histories. In: Krebs JR and Davies NB (eds.) *Behavioural ecology: An evolutionary approach*, 3rd edn., pp. 32–68. Oxford: Blackwell Scientific Publications.

Miller RA, Cornelius T, Morimoto RI, and Kelly JW (1996) The aging immune system: Primer and prospectus. *Science* 273(5271): 70–74.

Palacios MG, Winkler DW, Klasing KC, Hasselquist D, and Vleck CM (2011) Consequences of immune system aging in nature: A study of immunosenescence costs in free-living tree swallows. *Ecology* 92(4): 952–966.

Pugliatti M, Sotgiu S, and Rosati G (2002) Clinical neurology and neurosurgery the worldwide prevalence of multiple sclerosis. *Clinical Neurology and Neurosurgery* 104: 182–191.

Sheldon BC and Verhulst S (1996) Ecological immunology: Costly parasite defences and trade-offs in evolutionary ecology. *Trends in Ecology & Evolution* 11(8): 317–321.

Solana R, Tarazona R, Gayoso I, Lesur O, Dupuis G, and Fulop T (2012) Innate immunosenescence: Effect of aging on cells and receptors of the innate immune system in humans. *Seminars in Immunology* 24(5): 331–341.

Waller LA (2008) Exploratory spatial analysis in disease ecology. In: *Encyclopedia of GIS*, pp. 297–301. Boston, MA: Springer US.

Williams GC (1957) Pleiotropy, natural selection, and the evolution of senescence. *Evolution* 11(4): 398–411.

Ecological Niche☆

Jitka Polechová, University of Tennessee, Knoxville, TN, United States; BioSS, Edinburgh, United Kingdom; University of Vienna, Vienna, Austria
David Storch, Charles University, Prague, Czech Republic

Introduction

Ecological niche is a term for the position of a species within an ecosystem, describing both the range of conditions necessary for persistence of the species, and its ecological role in the ecosystem. Ecological niche subsumes all of the interactions between a species and the biotic and abiotic environment, and thus represents a very basic and fundamental ecological concept. The tentative definition presented above indicates that the concept of niche has two sides which are not so tightly related: one concerns the effects environment has on a species, the other the effects a species has on the environment. In most of ecological thinking, however, both meanings are implicitly or explicitly mixed. The reason is that ecology is about interactions between organisms, and if persistence of a species is determined by the presence of other species (food sources, competitors, predators, etc.), all species are naturally both affected by environment, and at the same time affect the environment for other species.

If we want to treat both of these aspects of ecological niche within one framework, we can define it more formally as the part of ecological space (defined by all combinations of biotic and abiotic environmental conditions) where the species population can persist and thus utilize resources and impact its environment. It is useful, however, to distinguish three main approaches to the niche. The first approach emphasizes environmental conditions necessary for a species presence and maintenance of its population, the second approach stresses the functional role of species within ecosystems, and the third one a dynamic position of species within a local community, shaped by species' biotic and abiotic requirements and by coexistence with other species.

Concepts of Niche

Niche as the Description of a Species' Habitat Requirements

The first formulations of the concept of an ecological niche were close to the general meaning of the term: the ecological niche was defined by a place a species can take in nature, determined by its abiotic requirements, food preferences, microhabitat characteristics (e.g., a foliage layer), diurnal and seasonal specialization, or predation avoidance. This concept is associated mostly with Joseph Grinnell, who first introduced the term. He was especially interested in factors determining where we can find a given species and how niches, generated by the environment, are filled. The knowledge of a species niche determined by its habitat requirements is essential for understanding and even predicting its geographic distribution; this concept of the niche is thus more relevant in biogeography and macroecology than in community or ecosystem ecology.

Niche as Ecological Function of the Species

In this concept of a niche, each species has a particular role in an ecosystem and its dynamics, and one such role can be fulfilled by different species in different places. The observation of distant species adapted to equivalent ecological roles (the resemblance

☆*Change History:* February 2018. Polechová updated Figure 1, minor changes throughout (ambiguities, English - suggestions by Swati Patel), second paragraph in Evolution of Niche Width ("swamping" by gene flow) and checked by Polechová and Storch.

between jerboa and kangaroo rat, between many eutherian and marsupial species, or the Galapagos finches diversifying to highly specialized roles including those normally taken by woodpeckers) was clearly influential to Charles Elton, who emphasized the functional roles of species. According to Elton, there is the niche of burrowing detritivores, the niche of animals specializing in cleaning ticks or other parasites, or the pollination niche. Elton's niche can apply to several species, for example, "the niche filled by birds of prey which eat small mammals". This *functional niche* therefore refers to a species position in trophic chains and food webs, and the concept is thus especially relevant for ecosystem ecology.

Niche as a Species Position in a Community: Formalization of Ecological Niche Concept

The emphasis on the diversity of ecological communities and interspecific competition within them in the second half of 20th century has lead to the formalization of niche concept, and an emphasis on the properties of the niches which enable species coexistence within a habitat. George Evelyn Hutchinson postulated that niche is a "hypervolume" in multidimensional ecological space, determined by a species requirements to reproduce and survive. Each dimension in the niche space represents an environmental variable potentially or actually important for a species' persistence. These variables are both abiotic and biotic, and can be represented by simple physical attributes as temperature, light intensity or humidity, but also more sophisticated attributes such as soil texture, ruggedness of the terrain, vegetation complexity or various measures of resource characteristics. This could be viewed simply as a formalization of original Grinnellian niche, that is, the exact descriptions of a species habitat requirements. However, in the Hutchinsonian view ecological niches are dynamic, as the presence of one species constrains the presence of another species by interspecific competition, modifying the position of species' niches within the multidimensional space. This concept therefore combines the ecological requirements of the species with its functional role in the local community.

Fundamental and Realized Niche

Hutchinson recognizes a species' *fundamental niche*, a multidimensional 'cloud' of favorable conditions determined by all environmental (abiotic and biotic) variables where the species can reproduce and survive, and the *realized niche*, which is a subset of the abstract fundamental niche, where the species can persist given the presence of other species competing for the same resources. Thus, a realized niche always has a narrower extent along respective dimensions; a species which could potentially live in a broad range of humidity conditions, for instance, may occupy a much narrower range of these conditions in an environment with competing species, since its population growth rate decreases to negative values in some conditions. To a good approximation, if we ignore stochastic sampling from a heterogeneous species' population, species does fill its realized niche.

According to Hutchinson's formalization, niches of different species can be separated along any of these dimensions or by a combination of them (i.e., their interaction) (Fig. 1). Although this formal model of the niche has quite straightforward theoretical

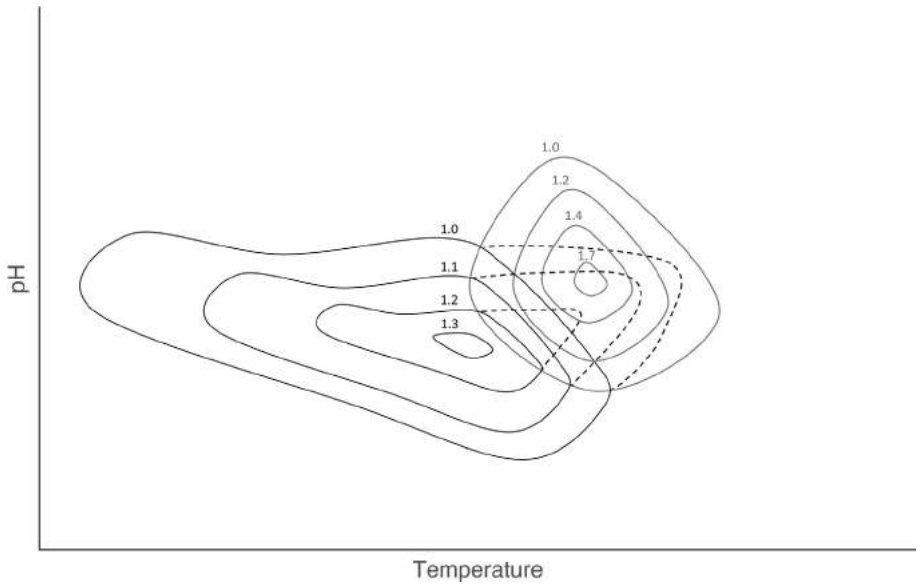


Fig. 1 A hypothetical example of species niches and their interaction. Two niche axes are shown, pH and temperature. The two species of hypothetical protists have different optima, and one (left) is adapted to wider range of conditions and has overall lower growth rate (measured at a given low density). When the niches of the two species overlap, the growth rates are expected to decrease, potentially to the point where the population cannot be sustained. Here, one of the species if fully dominant in its niche. Therefore, the other species can sustain its population only on a part of its fundamental niche, as its growth rate decreases at the overlapping areas. The growth rate isoclines are shown, with the dashed lines depicting the growth rate isoclines of the fundamental niche (i.e., had the other species been absent).

consequences, in practice it can be quite difficult to describe properly the ecological niches of real species, because the number of niche dimensions is potentially infinite, and the significant niche axes (and appropriate measurements) may be rather hard to find; a niche overlap among species may mean we did not succeed in determining the crucial niche axes of separation. However, often a few variables are sufficient to separate species' realized niches, and they or their correlates can be inferred assuming we understand the species' biology reasonably well. For example, five species of warblers, analyzed by Robert MacArthur, showed significant (though incomplete) separation along only three niche axes (feeding behavior, feeding height and nesting time).

The difficulties in determining appropriate niche axes, however, still considerably limit the usefulness of the concept in empirical research. Even if we know the important resources, it is still problematic to decide which characteristics to measure. A further problem, albeit rather technical, is posed by including discrete categories: the width of the cloud in the respective dimension would be reduced to zero, and its position can be arbitrary. More importantly, although species can often potentially live in a much broader range of environmental conditions than where they do actually live, the distinction between the "fundamental" and "realized" niche is slightly arbitrary, driven purely by the interest in coexistence of species sharing resources. As the dimensions of the fundamental niche are both abiotic and biotic, there is no a priori reason to exclude presence of competitors from the dimensions characterizing fundamental niche. The distinction between the fundamental and realized niche may also be blurred, as species' interactions need not fit to our discrete categories—for example, competitors may act also as predators.

Due to the difficulties with the concept, and for tractability, a considerable part of the theory actually dealing with species coexistence works with a one-dimensional approximation of the *trophic niche, a resource utilization function*—given by the frequency distribution of an important characteristic of utilized resource (e.g., a prey size).

Competitive Exclusion

Historical development of the niche theory is very closely related to one of the most important topics of ecology, that is, the problem of species competition and coexistence. Since the beginning of the ecological niche concept, it has been assumed that no two species sharing a single niche could locally coexist. Originally, the later *Volterra–Gause principle* states that "under constant conditions, no two species utilizing, and limited by, a single resource can coexist in a limited system" and was formulated and proved by Vito Volterra, whilst Alfred Gause showed experimental evidence of competitive exclusion in an undiversified environment. The explanation of the competitive exclusion lies in the fact that utilization of a limited resource leads to its depletion, and the population growth, therefore, necessarily leads to a moment when the resource level is insufficient for further growth. If only one population utilizes the resource, this situation leads to simple negative feedback, causing the decrease of population growth rate and thus a release of resource consumption, stabilizing the population size. However, in the case of two species sharing the resource, there will likely exist a resource level when the first species can still grow up even if the second cannot, leading to further decrease of population growth rate of the second species, and eventually to its extinction. Even if two species sharing several resources have exactly the same requirements and ability to utilize them, the coexistence of such species is not stable due to environmental or demographic stochasticity: over time, one of the species would ultimately become extinct by chance.

The "competitive exclusion principle" is the core principle in community ecology, and much of this field has been devoted to study how species with similar ecological requirements can coexist. This question has transformed into the problem of "limiting similarity": how similar can ecological niches be to still ensure local coexistence.

Limiting Similarity, Species Packing: How Close Can Species Be to Each Other?

Hutchinson states that a species' realized niche is exclusive, that is, no two species can share a single niche and no overlap in the realized niches is possible in a stable environment. In other words, were there to be an overlap in, say, the trophic "dimension" of the niche, species would differ in other dimensions—for example, in their tolerance to abiotic factors, or avoidance of predators. Now, the (rather vague) consensus is that a little overlap between niches is consistent with coexistence, whereas somewhat larger overlap is not. The theory of *limiting similarity*, formalized by Robert MacArthur and Richard Levins, predicts the minimum permissible degree of overlap in the resource utilization curve. They showed that coexistence between species utilizing a continuous resource is possible when the ratio between the niche width (see Box 1) and the distance between species' optima is approximately unity or smaller. (This has been derived using the Lotka–Volterra equations describing the growth rates and hence stability of populations of competing species, where the competition coefficients were determined by the proximity of species' bell-shaped utilization curves.) However, the result is sensitive to the assumptions about the form of the resource utilization function and population growth rate; notably, highly peaked resource utilization functions show actually almost no limits to coexistence (as their overlap is always minute) and niches can overlap broadly when fitness increases as the frequency of individuals carrying the respective trait decreases—under negative frequency-dependence. Also, coexistence between species can be facilitated by nonlinear responses of the competitors to common fluctuations in the environment. Note that the predictions of the theory of limiting similarity cannot be directly corroborated by observation: by definition, the population density of one of the species is close to zero if the species pair is close to limiting similarity, and thus the utilization functions are not observable in such a situation. On the other hand, finding a similarity higher than predicted would clearly indicate that some of the assumptions of the model are violated.

The spacing between species in niche space, resulting from partitioning the available resources (*species packing*), differs considerably between sexually and asexually reproducing species. In sexually reproducing populations, recombination generates

Box 1 Niche width

Niche width describes the dispersion of population resource use along a niche dimension. As such, it is very laborious to measure: more often, we get estimates of niche width from the morphological traits related to the resource use: for example, beak dimensions, jaws or teeth size. However, this measure delivers only a part of the information: both phenotypic variation in the traits important for food gathering and the ability of an individual to exploit a range of resources generally contribute to the niche width. For example, the niche breadth of *Anolis* lizards, studied by Joan Roughgarden, is mostly determined by variation in jaw size within species, but any individual still contributes to the total niche width, having its own range of prey sizes. Importantly, Rougharden shows that a measure of the total niche width can be calculated as a sum of a *within-phenotype component*, the average variance of the individual's utilization function, and a *between-phenotype component*, the variance in population resource utilization function. Often, the range of two standard deviations (twice the square root of the sum), comprising about 95% of resource used, is denoted as the niche width.

The related term *niche breadth* is originally due to Richard Levins. Levin's measures of niche breadth reflect the diversity of species' use of available resources: niche breadth is determined by the Shannon index (i.e., information entropy), or Simpson's index (i.e., the inverse of the sum of squared frequencies of the focal species over all resources). Although niche breadth intuitively captures differences between generalists and specialists, the measure is very sensitive to the categorization of resources and their frequency distribution.

all possible combinations from the common gene pool, although some of these may be maladaptive. Trade-offs in utilizing the resource spectrum may—or may not—generate disruptive selection strong enough to drive evolution of reproductive isolation and evolution of distinct species, thus eliminating maladapted recombinant genotypes. In contrast, in asexual species, clones bearing favorable combinations do not recombine, and therefore those adapted to the various resource combinations can be arbitrarily spaced in the niche space. Due to the necessity of finding a mating partner, population growth rate of sexual populations can sharply decrease at low densities (*Allee effect*), limiting both adaptation to marginal conditions and invasion to a new area. Both these effects contribute to discontinuities in distribution of resource use of sexually reproducing species.

Modes of Species Coexistence

Species coexistence is often ensured by niche separation. The *niche shift* can follow from the competitive exclusion of one species from the part of ecological space where the niches overlap, or from coevolution of competing species, favoring in each species phenotypes differing from the phenotype of the competitor. Current niche segregation can be due to the processes that took place in distant evolutionary past—sometimes emphasized by the term "the *ghost of competition past*". When phenotypic differences arose due to divergent evolution of sympatric competitors, we speak of *character displacement*. Typically, sympatric populations of competing species evolve towards more different sizes of characters associated with food consumption (beaks, teeth) than allopatric populations – if there is an island with only one species of Galapagos finches, it has an intermediate beak size enabling it to utilize a wide spectrum of seed sizes, whereas if there is an island with two species, one has a bigger and the other a smaller beak than the species occurring without competitors. If there are more than two locally coexisting species, we often observe regularly spaced sizes of morphological characters, again indicating past competition leading to maximum niche separation.

Simple separation of niche optima is not, however, the only way that stable local coexistence of species is attained. Many species pairs, for instance, consist of one species which is competitively dominant, and the other species which is less specialized and can thrive in a broader range of ecological conditions. An example is the pair of two closely related species of redstarts, where the black redstart *Phoenicurus ochruros* is bigger and more aggressive, but the common redstart *Phoenicurus phoenicurus* can utilize a wider spectrum of habitats, such that it always has an option to thrive outside of the range of conditions preferred by the black redstart. Such niche division between dominant aggressive specialist and subordinate generalist has also been observed in many mammal species, and is apparently stable. In plants, competitively inferior species are often those with higher rates of spreading and growth, which enable them to quickly occupy empty places before arrival and eventual overgrowth of competitively superior species. In this case, we speak about the *regenerative niche*, representing a time window for the success of competitively inferior, but fast spreading and fast growing species, thus ensuring long-term coexistence of competitors in the same habitat.

If species are very similar to each other, such that they do not differ substantially in their utilization of resources, the competitive exclusion can take a very long time. If the replacement of old individuals by young ones is basically a random process, that is, all individuals regardless of species identity have equal chances to give birth to their descendants within an environment, populations of all involved species will fluctuate randomly and the prevalence of a particular species is just a matter of chance. However, due to these stochastic fluctuations and due to the fact that the species which incidentally prevails in a time step will have higher probability to further increase its abundance, this process will finally lead to apparent competitive exclusion. This process, called *community drift*, can be relatively slow and may be further slowed down by dispersal limitations (leading to random prevalence of different species in different local communities isolated by migration barriers) and balanced by the emergence of new species (i.e., speciation or migration from elsewhere).

Communities where dispersal limitation and community drift play a major role are called *dispersal-assembled communities*, in contrast to *niche-assembled communities* where niche differences play a major role in determining species distributions and abundances. Trees in tropical forests represent a very good candidate for dispersal-assembled communities. Most tropical tree

species are very similar in terms of their ecology and growth characteristics, and it has been documented that for their recruitment the proportion of parent individuals in a given locality (i.e., dispersal limitation of more distant individuals) is much more important than any habitat characteristics. Still, an incredible number of species can coexist locally. It is hardly believable that there are several hundreds of different narrow ecological niches (i.e., combinations of environmental characteristics) on a hectare of tropical forest to enable coexistence of several hundreds tree species on the basis of their niche differences—the dispersal assembly and coexistence without significant niche differentiation seems more likely. However, an unusual aspect of niche differences can still be involved in this classical case of species coexistence. It has been demonstrated that coexistence of tropical trees is facilitated by frequency-dependence, where relatively rare species have an advantage of not being so severely attacked by natural enemies which strongly limit recruitment of more common species on which they specialize. In a sense, all species compete for "enemy-free space", and this "niche" for a given species is open only if the species is not too abundant to allow population growth of specialized natural enemies. Separation of "niches" of tropical trees seems thus to be determined by the community of species-specific pathogens.

In conclusion, coexistence among species can be certainly maintained both by niche differences and—at least in a none-quilibrium world—by niche similarity. Coexistence of species with similar niches maintained by dispersal-assembly processes could be a reason why we often observe that species are not regularly distributed in a niche space, but form clumps of species whose niches are closer to each other than to other species.

How Many Ecological Niches Are There?

The notion that ecological niches cannot be infinitely similar to each other, and the knowledge that ecological space is heterogeneous and that the distribution of resources available to a community is always limited, has led to an idea that for a given environment there is a limited number of available niches which could be potentially occupied. An environment then could be seen as a set of empty niches, which could—but may not—be filled with species. Consequently, we might ask whether in a particular case the niche space is or is not saturated with species.

There are two facets of the problem, which are sometimes confused. First, there is no doubt that the limited amount of resources in an ecosystem can sustain only limited total number of individuals (assuming a given body size distribution). Therefore, there is always a limited potential for the whole community size determined by total amount of resources, and thus also for a limited number of species (given that each species needs some viable population size). If this potential is fully utilized, we speak about biotic saturation of the community. However, biotic saturation does not imply that the number of ecological niches is fixed and that all possible niches are occupied. Such a statement would be much stronger and would require at least some level of discreteness of ecological niches, that is, that ecological space cannot be divided into an infinite number of subtly different niches with arbitrary positions. Is there any reason to believe that niches are discrete and their number within an environment is limited?

Apparently, there is a considerable level of environmental heterogeneity in resource distribution and abundance; resources are more abundant for some combinations of parameters than for other. Environmental heterogeneity would not be, however, a sufficient condition for discreteness of ecological niches if species could utilize equally easily several different resources. The discreteness of ecological niches comes out from the existence of *trade-offs* in resource utilization: resources can always potentially be utilized in many ways, but a species which utilizes many resources typically does so with a lower efficiency compared to a specialist. A Galapagos finch from the genus *Geospiza* can have either a big beak appropriate for cracking big seeds, but then it can crack small seeds with much more difficulty—and vice versa. Some strategies in resource utilization are mutually exclusive: a plant can either invest to its rapid growth and so quickly utilize resources, or it can invest into woody trunk which enables it to grow higher and sustain longer—but at a cost associated with slower growth. Whilst natural selection supports phenotypes that are better at utilizing available resources, it can only explore the vicinity of the current strategy, selecting from the available phenotypes. Consequently, evolution leads to utilization of only a restricted spectrum of resources.

In the presence of trade-offs, there is only a limited number of mutually exclusive ways to utilize resources, and thus a limited number of available niches. However, as the discreteness of niches follows from the trade-offs between adaptations, and since all the trade-offs are determined by unique properties and constraints of given organisms, it makes sense to speak about available niches only in relation to organisms which already inhabit the environment. A habitat without its inhabitants can provide a potentially infinite number of opportunities for existence, and this landscape of opportunities changes with each new inhabitant. For the organism in an environment, the number of possible niches is determined by the number of possible ways to utilize the resource—with all constraints and trade-offs of the given organism. Therefore, it is likely that there are always more niches than the current number of species, because each species has several mutually exclusive possibilities of future adaptive evolution arising from the trade-offs—unless all niche changes require a corresponding niche change in other species.

In some cases, the number of available niches can be predicted from the knowledge of resource heterogeneity and the possibilities of resource utilization for given taxon. The number of Galapagos finches occurring on each island is reasonably well predicted by the number of peaks of the "landscape" constructed using the knowledge of frequency distribution of seed size, the general relation between finch and seed biomass, and the relation between preferred seed size and beak depth (Fig. 2). Similarly, using the knowledge of the relationship between beak shape of crossbills (*Loxia curvirostra*) and their foraging efficiency in obtaining cone seeds from cones of various coniferous tree species, it is possible to construct a resource utilization function related to different morphologies, and find out how many optimal shapes do exist. And again, it has been found that there are several ecomorphs of crossbills, each of them occupying one adaptive peak (optimum) in the morphological space.

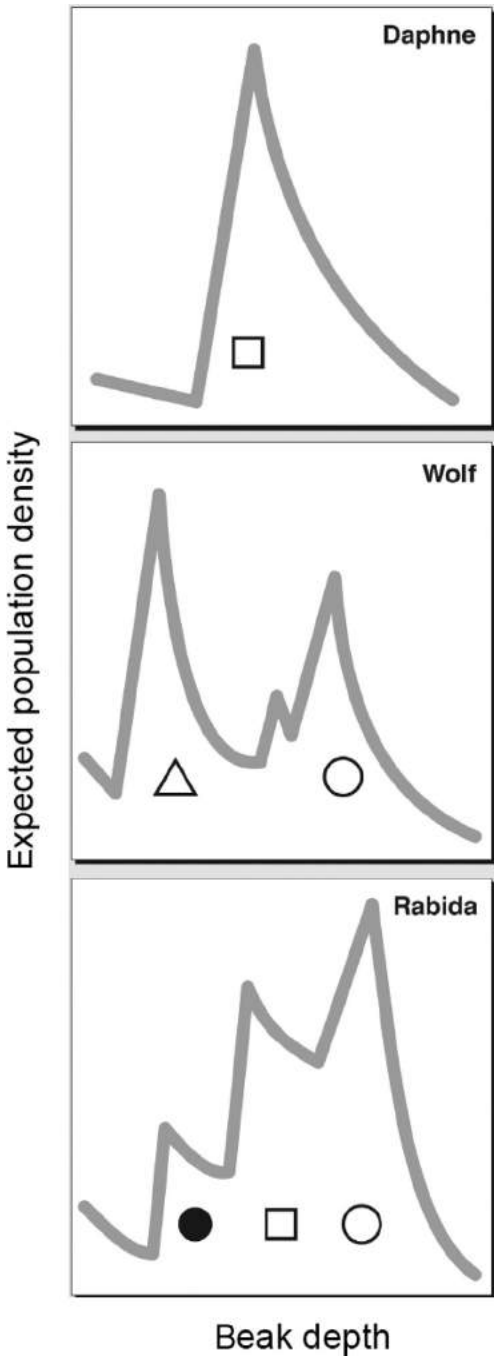


Fig. 2 In this example of Galapagos finches on three different islands, the number of niches can be predicted from the peaks in the expected finch density. The expected finch density is calculated from distribution of seed biomass converted to finch numbers, using preferred seed size estimated from the mean size of the beak. The beak depth of the finches occurring on each islands corresponds well to the maxima of the curve. Position of the symbols marks mean beak depth of male ground finch on each of the three islands: *Geospiza fortis* (□), *G. difficilis* (△), *G. magnirostris* (○) and *G. fuliginosa* (●). The beak depth scale is kept the same for the three pictures; the population density is scaled to the maximum. Modified from Schluter, D. and Grant, P. R. (1984). Determinants of morphological patterns in communities of Darwin's finches. *American Naturalist* **123**, 175–196.

There is other evidence that ecological niches are partially predictable—the phenomenon of community convergence. Animal or plant communities occurring on different continents or biotic provinces often comprise similar morphological types utilizing similar types of resources. *Anolis* lizards, for instance, have evolved independently into several well-recognizable ecomorphs on each Caribbean island, with known sequence of this evolution, repeated on every island. However, there can be more than one species within each ecomorph, and thus this convergence does not imply that the number of species-specific niches is predictable. This is quite typical for most cases of community convergences: they provide a clue to our understanding of how many possibilities are

there for utilizing resources within a given habitat and for a given taxon, but not to the prediction of how many species can actually coexist there. The total potential number of species within an environment is given by the total amount of resources determining the total number of all individuals, regardless the level of discreteness of ecological niches.

Ecological Niches and Patterns in Species Abundance and Distribution

Species spatial distributions as well as their abundances are often attributed to the breadth and position of their niches. A species occurs in places where its requirements are fulfilled, that is, where it finds its niche. However, the "presence of the niche" is not a sufficient condition for the presence of a species, and in special cases it may not be even the necessary condition. Spatial population dynamics driven by dispersal and spatial distribution of available habitat patches is equally important. Consequently, species may be absent even in sites containing habitat that fulfills its niche requirements if the site is far away from other occupied sites and the dispersal distance of the organism in concern is relatively small, hindering immigration into the site. On the other hand, a species may be present even in a site where its niche requirements are not fulfilled and population growth is negative if the population is maintained by a continuous supply of individuals from neighboring sites with positive population growth (so-called source-sink population dynamics). Therefore, species spatial distributions are determined by species niches and available habitat distributions, as well as by spatial population dynamics and dispersal limitation.

In a similar line, it has been argued that a significant proportion of the variation of species' abundances can be explained by the breadth of species' niches (Box 1). It is reasonable to assume that species which are able to utilize wider spectrum of resources can attain higher population abundances and also can occupy more sites. Local population densities are mostly positively correlated with species range sizes, which can be taken as evidence of such niche differences. However, patterns in species abundances can be often well explained by spatial population dynamics—for instance species which were incidentally able to spread to more sites have higher chance to colonize further sites and to further increase local population densities by immigration (this is the nonlinearity of the dynamics of metapopulations). Moreover, the statistical relationship between niche breadth and abundance can actually have a reversed causality, as abundant species are forced to utilize a wider range of resources due to intraspecific competition. More abundant species can also be those that do not utilize a broader range of resources, but are specialized on resources which are relatively more abundant, or may simply have higher population growth and/or dispersal rate (although these features can be understood as niche properties).

One of the most prominent ecological patterns is the frequency distribution of abundance of individual species within local communities or regional species assemblages—the so called species-abundance distribution. It is always highly unequal, the majority of species having low abundance and only a few being common (the frequency distribution is often close to lognormal, though other models may fit the observed species-abundance distribution better in particular situations). This distribution has been modeled as a stepwise division of niche space, where each newly arriving species obtains some (random) proportion of niche space previously utilized by other species. One of these models, based on sequential resource partitioning, predicts observed species-abundance distribution quite well (Box 2). However, models based on spatial dynamics and dispersal limitations—especially those involving "community drift"; see above—can provide equally good predictions of species-abundance distribution. This again indicates the complementarity between niche-based and dispersal-based explanation of ecological patterns, and supports our consideration of both niche differences and spatial population dynamics as essential drivers of species distribution and abundance.

Niche Divergence and Resource Specialization

The diversity of ecological niches even among closely related species is enormous and demands explanation. What is the reason for such diversity? We have already mentioned one of the most important factors—interspecific competition, which pushes ecological niches of species far away, to avoid niche overlap. More specifically, natural selection prefers such phenotypes of competing species which utilize different resources than those which share them. Competition thus leads to the increase of the resource range utilized by a given taxon, and this process is faster when other taxa do not constrain this diversification. Indeed, the increase of the breadth of utilized resources in the course of evolution is fastest in such situations where other taxa with similar requirements are absent. For example, ecomorphological diversification of Galapagos finches and Hawaiian honeycreepers has been much faster than the

Box 2 Sequential resource partitioning

It appears that relative species abundances within taxa can be reasonably well explained by a simple null model of resource partitioning between species, proposed by Mutsunori Tokeshi. A common resource, represented by a "stick", is divided once at a random location chosen uniformly along its length, and for further partitioning one part is chosen with a probability proportional to its length raised to a power of K, where K is a parameter between 0 and 1 (e.g., 0.05), and the division and selection process continues to distribute the "niche" among all the species within the taxon. The model seems to describe well relative abundances of species within taxa, across a large range of their species richness.

diversification of related taxa on the mainland, where the utilization of new resources was constrained by other taxa already utilizing them.

Interspecific competition is not, however, the only force driving niche diversification. Each species has its own evolutionary history, and thus can adapt to different resources by an independent process of evolutionary optimization, as phenotypes which are more efficient in transforming obtained energy into offspring are favored by natural selection. If there are several mutually exclusive ways to achieve this, it is likely that each species will go by a different route due to evolutionary contingency, and niche diversification will follow without competition. Notably, optimization does not lead to an advantage of the whole species in terms of the resource utilization but only to an individual advantage regardless of the evolutionary fate of the whole species: when the niche becomes narrower, the species' range and hence total population size decreases. As evolution is opportunistic, species can evolve to extremely specialized forms in terms of either habitat utilization or food preference, which is apparently disadvantageous for future species persistence in ever changing world.

Evolution of Niche Width

Progressive specialization, that is, narrowing of niche width in the course of evolution, is forced by interspecific competition (when the niches overlap) and intraspecific optimization, and thus represents an expected evolutionary trend. The opposite process, that is, an extension of niche width, is observed mostly after entering a new environment without competitors, allowing utilization of a wider spectrum of resources. This process is called *ecological release* and may be underlined both by the extension in within- and between-phenotype component of species ecological variation (importance of the two contributing modes vary widely among species). Species niches can widen also because of *phenotypic plasticity* (heritable genotype–environment interactions directing the trait in the early ontogenesis), and can vary even purely behaviorally, as an immediate response to an altered resource or species structure.

Although sometimes there is an obvious constraint on expanding a species' niche—for example, physiological constraints like freezing of body fluids or presence of a competing species—we often see no apparent reason why species niches stay restricted to a fraction of a resource which continuously varies in space. However, even a gradual change in the environment can lead to a sharp range margin, as "swamping" by gene flow from central to marginal areas can create a positive feedback between the reduction in fitness, where population size decreases due to maladaptation, and the erosion of genetic variance by genetic drift. When local genetic variance is too low, continuous adaptation to a spatially variable environment fails. Moreover, when there is a large asymmetry in the carrying capacity across the habitats, even alleles which neutral or nearly neutral in the main population and deleterious in the marginal populations, can sweep through the small marginal population, thus preventing the adaptation.

Asexual reproduction or self-fertilization can hence provide an advantage in adapting to marginal conditions—both because small populations are still viable (as there is no need to find a mating partner) and because gene flow does not restrict adaptation to marginal conditions. Indeed, it is found in many plants and animals adapting to extreme, marginal habitats (classic animal examples are *Daphnia pulex* or freshwater snail *Campeloma*). However, although lack of recombination in asexuals means that locally favorable gene combinations are maintained, adaptive evolution in asexual species is significantly slowed down as beneficial combinations have to arise in each strain independently. It appears that high levels of, but not obligatory, self-fertilization or asexual reproduction (parthenogenesis and vegetative reproduction) are commonly advantageous for adaptation to marginal habitats.

In some cases, we observe an apparent regularity in the evolution of niche width and position. The classical example represents cycles of species dispersal, specialization and local adaptation (and eventual extinction) observed on various archipelagoes, called *taxon cycles*. They were originally described by Edward O. Wilson on Melanesian ants, but were best documented by Robert Ricklefs and his coworkers on Caribbean birds. In the first stage, an immigrant, which is mostly a species with a high dispersal ability, colonizes coastal or disturbed areas. Then the species spreads across the island, adapting to the new resources and expanding its niche (quite likely as a consequence of a release from competitors, predators and parasites). In the next step the species becomes more specialized, and its distribution becomes spottier. The narrowing of its niche may be driven by an immediate advantage of an adaptation to a local resource or immigration of new generalist competitor. Finally, species distribution becomes very fragmented, ending in local endemism, and ultimately extinction.

Changing the Niche Space, Niche Construction and Coevolution

Both environment and species change in the course of time, and thus ecological niches are not stable and given forever. Species not only respond to environmental changes, but also actively change their biotic and abiotic environment, affecting both their own niche and the niches of other organisms. The importance of competition, predation or mutualism has been already stressed. Moreover, organisms often make niche space for other organisms available in the environment—think of the resource space generated by an emerging tree or of successive colonization of an island, where first colonizers modify the environment for their successors. Some organisms strongly directly affect abiotic environment, determining possible niches for a whole community of species. Beavers building dams, earthworms altering soil structure or, on a larger scale, plants providing oxygen are classical examples. Organisms substantially affecting abiotic environment are often called *ecosystem engineers* and the process in which an

organism systematically modifies its own niche (both biotic and abiotic components), is called *niche construction*. Obviously, this process is most pronounced in *Homo sapiens*, which is currently the most conspicuous ecosystem engineer.

If a species can change its environment as well as to adapt to it, coevolution between a species and its niche can follow, based on the continuous feedback between a species' *niche construction* and its adaptation. Since species continuously change the environment for themselves as well as for other species, species' niches can be very dynamic. Often, however, species' ecological requirements are quite stable over evolutionary time, so that it is even possible to reliably reconstruct an ancient environment on the basis of presence of particular species in the fossil record and the knowledge of their contemporary ecological niches. This can be attributed to the fact that it is often easier to search for appropriate habitat elsewhere if an environment within a locality is changing than to adapt to it. All species have some dispersal abilities, and are thus able to track spatiotemporally changing habitat availability by migration rather than adapt to different conditions by mutation-selection process. The other reason for the apparent niche conservatism is the existence of evolutionary constraints and consequent trade-offs: species often cannot easily change their traits in a particular direction if these traits are associated with other traits whose change is not advantageous.

As species undertake evolutionary changes, their functional niches can change, leading to changes in the overall "ecological space" in an ecosystem, and promoting further changes in species traits. On the other hand, some functional niches, that is, particular ecological roles, can be rather stable even if species evolve, go extinct and new species emerge—similar functional role can be progressively fulfilled by different species. Community evolution can be therefore viewed as a coevolution of ecological niches rather than of species themselves.

Further Reading

Abrams P (1983) The theory of limiting similarity. *Annual Review of Ecology and Systematics* 14: 359–376.
Chase JM and Leibold MA (2003) *Ecological niches: Linking classical and contemporary approaches*. University of Chicago Press.
Elton C (1927) *Animal ecology*. London: Sidgwick & Jackson.
Grinnell J (1917) The niche-relationships of the California thrasher. *The Auk* 34: 427–433.
Hutchinson GE (1958) Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology* 22: 415–427.
Hutchinson GE (1959) Homage to Santa Rosalia, or why are there so many kinds of animals? *American Naturalist* 93: 45–159.
Losos JB (1992) The evolution of convergent community structure in Caribbean Anolis communities. *Systematic Biology* 41: 403–420.
MacArthur RH (1958) Population ecology of some warblers of northeastern coniferous forests. *Ecology* 39: 599–619.
MacArthur R and Levins R (1967) The limiting similarity, convergence, and divergence of coexisting species. *American Naturalist* 101: 377–385.
Odling-Smee FJ, Laland KN, and Feldman MW (2003) *Niche construction: The neglected process in evolution*. Princeton: Princeton University Press.
Ricklefs RE and Bermingham E (2002) The concept of the taxon cycle in biogeography. *Global Ecology and Biogeography* 11: 353–361.
Roughgarden J (1974) Niche width: Biogeographic patterns among anolis lizard populations. *American Naturalist* 108: 429–442.
Schluter D (2000) *The ecology of adaptive radiation*. Oxford: Oxford University Press.
Schluter D and Grant PR (1984) Determinants of morphological patterns in communities of Darwin's finches. *American Naturalist* 123: 175–196.
Schoener T (1989) The ecological niche. In: Cherrett J (ed.) *Ecological concepts*, pp. 79–113. Oxford: Blackwell Scientific Publications.
Tokeshi M (1999) *Species coexistence: Ecological and evolutionary perspectives*. Oxford: Blackwell Science.

Endemism

JJ Morrone, UNAM, Mexico City, Mexico

Introduction

The term 'endemism' refers to a taxon restricted to a particular geographical area of the world. Such taxon is said to be 'endemic' to that area. Areas where distributional areas of two or more taxa overlap are called 'areas of endemism'. The concept of endemism dates back to Augustin Pyramus de Candolle's *Geographie Botanique*, in the early nineteenth century. The biogeographical, evolutionary, and conservation perspectives of endemism are explored herein.

What Is Endemism?

Endemism describes taxa that are distributed on particular areas. It represents a basic feature of geographic distributions: species are rarely cosmopolitan and most species and even supraspecific taxa are confined to restricted regions. Endemism occurs on a variety of spatial scales, from areas as large as continents to small areas as islands or mountain tops. The puma (*Puma concolor*) is endemic to the Americas, occurring from Canada to Patagonia. Several plant families are endemic to the Neotropics. The volcano rabbit or teporingo (*Romerolagus diazi*) is endemic to some volcano tops (2800–4250 m) from the central part of the Transmexican Volcanic Belt (Mexico). Many researchers in the past focused on species endemic to small areas and thus endemism has been incorrectly associated with 'rarity'.

Organisms can be endemic on different taxonomic levels; usually the size of the area depends on the category of the taxon, with genera having larger areas than species, and families having larger areas than genera. This situation, however, is not comparable between different taxa: the distribution of a plant species may correspond to the distribution of an insect family. Conventionally, some authors have used the term 'endemic' for taxa restricted to a single biogeographical region, 'characteristic' for those taxa shared by two regions, 'semicosmopolitan' for taxa inhabiting three or four regions, and 'cosmopolitan' for taxa inhabiting five or more regions.

The restriction of taxa to particular areas is a consequence of both historical and ecological factors. Historical events are invoked to explain how a taxon became confined to its present range. Vicariant events caused by drifting continents, long-distance dispersal, and extinction are some of these events. On the other hand, ecological explanations are invoked to explain the present limits of endemic taxa. Abiotic (temperature, altitude, soil) and biotic factors are commonly considered.

Classification of Endemism

Endemic taxa can be classified into some categories, based on their distribution, origin, age, and taxonomy.

Autochtonous endemics. Taxa that evolved in the areas where they are currently found.

Allochtonous endemics. Taxa that evolved in a different area from where they are found today.

Taxonomic relicts. Taxa that are the sole survivors of a once diverse group.

Biogeographic relicts. Taxa that are the narrowly endemic descendants of a once widespread taxon. They are also known as living fossils or narrow endemics. Tuataras (*Sphenodon punctatus* and *S. guntheri*), endemic to New Zealand, are a good example of biogeographic relicts.

Neoendemics. Taxa that have evolved relatively recently and may be restricted in their distribution because they have not had yet time to disperse further. The plant species *Aquilegia barbaricina* (Ranunculaceae), which grows only along water courses at 1300–1400 m in Sardinia, is an example of a neoendemic taxon.

Paleoendemics. Taxa that have a long evolutionary history and usually are restricted by barriers to dispersal or by extensive extinction in the remaining areas where they were distributed in the past. *Bupleurum dianthifolium* (Apiaceae), which today grows only in the small island of Marettimo, west of Sicily, but was widespread in mountains of the Mediterranean area when there were more tropical conditions, represents a paleoendemic species.

Areas of Endemism

Areas where the distributional areas of two or more taxa overlap are called areas of endemism. The congruence of the distributional areas of different taxa gives identity to areas of endemism. It is interpreted as primary biogeographical homology, which means a conjecture on a common biogeographic history, namely, they belonged to the same ancestral biotic component.

If we map the distributional ranges of relatively well-known taxa, the substantial overlapping in their ranges determines an area of endemism. When dealing with few taxa, this represents an easy task, but when having a high number of taxa to analyze, difficulties may arise. In order to provide ways to choose objectively which taxa to map and maximize the number of taxa

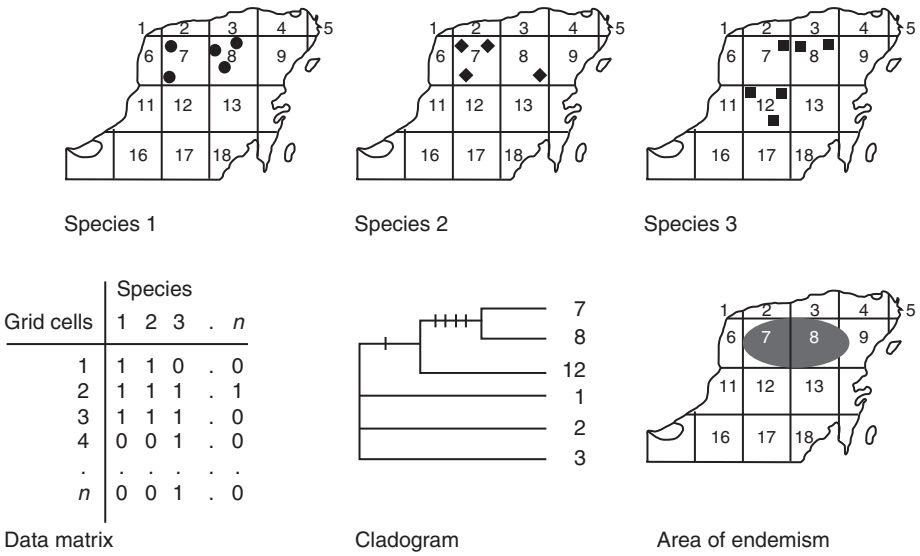


Fig. 1 Steps of a PAE, procedure that allows to identify areas of endemism.

contributing to the areas of endemism recognized, several procedures have been developed during the last decades. The most popular is PAE, formulated originally by the paleontologist Brian Rosen. It comprises the following steps (Fig. 1):

1. Draw grid cells on a map of the region to be analyzed, considering grid cells only where at least one locality of one taxon exists.
2. Construct an $r \times c$ data matrix, where r (rows) represent the grid cells and c (columns) the taxa. Entries are 1 if a taxon is present and 0 if it is absent. A hypothetical area coded 0 for all columns is added to root of the tree.
3. Perform a parsimony analysis of the data matrix; if several cladograms result, obtain a strict consensus cladogram.
4. Delimit in the cladogram obtained the groups of grid cells defined by at least two species.
5. Superimpose the groups onto the grid cells and map the species endemic to each group of grid cells, in order to delineate the boundaries of each area.

Although the extension of the grid cells to be considered in step 1 is difficult to assess *a priori*, an additional step could be added when the five steps have been completed. In this step, the original grid cells involved in conflictive relationships with more than one of the delimited areas, could be further subdivided in smaller units, and then the procedure restarted. When dealing with distributions of widespread species, they could overlap and generate many equally parsimonious cladograms. The strict consensus cladogram, however, preserves the most robust groupings of grid-cells, thus minimizing the influence of widespread species.

Areas of endemism are successively nested, which means that within larger areas of endemism smaller ones are recognized, and within the latter there are even smaller ones. This allows proposing a hierarchic biogeographical classification that parallels the taxonomic Linnaean hierarchy, employing the following subdivisions (in decreasing size): realms (or kingdoms), regions, dominions, provinces, and districts (Fig. 2).

Biogeographical Classification of the World

Modern biogeographical classification began with De Candolle, Sclater, and Wallace in the nineteenth century. Based on recent studies, Morrone proposed the following biogeographical system for the world:

1. *Holarctic realm*. It comprises Europe, Asia north of the Himalayan mountains, northern Africa, North America (excluding southern Florida), and Greenland. From a paleogeographic viewpoint, it corresponds to the paleocontinent of Laurasia.
   1.1. *Nearctic region*. It corresponds to the New World, in Canada, most of the USA, and northern Mexico.
   1.2. *Palearctic region*. It corresponds to the Old World, in Eurasia and Africa north of the Sahara.
2. *Holotropical realm*. Basically the tropical areas of the world, between 30° south latitude and 30° north latitude. The Holotropical region has been previously recognized by Rapoport and would correspond to the eastern portion of the Gondwanaland paleocontinent.
   2.1. *Neotropical region*. Tropical South America, Central America, south-central Mexico, the West Indies, and southern Florida.
   2.2. *Afrotropical region*. Central Africa, the Arabian Peninsula, Madagascar, and the West Indian Ocean islands.

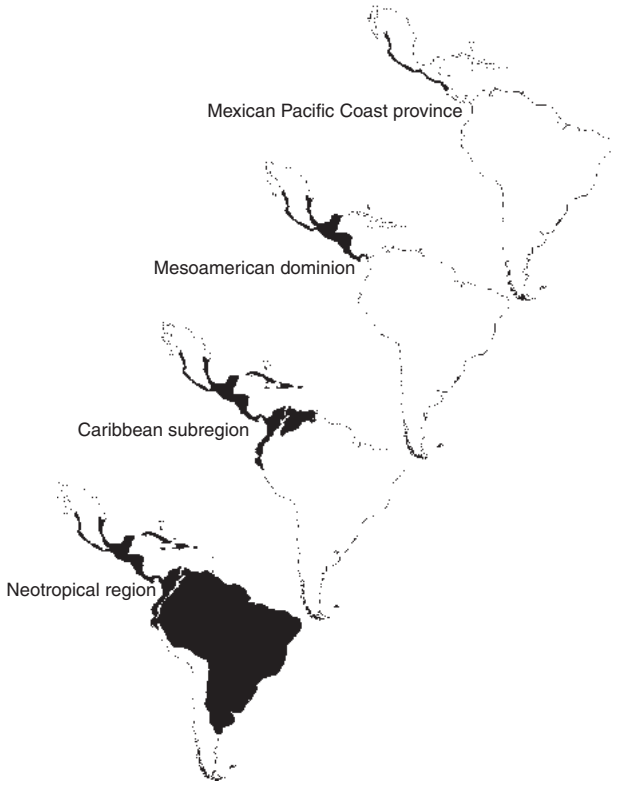


Fig. 2 Nestedness of the areas of endemism, which is the basis for biogeographical classification.

2.3. *Oriental region*. India, Himalaya, Burma, Malaysia, Indonesia, the Philippines, and the Pacific islands. In spite of the obvious tropical biotic elements of this region, it has been placed in earlier paleogeographic reconstructions as part of Laurasia. Recent authors have postulated that this area was part of Gondwanaland.
2.4. *Australian Tropical region*. Northwestern Australia.

3. *Austral realm*. It comprises the southern temperate areas in South America, South Africa, Australasia, and Antarctica. This region has been recognized previously by Kuschel and Rapoport and would correspond to the western portion of the paleocontinent of Gondwanaland.
   3.1. *Andean region*. Southern South America below 30° south latitude, extending through the Andean highlands north of this latitude, to the Puna and North Andean Paramo.
   3.2. *Antarctic region*. Antarctica.
   3.3. *Cape or Afrotemperate region*. South Africa.
   3.4. *Neoguinean region*. New Guinea plus New Caledonia.
   3.5. *Australian Temperate region*. Southeastern Australia.
   3.6. *Neozelandic region*. New Zealand.

Evolution

In order to reconstruct the historical relationships of areas of endemism, Donn Rosen, Gareth Nelson, Norman Platnick, and Ed Wiley developed cladistic or vicariance biogeography. It assumes that the correspondence between taxonomic relationships and area relationships is biogeographically informative. It is based on an analogy between biogeography and systematics, where taxa are treated as characters.

A cladistic biogeographic analysis is comprised of three basic steps (**Fig. 3**):

1. Construction of taxon-area cladograms from taxon cladograms, by replacing the terminal taxa by the area(s) of endemism inhabited where they are found.
2. Conversion of taxon-area cladograms into resolved area cladograms, by resolving problems due to widespread taxa, redundant distributions, and missing areas.
3. Derivation of general area cladogram(s), that represent(s) the most logical solution for all the taxa analyzed.

Patterns of area relationship derived from a cladistic biogeographic analysis are interpreted as secondary biogeographical homology. This is the cladistic test of the primary biogeographical homology formerly recognized.

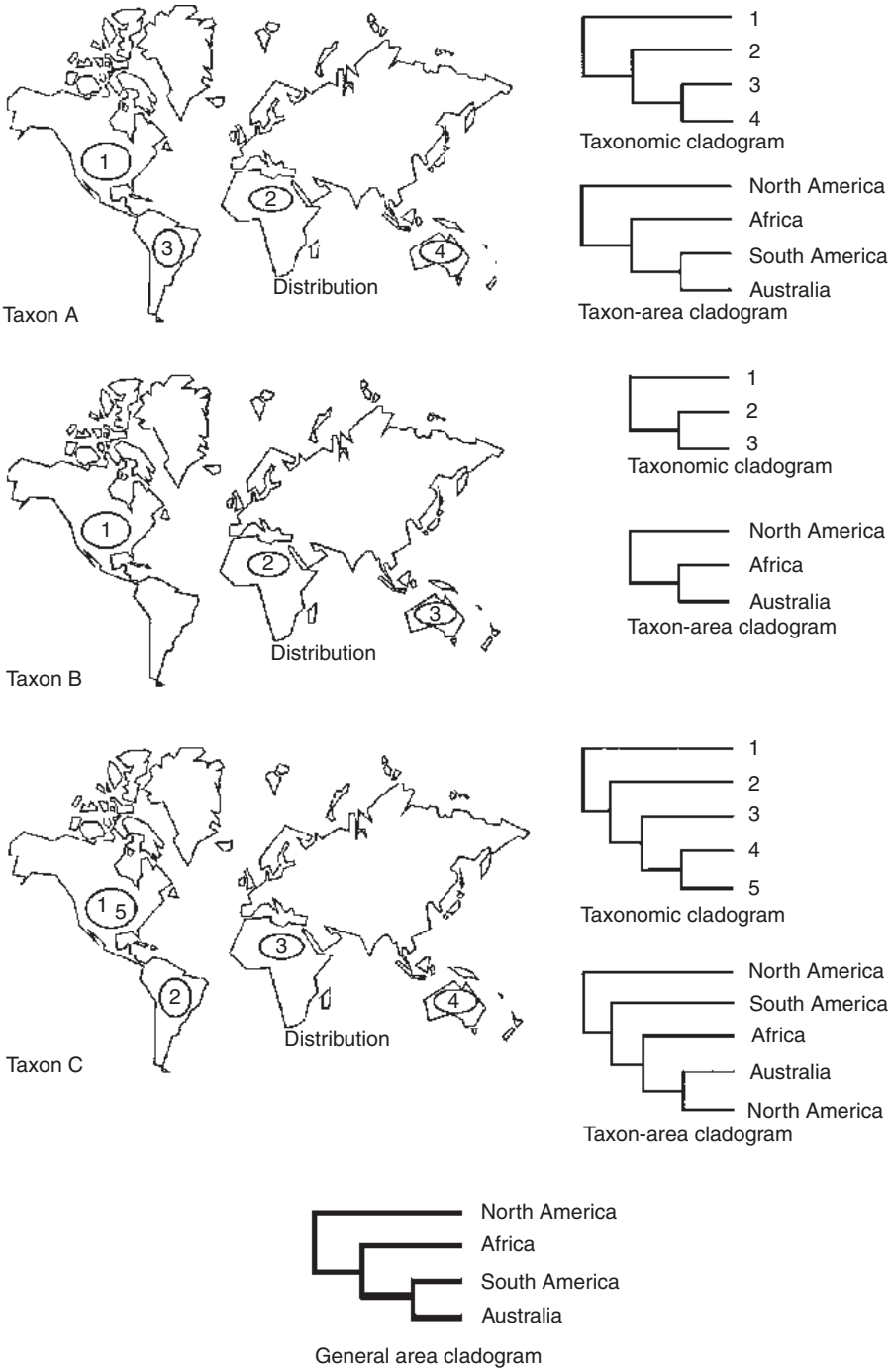


Fig. 3 Steps of a cladistic biogeographic analysis, where distributional and phylogenetic data of three different taxa allow to obtain a general area cladogram.

Conservation

Biodiversity is in global crisis. One of the major goals of conservation is the maintenance of as much as possible of the diversity of life. Thus we have to measure and compare priorities regarding areas to be protected, and we need to measure and compare local biodiversity, taking into account not only the number of species, but also the degree of difference among them.

One criterion for measuring biodiversity is species richness, exemplified by the 'megadiversity countries' concept; however, many important components of biodiversity are not represented in countries with highest values of diversity, and some areas could harbor a large number of widespread species, with no great conservation concern. Ecologists have developed diversity measures that combine species richness with information about abundance among species, and information on the vulnerability

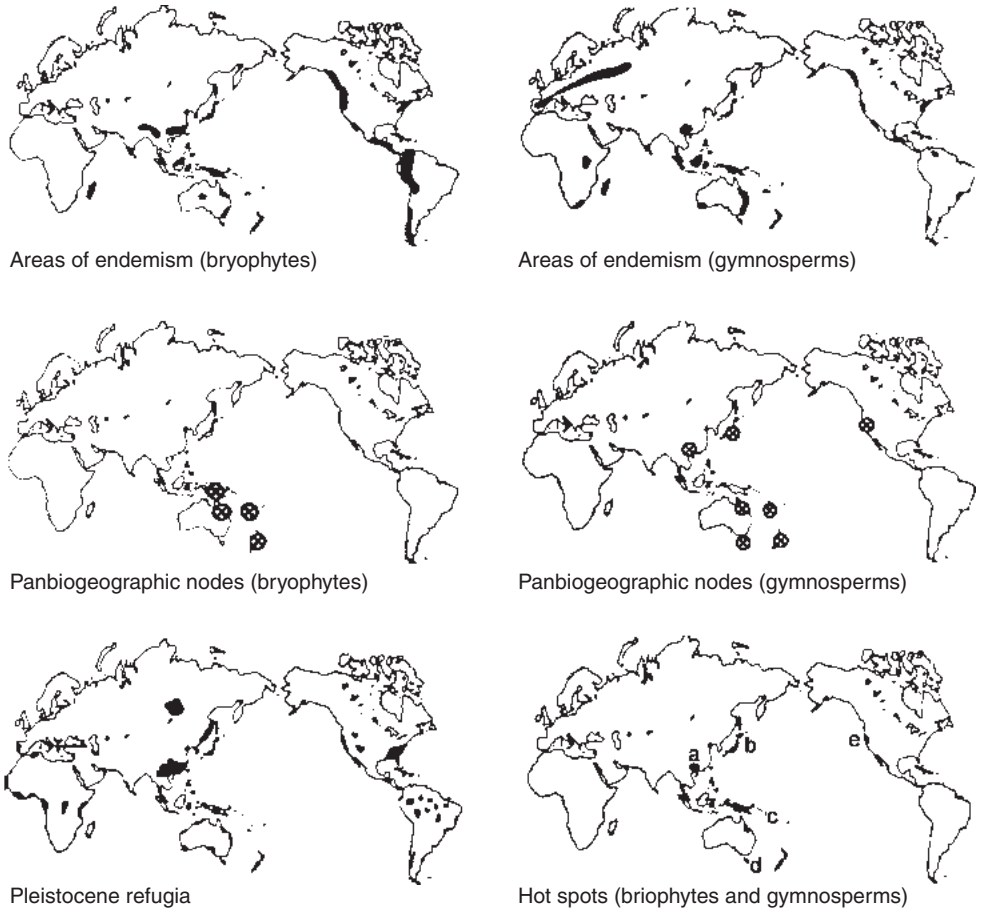


Fig. 4 Biogeographic analysis of gymnosperms of the world, with the hot spots that resulted from overlap among areas of endemism, panbiogeographic nodes, and Pleistocene refugia. a, Southeastern China; b, Japan; c, New Caledonia; d, Tasmania; e, western North America.

of those species, including the 'hot spots' analysis. Some authors have suggested that endemicity may help determine priorities for biodiversity conservation, whereas others have argued that it is not an appropriate measure of diversity, and is an ineffective mean for selecting areas for conservation.

A recent analysis of the distributional patterns of bryophyte and gymnosperm taxa of the world has found a remarkable congruence among areas of endemism, panbiogeographic nodes, and refugia in western North America, Japan, southeastern China, Tasmania, and New Caledonia (**Fig. 4**). It was concluded that these areas deserve special status for conservation, being significant candidates for designation as 'hot spots'.

See also: Conservation Ecology: Endangered Species; Conservation Genetics; Colonization; Isolation; Phylogenomics and Phylogenetics. General Ecology: Biodiversity. Terrestrial and Landscape Ecology: Island Biogeography; Spatial Distribution

Further Reading

Anderson, S., 1994. Area and endemism. Quarterly Review of Biology 69, 451–471.
Brown, J.H., Lomolino, M.V., 1998. Biogeography, 2nd edn. Sunderland, MA: Sinauer Associates.
Contreras-Medina, R., Morrone, J.J., Luna-Vega, I., 2001. Biogeographic methods identify gymnosperm biodiversity hotspots. Naturwissenschaften 88, 427–430.
Contreras-Medina, R., Morrone, J.J., Luna Vega, I., 2003. Uso de herramientas biogeográficas para el reconocimiento de 'hotspots': Un ejemplo de aplicación con briofitas y gimnospermas. In: Morrone, J.J., Llorente, J. (Eds.), Una perspectiva latinoamericana de la biogeografía. Mexico City: Las Prensas de Ciencias, UNAM, pp. 155–158.
Cox, C.B., 2001. The biogeographic regions reconsidered. Journal of Biogeography 28, 511–523.
Espinosa Organista, D., Aguilar Zúñiga, C., Escalante Espinosa, T., 2001. Endemismo, áreas de endemismo y regionalización biogeográfica. In: Llorente Bousquets, J., Morrone, J.J. (Eds.), Introducción a la biogeografía en Latinoamérica: Teorías, conceptos, métodos y aplicaciones. México City: Las Prensas de Ciencias, UNAM, pp. 31–37.
Harold, A.S., Mooi, R.D., 1994. Areas of endemism: Definition and recognition criteria. Systematic Biology 43, 261–266.
Morrone, J.J., 2001. Homology, biogeography and areas of endemism. Diversity and Distributions 7, 297–300.
Morrone, J.J., 2002. Biogeographic regions under track and cladistic scrutiny. Journal of Biogeography 29, 149–152.
Morrone, J.J., 2004. Homología biogeográfica: Las coordenadas espaciales de la vida. Mexico City: Cuadernos del Instituto de Biología 37, Instituto de Biología, UNAM.

Morrone, J.J., 2005. Cladistic biogeography: Identity and place. Journal of Biogeography 32, 1281–1284.
Morrone, J.J., Carpenter, J.M., 1994. In search of a method for cladistic biogeography: An empirical comparison of component analysis, Brooks parsimony analysis, and three area statements. Cladistics 10 (2), 99–153.
Morrone, J.J., Crisci, J., 1995. Historical biogeography: Introduction to methods. Annual Review of Ecology and Systematics 26, 373–401.
Morrone, J.J., Escalante, T., 2002. Parsimony analysis of endemicity (PAE) of Mexican terrestrial mammals at different area units: When size matters. Journal of Biogeography 29, 1095–1104.
Nelson, G., Platnick, N.I., 1981. Systematics and biogeography: Cladistics and vicariance. New York: Columbia University Press.
Rosen, B.R., 1988. From fossils to earth history: Applied historical biogeography. In: Myers, A.A., Giller, P.S. (Eds.), Analytical biogeography. London: Chapman and Hall, pp. 437–481.
Wiley, E.O., 1988. Parsimony analysis and vicariance biogeography. Systematic Zoology 37, 271–290.
Williams, P.H., Humphries, C.J., 1994. Biodiversity, taxonomic relatedness, and endemism in conservation. In: Forey, P.L., Humphries, C.J., Vane-Wright, R.I. (Eds.), Systematics Association, Special Volume 50: Systematics and Conservation Evaluation. Oxford: Clarendon Press, pp. 269–287.
Zunino, M., Zullini, A., 2003. Biogeografía: La dimensión espacial de la evolución. Mexico City: Fondo de Cultura Económica.

Evolutionary Ecology

Bregje Wertheim, University of Groningen, Groningen, The Netherlands

Glossary

Genome The complete genetic information of an individual, including all its genes and noncoding DNA.
Genotype The combination of alleles of an individual at one or more positions in the genome.
Habitat The natural environment in which an organism lives.
Life history The age-, size-, or stage-specific resource allocation of organisms to development, growth, maturation, maintenance, survival and reproduction.
Meta-community Local communities that are linked by dispersal.
Phenotype The set of traits that an individual exhibits.
Taxa Taxonomic categories or groups, such as phyla, orders or species.

Introduction

Evolutionary ecology is the field of science that studies biological diversity as the outcome of evolution and ecology. Diversity is used here in its broadest sense of the word, encompassing species diversity as the product of evolutionary processes (including diversification, speciation and extinctions) and ecological interactions in the habitat or (meta-)community, trait diversity (including life history strategies, morphology, behavior and physiology) that evolved through natural selection, and genetic diversity that can be studied to infer the occurrence and influence of evolutionary and ecological processes. Evolutionary ecology aims to explain the distributions, abundance and characteristics of organisms from ecological and evolutionary processes, both at historical and contemporary timescales. It seeks to explain the ecology of organisms in the context of evolution, as well as the patterns of evolution as explained by ecological processes.

Among the evolutionary processes, natural selection is what causes organisms to adapt to their environment. Natural selection is the difference in reproduction and survival (fitness) among individuals with different traits or phenotypes. When these trait or phenotypic differences are (partly) heritable and not all individuals are equally successful in reproducing, then natural selection over successive generations leads to an increased frequency of individuals with advantageous traits or phenotypes in the population. Natural selection thus accounts for the traits and adaptations that equip organisms to survive and reproduce in a particular habitat. It also has lead to the present-day biodiversity that arose from a single common ancestor, with many different organisms now inhabiting every environment on Earth. However, not all evolution is adaptive, and also nonadaptive evolutionary processes, such as genetic drift, mutation and migration can largely contribute to trait diversity and species distributions. The combined adaptive and nonadaptive processes can eventually lead to the formation of new species (speciation), new traits (evolutionary innovations) or trait optimizations, as well as the loss of species (extinctions) or trait loss.

The ecological processes that affect diversity include the match of individuals with their environment, the interactions of individuals with members of their own populations, the interactions between populations of different species, and the exchange among communities at both local and regional scales. The match with the environment comprises the ability of an organism to cope with both the prevailing abiotic conditions (e.g., temperature, salinity and humidity) and biotic conditions (i.e., all the other organisms that exist in the environment). Whether a species can persist in an environment is determined by its own ability to survive and reproduce under the particular conditions of that environment, the abundance and characteristics of food, competitors, mutualists and natural enemies, its ability to disperse to/from other populations, as well as historic and regional events that shaped the species diversity in the meta-community.

The premise that diversity is the outcome of both ecology and evolution makes for a vast, nearly all-encompassing, research field. To allow for more tractable research domains, many subdisciplines developed where evolutionary phenomena are investigated in the context of ecology, or vice versa. For example, research that focuses on natural selection and fitness optimization to explain trait diversity include, among others, behavioral ecology, life history theory and ecophysiology. Research domains that focus on species abundance or diversity in (meta-)communities include, for example, population dynamics, predator–prey interactions, facilitation, coevolution, community ecology and biogeography. Research domains that primarily focus at causes and consequences of genetic diversity within species include, for example, molecular ecology, population genetics, conservation biology and ecological genetics. To address the rich variety of research questions in evolutionary ecology, many different research approaches are applied. These comprise observational, descriptive and comparative research, laboratory and field experiments, molecular and genomics analysis, as well as mathematical simulation and theoretical modeling. The subdisciplines share the overarching aim to explain the origin and maintenance of biological diversity, but they tend to focus on different subsets of processes or spatio-temporal scales. By integrating knowledge from the different subdisciplines and approaches, evolutionary

ecology has contributed largely to the development of a strong synthesis of the many factors that can shape biological diversity. This has lead to conceptual theories (e.g., optimality theory, evolutionary game theory, life history theory, foraging theory, adaptive dynamics) that have been adopted throughout biology, and some have also permeated in economics, psychology and medical sciences.

There is large consensus on the interdependence of ecology and evolution—without ecology we cannot understand adaptive and diversifying evolution, and without evolution we cannot explain the properties of organisms for their interaction with the environment. However, ecologists and evolutionary biologists have had different perspectives on the timescale at which some of the processes occur, the extent to which evolution can affect short-term ecological dynamics, and the extent to which (localized) ecological processes and complexity are influential for the long-term outcome of evolutionary processes. Many ecologists considered evolution as a process that is too slow to have a direct impact on contemporary population dynamics or species interactions. Implicitly, species traits were considered to be "static" at ecological timescales. This applies indeed for the origins of new species or new traits, but it does not take into account the ample (heritable) trait variation that natural selection acts upon. When we use the population genetics definition of evolution, that is, the change in the genetic composition of a population over time, then evolutionary and ecological processes can operate at similar spatial and temporal scales. The appreciation for the potential of evolution to act rapidly and reciprocally between interacting species, and its impact on ecological dynamics, has substituted the static perception of species traits in the field of eco-evolutionary dynamics. Conversely, many evolutionary biologists have disregarded the patchiness and local variation that is paramount in the environment. They typically acknowledge this as potentially relevant for processes such as local adaptation, but tend to limit ecological complexity in reductionist experiments to dissect how a single ecological factor leads to evolutionary change. Ecological complexity, however, can lead to very different ecological and evolutionary dynamics, that may not be captured in reductionist experiments. While this may not be easily resolved in experimental laboratory settings, it can be addressed in carefully designed field experiments. Additionally, the development of genomics technology provides new opportunities to investigate the interface between ecological and evolutionary processes, both at ecological and evolutionary timescales.

History of Evolutionary Ecology

In essence, the field of evolutionary ecology came into existence the moment that scientists started to realize that life is not static, that is, the end product created by God, but that life in all its forms changes in response to the interaction with its natural environment. Since the Middle Ages, the Western scientific society had developed a strong theological and teleological worldview, also based on the manuscripts of Plato and Aristotle. These manuscripts, in combination with the belief in a Creator of the universe and all species, fuelled the notion that God had created perfect forms that remained forever unchanged, in the same (perfect) form as when they were created. For many centuries, natural scientists worked toward describing and cataloguing the grand diversity as a homage to this great design by God, with naturalists characterizing the striking biodiversity, astronomers and physicists describing the stars and the universe, and geologists describing the composition of the Earth. This worldview came under debate toward the end of the 17th century during the Scientific Revolution, when geologists started to report evidence that Earth had undergone profound changes through time, astronomers that the universe was not perfectly shaped around the Earth, and naturalists that life was not static but changeable. These scientists included Erasmus Darwin, the grandfather of Charles Darwin, who propositioned that "all warm-blooded animals have arisen from one living filament […] with the power of acquiring new parts, attended with new propensities" (Darwin, 1794, http://www.gutenberg.org/files/15707/15707-h/15707-h.htm). It also included Chevalier de Lamarck, who was a clear advocate for the nonstatic worldview, and proposed that new species could arise (from nonliving matter), and that traits could change due to the needs that a species had. It culminated with the seminal book of Charles Darwin on the "Origin of Species by Means of Natural Selection" (1859), in which he sets out two major theories, (i) the theory of descent with modification of all living species from common ancestors by incremental and small changes, and (ii) the theory of natural selection as cause for the evolutionary changes that leads to adaptations of organisms to their environment. The theory of natural selection was simultaneously and independently conceived by Alfred Russel Wallace.

Ernst Haeckel became inspired by the work of Darwin on the theory of descent with modification through natural selection, and how this may alter the systematic study and classification of the morphology of organisms. In his book, "*Generelle Morphologie der Organismen. Allgemeine Grundzüge der Organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte Descendenz-Theorie*" (1866) he was the first to introduce the term "*ecology,*" and he did so explicitly in the context of evolution. He defined ecology as "the whole science of the relations of the organism to the environment including, in the broad sense, all the conditions of existence. These are partly organic, partly inorganic in nature; both, as we have shown, are of the greatest significance for the form of organisms, for they force them to become adapted. […] Thus the theory of evolution explains the housekeeping relations of organisms mechanistically as the necessary consequences of effectual causes and so forms the monistic groundwork of ecology" (Haeckel, 1866; translated by Stauffer, 1957).

Darwin's theories inspired many, but they were also received with skepticism by scientists and the general public, both because it conflicted with the reigning worldview, and because of serious discrepancies with the then prevailing incorrect ideas on "blending" inheritance of traits. These ideas were incompatible with the proposed action of natural selection on individual variation, and the maintenance of individual variation. The incorrect view on inheritance were resolved by the discoveries of Gregor Mendel, a contemporary of Charles Darwin, who developed the fundamental laws of inheritance: (1) the law of

segregation, (2) the law of independent assortment and (3) the law of dominance. However, Darwin was unaware of the important work by Mendel on "particulate" inheritance. It took until the 1930–1940s when Mendel's laws were rediscovered by the mathematicians Ronald Fisher, John Haldane and Sewell Wright, who reconciled it with Darwin's theory of natural selection. Theodosius Dobzhansky then wrote the book "Genetics and the Origin of Species" (1937), where the population genetics models of Fisher, Haldane and Wright were applied to the study of natural populations, making it more accessible to other biologists. This "Modern Synthesis" provided a solid genetic mechanism for natural selection, and now forms the foundation for modern evolutionary theory.

The Study of Biological Diversity

Species Diversity

The species diversity in a large area can be partitioned into the local diversity and the regional diversity, where the latter allows for exchange, or turnover, of species between localities. Local diversity is constraint by the local characteristics of the habitat (e.g., productivity, habitat heterogeneity, climate and other abiotic conditions), species interactions (e.g., competition, predators and parasites) and chance events (e.g., habitat disturbance or destruction). Saturation of the local species diversity is considered to be largely due to antagonistic species interactions, primarily competition but also parasite–host interactions. However, species interactions can also facilitate coexistence of species through facilitation or mutualisms. Evolutionary processes can contribute to species coexistence at local scales when the populations becomes better at coping with the species interactions and/or the environmental challenges that they encounter in their habitat. For example, species can develop behavioral preferences or character divergence to reduce competition (resource partitioning, ecological specialization), or evolve a higher resistance against the local parasites that they are exposed to. Additionally, spatial and temporal heterogeneity can facilitate coexistence of species, because it diversifies the direction and strength of fitness effects caused by species interactions. Similarly, adaptations to a particular condition in the habitat may trade-off with other traits that influence the performance under different conditions, such that species perform better under some, but not all, environmental conditions. Although adaptive processes were originally considered to be too slow to matter for local diversity, nowadays it is recognized that even at ecological timescales, such evolutionary dynamics can occur and can indeed affect ecological dynamics.

Local diversity can increase by immigration from the regional species pool. Regional species diversity is also to some extent governed by the ephemeral processes and species interactions that affect local communities. At intermediate time scales, the generation and maintenance of species diversity is due to processes that operate within the regional landscape, such as dispersal, colonization, diversification, speciation and extinction. At very long timescales, historical processes (e.g., glaciation) and geographic events (e.g., plate tectonics) are important drivers for the accumulation or loss of species over time. It can be informative to retrace the historical development of regional species diversity, for example, to explain the adaptive evolution of traits in the context of species interactions, or to relate patterns of adaptive radiation to nutrient availability or release from predators or parasites. This requires a reconstruction, for example, of the composition of the historic (meta-)communities, including the associations and coevolution within these communities or the historic patterns of resource use, or on the phylogenetic origin of species and the rates of clade diversification.

One major question in evolutionary community ecology is what caused and maintained the differences in species composition and species diversity of local communities. This has become of a matter of intense debate. One extreme position is that these differences are largely driven by local species interactions, primarily competition, which results in ecological diversity and niche differentiation through adaptive evolution. The other extreme position is that it is largely determined by random sampling from the regional species pool (i.e., ecological drift), essentially assuming that species are ecologically identical and the assembly of species is a purely stochastic process relying on random death, dispersal and speciation events. The latter has been formalized in the neutral theory of species diversity. While the neutral theory initially received much criticism as too much a simplification of reality, it is now considered a useful null-model to contrast different ecological and evolutionary scenarios that may explain species diversity. The vast number of processes that can govern community composition and contribute to the uniqueness of each study system, was synthesized into four general classes by Mark Vellend: (i) speciation and (ii) dispersal add species to communities and (iii) drift and (iv) selection shape the relative abundances of these species, while ongoing dispersal drives the community dynamics (Vellend, 2010).

Trait Diversity

The living world exhibits a rich diversity in forms, functions, strategies and behaviors. This phenotypic variation is prominent both among and within species. Trait diversity describes the alternatives in behavior, personality, physiology, morphology and life history strategies. In evolutionary ecology, trait diversity is studied both within and among species, mostly in the context of adaptive evolution in response to biotic and abiotic conditions of the current or past environment. It addresses how natural selection and/or sexual selection have lead to a optimization of a trait, or maintained trait variation, to maximize reproduction and survival in the context of limited resources, within-species interactions, between-species interactions, the abiotic conditions of the environment and the spatio-temporal variability that may exist in any or all of these factors. When the trait is a new

characteristic that enhances the survival or reproduction of an organism relative to ancestral character states, it is called an adaptation. Natural selection is the only mechanism known to cause the evolution of adaptations.

Optimization of a trait in response to an environmental challenge can take very different forms, depending on the nature of the selection pressure, as well as trade-offs and constraints. Under directional selection, for example, for larger body size, the average trait value will shift upward within the population until ecological and physiological constraints restrict further increases. Under negative frequency dependent selection, however, rare alleles have a selective advantage, and the optimal trait value thus depends on the allele frequency in the rest of the population. This will lead to ongoing evolutionary dynamics, but in constantly changing directions. This may occur, for example, in host-parasite coevolution, with continual adaptations and counter-adaptations that alter resistance and virulence. Organisms have to optimize a variety of traits to maximize their fitness, and may thus need to economize in the allocation of resources to some traits to increase the investments in others (trade-offs). Moreover, the expression of many traits, for example, body size, is also strongly influenced by environmental conditions, such as the quantity and quality of nutrients, and the density of competitors. Finally, natural selection and sexual selection may form opposing selection pressures on a single trait; for example, increased coloration may enhance mating success but simultaneously increase the risk of predation.

Genetic Diversity

Genetic diversity describes the number of DNA sequence variations between individuals. The variations arise through random mutation and recombination. Most mutations do not result in detectable effects on phenotypic traits or fitness and are effectively neutral. Neutral and nearly-neutral mutations accumulate over time at a rate that depends mostly on the mutation rate and genome size. The mutation rate and genome sizes vary considerably among taxa, differing by several orders of magnitude. In general, however, the genetic diversity among individuals is huge, in the order of 1 sequence variation in every 50–1000 nucleotides in the genome. Any two unrelated individuals therefore differ in their DNA sequence in up to hundreds of thousands of nucleotides. Moreover, these neutral sequence variations evolve mostly under genetic drift, which is a random process that is not affected by environmental conditions and operates independently in separate populations. The factors that do affect genetic drift are population size and migration. Some of the genetic diversity, however, is not neutral. When a mutation results in a phenotypic effect, this is most often deleterious and only rarely beneficial for fitness. Natural selection against or for individuals with these DNA sequence variations then leads, over time, to a change in the frequency of these alleles in the population.

Genetic diversity may be investigated to study the ecological and evolutionary factors that drive a species' distribution. Species are usually subdivided into separate or partially overlapping populations across the geographic areas that they inhabit. The genotype of individuals for various sequence variations can easily be measured from field-collected samples. The patterns in genetic diversity of the neutral sequence variations reveal information on, for example, the connectivity of populations and their genealogical structure, the mating structure of the species, recent and historic changes in population size (constrictions and expansions), and the frequency of migration or gene flow among populations. Population genetics analysis is widely used to infer these patterns. When these patterns are combined with knowledge on concurrent ecological or historical events, this can reveal which processes have likely contributed to the current distribution and abundance of species. It can also be used in conservation management of threatened species, or to help identify populations that have become isolated from other populations.

Genetic diversity can also be used to study adaptive evolution by comparing the genetic diversity among species and/or within species. In these comparisons, low genetic diversity in particular regions of the genome may indicate evolutionary conservation or negative selection, indicating that sequence variation cannot be tolerated without negative consequences for fitness. High genetic divergence is a signature of positive selection and can indicate adaptation to a habitat. Populations or species that occupy multiple habitats with different selection pressures can evolve along different trajectories due to the local combinations of selection pressures. This is associated with genetic divergence for the genes that are associated with traits that affect fitness. When comparing unrelated taxa across long phylogenetic distances, these comparisons can also provide indications of convergent evolution, where multiple species have evolved the same solution to a common ecological problem.

Ecological and Evolutionary Genomics

A major transition in evolutionary ecological research was initiated by the development of genomics technology at the start of this century. Whereas we previously were restricted to study fragments of DNA and the expression of a small subset of genes or protein variants, we now have access to whole genome sequences, genome-wide gene expression profiles (transcriptome) and the total protein composition of a cell, tissue or organism (proteome). Importantly, we can now generate these data for many species and for many individuals of each species, which allows us to characterize both the similarities and differences among individuals, populations and species. These technologies, when combined with experiments on ecological interactions or information on the natural history of individuals, populations and species, has added a new dimension to evolutionary ecological research. It enables us to investigate long-standing fundamental questions on diversity, diversification and the evolutionary and ecological processes that shape variation. It also enables us to formulate new research questions and hypotheses on the molecular mechanisms of evolutionary processes and ecological interactions.

All of life stems from a single ancestor, and every trait and every species is the outcome of ecology and evolution. This diversity and diversification can be partially retraced or studied by investigating the genome sequences. Information from DNA enables us

to connect changes in individual genomes to population level processes to the generation of biodiversity. Extracting this information from the DNA allows us to reach back in time to infer past evolution, but also to directly observe the dynamics of evolution at ecological timescales. We can map the changes in the DNA in populations that occur during evolutionary adaptation, for example, in experimental populations that are exposed to a specific environmental condition for a number of successive generations (experimental evolution), or by monitoring natural populations over time to measure their evolutionary response to environmental change. We can also align and compare genomes from populations (population genomics) or from different species (comparative genomics) to identify genes and genomics regions that experienced positive and negative selection, to study speciation events, to trace ancestry and to determine genealogical or phylogenetic relations. With genomics technology, we can test hypotheses on individuals, traits, populations and species, both from existing theories as well as to construct new theories.

Genomics technology is also applied to measure species diversity in metacommunities by studying the genetic material from entire metacommunities or environmental samples (meta-genomics, environmental genomics or ecogenomics). This approach has had an enormous impact on microbial ecology, as it revealed the hidden diversity of microscopic life. Previously, the characterization of microbes had been restricted to those that could be cultured in the laboratory. Genomics technology uncovered the astronomical diversity and complexity of microbiota. As a consequence, the tree of life had to be largely redrawn (Hug *et al.*, 2016). What makes this particularly relevant in the context of evolutionary ecology is that microbes are essential for all life on Earth. They are key to the geochemical and nutrient cycles that convert carbon, nitrogen, oxygen and sulfur into biologically accessible forms, and form close and necessary associations with every species on Earth. Among the ecological and evolutionary processes that have been shaping biological diversity, the direct and indirect interactions with microbes have been enormously important.

Also studies on trait diversity have profited from genomics technology. The earlier overview on trait diversity was mostly restricted to the functions of traits and why they evolved, for example, which ecological processes resulted in natural selection for new adaptations or trait optimization. Another important question is how traits evolve, that is, how natural selection on individual phenotypes results in a change in the genetic composition of the population. This requires bridging the gap between the phenotype and the genotype of an individual. For most traits that are relevant in the ecology of an organism, it turns out that complex gene interaction networks underlie those traits, with many genes and gene interactions that contribute to the trait variation (the genotype–phenotype map). To identify the genes that matter for a trait, transcriptomics has been widely used to measure which genes change in expression in response to an environmental challenge or ecological stimulus (e.g., cold or heat stress, desiccation or drought, salinity, feeding by herbivorous insects, infection by a pathogen). Another approach relies on measuring individual variations in both the phenotypic trait and the genotype for many individuals, and then statistically associating phenotypic or trait variations with the genetic variation among those individuals (GWAS, or genome-wide association studies). These studies have provided much insight on the complex molecular and genetic mechanisms of trait diversity, for example, by showing that evolution of gene expression regulation (regulatory evolution) is a major contributor to trait and species diversity. Nonetheless, the genotype–phenotype map remains largely elusive for most traits. The complexity of the genetic architectures of trait can to some extent impede our ability to predict how traits will evolve in response to ecological changes and challenges. Yet, these complex genetic networks also provide a mechanism for robustness to produce a functional phenotype under a range of ecological conditions. Additionally, genetic variation anywhere in the network may contribute to phenotypic variation, which provides the adaptability of organisms to changing conditions.

Concluding Remarks

Evolutionary ecology is a field of science that addresses research questions on diversity across all levels of biological organization—from genetics to metacommunities—and on all levels of spatial and temporal resolution. The range of phenomena that is being investigated spans all aspects of life. Evolutionary ecologists typically specialize their research toward a subset of subdisciplines, based on personal preferences but also by necessity. Mastering the expertise required for each subdiscipline is simply impossible. Nonetheless, the integration of knowledge from the full breadth of the field is important, as evolutionary ecology is effectively "the theory of everything." Genomics technology is providing a new stimulus for further integration of studies on ecological processes (including their complexity) and evolutionary processes (including the direct observation of evolution in real-time) that shape the diversity of life.

Research in evolutionary ecology addresses fundamental scientific questions and is driven by curiosity on the processes that have contributed to and shaped the bewildering diversity in all of life around us. The knowledge, insight and theories that are being developed within the field of evolutionary ecology also have broad societal relevance. This knowledge-base provides important bearings for facing some of the grand challenges of our time, including the development of realistic plans for coping with the implications of global change, for stopping the alarming biodiversity loss, for designing evolutionary stable strategies to set-up a sustainable society with food security for an expanding human population, and for developing evolution-informed medical treatment.

See also: Behavioral Ecology: Dispersal–Migration; Competition. Evolutionary Ecology: Macroevolution; Life-History Patterns; Metagenomics; Natural Selection; Fitness; Metacommunities; Adaptation; Coexistence. General Ecology: Philosophy of Ecology: Overview; Abundance; History of Ecology; Biodiversity. Terrestrial and Landscape Ecology: Spatial Distribution

References

Darwin, E., 1794. Zoonomia, Vol. I or, the Laws of organic life. London: Johnson.
Darwin, C., 1859. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. London: Watts.
Dobzhansky, T.G., 1937. Genetics and the origin of species. New York: Columbia University Press.
Haeckel, E., 1866. Generelle Morphologie der Organismen. Allgemeine Grundzüge der Organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte Descendenz-Theorie. Berlin: Reimer.
Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., *et al.*, 2016. A new view of the tree of life. Nature Microbiology 1, 16048.
Stauffer, R.C., 1957. Haeckel, Darwin, and ecology. The Quarterly Review of Biology 32 (2), 138–144.
Vellend, M., 2010. Conceptual synthesis in community ecology. The Quarterly Review of Biology 85 (2), 183–206.

Further Reading

Bromham, L., 2016. An introduction to molecular evolution and phylogenetics. Oxford: Oxford University Press.
Futuyma, D.J., 2013. Evolution, 3rd edn. Sunderland: Sinnauer Associates.
Losos, J.B., Arnold, S.J., Bejerano, G., Brodie III, E.D., Hibbett, D., Hoekstra, H.E., *et al.*, 2013. Evolutionary biology for the 21st century. PLoS Biology 11 (1), e1001466
Mayhew, P.J., 2006. Discovering evolutionary ecology: Bringing together ecology and evolution. Oxford: Oxford University Press.
Pianka, E.R., 2011. Evolutionary ecology, 7th edn. eBook from Google.
Ricklefs, R.E., Schulter, D., 1993. Species diversity in ecological communities. Historical and geographic perspectives. Chicago: The University of Chicago press.

Fecundity

CJA Bradshaw, Charles Darwin University, Darwin, NT, Australia
CR McMahon, University of Wales Swansea, Swansea, UK

Introduction

Organisms exist to reproduce and fecundity is the measure of an individual's (or population's) reproductive performance. As such, fecundity (in combination with its sister concepts – see below) can be viewed as one of the two cornerstones of population biology, the other being the ability of an individual to survive in order to reproduce. Mortality (proportion of surviving individuals) expresses the proportion of a population that dies within a defined period of time and fecundity quantifies the new individuals added to the population. A population remains stable when the number of individuals dying is equivalent to the number of new individuals added, and it decreases when mortality exceeds replacement (births and immigration). Although these themes may appear straightforward, the relationships between reproduction, age, population density, and environmental stochasticity (random variation due to environmental changes) are complex and not fully understood. Furthermore, these relationships must be contextualized within the major reproductive strategies that have evolved as a result of different selective pressures, and they must be examined while taking into account the multitude of tradeoffs that occur between reproduction and individual survival.

In ecology, 'fecundity' can be described simply as the physiological maximum potential reproductive output of an individual over its lifetime, and it is usually applied to the female's (rather than a male's) output, although there is no inherent reason why fecundity should be restricted to females. Regardless, ecologists have defaulted to using the gender-general term of 'reproductive success' to describe the reproductive output of both males and females. We adhere to these traditions and restrict our definition of fecundity to apply to females only. However, reproductive success is not equivalent to fecundity because the former is a measure of an individual's genetic contribution to subsequent generations. As such, reproductive success can vary among individuals due to the effects of maternal age, fitness of progeny from clutches of different size, and other life-history characteristics. Fecundity is therefore a genetical and developmental trait that evolves within a particular selective framework. The term 'fertility' differs from fecundity in that it describes the actual (or current) reproductive performance of (typically) a female, and it is a generalization of the terms 'maternity', 'birth rate' and 'natality' which refer to the average number of offspring produced by an individual female of a particular age per unit time. 'Net reproductive rate' is defined as the average number of daughters produced by mothers over a lifetime. Fertility varies primarily as a function of environmental stochasticity and demography, and generally fluctuates temporally and spatially among individuals and populations.

In this article we have chosen to discuss both fecundity and fertility given their obvious connection and importance to the discipline of ecology. Henceforth the usage of each term relates specifically to the developmental and evolutionary (fecundity) or environmental (fertility) contexts shaping these fundamental parameters of population biology.

Sexual and Asexual Reproduction

The concept of fecundity should, in practice, be applied equally to the two major modes of reproduction – sexual and asexual. Discussions centered on the evolution of sexual reproduction must reconcile the theoretical advantage asexual reproduction presents to an individual – the possibility of a twofold fitness (fecundity) advantage over their sexual counterparts. The relatively higher cost of sexual reproduction arises from anisogamy (gamete dimorphism), but sexual selection can reduce or even eliminate this cost. Although there is no difference in the fecundity of sexual females and asexuals of the same genotype, the equilibrium frequency of deleterious mutations is lower in sexual populations, usually giving rise to higher fitness in sexual females.

Despite the advantages of sexual reproduction, the prevalence of clonal or asexual reproduction in the life histories of many plant and animal species suggests that this strategy does indeed provide advantages at least in some circumstances. For example, clonal reproduction may allow the clone to survive many reproductive events by reducing a potentially catastrophic local extinction associated with a restricted distribution. A large number of clone-mates may also increase the number of bodies producing gametes and improve the probability of them finding mates once sexual reproduction is invoked. Thus, variation in the costs and benefits of clones should depend on an individual's sexual neighborhood (relative abundance of potential mates) and the degree of competitive stress induced by the environment.

Inclusive Fitness

The concept of inclusive fitness has become the foundation of modern biology and may be one of the most important advances in evolutionary biology since Darwin introduced the theory of natural selection. If natural selection occurs at the individual level

(i.e., individuals with higher reproductive success have the greatest genetic representation in future generations), how do cooperative breeding systems evolve? The apparent paradox of some individuals within a population foregoing the opportunity to breed can be explained if some individuals help their close relatives reproduce. By assisting close relatives, altruists still manage to pass on their genes indirectly. This behavior can arise when the genetic relatedness between an altruist and beneficiary (r) multiplied by the fitness gain to the beneficiary (b) minus the fitness cost to the altruist (c) is greater than zero; that is, $(rb - c) > 0$.

Consequently, the degree of relatedness between the altruist and the beneficiary determines the probability that they share a particular gene, and therefore, the probability that the altruistic gene will be passed on to the next generation. Sterile castes in social insects are an extreme case of quasi-altruism (worker females give up reproduction and raise their mother's offspring). This occurs most commonly in the Hymenoptera (bees, wasps, ants), and is probably the result of their haplodiploid chromosomal organization where males have single chromosome copies and develop parthenogenetically from unfertilized eggs and females have two copies of each chromosome. Because males are haploid, there is no meiosis and all daughters of a particular male have identical paternal genes. Nonreproductive females have a greater chance of passing on their own alleles via their reproductive sisters than they do by reproducing themselves, so zero direct fecundity for these individuals results in the highest indirect lifetime reproductive success.

Eusocial Systems

Eusocial systems are an evolutionarily advanced level of colonial existence in which adult colonial members (1) belong to two or more overlapping generations, (2) care cooperatively for the offspring, and (3) are divided into either reproductive and nonreproductive (or less-reproductive) castes. The presence of castes is an integral requirement of eusociality, and it is most commonly found in social insects, especially ants, bees, wasps, and termites. As a reproductive strategy it is rarely observed in nature, but it appears to be an important mode of reproduction nonetheless. Although eusocial ants comprise only 2% of the *c.* 900 000 insect species known globally, this taxon comprises more than half of the global insect biomass. The strategy is obviously successful in evolutionary terms, so why then, despite this apparent success, has eusociality rarely evolved? It is generally thought that eusociality is the result of some extraordinary environmental circumstances that existed in the evolutionary past when a particular genome coding for group breeding and subordinate reproductive suppression was selected. Selection of this genome consequently led to a cooperative breeding system and was maintained not by the inability of individuals to breed, but because individuals are outcompeted by breeding conspecifics in well-integrated communal colonies.

Implicit in understanding fecundity patterns in eusocial systems is recognizing the factors that regulate body size and growth. The eusocial female that does all or most of the breeding is usually the biggest and most dominant. As such, in species of cooperative insects living in large groups, selection for increased fecundity has repeatedly resulted in the evolution of an increased body size for females. Whether this is also true of cooperative vertebrates is currently debatable; however, increased size of dominant breeding females has been documented in naked mole rats (*Heterocephalus glaber*) and meerkats (*Suricata suricatta*). Eusociality and the resultant reduction or eradication of fecundity in some individuals among diploid animals is rare and appears to occur only when (1) species live in burrows, (2) food is abundant, (3) parents care for offspring, and (4) mechanisms exist for mothers to manipulate other females, such as pheromones that inhibit their breeding.

Fecundity Patterns

The term 'parity' is used to describe a breeding event, and species can assume various modes of reproductive timing and frequency depending on their evolutionary past and the constraints of their current environment. There are too many variants of life-history strategies to describe succinctly given the large number of possible combinations of longevity, breeding frequency, and offspring number, but there are some major groups of strategies that have evolved. The two major fecundity patterns are semelparous and iteroparous reproduction.

Semelparity

Semelparous organisms reproduce only once during their lifetime. This may occur at the age of only 20 min in certain bacteria, a few hours in many protozoa, or up to a few weeks or months in some insects and mammals. Many semelparous species are annuals (live only one year), but some reproduce only after several years of maturation. Early work predicted that greater temporal variation in adult survival relative to juvenile survival favored the evolution of semelparity, although more recently it has been demonstrated that this also depends on the age structure of the population under selection. Semelparity may also provide other advantages over iteroparity in terms of offspring body size, leading possibly to increased juvenile growth rates and survival.

Semelparity in mammals is restricted to two marsupial families (Didelphidae and Dasyuridae), where all species demonstrate high post-reproductive senescence, but not all are semelparous. In some of the semelparous species, only the males die after the short, highly synchronous mating season, whereas in others, male die-offs are facultative. Some have argued that because the interval between conception and weaning is short and that these marsupials tend to live in highly predictable seasonal environments, there should be selection for a monoestrous reproductive pattern, high estrus synchrony, and a short mating season. These factors should therefore

induce intense male–male competition and a low probability of multiseason male survival giving rise to the evolution of male semelparity. Additionally, the high rate of female mortality resulting from long lactation periods in some species also demonstrating female semelparity selects for a bet-hedging strategy by males and the evolution of male semelparity.

Iteroparity

Species that reproduce more than once during their lifetime are iteroparous. The time of maturation preceding first reproduction may vary from a few days in small crustaceans to greater than 100 years in some trees. The frequency of reproduction can also vary markedly – daily (e.g., some tapeworms), semiannually, annually, biennially, or irregularly (e.g., humans). Some have argued that iteroparity is favored over semelparity when high environmental variability induces large variation in offspring production and offspring survival. However, it has been demonstrated more recently that the evolution of either strategy does not depend on environmental variability alone, but also on the strength of intrinsic regulation (a tradeoff between fertility and survival – see below) operating on a population and the demographic rates most affected by extrinsic variation.

In many iteroparous species there are a certain proportion of individuals that reproduce only in alternate seasons (or less often), thereby foregoing a proportion of their potential lifetime reproductive success. Known formally as 'low frequency of reproduction', this phenomenon is most common among vertebrate ectotherms, but it has also been documented in endotherms (e.g., willow tits (*Parus montanus*); kittiwakes (*Rissa tridactyla*); fat dormice (*Myoxus glis*)). There is also considerable phenotypic variation within a species, with some individuals opting for lower-frequency reproduction than others. Ectothermic species employing this strategy appear to have accessory activity associated with reproduction (but independent of fecundity) such as breeding migrations, egg brooding through incubation, and live bearing. With its apparent ubiquity, low-frequency reproduction appears to offer some selective advantage in certain systems. It has been suggested that the strategy could evolve only under two scenarios: (1) when reproduction does occur, it confers a much higher fecundity benefit than would a regular-frequency breeding event; or (2) the probability of surviving the interval between reproduction is much higher than it would be during the same interval of regular-frequency breeding. Thus, lower-frequency reproduction may actually result in higher average fecundity over the individual's lifetime when at least one of the two previous conditions is met. There also appears to be a relationship between the occurrence of low-frequency breeding with habitat quality. Here, regular reproduction may become less favorable as the habitat becomes less suitable because resource availability declines and accessory activity cannot decrease to the same extent.

Age Specificity

In iteroparous organisms, fecundity often increases with age following reproductive maturity and then can decline at older ages. As such, age specificity of reproduction is an essential element for understanding the evolution of life-history strategies. Reproductive rate may increase with age when an individual devotes relatively more resources to reproduction with increasing age. For example, growth rate often declines after reproductive maturity is achieved, so more resources can thereafter be directed to reproduction, thereby increasing age-specific fecundity. This introduces the concept of 'primiparity', which is generally defined as the age of first reproduction. In some species, this age is rather constant suggesting genetic control. However, in other species (especially ectothermic organisms), it is environmentally mediated and may demonstrate high phenotypic plasticity. The extent to which primiparity is delayed (i.e., the length of the pre-reproductive period) also depends on type of niche exploited by a species and the particular demographic configuration of a population.

Increasing fecundity is also observed in many organisms that do not grow following maturity, giving rise to three hypotheses to explain the pattern: (1) less-fit individuals are constantly eliminated from a population so that the average fecundity of surviving individuals increases; (2) increasing fecundity is a reflection of the gradual improvement in the competence of older, more experienced, and often higher-ranking individuals; and (3) as life expectancy decreases with age, individuals allocate more and more resources to reproduction at the expense of survival. The decline in fecundity with age is often referred to as 'reproductive senescence'. One hypothesis to explain this pattern is as an individual ages its resource acquisition rate deteriorates due to physiological ageing.

Recent work has attempted to combine the various hypotheses for age specificity in fecundity into a single conceptual model that incorporates the other key life-history parameter – age-specific survival. When extrinsic mortality is high, early maturity is favored allowing for high reproductive investment and fecundity early in life (**Fig. 1**). Under this scenario, allocating resources to repair cumulative somatic damage is not beneficial when life expectancy is already low, and fecundity will still diminish with time due to physiological ageing. When extrinsic mortality is low, investing in somatic repair at the cost of reduced fecundity is beneficial because it increases life expectancy. Thus, reproductive rate is low early in life, but gradually increases toward a late-life maximum, and then inevitably declines again due to physiological deterioration late in life (**Fig. 1**).

Allometric Scaling

Allometry is the relation between the size of an organism and aspects of its physiology, morphology, and life history. Typically, variation in body mass among individuals or species can be used to predict traits such as metabolic rate, dispersal capacity, survival

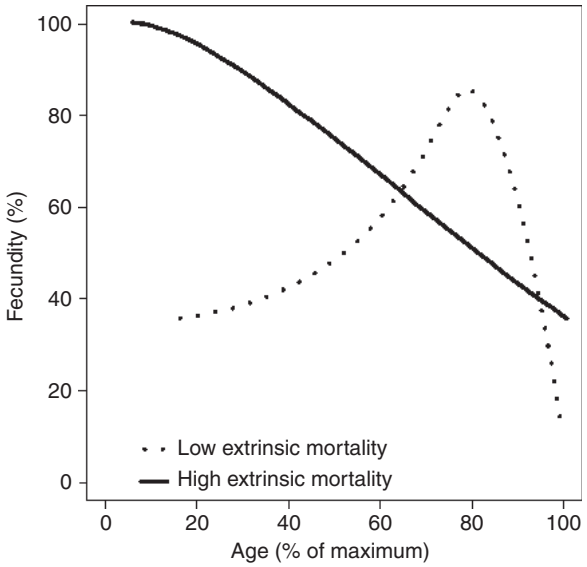


Fig. 1 Patterns of age-specific fertility fecundity expressed as a percentage of maximum age under two different rates of extrinsic mortality. Reproduced from Cichoń M (2001) Diversity of age-specific reproductive rates may result from ageing and optimal resource allocation. *Journal of Evolutionary Biology* 14: 180–185.

probability, and fecundity. The ratio of combined offspring mass to maternal mass tends to remain approximately constant over a broad range of maternal sizes within a species. As such the ratio of offspring mass at independence to average adult body mass remains stable within major taxonomic groups. In other words, larger females tend to have more or larger offspring, especially among invertebrates and ectothermic vertebrates. An increase in fecundity with increasing female body size may constitute a selective advantage toward large female body size (the so-called 'fecundity advantage' model) in some taxa, especially in those for which energy is not limiting. However, this does not necessarily mean that larger females will always have higher reproductive success because of the increased risks due to predation, limited resources, and environmental uncertainty.

Density Dependence

It is in every individual's interest to maximize its reproductive output without compromising its own survival prospects; however, the production of too many individuals can lead to overcrowding and a reduction in per capita resources such as food availability, breeding sites, and territories. Under such circumstances, it is clearly unprofitable for individuals to engage in potentially wasteful energy expenditure associated with fertility if the probability of their progeny surviving to breeding age is low. This observation is supported by the ample evidence that population density affects fertility negatively. In general, as population size increases, either the average number of offspring produced per female decreases (in litter or clutch sizes greater than one), or average offspring size decreases. However, the relationship between density and fertility is not necessarily linear. For example, for many long-lived, iteroparous species, the negative effects of population density are not expressed in average fertility until the population approaches its environmentally mediated carrying capacity. In many semelparous species with short life expectancies the opposite is true – fertility declines rapidly as soon as the population begins to grow. **Fig. 2** shows examples of fertility patterns relative to population density in different species.

Increasing density does not always result in depressed fertility. In many species, fertility is relatively invariant over the range of population densities experienced so that intrinsic regulation operates almost exclusively through the modification of age-specific survival rates. Some species do not appear to respond to density by adjusting fertility *per se*; rather, the age of first reproduction (primiparity) may change according to the per capita resources available. In other species, high-density living can actually stimulate production of offspring (known as inverse density dependence; e.g., **Fig. 2c**). For example, many plant species require high population densities to attract grazers or pollinators that assist in cross-pollination and propagule dispersal. This Allee effect (a term used to describe a reduction in reproductive output with decreasing population density, thus affecting population growth rate) is another example of inverse density dependence that occurs when populations existing at low densities suffer from stochastic demographic events that, for example, skew the tertiary sex ratio (the ratio of the rarer sex to the total number of breeding adults). In these cases, the availability of potential mates for the more common sex is reduced so that not all reproductively capable individuals succeed in fertilization and the production of offspring. The type and strength of density dependence operating on a population also have implications for the patterns of age-specific fecundity. A reduction in extrinsic mortality favoring the evolution of reduced reproductive senescence is predicted only when density dependence acts principally on fertility and without differentially decreasing late-age fecundity. Alternatively, a reduction in extrinsic mortality can favor the evolution of faster senescence if density dependence acts mainly on the survival of older age classes.

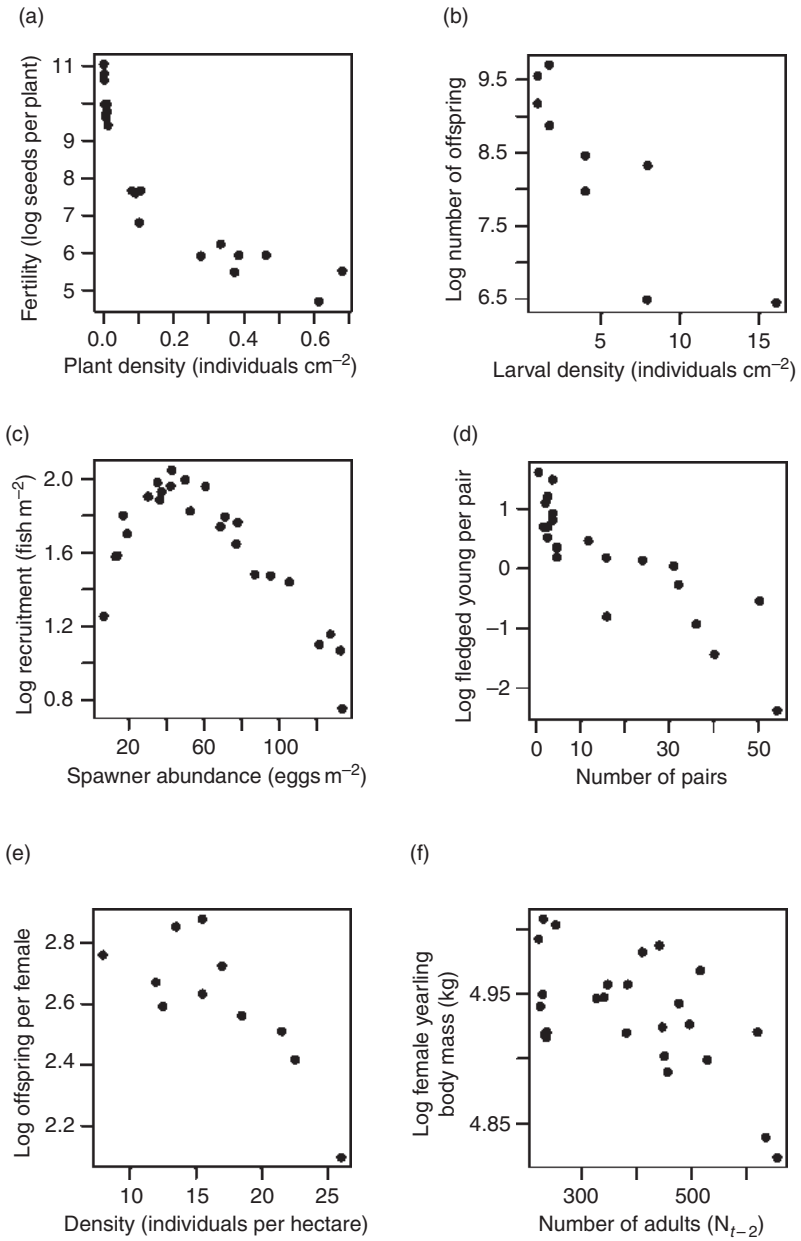


Fig. 2 Example patterns of fecundity fertility with changes in population density for (a) scentless chamomile (*Tripleurospermum perforatum*); (b) midge (*Chironomus riparius*); (c) brown trout (*Salmo trutta*); (d) muted swan (*Cygnus olor*); (e) European rabbit (*Oryctolagus cuniculus*); and (f) moose (*Alces alces*). (a) Reproduced from Buckley YM, Hinz HL, Matthies D, and Rees M (2001) Interactions between density-dependent processes, population dynamics and control of an invasive plant species, *Tripleurospermum perforatum* (scentless chamomile). *Ecology Letters* 4: 551–558. (b) Reproduced from Hooper HL, Sibly RM, Hutchinson TH, and Maund SJ. (2003) The influence of larval density, food availability and habitat longevity on the life history and population growth rate of the midge *Chironomus riparius*. *Oikos* 102(3): 515–524. (c) Reproduced from Myers RA (2001) ICES *Journal of Marine Science* 58: 937–951. (d) Reproduced from Nummi P and Saari L (2003) Density-dependent decline of breeding success in an introduced, increasing mute swan *Cygus olor* population. *Journal of Avian Biology* 34 (1): 105–111. (e) Reproduced from Rödel H, Bora A, Kaiser J, *et al.* (2004) Density-dependent reproduction in the European rabbit: A consequence of individual response and age-dependent reproductive performance. *Oikos* 104 (3): 529–539. (f) Reproduced from Solberg EJ, Saether B-E, Strand O, Loison A (1999) Dynamics of a harvested moose population in a variable environment. *Journal of Animal Ecology* 68 (1): 186–204.

Tradeoffs

A particular fecundity pattern evolves as an adaptation to the specific constraints imposed by the niche exploited by a species. For example, the high fecundity of parasites and many marine organisms is thought to have evolved because the probability that any particular individual will establish itself and reproduce successfully is low. As such, life-history theory predicts that the degree to

which parents place themselves or their offspring at risk of mortality when threatened with predation depends on the offspring number and survival probability of parents. This type of tradeoff will also vary between the sexes in relation to the degree of investment in the offspring. A species with low adult survival and high fecundity should tolerate greater risk from predation given the low probability of surviving to reproduce in the future. In contrast, species with high adult survival and lower fecundity should tolerate less risk even at a cost to their offspring. This fundamental tradeoff essentially expresses an organism's capacity to maximize its potential lifetime reproductive success by assessing the best strategy to achieve the highest potential fecundity. Thus, tradeoffs are the benefits accrued from one life-history process that are purchased at the expense of another. For example, pre-breeding migrations requiring the use of stored resources can result in less energy available for the production of offspring. Other processes that present an individual with additional energy costs can also affect reproductive output, including exposure to antigens or parasites that increase the costs associated with mounting an immune reaction.

This introduces the concept of 'reproductive allocation' (or 'reproductive effort') which is the proportion of an individual's acquired resources that is allocated to reproduction over a defined interval of time. Because resources are finite, reproduction itself can incur a cost in that it can reduce the survival probability or growth rate of the individual. As such, the rules governing reproductive allocation are viewed in a benefit–cost framework. Dimensionless ratios of the benefits and costs of reproduction, such as the ratio of the life span spent in reproduction (E) to the length of the pre-reproductive period (α), and the ratio of the fraction of adult mass (m) devoted to reproduction (R/m) to the inverse of the life span spent in reproduction (E^{-1}), provide useful indices to classify life histories because they tend to be invariant within major taxonomic groups.

Environmental Stochasticity

Climate change is perhaps the most pressing and urgent environmental issue facing the world today. Despite the limitations in our ability to predict and quantify the consequences of this change, there is ample evidence that changes in climate have profound effects on the bios. Indeed, given that fertility is one of the key life-history parameters determining population abundance and persistence, it is important to understand how it responds to environmental stochasticity. Prevailing environmental conditions can affect fertility in four principal ways by altering (1) maternal body condition, (2) maternal survival, (3) offspring body condition, and (4) offspring survival. However, quantifying these effects can be challenging. The usual way of collecting such information on vital rate variability is to census a population in different age classes or stages over time and to compare measured vital rates (e.g., fertility) against some environmental variable (see **Fig. 3** for examples). Although repeated measures provide estimates of temporal variance in vital rates, this total variance also includes two other sources: demographic stochasticity and measurement/sampling error. The true variance due to environmental stochasticity must be isolated from other sources of variability or risks of decline and extinction may be biased. Several methods exist for partitioning these variances for fertility to examine the role of environmental stochasticity on this life-history parameter within population models (see also the section titled 'Fecundity and fertility in population dynamical models').

Sex-Allocation Theory

Although sex ratios are discussed elsewhere in this encyclopedia, it is important to understand the basics of sex-allocation theory with respect to fecundity. The discussions centered on the allocation of resources toward reproduction have thus far ignored the notion that the long-term value of producing offspring of one sex versus another depends on the complex relationship between current and future reproduction. Just as various life-history strategies have evolved with respect to the timing, frequency, and magnitude of reproductive effort, there are various optimality models that exist to explain sex ratio variation. Early theoretical work argued that equal sex ratios are evolutionarily stable because parental investment in sons and daughters should be identical under stable environmental conditions and population-wide random mating. However, skewed sex ratios are commonly observed due to factors such as nutritional stress, age, condition, and social rank of mothers, litter size, population density, and changes in resource availability. Optimality models focus on the idea that parents should adjust the sex ratio of their offspring in response to factors affecting their own and their offspring's future reproductive success. For example, one highly cited model used to explain sex ratio variation in mammals predicts that in polygynous species with marked sexual dimorphism, good-condition mothers should invest in the sex showing the highest variation in reproductive success because investment in that sex can potentially provide a better lifetime return on maternal investment. The apparent contradiction to the evolutionary stability of equal sex ratios invoked by adaptive sex ratio manipulation models may be explained by the existence of negative temporal autocorrelation in sex ratios between breeding events over the lifetime of the individual. In other words, equal investment in the sexes may only operate when taking total lifetime reproductive output into consideration.

Mate Choice

In the case of sexual reproduction, the optimal allocation of resources to reproduction actually begins prior to fertilization itself. The theory of mate selection by females is usually described in the context of sexual selection so that the evolution of female mating preferences centers on the genetic consequences of gamete union. Hypotheses to explain the patterns range from females

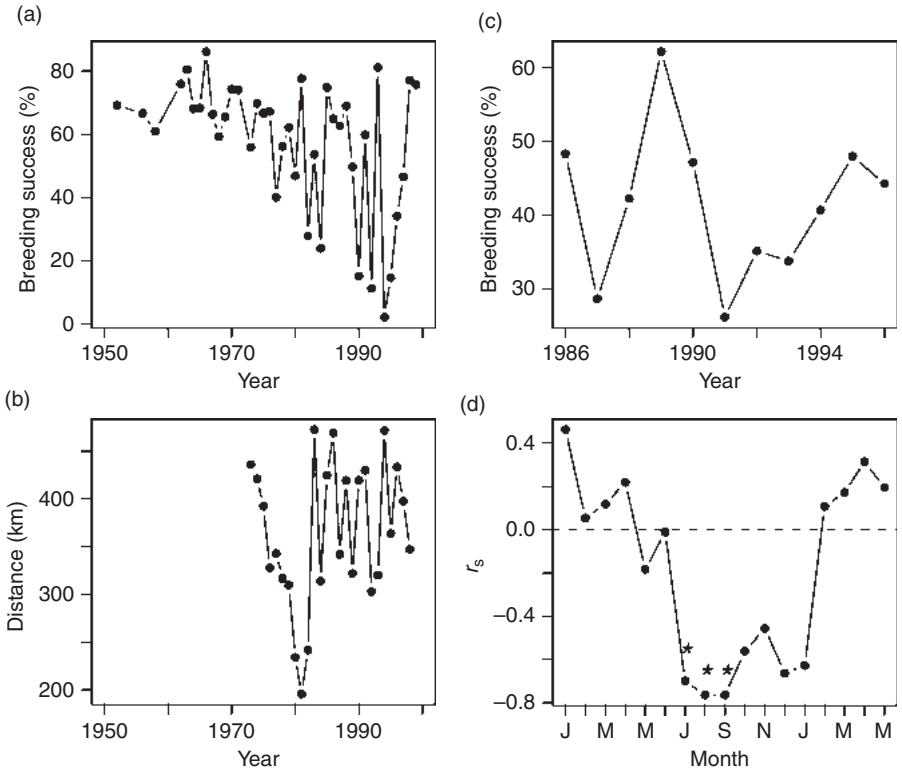


Fig. 3 Examples of fecundity fertility measures fluctuating with environmental stochasticity. (a) Breeding success (% chicks fledged of eggs laid) of emperor penguin (*Aptenodytes forsteri*) pairs from 1952 to 1999. Variation in breeding success increased progressively since the 1970s resulting from complete or extensive breeding failures in some years due to early breakout of the sea ice (b) or from prolonged blizzards during the early chick-rearing period. (c) Breeding success (% chicks fledged of eggs laid) of blue petrel (*Halobaena caerulea*) pairs from 1985 to 1996. (d) Spearman's rank correlation coefficients between sea surface temperature (SST) anomalies and blue petrel breeding success for the period ranging from February preceding the reproductive period until May the following year. Months where the correlation was nonzero are indicated by an asterisk (*). (a) Reproduced from Barbraud C and Weimerskirch H (2001) Emperor penguins and climate change. *Nature* 411: 183–186. (c) From Inchausti P, Guinet C, Koudil M, Durbec J-P, *et al.* (2003) Inter-annual variability in the breeding performance of seabirds in relation to oceanographic anomalies that affect the Crozet and the Kerguelen sectors of the Southern Ocean. *Journal of Avian Biology* 34(2): 170–176.

seeking to improve the genetic quality of their offspring by choosing genetically superior male mates to the existence of genetic correlations between female preference and male traits. Indeed, the direct benefits of female choice can be expressed in terms of increased fertility during a breeding season, or the evolution toward selecting particular mate characteristics to increase fecundity. Mate choice is therefore seen as an important component in the evolution of particular life-history strategies.

Models constructed to investigate mate choice can be complicated by the existence of multiple mating. In many species (e.g., the majority of insects), females permit or actively seek multiple mating with different males, even though several female fitness costs of excess mating are known (e.g., additional time and energy devoted to mating, increased predation risk, higher chance of injury, higher probability of disease or parasite transmission, male-originating chemicals reducing female longevity or female fecundity). Nonetheless, the ubiquity and magnitude of this trait suggest large positive effects on female fitness. For example, a fresh supply of sperm can maintain egg fertility, accessory substances can increase egg production rate, and the mating act itself can stimulate egg production.

However, the most studied component of multiple mating concerns sperm competition. This widespread phenomenon occurs when sperm from two or more males compete for a female's ova. Relative sperm numbers are important for sperm competitive success, so that species experiencing higher rates of sperm competition have males with larger testes that produce more sperm. Sperm competition is therefore a selective force shaping optimal ejaculate structure, although spermatogenesis is not unlimited. While males usually possess a greater reproductive potential than females, males have evolved mechanisms for the optimal allocation of finite sperm among females to maximize lifetime reproductive success.

Fecundity and Fertility in Population Dynamical Models

Being a cornerstone of population biology, quantifying fecundity and fertility patterns is a central component of models that attempt to describe complex population dynamics and structure. Many models exist exclusively to provide direction in the

understanding of the evolution of life-history strategies, while others have a more utilitarian function such as the estimation of a population's future status for conservation purposes. One of the most common methods used to summarize population behavior mathematically is the stage-structured matrix population model that provides a link between the individual (and its selective pressures) and the population using stage-specific estimates of vital rates (rates of birth, growth, maturation, fertility, and mortality). A more detailed description of matrix models is provided in this encyclopedia; however, there are some important aspects of these models that deal with fecundity and fertility specifically.

Stage-structured models provide an estimate of the stable stage distribution, which is the theoretical proportional allocation of individuals within the population to the defined stages (e.g., age groups, size classes, developmental stages) resulting from constant demographic rates. The stable stage distribution is a useful metric because it provides the theoretical composition of a population exhibiting a fixed birth rate, so factors such as environmental variation or intrinsic regulation that alter this theoretical distribution can be ranked for their effects on the future composition of a population. This introduces the concept of 'reproductive value' – a measure of the combined effects of fecundity, fertility, and survival that takes an individual's proportional contribution to the future status of its population into account. It is formally expressed as the sum of the current and future reproductive values, and is considered the currency used by natural selection to produce a particular life-history strategy.

Because life histories evolve to maximize reproductive success, fecundity is linked formally to population models by using the measure of reproductive value. In matrix models the relative effects of proportional changes in fertility (and survival) to population growth provide a sensitivity analysis for stage-specific terms. Here, the reproductive value for a particular stage is calculated as the product of the sensitivity of all matrix elements that contain that stage and the stable stage proportion. For example, short-lived species tend to demonstrate a relatively higher sensitivity for fertility than survival, while the reverse is often true for long-lived organisms. Thus, sensitivity analysis can be used in combination with observed variation in demographic rates to determine which factors have the largest impact on population growth and thus, the future composition of the population. Lifetime reproduction patterns expressed as reproductive allocation and ranked relative to other demographic rates therefore provide a means to compare the evolution of life-history strategies.

Genetic Considerations

An important consideration for population studies is the effect of inbreeding on reproductive fitness (the combination of fecundity, fertility, and survival). There is strong evidence to demonstrate that inbreeding reduces reproduction and survival in naturally outbreeding species and to a lesser extent in self-fertilization species, whereas increasing the incidence of outbreeding can reverse these deleterious effects. Not only does inbreeding reduce individual reproductive fitness, a population of inbred individuals often results in a declining population and an increased rate of inbreeding. As such, inbreeding is also expected to increase the risk of population extinction. This process begs the question: how much inbreeding can be tolerated without inbreeding depression? All finite closed populations will eventually become inbred without outbreeding, and even a low frequency of inbreeding is expected to result in some inbreeding depression. Effective population sizes of 50 or greater have been suggested as sufficient to avoid inbreeding depression, although it is generally unknown how large populations must be to avoid inbreeding depression in the long term. This number will also be highly variable among populations as a result of local dynamical processes and extrinsic forces.

See also: Behavioral Ecology: Kin Selection; Social Behavior and Interactions; Sexual Selection and Sexual Conflict; Age Structure and Population Dynamics; Mating Systems. Ecological Data Analysis and Modelling: Forest Models. Ecological Processes: Allometric Theory: Extrapolations From Individuals to Ecosystems. Evolutionary Ecology: Units of Selection; Life-History Patterns; Fitness. General Ecology: Demography; Generation Time; Age-Class Models

Further Reading

Agrawal, A.F., 2001. Sexual selection and the maintenance of sexual reproduction. Nature 411, 692–695.
Barbraud, C., Weimerskirch, H., 2001. Emperor penguins and climate change. Nature 411, 183–186.
Benton, T.G., Grant, A., 1999. Optimal reproductive effort in stochastic, density-dependent environments. Evolution 53, 677–688.
Buckley, Y.M., Hinz, H.L., Matthies, D., Rees, M., 2001. Interactions between density-dependent processes, population dynamics and control of an invasive plant species, *Tripleurospermum perforatum* (scentless chamomile). Ecology Letters 4, 551–558.
Bull, J.J., Shine, R., 1979. Iteroparous animals that skip opportunities for reproduction. American Naturalist 114, 296–303.
Charnov, E.L., Schaffer, W.M., 1973. Life history consequences of natural selection: Cole's result revisited. American Naturalist 107, 791–793.
Charnov, E.L., 2002. Reproductive effort, offspring size and benefit–cost ratios in the classification of life histories. Evolutionary Ecology Research 4, 749–758.
Cichoń, M., 2001. Diversity of age-specific reproductive rates may result from ageing and optimal resource allocation. Journal of Evolutionary Biology 14, 180–185.
Clutton-Brock, T.H., 1988. Reproductive Success. Studies of Individual Variation in Contrasting Breeding Systems. Chicago: University of Chicago Press.
Cole, L.C., 1954. The population consequences of life history phenomena. Quarterly Review of Biology 29, 103–137.
Hooper, H.L., Sibly, R.M., Hutchinson, T.H., Maund, S.J., 2003. The influence of larval density, food availability and habitat longevity on the life history and population growth rate of the midge *Chironomus riparius*. Oikos 102 (3), 515–524.

Inchausti, P., Guinet, C., Koudil, M., *et al*., 2003. Inter-annual variability in the breeding performance of seabirds in relation to oceanographic anomalies that affect the Crozet and the Kerguelen sectors of the Southern Ocean. Journal of Avian Biology 34 (2), 170–176.
McGovern, T.M., 2003. Plastic reproductive strategies in a clonal marine invertebrate. Proceedings of the Royal Society of London Series B – Biological Sciences 270, 2517–2522.
Myers, R.A., 2001. ICES. Journal of Marine Science 58, 937–951.
Nummi, P., Saari, L., 2003. Density-dependent decline of breeding success in an introduced, increasing mute swan *Cygus olor* population. Journal of Avian Biology 34 (1), 105–111.
Orzack, S.H., Tuljapurkar, S., 1989. Population dynamics in variable environments. VII. The demography and evolution of iteroparity. American Naturalist 133, 901–923.
Ranta, E., Tesar, D., Kaitala, A., 2002. Environmental variability and semelparity vs. iteroparity as life histories. Journal of Theoretical Biology 217, 391–396.
Rödel, H., Bora, A., Kaiser, J., *et al*., 2004. Density-dependent reproduction in the European rabbit: A consequence of individual response and age-dependent reproductive performance. Oikos 104 (3), 529–539.
Roff, D.A., 1992. The Evolution of Life Histories: Theory and Analyses. New York: Chapman and Hall.
Schaffer, W.M., 1974. Optimal reproductive effort in fluctuating environments. American Naturalist 108, 783–790.
Shine, R., 1988. The evolution of large body size in females: A critique of Darwin's 'fecundity advantage' model. American Naturalist 131, 124–131.
Solberg, E.J., Saether, B.-E., Strand, O., Loison, A., 1999. Dynamics of a harvested moose population in a variable environment. Journal of Animal Ecology 68 (1), 186–204.
Stearns, S.C., 1992. The Evolution of Life Histories. Oxford: Oxford University Press.

Fitness☆

Xia Hua and Lindell Bromham, Australian National University, Canberra, ACT, Australia

Nomenclature

W	Absolute fitness of a class of variant
w	Relative fitness of a class of variant
s	Selection coefficient of a class of variant
$\overline{w}$	Mean fitness of a population
ϕ	Value of a continuous trait or the presence of a class of variants
$\overline{\phi}$	Mean value of a continuous trait or the relative frequency of a class of variant in the population
δ	Change in ϕ during a time period that is due to factors other than directional selection
N	A vector of the number of individuals with each state in a class of variants
A	A matrix of the probability of a transition from each state to each of the other states in a class of variants
B	A matrix of the per capita rate of a transition from each state to each of the other states in a class of variants
C and D	Matrices of the probability of a transition from one class to another class in a population
$s_x(y)$	Invasive fitness of a new class x under the environment of a resident population y
$H(x)$	Shannon's entropy of the uncertain system x
$I(x;y)$	Mutual information of y on uncertain system x
x_t	Environmental state at time t
π_t	Representation of different classes of variants in a population at time t
π	The value that π_t averaged over time approaches as time goes on
G_{max}	The highest conceivable growth rate of a population in the uncertain environment
G_{min}	Long-term growth rate of a population with no information on the uncertain environment
$H'(x)$	Fitness cost of uncertainty on the environment x
$I'(x;y)$	Fitness value of information y on environment x
$D(\hat{\pi}\|\pi)$	Kullback-Leibler divergence from π to $\hat{\pi}$.
c	Cost of the behavior to the actor in kin selection
b	Benefit to recipient of the behavior in kin selection
r	Genetic relatedness between the actor and the recipient in kin selection

Glossary

Allele One of two or more alternative versions of a heritable trait, often corresponding to a DNA sequence difference at a particular locus (place) in the genome.

Class General term for an identifiable variant of a heritable trait, which may be present in multiple copies in the population. For example, a given allele is a class representing a variant DNA sequence for a particular locus (place) in the genome. There may be many copies of the allele in many individuals, but they all belong to the same class.

Evolutionarily stable strategy (ESS) A class (typically representing some form of strategy) or a particular mix of classes that cannot be replaced by any other classes because no other class or combination of classes will have a fitness advantage over it (relative to an environment).

Genotype While this term has several possible meanings, here we use it to mean the particular genetic variants carried by an individual: that is which allele or alleles it has in its genome.

Kin selection Selection based on inclusive fitness, which includes the fitness advantage due to helping relatives to reproduce so that they increase the representation of shared alleles in the population.

Malthusian parameter Per capita growth rate of the number of copies in a class (e.g., individuals with a particular heritable trait, or number of copies of a given allele in a population).

Polymorphism Occurrence of more than one class of variants (e.g., multiple alleles of a gene, or different strategies for a behavioral response) in a population.

Background

The term "fitness" can have a range of meanings in biology, sometimes referring to a relatively general concept, at other times as a label for a specific parameter. This variation in meaning can lead to some confusion, partly due to the changing roles the term fitness has played in evolutionary biology over time (Dawkins, 1982). Fitness had long been used to describe the way that organisms are evidently suited to their particular mode of life, or the way that parts of an organism work together (Paley, 1809). Herbert Spencer introduced the phrase "survival of the fittest" as equivalent to "natural selection" (Spencer, 1896). Alfred Russel

☆*Change History*: February 2018. X. Hua made minor changes to the text, figures, and references.

This is an update of J.A.J Metz, Fitness, In Reference Module in Earth Systems and Environmental Sciences, Elsevier, 2014.

 doi:10.1016/B978-0-12-409548-9.09361-1

Wallace adopted the phrase, equating it to Darwin's description "preservation of favoured races in the struggle for life" (Wallace, 1867), and he recommended Darwin do the same, in order to avoid some of the misunderstandings associated with the phrase natural selection, which some readers took to imply a conscious choice (akin to artificial selection). Darwin concurred, and used the phrase survival of the fittest in some of his later works, though he preferred the term natural selection (Darwin, 1868). So, at its most basic, fitness is the capacity of organisms to survive and reproduce in the environment in which they find themselves.

Under this broad concept, the fitness of an individual is considered to be reflected in the number of its descendants. But it is not clear how this broad concept of fitness should be measured in practise, because there is no obvious point in time at which descendants should be tallied. In a finite population every allele in the genome can be traced to a single ancestor at some point in the past, leaving little obvious leverage for comparing fitness between individuals or genetic variants. An alternative and more practical measure would be to measure fitness differences by comparing differences in lifetime production of offspring. However, the actual reproductive output of an individual is dependent not only on its own characteristics but also chance events, therefore fitness should be expressed as the probability distribution of the number of offspring. Classical models in population genetics suggest that natural selection maximizes the expected number of offspring as long as the relative expected number of offspring of different genetic variants stay constant over time and space, which usually requires constant (or infinitely large) population size and a stable environment (Fisher, 1930). When these assumptions are relaxed, natural selection no longer maximizes the expected number of offspring, but maximizes some more complicated form of fitness. For instance, under fluctuating environment, natural selection maximizes the geometric mean of the expected number of offspring over generations (Orr, 2009).

In most usages in evolutionary theory, fitness does not describe a property of a unique individual, but the property for a class of variants. Here we mean "class" in the sense of a group that all share the same feature, so our focus might on a specific allele, or a haplotype consisting of several base changes inherited as a unit (e.g., a mitochondrial variant), or individuals with the same trait of interest, or a phenotype consisting of several traits, or a life history strategy. For example, the use of antimalarial drugs places selective pressure on *Plasmodium* (malaria) parasites: any allele that confers resistance will be favored by selection, so individuals carrying this allele have a fitness advantage over susceptible individuals. Under these conditions, we expect the allele with the fitness advantage to steadily increase in the population (Roper *et al.*, 2004). Note that this is a statement about expected outcomes. We cannot predict the reproductive success of a particular *Plasmodium* individual based on whether or not it has the resistant allele —a susceptible individual may thrive if it is lucky enough to avoid being exposed to pyrimethamine, and a resistant individual may fail to reproduce when the mosquito it is carried in gets squashed by an irate human before it can achieve infection. But we can say that on balance, when considered across the whole population, over the period of observation, the favored allele had a higher rate of inclusion in subsequent generations that would be expected under random sample (Nair *et al.*, 2003).

Influence of Environment on Fitness

Fitness is always considered relative to a given environment. We would rarely consider an allele to have an absolute advantage under all sets of environmental conditions. For example, the fitness advantage of the resistance allele in populations of *Plasmodium* would disappear if pyrimethamine were not present in the environment (and may even be reversed if there is a cost to resistance). The "environment" of a class of variants can be considered to have many different levels.

On the genomic level, when a class is a specific allele, the "environment" of the class includes the other alleles in the genome in which it resides. No gene acts in isolation, but in combination with other genes, so the selective advantage of an allele depends on the presence of other alleles. Selection on one allele can carry linked neutral or nearly neutral alleles to fixation (hitchhiking: Maynard Smith and Haigh, 1974). Similarly, selection against a deleterious allele will removed linked alleles from the population (background selection: Charlesworth *et al.*, 1993). Both hitchhiking and background selection will result in a reduction in observed variation around a site under selection, thus one way of detecting alleles under selection is to detect chromosomal regions of reduced variability, which is interpreted as sign of a "selective sweep" that has driven linked alleles to fixation. For example, selection on the pyrimethamine resistant allele has led to reduced variation for 100 kb of linked sequence (Nair *et al.*, 2003).

The fitness of an allele can also depend on unlinked alleles in the same genome, or between nuclear and organelle genomes in the same cell, due to interactions between gene products. For example, some mitochondrial genes will be advantageous in combination with particular nuclear alleles, but have fitness costs when combined with other nuclear alleles, which can lead to cytonuclear conflict (Rand *et al.*, 2004). There may even be sex-specific differences in fitness costs and benefits of alleles. For example, some mitochondrial alleles lead to poor sperm performance, but because mitochondria are passed through the female line, alleles deleterious in males can persist for many generations (Gemmell *et al.*, 2004).

Because interactions with other alleles can have positive or negative effects on an allele's fitness, the "environment" of an allele against which fitness is tested will also include other alleles segregating in the population. In a sexual population, each generation is formed by sampling the alleles from one generation and reassorting them into individuals in the next generation. But alleles are not simply filtered through a "substitutional sieve" that preferentially lets those of higher fitness through to the next generation (Dobzhansky, 1937). The passage of an allele to the next generation will depend on whether it tends to have a fitness advantage when combined with other common alleles in the population, such that particular combinations of alleles will tend to result in higher rates of reproduction. This means that alleles that do not work well with other common alleles will tend to have reduced representation in subsequent generations. As a consequence, co-adapted alleles tend to accumulate in populations. Alleles within a

population are selected to work well with each other, but if lack of gene flow prevents those alleles being tested against variants in a neighboring population, then alleles that have high fitness in one population may have low fitness when combined with alleles in another population. This leads to hybrid individuals, who have a parent from each population, to have lower fitness than the offspring of parents from the same population, potentially generating selection for isolating mechanisms that prevent interbreeding, thus leading to the formation of genetically isolated species (see Hua and Bromham, 2017).

The fitness of some alleles may depend on characteristics of the population in which they are found. Some alleles may be beneficial under high population density, but lose this fitness advantage under low densities. Furthermore, when considering the fitness of particular behavioral strategies, the "environment" includes the frequency of other strategies in the same population. For example, the fitness of alleles influencing foraging behaviour in *Drosophila melanogaster* can be frequency dependent: in low nutrient environments, alternative alleles of the same gene that promote "roving" (moving around) or "sitting" (remaining) are both advantageous when rare, as rare strategies allow access to resources with less competition from others following the same strategy (Fitzpatrick *et al.*, 2007).

Three Alternative Ways of Defining Fitness

We have discussed general conceptions of fitness, and the way that fitness is influenced by the complicated web of interactions with other alleles, other individuals in the population and many aspects of the environment. Formal concepts of fitness are mostly mathematical and statistical, and there is a range of different parameters that have been labeled as fitness. These mathematical definitions of different fitness parameters differ from the conceptual definition of fitness that we have just discussed (Otsuka, 2016). Below, we describe some bases of the most popular theories: the Price equation, Evolutionarily Stable Strategy (ESS) theory, and information theory. To make this discussion general to any population that can undergo evolution by natural selection, we are going to refer to the variants under selection as a "class." The "class" might refer to alleles, strategies, lineages, groups, or any other unit where we can compare the relative reproductive success of one class of heritable variants against others.

The Price Equation

The Price equation was derived by Price (1970), originally for the purpose of re-deriving W.D. Hamilton's work on kin selection (Hamilton, 1964), but it is now considered as a fundamental theorem that summarizes all the equations and theorems in population genetics. In population genetics, fitness, or more specifically, the absolute fitness (W) of a class of variants typically refers to the class's expected number of progeny multiplied by its chance of survival (Orr, 2009). Often, population geneticists normalize this fitness by the absolute fitness of the fittest class and call this relative fitness (w), so that the fittest class has a relative fitness of 1. Now the relative fitness of all the other classes can be written in a form of $1 - s$, where s is called the selection coefficient. The mean fitness of a population ($\overline{w}$) is then the weighted sum of the relative fitness of each class in the population and the weight is the frequency of each class. For example, if a class is a genotype and a population contains two different genotypes, then the mean fitness of the population is $\overline{w} = p_1 w_1 + p_2 w_2$, where p_1 and w_1 are the frequency and the relative fitness of one genotype and p_2 and w_2 are the frequency and the relative fitness of the other genotype. Similarly, if a class is a given value of a continuous trait (z), then the relative fitness is described as a function of trait value, $w(z)$, and the mean fitness of a population integrates the fitness over the total range of values that the trait could take $\overline{w} = \int p(z)w(z)dz$.

The Price equation describes the average change in the representation of a class ϕ in a population during a time period ($\Delta\overline{\phi}$). The class can be a given value of a continuous trait, such as body size or some measure of fitness itself, and $\overline{\phi}$ is the mean value of the trait or the mean fitness of the population. The class can also be a categorical type, such as the allele of interest, and $\overline{\phi}$ is the relative frequency of the category in the population. According to the Price's equation, $\Delta\overline{\phi}$ is the sum of two parts (Queller, 2017):

$$\Delta\overline{\phi} = \frac{1}{\overline{w}}[\mathrm{Cov}(w_i, , \phi_i) + E(w_i\delta_i)]$$

where i denotes a class of variants. The first part, $\frac{1}{\overline{w}}\mathrm{Cov}(w_i, , \phi_i)$, represents change in $\overline{\phi}$ due to directional natural selection. When ϕ is the relative fitness itself: $w_i = \phi_i$, the covariance between fitness and the class $\mathrm{Cov}(w_i, \phi_i)$ becomes the variance for fitness, and so the first part of the Price equation captures Fisher's fundamental theorem of natural selection that states natural selection increases the mean fitness of a population at a rate equal to the additive genetic variance for fitness (Fisher, 1930). The second part, $\frac{1}{\overline{w}}E(w_i\delta_i)$, where δ_i is the change in the value of ϕ_i during the time period, describes the expected change in $\overline{\phi}$ due to factors other than directional selection, such as mutation, migration, a change in the environment, and so on. Classical models in population genetics show that natural selection maximizes the expected number of offspring of a population $\overline{w}$, because these models assume constant relative fitness, that is, $\delta_i = 0$; in other words, the second part of the Price's equation is ignored.

In principle, the Price equation can predict changes in the representation of any kind of class in a population under natural selection as long as the class has heritability (Queller, 2017). For example, a class can be a suit of co-adapted alleles or traits, so that fitness depends on multiple alleles or traits. Then $\mathrm{Cov}(w_i, \phi_i)$ in the first part not only accounts for the covariance between fitness and each trait, but also the association among traits. The second part of the Price equation can be modified to account for changes in $\overline{\phi}$ due to various factors other than directional natural selection. For example, consider frequency-dependent selection on sex ratio. Class ϕ is a given value of the number of male offpsring relative to female offspring. Starting with an initial condition

where there are more females than males in the population, females who have more male offspring should have higher fitness in a random, multiple mating population, because their offspring will have higher reproductive success in total. However, when more males are produced, environment for these males changes because there is stronger competition for mates; in other words, δ_i becomes negative because some male offspring will not reproduce. As a result, there may be a point where the two parts of Price's equation cancels out, that is, $\Delta\overline{\phi} = 0$. Under this condition, natural selection keeps ϕ at an optimal point.

Evolutionarily Stable Strategy (ESS) Theory

The concept of the evolutionarily stable strategy was first formulated by John Maynard Smith, who applied game theory to study the evolution of animal behaviors (Maynard Smith, 1974). Game theory essentially looks for the existence of strategy equilibria given the expected payoff of each strategy. Here, the payoff is the fitness of the strategy, often estimated by the Malthusian parameter that was proposed by Ronald Fisher to describe the rate of increase in the number of individuals as an aggregate of individual fecundity and mortality (Fisher, 1930). This is also a measure on the logarithm scale of the absolute fitness used in population genetics. In principle, ESS theory can be generalized to any class of variants, such as a new mutation, a new trait value, or a new strategy, so hereafter we use "class" instead of "strategy."

We can write changes in the number of individuals for discrete time as:

$$\mathbf{N}(t+1) = \mathbf{A}\mathbf{N}(t)$$

or for continuous time as:

$$\frac{d}{dt}\mathbf{N}(t) = \mathbf{B}\mathbf{N}(t)$$

For a class of variants where each variant has multiple states, for example, age state often matters for a reproductive strategy because only mature individuals can reproduce, $\mathbf{N}(t)$ records the number of individuals at each state, and components of matrix **A** are demographic parameters, with the i, j-th components of matrix **A** describing the probability of a transition from state j to state i. For age state, matrix **A** is the classic Leslie matrix, in which the first row is the average number of female offspring born from mother of each age state and diagonal of the rest of the matrix is the fraction of individuals that survive from an age state to the next age state. Matrix **B** is similar to matrix **A**, except that its components are per capita rates. The Malthusian parameter for a given class is then the natural log of the dominant eigenvalue of matrix **A** or the dominant eigenvalue of matrix **B** (Metz *et al.*, 1992).

We can also calculate the Malthusian parameter of a population that consists of multiple classes. This is analogous to the mean fitness of a population used in population genetics. To do this, we combine the matrix **A** or **B** of each class into a larger matrix. For example, if there are two classes in the population, class 1 has matrix **A1** and class 2 has matrix **A2**, they are combined into one matrix as:

$$\begin{bmatrix} \mathbf{A1} & \mathbf{D} \\ \mathbf{C} & \mathbf{A2} \end{bmatrix}$$

where matrix **C** describes the probability of a transition from class 1 to class 2, and matrix **D** describes the probability of a transition from class 2 to class 1. These transition probabilities can have various forms, depending on the genetic bases of the two classes. For example, in a haploid population, the transition probabilities may reflect the mutation rate between classes. When the population can grow exponentially, natural selection will maximize the Malthusian parameter of the population; otherwise the population size will eventually reach equilibrium where the Malthusian parameter equals 0.

ESS theory is often used to ask two questions related to the relative fitness of classes in a population. First, can a new class invade an existing population? Second, if it can, will the new class replace or be replaced by any existing classes in the resident population, or will it coexist with the resident? To study the first problem, the theory introduces a new parameter $s_x(y)$, the invasive fitness of a new class x under the environment of a resident population y. Assuming that mutations are rare so that the resident population is always at its strategy equilibrium, when a new mutation enters the population, $s_x(y)$ equals the Malthusian parameter of the population that consists of both the new and the existing classes. If $s_x(y) > 0$, the new class has a positive probability to invade the resident population. If $s_x(y) < 0$, the new class is doomed to extinction (Metz, 2008).

As the new class forms a greater proportion of the population, its influence to the environment of the population gets larger. So to predict the fate of the new class in long run, we not only account for the influence of the resident classes to the environment, but also the influence of the new class on the environment. Organism's influence on the environment can be modeled by expressing the demographic parameters in the matrix **A** as a function of "feedback variables" that describe the source of the influence (Rueffler *et al.*, 2013). For example, under density dependence, the fraction of individuals that survives from an age state to the next age state may depend on the total number of individuals in the population due to competition. Here, the total number of individuals in the population is the feedback variable. At the strategy equilibria, the Malthusian parameter of the population equals 0, so one way to look for the existence of strategy equilibria is to prove that there are finite number of conditions under which the Malthusian parameter of the population equals 0 (Diekmann, 2004).

An ESS view of fitness is particularly helpful in cases where the fitness effects of classes have complicated interactions. For example, a non-transitive set of behaviors can form a loop, in which each class has higher fitness that one alternative class but lower fitness than another alternative class. This is like a "rock-scissors-paper" game (rock beats scissors, scissors beats paper, paper beats rock). In other words, if $s_x(y) > 0$ and $s_y(z) > 0$, no one class can replace the others when $s_x(z) < 0$ (Gyllenberg and Service,

2011). For example, *Escherichia coli* that can produce the antibacterial compound colicin (C) can have a fitness advantage over colicin-sensitive strains (S), but colicin-resistant strains (R) have a reproductive advantage over C by avoiding the cost of producing colicin. But the resistant strains (R) have a slower growth rate than sensitive strains (S). This is a non-transitive fitness loop because C beats S, S beats R and R beats C, leading to the potential for stable co-existence of the three strategies (Kerr *et al.*, 2002).

Another condition under which no one class can replace the others is when there are more than one "feedback variable" (Diekmann, 2004). For example, to study the evolution of annual versus biannual strategies in a population, we may express the Malthusian parameter of the population as a function of the relative frequency of the annual strategy in the population. If the survivability over the first winter is affected by the number of newborns and the survivability over the second winter is affected by the number of one-year-old individuals, then there are two feedback variables: the number of newborns and the number of one-year-old individuals. It has been shown that under this condition, populations with different frequencies of the annual strategy perform equally well (Diekmann, 2004). In contrast to the wide interest in finding a good proxy for fitness, the ESS theory suggests that the conditions under which natural selection will maximize the fitness of a population are actually limited.

Information Theory

The fundamental question in the information theory is how to reduce the error rate of data communication over noisy channels, or more abstractly, how to increase the amount of information on uncertain systems. To answer this question, we first need a measure of uncertainty: Shannon's entropy (Shannon, 1948). Given a variable x that could take N states, each with a probability p_i, the entropy of the variable is:

$$H(x) = -\sum_{i=1}^{N} p_i \log\ p_i$$

which counts the expected number of bits needed to code the variable x. Now we have a piece of information y on x. The value to attain the information is measured by how much uncertainty it reduces or how many bits it saves to code variable x, which we called mutual information (Cover and Thomas, 1991):

$$I(x;y) = H(x) - H(x|y)$$

If we treat environment as the uncertain system x and measure the value of information y on the environment in terms of the amount of increase in fitness, then it is more straightforward to account for uncertainty in the environment using information theory than the Price equation or ESS theory. With environmental uncertainty, long-term growth rate of a population becomes a more relevant proxy of fitness than the instantaneous Malthusian parameter because maximizing the growth rate at one time does not guarantee positive growth rate at another time. The long-term growth rate of a population can be intuitively defined as the arithmetic mean of the Malthusian parameter during each time step that is small enough that the environment during a time step is assumed constant (Donaldson-Matasci *et al.*, 2010). Mathematically, the long-term growth rate is the dominant Lyapunov exponent that describes the asymptotic rate of increase in population size (Metz, 2008).

Early work by Haldane (1957) and Kimura (1961) and recent studies on evolution under fluctuating environments (Donaldson-Matasci *et al.*, 2010; Rivoire and Leibler, 2011) suggest a tight link between fitness as formulated by information theory and the long-term growth rate of a population. To illustrate the link, we denote the environmental state at time t as x_t. The population will increase the representation of the class of variants that has the highest absolute fitness at time t in the population based on the information y_t they attain on x_t. The representation of different classes of variants at time t is denoted as π_t. The information can be directly acquired from the current environment, if π_t are mediated by phenotypic plasticity as a response to the current environmental cues, such that π_t does not depend on π_{t-1}. But more often, the information is inherited. The information on the current environment comes through the selective transmission of alleles or traits that are better suited to the previous environments than others, such that π_t depends on both π_{t-1} and the current environment. As time goes on, π_t averaged over time will approach a real number π. Our goal is to find the π that maximize the fitness or the long-term growth rate of the population.

It is intuitive that the maximum conceivable growth rate for a population occurs when the class of variants with the highest absolute fitness has 100% representation in the population at any point in time, so that the population is tuned perfectly to x_t. The worst situation is when traits associated with high fitness in past generations provide no information on likely success in the current environment. Under this situation, a population is most likely to survive over time if the relative frequency of different classes in the population is proportional to the relative frequency of different environmental states over time. Now, we denote the maximum conceivable growth rate as G_{max} and the long-term growth rate in the absence of information as G_{min}. The relationship between G_{max} and G_{min} follows (Rivoire and Leibler, 2011):

$$G_{min} = G_{max} - H'(x)$$

where $H'(x)$ is the fitness cost of uncertainty on the environment. When the information on the environment is imperfect, that is a population with π_{t-1} does not behave optimally under x_t, the maximum long-term growth rate of the population with $\hat{\pi}$, the optimal value of π is (Rivoire and Leibler, 2011):

$$G(\hat{\pi}) = G_{max} - H'(x) + I'(x;y)$$

where y is the imperfect information on the environment x and $I'(x;y)$ is the fitness value of information y on x. The selective

disadvantage of a population with an arbitrary value of π is related to $D(\hat{\pi}\|\pi)$ (Frank, 2012), the Kullback-Leibler divergence from π to $\hat{\pi}$, which is used in the information theory to measure the number of extra bits required to code variable π using a code optimized for $\hat{\pi}$ rather than the code optimized for π.

It has been shown that when π_t does not depend on π_{t-1}, and when the absolute fitness of the population in x_t is nonzero for only one class of variants, $H'(x)$ equals $H(x)$, the Shannon's entropy of the environment, $I'(x;y)$ equals $I(x;y)$, the mutual information on the environment, and the selective disadvantage of π equals $D(\hat{\pi}\|\pi)$ (Donaldson-Matasci *et al.*, 2010; Rivoire and Leibler, 2011). Under other conditions, the forms of $H'(x)$ and $I'(x;y)$ are more complicated, but $H(x)$ and $I(x;y)$ still defines the upper bounds of these values (Donaldson-Matasci *et al.*, 2010; Rivoire and Leibler, 2011).

Individual Fitness

The three different formulations we have discussed are related, in that each can be derived from each of the others (Demetrius and Ziehe, 2007; Frank, 2012). But these different formulations may be suited to answering different empirical questions. For example, the Price equation is a convenient way to investigate which features of organisms are correlated with fitness. ESS theory is well-suited to examining why populations may contain one or more alternative strategies. Information theory allows us to consider how environmental uncertainty affects fitness in the long run. These are all questions that biologists may wish to ask of real populations. In order to do so, we need to be able to measure the fitness relevant parameters, such as the Malthusian parameter, in real populations, in the lab or the field.

Field ecologists often measure fecundity and survivability by tracking marked individuals through their life span. So the major problem of estimating the Malthusian parameter of a class of variant or of a population in the field is how to translate discrete individual events—the birth, death, and reproduction of an individual—into the growth rate of the class or the population. For example, while any given individual will have a specific value for age at death that reflects both fitness and chance events, the survivability measures the proportion of individuals that survive to a given age in the population, so reflects the fitness of a given class of individuals.

Fisher (1930) showed that, without demographic and environmental stochasticity, the sum of the reproductive value of all individuals in the population increases at the population's asymptotic growth rate, regardless of whether the population is in stable age distribution. This suggests that if there is a measure of the reproductive contribution of each individual to the class or the population (Fisher, 1930), sometimes called the individual fitness (Sæther and Engen, 2015), then we can estimate the fitness of the class or the population by tracking some individuals of the class or the population through their life span. Using individual fitness is also a logic way to accounting for demographic and environmental stochasticity, because these sources of stochasticity must be reflected in the variation of individual fitness (Engen *et al.*, 2009).

Traditionally, individual fitness is measured by the number of offspring produced by an individual during its lifetime. However, using this individual fitness to estimate the fitness of the class or the population that the individual belongs to require additional assumptions, including constant population size and stable age distribution (Sæther and Engen, 2015). To avoid these assumptions, at least two new measures of individual fitness have been proposed. One measure defines individual fitness as the sum of the number of offspring an individual produced during a given time step (divided by 2 if sexual reproduction) and a dummy variable describing whether the individual is alive or not during the time step (Sæther and Engen, 2015). This measure is readily applicable to the Price equation by replacing w_i with the individual fitness and ϕ_i with the feature value of the individual. Another measure calculates individual fitness as the dominant eigenvalue of a matrix that is similar to the Leslie matrix, where the first row is the number of offspring produced at each age of the individual (divided by 2 if sexual reproduction) and the diagonal of the rest of the matrix is 1 if the individual is alive at the age and 0 if not alive (McGraw and Caswell, 1996). The geometric mean of individual fitness converges to a value as the number of individual increases, but the value is a biased estimate of the Malthusian population growth rate (McGraw and Caswell, 1996).

There are other measures of individual fitness, but they require estimating population growth rate or some demographic parameters of the population first. For example, one measure calculates individual contributions to population growth by comparing population growth rate estimated from datasets before and after removing the individual and its offspring that are produced during a given time step (Coulson *et al.*, 2006). These measures are still useful to remove demographic and environmental uncertainties because they can capture the variation in individual fitness either due to chance events that happen to different individuals during their lifespan or variation in the environment around an individual, such that some individuals happen to live in a good year and others in bad years.

Fitness Beyond the Individual

The fitness of a class of heritable variants, whether defined at the level of the genotype or the phenotype, is due to it having properties that increase its representation in the next generation. But, given that fitness concerns the expected outcomes for a class of heritable variants, not the fates of any given individual carrying the trait, an individual does not have to reproduce to increase the fitness of the class of variants. If an allele causes its carrier to help others with the same allele, even at personal cost, then it may result in a net fitness advantage if more copies of the allele are included in the next generation than would have been without the

altruistic action (Haldane, 1955). The more closely individuals are related, the more alleles they will share in common, hence the famous quote attributed to J.B.S. Haldane that he would risk his life to save two brothers from drowning, or eight cousins (Maynard Smith, 1975). This idea is captured in Hamilton's (1964) inclusive fitness, which writes the fitness effect of having a particular co-operative behaviour as

$$w = 1 - c + br$$

where c is cost of the behavior to the actor, b is the benefit to recipient, r is the genetic relatedness between the actor and the recipient. "1" is the baseline fitness as used in population genetics. In the ESS theory, this is the fitness of the resident population, but on the logarithm scale $\log(1) = 0$. The ESS theory predicts that a new variant can invade the resident population if its fitness is greater than the fitness of the resident population. So natural selection favors a particular co-operative behaviour when $1 - c + br > 1$, that is $br > c$ (Hamilton, 1964).

But individuals do not have to be close kin to share an allele whose fitness is increased by co-operation. If an allele within a population leads carriers of the same allele to co-operate to raise their net reproductive output, it will have a selective advantage. Such a situation is referred to as a "green beard" phenomenon: an allele that causes both green beards and propensity to help others with green beards could lead to evolution of altruism by providing a heritable basis for the fitness benefits of co-operation (Dawkins, 1979). For example, in the highly invasive fire ant *Solenopsis invicta*, a gene that codes for an odor-binding protein acts as a green beard: workers carrying the b allele help queens with the b allele to reproduce, and tend to kill queens that lack b (Keller and Ross, 1998). This is not strictly a case of kin selection, because it is irrelevant in this case whether the queens with the b allele are related to the workers or not, it is the presence or absence of the "green beard" olfactory cue allele that determines co-operation.

Individuals often inherit more than just genes from their parents. For example, if offspring tend to be located closer to their parental home than other members of the population are, then they inherit their environment as well as their genes. Since organisms also influence their own environment, natural selection can create feedback between the environment and the organism through niche construction. Organisms may change the physical structure of the world (e.g., termite mounds or birds nests), the physical form of other organisms (e.g., galls induced by wasps infecting trees), or the behaviour of other organisms (e.g., parasite manipulation of host behaviour). Influences on the environment may be neutral with respect to fitness or even deleterious (e.g., exhaustion of resources). But if there are heritable influences on environment that increase the fitness of a class of heritable variants by increasing its relative representation in subsequent generations, then we expect this "extended phenotype" to be under natural selection (Dawkins, 1982).

Because the fitness effect may extend beyond the physical boundaries of an individual organism, the fitness of a class of heritable variants depends not only on features of other classes of variants in the conspecific population, but also on the heritable features found in other species. These tangled fitness relationships blur the line of demarcation between individuals, and even between species. It is not always obvious when a mutually beneficial relationship between endosymbiont and host (e.g., termites and the bacteria that allow them to digest cellulose) becomes a dual genetic inheritance system (e.g., nuclear and mitochondrial genes). For example, braconid wasps reproduce by parasitizing caterpillars, but cannot do so without genes encoded in a virus that is amplified and transmitted in the wasps' eggs. The fitness of any alleles in the wasp genome depends on their co-transmission with viral genes, and vice versa (Louis *et al.*, 2013).

The consideration of combined fitness effects of alleles in different organisms leads us to the question of whether the concept of fitness can be applied to communities or ecosystems. In the informal sense, we might consider communities to have features that aid survival and propagation, such as robustness or resilience, and these features could be ultimately due to the net genetic information in the system (Van Valen, 1989). However, the formal theoretical sense of fitness concerns relative reproduction rate, and it is not clear what we might be comparing the fitness of an ecosystem to. To apply the theoretical framework of fitness to multi-species groups we would need to be able to define a system where communities compete for some key resource, that they have the capacity for reproduction (not just persistence) and that they have heritable traits that influence their chances of representation in subsequent generations.

See also: Behavioral Ecology: Kin Selection; Environmental Stress and Evolutionary Change. Evolutionary Ecology: Fecundity; Units of Selection; Natural Selection; Adaptation. General Ecology: Age-Class Models

References

Charlesworth, B., Morgan, M.T., Charlesworth, D., 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134, 1289–1303.
Coulson, T., Benton, T.G., Lundberg, P., Dall, S.R.X., Kendall, B.E., Gaillard, J.M., 2006. Estimating individual contributions to population growth: Evolutionary fitness in ecological time. Proceedings of the Royal Society B: Biological Sciences 273, 547–555.
Cover, T.M., Thomas, J.A., 1991. Elements of information theory. New York: Wiley-Interscience.
Darwin, C., 1868. The variation of animals and plants under domestication. London: John Murray.
Dawkins, R., 1979. The selfish gene. Oxford: Oxford University Press.
Dawkins, R., 1982. The extended phenotype. Oxford: Oxford University Press.
Demetrius, L., Ziehe, M., 2007. Darwinian fitness. Theoretical Population Biology 72, 323–345.
Diekmann, O., 2004. A beginners guide to adaptive dynamics. In: Rudnicki, R. (Ed.), Mathematical modelling of population dynamics. Volume 63 of Banach Center Publications. Warsaw: Polish Academy of Sciences, pp. 47–86.

Dobzhansky, T.G., 1937. Genetics and the origin of species. New York: Columbia University Press.
Donaldson-Matasci, M.C., Bergstrom, C.T., Lachmann, M., 2010. The fitness value of information. Oikos 119, 2197–2230.
Engen, S., Lande, R., Sæther, B., Dobson, F.S., 2009. Reproductive value and the stochastic demography of age-structured populations. The American Naturalist 174, 795–804.
Fisher, R.A., 1930. The genetical theory of natural selection. Oxford: Oxford University Press.
Fitzpatrick, M.J., Feder, E., Rowe, L., Sokolowski, M.B., 2007. Maintaining a behaviour polymorphism by frequency-dependent selection on a single gene. Nature 447, 210–212.
Frank, S.A., 2012. Natural selection. V. How to read the fundamental equations of evolutionary change in terms of information theory. Journal of Evolutionary Biology 25, 2377–2396.
Gemmell, N.J., Metcalf, V.J., Allendorf, F.W., 2004. Mother's curse: The effect of mtDNA on individual fitness and population viability. Trends in Ecology & Evolution 19, 238–244.
Gyllenberg, M., Service, R., 2011. Necessary and sufficient conditions for the existence of an optimisation principle in evolution. Journal of Mathematical Biology 62, 359–369.
Haldane, J.B.S., 1955. Population genetics. New Biology 18, 34–51.
Haldane, J.B.S., 1957. The cost of natural selection. Journal of Genetics 55, 511–524.
Hamilton, W.D., 1964. The genetical evolution of social behaviour. II. Journal of Theoretical Biology 7, 17–52.
Hua, X., Bromham, L., 2017. Darwinism for the genomic age: Connecting mutation to diversification. Frontiers in Genetics 8, 12.
Keller, L., Ross, K.G., 1998. Selfish genes: A green beard in the red fire ant. Nature 394, 573–575.
Kerr, B., Riley, M.A., Feldman, M.W., Bohannan, B.J.M., 2002. Local dispersal promotes biodiversity in a real-life game of rock-paper-scissors. Nature 418, 171–174.
Kimura, M., 1961. Natural selection as process of accumulating genetic information in adaptive evolution. Genetics Research 2, 127–140.
Louis, F., Bézier, A., Periquet, G., Ferras, C., Drezen, J.-M., Dupuy, C., 2013. The bracovirus genome of the parasitoid wasp *Cotesia congregata* is amplified within 13 replication units, including sequences not packaged in the particles. Journal of Virology 87, 9649–9660.
Maynard Smith, J., 1974. The theory of games and the evolution of animal conflicts. Journal of Theoretical Biology 47, 209–221.
Maynard Smith, J., 1975. Survival through suicide. New Scientist 28, 496–497.
Maynard Smith, J., Haigh, J., 1974. The hitch-hiking effect of a favourable gene. Genetical Research 23, 23–35.
McGraw, J.B., Caswell, H., 1996. Estimation of individual fitness from life-history data. The American Naturalist 147, 47–64.
Metz, J.A.J., 2008. Fitness. In: Jørgensen, S., Fath, B. (Eds.), Evolutionary ecology. Volume 2 of Encyclopedia of Ecology. Oxford: Elsevier, pp. 1599–1612.
Metz, J.A.J., Nisbet, R.M., Geritz, S.A.H., 1992. How should we define "fitness" for general ecological scenarios? Trends in Ecology & Evolution 7, 198–202.
Nair, S., Williams, J.T., Brockman, A., Paiphun, L., Mayxay, M., Newton, P.N., Guthmann, J.-P., Smithuis, F.M., Hien, T.T., White, N.J., 2003. A selective sweep driven by pyrimethamine treatment in southeast asian malaria parasites. Molecular Biology and Evolution 20, 1526–1536.
Orr, H.A., 2009. Fitness and its role in evolutionary genetics. Nature Reviews Genetics 10, 531–539.
Otsuka, J., 2016. A critical review of the statisticalist debate. Biology and Philosophy 31, 459–482.
Paley, W., 1809. Natural theology or evidences of the existence and attributes of the deity, 12th edn. London: J. Faulder.
Price, G.R., 1970. Selection and covariance. Nature 227, 520–521.
Queller, D.C., 2017. Fundamental theorems of evolution. The American Naturalist 189 (4), 345–353. (early view).
Rand, D.M., Haney, R.A., Fry, A.J., 2004. Cytonuclear coevolution: The genomics of cooperation. Trends in Ecology & Evolution 19, 645–653.
Rivoire, O., Leibler, S., 2011. The value of information for populations in varying environments. Journal of Statistical Physics 142, 1124–1166.
Roper, C., Pearce, R., Nair, S., Sharp, B., Nosten, F., Anderson, T., 2004. Intercontinental spread of pyrimethamine-resistant malaria. Science 305, 1124.
Rueffler, C., Metz, J.A.J., Van Dooren, T.J.M., 2013. What life cycle graphs can tell about the evolution of life histories. Journal of Mathematical Biology 66, 225–279.
Sæther, B., Engen, S., 2015. The concept of fitness in fluctuating environments. Trends in Ecology & Evoluton 30, 273–281.
Shannon, C.E., 1948. A mathematical theory of communication. Bell System Technical Journal 27.379–423 & 623–656.
Spencer, H., 1896. The principles of biology. New York and London: D. Appleton.
Van Valen, L.M., 1989. Three paradigms of evolution. Evolutionary Theory 9, 1–17.
Wallace, A.R., 1867. Mimicry, and other protective resemblances among animals. Westminster and foreign quarterly review 32, 1–43.

Gause's Competitive Exclusion Principle☆

Jamie M Kneitel, California State University, Sacramento, CA, United States

Introduction

The "competitive exclusion principle" (CEP), usually attributed to Gause (1934), states: "Two species with identical niches (and compete for a single resource) cannot coexist together indefinitely." This statement, in its simplicity, represents one of the most influential, compelling, and controversial concepts in ecology. Indeed, it pervades almost all theoretical and empirical approaches to species coexistence and diversity. Further, it is a conceptually centralizing paradox in ecology: if similar competing species cannot coexist, then how do we explain the great patterns of diversity that we observe in nature? If species living together cannot occupy the same niche indefinitely, then how do competitors coexist? Community ecology strives to reconcile this paradox and explain the mechanisms of species coexistence: niche differences among species have been the primary explanation for the maintenance of species coexistence.

Conceptual Beginnings

Since the beginning of ecology in the 19th century, research has centered on resource competition as the primary interaction regulating and structuring populations and communities. As far back as Darwin, ecologists have acknowledged that species occupied roles in their environment and that ecologically similar species, especially closely related species, will compete strongly and result in the extinction of poor competitors. Alternatively, species may ecologically or evolutionarily diverge in their resource use, or niche, resulting in reduced niche overlap. These interactions represented the perceived driving mechanisms for speciation and much of the diversity of life on this planet. Early in the 20th century, Grinnell (1917) had coined the term "niche" to refer to species' requirements and their use of habitat. At that time, he had been expressing for over a decade that no two species can occupy the same niche. A decade later, Elton elaborated on the term to also mean a species' role and effects in its habitat.

"The Struggle of Existence"

The concept of "struggle for existence" has been around for millennia, but it was coined by Darwin (1859). Darwin used it in regard to individuals of a species or different species interacting strongly for limited resources; however, we observe species persistence in communities with the plethora of these strong interactions. This leads to numerous questions regarding community level patterns, including the nature of species dominance, persistence, and extinction; the structure and assembly of communities; the patterns of biodiversity; and ultimately the mechanisms leading to observed community patterns. Early in the 20th century, Lotka and Volterra separately developed mathematical equations that addressed the population dynamics of competing species and the conditions when extinction or coexistence will occur in these populations. For example, on a single resource, only one species can persist (Fig. 1), but if intraspecific competition is greater than interspecific competition in the two competing populations, then similar species may stably coexist. In 1934, Gause published *The Struggle for Existence*, which consisted of the first experimental tests of Lotka–Volterra equations. Specifically, Gause used unicellular organisms (protozoa and yeast) in a series of laboratory experiments to quantify species interactions in competition to determine the conditions necessary for coexistence. One of Gause's competition experiments showed that two *Paramecium* species exhibited logistic growth when grown alone, but when grown together, one was always driven to extinction (Fig. 1). Naturalists and theoreticians had previously mentioned this pattern of competitive exclusion; Gause united the niche and competitive interactions conceptually and empirically.

☆*Change History*: March 2018. Jamie Kneitel made minor changes to the references and text.

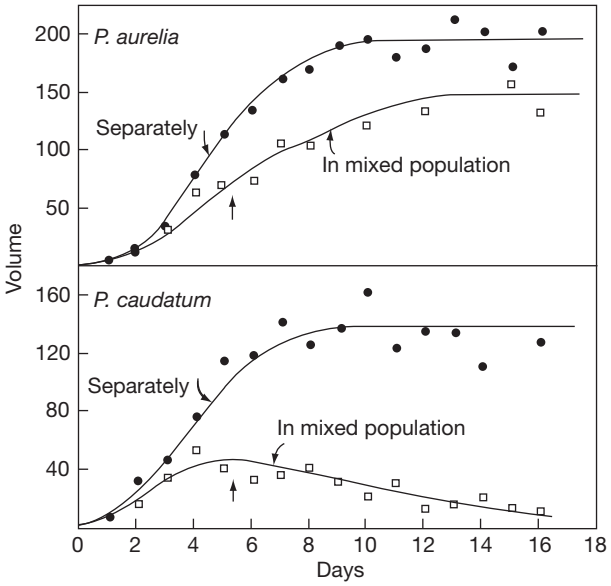


Fig. 1 Population growth for *Paramecium aurelia* and *Paramecium caudatum* when grown alone ("separately") and in competition ("in mixed population"). This is an example of competitive exclusion between two similar species. From Gause, G.F. (1934). *The Struggle for Existence.* Baltimore: Williams and Wilkins.

Gause's Contributions

Gause's book was influential on future studies in the areas of competition, the niche, and species coexistence. His approach of experimentally testing mathematical models has become a standard for ecologists, and his experimental results have become common textbook examples of competition and predator–prey interactions. In the years following publication of his book, various authors referred to the "competitive exclusion principle" as Gause's hypothesis, axiom, law, postulate, contention, or thesis. However, it was not until Hardin (1960) that the term "competitive exclusion principle" was coined. The article summarized the historical development of the principle and its various uses in ecology. Hardin also provided a critique of the principle and ecologists' ability to test it. For example, the CEP is not falsifiable: extinction between competitors provides support for it, but if species coexist, then the conclusion is that the species must somehow be ecologically different (i.e., different niches). This naturally leads to the question of "how different is different?" (see also excellent discussion in Palmer (1994)). Nonetheless, competition experiments dominate the ecological literature to this day, and the approaches have become more quantitative and rigorous in experimental designs.

Species Coexistence

The study of biodiversity is the study of how competitive exclusion is foiled—through the exploitation of heterogeneity and pattern in the environment, and through the evolutionary displacement of the ways species see and utilize the environment (Levin, 2000).

Ecologists have long sought to determine the factors that allow competing species to coexist in nature. Competitive exclusion (extinction) is one possible outcome of competitive interactions, and ecologists have determined several potential outcomes and mechanisms for species coexistence, typically associated with trait differences among species as implied by CEP. As mentioned previously, these differences result in intraspecific interactions being stronger than interspecific interactions. For the past several decades, Lotka–Volterra equations have been further developed and elaborated to identify conditions in which coexistence is possible. One conclusion from several theoretical studies in the 1960s pointed out that n species can coexist only when the number of resources or limiting factors is greater than the number of species, $k > n$. Other work has pointed to nonlinear functional response relationships with resources, varying densities, and temporal variation in resources as other conditions facilitating coexistence.

Early studies by Tansley and later by Park showed that the nature of competitive interactions may differ, in fact may be reversed, under different environmental conditions. In Tansley, conspecifics of the plant genus *Galium* differentially dominated in different soil types, and Park found *Tribolium* beetle species to differentially dominate under different temperature and humidity conditions. Recent examples of this common pattern highlight that the strength of competitive interactions can vary with habitat conditions in space or time, and coexistence may be occurring because of habitat heterogeneity along spatial and/or temporal scales. Therefore, species coexistence and exclusion are contingent on habitat conditions and spatiotemporal scale.

Resource competition can also lead to character displacement, whereby competitors evolutionarily diverge in their resource use, reduce niche overlap, and thereby coexist. This competition-driven process assumes that individuals of species who overlap in their niche (and this being directly related to the degree of competition) will have lowered fitness, causing directional selection. Species will then evolve decreased niche overlap and competition, which may also be reflected in morphological differences. The evidence

required to support the presence of character displacement should include (1) nonrandom size differences, (2) genetic basis for trait differences, (3) differences resulting from evolutionary shifts, (4) morphological differences resulting in differential resource use, and (5) competitive interactions must be established.

The ecologist G. E. Hutchinson contributed immeasurably to the development of the niche concept. Hutchinson's "Homage to Santa Rosalia" article addressed why so many species were able to coexist and was a pivotal landmark in bridging the relationship between the niche and competition. He pointed out that species that were too similar morphologically (e.g., body size) and ecologically will result in competitive exclusion of one species. However, coexistence between closely related species will occur when the species are sufficiently different in body size and therefore their resource use. This minimum difference in body size was estimated to be 130% or a factor of 1.3, which came to be known as the Hutchinsonian ratio; the measuring of size ratios as an estimate of prior character displacement became the focus of much research for a generation of ecologists trying to understand community structure and species coexistence. Ultimately, this approach faced an array of criticisms based on statistical and inferential grounds. For example, in many communities, the size differences were not different than randomly selected sets of species, the patterns were not different than random distributions of size differences. Further, species that differ in body size may not always compete less than similarly sized species.

For several decades, there was a decline in the use of the term "niche," but the tide has turned. In recent years, there has been a resurgence and renewed interest in the niche as a central concept and explanation for species coexistence. This has also resulted in the conceptual development of how the niche is measured by using traits. Further, the "niche" has been enhanced (e.g., species traits) and promoted as a central theme for the future of ecology, reflecting previous work on fundamental and realized niche. Another update on the view of the niche has been the concept of niche tradeoffs as a mechanism for species coexistence (Fig. 2).

One criticism of the niche explanations for coexistence is that they have difficulty explaining high levels of species diversity. In recent years, the metacommunity concept has emerged to address community dynamics at different spatial scales, local and regional. Processes at the local community scale are the smaller-scale factors affecting communities, including species interactions with each other (e.g., competition, predation) and their environment. At the regional community scale, processes among local communities dominate, including dispersal and habitat heterogeneity. The processes at local and regional scales contribute to community dynamics and patterns of abundance and diversity. A number of perspectives have emerged within this framework for explaining diversity and abundance patterns. One explanation for diversity patterns has been the neutral theory of species diversity. In these models, species are assumed to be identical ecologically and demographically, and diversity and relative abundance is driven by metacommunity size, and random (stochastic) variation in demographic (births and deaths), dispersal, and speciation rates. Empirical support has been variable for neutral theory, but it has proven to be an influential alternative to the niche perspective. It appears that both niche (deterministic) and neutral (stochastic) processes act in community dynamics. While most approaches to the niche viewed differentiation on a single dimension (except for Hutchinson's *n*-dimensional hypervolume), niche tradeoffs view two or more niche axes where species traits have a negative functional relationship (Fig. 2). For example, a species' ability to compete may come at the cost of defending against predators (Fig. 2), which has been documented in numerous aquatic and terrestrial communities. Further, ecologists are increasingly becoming aware of the importance of spatial dynamics that can contribute to species coexistence. The competitive–colonization ability tradeoff has been hypothesized as a mechanism for species coexistence in space, with some empirical support. Consequently, species niche differences can be exhibited along numerous niche axes along on spatial and temporal planes, potentially explaining higher levels of diversity.

In *The Origin of Species*, Darwin suggested that closely related species should compete strongly since they are more likely to be similar in their resource use (i.e., niche). Empirical (experimental and observational) support for this observation has been rather weak, but new molecular and computational approaches have facilitated hypotheses on the relative importance of evolutionary history and ecological traits for species coexistence. Two alternative hypotheses, overdispersion or clustering, are proposed for relationships among species occupying local communities. Phylogenetic overdispersion is found when coexisting species are less related than expected by chance, a pattern that assumes competitive exclusion and consistent with Darwin's observation. In contrast, phylogenetic clustering is found when coexisting species are more related than expected by chance, which implies environmental

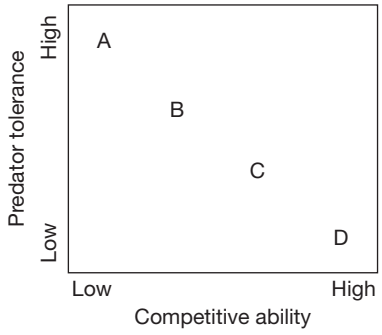


Fig. 2 An example of tradeoffs among species A–D along two niche axes, competitive ability and predator tolerance, which will hypothetically facilitate coexistence.

filtering is more important than competition in determining community patterns. Empirical studies have found support for both hypotheses. Additionally, factors other than competition (e.g., herbivore effects) are often found to influence these patterns.

The Competitive Exclusion Principle is one of ecology's most influential concepts. Empirical evaluations have been diverse and support has been varied for the importance of competitive exclusion in structuring communities. In addition, other explanations for coexistence patterns include stochastic processes, dispersal limitation, apparent competition, and abiotic conditions. Nonetheless, CEP and its modern descendants will continue to inspire ecological research and develop with technological, empirical, and theoretical advances.

References

Darwin CR (1859) *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: Murray.
Gause GF (1934) *The struggle for existence*. Baltimore, MD: Williams and Wilkins.
Grinnell J (1917) The niche-relationships of the California thrasher. *The Auk* 34: 427–433.
Hardin G (1960) The competitive exclusion principle. *Science* 131: 1292–1297.
Levin SA (2000) Multiple scales and the maintenance of biodiversity. *Ecosystems* 3: 498–506.
Palmer MW (1994) Variation in species richness: Towards a unification of hypotheses. *Folia Geobotanica et Phytotaxonomica* 29: 511–530.

Further Reading

Armstrong RA and McGehee R (1980) Competitive exclusion. *American Naturalist* 115: 151–170.
Cavender-Bares J, Kozak KH, Fine PVA, and Kembel SW (2009) The merging of community ecology and phylogenetic biology. *Ecology Letters* 12: 693–715.
Chase JM and Leibold MA (2003) *Ecological Niches*. Chicago: University of Chicago Press.
Chave J (2004) Neutral theory and community ecology. *Ecology Letters* 7: 241–253.
Chesson P (2000) Mechanisms of maintenance of species diversity. *Annual Review of Ecology and Systematics* 31: 343–366.
Dayan T and Simberloff D (2005) Ecological and community-wide character displacement: The next generation. *Ecology Letters* 8: 875–894.
Elton CS (1927) *Animal Ecology*. London: Sidgwick and Jackson.
Hubbell SP (2001) *The unified neutral theory of biodiversity and biogeography*. Princeton, NJ: Princeton University Press.
Hutchinson GE (1959) Homage to Santa Rosalia or why are there so many kinds of animals? *American Naturalist* 93: 145–159.
Kingsland SE (1995) *Modeling nature: Episodes in the history of population ecology*. Chicago: University of Chicago Press.
Kneitel JM and Chase JM (2004) Trade-offs in community ecology: Linking spatial scales and species coexistence. *Ecology Letters* 7: 69–80.
Lack D (1949) *Darwin's finches*. London: Cambridge University Press.
Levin SA (1970) Community equilibria and stability, and an extension of the competitive exclusion principle. *American Naturalist* 104: 413–423.
Lotka AJ (1925) *Elements of physical biology*. Baltimore: Williams and Wilkins.
MacArthur R and Levins R (1964) Competition, habitat selection, and character displacement in a patchy environment. *Proceedings of the National Academy of Sciences of the United States of America* 51: 1207–1210.
Mayfield MM and Levine JM (2010) Opposing effects of competitive exclusion on the phylogenetic structure of communities. *Ecology Letters* 13: 1085–1093.
Tilman D (1982) *Resource competition and community structure*. Princeton, NJ: Princeton University Press.
Tilman D (2004) Niche tradeoffs, neutrality, and community structure: A stochastic theory of resource competition, invasion, and community assembly. *Proceedings of the National Academy of Sciences of the United States of America* 101: 10854–10861.
Udvardy MFD (1959) Notes on the ecological concepts of habitat, biotope and niche. *Ecology* 40: 725–728.
Vamosi SM, Heard SB, Vamosi JC, and Webb CO (2009) Emerging patterns in the comparative analysis of phylogenetic community structure. *Molecular Ecology* 18: 572–592.
V. Volterra, Variations and fluctuations of the number of individuals in animal species living together. Animal Ecology. (1928) McGraw-Hill, New York; In: Chapman RN (trans) (1931).
Wake DB, Hadly EA, and Ackerly DD (2009) Biogeography, changing climates, and niche evolution: Biogeography, changing climates, and niche evolution. *Proceedings of the National Academy of Sciences of the United States of America* 106: 19631–19636.

Genetic Drift

Olivier Honnay, University of Leuven, Heverlee, Belgium

Glossary

Allele One form of a gene or a specific chromosome region (locus). For one diploid organism, two forms are possible.

Effective population size The number of individuals in an ideal population that has the same properties with respect to genetic drift as the actual population does. The effective population size is usually smaller than the census population size because not all individuals contribute to the offspring to the same degree.

Genetic differentiation The amount of the genetic variation in an organism that is present among different spatially separated populations. Genetic differentiation may vary between 0.00 (all populations possess exactly the same alleles) and 1.00 (all populations have no alleles in common).

Heterozygote An organism possessing two (in the case of a diploid organism) different forms of a particular gene or specific chromosome region (locus), one inherited from each parent.

Introduction

Genetic drift, also known as the 'Sewall Wright effect', is one of four factors (next to mutation, gene flow, and natural selection) causing a gene pool to change over time. Genetic drift is the random variation in allele frequencies between generations due to sampling error in finite populations. As an example consider a single locus with two alleles, A and a, with equal frequencies $p=0.5$ and $q=0.5$, in a population of 10 diploid parents that interbreed and produce 10 offspring. Although the probability of drawing either allele a or A is 0.5, it is likely that a random sampling of 20 alleles from the available allele pool will yield a slightly different allele frequency in the offspring. For example, allele frequencies in the second generation might have shifted to $p=0.4$ and $q=0.6$, purely due to chance effects. Continuing with these second-generation allele frequencies, and after a second random sampling of 20 alleles, in the third generation deviations from the initial 0.5 frequencies will become even more likely. Note that this process of genetic drift shows that the Hardy–Weinberg equilibrium, predicting constant allele frequencies over time, does not hold in finite populations. Genetic drift is also nondirectional and as likely to decrease as to increase the frequency of one particular allele. Although genetic drift is an evolutionary process (because allele frequencies are changing), it does not directly change the degree of adaptation of an individual or a population.

Example

In a classic experiment with 107 small ($N=16$) populations of the fruit fly *Drosophila melanogaster*, each consisting of eight males and eight females, Peter Buri in 1956 has empirically shown how initial allele frequencies ($p=q=0.5$) of a gene coding for eye color can dramatically change after relatively few generations. The evolution of the frequency distribution of the alleles, over all experimental populations, can be seen in **Fig. 1**. Each population can be considered as a random sampling of 32 alleles from the available pool. After 19 generations, one of the alleles has reached a frequency of 1.0 in almost half of the 107 experimental populations (it is said that the allele has become 'fixed' in that population), implying that the other allele was lost from the population. This illustrates a first very important consequence of genetic drift: The loss of genetic diversity. The proportion of populations where fixation for a certain allele is expected to occur is equal to the initial frequency of that allele. **Fig. 1** also shows a second important consequence of genetic drift. Because allele frequencies are changing in different directions in each population (one allele becomes more frequent in one population, whereas it decreases in frequency in another population), the populations start to differentiate from each other. It is said that genetic differentiation among populations is increasing. A third consequence of genetic drift is that as the population becomes fixed for one allele, the heterozygosity (H_t) in the population (the frequency of heterozygotes) in generation t is expected to decline each generation. Starting from an initial heterozygosity (H_0), this happens according to the rule: $H_t=(1-1/2N_e)^tH_0$. This means that heterozygosity will decrease much faster in populations with small effective population size (N_e).

Genetic Drift and Population Size

Drift is much more important in small populations because the sampling error effect causing changes in allele frequencies is relatively small in large populations. In the absence of dominance or epistatic variance, genetic variance will be lost at the same

 doi:10.1016/B978-0-444-63768-0.00616-8

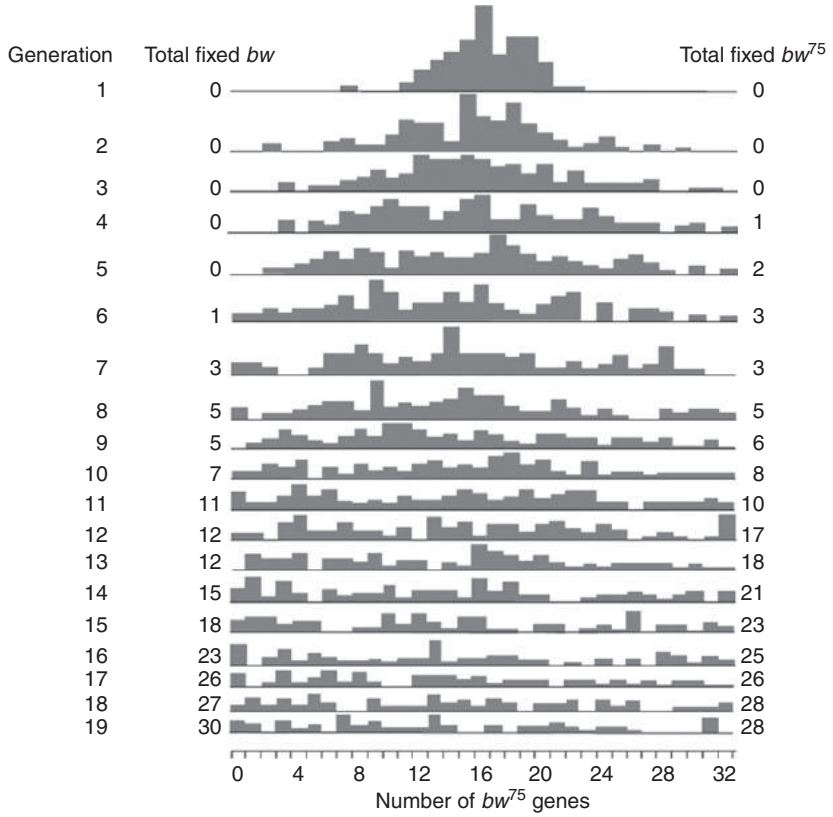


Fig. 1 Empirical histogram of allele frequencies of two alleles in 107 experimental populations of the fruit fly *D. melanogaster* over 19 generations. All populations started with an allele frequency of $p=0.5$. Redrawn from Buri, P., 1956. Gene frequency in small populations of mutant *Drosophila*. Evolution 10, 367–402.

rate as heterozygosity, that is, at a rate of $1/2N_e$ per generation. **Fig. 2** shows simulations of the random change in the frequency distribution of a single hypothetical allele over 20 generations for two different-sized populations. In the smallest population, alleles are driven to fixation in a very short time, while in the larger population allele frequencies remain more constant. Note that on average, the frequencies of the alleles remain constant across all populations, but that the frequencies start to diverge among populations. Kimura and Ohta in 1971 have presented an expression for the mean time until fixation (T) of an allele with an initial frequency of q: $T=(-4N_e(1-q)\ln(1-q))/q$. It can be seen that the time until fixation (and hence until the loss of genetic variation) is a linear function of the effective population size (N_e).

Bottleneck and Founder Effects

Even populations that are currently large may be genetically impoverished because genetic drift has been important in the past. This may be the case when the population went through a bottleneck (the bottleneck effect), or when it is derived from a small number of founding parents (the founder effect). Genetic bottlenecks occur when populations show a drastic reduction in population size, for example, due to temporally deteriorating environmental conditions, followed by population growth when the environmental conditions improve again. Bottleneck effects and founder effects imply that also the genetic properties of currently large populations can be shaped by genetic drift.

Minimal Viable Population Size

Genetic drift is an important concept in conservation biology where often small and spatially isolated populations of threatened plant and animal species are studied. Often crucial is the estimation of the minimum viable population size. The minimum viable population size is an estimate of the number of individuals required for a high probability of survival of a population over a given period of time. A commonly used definition is a higher than 95% probability of persistence over 100 years. One important requirement to prevent extinction of a population is to maintain sufficient population genetic diversity to allow for adaptation to changing environmental conditions. Genetic drift should then be minimized, and at most equal the mutation rate within the

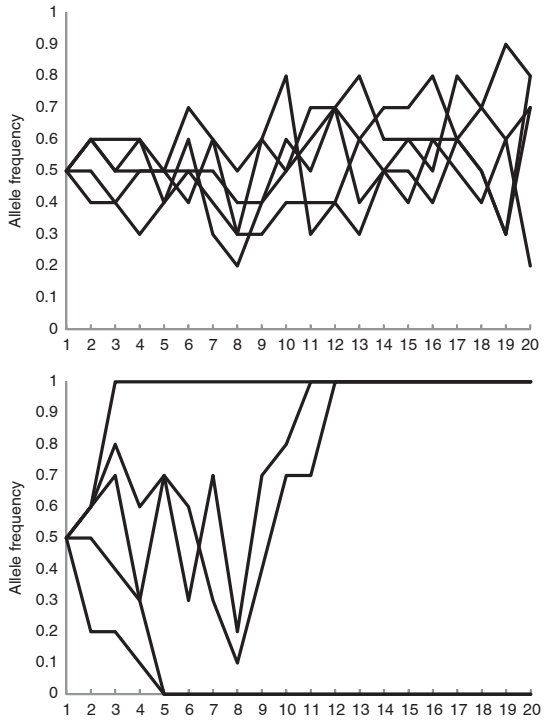


Fig. 2 Hypothetical frequency of an allele A for a number of replicate small (10 individuals, below) and large (100 individuals, above) populations across 20 generations.

population. In 1980, Ian Franklin has proposed that there is an equilibrium between loss of genetic diversity through drift and gain of genetic diversity through mutation, when $1/2N_e = 0.001$, this is when $N_e = 500$ individuals. This rule of thumb for minimum viable population size has become widespread among conservation biologists, although recent meta-analyses have put forward minimum viable population sizes of at least a few thousand individuals. Note that the census population size should usually be much larger than the effective population size, because not all individuals in a population are contributing to the genetic diversity of the offspring. Many factors can increase the difference between the effective population size and the census population size: An unequal numbers of males and females, variance between individuals in likelihood to produce offspring, high variance in the number of offspring between individuals, nonrandom mating, and fluctuations in the number of breeding individuals from one generation to the next. Populations where N_e equals the census size are said to be ideal.

Genetic Drift and Gene Flow

The loss of alleles through drift from a population, and the concomitantly increasing population genetic differentiation, can be counteracted if there is sufficient gene flow among populations. Gene flow can occur through the active migration of individuals or through the passive dispersal of seeds and pollen. It can be shown that for a large number of ideal populations: $F_{ST} = 1/(4N_e m + 1)$, where F_{ST} is a measure of the genetic differentiation between populations, m is the proportion of individuals that migrate between populations per generation, and N_e is the effective population size of the populations. This implies that one incoming migrant per generation in each population ($N_e m = 1$) results in a population genetic differentiation of 0.20. This is considered as still acceptable and has been coined the 'one-migrant-per-generation-rule', which states that receiving one migrant per generation is sufficient to prevent genetic drift from reducing the population genetic variation and increasing the genetic differentiation. It should be noted that a genetic differentiation of 0.20 is a relatively arbitrary threshold, and that in the case of nonideal populations, much more migrants are required to counteract the effects of genetic drift.

Genetic Drift and Evolutionary Theory

Genetic drift is at the core of the shifting-balance theory of evolution coined by Sewall Wright where it is part of a two-phase process of adaptation of a subdivided population. In the first phase, genetic drift causes each subdivision to undergo a random walk in allele frequencies to explore new combinations of genes. In the second phase, a new favorable combination of alleles is fixed in the subpopulation by natural selection and is exported to other demes by factors like migration between populations.

Much of the basic theory of genetic drift was developed in the context of understanding the shifting-balance theory of evolution. Genetic drift has also a fundamental role in the neutral theory of molecular evolution proposed by the population geneticist Motoo Kimura. In this theory, most of the genetic variation in DNA and protein sequences is explained by a balance between mutation and genetic drift. Mutation slowly creates new allelic variation in DNA and proteins, and genetic drift slowly eliminates this variability, thereby achieving a steady state. A fundamental prediction of genetic drift theory is that the substitution rate in genes is constant, and equal to the mutation rate.

See also: Behavioral Ecology: Dispersal–Migration. Conservation Ecology: Conservation Genetics; Connectivity and Ecological Networks. Ecological Complexity: Ecological Indicators: Connectance and Connectivity. Ecological Data Analysis and Modelling: Modeling Dispersal Processes for Ecological Systems. Evolutionary Ecology: Hardy–Weinberg Equilibrium. General Ecology: Migration and Movement

Further Reading

Franklin, I.R., 1980. Evolutionary change in small populations. In: Soule, M.E., Wilcox, B.A. (Eds.), Conservation Biology: An Evolutionary–Ecological Perspective, first ed. Sunderland, MA: Sinauer, pp. 135–169.
Franklin, I.R., Frankham, R., 1998. How large must populations be to retain evolutionary potential? Animal Conservation 1, 69–70.
Jacquemyn, H., Vandepitte, K., Roldán-Ruiz, I., Honnay, O., 2009. Rapid loss of genetic variation in a founding population of *Primula elatior* (Primulaceae) after colonization. Annals of Botany 103, 777–783.
Kimura, M., 1983. The Neutral Theory of Molecular Evolution. New York: Cambridge University Press.
Kimura, M., Ohta, T., 1969. The average number of generations until fixation of a mutant gene in a finite population. Genetics 61, 763–771.
Mills, L.S., Allendorf, F.W., 1996. The one-migrant-per-generation rule in conservation and management. Conservation Biology 10, 1509–1518.
Slatkin, M., 1987. Gene flow and the geographic structure of natural populations. Science 236, 787–792.
Templeton, A.R., 2006. Population Genetics and Microevolutionary Theory. Hoboken, NJ: Wiley.
Traill, L.W., Bradshaw, J.A., Brook, B.W., 2007. Minimum viable population size: A meta-analysis of 30 years of published estimates. Biological Conservation 139, 159–166.
Wright, S., 1931. Evolution in Mendelian populations. Genetics 16, 97–159.
Wright, S., 1969. Evolution and the Genetics of Populations, Vol. II: The Theory of Gene Frequencies. Chicago, IL: University of Chicago Press.

Relevant Websites

http://science.jrank.org— Science Encyclopedia, Genetic Drift.
http://www.nature.com/scitable/topicpage/genetic-drift-and-effective-population-size-772523— Scitable, Genetic Drift and Effective Population Size.

Hardy–Weinberg Equilibrium

Patrick G Meirmans, University of Amsterdam, Amsterdam, Netherlands

Glossary

Allele A variant form of a gene or genetic marker; a gene or genetic marker is only regarded to be variable when there are multiple alleles present in a species or population.

Genetic drift Change in the frequency of an allele resulting from random sampling of parents and random sampling of gametes to produce offspring. The strength of genetic drift is inversely related to population size.

Inbreeding coefficient Summary statistic that quantifies the level of inbreeding in a population or, more generally, the degree of deviation from Hardy–Weinberg equilibrium. Also called F_{IS} F, f, or the fixation index.

Locus The location of a gene or genetic marker on a chromosome.

Type I error In statistical significance testing this is the probability of rejecting the null hypothesis when it is actually true.

Type II error In statistical significance testing this is the probability of not rejecting the null hypothesis when it is actually false.

Introduction

The Hardy–Weinberg (HW) principle is one of the cornerstones of evolutionary theory. Arguably, the whole field of population genetics started with the first description of the HW principle (Crow, 1988). After all, this was the first time that Mendelian genetics was scaled up from describing the inheritance of alleles in pedigrees to a quantitative description of genotype frequencies in populations. The fundamental points of the HW principle can be neatly summarized in two postulates (following Waples, 2015):

1. After a single generation of random mating, the frequencies of the different genotypes at a locus can be expressed as a simple function of the allele frequencies.
2. In the absence of genetic drift, selection, mutation and migration, the allele and genotype frequencies remain stable over time.

Even though, historically, the HW principle arose from a discussion about the stability of allele frequencies under Mendelian inheritance (postulate 2), current applications of the HW principle almost exclusively concern the calculation of the expected genotype frequencies (postulate 1).

For a single biallelic locus with alleles A and a, there are three possible genotypes: AA, Aa, and aa. When the frequency of allele A in a population is p and that of allele a is q (where $q = 1 - p$), the expected frequencies of these genotypes are p^2, $2pq$, and q^2, respectively, where $p^2 + 2pq + q^2 = 1$ (see **Fig. 1**). A similar expansion can be made for multi-allelic loci, though the number of terms in the equation gets rather cumbersome for loci with many alleles. When the observed genotype frequencies in a population match the expected frequencies, the population is said to be in Hardy–Weinberg equilibrium (HWE). Note, however, that this is a special kind of "equilibrium" as for diploids it is reached within a single episode of random mating. This sets it apart from other equilibria in genetics, such as linkage equilibrium and migration-drift equilibrium, which are approached asymptotically over time.

Testing for deviation of observed genotype frequencies from the HWE expectations is often the first step in the analysis of genetic data. This is because many subsequent analyses rely on an assumption of HWE in order to work correctly. Testing for HWE deviation is furthermore often employed to check for the possible presence of genotyping errors that can be removed before proceeding with the further analyses.

Quantifying the degree of deviation from HWE is often done using the summary statistic F_{IS} (also called F, f, the fixation index, or the inbreeding coefficient; Wright, 1922). F_{IS} compares the observed frequency of heterozygotes (H_O) with the expected frequency under HWE (H_E, also frequently indicated as H_S): $F_{IS} = 1 - H_O/H_E$. When a population is in HWE, the observed and expected heterozygosity are equal and F_{IS} has a value of zero. However, the range of the statistic goes from -1 to 1, with negative values indicating an excess of heterozygotes and positive values a shortage of heterozygotes. Note that a value of -1 can only be reached when there are only heterozygotes; for a biallelic locus this is only possible when $p = q = 0.5$. Comparing F_{IS} values among populations or species can be highly informative on how differences in mating system or demography can influence the distribution of genetic variation.

History

Mendelians Versus Biometricians

After the rediscovery of Mendel's laws in 1900, evolutionary biologists in England were engaged in a fierce debate that dominated the field for over a decade. On the one side were the "Mendelians," headed by William Bateson. The mendelians thought that only discontinuous

 doi:10.1016/B978-0-12-409548-9.10555-X

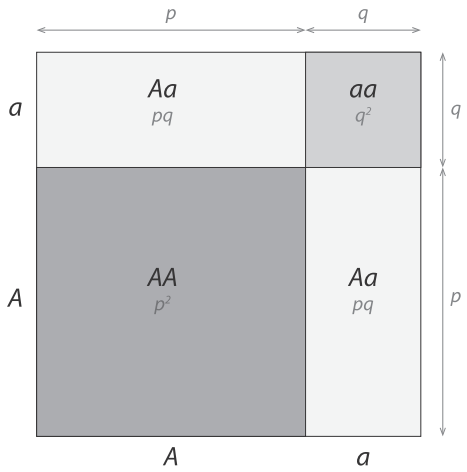


Fig. 1 "Punnet square" diagram illustrating how the frequencies of the three genotypes of a biallelic loci relate to the allele frequencies. The length of the sections labeled "*A*" and "*a*" on the *X* and *Y*-axes correspond to their frequencies ($p = 0.7$ and $q = 0.3$). The area of the rectangles in the plot thus corresponds to the genotype frequencies.

characters could lead to evolution and saw an important role for Mendel's laws of inheritance for this. On the other side were the "biometricians," headed by Karl Pearson. The biometricians put most emphasis on continuous, quantitative, traits. Though they did recognize the validity of Mendel's experiments, they thought that his results were only of moderate interest as they did not see how they could explain the quantitative variation seen in nature. This argument was only fully resolved in 1918 when Fisher published his landmark paper (Fisher, 1918) on combining quantitative and mendelians genetics. In any case, both camps were studying Mendel's laws. The Mendelians did so in order to show how it could explain evolutionary significant variation; the biometricians did so to show that it could not explain what they thought were the most important evolutionary processes.

Among the biometricians was G. Udny Yule, who had joined the group around Pearson in 1893. In 1902, Yule published a long-winded paper (Yule, 1902) that he started with a diatribe against Bateson and his writing style, before explaining his "law of ancestral heredity." In the second half of the paper he discusses Mendel's laws and in doing so actually reveals the HW genotype ratios for $p = q = 0.5$. In a later paper, Pearson (1904) does the same, though without acknowledging the prior work of his co-worker. Surprisingly, both scientists failed to see how these genotype ratios could be generalized beyond allele frequencies of exact one half.

In February 1908, Yule attended a lecture by the mendelian R.C. Punnett (who lent his name to the Punnett square, see **Fig. 1**) titled "Mendelism in relation to disease." In this lecture, Punnett gave an extensive explanation of Mendel's laws and Bateson's work on it, adapted for a medical audience. In the subsequent discussion, Yule led the argument against the mendelian school. Yule argued that for a dominant trait, such as brachydactyly, one would expect, over the course of time and assuming random mating, to have three brachydactylous persons for every normal person. Yule pointed out that this was obviously not the case in the human population. Punnet was taken aback by this question and could not readily provide a satisfactory answer (Edwards, 2008).

Geoffrey H. Hardy

G.H. Hardy was the leading mathematician of his time. In a way he can be regarded as the prototypical English gentleman scholar; a staunch bachelor, he devoted his time between mathematics, cricket, and conversation at Cambridge University's "High Table". Punnet and Hardy played cricket together, so naturally Punnett posed Yule's problem to Hardy. Hardy quickly grasped the concept and provided Punnett with a satisfactory answer. He then proceeded to write a short letter to the editor of Science (Hardy, 1908), which was published in July 1908. In it, he expressed both his reluctance "to intrude in a discussion concerning matters of which I have no expert knowledge" and his exasperation about the apparent lack of ability of biologists to perform "mathematics of the multiplication-table type." He then pointed out the stability of both allele and genotype frequencies, which afterwards became known as "Hardy's law."

Wilhelm Weinberg

Meanwhile in Stuttgart, Germany, a physician was persistently working on a large volume of papers on a wide range of subjects, but mostly focusing on genetics. Wilhelm Weinberg spent all his working life as both a private physician and a volunteer physician to the poor. His specialism being obstetrics, he is said to have delivered no <3500 babies (Crow, 1999). Among these must have been a good number of twins and children with hereditary defects, which may have triggered an interest in these subjects. In his spare time Weinberg wrote around 160 scientific papers.

In January 1908—3 months before Hardy wrote his letter to the editor—Weinberg gave a lecture titled "Über den Nachweis der Vererbung beim Menschen," to the society for natural sciences of the state of Württemberg. A corresponding paper was published in the yearbook of the society in the fall of the same year (Weinberg, 1908). In this paper Weinberg discusses several subjects

regarding human genetics, most notably the heredity and genetics of twinning. At the end of the paper, Weinberg explained the expected genotype frequencies under random mating and their stability, doing so in a much clearer and more extensive fashion than Hardy in his short letter. Similar to Mendel's paper 40 years earlier, Weinberg's paper was written in German and published in an obscure journal and consequently completely ignored by the scientific establishment. After all, the field of genetics was largely dominated by English speakers. It was not until 1943—6 years after Weinberg's death—that Curt Stern, a German geneticist working in the U.S. called attention to Weinberg's paper (Stern, 1943) and declared that the rule regarding the genotype frequencies should be known as the "Hardy–Weinberg law."

Assumptions

The HW principle as posited above is only valid in a highly idealized population where there are no perturbations affecting the genotype and allele frequencies (Waples, 2015). Therefore, the principle as a whole makes a number of assumptions: random mating, infinite population size, no mutation, no selection, a single population, no migration, non-overlapping generations, and diploid inheritance. However, each has different effects on the two postulates of the HW principle. Most importantly, factors that affect the stability of allele frequencies (postulate 2) do not necessarily have an effect on the expected genotype frequencies (postulate 1). Here, I discuss violations of these assumptions one by one and for each explain whether and how it affects the two postulates. Note that all these assumptions regard biological processes; in many cases, deviation from HWE is caused by genotyping errors, a topic that is discussed later.

Non-Random Mating

In an infinitely large population, non-random mating has no effect on the allele frequencies and so does not invalidate postulate 2. On the other hand, the genotype frequencies (postulate 1) are more strongly affected by the violation of the assumption of random mating than by the violation of any of the other assumptions. Truly random mating is rare, and there are many reasons why organisms may not mate randomly. In animals, there may be polygyny, where one male has multiple females, or polyandry, where one female has multiple males, which causes the dominant adults to have a disproportionate number of offspring. A similar type of non-random mating occurs when the sex-ratio is biased. Another type of non-random mating is assortative mating: preferentially mating among individuals with a similar phenotype. If the phenotype is determined by a simple Mendelian gene, with the two homozygotes having two contrasting phenotypes, assortative mating results in increased homozygosity at that gene, relative to HW expectations. Much rarer than assortative mating is its opposite, disassortative mating (or negative assortative mating), the preferential mating among dissimilar genotypes. Disassortative mating results in an excess of heterozygotes at the corresponding locus; when there are overall differences in allele frequencies between the phenotypes (e.g., as a result of genetic drift), the excess of heterozygotes may also be visible at unlinked loci. There is also non-random mating when dispersal is limited so individuals that are geographically close are more likely to mate with each other even though these are also more likely to be genetically related. In plants this happens as pollinators tend to subsequently visit flowers that are in close proximity. The most extreme form of inbreeding is self-fertilization, which occurs in many hermaphroditic species. All these forms of inbreeding lead to an excess of homozygotes in a population ($F_{IS} > 0$) that increases over generations until it asymptotically reaches an equilibrium. In the case of complete self-fertilization, there are no longer any heterozygotes left when this equilibrium is reached, and consequently $F_{IS} = 1$. However, one generation of random mating is enough to restore the genotype frequencies to the HW proportions.

Finite Population Size

Finite population sizes lead to genetic drift, which is widely recognized as a major evolutionary force by itself. The presence of genetic drift invalidates the stability of the allele frequencies (postulate 2). On the other hand, genetic drift only affects allele frequencies from one generation to the next and is therefore not visible when looking at a single generation. As a result, genetic drift does not affect the expected genotype frequencies (postulate 1). However, population size does have an actual effect on the genotype proportions that is unrelated to drift. Rather, population size affects HWE because of the non-random mating occurring in finite populations. Truly random mating is unlikely in small populations, especially in species with two separate sexes or when self-fertilization is not possible. Such non-random mating leads to an excess of heterozygotes and therefore to negative values of F_{IS}. The expected effect on F_{IS} is proportional to the inverse of the effective population size (Balloux, 2004; Waples, 2015):

$$E(F_{IS}) = -1/(2N_e + 1)$$

For most values of the effective population size, this effect is negligibly small. Even when N_e equals 100 individuals—small enough to make genetic drift very strong—the expected value of F_{IS} is only about -0.005; this means that in practice the effect of population size will be virtually undetectable in tests of HWE.

Mutation

Mutation induces a change in allele frequencies and thus presents a violation of postulate 2. However, since mutation rates are typically very low, the effect of mutation itself is very slight; mutation only leads to appreciable allele frequency changes when

paired with genetic drift or selection. When testing for deviation of genotype data to HW expectations (postulate 1), mutation actually has no effect. This is because mutation most importantly takes place during gametogenesis and therefore does not affect the genotype frequencies after random union of the produced gametes. So even when a sample would contain one or multiple mutants, the newly introduced alleles still result in genotype frequencies that are in HWE.

Selection

Selection is the most potent evolutionary force, and as such can affect both the allele frequencies and the genotype frequencies, so both postulates 1 and 2. However, not all types of selection are equal. For example, balancing selection does not change the allele frequencies (postulate 2), and even makes them impervious to genetic drift. Similarly, even strong selection does not necessarily lead to genotype frequencies that deviate from HW expectations (postulate 1). This is the case when there are differences in fertility among genotypes. Assume that genotype *aa* produces only half the number of gametes compared to genotypes *AA* and *Aa*, resulting in a drastic reduction in the frequency of allele *a* from one generation to the next. Nevertheless, when the produced gametes all unite at random, the resulting genotype frequencies conform perfectly to HW proportions, despite the strong selection. The effects of selection will also go unnoticed when sampling takes place before selection can act. For example, when selection at a locus for mimicry affects butterfly survival, the genotypes of the caterpillars will conform to HW expectations (provided other relevant assumptions are met). Finally, there are some very specific combinations of values for the genotype fitnesses that, for mathematical reasons, will not lead to deviation from HW expectations even though selection is strong (Lewontin and Cockerham, 1959). This is the case when $W_1{*}W_3 = W_2^2$, where W_1, W_2, and W_3 are the fitness of genotypes *AA*, *Aa*, and *aa* respectively. So even if we have independent proof that a locus is under selection, this does not necessarily mean that the locus is not in HWE.

Multiple Populations

The HW principle concerns only a single population, so any case where there are multiple interconnected populations may lead to a violation of both postulate 1 and 2. Unknowingly sampling individuals from more than one population may occur because of several reasons. For example, individuals may be philopatric with respect to breeding grounds, but still share feeding grounds. Taking a sample from the feeding grounds then yields a mixture of multiple genetic clusters. In other cases, it may simply not be clear how a species is distributed into populations as the main dispersal barriers may not be discernible. In such situations, it may be prudent to perform a clustering analysis to detect any possible population structure, before drawing any conclusions about HW conformity (De Meeûs, 2017). Taking samples from multiple populations will lead to an excess of homozygotes relative to HW expectation ($H_E > H_O$, and consequently $F_{IS} > 0$). The strength of this so-called "Wahlund effect" depends on the proportions at which the populations are mixed and on the differences in the allele frequencies between the populations (population differentiation). The most extreme case for two populations is when they are each fixed for a different allele at a locus. In this case, there are no heterozygotes at all though the mixed sample is expected to harbor 50% heterozygotes when the populations are mixed evenly.

Migration

Migration can be seen as a special case of sampling from multiple populations, though it merits discussion on its own. Migration into a population changes its allele frequencies and therefore affects postulate 2. When migrants are included in a sample when testing for deviation from HW genotype proportions (postulate 1), this causes a Wahlund effect: an excess of heterozygotes. Nevertheless, in most cases we can expect that migration has only a negligible effect on HWE tests. This is because migration only affects HWE testing when there is a substantial proportion of migrants in the sample, that is, when the migration rate is high. However, such high migration rates very quickly even out the allele frequencies across populations. Conversely, in order to get substantial genetic differentiation among populations the migration rate between them should be so low that few, if any, migrants will be included in a typically-sized population sample. So migration is only expected to affect tests of HWE deviation when there is a high rate of migration between previously separated populations, for example, in cases of human induced secondary contact.

Overlapping Generations

In many long-lived species, mating is non-random with respect to age, if only because juveniles do not mate. When there are differences in allele frequencies among age classes, this will lead to a Wahlund effect and thus give rise to a lack of heterozygotes compared to HW expectations (postulate 1). Note that the development of allele frequency differences between age classes can only occur when at least one other assumption is violated, such as when there is genetic drift or selection. Therefore, overlapping generations in an otherwise ideal population does not lead to invalidation of either postulate 1 or 2.

Non-Diploid Inheritance

Both Hardy and Weinberg phrased the expected genotype frequencies in terms of diploid genotypes, and so did I in the introduction. This is no wonder since most species have a diploid genome. However, not all genomes are diploid: many species,

especially plants, are polyploid; other species, including some major insect groups, are haplo-diploid. Even predominantly diploid genomes may show variation in ploidy: sex chromosomes have a different ploidy in males than in females. The non-diploid inheritance does not mean a violation of the Hardy Weinberg principle per se, but only with how it was originally defined; ploidy level does not affect the stability of the allele frequencies (postulate 1) and the equations for the expected genotype frequencies (postulate 2) can be easily calculated for any ploidy level (Meirmans *et al.*, 2018). For example, for a hypothetical triploid the four possible genotypes at a biallelic locus are *AAA*, *AAa*, *Aaa*, and *aaa*; these have expected frequencies of p^3, $3p^2q$, $3pq^2$, and q^3, respectively. Similar adjustments are possible for sex-chromosomes and haplo-diploids. When the ploidy level or mode of inheritance is known, non-diploid inheritance should not pose a problem for tests of HWE deviation. However, there are possible complications. Sex chromosomes may contain pseudo-autosomal regions, where the degree of deviation from HWE may differ between loci depending on their degree of linkage with the sex-determining gene. In polyploid species, there may be variation in ploidy among or even within populations, and the exact ploidy level may not be known for every individual. Furthermore, in polyploids there may be a process called double reduction, which leads to a higher frequency of homozygote gametes than expected. This will lead to an increase in homozygosity compared to HW expectations. The rate of double reduction often differs among loci, depending on their location on the chromosome. Another important point that should be mentioned for polyploids is that the equilibrium genotype frequencies are not restored after a single generation of random mating. Instead, equilibrium is reached asymptotically over multiple generations, at a speed that depends on the ploidy level.

Testing

Testing whether genotype frequencies for a given sample of individuals match the expectations under HWE is a standard procedure in genetic data analysis. At first glance it seems like a problem out of a first-year statistics course: observations neatly fall into a limited number of classes (the genotypes) and the expected value for each class is easily calculated. So it basically comes down to comparing the observed and expected genotype frequencies. Despite this apparent simplicity, a large number of approaches has been developed for testing HWE. These approaches fall into four major classes: goodness-of-fit tests, exact tests, permutation tests, and Bayesian inferences. Here, I will first discuss the main aspects of these approaches before generally discussing issues of power, multiple testing and interpretation.

Goodness-of-Fit Tests

The simplest approach for HWE-testing is the chi-square test, which, despite its major shortcomings, remains widely used. In this test, a test statistic is calculated by for each genotype comparing the expected (n_e) and observed (n_o) number of individuals in the sample as: $(n_o - n_e)^2/n_e$. The overall value of the test statistic is then obtained by summing over all possible (both observed and unobserved) genotypes at a locus. This test statistic asymptotically has a chi-square distribution, which gives the test its name (it is also, more formally, known as "Pearson's goodness-of-fit test"). The *P*-value can thus be obtained by comparing the value of the test statistic to the chi-square distribution for the appropriate number of degrees of freedom. This number can be calculated by taking the number of genotypes minus one, and then subtracting the number of parameters estimated from the sampled data that are used to calculate the expected frequencies. This latter number equals the number of alleles at a locus minus one. So in practice the degrees of freedom can be obtained by taking the number of possible genotypes minus the number of alleles. For a locus with two alleles, there are three genotypes and therefore there is a single degree of freedom.

The main problem of the chi-square test is that for small sample sizes, the actual distribution of the test statistic is discrete; for small samples there are only a limited number of possible genotype combinations and thus a few possible values of the test statistic. This discrete nature of the test statistic does not match the continuous chi-square distribution. This non-continuity problem is especially prominent when one or more of the genotypes has a very low number of expected occurrences (<5), as it causes the value of the test statistic to become inflated. In genetic studies, this is quickly the case as alleles with a low frequency are common and sample sizes are usually small to moderate. For example, when the frequency of the minor allele is 10% and 50 individuals were sampled, the expected number of individuals with the corresponding homozygote genotype is only 0.5. The inflation of the test statistic is especially troublesome for multi-allelic loci such as microsatellites as these usually have several alleles that occur in only a few individuals. This problem can only partly be solved by lumping rare alleles together and/or by applying a continuity correction (Levene, 1949). The log-likelihood ratio test, another goodness-of-fit test with a test statistic that is asymptotically distributed as chi-square, suffers from this same problem. As a result, both these tests are best avoided and one of the other approaches described below should be used instead.

Exact Tests

An exact test is based on calculating the exact probability of obtaining the observed deviation from HWE proportions or a more extreme deviation. An exact test therefore requires a complete enumeration of all possible genotype arrays that have a more extreme deviation than the observed array. This enumeration is straightforward for loci with only two alleles, but with multiple alleles it quickly gets complex, even on modern multi-core computers. So for multi-allelic loci an approximation is needed, which can be done by using either random permutation of the data or by using a Markov Chain (Guo and Thompson, 1992). Either way, when such an

approximation is used, the resulting *P*-value is only an estimate of the exact *P*-value, so the test loses some of its "exactness." However, this should not be a problem when enough permutations or Markov Chain steps are used. Exact tests do not suffer from the non-continuity problems of the goodness-of-fit tests and are therefore preferred for testing for HW deviation for single loci.

Permutation of F_{IS}

As explained above, the F_{IS} statistic can be used to quantify the degree of deviation from HWE. It is therefore intuitive to use F_{IS} as a test statistic in a HWE test. This can be done using standard Monte Carlo permutations. First, the value of F_{IS} is calculated based on the observed genotype frequencies in the sample. Then a permuted dataset is generated by randomly combining alleles into genotypes, and then F_{IS} is calculated for the permuted dataset. When a large enough number of such permuted F_{IS} values are generated, these form a null distribution against which the observed F_{IS} can be compared. When the observed value falls within one of the 2.5% tails of the generated null distribution (or either the left or right 5% tail, in case of one-tailed testing), the deviation from HWE is deemed significant.

One problem arises when using the permutation test for loci with four alleles or more: in such cases it is possible to have genotype arrays that deviate from HWE and still have a low F_{IS}-value. For example, take a locus with four alleles *A*, *B*, *C*, and *D*, and a population sample that consists of 5 individuals each of genotypes *AA*, *BB*, *CC*, and *DD*, and 15 individuals each of genotypes *AC*, *AD*, *BC*, and *BD*. This sample is obviously not in HWE as genotypes *AB* and *CD* are completely lacking. Nevertheless the observed frequency of heterozygotes is exactly equal to the expected frequency of 0.75, so F_{IS} equals zero. Because of this problem permutation tests using F_{IS} are not recommended for single loci with multiple alleles. Their most important use, however, is for analyses involving multiple loci. A multilocus value of F_{IS} can easily be calculated by first averaging the H_O and H_E across loci, and then using these averages for calculating F_{IS}. The above problem is already unlikely to occur at a single locus, so it will almost certainly not happen across multiple loci simultaneously. Therefore, the multilocus permutation test is a suitable approach to get an single test of whether populations are in HWE. Averaging over populations is also possible to obtain a species wide estimate of the degree of deviation from HWE and a species-wide *P*-value.

Bayesian Inferences

Various Bayesian approaches to assessing deviation from HWE expectation have been proposed, but none of these has become very popular. The reason for this is probably that the HWE problem itself is relatively simple and the above frequentist methods are appropriate and straightforward. The most popular Bayesian method is the one proposed by Shoemaker *et al.* (1998), who approach the problem as one of parameter estimation. They developed a summary statistic for the degree of deviation from HWE, related to, but not analogous to, F_{IS}. They then use the Bayesian framework for obtaining the posterior distribution for this parameter, which can be studied to see whether there is any meaningful deviation from HWE. In this and several other developed Bayesian approaches, the choice of the prior distribution has a large effect on the results. Therefore great care has to be placed in the choice of prior.

Power

In statistical significance testing (here: the goodness-of-fit test, exact test and permutation test), there are two main types of errors: rejecting the null hypothesis when it is in fact true (type I error), and not rejecting the null hypothesis when it is in fact false (type II error). The rate at which the type I error occurs (α) can be controlled by setting the significance level in advance; usually this is set at $\alpha = 0.05$ for a single test. The rate at which the type II error occurs (β), however, cannot be set by the researcher, but is dependent on the selected significance level, the sample size, and the properties of the used test. The power of the test—the rate at which the test is expected to correctly reject a false null hypothesis—is then defined as $1 - \beta$. The concepts of type I error, type II error and power is only applicable to the frequentist analyses (goodness-of-fit, exact test, and permutation test), because Bayesian inferences do not rely on *P*-values.

Tests of Hardy–Weinberg equilibrium generally have a low statistical power, unless the deviation is strong or the sample size is large (Wittke-Thompson *et al.*, 2005; Graffelman, 2010). In evolutionary or ecological genetic studies, one usually only samples a limited number of individuals per population; a sample size of 50 individuals is already slightly above the standard. Small deviations from HWE, on the order of $F_{IS} = 0.25$, are difficult to detect using such small sample sizes. Even smaller deviations from HWE, on the order of $F_{IS} = 0.1$, are nearly undetectable, even when sample sizes are an order of magnitude larger. The power of HWE tests also depends on the minor allele frequency; when the minor allele frequency is low, the power also gets reduced. In general, the power of the goodness-of-fit test is much lower than that of the other two methods, especially at small sample sizes. This is in addition to the non-continuity problem discussed above, which leads to an inflation of the type I error (Wittke-Thompson *et al.*, 2005).

Multiple Testing

When analyzing data from multiple loci and/or populations, many separate tests for HWE will be performed. This multiple testing means that the probability increases that across all those tests there will be at least one locus for which the null hypothesis is falsely

rejected. In other words: the experiment-wide type I error rate increases with the number of tests. Many different methods have been developed to correct for this (e.g., Bonferroni, sequential Bonferroni (Rice, 1989), False Discovery Rate (Benjamini and Hochberg, 1995)) and it is generally advised that such a correction is applied when testing for HWE.

One major disadvantage of correction for multiple testing is that it greatly decreases the power of the test. In a standard Bonferroni correction—which is the most conservative of the available methods—the significance level is divided by the number of tests. So with 50 loci and two populations, and thus 100 tests, the corrected alpha becomes 0.0005. For $F_{IS} = 0.4$, $p = 0.5$ and a sample of 50 individuals per population, this means that the power drops from 78% to only 22%. Whether it is worth trading off type I error for such a high type II error depends on the type of question being asked and the reason why HWE tests are being performed. When HWE tests are used to weed out genotyping errors to prevent errors in later analyses, it would be unwise to have such a low power. With modern sequencing methods, it is easy to obtain tens of thousands of loci. In such cases it may be more prudent to sacrifice some loci to type I error rather than to leave in a lot of truly erratic loci.

Applications in Ecology and Evolution

As should be clear from the above, there are many ways to test for HWE deviation and those methods are also readily available in easy-to-use software packages (see the list of websites for some examples). In contrast with the ease of calculation, the interpretation of the results can be difficult (Waples, 2015). This is because the HW principle depends on so many assumptions that it is generally difficult to tell anything about the underlying cause of a significant deviation for a particular locus. It is also important to keep in mind that the absence of a significant deviation does not necessarily mean that all assumptions are met. Nevertheless, testing for deviation from HWE is an important part of any population genetic study. Its use falls into two main categories: checking for aberrant loci and testing evolutionary and ecological hypotheses.

Removing Genotyping Errors and Aberrant Loci

Despite the simple discrete nature of genotypic data, there are a lot of things that can go wrong in genetic studies: there may be null alleles, strand slippage, preferential amplification of alleles, stutter bands, dosage compensation, low coverage, genotype calling errors, etc. So, notwithstanding the many assumptions of the HW principle, genotyping errors are probably the most common cause of HWE deviation. Accordingly, this means that testing for HWE is a simple way of checking for genotyping errors. Besides genotyping errors, there are other locus-specific reasons why a certain locus may not conform to HW expectations and has to be excluded from further analysis. The locus may for example be under selection, be present in a duplicated gene, or it may be located on a sex chromosome, b-chromosome or on a cytoplasmic genome.

There are several important issues to keep in mind when using HWE tests for this purpose. First, the approach only works when the sampled population is indeed in HWE. A slight deviation from HWE might be difficult to detect, but may still lead to too many loci being discarded. Plus, after discarding these loci, the population may give a false overall impression of conforming to HWE. A second issue is that several bioinformatics pipelines for allele calling already make the assumption of HWE conformation. This is especially the case for pipelines that were developed for human genetics. This assumption makes subsequent testing for HWE on the resulting genotypes both redundant and biased; these pipelines are therefore unsuitable for species or populations where deviation from HWE can be expected. A third issue has already been discussed above: the low power of HWE tests combined with correction for multiple testing makes these tests highly inefficient.

Testing Evolutionary and Ecological Hypotheses

HW theory is not only suitable for weeding out erratic loci, but can also be put to much more interesting use, namely to test evolutionary and ecological hypotheses. Most of this use revolves around the analysis of non-random mating and inbreeding, and in this light it makes sense that the F_{IS} statistic is also called the inbreeding coefficient. When viewed across species, it becomes clear that species that are known to have strong inbreeding (e.g., self-pollinating plants) have much higher values of F_{IS} than species that are obligatory outcrossers (e.g., plants with a self-incompatibility system). For a species where the degree of inbreeding is not known, F_{IS} can be used to get a first insight into the mating system.

The use of F_{IS} is not limited to comparisons among species; also within species there may be differences in the degree of deviation from HWE. There are several ecological and demographic scenarios that can cause such differences. In animal-pollinated plants, the rate of outcrossing depends on the availability and efficiency of its pollinators. Therefore the inbreeding coefficient will be higher where insect diversity and activity is lower, such as at higher latitudes and altitudes and in less-suitable habitats. Small populations of a plant species may attract less pollinators and thus have increased inbreeding, causing a correlation of F_{IS} with population size. In animals, populations may differ in sex ratio, and more skewed sex ratios may lead to a higher deviation from HWE. Inbred individuals often have lower fitness than non-inbred ones, causing a correlation between individual heterozygosity and fitness. In long-lived species, the individuals with the highest heterozygosity may thus have the

longest lifespan. A comparison of F_{IS} across different age classes may then reveal that F_{IS} values decrease with age, being negative for the older age classes.

Generally, the use of HWE for testing evolutionary or ecological hypotheses can be done by calculating F_{IS} for every population or group and then correlate these F_{IS} values to a parameter of interest, such as population size, latitude, or sex ratio. In such cases, it is important that multi-locus F_{IS}-values are used for estimation and testing. Multi-locus values are relatively robust to locus-specific effects such as selection and genotyping errors. Nevertheless, it remains important to also calculate single-locus F_{IS}-values to check for consistency among loci. This is because the effects of non-random mating are expected to be similar across loci. Another advantage is that performing a single test that combines information across all loci gives much greater power compared to doing multiple single-locus tests. However, also here it is important to check that the results are not influenced by one or a few outlier loci.

Perspective

Overall, the HW principle presents a very complete theory with a well-described statistical toolkit. This toolkit can both be used for error-checking and for hypothesis testing. However, one disadvantage of the current approaches is that they cannot do both at the same time. HWE tests for error-checking are only possible when a population is truly in HWE. When using HWE for testing ecological hypotheses, they cannot simultaneously be used for error-checking. Therefore, there is room for an additional type of analysis that combines the two uses and that distinguishes locus-specific effects (that are due to genotyping errors) from population-specific effects (that are due to differences in mating patterns). Bayesian analysis would be a perfect framework for such a type of combined inference.

See also: Behavioral Ecology: Dispersal–Migration; Mating Systems. Conservation Ecology: Conservation Genetics. Ecological Data Analysis and Modelling: Metapopulation Models. Evolutionary Ecology: Isolation; Genetic Drift; Natural Selection. General Ecology: Migration and Movement

Reference

Balloux, F., 2004. Heterozygote excess in small populations and the heterozygote-excess effective population size. Evolution 58, 1891–1900.
Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate – A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B – Methodological 57, 289–300.
Crow, J.F., 1988. 80 years ago – The beginnings of population-genetics. Genetics 119, 473–476.
Crow, J.F., 1999. Hardy, Weinberg and language impediments. Genetics 152, 821–825.
de Meeûs, T., 2017. Revisiting F_{IS}, F_{ST}, Wahlund effects, and null alleles. Journal of Heredity 2017. doi:10.1093/jhered/esv019.
Edwards, A.W.F., 2008. G.H. Hardy (1908) and Hardy–Weinberg equilibrium. Genetics 179, 1143–1150.
Fisher, R., 1918. The correlation between relatives on the supposition of Mendelian inheritance. Philosophical Transactions of the Royal Society of Edinburgh 52, 399–433.
Graffelman, J., 2010. The number of markers in the Hapmap project: Some notes on Chi-square and exact tests for Hardy-Weinberg equilibrium. The American Journal of Human Genetics 86, 813–818.
Guo, S.W., Thompson, E.A., 1992. Performing the exact test of Hardy–Weinberg proportion for multiple alleles. Biometrics 48, 361–372.
Hardy, G.H., 1908. Mendelian proportions in a mixed population. Science 28, 49–50.
Levene, H., 1949. On a matching problem arising in genetics. Annals of Mathematical Statistics 20, 91–94.
Lewontin, R.C., Cockerham, C.C., 1959. The Goodness-of-Fit test for detecting natural selection in random mating populations. Evolution 13, 561–564.
Meirmans, P.G., Liu, S., Van Tienderen, P.H., 2018. The analysis of polyploid genetic data. Journal of Heredity. doi:10.1111/1755-0998.12496.
Pearson, K., 1904. Mathematical contributions to the theory of evolution. XII. On a generalised theory of alternative inheritance, with special reference to Mendel's laws. Philosophical Transactions of the Royal Society of London Series A 203, 53–86.
Rice, W., 1989. Analyzing tables of statistical tests. Evolution 43, 223–225.
Shoemaker, J., Painter, I., Weir, B.S., 1998. A Bayesian characterization of Hardy–Weinberg disequilibrium. Genetics 149, 2079–2088.
Stern, C., 1943. The Hardy–Weinberg law. Science 97, 137–138.
Waples, R.S., 2015. Testing for Hardy–Weinberg proportions: Have we lost the plot? Journal of Heredity 106, 1–19.
Weinberg, W., 1908. Über den Nachweis der Vererbung beim Menschen. Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg 64, 368–382.
Wittke-Thompson, J.K., Pluzhnikov, A., Cox, N.J., 2005. Rational inferences about departures from Hardy-Weinberg equilibrium. The American Journal of Human Genetics 76, 967–986.
Wright, S., 1922. Coefficients of inbreeding and relationship. The American Naturalist 56, 330–338.
Yule, G.U., 1902. Mendel's laws and their probable relations to intra-racial heredity. New Phytologist 1 (9), 222–238.

Further Reading

Hedrick, P.W., 2011. Genetics of populations. Sudbury, MA: Jones & Bartlett Publishers.
Masel, J., 2012. Rethinking Hardy–Weinberg and genetic drift in undergraduate biology. BioEssays 34, 701–710.
Rousset, F., Raymond, M., 1995. Testing heterozygote excess and deficiency. Genetics 140, 1413–1419.
Salanti, G., Amountza, G., Ntzani, E.E., Ioannidis, J.P.A., 2005. Hardy–Weinberg equilibrium in genetic association studies: An empirical evaluation of reporting, deviations, and power. European Journal of Human Genetics 13, 840–848.
Wakefield, J., 2009. Bayesian methods for examining Hardy–Weinberg equilibrium. Biometrics 66, 257–265.

Relevant Websites

Arlequin, n.d., http://cmpg.unibe.ch/software/arlequin35/—Arlequin, program for HWE testing.
GenePop, n.d., http://genepop.curtin.edu.au—GenePop, program for HWE testing.
GenoDive, n.d., http://www.patrickmeirmans.com/software/GenoDive.html—GenoDive, program for HWE testing.
R-package, n.d., https://cran.r-project.org/web/packages/HardyWeinberg/HardyWeinberg.pdf—R-package "HardyWeinberg".
Virtual Biology Lab, n.d., http://virtualbiologylab.org/PopGenFishbowl.htm—Virtual Biology Lab HWE simulation.

Isolation

JP Wares and TM Bell, University of Georgia, Athens, GA, USA

An important consideration in the design of ecological reserves and the potential for local adaptation within a species is the occurrence and degree of isolation between populations. Isolation can occur by many mechanisms. Individuals can be isolated 'physically' due to geographic or environmental barriers that limit effective dispersal between regions; 'numerically' (i.e., Allee effects in areas of low population size or reproductive success); 'reproductively' (depending on both recognition factors and genetic factors); or 'ecologically'. Isolation should primarily be considered in terms of demography – that is, individuals from different populations may come into contact with one another, but if that contact does not result in successful reproduction then it may indicate a form of isolation.

Isolation is a quantitative and potentially dynamic trait of populations. While the two extremes of this trait might be complete isolation (no interpopulation mating, gene flow, or interaction) or complete mixing of individuals and/or genomes from each population (panmixia), the greater challenge is to understand the mechanisms underlying intermediate patterns of isolation. Situations where isolation has not been completely achieved are a topic of great interest and in some cases are studied as examples of incipient speciation events.

Measuring Isolation

For example, two populations (e.g., distinguished by geographic separation) might be characterized as having only sporadic events of migration in which an individual successfully becomes established and reproduces in a non-natal population. These events may be measured directly, meaning that propagules must be followed for at least one generation; or indirectly, as with stable isotope markers or genetic markers. Genetic markers are frequently used to identify the relatedness of individuals from different populations, as the propagules of many species are difficult to track in terms of location and subsequent reproductive success.

Most genetic markers are assumed to be selectively neutral – that is, not under the influence of natural selection. Thus, the diversity and distribution of selectively neutral alleles at any given genetic locus are governed solely by the equilibrium between mutation (which adds novel alleles, or distinct markers), genetic drift (which stochastically eliminates alleles over time), and migration (which may introduce alleles that originally arose in a different population). The assumption of neutrality is important for understanding the degree to which populations are isolated from one another, because the diversity of genetic markers then reflects only the demographic characteristics of the populations, such as migration and reproduction.

For a given population size, the relative contribution of the mutation rate (μ) and the migration rate (m) to an inference of isolation will be important. If $\mu \gg m$, new alleles will arise in a population faster than they arrive from other populations, and the diversity in that population will become relatively distinct from that found in other regions. If $m \gg \mu$, then migration acts as a homogenizing force among populations and they will be composed of a similar set of allelic diversity. As an example, the populations shown in **Fig. 1** represent a continuum of isolation. If the individuals are distinguishable based on genetic markers, we can measure isolation by quantifying the amount of diversity within each population (d_w) relative to among-population (d_a) diversity. A class of statistics used to measure isolation in this case (Wright's F statistics) use a ratio of

$$(d_a - d_w)/d_a$$

to generate a value from 0 to 1, with 1 representing complete isolation. This class of statistics can be used to compare diversity among any hierarchical set of populations. Essentially, if there is equivalent diversity within a given group as among groups, there is little evidence for isolation. The top panel in **Fig. 1** illustrates 'complete isolation'; the diversity among populations (d_a) could be represented as 50% based on the frequency of two alleles across populations, while d_w would be 0% (there is no variation within either populations). Thus, the isolation for this set of populations is 1. Using similar rough methods, the isolation of the remaining illustrated populations is $(0.50 - (1/7))/0.50 = 0.714$ (significant isolation), 0.143 (some isolation), and 0 (no isolation). Similar statistics can be calculated for other forms of isolation (e.g., for measures of sexual isolation, the frequency of homospecific to heterospecific mating attempts, or successes compared to the total number of mating opportunities).

Effects of Isolation

Migration itself does not guarantee gene flow. If individuals are locally adapted to their natal environment, the reproductive fitness of those individuals in a different population residing in a new environment may be quite low, leading to an effective isolation of the populations. Selection may act differentially on parts of the genome that are most directly affected by environmental differences. For example, if an individual migrates into a population and successfully reproduces with a native individual, sexual recombination of gametes will generate offspring with combinations of 'native' and immigrant gene copies. Considering two physically unlinked genes A and B, where different copies at the A gene are selectively neutral but alleles at the B gene are not,

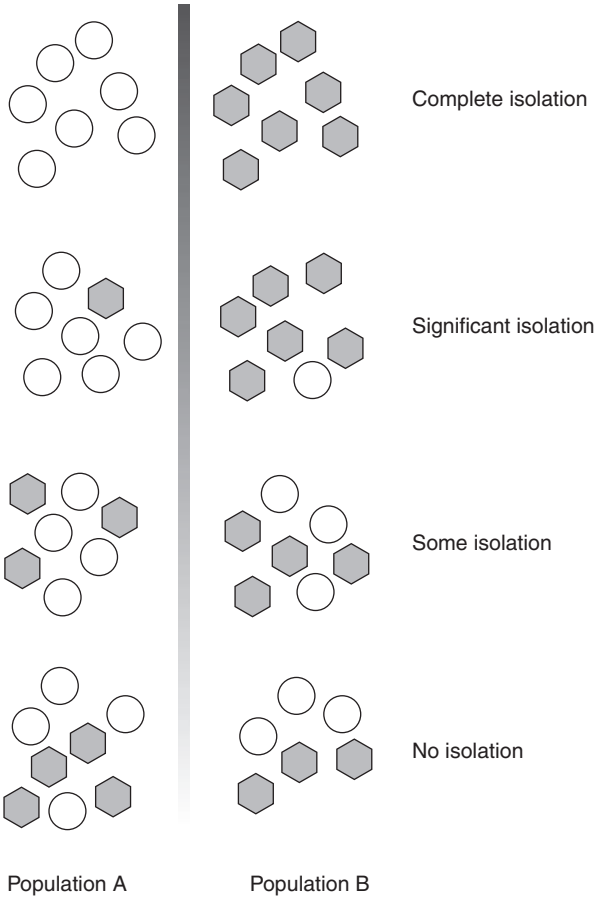


Fig. 1 Isolation is measured as the tendency toward different identity. In the case of sexual isolation, the ratio of homospecific to heterospecific matings is measured, considering all available interactions. In the case of using genetic markers to identify demographic isolation, the level of identity by state (e.g., sharing the same allele) is considered within a population relative to among populations. Ecological factors such as phenology differences among populations may contribute to both types of isolation. Here, two populations are shown. They may be defined based on their geographic location or any other distinguishing traits, and symbols within each population represent distinguishable traits of individuals, such as genetic markers, otoliths, or stable isotopic signatures.

many offspring may survive in the population that carry the immigrant allele at the A gene, but few or none may survive if they carry the immigrant allele at the B gene. Here, isolation is again quantitative in that the isolation may only be for particular elements of an organism's genome, rather than isolating two populations or species entirely.

The net result of these varying types of isolation is that to a certain extent, indirect techniques such as measuring allelic diversity at a variety of genetic loci in multiple populations can be used to characterize the degree to which populations fit the equilibrium neutral model governed by mutation, drift, and migration. There are a number of ways in which this model may be violated, particularly in cases where one or more populations being compared is or has recently expanded from a founder population (e.g., range expansions or species introductions); nevertheless, the comparison of isolation measures across many markers may be indicative of both demographic and selective forces that promote the evolutionary divergence of populations in isolation.

Isolation is not a static characteristic of populations. The Isthmus of Panama is an excellent case of complete contemporary isolation of marine populations on either side; there is no gene flow, no migration, and no interaction between individuals from the tropical eastern Pacific and the Caribbean. However, as the Isthmus formed over the course of a million years or more, populations that were initially freely mixing were slowly but increasingly isolated by the uplift of land masses and reduction of currents by freshwater inflows and mangrove swamps, and the final reduction in the number of pathways by which individuals could travel from one side to the other. This more realistic scenario of isolation allows for migration to persist beyond the initiation of the isolating event. Known as the isolation-migration model, this statistical approach allows more complete description of the cause and timing of isolation between populations. Without considering both factors, it is difficult to distinguish whether two populations share many alleles due to high migration, or recent (complete) isolation, or a mixture of intermediate migration and intermediate levels of isolation.

Complex scenarios may arise due to the interaction between historical and contemporary causes of isolation. While the variance in allele frequencies from one population to another is a standard method of measuring isolation it is not always clear whether that isolation is truly due to an equilibrium of mutation, drift, and migration – the isolation by distance model. It is also possible that ancestral events could separate a single population into two or more disjunct populations; subsequent environmental change, again permitting migration between the two areas, would cause a pattern of secondary gene flow. Geographic areas where two genetically or morphologically distinct groups interact, with incomplete isolation, are called clines. Many well-studied clines are caused by the secondary interaction between historically isolated populations.

One case study that illustrates both isolation by distance and the interaction of historically isolated lineages involves a 'ring species' of warblers in Asia. The geographic range of *Phylloscopus trochiloides* wraps around the Tibetan Plateau, and limited dispersal from the nest site results in populations from western Siberia through Tibet that exhibit a strong correlation of the geographic distance between sampled sites and the genetic distance measured at a mitochondrial gene. With greater geographic distance around the Tibetan Plateau, genetic distance gradually increases. However, there is a zone of overlap in central Siberia in which populations at one end of the 'ring' encounter populations from the other end – but in this region, the birds differ significantly in terms of plumage and song. This case is made more interesting by the isolation of eastern Siberian populations from those further south on the Tibetan Plateau due to deforestation – eventually, without any means of dispersal between the two sites, the east Siberia populations of *P. trochiloides* could become completely distinct (ecologically and demographically isolated).

Isolation and its subsequent effect on the evolutionary trajectory of a species has been the central theme of speciation literature over the past century. Only recently has this topic begun to be regularly incorporated into the study and theory of ecology and demography. Much of the diversity in form and function observed in our present-day environment can be thought to be the result of past isolation events and should therefore be considered a topic of much importance in fields of study other than evolutionary biology. Isolation events and their underlying mechanisms can have wide-ranging effects on a species. The effects of isolation may be detectable in genetic, ecological, and demographic patterns. Therefore, these topics should be of increasing relevance to current research focused on population ecology, local adaptive processes, and conservation biology.

See also: Conservation Ecology: Endangered Species; Conservation Genetics; Connectivity and Ecological Networks; Ecological Risk Assessment. Ecological Complexity: Ecological Indicators: Connectance and Connectivity. Ecological Data Analysis and Modelling: Modeling Dispersal Processes for Ecological Systems. Human Ecology and Sustainability: Environmental Protection and Ecology

Further Reading

Coyne, J.A., Orr, H.A., 2004. Speciation. Sunderland, MA: Sinauer.
Irwin, D.E., Bensch, S., Price, T.D., 2001. Speciation in a ring. Nature 409, 333–337.
Pérez-Figueroa, A., Caballero, E., Rolán-Alvarez, E., 2005. Comparing the estimation properties of different statistics for measuring sexual isolation from mating frequencies. Biological Journal of the Linnean Society 85 (3), 307–318.
Slatkin, M., 1985. Gene flow in natural populations. Annual Review of Ecology and Systematics 16, 393–430.

Life-History Patterns

SH Alonzo and HK Kindsvater, Yale University, New Haven, CT, USA

Introduction

A life history can be defined as the timing of and relative investment of energy in survival, growth, and reproduction (Fig. 1). However, such a simple definition belies the complexity of life-history patterns we observe in nature. For example, trees may survive for hundreds of years while many insects live for only days. Marine mammals may have only a few offspring in a lifetime while some marine fish can produce hundreds of thousands of eggs every time they reproduce. This is the diversity of patterns that the study of life histories attempts to explain.

Ecological and evolutionary processes interact to produce variation in life history. For example, predator–prey interactions depend on the population dynamics of both predators and prey. Yet predation, as a source of mortality, influences the evolution of traits affecting body size and reproduction in prey. Prey reproduction will in turn affect prey population growth rate and thus the biomass of prey available to predators affecting predator reproduction and population size. When thinking about life-history patterns, it is impossible to disentangle evolution from ecology and ecology from evolution.

The study of life-history patterns has a long history of theoretical and empirical research. We generally understand the selective pressures that determine basic life-history traits such as age at first reproduction, number of reproductive bouts, partitioning of reproductive effort, and longevity. Yet many questions remain unanswered, such as how environmental variation and climate change may affect life histories and how the details of life-history variation affect ecological communities. In addition, new methods (such as novel genetic techniques and modeling approaches) continue to advance our knowledge and raise new questions.

Approaches and Paradigms

Tradeoffs

At some level all organisms experience constraints: genetic, energetic, physiological, developmental, and ecological. As a result, life-history theory is founded on a basic paradigm of tradeoffs. For example, given a set amount of energy, organisms must tradeoff the investment of that limited energy between growth and reproduction. Similarly, genetic correlations can cause changes in one trait (such as growth rate) due to selection on another trait (such as fecundity). Thus, a genetic correlation can limit the evolutionary response in both traits.

Phenotypic Models

The most common approach to modeling life-history patterns is optimality theory. This approach is based on the idea that natural selection will favor individuals or genes that confer the greatest fitness and thus asks which of all possible phenotypes is expected

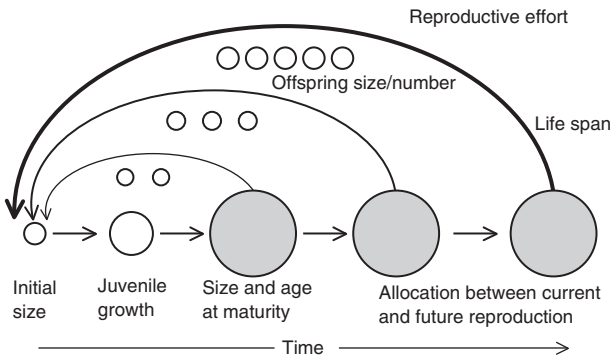


Fig. 1 An organism's life history can be represented by its survival, growth, and reproduction through time. However, a life history is actually a composite of many individual traits (such as initial size) and is influenced by multiple life-history tradeoffs (such as between offspring size and number). This figure presents one possible life-history pattern where each circle is a different age and the size of the circle represents body size. Arrow width represents survival between age classes and the contribution through reproduction to future generations. In this example, the hypothetical organism has an equal probability of survival across all age classes, does not grow following maturity, and total reproductive effort (offspring size and number) increases with age.

 doi:10.1016/B978-0-444-63768-0.00856-8

to confer the greatest fitness. Of course, this approach only makes sense in light of the tradeoffs or constraints discussed above. Otherwise a 'Darwinian Demon' (an organism that matures immediately, produces an infinite number of tiny offspring, and lives forever) would be predicted. Typically, phenotypic models examine a single tradeoff (e.g., between growth and reproduction) and predict the trait or traits expected to maximize fitness for that set of circumstances. The most common approach has been to examine how traits such as age or size at maturity (through their effect on birth rates) influence population growth rate (often referred to as the Malthusian parameter r) and find the life-history pattern that maximizes r. This links life history explicitly to population dynamics.

Another type of phenotypic model that has been influential in evolutionary ecology is game theory. This approach to modeling also predicts the phenotype that will be favored by natural selection but considers the situation where fitness depends on the traits of other individuals in the population. For example, if offspring survival depends on relative rather than absolute size, the size of offspring that maximizes individual fitness will depend on the size of offspring produced by other individuals in the population.

Both phenotypic approaches have been criticized for assuming any phenotype is possible, that organisms will be optimally adapted to their environment, and for ignoring the underlying genetic basis and dynamics of trait evolution. Despite these simplifying assumptions, optimality and game theory have successfully predicted observed patterns.

Genetic Models

In contrast, genetic models focus on how selection and the underlying genetic basis of traits influence life-history evolution. Population genetics examines how the differential fitness of genotypes (resulting from a few alleles and loci) leads to changes in allele frequencies over time. Quantitative genetics assumes that many genes of small effect contribute to the traits under consideration and that phenotypes are the result of an interaction between genetic and environmental factors. One advantage of quantitative genetics for life-history theory is its ability to examine the evolution of multiple correlated traits and to consider how various forms of environmental variation and genetic constraints affect the evolution of traits. In addition, the use of quantitative genetics has a strong empirical component where key parameters can be measured empirically and predictions can be readily compared to observed patterns in the lab and field (though many of the underlying genetic parameters are not known and are difficult to measure in wild populations). Research by Derek Roff and colleagues has illustrated the immense power of using quantitative genetic models and empirical studies in combination to illuminate our understanding of life-history evolution.

Summary of Phenotypic and Genetic Models

While phenotypic and genetic approaches each have strengths and weaknesses, they have also both made important contributions to our understanding of life-history patterns. We have learned the most about the evolution and ecology of life-history patterns by applying multiple methods to the same general questions.

Key Questions and Case Studies

Because of their inherent relationship to both fitness and population ecology, the study of life-history patterns has influenced many areas of research within evolution and ecology including: mating systems, sexual selection, migration, species interactions, and community dynamics. It is impossible to cover such broad influences and applications here. Therefore, we focus on a few classic life-history questions and discuss some new directions of research.

The Timing of Reproduction: Age and Size at Maturity

One of the most basic questions in life-history theory is when an organism should mature and reproduce. In general, size is positively correlated with fecundity and fertility. Because individuals tend to grow more slowly (if at all) when reproductive, there is an advantage to growing as large as possible before maturity. However, delayed reproduction will not be favored by selection if mortality risk is high. Additionally, resource availability also interacts with the strength of selection on growth and maturation. Classic predictions come from finding the age at maturity α that maximizes population growth rate (r) in the Euler–Lotka equation

$$1 = \int_{\alpha}^{\infty} e^{-rx} l_x m_x dx$$

where l_x represents expected survival to age x, α age at first reproduction, and m_x fecundity at age x. This approach makes the assumption that the population is at equilibrium. The life history that maximizes r and satisfies the above equation is predicted to be favored by natural selection. It is possible to examine different patterns of survival and fecundity with age (e.g., different functions for l_x and m_x with respect to α) and determine how they affect the optimal age at maturity. Theory and empirical studies have shown that higher adult mortality favors early age and size at maturity. Similarly, greater size-dependent reproductive advantages favor later maturity (all else being equal).

An interesting and relatively new application of this theory is in fisheries management. New research suggests that fishing mortality can act as a selective force leading to both plastic responses in individual traits and evolutionary change. For example, in

populations of Atlantic cod (*Gadus morhua*), a decline in observed age and size at maturity is associated with a long history of fishing (Fig. 2). Whether this represents an evolved or plastic response is still debated. However, an evolutionary change in size and age at maturity of cod may explain the fact that cod have not recovered despite reduced fishing mortality. Dave Conover and Steve Munch have shown a similar pattern experimentally in artificially selected populations of a small marine fish (*Menidia menidia*).

Current Versus Future Reproduction and the Evolution of Life Span

Another key tradeoff experienced by organisms is the tradeoff between current and future reproduction. For example, if a plant or animal invests most of their energy in their current reproductive event (e.g., produces more gametes or invests more in parental care), they may invest less energy in survival, growth, or future reproduction (Fig. 3). This has led to the question of how many times an organism should reproduce once mature. Organisms that reproduce only once are known as annual (plants) or semelparous (animals). Organisms with multiple reproductive bouts are known as perennials (plants) or iteroparous (animals). The probability of survival at each age or size class, combined with the relationship between offspring success and size or age affects the degree of iteroparity.

Longevity and senescence (increases in age-specific mortality or decreases in reproductive rates at higher ages) are related to the tradeoff between current and future reproduction. Immense variation in life span exists, with maximum age varying across taxa from days to hundreds of years. Some of this variation can be explained by selection on current versus future reproduction and the costs and benefits associated with delayed maturity. For example, if larger and older individuals have much higher reproductive rates and mortality is relatively low, theory predicts that organisms will be selected to delay maturity and invest in future growth and survival resulting in long life span. A more puzzling question, however, is why most organisms exhibit senescence. One body of theory explains senescence as the result of tissues accumulating damage over time. Another explanation has been that selection is much stronger on earlier age classes, and thus traits expressed later in life are less strongly selected. However, the evolution of life

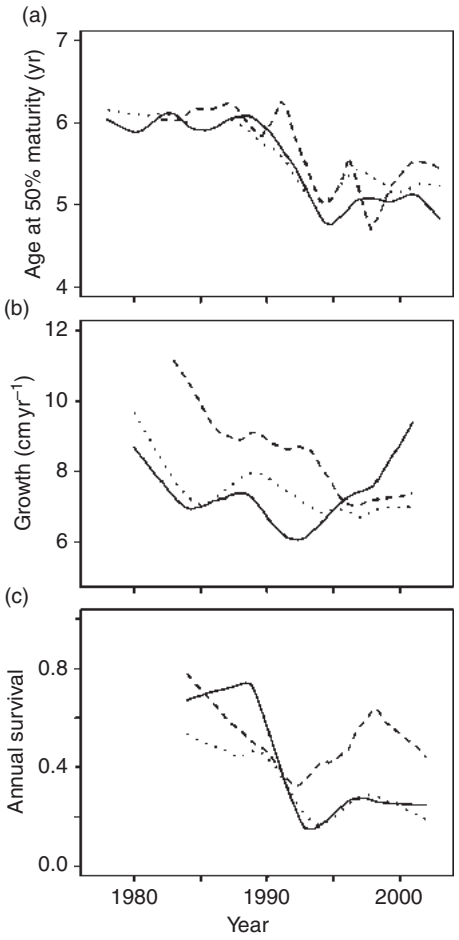


Fig. 2 Age at maturity (a) and annual growth based on length increments (b) have decreased in Atlantic cod over time in populations that have also experienced decreased expected annual adult survival (c) due to fishing. Each line represents a different region or division surveyed by the Northwest Atlantic Fisheries Organization (NAFO). Reprinted by permission from Macmillan Publishers Ltd: *Nature*, (Olsen EB, Heino M, Lilly GR, *et al.* (2004). Maturation trends indicative of rapid evolution preceded the collapse of northern cod. *Nature* 428: 932–935), Copyright (2004).

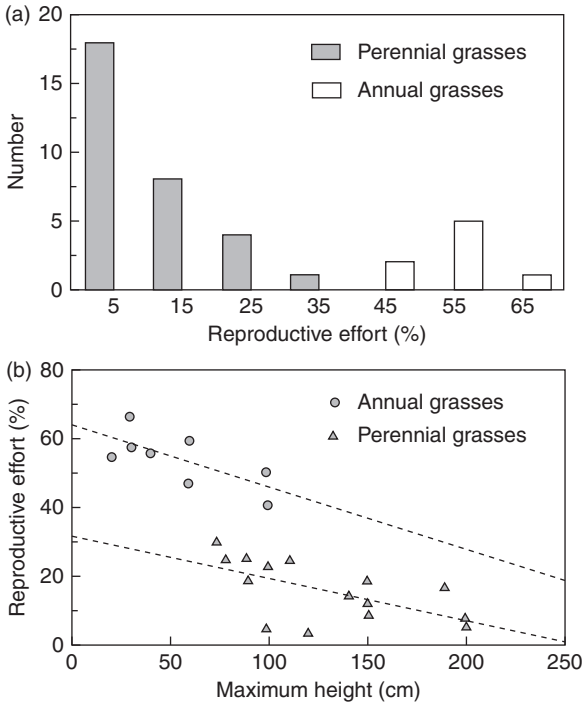


Fig. 3 Reproductive investment in annual versus perennial grasses. In general, life-history theory predicts a tradeoff between current and future reproduction. (a) A survey of 40 different species of British grasses found that reproductive effort in annual species was higher than for perennial grasses. (b) The relationship between plant size and reproductive effort also differed between annual and perennial species; annual species invest a larger portion of their energy in reproduction per year than perennial species. Reproduced from Roff DA (1992). *The Evolution of Life Histories*: *Theory and Analysis*. New York: Chapman and Hall. Redrawn from Wilson AM and Thompson K (1989). A comparative study of reproductive allocation in 40 British grasses. *Functional Ecology* 3: 297–302, with kind permission from Springer Science and Business Media and Blackwell Science.

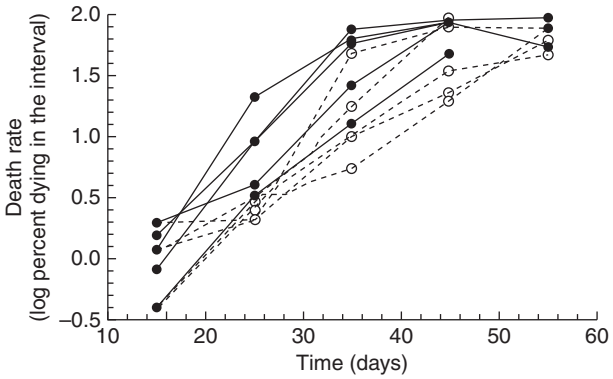


Fig. 4 The evolution of life span and mortality patterns. Linda Partridge and colleagues conducted artificial selection experiments on laboratory populations of *Drosophila melanogaster* (fruitflies). 'Young' lines of flies (filled circles and solid lines) were selected for early age at reproduction while 'old' lines of flies (open circles and dashed lines) were selected for late maturity. After 31 generations of selection, age-specific differences in mortality rates evolved. 'Old' lines exhibit lower mortality at intermediate ages demonstrating the evolution of mortality patterns and expected life span in response to selection. Reproduced from Partridge L, Prowse N, and Pignatelli P (1999). Another set of responses and correlated responses to selection on age at reproduction in *Drosophila melanogaster*. *Proceedings of the Royal Society of London series B* 266: 255–261, Fig. 3b with kind permission from The Royal Society of London.

span and the explanation of senescence itself remains an area of life-history evolution that is not fully understood. Some of the most promising research on life span and senescence has been conducted with artificial selection studies especially in *Drosophila melanogaster*. Linda Partridge and colleagues have shown that mortality patterns (and thus also life span) can evolve in response to differential selection on age at maturity (Fig. 4).

Investment Per Offspring and the Size/Number Tradeoff

A tradeoff between energetic effort per individual offspring and total offspring number is expected because resources are limiting for virtually every organism. Furthermore, an intuitive physiological constraint on maximum gonad or brood size exists for any body size. Larger eggs or offspring may have a higher per capita chance of survival or greater expected fitness later in life. Therefore, selection on offspring number and on offspring size is expected to determine the optimal effort per offspring in a given environment (Fig. 5).

The ornithologist David Lack approached this problem by assuming that as clutch size increases individual offspring survival must decrease, given that parents must feed or provision each offspring. He argued that the optimal clutch size in birds would depend on this tradeoff between offspring number and individual survival. To determine the 'Lack clutch', let N_e represent clutch size (or number of eggs laid), l_f the proportion of eggs surviving, and N_f the number of offspring that survive (or the number fledged). Imagine, however, that offspring survival decreases with clutch size such that $l_f = 1 - cN_e$. Then the expected number of surviving offspring (a proxy for individual fitness) is given by

$$N_f = l_f N_e = (1 - cN_e)N_e = N_e - cN_e^2$$

Using basic calculus, it can be shown that $N_e = 1/2c$ maximizes N_f (Fig. 6). Experiments manipulating clutch size in birds found that Lack's optimal clutch size could predict between species variation in clutch size qualitatively but that actual clutch sizes were often slightly lower than predicted. The primary explanation for this pattern has been that the tradeoff between current and future reproduction limits the optimal clutch size even further.

Sex Ratio and Sex Allocation

Another basic question in life-history theory is the relative allocation to sons or daughters in separate-sexed species and into male versus female gamete production in hermaphrodites. In general, most populations exhibit an equal primary sex ratio at birth (i.e., equal numbers of sons and daughters on average). In 1930, Fisher developed a model explaining this observation. If sons and daughters take an equal amount of energy to produce and every zygote is produced by the fusion of one male and female gamete, then it can be shown that sons have an advantage when males are rare and daughter have an advantage when females are rare. The evolutionary stable point occurs where on average individuals produce an equal number of sons and daughters at birth (Fig. 7). However, deviations from this classic expectation are observed. For example, studies in fig wasps (where females lay eggs in figs where offspring mate before leaving) have shown that females produce more daughters than sons when they are the only females laying eggs in a single fig (thus decreasing competition among their sons to fertilize their daughters). In addition, studies on red

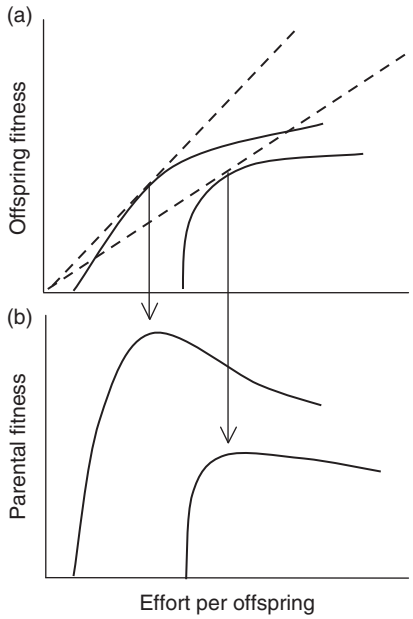


Fig. 5 A graphical model presented by Smith and Fretwell in 1974 that predicts the optimal tradeoff between offspring effort and offspring number for a given level of parental reproductive effort. (a) Offspring fitness for two environments. The model assumes offspring fitness is a convex function of parental effort. (b) Corresponding parental fitness for the two environments. Parental fitness is a function of offspring effort; the maximum occurs where the tangent of the curve in (a) intersects with zero. The shape of the curve in (b) is the product of the curve in (a) and by the shape of the tradeoff between offspring size and number (not shown). The model predicts a single optimal offspring size, set by the environment, that maximizes parental fitness, and that this size is 'less' than the size which maximizes offspring fitness. Much of the research on the offspring-size and -number tradeoff is rooted in these simple predictions.

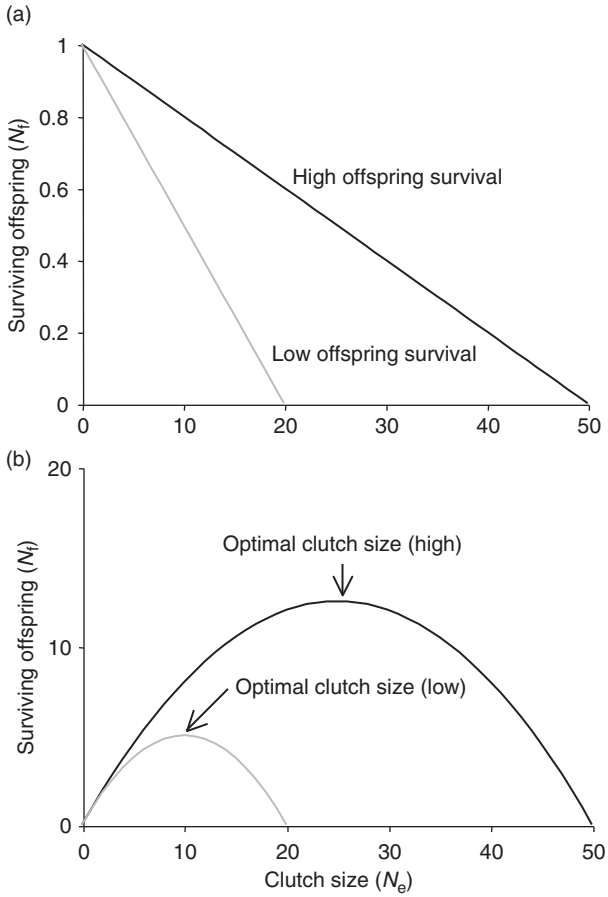


Fig. 6 Optimal clutch size. (a) Offspring survival as a function of clutch size for $c=0.02$ (high, black line) and $c=0.05$ (low, gray line). (b) Number of offspring fledged as a function of clutch size. Optimal clutch size is lower ($N_e=10$) for low offspring survival than for high offspring survival ($N_e=25$). For a description of the model and equations used see the text.

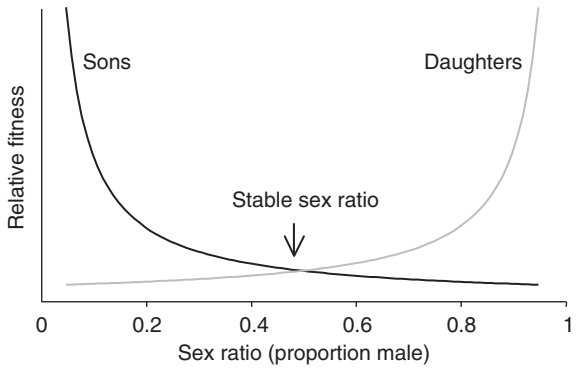


Fig. 7 Sex-allocation theory argues that the fitness of males and females is negatively frequency-dependent meaning that sons will have a fitness advantage when males are rare in the population while daughters will have greater relative fitness when females are rare in the population. The Fisherian evolutionary stable sex ratio is predicted to occur at a sex ratio of equal males and females (proportion male = 0.50).

deer have found that females in high condition are more likely to produce sons and females in poor condition are more likely to produce daughters. This pattern can be explained by the fact that daughters are able to produce offspring even if small, while small males from low-condition mothers tend to have no reproductive success (as a result of strong competition among males for access to females). Sex-allocation theory has also been able to explain the relative allocation to male and female gamete production in hermaphrodites (mainly in fish and plants) but has been less successful in explaining why any individual species is hermaphroditic instead of dioecious (separate-sexed).

Departures from Classic Theory

The power of life-history theory to predict observed patterns has been immense. Life-history models have been able to predict patterns of variation in traits such as age at maturity, number of offspring produced, life span, and the relative production of sons versus daughters. However, empirical support for predicted life-history patterns is sometimes equivocal. Phylogenetic history of a taxonomic group will often constrain the response to selection. For example, we would not expect to see an oak produce a single large acorn or a whale produce many tiny offspring. Furthermore, the detection of tradeoffs can be elusive, particularly in organisms with complex life-history traits (such as indeterminate growth, metamorphosis, or alternative reproductive tactics). In many cases, tradeoffs occur among a suite of life-history traits, which may become empirically intractable.

Environmental and genetic variation can also mask expected tradeoffs. In 1986, van Noordwijk and de Jong showed that variation in acquisition of resources will lead to undetected tradeoffs if that variation is greater than the variation in allocation. A common metaphor for this argument is the observation that people with big houses often drive expensive cars. We do not detect a tradeoff between size of house and size of car because variation in income in the population is so great. This idea has been largely successful in explaining deviations from expected life-history tradeoffs.

Future Directions

While life histories arise at the intersection of evolutionary and ecological dynamics, most of current life-history theory does not include ecological interactions (such as predator–prey or host–disease dynamics) and little empirical evidence exists examining the effect of life-history evolution on ecological interactions. Life-history theory must move from a dichotomous tradeoff approach to a more integrative concept of individual fitness in an ecological context. An understanding of how life-history patterns respond or do not respond to environmental change will be essential to species management and conservation in the future as will an understanding of how those life-history responses affect ecological- and ecosystem-level dynamics.

See also: Aquatic Ecology: Microbial Communities. Behavioral Ecology: Age Structure and Population Dynamics; Mating Systems. Ecological Data Analysis and Modelling: Forest Models. Evolutionary Ecology: Fecundity; Body Size, Energetics, and Evolution; Fitness; Evolutionary Ecology

Further Reading

Charnov, E.L., 1982. The Theory of Sex Allocation. Princeton, NJ: Princeton University Press.
Conover, D.O., Munch, S.B., 2002. Sustaining fisheries yields over evolutionary time scales. Science 297, 94–96.
de Jong, T., Klinkhamer, P., 2005. Evolutionary Ecology of Plant Reproductive Strategies. Cambridge: Cambridge University Press.
Fox, C.W., Roff, D.A., Fairbairn, D.J. (Eds.), 2001. Evolutionary Ecology: Concepts and Case Studies. Oxford: Oxford University Press.
Olsen, E.B., Heino, M., Lilly, G.R., *et al*., 2004. Maturation trends indicative of rapid evolution preceded the collapse of northern cod. Nature 428, 932–935.
Partridge, L., Mangel, M., 1999. Messages from mortality: The evolution of death rates in the old. Trends in Ecology and Evolution 14, 438–442.
Partridge, L., Prowse, N., Pignatelli, P., 1999. Another set of responses and correlated responses to selection on age at reproduction in *Drosophila melanogaster*. Proceedings of the Royal Society of London series B 266, 255–261.
Roff, D.A., 1992. The Evolution of Life Histories: Theory and Analysis. New York: Chapman and Hall.
Schaffer, W.M., 1974. Optimal reproductive effort in fluctuating environments. American Naturalist 108, 783–790.
Smith, C.C., Fretwell, S.D., 1974. Optimal balance between size and number of offspring. American Naturalist 108, 499–506.
Stearns, S.C., 1992. The Evolution of Life Histories. Oxford: Oxford University Press.
Van Noordwijk, A.J., de Jong, G., 1986. Acquisition and allocation of resources: Their influence on variation in life history tactics. American Naturalist 128, 137–142.
Wilson, A.M., Thompson, K., 1989. A comparative study of reproductive allocation in 40 British grasses. Functional Ecology 3, 297–302.

Limiting Factors and Liebig's Principle

K Mengel, Justus-Liebig-Universität Gießen, Pohlheim, Germany

Introduction

One of the most important resource for mankind is the fertile soil. Already in ancient times men knew that fertility is of utmost importance for living. The term fertility was associated with gods, the fertility of soils and animals including women. It was only in recent times due to the impact of modern philosophy, particularly the Enlightenment, men tried to explain fertility and also soil fertility in modern scientific terms. One of the first pioneers of this modern thinking was the French chemist Antoine Laurent de Lavoisier (1743–94). He carried out numerous field trials and came to the conclusion that the excrements of animals were the matter from which the humus was formed in soils and that the humus was the basis of soil fertility. It was shown that the incorporation of animal excrements into the soil improved crop yields. Based on these experiments the humus theory was formed which in the nineteenth century had a favorable impact on farming in France and Central Europe. Humus was considered as the nutrient of plants (humus theory). According to this theory animals would compete with men for fertile soils since animals were required for producing the matter by which soil fertility was maintained. This humus theory was accepted by farmers and had a favorable influence on farming in West and Central Europe.

Carl Sprengel (1787–1859), a German chemist at the beginning of the nineteenth century, analyzed plants and soils on their chemical elements and came to the conclusion that the actual plant nutrients were chemical elements. He disproved the humus theory.

Liebig's Principle

Justus Liebig (1803–73) carried out numerous analyses in plants and soils and he also came to the conclusion that plants feed from inorganic nutrients. According to his statement the first sources of plant nutrients are exclusively of inorganic nature. Liebig integrated his statement in a greater context as he speculated that plants in nature play a unique role in transferring chemical elements into organic plant matter. This organic matter when brought back into the soil is decomposed – in those days microbes were not yet known – and inorganic elements are released which again are taken up by plants as nutrients. To our knowledge Liebig was the first assuming a cyclic process in nature. According to this concept the limiting process of producing food for mankind was not the humus, and animals did not compete with man for arable land. He claimed that "the first sources of the plant nutrients are provided exclusively by inorganic nature" and he further wrote that "as the principle of agriculture must be seen that soils must receive in full measure what was taken from it." These were the inorganic plant nutrients and this was the start for producing and applying chemical fertilizers.

According to Liebig's concept, crop yield per hectare in Germany was increased by a factor of *c.* 3 in the last half of the nineteenth century. In addition, infertile soils, particularly sandy soils, were rendered fertile by the application of lime, phosphate, and potassium. This not only was the case in Germany but also in other European countries and in North America which applied Liebig's principle. Thus area and quality of soils limiting food production were increased.

In the beginning Liebig propagated that N fertilizer application was not necessary since according to his investigations ammonium N was taken up by plant leaves. This assumption contrasted with results of the French scientist Boussingault (1802–87) who according to his investigations held the view that fertilizing inorganic N was of great importance. Only later Liebig became convinced – particularly by the Rothamsted field trials – that inorganic nitrogen also should be fertilized. Liebig, however, immediately realized that the amount of potential inorganic fertilizer N, mainly guano, is limited and would run out quickly when farmers apply it in various countries. Liebig was interested when he heard that electrical discharges (lightning) oxidized the molecular nitrogen in the air and he speculated whether this could be a process by which nitrogen fertilizer could be produced. Liebig realized that nitrogen was a limiting factor in plant nutrition and he emphasized that farmers should carefully handle their farmyard manure and slurry thus avoiding N losses.

Liebig propagated the law of the minimum, originally going back on Carl Sprengel's idea that the limiting factor is controlling the crop yield. In the nineteenth century, nitrogen was limiting crop growth in many cases. Liebig's recommendation to fertilize soils with lime, phosphate, and potassium, however, had indirectly also a positive impact on soil nitrogen because lime, phosphate, and potassium are particularly required for the growth of leguminous species. The numerous attempts to grow clover mostly failed because soil pH was too low and potassium and phosphate were not available in sufficient quantities. These limiting factors were overcome by the application of lime, phosphate, and potassium, and leguminous crops thrived vigorously. These crop species living in symbiosis with N_2-fixing bacteria increased the level of available nitrogen in soils. Thus a further crop production limitation was overcome. This promoted farming enormously since farmers had more forage. They could feed their animals better and also increase the number of animals which produced more farmyard manure and slurry in the rotation which was brought back on fields and meadows.

In those days microbial dinitrogen fixation was not yet known until 1886. Liebig assumed that the broad leaves of leguminous species were able to absorb the ammonium from the atmosphere which actually is true but the quantities are too low for obtaining

satisfactory yields. Liebig recommended a rotation with two cereals followed by legumes or other dicotyledonous species (beets, potatoes). By this strategy infertile soils were rendered fertile and thus the area of fertile soils was enlarged. By increasing soil pH through liming, microbial activity also was promoted since most microbes require a neutral soil pH. Hence the decomposition of organic matter and thus the microbial production of ammonium and nitrate in soils was stimulated by which N nutrients are selectively taken up by plant roots. In addition, N_2 assimilation by free-living N_2 fixers and by microbes such as *Rhizobium* species living in symbiosis with higher plants profited from soil liming.

Synthetic Nitrogen Fertilizers

According to Liebig's principle, N supply of soils required leguminous crop species which not only supplied crops directly required for men but a substantial part of leguminous crops was required for animal forage. Hence men and animals competed for fertile soils. This situation changed drastically when the chemical production of N fertilizers was invented (Haber–Bosch technique). Now farming was possible without animals, which was the preposition for the separation of crop farming from animal farming. This had an enormous impact on the economics of farming and a disastrous one on ecology. In many animal farming systems land is overdressed with farmyard manure or slurry. In the Netherlands, the average surplus of mineral N, mainly nitrate, for all dairy farms is in the range of 200–400 kg N ha^{-1}. This quantity of plant-available N would suffice to produce 6000–12 000 kg wheat grain per ha per year. This surplus of N leads to an enormous nitrate leaching with a risk for the groundwater and also for denitrification by which N_2O, a greenhouse gas, is produced.

Australian sheep farmers did not follow Liebig's recommendation as they did not lime their pastures. Consequently, during a period of 50 years, soil pH gradually decreased from about 6 to 4.8. At this low pH, soluble and exchangeable Al and Mn oxides/hydroxides are formed which are toxic to plant roots, particularly to leguminous species and affect their growth, particularly that of *Trifolium subterraneum* which is the most important symbiotic N_2 fixer in Australian pastures. Thus the soils were rendered infertile. By this one of the most important worldwide resource, limiting resource, the fertile soil, is diminished. There are other soil processes which are less spectacular but still have a serious impact on soil fertility.

Potassium is an indispensable plant nutrient. In addition, it is indispensable in most fertile soils as an essential element of the structure of 2:1 clay minerals. Tributh *et al.* found that cropping fertile soils containing 2:1 clay minerals without K fertilizer application lose gradually their fertility because plant roots take up K^+ from the 2:1 silicate structure and thus finally destroy the mineral structure which means a loss of fertility. These minerals are capable to store K^+ and NH_4^+ and protect these important plant nutrients against leaching. Plant roots feed from thus-bound K^+ and NH_4^+. The nutrients thus taken up by the crop must be replenished by fertilizer application in order to maintain soil fertility as according to Liebig's principle what was taken from the soil should be given back. This degradation of a 'fertility mechanism' is a slow but irreversible process which renders eventually fertile soils infertile.

Worldwide Effects

For livestock farms the area of grassland and arable land has no major meaning if cheap imported concentrates, mainly soya, are available. This leads to serious environmental problems as already discussed above for intensive dairy farms. A great proportion of soya used for animal feeding is imported from Brasilia by European countries. In Brasilia large areas of the tropical forest are cut and the land is mainly used for the cultivation of soya. Under the tropical climate conditions, however, the humus of these soils is quickly oxidized and then soils are exposed to water and wind erosion and become irreversibly infertile within a few decades.

The surplus of nitrate in soils of Dutch farmers means also an accumulation of phosphate in soils as the animal excrements contain a substantial amount of phosphate. The latter, however, is a limiting resource which will be exhausted in a few centuries if mined with the actual rate. Therefore farmers should fertilize their soils according to soil analysis on available P. The same is true for N as a surplus of fertilizer N is not only an environmental hazard but also a waste of fossil energy required for the N fertilizer production. The production of NH_3 is associated with the release of CO_2. Precise N fertilizer rates based on soil analyses which take into account the inorganic soil N and a part of the organic N in the upper soil layer are highly profitable for farmers as recently shown by Mengel *et al.* Not considering the organic soil N leads to an overfertilization with N fertilizer with a disastrous impact on the environment.

Fresh water

In most countries where irrigation is applied, irrigation constitutes a major part of freshwater consumption. The efficiency of irrigation (percentage of water taken up by crops and total water applied) ranges between 50% and 90%, depending to a large extent on the type of irrigation. With drip irrigation, where the water drips through plastic types to individual plants, the irrigation efficiency is in the range of 90% whereas in furrow irrigation it is about 50%. The problem is not only the consumption of water but also the quantities of salt brought on the field with the water under arid conditions which finally results in soil salinity and thus soils become infertile. Already in ancient times, in North Africa and the Euphrates/Tigris lowlands, large areas became saline through irrigation and

went out of crop production. The drip irrigation cannot be applied to small crops such as cereals. These crops, however, are grown on a large area. Recently a new subsoil irrigation system was described which is based on irrigation techniques from ancient Persia. The clay types used are imbedded into the soil and have a porous structure which releases water into the soil according to the water requirement of plants. The moist zone around the pipe does not reach the soil surface. Hence no evaporation of water occurs and the risk of soil salinization is reduced to a minimum. The water-use efficiency (kg wheat grain/m^3 irrigation water) of the furrow irrigation system was on average 85 and for the subsoil irrigation 250.

High Na^+ concentration in soils affects soil structure and the uptake of other cationic species, particularly K^+. Hence plants suffer from K deficiency which particularly depresses protein synthesis and thus growth. Field experiments carried out under arid conditions in Turkey showed that the negative effect of saline water on the leaf growth of satsuma mandarins was efficiently counteracted by high K fertilizer rates.

See also: Ecological Data Analysis and Modelling: Carbon Biogeochemical Cycle and Consequences of Climate Changes. Ecological Processes: Nitrification. General Ecology: Tolerance Range; Plant Physiology; Water Availability; Growth Models; Growth Constraints: Michaelis–Menten Equation and Liebig?s Law. Global Change Ecology: Nitrogen Cycle; Phosphorus Cycle; Sustainable Cropping Systems. Human Ecology and Sustainability: Nitrogen Footprints; The Sustainable Development Goals

Further Reading

Banedjschafie S, Bastani S, Widmoser P, and Mengel K (in press) Improvement of water use and N fertilizer efficiency by subsoil irrigation of winter wheat. *European Journal of Agronomy*.

Boulaine, J., 1987. Modernité de Jean Baptiste Boussingault en matière de sciences de la terre. CR Academy of Agriculture of France 73 (6), 11–20.

Bromfield, S.M., Cumming, R.W., Davis, D., Williams, C.H., 1983. Change in soil pH, manganese and aluminium under subterranean clover. Australian Journal Experimental of Agriculture and Animal Husbandry 23, 181–191.

Hanegraf, M.C., de Boer, D.J., 2003. Perspectives and limitations of the Dutch minerals accounting system. European Journal of Agronomy 20, 25–31.

Johnston, A.E., 1992. Liebig and the Rothamsted experiments. In: Judel, G.K., Winnewisser, M. (Eds.), 150 Jahre Agrikulturchemie. Giessen, Germany: Symposium der Justus Liebig Gesellschaft, pp. 37–64.

Liebig, J., 1841. Die organische Chemie in ihrer Anwendung auf Agrikultur und Physiolgie. Brunswick, Germany: Verlag Viehweg.

Mengel, K., 1997. Agronomic measures for better utilization of soil and fertilizer phosphates. European Journal of Agronomy 7, 221–233.

Mengel, K., Hütsch, B., Kane, Y., 2006. Nitrogen fertilizer application rates on cereal crops according to available mineral and organic soil nitrogen. European Journal of Agronomy 24, 343–348.

Mulvaney, R.L., Khan, S.A., Ellsworth, T.R., 2006. Need for a soil based approach in managing nitrogen fertilizers for profitable corn production. Soil Science Society of America Journal 70, 172–182.

Picard, D., 1994. New approaches for cropping system studies in the tropics. In: Borin, M., Sattin, M. (Eds.), Third Congress of the European Society of Agronomy. Colmar, France: European Society of Agronomy, pp. 30–36.

Tributh, H., von Boguslawski, E., von Liers, A., Steffens, S., Mengel, K., 1987. Effect of potassium removal by crops on transformation of illitic clay minerals. Soil Science 143, 404–409.

Macroevolution

M Shpak, University of Texas at El Paso, El Paso, TX, USA

The term 'macroevolution' refers, broadly speaking, to evolutionary changes above the species level. The term can be used to refer to phenomena such as the origin of morphological (or biochemical) novelty, the origin and subsequent diversity dynamics of higher taxa due to changes in origination or extinction rates, and the role of regulatory genes and developmental constraints on the nature and rate of morphological change within and between clades.

There have been two schools of thought on the nature of macroevolution. While nearly all biologists recognize macroevolutionary phenomena as real patterns observable in the fossil record and from the reconstruction of phylogenies, it has been debated from the time of Darwin and the rediscovery of Mendel's laws whether the microevolutionary processes at the population level extrapolated over time account for all macroevolutionary patterns, or whether there are special evolutionary mechanisms, fundamentally different from those acting within species and populations, responsible for the major events in the history of higher taxa.

Historically, this debate has taken several forms since the resurrection of Mendelism in the early twentieth century. The first was the conflict between the 'gradualist' Darwinian school and the 'mutationist' school represented by geneticists such as H. de Vries and later R. Goldschmidt. While the work of R. A. Fisher reconciled the alleged conflict between Mendelian genetics and the quantitative traits studied by the biometricians, it left open the debate of whether the most important traits in evolution were polygenic quantitative traits (nearly normally distributed due to the small effects of many genes) or discrete traits under the control of small numbers of genes. The macromutationist view emphasized the importance of mutations with large effects for the origin of the type of evolutionary novelties that define higher taxa and allow organisms to invade new regions of ecospace, limiting the importance of gradual Darwinian adaptation to phenomena at the intraspecific population level.

Since that time, 'mutationism' in the form advocated by de Vries and Goldschmidt has fallen into disrepute due to the simple empirical observation that almost every observed macromutation severely reduces the fitness of an organism, while potentially beneficial macromutations are sufficiently rare to be of limited evolutionary importance. Nevertheless, the spirit of mutationist thinking, which emphasizes the importance of rare, drastic changes in phenotype and the importance of internal genetic and developmental factors rather than molding by selection, lives on as an active source of debate.

Neo-Darwinism and the 'New Synthesis' Paradigm

The groundwork of classical population genetics theory by Fisher, Wright, Haldane, and others formulated in mathematical language the forces underlying microevolution – mutation, natural selection, genetic drift, assortative mating, and migration. With later contributions, the theory would also incorporate the genetics of recombination and linkage, as well as the selective dynamics of frequency dependence and other complicating processes.

The question that remained to be answered was whether this impressive body of theory could account for all the phenomena of interest to evolutionary biologists. In terms of explaining the patterns of variation and rates of change in heritable traits within populations, microevolutionary theory made enormous headway. In the modeling and prediction of adaptive intraspecific change and variation, Neo-Darwinian theory was certainly a triumph and to this day is the foundation for active research programs directed toward explaining the enormous amount of information about intraspecific genetic variation at the molecular level that has recently become available.

Where traditional population genetic models were (at least initially) found wanting was in describing the origin of species and higher taxa. The first question, that of speciation, was reconciled with population genetics theory at a qualitative level in the writings of the 'new synthesis', most significantly in the writings of the new synthesis, most significantly in the work of T. Dobzhansky. More recently, the mechanisms of both allopatric and sympatric speciation have been given a firm mathematical foundation, one which represents the complicated interactions between assortative mating, migration, and frequency-dependent selection on phenotypic traits.

While the importance of allopatric versus sympatric speciation was broadly debated and the importance of genetic drift versus adaptation in species formation was a source of contention, it was widely recognized that the mechanisms driving speciation were entirely consistent with the processes of natural selection, assortative mating, and genetic drift in classical population genetics theory. Speciation could basically be understood as the origin of pre- or postzygotic isolating mechanisms through the standard processes of microevolution. Consequently, in the Neo-Darwinian worldview, speciation is a special case of microevolutionary change, and higher taxa are defined by the phenotypic changes associated with the origin of their ancestral species (i.e., since all monophyletic higher taxa begin as single species and the morphological or molecular traits that define that higher taxon originated in the 'stem' species of the clade).

Nevertheless, speciation introduces a new dimension into the evolutionary process. Since the local adaptive evolution during the history of an ancestral species can potentially be decoupled from the changes that take place during speciation (particularly in peripatric speciation where the incipient species consists of a small, isolated subpopulation separated from the ancestor), some of

 doi:10.1016/B978-0-444-63768-0.00806-4

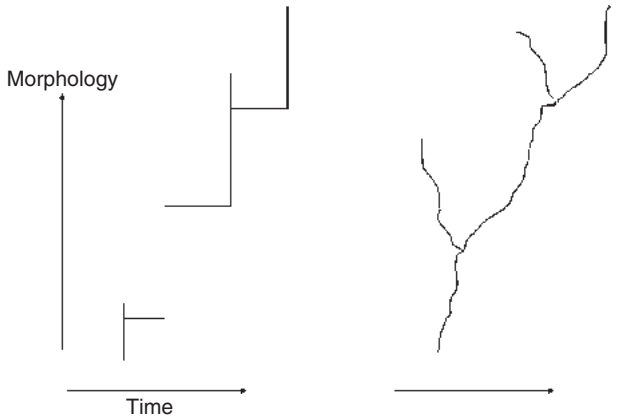


Fig. 1 A schematic illustrating the assumptions of how morphology changes through time under punctuated equilibrium (left) versus phyletic gradualism (right). In the former, most change in morphology is concentrated at speciation events; in the latter, speciation is a split in a lineage whose morphology has been evolving continuously.

the most visible recent challenges to the Neo-Darwinian worldview have focused on temporal patterns of speciation and cladogenesis in the fossil record.

Punctuated Equilibrium

It has long been observed in the fossil record that discontinuities between proposed ancestor and descendant species (and between closely related higher taxa) are quite prevalent. Paleontologists, following Darwin's own treatment of the subject, had traditionally assumed that this was due to the incompleteness of the fossil record itself.

Taking into consideration the implications of the peripatric model of speciation, S. J. Gould and N. Eldredge proposed that the observed discontinuities are often not an artifact of gaps in the stratigraphic record, but rather reflect the logical consequences of a supposedly prevalent mode of speciation. If most new species arise from small peripheral isolates rather than through a gradual transformation of a large population, then the change from ancestral to descendant morphology can be independent of the adaptive trends within the mainstream of the ancestral species' population. Furthermore, because the speciation event will involve a comparatively small marginal subpopulation, the transition is unlikely to be captured in the fossil record. What one observes instead is an ancestral species with either a static morphology, or a history characterized by random changes in phenotype unrelated to the differences between it and its descendant form. The phenomenon, combined by the subsequent sudden appearance of the descendant species, was referred to as 'punctuated equilibrium' by Eldredge and Gould.

As an empirical observation, punctuated equilibrium has been supported by a number of examples in the fossil record, although cases of phyletic gradualism (gradual transformative phenotypic evolution, as contrasted in Fig. 1) from one recognized morphology-defined species to another have also been documented in many fossil taxa.

The question remains whether this empirical observation constitutes a new model for evolutionary change inconsistent with population genetics theory. Following the arguments of C. H. Waddington on canalization and genetic assimilation, Eldredge and Gould postulated that the long periods of stasis reflect internal developmental homeostasis while the punctuations suggest a breaking of normally canalized development. Punctuated equilibrium proponents also argue that because of the supposed importance of non- or counteradaptive changes due to genetic drift during speciation (so-called 'founder flush' speciation leading to 'genetic revolutions'), the fact that most long-term trends in a lineage are due to speciation processes rather than a clade's history would imply that adaptation plays a limited role in macroevolutionary trends.

While acknowledging the empirical reality of punctuation in the fossil record, many population genetics theorists have convincingly argued that the punctuated equilibrium pattern does not require any exotic, hitherto-unknown processes in genetics or development. Though perhaps a small number of paleontologists and stratigraphers may have naively assumed that Darwinian evolution requires gradual, constant-rate changes through time, there is nothing in the Neo-Darwinian paradigm or population genetics theory that requires this to be the case, making the notion of a constant rate of evolution a kind of straw man.

During their history, populations can be subject to conflicting selective pressures. A relatively stable environment (or an ability to track a changing environment by changes in distribution) imposes stabilizing selection and hence stasis on phenotype, while the selective pressures that drive speciation generally involve directional selection on isolates or disruptive selection during sympatry. Indeed, the very nature of species as self-contained gene pools imposes stasis as a default and speciation as a comparatively infrequent phenomenon. Counter to the claim that evolutionary trends are nonadaptive because speciation is nonadaptive, speciation due to genetic drift alone has been demonstrated to be improbable and at best a rare event from both theoretical and empirical considerations. The evolutionary changes that lead to allopatric and sympatric speciation alike are generally driven by natural or sexual selection, with genetic drift usually (though not necessarily) playing a confounding role.

Thus the mechanisms involved in speciation, allopatric or otherwise, do not demand 'genetic revolutions' (for which there is no real empirical evidence) or any assumptions about developmental constraints, so as such the phenomenon of punctuated equilibrium does not require an explanation beyond what standard evolutionary theory already provided. In fact, a number of the original proponents of the 'strong' form of punctuated equilibrium, such as Eldredge himself, have in recent writings shown favor of this more conventional, Neo-Darwinian understanding of stasis and punctuation.

Nevertheless, the discussion of punctuated equilibrium and the distinction between 'phyletic' and 'cladogenetic' evolutionary change did lead to an evolutionary model that could not be so readily reconciled with Neo-Darwinism: the notion that species, and possibly higher taxa as well, are themselves units of selection and evolution.

Species Selection

Implicit in the work of Eldredge and Gould's original papers and explicit in S. M. Stanley's treatment of the subject is the connection of punctuated equilibrium with the concept of 'species selection'. The basic argument is that just like genes and individual organisms, species can reproduce (through speciation) and exhibit heritable variation (i.e., covariance between ancestral and descendant species traits). The basic idea is that certain traits relating to dispersal ability, tendencies toward habitat specificity, etc., predispose some taxa to have high rates of speciation, while other taxa have ecological or behavioral traits that make speciation infrequent.

From this perspective, the reason some clades have been species rich throughout their history and others species poor is not due to members of one clade having traits that lead to higher Darwinian fitness at the individual or population level, but rather because the species-rich clade consists of species that themselves are likely to speciate at an enhanced rate. At some level, the argument has intuitive appeal – it makes little sense to argue that individuals from species-rare 'living fossils' such as horseshoe crabs are 'unfit' when compared to individuals from, for example, species-rich beetles. It is more likely that there is something about the ecology of phytophagous insects that makes them speciate at higher rates than generalist benthic marine invertebrates. D. Jablonski and other paleontologists have proposed that generalist or eurytopic taxa will tend to be less species rich than specialist or stenotopic taxa. Specific habitat requirements tend to lead to strong competitive exclusion and character displacement, low rates of gene flow, and consequently high rates of speciation rates in specialists relative to generalists. It could also be argued that many observed long-term evolutionary trends, such as the increase in average body size observed in many lineages, occur not because of strong selection for the trait in question during the history of individual species or during speciation, but because taxa with the trait in question simply speciate at a higher rate.

Species selection has certainly been demonstrated to be a theoretical possibility based on the fact that any kind of heritable variation can be selected. However, species selection as a major force in evolution runs into the same problems as group or interdemic selection: the generation time of individual organisms is usually several orders of magnitude shorter than the 'generation time' for species. The former is generally on the scale of months or years, the latter on the scale of hundreds of thousands if not millions of years. The very argument that suggests species selection as a theoretical possibility, namely a decomposition of heritable variation into components following Price's equation, is what makes it of limited scope in practice. Specifically, rapid natural selection at the intraspecific level eliminates variance in trait at the interspecific level before higher-level selection can act.

Consequently, if a certain trait in an organism leads to a higher propensity to speciation, but is selected against among individuals in a population, then the trait in question will generally be lost prior to any speciation events. In recent studies, such as those of Sean Rice, it has been shown that under conditions when individual generation times are very long and speciation rates are very high, species selection can run counter to intraspecific selection, but these circumstances are probably exceptional.

Realistically, species selection may be important for traits that are on average neutral at the individual and genic (intraspecific) level. The possibility of species selection being important in such neutral traits remains an interesting empirical question, albeit one that is difficult to answer because in many cases it is impossible to disentangle the history of selection on a trait from its fossil or phylogenetic history.

Mass Extinction and Other Challenges from the Fossil Record

A number of other challenges to the view that macroevolution was simply extrapolated microevolutionary change have come from paleontology. While the process of speciation and its genetic explanation forge a strong conceptual link between microevolution and many of the morphological changes observed in fossil lineages, there are a number of events in the history of life that seem so striking and catastrophic that certainly no population genetic models, even with fluctuating selection intensity or strong frequency dependence, can account for them.

The first such case is that of mass extinction. It has been known since the inception of geology as a field of scientific inquiry that the major periods in geological time are defined stratigraphically by complexes of fossil organisms that are distinct from those in higher or lower strata. These complexes can be defined over many timescales, with eras being subdivided into periods and periods being subdivided into lower units still, most of which are defined by the presences or absences of characteristic flaural and faunal assemblages. One reason that such boundaries defined by distinct sets of organisms exist is that the history of life has been defined by catastrophic mass extinction events.

For example, it has been estimated that the mass extinction at the Permian–Triassic boundary was responsible for the extinction of nearly 97% of marine invertebrate species (~85% of genera) and strongly impacted terrestrial plants, insects, and vertebrates as well. The extinction at the Cretaceous–Tertiary boundary took a somewhat smaller but still catastrophic toll of *c.* 50% of marine invertebrate genera, as well as (famously) leading to the extinction of the dinosaurs and many other terrestrial animals and plants. Other significant mass extinctions characterized the Late Devonian and the Late Triassic, with lesser ones defining other stratigraphic boundaries. In the 1980s, D. Raup and J. J. Sepkoski argued from surveys of various fossil taxa that mass extinctions are periodic, on ~26-million-year cycles. Since then the notion has been attacked both on statistical grounds (i.e., a frequent lack of resolution of dating below 5-million-year intervals) and on the basis that no single geophysical cause accounts for all or even most mass extinctions.

While much debate surrounds the causes of mass extinctions, it appears that no one causal explanation can account for the half dozen or so documented mass dyings. Based on evidence of elevated iridium levels in the boundary strata, it is likely a comet impact that was responsible for the K–T boundary extinctions. On the other hand, there is little evidence that extraterrestrial impacts had anything to do with the much greater P–T extinction, which may have been driven by drastic changes in sea level, volcanism, and changes in methane and carbon dioxide levels in both terrestrial and marine environments. The earliest documented mass extinction, one in the Late Proterozoic (the Varangian ice age of *c.* 850 million years ago), which hit unicellular eukaryotes in the such as acritarchs, may have been the most catastrophic of all, as some geologists believe that it was caused by a severe and global cooling event ('snowball Earth') that froze most of the oceans and nearly wiped out all living organisms apart from those sustained near volcanic vents and similar warm refugia.

Regardless of specific causes, it is clear that such massive, sudden extinctions of species cannot be modeled meaningfully at the microevolutionary level, and what lineages persist versus which ones become extinct can be arguable due to matters of chance rather than matters of adaptation or fitness. It seems to be the case that specialist taxa are more vulnerable than generalist taxa, which again is a statement of 'selection' properties above the species level, one that introduces an intrinsically biological, rather than simply extrinsic, explanation for the outcomes of mass extinctions.

Whether ecological catastrophes caused by abiotic events can be considered evolutionary processes as such is a matter of semantics, but it is clear that to address questions such as 'why are brachiopods prevalent in the Paleozoic while bivalve mollusks are prevalent in the Mesozoic and Paleozoic?' or 'why are there no living trilobites or ammonites?' arguments based on superior or inferior Darwinian fitness are not the most meaningful way to seek answers. For such questions, an understanding of geology and random mass extinction is probably more informative than an understanding of Darwinian fitness in Mendelian populations.

But what of the other aspect of biostratigraphy? Geological boundaries are defined not only by the extinction of groups of taxa, but by the relatively sudden origin of many others. The extinction of large amphibians and pelycosaurs in the Permian was followed by a dinosaur radiation in the Mesozoic, just as the extinction of dinosaurs was followed by a radiation of mammals in the Cenozoic. The history of life contains many even more spectacular 'adaptive radiations' – the most dramatic being the Cambrian explosion, which was characterized by the rapid appearance of most of the major metazoan phyla in the Early Cambrian after the near absence of any but the most basal invertebrate taxa in the latest Proterozoic.

As with mass extinctions, such rapid evolutionary radiations are often due to major environmental changes (biotic or abiotic) in Earth history. For example, among the many competing explanations for the Cambrian explosion, one of the leading hypotheses involves the likely increase in atmospheric free oxygen in the Late Proterozoic. Other major radiations are to a large part due to the ecological niches vacated by the mass extinction of competing taxa; for example, large Cenozoic mammals filled niches once occupied by dinosaurs.

The issue of whether adaptive radiations are due to processes outside the scope of classical Neo-Darwinism is a question having two facets. The first is whether the phenomenon of high rates of speciation and adaptive evolution can be explained through classic Neo-Darwinian (population genetic) models; the second is whether the extensive changes in morphology characterizing the origin of higher taxa (as in the Cambrian explosion) require special genetic or developmental explanations.

Adaptive Radiations

The seemingly accelerated rates of evolution that occur in adaptive radiations have long been recognized by biologists and paleontologists. G. G. Simpson, who did much to reconcile paleontological data with Neo-Darwinian notions of adaptive change and speciation, referred to the phenomenon as 'quantum evolution', to distinguish it from the 'background' processes of adaptation and speciation that characterize most of a lineage's history. However, Simpson emphasized that quantum evolution did not entail any novel evolutionary mechanism. Rather, he saw the process as a consequence of the fact that external circumstances (such as a vacated ecological niche) imposed much stronger diversifying selection than that encountered by most taxa during their normative history.

That adaptive radiations can be seen as adaptation and speciation under special external circumstances has support from classical population genetics theory. In an important early paper, R. A. Fisher demonstrated that mutations with large fitness consequences have a higher probability of fixation early in the history of a species, with adaptive change having 'diminishing returns' as time progressed. This is the basic pattern with many evolutionary radiations – large-scale differences are established first, followed by lower-level changes later in the history of a lineage.

The basic argument is that if the space of possible changes in phenotype is represented in a multidimensional Cartesian coordinate system. In this multidimensional space, an ancestral genotype is some distance from an optimum. At first, mutations

with large phenotypic effects have a high probability of moving the population closer to the optimum and increasing mean fitness. However, as the population approaches the optimum, mutations with large phenotype effects are more likely to move the population further from the optimum and reduce fitness, so that only mutations with small effects have a high probability of fixation and increasing mean fitness. While the details of Fisher's analysis have been disputed, this general result has been vindicated in a number of recent papers. A related argument, based on the declining probability of moving to higher adaptive 'peaks' on a fitness landscape with multiple local optima, was advanced by S. Kauffman and other 'complexity' theorists.

The same principles that apply to tracking optima with intraspecific variation are applicable to differences between species. In the case of an ancestral species encountering a large number of open niches, each new niche colonized represents a large distance to be reached in genotype and phenotype space. Since there are many such niches, initially mutations with large-scale phenotypic effects are likely to be favored because they are likely to take the genotype to a better evolutionary place. Afterward, when at the 'coarse-grained' level the descendant lineages have occupied the new niches, mutations with large effects are more likely to move them into regions of inferior fitness. At that point, mutations with small phenotypic effects are likely to be fixed as the populations optimize their fitness within the new niche (i.e., the type of 'fine-grained' adaptation that constitutes normal adaptive evolution).

This dynamics has been observed in a number of computer simulations of adaptive evolution, and is consistent with the observed pattern in the phylogenies of many higher taxa where the deeper nodes define major changes in phenotype associated with colonizing new regions of ecological space, while the differences among lower-level taxa (e.g., interspecific differences within a genus) usually involve relatively minor differences in morphology and ecology.

In other words, adaptive radiations reflect adaptive evolution and speciation on fitness landscapes that are intrinsically coarser grained than those typically encountered by most evolving lineages. However, while such a perspective accounts for the types of adaptive radiations typically discussed in the literature (e.g., differences in trophic modes and coloration in African lake cichlids, beak morphology in finches on the Galapagos, etc.), it still leaves unanswered the other aspect of evolutionary radiations seen in the early fossil record – the origin of truly novel morphologies.

Evolutionary Novelty and Innovation

It has often been argued that the differences in morphology seen in the Cambrian explosion seem more 'fundamental' than the differences in body shape and color seen in recent adaptive radiations because of a retrospective fallacy, that is, we define certain characters as more fundamental simply because they are older. While this may be true for certain characters (e.g., traits that define higher taxa in some groups vary intraspecifically in others), it is hard to escape from the view that the presence or absence of a coelom, body segments, and limbs reflect greater and more fundamental differences in genetics, development, and ecology than the shape of a finch's beak or a cichlid's teeth. Therefore, evolutionists are left with the very real question of why certain periods in the history of life are characterized by major novelties in morphology and modes of life.

It was because of this question that macromutationism and Goldschmidt's "hopeful monsters" had such appeal. The theoretical edifice of Neo-Darwinism, population and quantitative genetics, assumed a fixed space defined by preexisting quantitative traits and loci. The theory of mutation, selection, and genetic drift described the dynamics of populations in this space, but had little to say about how the space itself changes through the addition of new genes and new characters.

The first steps toward ameliorating the lack of a theory of evolutionary novelty came from work on gene duplication, which looked at the origin of new genes rather than allelic variation at existing loci. The mathematical structure of gene duplication models and general representations of novel characters is quite different from the representations used in classical population and quantitative genetics, and recent progress has been made in defining the properties of 'evolutionary space' where the number of dimensions (characters, as opposed to states of a fixed number of characters) is itself fluid.

However, because the mathematical language of Neo-Darwinism was incomplete from the standpoint of describing evolutionary novelty, this does not imply that the biological mechanisms involved in morphological and genetic innovations are unknown and outside the scope of genetics. In a sense, the problem is analogous to the state of speciation theory in the first half of the twentieth century – while speciation (and evolutionary novelty) may not required any novel genetic or evolutionary mechanisms as explainations, the existing paradigm failed to incorporate these phenomena, because in the first case classical population genetics did not ask questions about the origin of isolating mechanisms, and in the second case it did not ask questions about how to represent the loss or gain of traits in the traditional framework.

As to the underlying mechanisms involved in the origin of evolutionary novelty, most of the current explanations focus on two classes of genetic phenomena: gene duplication and developmental regulatory genes. In the former case, the duplication of a gene through various means (such as unequal crossover or gene conversion) creates genetic redundancy that at least in principle permits one copy to take on novel functions while the duplicate maintains its original function. This allows for the origin of novelty without counteradaptive macromutational changes. A noteworthy and familiar example is the globin gene family in vertebrates, where gene duplication has allowed these oxygen-carrying molecules to evolve specialized functions such as fetal hemoglobin in mammals.

The other main empirical contribution to the understanding of evolutionary novelty is the discovery of regulatory genes (such as Hox genes in animals) that control the activation and timing of downstream regulatory cascades. Such genes have the property that a single locus mutation can have drastic phenotypic consequences. The evolution of Hox and other regulatory genes is also full of examples where one regulatory gene co-opts the function of others or is recruited in an entirely new function.

While many mutations in such regulatory genes are deleterious if not outright lethal, a number of experimentally induced Hox mutations in *Drosophila* and other laboratory animals have produced morphological variants (such as additional pairs of wings in normally two-winged flies, or limbs in place of antennae) that resemble the morphologies of distantly related taxa. For example, it is likely that the small amount of overall genetic difference between humans and great apes, which is observed in spite of the great morphological and behavioral differences, is quite typical of macroevolutionary changes in morphology. Specifically, mutations in a small number of regulatory genes lead to significant phenotypic divergence, including the losses of key traits such as appendages, change in their position, and sometimes even the seemingly *de novo* acquisition of traits.

Consequently, just as adaptive radiations can be seen as a special case of speciation and adaptive evolution under certain fitness landscapes, so large-scale changes in phenotype can be seen as special case of mutation and selection in regulatory genes.

Concluding Remarks

Much of the debate between the sufficiency of microevolutionary mechanisms to account for macroevolutionary phenomena stems from looking at problems at different spatial or temporal scales. By analogy, while nobody doubts that macroeconomic phenomena ultimately arise as a consequence of individual decision-making processes, nobody argues that macroeconomic phenomena are not real and that large-scale human institutions have emergent properties that can be studied and understood in their own right. In the same way, biological phenomena are 'real' and need not be reduced to physics and chemistry to be understood, even though no biological processes contradict the laws of chemistry or physics. In fact, it is often more instructive to study biological phenomena at their own scale and level of organization than by reducing them to the underlying physicochemical interactions.

By analogy, although most if not all patterns of interest of macroevolutionists (apart from catastrophic environmental changes that are more questions of geology than of biology) have their origins in the genetics and variational properties of ancestral species, this does not imply that problems such as the invasion of new ecological niches, or the distribution of characters in phylogenies, are best understood by studying population genetics. It simply means that these phenomena are not in violation of microevolutionary processes, and do not require the discovery of previously unknown mechanisms any more than the emergent properties of living organisms violate the laws of chemistry and physics.

Along similar lines, changes in regulatory genes that have major effects on phenotype differ quantitatively and qualitatively from those genes that have minor effects on size or color within species, yet it would be difficult to justify the claim that such major changes in phenotype are caused by fundamentally different processes at the molecular level than those that lead to the more typical material of intraspecific variation. Yet even though such variants are subject to the same underlying population biology and the same forces of natural selection, one can probably learn more about these processes through an understanding of developmental genetics than from population genetics theory. One must distinguish the claim that macroevolutionary phenomena do not contradict microevolutionary (almost certainly the case) from the claim that macroevolutionary events are simply epiphenomena of the former, which is harder to justify given the qualitative differences in time and scope.

See also: Evolutionary Ecology: Genetic Drift; Units of Selection; Natural Selection; Phylogenomics and Phylogenetics; Ecological Niche. General Ecology: Allopatry. Global Change Ecology: Paleoclimatology

Further Reading

Arnold, A.J., Fristrup, K., 1982. A hierarchical expansion of the theory of evolution by natural selection. Paleobiology 8, 113–129.
Charlesworth, B., Lande, R., Slatkin, M., 1982. A Neo-Darwinian commentary on macroevolution. Evolution 36, 474–498.
Eldredge, N., 1990. Macroevolutionary Dynamics: Species, Niches, and Adaptive Peaks. New York: McGraw-Hill.
Gavrilets, S., 2004. Fitness Landscapes and the Origin of Species. Princeton, NJ: Princeton University Press.
Gavrilets, S., Vose, A., 2005. Dynamic patterns of adaptive radiation. Proceedings of the National Academy of Sciences of the United States of America 102, 18040–18045.
Gerhard, J., Kirchner, M., 1997. Cells, Embryos and Evolution: Toward a Cellular and Developmental Understanding of Phenotypic Variation and Evolutionary Adaptability. New York: Blackwell Scientific.
Gould, S.J., 2002. The Structure of Evolutionary Theory. Cambridge, MA: Harvard University Press.
Kauffman, S.A., 1993. The Origins of Order: Self-Organization and Selection in Evolution. Oxford: Oxford University Press.
Knoll, A.H., 2004. Life on a Young Planet: The First Three Billion Years of Evolution of Life on Earth. Princeton, NJ: Princeton University Press.
Levinton, J.S., 2001. Genetics, Paleontology, and Macroevolution, 2nd edn. Cambridge: Cambridge University Press.
Maynard Smith, J., Szathmary, E., 1998. The Major Transitions in Evolution. Oxford: Oxford University Press.
Newman, C.M., Cohen, J.E., Kipnis, C., 1985. Neo-Darwinian evolution implies punctuated equilibria. Nature 315, 400–401.
Nitecki, M.H. (Ed.), 1990. Evolutionary Innovations. Chicago: University of Chicago Press.
Raff, R.A., 1996. The Shape of Life: Genes, Development, and the Evolution of Animal Form. Chicago: University of Chicago Press.
Raup, D.M., 1992. Extinction: Bad Genes or Bad Luck? New York: W. W. Norton.
Rice, S.H., 1995. A genetical theory of species selection. Journal of Theoretical Biology 143, 319–342.
Simpson, G.G., 1944. Tempo and Mode in Evolution. New York: Columbia University Press.
Stanley, S.M., 1979. Macroevolution: Pattern and Process. Baltimore, MD: The Johns Hopkins University Press.
Valentine, J.W. (Ed.), 1986. Phanerozoic Diversity Patterns: Profiles in Macroevolution. Princeton, NJ: Princeton University Press.

Metacommunities☆

Marcel Holyoak, University of California, Davis, CA, United States

Glossary

Competition-colonization tradeoffs Occur when within a community or metacommunity the competitive ability (or rank) are negatively related to the dispersal ability of species. With either local disturbances that open up vacant habitat or through simple birth and death events opening up spaces within communities, species can then persist locally either by being good competitors or regionally by being good at dispersing to vacant habitat.

Lottery competition Is when individuals compete for the occupancy of sites or places within a community. The winner of competition is random with respect to individual and species identity.

Mass effect Represents a flux of individuals from a large community to a small community. Analogous to gravity larger bodies exert more effect than smaller bodies, and so larger communities contain more individuals summed across species and these are likely to spillover into smaller communities and enhance abundance, and potentially species richness.

Phylogenetic dispersion Represents the degree to which taxa within a community are clustered versus dispersed in their degree of relatedness. Filtering by the local environment would produce clustered communities, whereas competition between species would produce dispersed communities.

Rescue effect Is when individuals immigrate into a small population and augment abundance, thereby forestalling extinction of the small population.

Source-sink dynamic A flux from populations in a high quality (source) habitat to a poor quality (sink) habitat, which allows populations in the source habitat to persist. Rescue effects prevent extinction of the source population.

Spatial subsidy An ecosystem concept representing the movement of materials or individuals from one ecosystem to another. It includes dead organisms, sediments, and living organisms that bring energy, chemicals and nutrients from one system to another.

Introduction

A metacommunity is defined as a group of communities that are connected by the dispersal of one or more interacting species. The term was first used in 1991 and since then has grown into an important concept for studies of species diversity and community structure. The concept enlarges the scale at which community dynamics are considered. Metacommunities have several distinctions that make them a valuable unit for consideration in both theoretical and applied ecology, such as in considering the effects of habitat fragmentation on biodiversity and ecological communities.

First, traditional community ecology often relies on the assumption that communities are closed, isolated entities. This assumption arises from consideration of mathematical models such as the Lotka and Voterra competition equations, which are simpler if the community is assumed to be closed to movement. By contrast, theories like source and sink dynamics propose that immigration allows some species to be present even in sink habitats where conditions are insufficient to support viable populations in the absence of immigration, either due to low resources or the presence of predators or competitors. Therefore opening communities to immigration may change the species present in local communities, such that the conditions assumed in closed models do not apply to open communities. Movement of individuals might also modify species interactions and the local dynamics of individual species. It is widely acknowledged that numbers of individuals moving and the distances they move are some of the most difficult parameters to measure in ecology, and therefore our knowledge of the degree to which real communities are open or closed is limited. Robert Ricklefs argues that the concept of a local community is invalid, with assemblages being little more than a representation of the overlapping distributions of various species and the result of the interactions between them.

Second, the responses of species diversity to habitat change including partial destruction and fragmentation may arise either because of the responses of individual species or because species influence one-another. Hence mechanisms may arise at levels of populations, metapopulations, communities, and metacommunities. If we consider only some of these levels we will have at best an incomplete idea of the effects of habitat change.

Third, it is useful to recognize that regional metacommunity-level species diversity is determined by the sum of both local and regional processes. The local and regional parts merit elaboration. The local contribution to species diversity includes the structure

☆*Change History*: March 2018. M. Holyoak added a new Figure 2, Glossary, and new section "Combining Functional Traits and Phylogenetic Information with Species Composition Patterns". He updated the evidence for different kinds of metacommunity dynamics and made minor edits throughout the text.

This is an update of M. Holyoak and T.M. Mata, Metacommunities, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 2313–2318.

 doi:10.1016/B978-0-12-409548-9.11114-5

of the local community, ideas that are recognized by traditional community ecology including niche theory. A variety of regional processes are possible: (1) there may be a balance between the extinction of species from local communities and their (re) colonization, creating a balance that allows species to persist regionally even if they do not have a predictable place in any given local community. Such dynamics may be rendered more likely if there are tradeoffs in the abilities of species to colonize new or vacant habitat and their ability to compete locally. (2) Immigration may forestall extinction from local communities, as in source and sink models and through so-called rescue effects where immigration raises population size and rescues local populations from extinction. (3) Differences in habitat type among local communities might create different niches for different species. (4) The potential for (1) to (3) may depend on the degree to which dynamics are independent (or asynchronous) in different local communities. There is also the possibility that immigration and emigration modify either the richness and composition of the community or the interactions of species within that community, which represent an interaction between local and regional processes. Although two spatial scales are recognized, local and regional, this is arbitrary and more spatial scales are often represented in metacommunity models. The feedback between local and regional scales differs sharply from the equilibrium theory of island biogeography, where species diversity is viewed as being fixed by a permanent mainland pool of species and individual habitat islands contain arbitrary subsets of this species pool.

Fourth, interaction among species means that the metacommunity is not simply a collection of metapopulations where species are largely independent from one-another. Single species, or those independent from others in communities, and pairs of interacting species have typically been considered in the literature about metapopulations. Whereas, larger numbers of species are increasingly considered to be the domain of metacommunity theory. Many-species spatial competition models were not originally termed metacommunity models but best fit this categorization. Ideas such as keystone species recognize that, relative to their abundance or biomass, some species may have disproportionate effects on other community members.

Empirical work has long recognized a spatial component to species diversity, through the division of regional (gamma) diversity into local (alpha) diversity and the turnover of species among sites (beta diversity). The existence of differences in species composition among sites (beta diversity) creates the potential for movement to alter local community composition if species are able to move among local communities. The concept of alpha, beta and gamma diversity also recognizes that regional diversity is made up of local diversity and the differences among local communities. Such diversity patterns do not, however, distinguish what is creating the diversity, in other words, whether local diversity contributes to regional diversity or whether it is a subset of regional diversity.

In the following sections we describe models of species diversity, and then describe some of the factors other than species diversity that are influenced by a metacommunity structure.

Mechanisms of Metacommunity Dynamics

Four broad kinds of metacommunity model have been described to describe the assembly and maintenance of communities at local and regional scales, each emphasizing different persistence mechanisms. These are termed neutral community models, patch dynamics, species sorting and mass effects models. It should be borne in mind that although the discussion is often in terms of patches, the same kinds of dynamics can occur in spatially continuous habitats without discrete patches.

The best known metacommunity models are neutral community models, which assume that all individuals have identical fitness. Neutral models were popularized by the publication in 2001 of Stephen Hubbell's book *The Unified Neutral Theory of Biodiversity and Biogeography*. Hubbell's model includes lottery competition among individuals that are all equal competitors on a spatially uniform habitat plane (Fig. 1). Competition in a uniform environment with no differences among species causes species to be transient, and for species diversity to drift through time (termed ecological drift by Hubbell). In Hubbell's model all individuals also have limited movement ability (Fig. 1), which causes different species to cluster in different parts of the habitat plane, slowing down competitive exclusion but not preventing it. To replace species that were lost, Hubbell included speciation as a source of new species. The mechanisms maintaining diversity in Hubbell's model are localized dispersal, and speciation that

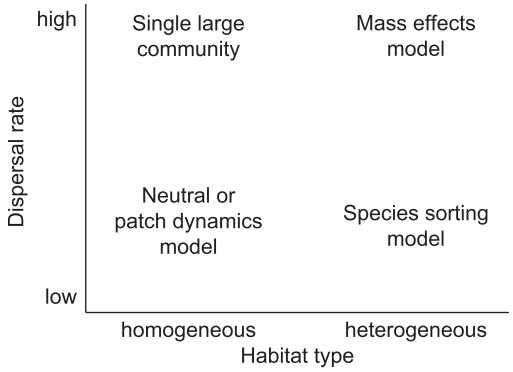


Fig. 1 Theoretical types of communities and metacommunities as a response to the similarity of habitat patches and the dispersal rate of species between communities.

replaces lost species. Hubbell's model was novel in including both ecological and evolutionary processes that might account for species diversity. Neutral models by other authors have often started with a fixed number of species and limited the time span of simulations so that numerous species are still present. Other, so-called neutral models have a fixed pool of species and all individuals are identical, but species cannot go extinct from the regional metacommunity. In such models persistence is determined by the permanent pool of species (like in the equilibrium theory of island biogeography), and this pool of species prevents neutral dynamics from determining diversity. Hence although individuals are neutral, the mechanisms determining diversity that are present in neutral models without permanent species pools cannot play out. Neutral models have surprised ecologists by predicting a number of patterns that are widely observed, such as the distribution of the rank abundance of species, or species-area relationships. Original tests were primarily in tropical trees, but neutral community models have been tested in a variety of taxa, from sedges, to bacteria, to gut parasites, and to bird assemblages. Unfortunately most work has concentrated on testing between neutral models and a single alternate null model, rather than accepting that there might be elements of neutral dynamics that occur along with other kinds of dynamics. For example, neutral models emphasize stochastic and transient dynamics, but such dynamics are also features of other kinds of models. In tests of a single alternative model, the best evidence leads to the rejection of neutral dynamics in all cases.

A second type of model has been termed patch dynamics, and differs from neutral dynamics because species differ in traits like colonization or competitive ability. Patch dynamics models are multispecies versions of classical metapopulation models. For large numbers of species to coexist permanently there needs to be a trade-off among appropriate traits. Most commonly they consider competing species that have a trade-off such that those that are better competitors are poorer dispersers. Inferior competitors can survive by living a fugitive existence where they arrive early at patches unoccupied by superior competitors. The habitat is uniform (**Fig. 1**), lacking environmental differences and can either be continuous or subdivided into separate patches. For coexistence a trade-off would need to be present, and species would need to have localized dispersal (like in neutral community models). Too much dispersal would cause species to show synchronous dynamics across all patches and inferior competitors would occur in the same patches as superior competitors until they were driven regionally extinct and a large single local community is left (**Fig. 1**). The model is hard to prove because it relies on the absence of environmental effects (that are present in the two models below) and it is difficult to demonstrate trade-offs for multiple species. There are recent demonstrations of such a trade-off among coexisting species in simplified laboratory systems where the environment is uniform but subdivided so that movement is limited.

A third type of model, termed species sorting, is based on classical niche theory, where niches are present either in different habitat types or at points along a habitat gradient (**Fig. 1**). Differences in habitats could be due to physical or biotic factors. Physical factors could include things like temperature, nutrient availability, resource availability, and tidal exposure, whereas biotic factors include the presence of habitat-forming species (e.g., trees or corals) or keystone species that exert disproportionate effects on other species. For species sorting to maintain diversity, species would need to be capable of reaching habitat to recolonize if they had gone extinct. If species dispersed so frequently that they often left good habitat, dynamics would become more like the next model, which involves mass effects (**Fig. 1**). For species sorting to operate, species could either select habitat actively and behaviorally, or they could have differential success in different habitats. The consequences of either of these would be that species survive and reproduce better in some habitats than others. Species functional traits that relate to responses to the environment are therefore an important part of community structure. The persistence mechanism in species sorting models is classical niche partitioning, in this case where niches are in different spatial locations. Niches are often contrasted with neutral models because individual responses to the environment would not be consistent with neutral models where all individuals are equivalent. Therefore species composition varying in response to environmental conditions would be evidence against neutral community models. Some of the best evidence for species sorting comes from differences in zooplankton composition among interconnected ponds with and without predatory fish. Dispersal of zooplankton between ponds seems to be somewhat frequent in such systems but species composition is strongly related to the presence of fish which do not move readily between ponds.

The fourth model, mass effects, results when there are different habitat types present and species are highly dispersive, so that they spillover from the habitats in which they reproduce and survive best to poorer habitats. These kinds of dynamics are described by source and sink models for individual species and are termed a mass effects for multiple species, where these is a mass flow of individuals from one place to another (**Fig. 1**). Habitat areas that are suboptimal and more distant in space from good habitat would be less likely to contain species that cannot maintain viable populations in the suboptimal habitat than for similar areas that are close to good habitat areas. This effect arises because species are likely to be limited in their dispersal ability. Were such dynamics present, there should also be a correlation between population productivity and the environmental factors that make habitat good or bad. The regional persistence mechanism is best described as a spatial storage effect, which Peter Chesson describes as differential survival and reproduction in different locations averaging out the success (finite growth rate) of a species such that the species can maintain a viable population within the metacommunity. Such dynamics have been modeled by Nicolas Mouquet and Michele Loreau in models with differences in habitat among patches and many competing species that differ in their response to environmental conditions.

If dispersal among habitat patches is high and habitats are relatively uniform across space, a single large local community is likely to result (**Fig. 1**). In this case it is not clear that anything is to be gained by considering the spatial dynamics of such a community. That is, the system is structured as a community rather than a metacommunity.

The most general form of evidence for and against the various metacommunity models comes from examining whether differences in community composition can be explained by differences in environmental factors or by the distance between the

local communities. Species sorting predicts a purely environmental effect, whereas mass effects predict both environmental effects and effects of distance. Both neutral and patch dynamic models predict effects of distance but not of environmental factors. A complicating factor in such analyses is that the environment may also be spatially structured. A survey of 158 natural community datasets by Karl Cottenie found that 22% of the total variation in community structure was explained by the pure environmental fraction, 16% by the pure spatial component and 10% by the spatially structured environmental fraction. From these data Cottenie concluded that 69 metacommunities (44%) best fit the species sorting model, 46 (29%) a mixed species sorting and mass effect model, and only 13 (or 8%) the neutral or patch dynamics models; 19 data sets could not be associated with these models, and 11 had no significant components. Therefore it is likely that species sorting and mass effects are relatively frequent in the natural world, but patch and neutral dynamics are less frequent. To date there are no natural communities that have been extensively examined and where we can conclude that the dynamics best fit neutral community models. A book edited in 2005 by Holyoak, Leibold and Holt, on metacommunities also found that more detailed studies most frequently identified species sorting, followed by mass effects and with patch dynamics being less common. There was no system that clearly showed neutral community dynamics. Likewise a 2011 review found that species sorting and mass effects were both the most commonly tested and supported metacommunity models.

It should be noted that the metacommunity models described above are useful as a starting point, but that they are theoretical representations of what is likely a continuum between different kinds of dynamics. A review of evidence for metacommunity dynamics in 2011 found that individual empirical systems often showed a mix of two of the four types of metacommunity dynamics. Currently the literature lacks models that can represent all of these kinds of dynamics within a single mathematical model, which would be useful for showing how the four different kinds of dynamics relate to each other. Consequently, more empirical and theoretical investigations are needed to ascertain the kinds of dynamics that are found and whether these relate to particular ecosystems or kinds of species. The above discussion also relates primarily to species diversity whereas in the following section we consider other aspects of community structure.

Metacommunity Topics Extending Beyond Species Diversity

Combining Functional Traits and Phylogenetic Information With Species Composition Patterns

Beyond the four metacommunity models described above, a wide range of studies use a combination of functional traits, environmental information and phylogenetic information to study patterns of species co-occurrence or species composition (Fig. 2). Functional traits are expected to be related to environmental conditions and/or species interactions, often stated as a dichotomy between the assembly of communities resulting from environmental filtering versus competition. Environmental conditions can include both abiotic conditions and biotic conditions such as habitats or ecosystem types, and interactions span the full range of species interactions not just competition. Analyses of community data such as "RLQ analyses," look at the correlations and partial correlations between species occurrence, environmental conditions and functional traits. Traits often include those related to dispersal (movement) capacity of species, making these essentially metacommunity analyses.

Phylogenetic information as a proxy for evolutionary history is used to exam hypotheses of niche conservatism, environmental filtering, or competition. If niches are conserved, certain refuge habitats would be expected to contain old lineages of species, and local communities are expected to be phylogenetically clustered. Environmental filtering would produce strong associations between certain taxa, with particular traits in certain habitats (or ecosystems); trait x phylogeny correlations are called the phylogenetic signal. Competition is expected to produce communities that show phylogenetic dispersion and dispersion of functional traits because similar species outcompete one another, or exclude each other through other kinds of species interactions. Hence, studying the relationships between communities or species co-occurrence patterns, functional traits, environmental conditions and phylogenetic patterns can help elucidate the mechanisms by which the composition of local communities are maintained. Pillar and Duarte developed methods that look at correlations between species composition, functional traits,

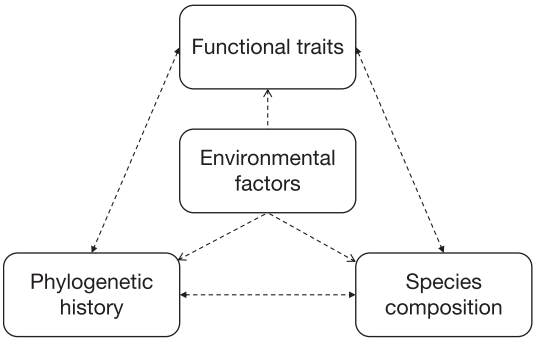


Fig. 2 The relationships between functional traits, evolutionary (phylogenetic) history and species composition, which are all modified by environmental factors.

environmental characteristics and phylogeny at both community and metacommunity scales. Such methods have the capacity to answer questions about community and metacommunity assembly mechanisms in systems which are hard to manipulate to study metacommunity dynamics. There are few applications to date, which limits our current insights.

Food Webs

Historically, food web research has focused on how local deterministic interactions will influence web structure. Considering food webs in a spatial context, however, provides alternative mechanisms for observed phenomena. When species do not interact, we expect that subdivision will only decrease population sizes and make species more vulnerable to stochasticity. Interacting, species, however, can use space as a mechanism of persistence and coexistence that promotes a more complex food web structure.

Ultimately, the degree to which metacommunity dynamics influence food web properties over local community dynamics will depend on dispersal, patchiness, and the interaction between the two. Spatial subsidies of immigrants, energy or materials can stabilize or destabilize local interactions that would otherwise be determined by environmental factors. Patchiness in the form of environmental heterogeneity will allow for site-specific demography and alternative food web states within patches, though the resultant food web structure will also depend on the rate of dispersal of and colonization by both competitors and their consumers. Lastly, patchiness in the form of the size and distribution of habitat patches will influence the extent to which species of varying life histories and dispersal abilities experience subdivision. Together, these factors will influence interactions within and between trophic levels. Where they contribute to the stabilization of local interactions, we expect to see longer food chains or more complex food webs.

Though there is a long history of seeking to explain food chain length with productivity, recent work has shown that ecosystem size may be a more consistent predictor of chain length. This could be a result of metacommunity dynamics. If a food web is under donor-control, the criteria for the existence of higher trophic levels will become stricter for each succeeding level. If basal species have low colonization and high extinction rates, and their habitat is rare, we might expect short food chains because, by virtue of being dependent on lower trophic levels for resources, each succeeding trophic level must have extinction rates equal to or higher than the previous trophic level. Increasing colonization rates or the number of habitat patches, or decreasing extinction should enhance regional mechanisms of persistence and coexistence, stabilize local interactions, and allow for greater trophic diversity.

Spatial structure can also influence the impact of consumers on their resources. A consumer's ability to optimize energy intake by switching prey and feeding location can dampen fluctuations in prey populations, stabilizing local food webs. Even when predators are not actively selecting prey and habitat, space can create refugia for prey populations and, in some systems, decrease top predator density and allow other predators to coexist. Dispersal by consumers can also link resources and habitats in ways that are detrimental for some prey species. For example, shared predation by a mobile consumer may cause two species that never co-occur locally to exhibit apparent competition, or, oppositely, may facilitate the coexistence of two species that otherwise could not coexist locally. All of these factors affect the stability of local interactions and thus the complexity of the food web.

Ecosystem Functioning

Ecosystem functioning reflects biotic effects on the physical and chemical properties of the environment and an emerging theme has been to link this functioning to metacommunity properties. A common theme in research on ecosystem functioning is the diversity-productivity relationship. Non-spatial models generally predict a positive relationship between diversity and productivity due to increasing niche complementarity, with eventual deceleration due to niche overlap (**Fig. 3**A). Though there is experimental evidence to support these predictions, there are also studies that show a neutral, unimodal, or negative relationship between diversity and productivity. Some of these conflicting results may be reconciled through a metacommunity framework that accounts for the spatial exchanges known to occur in real ecosystems.

Varying rates of dispersal between patches may explain a unimodal diversity-ecosystem functioning relationship at local scales and a positive linear relationship at regional scales. At the local level, diversity may increase productivity because of niche complementarity or sampling effects. This relationship may peak and become negative, however, if dispersal allows inferior

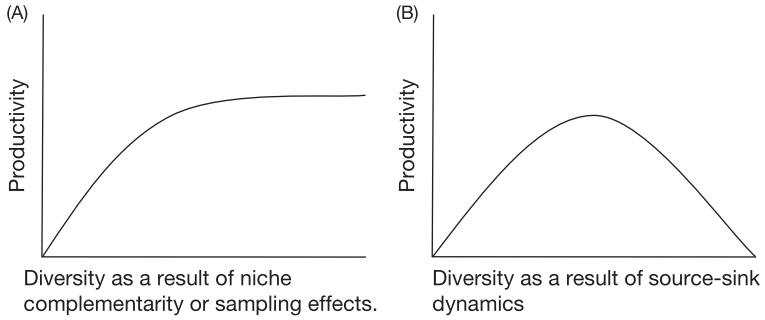


Fig. 3 Theoretical relationships between productivity and species diversity.

competitors to interfere with superior competitors through source-sink dynamics (Fig. 3B). At larger scales, increased diversity may allow species to coexist and complement each other regionally, increasing productivity linearly so long as the number of species does not exceed the number of limiting resources. A study examining productivity at the pond (local) and watershed (regional) scale found this scale-dependent relationship between diversity and productivity in nature, evidenced by low local diversity and increasing species dissimilarity between ponds in watersheds with higher productivity.

A form of spatial insurance could also result in a scale-dependent diversity-productivity relationship. If the dynamics of local communities and environments fluctuate asynchronously, dispersal will supply immigrants to localities where the environment has changed to be unsuitable for the current residents. Increasing regional diversity will increase the likelihood that there will be species appropriate for the various environmental conditions, which will enhance resource use efficiency and productivity. When levels of dispersal are intermediate, supplying enough immigrants to fill vacancies without leading to dominance by a species with intermediate traits, the temporal mean of productivity across the region will be higher and exhibit lower variability. When the system involves more than one trophic level, consumers that track resources can increase variability in productivity at a local scale, but decrease variability at the metacommunity scale by creating a heterogeneous environment that hosts a more diverse community capable of spatial insurance.

Evolution

Evolution influences communities on many timescales. In a metacommunity context, whether local adaptation promotes or deters coexistence will depend on how influential the change is on a given interspecific interaction and the level of immigration among communities.

Over long periods of time, evolution can promote coexistence through niche partitioning, competition-colonization trade-offs or mechanisms that make species regionally similar in fitness despite local competitive asymmetries. In addition to shaping the dynamics of interspecific interactions, evolution will impact the connectedness of a community, and thus the frequency of those interactions, by molding dispersal capabilities and habitat preference.

Empirical evidence also suggests that evolution can influence community dynamics within a short, ecologically relevant timeframe. Asymmetrical gene flow from source to sink populations can prevent efficient natural selection, leading to maladaptation in the sink and creating range limits. This type of source-sink dynamics can also decrease niche breadth and inhibit a species' ability to respond to changing biotic and abiotic environmental conditions. In the same way that intermediate dispersal can promote species diversity, intermediate gene flow can promote genetic diversity by supplying new alleles without extirpating the local genetic framework. If environmental conditions in a patch change, the genetic storage effect can provide immigrant genes adapted to the new conditions that can rescue maladapted populations. Therefore, maintaining genetic diversity on a regional scale promotes overall species diversity.

Summary

Metacommunity models offer the most complete view of the factors maintaining species diversity that has been described to date. The metacommunity concept realistically describes how both local and regional forces can contribute to species diversity, and how the structure of local communities can be altered by immigration from other communities within the region. To date, metacommunity models consist of four broad kinds, termed neutral, patch dynamics, species sorting and mass effects (or source-sink). There is a variety of empirical evidence for species sorting and mass effects models, whereas patch dynamics appear less common, and there is no good example of a system with neutral community dynamics. Often systems display a mix of two kinds of metacommunity dynamics, and further empirical and theoretical work is needed to understand the factors that produce a mix of the four metacommunity paradigms. A metacommunity structure is likely to alter the composition of local communities, food web structure, species abundances, and the potential for evolution. Studying the relationships between community composition, species' functional traits, environmental factors and phylogenetic relationships provides a rich source of information about community assembly at a metacommunity scale. Yet most studies of such relationships have not explicitly recognized the role of spatial structure, although methods exist to do so.

See also: Behavioral Ecology: Dispersal–Migration; Competition. Conservation Ecology: Spatial Subsidy; Biodiversity Indices; Source–Sink Landscape. Ecological Data Analysis and Modelling: Species Distribution Modeling; Modeling Dispersal Processes for Ecological Systems; Grassland Models; Spatial Models and Geographic Information Systems. Evolutionary Ecology: Colonization; Ecological Niche; Coexistence. General Ecology: Community; Biodiversity. Terrestrial and Landscape Ecology: Island Biogeography; Spatial Distribution

Further Reading

Cadotte, M.W., Fukami, T., 2005. Dispersal, spatial scale, and species diversity in a hierarchically structured experimental landscape. Ecology Letters 8, 548–557.
Chase, J.M., Leibold, M.A., 2003. Ecological niches: Linking classical and contemporary approaches. Chicago: University of Chicago Press.
Chave, J., 2004. Neutral theory and community ecology. Ecology Letters 7, 241–253.

Chesson, P., 2000. Mechanisms of maintenance of species diversity. Annual Review of Ecology and Systematics 31, 343–366.
Cottenie, K., 2005. Integrating environmental and spatial processes in ecological community dynamics. Ecology Letters 8, 1175–1182.
France, K.E., Duffy, J.E., 2006. Diversity and dispersal interactively affect predictability of ecosystem function. Nature 441, 1139–1143.
Fukami, T., 2005. Integrating internal and external dispersal in metacommunity assembly: Preliminary theoretical analyses. Ecological Research 20, 623–631.
Hairston, N.G., Ellner, S.P., Geber, M.A., Yoshida, T., Fox, J.A., 2005. Rapid evolution and the convergence of ecological and evolutionary time. Ecology Letters 8, 1114–1127.
Holyoak, M., Leibold, M.A., Holt, R.D. (Eds.), 2005. Metacommunities: Spatial dynamics and ecological communities. Chicago, Illinois: University of Chicago Press.
Hubbell, S.P., 2001. The unified neutral theory of biodiversity and biogeography. Princeton, N.J.: Princeton University Press.
Leibold, M.A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J.M., Hoopes, M.F., Holt, R.D., Shurin, J.B., Law, R., Tilman, D., Loreau, M., Gonzalez, A., 2004. The metacommunity concept: A framework for multi-scale community ecology. Ecology Letters 7, 601–613.
Logue, J.B., Mouquet, N., Peter, H., Hillebrand, H., 2011. Empirical approaches to metacommunities: A review and comparison with theory. Trends in Ecology & Evolution 26, 482–491.
Loreau, M., Mouquet, N., Gonzalez, A., 2003. Biodiversity as spatial insurance in heterogeneous landscapes. Proceedings of the National Academy of Sciences of the United States of America 100, 12765–12770.
Losos, J.B., 2008. Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. Ecology Letters 11, 995–1003.
Mathiessen, B., Hillebrand, H., 2006. Dispersal frequency affects local biomass production by controlling local diversity. Ecology Letters 9, 652–662.
McCann, K.S., Rasmussen, J.B., Umbanhowar, J., 2005. The dynamics of spatially coupled food webs. Ecology Letters 8, 513–523.
McGill, B.J., Maurer, B.A., Weiser, M.D., 2006. Empirical evaluation of the neutral theory. Ecology 87, 1411–1423.
Mouquet, N., Loreau, M., 2002. Coexistence in metacommunities: The regional similarity hypothesis. American Naturalist 159, 420–426.
Pillar, V.D., Duarte, L.D.S., 2010. A framework for metacommunity analysis of phylogenetic structure. Ecology Letters 13, 587–596.
Urban, M.C., 2006. Maladaptation and mass effects in a metacommunity: Consequences in species coexistence. American Naturalist 168, 28–40.
Urban, M.C., Skelly, D.K., 2006. Evolving metacommunities: Toward an evolutionary perspective on metacommunities. Ecology 87, 1616–1626.
Vellend, M., 2006. The consequences of genetic diversity in competitive communities. Ecology 87, 304–311.
Webb, C.O., Ackerly, D.D., McPeek, M.A., Donoghue, M.J., 2002. Phylogenies and community ecology. Annual Review of Ecology and Systematics 33, 475–505.

Metagenomics

Matthew Haynes, Joint Genome Institute, Walnut Creek, CA, United States

Glossary

Assembly Combination of multiple sequences from a metagenome based on overlaps of a defined size.
Community A collection of multiple species in close proximity.
Contig A DNA sequence assembled from smaller overlapping sequences.
Coverage The degree to which a metagenome contains all of the sequence information representing a given genome or gene.
Diversity The degree of complexity of a metagenome based on the number and relative abundance of different sequences.
Library A synthetic collection of DNA fragments that have been processed to facilitate DNA sequencing.
Metadata Information other than DNA sequences associated with a metagenome.
Metagenome A collection of DNA sequences derived from a sample containing a community of organisms.
Operational taxonomic unit (OTU) The genotype classification used to define species based on ribosomal RNA sequences.
Population A collection of organisms of the same species living in close proximity.

Metagenomics is the study of the combined genomes of multiple microbial or viral species within a sample. A metagenome can therefore be regarded as a community genome. The cumulative genetic information of an entire community of microorganisms will often have functional properties that transcend those of any individual species. Metagenomics requires extraction and sequencing of genomic nucleic acids of all organisms in a specific community or biome. Depending on the sample being analyzed, the organisms may be filtered or otherwise separated to yield a metagenome representing only a particular component (e.g., viruses or microbes). Because culturing all organisms within a complex sample is usually impossible, metagenomics requires an uncultured approach. Genomic nucleic acids (DNA or RNA) are directly isolated from a sample and sequenced. High-throughput sequencing is required to characterize a sample of significant diversity. Although obtaining sufficient material for sequencing may be an obstacle, the most challenging aspect of metagenomics is usually the data analysis.

Samples

Metagenomes have been obtained from a wide variety of sources, from seawater to soil to human body fluids and tissues. Procedures such as filtration and centrifugation may be used to fractionate the sample in order to produce a metagenome for microorganisms within a given size or density range. For example, it is often desirable to separate viruses from microbes.

DNA/RNA Isolation

Nucleic acids are isolated directly from the sample; attempts to culture microorganisms from most samples will inevitably result in the loss of genetic information for the vast majority of organisms. Small amounts of DNA may require sequence-independent amplification to produce a sufficient quantity for sequencing. Sometimes this method can alter the original relative abundance of taxa (sequence bias).

Sequencing

Modern high-throughput sequencing methods are a necessity to obtain the very large number of sequences required for accurate assessment of diverse natural communities. High-throughput sequencing technologies currently require the construction of some form of DNA library from the sample DNA to facilitate the sequencing process. In a shotgun library, genomic DNA is fragmented into pieces of a usable size and modified with sequencing primer adapters or other means appropriate for the chosen method. A shotgun library will yield sequences that randomly cover all of the DNA contained within the original sample. In order to sequence metagenomes from a number of samples simultaneously, a multiplex strategy may be used in which DNA from each sample is individually tagged with a short synthetic sequence that is later used to sort the data (Fig. 1). Depending on the complexity of the sample and the relative abundance of individual organisms, whole genomes may or may not be recovered. For ribosomal DNA (rDNA) metagenomes, a polymerase chain reaction (PCR) amplicon is derived from the desired rRNA gene (usually the small

subunit 16S gene), and this amplicon library is sequenced. Sequencing technology is advancing rapidly, and, soon, more sophisticated single-molecule DNA sequencing methods may become the standard.

Special Cases

A different approach is to create an amplicon library, using PCR to amplify specific ubiquitous and conserved sequences such as rRNA genes that can be used as a proxy for microbial species (operational taxonomic units (OTUs)). The typical microbial amplicon library targets some subregion of the 16S rRNA gene, although many other variations are possible. In addition to genomic DNA, viral genomic RNA can be isolated and sequenced after reverse transcription into DNA. Microbial messenger RNA (mRNA) can be isolated from a sample to yield a metatranscriptome that will reflect transcription at the community level.

Data Analysis

Bioinformatics methods are used to analyze metagenomic data. A rarefaction curve (Fig. 2) can be used to determine whether a sample has been sequenced to an extent sufficient to represent its true diversity. The overall diversity of a metagenome can be

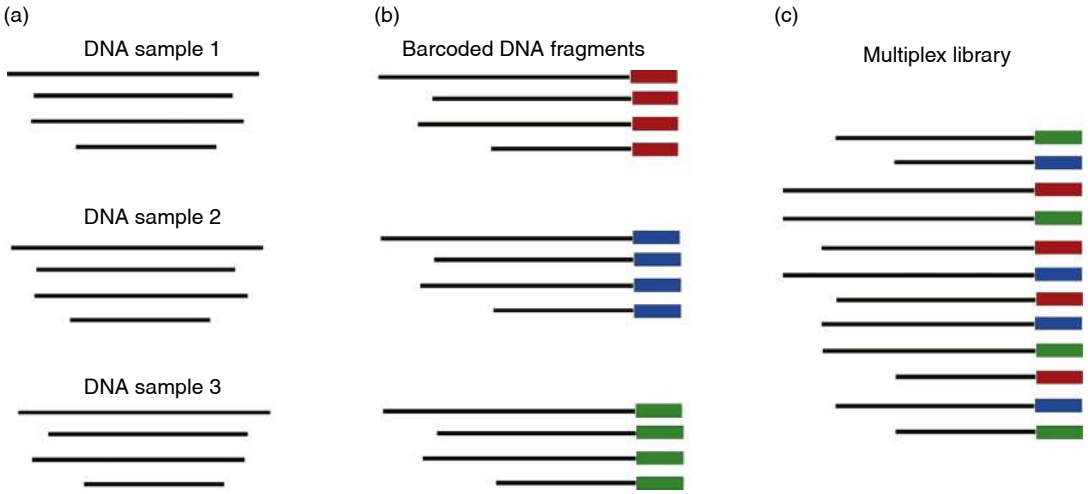


Fig. 1 Overview of multiplex sequencing strategy. (a) Viral DNA is isolated from each sample. (b) Each molecule of the individual DNA samples is tagged with a short sequence during the library preparation for sequencing. (c) Samples are pooled for pyrosequencing, but the sequence tags permit identification of the source of each sequence.

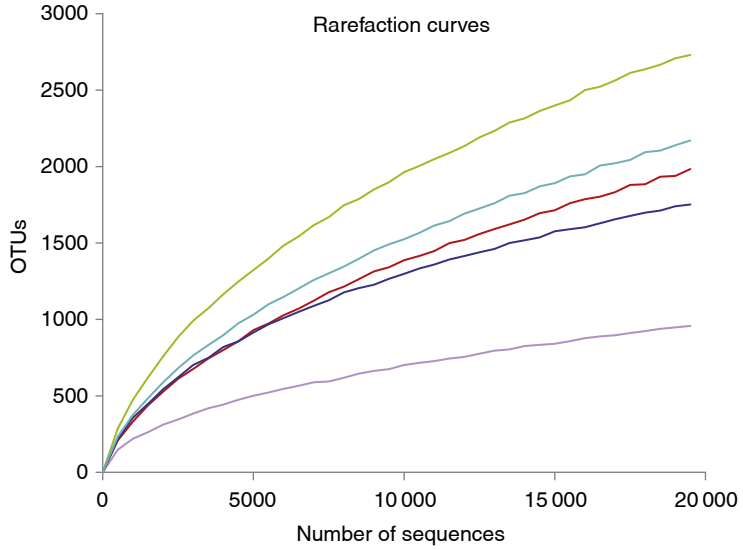


Fig. 2 Rarefaction curves. As the number of sequences from a sample increases, the number of species or OTUs converges on the true diversity.

viewed as the combination of the number of types of sequences present and the relative abundance of these sequences. The sequences in a typical metagenome, reflecting the distribution of species in most natural communities, are present in very unequal abundances. A small number of sequences (species) generally dominate, with a much larger number present in significantly smaller abundances (Fig. 3). There are three levels of classifying diversity: α diversity, which refers to diversity within a sample; β diversity, which refers to variation between two samples; and γ diversity, which is the diversity within a broad geographical area.

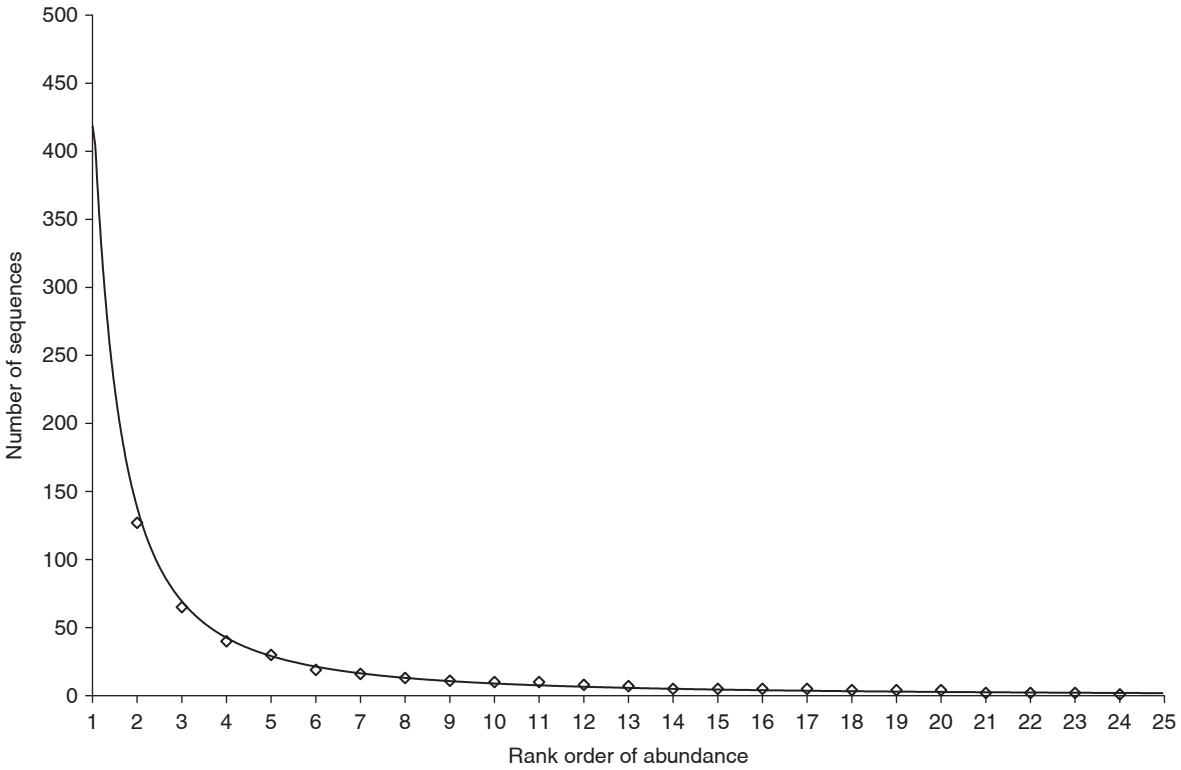


Fig. 3 Rank-abundance curve. A typical metagenome is dominated by sequences representing a few of the most abundant species. A much larger number of lower-abundance species yield only one or a few detectable sequences.

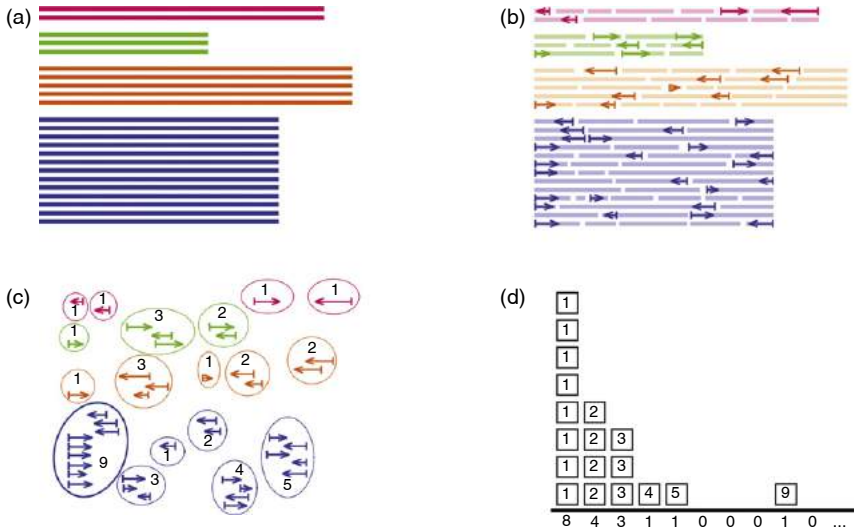


Fig. 4 Contig spectrum. (a) Four species of varying abundance in a sample. (b) The DNA is sequenced randomly. (c) Sequences that overlap are assembled. Sequences that are more abundant produce contigs made up of larger numbers of sequences. (d) Contigs are ranked by the number of sequences they contain (1-contigs, 2-contigs, etc.), and the number of contigs in each category gives the contig spectrum. The contig spectrum is a measure of the diversity of the original sample that does not require identification of any of the sequences.

Two or more sequences can be assembled into larger contigs (contiguous sequences) if the length and similarity of overlaps are sufficient. Sequences may be characterized on the basis of taxonomy or by encoded metabolic potential. Identification of genes that encode particular proteins can be used to create a profile of community metabolic potential. In some cases, the metabolic activity of an entire microbial community is more stable over time than its taxonomic composition. Certain environments or sample types (e.g., viruses) may yield a large number of sequences with no homology to any known sequences, for which taxonomic analysis is therefore impossible. Bioinformatics methods include several approaches that do not require identification by similarity to known sequences. The ability to assemble contigs of various sizes from a metagenome (contig spectrum) reflects the diversity of the metagenome and can thus be used to derive diversity information even when a large proportion of the sequences cannot be taxonomically identified (**Fig. 4**).

See also: Aquatic Ecology: Microbial Communities. Conservation Ecology: Biodiversity Indices. Evolutionary Ecology: Phylogenomics and Phylogenetics; Microbiomes and Holobionts. General Ecology: Soil Ecology; Rhizosphere Ecology. Global Change Ecology: Biogeocoenosis as an Elementary Unit of Biogeochemical Work in the Biosphere; Microbial Cycles

Further Reading

Committee on Metagenomics: Challenges and Functional Applications, National Research Council, National Research Council, 2007. The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet. Washington, DC: National Academies Press.
Marco, D. (Ed.), 2010. Metagenomics: Theory, Methods and Applications. Norfolk: Caister Academic Press.
Marco, D. (Ed.), 2011. Metagenomics: Current Innovations and Future Trends. Norfolk: Caister Academic Press.
Nelson, K.E. (Ed.), 2011. Metagenomics of the Human Body. New York: Springer.
Wolfgang, R.R.S., Rolf, D. (Eds.), 2010. Methods in Molecular Biology, vol. 668: Metagenomics: Methods and Protocols. New York: Humana Press.

Relevant Websites

http://www.camera.calit2.net—CAMERA.
http://www.genome.jp—GenomeNet.
http://www.theseed.org—Home of the SEED - TheSeed.
http://metagenomics.anl.gov—MG-RAST.
http://dels-old.nas.edu—National Academy of Sciences.
http://www.ncbi.nlm.nih.gov—NCBI.

Microbiomes and Holobionts

Derek Skillings and Katarzyna Hooks, University of Bordeaux, Bordeaux, France

Glossary

16S rRNA Universally conserved bacterial ribosomal RNA gene. 16S rDNA is the most commonly used marker in microbial diversity studies.
Holobiont A host plus all of its symbiotic microbiota.
Hologenome The sum total of all genetic material from all members of a holobiont.
Metabolome Profile of existing intermediate and end products of metabolic pathways within a system under a given set of conditions.
Metagenomics The study of the combined genetic material extracted from environmental samples.
Metatranscriptome All RNA molecules, including mRNA, rRNA, tRNA, and other noncoding RNA transcribed in a population.
Microbiome (1) All the microbes in a community. More recently, (2) all of the genes of all of the microbes in a community.
Microbiota The community of commensal, symbiotic, and pathogenic microorganisms in a particular site, habitat, or geological period.
Mycobiome All the fungi in a community.
Symbiosis Two or more species living closely together in a long-term relationship.
Virome Collection of viral nucleic acids in a community.

Introduction

Microbes, or microorganisms, include microscopic organisms such as bacteria, archaea, protozoa, and unicellular fungi and algae. This is sometimes extended to include viruses. "Microbiome" started as primarily an ecological term, defined as a characteristic microbial community occupying a reasonably well-defined habitat, which has distinct physiochemical properties. It referred to the microorganisms, their activities and their immediate environment. The "-ome" in "microbiome" was related to the term "biome." Later usage started to distinguish the microbiota (the organisms) from the microbiome (their collective genome). Here the "-ome" is related to "genome" and "chromosome." The result is that current practice refers to both organisms and their molecular constituents as microbiomes. The methodology for studying such systems is similar and, thus, we define microbiome research as the analysis of the aggregated molecular components of a defined microbial community. When used in its most inclusive sense, the microbiome might also include the resident virome and mycobiome of the community under study. The virome is the collection of viral nucleic acids in a community, including both bacteriophages (bacterial viruses) and DNA- and RNA-encoded eukaryotic viruses, and the mycobiome is the microscopic fungal component of the community.

History

From Cultures to Sequencing

Microbiome research as it exists today is a recent development in the life sciences. It combines techniques from both microbiology and molecular genetics and has applications that go beyond the traditional domain of microbiology. Microbiology started as a laboratory-based science that focused on isolating microorganisms in pure cultures. This methodology dominated microbial studies because of the focus on isolating pathogens or other undesirable organisms. Microbial causality of disease states or food spoilage has been historically established using Koch's postulates. This procedure establishes a sequence of isolating, culturing and experimenting in hosts. Four criteria must be met in order to fulfill Koch's postulates and establish a microorganism is causal: (1) the microbe must be present in all sick individuals and absent from all healthy individuals; (2) the microbe must be isolated from sick individuals; (3) a pure culture of that microbe must cause the illness when administered to a healthy individual; and (4) the microbe must be re-isolated from the sick test individual and shown to be identical to the originally isolated microbe.

Koch's postulates have been adapted numerous times since their introduction in the late 19th century. Later techniques were imported from biochemistry in order to identify the biochemical pathways or important molecules that were produced by isolated cultures of microbes and responsible for the effect of interest. When successful, this narrowed the causal attributions from individual organisms to particular toxins or other molecular compounds. A biochemical approach was also used to explore the signature of microbial activity in natural environments outside of pure cultures in the lab. Environmental microbiology was transformed by the advent of molecular genetics. DNA sequencing allowed for the identification of microbes without culturing in isolation. The first studies focused on the genes for ribosomal RNA (rRNA), which had been established as a useful marker for phylogenetic analyses. This technique, when applied to environmental samples, uncovered an unprecedented amount of previously unrecognized microbial diversity.

Metagenomics

Improvements in sequencing technology and large-scale data analyses caused a shift from single genes to whole microbial genomes. Metagenomics started as the whole-genome sequencing of samples extracted from natural environments. The first use of the term "metagenome" was in 1998 by Jo Handelsman and colleagues, in reference to the combined genomes found in an environmental soil sample. The shift here was to treat the collection of genes recovered from an environmental sample in a way that is analogous to the study of a single genome. Metagenomics tells who is there, and early whole microbiome research was made up of primarily of microbial composition studies. As the field developed it started determining functionality and what the microbes are doing by looking at gene expression profiles and metabolite production (metatranscriptomics and metabolomics, respectively). The next step has been to identify how the microbes interact each other when in their natural habitats.

Holobionts and Hologenomes

It was becoming increasingly clear that the living world is dominated by, and completely dependent on, complex microbial communities. Evidence was also mounting that most microbes do not live independently, they interact in interconnected networks of communication, cross-feeding, genetic recombination, and coevolution. These interactions extend to the level of multicellular organisms, suggesting that multicellular organisms have been engaged in symbiotic relationships with microorganisms throughout their evolutionary history. It was long thought that all macroorganisms are routinely colonized by a large number of microorganisms, but the details and extent of macrobe–microbe interactions remained difficult to uncover. As the molecular tools for identifying microbial diversity grew, attention turned to multicellular organisms as sites of new and diverse microbial ecosystems with multiple environmental niches.

It is now widely accepted that microorganisms have always played many important roles in the lives of plants and animals. Symbiotic interactions between microbes and multicellular organisms have been documented across a diverse swath of life, and many researchers maintain that all large organisms engage in symbiotic interactions with microbes in natural settings. The term "holobiont" was coined by Lynn Margulis and used to refer to symbiotic associations that last throughout a significant portion of an organism's lifetime. It is derived from the Greek word *holos*, which means whole. The term first found wide usage in coral biology where it was defined as a coral colony and its associated photosynthetic algal symbionts and bacterial communities. The recent massive influx of interest in host-microbiome relations has led to a proliferation of the term "holobiont," now most often understood as a host and *all* of its associated microbiota, including bacteria, archaea, viruses, protists, fungi, and microscopic multicellular animals such as nematodes. Because the holobiont includes *all* associated microbiota, the interactions between holobiont partners may be harmful, beneficial or of no consequence.

The term "hologenome" originated more recently. It is defined as the collective unit made up of all of the host and microbial genomes of the holobiont, or in other words, the host genome plus the resident microbiome, virome, mycobiome and all other genomic material associated with the host. Hologenome is often used in place of the more familiar host + microbiome when researchers are trying to emphasize the interaction and integration of host and microbial genes in the formation of host traits or phenotypes.

Although there have been significant recent advances in our ability to understand the microbial world, converting microbiome data into meaningful biological insights remains very challenging. Barriers to a more thorough understanding include the difficulty in decoding the functional relevance of microbiota at appropriate scales, the difficulty in establishing causality in complex microbial networks, and our limited ability to make informed manipulations that lead to predictable outcomes in natural systems.

Methodology

Sampling Strategies

Microbiota research is interested in two general types of microbial communities: environmental and host-associated. Some of the microbes and their functions in the natural environments are known and well described (e.g., decomposition of organic matter and cycling of nutrients, especially nitrogen). There is an increasing interest in the built environments (buildings, vehicles, water systems, etc.) and their bacterial composition. Host-associated microbiota studies range from investigating microbes found in and on agriculturally important plants, marine invertebrates and different human body sites.

Early Methods

Traditionally identification of most members of a complex microbial community was challenging. It was especially true for the gut microbiome—a highly diverse and densely populated microbial community with only a small percentage of microbes that could be cultured.

Culture and culturomics

The early microbiome studies were stemming from traditional microbiology and involved pure culture or co-culture methods. Many bacterial species were excluded from such analyses since the optimal conditions for their growth were unknown. Cultivation-based approaches to study microbial communities are still important today, especially in the form of microbial culturomics (see Lagier et al., 2016). However, since culture methods are performed outside of the ecological environment, they fail to inform us about community dynamics and species interactions.

Hybridization methods

One of the first culture-independent methods for microbiome study relied on hybridization of nucleic acids, either in the form of in situ hybridization or Southern blotting. Usually the probes were designed to recognize 16S rDNA with specificity varying from the whole kingdom to species level. In situ hybridization is still in use in the field of environmental microbiology to visualize the spatial distribution of species (e.g., in biofilm).

Amplification

Amplification methods such as PCR and qPCR are able to detect and quantify particular taxa when taxa-specific primers are used. PCR can be also used to create a library, for example, by using a universal primer for 16S or 18S rRNA. Such libraries then can be used as an input for fingerprinting methods or more recently tag sequencing (see below). In both cases the success of the method depends on the specificity and sensitivity of the primers used and the prior knowledge of the target sequences.

Fingerprinting methods

Most popular fingerprinting methods were terminal-restriction fragment length polymorphism (T-RFLP), denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE). They allowed comparisons of the whole communities but offered little information regarding taxonomic composition driving the change. Sometimes low-throughput sequencing was used to complement the DGGE or TGGE analysis, where visibly differing bands were excised, sequenced and compared with bacterial sequences in the databases.

State of the Art Technology

Initially sequencing technology was slow and expensive. It involved cloning interesting fragments of DNA, like gene encoding 16S rRNA, cloning it into plasmids, transforming suitable hosts (usually *Escherichia coli*) and sequencing by Sanger method. However, even then it allowed for more precise and rapid taxonomic identification of bacterial communities than culture-based methods. Real revolution came with the advent of massively parallel next-generation sequencing and other –omics technologies (**Fig. 1**).

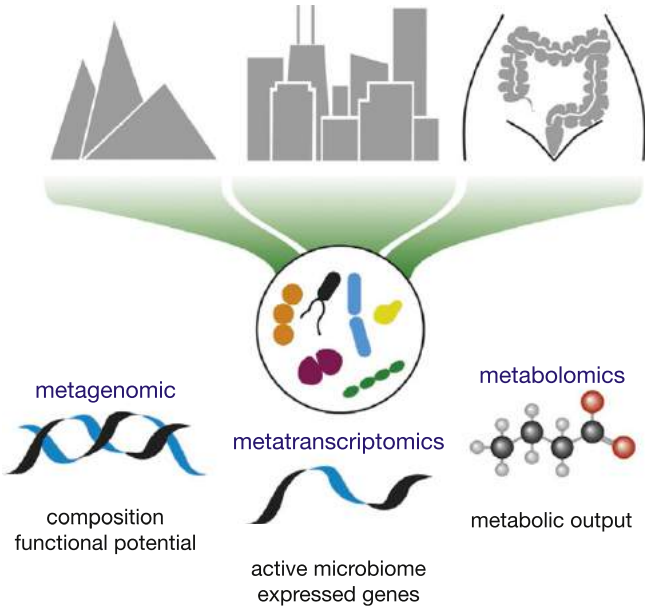


Fig. 1 Microbial community samples can be taken from natural, built, or living environments. Different analyses target different molecular components of the sample and provide complementary information on the identity and activity of the microbes in the sample.

Tag sequencing

In tag sequencing total DNA is isolated from samples, regions within universally conserved 16S/18S rRNA genes are amplified by PCR and amplicons are sequenced by one of the high-throughput methods. Identification of variations in specific regions of this gene allows for classification of bacterial taxa. Popular pipelines for bioinformatics analysis of 16S sequencing data include Quantitative Insights Into Microbial Ecology (QIIME) and Mothur.

With the ability to enrich templates with initial concentrations too low to detect, PCR-based sequencing technology allows identification of individuals with very low populations within complex communities. Tag sequencing is relatively inexpensive to perform but has limited resolution: not all bacteria can be classified and species-level identification may not be possible, nor can it detect viruses and eukaryotic communities. Questions are also raised about the biases introduced by the initial PCR step.

16S rRNA sequencing has generated a wealth of data on microbiome composition from different environments and conditions. Using this method the Human Microbiome Consortium published in 2012 reference metagenomes of microbes present within healthy humans.

Shotgun sequencing

Metagenome shotgun sequencing (MGS) extends the information provided by 16S/18S rDNA amplicon sequencing. It uses the entire DNA available in the microbiome sample to prepare a sequencing library. Additional amplification step might be necessary for environmental samples yielding small DNA amount. The resulting library can be sequenced either with the use of short (2×250 or 2×300 nt) or longer reads (500–4000 nt). Number of reads mapping to marker genes can be then used to quantify specific taxa and bacterial strains. Multiple different algorithms exist for metagenome assembly. There are also dedicated pipelines like EBI Metagenomics, CAMERA or MEGAN 6 performing the analysis and visualization of the MGS data.

MGS allows identification of not only bacterial, but also viral, fungal, and protozoan DNA. It produces a much better resolution of bacteria at the species level and allows for annotation of bacterial gene clusters and pathways based on direct sequencing of bacterial genes. The downsides of MGS are higher sequencing costs, higher bioinformatics load due to the large number of sequence reads produced, and the difficulty to analyze genomes absent in the reference databases or genes with unrecognized function. Contamination by host DNA is another challenge in MGS when biopsy or mucosal material is being collected.

Metatranscriptomics

Instead of studying the DNA present in the sample it is also possible to study the activity of microbes within complex communities by metatranscriptomics. Instead of describing the potential of metagenome to encode certain pathways, RNA sequencing (RNAseq) directly provides information about the bacterial gene expression. Such studies can complement compositional data and have a potential to show differences in gene expression before the compositional changes can be observed through metagenome analysis.

Metabolomics

To better describe the bacterial community, the totality of their metabolic products can be described by metabolomics analyses. Such information extends the knowledge about the functions performed by the microbes by measuring the quantities of products of their pathways. The most frequently used metabolomics methods are spectrometry (MS) and nuclear magnetic resonance (NMR). Further advances were offered by Matrix-Assisted Laser Desorption Ionization time of flight (MALDI-TOF), Secondary Ion Mass Spectrometry (SIMS), and Fourier transform ion cyclotron resonance MS that greatly improved the accuracy and throughput in metabolomics.

Extensive microbiome profiling led to saturation of information about the bacterial composition in certain environments, for example, human gut. In contrast, metabolomics profiles are not well described even for the routinely grown cell cultures exposing the gap in the knowledge about complex biological systems. Considering that the biosynthesis of oligosaccharides and lipids allows for many permutations in combining simpler molecules, the number of metabolome components is at least an order of magnitude higher than that of the protein-coding genes.

Data Analysis and Bioinformatics

The new high-throughput technologies generated a wealth of data that require a careful analysis. With the constantly decreasing price of sequencing the bottleneck lies now in the storage, analysis and interpretation of the results.

Typical microbiota study involves multiple steps of data analysis after sequencing. First the resulting reads undergo a quality checking and filtering. If the tag sequencing was performed, the reads are usually clustered into the Operational Taxonomic Units (OTUs)—a group of individual bacteria whose sequences are at least as similar as a chosen threshold (frequently 99% or 97% for species-level resolution when considering variable rRNA regions). OTUs can then be attributed to specific species or a group of species and used to describe the composition of the sample. Statistical methods borrowed from the ecological studies allow performing alpha and beta diversity analyses on the microbial composition matrices. Bioinformatics methods aid in the exploratory data analysis and visualization of taxonomic composition. They also make it possible to predict the genes present in

the microbiome using the taxonomic assignments, for example, with the Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICRUSt).

The bioinformatics analysis of the metagenome shotgun sequencing is more complex but it can also yield much more robust results. With the information about the all DNA present in a sample, it is possible not only to find taxonomic composition of the sample but also describe the functionality of microbes within the community by comparing identified genes and predicted proteins from sequence data to databases.

With different big-scale microbiome projects completed, additional bioinformatics analyses and improved statistical methods can lead to new discoveries overcoming the spurious conclusions of small studies. Comprehensive meta-analyses combining in a principled manner taxonomic and functional profiling results can lead to meaningful synergy.

Mouse Models and Causal Inference Via Whole-Community Transfers

One of the experimental methods of manipulating and investigating the whole microbiomes has to do with the use of germ free rodents. Such animals are born through caesarian section and kept in the sterile environment including feeding with irradiated food and water that prevents them from being colonized by any bacteria. Germ free mice can then be infected with a specific microbiota—either from other mice, humans or artificially constructed communities. The microbiota transfer can happen in two ways—by relying on the mice coprophagy or by directly administrating microbiota orally through gavage feeding. This technique allows also observing in vivo effect of difficult to grow bacteria, interaction between microbes and host and within microbial communities. Researchers described physiological differences in the host after the administration of altered microbiomes. In particular, transfer of the whole organism phenotype like obesity, suggests that microbiota might be the cause of such changes.

Research Areas Associated With Microbiome Studies

Biogeochemistry

Microorganisms are the biogeochemical engines that support all life on earth. Phytoplankton (single-celled algae and photosynthetic bacteria) in the earth's oceans are major drivers of the carbon cycle, together responsible for half of the global atmospheric carbon that is fixed each year. Microbes are also primarily responsible for stabilizing and recycling the fixed carbon into terrestrial carbon stocks. Microbes are also responsible for nitrogen fixation. This process is so valuable that many plants and insects have formed long-lasting symbioses with nitrogen-fixing bacteria. Beyond being the backbone of the carbon and nitrogen cycles, microbes are also important for controlling and stabilizing the flux of important nutrients like nitrogen and phosphorus, leading to stable environments and complex ecosystems. How that is done is often unknown. The functions of most microbial genes are still mysterious, as are the products of those genes, the functions of those products, and how their production is regulated in natural environments. Large-scale high-throughput microbiome studies are necessary to unravel the microbial diversity and biochemical pathways underpinning these important biogeochemical processes.

Microbial Ecology

Microbial ecology aims to determine the patterns and drivers of microbial community distribution, interaction, and assembly. It has been well demonstrated that microbial community composition changes across most environmental gradients, such as geographic distance, nutrients, temperature, moisture levels, salinity, oxygen availability, pH, and day length. Broader patterns have been difficult to track because of the necessarily small-scale of individual studies. Initiatives like the Earth Microbiome Project bring together regional studies that tend to focus on one type of environmental sample in one locality in order to investigate microbial community dynamics at larger regional and global scales. Environmental microbiome samples from diverse environments are paired with metadata consisting of spatial, temporal, physiochemical, and animal/plant association variables. Initial studies looked at covariation between microbial diversity, measured in terms of operational taxonomic units (OTUs) clustered by single-gene sequencing (usually 16S rRNA), and environmental variables. These approaches have expanded to take advantage of whole-genome methods, strain-level diversity, and more fine-grained sampling. Initial results show a general pattern of hierarchical organization, with less-rich communities nested within richer communities at higher taxonomic levels.

Host Biology and Holobionts

Host-microbial symbioses are common in nature and have a long research history. Until recently, little was known about the details of these relationships and research was usually limited to the study of pairwise associations in a few familiar cases, such as in termites, ruminants, corals, and legumes. Metagenomics studies on the internal environments of eukaryotic organisms soon uncovered that nearly all visible organisms are hosts to a wide assortment of microbes. Furthermore, these microbes often have an influence on the physiology, anatomy, behavior, reproduction, and a fitness of their hosts. The ubiquity of host-microbiome interactions and the intimate ways by which microbes are integrated into host biology is what motivates thinking of the entire system as a holobiont—a new unit of biological organization.

Interactions between hosts and their microbes vary across the entire spectrum of possibilities. Relationships between partners may be coevolved or opportunistic, competitive or cooperative, obligatory or facultative. Some partner interactions are best considered as symbioses—ranging from mutualism to parasitism—where the partners mutually form a part of each other's environments. Other interactions long ago bound the individual lineages together into a higher-level lineage. This is what happened during eukaryogenesis—the origin of mitochondria and the birth of all eukaryotes. This new lineage was formed when an archaeon enveloped but did not consume its bacterial prey and the two continued to live together and reproduce as one higher-level unit.

Holobionts are an interesting level of biological organization because they can share features of both organisms and communities. Focusing on the processes, interactions, and relations that occur between holobiont partners has opened up a suite of research questions. What kinds of interactions and traits lead to increased metabolic integration and alignment of fitness? To what degree are holobiont partnerships species-specific coevolved consortiums versus generalist assemblages taking advantage of leaky products or stable environments? What is the relationship between different biological parameters: mode of transmission versus alignment of fitness versus degree of metabolic integration?

The Human Microbiome

Humans are associated with up to 10^{14} microorganisms. The vast majority of the microorganisms that make up the human microbiome are bacteria residing within the gastrointestinal tract. The human microbiome literature, in turn, tends to focus exclusively on bacteria. Many gut microbes can now be cultured in special conditions, but very little is known about what these microbes are doing and how they are interacting, especially regarding the anaerobes. After the success of the Human Genome Project, work began on using metagenomics tools to characterize the "the second human genome," the microorganisms making up the human microbiome. One such initiative is the Human Microbiome Project (Human Microbiome Project Consortium, 2012). By the numbers, it is likely that the largest portion of contemporary microbiome research is on the human microbiome, and in particular, the gut microbiome.

Although it was known that there can be extremely fine-grained microbial diversity at the species and strain level, much of the early human microbiome work was done at the coarse phylum level, mainly due to technical and database limitations. Most of the guts of healthy human so far sampled are dominated by only two phyla: Firmicutes and Bacteroidetes. Together they comprise about 90% of human microbial gut diversity. Based on this information it was assumed at the start of the Human Microbiome Project that humans would share a large core of microbial lineages, and there would a large diversity of rare lineages that make each person unique. This picture of the human microbiome will probably end up being wrong. The first 5 years of results using higher taxonomic resolution have consistently found that the differences between individuals are substantially greater than the differences within an individual at different environments along the gut. Other studies have routinely shown that healthy Western adults can be more than 90% different in terms of gut microbes at the species level. Even with the dominance at the coarse-grained taxonomic level, some studies have suggested that no bacterial species is present within 100% of subjects.

What explains the dominance of Firmicutes and Bacteroidetes then? It is likely because microbial diversity at the taxonomic level does not neatly map on to functional diversity. Microbiomes may be very diverse in terms of which species are present, but less diverse in terms of which higher-level taxa and metabolic functions are present. Different microbial communities appear to often converge on similar functional profiles.

The Human Microbiome and Its Relationship to Human Health

Once the composition of a host-associated microbiome has been roughly described and categorized it is often employed to look at links between composition and host function. Host-associated microbiomes have been shown to play key roles in the development, metabolism, homeostasis and immunity of the host. Particular compositional patterns have also been associated with disease, obesity, diabetes, cancer, and allergies. Most human microbiome research targets the links between the microbiome and human health or development in some way. Although these comparisons are done for numerous body niches, the colon and lower gut are the focus of much of this research. This is likely due to the ease of access to the gut microbiota, which is commonly done by proxy via fecal samples.

Some unhealthy states are correlated with signatures in the microbiome, and thus the composition of the microbial communities might serve some important diagnostic and predictive purposes. Numerous studies have shown that an elevated proportion of Firmicutes is associated with host obesity and the change in metabolism accompanying it. Crohn's disease is associated with a significant decrease in Bifidobacteriaceae populations and an increase in groups containing potential pathogens. A wide range of other diseases, including brain and behavioral disorders has now been associated with changes in microbiota diversity. Phylum-level patterns associated with conditions such as human obesity have been described as "highly conserved bacterial traits" that affect host phenotype. Many of these association studies do not, however, go on to examine any causal implications very thoroughly (i.e., whether changes in composition are causes or effects of disease states). There have been also attempts at genome-wide association studies pinpointing the genetic and nongenetic host factors influencing the microbiota. In order to effectively understand the impact of the microbiome on the host, it is critical to connect compositional and correlational studies to functional studies and the causal mechanisms underlying those functions.

(Micro)Biological Engineering

Microbes can live in nearly every environment on Earth and produce enzymes for every major biochemical transformation of organic or inorganic matter on the planet. The global microbiome is a reservoir of billions to trillions of genes that could be drawn upon to build valuable pathways and products. There is a long history of harnessing the wild capabilities of microbes to generate desirable products, from bread and beer to antibiotics and fuel. Microbiome methods have opened up the search for microbes, in isolation or in consortia that can synthesize compounds with environmental, industrial, and pharmaceutical value. Isolating and identifying the functions of single genes and biosynthesis pathways on large scales is also necessary for the construction of synthetic microbes. Microbe-based fuel manufacturing holds considerable promise for the future generation of alternative fuels with less waste and fewer toxic by-products.

Controversies

Dysbiosis

Gut microbiota has been investigated for many healthy and diseased people and significant differences have been observed in the composition of their bacterial communities. Because such remarkable changes cannot be always explained by specific taxa, the idea of a whole-community dysfunction was proposed instead. Such unusual states of the microbiome are sometimes referred to as "dysbiosis," a very loosely defined change in the microbial composition. Many microbiota researchers criticized the term due to its ambiguity, difficulties in defining the reference healthy or normal state and the uncertainty if observed dysbiosis is a cause or effect of the disease (see Hooks and O'Malley, 2017). More worryingly dysbiosis is sometimes reported as a result of a study rather than taken as a sign that a further research is needed to elucidate the mechanisms, by which the microbiota is interacting with and responding to the altered host environment.

The Hologenome Theory of Evolution

Recognition that holobionts are common in nature has led many researchers to reassess their views about various processes and concepts that are foundational in biological thinking (Skillings, 2016 has an overview of this debate). One suggestion is that there is now a need to upgrade fundamental theories about the action of natural selection because holobiont systems "raise the discussion of individuality and organismality beyond its historical perspective to a level that challenges and extends current thinking" (Theis et al., 2016). Several researchers now claim that holobionts, or similar multilineage assemblages of macrobes and microbes, constitute at least one level of organization at which natural selection acts. It is unlikely that there is any holobiont that is also a unit of selection or organism if the holobiont is defined as a macrobe host and *all* of its associated microorganisms. It is not impossible that a host and its symbionts could form a unit of selection, it is just that the conditions are unlikely to obtain. High partner fidelity and alignment of fitness are necessary. This is achieved by vertical inheritance or by strong mutual partner choice. Such high-fidelity associations are unlikely to occur across all of the partnerships within a holobiont. Where it does not, selective pressures at the level of the individual lineages will tend to put the partners into direct competition or active exploitation. Neither reducing the holobiont to a set of pairwise interactions between symbiont partners nor treating the entire community as a single biological individual is a universally appropriate approach for holobiont research.

See also: Aquatic Ecology: Microbial Communities. Conservation Ecology: Biodiversity Indices. Ecological Data Analysis and Modelling: Forest Models. Evolutionary Ecology: Units of Selection; Metagenomics; Phylogenomics and Phylogenetics; Association. General Ecology: Parasites. Global Change Ecology: Biogeocoenosis as an Elementary Unit of Biogeochemical Work in the Biosphere; Microbial Cycles

References

Hooks, K.B., O'Malley, M.A., 2017. Dysbiosis and its discontents. MBio 8.e01492-17
Human Microbiome Project Consortium, 2012. Structure, function and diversity of the healthy human microbiome. Nature 486, 207–214.
Lagier, J.C., Khelaifia, S., Alou, M.T., *et al.*, 2016. Culture of previously uncultured members of the human gut microbiota by culturomics. Nature Microbiology 7, 16203.
Skillings, D., 2016. Holobionts and the ecology of organisms: Multi-species communities or integrated individuals? Biology and Philosophy 31, 875–892.
Theis, K.R., Dheilly, N.M., Klassen, J.L., *et al.*, 2016. Getting the hologenome concept right: An eco-evolutionary framework for hosts and their microbiomes. mSystems 1. e00028-16

Further Reading

Blaser, M.J., Cardon, Z.G., Cho, M.K., *et al.*, 2016. Toward a predictive understanding of the microbiome. MBio 7.e00714-16.
Dorrestein, P.C., Mazmanian, S.K., Knight, R., 2014. Finding the missing links among metabolites, microbes, and the host. Immunity 40, 824–832.

Goodrich, J.K., Di Rienzi, S.C., Poole, A.C., *et al.*, 2014. Conducting a microbiome study. Cell 158, 250–262.
Handelsman, J., Rondon, M.R., Brady, S.F., *et al.*, 1998. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. Chemistry & Biology 5, R245–R249.
Kurilshikov, A., Wijmenga, C., Fu, J., Zhernakova, A., 2017. Host genetics and gut microbiome—Challenges and perspectives. Trends in Immunology 38, 633–647.
Lloyd-Price, J., Mahurkar, A., Rahnavard, G., *et al.*, 2017. Strains, functions and dynamics in the expanded human microbiome project. Nature 550, 61–66.
McFall-Ngai, M., Hadfield, M.G., Bosch, T.C., *et al.*, 2013. Animals in a bacterial world, a new imperative for the life sciences. Proceedings of the National Academy of Sciences of the United States of America 110, 3229–3236.
Thompson, L.R., Sanders, J.G., McDonald, D., *et al.*, 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. doi:10.1038/nature24621.
Zhou, J., He, Z., Yang, Y., *et al.*, 2015. High-throughput metagenomic technologies for complex microbial community analysis: Open and closed formats. MBio 6.e02288-14

Relevant Websites

www.earthmicrobiome.org—Earth Microbiome Project.
www.ebi.ac.uk/metagenomics—EBI Metagenomics service.
www.metahit.eu—Metagenomics of the Human Intestinal Tract.
nas-sites.org/builtmicrobiome—Microbiomes of the Built Environment.
www.mothur.org—Mothur microbiome bioinformatics software.
www.hmpdacc.org—NIH Human Microbiome Project.
qiime2.org—QIIME microbiome bioinformatics platform.

Natural Selection

Kent E Holsinger, University of Connecticut, Storrs, CT, United States

Darwin, Wallace, and the Theory of Natural Selection

Charles Darwin (1809–82) spent nearly 5 years (December 1831 to October 1836) as a naturalist aboard the *Beagle* during its expedition around the world, in addition to spending many years studying natural history in southeastern England. Alfred Russel Wallace (1823–1913) spent 4 years in the tropical forests of Brazil and 8 more years in the forests of southeastern Asia. Their years of work and that of many biologists who preceded them revealed many examples of plants and animals whose physiology and habits made them well adapted to their environment. Darwin had completed the first draft of a book on species and speciation when he received a letter from Wallace including the draft of a paper entitled 'On the Tendency of Varieties to Depart Indefinitely from the Original Type'. In this paper, Wallace described the theory of natural selection, a theory Darwin had independently discovered 20 years earlier but was only now preparing for publication. In 1858, Wallace's paper was published in the *Journal of the Linnean Society* together with extracts from an essay Darwin wrote in 1844 but never published. One year later, Darwin's *On the Origin of Species* appeared.

The theory of natural selection that Darwin and Wallace presented is beguiling in its simplicity, yet it is sufficient to explain the many intricate adaptations of plants and animals.

(1) In every population of every species on Earth, more individuals are born than survive to reproduce.
(2) In most populations, individuals differ from one another in characteristics that cause them to differ in their chances of survival, in the number of offspring they produce if they survive to reproduce, or in both.
(3) Offspring tend to resemble their parents.

From these simple observations follows the obvious conclusion: Any characteristic that increases an individual's chance of survival or its fecundity will tend to become more common. Similarly, any characteristic that lessens an individual's chance of survival or reduces its fecundity will tend to become less common. Thus, individuals will tend to be well adapted to the circumstances in which they find themselves.

Mutations introduce variation into populations that may lessen the chance that some individuals survive. Genetic drift may have a greater influence than natural selection on the transmission of genetic variation from one generation to the next in small populations. A population may not have had time to adapt to recent changes in its environment, or its environment may be constantly fluctuating so that no single characteristic is always favored. For these and other reasons, organisms are not perfectly adapted to their environment. But the process of evolution by natural selection guarantees that most characteristics of most organisms will be well suited to the conditions in which they are found most of the time.

Genetic Consequences of Natural Selection

The genetic consequences of natural selection are the easiest to understand if we study how allele frequencies at one locus with two alleles change when genotypes differ in their probability of survival. An individual's fitness is its contribution to the composition of later generations, relative to the contribution of other individuals in the same population. Fitness differences may arise because individuals differ in their probability of survival, in their ability to find mates, in the number of offspring they produce when mated, and in many other ways. Differences in probability of survival are the easiest to understand. Fortunately, many of the genetic consequences of natural selection do not depend on whether fitness differences arise from differences in survival probability or from differences in some other component of fitness.

Suppose an individual with genotype A_1A_1 survives to reproduce with probability w_{11}, and suppose that the survival probabilities for genotypes A_1A_2 and A_2A_2 are w_{12} and w_{22}, respectively. If individuals choose their mates at random, then genotypes in newly formed zygotes will be found in Hardy–Weinberg proportions. So if the frequency of allele A_1 is p and that of allele A_2 is q, we can calculate the frequency of the three genotypes in zygotes and adults as follows in **Table 1**. $\overline{w}$, which is equal to $p_i^t w_{11} + 2p_t q_t w_{12} + q_i^t w_{22}$, is known as the mean fitness. It is the average survival probability in the population. From the adult frequencies in the last row of **Table 1**, we can calculate the allele frequency among newly formed zygotes of the next generation, namely,

$$p_t + 1 = \frac{(p_t^i w_{11} + p_t q_t w_{12})}{\overline{w}}$$

Suppose that the frequency of the A_1 allele in newly formed zygotes is 0.4 and that $w_{11}=0.9$, $w_{12}=0.8$, and $w_{22}=0.7$, then the above equation allows us to predict that the frequency of A_1 in newly formed zygotes of the next generation will be 0.43. Now

Table 1 .

Genotype	A_1A_1	A_1A_2	A_2A_2
Zygote frequencies	p_t^i	$2p_tq_t$	$q_{t>}^i$
Probability of survival	w_{11}	w_{12}	w_{22}
Adult frequencies	$p_t^i w_{11}$/*not found*	$2p_tq_tw_{12}$/*not found*	$q_t^i w_{12}$/*not found*

suppose that the survival probabilities were all cut in half, that is, $w_{11}=0.45$, $w_{12}=0.4$, and $w_{22}=0.35$. Then, we can use the above equation again to predict the frequency of A_1 in newly formed zygotes of the next generation, namely 0.43, exactly what we predicted before. These calculations illustrate a very important fact about natural selection: The change in allele frequency from one generation to the next as a result of natural selection depends only on the fitness of genotypes relative to one another. Even a genotype with a low probability of survival can be favored by natural selection if its probability of survival is higher than that of other genotypes in the population.

Since natural selection favors characteristics that increase the probability of survival, it is not surprising that the mean fitness of the new progeny generation is greater than that of the one that preceded it, unless the mean fitness is as great as it can be under the current conditions. When mean fitness is at a maximum, the allele frequency will not change from one generation to the next, even though genotype frequencies will differ between newly formed zygotes and reproductive adults. The population is at equilibrium.

We can predict characteristics of the population at equilibrium simply by knowing which genotype is most likely to survive, which is least likely to survive, and which has an intermediate probability of survival. Three patterns of selection are possible:

$w_{11}>w_{12}>w_{22}$ – Directional selection
$w_{11}<w_{12}<w_{22}$ – Directional selection
$w_{11}>w_{12}$ and $w_{22}>w_{12}$ – Disruptive selection (heterozygote disadvantage)
$w_{11}<w_{12}$ and $w_{22}<w_{12}$ – Stabilizing selection (heterozygote advantage).

Directional Selection

Directional selection occurs when individuals homozygous for one allele have a fitness greater than individuals with other genotypes and individuals homozygous for the other allele that have a fitness less than individuals with other genotypes. At equilibrium, the population will be composed entirely of individuals that are homozygous for the allele associated with the highest probability of survival. The rate at which the population approaches this equilibrium depends on whether the favored allele is dominant, partially dominant, or recessive with respect to survival probability. An allele is dominant with respect to survival probability if heterozygotes have the same survival probability as homozygotes for the favored allele, and it is recessive if heterozygotes have the same survival probability as homozygotes for the disfavored allele. An allele is partially dominant with respect to survival probability if heterozygotes are intermediate between the two homozygotes in survival probability. This pattern of selection is referred to as directional selection because one of the two alleles is always increasing in frequency and the other is always decreasing in frequency.

When a dominant favored allele is rare, most individuals carrying it are heterozygous, and the large fitness difference between heterozygotes and disfavored homozygotes causes rapid changes in allele frequency. When the favored allele becomes common, most individuals carrying the disfavored allele are heterozygous, and the small fitness difference between favored homozygotes and heterozygotes causes allele frequencies to change much more slowly (**Fig. 1**). For the same reason, changes in allele frequency occur slowly when an allele with recessive fitness effects is rare and much more rapidly when it is common. A deleterious recessive allele may be found in different frequencies in isolated populations even if it has the same fitness effect in every population, because natural selection is relatively inefficient when recessive alleles become rare, allowing the frequency to fluctuate randomly as a result of genetic drift.

Disruptive Selection

Disruptive selection occurs when heterozygous individuals are the least likely to survive. For that reason, this fitness pattern is also referred to as heterozygote disadvantage. If a population happened to start with an allele frequency exactly equal to $p^*=(w_{12}-w_{11}-w_{22})$, the allele frequency would not change, that is, the population would be at equilibrium. But the equilibrium is not stable. Selection magnifies even a tiny change in allele frequency until eventually one allele or the other is lost from the population. Which allele is lost depends on whether the initial allele frequency is greater or less than p^*. If the initial allele frequency is greater than p^*, A_2 will be lost and the population will be composed entirely of A_1 homozygotes at equilibrium. If the initial allele frequency is less than p^*, A_1 will be lost and the population will be composed entirely of A_2 homozygotes at equilibrium. This pattern of selection is referred to as disruptive selection because selection will cause two populations with similar allele frequencies to evolve in opposite directions if one has an allele frequency slightly less than p^* and the other has an allele frequency slightly greater than p^*

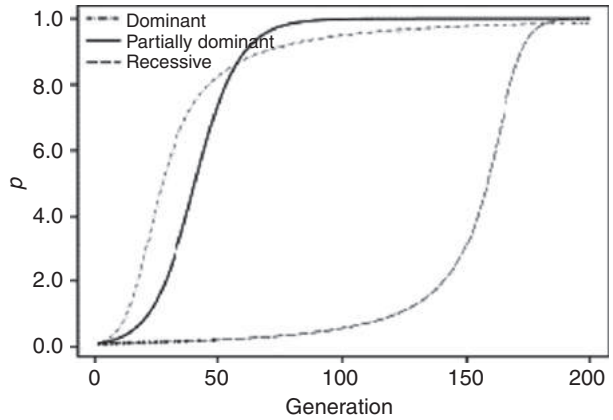


Fig. 1 Dynamics of directional selection.

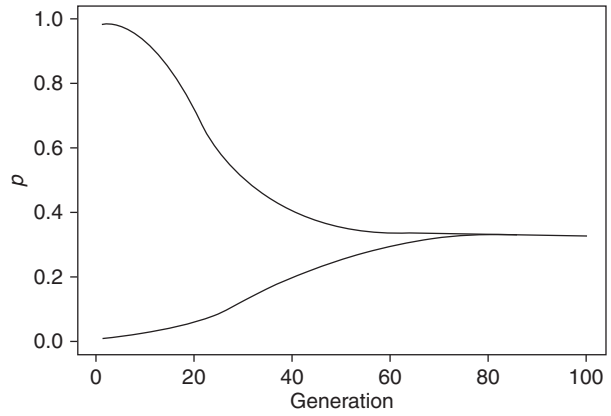


Fig. 2 Dynamics of heterozygote advantage.

Stabilizing Selection

Stabilizing selection occurs when heterozygous individuals are the most likely to survive. For that reason, this fitness pattern is also referred to as heterozygote advantage. As with disruptive selection, if a population happened to start with an allele frequency exactly equal to $p^* = (w_{12} - w_{11} - w_{22})$, the allele frequency would not change. When heterozygotes are more likely to survive than either homozygote, however, p^* is a stable equilibrium. Selection causes small departures from p^* to become even smaller with time. Moreover, the allele frequency in the population will evolve toward p^* regardless of the initial allele frequency, as long as both alleles are initially present. In Fig. 2, for example, $w_{11} = 0.72$, $w_{12} = 0.9$, and $w_{22} = 0.81$, and the population evolves toward $p^* = 0.33$ regardless of whether the initial allele frequency is 0.01 or 0.99.

Selection on Continuous Traits

Darwin and Wallace proposed the theory of natural selection almost 50 years before Mendel's rules were rediscovered. The logic is incontrovertible. If there are heritable differences among individuals that cause differences in reproduction and survival, then traits that increase the probability of survival and reproduction will become more common and those that decrease the probability of survival and reproduction will become less common. When geneticists study the inheritance of a continuous trait, they use the heritability of that trait to describe the extent to which offspring resemble their parents.

Response to Selection

Differences among individuals may arise because they have the same genotype but were exposed to different environments, because they were exposed to the same environment but have different genotypes, or because they have different genotypes and were exposed to different environments. The heritability of a trait is the proportion of phenotypic variation that can be transmitted from parents to offspring. In Fig. 3, the *x*-axis is half the summed body weight (in grams) of paired male and female laboratory

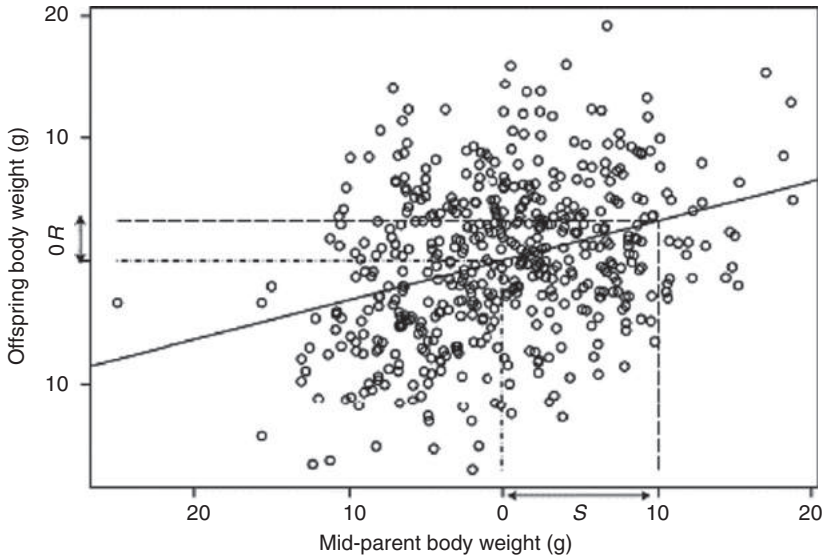


Fig. 3 Selection on body weight in laboratory mice.

mice. (This quantity is known as the mid-parent body weight.) The y-axis is the body weight (in grams) of offspring. The slope of the regression line running diagonally through the figure is equal to the heritability of body weight in this population of mice.

The regression line allows us to predict the body weight of offspring from the body weight of parents. Specifically, if we let x_s be the mid-parent body weight of a particular pair of parents in the population and $\bar{x}$ be the average mid-parent body weight in the population as a whole, then the expected body weight of their offspring is

$$x_0 = h^2(x_s - \bar{x}) + x$$

Suppose that natural selection causes a difference between the mean of a trait in those individuals that reproduce and the mean in a population as a whole. We can apply the above equation to the whole population. Now we interpret $\bar{x}$ as the mean mid-parent body weight 'before selection' and x_s as the mean mid-parent body weight 'after selection'. We can also rearrange the equation so that it directly predicts how much the mean body weight will change from parents before selection to offspring before selection (in the next generation):

$$x_0 - x = h^2(x_s - \bar{x})$$

The quantity $x_s - \bar{x}$ is called the selection differential, S. The quantity $x_0 - x$ is called the response to selection, R. The response of a trait to selection depends on both the heritability of the trait and the selection differential: $R = h^2S$. If a trait lacks heritable variation, $h^2 = 0$, there will be no change in response to selection no matter how strong the selection is.

Fisher's Fundamental Theorem of Natural Selection

We can use these results to study how fitness itself changes from one generation to the next. Intuition tells us that if fitness-enhancing traits become more common, the average fitness in the whole population should also increase over time, and that is just what we find. Let w_t be the mean fitness before selection, $\overline{w}_t^*$ be the mean fitness after selection, and w_{t+1} be the mean fitness in the next generation. Then,

$$w_{t+1} - w_t = h^2(\overline{w}_t^* - w_t)$$

The mean fitness of a population after selection ($\overline{w}_t^*$) is greater than that before selection (w_t) because individuals with a high probability of survival are more common in the population after selection than they were before selection. In addition, h^2 is necessarily positive. Thus, w_{t+1} must be greater than w_t. The only time when this inequality will not hold is when $h^2 = 0$, which means that the population has reached the maximum possible fitness. This equation embodies Fisher's fundamental theorem of natural selection, and it implies three things about the process of evolution by natural selection:

(1) The change in mean fitness between generations is proportional to the heritability of fitness.
(2) The mean fitness of a population will never decrease from one generation to the next and will remain constant only when the population has reached the maximum possible fitness.
(3) Evolution by natural selection gradually depletes the heritable variation in fitness that is required for it to continue.

Fisher's fundamental theorem of natural selection has many exceptions. In small populations, genetic drift may be a more important influence on the evolution of populations than natural selection, selection on fecundity differences need not follow the

same pattern as selection on viability differences, and fitness differences are rarely the same for more than a few generations, for example. Nonetheless, Fisher's fundamental theorem validates our intuition: As fitness-enhancing traits become more common, the average level of adaptation in a population will increase.

Detecting Natural Selection

When biologists say that a trait is subject to natural selection, they mean that

- Individuals differ in the probability with which they survive (viability selection);
- Individuals differ in their ability to attract mates (sexual selection);
- Pairs of individuals differ in the number of offspring they produce (fecundity selection); or
- Individual alleles differ in the probability with which they are incorporated into gametes (gametic selection).

Viability Selection

Differences in probability of survival are most often associated with natural selection. When we say that an organism is well adapted to its environment, we often mean that it has a high probability of survival. Detecting differences in survival probability is relatively straightforward.

When a wild-type female of *Drosophila melanogaster* heterozygous for the allele causing white eye color is crossed with a white-eyed male, we expect half of the offspring to be white eyed and half to be wild type, regardless of sex. In one such set of crosses, experimenters obtained the results shown in **Table 2**.

Since the rules of Mendelian genetics tell us that there were equal numbers of wild-type and white-eyed zygotes formed, the deficiency of white-eyed flies must be the result of a lower probability of survival. This experiment shows that there was viability selection against white eyes in this laboratory population of *D. melanogaster*.

Sexual Selection

The most obvious way in which individuals differ in their ability to attract mates is when males compete for control of a harem, as in North American elk (*Cervus canadensis*). This type of sexual selection is known as male–male competition. Males and females of most species differ in many characteristics that are not directly related to the reproductive process. Such characteristics are known as secondary sexual characteristics. When differences among males in secondary sexual characteristics cause females to choose some for mates in preference to others, it is a type of sexual selection known as female choice.

The pied flycatcher (*Ficedula hypoleuca*) breeds in Europe, northern Africa, and western Asia. Two color forms of male pied flycatchers are found in Europe. The black-and-white form has black feathers on its head, the nape of its neck, and its back. It has white feathers on its chin, the front of its neck, and its underside. The brown form has brown feathers instead of black ones on its head, the nape of its neck, and its back.

To determine whether a male's color affects the female's choice of mate, experimenters placed pairs of males, one black and white and one brown, in outdoor aviaries. The aviaries contained three compartments. Each male was placed in a separate compartment and prevented from seeing the other male. After the males were habituated to the aviary, a female was placed in the third compartment with two nest boxes, one close to each of the males. The female could see both males. When the female built a nest, the experimenters noted whether she built it in the nest box associated with the black-and-white male or in the one associated with the brown male.

Females for the experiment were collected from an area in central Europe where the closely related black-and-white collared flycatcher (*Ficedula albicollis*) also occurs. Out of 12, 10 females (5:1) built their nests in boxes associated with the brown male, showing that female choice leads to sexual selection in favor of the brown form in this region. In areas where the black-and-white collared flycatcher does not occur, similar experiments showed that sexual selection through female choice favors the black-and-white form of the pied flycatcher.

Fecundity Selection

Experiments in laboratory populations of *D. melanogaster* have repeatedly shown that differences among individuals in the number of offspring they produce contribute more to fitness differences among individuals than to differences in the probability of

Table 2 .

	Red eyed	*White eyed*	*Total*
Observed	2652	2088	4740
Expected	2370	2370	4740

survival. Experiments measure the magnitude of fecundity selection by counting the number of offspring produced from different types of matings.

Adults heterozygous for *Cy* have curled wings when pupae are raised at 25°C. To determine whether females with curled wings produce fewer offspring than those with normal wings, experimenters allowed each female to mate with one male and calculated the mean number of adult offspring each type of mating had produced 18 days after mating (**Table 3**).

The fecundity of *Cy* females was 95% of the fecundity of wild-type females when mated with a wild-type male and only 79% of the fecundity of wild-type females when mated with a *Cy* male. Similarly, *Cy* males produced fewer offspring than wild-type males regardless of whether they were mated with *Cy* or wild-type females. Fecundity selection favors wild type in both females and males. Moreover, the differences between *Cy* and wild type in female fecundity (5%–20%) are much greater than those in viability ($<1\%$).

Gametic Selection

Mendel's rules tell us that half of the gametes produced by a heterozygous individual will carry one allele and half will carry the other, but Mendel's rules are sometimes broken. In mice (both *Mus musculus* and *M. domesticus*), the genes of the major histocompatibility complex and many others are tightly linked in a region near the centromere of chromosome 17. Because recombination between these genes is rare, the entire region is usually transmitted as if it were a single Mendelian gene. Mutations in the *t* complex, as this chromosomal region is known, often affect viability. In addition, over 90% of the gametes transmitted by males heterozygous for a 'complete' *t* and a wild-type *t* haplotype carry the complete *t* haplotypes. The great excess of complete *t* haplotypes in the progeny of heterozygous males shows that gametic selection favors the complete *t* haplotypes over wild-type *t* haplotypes.

Levels of Selection

It is natural to refer to organisms when discussing natural selection and its consequences. But the example of the *t* haplotype in mice illustrates that natural selection can operate at the level of gametes too. In fact, the *t* haplotype illustrates that natural selection may act simultaneously at several different levels of biological organization, the gamete or the gene, the individual organism, and the population or group.

We have just seen how biased segregation in favor of the complete *t* haplotype leads to gametic selection in favor of the complete *t* haplotype. If gametic selection were the only evolutionary force affecting this trait, the complete *t* allele would rapidly sweep through mouse populations and eliminate the wild-type allele. But gametic selection is not the only force. Many complete *t* haplotypes carry recessive lethals, and many of those that do not carry lethals cause sterility when homozygous. Selection at the level of the individual organism favors the wild-type haplotype. If it were the only evolutionary force affecting this trait, the complete *t* allele would be rapidly eliminated from mouse populations. The balance between these opposing forces leads to maintenance of both alleles in mouse populations.

Mathematical models that combine gametic and individual selection, however, predict that the complete *t* haplotype should be found much more frequently than it is. Selection at the level of groups or populations may be responsible for the discrepancy.

Mouse populations are often small and founded by only a few individuals. As a result, genetic drift may have a large influence on allele frequencies within them. In a few populations, the complete *t* haplotype may become very common. When it does, there is also a possibility that, by chance, all the offspring produced will be homozygous for a complete *t* haplotype. If they are, the population is doomed to extinction. If the population is recolonized, the new colonists will probably have a lower frequency of the complete *t* haplotype. Selection among groups favors groups with a low frequency of the complete *t* haplotype, reinforcing selection at the level of individual organisms.

Evolution by natural selection among groups or populations is possible when

(1) Groups differ from one another in their probability of extinction, in the probability that migrants from them found new populations, or in the probability that migrants from them are incorporated into existing groups; and
(2) Migrants that form new groups or are incorporated into existing groups resemble the groups from which they were drawn. As the example of the complete *t* haplotype makes clear, evolution by natural selection need not produce the best possible results for a whole species or even for individual populations. The 'best' result for mouse populations would be if the

Table 3 Fecundity selection in *Drosophila melanogaster*

Female genotype	*Male genotype*	
	Cy	*Wild type*
Cy	90.0	111.8
Wild type	114.2	117.1

complete *t* haplotype were completely eliminated. Selection in favor of the complete *t* haplotype at the level of gametes ensures that this will not happen, and the result is an equilibrium (a compromise) between two extremes.

See also: Aquatic Ecology: Microbial Communities. Behavioral Ecology: Altruism; Kin Selection; Optimal Foraging Theory; Environmental Stress and Evolutionary Change; Anti-Predation Behavior; Sexual Selection and Sexual Conflict. Evolutionary Ecology: Units of Selection; Fitness; Eco-Evolutionary Dynamics; Evolutionary Ecology. General Ecology: Ecophysiology

Further Reading

Clark, A.G., Feldman, M.W., 1981. The estimation of epistasis in components of fitness in experimental populations of *Drosophila melanogaster*. II. Assessment of meiotic drive, viability, fecundity, and sexual selection. Heredity 46, 347–377.
Darwin, C., 1859. On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life. London: John Murray.
Endler, J.A., 1986. Natural Selection in the Wild. Princeton, NJ: Princeton University Press.
Hartl, D.L., Clark, A.G., 1997. Principles of Population Genetics, third ed. Sunderland, MA: Sinauer Associates.
Prout, T., 1965. The estimation of fitnesses from genotypic frequencies. Evolution 19, 546–551.
Sober, E., 1984. The nature of selection: Evolutionary theory in philosophical focus. Cambridge, MA: MIT Press.
Stre, G.-P., Moum, T., Bureŝ, S., *et al*., 1997. A sexually selected character displacement in flycatchers reinforces premating isolation. Nature 387, 589–592.

Phylogenomics and Phylogenetics

Rodney L Honeycutt, Pepperdine University, Malibu, CA, United States

Glossary

Character Attribute of an organism useful for distinguishing taxa and inferring relationships among groups of organisms.

Character state Variants of a character.

Clade Branch of phylogenetic tree considered to be monophyletic.

Cladogram Diagrammatic representation of groupings of organisms based on shared-derived characters.

Classification Hierarchal ordering of taxa into categories of different ranks with the hierarchy, such as species, genus, family, order, class.

Homology Similar characteristics resulting from a shared ancestry.

Homoplasy Similarity of a characteristic that is not the result of shared ancestry.

Monophyletic group Grouping of taxa based on a shared common ancestor.

Orthologous Homologous genes derived from a common ancestor.

Outgroup Taxon used in a phylogenetic analysis to establish the placement of the root of a phylogenetic network of a group being studied and to determine the polarity of characters (primitive to derived states) used to construct the phylogeny.

Paraphyly Group in a classification that does not contain all the descendants of the most recent common ancestor.

Paralogous Term describing duplicated genes that produce gene trees different from species trees.

Phylogeny Tree that represents the evolutionary history of a group of organisms.

Sister groups Two groups descended from a common ancestor.

Taxon (pl. taxa) A group recognized by taxonomists as being a unit based on the shared and uniquely derived characters; the rank ranges from groups of populations to higher taxonomic categories.

Taxonomic characters Morphological or molecular attributes used to identify similarities and differences among species or groups of higher taxonomic rank.

Tree of Life

"As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever-branching and beautiful ramifications." This metaphor by Darwin (1859) reflects his view of all living organisms sharing a common evolutionary history characterized by a branching pattern of ancestor/descendant relationships that change through time as new species arise and others go extinct. Today, Darwin's Tree of Life is referred to as a phylogeny, an evolutionary hierarchy displaying ancestor/descendent relationships of groups of organisms (clades or monophyletic groups) sharing uniquely derived characteristics and a common ancestor. Once constructed, a phylogeny provides an interpretive framework for the development of a natural classification of organisms and for detailed examination of biological processes. As a result, phylogenies are being used in many disciplines of the life and earth sciences, anthropology, behavioral sciences, and applied sciences such as medicine, agriculture, and forensics. Even the origin of our domestic plants and animals is now being investigated through the lens of a phylogenetic framework.

Systematic Biology

Systematic biology focuses on organismal diversity, including the distribution of and relationships among populations and species. As a discipline, systematics is foundational to a broad spectrum of biology because it provides an ordered way of thinking about biological diversity and the processes responsible for that diversity. It is a comparative science that evaluates patterns of variation and uses that variation to recognize species and to infer relationships among populations and species. This latter component of systematics is known as phylogenetics. Another facet of systematic biology deals with taxonomy, which discovers and describes species and establishes classifications that reflect a hierarchy of groups based on their phylogenetic relationships. Phylogenetic classifications assign names (e.g., kingdom, phylum, class, order, family, genus) to groups of related organisms (clades) sharing a common ancestry (monophyly).

Reading a Phylogeny

Relationships among organisms, denoted 1–4 in Fig. 1, display a branching pattern, known as a phylogeny. Branches can be of two types, internal branches occurring between two nodes and branches connecting the tip with an ancestral node. Rooted phylogenies

 doi:10.1016/B978-0-12-409548-9.10578-0

also infer a time dimension with the oldest nodes and branches near the base of the tree with more recent branching patterns progressing toward the tips (Fig. 1A). Nodes of a tree represent hypothetical common ancestors, with bifurcation at a node representing points of divergence or speciation (Fig. 1B). Character states shared between the ancestor and descendants are considered primitive (plesiomorphic), denoted by the *dotted line*, whereas shared derived traits (synapomorphic) are denoted by the dashed line and unique changes along the solid lines are considered apomorphies (Fig. 1C). These shared derived traits define groups or clades sharing a common ancestor (monophyletic groups). In this figure, clade containing 1, 2, and 3 are defined by shared derived traits, while another clade (3, 4) share a separate set of derived traits. As a result of differences in the overall rates of evolution along particular lineages, branches can vary in length. When internode lengths are short, phylogeny reconstruction becomes more challenging. Outgroups are used to define the root of a tree containing a monophyletic group that is sister to the outgroup (Fig. 1D). Clades or sister groups are defined by a shared common ancestor (e.g., clades 1/2/3/4, 2/3/4, 3/4).

Phylogenetics

The goal of phylogenetics is to infer relationships among lineages (genes, individuals, populations, and species), with the underlying assumption that there is only one true phylogeny. Approximating that phylogeny requires an objective means of evaluating alternative phylogenetic hypotheses.

Early attempts at inferring phylogenetic relationships were based on morphological traits shared by extant (living) and extinct (fossil) forms of life. Approaches to phylogeny reconstruction have varied through time, and in many cases debate over the correct procedure has been contentious. Even with this disagreement, all systematic biologists agree that phylogenies should be derived from homologous characters (similarities derived from a common ancestor) rather than characters that are similar but not the result of shared ancestry (considered forms of homoplasy). A classic example of a homoplastic character is similarity (convergence) as a result of function rather than shared ancestry. For instance, the four chambered heart and homeothermy (ability to regulate body metabolically) are shared between birds and mammals, yet these traits should not be used to infer a direct common ancestry because they represent independent adaptations to an active lifestyle. An example of convergence at the molecular level is the shared similarity of amino acid sequences seen in the lysozyme c gene of leaf-eating monkeys and ruminant ungulates, two mammalian lineages that do not share a common ancestor. In this case, these similarities are in response to structural/functional changes associated with the activity of lysozyme c in foregut fermentation systems. Another form of homoplasy is reversal of a character to an ancestral state. Reversals are common with molecular data, especially nucleotide substitutions, as a result of the limited number of possible changes allowed. When referring to molecular data, the term orthology is used rather than homology, while paralogous characters reflect gene genealogies that run counter to the genealogies of species. For instance, the gene genealogy of gene duplications would fail to correspond to a species phylogeny.

Methods of Phylogeny Reconstruction

Several different methods have been used to infer phylogenetic relationships. These methods include distance-based approaches, maximum parsimony, maximum likelihood, and Bayesian inference. All these methods begin with a suite of characters (traits used

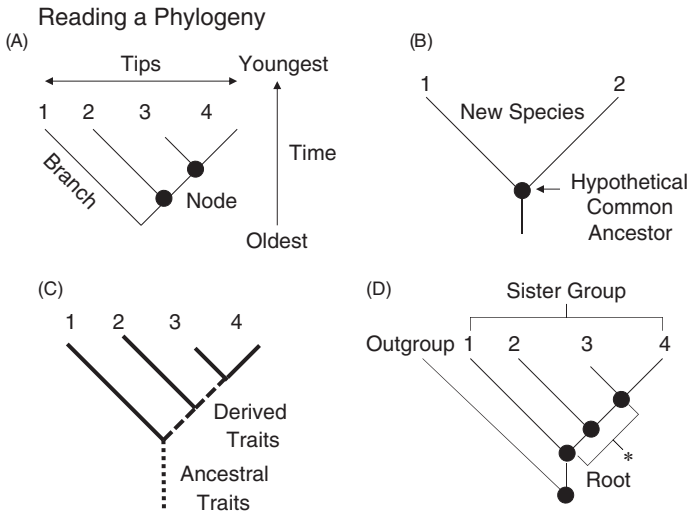


Fig. 1 (A) Components of a phylogeny with oldest being at the base of the tree; (B) species diversification from a common ancestor; (C) branching pattern characterized by ancestral (plesiomorphic) states denoted by *dotted line*, derived states (synaporphies) *dashed line*, and uniquely derived states (autapomorphies) denoted by solid line; (D) phylogeny showing the outgroup that defines the root as well as internal nodes of the tree (*).

to infer relationships such as a morphological feature or nucleotide and amino acid sites of a gene) used to create a data matrix, containing all taxa, characters, and character states (variants of each character) or in the case of distance-based approaches, a pairwise distance matrix. For some morphological traits, character states may be binary (present or absent), whereas molecular data can be represented as specific nucleotides (A, G, C, T), one of the 20 possible amino acids at each site, and an estimate of genetic distance based on the number of substitutions.

Phylogenies can be represented as either rooted or unrooted. An outgroup represents a taxon or taxa hopefully closely related (sister) to the group under study (the ingroup). The outgroup serves two purposes. First, it allows for the establishment of the root of the tree, and second, the polarity of character state transformations can be established. For instance, any character state shared between the outgroup and ingroup is considered an ancestral or shared primitive state (plesiomorphic). Derived (apomorphic) character states represent transformations from the ancestral state.

Finding the most accurate phylogeny becomes more complex as the number of taxa increases. For instance, the number of possible bifurcating trees for a matrix containing 10 taxa is over 34 million possibilities. Above 20 taxa, there is no exact solution, thus various methods of phylogenetic reconstruction use algorithms that perform some type of heuristic search of tree space to find the most optimal solution. How does one establish confidence in a branching pattern obtained from a particular analysis? All phylogenetic methods employ some type of analysis to establish confidence with respect to the interior nodes that join taxa into distinct groups or clades. Bootstrap methods are used for distance-based methods, maximum parsimony, and maximum likelihood to estimate support for particular branches of the phylogeny. The procedure takes random subsamples with replacement from a data matrix and performs replications of phylogeny reconstruction on the subsamples. This analysis is repeated hundreds or thousands of time, and the frequency at which the same nodes connecting particular taxa provides an estimate of support for that particular node. For instance, if 900 of 1000 iterations of bootstrap replications supported a node connecting guinea pig and mouse, the bootstrap support for that node would be 90%. Simulations and empirical studies suggest that a bootstrap value $\geq$70% is considered reasonable support. Bayesian inference estimates posterior probabilities denoting the level of support for a particular clade. These estimates are based on a majority rule consensus of all trees sampled during a run.

Distance-Based Methods

Over the years, methods of phylogenetic inference have varied. Phenetic approaches, using an estimate of overall similarity or distance to identify clusters or groups organisms, were some of the first to provide a more quantitative approach for inferring relationships. This particular method makes no distinction between ancestral and derived characters, and earlier methods, such as UPGMA (unweighted pair group method) assumed rate constancy. More recent methods employing pairwise distance matrices and clustering, such as neighbor-joining, do not assume rate constancy. For molecular data, this approach can allow for incorporation of models that correct for some variance associated with nucleotide sequence data. This method is still widely used in phylogenetic studies, and it is computationally fast and provides branch lengths. Two drawbacks are that it does not allow for an evaluation of discrete characters but relies on overall morphological or genetic distance among taxa, and distance estimates from highly divergent taxa are more likely to be inaccurate.

Maximum Parsimony

Maximum parsimony (MP) is a model-free method that uses discrete character states for inferring phylogenies. This particular method is advocated by those promoting a cladistic approach, whereby shared-derived characters (synapomorphies) are considered phylogenetically informative in terms of defining clades (monophyletic groups of organism). Shared ancestral character states (plesiomorphies or primitive) and autapomorphies (character states unique to a single linage or taxon) are both considered uninformative. A cladogram represents a dichotomous branching pattern characterized by the distribution of shared derived characters that identify monophyletic groups derived from a common ancestor. This approach argues that classifications should reflect the phylogeny.

The basic assumption of maximum parsimony is that the simplest explanation is the tree that most effectively minimizes homoplasy (convergence, parallelism, and reversal). Therefore, the tree with the shortest length is the most parsimonious tree. Maximum parsimony has been used extensively for the analysis of morphological data, and for more recently diverged taxa, it works well for molecular data. However, MP underestimates branch lengths, and it fails to provide a means for addressing evolutionary processes associated with molecular data, such a base composition bias, among-site rate variation, and multiple substitutions at a single site. This can result in inaccurate tree construction for deeper levels of molecular divergence as well as when branches differ in length and are connected by short internodes.

Maximum Likelihood

Maximum likelihood is a model-based method that allows for the use of models to address differences in the probability of some types of nucleotide substitutions. These models can range from the treatment of all substitutions and base frequencies as being equal to those that accommodate variance in base frequencies as well as all six classes of nucleotide substitutions and among site rate variation. An increase in model complexity increases the number of parameters that need to be assessed, and over

parameterization can result in inaccurate phylogeny reconstruction. Therefore, model selection is important prior to initiating maximum likelihood analysis. Maximum likelihood is computationally expensive because of its method of evaluating trees, and as such, its use for large molecular datasets is limited.

Bayesian Inference

Another probabilistic method used in phylogenetics is Bayesian inference, which is becoming very popular for analyzing large molecular datasets currently produced in the era of phylogenomics. Like maximum likelihood, Bayesian inference requires selection of a model of molecular evolution and uses a likelihood function, and it allows for a range of model complexity. Unlike maximum likelihood, which determines likelihood scores for all parameters of a particular model, Bayesian inference uses a prior distribution of trees in combination with the Markov Chain Monte Carlo (MCMC) method and parameters set by the selected model to estimate the posterior probability of trees. This method allows for a random search in tree space with sample trees taken at a designated number of steps. Assuming a prior distribution for phylogenetic studies is difficult. Therefore, most molecular phylogenetic studies assume a uniform prior distribution, thus allowing tree differences to be based on the likelihood of fit with the chosen model's parameters. In addition to phylogenetic inference, Bayesian methods are also used to estimate divergence times from molecular data. Finally, as with maximum likelihood, model selection is important for retrieval of accurate phylogenies and estimates of divergence times.

Phylogeography

Phylogeography combines gene genealogies with geography to explain the current distribution of populations and species resulting from various historical events. Phylogeographic patterns, especially those that are similar across unrelated groups (genealogical concordant), allow for an investigation of the processes (dispersal and vicariance) responsible for the biogeographic distribution of distinct gene lineages. In some cases, this approach helps identify common geographic events responsible for the patterns observed. A classic example is the concordant subdivision of unrelated freshwater fish populations into distinct haplogroups (phylogenetically related mitochondrial haplotypes) occurring in the southeastern United States and Florida relative to more western groups along the Gulf Coast. Such patterns suggest a common cause for the separation of these two major lineages.

The most common molecular marker used in phylogeographic studies is mitochondrial DNA (mtDNA), a maternally inherited genome derived from the mitochondrion. In comparison to the nuclear genome, the mitogenome evolves considerably faster, and this rapid rate of mutation in combination with its lack of recombination (resulting in haplotypes) provides strong geographic signal for divergence of populations as a result of isolation by distance or barriers to dispersal. Additionally, the mitochondrial clock is used to determine the geological date of events that occurred throughout the evolutionary history of a species or group of species.

Early studies of mtDNA variation relied on the estimation of genetic distances derived from an examination of restriction fragment length polymorphisms. Later, nucleotide sequencing of specific mitochondrial genes or the noncoding control region was used to infer mitochondrial phylogenies and estimate divergence times. Rapid and less expensive sequencing technologies have allowed for whole mitochondrial genomes of individuals to be sequenced. These mitochondrial genome sequences (termed mitogenomics) are being used to infer phylogenetic relationships across more divergent lineages, such as mammalian orders and families.

An example of the utility of mitogenomics can be seen from studies of human origins. Mitochondrial DNA appears to be better preserved in ancient samples, and today with genomic techniques, whole mitochondrial genomes have been sequenced from early modern humans and other hominins. Some of the hominin fossils have archeological dates as old as 300,000 years. This information has helped confirm the African origin of modern *Homo sapiens* as well as the dispersal patterns used by humans as they colonized regions out of Africa. Moreover, mitogenomics has allowed for the investigation of phylogenetic patterns of modern humans relative to Neanderthals and Denisovans, and comparisons based on nuclear genomes reveal evidence of genetic exchange among humans and these early hominins.

The current trend in phylogeography is to use a multilocus approach that combines information from both the nuclear (sequences from the Y-chromosome and autosomes) and mitochondrial genomes. In many cases, nuclear and mitochondrial markers provide concordant geographic patterns. For phylogeographic studies of human populations, variation in non-recombining regions of the Y-chromosome is being compared to patterns seen for mitochondrial DNA haplotypes. The phylogeographic trend observed for these paternal and maternal markers suggests a similar pattern of the movement of human populations out of Africa.

Phylogenomics

The advent of next-generation sequencing is revolutionizing studies designed to investigate the architecture of the genome and the function of various components of the genome. These same types of data also are being used to enhance our knowledge of molecular evolution and molecular phylogenetics. Genome level databases can be obtained at a small portion of the cost relative

to that formally required for the Human Genome Project, and this is allowing for detailed studies of broad areas of the Tree of Life. Phylogenomics emphasizes the use of genome-scale sequence data to construct phylogenetic trees depicting the relationships among organisms and to examine patterns of genome evolution and the function of genes.

The impact of phylogenomic approaches to phylogeny reconstruction is obvious from the plethora of recent students on many divergent groups ranging from prokaryotes to mammals. Large sets of data, derived from sequencing genes from both mitochondrial and nuclear genomes, are providing a better perspective of the placental mammal Tree of Life, which is now divided into four major superorders. One particular superorder, Afrotheria, contains a group of seven morphologically divergent and old lineages (elephants, sea cows, hyraxes, aardvarks, elephant shrews, golden moles, and tenrecs) that have an African origin. At the other end of the scale of divergence, phylogenomic approaches are providing insight into the genealogical consequences of adaptive radiations, characterized by the rapid origin of species connected by short branch lengths. For instance, Darwin's finches located in the Galapagos Islands represent a classic case of adaptive radiation, characterized by differences in the beak and ecology of the various species occupying these islands. Rather than a simply phylogeny, Darwin's finches reveal evidence of both hybridization (previously documented from ecological studies) between lineages and polymorphisms shared prior to divergence of lineages, resulting in incomplete lineage sorting. In addition, variation at one particular locus appears to associate with changes in shape of the beak, thus linking an adaptive phenotypic trait with underlying patterns of genetic variation. Lake Victoria cichlid fishes represent another group of species that experienced a recent adaptive radiation, and determining the phylogenetic relationships among this swarm of species has proven rather intractable. Phylogenomic methods, employing genome-wide sequencing, appear to be making progress in resolving relationships among sympatric groups of species.

Conducting phylogenomic research

Application of phylogenomics is the direct result of high-throughput methods for sequencing portions of or entire genomes in a fast and cost effective manner. Some of these next-generation sequencing platforms include Illumina, Ion Torrent, and Roche 454. In comparison to traditional Sanger sequencing, these platforms produce shorter sequence reads (a few 100 base pairs) with higher error, which requires multiple reads of the same sequence to reduce the rate of error. Third generation sequencing, such as single molecule sequencing in real time, offers the ability to provide sequence reads in the tens of thousands of base pairs. As this technology develops, genomics will enter a whole new era in terms of speed, accuracy, and cost.

A number of methods can provide genetic markers that span the entire genome, while enhancing the size of datasets used in population genetics, phylogeography, and phylogenetics. For species with a reference genome sequence, large scale deep genome sequencing is possible. Tens of thousands of human genomes are being sequenced with high levels of sequence coverage ($30–40\times$). This information offers opportunities for detailed studies of genetic disease as well as trends of human population divergence and movements. Whole microbial genomes are being sequenced, and the current results thus far suggest a tangled ancestry near the base of the Tree of Life. Whole genome sequencing of organisms that do not have an appropriate reference genome sequence (termed de novo sequencing) present more challenges in terms of assembling and annotating the new genome sequence. They also require considerably higher sequence coverage.

Other than whole genome sequences, there are a number of markers that can be used in phylogenomics. Transcriptome sequencing provides sequences from expressed genes (exons only). Use of this approach requires high quality mRNA, which is reverse transcribed into cDNA, followed by sequencing the paired ends of each cDNA. Sequences from these paired ends can be used to assemble contigs (contiguous sequence) representing overlapping reads that allow for construction of a consensus sequence for a particular gene. Alternatively, expressed sequence tags (ESTs) can be obtained from randomly sequencing the end of cDNA clones. Single nucleotide polymorphisms (SNPs) are widely distributed and abundant markers occurring in plant and animal genomes. Variation at these sites tends to be biallelic, and given their abundant distribution, they provide thousands of potential markers. These markers appear useful over a broad range of evolutionary time, being informative to investigating the origin of the domesticated dog as well as providing phylogenetic information for divergent ruminant artiodactyls including extinct and extant species.

Ultraconserved elements (UCEs) represent conserved regions of the genome that can be found across divergent taxa. Identification of these markers requires comparisons of genome alignments in search of regions that are highly similar. Sequences from these elements as well as more variable flanking regions provide larger numbers of orthologous loci that can be used to infer phylogenetic relationships among highly divergent taxa. For instance, the phylogenetic position of turtles represents a long-standing problem for those interested in the evolution of reptiles. Recently, a phylogenetic study using UCEs provided strong support for turtles being sister to the crocodile/bird clade.

Restriction site-associated DNA sequencing (RADSeq) is another method that produces large number of markers especially useful for population genetic and phylogeographic studies as well as for discovery of SNPs. These markers can be obtained from organisms for which little is known about the genome sequence, and large numbers of individuals can be screened for polymorphisms. Finally, the genome offers great potential for identifying other types of phylogenetically informative markers. The presence/absence of retrotransposons distributed throughout the genome and genome organization (as seen in comparisons of mitochondrial DNA) offer markers that may be useful in cases where standard sequencing fails to resolve relationships.

Processing phylogenomic data

Regardless of whether a phylogeny is constructed from portions of genes, single genes or whole genomes, there are several procedures that must be followed prior to organizing a data matrix for phylogenetic inference. Many of these procedures rely

heavily on bioinformatic techniques. Two of the first steps in processing sequence data prior to phylogenetic analysis are the elimination of sequence reads containing potential errors and the assembly of small sequence reads into a single contig. Such assembly can be computationally demanding, and sequencing errors must be minimized prior to sequence assembly.

As noted early, accurate inference of phylogenetic relationships requires the use of orthologous (homologous) sequences and annotation or gene prediction of those sequences. This means that proper identification and validation procedures must be performed prior to creation of a data matrix. Without a reference sequence based on whole genome sequences or sequences from transcriptomes, identification and annotation of orthologous sequences is less straightforward, but there are a number of analytical tools (e.g., tools found at the National Center for Biotechnology Information or NCBI) that can be used in the annotation and identification of orthologous sequences. One means of identifying orthologous sequences is to conduct a search of sequence databases. BLAST searches (implemented at NCBI) employ alignment procedures that allow for comparison of an unknown sequence to known sequences in the GenBank database. Positive matches are displayed in order of the overall significance of the match, and alignments for each match are provided. There are various options that can be used to optimize a search and allow for different degrees of similarity between the unknown and known sequences. In addition, different types of queries can be made including blastn (nucleotide search), blastp (amino acid sequences of proteins), and blastx (translates a nucleotide sequence into all six reading frames). In the case of blastx, open reading frames for translation of nucleotide sequences to amino acid sequences can be identified, thus allowing for the elimination of potential pseudogenes. Another method for identifying orthologous sequences is to perform a phylogenetic analysis, such as clustering, that can be used to compare gene trees with species trees.

The accuracy of inferring phylogenetic trees derived from either nucleotide or amino acid sequences depends on the multiple alignment used to establish positional homology of the individual nucleotides or amino acids being compared across species. There are several bioinformatic techniques available for aligning sequences. Clustal is a widely used program for multiple alignment, which uses estimates of pairwise distance between sequence pairs to construct a tree based on a clustering method that identifies the most closely aligned sequences. Alignment scores are provided, and newer versions of the program allow for selection of different gap penalties and other forms of weighting. As organisms become more divergent, establishment of positional homology becomes more challenging, thus requiring the alignment of amino acid sequences rather than nucleotide sequences. Another popular program is MUSCLE, which provides a fast algorithm that is useful for large sets of data.

Matrices containing pairwise comparisons of either nucleotide or amino acid sequences of taxa are constructed from multiple alignments. In most cases, sequences from all genes for each taxon are concatenated into a supermatrix, which provides the baseline for subsequent phylogenetic analyses. Nevertheless, inferring relationships from a supermatrix does not guarantee phylogenetic accuracy. Placement of some taxa may either remain ambiguous or display incompatibility with phylogenies obtained from other studies based on a different dataset of gene sequences, morphological characters or both. There are two perspectives as to how accuracy can be increased, and they are not mutually exclusive because both involve increasing the size of the supermatrix.

Some argue that taxon sampling is important for increasing accuracy, whereas others suggest that increasing the amount of sequence data is more useful for enhancing accuracy and addressing unresolved relationships. Both simulations as well as empirical approaches have been used to address this issue. It is clear that the use of large amounts of data with only a few representative taxa can produce misleading results. One example pertains to the use of a small number taxa to argue for the mammal order Rodentia not being monophyletic, with the guinea pig placed in a clade separate from rodents. The basis for this claim used complete mitochondrial genome sequences but only three rodents (guinea pig, mouse, and rat). This finding highly contradicts all the morphological and paleontological evidence that clearly identified all rodents, including the guinea pig, as monophyletic. Later molecular phylogenetic studies, using larger taxon sampling as well as nuclear gene sequences, provided strong support for rodent monophyly. Another reason for increased taxon sampling pertains to the elimination of long-branch attraction, which is a consequence of grouping lineages with a mixture of long and short terminal branches resulting from either unequal rates of change or missing taxa. In the latter case, inclusion of more taxa increases the probability of breaking up long branches, thus increasing accuracy. At the same time, inclusion of increased amounts of molecular data has the potential to resolve ambiguities associated with taxa connected by short internodes. For instance, lineages that separated over short periods of time as a result of rapid speciation are likely to be connected by short periods of shared ancestry.

Other than taxon sampling, the increase in taxa and amount of data can result in an increase in missing data for some taxa, and this represents another factor that might influence the accuracy of phylogenetic inference. Again, both empirical studies and computer simulations have been used to investigate the impact of missing data. Most studies suggest that incomplete sets of data do not necessarily decrease the placement of taxa, and the inclusion of taxa with incomplete datasets may help prevent systematic error from long-branch effects. However, too much asymmetry in missing data still has the potential to reduce phylogenetically informative character states required for determining relationships among some taxa.

Challenges to phylogenomics

Phylogenies derived from both the nuclear and mitochondrial genomes do not always agree (are incongruent) in terms of depicting the species tree. There are many explanations for why gene trees and species trees can provide contradictory patterns. Some of the more obvious molecular evolutionary processes that can result in incongruence include lateral (or horizontal) gene transfer, incomplete lineage sorting, and gene duplication (**Fig. 2**).

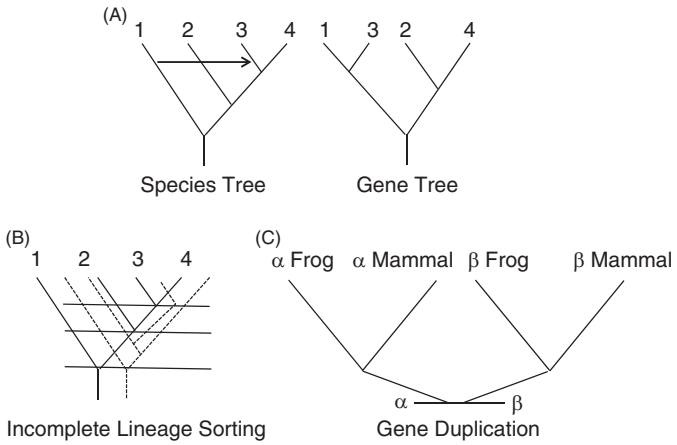


Fig. 2 (A) Species tree with lateral gene transfer shown with *arrow*, resulting in a different gene tree; (B) species tree denoted by the solid line phylogeny and relationships resulting from incomplete lineage sorting denoted by dashed lines, horizontal lines represent speciation events; (C) phylogeny of duplicated globin genes.

Lateral gene transfer (LGT), a phenomenon more common in prokaryotes, is a process that results in the exchange of genes between species that are not the result of vertical transmission from ancestor to descendent (**Fig. 2**A). In the case of groups that have experienced widespread horizontal gene transfer, a branching Tree of Life may not be the best representation of that group's evolutionary history. This process appears more pronounced in prokaryotes, resulting in some genes being positively misleading in terms of reflecting species trees. Views on the extent of LGT in prokaryotes differ in that some argue that it occurs extensively, whereas others suggest that the frequency is not extensive enough to preclude retrieval of the species tree. Obviously for prokaryotes, LGT among phylogenetically unrelated groups would yield a complex network displaying extensive reticulation. The frequency of LGT between either prokaryotes and eukaryotes or different eukaryotes is less clear. Endosymbiosis resulting in transfer of mitochondrial and chloroplast genomes from bacteria to animals and plants clearly represents an early exchange. There appears to be other evidence for LGT between bacteria and herbivorous insects and round worms. Within the eukaryotic genome, plant mitochondrial genomes show evidence of receiving genes from the nuclear genome, and segments of the mitochondrial genome are known to transfer to the nuclear genome of animals. Although some related fungi exchange genes, the frequency of LGT between eukaryote taxa is less clear. One reason for disagreement over the frequency of LGT in both prokaryotes and eukaryotes relates to the methods available for identifying cases of LGT. Most methods use phylogenetic incongruence between trees of the same taxa, unusual base composition in some genes relative to that seen in the whole genome, and high sequence similarity between genes from divergent lineages. Unfortunately, other processes (positive or purifying selection, hybridization, incomplete lineage sorting, gene duplication) can just as easily explain these anomalies, especially in eukaryotes.

Incomplete lineage sorting (ISL) occurs when an ancestral polymorphism is retained in lineages after speciation. If multiple lineages arise rapidly, the internode connecting those lineages will be short. As variation along each lineage coalesces, similar alleles may be retained in lineages not sharing a direct common ancestor (**Fig. 2**B). The consequence of ISL is incongruence between gene trees and species trees. For instance, resolving relationships among human, chimpanzee, and gorilla has been a challenge for decades, primarily because divergence of these three lineages occurred over short periods of time relative to their closest relative, the orangutan (**Fig. 3**). Morphologically, chimpanzee and gorilla are similar, yet most extensive genomic data (from DNA hybridization and sequencing) support a human/chimpanzee sister group. Nevertheless, genome wide comparisons suggest that a percentage of genomes support different topologies (**Fig. 3**A–C). This result is consistent with ISL.

Whether or not ISL is responsible for incongruence of genes at deeper levels of divergence has stimulated considerable debate, especially with respect to resolution of branching patterns in the mammalian Tree of Life. Some researchers suggest that algorithms that address coalescence and ISL provide better support for portions of the placental mammal phylogeny, especially when using a supermatrix of concatenated sequences that may have different gene genealogies. These coalescent methods attempt to deal with phylogenetic incongruence among individual gene trees. Others argue that ISL is a minor cause of incongruence, and that large molecular datasets of gene sequences should offset the effect of ISL.

Finally, gene duplication and loss of genes can influence phylogenetic inference. As indicated earlier, orthologous rather than paralogous sequences are required for inferring species trees. Gene trees can represent a family of genes that consisting of paralogs that are the result of gene duplication and serve similar functions, and mechanisms responsible for gene duplication include unequal crossing over and retrotransposition. The inability to correct for inclusion of duplicated loci can result in incongruence between gene trees and species trees. For instance, globin gene evolution in vertebrates involved gene duplication. Therefore, paralogous genes provide a history of the gene genealogy rather than the species genealogy (**Fig. 2**D). This implies that accurate identification of orthologous genes is necessary for retrieval of the species tree.

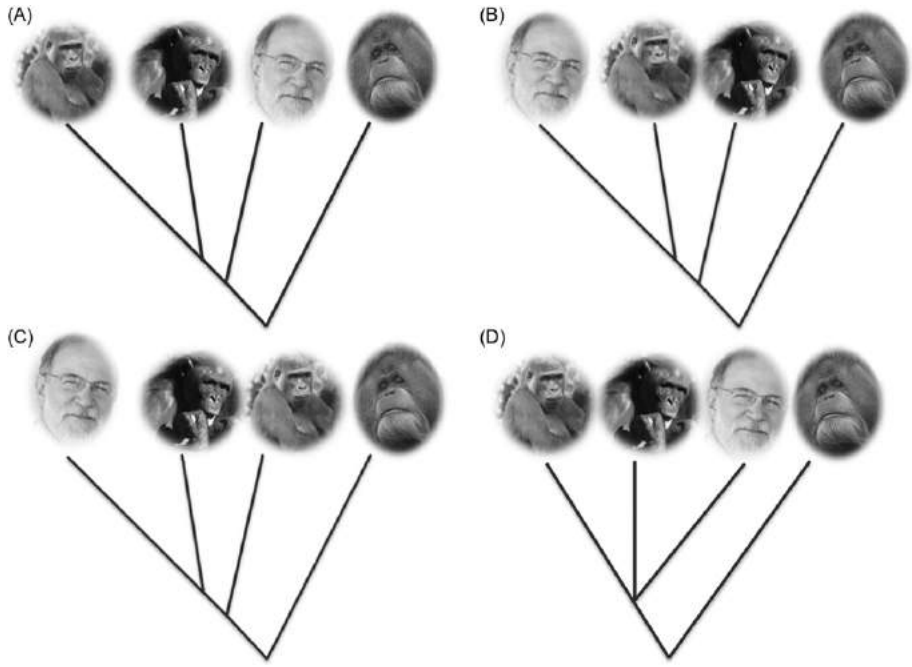


Fig. 3 There are three possible arrangement of human, gorilla, and chimpanzee. Although some early anthropological studies supported an unresolved polytomy as seen in (D), molecular data support the relationships shown in (C). Organutan is considered an outgroup.

Current and Future Uses of Phylogenies

As indicated earlier, a phylogeny provides an interpretive framework necessary for testing biological hypotheses associated with biogeography, ecology, functional anatomy, gene function, and molecular evolution. Therefore, an accurate phylogeny has many uses other than determination of relationships among taxa.

One of the original suggestions of a phylogenetic approach emphasized the use of a phylogeny for predicting the function of unknown genes. In this approach, gene trees provide a framework for mapping the function of known genes relative in an effort to infer function of unknown genes.

Phylogenies allow for comparison of groups of species that show as an evolutionary association, such as parasites and hosts, plants and pollinators, hosts and pathogens. Such comparisons allow for an examination of host switching in terms of parasites and diseases, as well as an evaluation of cospeciation between interacting organisms. For instance, *Mycobacterium tuberculosis* represents a disease of modern humans that shows congruent phylogeographic patterns to its host, suggesting that humans and tuberculosis coevolved and followed a similar migration pattern out of Africa.

Traits representing key innovations associated with an adaptive radiation can be identified with the species tree. The beak of the finch is a good example as well as changes in the feeding mechanism of cichlid fishes. This allows for an evaluation of the trait and patterns associated with the phylogeny, especially adaptive traits resulting from response to changing environmental conditions. Such traits covary in a phylogenetic context.

In recent years, interest in phylogenetic community ecology has increased. Such an approach examines how communities are structured based on the occurrence of phylogenetically related and phenotypically similar versus dissimilar lineages or species. Do phylogenetically related lineages share similar life histories and ecologies? What is the phylogenetic relatedness of members of a community? Such estimates of phylogenetic relatedness provide a metric for the structure of communities and the formation of island communities. Phylogenetic relatedness is also a proxy for levels of functional differentiation related to ecosystems.

Phylogenies provide information pertaining to the origin of diseases as well as the sources of particular infectious agents, and they can aid in drug discovery by identifying groups of organisms that may offer potential chemical agents similar to that found in a single representative from a particular clade. They also allow for examination of the spread of pathogens and to identify changes in the genomes of pathogens as they become more virulent and resistant to various drugs. Additionally, phylogenetic methods can help determine the origin of emerging diseases as well as the dispersal of diseases from their origin. For instance, HIV is now known to be derived from two different primate hosts.

Forensics is another area where phylogenetics is making inroads, and phylogenetic methods are used in criminal cases. One example is the use of phylogenetics to identify the source of anthrax sent through the postal system.

Finally, aside from inferring phylogenetic relationships, molecular data are used to estimate divergence times for various parts of the Tree of Life. Some of these temporal estimates provide information otherwise unavailable from direct evidence of fossils (e.g., estimating chimpanzee/human divergence). Since the 1960s, the concept of a molecular clock has fostered a plethora of studies

designed to estimate both recent (e.g., modern *Homo sapiens*) and ancient (e.g., Metazoa) origins. Ultimately, fossils do provide the primary means of calibrating the molecular clock. This requires calibration points derived from fossils that are reliable. Sometimes, fossils and molecules disagree, with the former suggesting considerably younger dates than the latter. A good example is the disagreement of data obtained for the origin of placental mammals. Some of the first dates based on molecules suggested the origin of placental mammals in the late Jurassic, which is 50–60 million older than the oldest fossil. The inclusion of more complex models that deal with rate heterogeneity, as well as more well-verified fossil calibration points, has resulted in estimates of divergence dates for placental mammals that are consistent with dates provided by fossils.

Future of Phylogenomics

Separating signal from noise in phylogenetics is a challenge when using any type of character. The real issue with phylogenomics is that with the ever-increasing size of datasets comes an almost exponentially increasing level of complexity. As we know, the genome is a mixture of coding and noncoding sequences with sometimes very different evolutionary histories in terms of the degree of selective constraint, recombination, gene conversion, duplication, horizontal gene transfer, chromosome rearrangement, unequal crossing over, and other genetic and demographic (e.g., generation time and effective population size) processes that create a nonhomogenous suite of characters. Separating the wheat from the chaff in an attempt to accurately depict the Tree of Life is computationally challenging, especially when attempting to include the deeper nodes of the tree near its roots. How does one increase reliability in the search for the most optimal solution, especially when a particular tree is positively misleading, yet well-supported by the optimality criterion selected? Is it better to use complex models that attempt to either correct for or identify sources of heterogeneity in the way molecules evolve? Is it better to remove potentially ambiguous genes or is it better to include all of the data with the assumption that the real phylogenetic signal will be obtained? These are not trivial questions, and to date no final solution exists. Rather, we are in a period of debate as to the best approach, which is reminiscent of earlier debates over many of the same issues. Nothing is new under the Sun.

See also: Evolutionary Ecology: Macroevolution; Metagenomics. General Ecology: Biodiversity; Allopatry

Further Reading

Avise, J.C., 2000. Phylogeography: The history and formation of species. Cambridge, MA: Harvard University Press.
Comas, I., Coscolla, M., Luo, T., Borrell, S., Hold, K.E., 2013. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nature Genetics 45, 1176–1182.
Darwin, C.R., 1859. On the origin of species by means of natural selection, or the preservation of favored races in the struggle for life. London: John Murry.
DeSalle, R., Rosenfeld, J., 2012. Phylogenomics: A primer. New York/London: Garland Science.
Donoghue, P.C.J., Benton, M.J., 2007. Rocks and clocks: Calibrating the tree of life using fossils and molecules. Trends in Ecology & Evolution 22, 424–431.
Eisen, J.A., 1998. Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Research 8, 163–167.
Felsenstein, J., 2004. Inferring phylogenies. Sunderland, MA: Sinauer Associates, Inc.
Holder, M., Lewis, P.O., 2003. Phylogeny estimation: Traditional and Bayesian approaches. Nature Reviews. Genetics 4, 275–284.
Keeling, P.J., Palmer, J.D., 2008. Horizontal gene transfer in eukaryotic evolution. Nature Reviews. Genetics 9, 605–618.
Kumar, S., Filipski, A.J., Battistuzzi, F.U., Pond, S.L.K., Tamura, K., 2011. Statistics and truth in phylogenomics. Molecular Biology and Evolution 29, 457–472.
Lee, M.S.Y., Ho, S.Y.W., 2016. Molecular clocks. Current Biology 26, R399–R402.
McCormack, J.E., Hird, S.M., Zellmar, A.J., Carstens, B.C., Brumfield, R.T., 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular Phylogenetics and Evolution 66, 526–538.
Mindell, D.P., 2006. The evolving world: Evolution in everyday life. Cambridge, MA: Harvard University Press.
dos Reis, M., Inoue, J., Hasegawa, M., Asher, R.J., Donoghue, P.C.J., Yang, Z., 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proceedings of the Royal Society B 279, 3491–3500.
Veeramah, K.R., Hammer, M.F., 2014. The impact of whole-genome sequencing on the reconstruction of human population history. Nature Reviews. Genetics 15, 149–162.
Weber, M.G., Agrawal, A.A., 2012. Phylogeny, ecology, and the coupling of comparative and experimental approaches. Trends in Ecology & Evolution 27, 394–403.
Yang, Z., Rannala, B., 2012. Molecular phylogenetics: Principles and practice. Nature Reviews. Genetics 13, 303–314.

Pioneer Species

JW Dalling, University of Illinois Urbana-Champaign, Urbana, IL, USA

Introduction

In early ecological literature, the term pioneer was used to describe those plant species that initiate community development on bare substrate (primary succession). More recently, usage of the term has included microbial and invertebrate taxa, and describes the first colonists of sites affected by less extreme disturbance which undergo secondary succession. Pioneers of primary and secondary successions share some traits; in both cases colonization of new habitat depends on effective dispersal, which generally selects for high reproductive output and small propagule size. However, differences in resource availability between these habitat types result in different opportunities for growth and reproduction. Few species can be successful on both primary and secondary successions.

Pioneers in Primary Succession

Primary succession occurs when extreme disturbances, such as landslides and volcanic eruptions, create new habitats by removing or covering existing vegetation and soil. Pioneers that initiate primary succession must be able to establish and grow on substrates that are nutrient poor and that often have unfavorable moisture conditions. The most extreme sites are exposed unweathered rock surfaces. Here, colonization may be limited to cyanobacteria ('blue-green algae'), lichens, and bryophytes, with no further vegetation development. Somewhat more nutrient-rich conditions associated with weathered or fragmented bedrock surfaces, such as the scree slopes of landslides, are often dominated by tree species. Sites still richer in mineral nutrients, which may contain some residual organic soil, such as the depositional zones of glacial moraines, in turn are often colonized by herbaceous species and grasses with faster growth rates (Fig. 1).

For pioneers in primary successions, nitrogen is often the most limiting resource. Unlike other mineral nutrients that can be released through weathering of underlying rock, nitrogen must either be transported to primary successions through leaching and deposition, or fixed *in situ*. Some of the most inconspicuous pioneers on exposed rock faces are nitrogen-fixing cyanobacteria. Rates of nitrogen fixation by cyanobacterial 'biofilms' on rock surfaces may be considerable; thus, nitrogen-rich leachate from these surfaces may affect community development at down-slope sites. Cyanobacteria may also form symbiotic associations with lichens (e.g., *Stereocaulon* spp.). These lichens are among the first colonists of landslides and lava flows (Fig. 2). Nitrogen fixation by lichens on these sites (0.2–0.4 kg N ha^{-1} yr^{-1}) may be important in facilitating the later colonization of these sites by vascular plants.

Relatively few vascular plant pioneers are nitrogen fixing. An exception is the perrenial lupine (*Lupinus lepidus*). Lupine is a conspicuous pioneer of the extensive ash and pumice fields that were created by the eruption of Mt. St. Helens (Washington State, USA) in 1980. In the first decade after the eruption lupine patches spread rapidly, increasing available nitrogen in the soil tenfold, potentially facilitating the growth of other colonizing plant species. More recently, however, the spread of lupine patches has slowed as specialist insect herbivores have colonized the plants. The patchiness and unpredictability of vegetation colonization on Mt. St. Helens also highlights the importance of constraints other than nutrient availability that limit recruitment success. Dispersal limitation, described as the failure of seeds to arrive at suitable establishment sites, may limit the rate at which pioneers colonize available substrate, and may be important in shaping the trajectory of successional change. Similarly, requirements for safe sites that provide favorable conditions for seedling establishment may account for the nonrandom distribution of pioneers on substrates such as glacial till. Small seed size and wind dispersal are particularly common traits among vascular plant pioneers.

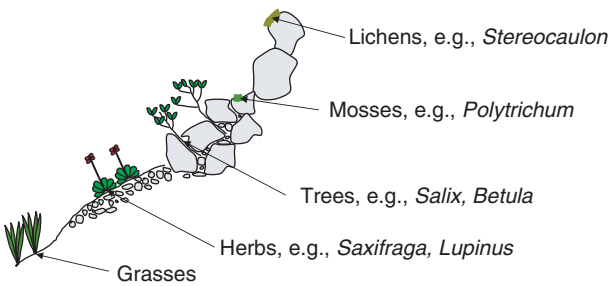


Fig. 1 Pioneer species typical of substrate types exposed following deglaciation. More fertile (nitrogen-rich) substrates often support herbaceous species and grasses, whereas rock surfaces and glacial till support lichens, bryophytes, and woody species. Modified from Grubb PJ (1986) The ecology of establishment. In: Bradshaw AD, Goode DA, and Thorp E (eds.) *Symposium of the British Ecological Society, Vol. 24: Ecology and Design in Landscape*, pp. 83–97. Oxford: Blackwell.

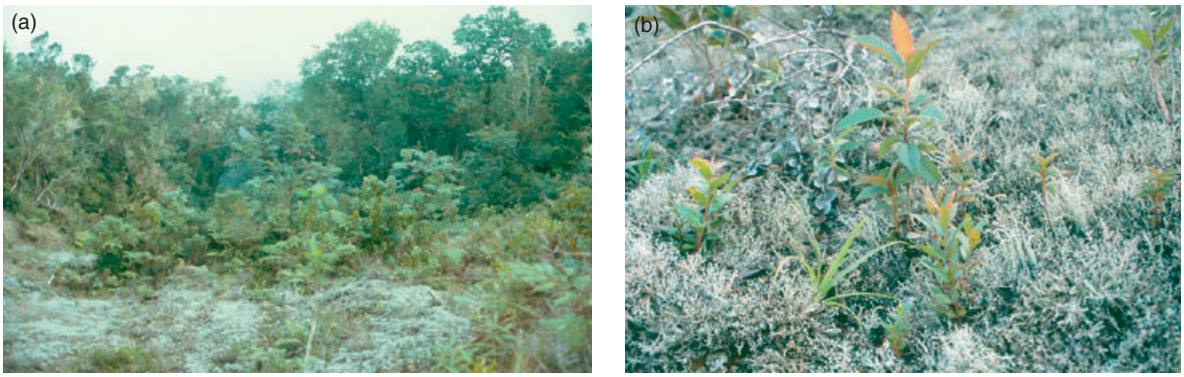


Fig. 2 (a) A 15-year-old landslide scar at 1500 m in the Blue Mountains, Jamaica. (b) The whitish appearance of the landslide surface results from its coverage by a dense mat of *Stereocaulon* lichen (note recruitment of woody species).

Table 1 Characteristics of pioneer tree species in tropical forests that distinguish them from nonpioneer species

Pioneer species	*Nonpioneer species*
1. Juveniles recruit from seed following disturbance; seedlings are unable to survive beneath a forest canopy	Seedlings and saplings persist in the shade of a forest canopy
2. Seeds germinate in response to cues provided by changes in light, temperature, or soil nitrate concentrations indicating disturbance to canopy vegetation	Seeds germinate immediately after dispersal or seasonally during periods favorable for establishment
3. Seeds generally small; frequently dispersed by wind	Seeds may be large; frequently dispersed by vertebrates
4. Seeds often persist in the soil (weeks to decades after dispersal)	Seeds lack dormancy or remain in the soil for less than a year
5. High height growth rate and juvenile mortality rate	Lower height growth, crowns often show lateral spread in the shade
6. High maximal photosynthetic rate, light compensation point, and foliar nutrient concentrations	Low maximal photosynthetic rate, light compensation point, and foliar nutrient concentrations
7. Short-lived leaves with high leaf area per unit leaf mass	Leaves of juvenile plants may persist for several years with low leaf area per unit leaf mass
8. Open canopies with sparse branching	Closed canopies
9. Low wood density	Medium–high wood density
10. Low investment in chemical anti-herbivore defense	High investment in chemical and structural defenses
11. Often form defensive mutualisms with ants	Defensive mutualisms uncommon
12. Adult lifespan typically <100 years	Adult lifespan up to 500 years
13. Wide geographic and ecological range	Often restricted geographic range and habitat requirements

Adapted from Swaine MD and Whitmore TC (1988) On the definition of ecological species groups in tropical rain forests. *Vegetatio* 75: 81–86.

Wind dispersal is favored when animal dispersal vectors are rare. Small seed size may increase the probability of seeds becoming trapped in cracks and small depressions where germination and seedling survival are likely to be enhanced. Conversely, it has been suggested that small seed size limits the initial nutrient resource supply available to the plant and may prevent seedlings from developing mutualisms with nitrogen-fixing bacteria.

Pioneers in Secondary Succession

Secondary succession occurs when the severity of disturbance is insufficient to remove all the existing vegetation and soil from a site. Many different kinds of disturbances, such as fire, flooding, windstorms, and human activities (e.g., logging of forests) can initiate secondary succession. Pioneers of secondary successions face quite different conditions from those that accompany primary succession. Secondary successions often start with resource-rich conditions associated with high light availability and reduced competition for nutrients and moisture. Disturbances may also be short-lived; for example, gaps created in forest canopies close as the crowns of surrounding trees expand and as seedlings and saplings in the understory grow up in response to increased light. Pioneers rely on recruitment from propagules present in the soil, or that disperse into the site after disturbance occurs. Pioneers are able to outcompete established vegetation that survived the disturbance by maintaining high juvenile growth rates. Some of the fastest growing trees are pioneers in tropical rain forests. Individuals of the balsa tree *Ochroma pyramidale*, for example, can grow from seedlings to adults with >30 cm trunk diameter in <10 years.

The difference between pioneer and nonpioneer species is difficult to delineate (**Table 1**). Attempts to define distinct life-history strategies (implying coordinated evolution of life-history traits) are confounded because key traits such as propagule size and

juvenile growth rate can vary over several orders of magnitude within a community and show broad overlap among species with contrasting habitat requirements. Nonetheless, interactions among traits can be used to describe some life-history tradeoffs that largely constrain the habitat requirements of pioneers. For vascular plants, paramount among these is a tradeoff between growth in the sun and survival in the shade (**Fig. 3**). The high growth rates of pioneers are maintained by allocating a large fraction of the plant's resources to new leaf area production, and by investing in nutrient-rich leaf tissue that can attain high-maximum photosynthetic rates. A consequence of preferential allocation to leaf production is that few resources remain that can be used to defend the plant against herbivores and pathogens, or to recover from physical damage. This results in high mortality, particularly in the shade, where resources needed for tissue replacement are most limiting.

For pioneers growing in high light environments, abundant supplies of carbohydrate fixed through photosynthesis can be used to co-opt the services of predaceous insects that defend the plant against herbivores. Many pioneers have extra-floral nectaries that provide food for insect mutualists. Two of the dominant genera of pioneers in tropical forests – *Cecropia* (Urticaceae) in the neotropics and *Macaranga* (Euphorbiaceae) in the Asian tropics – have developed a more elaborate mutualism that provides a striking example of convergent evolution in morphological traits. In both genera the hollow stems of saplings are colonized by queen ants (*Crematogaster* in *Macaranga; Azteca* in *Cecropia*). The ant colonies are then provisioned with carbohydrate and lipid-rich food bodies produced on leaf surfaces, stipules, or petioles (**Fig. 4**).

The transient and unpredictable occurrence of secondary successional habitats has selected for high dispersal ability among pioneers. Typically, pioneers are small-seeded reflecting selection for high reproductive output. Even so, seed mass may vary over four orders of magnitude among pioneers within a plant community, reflecting a second life-history tradeoff between colonization success (selecting for small seeds) and establishment success (selecting for larger seed reserves). For pioneers with limited dispersal, the probability of colonizing disturbances can be increased by maintaining populations of viable seeds in the soil. These soil seed banks may be transient, with seeds lasting a few weeks or months following dispersal, or may be persistent with seed surviving for decades. In temperate forests most seed bank-forming species are annual or perennial herbs. These are typically small-seeded species (<1 mg seed mass) that germinate in response to an increase in the intensity or red:far-red ratio of light associated with openings in the canopy or in the litter layer. In tropical forests both trees and herbs form seed banks with greater seed persistence common among the larger-seeded species (1–100 mg seed mass). Many of these species germinate in response to diurnal temperature fluctuations in the soil associated with large canopy gaps.

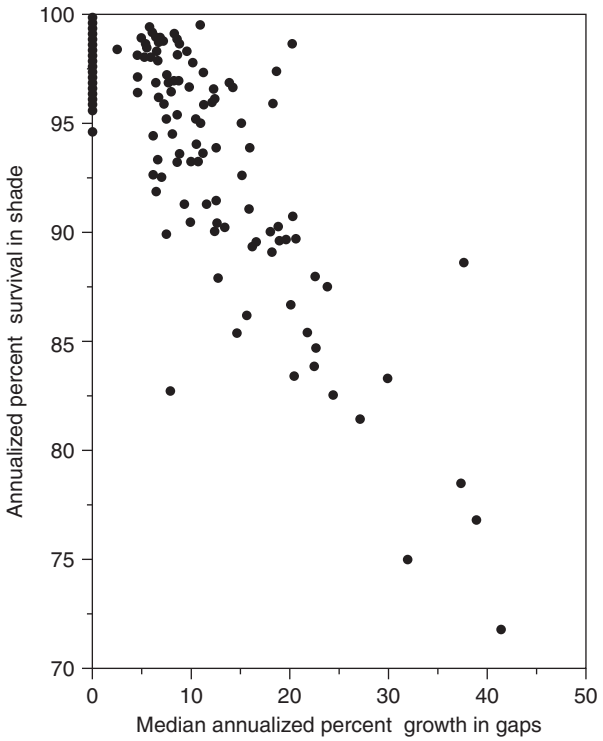


Fig. 3 The negative correlation between annual survival of rate of saplings 1–4 cm diameter at breast height in understory shade versus the median annual growth in the sun in tree fall gaps. Data are for canopy and mid-story tree species growing in semi-deciduous tropical forest in the 50 ha forest dynamics plot on Barro Colorado Island, Panama. Each data point is an individual species. Pioneer species have high growth rates in gaps and low survival in shade. Note that there is a continuum of responses to sun and shade that prevents a clear delineation of the pioneer guild. Reproduced from Hubbell SP and Foster RB (1992) Short-term dynamics of a tropical forest: Why ecological research matters to tropical conservation and management. *Oikos* 63: 48–61, with permission from Blackwell Publishing.

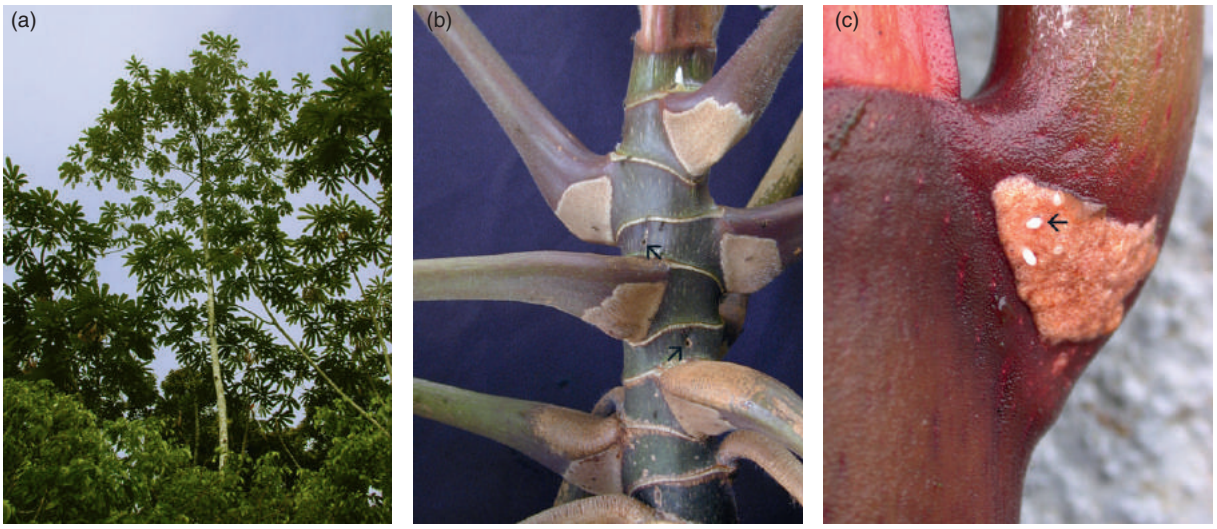


Fig. 4 (a) *Cecropia engleriana*, growing along a roadside in Yasuní National Park, Ecuador. *Cecropia* spp. are the dominant pioneers of young secondary forests in the neotropics. (b) The hollow stems of most *Cecropia* species are inhabited by aggressive ants (*Azteca* spp.) that predate insect herbivores. Arrow shows nest entrance. (c) In return, the plant provides the ants with Mullerian food bodies (shown by arrow) produced on the trichilium, a structure at the base of the petiole.

Changing land-use patterns have led to large increases in the abundance and distribution of many pioneer species. Many of the herbaceous pioneers that were originally restricted to forest gaps, or marginal habitats such as stream banks, have now become economically important weeds in agricultural systems. Similarly, in the tropics, clearance of old-growth forests, and abandonment of unproductive agricultural land has provided new habitats for pioneer tree species. Some of these pioneers can be quite long-lived and can produce valuable timber (e.g., teak, *Tectona grandis;* and laurel, *Cordia alliodora*).

See also: Ecological Processes: Succession and Colonization. Evolutionary Ecology: Colonization; Life-History Patterns; Ecological Niche; r-Strategists/K-Strategists. General Ecology: The Intermediate Disturbance Hypothesis. Terrestrial and Landscape Ecology: Anthropogenic Landscapes

Further Reading

Dale, V.H., Sawnson, F.J., Crisafulli, C.M., 2005. Ecological Responses to the 1980 Eruption of Mount St. Helens. New York: Springer.
Dalling, J.W., Burlsem, D.F.R.P., 2005. Role of life-history trade-offs in the equalization and differentiation of tropical tree species. In: Burslem, D., Pinard, M., Hartley, S. (Eds.), Biotic Interactions in the Tropics. Cambridge: Cambridge University Press, pp. 65–88.
Fastie, C.L., 1995. Causes and ecosystem consequences of multiple pathways of primary succession at Glacier Bay, Alaska. Ecology 76, 1899–1916.
Grubb, P.J., 1986. The ecology of establishment. In: Bradshaw, A.D., Goode, D.A., Thorp, E. (Eds.), Symposium of the British Ecological Society, vol. 24: Ecology and Design in Landscape. Oxford: Blackwell, pp. 83–97.
Hubbell, S.P., Foster, R.B., 1992. Short-term dynamics of a tropical forest: Why ecological research matters to tropical conservation and management. Oikos 63, 48–61.
Miles, J., Walton, D.H., 1993. Primary Succession on Land. Oxford: Blackwell.
Stearns, S.C., 1992. The Evolution of Life Histories. Oxford: Oxford University Press.
Swaine, M.D., Whitmore, T.C., 1988. On the definition of ecological species groups in tropical rain forests. Vegetatio 75, 81–86.
Thompson, K., 2000. The functional ecology of soil seed banks. In: Fenner, M. (Ed.), The Ecology of Regeneration in Plant Communties. Wallingford: CAB International, pp. 215–235.

Red Queen Dynamics

Ellen Decaestecker, KU Leuven, Science & Technology, IRF Life Sciences, Kortrijk, Belgium
Kayla King, University of Oxford, Oxford, United Kingdom

Glossary

Coevolution Changes in the genotypes of two or more species that are a direct consequence of the species' interaction with one another.

Coevolutionary arms race Occurs when an adaptation in one species reduces the fitness of individuals in a second species, thereby selecting in favor of counter-adaptations in the second species. These counter-adaptations, in turn, select in favor of new adaptations in the first species, and so on.

Eco-coevolutionary dynamics eco-evolutionary dynamics where coevolution is the central evolutionary process. In eco-evolution, ecological and evolutionary processes occur at the same time and pace and therefore interact with each other.

Genetic specificity in host parasite interactions—each host genotype resists only specific pathogen genotypes or species and each pathogen genotype or species infects only specific host genotypes.

Genetic polymorphism The long-term occurrence in a population of two or more genotypes in frequencies that cannot be accounted for by recurrent mutation. Such polymorphism may be due to mutations that are (a) advantageous at certain times and under certain conditions and (b) disadvantageous under other circumstances, and which exist in habitats where situations (a) and (b) are encountered frequently. Genetic polymorphism may also result if genotypes heterozygous at numerous loci are generally superior to any homozygous genotype.

Matching-allele type of model A model that explains the genetic background of parasite infectivity and host resistance, and states that genetic resistance in host–parasite interactions involves multiple genes: a host is resistant when the alleles of the host resistance genes match the alleles of the parasite virulence (infectivity) genes. It is mostly used to model host resistance and parasite infectivity in animal-pathogen associations. The MHC-system in vertebrates is a typical example of a system which is modeled by a matching-allele model.

Metapopulation A group of partially isolated populations belonging to the same species; subpopulations of natural populations that are partially isolated one from another and are connected by pathways of immigration and emigration.

Red Queen hypothesis The hypothesis that states that the adaptive importance of genetic recombination (and sex) can be found in its ability to create genetic variation among the offspring, which is important in confrontation with fast-evolving environmental conditions, including the coevolving parasites.

Transmission The process by which a *parasite* passes from a source of infection to a new *host*. Horizontal transmission is transmission by direct contact between infected and susceptible individuals or between disease vectors and susceptible individuals. Vertical transmission occurs when a parent conveys an infection to its unborn offspring, as in HIV in humans.

Virulence The morbidity and mortality of a host that is caused by parasites and pathogens.

Introduction

If we think of evolution, we often think of macro-evolutionary processes, long-term processes, which are difficult to imagine. Nevertheless, adaptations shaped by natural selection often occur over very short time frames. These adaptations lead to fast genetic changes over time. One type of interaction that plays a key role in ecological and fast evolutionary processes is the antagonistic interaction between hosts and parasites. Parasites are widely defined as organisms living in or on another organism—the host—, feeding on it, showing some degree of structural adaptation to exploiting it, and causing it some harm. Parasitism is the most common life style on earth: considering the high abundance and diversity of parasites, we can fairly say that every species is affected by parasites (Lafferty *et al.*, 2006). Empirical evidence has shown that new crop varieties have to be frequently renewed, as old ones are susceptible to their evolving pest populations. New diseases emerge and mutate (e.g., HIV), and the genetic composition of traditional diseases (influenza, malaria) changes continuously (Little, 2002). An example of a fast evolutionary process is the development of antibacterial resistance, for example, MRSA. Before the use of antibiotics, a large number of bacterial types were present, some were susceptible, others were resistant to antibiotics. Once antibiotics resistance occurred, only the most resistant bacterial strains remained, which became difficult to eradicate. It is this process that has lead to the fact that *Staphylococcus aureus* became resistant to methicillin such that due to mutations, methicillin can no longer block the cell wall synthesis of the bacteria. A next step in the antagonistic interaction between the pathogens and humans, is to find a new antibiotic or technique to eradicate these resistant bacterial strains.

The specificity of host–parasite interactions, their density-dependent transmission and the negative influence of parasites on the survival and reproduction of the host enable them to have an important influence on host population density and dynamics. The intense and reciprocal interactions between hosts and parasites may lead to coevolution: genetic changes that two (or more) species undergo as answer on their mutual interactions. The interaction between pathogenic microorganisms and their hosts is often compared with a "coevolutionary arms race," referring to the fact that the parasite searches for a new tool to attack and the host a new tool to defend. As microparasites adapt fast to their host genotype, it is beneficial for the host to pass its genes to genetically diverse offspring. Consequently, the host–parasite arms race is important for the maintenance of genetic polymorphism in natural populations (Otto and Michalakis, 1998), and idea posited by The Red Queen Hypothesis. This hypothesis states that the adaptive importance of genetic recombination (and sexual recombination, Jaenike, 1978; Bell, 1982) can be found in its production of genetic variation among the offspring, which is important in confrontation with fast-evolving environmental conditions, *in casu* fast genetic adaptation of parasites to their hosts (Maynard Smith, 1989). Parasites here can evolve quickly and continuously in response to the ongoing evolution of the defenses in the host (Paterson *et al.*, 2010). New mutations that influence host resistance and the possibility of the parasite to overcome that resistance are the fuel of host–parasite coevolution. When the environment changes, previously neutral or deleterious alleles can become favorable. If the environment changes sufficiently rapidly over time (i.e., between different generations), sexual reproduction reintroducing these alleles can be advantageous. This is especially so in systems in which parasites continuously track specific and common host genotypes, resulting in parasite driven time-lagged, negative frequency-dependent selection. The selective value of a host resistance allele is dependent on its frequency of occurrence in the population. In a host–parasite arms race, the parasite will increase its fitness by specializing on the most common host genotypes. Rare host genotypes have a selective advantage, until they become abundant and are no longer resistant. The production of genetic variation among offspring provided by sexual recombination is thus important in confrontation with the fast and specific genetic adaptation of parasites to their hosts. The genetic changes of trait values, hosts and parasites use to cope with each other in their evolutionary race are not always advantageous in the absolute sense, that is, as a remaining and increased impact on fitness. Rather, there is a constant need to adapt to avoid extinction driving by antagonism (cf. "in this place, it takes all the running you can do, to keep in the same place," said the Red Queen in Through the Looking Glass by Lewis Carroll; hence Red Queen hypothesis).

The consequence of these fast reciprocal interactions is fluctuating frequencies of allelic variants of the host and the parasite genotypes over time. These fluctuations in allele frequencies of host and parasite genotypes result in the above mentioned "coevolutionary arms race," where there is no net change or remaining increase in the fitness of both partners over time. Thus, who stands still, looses the game, but, no one progresses neither, at least not in terms of absolute fitness. The negative, frequency-dependent selection by parasites results in cyclic dynamics of host and parasite genotypes, each with their own resistance and infectivity, the so-called Red Queen dynamics. Host–parasite Red Queen coevolutionary dynamics are expected to be evolutionary unstable, as they are characterized by fluctuations in the frequencies of alleles of resistance and virulence associated traits. These coevolutionary "Red Queen dynamics" have been suggested to play a role in widespread biological phenomena such as the evolution and maintenance of sexual reproduction, and to shape the structure of natural populations and communities with consequences for biodiversity (Wood *et al.*, 2007). Rapid evolution and coevolution must therefore be one of the working hypotheses for explaining even short-term patterns and processes in the ever-changing web of life. There is no reason to think that the rapid evolutionary changes found in Darwin's finches, some invasive plants and insects, microbes in laboratory microcosms and responsible for antibiotics resistance are in any way exceptions (Thompson, 2009).

Experimental Model Systems of Red Queen Dynamics

Daphnia–Parasite Interactions

Coevolutionary interactions between the invertebrate water flea *Daphnia* and its micro-parasites have been studied intensively in the last decade. The *Daphnia*-parasite system is becoming a model system for biomedical questions and, more particularly, as an evolutionary framework for epidemiology (Ebert, 2005, 2011; Decaestecker *et al.*, 2009). *Daphnia* pathogens belong to a multitude of taxa. The main focus is on bacterial and microsporidian species, as these pathogens infect the *Daphnia* internally and have been shown to induce reduced fecundity and survival in the *Daphnia*. These bacterial and microsporidian pathogens are obligate intracellular pathogens that are mostly transmitted horizontally by means of infective spores (Ebert, 2005). It has been shown that *Daphnia* and its micro-parasites show characteristics that are expected to lead to strong coevolutionary responses: there is ample genetic variation in *Daphnia* resistance to parasites; host–parasite interactions are genotype specific (Carius *et al.*, 2001; Decaestecker *et al.*, 2003); and the genetic structure of *Daphnia* shifts during epidemics. There is also evidence of local parasite adaptation and short-term parasite-mediated selection in *Daphnia* (Haag and Ebert, 2004), which affects host–parasite dynamics (Duffy and Sivars-Becker, 2007). Also, on a higher taxonomic level, it has been shown that an under-infected taxon of a *Daphnia* hybrid system became over-infected after an increase in frequency and that this over-infection had a genetic basis (Wolinska *et al.*, 2006).

It is, however, notoriously difficult to study coevolutionary dynamics and the under-laying genetic mechanisms in nature, because often long-time series are needed. So far, direct empirical proof for the process of host–parasite coevolution is rare. In the *Daphnia*-parasite system, it is possible to circumvent this time constraint, as *Daphnia* and its parasites produce dormant stages resulting in

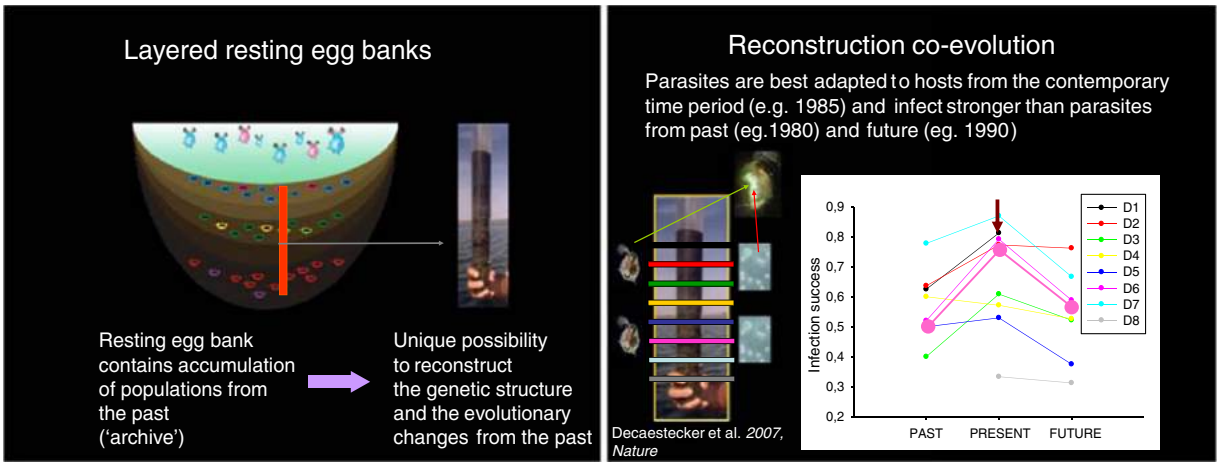


Fig. 1 "Resurrection ecology" approach to reconstruct Red Queen coevolutionary dynamics of *Daphnia*-parasite interactions (Decaestecker *et al.*, 2007; Houwenhuyse *et al.*, 2017).

dormant stage banks in layered pond sediments. Dormant stage banks confer a reflection of the populations from the past and provide a unique archive of past gene pools of both antagonists that can be reconstructed via the hatching of dormant stages from different sediment depths (this discipline is referred to as "resurrection ecology," Kerfoot and Weider, 2004, Orsini *et al.*, 2013). As such, genetic structure and evolutionary changes from the past can be reconstructed. In an earlier phenotypic study, a historical reconstruction of the coevolutionary process between *Daphnia magna* and its bacterial *endo*-parasite *Pasteuria ramosa* in a natural setting was made (Decaestecker *et al.*, 2007, 2013). In this study, coevolution was documented based on changes in host and parasite fitness parameters in infection experiments, where *Daphnia* clones were confronted with parasite isolates, both isolated out of different depths of sediment cores, reflecting different time frames of this interaction (cross infection experiments with parasites of the "past," "present" and "future", **Fig. 1**). Here, *Daphnia* and its *endo*-bacterial parasite *Pasteuria* were shown to adapt relatively fast and continuously, resulting in no net increase in parasite infectivity over time. This result reflects Red Queen dynamics, suggesting that the *Daphnia–Pasteuria* interaction is characterized by highly specific interactions and frequency-dependent selection across evolutionary time.

Potamopyrgus antipodarum Snail-Trematode Parasite Interactions

The New Zealand freshwater snail *Potamopyrgus antipodarum* and its coevolving sterilizing trematode parasite *Microphallus "iveyi"* is a powerful system to study Red Queen dynamics in nature. The snail exhibits polymorphism in its reproductive strategies, being either obligately parthenogenetic or sexual and dioecious; asexual *P. antipodarum* lineages are derived on multiple separate occasions from sexual progenitors (Neiman *et al.*, 2005). Infection by the castrating trematode generates negative frequency-dependent selection favoring sexual forms of *P. antipodarum* and diversity among the clonal forms all via Red Queen dynamics (Koskella and Lively, 2009; Jokela *et al.*, 2009; King *et al.*, 2009).

Multiple studies suggest that snail-trematode interactions here fit the "matching alleles" infection genetics model (Agrawal and Lively, 2001; Lively *et al.*, 2004; King *et al.*, 2009). This is an essential assumption underlying Red Queen dynamics, whereby infection is determined by whether a particular parasite genotype matches host genotypes that code for resistance. This match between host and parasite alleles enables *M. livelyi* to appear to the snail as "self" and evade immune defenses, ultimately establishing a successful infection. Strong selection is imposed reciprocally in this interaction because infected *P. antipodarum* are sterilized, and snails kill trematodes that cannot infect (King *et al.*, 2011a).

These snails and their parasites occur across multiple glacial lakes on the South Island of New Zealand, thereby providing the opportunity to examine Red Queen processes over space. The reciprocal selection for host resistance and parasite infectivity generated by *P. antipodarum–M. livelyi* interactions varies across these lakes, with parasite infection frequencies highest (and constant) in lakes where sexual snails are maintained in the face of competition with asexual conspecifics (Lively, 1987; Lively and Jokela, 2002). A similar pattern between sex and infection is observed within lakes whereby shallow regions, where parasites can be recycled and duck final hosts commonly forage, are coevolutionary hotspots (King *et al.*, 2009, 2011a). Deeper regions of lakes are coevolutionary coldspots where infection is rare and asexual snails dominate. In addition, the independent coevolutionary trajectories of each population are demonstrated by the strong patterns of parasite local adaptation to *P. antipodarum* populations, particularly in those where sex is maintained. There is a drastically higher infection rate in snails experimentally exposed to sympatric (same lake) vs. allopatric (different lake) parasites (King *et al.*, 2009).

Theoretical Models of Red Queen Dynamics

Mathematical models have confirmed that coevolving parasites should promote host genetic diversity, with immune system genes often showing striking levels of polymorphism, as well as species-level diversity. As such, antagonistic coevolution may be a key

driver of biodiversity in nature. However, despite extensive knowledge that variation in susceptibility has a genetic basis, we lack understanding of the genes involved, and of the underlying evolutionary and coevolutionary dynamics that these genes experience (Paterson *et al.*, 2010). Without such understanding, we lack the capacity to determine how adaptation arises at defense-related genes, how rapidly evolution can occur, and what drives the evolution of parasite susceptibility.

The most important models in host–parasite coevolution describing the underlying genetic control of infections are the matching-allele type of model and the gene-for-gene type of model. Host–parasite interactions explained by the gene-for-gene model are most commonly observed in plant-pathogen and bacteria-viral parasite interactions. The most important feature of this model is that one parasite genotype has universal virulence, such that it can infect all host genotypes. It is a model that describes generalist parasite and host strategies. A cost of virulence is needed to prevent this general genotype spreading throughout the parasite population. Thus, in a gene-for-gene interaction, there are susceptible and resistant host alleles and avirulent (non-infective) and virulent (infective) parasite alleles. Resistant hosts can only be infected by virulent parasites. Once a dominant resistance allele is present in the host population, it will spread (given that resistant hosts can defend themselves in the presence of avirulent parasites) and reciprocally lead to the spread of a virulent parasite allele (virulent parasites have an advantage in the presence of resistant hosts). Fluctuations in allele frequencies continue as once the parasite virulence allele is fixed in the population, the resistant allele of the host has no longer use, and the frequency of it will decline via selection (if resistance contains a cost), mutation or genetic drift.

A second type of model leading to allele frequency cycling more befitting the Red Queen is that assuming a highly specific, matching-alleles type of infection genetics. Matching-allele type of models are based on "self" versus "nonself" recognition in invertebrates, and describe the interactions of many animal-parasite interactions. The most important feature of these models is that infection and resistance contains a perfect "match" between the host and the parasite genotypes, such that parasites are recognized as "self" by the host and thus avoid detection. "Universal" virulence is not possible in these models and polymorphism can be sustained via negative frequency-dependent selection. In a matching-allele model, the parasite "wins" if it perfectly "matches" the host genotype. Parasites that match the host genotype, will have a higher fitness by which the frequency of the host genotype decreases, which leads to an increase of other host genotypes and consequently to an increase of other parasite genotypes.

Spatial Patterns of Red Queen Dynamics

The constant process of adaptation and counter-adaptation may lead to patterns of local adaptation dependent on the evolutionary potential of hosts and parasites. Parasites are expected to have a higher evolutionary potential than their hosts, because of similar mutation and recombination rate, combined with a much shorter generation time. Natural populations of many species are organized as a geographic structure existing of local populations between which migration (of parasites and/or host individuals) takes place and which have a largely independent population dynamic and chance of extinction (metapopulation structure, Olivieri *et al.*, 1990, Hanski and Simberloff, 1997). This has resulted in a growing interest of interactions between local populations on host–parasite relationships (Thompson, 1999). The interactions within local host–parasite interactions can be dependent on neighboring populations because populations might differ in their resistance and virulence genes. Through gene flow, new resistance and virulence genes can enter a given local population (Gandon *et al.*, 1998). On the one hand, gene flow can increase local genetic diversity in both parasites and hosts, but on the other hand, gene flow may also homogenize regional genetic variation and prevent local adaptation of both host and parasite (Little, 2002). In a temporal variable environment, gene flow is not necessarily maladaptive. A relatively high level of colonization can also restore the loss of genetic diversity through extinction. Moreover, a combination of local interactions, high virulence and limited gene flow can enhance genetic differentiation between populations, resulting in local adaptation (Gandon *et al.*, 1998; Lively, 1999). In this way, it is possible for the host to increase its chances in the coevolutionary arms race with the fast-evolving parasites (Jousimo *et al.*, 2014). For the host, gene flow may then act as a second mechanism, next to sexual recombination, to reduce the disadvantage of its long generation time and low evolutionary potential in comparison with its parasite. With its different, sometimes conflicting effects, gene flow is an important aspect of the geographic mosaic of coevolution. Genetic drift, founder effects, and local extinctions can also shape the interaction and contribute to the large-scale picture (Thompson, 1999).

Environmental Variation in Red Queen Dynamics

Awareness has grown that host–parasite coevolutionary interactions are strongly influenced by changing environments, both abiotic (Wolinska and King, 2009) and biotic (Betts *et al.*, 2016). This is a concept that originated in the "geographical mosaic of coevolution" concept (Thompson, 2005) that focuses on coevolutionary interactions in a spatial context and that is based on the observation that interspecific interactions vary depending upon genetic factors as well as environmental variation. Environmental modification can be the driver in the spread of allelic variants, for example a host genotype that is the least susceptible at one temperature, may be the most susceptible at another ($G \times E$ interactions, Vale and Little, 2009). As such, these processes will influence host–parasite frequency-dependent coevolutionary dynamics. Furthermore, the environment in which hosts and parasites interact may also substantially affect the strength and specificity of selection ($G \times G \times E$ interactions). In addition, various components of host–parasite fitness, for example, costs of resistance and infectivity can be differentially altered by the

environment (Wertheim *et al.*, 2003; Morgan *et al.*, 2009). Nevertheless, environmental fluctuations are often excluded from experimental coevolutionary studies and theoretical models as "noise." Because most host–parasite interactions exist in heterogeneous environments, it is argued that there is a need to incorporate fluctuating environments into future empirical and theoretical work on host–parasite coevolution (Wolinska and King, 2009).

The expected pattern of host–parasite coevolution depends, among other things, on the time lag between the host and parasite coevolutionary dynamics, on the time scale separating the various samples of the host populations and on the number of generations of hosts and parasites considered (Gandon *et al.*, 2008). For instance, seasonal changes in temperature and predation alter the interaction between the water flea *Daphnia* and its parasites such that differences in temperature and predation pressure will change parasite mediated selection, resulting in the maintenance of genetic variation of the traits involved (Duffy *et al.*, 2005; Hall *et al.*, 2006; Vale *et al.*, 2008). Also changes in environmental resource quality can modify the epidemiology of host–parasite interactions, but are also suggested to change the strength and direction of selection in host–parasite interactions. If environmental nutrient enrichment is translated into higher population sizes, then increased host–parasite interactions, in terms of within-host competition and parasite transmission, may intensify the coevolutionary arms race between hosts and parasites. Evolutionary host population responses with respect to resistance toward parasites are assumed to be strongest in highly productive systems. Larger host and parasite populations broaden the genetic variation available for selection and increase encounter rates between hosts and parasites. Higher encounter and transmission rates are associated with more virulent parasite strains, which pose stronger selection pressures for host defenses, further increasing the selection pressure on parasites for counter-defense. High environmental nutrient levels are found to favor rapidly growing species (growth rate hypothesis, GRH) which would then offer high-quality resources for parasites, and this could again favor rapidly growing parasites that have high nutrient demands and consequently express high virulence. Fast growth rates can thus be expected to intensify coevolutionary cycles via accelerating population turnover and increasing the number of mutations. Thus, environmental nutrient enrichment is expected to direct host evolution toward higher resistance, and parasite evolution toward higher virulence, intensifying the coevolutionary arms race and reducing the period of coevolutionary frequency-dependent oscillations (Decaestecker *et al.*, 2013; Aalto *et al.*, 2015).

The Red Queen in Eco-Coevolutionary Dynamics

It has become clear that evolutionary and ecological processes may occur at the same time and pace and therefore interact with each other ("eco-evolutionary dynamics," as defined in Hairston *et al.*, 2005, Hendry, 2016, De Meester *et al.*, 2016). This notion has led to many discussion papers about the need to integrate both evolution and ecology if we really want to understand ecosystem dynamics (Matthews *et al.*, 2011). The future challenge is going beyond understanding one-way interactions between ecology and evolution and to study how the reciprocal effects of organisms change their biotic and abiotic environment (Urban *et al.*, 2012). In a full "eco-evolutionary feedback loop," evolutionary processes significantly alter ecological dynamics and in return the altered ecological dynamics influence the evolutionary processes (Fussman *et al.*, 2007; Becks *et al.*, 2012; Turcotte *et al.*, 2013). Recently, there has been considerable interest in the interaction of ecological and evolutionary dynamics in an attempt to understand them as coupled "eco-evo" processes. Such eco-evolutionary feedbacks can occur at multiple levels, such as in demographic parameters, community composition, food webs, nutrient cycling and productivity (Hairston *et al.*, 2005; Urban and De Meester, 2009; Hendry, 2016). At the population level, natural selection and population dynamics are closely linked because both are affected by the birth and death of individuals. Thus, if natural selection acts on a trait through survival or reproductive success, it will leave a population dynamical signature. At a larger scale, changes in the genetic composition of a species can affect its fitness dependencies with other species (e.g., through trophic interactions or competition) and hence alter the ecological dynamics of an ecosystem, and vice versa.

Parasites play a crucial role in regulating and shaping host populations by altering their density, genetic structure and diversity. Such population level effects can mediate changes up to the community level through alteration of interspecific competition or through changes in host behavior (Penczykowski *et al.*, 2014). Finally, one can expect that effects of parasites on community structure may be propagated throughout the whole ecosystem by changing food web interactions and dynamics, thereby influencing energy flow and nutrient cycles. Parasites can greatly add to our knowledge on the role of "eco-coevolutionary dynamics" (i.e., eco-evolutionary dynamics in which coevolutionary processes are the crucial drivers) for ecosystem functioning with respect to biogeochemical processes. Rapid coevolutionary responses between hosts and parasites may play a critical role in shaping ecological processes (Hiltunen and Becks, 2014). Parasites have relatively short generation times and show strong and fast adaptation (Decaestecker *et al.*, 2007), increasing the potential of time scales for ecological and evolutionary responses to be congruent. The evolution of parasite phenotypes have been shown to induce changes in important ecological processes, such as nutrient cycling (Lennon and Martiny, 2008), but we lack demonstrations of a complete feedback loop between coevolutionary and ecological dynamics (Hiltunen and Becks, 2014).

Beyond the Pairwise: Red Queen and Microbiota

While pathogens may drive evolution with a stick, mutualistic symbionts can achieve similar feats by offering the carrot. Mutualists are facilitators of niche adaptation and are thought to have prominent roles in host speciation (Vavre and Kremer, 2014). While

examples of species evolution driven by specific pathogens or mutualists are numerous, the full picture is likely more complex, as animals (and plants) harbor complex communities comprising diverse microbes, some more adapted to their host, others generalists, or transient, representing a broad spectrum of potential contributions (Shapira, 2016, Macke *et al.*, 2017a).

By mediating interactions between hosts and other organisms, such as parasites, symbionts that protect hosts from infection ("defensive" or "protective" microbes) can play a direct role in the process of host–parasite coevolution (Kwiatkowski *et al.*, 2012; Parker *et al.*, 2011; Ford and King, 2016). So far, most models and laboratory experiments investigating coevolution processes, such as the Red Queen Hypothesis, have been mainly based on pairwise-species interactions (Lively *et al.*, 2014). However, in natural environments a lot of factors, such as diverse species interaction networks may affect coevolution, and should thus be more systematically considered in future research (Koskella and Brockhurst, 2014, Betts *et al.*, 2016, Macke *et al.*, 2017b). Whole microbial communities and individual microbial symbiont species can be associated with specific host genotypes driving patterns of host–parasite interaction specificity (Koch and Paul Schmid-Hempel, 2012; Rouchet and Vorburger, 2012), a fundamental assumption of coevolutionary dynamics (Agrawal and Lively, 2002). If parasites adapt to these genotype-specific microbes then defensive symbionts have the potential to mediate host–parasite coevolutionary interactions (Kwiatkowski *et al.*, 2012) and even evolve themselves (Ford *et al.*, 2016a,b). By integrating defensive microbial symbionts into host–parasite interactions at the mechanistic level with evolutionary theory, Ford and King (2016) predict how defensive microbes might alter the evolution of host and parasite traits, such as resistance and virulence, which in turn might greatly affect host population dynamics. First, direct coevolution between defensive microbes and parasites would provide "real time" control of the infection, whereby evolutionary changes in parasites are met by rapid reciprocal evolution in defensive microbes (Ford *et al.*, 2016a). Second, given that defensive microbes protect hosts from parasite-induced fitness costs, they could reduce selection for costly immune or behavioral defense mechanisms in the host. Over evolutionary time, a host may thus become dependent on microbe-mediated protection, a hypothesis that has been invoked to explain the loss of immune genes in pea aphids and honeybees (Gerardo *et al.*, 2010; Kaltenpoth and Engl, 2014). Finally, defensive symbionts can shape the evolution of parasite virulence (Ford *et al.*, 2016b), through mechanisms similar to interactions occurring between co-infecting parasites, such as resource competition, interference competition or immune mediation (Macke *et al.*, 2017a).

These observations can in principle be extended to interactions other than host–parasite coevolution, such as plant–insect coevolution, and may be valuable for a more realistic understanding of coevolutionary processes. Through the horizontal transfer of bacteria, species or populations of the same species can affect the fitness, and thus potentially the evolution, of each other (Feldhaar, 2011). Gut symbionts can also act as ecosystem engineers and contribute to modifying the biotic and abiotic environment of their host, potentially affecting other species of the community. For example, by contributing to food digestion, gut symbionts play a major role in the food web and can contribute to the stability of the whole community.

See also: Behavioral Ecology: Environmental Stress and Evolutionary Change. Ecological Data Analysis and Modelling: Metapopulation Models. Ecological Processes: Evolution of Parasitism. Evolutionary Ecology: Coevolution; Eco-Evolutionary Dynamics; Microbiomes and Holobionts; Eco-Immunology: Past, Present, and Future. General Ecology: Parasites; Ecophysiology. Global Change Ecology: Xenobiotic (Pesticides, PCB, Dioxins) Cycles

References

Aalto, S., Decaestecker, E., Pulkkinen, K., 2015. A three-way perspective of stoichiometric changes on host–parasite interactions. Trends in Parasitology 31, 333–340.
Agrawal, A.F., Lively, C.M., 2001. Parasites and the evolution of self-fertilization. Evolution 55, 869–879.
Agrawal, A., Lively, C.M., 2002. Infection genetics: Gene-for-gene versus matching-alleles models and all points in between. Evolutionary Ecology Research 4, 91–107.
Becks, L., Ellner, S.P., Jones, L.E., Hairston, N.G., 2012. The functional genomics of an eco-evolutionary feedback loop: Linking gene expression, trait evolution, and community dynamics. Ecology Letters 15, 492–501.
Bell, G., 1982. The masterpiece of nature: The evolution and genetics of sexuality. Berkeley: University of California Press.
Betts, A., Rafaluk, C., King, K.C., 2016. Host and parasite evolution in a tangled bank. Trends in Parasitology 32, 863–873.
Carius, H.J., Little, T., Ebert, D., 2001. Genetic variation in a host–parasite association: Potential for coevolution and frequency-dependent selection. Evolution 55, 1146–1152.
De Meester, L., Vanoverbeke, J., Kilsdonk, L.J., Urban, M.C., 2016. Evolving perspectives on monopolization and priority effects. Trends in Ecology & Evolution 31, 136–146.
Decaestecker, E., Vergote, A., Ebert, D., De Meester, L., 2003. Evidence for strong host clone-parasite species interactions in the Daphnia microparasite system. Evolution 57, 784–792.
Decaestecker, E., Gaba, S., Raeymaekers, J.A.M., Stoks, R., Van Kerckhoven, L., Ebert, D., De Meester, L., 2007. Host–parasite 'Red Queen' dynamics archived in pond sediment. Nature 450, 870–874.
Decaestecker, E., De Meester, L., Mergeay, J., 2009. Cyclical parthenogenesis in Daphnia: Sexual versus asexual reproduction. In: Schön, I., Martens, K., van Dijk, P. (Eds.), Lost sex. The evolutionary biology of parthenogenesis. Dordrecht, Heidelberg, London, New York: Springer, pp. 295–316.
Decaestecker, E., De Gersem, H., Michalakis, Y., Raeymaekers, J.A.M., 2013. Damped long-term host–parasite Red Queen coevolutionary dynamics: A reflection of dilution effects? Ecology Letters 16, 1455–1462.
Duffy, M.A., Sivars-Becker, L., 2007. Rapid evolution and ecological host–parasite dynamics. Ecology Letters 10, 44–53.
Duffy, M.A., Hall, S.R., Tessier, A.J., Huebner, M., 2005. Selective predators and their parasitized prey: Are epidemics in zooplankton under top-down control? Limnology and Oceanography 50, 412–420.
Ebert, D., 2005. Ecology, epidemiology, and evolution of parasitism in Daphnia [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information, Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books.
Ebert, D., 2011. A genome for the environment. Science 331, 539–540.
Feldhaar, H., 2011. Bacterial symbionts as mediators of ecologically important traits of insect hosts. Ecological Entomology 36, 533–543.

Ford, S.A., King, K.C., 2016. Harnessing the power of defensive microbes: Evolutionary implications in nature and disease control. PLoS Pathogens 12.e1005465.
Ford, S.A., Williams, D., Paterson, S., King, K.C., 2016a. Co-evolutionary dynamics between a defensive microbe and a pathogen driven by fluctuating selection. Molecular Ecology. doi:10.1111/mec.13906.
Ford, S.A., Kao, D., Williams, D., King, K.C., 2016b. Microbe-mediated host defence drives the evolution of reduced pathogen virulence. Nature Communications 7.13430
Fussman, G.F., Loreau, M., Abrams, P.A., 2007. Eco-evolutionary dynamics of communities and ecosystems. Functional Ecology 21, 465–477.
Gandon, S., Ebert, D., Olivieri, I., Michalakis, Y., 1998. Differential adaptation in spatially heterogeneous environments and host-parasite coevolution. In: Mopper, S., Strauss, S. Y. (Eds.), Genetic structure and local adaptation in natural insect populations: Effects of ecology, life history and behavior. New York: Chapman and Hall, pp. 325–342.
Gandon, S., Buckling, A., Decaestecker, E., Day, T., 2008. Host-parasite coevolution and patterns of adaptation across time and space. Journal of Evolutionary Biology 21, 1861–1866.
Gerardo, N.M., Altincicek, B., Anselme, C., Atamian, H., Barribeau, S.M., De Vos, M., Duncan, E.J., Evans, J.D., Gabaldón, T., Ghanim, M., Heddi, A., Kaloshian, I., Latorre, A., Moya, A., Nakabachi, A., Parker, B.J., Pérez-Brocal, V., Pignatelli, M., Rahbé, Y., Ramsey, J.S., Spragg, C.J., Tamames, J., Tamarit, D., Tamborindeguy, C., Vincent-Monegat, C., Vilcinskas, A., 2010. Immunity and other defenses in pea aphids, *Acyrthosiphon pisum*. Genome Biology 11, R21.
Haag, C., Ebert, D., 2004. Parasite-mediated selection in experimental metapopulations of *Daphnia magna*. Proceedings of the Royal Society of London B: Biological Sciences 271, 2149–2155.
Hairston, N.G., Ellner, S.P., Geber, M.A., Yoshida, T., Fox, J.A., 2005. Rapid evolution and the convergence of ecological and evolutionary time. Ecology Letters 8, 1114–1127.
Hall, S.R., Tessier, A.J., Duffy, M.A., Huebner, M., Cáceres, C.E., 2006. Warmer does not have to mean sicker: temperature and predators can jointly drive timing of epidemics. Ecology 87 (7), 1684–1695.
Hanski, I., Simberloff, D., 1997. The metapopulation approach, its history, conceptual domain and application to conservation. In: Hanski, I., Gilpin, M. (Eds.), Metapopulation biology: Ecology, genetics and evolution. London: Academic Press, pp. 5–26.
Hendry, A.P., 2016. Eco-evolutionary dynamics. New Jersey, Oxfordshire: Princeton University Press.
Hiltunen, T., Becks, L., 2014. Consumer co-evolution as an important component of the eco-evolutionary feedback. Nature Communications 5, 5226.
Houwenhuyse, S., Make, E., Bulteel, L., Decaestecker, E., 2017. Back to the future in a petri dish: Origin and impact of resurrected microbes in natural populations. Evolutionary Applications 11 (1), 29–41.
Jaenike, J., 1978. A hypothesis to account for the maintenance of sex within populations. Evolutionary Theory 3, 191–194.
Jokela, J., Dybdahl, M.D., Lively, C.M., 2009. The maintenance of sex, clonal dynamics, and host-parasite coevolution in a mixed population of sexual and asexual snails. The American Naturalist 174, S43–S53.
Jousimo, J., Tack, A.J.M., Ovaskainen, O., Mononen, T., Susi, H., Tollenaere, C., Laine, A., 2014. Disease ecology. Ecological and evolutionary effects of fragmentation on infectious disease dynamics. Science 344, 1289–1293.
Kaltenpoth, M., Engl, T., 2014. Defensive microbial symbionts in Hymenoptera. Functional Ecology 28, 315–327.
Kerfoot, W.C., Weider, L.J., 2004. Experimental paleoecology (resurrection ecology): Chasing van Valen's Red Queen hypothesis. Limnology and Oceanography 49, 1300–1316.
King, K.C., Delph, L.F., Jokela, J., Lively, C.M., 2009. The geographic mosaic of sex and the Red Queen. Current Biology 19, 1438–1441.
King, K.C., Jokela, J., Lively, C.M., 2011a. Trematode parasites infect or die in snail hosts. Biology Letters 7, 265–268.
Koch, H., Paul Schmid-Hempel, P., 2012. Gut microbiota instead of host genotype drive the specificity in the interaction of a natural host-parasite system. Ecology Letters 15, 1095–1103.
Koskella, B., Brockhurst, M.A., 2014. Bacteria–phage coevolution as a driver of ecological and evolutionary processes in microbial communities. FEMS Microbiology Reviews 38, 916–931.
Koskella, B., Lively, C.M., 2009. Evidence for negative frequency-dependent selection during experimental coevolution of a freshwater snail and a sterilizing trematode. Evolution 63, 2213–2221.
Kwiatkowski, M., Engelstädter, J., Vorburger, C., 2012. On genetic specificity in symbiont-mediated host-parasite coevolution. PLoS Computational Biology 8.e1002633
Lafferty, K.D., Dobson, A.P., Kuris, A.M., 2006. Parasites dominate food web links. Proceedings of the National Academy of Sciences of the United States of America 103, 11211–11216.
Lennon, J.T., Martiny, J.B.H., 2008. Rapid evolution buffers ecosystem impacts of viruses in a microbial food web. Ecology Letters 11, 1178–1188.
Little, T.J., 2002. The evolutionary significance of parasitism: Do parasite-driven genetic dynamics occur ex silico? Journal of Evolutionary Biology 15, 1–9.
Lively, C.M., 1987. Evidence from a New Zealand snail for the maintenance of sex by parasitism. Nature 328, 519–521.
Lively, C.M., 1999. Migration, virulence, and the geographic mosaic of adaptation by parasites. The American Naturalist 153, S35–S47.
Lively, C.M., Jokela, J., 2002. Temporal and spatial distributions of parasites and sex in a freshwater snail. Evolutionary Ecology Research 4, 219–226.
Lively, C.M., Dybdahl, M.F., Jokela, J., Osnas, E.E., Delph, L.F., 2004. Host sex and local adaptation by parasites in a snail-trematode interaction. The American Naturalist 164, S6–S18.
Lively, C.M., de Roode, J.C., Duffy, M.A., Graham, A.L., Koskella, B., 2014. Interesting open questions in disease ecology and evolution. The American Naturalist 184, S1–S8.
Macke, E., Callens, M., De Meester, L., Decaestecker, E., 2017a. Host genotype-dependent gut microbiota drives zooplankton tolerance to toxic cyanobacteria. Nature Communications 8.Art. No. 1608.
Macke, E., Macke, E., Tasiemski, A., Massol, F., Callens, M., Decaestecker, E., 2017b. Life history and eco-evolutionary dynamics in light of the gut microbiota. Oikos 126, 508–531.
Matthews, B., Narwani, A., Hausch, S., Nonaka, E., Peter, H., Yamamichi, M., Sullam, K.E., Bird, K.C., Thomas, M.K., Hanley, T.C., Turner, C.B., 2011. Toward an integration of evolutionary biology and ecosystem science. Ecology Letters 14, 690–701.
Maynard Smith, J., 1989. Evolutionary genetics. Oxford: Oxford University Press.
Morgan, E.R., Jefferies, R., Krajewski, M., Ward, P., Shaw, S.E., 2009. Canine pulmonary angiostrongylosis: The influence of climate on parasite distribution. Parasitology International 58, 406–410.
Neiman, M., Jokela, J., Lively, C.M., 2005. Variation in asexual lineage in *Potamopyrgus antipodarium*, a New Zealand snail. Evolution 59, 945–1952.
Olivieri, I., Couvet, D., Gauyon, P.H., 1990. The genetics of transient populations: Research at the metapopulation level. Trends in Ecology & Evolution 5, 207–210.
Orsini, L., Schwenk, K., De Meester, L., Colbourne, J.K., Pfrender, M.E., Weider, L.J., 2013. The evolutionary time machine: Using dormant propagules to forecast how populations can adapt to changing environments. Trends in Ecology & Evolution 28, 274–282.
Otto, S.P., Michalakis, Y., 1998. The evolution of recombination in changing environments. Trends in Ecology & Evolution 13, 145–151.
Parker, B.J., Barribeau, S.M., Laughton, A.M., de Roode, J.C., Gerardo, N.M., 2011. Non-immunological defense in an evolutionary framework. Trends in Ecology & Evolution 26, 242–248.
Paterson, S., Vogwill, T., Buckling, A., Benmayor, R., Spiers, A.J., Thomson, N.R., Quail, M., Smith, F., Walker, D., Libberton, B., Fenton, A., Hall, N., Brockhurst, M.A., 2010. Antagonistic coevolution accelerates molecular evolution. Nature 464, 275–278.
Penczykowski, R.M., Hall, S.R., Civitello, D.V., Duffy, M.A., 2014. Habitat structure and ecological drivers of disease. Limnology and Oceanography 59, 340–348.
Rouchet, R., Vorburger, C., 2012. Strong specificity in the interaction between parasitoids and symbiont-protected hosts. Journal of Evolutionary Biology 25, 2369–2375.
Shapira, M., 2016. Gut microbiotas and host evolution: Scaling up symbiosis. Trends in Ecology & Evolution 31, 539–549.
Thompson, J.N., 1999. Specific hypothesis on the geographic mosaic of coevolution. American Naturalist 153, S1–14.
Thompson, J.N., 2005. Coevolution: The geographic mosaic of coevolutionary arms races. Current Biology 15, R992–R994.
Thompson, J.N., 2009. The coevolving web of life. The American Naturalist 173, 125–140.

Turcotte, M.M., Reznick, D.N., Daniel, H.J., 2013. Experimental test of an eco-evolutionary dynamic feedback loop between evolution and population density in the green peach aphid. The American Naturalist 181, S46–S57.
Urban, M., De Meester, L., 2009. Community monopolization: Local adaptation enhances priority effects in an evolving metacommunity. Proceedings of the Royal Society of London B: Biological Sciences B 276, 4129–4138.
Urban, M.C., De Meester, L., Vellend, M., Stoks, R., Vanoverbeke, J., 2012. A crucial step toward realism: Responses to climate change from an evolving metacommunity perspective. Evolutionary Applications 5, 154–167.
Vale, P., Little, T.J., 2009. Measuring parasite fitness under genetic and thermal variation. Heredity 103, 102–109.
Vale, P., Stjernman, M., Little, T.J., 2008. Temperature-dependent costs of parasitism and maintenance of polymorphism under genotype-by-environment interactions. Journal of Evolutionary Biology 21, 1418–1427.
Vavre, F., Kremer, N., 2014. Microbial impacts on insect evolutionary diversification: From patterns to mechanisms. Current Opinion in Insect Science 4, 29–34.
Wertheim, B., Vet, L.E.M., Dicke, M., 2003. Increased risk of parasitism as ecological costs of using aggregation pheromones: Laboratory and field study of *Drosophila-Leptopilina* interaction. Oikos 100, 269–282.
Wolinska, J., King, K.C., 2009. Environment can alter selection in host–parasite interactions. Trends in Parasitology 25, 236–244.
Wolinska, J., Bittner, K., Ebert, D., Spaak, P., 2006. The coexistence of hybrid and parental *Daphnia*: The role of parasites. Proceedings of the Royal Society of London B: Biological Sciences 273, 1977–1983.
Wood, C.B., Byers, J.E., Cottingham, K.L., Altman, I., Donahue, M.J., Blakeslee, A.M.H., 2007. Parasites alter community structure. Proceedings of the National Academy of Sciences of the United States of America 104, 9335–9339.

Further Reading

King, K.C., Delph, L.F., Jokela, J., Lively, C.M., 2011b. Coevolutionary hotspots and coldspots for host sex and parasite local adaptation in a snail–trematode interaction. Oikos 120, 1335–1340.

r-Strategists/*K*-Strategists☆

Jonathan M Jeschke, Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB), Berlin, Germany; Freie Universität Berlin, Institute of Biology, Berlin, Germany; and Berlin-Brandenburg Institute of Advanced Biodiversity Research (BBIB), Berlin, Germany
Wilfried Gabriel, Ludwig-Maximilians-University Munich, Planegg-Martinsried, Germany
Hanna Kokko, University of Zurich, Zurich, Switzerland

Glossary

Carrying capacity *K* The maximum number of individuals from a given population that a given environment can sustain under constant conditions.

Fast–slow continuum A continuum of species from fast to slow life histories, which are typically characterized by early reproduction (slow life histories: late reproduction), a short (long) interbirth interval, high (low) fecundity, small (large) offspring, small (large) adult body size and a short (long) lifespan.

Intrinsic growth rate *r* The difference between per capita birth rate and death rate at very low population densities.

***K*-selection** The *K*-endpoint of the *r*-*K* continuum, representing the qualitative extreme: density effects are maximal and the environment is saturated with organisms; competition is keen and the optimal strategy is to allocate resources to maintenance and the production of a few extremely fit offspring (based on Pianka, 1970).

Life history An organism's lifetime pattern of growth, differentiation, storage, reproduction, and survival. Life-history traits determine individual fitness and are often phenotypically plastic. Complex theories have been developed to predict how selection shapes life histories.

***r*-selection** The *r*-endpoint of the *r*-*K* continuum, representing the quantitative extreme: a perfect ecological vacuum, with no density effects and no competition; the optimal strategy is to allocate resources to reproduction, producing as many offspring as possible (based on Pianka, 1970).

Introduction

The concept of *r*-strategists and *K*-strategists lies at the interface between ecology and evolution. It was developed in the 1960s and 70s mainly by the three US-American scientists Robert H. MacArthur (1930–1972), Edward O. Wilson (1929–), and Eric R. Pianka (1939–). The concept was especially important in the 1970s. One short paper by Pianka from 1970 titled "On *r*- and *K*-selection" has been cited more than 3000 times according to Google Scholar. Although the concept as a whole is not seen as accurate anymore today, parts of it still are.

In this article, we outline the historical development of the *r*/*K* concept, followed by its problems as seen today. We then describe its aspects that are still in use, namely the observation that life histories show patterns within and among species and the idea that selection regimes vary with population density.

Historical Development of the *r*/*K* Concept

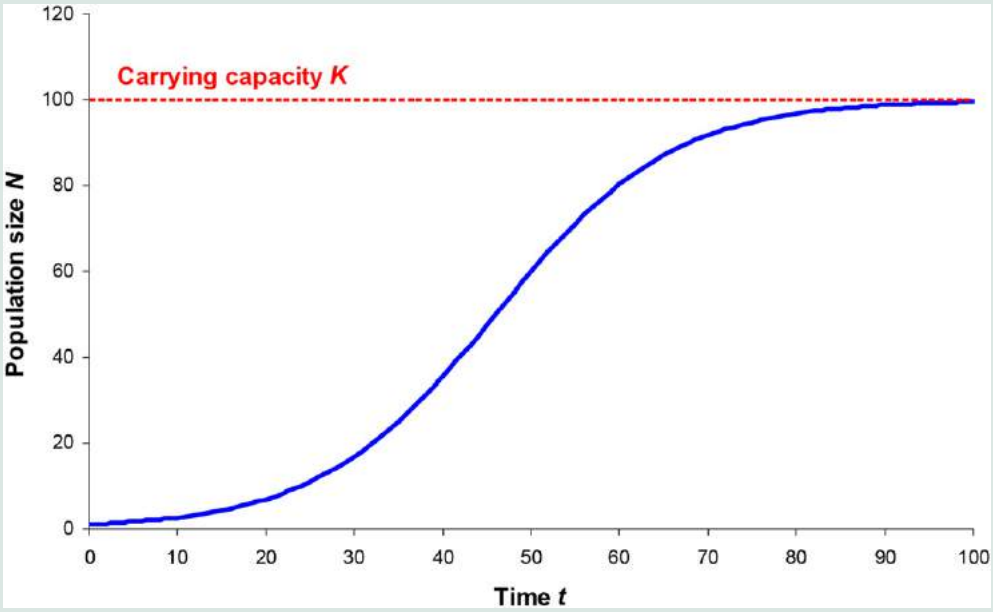
The *r*/*K* concept is based on the idea that environments differing in population abundance and fluctuation should select for different phenotypes. In a paper published in 1950, Theodosius Dobzhansky compared evolution in the tropics to evolution in temperate environments. The tropics are more stable and populated by different species than temperate environments, so "interrelationships between competing and symbiotic species become the paramount adaptive problem." (p. 220 in Dobzhansky, 1950) On the other hand, "Physically harsh environments, such as arctic tundras or high alpine zones of mountain ranges, are inhabited by few species of organisms. The success of these species in colonizing such environments is due simply to the ability to withstand low temperatures or to develop and reproduce during the short growing season." (p. 220).

The idea that environments differing in stability and population select for different phenotypes was formalized by MacArthur and Wilson (1967) in their landmark book "The theory of island biogeography" from 1967. In contrast to Dobzhansky (1950), however, MacArthur and Wilson did not look at the population of environments by different species (i.e., biodiversity) but at the population density of species. Given the title of their book, it is no surprise that they focused on species colonizing islands. They formally found that successful colonizers should have a high intrinsic growth rate *r*, which is the difference between per capita birth rate and death rate at very low population densities. Looking at empirical evidence, they concluded: "The evidence for birds and ants […] points to a preference for unstable, scattered habitats as a preadaptation to successful colonization" (p. 82). Regarding population persistence, they found that a

☆*Change History*: January 2018. Jonathan M Jeschke, Wilfried Gabriel, and Hanna Kokko updated the text and bibliography, and added the glossary and relevant websites.

This is an update of J.M. Jeschke, W. Gabriel and H. Kokko, r–Strategist/K–Strategists, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 3113–3122.

Box 1 The logistic equation with its parameters *r* and *K*



Visualization of the logistic equation $\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right)$, where N is population size, t is time, r is intrinsic growth rate, and K is carrying capacity. The time-discrete analog is $N_{t+1} = \frac{e^r N_t}{1+aN_t}$ with $a = \left(\frac{e^r - 1}{K}\right)$.

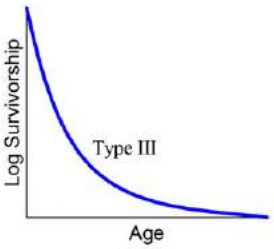
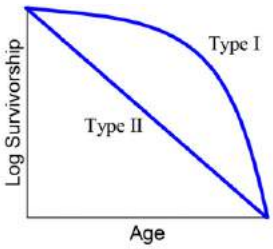
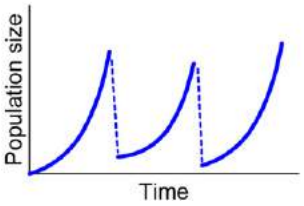
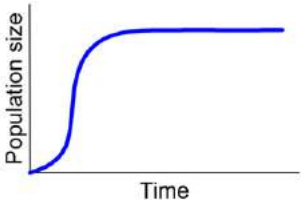
The parameter r, intrinsic growth rate, is the difference between per capita birth rate and death rate at very low population densities. It is part of the logistic equation and also of the Euler–Lotka equation: $1 = \sum_x e^{-rx} l_x m_x$, where l_x is the probability of surviving from birth to age x and m_x is the number of daughters per female at age x. The Euler–Lotka equation links the parameter r with the life history of individuals. Hence, r is an individual trait that can be selected.

The parameter K, carrying capacity, can however hardly be considered an individual trait. It is really not more than a parameter representing density dependence in the logistic equation. It is phenomenological, thus not directly biologically interpretable.

high carrying capacity K equals a long expected persistence time, where K is the number of individuals the island maximally can hold in equilibrium. The two parameters r and K form the basis of the logistic equation (Box 1). MacArthur and Wilson then extended their findings to populations beyond islands. Based on work by Fisher, Haldane, and Wright, it was already known that r generally is an appropriate measure of fitness at low and increasing population densities. MacArthur and Wilson added that K is an appropriate measure of fitness at high densities and accordingly coined the terms r-selection and K-selection: in fluctuating environments, populations are repeatedly diminished, so r-selection will dominate. In stable environments, on the other hand, populations will grow to a rather constantly high size where K-selection will dominate. They argued that r-selection tends to lead to "a shorter developmental time, a longer reproductive life, and greater fecundity, in that order of probability" (p. 157). In other words, r-selection should lead to high and fast productivity, whereas K-selection should lead to efficiency, especially of resource utilization.

In a short note published in 1970, Pianka made the connection between r-selection, K-selection, and life history more explicit and thereby gave the r/K concept its final form (**Table 1**). He wrote: "Certainly, no organism is completely 'r-selected' or completely 'K-selected,' but all must reach some compromise between the two extremes.[…] We can visualize an r-K continuum, and a particular organism's position along it. The r-endpoint represents the quantitative extreme – a perfect ecologic vacuum, with no density effects and no competition. Under this situation, the optimal strategy is to put all possible matter and energy into reproduction, with the smallest practicable amount into each individual offspring, and to produce as many total progeny as possible. Hence r-selection leads to high productivity. The K-endpoint represents the qualitative extreme – density effects are maximal and the environment is saturated with organisms. Competition is keen and the optimal strategy is to channel all available matter and energy into maintenance and the production of a few extremely fit offspring. Replacement is the keynote here. K-selection leads to increasing efficiency of utilization of environmental resources" (p. 592 in Pianka, 1970). This is the reasoning behind **Table 1**, so Pianka gave no formal justification of the attributes of r-strategists and K-strategists. Synonyms for r-strategists are r-selected species, fugitive species, and opportunistic species. For K-strategists, the synonyms K-selected species and stable species have been used. As typical r-strategists, Pianka mentioned insects (with the exception of 17-year cicadas and similar species), whereas vertebrates were typical K-strategists (except some amphibians). Within each taxon, of course, some species are more on the r-end of the continuum while others are more on the K-end.

Table 1 Attributes of *r/K*-strategists

	r-strategists	*K-strategists*
Climate	Variable and/or unpredictable	Fairly constant and/or predictable
Mortality	Often catastrophic, nondirected, density-independent	More directed, density-dependent
Survivorship		
Population size	Variable in time, nonequilibrium: 	Fairly constant, equilibrium: 
Intra- and interspecific competition	Variable, often lax	Usually keen
Life history	Rapid development	Slow development, greater competitive ability
	High r_{max}	Lower resource thresholds
	Early reproduction	Delayed reproduction
	Small body size	Large body size
	Semelparity: single reproduction	Iteroparity: repeated reproduction
	Short lifespan	Long lifespan

Problems of the *r/K* Concept

"The theory of *r*-selection and *K*-selection […] helped to galvanize the empirical field of comparative life-history and dominated thinking on the subject from the late 1960s through the 1970s. […] By the early 1980s, sentiment about the theory had changed so completely that a proposal to test it or the use of it to interpret empirical results would likely be viewed as archaic and naïve" (Reznick *et al.*, 2002, p. 1509). Similarly, Roff (2002, p. 79) wrote: "it may be preferable to avoid use of the terms [*r*- and *K*-selection] altogether." As these quotes show, the *r/K* concept was very important in the past but today, it is in its whole basically not used anymore. It has fallen into disfavor due to the recognition of several problems, in particular: (1) the concept's assumption of a trade-off between *r* and *K* is often not valid; (2) the parameter *K* is not directly biologically interpretable; and (3) the life-history traits attributed to *K*-strategists are not justified, neither theoretically nor empirically.

(1) The assumption of a trade-off between *r* and *K* is often not valid. The *r/K* concept assumes that *r*-selection and *K*-selection are in opposition although there is no logical necessity to this assumption. A trade-off between *r* and *K* has not often been found empirically, either. Although it has received some support in experiments with *Drosophila melanogaster*, this was not the case for experiments with *Escherichia coli*, the rotifer species *Asplanchna brightwellii*, or the cladoceran *Bosmina longirostris*.

(2) The parameter *K* is not directly biologically interpretable. While *r* is the difference between per capita birth rate and death rate at very low population densities and can be directly related to the life history of individuals, *K* is a quite complex parameter: it is meant to give the maximum number of individuals that a given environment can sustain under constant conditions. This phenomenological parameter cannot be determined in natural populations and is thus not directly biologically interpretable. In models, *K* is defined as the unstable or stable point of equilibrium where death rates equal birth rates ($dN/dt=0$ in time-continuous models, $N_{t+1}=N_t$ in time-discrete models). In real populations, such points of equilibrium are rarely constant over time. How *K* relates to life-history traits is indefinable, too. Stearns (1977, p. 155) wrote: "*K* is not a population parameter, but a composite of a population, its resources, and their interaction. Calling *K* a population trait is an artifact of logistic thinking, an example of Whitehead's Fallacy of Misplaced Concreteness. Thus *r* and *K* cannot be reduced to units of common currency." In other words, the *r/K* concept is comparing apples and oranges.

(3) The life-history traits attributed to *K*-strategists are not justified. There is no reason why species living in constant environments should have the combination of traits proposed by Pianka and reproduced in **Table 1**. The linkage between the environment and

the life history made by Pianka is at the heart of the *r*/*K* concept but has never been theoretically justified, neither by Pianka nor anyone else. To achieve a high *r*, a species can either maximize its birth rate and/or minimize its death rate, and the corresponding strategies will result in different life-history traits. This ambiguity questions the life-history attributes of *r*-strategists. But while these attributes can still be logically defended, the main reason for the traits of *K*-strategists seems to have been the intuitive assumption that they should be the opposite of those of *r*-strategists. As mentioned above, however, there is no necessary trade-off between *r* and *K*. When we take a closer look at the attributes, we may for example ask why, as claimed by Pianka, a population of large aggressive individuals should have a higher carrying capacity than a population of small peaceful individuals. Larger individuals need more resources than smaller ones, so a given amount of resources provided by the environment can be used either by a small number of large individuals or a large number of small individuals. Under many circumstances we can therefore expect a smaller carrying capacity for larger individuals, while the *r*/*K* concept claims the opposite. In defending the concept, we could reinterpret *K* and measure it in biomass rather than individuals. This trick does not help with the problem of aggressiveness, however. Intraspecific aggression should often lead to a smaller population size and thus a smaller carrying capacity, again in contrast to the *r*/*K* concept. Of course, there are situations where selection on individuals leads to at least a specific life-history stage being maximized (i.e., the number of individuals in it), but this is not interpretable as giving support to a broad-brush categorization of traits like in the Pianka scheme.

The linkage between the environment and life history made by Pianka also lacks empirical support. For example, when fruit flies (*Drosophila melanogaster*) were reared at low or high densities, the low-density lines evolved a higher capability to increase in population size at low densities but a lower capability to increase at high densities. In the high-density lines, the fly larvae were more competitive due to a higher feeding rate and pupation at a greater height above the medium compared to low-density lines. These experimental results are in accordance to the general predictions of the *r*/*K* concept about the differences between selection at low versus high densities. However, they are not in accordance to the explicit predictions about the linkage of these differences to specific life-history traits. The same is true for experiments with pitcher-plant mosquitoes (*Wyeomyia smithii*) where differences in population densities again led to differences in competitive ability but not to differences in life-history traits. Although some supportive empirical data do exist (Sæther *et al.*, 2016), a direct relationship between population density or fluctuation on the one hand and life-history characteristics on the other hand, as proposed by the concept, has not been generally established. It is true that life-history patterns exist, but apparently the *r*/*K* concept cannot explain them.

Aspects of the *r/K* Concept Used Today

As the last section has made clear, the reason why the *r*/*K* concept is included in today's ecological textbooks and this encyclopedia is not because it is still widely used or considered to be correct. The reasons are that the concept is historically important and that two of its aspects are still in use. The first one is that life-history traits show patterns within and among species: they do not vary randomly but correlate to each other, an observation that has led to the concept of fast and slow life histories. The second preserved aspect is the basic idea of Dobzhansky, MacArthur, and Wilson that heavily populated and stable environments select for different traits than less populated and fluctuating environments. We shall comment on both aspects below.

Fast and Slow Life Histories

Observed patterns

According to the concept of fast and slow life histories, species with fast (slow) life histories have certain life-history characteristics that are similar to those of *r*-strategists (*K*-strategists) in the *r*/*K* concept (**Table 2**). In contrast to the *r*/*K* concept, however, fast and slow life histories are primarily observed patterns and not necessarily connected to an explanation.

The fast–slow concept is based on the observation that certain life-history traits of species often correlate to each other in a similar way. These traits are age of first reproduction, interbirth interval, fecundity, offspring size, adult body size (usually measured as body mass), and lifespan (**Table 2**). Based on this suite of correlated traits, many species fit on a continuum from fast to slow life histories. For example, the bank vole (*Clethrionomys glareolus*) and the house mouse (*Mus musculus*) are mammals with fast life histories: they reproduce early, have a short interbirth interval, a high fecundity, small offspring, are also small as adults,

Table 2 Life histories of species on the fast–slow continuum with body size included

Fast	*Slow*
Early reproduction	Late reproduction
Short interbirth interval	Long interbirth interval
High fecundity	Low fecundity
Small offspring size	Large offspring size
Small adult body size	Large adult body size
Short lifespan	Long lifespan

Note: When body size is factored out, a second continuum can be observed among the remaining variables.

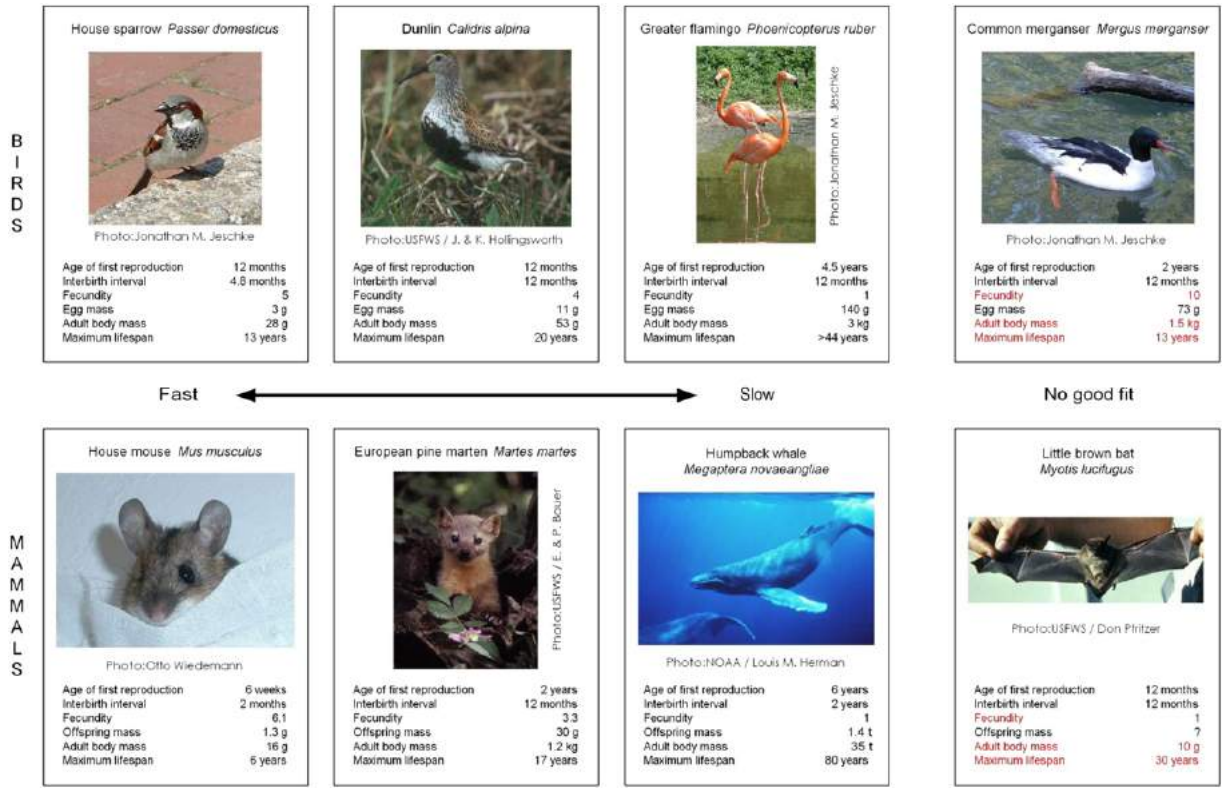


Fig. 1 Examples of species on the fast–slow continuum according to **Table 2** and of species that do not fit well on the continuum.

and have a short lifespan (**Fig. 1**, **Table 2**). The opposite is true for *Homo sapiens* and the humpback whale (*Megaptera novaeangliae*) which are mammals with slow life histories. Naturally, many species are on neither extreme end of the fast–slow continuum but in between, for example, the arctic hare (*Lepus arcticus*) or the European pine marten (*Martes martes*) (**Fig. 1**). The position of a species on the continuum depends on the focal taxonomic level. For example, on the level of all mammals, the bat-eared fox (*Otocyon megalotis*) has a rather slow life history, but on the level of carnivores, it has a relatively fast life history. Within birds, examples for species with fast life histories include the blue tit (*Parus caeruleus*) and the house sparrow (*Passer domesticus*). Birds on the slow end of the continuum include the California condor (*Gymnogyps californianus*) and the greater flamingo (*Phoenicopterus ruber*). Finally, examples for birds with intermediate life histories are the dunlin (*Calidris alpina*) and the magpie (*Pica pica*) (**Fig. 1**).

Of course, the fast–slow continuum cannot explain all the variation. Many species do not neatly fit on the continuum, for example, the little brown bat (*Myotis lucifugus*) which has an exceptionally low fecundity and long lifespan for its body size (**Fig. 1**). Other bat species share this deviation from the overall mammalian pattern, however, so the little brown bat is not exceptional within bats. The European rabbit (*Oryctolagus cuniculus*), on the other hand, has a too short interbirth interval for its size to neatly fit on the overall fast–slow mammal continuum: its average adult body mass is 1.6 kg and its interbirth interval 2.4 months. Within birds, the common merganser (*Mergus merganser*) has an exceptionally high fecundity and short lifespan for its size (**Fig. 1**), while Costa's hummingbird (*Archilochus costae*) has an exceptionally low fecundity (2 eggs) and long interbirth interval (12 months) for its size (average adults only weigh about 3 g). Many other hummingbirds and very small organisms in general seem to more often deviate from the fast–slow continuum than medium-sized and large organisms, which might be caused by size constraints at the low end of the body size distribution: when you are tiny, it may, for example, be hard to produce a huge number of offspring simultaneously, so your fecundity may be limited.

Thus far, we focused on the fast–slow continuum according to the traits given in **Table 2**. When body size is factored out, a second fast–slow continuum can be found among the remaining variables: age of first reproduction, interbirth interval, fecundity, offspring size, and lifespan. The two fast–slow continua are different, so some species that have a fast life history when body size is included have a slower life history when body size is factored out and vice versa. Also, some species fit better on the continuum when body size is factored out and vice versa. For example, the European rabbit fits well on the fast end of the continuum when body size is factored out, while it does not fit well on the continuum when body size is included (see above). Both fast–slow continua share the same name and are thus rarely discriminated, which is fine in some contexts. But the discrimination would be necessary in other contexts. In this article, we refer to "fast–slow continua" when we consider the discrimination not necessary but otherwise indicate the continuum we are referring to.

The existence of fast–slow continua has been established in vertebrates by a number of studies. For invertebrates, less information on life histories is available which makes comprehensive tests difficult. Furthermore, the concept of fast and slow life

histories has mainly attracted the attention of zoologists, so tests beyond the kingdom of Animalia are relatively rare as well. Particularly unknown is the general applicability of the concept to intraspecific differences in life history.

It is also uncertain whether and, if yes, how fast–slow continua are connected to other species traits. For example, it has been proposed that species with fast life histories have higher population densities than species with slow life histories, but empirical data have been inconclusive so far. In birds, there is some evidence that fast species have more variable population dynamics than slow species. The fast–slow continua may be helpful in conservation biology, for species with slow life histories tend to be more endangered of extinction (Jeschke and Strayer, 2008). On the other hand, nonnative species with a fast life history do not seem to have a higher invasion success than nonnative species with a slow life history, although this is frequently assumed (Jeschke and Strayer, 2008).

Some researchers have combined the fast–slow concept with other concepts, for example, the one of wasteful and frugal strategies: wasteful species are those that are adapted to good environmental conditions; they have a low production efficiency but a high mass-specific metabolic rate, and grow fast. Frugal species, on the other hand, are adapted to poor environmental conditions; they have a high production efficiency but a low mass-specific metabolic rate, and grow slowly. As fast and slow, wasteful and frugal are considered to be two ends of a continuum. Combining wasteful–frugal with fast–slow leads to a square with the four corners wasteful-fast, wasteful-slow, frugal-fast, and frugal-slow.

The concept of fast and slow life histories describes patterns but does not, as such, link the patterns to an explanation. The observation certainly hints at a trade-off between traits: it appears, for example, easier to achieve high fecundity at the expense of lifespan, or vice versa, than to maximize both simultaneously. The challenge of life-history theory is therefore to identify the important trade-offs and to explain why certain environmental conditions favor particular solutions along them. Most explanations have focused on explaining a single trait, for example, clutch size, but other approaches have focused on suites of traits, for example, fast or slow life histories. One of these approaches was the r/K concept but it failed in its pure form, as outlined above. Another approach are age-specific demographic models in which fluctuating juvenile mortalities select for slow traits, for example, delayed reproduction or low reproductive effort, while fluctuating adult mortality leads to fast traits. These models did originally not include density-dependent selection but have later been extended to do so. They now consider density in a much more precise way than the r/K concept: for example, a general result by Charlesworth (1980) shows that natural selection will under very general conditions maximize the number of individuals in the life stage that is subject to density dependence, and this obviously corresponds to increasing K if the population consists of identical individuals. Although models have usually focused on a single trait, they may also have the potential to explain suites of traits, as we will outline below. Before, we shall explain allometric scaling models that also shed light on fast and slow life histories.

Allometric models

Allometric models try to explain one or more of the many allometric relationships that have been found among species traits. An allometric relationship between x and y can be expressed in power form as $y = a\,x^b$ or logarithmically as $\log y = \log a + b \log x$, where a is a constant and b is a scaling exponent. If b equals 1, the relationship is called isometric. The interest of science in allometry dates back to the 17th century when Galileo noted that large animals have limb bones that are proportionally thicker than small animals. In other words, b is larger than 1 in this case which is called a positive allometry. Negative allometries, on the other hand, are those where b is smaller than 1. An example is the relationship between metabolic rate and body mass which has been found about two centuries ago. There has been much debate about the "true" value of the scaling exponent b in this relationship: whether it is 2/3 or 3/4 or neither of these two attractive numbers. It is undoubted, however, that b is smaller than 1, so although heavy animals have a higher metabolic rate than lighter animals in absolute terms, their metabolic rate per g-body mass is lower than that of lighter animals. For example, an average adult humpback whale weighs about 2 million times more than an average adult house mouse (**Fig. 1**), but its metabolic rate is not 2 million times higher than that of the house mouse. Per g-body mass, the humpback whale's metabolic rate is much lower than that of the house mouse. Besides the thickness of limb bones and the metabolic rate, there are many other species traits that scale allometrically to body mass. Originally, only single relationships have been addressed by allometric models, but their use now has been extended to suites of relationships, for example, fast or slow life histories. Allometric models now connect the traits body mass, age of first reproduction, interbirth interval, and lifespan.

The allometric models developed by James H. Brown and colleagues are currently the best known. Their so-called "metabolic theory" (Brown *et al.*, 2004) looks at the structure of biological networks, for example, blood vessels in vertebrates, and assumes that: (a) these networks are space-filling and branch hierarchically to supply all parts of the body; (b) the terminal tubes of these fractal networks do not vary with body size; (c) metabolic rate equals the rate at which the networks transport resources; and (d) evolution has minimized the time and energy needed for this transport. Based on these assumptions, the theory predicts b to be 3/4 for the scaling of metabolic rate to body mass, a prediction that appears to match empirical data, although this is controversial (see above). Because the mass-specific rate of metabolism is the metabolic rate divided by body mass, it is predicted to scale to body mass with $b = -1/4$. Brown *et al.* wrote: "the metabolic rate is the fundamental biological rate, because it is the rate of energy uptake, transformation, and allocation" (Brown *et al.*, 2004, p. 1772). The theory therefore predicts that other biological rates scale to body mass with $b = -1/4$ as well, for example, heart rate, which again seems to match empirical data. Furthermore, because times are the reciprocals of rates, biological times are predicted to scale to body mass with $b = 1/4$, for example, time to maturity, interbirth interval, or lifespan which are relevant to the fast–slow concept. Here again, predictions appear to match data. The basic models of this theory apply to endotherms only, but later models include temperature as the second determinate besides body mass and hence also apply to ectotherms.

The metabolic theory is broader than depicted here and was not specifically designed to explain fast or slow life histories. It has several limitations and weaknesses. For example, the network structure assumed by the theory does not match all network structures realized in nature, for example, the insect tracheal system, but the theory nonetheless correctly predicts metabolic rates of such organisms. Critiques say this suggests that another mechanism than the one proposed by the metabolic theory causes the scaling of metabolic rate to body mass. As already mentioned above, critiques also question that the quarter-power scaling predicted by the theory (that b is a multitude of 1/4) is empirically as universal as claimed by Brown *et al.* Depending on the statistical method, multitudes of 1/3 are observed, too, and other numbers in between as well. Kozłowski and Konarzewski questioned the consistency of the theory. To make it consistent, they argued, either metabolic rate had to scale isometrically to body mass or the assumption that the terminal network tubes are size-invariant needed to be relaxed. The first option does not agree to empirical data, but Kozłowski and Konarzewski cite empirical studies that challenge the assumption of size-invariance of the terminal tubes. Opponents of the theory have raised more critical points but an extensive discussion is beyond the scope of this article.

Various other models have been developed by different researchers to explain the allometric relationship between metabolic rate and body mass. These models differ in their assumptions and math but combined with the above-mentioned reasoning by Brown *et al.*, they all offer a linkage between body mass and—via metabolism—age of first reproduction, interbirth interval, and lifespan.

Density-Dependent Selection

Scientists today agree that environments differing in population abundance or fluctuation select for different strategies. This insight is, besides fast–slow continua, the second aspect that has been preserved from the r/K concept. An empirical example comes from the island of St. Kilda, Scotland, where Soay sheep (*Ovis aries*) with a darker coat had a higher survival rate at high densities compared to sheep with a lighter coat. Similarly, sheep with unscurred horns survived better at intermediate densities, while sheep with scurred horns did better at high densities. Other empirical examples for density-dependent selection come from *Drosophila* flies reared in the laboratory, tadpoles in temporary ponds, guppies (*Poecilia reticulata*) in Trinidad, or great tits (*Parus major*) in the Netherlands.

The r/K concept suffered from not deriving an explicit (and correct) link between density-dependent selection and the life-history traits favored by selection. In modern models that consider density-dependent regulation, such a link is made specific: an increase in population density affects the life history which feeds back negatively on population growth. For example, an increase in population density could lead to a lower fecundity. Which trait or traits change and how they change exactly depends on the nature of the specific model.

The procedure used in such models is clear in principle: a population consisting of individuals with particular life-history traits will obey characteristic fluctuations that depend on the traits themselves but also on the resources available (resource use being perhaps a function of the life-history traits in question), as well as any environmental fluctuations. An alternative life history can invade if it enters the population in a way that eventually creates more descendants than what the original strategy was able to achieve. "Eventually" here refers to the fact that the number of descendants left in one generation is not necessarily maximized; for example, if fast reproduction is essential, then a strategy that leaves few but early offspring may eventually win over others. Counting the "eventual" offspring is made mathematically precise by calculating the so-called invasion theta, or Lyapunov exponent, of the linearized system describing the invasion of a population.

Calculating the Lyapunov exponent is not a trivial procedure, however: one can liken it to a tool that is so general that it sometimes resembles a hammer while at other times it provides the scissors needed to dissect a given problem. Put more precisely, the effect that density regulation has on the optimal life history has been shown to depend strongly on the exact type of life cycle and on the way stochasticity is incorporated. It would consequently be useful to know what shape this tool takes for biologically relevant questions, such as the traits that underlie fast–slow continua. Some progress has been made in this direction. An important paper published by Mylius and Diekmann (1995) showed that there are conditions under which natural selection will always favor strategies that maximize r: an example is density regulation that increases the mortality of all age classes. Density-dependent juvenile mortality, on the other hand, implies that the correct measure of fitness is R_0, the expected lifetime reproductive success. However, for many situations neither r nor R_0 are correct measures of fitness and there is then no other escape than to have exact knowledge of the type of density regulation present—or at least make the assumption explicit—while deriving the invasion prospects by calculating the demographic consequences of each strategy.

That the mathematics of fitness measurements produces such nontrivial conclusions regarding individual fitness estimation is bad news for empiricists, who are rarely blessed with easy estimates of density regulation in the populations they study. In terms of ease of calculation, the two most practical measures of individual fitness are λ, which is the discrete-time version of r ($\lambda = e^r$ if λ and r are measured at equivalent time scales), and the lifetime reproductive success R_0, which simply equals the number of offspring and is often abbreviated as LRS. A typical pragmatic approach is to draw conclusions based on detailed analysis of one measure or to calculate both λ and LRS and compare the results. There are also examples where more detailed data on individual performance in density-regulated populations allows building tailor-made models, giving therefore much more power to draw inferences regarding optimal strategies in a particular species. However, it has also been argued that for many practical applications, simple measures such as the lifetime reproductive success perform quite well.

It is probably fair to say that theoretical work is currently better equipped to answer questions about a single trait, for example, age of first reproduction, than about suites of traits such as fast or slow life histories. Clearly, trade-offs between current and future investment must lie behind such suites but we do not know enough about how they evolve exactly according to prevailing environmental conditions, including the different stochastic components of life cycles, as exemplified by those determining recruitment of birds into territorial populations. Some theoretical results exist where variable nonequilibrium dynamics select for

slower rather than faster life history, and they clearly pose a puzzle to be solved in this context. The solution may lie in the fact that detailed knowledge of the density regulation really matters. Indeed, alternative assumptions concerning variability and temporal correlations of vital rates lead easily to opposite conclusions regarding the effect of variability on the "speed" of a life history. Thus, it is unclear whether or under which conditions different types of density regulation evolve and how density regulation is related to life-history traits. Future work in this area might help to find explanations for observed life-history patterns.

Summary

The concept of *r*-strategists and *K*-strategists links population dynamics to life history: strongly fluctuating environments lead to a strongly fluctuating population density that is low on average. According to the concept, these circumstances select for a high intrinsic growth rate *r* which is achieved by a distinctive life-history strategy consisting of rapid development, a small body size, early reproduction, semelparity, and a short lifespan. Species with these characteristics are called *r*-strategists. On the other hand, relatively constant environmental conditions allow a population to reach its carrying capacity *K* and thus a high average population density. The concept says that these circumstances select for a high *K* which is achieved by slow development (associated with great competitive ability), a large body size, delayed reproduction, iteroparity, and a long lifespan. Species with these characteristics are called *K*-strategists. The concept proposes to classify natural species on a continuum from *r*-strategists to *K*-strategists. It was developed from the 1960s to the 70s and very popular at that time. However, its popularity has vanished due to the recognition of serious problems, for example, the lack of a theoretical or empirical justification for the proposed life-history traits of *K*-strategists.

The *r*/*K* concept nonetheless is not only of historical importance, for two of its aspects are still in use today: the first one is the observation that life histories show patterns within and among species, and the second is that selection regimes vary with population density. The best known life-history patterns are continua from fast to slow life histories: species with a fast (slow) life history have characteristics similar to those that were proposed for *r*-strategists (*K*-strategists). Explanations for fast and slow life histories offered by the literature include allometric scaling and demographic models. The latter often include density-dependent selection in a way that is much more precise than in the *r*/*K* concept.

See also: Behavioral Ecology: Competition. Ecological Processes: Allometric Theory: Extrapolations From Individuals to Ecosystems. Evolutionary Ecology: Colonization; Pioneer Species; Life-History Patterns. General Ecology: Carrying Capacity; Carrying Capacity; Succession; Demography; Metabolic Theories in Ecology: The Dynamic Energy Budget Theory and the Metabolic Theory of Ecology

References

Brown, J.H., Gillooly, J.F., Allen, A.P., Savage, V.M., West, G.B., 2004. Toward a metabolic theory of ecology. Ecology 85, 1771–1789.
Dobzhansky, T., 1950. Evolution in the tropics. American Scientist 38, 209–221.
Charlesworth, B., 1980. Evolution in age-structured populations. Cambridge: Cambridge University Press.
Jeschke, J.M., Strayer, D.L., 2008. Are threat status and invasion success two sides of the same coin? Ecography 31, 124–130.
MacArthur, R.H., Wilson, E.O., 1967. The theory of island biogeography. Princeton: Princeton University Press.
Mylius, S.D., Diekmann, O., 1995. On evolutionarily stable life histories, optimization and the need to be specific about density dependence. Oikos 74, 218–224.
Pianka, E.R., 1970. On *r*- and *K*-selection. American Naturalist 104, 592–597.
Reznick, D., Bryant, M.J., Bashey, F., 2002. *r*- and *K*-selection revisited: The role of population regulation in life-history evolution. Ecology 83, 1509–1520.
Roff, D.A., 2002. Life history evolution. Sunderland: Sinauer.
Sæther, B.-E., Visser, M.E., Grøtan, V., Engen, S., 2016. Evidence for *r*- and *K*-selection in a wild bird population: A reciprocal link between ecology and evolution. Proceedings of the Royal Society B 283.20152411.
Stearns, S.C., 1977. The evolution of life history traits: A critique of the theory and a review of the data. Annual Review of Ecology and Systematics 8, 145–171.

Further Reading

Jeschke, J.M., Kokko, H., 2009. The roles of body size and phylogeny in fast and slow life histories. Evolutionary Ecology 23, 867–878.
Kozłowski, J., Konarzewski, M., 2004. Is West, Brown and Enquist's model of allometric scaling mathematically correct and biologically relevant? Functional Ecology 18, 283–289.
Promislow, D.E.L., Harvey, P.H., 1990. Living fast and dying young: A comparative analysis of life-history variation among mammals. Journal of Zoology 220, 417–437.
Reynolds, J.D., 2003. Life histories and extinction risk. In: Blackburn, T.M., Gaston, K.J. (Eds.), Macroecology: Concepts and consequences. Oxford: Blackwell, pp. 195–217.

Relevant Websites

http://genomics.senescence.info/species/—AnAge: Animal Ageing and Longevity Database.
http://animaldiversity.org—Animal Diversity Web.
http://esapubs.org/archive/ecol/E088/096/—Avian body sizes in relation to fecundity, mating system, display behavior, and resource sharing.

www.compadre-db.org—COMADRE: a global data base of animal demography.
https://ecologicaldata.org—Ecological Data Wiki.
www.fishbase.org—FishBase.
www.freshwaterplatform.eu—Freshwater Information Platform.
www.demogr.mpg.de/longevityrecords/—Longevity Records.
http://esapubs.org/archive/ecol/E090/184/—PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals.
www.vertnet.org—VertNet.

Units of Selection

M Shpak, University of Texas at El Paso, El Paso, TX, USA

Natural and artificial selection act on the heritable variation in traits that correlate with fitness (defined in terms of viability and fecundity of the entity in question). Biological entities replicate and survive at different rates and probabilities, and when these rates and probabilities are correlated with the measured values of heritable traits, those trait values will evolve through the action of selection.

The most obvious and familiar instance of the process involves selection on differential fitness between individual organisms in a population and the subsequent change in phenotype or genotype frequencies. However, because selection acts on any form of covariation between fitness and heritable traits, it has been argued that entities above and below the level of the individual organism (or genome) in a population can be units of selection. Selfish genetic elements which increase in frequency at the expense of the fitness of the genome represent units of selection below the individual, while various interpretations of kin, interdemic, and species selection suggest at least the theoretical possibility of selection at levels above the individual genotype.

To see how a self-reproducing entity at any level of organization can (at least in theory) respond to selection, it is instructive to look at the equations for genic selection as special cases of a more general expression. The standard Fisher–Wright equation for the selection on the frequency of the ith allele p_i at a diploid locus is

$$\Delta p_i = \frac{p(W_i^* - \overline{W})}{\overline{W}} = \frac{p_i(1-p_i)}{2\overline{W}}\frac{\partial \overline{W}}{\partial p_i} \quad [1]$$

where $\overline{W}$ is the mean and W^* is the marginal fitness at the ith allele; $\overline{W} = \sum_i p_i W_i^*$, $W_i^* = \sum_j p_j W_{ij}$. To see how this can be derived from a more inclusive model of selection, we note that for any trait X (which can be the measured value of a gene, genotype, phenotype, higher moments of trait or frequency values), the rate of change in the mean value of X can be described by Price's theorem:

$$\Delta \overline{X} = \frac{1}{\overline{W}}(\mathrm{cov}[W, X] + E[W\overline{\delta}]) \quad [2]$$

In the second term on the right-hand side, $\overline{\delta}$ is a measure of average deviation between an individual's (or any reproducing entity's) phenotype X and that of its offspring (due to mutation, recombination, or any number of other factors). In the absence of this second term, from the definition of covariance and regression,

$$\Delta \overline{X} = \frac{1}{\overline{W}}(\mathrm{cov}[W, X]) = \frac{1}{\overline{W}}\beta_{W,X}\sigma_X^2 \quad [3]$$

where $\beta_{W,X} = \partial \overline{W}/\partial \overline{X}$ (the regression coefficient of trait value X on fitness W, i.e. the slope of predicted mean value of W given a mean trait value X) and σ_X^2 is the variance in trait X. In previous example, X is the value (0 or 1) of an allele with two states, 0 (absence) and 1 (presence), $\overline{X} = p$, the variance in X is the variance of a Bernoulli random variable, $p(1-p)$. Making the substitutions, one obtains the Fisher–Wright equation.

Note that in deriving eqn [1] from [2], it was assumed that the second term (representing 'transmission') on the right-hand side $E[W\overline{\delta}]$ was equal to zero. This basically states that an individual of a given genotype or phenotype produces offspring identical to itself. If one introduces multiple loci and recombination, of course, this condition is not satisfied and any derivation of the rate of change in mean trait value has to incorporate transmission values. These considerations arise when one wishes to derive equations for selection acting on genotypes defined by the allelic states at two or more loci. It is instructive to look at this example, as it illustrates the conflict between selection at the level of genes and at the level of genotypes.

Are Genes the Units of Selection?

R. C. Lewontin and K. I. Kojima were the first to derive the state equations for selection–recombination dynamics at two loci, which since then has been extended by various authors to more general multilocus models. The equations for selection on the frequency at two recombining diploid loci i, k (with alleles i=A or a, k=B or b) are

$$\begin{aligned} \Delta p_{i,k} &= \frac{1}{\overline{W}}(\mathrm{cov}[W, X_{i,k}] \pm rDW_{ABab}) \\ &= \frac{1}{\overline{W}}\left(\left(W_{i,k}^* - \overline{W}\right)p_{i,k} \pm rDW_{ABab}\right) \end{aligned} \quad [4]$$

Here r is the recombination rate (r=0.5 in the absence of physical linkage), $W_{ij}^* = \sum_{j,i} W_{ij,kl}$ is the marginal fitness of the i,kth haplotype (averaged over all possible diploid backgrounds j,l), and D is the linkage disequilibrium, defined as a measure of deviation from random association of alleles at different loci,

 doi:10.1016/B978-0-444-63768-0.00810-6

$$D = p_{i,k} - p_i p_k \quad [5]$$

It was also shown by Lewontin and Kojima that when the fitness of a two-locus genotype is nonadditive (i.e., the fitness of genotype AB cannot be written as a sum $W_A + W_B$) the selection–recombination system will evolve toward nonzero values of D. What this means is that in general one cannot determine genotype frequencies in the system as the product of allele frequencies at individual loci.

Furthermore, implicit in the equations for selection on a single locus [1] is that the measured fitness of an allele is a marginal fitness, not only against the corresponding sister allele in the diploid state, but against all possible genetic backgrounds at other loci. In other words, the fitness values of W_{ij} (of genotype ij at the first locus) are themselves marginal fitness values that depend on the frequency and fitness interactions at a second locus (or in general, over all other loci), that is,

$$W_{ij} = \sum_{kl} p_{ij,kl} W_{ij,kl} \quad [6]$$

In general, the fitness of an allelic combination at a given locus will depend on the frequencies of alleles at other loci. If allele frequencies at the second locus change, W_{ij} will not be constant for genotype ij from one time step to the next. In fact, it can be shown that one may only calculate [1] with W_{ij} as constant parameters in the special case when genotype fitnesses are multiplicative across loci, that is, when $W_{ij,kl} = W_{ij} W_{lk}$. The use of these marginal fitness values as constants also requires linkage equilibrium, when $p_{ij,kl} = p_{ij} p_{kl}$.

So one is left with a situation where [1] is dynamically sufficient (in the sense of being able to predict the frequency of an allele frequency in the next generation from the current state at that locus alone), only when $D = 0$ and the fitness contributions at each locus are multiplicative. However, when the fitness effects across loci are multiplicative, there are equilibria where $D \neq$ zero, while the additive model where the only equilibrium state has $D = 0$ requires one to calculate marginal fitnesses from the effect of other loci. Consequently, the single locus approximation should only be valid under relatively weak selection.

If one equates units of selection with dynamically sufficient state variables it is not meaningful to refer to individual genes (allele frequencies at a single locus) as the unit of selection. In contrast to the genic selection interpretation of eqn [1], selection acts on the entire genome, complete with epistatic interactions and linkage disequilibria between its myriad genes. Genic selection is only an approximation that applies under weak epistasis, when additive and multiplicative interactions are essentially equivalent.

These results are in a sense congruent with the philosophical debate that focused on the ontology of the selection process. Contrary to the claims in the popular writings of R. Dawkins and other advocates of genic selection, Eliot Sober and colleagues have argued that a unit of selection must be both a replicator (i.e., can reproduce itself as an independent entity) and an interactor (i.e., compete or otherwise interact with entities like itself). Under most circumstances, individual genes fail to fulfill these conditions, since they only replicate in the context of a whole genome, and, as demonstrated by the multilocus equations, only 'interact' as components of genomes and organisms. Of course, under this criterion, the case where [1] is dynamically sufficient is simply a special case of selection at the individual/genomic level where certain symmetry properties are satisfied, as opposed to genic selection as such.

When Genes Are Units of Selection: Selfish Genetic Elements

There are cases where genes can replicate independently of the genome, with retrotransposons being the most obvious example. While DNA transposons are mobile elements that can self-excise and insert to new sites, retrotransposons can make copies of themselves and proliferate within the genome through reverse transcription mechanisms. They not only replicate independently of the cell's mitosis mechanism, they often do so at the expense of the rest of the genome and the organism's fitness by inducing mutations at the insertion sites and bloating the size of genomes (imposing constraints on cell division rate and cell size, most notably in amphibians). Thus transposons are truly examples of 'selfish genes'. A similar example comes from plasmids in bacteria, which can propagate independently of the rest of the bacterial genome via conjugation.

In addition to transposons (a clear case of an entity that is both a 'replicator' and an 'interactor'), there are a number of other genes that exhibit selfish behavior and seem to satisfy Sober's ontological conditions for being units of selection. Among these are genes that exhibit meiotic drive. Unlike transposons, such genes still require the molecular machinery of meiosis and mitosis to replicate. However, they use a number of mechanisms (such as the killing of gametes containing sister chromatids by the well-studied t-locus in mice) to distort Mendelian segregation ratios, so that they effectively replicate at higher rates than the rest of the genes on the chromosome.

Another example of selfish genetic elements comes from the genomes of plastids (mitochondria and chloroplasts). While normally plastids replicate synchronously with their 'host' cell, there are mutations that allow for selfish replication of mitochondria within a cell, or else mitochondria that induce male sterility to guarantee their propagation through female lines.

The conflict between (true) genic selection on the selfish element and selection on the genome can be best understood from Price's equation. If in [2] the variable X refers to the presence or absence of a 'selfish' allele at a given locus, the first term on the right-hand side reflects genomic (individual) selection against the allele (negative covariance between X and fitness) while the second term $\overline{\delta}$ will be positive due to the enhanced replication rate of the selfish gene. Rearranging the terms in [2], one obtains

$$\Delta\overline{X} = \frac{1}{\overline{W}}(\mathrm{cov}[W, X + \overline{\delta}]) + E[\overline{\delta}] \quad [7a]$$

If $\overline{\delta}$ represents the transmission bias due to selfish replication, one can see clearly how selection at the lower level (positive $\overline{\delta}$) runs counter to selection at the genomic or individual level (negative selection against an increase in X).

The drawback of Price's equation in this form is that it treats the $\overline{\delta}$ term as a black box of 'transmission bias', even in cases where it is in fact due to selection at another level. In the case of selfish genetic elements, what is transmission bias with respect to genome (individual)-level selection is actually selection at the genic level (which can further be decomposed recursively into a selection and bias term).

The results of conflicts between levels of selection depend on the strength of genomic versus genic selection, and on the rate at which the genes replicate relative to the generation time of the organism (since the 'transmission bias' term is averaged over the timescale relevant to individual selection). To see that this is the case, note how $\overline{\delta}$ from [7a] can be further expanded to model selection at the lower (genic) level as the source of transmission bias at the organismal level, that is,

$$\Delta\overline{X} = \frac{1}{\overline{W}}\left(\mathrm{cov}[W, X] + E\left\langle W, \frac{1}{\overline{w}}E\{\mathrm{cov}[w_i, X_i] + E[w_i\overline{\delta}_i]\}\right\rangle\right) \quad [7b]$$

The first term in the outer parentheses represents selection at the individual level, while the second term (w_i, X_i) represents selection at the ith lower level entity explicitly, with $\overline{\delta}_i$ being the transmission bias among individuals at the lower level, and $\overline{W}$ the mean fitness of lower-level entities. For example, if X represents the number of transposons in a genome, the first covariance term represents selection on the genome, and the second set of terms to intragenomic selection of 'selfish' transposons.

Note that an expansion of the form [7b] can include an arbitrarily high number of levels and represent any situation where there is a conflict between these levels of selection, since the respective covariance terms need not have the same sign. Furthermore, it can be used to model different modes of selection across levels: genomes and selfish genes, groups and individuals, and so forth.

Kin Selection and Group Selection

Just as the discovery of selfish genetic elements challenged the primacy of organism and genic selection from below, the phenomenon of altruism challenged it from above. It is well known that organisms will make sacrifices to their own fitness to benefit near relatives (in extreme cases such as worker castes in social insects, forfeiting reproduction). Darwin saw this as a challenge to his own theory which was based on individual-level selection, even though he had the largely the correct intuition about tradeoffs between benefits to self and benefits to kin. These issues were formally addressed by W. D. Hamilton in his work on kin selection and inclusive fitness.

Hamilton's solution to the problem of altruism was conceived in terms of genic selection, that is, from the analysis of pedigrees one can calculate the coefficient of relatedness ρ, which is the probability that an allele at a given locus is shared by an individual and its relatives ($\rho = 1/2$ for siblings and for parents/offspring, 1/4 for half-sibs, 1/8 for first cousins, etc.). Hamilton showed that an allele for altruism will be selected if the cost to the altruist is outweighed by the benefit to recipients in inverse proportion to their relatedness, that is,

$$\rho B > C \quad [8]$$

where C is the cost to the altruist, B the benefit to the relative. The theory of kin selection helped resolve many questions dealing with the evolution of social behavior, and was shown to give a robust prediction of sex ratios in hymenopteran (ant, bee, and wasp) colonies. From the above inequality, it was argued that kin selection (for altruism in family groups) was reducible to individual or genic selection.

However, it can be shown that whether an altruist allele increases in frequency depends critically on the structure of the social group in which an organism interacts. As a predictor of whether an altruistic allele can invade and spread in a population, [8] is incorrect – it is a criterion for the increase in the absolute number of altruists, not in their relative frequency. In order for an altruistic genotype to increase in frequency in a population at the expense of a selfish type, the altruists must interact preferentially with their fellow altruists. If altruists and selfish individuals are distributed randomly, altruism can be shown to always decrease in frequency even when [8] is satisfied. In other words, selection for altruism requires not only variance in individuals within demes, but also variance between demes. **Box 1** illustrates how intrademic selection will decrease the frequency of altruists regardless of whether Hamilton's inequality is satisfied, while interdemic selection can increase its overall frequency.

Interpreted from the standpoint of Price's equation, kin selection acts on covariance between group differences in frequency and fitness. The correct form of Hamilton's inequality can be derived from

$$\Delta\overline{X}_D \frac{n(\mathrm{var}[X_i])}{\overline{W}}[\beta_{x_i,x_D}\beta_{\overline{W}_D,X_D} + \beta_{W_i,X_i}] \quad [9]$$

where n is the number of individuals in a deme, W_i is individual fitness, X_i the ith individual phenotype, X_D the trait value in deme D (e.g., the frequency of the trait in a given deme), and $\overline{W}_D$ the mean fitness of the Dth deme. The first terms in the brackets represent selection on the trait at the demic level (weighted by the regression value of individual traits against deme means), the second represents individual selection within demes. Interpreted in terms of [2], the term $\overline{\delta}$ corresponds to selection at the lower

level (individual selection) biasing trait values at the higher (group or deme) level, as opposed to being driven by transmission dynamics as such.

The condition for increase of the altruist allele or any other trait subject to both interdemic and individual selection is that

$$\beta_{X_i,X_D}\beta_{\overline{W}_D,X_D} + \beta_{W_i,X_i} > 0 \qquad [10]$$

Traits such as altruism will be selected against at the intrademic level, so that the second term will be negative. Selection for altruism requires that there is strong interdemic selection and an association between mean deme value and trait value (i.e., altruists tend to associate in demes with other altruists, as do selfish individuals, as given by the first coefficient of the first term in [9]). From this interpretation, kin selection is actually an extension not of individual or genic selection but of group selection, because the covariance relationship that determines the magnitude of eqn [10] is a group property, one that is the ultimate target of selection.

Group selection has been a source of controversy since V. C. Wynne-Edwards proposed that behavioral traits such as territoriality evolved for the good of populations or species, as a means of preventing overcrowding and population crashes. It was demonstrated in the critical writings of J. Maynard Smith and G. C. Williams that these population regulatory mechanisms were a by-product of selection at the individual level, and that group selection, while theoretically a possibility, was in practice improbable because selection at the individual level acts over much shorter time spans (individual generation time) than the 'generation time' (colonization and extinction rates) of demes. Therefore, individual selection would eliminate variance in a 'good of the species' trait before interdemic selection could act in favor of it.

Since that time, group selection has been proposed to be much wider in scope and potential importance than was initially recognized, even though the scenarios proposed by Wynne-Edwards are untenable. For example, it can be seen from eqn [9] that interaction between relatives is not necessary for interdemic selection to operate. All that is necessary is a correlation between individual genotype/phenotype and group trait value, $\beta_{X_i,X_D} > 0$, so that individuals of similar trait values tend to group together. Since many organisms interact in family groups, kin selection may be a particularly effective form of group selection. Furthermore, it is not necessarily true that the 'generation time' for groups is on a vastly different timescale than individual selection, as the generation time for many animal colonies and societies can operate on the same general timescale as the generation time for constitutive individuals.

Equations [9] and [10] also demonstrate that selection on group trait value (even counter to individual selection) can be effective under a number of other scenarios where there is strong covariance between deme trait values and their rate of propagation. The most familiar example comes from the evolution of parasite virulence. Individual selection favors high virulence in parasites under many circumstances; however, if all parasites in a host are maximally virulent, they will kill the host and themselves in the process. Thus strong group selection will override individual selection for virulence, since 'benign' groups will survive and propagate for longer periods of time along with their hosts, while virulent groups will become extinct at a high rate.

To consider another example, R. A. Fisher demonstrated that individual-level frequency-dependent selection will lead to 1:1 sex ratios in organisms (because the rarer sex has higher fitness with unequal ratios), which is the rule in most organisms. However, strong deviations of this sex ratio in favor of females have been observed in a number of organisms, such as parasitoid wasps. The reason for this deviation is that populations with large numbers of females have a higher rate of propagation than those with an even sex ratio, so if interdemic selection is strong (as it is for many parasites and parasitoids within a host animal or plant), selection on the group trait for high propagation can run counter to individual selection for equal sex ratios.

Interestingly, the question of the origin of eusociality in insects, which has long served as the textbook example of kin selection, has been recently challenged on a number of fronts. Recently, E. O. Wilson and other social insect specialists have suggested that because many primitively social insects have colonies founded by nonrelated females (and because kin selection generate strong parent–offspring conflicts that select against colony altruism), group selection on colony fecundity and viability may have played a more important role in driving the evolution of advanced sociality and altruism than has kin selection.

It should also be noted that Price's theorem and equations derived from it do not require that a trait (group or individual) under selection be genetic in nature, simply that it be heritable and covary with fitness. It has been argued by D. S. Wilson and colleagues that group selection probably plays an important role in the evolution of human culture, since cultural traits are transmitted at the individual and group level and exhibit high variance (and differential propagation) both within and among groups. Certainly, altruism in humans extends far beyond family units to encompass groups defined by nationality, religion, and political beliefs, so an evolutionary model other than kin selection is necessary to account for these social phenomena.

Species Selection

A number of paleontologists (most notably S. Stanley and S. J. Gould) have made the intriguing suggestion that species-level traits are subject to selection and that this plays an important role in various trends observed in the fossil record. The basic argument is that certain traits in an organism related to habitat specificity and low dispersal ability lead to enhanced rates of speciation and therefore greater representation of these traits among species, even when they are non- or maladaptive at the individual level.

The theoretical possibility of species selection can be demonstrated through applications of Price's theorem, since the focal point of selection can be any replicating entity with the property of heritability (as species do have, since they give rise to new species with properties that correlate with those of the ancestor). Unlike eqn [9], which combines group and individual selection

acting on the same timescale, models of species selection (or any kind of hierarchical selection where the timescale at one level is orders of magnitude different from the other) require that the effects of rapid individual-level selection be averaged with respect to the higher-level timescale. If X is a species-level trait also subject to individual selection, the equations for the rate of change have the same form as eqn [7b], where w_i now represents selection at the individual level and W at the species level.

Since the focal point of selection here is the species level, the expectation within the brackets representing the 'transmission bias' $\overline{\delta}$ at the species level due to lower-level selection, $\overline{\delta} = E[(1/\overline{w})\mathrm{cov}[w_i, X_i] + \overline{\delta}_i]$, is calculated by averaging over the waiting time to speciation, that is, it measures the expected change in X until the time of a speciation or extinction event. The same arguments used against Wynne-Edwards apply here, that is, it can be shown that if selection is strong and the waiting time is large, the trait value will be fixed in the species long before species selection can act.

It has been shown by S. Rice and others that species selection (or any form of group selection where there is a large difference in generation time of the group vs. the individual) is indeed weak when selection at the individual level is strong and in a consistent direction. On the other hand, if selection fluctuates so that some species (or populations) are fixed for one trait value and others for the opposite, individual-level selection can actually facilitate species or group selection by generating variance at the higher level. As expected, species selection would be most effective in taxa with long individual generation times and high speciation rates, or in those where population sizes are small so that high interspecific variance can be generated through genetic drift.

Hierarchical Selection and Evolutionary Novelty

Because any form of heritable variation at any level of biological organization may be the target of selection, there is much potential for conflict between the demands of selection at different levels. The resolution of such conflicts has evidently played an important role in a number of key events in the history of life.

Selfish genetic elements demonstrate the potential conflict of interest between a genome (which is effectively a 'colony' or group of genes) and individual genetic elements. It is likely the case that the DNA replicative machinery and the complex processes of mitosis and meiosis evolved as the genome's mechanisms of keeping selfish genetic elements in check, with segregation distorters and transposons finding ways around these mechanisms to advance their own fitness at the expense of the rest of the genome.

Similar problems arise in the symbioses that were the likely origin of mitochondria, choloroplasts, and other eukaryotic organelles. In order to prevent their proliferation at the expense of the host cell, the hosts gradually co-opted much of the organelle genomes within the cell nucleus and evolved mechanisms to limit the genetic variance and independence of plastid elements. Specifically, it is likely that anisogamy evolved as a mechanism to prevent the proliferation of deleterious selfish mitochondria, as it reduces the genetic variation within the cell, which restricts the potential for selection at the level of selfish plastids.

At a yet higher level of organization, multicellular organisms are composed of entities – cells that can themselves reproduce via mitosis. In order to minimize conflicts caused by the proliferation of 'selfish' cell lines, various mechanisms such as germ line segregation and development from single (as opposed to multiple) zygotes minimized the genetic variation among cells within an organism. Since natural selection requires covariance between fitness and trait value, having nearly identical genomes in all cells makes intraorganismal selection weak when compared to the strong selection (due to high variance in phenotype) within the colony or organism. Cancers can be regarded as a breakdown of the safeguards against selfish cells in the same way that transposons and meiotic distorters break down the genome's safeguards against selfish DNA.

Finally, there are the conflicts that arise within societies and colonies, including human societies. Since there is generally much more genetic and phenotypic heterogeneity among animals in societies (excepting clonally reproducing invertebrates), selection at the individual level remains strong even as selection at the higher group level can be seen to emerge. From the perspective of Price's equation, both the higher-level selection and 'transmission bias' due to lower-level (individual) selection are significant, without any one unit or level of selection dominating others.

See also: Aquatic Ecology: Microbial Communities. Behavioral Ecology: Altruism; Kin Selection. Evolutionary Ecology: Genetic Drift; Macroevolution; Natural Selection; Fitness

Further Reading

Arnold, A.J., Fristrup, K., 1982. A hierarchical expansion of the theory of evolution by natural selection. Paleobiology 8, 113–129.
Burt, A., Trivers, R., 2005. Genes in Conflict: The Biology of Selfish Genetic Elements. Cambridge, MA: Harvard University Press.
Buss, L.W., 1988. The Evolution of Individuality. Princeton, NJ: Princeton University Press.
Dawkins, R., 1976. The Selfish Gene. Oxford: Oxford University Press.
Frank, S.A., 1995. George Price's contribution to evolutionary genetics. Journal of Theoretical Biology 175, 373–388.
Gould, S.J., 2002. The Structure of Evolutionary Theory. Cambridge, MA: Harvard University Press.
Hamilton, W.D., 1998. Narrow Roads in Gene Land: The Collected Papers of W. D. Hamilton, vol. 1: Evolution of Social Behavior. Oxford: Oxford University Press.
Lewontin, R.C., 1970. The units of selection. Annual Review of Ecology and Systematics 1, 1–18.
Lewontin, R.C., Kojima, K.-I., 1960. The evolutionary dynamics of complex polymorphisms. Evolution 14, 458–572.

Maynard, S.J., Szathmary, E., 1994. The Major Transitions in Evolution. Oxford: Oxford University Press.
Price, G.R., 1970. Selection and covariance. Nature 227, 520–521.
Rice, S.H., 2004. Evolutionary Theory. Sunderland, MA: Sinauer Associates.
Salthe, S.N., 1985. Evolving Hierarchical Systems. New York: Columbia University Press.
Sober, E., Wilson, D.S., 1999. Unto Others: The Evolution and Psychology of Unselfish Behavior. Cambridge, MA: Harvard University Press.
Stanley, S.M., 1979. Macroevolution: Pattern and Process. Baltimore, MD: The Johns Hopkins University Press.
Williams, G.C., 1966. Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought. Princeton, NJ: Princeton University Press.
Williams, G.C. (Ed.), 1971. Group Selection. Chicago: Aldine and Atherton.

Abundance

JT Harvey, Moss Landing Marine Laboratories, Moss Landing, CA, USA

Introduction

The abundance of an organism, often considered as total population size or the number of organisms in a particular area (density), is one of the basic measures in ecology. Ecologists often are interested in the abundance and distribution of organisms because the number and spatial extent of an organism reflects the influences of many factors such as patterns in nutrients (fuel), predators or herbivores, competitors, dispersal, and physical conditions. Organisms generally are more abundant where conditions are favorable, such as locations with sufficient quantity and quality of food or nutrients, fewer herbivores or predators, fewer competitors, and optimal physical features. The physical features that affect abundance could be substrate type, moisture, light, temperature, pH, salinity, oxygen or CO_2, wind, or currents. Ultimately, the abundance of an organism is dependent on the number of individuals that survive and reproduce. Therefore, any factors that affect survival or reproduction will affect abundance.

Abundance can be measured at many levels, such as the number of individuals of a certain sex or age within a population, the number in a certain geographical region, the number in a certain population (possibly defined as the interbreeding individuals of the same species in a certain geographical area), or the number of individuals of a certain species. Species or populations have different levels of abundance and different population dynamics because of inherent biological characteristics (vital rates), such as the number of young produced per individual, longevity, and survival, and because the species may be adapted and exposed to various environmental conditions. Estimating abundance, however, can be difficult depending on the distribution, visibility, density, and behaviors of the organism.

Estimates of abundance can be obtained by counting all individuals in the population or sampling some portion of the population. A census or total count of all individuals is a common technique used to assess abundance of organisms that are relatively rare and easily observed. If the organism is too numerous or not easily counted then a representative portion of the population is sampled using various techniques such as (1) counts within randomly selected sampling units (e.g., quadrats, cores, nets, or traps); (2) mark-recapture; (3) strip or line transects, which is essentially sampling a long thin quadrat; and (4) distance methods (e.g., nearest neighbor). Most of these methods have a well-developed theoretical and analytical basis. Based on whether the organism is numerous and relatively stationary (e.g., plants), or rare and mobile (e.g., many vertebrates) certain techniques are appropriate. Numbers of individuals within a sample can be determined directly by visually counting individuals or indirectly using acoustics, such as hydroacoustics for assessing fishes or counting calls of bird or whales. Other indirect methods include counting the number of eggs or juveniles, which is an indication of the number of adults (sometimes used to assess fish abundance) or counting nests (such as used for birds). Recently the amount of genetic variation in a population has been used to estimate abundance.

If an actual number of individuals cannot be determined, scientists have used indices of abundance, such as changes through time, percentage cover or harvested biomass (e.g., for plants), and catch per unit effort (e.g., for fishes). Ecologists are always striving for an accurate and precise estimate of abundance; therefore, a thorough knowledge of the organism and its environment is necessary to design the proper sample unit and best allocation of that sample unit in space and time. Estimates of abundance can be made more accurate or at least accuracy assessed by eliminating or decreasing biases and by using various methods to determine abundance. Determining whether there is an accurate estimate of abundance is difficult because usually the true abundance is unknown and the estimate may contain unknown biases. Being aware of the potential biases and striving to minimize and investigate biases will increase the chances of an accurate abundance estimate. Different sampling designs will help ensure a representative sample is obtained that also will increase accuracy. Variability in the estimate of abundance, or precision, is affected by the natural variability in abundance among the samples and by the number of samples. Because natural variability cannot be controlled, the single best means of increasing the precision of the abundance estimate is to increase sample size (e.g., number of transects, cores, marked individuals). An understanding of sampling design (or the observation of sample units in space and time) can help determine whether there are enough independent and representative sample units to provide an accurate and precise estimate of abundance.

The spatial and temporal patterns of abundance (i.e., dispersion) often indicate fluctuations in physical or biological factors. Abundance is a measure of how many organisms are within an area whereas dispersion is how those organisms are arranged within the area. We usually recognize three basic patterns of abundance in space or time: uniform, random, or aggregated (**Fig. 1**). A uniform abundance in space or time is one where the organism is spaced evenly. Rarely is this the case for organisms because biological factors (e.g., attraction, aggression, competition) will cause nonuniformity and most environmental features that affect organisms are not uniformly distributed. With a random distribution we assume that the probability that an organism can inhabit any location or time is equal. This also is rare for the same reasons that organisms are not uniformly distributed. Finally, organisms can be aggregated if they occupy very specific locations, such that in some locations or times the probability of encountering that organism is nearly one and in

 doi:10.1016/B978-0-444-63768-0.00454-6

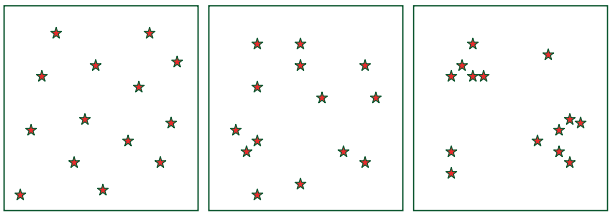


Fig. 1 Various forms of dispersion in space (uniform, random, and aggregated) are depicted using red stars as individual organisms. In reality, there is a continuum of dispersion from the extremely uniform to greatly aggregated, with random in between these two forms.

other locations or times it is zero. At some spatial scale all organisms are grouped or aggregated. Sea bird nests may be uniformly distributed within a colony because the birds place their nests just far enough apart to not be pecked by their neighbor, but the nests are very much aggregated because some sea birds only nest on isolated islands. If you focused your attention at the scale of the colony, the bird's nests would be distributed uniformly, but at the scale of the world, the nests are aggregated. Aggregations typically occur where local conditions are optimum for survival and reproduction. For instance, certain plants require specific soil types, light exposure, moisture, and nutrients. Through adaptive radiation, species have evolved specific requirements; hence, they cannot live just anywhere. The aggregated distribution of organisms implies that individuals of a population will be abundant in some locations and rare or absent from other locations. The same patterns and rationale can be used to assess abundance patterns in time. Certain periods of time are more conducive to some species; hence they are more abundant, than other times. The timescales that affect abundance can be days for short-lived organisms like insects, or thousands of years like large trees. Natural and anthropogenic changes in environmental conditions will cause changes in abundance for all organisms, and these changes can be predicted using mathematical models of population dynamics. The actual patterns of dispersion in space and time form a continuum, where populations can have varying levels of uniform, random, or aggregated patterns. Because organisms in space and time have an infinite array of patterns it makes it difficult to accurately model and test patterns of dispersion.

Population Dynamics and Growth Models

Population growth rate is defined as the change in number of individuals in the population through a certain amount of time. Changes in abundance often are assessed at the population level, because that is where the effects of biological and environmental conditions are most evident. Later the modeling of metapopulations (groups of populations that share individuals) will be discussed. Often ecologists measure population changes using density (the number of individuals per area) because interactions among individuals and between an individual and its environment are more affected by density than actual population abundance. Changes in abundance of a population, called population dynamics, are caused by many factors, but at the most fundamental level the number of individuals in the population at some later time period (N_{t+1}) is simply the number of individuals at the beginning of the time period (N_t) plus the number of births (B) and immigrants (I) minus the number of deaths (D) and emigrants (E) that occur during the time period:

$$N_{t+1} = N_t + B - D + I - E \quad [1]$$

We can rearrange the terms to determine the change in population abundance ΔN:

$$(N_{t+1} - N_t) = \Delta N = B - D + I - E \quad [2]$$

If we assume the population is closed, that is, there is no movement of individuals into the population (immigration) or movement out of the population (emigration), then the equation becomes simplified. In this case, we are modeling only the changes that occur within a population. Although most populations are not closed (i.e., there is immigration and emigration), it allows for some easier calculations and allows us to provide details on a specific population. For a closed population, with $I=0$ and $E=0$, the equation simplifies to

$$\Delta N = B - D \quad [3]$$

If we also assume that the period of time between estimates of population abundance is extremely small (i.e., t is nearly 0), then we can treat population growth as continuous. If we estimated population abundance only after a large period of time then population growth would not be modeled as continuous but would be a discrete function. Modeled as a discrete function the estimate of the population abundance would be the same or changing during the time period, then at the end of the period it would change to a new level, stepping from one abundance level to another after each time period (**Fig. 2**). By assuming population growth is continuous we can use a differential equation to describe the change in population size (dN) that occurs during an extremely small period of time (dt):

$$\frac{dN}{dt} = B - D \quad [4]$$

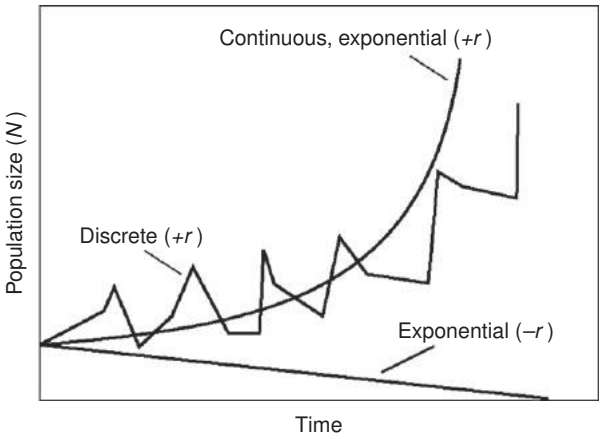


Fig. 2 Various models of population change: a continuous exponential population increase (when r is positive), a continuous exponential decrease (when r is negative), and a discrete model (when r is positive).

B and D now represent rates, or the number of births or deaths in this population during these short periods of time. The birth rate (B) can be thought of as the number of births per individual per time period (b), called the instantaneous birth rate, times the number of individuals in the population (N) at that time. The death rate (D) also can be calculated using the instantaneous death rate (d) times the population abundance (N). The continuous differential equation is now

$$\frac{dN}{dt} = bN - dN = (b - d)N \quad [5]$$

This is very much a simplification of the real world (a model) because it assumes that the instantaneous birth rate and death rate are constants, thus per capita birth and death rates are the same no matter what the population abundance. We know this is not true, because for many populations the birthrate will decrease and the death rate increase when populations become larger. Thus the factors affecting population growth are dependent on the population abundance or density, hence are called density-dependent factors. The concept of density-dependent factors will be discussed later.

The instantaneous rate of increase of the population (r), often called the intrinsic rate of increase, is $b - d$, so substituting r for $(b - d)$ produces a new model for population growth that is

$$\frac{dN}{dt} = rN \quad [6]$$

This simple equation allows us to predict that the change in population abundance per unit time is exponential and proportional to a constant value of r (**Fig. 2**). If the intrinsic rate of increase is zero then the reproductive rate and mortality rate are equal ($b - d = 0$). If r is positive or greater than 1 (per capita birth rate exceeds per capita death rate) then the population would increase proportional to the population abundance (N) in an exponential fashion. If r is negative or less than 1 (death rate exceeds the birth rate), then the population would decrease exponentially (**Fig. 2**). The greater the value of r the more rapidly the population will change. The intrinsic rate of increase can be used to model or predict future population abundance (N_{t+1}) by integrating the population growth model and knowing the time interval (t) over which you want to predict abundance and the initial population abundance (N_t):

$$N_{t+1} = N_t e^{rt} \quad [7]$$

This simple exponential model has many assumptions. We assumed the population is closed, a simplification to eliminate the effects of immigration and emigration. This assumption will be removed during a discussion of metapopulation models. We assumed that per capita birth and death rates were constant, but they do change with changes in population size, and with different age and sex classes. We also assumed that population growth was continuous; however, many species have discrete generations, where births and deaths are measured not instantaneously but at certain longer time frames. Population dynamics of species that reproduce once per year, for instance, often are modeled not with a continuous but with a discrete model. Finally, this model is deterministic; it will always predict the same abundance given an initial population size and a specific r and t. Adding variability to the model produces a stochastic model (i.e., the effects of random variation that may be more realistic), and predicted abundances are affected by the variability in the intrinsic rate of increase. If r is positive but has a great deal of variability, the population will increase and the predicted abundances will have greater variability through time. The consequence is that through random chance, the population could go extinct because of an extreme negative random effect although the population was increasing previously.

As a population begins, growth rate is minimal because there are few individuals and many are probably young and not producing offspring. After the early stages, growth becomes more rapid, often increasing exponentially, because there are now more individuals that are reproducing and factors that may restrict population growth, such as food resources, space, predators,

diseases, are not affecting the population in a large manner. During this phase as the exponential model would predict, population growth is proportional to the population size and magnitude of r. Eventually, as the population increases in size, nutrients become more difficult to locate or obtain, predators or herbivores are more abundant, competition among individuals within the population increases, mortality, because of disease, parasites, and movements into suboptimal areas increases, and space or shelter becomes limiting. These factors to some extent control the ultimate size of the population, causing density-dependent regulation. The maximum size that the population can achieve is often referred to as the carrying capacity (labeled K). Carrying capacity is not really constant nor is it the maximum capacity but the population often fluctuates through time as various conditions within and outside the population change. The carrying capacity is more realistically determined by examining the long-term average of abundance. Species that live in relatively stable environments (e.g., invertebrates in the deep sea) may have relatively constant population numbers around the carrying capacity, whereas those populations in unstable environments (e.g., nearshore invertebrates) may have variable numbers through time. Large regime shifts or major changes in environmental conditions caused by natural or anthropogenic factors can produce a shift in the carrying capacity for the population.

Because organisms cannot increase exponentially forever, the concept of a carrying capacity (K) was developed for the population growth models. The simplest relationship is to assume that changes in per capita birth rate (b) or death rate (d) are linearly related to population size. As population size increases b decreases linearly and d increases linearly. The concept makes sense; as the population increases, reduced resources, competition, predation, and other density-dependent factors would cause an increase in deaths and decrease in births. It probably is intuitive also that the relationship between b and d with population size is not linear, but more likely some nonlinear function, with the changes in b and d becoming more pronounced as density becomes extreme. Carrying capacity occurs when the population abundance reaches a maximum equilibrium ($N=K$), when birth rate and death rate are equal. The population model that depicts the effect of carrying capacity on population growth is the logistic model:

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right) \quad [8]$$

This is basically the exponential growth model with another term $(1 - N/K)$, that is, the portion of the carrying capacity that can still be filled with the population. When the population is in its infancy (N is extremely small compared with K), then N/K is small and $(1 - N/K)$ is nearly equal to 1. At this early stage the population grows exponentially (rN). As the population increases and approaches K, the ratio of N/K approaches the value of one, $(1 - N/K)$ approaches zero, and the change in population size (dN/dt) approaches zero (**Fig. 3**). Thus as the population approaches K the rate of change in the population slows or decreases, effectively stopping exponential positive growth (**Fig. 4**). If the population should exceed K, then $(1 - N/K)$ is negative and the rate of change will be negative and the population will decrease, presumably toward K. In the logistic model, the greatest population growth rate is achieved at half the carrying capacity ($K/2$), and populations with a greater value of r will reach K more rapidly (in less time) than populations with lesser values of r.

Because the various density-dependent factors do not affect the population growth rate immediately, often there is a time lag between the change in population size and effects on population growth. This lag time effectively creates an oscillation of the population abundance around carry capacity, because the population adjustment by the carrying capacity does not take effect immediately; so abundance will go past K, and then dip below K. If the combined effects of lag time and response time $(1/r)$ are small then the population will increase in a smooth logistic fashion; if the effects are moderate, the population will oscillate around K, eventually decreasing in oscillation until the population is at K. If the combined effects of lag time and r are large, the population abundance will oscillate around K without reaching an equilibrium value. The amplitude of the oscillation around K will be greater if the population grows rapidly (r is great) and the time lag is great. Many populations oscillate around K because the effects of reaching carrying capacity, such as increased number of predators, decreased reproductive output, and increased

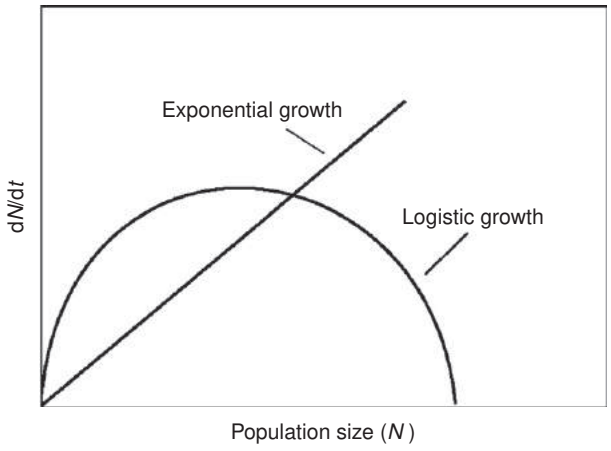


Fig. 3 The change in population per unit time (dN/dt) relative to population size (N) for a population increasing exponentially and one using a logistic model.

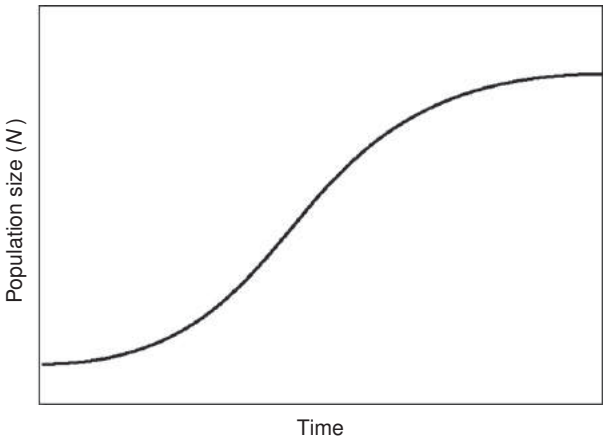


Fig. 4 The logistic population model plotting the population size (N) through time.

aggression, take some time before they have an effect (there is a substantial lag time). With greater levels of r, population dynamics can become chaotic, in the sense that there is no stable equilibrium and no stable cycles. In a chaotic system you cannot predict the cycle, but the population could still fluctuate around a long-term mean.

The previous models best approximate population growth rates for smaller organisms (e.g., bacteria), where the organisms reproduce rapidly (r is great) and have short generation times. Many larger organisms, however, can be modeled using age- or stage-based models often called life table models (**Table 1**). In these models, probability of survival to the next age class (l_x) and mean number of female offspring per female of an age class (m_x) are computed for each age class and used to compute net reproductive rate (R_o):

$$R_o = \sum_{x=0}^{k} l_x m_x \qquad [9]$$

where x=age class, and k=final age class.

R_o is the potential number of female offspring produced by a female during her lifetime. If R_o is greater than 1 then the population will increase because a single female is producing more offspring than she needs to replace herself. If R_o is less than 1 then the population will decrease, and if R_o is one the population is stable. Generation time is the average age difference between a cohort (a group of individuals born at the same time) and their parents or the amount of time it takes a cohort to replace another cohort. Generation time is calculated as

$$G = \frac{\sum_{x=0}^{k} l_x m_x x}{\sum_{x=0}^{k} l_x m_x} \qquad [10]$$

Using these age-based models, r is determined by solving in an iterative fashion, the following equation:

$$1 = \sum_{x=0}^{k} e^{-rx} l_x m_x \qquad [11]$$

The value of r can then be used in the logistic equation. The advantage of using age-based models is that the number of individuals in each age class can be estimated. Often the age-based life table is converted to an age-class matrix (a Leslie matrix) that is used to model changes in the age structure with population growth. In these models, with a constant birth and death rate, the population will attain a stable age distribution (i.e., relative numbers in each age class will remain equal as the population increases), no matter what the initial age structure.

These previous models have depicted the changes in abundance for a single isolated population, but most populations are not isolated and are affected by the immigration and emigration of individuals (a open population). Certain population models have been determined that allow for open populations, and treat the system as a metapopulation (a group of populations that are linked by individuals that emigrate from one population to be immigrants in other populations). The long-term existence of the metapopulation is dependent on the probability of extinction of populations and probability of colonizing new spots (based on emigration rates). If there is a reasonable amount of emigration, even large probabilities of population extinctions usually predict the continued existence of the metapopulation because new recruits always are arriving to colonize new locations. The continuation of the metpopulation also is positively related to the emigration of individuals into populations that already exist, which increases numbers within existing populations, thereby decreasing the probability of population extinction. The theories associated with mark-recapture techniques are fairly well developed, allowing the observations of marked individuals in future sampling to be used to model immigration and emigration and population size.

Table 1 A theoretical life table, perhaps of a mammal, because mortality is greatest in the early years, less during midlife, and increases again near maximum age

Age, x (years)	*No. alive at beginning of age interval, n_x*	*Proportion surviving until beginning of age interval, l_x*	*No. dying within age interval (from x to x+1), d_x*	*Finite rate of mortality, q_x*	*Finite rate of survival, p_x*	*Mean no. of female offspring produced within age interval, m_x*	$l_x m_x$
1	1000	1.000	480	0.480	0.520	0	0
2	520	0.520	300	0.577	0.423	0	0
3	220	0.220	120	0.545	0.454	0	0
4	100	0.100	15	0.150	0.850	0.5	0.050
5	85	0.085	23	0.271	0.729	1.5	0.128
6	62	0.062	21	0.339	0.661	2.2	0.136
7	41	0.041	13	0.317	0.683	3.3	0.135
8	28	0.028	11	0.393	0.607	3.1	0.087
9	17	0.017	10	0.588	0.607	2.8	0.048
10	7	0.007	7	1.000	0.000	1.8	0.013
11	0	0.000					
						$R_0 = \sum l_x m_x = 0.597$	

In this theoretical case, the population would be expected to be decreasing because R_0 is less than one ($R_0 = 0.597$).

r-Selected versus *K*-Selected Organisms

If the intrinsic rate of increase (r) is large, as is true for many small organisms like many species of insects, populations can increase rapidly in an exponential fashion for some period of time. Likewise, the greater duration of time (t) would predict greater numbers of individuals in the future. Of course, no organism can increase exponentially forever; various density-dependent factors cause a slowing of population growth and eventual possible stabilization and variability around a carrying capacity. As the population reaches maximum numbers, various factors act to decrease reproductive rates and increase mortality rates, hence the per capita growth rate diminishes. Species with a greater intrinsic rate of increase generally have lesser generation times (average time between initiation of parent and initiation of their offspring) and greater reproductive potential, such as increased fecundity and reproduction starting at earlier ages. Abundance of these organisms can change rapidly, either increasing or decreasing. They have been called *r*-selected species because they have evolved traits that enhance their intrinsic rate of increase, and allow them to exploit times or areas of optimum conditions. *r*-selected species also are short-lived because the traits that increase reproductive potential and decrease generation time have a negative influence on survival. Therefore, there is an assumed tradeoff between increased reproductive capacity and survival and longevity. Organisms that put a large amount of energy into reproduction, via decreased age at first reproduction and increased fecundity will have less energy placed in growth and survival. These species have greater values of r and perhaps greater lag times so their oscillations around K may be great. At the other end of the continuum are so called *K*-selected species that have a greater body size, greater generation time, lesser reproductive potential, and greater longevity. These species presumably are adapted to maintain abundance in regions with minimal environmental variability, so their abundance is less variable around the carrying capacity than *r*-selected species. Life-history traits might also be affected by physiological, physical, or behavioral limitations. For instance, many large vertebrates must reach a certain size, maturity, and understanding of the reproductive system before they can defend territories or gain access to females for mating. Because the intrinsic rate of increase is less for *K*-selected species and lag time may be less, the oscillations around the K may be more dampened than *r*-selected species. These are all generalities, for instance, not all species that are *K*-selected have relatively invariant abundances, and their populations can change dramatically with changes in intrinsic and extrinsic forces. Recognize also that there is a continuum of *r*-selected and *K*-selected species, and that many species have combinations of *r*-selected and *K*-selected traits that affect their population growth rate. The idea of *r*–*K* selection has not really proven useful for predicting life-history traits; hence some researchers have abandoned the concept. The idea that environmental conditions affect life-history traits, however, is useful.

Factors Affecting Abundance

The upper boundary of abundance (carrying capacity) or abundance in general is controlled by numerous factors; however, determining how these factors control populations can be difficult. Internal factors (e.g., interspecific competition for space, food, or light, life-history traits, and cannibalism) can control abundance as can external factors (e.g., environmental conditions). As an organism becomes abundant it can drive down the abundance of the nutrients or energy that they require. Increasing numbers of plants can decrease the quantity of nitrogen, phosphorus, potassium, and other essential elements. Increasing number of heterotrophs (organisms that ingest other organisms or organic particles) can reduce their food supply. Hence, with increasing abundance these organisms can no longer increase at a rapid rate because their supply of fuel is decreasing. Competition between organisms that are using the same fuel resources or space/shelter that become in short supply will lead to reduction in population

growth. The manner in which competition can affect abundance of animals is through the need for food, territories, mates, or space, whereas plants may compete for light, nutrients, water, or space. As organisms become more abundant generally, there is an increase in predation or herbivory, disease, and in animal's antagonistic behavior. All of these factors by themselves or often in combination serve to increase mortality, hence decreasing population growth rates. These were examples of density-dependent controls on population changes but there are examples of density-independent factors. For example, in some instances bycatch of fishes could be considered density independent because the number of fishes caught in bycatch may not be determined by the size of the fish population. Mathematical modelers have predicted that when a population's growth rate is greatly affected by density-dependent factors, its abundance is more likely to fluctuate.

The effects of competition can be amazingly complex. Lotka and Volterra (in the mid-1920s) developed a simple model to predict the competition between two species. The change in abundance for each species was dependent on r, N, K for each species and a competition coefficient (i.e., the effect of one species on the growth of the other species). Depending on the strength and nature of the interactions, the two species could exist with stable populations, one or the other species dominates forcing the extinction of the other, or lastly that they could both exist in an unstable equilibrium. Competition will ultimately affect abundances negatively, by reducing the amount of resources to each of the competitors, and this interaction becomes increasingly complex when there are more than two competitors, as is the norm in nature.

Predation also affects abundance, and is often a density-dependent factor. As abundance of a species increases, organisms that consume that species likely increase. Hence, herbivores respond to an increase in a species they consume, as would a predator population increase with the increase of a prey species. The predator–prey relationship is another mechanism that regulates to some extent the size of populations in a density-dependent manner.

Species, through the process of natural selection, have developed certain life-history traits that affect their abundance. Population change is affected by individual traits, such as reproductive output, age at first reproduction, and survival. Species have evolved these traits to maximize their fitness (i.e., maximize their genetic contribution to future generations), which affects the intrinsic rate of increase of the species. Population change also is affected by characteristics affecting the population, such as many of the density-dependent factors. Abundance is affected by factors that affect metapopulations, such as immigration, emigration, and genetic exchange. The study of abundance is fascinating because it incorporates so many of the important factors of a species: individual traits, population controls, and metapopulation dynamics.

See also: General Ecology: Biomass

Further Reading

Andrewartha, H.G., Birch, L.C., 1954. The Distribution and Abundance of Animals. Chicago: University of Chicago Press.
Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., 1993. Distance Sampling: Estimating Abundance of Biological Populations. London: Chapman and Hall.
Caswell, H., 1982. Life history and the equilibrium status of populations. American Naturalist 120, 317–339.
Dodson, S.I., Allen, T.F.H., Carpenter, S.R., *et al*., 1998. Ecology. New York: Oxford University Press.
Emlen, J.M., 1984. Population Biology: The Coevolution of Population Dynamics and Behavior. New York: MacMillan.
Gotelli, N.J., 2001. A Primer of Ecology. Sunderland: Sinauer Association.
Krebs, C.J., 1999. Ecological Methodology. Menlo Park: Addison Wesley.
Lack, D., 1954. The Natural Regulation of Animal Numbers. Oxford: Oxford University Press.
Seber, G.A.F., 1973. The Estimation of Animal Abundance. New York: Hafner.
Slobodkin, L.B., 1980. Growth and Regulation of Animal Populations. New York: Dover.

Age-Class Models☆

David H LaFever, Texas A&M University, College Station, TX, United States

Glossary

Age structure The number or proportion of individuals in each age group.

Cohort A group of organisms of the same species and roughly similar ages.

Compartment models Mathematical models commonly used to represent many different ecological systems.

Fecundity Reproductive rate of an organisms or population, typically measured by the number of present gametes (eggs), seeds, or asexual propagules.

Matrix models A specific type of population model that uses a form of algebraic shorthand for summarizing a larger number of often repetitious and tedious algebraic computations (matrix algebra).

Stochasic Random processes are used to represent systems or phenomena that seem to change in a random way.

Introduction

This article presents an overview of single-species age-class models. The first section discusses the construction of life tables, which are then used to construct age-class models. A life table is an age-specific summary of mortality rates operating on a cohort of individuals. The second section focuses on compartment models, which can consist of systems of differential equations, integral equations, or matrix models with one equation or matrix for each compartment. The third section discusses matrix models including age- and stage-based matrix models. The fourth and final section focuses on stochastic models that incorporate the probabilistic nature of biological systems by treating one or more parameters as random variables.

Life Tables

A life table is an age-specific summary of mortality rates operating on a cohort of individuals first developed by human demographers and introduced to ecologists by Raymond Pearl in 1921. The mortality schedule is generally calculated based on the known number of survivors in each age class. In order to create a life table, an age interval must be decided upon in which to group the data. For longer-lived species, such as trees or turtles, an age interval of several years may be appropriate, whereas for shorter-lived species, such as birds, some plants, or insects, 1 year or less may be appropriate. These age intervals are known as age classes. Two different types of life tables can be calculated: a static life table (also called a stationary, time-specific, current, or vertical life table) and a cohort life table (also called a dynamic, generation, or horizontal life table). A static life table is calculated based on a cross section of a population at a specific time. In this case, the mortality schedule would be calculated for each age class at a specific time. A cohort life table is calculated for a cohort of organisms followed throughout life. In this case, one age cohort would be followed throughout their life (**Table 1**).

The columns of a life table include x (age), n_x (number alive at age x), l_x (proportion of organisms surviving from the start of the life table to age x), d_x (number dying during the age interval x to $x + 1$), and q_x (per capita rate of mortality during the age interval x to $x + 1$). Given x and n_x, the following equations can be used to obtain the remaining survivorship information:

$$l_x = n_x/n_0 \tag{1}$$

$$d_x = n_x - n_x + 1 \tag{2}$$

$$q_x = d_x/n_x \tag{3}$$

Fertility data can also be included in a life table. Additional columns include b_x (fertility schedule or the average number of female offspring produced per female aged x during time x), and l_xb_x and $x(l_xb_x)$. The fertility schedule is observed in the field or laboratory, while l_xb_x and $x(l_xb_x)$ are used to calculate the net reproductive rate (R_0), which is the multiplication rate per generation of the population:

$$R_0 = \sum_0^\infty l_xb_x \tag{4}$$

☆*Change History*: March 2018. H.R. Pethybridge included glossary, keywords, extended text and updated references.

This is an update of D.H. LaFever, Age–Class Models, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 73–75.

Table 1 Hypothetical cohort life table including both mortality and fertility data

x	n_x	l_x	d_x	*Mortality rate* (q_x)	*Fertility schedule* (b_x)	l_xb_x	$x(l_xb_x)$
0	500	1.00	77	0.154	0.000	0.00	0
1	423	0.846	208	0.492	1.000	0.846	0.846
2	215	0.430	107	0.498	1.000	0.430	0.860
3	108	0.216	36	0.333	1.000	0.216	0.648
4	72	0.144	53	0.736	1.000	0.144	0.576
5	19	0.038			1.000	0.038	0.19

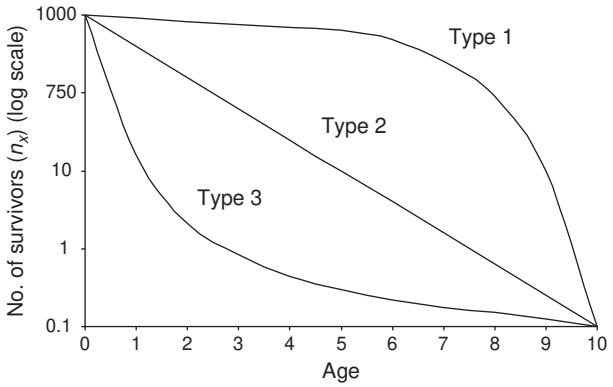


Fig. 1 Hypothetical survivorship curves (n_x). Adapted from Pearl, R. 1928. The Rate of Living. New York: Knopf.

The mean length of a generation:

$$G = \sum l_x b_x x / R_0 \tag{5}$$

The intrinsic capacity for increase (r), derived by Alfred Lotka in 1925, combines the natality and mortality demographic parameters. Lotka showed that a population subject to constant mortality and natality rates would gradually approach a fixed or stable age distribution. When a population has reached a stable age distribution, it will increase according to the differential equation:

$$\frac{dN}{dT} = rN \tag{6}$$

or in the integral form:

$$N_0 = N_0 e^{rt} \tag{7}$$

Based on the number of survivors, Raymond Pearl described three general types of survivorship curves on log-transformed scale (Fig. 1). A type 1 curve describes populations with low per capita mortality for most of the life span followed by high mortality of older organisms. A type 2 curve describes a linear survivorship curve that implies a constant per capita rate of mortality independent of age. Lastly, a type 3 curve describes a population with high per capita mortality early in life, followed by a period of much lower and relatively constant mortality. Humans in developed nations tend to follow the type 1 survivorship curve, while many birds exhibit a type 2 curve and many fish a type 3 curve. These curves are idealized and thus few, if any, populations have survivorship curves that exactly follow these curves.

Age-Structured Models

Compartmental Models

Compartment models are mathematical models and are commonly used to represent many different ecological systems. In age-structured compartment models, each age class is represented by a compartment or state variable. State variables are connected to the ages that precede and succeed it in time. Further, each age class can have separate mortality and natality rates. These models typically consist of systems of equations with an equation for each compartment. Individuals move through compartments by aging or are removed from the population by dying. These models are often used to determine the effects of different assumptions about the age, size, or spatial structure of a population on the dynamics of the population. Modeling approaches include using

(A)

$$\mathrm{M_A} = \begin{bmatrix} F_0 & F_1 & F_2 & \dots & F_{m-1} & F_m \\ P_0 & 0 & 0 & \dots & 0 & 0 \\ 0 & P_1 & 0 & \dots & 0 & 0 \\ 0 & 0 & P_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & P_{m-1} & 0 \end{bmatrix}$$

(B)

$$\mathrm{M_B} = \begin{bmatrix} G_1 & F_2 & F_3 & \dots & F_{m-1} & F_m \\ P_1 & G_2 & 0 & \dots & 0 & 0 \\ 0 & P_2 & G_3 & \dots & 0 & 0 \\ 0 & 0 & P_3 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & G_{m-1} & 0 \\ 0 & 0 & 0 & \dots & P_{m-1} & G_m \end{bmatrix}$$

Fig. 2 Population projection matrices. (A) An age-structured Leslie matrix ($\mathbf{M}_A$). (B) A stage-based Lefkovitch matrix $\mathbf{M}_B$. Subscripts represent the age or size class with m being the oldest or largest class, respectively.

differential equations, or difference equations. The type of mathematics used in the model depends on the questions asked by the researcher. Compartments may represent different parts of an ecosystem, stages of the development of an organism, or subdivisions of a population.

Matrix Models

Realized fecundity and survival rates change as organisms age. Matrix models separate populations into different age classes with each age class having the potential to possess different fecundity and survival rates. Matrix models are similar to compartment models, but matrix models are almost exclusively solved using linear algebra. The importance of specific population parameters for certain age classes can be tested using matrix models by investigating the sensitivity of the model to variations in parameter values for each age class. This is important for informing management decisions for target species, in particular, threatened and endangered species.

Leslie Matrix

The most common matrix model is an age-based matrix, also known as the Leslie matrix. In the age-based matrix, individuals in a population are separated into equal age classes (N_x) with each age class assigned a realized fecundity (F_x) and a probability of survival to the next age class (P_x). Within the transitional matrix $\mathbf{M}_A$, F_x is put along the top row and P_x along the subdiagonal, while all other elements in the matrix are assigned a value of zero (Fig. 2). Future population states can be projected by placing the current number of individuals in each age class into a vector $\mathbf{n}$ and obtaining the product of the $\mathbf{Mn}$ matrices:

$$\mathbf{M}\mathbf{n}_t = \mathbf{n}_{t+1} \tag{8}$$

where $\mathbf{n}_t$ is the vector whose elements are the number of individuals in each age class, and $\mathbf{n}_{t+1}$ is the vector containing the number of individuals in each age class at time $t + 1$.

Other Matrix Models

For some species, age is not a good indicator of condition. A stage-based matrix is a more general matrix model where life history stages replace age classes. The fundamental assumption of the stage-based matrix, also known as the Lefkovitch matrix, is that all individuals in a given stage are subject to identical mortality, growth, and fecundity schedules. The population projection matrix $\mathbf{M}_B$ describes the number of offspring born to each stage class that survive a given time period as well as the proportion of individuals in each stage class that survive and remain in that stage as compared with those that survive and enter a new stage (i.e., the transition probability). The elements of the matrix are F_x (stage-specific fecundity), G_x (probability of surviving and remaining in the same stage), and P_x (probability of surviving and growing to the next stage). Within the Lefkovitch matrix $\mathbf{M}_x$, F_x is put along the top row, P_x along the subdiagonal, and G_x along the diagonal while all other elements in the matrix are assigned a value of zero (Fig. 2). The primary differences between the Lefkovitch stage-class matrix and the Leslie age-class matrix are that the stages may incorporate several ages (i.e., they may differ in their duration and that individuals may remain in a stage from one time to the next).

Stochastic Models

The models so far discussed are deterministic models. This means that given certain initial conditions, there will only be one outcome. Stochastic models include the variation (i.e., the probabilistic nature) inherent in biological systems by representing one or more variables in the model as random variables. Stochastic variables can be based on a probability of occurrence (i.e., probability of breeding at any given time step, or probability of having a certain number of offspring during a given reproductive bout). These probabilities are generally assigned based on previous knowledge of the system. Random variables can also be chosen from a frequency distribution of historical data or from a statistical distribution fitted to historical data. With stochastic models and the addition of random variables one will not get the same output each time the model is run. When modeling an endangered species, for example, the population may grow after one simulation and decline to extinction after another.

See also: Behavioral Ecology: Age Structure and Population Dynamics. Ecological Data Analysis and Modelling: Ecological Models: Model Development and Analysis; Metapopulation Models; Parameterization; Ecological Models: Individual-Based Models. Evolutionary Ecology: Life-History Patterns. General Ecology: Demography; Generation Time; Growth Models

Further Reading

Brooks, E.N., Powers, J.E., Cortés, E., 2009. Analytical reference points for age-structured models: Application to data-poor fisheries. ICES Journal of Marine Science 67 (1), 165–175.
Caswell, H., 2001. Matrix population models: Construction, analysis, and interpretation, 2nd edn. Sunderland, MA: Sinauer Associates.
Hastings, A.e., 2013. Population biology: Concepts and models. In: Springer Science & Business Media.
Iannelli, M., Milner, F., 2017. The basic approach to age-structured population dynamics.
Krebs, C.J., 2001. Ecology, 5th edn. San Francisco, CA: Benjamin Cummings.
Leslie, P.H., 1948. Some further notes on the use of matrices in population mathematics. Biometrika 35, 213–245.
Masters, R.K., Reither, E.N., Powers, D.A., Yang, Y.C., Burger, A.E., Link, B.G., 2013. The impact of obesity on US mortality levels: The importance of age and cohort factors in population estimates. American Journal of Public Health 103 (10), 1895–1901.
Morris, J.A., Shertzer, K.W., Rice, J.A., 2011. A stage-based matrix population model of invasive lionfish with implications for control. Biological Invasions 13 (1), 7–12.
Schaub, M., Abadi, F., 2011. Integrated population models: A novel analysis framework for deeper insights into population dynamics. Journal of Ornithology 152 (1), 227–237.
Stott, I., Townley, S., Hodgson, D.J., 2012. A framework for studying transient dynamics of population projection matrix models. Ecology Letters 14 (9), 959–970.
Tuljapurkar, S., Caswell, H. (Eds.), . Structured-population models in marine, terrestrial, and freshwater systems,vol. 18. Springer Science & Business Media.
Yang, Y., Land, K.C., 2013. Age-period-cohort analysis: New models, methods, and empirical applications. CRC Press.

Relevant Websites

http://bandicoot.maths.adelaide.edu.au/Leslie_matrix/leslie.cgi
http://bioquest.org/esteem/esteem_details.php?product_id=210
https://www.r-bloggers.com/interactive-stage-structured-population-model-2/
http://slideplayer.com/slide/10772844/

Allopatry☆

Peter B Marko, University of Hawai'i at Manoa, Honolulu, HI, United States

Glossary

Adaptive A trait is adaptive if it results in greater fitness of individuals possessing that trait compared to individuals lacking that trait.

Clade A group of organisms that includes a common ancestor and all of the descendants of that ancestor. Synonymous with monophyletic.

Ecotypes A genetically distinct geographic form or type of a species which is adapted to a distinct habitat.

Fitness The ability of an individual to survive and reproduce in its environment.

Genetic drift Random changes in the genetic makeup of a population across generations.

Gene flow The movement and integration of alleles from one discrete population to another.

Molecular clock The concept that biomolecules, particularly DNA and RNA, evolve at a relatively constant rate, and that the number of molecular differences between species can be used to infer divergence times.

Natural selection Differences in the survival and reproduction of individuals possessing different traits or characteristics.

Reproductive isolation Intrinsic, genetically-based mechanisms that prevent individuals of different species from interbreeding. Isolating mechanisms operate either at prezygotic or postzygotic stages of sexual reproduction.

Sister-species Two species that are each others' closest relatives.

Introduction

Allopatry, meaning "in another place," describes populations or species that are physically isolated from other similar groups by a geographic barrier to dispersal (Fig. 1A). In the fields of ecology and evolutionary biology, allopatry is often used to describe populations that are geographically isolated to an extent that gene flow is severely reduced or absent, such that populations are evolutionarily independent. For that reason, allopatry is not usually defined by the characteristics of barriers per se, but by the effects of barriers on gene flow and genetic differentiation. For example, a wide river may restrict the dispersal of many small terrestrial mammals, but the same barrier will not hinder the movements of most birds.

"Allopatry" and "allopatric" are most often encountered in discussions of speciation, the process by which one species splits into two. In fact, allopatry is often used synonymously with "allopatric speciation," the process in which intrinsic (i.e., genetic) reproductive isolation evolves between geographically separated populations. Although new species may also arise from populations in sympatry (Fig. 1B, also see Sympatry) and parapatry (Fig. 1C), all evolutionary biologists agree that geographic isolation is a common, if not the most common, mechanism by which new species arise (Futuyma, 2013).

Origin and Early Use

The role of geographic isolation in the generation and maintenance of biological diversity has been long recognized by naturalists and evolutionary biologists, even well before publication of Charles Darwin's *The Origin of Species* in 1859. However, it was not until publication of seminal volumes by Theodosius Dobzhansky and Ernst Mayr, and subsequent research during the period of the mid-20th century known as the "Modern Synthesis" of evolutionary biology, that allopatry gained wide acceptance as a common mechanism of speciation.

By combining population genetics theory and observations of natural history, Dobzhansky's *Genetics and the Origins of Species* (Dobzhansky, 1937) described a basic model of speciation initiated by geographic separation: populations that were geographically isolated will, given enough time, accumulate genetic differences that would eventually cause them to lose their ability to interbreed. Later, in *Systematics and the Origins of Species* (Mayr, 1942), Mayr coined the term "allopatry" to describe two populations isolated in different geographic areas. Using examples and verbal theoretical arguments, Mayr argued persuasively that a severe reduction of gene flow between allopatric populations was typically the first step in the process of speciation.

Mayr called attention to the fact that most species exhibit substantial geographic variation, forming distinct races, varieties, and subspecies; the most distinct varieties were often geographically isolated with respect to the main range of the species. Mayr

☆*Change History*: January 2018. P. B. Marko made minor changes to the text, figures and references.

This is an update of P.B. Marko, Allopatry, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 131–138.

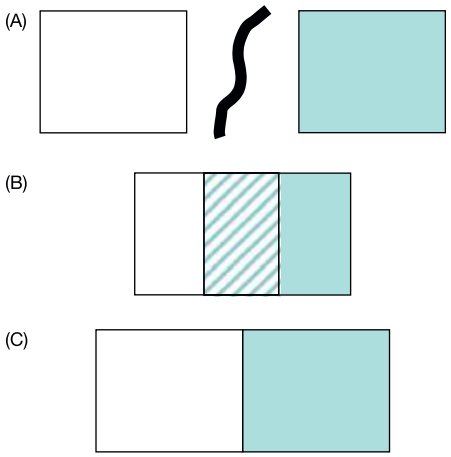


Fig. 1 Diagrammatic representation of (A) allopatric, (B) sympatric, and (C) parapatric distributions. Allopatric taxa are often separated by a physical barrier to dispersal, such as a river or mountain range. Sympatric taxa may have partially or completely overlapping geographic distributions, whereas parapatric taxa have abutting or continguous distributions.

asserted that geographic variation represented adaptation to local environments and that the features that distinguish local geographic variants are usually the same kind that typically differentiate reproductively isolated species. In other words, Mayr felt that distinct geographic variants represent intermediate stages in the process of speciation. Both Dobzhansky and Mayr reasoned that completely geographically isolated populations would ultimately evolve reproductive isolation as a byproduct of adaptive divergence to different geographic regions. Like Dobzhansky, Mayr hypothesized that natural selection, brought about by different environmental or ecological conditions, causes an accumulation of adaptive genetic changes between allopatric populations that eventually result in development incompatibilities that cause either sterility or inviability in hybrids. Therefore, if the barrier separating newly formed allopatric species broke down and the two species came into contact, they would remain genetically and phenotypically distinct due to the evolution of reproductive isolation, a byproduct of adaptive allopatric divergence. Hybrid inviability and sterility are examples of "postzygotic" isolating mechanisms because the isolating mechanism affects individuals after a zygote is formed. Alternatively, allopatric populations might evolve "prezygotic" isolating mechanisms that prevent successful mating or zygote formation, such as courtship differences or gametic incompatibilities.

Types of Allopatric Speciation

Vicariant Speciation

In vicariant allopatric speciation, a species is split into two relatively large populations by the formation of a barrier to dispersal (**Fig. 2**), such as mountain building, glaciation, changes in ocean circulation, or even human activities that disrupt gene flow between populations. Vicariant speciation corresponds to the geographic mode of speciation emphasized by Mayr in *Systematics*, and since then, a steady accumulation of case studies provide good evidence that the formation of geographic barriers are often involved in speciation. Most examples focus on sister-species found on opposite sides of geographic barriers in which the geological history of the formation of the barrier is well-understood. For example, drops in sea level during the Pleistocene created the Indo-Pacific barrier, a nearly continuous land barrier between Asia and Australia that subdivided many marine species into genetically distinct Indian and Pacific oceans lineages (e.g., Gaither *et al.*, 2011).

The most compelling examples of vicariant speciation involve species whose biogeographical histories are also temporally consistent with the history of barrier formation. For example, molecular-clock analyses of flowering plants show a striking temporal congruence in divergence times between many North American and Asian species pairs and the loss of continuous Northern Hemisphere mesophytic forest in the fossil record (Xiang *et al.*, 2000). Similarly, divergence times between sister-species of marine invertebrates and fish found on opposite sides of the Isthmus of Panama indicate that most sister-species were formed in the last 12 million years, corresponding to gradual isthmus uplift that incrementally interrupted interoceanic exchange during this period (O'Dea *et al.*, 2016). However, molecular data also indicate that roughly one-third of all transisthmian species pairs diverged before any restriction on water flow between the two oceans. This pseudo-congruence (Cunningham and Collins, 1994), a lack of temporal coincidence between divergence times and the time of formation of a geographic barrier, may be explained by different rates of molecular evolution in different taxa, but is more likely a consequence of extinction of one of the two daughter species created by the Isthmus (Marko *et al.*, 2015). Similarly, molecular data indicate that spatially congruent patterns of genetic differentiation in the southeastern USA (associated with the Atlantic vs. Gulf drainage pattern), likely arose at different times, possibly by different mechanisms (Near and Peck, 2005; Soltis *et al.*, 2006).

Because landscapes and seascapes are constantly changing, many species formed in allopatry will likely establish sympatric distributions in response to the breakdown of geographic barriers to dispersal. For that reason, Mayr argued that recently-diverged

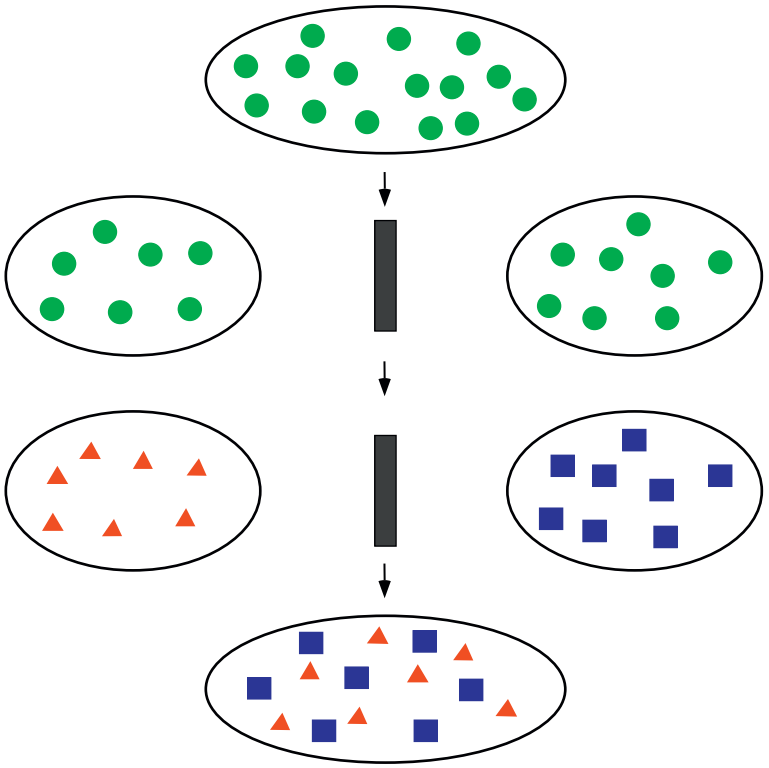


Fig. 2 Vicariant allopatric speciation. The appearance of a physical barrier divides a broadly distributed species into two physically isolated populations that diverge in traits that cause reproductive isolation in hybrids if the barrier breaks down and the two species come back into contact. The evolution of an intrinsic isolating mechanism causes both species to remain distinct when they develop overlapping geographic distributions.

species should be most informative about the geography of speciation because they are more likely (than relatively ancient speciation events) to reflect the spatial distribution of populations at the time of speciation (Mayr, 1954a). Using that rationale, molecular phylogenetic studies have compared species divergence times within clades to corresponding patterns of geographic overlap. For both terrestrial and marine organisms, these types of studies have found that the most sister-species—the most recently separated lineages within clades—are allopatric (Berlocher, 1998; Coyne and Price, 2000; Fitzpatrick and Turelli, 2006; Quenouille *et al.*, 2011; Choat *et al.*, 2012).

The Genetics of Vicariant Speciation

Natural selection is widely regarded as an important evolutionary force causing divergence in vicariant speciation, especially if populations are large (and have high genetic variation for selection to act on) at the time of speciation. How this actually happened was not immediately obvious to evolutionary biologists because it is difficult to explain how a beneficial mutation at one locus, present in only a single individual, could be maintained if it caused hybrid sterility or inviabilty given that the mutation would be initially only exist in heterozygotes? Dobzhansky (1937) and Muller (1942) provided a simple, yet elegant solution in the form of a two-locus model that approximates the kinds of genetic changes envisioned by Mayr, and can be easily generalized to multiple loci. Consider a population fixed for the two-locus genotype $A_1A_1B_1B_1$ that is split into two populations by the formation of a barrier restricting gene flow. If selection in one population favors replacement of the A_1 allele with a newly mutated A_2 allele, but in the other population B_1 is selectively replaced by B_2, the two populations will be fixed for the $A_2A_2B_1B_1$ and $A_1A_1B_2B_2$ genotypes, respectively. If the A_2 and B_2 alleles are incompatible, causing hybrid sterility or inviability (i.e., complete postzygotic isolation), the two populations will be reproductively isolated if they come back into contact. Many examples of Dobzhansky-Muller incompatibilities have been described in plants and animals, particularly those for which closely related species can be hybridized in the laboratory (e.g., Coyne and Charlesworth, 1986; Orr, 1995; Turelli, 1998).

Evidence of Natural Selection

Several investigators have shown that reproductive isolation has evolved as a by-product of adaptive divergence, a process called "ecological speciation." For example, the evolution of mimicry appears to have played an important role in speciation in the butterfly genus *Heliconius* (Jiggins *et al.*, 2001). The recently split sister species *H. melpomene* and *H. cydno* have diverged to mimic

the color patterns of different model taxa (Fig. 3A), but strong assortative mating (preferential mating among individuals with similar phenotypes) based on the mimetic coloration results in substantial prezygotic isolation (Fig. 3B). Rare hybridization events produce individuals with poorly adapted intermediate phenotypes, demonstrating that divergent patterns of mimicry also result in some postzygotic isolation. Among plants, species in the genus *Mimulus* have evolved highly divergent floral morphologies that appear to be adaptations to different types of pollinators (either bees or hummingbirds) that in turn effectively reproductively isolate sympatric populations (Schemske and Bradshaw, 1999). Using controlled laboratory experiments, several studies have also generated strong assortative mating among replicate lines of houseflies and species of *Drosophila* by subjecting them to artificial divergent selection on behavioral, morphological, and physiological traits (e.g., Dodd, 1989). Control populations exposed to the same selective pressures show little behavioral isolation, indicating that divergent selection, rather than genetic drift, caused the incidental evolution of prezygotic isolation.

Sexual selection is also probably important in allopatric speciation given that divergence in sexually selected traits will necessarily diminish interbreeding (Ritchie, 2007). Many experiments have shown that among species where a female chooses males to mate with, females make choices based on sexually dimorphic traits in males, such as large body size, bright coloration, large antlers, and elongated tail feathers. Sexual selection is likely involved in the evolution of gametic (sperm and egg) incompatibilities and isolation in some marine organisms that release their gametes into the water column in mass spawnings, a scenario that results in intense sperm competition among males. Analysis of the underlying DNA sequences for some sperm and egg surface proteins indicates that the evolution of these proteins is driven by selection (Galindo *et al.*, 2003). Rapid development of gametic incompatibility may explain how some populations evolve reproductive isolation during only brief periods of transient geographic isolation (Levitan and Ferrell, 2013).

Genome-wide patterns of genetic divergence between evolving species or ecotypes have provided numerous examples of the potential role of natural selection in speciation. Scanning thousands of loci across the genome allows identification of very strongly differentiated loci ("outliers") between geographically isolated populations, a pattern often interpretted as evidence of natural selection (Luikart *et al.*, 2003). However, as with any difference between recently-evolved species or ecotypes, it remains difficult to determine if outlier loci are a cause or consequence of reproductive isolation. Genome scans of allopatric ecotypes that have evolved more than once (repeated or parallel evolution) provide much greater power to identify the genetic loci responsible for adaptation and reproductive isolation between allopatric lineages. For example, in sticklebacks, whole-genome sequencing of

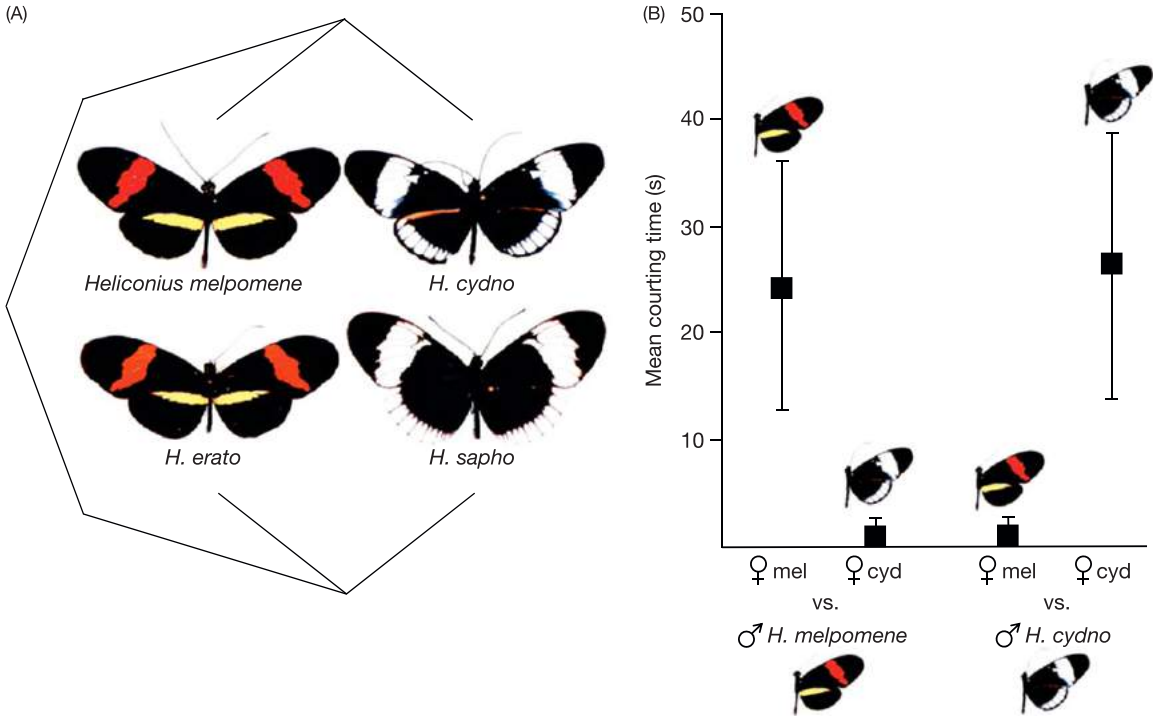


Fig. 3 Ecological speciation in *Heliconius* butterflies in which strong behavioral (i.e., prezygotic) isolation evolved as a by-product of mimicry. (A) Molecular phylogenetic relationships of *H. melpomene*, *H. cydno*, and their comimics *H. erato* and *H. sapho*. The molecular phylogeny establishes that similar coloration patterns involved the evolution of mimicry in one or both of the *H. melpomene*/*H. cydno* and *H. erato*/*H. sapho* lineages (rather than entirely as a consequence of common ancestry). (B) Time spent by males courting females with 95% confidence intervals for both *H. melpomene* and *H. cydno* from sympatric populations in Panama. Allopatric populations show weaker isolation, a pattern consistent with reinforcement of speciation. From Jiggins, C. D., Naisbit, R.E., Coe, R.L., and Mallet, J. (2001). Reproductive isolation caused by color pattern mimicry. *Nature* **411**, 302–305. Photos courtesy of C. Jiggins.

several freshwater ecotypes, that each repeatedly evolved independently from a marine ancestor, reveals that many outlier loci are shared among freshwater ecotypes (Jones *et al.*, 2012).

Reinforcement

During a period of allopatric isolation, populations may evolve only partial postzygotic isolation such that if they came back into contact they would produce viable hybrid offspring with lower fitness compared to pure genotypes. Dobzhansky introduced the concept that if hybrid genotypes have reduced fitness, natural selection should favor the evolution of enhanced mate discrimination. This process, in which prezygotic isolation evolves as an adaptation to minimize the costly production of low fitness hybrids, is called "reinforcement of speciation" or simply "reinforcement." Although intuitively appealing, several significant theoretical objections to reinforcement were initially raised and the idea quickly fell out of favor with most evolutionary biologists. However, the emergence of several compelling examples (e.g., Coyne and Orr, 1989) of reproductive character displacement, a pattern in which prezygotic isolation is stronger between species in sympatry than in allopatry, has fostered a revival of the hypothesis of reinforcement (Servedio and Noor, 2003). That said, even though the pattern of stronger behavioral isolation in sympatry than in allopatry is consistent with reinforcement, other processes can often explain the pattern of "reproductive character displacement." For example, populations with strong premating isolating mechanisms that evolved in allopatry may simply be capable of coexisting, whereas populations with weaker isolating mechanisms may freely interbreed and genetically fuse back into a single population when they come into contact. The expectation of population fusion, however, assumes that many evolving allopatric species pairs should have strong prezygotic isolation, a pattern that is not consistent with allopatric sister-species of *Drosophila* (Yukilevich, 2012). Similarly, prezygotic isolation appears to lag behind postzygotic isolation in snapping shrimp species isolated by the Isthmus of Panama (Knowlton *et al.*, 1993). Although recent experimental work with *Drosophila* confirms what theoretical studies have established, that reinforcement can evolve relatively quickly if gene flow is low or if selection against hybridization is strong (Matute, 2010), evidence for reinforcement in natural populations remains equivocal.

Peripatric Speciation

A second type of allopatric speciation is peripatric speciation, a theory developed by Mayr (1954b) more than 10 years after the publication of *Systematics*. From both biogeographical and genetic points of view, peripatric speciation represents a significant departure from the speciation process originally outlined by Mayr. Peripatric speciation, also called "founder effect speciation," occurs when a few individuals establish a small subpopulation on the periphery of the main range of a species and evolve reproductive isolation (**Fig. 4**). Under this model, the more widespread species is considered "parental" or "ancestral" with respect to the small, peripherally isolated "daughter" or "derived" species. Peripatric speciation was developed from observations by Mayr, mainly of birds on islands in the south Pacific, that small, peripherally isolated populations are often highly morphologically divergent despite existing under ecological conditions similar to those experienced by the parental population (Mayr, 1963).

The Genetics of Peripatric Speciation

An important component of Mayr's peripatric theory—and a significant shift in his opinion about the genetic basis of adaptation—is that gene flow, stabilizing selection, and complex epistatic interactions among numerous genetic loci all place constraints on phenotypic evolution. Therefore, Mayr felt that unusual genetic conditions might be necessary for speciation, suggesting that a new species may arise if the "sampling" of a small number of colonists leaves a small population with a highly nonrepresentative sample of genetic variation at some loci. The genetic drift associated with this sampling of individuals is called "the founder effect." As a result of epistatic interactions among different genes, Mayr argued that a substantial change in allele frequencies at a small number of loci will briefly change the relative selective values of alleles at many other loci across the genome, triggering massive genetic change and ultimately reproductive isolation (Mayr, 1954b, 1963).

Unlike the Dobzhansky-Muller model of adaptive divergence, in which ecological differences experienced by allopatric populations drive phenotypic divergence that leads to reproductive isolation, Mayr's original peripatric model contends that changes in the 'genetic environment' of the peripheral isolate, caused by the founder effect, allow natural selection to generate novel combinations of alleles at numerous interacting genetic loci ('adaptive gene complexes'). Mayr noted that due to geographical localization and rapid phenotypic evolution within peripheral isolates, intermediate phenotypes are probably rarely recorded in the fossil record. This last idea became an important component of Eldredge and Gould's (1972) highly influential but contentious theory of "punctuated equilibrium," which united Mayr's peripatric model with the observation that new species often appear suddenly in the fossil record, fully morphologically differentiated from their closest relative.

Evidence for and against Peripatric Speciation

Peripatric speciation was favored by many evolutionary biologists in the 1970s and 1980s. Widespread support for peripatric speciation was fostered in part by extensive research on Hawaiian picture-winged *Drosophila* (**Fig. 5**), led by Hampton Carson, Kenneth Kaneshiro, and Allan Templeton (Carson and Kaneshiro, 1976; Carson, 1983; Carson and Templeton, 1984). High

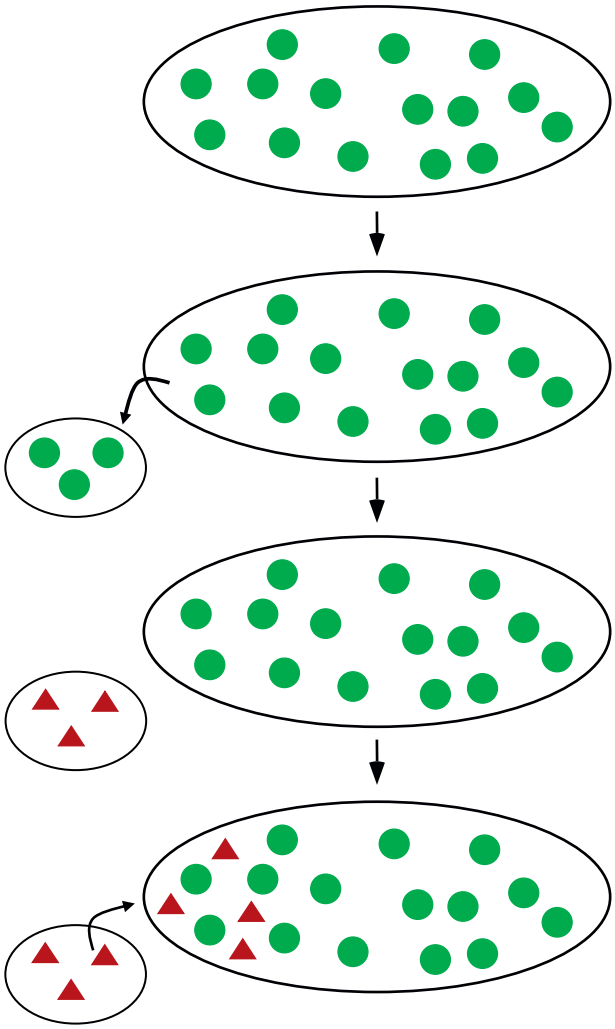


Fig. 4 Successive stages in the process of peripatric speciation. A small "daughter" population on the periphery of a more widespread "parental" population evolves reproductive isolation. Both species remain distinct if the peripherally isolated species invades the geographic range of the parental species. Under this model, the peripherally isolated population diverges from the parental population such that the latter remains unchanged.

endemnicity of species to single islands, with more recently derived species tending to occur on geologically younger islands, provides clear evidence that most speciation events in this group are associated with the colonization of newly formed islands from parental populations on older islands. Additional biogeographical evidence from a wide variety of island taxa provides a obvious link between speciation and dispersal, such as for flightless crickets in the Hawaiian genus *Laupala* (Mendelson and Shaw, 2005) and Darwin's finches in the Galapagos (Vincek *et al.*, 1997). The leaders of the *Drosophila* research program also developed their own genetic models of divergence. Although their models differed somewhat from Mayr's, most emphasized the role of the founder effect and the importance of changes in the genetic environment within a peripheral isolate.

Evidence of peripatric speciation is harder to come by for continental organisms, possibly because of a greater likelihood that the biogeographical evidence will be lost as a consequence of postspeciation range expansions. However, by comparing the relative sizes of the geographic distributions of species to their times of divergence, Barraclough and Vogler (2000) showed that many recently split lineages from a wide variety of taxa show unexpectedly high asymmetry with respect to their geographic ranges. Although other explanations are possible, a large asymmetry in geographic range between parental and daughter species at the time of speciation is consistent with peripatric speciation.

Phylogeographic studies also provide support for peripatric speciation if a putative daughter species possesses a small subset of the genetic diversity found in the parental species, indicating that the daughter species was formed from a relatively small population (Ovenden and White, 1990); the parental species' genome may also have many loci that are paraphyletic with respect to the daughter species, meaning that some alleles or haplotypes found in the parent are more closely related to those found in the daughter species, but all alleles or haplotypes in the daughter species are derived exclusively from a single lineage (Harrison, 1991; Marko, 1998). One of the most widely cited examples of peripatric speciation comes from a pair of species in the plant genus

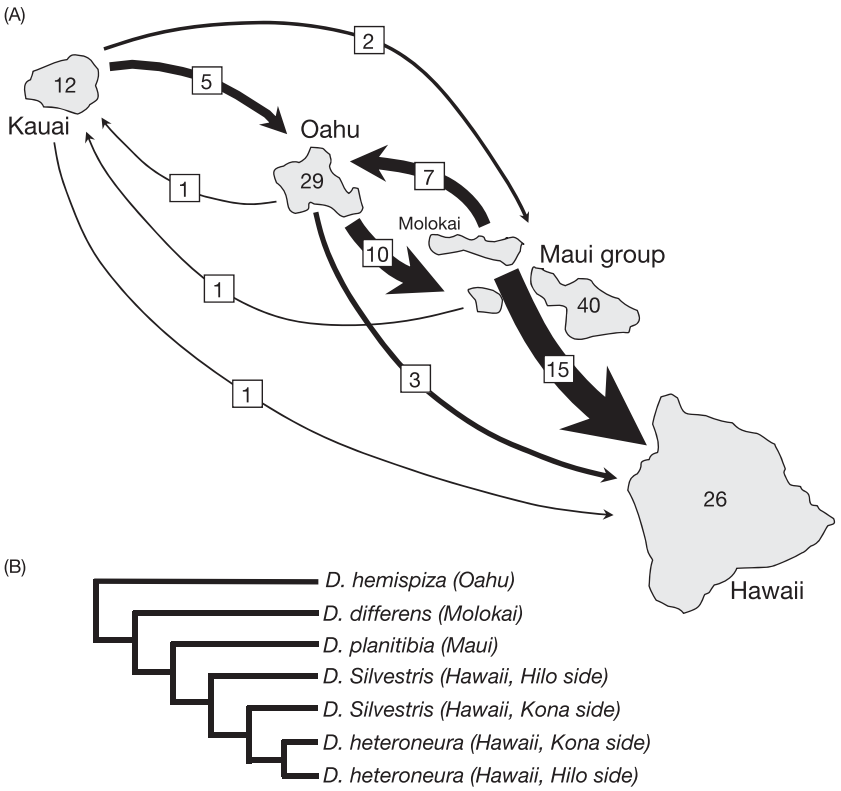


Fig. 5 Inter-island colonization and speciation in picture-winged *Drosophila* of Hawai'i. (A) Arrows indicate the minimum number of colonization events between islands, with the width of arrows reflecting the relative number of dispersal events (actual numbers are given in boxes). Note that colonization events are predominantly from older to younger islands. Numbers inside each island (or island group, in the case of Maui and Moloka'i) are the number of species found on each. (B) Phylogenetic relationships based on mitochondrial DNA for a subset of these species. More recently evolved species are found on younger islands, a pattern that corroborates the hypothesis that speciation is associated with dispersal between islands. Genetically distinct lineages of both *D. silvestris* and *D. heteroneura* on the Kona (southwest) and Hilo (northeast) sides of the island of Hawai'i suggest that intra-island speciation has also occurred in this group. (A) Modified from Carson, H. L. (1983). Chromosomal sequences and interisland colonizations in Hawaiian Drosophila. *Genetics* **103**, 465–482. (B) Reproduced from DeSalle, R. and Giddings, L. V. (1986). Discordance of nuclear and mitochondrial DNA phylogenies in Hawaiian Drosophila. *Proceedings of the National Academy of Sciences of the United States of America* **83**, 6902–6906, with permission.

Clarkia, in which the putative daughter species is restricted to only two sites at the edge of the geographic range of the parental species in central California; the daughter species possesses less phenotypic variation and only a subset of the genetic diversity of the parental species (Gottlieb, 2004). This mode of speciation could be common in plants, in which relatively small genetic changes have the potential to rapidly reproductively isolate a localized population.

Even if many species arise as a consequence of colonization events (the existence of many endemic species on isolated islands and archipelagos makes this conclusion self-evident), the mechanism driving genetic divergence in a peripheral isolate cannot necessarily be attributed to the founder effect. In fact, as biogeographical evidence supporting peripatric speciation has grown, support for the role of the founder effect has diminished, mainly for two reasons. First, few species putatively formed in peripatry show strong genetic evidence of a severe bottleneck in their recent past (Barton and Charlesworth, 1984). Furthermore, even if a putative daughter species has low genetic variation across its entire geographic range, it remains difficult to determine if a bottleneck happened at the same time as speciation or after (Harrison, 1991). For example, low genetic variation throughout nearly all of the geographic of the marine snail *N. emarginata* (relative to it's sister-species) in California is consistent with peripatric speciation, except that the southermost population of this species has as much genetic variation as any population in the parental species, indicating that a severe bottleneck was not involved in speciation (Marko, 1998). Extinction of (or failure to sample) the high diversity population, would easily lead to the erroneous conclusion of a founder effect. In this case, low diversity across most of the daughter species' range is likely a consequence of a postglacial population expansion that happened after a period of allopatry.

A second reason why the popularity of genetic models of peripatric speciation has waned is that theoretical population geneticists have developed persuasive arguments against the potential role of the founder effect (Barton and Charlesworth, 1984; Charlesworth, 1997). Although more recent models of founder effect speciation have been more successful demonstrating that bottlenecks may actually facilitate speciation (Gavrilets and Hastings, 1996), laboratory tests involving bottlenecked populations of fruit flies (*Drosophila*) provide equivocal results (Powell, 1978; Dodd and Powell, 1985). In fact, the mixture of outcomes from

these experiments have been interpreted by both proponents and critics of peripatric speciation as either favoring or refuting, respectively, the importance of the founder effect in speciation.

References

Barraclough, T.G., Vogler, A.P., 2000. Detecting the geographic pattern of speciation from species-level phylogenies. American Naturalist 155, 419–434.

Barton, N.H., Charlesworth, B., 1984. Genetic revolutions, founder effects, and speciation. Annual Review of Ecology and Systematics 15, 133–164.

Berlocher, S.H., 1998. Can sympatric speciation be proven from biogeographic and phylogenetic evidence? In: Howard, D.J., Berlocher, S.H. (Eds.), Endless forms: Species and speciation. New York: Oxford University Press, pp. 99–113.

Carson, H.L., Kaneshiro, K., 1976. *Drosophila* of Hawaii: Systematics and ecological genetics. Annual Review of Ecology and Systematics 7, 311–345.

Carson, H.L., 1983. Chromosomal sequences and interisland colonizations in Hawaiian *Drosophila*. Genetics 103, 465–482.

Carson, H.L., Templeton, A.R., 1984. Genetic revolutions in relation to speciation phenomena: The founding of new populations. Annual Review of Ecology and Systematics 15, 97–132.

Charlesworth, B., 1997. Is founder-flush speciation defensible? American Naturalist 149, 600–603.

Choat, J.H., Klanten, O.S., Van Herwerden, L., Robertson, D.R., Clements, K.D., 2012. Patterns and processes in the evolutionary history of parrotfished (Family Labridae). Biological Journal of the Linnean Society 107, 529–557.

Coyne, J.A., Charlesworth, B., 1986. Location of an X-linked factor causing male sterility in hybrids of *Drosophila simulans* and *D. mauritiana*. Heredity 57, 243–246.

Coyne, J.A., Orr, H.A., 1989. Patterns of speciation in *Drosophila*. Evolution 43, 362–381.

Coyne, J.A., Price, T., 2000. Little evidence for sympatric speciation in island birds. Evolution 54, 2166–2171.

Cunningham, C.W., Collins, T.M., 1994. Developing model systems for molecular biogeography: Vicariance and interchange in marine invertebrates. In: Schierwater, B., Streit, B., Wagner, G.P., DeSalle, R. (Eds.), Molecular ecology and evolution: Approaches and applications. Basel: Birkhauser Verlad, pp. 405–433.

Dobzhansky, T., 1937. Genetics and the origin of species. New York: Columbia University Press.

Dodd, D.M.B., 1989. Reproductive isolation as a consequence of adaptive divergence in *Drosophila pseudoobscura*. Evolution 43, 1308–1311.

Dodd, D.M., Powell, J.R., 1985. Founder-flush speciation, an update on experimental results with *Drosophila*. Evolution 39, 1388–1392.

Eldredge, N., Gould, S.J., 1972. Punctuated equilibria: An alternative to phyletic gradualism. In: Schopf, T.M. (Ed.), Models in Palaeobiology. San Francisco: Freeman Cooper, pp. 82–115.

Futuyma, D.J., 2013. Evolution, 3rd edn. Sunderland, MA: Sinauer.

Fitzpatrick, B.M., Turelli, M., 2006. The geography of mammalian speciation: Mixed signals from phylogenies and range maps. Evolution 60, 601–615.

Gaither, M.R., Bowen, B.W., Bordenave, T.-R., Rocha, L.A., Newman, S.J., Gomex, J.A., van Herwerden, L., Craig, M.T., 2011. Phylogeography of the reef fish *Cephalopholis argus* (Epinephelidae) indicates Pleistocene isolation across the indo-pacific barrier with contemporary overlap in the coral triangle. BMC Evolutionary Biology 11, 189.

Galindo, B.E., Vacquier, V.D., Swanson, W.J., 2003. Positive selection in the egg receptor for abalone sperm lysine. Proceedings of the National Academy of Sciences USA 100, 4639–4643.

Gavrilets, S., Hastings, A., 1996. Founder effect speciation: A theoretical assessment. The American Natural 147, 466–491.

Gottlieb, L.D., 2004. Rethinking classic examples of recent speciation in plants. New Phytologist 161, 71–82.

Harrison, R.G., 1991. Molecular changes at speciation. Annual Review of Ecology and Systematics 22, 281–308.

Jiggins, C.D., Naisbit, R.E., Coe, R.L., Mallet, J., 2001. Reproductive isolation caused by colour pattern mimicry. Nature 411, 302–305.

Jones, F.C., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M.C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M.C., Myers, R.M., Miller, C.T., Summers, B.R., Knecht, A.K., Brady, S.D., Zhang, H., Pollen, A.A., Howes, T., Amemiya, C., Broad Institute Genome Sequencing Platform & Whole Genome Assembly Team, Baldwin, J., Bloom, T., Jaffe, D.B., Nicol, R., Wilkinson, J., Lander, E.S., Di Palma, F., Lindblad-Toh, K., Kingsley, D.M., Jones, F.C., Grabherr, M.G., Chan, Y.F., *et al*., 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484, 55–61.

Knowlton, N., Weigt, L.A., Solorzano, L.A., Mills, D.K., Bermingham, E., 1993. Divergence in proteins, mitochondrial DNA, and reproductive compatibility across the isthmus of Panama. Science 260, 1629–1632.

Levitan, D.R., Ferrell, D.L., 2013. Selection on gamete recognition proteins depends on sex, density, and genotype frequency. Science 312, 267–269.

Luikart, G., England, P.R., Tallmon, D., Jordan, S., Taberlet, P., 2003. The power and promise of population genomics: From genotyping to genome typing. Nature Reviews Genetics. 981–994.

Marko, P.B., 1998. Historical allopatry and the biogeography of speciation in the prosobranch snail genus *Nucella*. Evolution 52, 757–774.

Marko, P.B., Eytan, R.I., Knowlton, N., 2015. Do large molecular sequence divergences imply an early closure of the isthmus of Panama? Proceedings of the National Academy of Science USA. doi:10.1073/pnas.1515048112.

Matute, D.R., 2010. Reinforcement can overcome gene flow during speciation in *Drosophila*. Current Biology 20, 2229–2233.

Mayr, E., 1942. Systematics and the origin of species. New York: Columbia University Press.

Mayr, E., 1954a. Speciation in tropical echinoids. Evolution 8, 1–18.

Mayr, E., 1954b. Changes in genetic environment and evolution. In: Huxley, J., Hardy, A.C., Ford, E.B. (Eds.), Evolution as a process. London: Allen and Unwin, pp. 157–180.

Mayr, E., 1963. Animal species and evolution. Cambridge, MA: Harvard University Press.

Mendelson, T.C., Shaw, K.L., 2005. Rapid speciation in an arthropod. Nature 433, 375–376.

Muller, H.J., 1942. Isolating mechanisms, evolution, and temperature. Biology Symposium 6, 71–125.

Near, T.J., Peck, B.P., 2005. Dispersal, vicariance, and timing of diversification in Nothonotus darters. Molecular Ecology 14, 3485–3496.

O'Dea, A., Lessios, H.A., Coates, A.G., Eytan, R.I., Restrepo-Moreno, S.A., Cione, A.L., Collins, L.S., de Queiroz, A., Farris, D.W., Norris, R.D., Stallard, R.F., Woodburne, M.O., Aguilera, O., Aubry, M.-P., Berggren, W.A., Budd, A.F., Cozzuol, M.A., Coppard, S.E., Duque-Caro, H., Finnegan, S., Gasparini, G.M., Grossman, E.L., Johnson, K.G., Keigwin, L.D., Knowlton, N., Leigh, E.G., Leonard-Pingel, J.S., Marko, P.B., Pyenson, N.D., Rachello-Dolmen, P.G., Soibelzon, E., Soibelzon, L., Todd, J.A., Vermeij, G.J., Jackson, J.B.C., 2016. Formation of the isthmus of Panama. Science Advances 2. e1600883doi:10.1126/sciadv.1600883.

Orr, H.A., 1995. The population genetics of speciation: The evolution of hybrid incompatibilities. Genetics 139, 1805–1813.

Ovenden, J.R., White, R.W.G., 1990. Mitochondrial and Allozyme Geneticsof incipient speciation in a landlocked population of *Galaxias truttaceus* (Pisces: Galaxiida). Genetics 124, 701–716.

Powell, J.R., 1978. The founder-flush theory speciation theory: An experimental approach. Evolution 32, 465–474.

Quenouille, B., Hubert, N., Bermingham, E., Planes, S., 2011. Speciation in tropical seas: Allopatry followed by range changes. Molecular Phylogenetics and Evolution 58, 546–552.

Ritchie, M.G., 2007. Sexual selection and speciation. Annual Review of Ecology, Evolution, and Systematics 38, 79–102.

Servedio, M.R., Noor, M.A.F., 2003. The role of reinforcement in speciation: Theory and data. Annual Review of Ecology, Evolution, and Systematics 34, 339–364.

Schemske, D.W., Bradshaw, H.D., 1999. Pollinator preference and the evolution of floral traits in monkeyflowers (*Mimulus*). Procedings of the National Academy of Sciences, USA 96, 11910–11915.

Soltis, D.E., Morris, A.B., McCLauchlan, J.S., Mano, P.S., Soltis, P.S., 2006. Comparative Phylogeography of unglaciated eastern North America. Molecular Ecology 15, 4261–4293.
Turelli, M., 1998. The causes of Haldane's rule. Science 282, 889–891.
Vincek, V., O'Huigin, C., Satta, Y., Takahata, Y., Boag, P.T., Grant, P.R., Grant, B.R., Klein, J., 1997. How large was the founding population of Darwin's finches? Proceedings of the Royal Society London B 264, 111–118.
Xiang, Q.-Y., Soltis, D.E., Soltis, P.S., Manchester, S.R., Crawford, D.J., 2000. Timing of the eastern Asian-eastern North American floristic disjunction: Molecular clocks confirm paleontological estimates. Molecular Phylogenetics and Evolution 15, 462–472.
Yukilevich, R., 2012. Asymmetrical patterns of speciation uniquely support reinforcement in *Drosophila*. Evolution 66, 1430–1446.

Further Reading

Coyne, J.A., Orr, H.A., 2004. Speciation. Sunderland, Massachusetts: Sinauer.
Cracraft, J., 1982. Geographic differentiation, cladistics, and vicariance biogeography: Reconstructing the tempo and mode of evolution. American Zoologist 22, 411–424.
Bull, J.W., Maron, M., 2016. How humans drive speciation as well as extinction. Proceedings of the Royal Society London B 283.20160600
Bush, G.L., 1975. Modes of animal speciation. Annual Review of Ecology and Systematics 6, 339–364.
Jordan, D.S., 1905. The origin of species through isolation. Science 22, 545–562.
Lande, R., 1980. Genetic variation and phenotypic evolution during allopatric speciation. The American Naturalist 116, 463–479.
Rice, W.R., Hostert, E.E., 1993. Laboratory experiments on speciation: What have we learned in 40 years? Ecography 47, 1637.
Schilthuizen, M., 2001. Frogs, flies, and dandelions: Speciation—The evolution of new species. New York: Oxford University Press.
Slatkin, M., 1996. In defense of founder-flush theories of speciation. American Naturalist 147, 493–505.
Wiens, J.J., 2004. What is speciation and how should we study it? The American Naturalist 163, 914–923.
Zimmer, C., Emlen, D.J., 2013. Evolution: Making sense of life. Greenwood Village, Colorado: Roberts and Company.

Relevant Websites

http://evolution.berkeley.edu/evolibrary/home.php—Understanding Evolution.
http://www.pbs.org/wgbh/evolution/index.html—PBS Evolution Library.
DeSalle, R., Giddings, L.V., 1986. Discordance of nuclear and mitochondrial DNA phylogenies in Hawaiian *Drosophila*. Proceedings of the National Academy of Sciences USA 83, 6902 6906.
Schluter, D., Conte, G., 2009. Genetics and ecological speciation. Proceedings of the National Academy of Sciences, USA 106, 9955 9962.

Animal Physiology

CE Cooper, Curtin University of Technology, Bentley, WA, Australia
PC Withers, University of Western Australia, Crawley, WA, Australia

Animal Physiology

Animal physiology is the study of how animals work, and investigates the biological processes that occur for animal life to exist. These processes can be studied at various levels of organization from membranes through to organelles, cells, organs, organ systems, and to the whole animal. Animal physiology examines how biological processes function, how they operate under various environmental conditions, and how these processes are regulated and integrated. The study of animal physiology is closely linked with anatomy (i.e., the relationship of function with structure) and with the basic physical and chemical laws that constrain living as well as nonliving systems. Although all animals must function within basic physical and chemical constraints, there is a diversity of mechanisms and processes by which different animals work. A comparative approach to animal physiology highlights underlying principles, and reveals diverse solutions to various environmental challenges. It can reveal similar solutions to a common problem, or modifications of a particular physiological system to function under diverse conditions. The discipline of animal physiology is diverse and here the major areas of research and investigation are outlined.

Homeostasis and Regulation

An important characteristic of animals is the ability to self-regulate the extracellular environment in which their cells are bathed and function. The extracellular environment is a buffer between the intracellular environment and the external environment of an animal, which consists of an aquatic or terrestrial environment in exchange with the atmosphere. These external environments can be highly variable with respect to their physical characteristics, which would affect the intracellular physiological processes necessary for animals to function. Therefore, some aspects of the intracellular environment of an animal are invariably kept different from their external environment. Consequently, an important role of homeostasis in animals is the regulation of aspects of the extracellular environment different from the external environment to provide an optimal internal environment in which the cells function.

Homeostasis is an underlying principle of animal physiology, and physiological systems are the means by which homeostasis is maintained. Homeostatic processes maintain the internal environment, although not all animals regulate all physiological variables to the same extent. Animals may conform with respect to some physiological variables, with the internal variable the same as for the external environment.

For both conformers and regulators, there is a range of environmental conditions over which the animal can survive. Beyond this range, conformers experience sufficient change in the internal environment that physiological processes no longer function effectively, and regulators can no longer regulate against the environmental gradient and their internal environment changes sufficiently to prevent normal physiological function.

Homeostasis does not necessarily require a regulatory mechanism. Equilibrium homeostasis and steady-state homeostasis are nonregulatory means by which an internal variable is kept constant; for example, a body fluid solute can remain relatively constant if the rate of excretion balances the rate of synthesis (**Fig. 1**). Many other homeostatic mechanisms, however, require regulation to maintain constancy. Negative feedback control is the most common regulatory system whereby a change in a variable is detected by a sensor and then counteracted by a response from an effector organ that is opposite to the perturbation (**Fig. 1**). Many physiological systems are controlled by several regulatory effectors, resulting in multiple control systems with greater overall precision of regulation. The nervous and endocrine systems are responsible for integrating physiological functions in an animal. They ensure that the physiological processes of different cells, tissues and organs occur in a controlled and coordinated manner, and result in whole-body homeostasis.

Nervous System

The nervous system integrates physiological functions and ensures that the physiological processes of different cells, tissues and organs occur in a controlled and coordinated manner, and result in whole-body regulation. It is responsible for coordinating rapid and precise responses to perturbations in the animal's internal and external environment by sensing changes in a physiological variable, integrating and interpreting the changes, and eliciting an effector response to counteract the change.

The nervous system consists of aggregations of two cell types, neurons that generate and conduct action potentials (a change in polarity of voltage across a cell membrane) and glial cells that are accessory cells which support and assist the function of neurons.

 doi:10.1016/B978-0-444-63768-0.00456-X

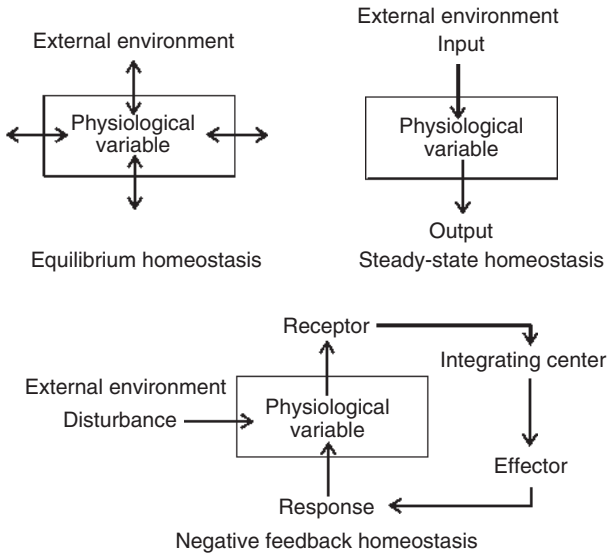


Fig. 1 Schematic of equilibrium, steady-state, and negative-feedback regulatory mechanisms.

Neurons can be classified as sensory (or afferent), inter (or internuncial) or motor (or efferent) and connect to one another, and to sensory or motor effector cells, via synapses (Fig. 2). In primitive animals (and simple reflexes in complex animals) there is a direct connection between sensory and effector cells by a single motor neuron, resulting in a simple three-cell sensory–motor circuit. However, in more advanced animals additional interneurons between the sensory and motor neurons allows for much greater complexity, permitting more complex integration and interpretation of sensory information, sophisticated motor control, and the development of complex behaviors.

The most primitive nervous systems are nerve nets; they occur in coelenterates and some flatworms. The development of cephalization (a head region) led to the concentration of neurons at the anterior end of the animal, forming the brain, and nerve cords consisting of concentrated groups of neurons transmitting information to other regions of the body. The nervous system is most highly developed in vertebrate animals. Here the brain and spinal nerve cord form the central nervous system while the peripheral nervous system consists of many paired nerves that run from the spinal cord to the peripheral regions of the body. These transmit sensory information to the central nervous system and return motor commands to the peripheral effectors. The somatic nervous system innervates efferent organs under conscious control (e.g., skeletal muscle), while the autonomic nervous system innervates involuntary visceral organs (e.g., gut and heart).

Endocrine Systems

Like the nervous system, the endocrine system regulates an animal's internal environment but it is a much slower control system. Chemical messengers (hormones) provide communication between sensory and effector cells. Hormone systems occur in all animals; they have become increasingly complex throughout evolutionary time compared to the basic neuron–endocrine systems of primitive animals. The endocrine system controls a wide range of physiological processes including reproduction, growth, development, metabolism, and osmo- and iono-regulation. It can respond to short- and long-term variations in internal and external environments and is important for the maintenance of homeostasis. Neuro-endocrine systems consist of neural sensory and interpretive pathways but instead of directly innervating an effector organ there is release of a chemical messenger into the blood at a hemal organ. This chemical messenger is then distributed to peripheral target organs where it has an effector action.

Hormones are secreted by endocrine glands (and neurohormones by nerve cells) in response to perturbing stimuli, and are then transported via the circulatory system or diffuse through tissues to target organs and cells. Thus a key characteristic of hormones is that they exert their action at a distance from the site of their secretion. Hormones do not initiate any unique cellular activities; rather they modify the rates of existing activities. Hormones may have inhibitory or excitatory effects on target cells, usually by inducing or repressing enzyme activity within cells, although they may act on the nucleus to influence the expression of genes or influence the permeability of cells to solutes.

Historically, hormones were considered to be chemicals released from endocrine glands (glands of internal secretion in contrast to exocrine glands such as salivary, sweat, and digestive glands that produce external secretions) but hormones may also be secreted by a variety of other tissues. Traditionally, hormones were considered to differ from neurotransmitters, which function only locally at the site of release (synapse) but this distinction is no longer so clear. Hormones function at very low concentration (e.g., 10^{-12}–10^{-9} M). Target organs have specificity for particular hormones due to the properties of receptors that are either on the surface of the cell membrane or inside the cell. Receptors reversibly bind the hormones with high specificity and affinity. Water-

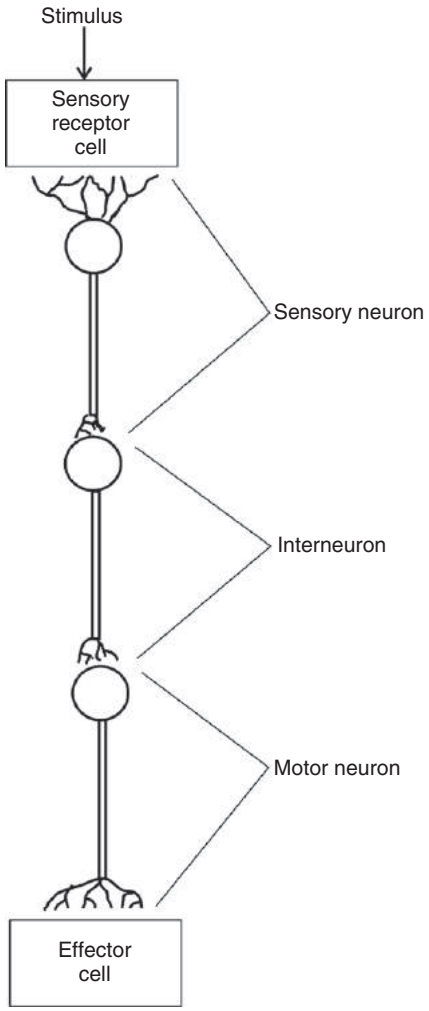


Fig. 2 Schematic sensory–motor neuron system consists of a sensory neuron input, integrative and interpretative interneuron, and an effector motor neuron.

soluble hormones are derivatives of amino acids (catecholamines, peptides, proteins) or fatty acids (eicosinoids). These interact with surface receptors that span the cell membrane. Often they trigger a secondary 'messenger' inside the cell. In contrast, lipid-soluble hormones such as steroids (adrenocortical and gonadal steroids in vertebrates, ecdysones, and juvenile hormones in invertebrates) and thyroid hormones usually pass through the cell membrane and interact with intracellular receptors. Some bind to membrane receptors which are then internalized. Many hormones that are transported in the circulatory system (in particular the lipid-soluble hormones) bind to a water-soluble carrier protein to aid transport.

Hormones are classified by the distance over which they travel to have their effect (Fig. 3). Autocrine hormones affect the cell that secreted them. They react with receptors on their own surface to produce a response and are usually involved in cell division. Paracrine hormones act over a very short distance, diffusing through extracellular fluid to affect local tissues. Endocrine hormones affect distant organs and tissues. They are secreted into the circulatory system and are transported by the hemolymph or blood. Pheromones are an additional form of chemical communication that occurs between rather than within individuals. They are highly volatile compounds released into the external environment and detected in small concentrations by receptors (usually on the nasal epithelium of vertebrates or antenna of insects) of another individual. Pheromones function to synchronize and induce reproductive activity, and to define territorial boundaries.

Water and Ion Balance

Maintaining water and ionic balance is a fundamental physiological process for animals because animal cells can only function effectively over a specific, relatively constant range of body fluid composition. For unicellular animals, the intercellular environment is juxtaposed with the external environment. For multicellular animals, the extracellular space is a buffer between the

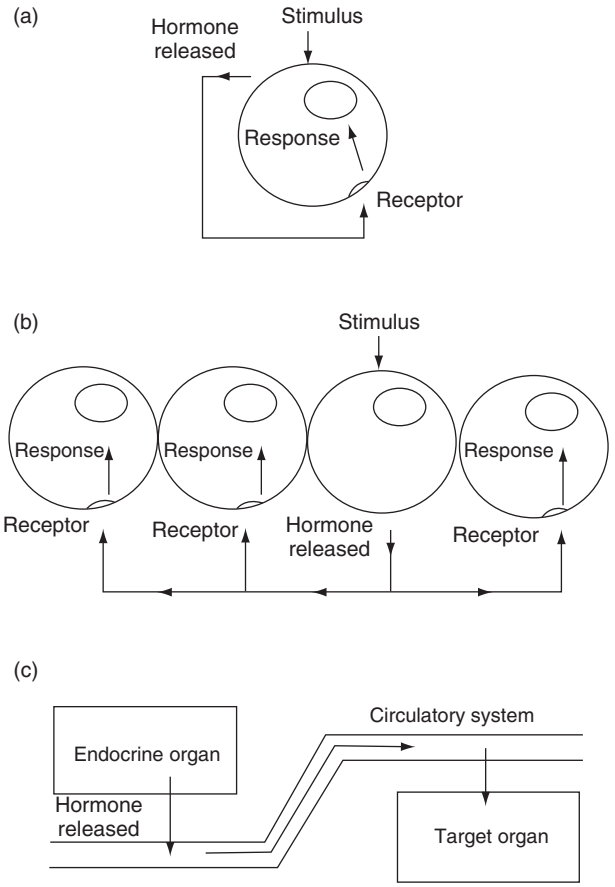


Fig. 3 Hormone systems. (a) Autocrine hormones affect the cell that secreted them. (b) Paracrine hormones affect local tissues. (c) Endocrine hormones are transported in the circulatory system to affect distant organs.

intracellar and external environments. In all animals, the intracellular environment has a different ionic composition from the external environment. For some animals there are osmotic differences, but their intracellular and extracellular environments must have the same osmotic concentration to maintain constant cellular volume (but invariably they have different ionic concentrations). The challenges associated with maintaining osmotic and ionic homeostatis differ with the external environment of the animal, and so there are various strategies for animals to maintain fluid and ion balance.

Aquatic Environments: Seawater

With respect to extracellular fluid, animals in marine environments either osmoconform (have the same osmotic concentration) to seawater (1000 mOsm; **Fig. 4**) or osmoregulate at a lower osmotic concentration (usually 300–400 mOsm). In addition, they either ionoconform with respect to their extracellular fluid, having the same ionic composition as seawater, or they ionoregulate and maintain different ionic concentrations. Animals that conform to seawater do not have to overcome the problem of continual osmotic loss of body water to and gain of ions from the environment, but high ion concentrations adversely influence cellular metabolic processes. Most marine invertebrates osmoconform and ionoconform to seawater, but a few osmo- and ionoregulate. Hagfish are the only vertebrates to both osmo- and ionoconform. Marine bony fish both osmo- and ionoregulate. Marine elasmobranchs osmoconform at 1000 mOsm but ionoregulate at about 600 mOsm; urea and trimethylamine oxide (TMAO; which counteracts the negative effects of the urea on proteins) make up most of the 400 mOsm osmotic gap between the ions and seawater (**Fig. 4**).

Aquatic Environments: Freshwater

Freshwater animals must both osmo- and ionoregulate as it is impossible to osmo- or ionoconform to such a dilute environment. Freshwater animals gain water by osmosis from their environment, and lose ions by diffusion. Excess water is eliminated as copious dilute urine and ions are obtained by active transport across the gills, skin, or gut.

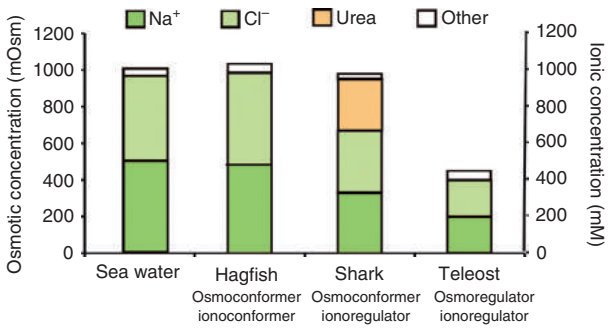


Fig. 4 Patterns of extracellular ion and osmotic regulation in vertebrate animals.

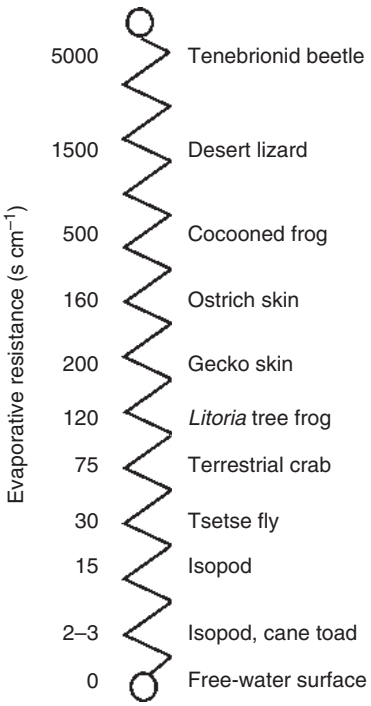


Fig. 5 Scale of resistance to evaporative water loss, from about 0 s cm^{-1} for a free-water surface to 5000 s cm^{-1} or more for animals that are very resistant to EWL.

Terrestrial Environments

Terrestrial environments are characterized by limited water availability, so dehydration is a major threat. Evaporative water loss (EWL) across the skin and respiratory tract is a major avenue of water loss by terrestrial animals. Water is also lost in feces and urine. Water is gained in a terrestrial environment via drinking, as preformed water in food, and as metabolic water production. Water may also be absorbed across the body surface. Ions are gained from food and by drinking, and are excreted in urine and feces and sometimes by salt glands.

Many invertebrates (e.g., mollusks, crustaceans) and amphibians are restricted to moist terrestrial habitats, at least when active, but many are more successful terrestrial animals because they have adaptations to minimize EWL. Arthropods have a chitinous exoskeleton, covered in a waxy cuticle that forms a barrier to evaporation. Birds, mammals, and especially reptiles have a cornified epithelium that increases resistance to EWL (Fig. 5). Insulating fur (mammals) or feathers (birds) is a further barrier. Nasal counter-current exchange of heat and water in the respiratory passages of reptiles, birds, and mammals reduces respiratory EWL. Arthropods, birds, and reptiles typically produce insoluble uric acid as their nitrogenous waste material, and the mixing of urine and feces in their hindgut (where water is reabsorbed) minimizes excretory water loss. Many desert reptiles and mammals survive without drinking, maintaining water balance with preformed and metabolic water alone. Most birds are able to travel long distances to obtain drinking water, although some can also survive without access to free water. Excess ions are lost by many reptiles and birds via cranial salt glands. Mammals do not have salt glands, and remove excess ions by producing urine that is

hyperosmotic to blood (up to 9000 mOsm). Some birds are also able to produce hyperosmotic urine to excrete excess ions, but not to the same extent as mammals.

Excretion

Excretory organs are essential for maintaining iono- and osmohomeostatis as they balance the gains and losses of water and solutes. They regulate the concentrations of ions and water in the body and play a vital role in excreting waste products including inorganic and organic solutes derived from the animal's diet, metabolic processes or foreign materials, preventing these wastes from accumulating to toxic levels. Thus excretory organs must selectively retain or remove a range of solutes from the body.

Simple animals rely on diffusion and membrane transport systems to remove wastes. However, the evolution of larger and more complex animals necessitated specialized excretory organs. Although in most animals the integument is relatively impermeable to water and solutes, specific epithelial regions can be specialized for the regulation of particular solutes or water. Tubular excretory organs are more generalized than these epithelial organs, and occur in most multicellular animals. They evolved primarily for water and solute excretion, but in a terrestrial environment they also play a crucial role in eliminating nitrogenous wastes and osmoconcentrating urine in some species.

Four major organ systems are responsible for excretion in animals. The respiratory system (lungs or gills) removes CO_2, and gills also play a vital role in ammonia, carbonate, and ion excretion, by both diffusion and active transport. The digestive system, in addition to eliminating undigested food, is also a site of ion and water absorption and excretion, and the vertebrate liver excretes bilirubin (derived from the breakdown of red blood cells) into the gut. The integument and various glands of animals may have a primary or secondary excretory function, for example, water and ion uptake by the skin of amphibians, salt glands of reptiles and birds, rectal glands of elasmobranchs, and sweat glands of mammals. Renal organs, including protonephridia, nephridia, Malpighian tubules, and coelomoducts (e.g., the vertebrate kidney) consist of tubules that filter body fluids and then selectively secrete or reabsorb water, organic molecules, and ions. The major functions of these excretory tubules are initial formation of excretory fluid, typically by filtration, then reabsorption of fluid and 'useful' solutes and secretion of specific 'waste' solutes. Only a few terrestrial animals are able to excrete urine that is more osmotically concentrated than their blood; the vertebrate kidney can excrete hypoosmotic or isoosmotic urine but only mammals and birds can excrete hyperosmotic urine due to the counter-current multiplication role of the renal medulla.

Gas Exchange

Most animals require oxygen to sustain their metabolic demands. Food is oxidized to produce adenosine triphosphate (ATP) and carbon dioxide is produced as a waste product, so animals must obtain oxygen from their environment and release carbon dioxide back into the environment. Gas exchange between the internal and external environment in all animals occurs through passive diffusion. For small, simple animals, diffusion across the body surface is sufficient to meet their metabolic demands. However, an evolutionary trend among animals for increased size and metabolic rate requires specialized surface regions for specific functions such as gas exchange (as well as locomotion, feeding, digestion, and sensory reception). So, a large body size and complexity necessitates specialized respiratory structures. Most respiratory structures require ventilation, the continual replacement of the external medium at the respiratory surface with fresh medium to maintain favorable concentration gradients for diffusion (**Fig. 6**). Animals are classified as air and/or water breathers. The physical characteristics of these two media constrain the ventilatory mechanisms necessary to maintain gas exchange across the respiratory surface, and therefore the nature of the surface itself.

Aquatic animals have gills, evaginated and highly folded external surfaces, for gas exchange. Water is dense and viscous (compared to air) so unidirectional flow over the gill surface is preferable. This also means that gills can have a counter-current flow of external medium (water) and internal fluid (blood/hemolymph) for very efficient O_2 extraction by counter-current exchange (**Fig. 6**). The O_2 concentration is also much lower for water (5–6 ml l^{-1}) than for air (210 ml l^{-1}) so a high efficiency of counter-current exchange is important. CO_2, however, is extremely soluble in water so its loss to the aquatic environment is not so problematic as O_2 uptake. Consequently, aquatic animals generally have low body fluid CO_2 levels.

Terrestrial animals have internalized respiratory structures, lungs or trachea, because avoiding desiccation is a major challenge. Moist externalized respiratory structures such as gills can have an excessively high EWL, but internalized structures have a lower EWL. Air is much less dense and viscous than water, and has a higher oxygen concentration, so lung ventilation by a tidal pool or cross-current system is not too inefficient or energetically restrictive. Lungs may be ventilated by positive pressure 'buccal pumping', as in amphibians, or by negative pressure inspiration, as in reptiles, mammals, and birds. Unlike the one-way tidal ventilatory pattern of most vertebrates, birds have a system of air sacs before and after the lung, which enables a one-way flow of air over the respiratory surface and allows a more efficient cross-current exchange system between the air and blood. The gas exchange system of arthropods consists of a series of air-filled tubes (tracheae) that infiltrate the body tissues and open to the external environment through spiracles at the body surface. Tracheal systems are generally not actively ventilated, relying on diffusion for gas exchange, a factor that limits the size of arthropods. The lungs of pulmonate snails are similarly diffusion driven.

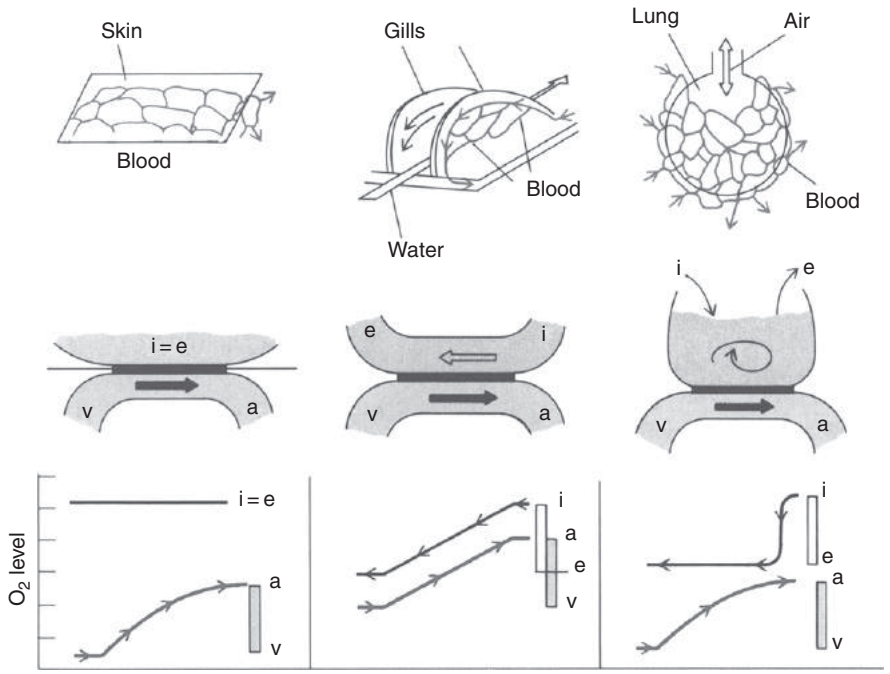


Fig. 6 Schematic diagrams of respiratory gas exchange across skin, gills, and lungs, showing patterns of fluid flow and O_2 exchange between the medium and blood, showing complete equilibration between the water (or air) and blood for skin exchange (left), typical counter-current arrangement of water and blood flow for gills (center), and a tidal pool of air for lungs (right). i, incurrent; e, excurrent; a, arterial; v, venous. Modified from Withers PC (1992) *Comparative Animal Physiology*. Philadelphia: Saunders College Publishing.

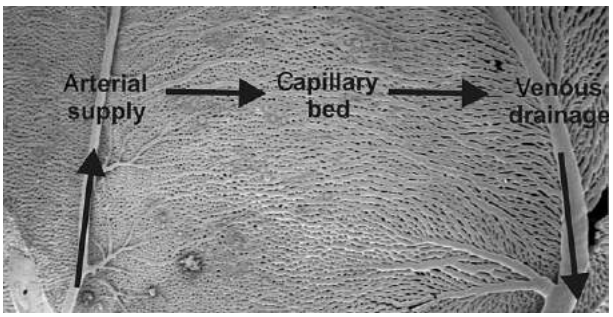


Fig. 7 Vascular arrangement for a closed circulatory system, showing arterial blood supply, capillary bed, and venous drainage, in the vascular supply to the eye of a marine toad (*Bufo marinus*). Photograph courtesy of T. Stewart, J. O'Shea, and S. Dunlop.

Circulation

Small, simple animals rely on diffusion to move solutes throughout their body. However, for larger and more complex animals the rate of diffusion is too slow so a circulatory system is needed for rapid transport of gases, nutrients, chemicals, and waste products. Circulatory systems may be open, where the circulating fluid is not always contained in vessels and is at times in direct contact with tissue cells (e.g., arthropods), or closed, where the circulating fluid is always contained inside vessels (e.g., vertebrates; **Fig. 7**). The circulating fluid is known as hemolymph (open systems) or blood (closed systems). It consists of plasma, a fluid containing water, ions and organic molecules, and various blood cells. These cells can be involved in transport of O_2 (erythrocytes), defense (leukocytes), or hemostasis (thromobcytes).

Blood and hemolymph flow is maintained by positive pressure created by the contraction of muscles in the body wall, or by the pumping of one or more hearts. Animal hearts are classified as neurogenic if they require innervation for contraction (e.g., arthropods), or myogenic if the contraction is spontaneous (e.g., mollusks and vertebrates). The complexity of animal hearts varies from the simple tubular hearts of insects that push blood by peristaltic contractions of the muscular wall, to the multichambered hearts of mollusks and vertebrates. Chambered hearts have a varying number of muscular-walled compartments, which contract in a coordinated manner to circulate blood. Generally circulatory systems transport oxygenated

blood from the respiratory surface(s) to the tissues and deoxygenated blood from the tissues to the respiratory surfaces. They can also be important in supplying nutrients to the tissues from the digestive system, transporting hormones from sites of synthesis to target cells, circulating cells of the immune system throughout the body, transporting heat, and generating a hydrostatic pressure.

Metabolism and Digestion

The use of chemical energy is a fundamental characteristic of living animals. It is necessary to maintain cellular order and is vital to almost all physiological processes. Catabolic metabolism breaks down macromolecules for production of usable energy by cellular processes such as active transport, muscle contraction, ciliary movement, and production of heat, electricity, or light. Most cellular reactions need 20–40 kJ of energy per mole of reactants, which is much less than the energy yield of the complete oxidation of a typical metabolic substrate. Therefore high-energy phosphate compounds (phosphagens) are used as intermediary chemical energy stores. ATP is the most common phosphagen. Free energy is released by the hydrolysis of its terminal phosphate to form adenosine diphosphate (ADP) and inorganic phosphate (P_i) that is, $ATP \leftrightarrow ADP + P_i + 30.5\ kJ\ mol^{-1}$. There is a cyclic formation of ATP from ADP (by cellular metabolism) and subsequent breakdown of ATP by energy-requiring processes.

Animals are heterotrophs, and as such are unable to synthesize their own organic compounds from inorganic molecules and so rely on other organisms for nutrients. Energy is obtained from nutrients such as carbohydrates, lipids, and sometimes proteins (amino acids are required for protein systhesis but also produce energy when oxidized). Essential vitamins, minerals, and fatty acids are also needed for proper cell functioning and must also be obtained via the diet. Single-celled animals and sponges ingest food particles by phagocytosis. These are chemically and enzymatically reduced within a food vacuole to a few constituent substances (e.g., monosaccharides, fatty acids, and amino acids) that are transported into the cytoplasm. Most multicellular animals have a digestive system specialized for extracellular digestion. Food particles enter the digestive system where a series of physical and chemical digestive processes break down food particles into constituent molecules that are absorbed and distributed to the cells. These molecules can then be used for energy metabolism, or for cell maintenance or growth.

Metabolism may be aerobic or anaerobic. Aerobic metabolism is the oxidation of carbohydrates, lipids, and proteins by oxygen to provide energy in the form of ATP. There are three major steps in the aerobic process: glycolysis, where glucose is converted to pyruvate with a net gain of 2 ATP (and 2 $NADH/H^+$), the citric acid (or Kreb's) cycle where pyruvate is converted to acetyl-CoA before undergoing a cycle of chemical reactions resulting in a further net gain of 2 ATP (and 6 $NADH/H^+$ and 2 $FADH_2$), and finally the mitochondrial electron transfer system. Ninety-five percent of the ATP is generated by electron transfer, where electrons from $NADH/H^+$ and $FADH_2$ are transferred to electron carrier proteins, passing through several protein complexes and generating 34 ATP. Oxygen is the final electron receptor in the chain, and water is formed as the end product.

Anaerobic metabolism is an alternative to aerobic metabolism, but it is very inefficient by comparison, forming as little as 2 ATP per glucose molecule. Consequently most large and complex animals rely on aerobic metabolism to meet their resting requirements, but they may use anaerobic metabolism for supplemental energy, for example, during intense activity or anoxia. Build-up of lactate as an anaerobic end product of glycolysis is a major inhibitory factor in the long-term use of anaerobic metabolism in tetrapod vertebrates. However some (e.g., carp) can convert pyruvate to ethanol as the end product, which can be easily excreted to the environment and therefore does not inhibit glycolysis.

Many factors affect the metabolic rate (MR) of animals, including temperature, developmental stage, diet, photoperiod, taxonomy, habit, environment, activity, and circadian rhythm. Body size is a major determinant of MR and is probably the best studied but least understood topic in animal physiology. Larger animals have a higher overall MR than small animals but have a lower MR per gram of body mass, so the relationship (eqn [1]) between mass (M) and MR

$$MR = aM^b \qquad [1]$$

does not scale isometrically (i.e., $b \neq 1$). Rather, $b < 1$ since small animals use proportionally more energy (i.e., per gram) than larger animals. This relationship is remarkably uniform for all animals, from single-celled protists to birds and mammals. Although there is some debate as to what the scaling coefficient actually is (and why), b appears to generally fall between 0.67 (the value expected if MR scales with surface area) and 1 (the value if MR is proportional to mass); b is typically about 0.75. The intercept of the scaling relationship (a) is lowest for unicellular organisms, higher for ectothermic animals, and highest for endothermic animals, but the slope is consistently about 0.75 (**Fig. 8**).

Temperature Relations

Body temperature has major significance for an animal's physiology. Temperature determines the state of matter and influences the rate of chemical reactions in biotic as well as abiotic systems. The body temperature of active animals generally ranges from $-2\ ^\circ C$ (freezing point of seawater) to $+50\ ^\circ C$ (where protein structure becomes unstable). This body temperature range can be even greater for animals in an inactive or dormant state; some can survive temperatures as low as $-200\ ^\circ C$ or as high as $120\ ^\circ C$!

All animals exchange heat with their environment. The vast majority of animals passively thermoconform to the temperature of their surroundings. However, some manipulate their thermal exchange to thermoregulate their body temperature within

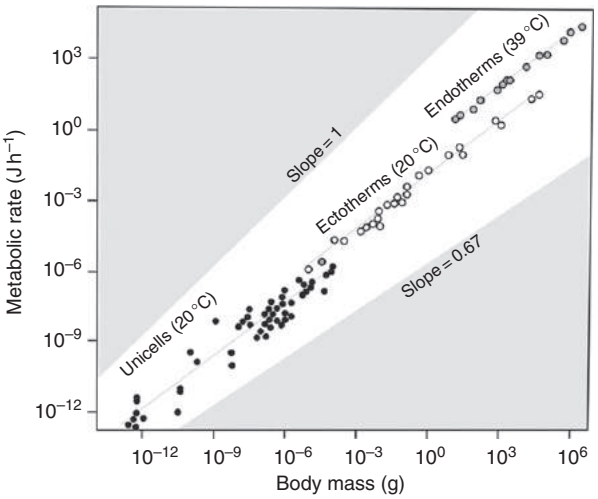


Fig. 8 Scaling of metabolic rate for unicellular organisms, and ectothermic animals (at 20 °C) and endothermic mammals and birds (at 39 °C). Modified from Hemmingsen AM (1950) The relation of standard (basal) energy metabolism to total fresh weight of living organisms. *Reports of the Steno Memorial Hospital and Nordic Insulin Laboratory* 4: 7–58.

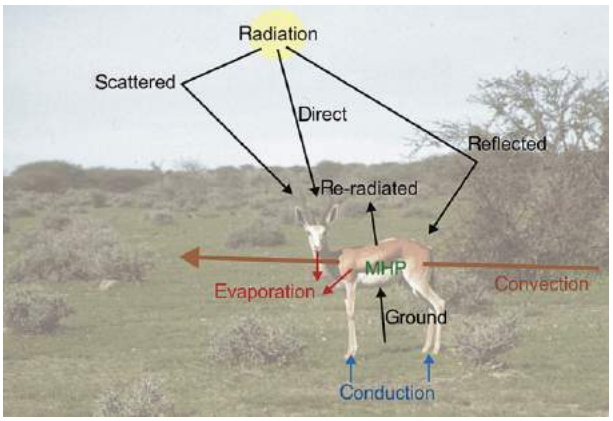


Fig. 9 Avenues of thermal exchange between an animal and its environment; conduction, convection, radiation, and evaporation. Photograph by P. Withers.

reasonably constant limits (typically 35–40 °C) and maintain an appreciable temperature gradient between themselves and the environment.

The thermal environment of an animal is complex. Heat exchange between an animal and its environment occurs by conduction, convection, radiation, and evaporation/condensation (**Fig. 9**). Conduction is direct heat transfer between two solid objects in physical contact. The rate of exchange depends on the area of physical contact, temperature difference, distance the heat must diffuse, and thermal conductive properties. Convection is transfer of heat by fluid movement (liquid or gas), and depends on the surface area, the temperature differential between the fluid and the surface of the solid, and the thickness and conductivity of the convective boundary layer. Forced convection occurs if the fluid movement is a result of external forces (e.g., wind), while free convection is induced by the temperature of the object itself. Radiation transfers heat between two objects that are not in physical contact by electromagnetic waves. The higher the surface temperature of an object, the greater is the radiative heat loss. Animals both emit and absorb radiation. Environmental sources of radiation for animals are complex and include direct solar radiation, diffuse scattered radiation, reflected radiation, and infrared radiation from surrounding objects and the ground. The structural and optical properties of an animal's surface are important determinants of its radiative heat load. Evaporative heat loss can be substantial because the latent heat of vaporization is about 2200 kJ g^{-1} (and condensation has an equivalent warming effect). Terrestrial animals lose heat via cutaneous and respiratory evaporation, and may have adaptations to reduce or augment this loss depending on environmental conditions.

Ectothermic animals have no physiological capacity to regulate their body temperature using internal metabolic heat production. They must either thermoconform to their environment, if their environmental temperature is relatively constant or if they can tolerate fluctuations in T_b, or use behavioral regulatory mechanisms to maintain a T_b that is somewhat independent of

environmental temperature. For example, many ectothermic reptiles remain largely independent of ambient temperature by using thermoregulatory behaviors such as basking, shuttling between warm and cool microhabitats, and postural adjustments to keep T_b about 36–38 °C. Endothermic animals, such as birds, mammals, and some insects, have physiological control of body temperature (typically 35–42 °C). They utilize heat produced as a by-product of metabolism to maintain their high and constant T_b independent of ambient conditions. Insulating fur and feathers reduces heat flux between endotherms and their environment. Endotherms may also employ behavioral thermoregulatory strategies to reduce the energetic costs of endogenous heat production, especially when the gradient between T_b and T_a is large.

See also: General Ecology: Ecophysiology

Further Reading

Campbell, N.A., 1993. Biology. Redwood City, CA: Benjamin/Cummings Publishing Company Inc.
Hemmingsen, A.M., 1950. The relation of standard (basal) energy metabolism to total fresh weight of living organisms. Reports of the Steno Memorial Hospital and Nordic Insulin Laboratory 4, 7–58.
Kay, I., 1998. Introduction to Animal Physiology. New York: BIOS Scientific Publishers Limited.
Knox, R.B., Ladiges, P., Evans, B., Saint, R., 2005. Biology: An Australian Focus. North Ryde, NSW: McGraw-Hill.
Louw, G.N., 1993. Physiological Animal Ecology. Harlow: Longman Scientific & Technical.
McNab, B.K., 2002. The Physiological Ecology of Vertebrates: A View from Energetics. New York: Cornell University Press.
Schmidt-Nielsen, K., 1997. Animal Physiology: Adaptation and Environment. Cambridge: Cambridge University Press.
Sherwood, L., Klandorf, H., Yancy, P.H., 2005. Animal Physiology: From Genes to Organisms. Cole California: Thompson Brooks.
Withers, P.C., 1992. Comparative Animal Physiology. Philadelphia: Saunders College Publishing.

Applied Ecology

A Georges, LJ Hone, and RH Norris, Institute for Applied Ecology, Canberra, ACT, Australia

Applied Ecology

The science of ecology involves the study of interactions between organisms and their environment, both biotic and abiotic, with particular focus on those interactions that determine their distribution and abundance. Applied ecology is the science of the application of ecology to contemporary problems in managing our biological resources. It includes scientific study of the effects of humans on the interactions between organisms and their environment, but excludes human ecology.

Applied ecology has two broad themes. The utilitarian theme concerns the interests of humans in their food, shelter, welfare, and health, that is, the material services the natural environment provides. Such ecosystem services, once compromised, can be very expensive to replace despite our technological advances. How do we bring ecology to bear in maintaining and improving these ecosystem services where they currently exist, in restoring or replacing them if they have been lost, or in mitigating the impact if those services are under threat? A second theme concerns nonconsumptive values of the biota, for recreation, tourism, psychological well-being, or simply because humans have an ethical responsibility as custodians of the natural environment and the species it contains. How do we bring ecology to bear in conserving these important nonconsumptive values?

These two broad themes overlap, since the nonconsumptive values of the environment are connected through biodiversity to the services healthy environments deliver. Naturally biodiverse systems are typically more resilient to human-induced perturbation than are systems that are highly modified, structurally simplified or degraded, and so are better able to sustainably provide the ecosystem services we expect.

Topics included under the broad discipline area of `applied ecology' are those where ecological knowledge and understanding are brought to bear on policy setting, decision making, and practice. The directions of the discipline are very much driven by the problems given priority by contemporary society, its governments and industry.

Some Iconic Examples

The scope of `applied ecology' is very broad indeed, possibly best illustrated by way of example. It is now widely accepted that the global climate is changing, that part of the cause is associated with industrial development, and that the impacts on human communities and the biota are potentially profound. Can we predict how natural ecosystems and the species they contain will respond, what have we done to constrain those responses (e.g., widespread fragmentation of habitat restricts range shifts), and what can we do to ameliorate the impacts of global climate change on the biota? It is in answering these questions that `applied ecology' complements climate change studies by meteorologists, geographers, and geologists in other disciplines more focused on studying the direct impacts of climate change on human society.

Land clearing for forestry, pastoralism, and agriculture and aquatic habitat destruction through land reclamation and water resource development are arguably the most serious threats to biodiversity today. The long-term consequences of such activity is often not realized at the time it is undertaken, and there is no good appreciation of how far the system can be pushed in meeting production goals before both ecological and economic sustainability are compromised. When land use and water resource development have overshot sustainable levels for production and for other land-use values such as biodiversity, what can be done to restore those values (restoration) or bring about change leading to an acceptable and sustainable condition (rehabilitation)? Both restoration and rehabilitation are important components of `applied ecology'.

Protected areas such as reserves and national parks make an important contribution to biodiversity conservation (**Fig. 1**), but are they adequate to sustain biodiversity in the long term? The overall goals are to conserve species, their genetic variability and potential to respond to environmental change, and the natural ecosystem processes that provide the ecological context in which they have evolved and continue to evolve. Protected area management, the inventory of values and selection of reserves, their design, and the management of threats to their values such as feral animals and weeds, fire management, impacts of human visitation are all topics addressed in part by `applied ecology'.

Globalized trade and associated movement of people and products leads inevitably to unwanted introduction of exotic species, some of which become established in the wild well outside their natural range. This is of major concern because feral populations can be reservoirs for disease that impacts on agricultural production. They can wreak havoc on native species through predation (stoats in New Zealand and foxes in Australia), competition (rabbits in Australia), or interference (zebra mussels in North America). Can we predict which species are most likely to establish, which are likely to cause the greatest impact, model the spread of exotic species when they arrive, and control their spread, distribution, and abundance in order to manage their impacts once they are established?

 doi:10.1016/B978-0-444-63768-0.00848-9

Fig. 1 Less than 1% of native temperate lowland grasslands of Australia remain intact, and that which remains is fragmented and under continual threat from agriculture, pastoralism and, in the Australian Capital Territory, urban expansion. Inset: The striped legless lizard (*Delma impas*), one of the many endangered species that rely on native grasslands. Photos: Sarah Sharp and Will Osborne.

These are examples of the broader societal context in which `applied ecology' does its work. The discipline is generally seen to add value to restoration ecology, habitat management and rehabilitation, management of invasive species (both native and exotic), conservation biology, wildlife utilization, protected area management, and agroecosystem management. The discipline also makes important contributions to environmental forensics, landscape architecture, ecotourism, and fisheries.

Foundation in the Fundamentals

The diversity of concepts drawn from ecology and applied in management of our natural resources is vast and stems from the multidisciplinary nature of natural resource management generally. In dealing with the impacts of climate change on the biota, for example, we need to know of the habitat requirements of species or `niche breadth', the extent of suitable habitat and connectivity that provides scope for `invasion' of new areas as the climate shifts. Coupled with this is the need to know of the limitations to their `physiological tolerances', their `dispersal capabilities' and the `demographic attributes' that will govern the speed of their response through range shifts. Many reptiles have temperature-dependent sex determination, and would appear appallingly vulnerable to climate change. What scope do they have to respond to climatic change, through `natural selection', which in part will depend on `genetic diversity' in the traits determining their `evolutionary responses' to changing climate. Will an evolutionary response be rapid enough? What scope do species have to respond through `phenotypic plasticity' rather than a direct evolutionary response. **Table 1** provides a link between fundamental ecological concepts, principles, and ideas, and the broad areas of application in `applied ecology'.

What Do Applied Ecologists Do?

Applied ecologists engage in their profession at a broader level than commonly recognized. On the spectrum of esoteric research (of no identifiable immediate relevance), through strategic research (of broad relevance) to tactical research (of immediate relevance), applied ecologists vary in their level of engagement. Some are practitioners at the coalface of application undertaking research in the immediate context of management problems, and addressing the immediate concerns of management. Their work is typically funded directly by resource management agencies or industry.

Others address research questions of more fundamental strategic value, in areas where improved knowledge, understanding, and techniques are likely to be of service in addressing contemporary problems as well as problems of the future, many of which are currently unforeseen. Their work is typically funded by research and development (R&D) organizations or by government agencies such as the US National Science Foundation, the Australian Research Council, the UK Natural Environment Research Council, or the NZ Marsden Scheme.

Application often draws support from unexpected quarters, and an important element of the development of the discipline of `applied ecology' is the need to provide tertiary education and research funding in a broad strategic context. There must not be too great a focus on immediate needs in funding applied ecological research, lest we risk passing by many opportunities to build the knowledge base from which solutions for the future can be drawn. At an individual level, it can be argued that to be a good applied ecologist, one must be a good ecologist with a broad research agenda, but also with a keen eye out for application and a willingness to engage in those applications when opportunities arise.

Table 1 Topics in `applied ecology' and concepts, ideas, and topics in ecology that are used in applied ecology

Topics in applied ecology	*Relavant ecological concepts, ideas, and topics*
Restoration ecology	Niche, succession, community dynamics, resilience
Habitat management and rehabilitation	Habitat selection, niche, community dynamics
Management of invasive species	Population dynamics, predator–prey relationships, competition, disease–host interactions, natural selection.
Conservation biology	Population dynamics, population genetics, population viability, biodiversity
Wildlife utilization	Population dynamics, sustained yield
Protected area management	Island biogeography, population viability, biodiversity, ecotones
Agroecosystem management	Competition, biodiversity, natural selection
Forensic sciences	Genetics, taxonomy
Landscape architecture	Connectivity, fragmentation, movements, metapopulations
Ecotourism	Population dynamics, thresholds, resilience
Fisheries	Population dynamics, sustained yield, food webs
Forestry	Population dynamics, sustained yield, demography
Urban development	Habitat, corridors
Ecosystem services	Nutrient cycling, biodiversity
Climate change	Niche, population dynamics
Pollution	Niche, assimilation, bioaccumulation, ecotoxicology
Energy generation and carbon management	Nutrient cycling, bioaccumulation
Water management	Niche, biodiversity assessment

Applied ecologists use one or more of the following approaches in conducting their science – observation, experimentation, and modeling. Any one study or topic may be studied and resolved using combinations of the approaches. For example, conservation of large kangaroos in Australia involves observational studies of kangaroo ecology – knowledge of reproductive cycles, diet, and behavior are all important to managing kangaroo populations. Experiments may be undertaken to explore causal relationships, perhaps involving exclosures and population manipulation to determine responses of vegetation, or to fine-tune survey and monitoring approaches. Modeling may be applied outside the scope of feasible experimentation to investigate the combined effects of environmental changes and human intervention on kangaroo populations as a tool to guide decision making. Some topics, such as large-scale climate change, can initially be studied by observation to quantify the changes that are or not occurring. Field experiments may be impossible, especially at large spatial scales, but small plot or laboratory experiments can provide useful information. Modeling provides a framework for integrating these observations and results of the limited experimentation that is possible to estimate likely changes in environmental conditions and responses by organisms to such changes. Future observations can be used to evaluate the accuracy of predictions of the modeling.

The mix of approaches that are used by applied ecologists is determined by their experience with each, the advantages and disadvantages of each including the costs, practicality, and the quality of data and hence the strength of conclusions obtained by each approach. For example, on the latter point, observations allow clear conclusions to be made about patterns in ecology. However experiments allow clearer conclusions about cause and effect in ecology; that is, about what causes changes in distribution and abundance of organisms compared with what changes have occurred. Modeling allows a great range of possible management actions or scenarios to be examined and a greater range than that can be examined by experiments. However the modeling results are hypothetical and require evaluation of their practical relevance.

Success through Communication and Engagement

Applied ecology measures its success in part by adoption of what it has to offer management. Successful adoption demands communication of results in a form that can be readily comprehended by resource managers and effectively brought to the table in policy setting and decision making (**Box 1**). There are a number of challenges to bring about effective communication.

The first challenge arises from the different cultural perspectives of scientist and manager. The core objective of a natural resource manager is to bring all available knowledge and understanding, scientific and otherwise, to bear on setting policy and making and implementing a decision. At the end of the process, the outcome is evaluated, and the decision confirmed as appropriate or not. For an ecologist, as a scientist, learning that the knowledge and understanding brought to the decision-making process was confirmed as appropriate is satisfying, but from the perspective of his/her discipline, it is potentially pedestrian. Science advances through failure, focuses on the causes of that failure, reevaluation of concepts and principles, collection of new data, and re-application. Managers want the problem solved – they get excited when it all goes well; scientists want to learn something new – they get excited when something unexpected happens. Lack of appreciation, and lack of reciprocal respect, for these differing perspectives can lead to breakdown in trust and with it, loss of communication.

Box 1 Melding ecological principles with urban planning

At the time of the first European settlement in Australia, lowland areas of southeastern Australia had one of the largest areas of native temperate grassland in the world. These grasslands are now among the most endangered natural communities in Australia (Fig. 1). The Australian Capital Territory (ACT) contains about 5% of the high-quality primary native grassland that occurred in the ACT prior to European settlement, home to a number of threatened animal species, including the legless lizard *Delma impar*, the mouthless moth *Synemon plana*, and the matchstick grasshopper *Keyacris scurra*. The expanding Australian capital city, Canberra, is placing continual and increasing pressure on these grasslands and presents city planners with the very great challenge of melding grassland conservation with the relentless expansion of suburban and rural urban development.

Planning for suburban development is a complex process, and planning decisions are made throughout the construction phase. Ecological theory is not in a form that can be used by urban planners, who continually need to assess the costs versus the benefits of planning decisions. Too often the cost–benefit analysis is driven entirely by financial considerations.

Applied ecologists were given the challenge of devising a set of principles that would govern the type and quality of ecological information brought to the planning process and that would enable planners to assess alternatives in the context of both financial and ecological considerations. The principles they devised are as follows.

1. Both regional and local objectives are required for conservation planning on the local scale.
2. Both species and functional communities need to be considered.
3. Knowledge of key life-history properties of species and dynamic processes within the ecological communities is essential for sound conservation planning.
4. Spatial scale is important when assessing the value of published knowledge of species and communities.
5. Common as well as rare species have a bearing on conservation planning.
6. The quality of available data and therefore its value to conservation planning, varies depending on its taxonomic and spatial resolution, seasonal biases, and temporal representation.
7. Areas considered for conservation should be those of the highest value for meeting local, regional, and national objectives.
8. Conservation value includes concepts of size (viability), diversity, representativeness, distinctiveness (rarity), and naturalness.
9. *Diversity*. Conservation areas that possess greater heterogeneity of environmental attributes (floristics, vegetation structure, abiotic components), within the bounds of those conditions known to support lowland grassland communities, are better than those that are largely homogeneous.
10. *Size*. Larger contiguous conservation zones are superior to smaller zones, or zones of equivalent size that are fragmented, all other considerations being equal.
11. *Shape*. Conservation zones that have a large area to perimeter ratio are better than those that are irregular in shape, elongated, or whose boundaries project into suboptimal habitat.
12. Replication of conservation areas in fragmented habitats is necessary as a hedge against catastrophic or stochastic local extinction.
13. Regional conservation planning based on remnants must consider the constraints and opportunities provided by the present and future land-use patterns.
14. Rehabilitation of fragmented habitats should be considered as a means of increasing overall size, buffering, and interconnection.
15. Integration of smaller systems within broader conservation systems increases their conservation value.
16. Consider alternative reserve structures in the light of constraints and opportunities provided by planned development.
17. Conservation zones are not isolated from external influences and careful consideration needs to be given to compatible adjacent land uses, and moderation of their impacts.
18. Include research-based management, monitoring and community participation.

Application of these principles led to the establishment of a series of outstanding urban native grassland reserves in the ACT, reserves that were established as an integral part of the planning and development of the new Gungahlin suburbs. For the applied ecologist, the exercise was communication of ecological principles in a form that could be readily adopted in the planning process, and engagement with planners in bringing about solutions to the challenges of conservation in an urban setting.

Adaptive management, that is, adopting an experimental approach to management intervention, provides a good framework in which scientists and natural managers can work together to achieve solutions to both management problems and advances in knowledge. Under this framework, management intervention is conducted in a rigorous experimental framework where the intervention is implemented as a scientific experiment. Due attention is paid to the fundamental tenets of experimental design and sampling – the use of temporal and spatial controls against which the effects of interventions can be measured, proper replication of experimental treatments, proper attention to sample unit selection and sample sizes. The management intervention occurs in the

context of a solid scientific foundation for the monitoring and evaluation that follows. The benefits of an experimental approach to management intervention are that the ecologists and natural resource managers are working together at all stages of the design and implementation of the intervention, the evaluation of the efficacy of the intervention is on a solid scientific footing and so the intervention can be recast in the light of the outcomes with confidence and, perhaps most importantly, knowledge is advanced both when the intervention is successful and when it is a failure. Adaptive management of natural resources is `applied ecology' at its best.

A second challenge faced in applying ecological knowledge to natural resource management is ensuring that all important information is available to management at the time of setting policy, making decisions, and putting policy into practice. Traditionally, communication of the results of ecological studies occurs in the presentation at learned conferences and by publication as scientific papers in leading journals such as the *Journal of Applied Ecology*, *Ecological Applications*, or *Biological Conservation*. While there have been efforts to better integrate scientists and natural resource managers into professional societies, the primary audience for these channels of communication remain ecologists and other scientists. The audience for refereed publications in journals is primarily comprised of applied ecologists and other scientists.

Many organizations responsible for natural resource management have limited in-house ecological capacity for accessing, evaluating, and adopting the results of research presented through formal scientific channels. This in turn can limit the information available to them at the time of making important decisions. Often, decisions are made on a very small base of available information and a limited network of trusted advisers. Many of the larger research organizations address the issue of broadening the base of information available to managers through the appointment of knowledge brokers – individuals employed and often placed within the natural resource agency whose sole responsibility is to broker exchange of management needs in one direction and ecological information in the other direction between managers and scientists, and to assist in providing that information in a useful form. Knowledge brokers and professional science communicators are also engaged to communicate the outcomes of science to the broader public through the media (television, radio, newspapers), community meetings, and websites of ecological associations. Knowledge brokers must have both scientific understanding and communication skills.

A third challenge is to bring ecologists, industry, and management together to build relationships, identify synergies, and achieve broad and lasting ownership over solutions to environmental problems. This is being addressed by governments in many developed countries by providing monetary incentives for science and industry to work together, placing conditions on industry and community participation in government-funded research, and establishing substantial cooperative research entities that bring industry, community groups, and researchers together in well-funded joint ventures. This has changed the face of `applied ecology', providing many more opportunities for ecologists to engage in research and application of immediate relevance to the economy and society.

Summary

In summary, `applied ecology' draws its strength from the commitment of ecologists to engage in the application of their science to natural resource management. The discipline relies upon a balance that maintains a strong commitment to broad enquiry going beyond the need for solutions to immediate problems at hand. Information important for solving the environmental challenges of the future will emerge from a broad base of fundamental ecological knowledge and understanding. The discipline also relies on effective communication between the diverse sectors responsible for bringing about effective action on the environment. Applied ecologists do not make policy or make decisions about how to manage the environment. Industry, government, and resource managers do that, and it is up to them to take or reject advice. Ecological knowledge and understanding must be brought to the table in a form that can be readily understood and adopted by industry, government, and management. The major challenge is for ecologists to recognize the contribution they can make to the triple bottom line of industry, and to decision making in government and nongovernment resource management agencies. Ecologists need to become and remain engaged in informing the process of policy formulation, decision making, and implementation by bringing the best-available science to the table. In a world where environmental challenges are increasing dramatically, responsible ecologists need to have a keen eye out for applications of their work and a commitment to engage with natural resource managers when opportunities to add value arise. This is applied ecology's *raison d'etre*.

Further Reading

Davis, W.S., 1995. Biological assessment and criteria: Building on the past. In: Davis, W., Simon, T. (Eds.), Biological Assessment and Criteria: Tools for Water Resource Planning and Decision Making. Ann Arbor, MI: Lewis Publishers, pp. 15–29.

Krebs, C.J., 2001. Ecology. The Experimental Analysis of Distribution and Abundance, 5th edn. San Francisco: Addison-Wesley.

Naiman, R.J., Magnuson, J.J., McKnight, D.M., Standford, J.A., 1995. The Freshwater Imperative. Washington: Island Press, 176pp.

Sutherland, W.J., Armstrong-Brown, S., Armsworth, P.R., *et al*., 2006. The identification of 100 ecological questions of high policy relevance in the UK. Journal of Applied Ecology 43, 617–627.

US EPA, 1998. Guidelines for Ecological Risk Assessment. Washington, DC: US: Environmental Protection Agency, EPA/630/R-95/002F.

Biodiversity☆

Rodolfo Dirzo and Eduardo Mendoza, Stanford University, Stanford, CA, United States

Glossary

Biodiversity offsets Measurable conservation outcomes resulting from actions designed to compensate for adverse impacts arising from project development.
Endemism The ecological state of being unique to a particular geographic location.
Genetic diversity The variation in the amount of genetic information within and among individuals of a population, a species, an assemblage, or a community.
Intrinsic value The inherent worth of something, independent of its value to anyone or anything else.
Species diversity Biodiversity at the species level, often combining aspects of species richness, their relative abundance, and their dissimilarity.
Species richness The number of species within a given sample, community, or area.

Introduction

This standard definition of biodiversity implies a logical functional link between the three levels it comprises and could be articulated, qualitatively, as follows: the organisms that make up a population of a given species behave and respond to their environment as they do as a result of the features determined by their genetic constitution; the contingent of such individuals in a population or group of populations constitutes a species, and the collection of species in a given region are constituents of communities that, together with their physical environment, form the ecosystems, landscapes, and, ultimately, the biomes of the Earth.

Beyond the intraspecific, interspecific, and ecosystem diversity levels, the biodiversity of a given region can be manifested by these additional facets: (1) morphofunctional diversity, represented by the variety of life forms among plants or, its rough zoological equivalent, the diversity of functional groups among animals; (2) the concentration of endemic organisms, that is the cooccurrence of taxa (species, genera, families) with a restricted geographic distribution; and (3) the agrobiodiversity, best reflected by the variety of domesticated and cultivated plants and animals and their wild relatives. Next, we briefly review each of these biodiversity facets.

Morphofunctional diversity is seldom recognized as an important facet of biodiversity, although its relevance is ecologically and evolutionarily crucial, for it represents the amazing variety of responses organisms have evolved to deal with the environmental pressures of their habitats. The examples of Figs. 1 and 2 highlight this facet of biodiversity. The panel of plants (Fig. 1) includes a diversity of life forms depicting a selected variety of ways in which plants from desert ecosystems deal with the crucial limiting factor, water. Thus the panel includes (clockwise arrangement): cacti of several types (e.g., a sahuaro and an opuntia), shrubs with tiny or absent leaves (e.g., ocotillo and creosote bush), succulent rosette-like shrubs and treelets (e.g., agaves), arborescent palm-like plants (e.g., yuccas), and ephemeral annual plants (e.g., dahlia). Such life-form diversity can be astonishing. For example, a description of the floristic diversity of Mexican desert ecosystems included 32 readily distinguishable life forms. The animal example of Fig. 2 shows the diversity of morphofunctional groups among one single group of mammals, bats, depicting the variety of evolutionary routes these animals have developed to solve the challenge of finding and using their food resources, including, clockwise in the diagram, fruit-feeding (e.g., *Artibeus*), nectar-feeding (e.g., *Choeroniscus*), blood-feeding (e.g., *Desmodus*), and insect-feeding (e.g., *Lonchorrhina*) bats. The diversity of adaptations to such variety of feeding habits is dramatically displayed in Fig. 2 by the variations in the animals' heads.

☆*Change History:* March 2018. H. R. Pethybridge included glossary, keywords, extended text and updated references.

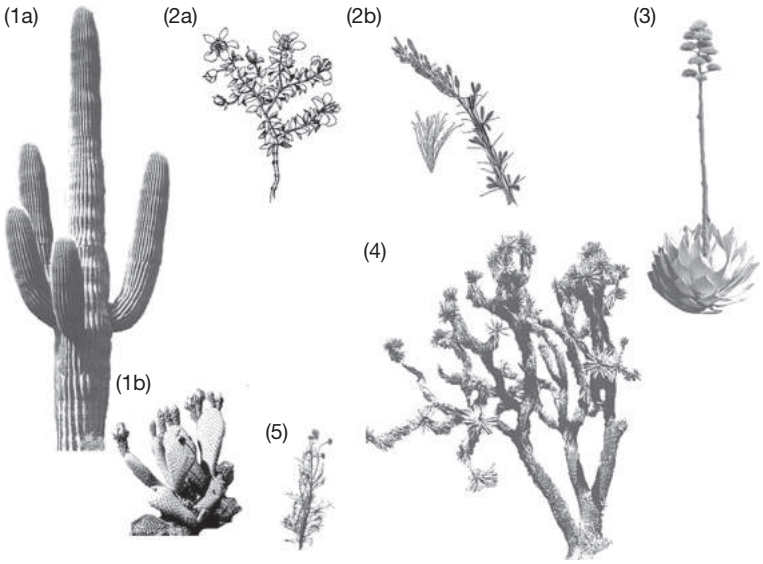


Fig. 1 Biodiversity expressed as the variety of plant life forms that can be found in a desert. Life forms include: (1A) columnar and (1B) prickly pear cacti; (2A) creosote bush and (2B) ocotillo shrub; (3) succulent rosette-like agave (with reproductive structure); (4) palm-like yucca (Joshua tree); and (5) annual herb (dahlia).

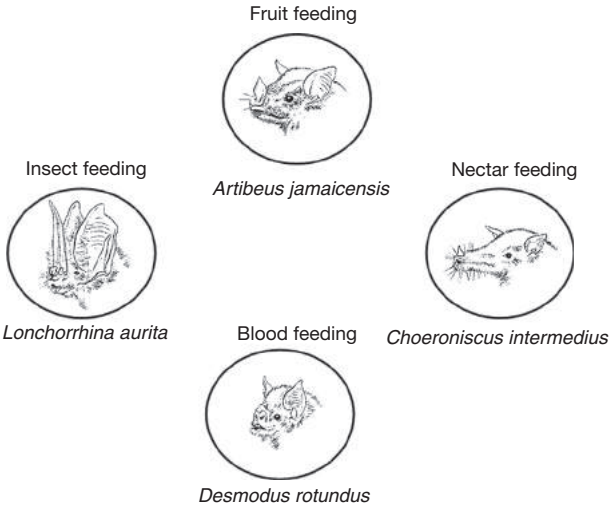


Fig. 2 Diversity of bat feeding habits. Head characteristics reflect functional attributes required to exploit specific feeding sources. For example, the very large ears and nose-leaf of Lonchorrhina aurita are associated to sensorial capacity needed to prey on flying insects. In *Choeroniscus intermedius*, the long snout and a long tongue allow reaching nectar at the base of flower tubes.

The concentration of taxa in a given geographical region, evolved and found nowhere else on the planet, that is, the concentration of endemisms, is an important facet of biodiversity, not only from the biogeographical (the geographic distribution of taxa) and evolutionary (the fact that they frequently represent evolutionary novelties) points of view, but also because regions of high levels of endemism become a priority in biodiversity conservation. This is because the alteration or destruction of regions with high concentrations of endemism (the so-called hot spots), would bring about the irreplaceable loss of unique products of evolution. Endemism can be expressed at different levels: species, genera, families, orders, or even phyla can be endemic to a given region. In Fig. 3, we show the restricted distribution of a pair of pine species (*Pinus* spp.) endemic to Mexico. These species, restricted to temperate forests of that region, evolved in situ and have not been found in other temperate forests of the globe, apart from those shown in the map. Furthermore, both species contrast in its range of distribution within the country. Populations of *Pinus herrerae* can be found along the mountains of western Sierra Madre and the Trans-Mexican Volcanic Belt, while *P. nelsonii* is restricted to a small arid area in north-east Mexico.

Agrobiodiversity is represented by the cultivated and domesticated plants and animals present in a given region; this facet includes as well the wild relatives from which traditional indigenous cultures have selected the forms adapted to the environments

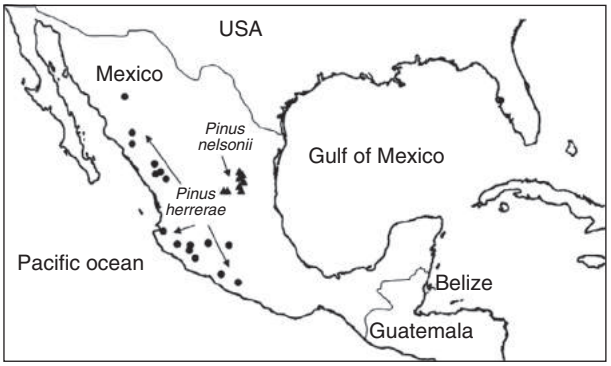


Fig. 3 Two contrasting levels of endemism illustrated by species in the genus *Pinus* within the territory of Mexico. While *P. herrerae* is distributed on the mountains of the western Sierra Madre and the Trans-Mexican Volcanic Belt, but cannot be found in any other country, *P. nelsonii* is restricted to an area of a few square kilometers in the arid lands of northwestern Mexico.

in which they are cultivated or propagated to satisfy diverse human needs. Although this facet of biodiversity represents a minor fraction of the global contingent of species of plants and animals, its importance is crucial, as it represents the vast majority of the food we consume. Agrobiodiversity emerges from the combination of two factors: availability of biodiversity in terms of the wild relatives, and the application of traditional management and knowledge leading to domestication. (Thus this facet of biodiversity includes also cultural diversity and traditional knowledge.) In the case of plants, this has determined the existence of a few regions of the world, popularly known as the Vavilovian domestication centers, or agrobiodiversity centers (Fig. 4). This handful of centers of origin and domestication of plant crops has contributed significantly to human development and wellbeing and, in addition, such centers are areas of great concern from the biodiversity conservation perspective. Among the roughly 2500 crop species, just 103 supply over 90% of the calories humans consume, directly or indirectly, and just three grasses (cereals), rice, wheat, and maize, supply about 60% of the total. In addition, some 15 plants are cultivated as sources of fiber, and thousands more are used as ornamentals or sources of medicines. Agrobiodiversity also illustrates intraspecific or genetic diversity. We do know that most crops are genetically diverse, including many land races that exist in the major cereals and crops such as bananas, cassava, potatoes, and tomatoes. Likewise, cabbage, cauliflowers, broccoli, kohlrabi, brussels sprouts, calabrese, and kale are all selected variants of *Brassica oleracea*. In the case of animals, from the approximately 50,000 described vertebrate species, 30–40 species of birds and mammals have been domesticated; a few of them have a global distribution (except for the Antarctic) and are extremely abundant: cattle (1300 million), sheep (1200 million), pigs (850 million), and chickens (10 billion). Again, the degree of intraspecific variation in animals is impressive, including about 800 distinct breeds each of cattle and sheep.

In summary, biodiversity is a very inclusive concept, involving several facets and levels of organization. Nevertheless, the level of species is the one that has received the greatest attention. Yet, as we describe in the next sections, even at this level our global knowledge is still limited.

Metrics of Biodiversity

The most commonly used method to describe the biodiversity occurring in a particular locality is based on the enumeration of the species present therein. This metric is known as species richness (S). The enumeration of the total number of species present in most habitats is a formidable task. For example, it has been suggested by a team of British ecologists that recording the total number of species in a single hectare of tropical forest would require employing up to 20% of the 7000 systematics active worldwide. Even when aimed at more specific groups (e.g., plants or insects), quantification of species richness might still be problematic if the group under inspection is particularly speciose, its taxonomy is poorly known, or is hard to survey in the field. Therefore, most of the characterizations of S are based on sampling protocols and the inferences that can be drawn from them. Here we provide an example of sampling to assess S.

An increasingly popular analytical procedure is that of species accumulation curves. These curves depict how species richness accumulate with increased sampling effort, as shown in Fig. 5, based on plant species recorded in ten 50 m 2 m transects established in the understory of a rain forest site in southeast Mexico. The survey included only woody plants of <1.5 m height, to assess the forest's regeneration potential, and was performed in homogeneous habitat (i.e., not including microtopographic variations such as ravines, river beds, or flooded areas; this would likely bring about different sets of species). Species richness associated with any given number of transects (sampling effort: two, three, up to ten) is represented as a mean value that was calculated using, in a random sequence, the entire set of the ten transects. Thus a sampling effort of two transects shows the mean of all possible combinations of two transects, and so on, up to a predetermined limit of 100 combinations. In this example, S was calculated applying an estimator that takes into account the relative representation of rare species, that is those represented by a single individual ("singletons"), and those represented by at least two individuals ("double-tons"). The ratio of singletons and doubletons will determine to what extent the observed number of species needs to be increased to estimate S. When such a ratio is zero, the

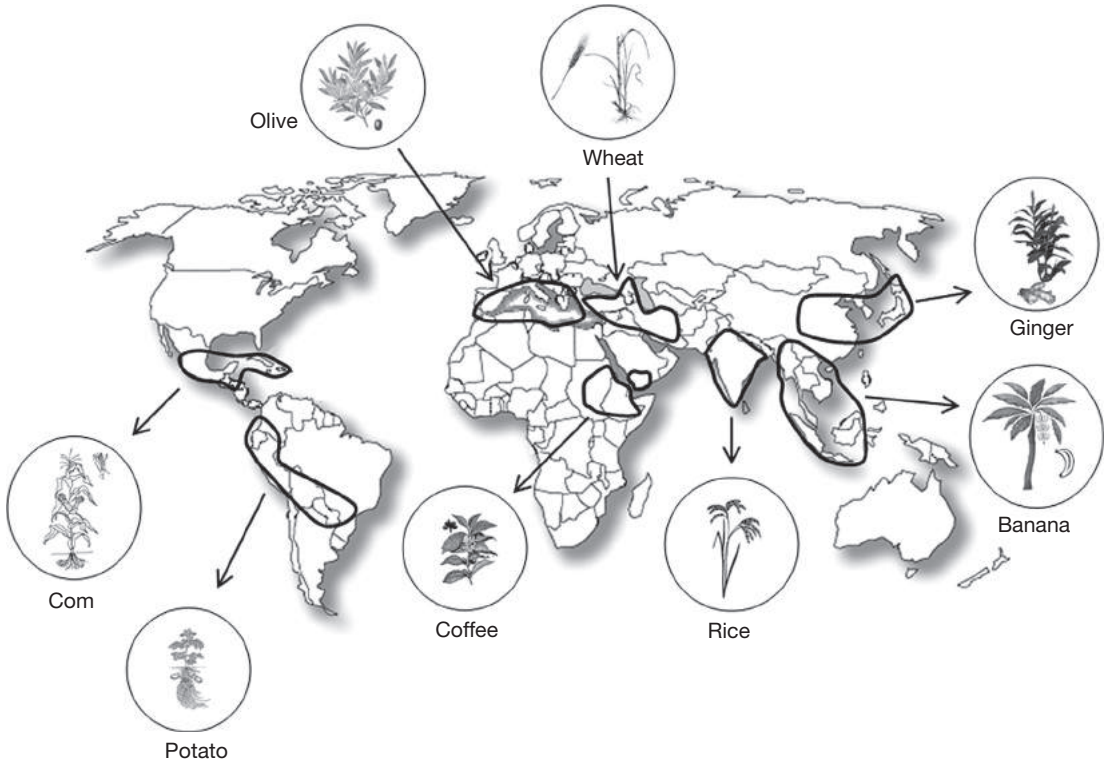


Fig. 4 The Vavilovian centers where traditional knowledge and human intervention on wild plants led to the origin of some of the most important crops.

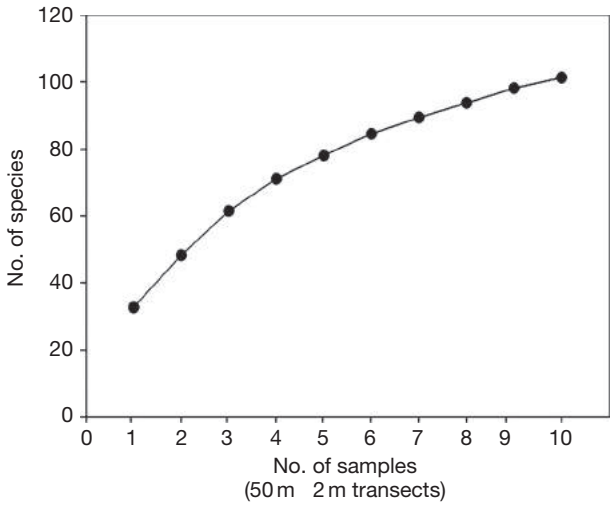


Fig. 5 Species accumulation curve derived from plant species sampled in ten independent transects established in rainforest understory in southeast Mexico. Each point is an estimate of the number of species expected for a given sampling effort (i.e., number of transects). As sampling effort increases, the slope of the curve becomes shallower, indicating that species inventory is approaching completeness (i.e., lesser new species are being added).

number of observed species can be considered to represent a complete inventory of species. The graph shows that at this site the estimated number of plant species in the understory habitat of the rainforest, based on the sampling effort of ten transects, is close to 110 and the rate of accumulation of additional species, as sampling effort increases, is very low. This approach, in addition to providing a trend of species accumulation (S), gives an idea of the completeness of the species inventory, whereby the shallower the slope of the curve at its right extreme, the lower the number of new species expected to be added with additional sampling. In this particular case, the known number of species in the understory microhabitat in this forest is close to the estimated value.

When information on species relative abundance is also available, more revealing descriptions of diversity can be undertaken by considering the distribution of individuals among species. One example of a common approach is the use of rank–abundance plots, also known as dominance–diversity curves (Fig. 6). In this type of plot, species are plotted in an increasing ranking sequence, from

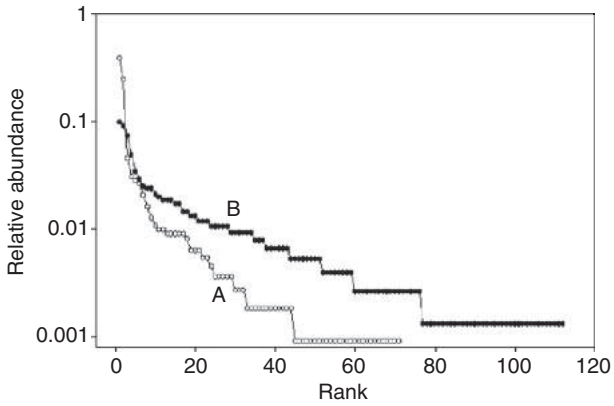


Fig. 6 Dominance–diversity curve comparing species richness and abundance of plant species sampled in two rainforest sites in southeast Mexico. Each species (represented by one circle) is ranked according to its abundance from more abundant (left) to less abundant (right) along the *x*-axis. Site B presents a more diverse species assemblage than Site A both in term of richness and in terms of the equitability in which individuals are distributed among species.

left to right on the *x*-axis, according to their abundance (far left the most abundant, far right the rarest). Proportional abundance is displayed on the *y*-axis using $\log_{10}$ scale, so that the resulting curve reflects the hierarchical distribution of abundances across species. Therefore, dominance/abundance plots readily display species equitability and comparisons among species assemblages with different species richness and equitabilities are easily accomplished. Fig. 6 shows the dominance/diversity curves for two tropical rainforest understory sites, A with 71 species, and B with 112 species (each species represented by open and closed circles, respectively). The plot highlights that while in Site A only a few species are clearly abundant, monopolizing a considerable proportion of the dominance (biologically interpreted as monopolization of the available resources), in Site B species are more evenly distributed. The contrast between Sites A and B shows that, in addition to the difference in the number of species, those of Site A are less equitably distributed than in Site B. Ecologists embrace the notion of equitability and use a metric to define it, evenness, with the implication that, complementary to species richness, the greater the evenness of a site, the greater its ecological diversity. The underlying idea is that if one (or few) species dominate the species assemblage, the community is not very diverse ("too much of the same"), whereas if no species dominate, the community is probably more diverse. Imagine these species represent prey of a predator looking for diversity of food resources in these two communities. Clearly the latter would appear as more diverse than the former to the predator.

The interest of having a synthetic measure of diversity based on the combination of two of its most distinctive attributes, species richness and species relative abundance, has favored the development of a wide variety of so-called diversity indices. These indices are sensitive to the number of species recorded but also to how contrasting their abundances are. Different indices are more or less sensitive to the characteristics of the data used for their calculation. For example, some indices are sensitive to variations in sample size while others are not. In addition, some diversity indices give greater weight to species richness (e.g., the Shannon–Wiener index) and others to the dominance of the commonest species (e.g., the Simpson index). Simpson's index assesses the probability that any two individuals taken at random from a community will belong to different species. It is easy to imagine how such probability would be high in a tropical forest in which tree species diversity is very high, compared to a low probability in a boreal forest where most trees belong to the same species.

In order to illustrate the importance of evenness, Box 1 shows the procedure to calculate the Simpson index ($1/D$) to compare the diversity of two hypothetical bird communities. The first step is to multiply the number of individuals in each species (n_i) by its frequency minus one ($n_i - 1$). The resulting products ($n_i\,(n_i - 1)$) are then divided by the product of the total number of individuals recorded (N) by $N - 1$. Finally, the resulting ratios ($n_i\,(n_i - 1)/N\,(N - 1)$) are added up over all the species to calculate D. The Simpson index is expressed as the inverse of D (i.e., $1/D$), in order to obtain increasing values of the index with increasing diversity. In the example presented in Box 1, Site 1 is about two times more diverse than Site 2, despite the fact that both sites have the same species richness ($S¼\ 8$). Differences in the relative abundance of species can be further examined by calculating the associated evenness index ($E_{1/D}$) that results from dividing the observed index by the maximum diversity, that is, when all species are represented by the same number of individuals. The $E_{1/D}$ values show that the bird assemblage of Site 1 is considerably more "even" that that of Site 2. This is consistent with the fact that the most abundant species of Site 2 accounts for 59% of all individuals in the assemblage (high dominance), while the most abundant species of Site 1 only represents 20% of the total.

The development of new methods to analyze and characterize diversity is a very active field of research. Environmental DNA (or DNA metabarcoding) in particular is an emerging tool for detection of animal species, and thus in monitoring past and present biodiversity. Recent approaches try to combine taxonomic and functional species traits with traditional information on species richness and abundance to generate more comprehensive descriptions of diversity. This may allow distinguishing between localities, even when they have the same number of species and similar species–abundance distributions. For example, if species in one locality belong to a few genera and species in a second locality span over a more ample variety of genera, the latter would be

Box 1

A hypothetical example showing how the distribution of individuals among species changes the diversity of bird assemblages occurring in two sites. Although the two sites have the same number of species, the more even distribution of individuals among species at Site 1 leads to a higher Simpson diversity index and evenness than at Site 2 (see text for details).

considered as more diverse. A similar situation would occur if species in one locality show a greater variety of functional roles (e.g., animal feeding guilds or plant life forms) than the species in the other locality. In sum, the ample variety of ways biodiversity manifests itself, calls for a diversity of criteria to characterize it.

Spatial Scales of Diversity

An essential aspect of biodiversity is that species contingents vary in their spatial distribution, sometimes abruptly, making it necessary to explicitly consider the spatial scale at which diversity is being analyzed. Fig. 7 depicts a simple representation of the different levels of spatial scale at which diversity can be analyzed, portraying a nested scheme of distribution of diversity, scaling up from homogeneous habitats (that we will refer to as localities in this exercise), nested within regions, which in turn are nested within a larger biogeographic region or province. In this scheme, the most basic, local measure of diversity is alpha diversity, which consists of the number of species recorded within a given locality.

In the four localities depicted in Fig. 7, diversity is the same: each locality contains three different species. However, the identity of each of the species in each of the contingents varies from locality to locality. That is, there is a high turnover of species as we move from one locality to another. Ecologists and conservation biologists use the term beta diversity to refer to species turnover. At the larger spatial scale of the region, we observe that the high turnover of species within each region determines a high level of regional diversity, corresponding to a high gamma diversity. Finally, given the high species turnover between regions, the overall biogeographic region exhibits a high delta diversity. It is worth noticing that the specific definitions of locality, region, or biogeographic region are somewhat subjective and vary with the investigator. In any case, the importance of the spatial scales of distribution of diversity for conservation can hardly be overemphasized: if a country, region, or biome is composed of a mosaic of localities of high diversity with a high species turnover (diversity), the protection of the country's biodiversity will demand the establishment of many preserves probably of different types, rather than a few or a single one, even of relatively large size.

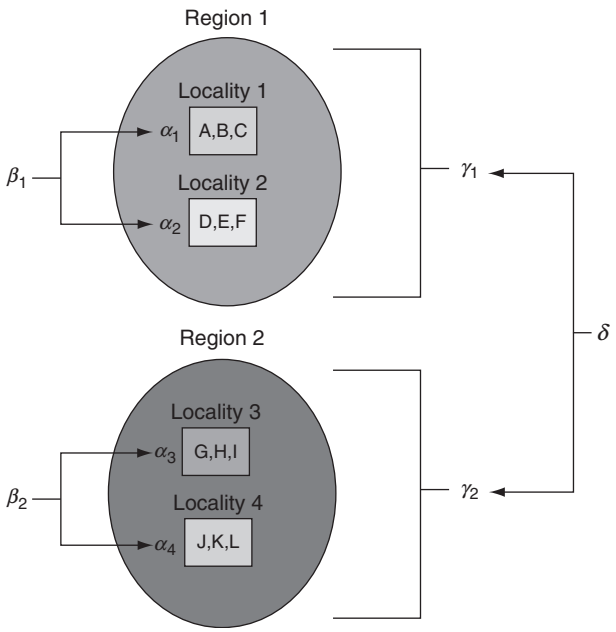


Fig. 7 Schematic representation of the different spatial scales at which diversity can be analyzed. The identity of the species found at each locality (diversity) changes (i.e., species turnover) across localities (diversity). The combination of and diversity makes up the diversity found at the regional level (diversity). Similarly, when moving between regions, a turnover in the species identity occurs (diversity).

Time Course of Biodiversity Through Geological Time

Our knowledge of the patterns of change of biological diversity through geologic time is mainly based on the information contained in the fossil record. The available fossil evidence suggests that the diversity of families of multicellular marine organisms (Fig. 8) rose steadily through the Cambrian and Ordovician period, attaining a plateau about 440 million years ago and then was punctuated by a great wave of extinction in the Permian (290–245 million years ago). After this, it steadily increased to the present.

Terrestrial organisms first appeared about 440 million years ago, with the invasion of the land by the ancestors of plants, fungi, vertebrate animals, and arthropods. In this period, each group increased rapidly in diversity from that time onward. At the species level, terrestrial vascular plants began to diversify markedly around 400 million years ago and declined during the worldwide Permian extinction event that also affected marine organisms profoundly. After this, similar to marine organisms, plants began to diversify around the Mid-Cretaceous Period, some 100 million years ago, with the flowering plants (angiosperms) becoming an extremely diverse and dominant group up to the present.

Thus the fossil records of both marine and terrestrial multicellular eukaryotes indicate maximum diversity at more likely than old rocks to be well preserved, and thus the most recent occurrences of species are more likely to be found than the older occurrences. On the other hand, a considerable fraction of the recent marine faunas are known from a few, restricted localities. Although paleontologists have amassed an impressive amount of information and analyzed it with sophisticated statistical methods, the patterns emerging from these analyses, although exciting and compelling, are still in need of further confirmatory work.

Global Magnitude of Biodiversity

Our understanding of the global magnitude of biodiversity, even considering the level for which we have the best information, species, is still very limited. Complete catalogs of the described, valid species exist for only a few groups of organisms, and the total can only be estimated. This figure is estimated to be around 1.5 million and includes exclusively eukaryotic organisms (basically

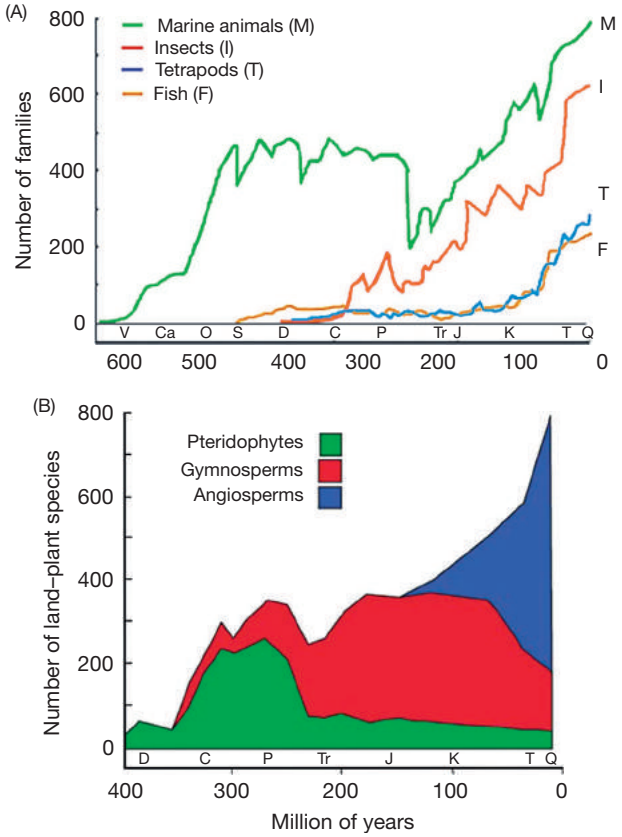


Fig. 8 Time course of (A) the number of families of marine animals, insects, tetrapods, and fish through the last 600 million years and (B) the number of land–plant fossil species in three major groups (angiosperms, gymnosperms, and pteridophytes) through the last 400 million years. *V*, Vendian; *Ca*, Cambrian; *O*, Ordovician; *S*, Silurian; *D*, Devonian; *C*, Carboniferous; *P*, Permian; *Tr*, Triassic; *J*, Jurassic; *K*, Cretaceous; *T*, Tertiary; and *Q*, Quaternary. Reprinted with permission from the Dirzo, R. and Mendoza, E. (2003). General ecology – Biodiversity. Annual Review of Environmental and Resources **28**, 374 by Annual Review.

plants, some groups of animals, and some fungi). The number and definition of prokaryotic species and viruses are still very limited. It is not surprising, therefore, that current estimates of the total number of species in the world span over 2 orders of magnitude (10^6–10^8). Methods employed to estimate global diversity can be classified in three broad categories: (1) ratios of known to unknown groups, (2) extrapolation from samples, and expert opinions by scientists who understand the level of diversity in a particular group of organisms well, or combinations thereof. An example of the first approach is the estimation carried out by ecologists Nigel Stork and Kevin Gaston in 1990. They first noted that the ratio of the number of species of butterflies to all species of insects on the well-known British fauna is 67:22,000. Assuming that such a relationship was robust overall, and based on an estimate of 15,000–20,000 species of butterflies world-wide, they scaled up to a total figure between 4.9 and 6.6 million species of insects. The second approach is illustrated by the study carried out by entomologist Terry Erwin at a rainforest site in Panama. Erwin performed a systematic sampling using a fogging protocol targeted on the beetle fauna living in the canopy of the tree *Luehea seemannii*. Based on the magnitude of the fauna collected and using a series of assumed relationships involving the ratio between beetles and other types of arthropod species, the degree of host specificity of arthropods, and the expected number of tree species to be found in the Tropics, Erwin came up with the astounding estimate of 30 million species of tropical arthropods. Robert May attempted an estimate based on the known relationship between body-size category and number of species of animals. Assuming that such relationship holds for individuals as small as 0.2 mm, May calculated that there might be about 10 million species of animals. All these estimations are based on a series of assumptions that await confirmation. However, there is some level of agreement that the plausible number of global species falls within a range of 5–15 million and a recent review suggests a best guess of around 7 million species.

Yet, new species are continuously described sometimes as a consequence of refinements in the taxonomic classifications but commonly as the result of real discoveries from surveys in nature. Current rates of publication of new species are significant, including some 13,000 animal species per year. This makes it evident that the task of describing the total number of species on Earth will not be completed for many decades, even assuming they can all be collected and analyzed by the appropriate experts before they go extinct due to anthropogenic impact. These discoveries are not restricted to small-sized organisms but also include species of large mammals such as monkeys and deer. Even more, exploration of remote areas sometimes results in the discovery of organisms that generate a biodiversity revolution because of their degree of biological novelty. Examples of this are the plant *Lacandonia schismatica*, found two decades ago in southeast Mexico. This plant was found to belong into an entirely new taxonomic family (Lacandonianceae), due to its spectacular characteristics, including that it presents its reproductive organs with a spatial disposition the other way around from any other known flowering plant: male structures toward the center of the flowers and female structures at the periphery! Equally spectacular as taxonomic novelty is the discovery, a few years ago, of an insect in Africa that combines characteristics of a cricket and a stick insect, so different to any other insect that it required the creation of a new taxonomic order (Mantophasmatodea). All these discoveries underscore the level of uncertainty we have about the magnitude of biodiversity. This uncertainty becomes more striking when we take into consideration that species richness is only one facet of biodiversity, as discussed earlier.

Geographic Distribution of Biodiversity

Among the many trends known about the geographic distributional patterns of biodiversity, the most evident is that of the increase in species diversity with latitude. This trend is consistent across many groups of organisms that have been well analyzed. Among insects this pattern is spectacular. For example, the number of species of ants in local regions increases from about 10 at latitude of 60 N, to some 2000 at latitudes near the equator. Fig. 9 highlights the case of plants and a representative group of vertebrates, birds. In both cases, the upper bound limit of species richness shows a dramatic peak at lower latitudes. In the case of plants, different analyses (e.g., local or regional floras, species density comparisons, standardized diversity samples, etc.) show a very consistent gradient of decrease in plant species diversity with latitude. In broad geographic terms, species densities range from over 5000 species/10,000 km^2 in tropical regions, to <100 in the highest latitudes. In terms of local species diversity, values range from an average of 270 species per 0.1 ha in Colombia to *c*. 15 near the United States–Canada border. Breeding birds increase from about 56 species in Greenland, to 105 in New York, to 1010 in Mexico, to 1395 in Colombia.

Another increasingly evident pattern of geographic variation in diversity has to do with intercontinental variation, specifically the patterns comparing the Neotropics (i.e., the tropics of the New World) with the Paleotropics. Data indicate that about 90,000 species of plants, approximately twice as many as in Africa, south of the Sahara, occur in the Neotropics and that the comparable area of Asia is roughly intermediate in this respect. Fogging sampling techniques using standardized protocols to compare data for canopy beetles (e.g., in terms of species per cubic meter) show the same tendency, although the values are even more contrasting than in the case of plants: 1.17 in Panama and 1.15 in Peru >0.29 in New Guinea >0.02 in Australia and Sulawesi. Similar trends have been observed in numerous other groups, including butterflies (Neotropics > Southeast Asia > Africa), frogs (Neotropics > Africa/Asia > Papua/Australia), and birds (Neotropics > Africa > Asia/Pacific > Australo-papuan). In mammals, the number depends on the particular group; specious groups such as bats are considerably more diverse in the Neotropics than in the Old World, while some groups with relatively few species in general, such as primates, show the opposite trend: Old World > New World.

Fig. 9 Gradients of latitudinal variation in species richness for birds (left) and plants (right). For both groups, species concentrates toward the Tropics peaking around the equator. The data for plants correspond to the number of species per 0.1 ha in different localities of the Western Hemisphere. The number of bird species is standardized per unit of sampling effort in different localities of the world.

Valuation of Biodiversity

The value of biodiversity has been considered from several points of view which can be classified in the following three broad categories: (1) esthetic, (2) ethical, and (3) economic. The esthetic point of view posits the idea that biodiversity includes a wealth of expressions of beauty equivalent to those found in the most esteemed collections of art work. Such an array of beauties ranges from vividly colored beetles and butterflies to whales and ancient forests. Moreover, these expressions of beauty are the result of very long evolutionary processes that exceed by far the age of the most ancient artwork.

The ethical point of view rests on the idea that biodiversity, by itself, has an intrinsic value. This point of view has its roots in philosophical beliefs and considerations that give other forms of life the same rights to exist and meet their needs as humans. This idea is complemented by the notion that *Homo sapiens*, the species currently monopolizing a large share of the energy and resources that support life on Earth, has the ethical responsibility to secure the preservation of other forms of life.

Economics criteria argue that biodiversity provides humanity with monetary revenues directly and indirectly. A classic example of a direct profit coming from biodiversity is illustrated by the variety of chemical compounds obtained from plants, animals, and microorganisms that function as a base for the active ingredients used in a large proportion of the available prescription drugs (e.g., digitalis, morphine, quinine, and antibiotics). In comparison, the notion of indirect profits of biodiversity rests in the realization that several organisms maintain and regulate processes that impact the quality of human life. For example, organisms inhabiting soil (e.g., earthworms and insects) are crucial for maintaining fertility and henceforth allow the growth of crops and forests. Another example is the case of plant pollinators. An important number of crops depend on the "service" provided by wild pollinators. Efforts have been focused to estimate the economical cost that the loss of such ecological services might involve. In the case of pollination by native insects in the United States, a study estimated that the ecological service they provide is worth $3.07 billion per year.

Recent interest in biodiversity valuation has increased in response to the threats it is facing. In this regard, the different criteria we presented have more or less potential to play a role in increasing the awareness about the relevance of conserving biodiversity. Esthetic appreciation of biodiversity has the caveat that it is, in some sense, biased toward the small subset of species that are considered "charismatic" such as whales or birds. Ethical considerations offer the most comprehensive valuation of biodiversity; however, it seems difficult that this type of philosophy will become internalized by a significant proportion of the humanity in the short term. Finally, the economic arguments are compelling and constitute a more tractable argument within the framework of formal markets. However, there are still a reduced number of cases where it has been possible to document with detail the economical value of the services provided by biodiversity. In the end, it is worth keeping in mind that the level of interrelatedness biological systems usually show, determines that the existence of charismatic, economical, or functionally valuable species depends on the maintenance of an unknown number of associated species and ecological processes.

An appealing approach, related to the three arguments referred to above, is that formulated by the Millennium Ecosystem Assessment (MEA). As a large coalition of international conservation and development organizations, governments, and a significant representation of the scientific international community, the MEA has compiled the most thorough assessment of the state of the planet's ecosystems, emphasizing the goods and services they provide, and the likely effects of potential pathways of human economic development on the provisioning of such goods and services (scenarios) and the interrelations thereof with human well-being. We can summarize the logic of the relationships articulated by the MEA, in brief, as follows: biodiversity, represented by the genes, populations, species communities, and biomes, generates a series of supporting services resulting from

ecosystem functioning. Such services, including primary production, nutrient cycling, and soil formation, are the basis for all other ecosystem services. The latter services belong to three major categories: (1) provisioning services, that is, products obtained from ecosystems, including food, fresh water, fuel wood, fiber, biochemical compounds, and a plethora of other genetic resources; (2) regulating services, those that produce benefits obtained from the regulation of ecosystem processes (e.g., climatic regulation, disease regulation, water regulation, air purification, pollination services, biological control); and (3) cultural services, the nonmaterial benefits obtained from ecosystems. These include the esthetic, inspirational, religious, and spiritual values offered by nature, recreation and tourism, educational services, and cultural heritages. The provisioning of such services impinges on human well-being in terms of affording basic materials for a dignified life, health, security, and good social relations. It is hoped that the framework of and the information summarized in the MEA, together with the formulation of future scenarios depending on different routes of economic development, will be used to guide policy regionally and globally. From a more ecological point of view, another interesting derivation of the MEA is that it can provide the framework to focus on the relevant research addressing the connections between biodiversity conservation and ecosystem services, and the influence of biodiversity on human well-being and vice versa.

Further Reading

Cardinale BJ, Srivastava DS, Duffy JE, Wright JP, Downing AL, Sankaran M, and Jouseau C (2006) Effects of biodiversity on the functioning of trophic groups and ecosystems. *Nature* 443(7114): 989.

Colwell RK and Elsensohn JE (2014) EstimateS turns 20: Statistical estimation of species richness and shared species from samples, with non-parametric extrapolation. *Ecography* 37(6): 609–613.

Heller NE and Zavaleta ES (2009) Biodiversity management in the face of climate change: A review of 22 years of recommendations. *Biological Conservation* 142(1): 14–32.

Henle K, Alard D, Clitherow J, Cobb P, Firbank L, Kull T, McCracken D, Moritz RF, Niemelä J, Rebane M, and Wascher D (2008) Identifying and managing the conflicts between agriculture and biodiversity conservation in Europe—A review. *Agriculture, Ecosystems & Environment* 124(1–2): 60–71.

Heyer R, Donnelly MA, Foster M, and Mcdiarmid R (eds.) (2014) *Measuring and monitoring biological diversity: Standard methods for amphibians*. Washington, DC: Smithsonian Institution.

Heywood VH (ed.) (1995) The current magnitude and distribution of biodiversity. In: *Global biodiversity assessment*. Cambridge: Cambridge University Press.

Hillebrand H and Matthiessen B (2009) Biodiversity in a complex world: Consolidation and progress in functional biodiversity research. *Ecology Letters* 12(12): 1405–1419.

Magurran AE (1988) Why diversity? In: *Ecological diversity and its measurement*, pp. 1–5. Dordrecht: Springer.

Magurran AE (2004) *Measuring biological diversity*. Oxford: Blackwell.

Mantyka-pringle CS, Martin TG, and Rhodes JR (2012) Interactions between climate and habitat loss effects on biodiversity: A systematic review and meta-analysis. *Global Change Biology* 18(4): 1239–1252.

Nicholls CI and Altieri MA (2013) Plant biodiversity enhances bees and other insect pollinators in agroecosystems. A review. *Agronomy for Sustainable Development* 33(2): 257–274.

Purvis A and Hector A (2000) Getting the measure of biodiversity. *Nature* 405: 212–219.

Rees HC, Maddison BC, Middleditch DJ, Patmore JR, and Gough KC (2014) The detection of aquatic animal species using environmental DNA—A review of eDNA as a survey tool in ecology. *Journal of Applied Ecology* 51(5): 1450–1459.

Rissman AR and Gillon S (2017) Where are ecology and biodiversity in social–ecological systems research? A review of research methods and applied recommendations. *Conservation Letters* 10(1): 86–93.

Stendera S, Adrian R, Bonada N, Cañedo-Argüelles M, Hugueny B, Januschke K, Pletterbauer F, and Hering D (2012) Drivers and stressors of freshwater biodiversity patterns across different ecosystems and scales: A review. *Hydrobiologia* 696(1): 1–28.

Taberlet P, Coissac E, Pompanon F, Brochmann C, and Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology* 21(8): 2045–2050.

Relevant Websites

https://www.greenfacts.org/en/biodiversity/l-3/1-define-biodiversity.htm – GreenFacts 2001–2018.
https://eatlas.org.au/content/what-biodiversity – eAtlas.

Biomass

RA Houghton, Woods Hole Research Center, Falmouth, MA, USA

Definition

Biomass refers to the mass of living organisms, including plants, animals, and microorganisms, or, from a biochemical perspective, cellulose, lignin, sugars, fats, and proteins. Biomass includes both the above- and belowground tissues of plants, for example, leaves, twigs, branches, boles, as well as roots of trees and rhizomes of grasses. Biomass is often reported as a mass per unit area ($g\,m^{-2}$ or $Mg\,ha^{-1}$) and usually as dry weight (water removed by drying). Unless otherwise specified, biomass usually includes only living material. For example, neither deadwood nor the organic matter of soils is considered biomass, although soils do contain biomass in the form of bacteria, fungi, and meiofauna. Generally, the biomass of soils (living and dead microbes) is $<5\%$ of soil organic matter.

This article deals almost exclusively with terrestrial biomass and particularly with the biomass of plants. Forests account for 70%–90% of terrestrial biomass, most of this biomass in trees. The distribution of terrestrial biomass among producers, consumers, and microbes is *c.* 0.90%, 0.001%, and 0.10%, globally, with considerable variation among ecosystems. Producers and microbes, for example, may account for similar fractions of biomass in some nonforest ecosystems. The estimates given here refer to the biomass of plants, or producers, only.

Importance

The biomass of plants is of interest for at least three reasons: ecosystem structure, timber, and carbon stocks. Biomass determines the structure of ecosystems. Terrestrial ecosystems generally have more biomass than aquatic ecosystems (compare trees with phytoplankton), and forests have a higher biomass than other types of ecosystems. As a result of the greater physical complexity in the distribution of this biomass, forests also provide a greater diversity of habitats.

The biomass of forests is particularly important as a source of wood for timber and pulp, and forests are managed to optimize production of wood. It is worth noting that the `production' of wood is not the same as the standing stock or amount of wood. Mature forests, for example, have high stocks of wood (high biomass), but `woody grasses' (e.g., sugar cane, bamboo, switchgrass) may be more productive of biomass and thus are receiving attention as potential biomass fuels.

A third reason for interest in biomass is that on a dry weight basis it is 50% carbon. Reductions in biomass through the clearing of forests for croplands, for example, release carbon (as CO_2) to the atmosphere, and reforestation withdraws carbon from the atmosphere. Reducing deforestation and increasing biomass (carbon sequestration – both terrestrial and oceanic) are both options important for managing carbon, and, thereby, the rate of climatic disruption.

Quantities of Biomass

The quantity of biomass per unit area varies spatially and temporally. Living biomass ranges over 2–3 orders of magnitude, from averages of $\sim 400\,Mg\,ha^{-1}$ in tropical forests to averages of less than $10\,Mg\,ha^{-1}$ in treeless grasslands, croplands, and deserts (**Table 1**). Biomass also varies considerably within ecosystem types. Some tropical forests and forests of the Pacific Northwest in North America may attain values in excess of $600\,Mg\,ha^{-1}$. In part this variability results from differences in environment (e.g., soil nutrients or the seasonal distribution of precipitation), and in part it results from disturbance and recovery. A recently burned forest has a living biomass of nearly zero, and as it recovers it reaccumulates carbon (**Fig. 1**). Forests do not accumulate biomass indefinitely, however, because stand-replacing disturbances keep turning old forests into young ones.

Most of the biomass in forests is in the stems, or boles, of trees, with branches, roots, and foliage accounting for lesser fractions. The fractions vary with growth, however. As forests grow larger, a greater and greater fraction of total biomass occurs in boles.

The estimation of forest biomass depends, in part, on spatial scale. At the level of a forest stand, biomass varies through time as a result of disturbances (and recovery). At the landscape scale, biomass varies through space because the ages (since the last disturbance) of patches vary across the landscape. **Fig. 2** shows the ages of forest stands in a $\sim 450\,km^2$ area in the province of Krasnoyarsk, Russia. The remarkable feature of this landscape is the degree to which the forest is a mosaic of different aged stands. The spatial variability in biomass is much greater than it appears from observations of the canopy from the air or from space and suggests that most forests are in the process of recovering from natural or human-induced disturbances.

The estimates of biomass shown in **Table 1** suggest a global total of ~ 1300 Mg, and forests account for more than 80% of that total. But the uncertainty is high. Another estimate suggests a global terrestrial biomass of only 770 Mg (see below).

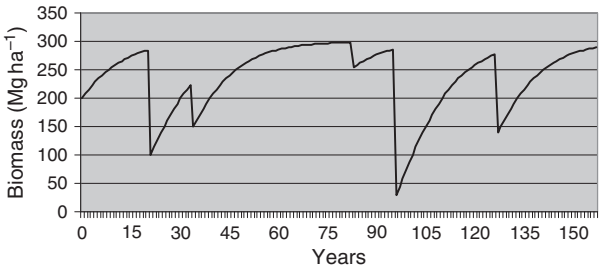


Fig. 1 Changes in the living biomass of a forest (Mg ha^{-1}) in response to disturbances (e.g., fire or logging) and recovery. Each disturbance removes living biomass, which subsequently reaccumulates as a consequence of growth.

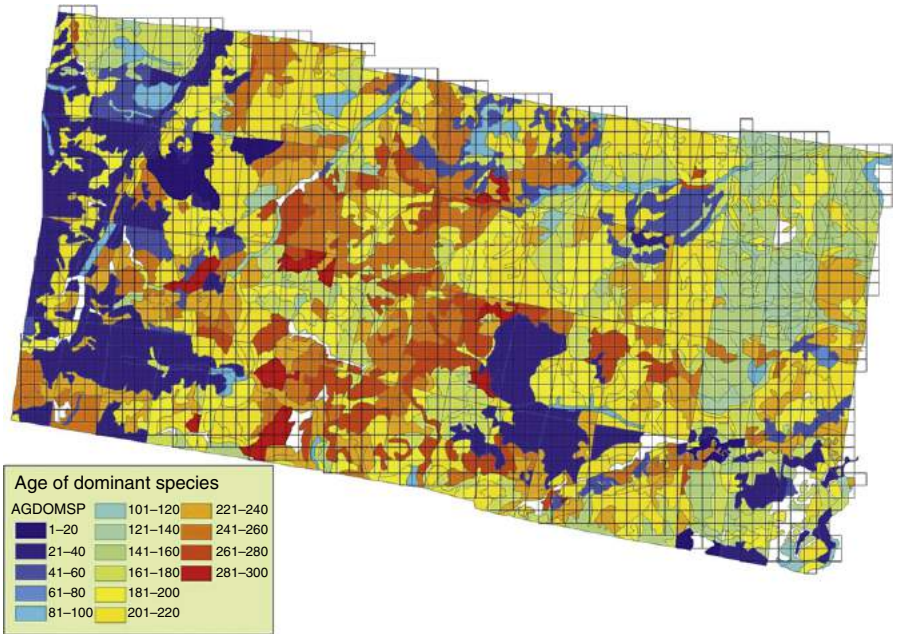


Fig. 2 Forest stands of different age in eastern Krasnoyarsk, central Siberia. Each cell is 500 m $\times$ 500 m.

Measurement of Biomass

There are two aspects to quantifying biomass: measurements at specific sites and methods for extrapolating the results from such sites to large areas. Measurement at a site is, in principle, straightforward. It involves harvesting all plant material within a defined area, drying it to a constant weight, and weighing it. In practice, this type of destructive measurement becomes more difficult if the belowground portions are included and if the vegetation includes large trees. Sorting roots from the soil–root matrix is difficult. The size of the plot is also important: small plots will either over- or underestimate the average biomass if they include or exclude large trees, respectively. Perhaps, more importantly, selection of sites is critical. Sites need to be representative of existing forests, not representative of some idealized notion of a forest. Sampling a large number of sites may help minimize biased selection.

Because destructive sampling is time consuming (expensive), foresters and ecologists have developed indirect methods for estimating biomass. The most common approach uses allometric equations that allow the estimation of tree biomass from more easily measured properties, such as diameter at breast height (dbh), height, and wood density. Systematic sampling (forest inventories) with such measurements and equations allow the biomass (or commercial wood volumes) to be obtained over large areas.

Because temperate zone and boreal forests have been inventoried more than once, their biomass is reasonably well known (Table 2). Accuracies vary among regions and countries but they are generally high for wood volumes. In the Southeastern US, for example, the error calculated for the total volume of growing stock was within 1.1%. On the other hand, `changes' in the stock had an error of $\pm 40\%$. The much larger error associated with change results from magnitude of the change relative to the magnitude of the stocks. Small differences between large numbers are difficult to determine precisely.

Table 1 Mean living biomass, area, and total living biomass of the world's major terrestrial ecosystems

Ecosystem type	*Area (10^6 ha)*	*Total biomass (Pg)*	*Mean biomass (Mg ha^{-1})*
Tropical forests	1750	680	390
Temperate forests	1040	280	270
Boreal forests	1370	110	83
Arctic tundra	560	4	7
Mediterranean shrublands	280	34	120
Croplands	1350	8	6
Tropical savannas and grasslands	2760	160	57
Temperate grasslands	1500	12	8
Deserts	2770	20	7
Ice	1550	0	0
Total	14930	1308	87

Adapted from Saugier B, Roy J, and Mooney HA (2001) Estimations of global terrestrial productivity: Converging toward a single number? In: Roy J, Saugier B, and Mooney HA (eds.) *Terrestrial Global Productivity*, pp. 543–557. San Diego, CA: Academic Press.

Table 2 Areas, total carbon stocks, and average carbon stocks in the biomass of forests and woodlands in the northern temperate and boreal zones in 1990

Region	*Forest area (10^6 ha)*	*Other woodland area (10^6 ha)*	*Forest living biomass (Pg)*	*Woodland living biomass (Pg)*	*Average forest biomass (Mg ha^{-1})*	*Average woodland biomass (Mg ha^{-1})*
Canada	316	88	26	3.2	82	36
United States	212	86	27	6.6	125	77
Europe	149	46	15	0.4	103	9
Russia	821	66	67	1.2	82	18
China	119	39	9.2	1.2	77	31
Other[a]	92	16	9.4	n/a	102	n/a
Total	1710	340	154	12.6	90	37
Forest and other woodland combined	2050		166.6		81	

[a]Countries included: Japan, North Korea, South Korea, Mongolia, Latvia, Lithuania, Estonia, and the Commonwealth of Independent States other than Russia.
Adapted from Goodale CL, Apps MJ, Birdsey RA, *et al.* (2002) Forest carbon sinks in the Northern Hemisphere. *Ecological Applications* 12: 891–899.

Table 3 Areas and average biomass of natural forests in tropical regions[a]

Region	*Forest area (10^6 ha)*	*Total biomass*[b]	*Average biomass*[b] (Mg ha^{-1})
Asia	264	37	140
Africa	636	85	134
America	952	225	236
Total	1852	348	188

[a]In Africa, three nontropical countries, Lesotho, South Africa, and Swaziland, are included, while the six countries of northern Africa are not. In Latin America, three nontropical countries, Argentina, Chile, and Uruguay, are included.
[b]Values of aboveground biomass were increased by a factor of 1.2 to include belowground biomass.
Adapted from FAO (2001) *Global Forest Resources Assessment 2000*. Rome: Food and Agriculture Organization of the United Nations.

Although forest inventories have been carried out in some tropical forests, large areas have rarely been inventoried. Nevertheless, estimates of the average tropical forest biomass exist (**Table 3**), based on one of three approaches: (1) dividing the forests into different classes (each type corresponding to an average biomass), (2) calculating biomass as a function of environmental parameters (e.g., mean annual temperature and the seasonal distribution of precipitation), and (3) `direct' measurement from remote sensing data. A comparison of seven estimates for the Brazilian Amazon revealed not only a wide range (greater than a factor of 2 between the lowest and highest estimates of total biomass), but also no agreement as to where the forests with the highest and lowest biomass existed (**Fig. 3**). More recent investigations with radar and lidar show much more consistent results, and satellites to measure aboveground biomass directly are being designed by space agencies in Japan, Canada, Europe, and the United States.

Such satellites, in concert with field measurements, are urgently needed. A comparison of **Tables 2** and **3** with **Table 1** suggests an uncertainty in biomass that is a factor of 2, especially if the differences for forests apply to other ecosystems. **Table 1** is a recent compilation of site-specific measurements in different types of ecosystems. In contrast, **Tables 2** and **3** are based on forest

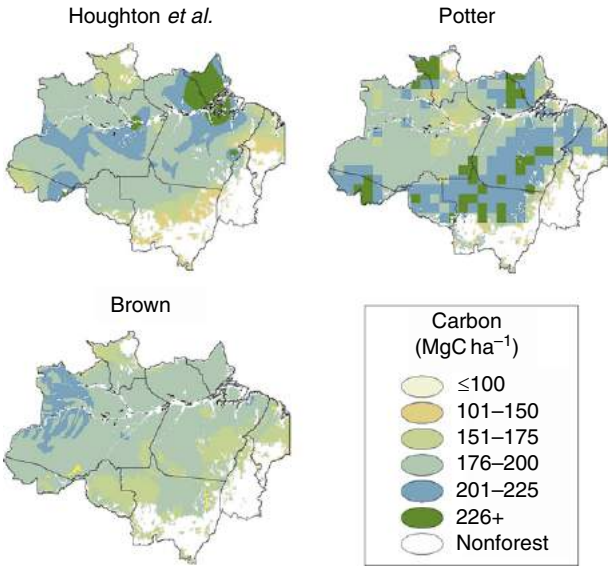


Fig. 3 Three estimates of the distribution of forest biomass in the Brazilian Amazon. Reproduced from Houghton RA, Lawrence KT, Hackler JL, and Brown S (2001) The spatial distribution of forest biomass in the Brazilian Amazon: A comparison of estimates. *Global Change Biology* 7: 731–746, with permission from Wiley-Blackwell Publishing Ltd.

Table 4 Area, total living biomass, and mean biomass of the world's major terrestrial ecosystems

Ecosystem type	*Area (10^6 ha)*	*Total biomass (Pg)*	*Mean biomass (Mg ha^{-1})*
Tropical forests[a]	1850	350	190
Temperate forests[b]	2450	185	75
Arctic tundra	560	4	7
Mediterranean shrublands	280	34	120
Croplands	1350	8	6
Tropical savannas and grasslands	2760	160	57
Temperate grasslands	1500	12	8
Deserts	2770	20	7
Ice	1550	0	0
Total	15070	770	50

[a]From **Table 3**.

[b]Includes temperate and boreal forests from **Table 2** and Australia's forests (from Dixon RK, Brown S, Houghton RA, *et al.* (1994) Carbon pools and flux of global forest ecosystems. *Science* 263: 185–190). All other rows are from **Table 1**.

inventories with a much larger number of measurements, presumably more representative of existing forests. Although some of the differences may be the result of changes wrought by human activity (natural forests degraded through use), most of them are probably the result of a biased selection of sites for **Table 1**.

Table 4 shows an alternative estimate of global biomass, where the estimates for forests are derived from **Tables 2** and **3** and the estimates for other ecosystems are the same as in **Table 1**. Although the areas of forests (and land ecosystems) are similar in **Tables 1** and **4**, global biomass from this new compilation is approximately half of that in **Table 1** (770 as against 1300 Pg). That is, global biomass is not known to even one significant figure. Similarly, estimates of the average biomass of forests vary by more than a factor of 2, a fact that is remarkable given that wood volumes are determined to within 1%–2% in national forest inventories. Some fraction of the discrepancy results from scaling wood volumes to total biomass, including not only roots, leaves, and branches, but also understory vegetation, noncommercial species, and trees smaller than those generally inventoried. Much of the difference between estimates, however, is probably the result of site selection.

Changes in Biomass

Biomass (dry weight) is approximately 50% carbon, and the amount of carbon in terrestrial biomass (385–650 PgC) is of the same order of magnitude as the amount of carbon in the atmosphere (~780 PgC) and in the surface layers of the ocean (~700 PgC). Thus, changes in terrestrial biomass have a significant effect on the concentration of CO_2 in the atmosphere.

The carbon released to the atmosphere as a result of deforestation and degradation of tropical forests (1–2 PgC yr^{-1}) is thought to account for 10–20% of anthropogenic emissions (8.4 PgC yr^{-1} were released in 2006 from fossil fuel combustion). In contrast, temperate zone and boreal forests are believed to account for a significant carbon sink in northern mid-latitudes ($\sim$2 PgC yr^{-1}).

Knowing `changes' in biomass is crucial for determining terrestrial sources and sinks of carbon. However, the carbon sink in northern mid-latitude forests calculated on the basis of data from forest inventories (0.65 PgC yr^{-1}) is smaller than the carbon sink inferred from the top-down approaches to evaluating carbon sources and sinks from atmospheric data and models ($\sim$2 PgC yr^{-1}). The difference might be explained by errors (both estimates have wide ranges of uncertainty), by ecosystems other than forests that are not systematically inventoried, or by the accumulation of carbon belowground (roots and soil carbon are not measured by forest inventories).

If the northern mid-latitude sink of $\sim$2.0 PgC yr^{-1} were evenly distributed over the forests of the region, the average annual increase in forest biomass would be 1.2 MgC ha^{-1} yr^{-1}, or 2.7%. Such a change would be difficult to observe with an error of $\pm$40%. However, the sink is not evenly distributed in space. Forest inventories indicate that Canadian and Russian forests lost living biomass $\sim$1990 (0.08 PgC), while forests in the US, Europe, China, and other northern regions gained it (a total of 0.28 PgC). The spatial variability raises the possibility that the major terrestrial sources and sinks of carbon may be much larger than the average sink. What if, for example, 90% of the net terrestrial flux of carbon occurs on lands where the annual changes are large, 12 MgC ha^{-1} yr^{-1}, instead of 1.2 MgC ha^{-1} yr^{-1}? At present, we do not know what fraction of the northern forests is growing and what fraction is already `grown´ (**Fig. 2**). We do not know whether the terrestrial carbon sink is distributed over very large areas (with small annual changes) or limited to areas characterized by rapid rates of change.

In the tropics, the general lack of repeated forest inventories makes it impossible to estimate the sources and sinks of carbon directly from observation. Rather, changes in land use have been used, together with the changes in biomass known to be associated with land-use change, to calculate sources and sinks of carbon. The calculated emissions from tropical deforestation and degradation vary between 1 and 2 PgC yr^{-1}. Uncertainties in biomass contribute about as much to this variability as uncertainties in rates of deforestation. Moreover, estimates of average biomass are insufficient for accurate estimates of carbon emissions. Regional averages, even if accurate, do not necessarily correspond to the biomass of the forests actually deforested.

Carbon emissions determined from deforestation and degradation in the tropics are consistent with the net source of carbon inferred from top-down analyses based on atmospheric data and models. On the other hand, repeated measurements of biomass on small plots of undisturbed forests throughout the tropics indicate an increase in the biomass of (undisturbed) Amazonian forests (although not in tropical Africa or Asia). If the increase applies over large areas of Amazonia, it represents a significant carbon sink. The reasons for the increase in Amazonian biomass are unclear. Are the forests recovering from earlier disturbances, or are they responding to a changing environment (e.g., higher concentrations of CO_2 in the atmosphere or less cloudiness)? Answering this question is important for understanding whether the accumulation of carbon in biomass is part of a steady-state system or a new carbon sink, and, if it is new, whether it can be expected to continue.

See also: General Ecology: Abundance

Further Reading

Brown S (1997) *FAO Forestry Paper 134: Estimating Biomass and Biomass Change of Tropical Forests – A Primer*. Rome: Food and Agriculture Organization of the United Nations.

Canadell, J.G., Quéré, C., Raupach, M.R., *et al.*, 2007. Contributions to accelerating atmospheric CO_2 growth from economic activity, carbon intensity, and efficiency of natural sinks. Proceedings of the National Academy of Sciences 104, 18866–18870.

Chave, J., Andalo, C., Brown, S., *et al.*, 2005. Tree allometry and improved estimation of carbon stocks and balance in tropical forests. Oecologia 145, 87–99.

Dixon, R.K., Brown, S., Houghton, R.A., *et al.*, 1994. Carbon pools and flux of global forest ecosystems. Science 263, 185–190.

Drake, J.B., Knox, R.G., Dubayah, R.O., *et al.*, 2003. Above-ground biomass estimation in closed-canopy Neotropical forests using lidar remote sensing: Factors affecting the generality of relationships. Global Ecology and Biogeography 12, 147–159.

FAO, 2001. Global Forest Resources Assessment 2000. Rome: Food and Agriculture Organization of the United Nations.

Goodale, C.L., Apps, M.J., Birdsey, R.A., *et al.*, 2002. Forest carbon sinks in the Northern Hemisphere. Ecological Applications 12, 891–899.

Houghton, R.A., 2005. Aboveground forest biomass and the global carbon balance. Global Change Biology 11, 945–958.

Houghton, R.A., Lawrence, K.T., Hackler, J.L., Brown, S., 2001. The spatial distribution of forest biomass in the Brazilian Amazon: A comparison of estimates. Global Change Biology 7, 731–746.

Jenkins, J.C., Chojnacky, D.C., Heath, L.S., Birdsey, R.A., 2003. National-scale biomass estimators for United States tree species. Forest Science 49, 12–35.

Lewis, S.L., Phillips, O.L., Baker, T.R., Malhi, Y., Lloyd, J., 2006. Tropical forests and atmospheric carbon dioxide: Current conditions and future scenarios. In: Schellnhuber, H.J. (Ed.), Avoiding Dangerous Climate Change. Cambridge: Cambridge University Press, pp. 147–153.

Saugier, B., Roy, J., Mooney, H.A., 2001. Estimations of global terrestrial productivity: Converging toward a single number? In: Roy, J., Saugier, B., Mooney, H.A. (Eds.), Terrestrial Global Productivity. San Diego, CA: Academic Press, pp. 543–557.

Carrying Capacity

MA Hixon, Oregon State University, Corvallis, OR, USA

Carrying capacity is typically defined as the maximum population size that can be supported indefinitely by a given environment. The simplicity of this definition belies the complexity of the concept and its application. There are at least four closely related but nonetheless different uses of the term in basic ecology, and at least half a dozen additional definitions in applied ecology.

Basic Ecology

Carrying capacity is most often presented in ecology textbooks as the constant K in the logistic population growth equation, derived and named by Pierre Verhulst in 1838, and rediscovered and published independently by Raymond Pearl and Lowell Reed in 1920:

$$N_t = \frac{K}{1 + e^{a-rt}} \text{(integral form)}$$

$$\frac{dN}{dt} = rN\left(\frac{K-N}{K}\right) \text{(differential form)}$$

where N is the population size or density, r is the intrinsic rate of natural increase (i.e., the maximum per capita growth rate in the absence of competition), t is time, and a is a constant of integration defining the position of the curve relative to the origin. The expression in brackets in the differential form is the density-dependent unused growth potential, which approaches 1 at low values of N, where logistic growth approaches exponential growth, and equals 0 when $N = K$, where population growth ceases. That is, the unused growth potential lowers the effective value of r (i.e., the per capita birth rate minus the per capita death rate) until the per capita growth rate equals zero (i.e., births = deaths) at K. The result is a sigmoid population growth curve (**Fig. 1**). Despite its use in ecological models, including basic fisheries and wildlife yield models, the logistic equation is highly simplistic and much more of heuristic than practical value; very few populations undergo logistic growth. Nonetheless, ecological models often include K to impose an upper limit on the size of hypothetical populations, thereby enhancing mathematical stability.

Of historical interest is that neither Verhulst nor Pearl and Reed used 'carrying capacity' to describe what they called the maximum population, upper limit, or asymptote of the logistic curve. In reality, the term 'carrying capacity' first appeared in range management literature of the late 1890s, quite independent of the development of theoretical ecology (see below). Carrying capacity was not explicitly associated with K of the logistic model until Eugene Odum published his classic textbook *Fundamentals of Ecology* in 1953.

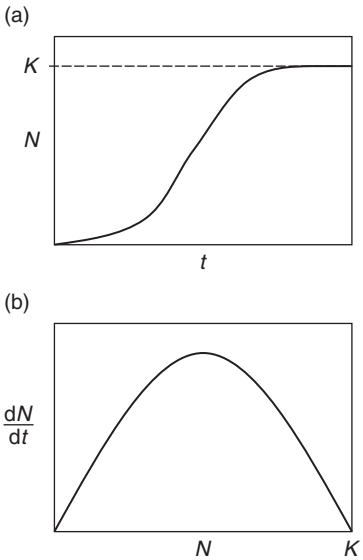


Fig. 1 The definition of carrying capacity most frequently used in basic ecology textbooks. (a) Logistic population growth model, showing how population size (N) eventually levels off at a fixed carrying capacity (K) through time (t). (b) Logistic population growth rate (dN/dt) as a function of population size. Note that the growth rate peaks at 0.5 K and equals zero at K.

 doi:10.1016/B978-0-444-63768-0.00468-6

The second use in basic ecology is broader than the logistic model and simply defines carrying capacity as the equilibrial population size or density where the birth rate equals the death rate due to directly density-dependent processes.

The third and even more general definition is that of a long-term average population size that is stable through time. In this case, the birth and death rates are not always equal, and there may be both immigration and emigration (unlike the logistic equation), yet despite population fluctuations, the long-term population trajectory through time has a slope of zero.

The fourth use is to define carrying capacity in terms of Justus Liebig's 1855 law of the minimum that population size is constrained by whatever resource is in the shortest supply. This concept is particularly difficult to apply to natural populations due to its simplifying assumptions of independent limiting factors and population size being directly proportional to whatever factor is most limiting. Moreover, unlike the other three definitions, the law of the minimum does not necessarily imply population regulation.

Note that none of these definitions from basic ecology explicitly acknowledges the fact that the population size of any species is affected by interactions with other species, including predators, parasites, diseases, competitors, mutualists, etc. Given that the biotic environment afforded by all other species in the ecosystem typically varies, as does the abiotic environment, the notion of carrying capacity as a fixed population size or density is highly unrealistic. Additionally, these definitions of carrying capacity ignore evolutionary change in species that may also affect population size within any particular environment.

Applied Ecology

The term carrying capacity may have first appeared in an 1898 publication by H. L. Bentley of the United States Department of Agriculture, with an original focus on maximizing production of domestic cattle on rangelands of the US southwest. The first use in wildlife management was apparently associated with classic studies of deer populations on the Kaibab Plateau in northern Arizona in the 1920s. The concept was popularized in wildlife ecology by Aldo Leopold and Paul Errington in the 1930s.

There have been four typical uses of carrying capacity in applied ecology, illustrated in Fig. 2: (1) the maximal steady-state number or biomass of animals an area can support in the absence of exploitation (the original use of carrying capacity, K); (2) the maximal sustainable yield (MSY) of biomass of animals an area can produce for exploitation, which equals $0.5K$ in the simplest form of the logistic model; (3) the maximal sustainable economic yield (MEY) of animals an area can produce for exploitation, which equals the maximum difference between yield value and cost of exploitation; and (4) the open-access equilibrium (OAE), where the value of the yield equals the cost of exploitation, which is the upper economic limit of exploitation in the absence of economic subsidies and restrictive management regulations. Note that open access, typical of historical marine fisheries, often leads to severe overexploitation because the population is reduced to sizes far below the other types of carrying capacity. Indeed, even the application of maximum sustainable yield in single-species fisheries management has proven elusive and often disastrous, as evidenced by the poor state of most marine fishery stocks so managed.

Two additional uses of carrying capacity in applied ecology focus on optimal stocking of rangeland with cattle, sheep, etc. The Society for Range Management defines the term as the maximum stocking rate possible which is consistent with maintaining or improving vegetation or related resources. A more general definition is the optimum stocking level to achieve specific objectives given specified management options. These practical definitions implicitly acknowledge that carrying capacity is not a constant, but rather is affected by a variety of environmental factors.

The elusive applied goal has been to determine number of animal-unit-days per unit area that produces a desired objective. A typical simplistic formulation follows:

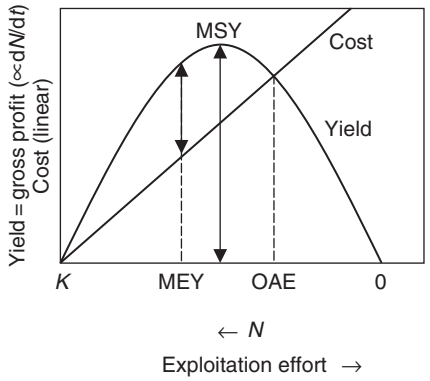


Fig. 2 Four definitions of carrying capacity used in applied ecology. Yield (or so-called surplus production, which directly translates to gross profit and which varies directly with the growth rate of the exploited population) initially increases and eventually decreases as exploitation effort increases, whereas the cost of exploitation presumably increases linearly with effort. The conventional carrying capacity (K) occurs in the absence of exploitation (i.e., zero effort). MSY occurs where total yield peaks (i.e., 0.5 K in the logistic model). MEY occurs where net profit (i.e., gross profit minus cost) is maximal (i.e., where the slopes of the cost and yield curves are identical). The OAE occurs where gross profit equals cost. Typically, as illustrated, $K > \text{MEY} > \text{MSY} > \text{OAE}$.

$$A = (B \times C)/D$$

where A is the number of animal-unit-days an area can support ($(\# \times d)$ per square kilometer), B is biomass of food in the area (kg km^{-2}), C is the metabolizable energy of that food (J kg^{-1}), and D is the metabolizable food energy required per animal unit per day (J/$(\# \times d)$). Obviously, such formulas ignore the reality of environmental variation, species interactions, etc.

A classic field study of wildlife carrying capacity was published by David Klein in 1968. In 1944, some two dozen reindeer were released on St. Matthew Island in the Bering Sea, where previously there had been none. Lichens were plentiful and the population increased at an average rate of 32% per year for the next 19 years, reaching a peak of about 6000 in 1963. During the severe winter of 1963–64, nearly all the animals died, leaving a wretched herd of 41 females and 1 male, all probably sterile. It was not so much the inclement weather that devastated the herd as it was a deficiency in food resources caused by overgrazing. After careful study, Klein concluded that 5 reindeer per square kilometer would have been the carrying capacity of an unspoiled St. Matthew Island. An animal census taken in 1957 gave 4 animals per square kilometer. A further 32% increase during the ensuing year brought the population to 5.3 per square kilometer, in excess of the predicted carrying capacity and a prelude to the eventual population crash.

Conclusions

Overall, the many and varied definitions of carrying capacity, typically stated in rather vague and ambiguous terms, render the concept to be most useful in theoretical ecology. Efforts to parametrize and measure carrying capacity in the field have proven problematic, such that the practical utility of the concept is questionable. This dilemma is especially true when considering the worldwide carrying capacity of humans, which seems better approached by the concept of ecological footprint. Nonetheless, the carrying capacity concept is clearly of heuristic value given the fundamental truth that no population can grow without limit, and especially given the fact that many human societies have behaved as if no limits exist.

See also: General Ecology: Abundance; Biomass

Further Reading

Bentley HL (1898) Cattle ranges of the Southwest: A history of the exhaustion of the pasturage and suggestions for its restoration. *US Department of Agriculture, Farmers' Bulletin Number 72*.
Caughley, G., 1976. Wildlife management and the dynamics of ungulate populations. Applied Biology 1, 183–246.
Caughley, G., 1979. What is this thing called carrying capacity? In: Boyce, M.S., Hayden-Wing, L.D. (Eds.), North American Elk: Ecology, Behavior, and Management. Laramie, WY: University of Wyoming Press, pp. 1–8.
Cohen, J.E., 1995. How Many People Can the Earth Support? New York: W. W. Norton and Company.
Dhondt, A.A., 1988. Carrying capacity: A confusing concept. Acta Oecologia (Oecologica Generalis) 9, 337–346.
Kingsland, S.E., 1995. Modeling Nature: Episodes in the History of Population Ecology, second ed. Chicago: University of Chicago Press.
Klein, D.R., 1968. The introduction, increase, and crash of reindeer on St. Matthew Island. Journal of Wildlife Management 32, 350–367.
Macnab, J., 1985. Carrying capacity and related slippery shibboleths. Wildlife Society Bulletin 13, 403–410.
Mautz, W.M., 1978. Nutrition and carrying capacity. In: Schmidt, J.L., Gilbert, D.L. (Eds.), Big Game of North America: Ecology and Management. Harrisburg, PA: Stackpole Books, pp. 321–348.
McLeod, S.R., 1997. Is the concept of carrying capacity useful in variable environments? Oikos 79, 529–542.
Myers, R.A., MacKenzie, B.R., Bowen, K.G., Barrowman, N.J., 2001. What is the carrying capacity for fish in the ocean? A meta-analysis of population dynamics of North Atlantic cod. Canadian Journal of Fisheries and Aquatic Sciences 58, 1464–1476.
Pearl, R., Reed, L.J., 1920. On the rate of growth of the population of the United States since 1790 and its mathematical representation. Proceedings of the National Academy of Sciences USA 6, 275–288.
Scarnecchia, D.L., 1990. Concepts of carrying capacity and substitution ratios: A systems viewpoint. Journal of Range Management 43, 553–555.
Sharkey, M.I., 1970. The carrying capacity of natural and improved land in different climatic zones. Mammalia 34, 564–572.
Verhulst, P.F., 1838. Notice sur la loi que la population suit dans son accroissement. Correspondances Mathématiques et Physiques 10, 113–121.
Wackernagel, M., Rees, W.E., 1996. Our Ecological Footprint: Reducing Human Impact on the Earth. Gabriola Island, BC: New Society Publishers.
Walters, C.J., Christensen, V., Martell, S.J., Kitchell, J.F., 2005. Possible ecosystem impacts of applying MSY policies from single-species assessments. ICES Journal of Marine Science 62, 558–568.
Young, C.C., 1998. Defining the range: The development of carrying capacity in management practice. Journal of the History of Biology 31, 61–83.

Communication☆

Peter K McGregor, Cornwall College, Newquay, United Kingdom

Introduction

Communication is a social behavior that mediates fundamental aspects of animals' lives. It is important during reproduction—playing a role in attracting another individual of the same species and coordinating mating, and extending to aspects of parental care. It is also important in many aspects of survival—from being alerted to the presence of predators by warning calls, through signaling prey defenses, to indicating food sources and defending them. Communication is also often conspicuous to the extent of being spectacular; examples are choruses of songbirds, cicadas, and frogs, and the coordinated displays of fireflies.

Communication is generally thought of as a characteristic of animals, but there are instances in which plants share features of animal communication. For example, the response of some plants to airborne chemicals indicating that a neighbor has been attacked has features in common with the warning or alarm calls of animals. However, as most information on communication comes from the animal kingdom, this article deals with animal communication. The article will begin by introducing key concepts in communication; it will then proceed to discuss the influence of physical and social environments on communication behavior, and it will end by looking at current issues linking behavior and ecology and will touch upon a role in applied ecology.

What Is Communication?

As we shall see below, defining communication is not straightforward and has been the subject of considerable debate. However, there are three readily identifiable components of communication: (1) the signal, (2) the signaler, and (3) the receiver.

Signal

The signal is the information carrier and is often categorized by the sense used to detect the signal (signal modality, see below). For example, a frog's call is an acoustic signal and the flash of a firefly is a visual signal.

Signaler

The signaler is the animal producing a signal, for example, a calling frog, or flashing firefly. Signalers are also referred to as senders, generators, and actors.

Receiver

The receiver is any animal picking up the signal, for example, another frog, or firefly, and they do not have to be the same species as the signaler. Receivers are also referred to as detectors or reactors.

The Relationship Between the Three Components

The relationship between these components is usually portrayed as the signaler produces the signal that is received by the receiver. This is the simplest possible form of communication, but this simple form probably only occurs rarely in natural communication systems. An important complexity of natural systems is that several individuals can be involved in communication, resulting in a communication network (Fig. 1), rather than a dyad of one signaler and one receiver. A second aspect of the complexity of natural systems is that an individual can change dynamically between the roles of signaler and receiver, and may be capable of simultaneous signaling and reception.

☆*Change History*: November 2015. PK McGregor updated this article. A new paragraph on anthropogenic effects was added to the "Towards a definition" section and the "Further Reading section was updated".

This is an update of P.K. McGregor, Communication, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 683–689.

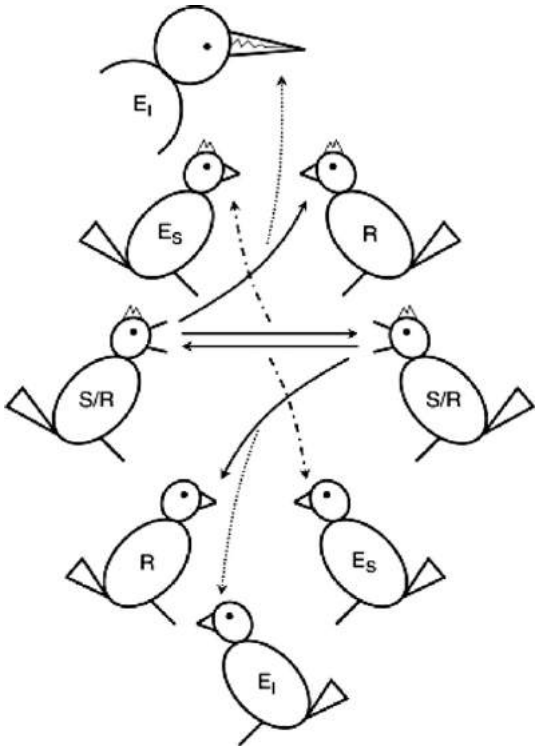


Fig. 1 A schematic representation of some possible patterns of signaling and receiving in a communication network. In the center of the figure, two males (distinguished from females by crest and longer tails) are involved in a signaling interaction (paired *arrowed lines*), that is, have signaling and receiving roles (S/R). The male on the right is also signaling to a female receiver (R, toward bottom left of figure) and another female (center bottom) is intercepting (*dotted line*) this signal (interceptive eavesdropping, E_I). A third female (toward bottom right) is social eavesdropping (E_S) on the interaction between the central males (*dot and dash line*), as is a male (toward top left). The central male on the left is also signaling to a male receiver (toward top right) and a predatory species (top left) is intercepting the signal.

Terms Commonly Used in Communication

The technical terms used when discussing the ecology of communication often originated in different subject areas, such as physics and applied mathematics, and the original meaning of some terms has altered. The meanings of terms listed below reflect their current use.

Transmission Medium

The transmission medium conveys the signal from the signaler to the receiver and, depending on the specific signal, can be air, water, or solid objects. For example, the transmission medium for frog calls and firefly flashes is air, whereas for whale song it is water.

Signal Modality

The signal modality (or channel) refers to the sense used to detect the signal. We are most familiar with senses of hearing, vision, and smell/taste. Three less common modalities use the vibration detection sense, the sense of touch, and the electrical sense. Some signals are referred to as multimodal because more than one sense is required to receive them.

Visual signals most commonly utilize reflected light, as few species are capable of generating light for signaling. Movement is often a feature of visual signals, for example, the claw waving display of fiddler crabs. Acoustic signals encode information in pressure differences of the transmission medium (i.e., air or water) and at extremely close range by displacement of the medium itself. Chemical signals use the sense of smell/taste and differ from other modalities in that they can persist in the absence of the signaler (e.g., scent-marked vegetation).

Seismic (vibration) signals transmit information through the substrate; an example is the foot-stamping signal used by rabbits. Tactile (touch) signals require the signaler and receiver to be in physical contact. Electrical signals are used by two taxonomic groups of freshwater fishes with a well-developed ability to detect electric fields and also organs specialized to generate electricity.

Each of these six modalities differs in transmission speed, persistence, inherent directionality, and the effect obstacles have on them. These differences strongly influence the form of signals used in different contexts and also how the environment affects signals during transmission.

We tend to consider humans good receivers of visual and acoustic signals; however, some animal signals in these modalities are outside of our perceptual range. Acoustic examples include the infrasonic rumbles of elephants and ultrasonic calls of bats; visual examples include the ultraviolet coloration of butterflies and the flashes of infrared light produced by some deep-sea fish.

Information

Information is a fundamental concept to communication because all communication involves information (however, not everything involving information is communication). Confusingly, information is used in two senses in the communication literature. In the everyday sense it means knowledge, but in a technical sense (derived from information theory) it means the reduction in a receiver's uncertainty following reception of a signal. (Uncertainty, H, is calculated by the Shannon–Weaver (or Shannon–Weiner) function: $H = -\sum pi \log_2 pi$ where i is the number of possible behavior patterns and pi is the probability that the ith will occur.) Often authors fail to mention which of the two meanings of information they are using. Where they do, information in the everyday sense is sometimes referred to as semantic information and information meaning reduction in uncertainty is referred to as statistical information. Information will be used in the everyday sense in this article.

Semiotics

Semiotics is the study of signs and symbols, with communication often considered under the heading of biosemiotics; it is dealt with elsewhere in this encyclopedia.

Toward a Definition of Communication

Communication is surprisingly hard to define. Part of the difficulty stems from the wide variety of contexts in which animals communicate and to some extent from the diversity of signals employed. Such diversity means that there can be important exceptions to each of the many definitions of communication. For example, one common definition is mutually beneficial information transfer; however, there are several instances where information transfer is not mutually beneficial (e.g., mimicry by edible species of inedible species, which is a deception to the receiver's detriment). Another aspect of the difficulty of definition is that communication is only a subset of information transfer; a mouse moving through fallen leaves and a tree bending in the breeze both transmit information (on location and wind direction, respectively) but would not be thought of as communication.

One approach is to focus on the role of signals in communication, that is, communication can be defined as behavior involving signals, where signals are defined as adaptations to transmit information. Therefore the call of a frog is an adaptation, for example, to attract females for mating, and likely contains information on species identity in order to function effectively. The rustling mouse generates information on location as an incidental by-product of movement; the rustling sound is probably maladaptive because it is likely to attract a predator.

Communication and the Environment

Communication is only one behavior in an animal's repertoire and whether or not to communicate at any instant is influenced by many factors outside the scope of this article. However, once the decision to communicate has been made, we would expect communication behavior to be as effective as possible. Both the physical environment and the social environment can have major influences on the effectiveness of communication; these are manifested as decisions concerning when and where a signal is broadcast and in the structure of the signal. A complication is that both types of environment change over a range of timescales and often vary cyclically. For example, vegetation changes over an annual cycle in temperate terrestrial habitats whereas the activities of other signalers can change from signal to signal. A limitation on our understanding of such matters is that signaling is a much more conspicuous behavior than receiving, and therefore most of the information on when and where to communicate relates to signaling behavior.

A further complication is the effect of anthropogenic noise; such noise is loud and virtually ubiquitous in all environments as a consequence of human activity, yet it is relatively recent on an evolutionary timescale. There are well-documented effects of anthropogenic noise on signal structure and communication behavior in both terrestrial and aquatic habitats. For example, urban bird song is generally high pitched and loud to reduce the interfering effects of traffic noise and whales show similar effects in the presence of ship sonar. The fitness consequences of such noise-induced changes have yet to be established.

In this section, we will consider the physical environment first because generally its effects are less complex and change less frequently than the effects of the social environment. However, it is important to remember that both physical and social environment contribute to the variation in communication behavior and signal structure that we observe.

Effects of the Physical Environment on Signal Form

All habitats modify signals during transmission. Signals are both attenuated and degraded, that is, signal energy is absorbed and signal details are distorted by the environment. Often degradation will limit the effective range of a signal before attenuation—it is a common personal experience to know that someone at a distance has said something, but to be unable to understand the words; the signal has been detected but the detailed information it contains cannot be discerned. Habitats can be thought of as imposing selection pressures on the form of signals and as a result most signals are adapted to some degree to their transmission environment. For example, birds found in open habitats (e.g., grassland) tend to sing songs with rapid temporal patterning (the songs contain more trills than whistles) and with a wide frequency range. By contrast, woodland bird songs tend to have more low frequency whistles than high frequency trills. This difference contributes to minimizing the effect of the habitat on signals during transmission; trills suffer less degradation from the irregular changes in air temperature and wind speed characteristic of open habitat, while low frequency whistles suffer less degradation induced by reflections from leaves and trunks in woodland.

Transmission conditions also change within a habitat on daily and seasonal timescales; for example, the effects on visual and acoustic signals of a deciduous forest in winter are very different from those in the same forest in summer. Such changes often affect to a greater extent when and where signals are produced than they affect signal structure.

Effects of the Physical Environment on When and Where to Communicate

The most effective time to signal is often related to cyclical changes in the physical environment. An example on a seasonal timescale is communication underlying breeding behavior; breeding is seasonal (e.g., spring in temperate zones and the rainy season in equatorial regions) and so are the long-range mate attraction and territory defense signals such as birdsong and the choruses of frogs and insects. Cycles occurring over shorter timescales are also accompanied by appropriate signaling behavior. Circalunar and tidal cycles are common in marine or intertidal environments and animals such as fiddler crabs produce their claw-waving visual signal at low tide. Daily (circadian) cycles give rise to choruses of bird song at dawn and of frogs and fireflies at dusk. It seems reasonable to expect cycles of receiver activity to coincide with those of signalers. However, it is possible that the amount of information available during periods of very high signaling activity (e.g., intense chorusing) could exceed receivers' processing abilities, with the result that receivers may avoid such periods.

The most effective place to signal from is affected by several factors including transmission conditions and the area to be encompassed by the signal. Environmental factors affecting signal transmission often vary from location to location within habitats. For example, studies that broadcast bird song and record it at different locations to investigate location effects on transmission have shown that the best perch height for maximizing song transmission (i.e., where signalers should be found) is lower than that which favors sound reception (where receivers should be found). As many visual signals rely on reflected light (see above), it is not surprising that signaler location is strongly affected by the incident light regimes in terrestrial and aquatic environments. Some species take things further and alter the properties of the signaling location to enhance signal transmission. For example, mole crickets excavate a calling burrow in an exponential horn shape to amplify their call and some tree crickets create acoustic baffles from leaves to modify the spreading pattern of their call. Signaling from a single location can be less effective than signaling from several locations, often because the signal does not encompass the required area. When several signaling locations are used, the signaler faces a series of decisions including how to apportion effort between locations and the interval before returning to each location. Often solutions to these issues based on a priori considerations will be overridden by the activity of adjacent signalers and receivers, that is, the social environment can modify and often override effects of the physical environment.

Effects of the Social Environment on Signal Form

The most noticeable effects of the social environment on signal form occur when other signalers and receivers have adverse influences that cannot be avoided by ceasing signaling or signaling from another location. One example involves the species specificity or species distinctiveness of signals and is a form of character displacement. If the signal characteristics of sympatric species overlap then we would expect selection for the signals to diverge because communicating with the appropriate species is important in many contexts. Character displacement is commonly assumed to be the cause of a lack of overlap of signal characteristics between species in sympatry when compared with the same species in allopatry.

A common influence of the social environment is an aspect of communication networks, that is, the presence of receivers in addition to the intended receivers (**Fig. 1**). These additional receivers can intercept signals intended for others, usually at a cost to the signaler. Often such interceptive eavesdroppers are predators or parasites that use broadcast signals to locate prey and hosts. Examples are predatory bats that feed on calling male frogs and flies that use calls to locate frog hosts for their parasitic larvae. Signalers rarely call in the presence of such eavesdroppers (i.e., eavesdroppers affect decisions on when and where to communicate; see below), but there are examples in which the form of the signal reduces the opportunity for interceptive eavesdropping. The clearest examples are signals that warn of predators because such warning calls must be given in a predator's presence and could be intercepted by the predator. For example, the warning call (seeet) given by some northern temperate species to an aerial predator has a frequency of about 8 kHz, to which the calling species is more sensitive than the eavesdropping predator.

Interceptive eavesdroppers may also be the same species as the signaler and therefore potential competitors and rivals of the signaler. For example, a male fiddler crab reacts by producing courtship waving signals when he intercepts the courtship signals produced by a male neighbor in response to the female, even though he cannot see an approaching female, and in doing so he becomes a competing suitor. During close-range, high-intensity courtship or aggressive interactions, some song birds change from a far-carrying advertising song to a form of song that carries less far as a consequence of its lower amplitude and greater high-frequency content. Such a change in signal form should restrict the potential for eavesdropping at times when the signaler can suffer high costs.

Effects of the Social Environment on When and Where to Communicate

The most readily observed effect of the social environment on communication is the avoidance of adverse influences (e.g., risk of predation, interference from other signalers) by ceasing signaling or signaling from another location. Interference from other signalers of the same species is particularly severe in high-density communication networks (e.g., breeding colonies, chorus aggregations) and can result in changes in communication behavior. For example, chorusing male frogs have less success attracting a female when another calling male is close by. In response to such interference males may vary the timing of their calls both in relation to close neighbors and the overall synchrony of the chorus. As males may fight to displace others calling from close by and as the larger male usually wins, small males often move away when a large male begins calling near them. It is likely that similar considerations apply to receivers.

Signaling interactions between others are one aspect of the social environment in which eavesdropping would be favored because information on relative aspects of the signalers is readily available from such interactions. In such social eavesdropping (cf. interceptive eavesdropping, above), the eavesdropper does not take part in the signaling interaction (Fig. 1). Social eavesdropping by both sexes has been demonstrated in laboratory experiments on fish using visual signals and in field experiments with territorial songbirds: males and females eavesdropped on aggressive signaling interactions between other males, and females eavesdropped on male–female courtship interactions. There are indications that receivers position themselves in space and time to facilitate social eavesdropping. The specific circumstances of an interaction and the individuals involved will determine whether the signalers involved will gain or lose from the presence of social eavesdroppers (e.g., the winner of an aggressive interaction may gain while the loser may not) and therefore how the eavesdroppers will affect when and where the interaction occurs.

Origin of Communication Behavior

The evolutionary origin of communication behavior is usually considered solely as a question of how signals originate, that is, how they evolved to become specialized to carry information. Two evolutionary origins for signals have been suggested; they differ in "who benefits from the information in the signal." If both signaler and receiver benefit from the communication, it is likely that signals originate as a consequence of selection for increasing efficiency of information transfer. This is usually termed ritualization, and it generates a signal by the simplification, exaggeration, repetition, and increased stereotypy of the signal precursor. Signal precursors are behaviors that incidentally contain some relevant information such as intention movements, displacement activities, and autonomic responses. A second evolutionary route for the origin of signals is usually referred to as sensory exploitation. Signals are selected to exploit "preexisting sensory biases" built into receivers. For example, male spiders use signals for mate attraction that stimulate females' prey-detection receptors.

If the evolutionary origin of communication is considered in a wider network context, a third possibility emerges. Information networks include communication networks (in which information transfer occurs via signals as described above), but also networks of information based on potential signal precursors. Semiochemicals (e.g., alarm pheromones) are examples of such potential signal precursors and receivers gathering information from them are said to be spying. Chemical signals could evolve from semiochemicals through spying (as production of the semiochemical by the signaler becomes specialized). In an analogous manner, communication networks could evolve from spying networks, reversing the presumed order of communication appearing before eavesdropping.

Current Issues and an Application

Honesty and Handicaps in Communication

Much of the recent research on communication in behavioral ecology has addressed the issue of whether signals are honest in the context of mate choice, resource defense, and predator–prey interactions. There are examples of communication involving deceit (such as the mimicry example mentioned above), but the wider issue is the selection pressures signalers and receivers exert on one another. When signalers and receivers have conflicting interests, the result is likely to be a coevolutionary arms race. Two such instances are potential prey signaling to deter predators from attacking and females using male signals to choose the best mate. Selection should favor signals that best achieve signalers' interests, even at the expense of receivers; therefore signals are unlikely to be honest. However, selection should also favor receivers adept at gathering information in their best interests. Such a co-evolutionary arms race seems likely to result in signals that are "honest on average."

One way it is thought that signal honesty can be ensured is if signals are costly. This idea is usually referred to as the handicap principle, where handicap means the conspicuous cost of the signal, that is, resources used for the signal that might otherwise have increased signaler fitness. A hypothetical example would be the energy used to produce a call indicating a larger than actual body

size—energy that could otherwise have been used to grow a larger body. The debate about the role of handicaps in communications is outside the scope of this article.

Persistence and Payoff Asymmetries

The persistence of communication even when signalers differ from receivers in benefits (i.e., there is a large payoff asymmetry) is a topic of current interest in behavioral ecology. Two factors that allow communication to persist despite payoff asymmetries are the rarity of the signal and the genetic relatedness of signaler and receiver.

Extreme payoff asymmetries are found when the signal is rare and communication occurs between species. For example, edible mimics are thought to gain protection from predators by resembling inedible or dangerous prey (models) only when mimics are rare relative to models.

Extremes of relatedness often characterize striking examples of cooperation and its associated signaling. For example, individuals of the same species sometimes cooperate in locating distant food sources through signals such as the dance language of honeybees and the pheromone trails of ants. Often such social insects are very close genetic relatives. Other examples of cooperation include foregoing reproduction to help a close relative reproduce (e.g., social insects, naked mole rats). The details of the signaling underlying such cooperation are still under study, but are likely to be important to the functioning of cooperation. When the signaler is a different species from the receiver, the lack of genetic relatedness may be the factor favoring cooperation and the associated signaling. Examples include the cleaner fish found on reefs that closely approach other, generally much larger species (clients) quite capable of eating the cleaner. However, the cleaner only approaches a client closely after an exchange of signals to indicate that the larger species is a client and not a predator. Recent research suggests that the signals also allow the client to minimize the risk that a cleaner will bite rather than clean.

Conservation and Communication

The long-range advertising signals in most signal modalities contain information on species identity. Such information is valuable to conservation efforts because such signals and signaling activity allow species to be identified and their abundance to be estimated. This application is particularly well developed for birds, allowing species to be identified, and counted even at night or when they are hidden by dense vegetation. If birds can be individually identified, then detailed information on survival, habitat use, and immigration rates can be collected. Most bird vocalizations are naturally individually distinctive and techniques have been developed to gather such information as a noninvasive alternative to catching and marking.

Summary

Communication is behavior involving signals. Signals are adaptations that transmit information and in this way are distinguishable from other aspects of the environment that incidentally contain information. Much communication occurs in a social environment of a network of several signalers and receivers. Signals are diverse, with several signal modalities (e.g., vision, hearing, smell) used for communication, some of which cannot be detected unaided by humans (e.g., ultraviolet color patterns of butterflies and the ultrasonic calls of bats).

The effectiveness of communication depends on conditions of both the physical and the social environment. These affect when and where communication occurs and the structure of the signal. The effects can occur over a range of timescales, from seconds to years, and often vary cyclically.

Signals can evolve through ritualization (of behaviors that are precursors of signals), through the use of preexisting sensory biases of receivers, and through information gathering in networks.

Communication has featured in two areas of current interest in behavioral ecology: whether signals are honest and how communication persists when there are large differences in the benefits between signalers and receivers. Animal signals also have a role as a census and monitoring tool in conservation.

Further Reading

Bradbury, J.W., Vehrencamp, S.L., 2011. The principles of animal communication, 2nd edn Sunderland, MA: Sinauer.
Brum, H. (Ed.), 2013. Animal communication and noise. Berlin, Heidelberg: Springer-Verlag.
Butlin, R.K., Guilford, T., Krebs, J.R., 1993. The evolution and design of animal signalling systems. Philosophical Transactions of the Royal Society of London, Series B 340, 161–225.
Catchpole, C.K., Slater, P.J.B., 2008. Bird song: biological themes and variations, 2nd edn. Cambridge: Cambridge University Press.
Cocroft, R.B., Gogala, M., Hill, P.S.M., Wessel, A., 2014. Studying vibrational communication. Berlin, Heidelberg: Springer-Verlag.
Dawkins, M.S., 1995. Unravelling animal behavior, 2nd edn. Harlow, UK: Longman Group.
Espmark, Y., Amundsen, T., Rosenqvist, G. (Eds.), 2001. Animal signals. Signalling and signal design in animal communication. Trondheim: Tapir Academic Press.
Gerhardt, H.C., Huber, F., 2002. Acoustic communication in insects and anurans: common problems and diverse solutions. Chicago: Chicago University Press.
Greenfield, M.D., 2002. Signalers and receivers: mechanisms and evolution of arthropod communication. Oxford: Oxford University Press.

Hedwig, B. (Ed.), 2014. Insect hearing and acoustic communication. Berlin, Heidelberg: Springer-Verlag.
Johnstone, R.A., 1997. The evolution of animal signals. In: Krebs, J.R., Davies, N.B. (Eds.), Behavioural ecology. Oxford: Blackwell Scientific Publications, pp. 155–178.
Kroodsma, D.E., Miller, E.H. (Eds.), 1996. Ecology and evolution of acoustic communication in birds. Ithaca, NY: Cornell University Press.
Maynard-Smith, J., Harper, D., 2003. Animal signals. Oxford: Oxford University Press.
McGregor, P.K. (Ed.), 1992. Playback and studies of animal communication. New York: Plenum.
McGregor, P.K. (Ed.), 2005. Animal communication networks. Cambridge: Cambridge University Press.
Searcy, W.A., Nowicki, S., 2005. The evolution of animal communication: reliability and deception in signaling systems. Princeton, NJ: Princeton University Press.
Stegmann, U. (Ed.), 2013. Animal communication theory: information and influence. Cambridge: Cambridge University Press.
Wyatt, T.D., 2014. Pheromones and animal behavior: chemical signals and signatures, 2nd edn. Cambridge: Cambridge University Press.

Community

AJ Underwood, University of Sydney, NSW, Australia

The Concept of a Community

Despite the widespread use of the term 'community', there has been remarkably little detailed ecological discussion of what it means. This remains a strange omission, given that communities are supposedly the actual units of study for many ecologists. There seem to be two, very different ways of viewing a community. One is that species exist in integrated communities that have persistent features through time (e.g., composition of species or structure of food webs) and are repeated in different places. Thus, the ecological community of species is organized, structured, and integrated. The species in the community are interdependently interactive and, often, it is presumed that the interactions are, at least in part, responsible for maintaining the entity.

The alternative view is much less about the structure of a community. It holds that communities are simply a human invention. The term community is used to describe the collection of organisms that are found in the same place at the same time. They may (or may not) exhibit interdependencies. They may or may not interact. They coexist because they have similar physiological responses to physical components of the environment and/or they have similar needs for resources of food and shelter. Alternatively, some of them may be present because they need others as prey (or as some other resource).

There is no doubt that these two views are very different and lead to very different forms of ecological study. If the former view is correct, the community is, indeed, a valid and necessary object of study. It will only be possible to understand the ecology of the component species, if their interactions and interdependencies are understood. There could be no argument that the populations (and, in fact, the individuals in the populations) that make up the community are not the only valid units to investigate and explain. The community, itself, has ecological features and properties.

In contrast, if the latter, more descriptive and less structured view is correct, then the community *per se* has no structured existence and is simply a collection of entities – the populations of the species present. These populations and the interactions among them are the important units to study.

Throughout the history of ecology, there have been periods when one or other view has been (or has been considered to be) dominant. There does not ever seem to be a time when one definition looks likely to prevail over the other. As a result, there is no particularly successful or acceptable definition of community that will satisfy all – or even most – ecologists. The most extreme views are a very tightly coevolved and integrated community versus a random collection of species that happen to co-occur. Neither view is particularly likely to be sufficient or even correct.

The reality is probably somewhere in between. Some components of a community may, in fact, be highly interdependent and coevolved. Others are not. At the same time, coevolved traits can almost certainly be caused or maintained even in the loosest co-occurring set of species. For example, it is not a necessary condition that coevolved features of inferior competitors for food must be caused by consistent interaction with the same dominantly competitive species. Such features could well result from consistent competition for resources by a range of different and changing superior competitors. The intensity of and responses to competition can be ongoing and progressive, even though the superior competitors change from one generation to the next (or faster) for inferior competitors. Competition can be extensive, but caused by different species in varying mixes from one part of a habitat to another.

To avoid some (but sadly not all) of the confusion associated with the term community, the two ecological approaches will here be referred to using the term 'community' for tightly knit, consistent sets of species. 'Assemblage' will be used for the more loosely associated set of co-occurring species, where the whole set of species is not a repeatable, identifiable set.

Community as a Superorganism

One of the consequences of considering a community to be structured has been the notion that the set of species making up the community were tightly interacting and coevolving. Plant ecologists had identified numerous different types of associations of various species of plants. Some of these associations, the communities, came to be considered as a completely integrated unit – a kind of superorganism.

In its most extreme form, the description of communities began to prevail that coexisting species occurred together because they have similar needs for resources and similar responses to environmental stresses. The community developed because of deterministic pathways of succession, leading to a stable climax. The analogy was that a community develops by processes like ontogenetic development of an individual animal or plant. It is predetermined from conception, that is, from the early establishment of species in the habitat in which the community develops. Interactions affecting each component species are evolved responses to the presence, abundances, and activities of the other species in the community. Such interactions lead to homeostatic development of an equilibrium in the climax community and therefore to predictable patterns of structure.

 doi:10.1016/B978-0-444-63768-0.00473-X

Another property of superorganismic communities is that wherever they come into contact with each other, there is only very limited possibility of overlap. Thus, where two plant communities, for example, a forest and an open grassland, exist side by side, their boundary is a fairly narrow ecotone. Sometimes, ecotones contain species not found in either adjacent community, but there cannot be broad existence of mixtures of species from the two communities because the species are closely adapted to the community where they 'belong'. Species from one community should therefore be unable to thrive in the absence of some members of that community or in the presence of species that do not belong with them.

Assemblages

The concept of communities as superorganisms received considerable criticism in the US and Russia. Much of this criticism was apparently ignored or otherwise had little impact. There were compelling reasons to maintain a body of theory about integrated communities. For example, without a widespread use of communities as objects to study, there was no compelling need for theories about the ecology of communities (such as some of the theories about deterministic processes and patterns in natural succession). Ecologists might apparently have had less work to do.

There were, nevertheless, important features of developing ecology that made it inevitable that concepts about communities would be questioned. One of these was undoubtedly that animal ecologists in the early decades of the 1900s were finding some difficulties not shared by plant ecologists. Animals, at least many of them, were actually more difficult to observe (they ran or flew away, they actively hid when disturbed, etc.). As a result, the passive, motionless structures of sets of plants (i.e., communities) were not so evident for sets of animals. Seeing apparent patterns was not such a preoccupation for ecologists studying animals.

A second area of developing ecology had this problem in a more severe form. Early marine ecologists could not actually see the animals under the water and were, willy-nilly, forced to examine them in dredges and grabs that would bring pieces of substratum, with attached or embedded animals to the surface. Any structure or relationship of species into groupings had to be inferred from these samples. Furthermore, early marine biologists were compelled to recognize that their units of study were not natural (nor even contrived descriptions of) communities. Instead, they were studying species in artificial and arbitrarily designed sampling units – dredges and grabs. This did not stop them describing an equilibrial type of community – a biocenosis. This term was used to describe the collection of organisms found together in an oyster bed. It was, however, important that the unit of study was clearly a sample or representation of reality and not a naturally defined collection of species. It took some years before the sample unit became the standard unit of terrestrial ecological study.

Inevitably, quantitative sampling pervaded more and more of terrestrial (and plant) ecology. Objectively sampled data along environmental gradients provided data that did not fit easily into a structure of defined plant communities. This was often resolved by designating the anomalous samples as transitional or atypical or as identifying 'mixed stands' of plants. The latter is a somewhat obscure way of dealing with anomalous results, given that the tightly integrated community cannot realistically be considered to be mixed with another community.

Eventually, sufficient evidence accumulated to indicate that evidence for communities was outweighed by contrary evidence. Whittaker's sampling of vegetation up the environmental gradient of height in the Great Smoky Mountains is a deserved classical study to demonstrate this. Instead of accepting the evidence of communities of species of plants, as was the widely held view, he sampled at intervals up the mountains, to determine which species were present and their vertical distribution. As a result, he found that most, if not all, species were distributed independently up the gradient. There were not tightly knit, coherent communities.

The result was definitely not the anticipated outcome. Communities up a mountainside should have consisted of a set of species which have (or, of which, most have) upper limits to distribution at about the same height. They should have reasonably coincident lower limits to distribution. Otherwise, there can be no organized community of co-occurring species. Instead, Whittaker in 1956 found that upper boundaries and lower boundaries were scattered, without any obvious order, up the mountain (see later, **Fig. 1**).

The interpretation of this result was difficult because the concept of communities as integrated units was so widespread that Whittaker had to resurrect the term 'individualistic concept' to describe the independent patterns of distribution of species.

Many subsequent analyses of assemblages of species, particularly across gradients, have described species being distributed according to their physiological tolerances. The distributions are, however, modified (usually, their extent or range is reduced) by interactions with other species (such as pathogens, competitors, and predators). Assemblages of such species intergrade along a gradient and do not display the abrupt, ecotonal changes supposedly identifying communities. Such analyses do not require *ad hoc* argumentation about the intergrading or missing or partial nature of communities.

So during the 1960s, work casting doubt on the superorganismic communities that had been popular with plant ecologists was being published. It is ironic that, at the same time, animal ecologists were developing new rationales for integrating the ecologies of species into integrated communities. A very sophisticated theory was beginning to be developed about coevolution and diversity of species in communities. 'Assembly rules' that were supposed to maintain the structure of communities began to appear.

A new unit of study – the ecosystem – was popularized. This was based not just on a set of integrated species being a community, but on a more holistic ecosystem, which included the biota and all the physical and chemical components of habitat within which they interact.

The definition of an ecosystem can be difficult if not impossible. It is supposedly a community, but not just consisting of an interactive, integrated set of species. Instead, an ecosystem is a set of species interacting intimately with all the inputs and outputs of resources and energy. Given that, for most assemblages, vital resources of food are autotrophs which require sunlight, the connection of use of solar energy must extend to all organisms on the planet that are dependent on solar energy. Thus, in an attempt to use a rational definition, all of life on Earth, except for chemotrophs (e.g., organisms around deep-sea vents of hot water), should be considered to be in a single ecosystem. In its most extreme form, the world as a single ecosystem has achieved considerable impetus from the well-publicized concept of Gaia. This is the Earth conceived to be responding to change as though it were a single entity. Despite this concept, it is, however, not clear that the whole planet has yet been demonstrated to be a necessary, nor useful unit of study for ecologists.

Why Are Communities Found in Nature?

One of the reasons for there being so many observations of communities in nature seems to be intimately related to receive wisdom (paradigms) about units of study. If an ecologist strongly believes that plants and/or animals are in communities, with few species overlapping from one community to another and with fairly abrupt boundaries, there can be problems for designing objective sampling. Suppose it is thought, or it has been described, that there are, say, three communities along some environmental gradient. Each of these communities is associated with particular dominant or abundant species (A and B; C; D and E; respectively). A study in some new area may well involve searching along the gradient until A and B are numerous together. This purportedly identifies the presence of the first community. The second community is identified to be where C are numerous, and the third where D and E are numerous together. Now, sampling around the three areas which have been already designated to be containing different communities will, inevitably, reveal great differences in the species making up the communities and in the densities or relative abundances of species in each community. Similarly, if two areas are defined to contain the same community, because they each have an abundance of the species that dominate that community, sampling in the two areas will inevitably reveal quite a lot of similarity.

Instead, sampling along the gradient should be at random or regular intervals, depending on the nature of the study and the hypotheses being tested. Then, at least, there is no prior dogma defining where to sample. Data about the identities or abundances of species at each point sampled can then be analyzed to test hypotheses.

Sampling without unquestioned acceptance of the existence of communities is, however, also fraught with difficulties, as illustrated in **Fig. 1**. In the first case (**Fig. 1a**), there are three very well-defined communities along an environmental gradient. Very few species transcend these structures. Sampling at the sites indicated will reveal different sets of species, with very little overlap. The data would reveal that communities exist because sample points contain quite different sets of species.

Unfortunately, if the species are entirely independently distributed along the gradient, this type of sampling will not be useful. Sampling at intervals along a gradient (as in **Fig. 1b**), where species have independent distributions, would also reveal apparent communities. There would be different sets of species at each place sampled. The data would be very similar whether or not communities really exist.

Independence of distributions is essentially the null hypothesis against which to test the hypothesis that species are clumped as communities, with nonrandom commonality of boundaries. There is, of course, a third alternative that species have boundaries that are more regularly spaced than is the case for randomly distributed species. This is, however, irrelevant to any attempt to identify the existence of communities.

This has led to the realization that identification of nonrandom patterns in arrangements of species along environmental gradients requires data about the actual distributions, rather than sampling of organisms at intervals across the habitat. One methodology uses contiguous quadrats across the gradient. The numbers of species that have a boundary in each quadrat are recorded (as in **Figs. 1c** and **1d**). So, transects of quadrats up a mountainside would be examined. As one progresses from the bottom upward, the first occurrence of a member of a particular species indicates its lower boundary. The number of these lower boundaries is recorded for each quadrat. Similarly, the quadrat in which the last member of a species is encountered denotes its upper boundary.

This is illustrated for the two distributions, each with 18 species, in **Fig. 1**. For the species in **Fig. 1a**, the numbers of lower and upper boundaries per quadrat are shown (**Fig. 1c**). The same data are shown for the distribution in **Fig. 1b**; there are clear differences. The clumped species have more boundaries in some quadrats (i.e., more 2's and 3's) than shown by the randomly scattered species. The former also have more and larger gaps between quadrats with boundaries.

Methods to analyze such data began to appear in ecological papers in the mid-1970s. Note that it is therefore a recent phenomenon to have statistical techniques to use in tests of hypotheses about the existence of communities where these are defined in terms of nonrandom coincidences of their boundaries. Prior to the last 30 years, it was very difficult to use the sorts of data collected by community ecologists to distinguish between communities (i.e., with clumped boundaries), random and regular patterns of distributions of species.

Collective Properties of Assemblages

One reason that it matters whether or not species coexist in structured communities is because of so-called 'emergent properties' of associations of species. Thus, communities were long thought to have properties, or characteristics, that were not discernible, nor

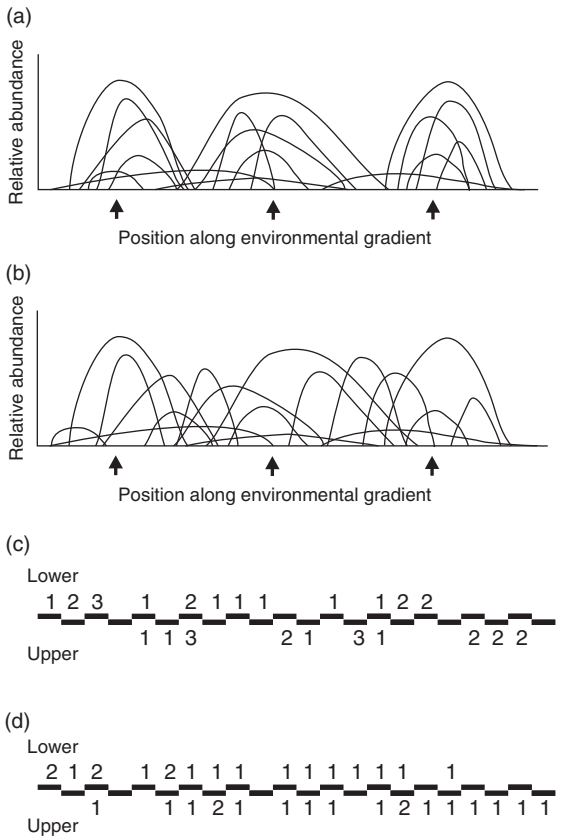


Fig. 1 Sampling along an environmental gradient (e.g., distance from a river, height up a mountain) to detect communities of species. (a) Most species are arranged in discrete communities, with very few occurring across boundaries of communities. Edges of distribution of the species in one community are coincident and there is little overlap with an adjacent community. (b) Species are arranged independent of each other and there are no communities. Sampling at intervals along the gradient (at the positions shown by arrows) would reveal different assemblages of species whether or not they were actually structured into communities. (c) and (d) The numbers of lower and upper boundaries per quadrat of the 18 species are shown for (a) and (b), respectively. The black bars represent the quadrats used along the gradient. Boundaries appear more clumped in (c) than in (d) (there are more boundaries in some quadrats in (c) than in (d); see text for details), identifying the difference between the two sets of distributions of species.

predictable, from knowledge of the species as individual entities. The properties of a set of species interacting with each other in a community were emergent in the sense that one had to study the community to know about, understand, or explain them. These characteristics of species 'emerged' from the community as the unit of study.

There has been widespread discussion of the concept of emergent properties in ecological investigations. Some of this discussion is, to say the least, confusing. For example, some properties of a set of species are simply collective properties due to grouping them into a set (whether or not such grouping is appropriate). Often, the diversity of species in a community is considered to be an emergent property, because it cannot exist for the individual species. In fact, this is a collective property that can be determined from the properties of the component species. An analogy is the average size of individuals of a population of deer. This is not a property of the individual deer, but can be calculated from their individual sizes. It does not 'require' study of the population as an entity.

Emergent properties may exist. For example, natural selection may operate on individuals because of their interactions with (and only with) other members of a community. If this were the case, clearly the community would have to be studied as a unit; the process (the property) could not be found (would not emerge) from study of only the component species.

As with other issues about communities, the concept of emergent properties is confusing, because the term means various and different things. To some, it means properties that become apparent when the scale of observation is changed, but this does not apply in the context of communities. The observations that are interpreted to be indicating the evidence of a community are at a particular scale; it makes no sense to change the scale.

To others, emergent properties are those that are unexpected to find in an assemblage because not enough information was available about its components. Using this definition, emergent properties are those that are found by studying communities, but that were not known from information about the component species. Ignorance about components does not make a more holistic study the best option – the community would still have to be understood in terms of its components for that ignorance to be dispelled. Apart from that, any such property could only possibly 'emerge' once. Once the property is known for one assemblage, it

can be predicted for another assemblage from similar knowledge about the components of that assemblage. This scarcely defines emergent properties of communities in a useful manner.

The third and, probably, philosophically most useful meaning of emergent properties is that they exist in a community, but cannot possibly be derived from knowledge of components of the community. Thus, reductionist studies of ecology will fail to provide understanding of communities because studying the components will not lead to understanding of the community as a whole.

So, for communities to be appropriate units of study, not only must they have some definable structure (as above), but also they must have emergent properties, so that subcomponents (i.e., the populations of different species) and their interactions are not appropriate units. This makes the practical usefulness of the concept of a community very difficult to demonstrate. Whether or not trying to demonstrate that existence of emergent properties is useful is a moot point, because it can distract from the real task of understanding the ecology of the observed assemblage.

A final comment on emergent properties is that many ecologists have holistic views about the way organisms operate, based on patterns of energy flow through ecosystems. Properties of ecosystems that make them appropriate units for study presumably include emergence. This makes the use of ecosystems even more problematic than the definitional problems briefly discussed earlier.

Do Ecologists Study Communities?

Apart from ecologists who have not been consistent in using any definition of a community, there are many who have attempted to define community and, perhaps, have attempted to demonstrate that what they study actually fits the definition. Searching the ecological literature for studies of communities is, however, revealing about the nature of community ecology. It is surprising to discover that many studies are, in fact, only about components of a community. The most common type of study is actually about a group of species that are closely related taxonomically, for examples, the warblers in a forest, or all the birds in a watershed, or all the beetles on a farm. Apart from being in different habitats, these three examples would usually involve increasing numbers of species. But, does this sort of study actually have anything to do with a community, where this has been defined in a useful or meaningful way?

The assemblage of fish on a patch-reef, or birds in a patch of forest, could only represent a community if no other species were interacting with them. Each group is generally interactive with other animals and plants as sources of food (and for other things). Such groupings are sometimes and should more generally be described as 'taxocoenes', that is, species that are related taxonomically and are found together in the same area.

Other ecologists study a 'guild', defined as an assemblage of species of similar or quite different taxonomic affinity that are found together and that use the same resources (for food, or shelter, etc.). For these, the term 'guild' is more informative than community (unless the definition of community is the same as guild, which makes the former redundant). An example would be to study, in some area, the set of birds, mammals, and ants that feed on seeds of the same species of plants.

So, most community ecologists do not actually study communities! There are, nevertheless, some cases where the unit of study is a community, where this means tightly coevolved species of different kinds of animals, living in close association, in the same arrangements at many times, in many places.

The classical examples are communities of host–parasite or host–pathogen relationships. The parasites are generally considered to persist because they adapt, immunologically, behaviorally, etc., to the host, despite the host's attempts to get rid of the parasites. Thus, nematodes and tapeworms are well adapted to their host's morphology, behavior, physiology, biochemistry, etc. If they were not so tightly integrated with the host, they might be likely to kill the host or to prevent it from reproducing. Both host and parasite would go extinct. Alternatively, the host would adapt to be able to rid itself of infestation and the parasite would not persist. Close coadaptation is therefore essential for the various species involved in a host–parasite community.

It is thus the case that some specialist communities do exist, can be objectively defined, and can be demonstrated to fit the definition. These are then the cases of communities which deserve the term as a description that identifies their usefulness as units of study. Caution is, however, still required. Not all cases of host–parasite (and similar) assemblages are so tightly structured. Many parasites have a complex life cycle, involving several quite different types of hosts (snails and sheep for some cestodes; invertebrates, birds, and fish for others). Under these circumstances, understanding the ecology of the assemblage of host and parasites requires understanding the direct and indirect trophic and competitive interactions between the hosts and other species that interact with them, but which are not involved with the parasites. This immediately creates all the problems of definition that were involved for any other apparent community.

Further Reading

Clements, F.E., 1916. Plant Succession: An Analysis of the Development of Vegetation. Washington: Carnegie Institute.
Edson, M.M., Foin, T.C., Knapp, C.M., 1981. Emergent properties' and ecological research. American Naturalist 118, 593–596.
Gleason, H.A., 1927. Further views on the succession concept. Ecology 8, 299–326.
Krebs, C.J., 1978. Ecology: The Experimental Analysis of Abundance and Distribution. New York: Harper and Row.
Lovelock, J.E., 1979. Gaia: A New Look at Life on Earth. Oxford: Oxford University Press.
McIntosh, R.P., 1985. The Background of Ecology. Cambridge: Cambridge University Press.
Odum, E.P., 1971. Fundamentals of Ecology, 3rd edn. Philadelphia: Saunders.

Salt, G.W., 1979. A comment on the use of the term 'emergent properties'. American Naturalist 113, 145–148.
Simberloff, D., 1980. A succession of paradigms in ecology: Essentialism, materialism and probabilism. In: Saarinen, E. (Ed.), Conceptual Issues in Ecology. Dordrecht, The Netherlands: Reidel, pp. 63–99.
Underwood, A.J., 1986. What is a community? In: Raup, D.M., Jablonski, D. (Eds.), Patterns and Processes in the History of Life. Berlin: Springer, pp. 351–367.
Whittaker, R.H., 1956. Vegetation of the Great Smoky Mountains. Ecological Monographs 26, 1–80.

Conbiota

Brian D Fath, Towson University, Towson, MD, United States and International Institute for Applied Systems Analysis, Laxenburg, Austria
Felix Müller, Christian-Albrechts-University, Kiel, Germany

Introduction

Textbooks described *Ecology* as the study of how organisms relate to other organisms and with their physical environment. More basically, this means the interactions of the living and nonliving aspects of the environment, or simply a study of the biotic and abiotic interactions in space and time. The breakthrough in this clarification, first proposed by German biologist, Ernst Haeckel in 1866 was the recognition and inclusion of the interactions and feedbacks of two broad scientific domains that previously were rarely considered in unison. The field of ecology has developed well over the past 150 years with little change in this basic direction or understanding.

However, this foundational insight raises some important questions. Just as biology addresses the fundamental question of "What is life?," we must also ask ourselves "What is nonlife?" The notion that we can distinguish between biota and abiota is perhaps a fallacy that should be explored more deeply.

By definition, abiota means without biota, or expressly, without life. Such conditions are easy enough to imagine in areas devoid of life, such as on the moon or other celestial surfaces known to be lifeless. However, any planet, such as earth that has life, it is clear that there is a constant and profound interplay between life and nonlife. The life actively selects elements from its environment and rearranges the chemical and physical conditions of that environment. Therefore, our contention is that it is inaccurate to refer to most nonliving aspects of the environment as abiotic. We introduce a new term, *conbiota*, to convey the sense that these environmental factors are not independent from life, but have coevolved *with life*. We provide several examples of this phenomenon and encourage readers to further explore and develop this concept of interdependence.

Global Examples of Coevolution Between Life and Environment

A useful perspective of the environment is to envision four interacting spheres: atmosphere, hydrosphere, lithosphere, and biosphere. The chemical composition of these four spheres is shown in **Table 1**. The current atmospheric balance, with a high concentration of reactive oxygen, is thermodynamically far from equilibrium, which is maintained by the renewal processes of the photosynthesis in the biosphere. This is evident in James Lovelock's famous insight that one way to detect if life is on a planet is to investigate the chemical composition of its atmosphere. In other words, basic atmospheric chemistry has evolved with life. The presence of oxygen not only was a requirement for the evolution of aerobic respiration, but also the precursor to the formation of the stratospheric ozone layer that provides protection from the sun's ultraviolet radiation. This has direct impact on the form of the lithosphere, and through indirect feedbacks, impact on the biosphere itself. Before the protective ozone layer was established, life was confined to water and near shore mudflats, which also provides some diffusive shield from the radiation. Thus, stratospheric ozone was a necessary step for the colonization of land by plants and animals. Before the advent of land plants, circa 600 million years ago, orogenic uplift activities would erode back to the sea "relatively" quickly, on a geological time scale of a few hundred million years. With the establishment of land plants, erosion rates were greatly reduced by the deep root action maintaining soil structure. Therefore, the lithosphere also is influenced, not only in composition but in form. The current topography of the planet is not an abiotic geological artifact but the coevolutionary result of life—nonlife interactions. **Fig. 1** shows possible pathways of exchange between the spheres. Note that while even a lifeless environment could experience exchanges, life is the active integrator of these spheres: the main drivers of transformation originate with life.

Local Examples of Coadaptation Between Life and Environment

At a local scale, scientists measure abiotic factors such as wind, light, temperature, humidity, runoff, soil chemistry, etc. It is clear to anyone that has walked through a forest that these factors are not independent of life. Quite the contrary, the measurements are highly adapted to the presence of living organisms. For example, under the canopy of a closed forest, the air is still, the temperature cool, the light subdued, the humidity high, and the rainfall largely intercepted. In an open field, one experiences more directly the sun, wind, rain, etc. Data from two experimental sites in Germany show explicitly the impact that biota can have on previously assumed abiotic factors.

The respective research sites are situated in the Bornhöved Lakes district and they have been investigated in an ecosystem research project for several years by several cooperating working groups. The detailed framework conditions and related results can be found a summarized form in Fränzle *et al.* (2008).

 doi:10.1016/B978-0-444-63768-0.10876-5

Table 1 Percentage atomic composition of the biosphere, lithosphere, hydrosphere, and atmosphere

Biosphere		*Lithosphere*		*Hydrosphere*		*Atmosphere*	
H	49.8	O	62.5	H	65.4	N	78.3
O	24.9	Si	21.22	O	33.0	O	21.0
C	24.9	Al	6.47	Cl	0.33	Ar	0.93
N	0.27	H	2.92	Na	0.28	C	0.04
Ca	0.073	Na	2.64	Mg	0.03	Ne	0.002
K	0.046	Ca	1.94	S	0.02		
Si	0.033	Fe	1.92	Ca	0.006		
Mg	0.031	Mg	1.84	K	0.006		
P	0.030	K	1.42	C	0.002		
S	0.017	Ti	0.27	B	0.0002		

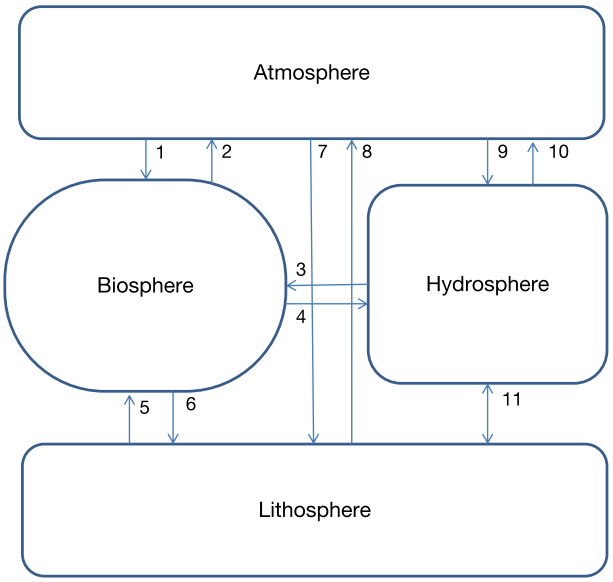


Fig. 1 Interactions of the four global spheres: (1) photosynthesis, (2) respiration, (3) water uptake, (4) evapotranspiration, (5) nutrient uptake, (6) decomposition, (7) dry deposition, (8) suspension, (9) precipitation, (10) evaporation, (11) ground water recharge and discharge.

The first example (Fig. 2A) demonstrates that organisms are manipulating the temperature regimes in their habitat, showing temperature regulations and buffer capacities related to microclimatic conditions. The figure includes a comparison of a beech forest ecosystem and the directly neighboring arable land ecosystem at the Bornhöved research area during one day in the summertime (June 11th). The measured temperatures in the air (0 to 36 m) and the soil (0 to −1 m) are depicted as hourly temperature gradients. The highest temperatures in the air appear at the 18:00 curve, the lowest at 5:00 in the beech forest. The grassland conditions are similar concerning the timing, but rather different concerning the amplitudes: 8–35°C in the air of the grassland, while the structure of the beech forests limits the span width to 10–24°C. In the soil, similar relations can be seen, whereby the average temperature of the grassland is much higher than in the beech forest, which shows an amplitude of about 2°C while 10°C are reached in the grassland.

The winter conditions are visible in Fig. 2B. Here the buffering capacity of the forest canopy is reduced strongly, therefore also the temperature amplitudes are high in beech forest. Also, the sequence of time steps is similar in both cases but the soils differ very much. While the winter temperature buffering capacity of beech forest is very high, higher frost challenges appear in the case of the grassland.

Similar to the temperature climate, ecosystems are also modifying their hydrological conditions on the base of living processes or living process products, such as soil organic matter or substrate heterogeneity. The plants are producing heterogeneity of hydrological pathways, for example, by stem flow or through fall distributions, and they are also strongly determining the seepage dynamics, the evapotranspiration and—by these factors—the whole water budget. Fig. 3 is demonstrating these conditions with reference to the two compared ecosystems "arable land" and "beech forest," which have also been analyzed in the temperature figures. The two ecosystems were both used as agricultural fields until 100 years ago when the forest was planted. Thus, all distinctions have been emerging due to different land use structures—and different conbiotic relations—for 100 years. Here, we can see interesting differences with respect to the pathways of water loss: while in the agricultural fields the seepage is dominating,

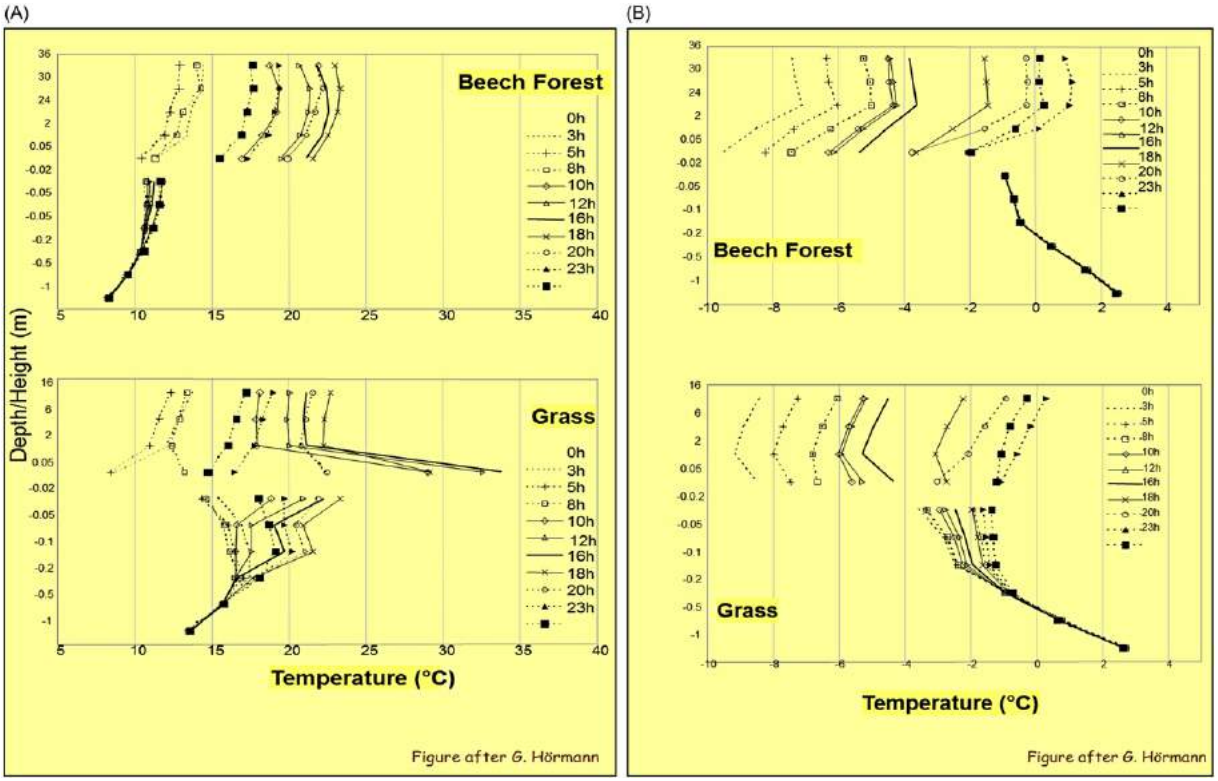


Fig. 2 (A) Temperature profiles of a beech forest ecosystem and the directly neighboring arable land ecosystem (planted with grass) at the Bornhöved research area in the summertime (June 11th). The measured temperatures in the air (0 to 36 m) and the soil (0 to −1 m) are linked as hourly temperature gradients. Measurements and data compilation were done by Hörmann and colleagues. (B) Temperature profiles of a beech forest ecosystem and the directly neighboring arable land ecosystem (planted with grass) at the Bornhöved research area in the winter (December 29th). The measured temperatures in the air (0 to 36 m) and the soil (0 to −1 m) are linked as hourly temperature gradients. Measurements and data compilation were done by Hörmann and colleagues.

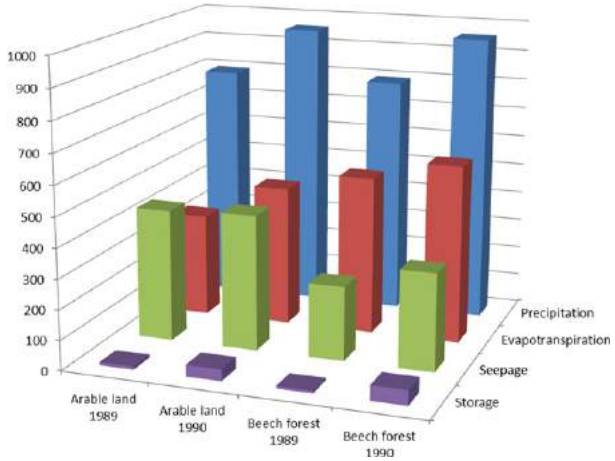


Fig. 3 Elements of the soil water balance of the two ecosystems "beech forest" and "arable land" in the Bornhöved Lakes area in mm/a, following Bornhöft (1993, p. 91).

producing a flow toward the groundwater layer, the forest flows are dominated by evapotranspiration flows, whereby the transpiration is the most important process.

In Fig. 4, the soil water balance of the arable land ecosystem is depicted as a gradient scheme from 1989 to 1998. Each row shows the development of the volumetric soil water content within 1 year (number of days at *x*-axis), and the colors are related to the percentage; deep blue demonstrates a high water content, while it is low in the case of the red colors. Obviously, at the measured field, the lower soil compartments are much dryer than the upper ones; around 50 cm depth, the plowing inputs have

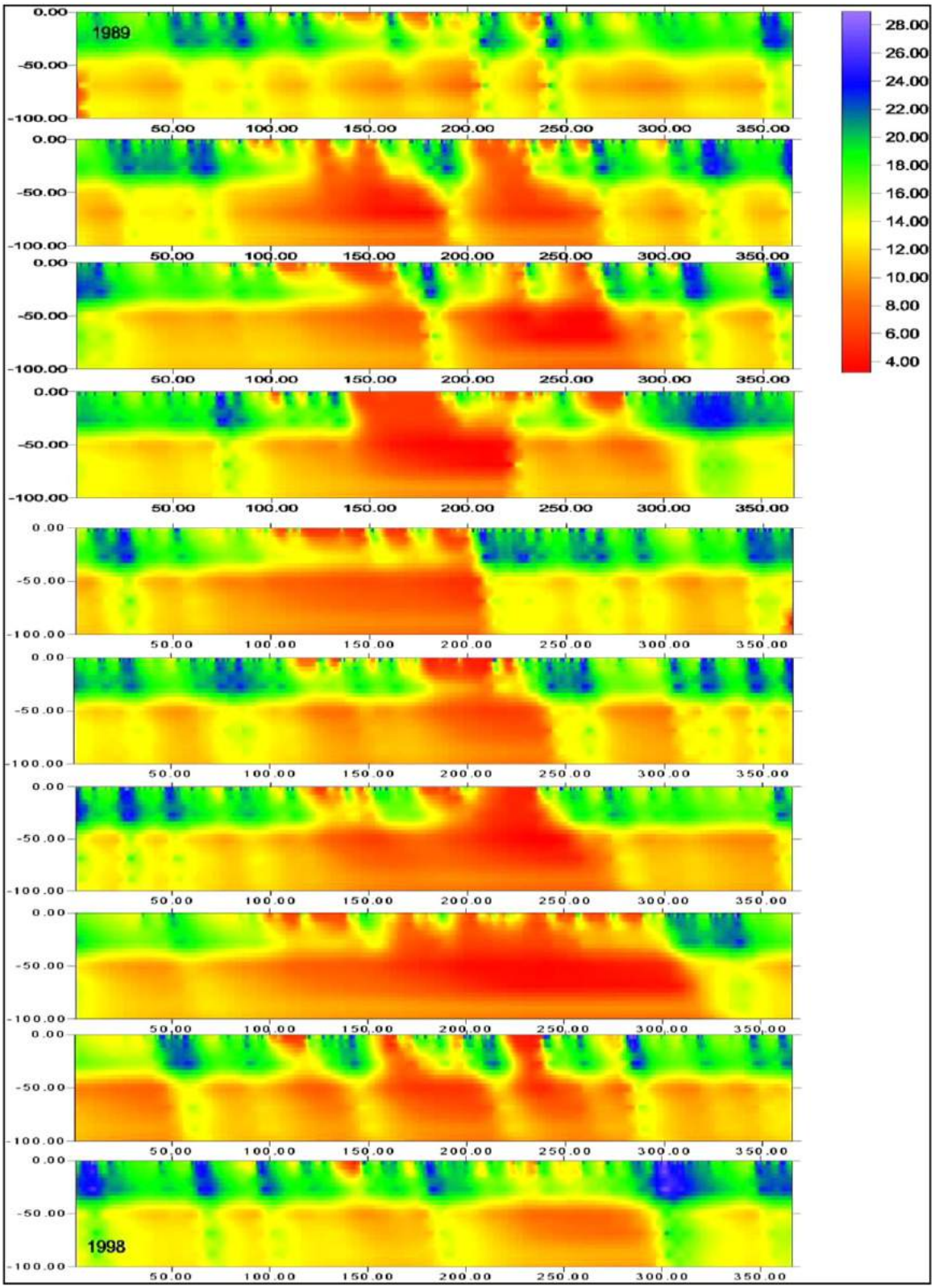


Fig. 4 Dynamics of volumetric soil moisture gradients (vol%) at the arable land site of the Bornhöved research area between 1989 and 1998. Data compilation and measurements by G. Hörmann and colleagues (see Fränzle *et al.*, 2008).

produced a strong gradient. This evidence of a plow pan can be found in all cases. The abiotic conditions are here regulating the water flows, which only in a few cases (blue intrusions) show rapid drainage phases, thus the "living space" with active plant roots

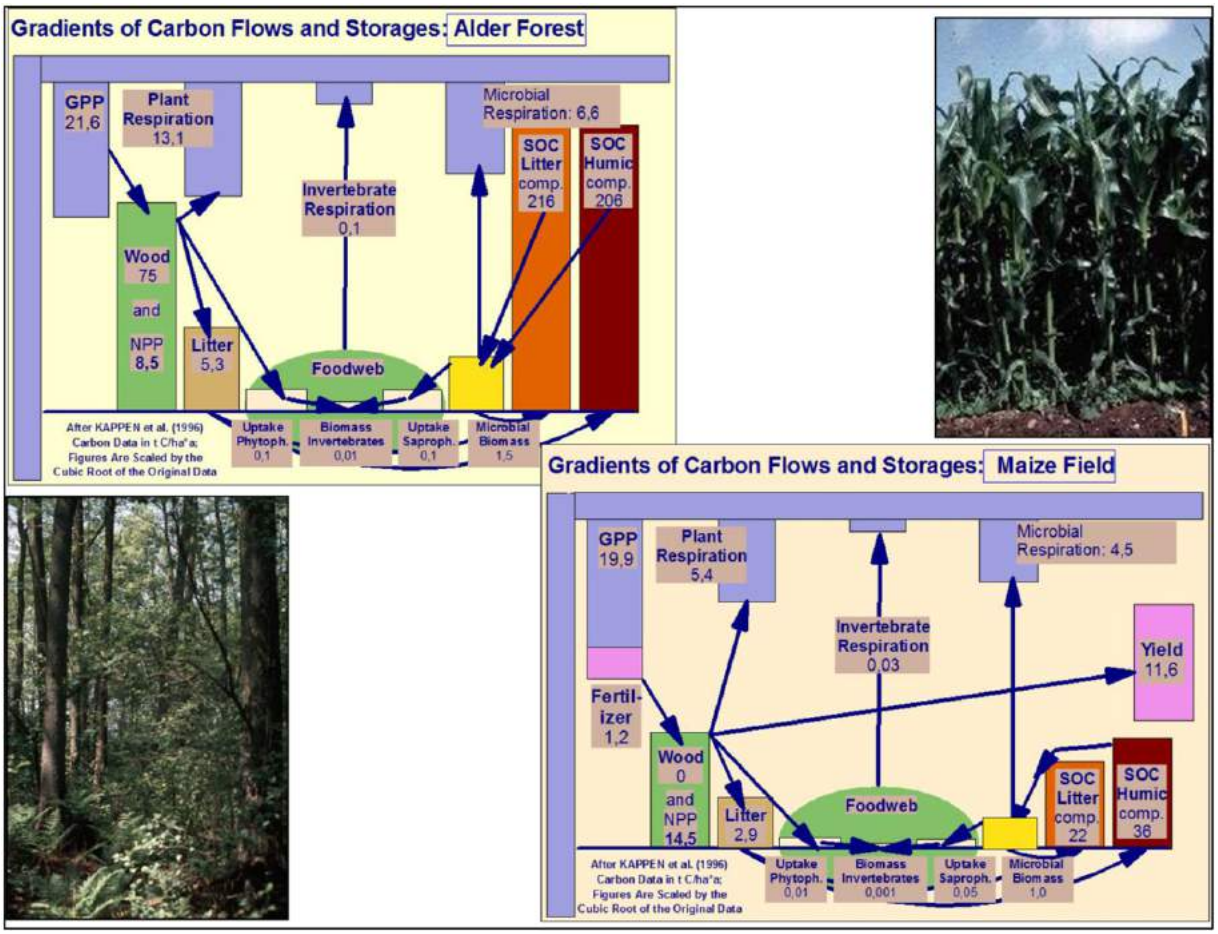


Fig. 5 Elements of the carbon budget of the alder forest ecosystem in comparison with the arable land test site at the Bornhöved research area. For original data and detailed descriptions see Fränzle *et al.* (2008).

is separated from the dry soil body in the deeper compartments. In this case, the biotic interference of the "abiotic" soil conditions was human use of the ecosystem for agriculture.

Another interesting observation can be related to the climate dynamics throughout the years of measuring activity: the summer drying period is steadily increasing in length and intensity between 1992 and 1996.

Fig. 5 also shows us how biotic and abiotic elements are cooperating and how the factors are mutually influenced. The flows of carbon are assessed for two ecosystems, the alder forest of the Bornhöved main research area and the neighboring maize field, which is the (meanwhile well-known) arable land ecosystem. While in this system, through agriculture, there are annual measures that reorganize it to a pioneer stage, the alder forest has had about 60 years to develop on the base of former wetlands. Thus, we can find extreme distinctions, for example, concerning the organic carbon contents in the soils, which are distinguished here into litter (easily decomposable) and humic compounds (hardly decomposable). Both fractions provide minimal values at the field ecosystem, which are not given the chance to store detritus as soil organic carbon in a higher intensity. These abiotic items are underlined by the water conditions: while the field is representing a rather dry state, the alder forest provides a steadily wet soil, which additionally is influenced by N fixing bacteria of the Frankia type. Furthermore, the forest has been enabled to build up a self-organized structure with many micro habitats which are missing on the field. Finally, the architecture of the forest is providing a well-buffered microclimate. In the end these "abiotic" factors which have been mainly achieved due to biotic activities, are providing habitat conditions for several species, thus the biomass of the fauna is 10 times higher than the biomass of the field ecosystem.

Finally, the biotic and abiotic conditions are summarized in **Fig. 6**, where integrity indicators are depicted in a comparison of the beech forest and the neighboring arable land ecosystem. The assessment of integrity has been based on the orientor approach (Müller and Leupelt, 1998) and ecosystem concepts, which try to synthesize ecosystem structures and functions, representing the ecosystem energy, water, and matter balances. Thus, we find a holistic representation of the degree and the capacity for complexifying ecological processes on the base of an accessible number of indicators. They also represent the basic trends of ecosystem development; thus, they show the developmental stage of an ecosystem or a landscape. They are totally based upon the conbiotic idea: biotic and abiotic elements are forming a mutual system of interactions with increasing interrelations as the developing sequence of succession continues. In this sequence, the system is optimizing the following indicators: biodiversity, abiotic heterogeneity, exergy capture, entropy export, storage capacity, reduction of nutrient loss, biotic water flows, and metabolic

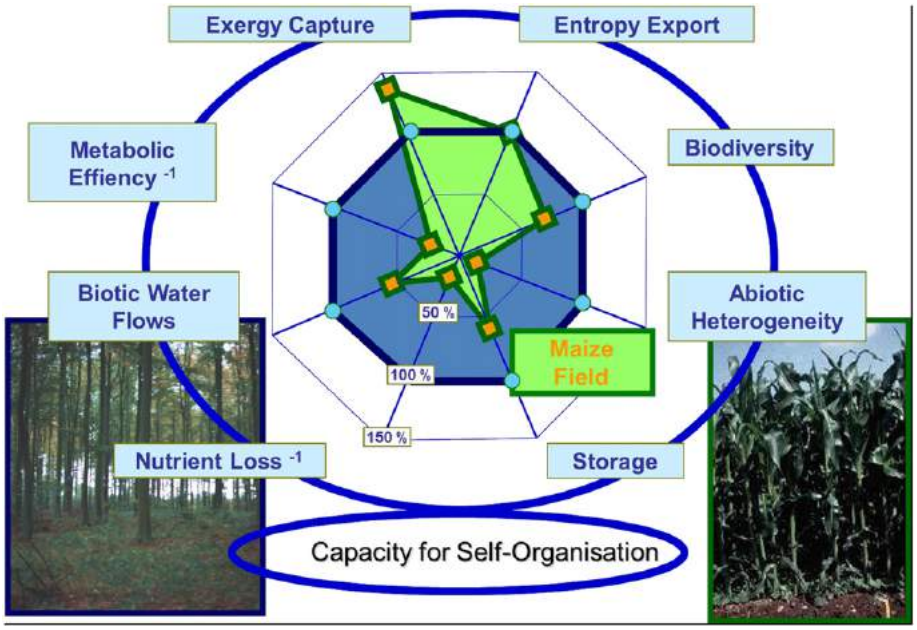


Fig. 6 Comparison of indicator values for ecosystem integrity referring to two ecosystems in Bornhöved Lakes research area. After Müller, F. (2005). Indicating ecosystem and landscape organization. *Ecological Indicators* **5**(4), 280–294.

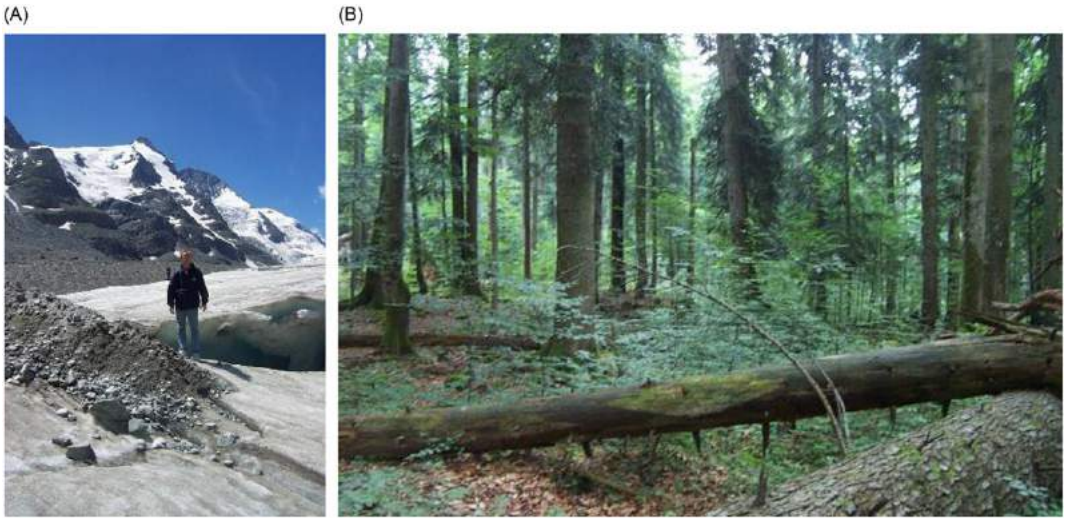


Fig. 7 (A) The Pasterze Glacier retreating in Austria. An r-selecting environment, where the conbiotic factors are weaker making it more ephemeral and more extreme. Photo by BD Fath. (B) An old growth forest in Slovenia. A K-selecting environment with strong conbiotic factors that regulate basic environmental conditions providing stability and predictability. Photo by BD Fath.

efficiency. All of them are produced as outcomes of biotic and abiotic activities, thus the basic features of ecological systems in fact are founded on conbiotic structures and functions.

Conbiota through a successional lens also leads to the recognition that environments are in varying stages of integration and development. The degree to which life modifies and shapes that environment is along the continuum from early states of succession to later stages of succession. Much emphasis is placed in succession theory on the traits of the "r-selected" species and "K-selected" species. The former are fast-growing, short-lived, with effective dispersal mechanisms such as wind borne seeds. They often employ vegetative or asexual reproduction, do not compete well with other species, and their numbers fluctuate widely, being strongly influenced of density-independent factors. The latter are slow-growing, long-lived, with low reproductive rates and dispersal rates. Populations are more stable due to diversion of production to defense and parental investment. These two characters present end points for idealized species to occupy a role. However, just as important as the species itself is the environment in which it inhabits. There are "r-selecting" environments and "K-selecting" environments. The former are more extreme, ephemeral, unpredictable (Fig. 7A), and the latter are more stable, homeostatic, and predictable (Fig. 7B). Furthermore,

they are that way precisely because the interactions between life and the environment. This is additional evidence that the earth is a self-regulating, feedback-driven system such that life creates conditions that are conducive to life.

Summary and Conclusion

The idea that life and environment are deeply intertwined is not new. However, language matters, and the current vocabulary that employs biota and abiota as strict opposites does not accurately reflect the reality of the situation. Therefore, here we suggest introducing the term *conbiota* to explicitly and formally recognize the coevolutionary and coadaptive ways in which life alters and shapes its environment. This term is a broad brush that acknowledges the role of life, but we assume that there is a gradient of influence of life on environment. Further refinement of this concept could lead to new classes of properties that measure the degree to which life plays an active role. This might spark a discussion not only into the question we posed earlier, "What is life?," but also into its counterpart "What is environment?"

References

Bornhöft, D., 1993. Untersuchungen zur Beschreibung und Modellierung des Bodenwasserhaushalts entlang einer Agrar- und einer Wald-Catena im Bereich der Bornhöveder Seenkette (Schleswig-Holstein). EcoSys Supply 6, 1–134.
Fränzle, O., Kappen, L., Blume, H.-P., Dierßen, K. (Eds.), 2008. Ecosystem organization of a complex landscape—Long-term research in the Bornhöved Lake District, Germany, Ecological studies, vol. 202. Berlin, Heidelberg/New York: Springer Science & Business Media.
Müller, F., Leupelt, M. (Eds.), 1998. Eco targets, goal functions, and Orientors. New York: Springer-Verlag.

Further Reading

Fath, B.D., 2014. Sustainable systems promote wholeness-extending transformations: The contributions of systems thinking. Ecological Modelling 293, 42–48.
Lovelock, J.E., 1972. Gaia as seen through the atmosphere. Atmospheric Environment 6 (8), 579–580.
Lovelock, J.E., Margulis, L., 1974. Atmospheric homeostasis by and for the biosphere: The Gaia hypothesis. Tellus 26 (1–2), 2–10.
Müller, F., 2005. Indicating ecosystem and landscape organization. Ecological Indicators 5 (4), 280–294.
White, I.D., Mottershead, D.N., Harrison, S.J., 1992. Environmental Systems, 2nd edn. London: Stanley Thornes, Ltd.

Cooperation☆

Raghavendra Gadagkar, Indian Institute of Science, Bangalore, India

Background

Not surprisingly, humans have displayed an absorbing fascination for examples of cooperation in the animal world, long before the evolutionary puzzle associated with them became evident. Indeed, freedom from evolutionary thinking accommodated all manner of untenable theories about cooperation, in the past. While T.H. Huxley believed that cooperation and altruism were only possible among close kin and P. Kropotkin saw "mutual aid" everywhere he looked, unconnected with any sort of kinship, both W.C. Allee and V.C. Wynne Edwards succumbed to a naive form of group selection, the notion that cooperation and self-sacrifice existed because they were good for the group and the species—never mind that they were harmful to the individuals displaying them.

Kin Selection

Modern evolutionary thinking by people such as J.B.S. Haldane, W.D. Hamilton, R.L. Trivers, J.M. Smith, and E.O. Wilson, which kept in mind the critical problem of the potential for a few cheaters to wreck any cooperative group, has given us our current theories of cooperation. These modern advances have depended in part of clear-cut demarcation and definition of different kinds of interactions possible among animals (see **Fig. 1**). The most significant advance in explaining the evolution of cooperation came from Hamilton's inclusive fitness theory. Not only does this theory provide a logical explanation for why cooperation evolves more easily among kin, it also shows why close kinship is not always essential. Kin selection or, more precisely, Hamilton's rule has three parameters, namely, cost to the actor, benefit to the recipient, and the coefficient of relatedness between actor and recipient. Given appropriately skewed cost/benefit ratios, it is easy to see that even rather low levels of relatedness can satisfy Hamilton's rule. Unfortunately, an excessive and often exclusive focus on measurement of relatedness and the neglect of the cost and benefit terms in empirical studies, has sometimes given the false impression that kin selection fails to explain cooperation.

When the cost and benefit terms have been adequately measured, Hamilton's rule has proved to be a powerful theoretical framework for understanding the evolution of cooperation and altruism in a wide variety of organisms from bacteria to man. To cite just two examples, studies on the white-fronted bee-eater in Kenya have shown that not only the presence of helpers at the nest but also the bizarre behavior of the father's harassing their sons to return and act as helpers, is consistent with the predictions of Hamilton's rule. Computation of the costs, benefits, and relatedness involved in different strategies shows that by harassing their sons and bringing them back to help rear additional offspring, fathers gain a substantial fitness advantage. In contrast, sons reap about the same fitness benefit whether they resist their father's harassment and carry on with their own family life or whether they succumb to the harassment and return to act as helpers. And in the primitively eusocial wasp *Ropalidia marginata*, Hamilton's rule correctly predicts that only about 5% of the individuals should opt for solitary life while the remaining should opt for altruistic, sterile worker roles.

In any case, kin selection is indeed inadequate when cooperation is directed toward nonrelatives, as it often is in human societies. Those who study humans use a somewhat more liberal use the word cooperation using it not only when there is cooperation, that is, both actor and recipient benefit from an interaction but also when there is altruism, that is, when the actor pays a cost and only the recipient benefits, cost and benefit being measured of course in terms of Darwinian fitness. To help explain the evolution of cooperation by natural selection in the absence of close genetic relatedness between actor and recipient, at least four additional potential mechanism of evolution have been suggested, in addition to kin selection (see **Fig. 2**). The first of these additional mechanisms is Direct Reciprocity, originally proposed by Robert Trivers as "Reciprocal Altruism." The idea here is that if a cost incurred now is retrieved at a later point in time, then both actor and recipient will benefit on the long run. As might be expected, such reciprocity will require individual recognition, long-term memory, and well developed cognitive abilities. For these reasons direct reciprocity is more likely in humans than in animals for which it was first proposed. And once we allow cognitive abilities involving recognition and memory, other more sophisticated and more subtle mechanisms can be imagined. For example, one can postulate that there is no need for the same individual who received help to reciprocate. As long as the helper received help in the future from someone, the same effect is achieved. This is possible if the act of helping someone by a helper is noticed by others in the population and the helper builds up a reputation for being a helper. Then others who may

☆*Change History*: March 2018. Raghavendra Gadagkar added new keywords, modified the section on kin selection, revised the conclusion section, and added several new items to the Further reading list, and added two new figures with legends.

This is an update of R. Gadagkar, Cooperation, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 776–777.

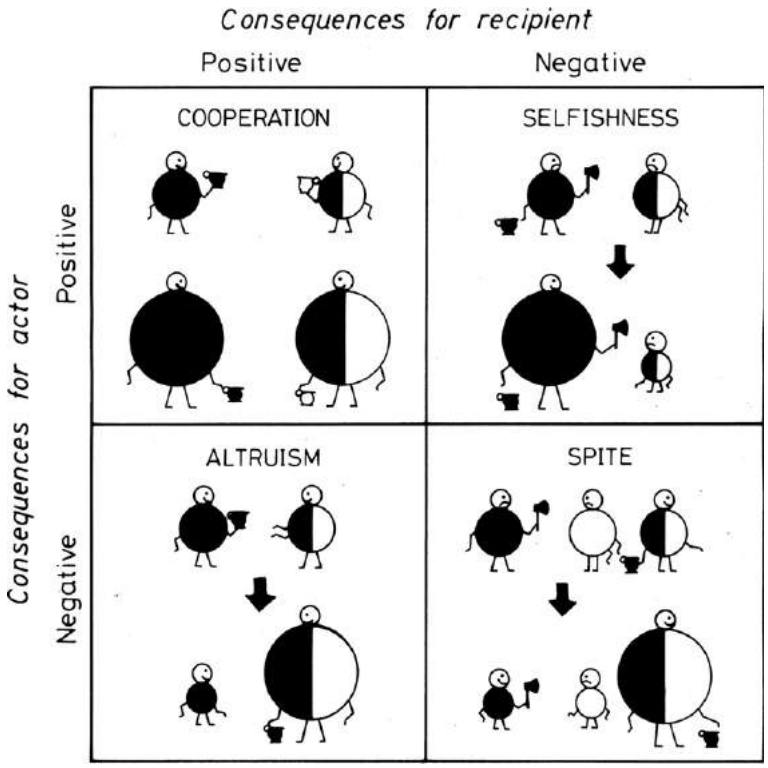


Fig. 1 The consequences of interaction between animals. The recipient here is the actor's brother and therefore shares 50% of his genes, as is indicated by the shading. Help of any kind (the offering of food or shelter, easing access to a mate, and so on) is indicated by a vessel, and harmful behavior by an ax. *Cooperation*: Both individuals benefit and such behavior will therefore evolve easily. *Altruism*: The altruist diminishes his own genetic fitness but raises his brother's fitness to the extent that the shared genes are actually increased in the next generation. *Selfishness*: The selfish individual reduces his brother's fitness but increases his own to an extent that more than equals the brother's loss. *Spite*: The spiteful individual lowers the fitness of an unrelated competitor (the unshaded figure) while reducing that of his own or at least not improving it; but the act increases the fitness of the brother to a degree that more than compensates for the actor's loss. Reproduced from Gadagkar, R. (1997). Survival strategies—Cooperation and conflict in animal societies. Cambridge, MA: Harvard University Press.

not necessarily have received help directly from this helper will also be likely to help her because she is a known helper. This kind of mechanism labeled as Indirect Reciprocity is postulated to explain why for example, people donate to charities. A more sophisticated form of indirect reciprocity may involve cooperators for spatially defined networks to keep cheaters at bay. This is labeled as Network Reciprocity and has been shown mathematically to help cooperation to evolve in populations of potentially selfish individuals. Finally there is Group Selection, a refined version that is cognizant of the problem of cheaters and postulates that although selfish individuals outcompete cooperative or altruistic individuals within groups, groups of cooperators outcompete groups of selfish individuals and the net effect depends on how fast the selfish individuals drive cooperators within group to extinction relative to how fast groups of cooperators drive to extinction groups of selfish individuals. Despite their apparent diversity, these five mechanisms for the evolution of cooperation have a rather beautiful mathematical unity (see Nowak, 2006).

Perhaps the most fascinating recent advance in the study of cooperation and altruism in humans has been due to the collaboration of evolutionary biologists, psychologists, and economists and the use of "games," such as the ultimatum game and the public goods game, to uncover patterns of human behavior. The main results of such studies are that people by and large do not behave and expect others to behave, in apparently rational, selfish ways traditionally predicted by theoretical economists. Instead, people behave in a fair manner and expect others to do the same. Even more interestingly, people appear to have an innate dislike for cheaters and are often willing to incur as cost to themselves to punish cheaters even if it yields them no direct benefit. The prevalence of such "altruistic punishment" is now thought to be the evolutionary force that maintains cooperation and altruism in human societies.

It must be mentioned that even in the context of evolution of sociality and cooperation in animals and insect societies such as those of ants, bees and wasps, kin selection itself has recently come under severe criticism. Claiming that kin selection, also known often as inclusive fitness theory or Hamilton's rule, is not sufficiently general and robust, some prominent researchers have turned back to a combination of individual and group selection to explain the evolution of cooperation, altruism, and

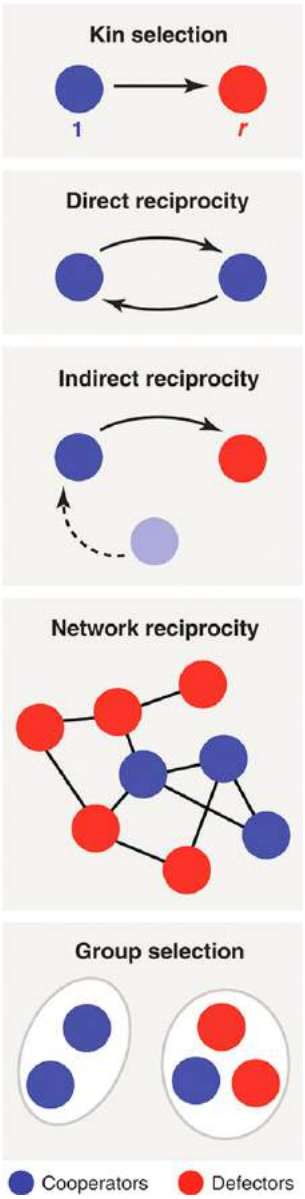


Fig. 2 Five mechanisms for the evolution of cooperation. Kin selection operates when the donor and the recipient of an altruistic act are genetic relatives. Direct reciprocity requires repeated encounters between the same two individuals. Indirect reciprocity is based on reputation; a helpful individual is more likely to receive help. Network reciprocity means that clusters of cooperators outcompete defectors. Group selection is the idea that competition is not only between individuals but also between groups. Reproduced from Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science* **314**, 1560–1563.

sociality in the animal kingdom as a whole. This effort is sometimes referred to as multilevel selection. While the defenders of kin selection have dismissed these criticism as unjustified and a rather fierce debate has ensued and deeply divided the disciple, it appears to me that the scientific study of cooperation in animals and humans is poised for even more exciting new developments, whichever way the debate for and against kin selection goes. This optimism is rooted in the current realization among biologists that although natural selection is rooted in competition it is cooperation among competing actors that is central to every major evolutionary innovation, every major transition in evolution, whether it was the origin of cells, of eukaryotes, of insect and other animal societies, or of language.

See also: Behavioral Ecology: Social Behavior and Interactions. General Ecology: Communication

Reference

Nowak, M.A., 2006. Five rules for the evolution of cooperation. Science 314, 1560–1563.

Further Reading

Abbot, P., Abe, J., Alcock, J., *et al*., 2011. Inclusive fitness theory and eusociality. Nature 471, E1–E4.
Bourke, A.F.G., 2011. Principles of social evolution. Oxford: Oxford University Press.
de Vladar, H.P., Szathmáry, E., 2017. Beyond Hamilton's rule. Science 356, 485–486.
Dugatkin, L.A., 2006. The altruism equation—Seven scientists search for the origins of goodness. Princeton, NJ: Princeton University Press.
Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415, 137–140.
Gadagkar, R., 1997. Survival strategies—Cooperation and conflict in animal societies. Cambridge, MA: Harvard University Press.
Gadagkar, R., 2001. The social biology of *Ropalidia marginata*. Cambridge, MA: Harvard University Press.
Gadagkar, R., 2010. Sociobiology in turmoil again. Current Science 99, 1036–1041.
Maynard Smith, J., Szathmáry, E., 1995. The major transitions in evolution. Oxford: W.H. Freeman and Company Ltd.
Nowak, M.A., Highfield, R., 2011. Super cooperators—Altruism, evolution, and why we need each other to succeed. New York/London: Free Press.
Nowak, M.A., Tarnita, C.E., Wilson, E.O., 2010. The evolution of eusociality. Nature 466, 1057–1062.
Rubenstein, D.R., Abbot, P.E., 2017. Comparative social evolution. New York: Cambridge University Press.
Székely, T., Moore, A.J., Komdeur, J., 2010. Social behaviour. Cambridge: Cambridge University Press.
Trivers, R.L., 1971. The evolution of reciprocal altruism. Quarterly Review of Biology 46, 35–57.
Wilson, E.O., 2012. Social conquest of earth. London: W.W. Norton & Company Ltd.

Demography

B-E Sæther, Centre for Conservation Biology, Norwegian University of Science and Technology, Trondheim, Norway

Introduction

Generation time, that is, the mean age of mothers of newborn individuals when stable age distribution is achieved, summarizes many key demographic features of a species. Long generation times occur when the survival rate is large and when the population growth rate or the net reproductive rate is small. Here we shall show that variation in generation time can explain many basic relationships between demography and population dynamical characteristics and life history.

Definition of Generation Time

The famous Euler–Lotka polynomial equation,

$$\sum_{x=1}^{\omega} \lambda^{-x} l_x m_x = 1$$

allows us to calculate the asymptotic population growth rate λ if the survival function l_x, which gives the probability of being alive from age 1 to age x, as well as the fecundity rate at age x, m_x, are known. This equation has ω roots, that is, equal to the maximum number of age classes, which can be obtained numerically. The generation time is then given by

$$T = \sum_{x=1}^{\omega} x\lambda^{-x} l_x m_x$$

Many species have demographic rates that are almost constant or independent of age. If α is the age of first reproduction, and assuming $\omega = \infty$, we get a simplified expression for the generation time $T = \alpha + s/(\lambda - s)$, where s is the adult survival rate. We see that in general, T decreases with λ and increases with the net reproductive rate.

The `Slow–Fast' Continuum of Life-History Variation

Comparative analyses of many taxonomic groups have shown a strong covariation among different life-history traits. At one end of this `slow–fast' continuum of life-history variation, we find species that mature early and have large litter sizes, but short life expectancy. At the other end of this continuum, species are located that start to reproduce first after several years, have a small litter size (often just a single offspring), but have high adult survival rates, which may exceed 95%.

The sensitivity of λ to a small change in a demographic trait can be derived by implicit differentiation of the Euler–Lotka equation trait, giving

$$\frac{\partial \lambda}{\partial m_x} = \frac{\lambda^{-x+1} l_x}{T} \text{ and } \frac{\partial \lambda}{\partial s_x} = \frac{\lambda^{-x} l_x v_{x+1}}{T}$$

where v_x is the reproductive value:

$$v_x = \frac{\lambda^x}{l_x} \sum_{u=x}^{\omega} l_u m_u \lambda^{-u}$$

Similarly, the elasticity of ln λ to a change in the logarithm of a life-history parameter is

$$e_p = \frac{\partial \ln \lambda}{\partial \ln p} = \frac{p}{\lambda} \frac{\partial \lambda}{\partial p}$$

The elasticity index can be used to compare the relative contribution of the trait to the population growth rate λ because elasticities sum up to 1. We can then examine how the same relative change in a trait, for example, in adult survival, will affect λ when we move along the `slow–fast' continuum of life-history variation.

In birds, some clear patterns appear in the interspecific distribution of elasticities. (1) The mean elasticity across species of adult survival rate was significantly larger than the mean elasticity of fecundity rate. (2) The distribution of the elasticities of fecundity rate was skewed against small values, whereas the elasticities of adult survival rate values were approximately normally distributed around the mean (Fig. 1). (3) The elasticity of adult survival increased with adult survival rate (Fig. 2a) and decreased with clutch size (Fig. 2c). In contrast, the elasticity of fecundity rate decreased strongly with adult survival rate (Fig. 2b), and hence age at maturity, but increased with clutch size (Fig. 2d). Thus, relative changes in adult survival of birds have larger impact on the

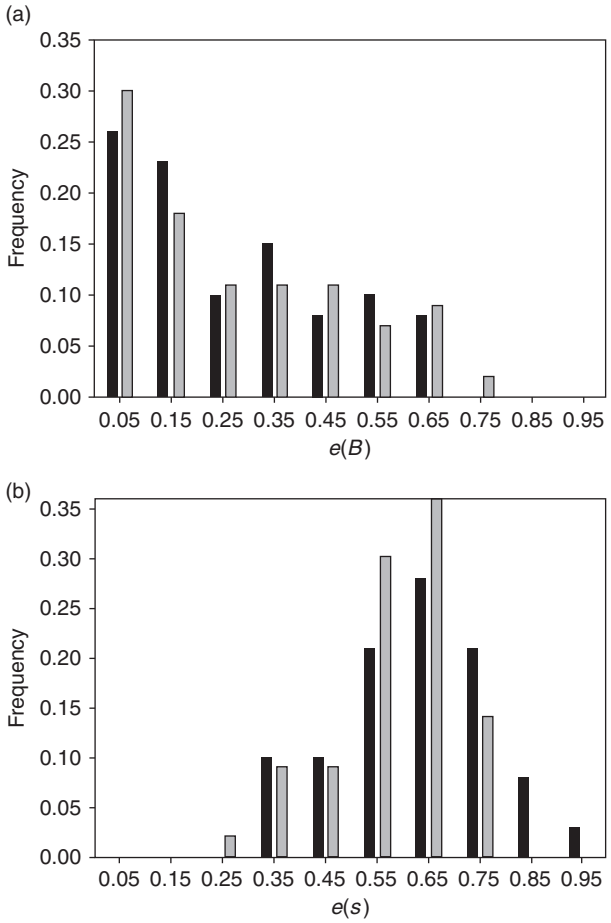


Fig. 1 The distribution of the elasticities of the (a) fecundity rate $e(B)$ and (b) adult survival rate $e(s)$ calculated for the asymptotic population growth rate λ estimated from the Leslie-matrix (solid columns) and when the juvenile survival rate s_{juv} is chosen to give $\lambda = 1$ (shaded columns).

population growth rate than changes in fecundity rate. Furthermore, the growth rate of species at the slow end of the `slow–fast' continuum of life-history variation is more sensitive to changes in adult survival rate than high-reproductive species. Similar patterns also seem to be present in other vertebrate taxa as well.

Decomposing Population Dynamics

Analyses of time series of population fluctuations show large interspecific differences in population dynamics. As a first step, it is convenient to classify them into time series that show a long-term increase or decrease, and time series that show fluctuations around some mean population size. In the first group, there is no density dependence, and the changes from one year to another are determined by the population growth rate and stochastic variation in the growth rate. In the second group, density dependence is also present, which produces a characteristic return time to equilibrium.

Stochastic variation in the population growth rate is due to demographic and environmental stochasticity. Demographic stochasticity is caused by random variation among individuals in fitness contributions due to independent chance events of individual survival and reproduction. This produces random fluctuations in population size, with variance proportional to σ_d^2/N. Environmental stochasticity affects the age-specific demographic rates of all individuals in the population similarly, producing a constant variance among years in the population growth rate σ_e^2, independent of population size. Thus, environmental stochasticity affects the population at all sizes, whereas demographic stochasticity most strongly influences the dynamics at smaller population sizes.

To understand the effects of generation time we must examine how it affects each of the components in the population dynamics.

Life-History Correlates of Demographic Stochasticity in Birds

Based on individual demographic data, the demographic variance σ_d^2 can be estimated from data on individual variation among females in their fitness contributions to the following generations. The total contribution of a female i in year t (R_i) is the `number of female offspring born during the year that survive for at least 1 year' plus 1 if the female survives to the next year. The

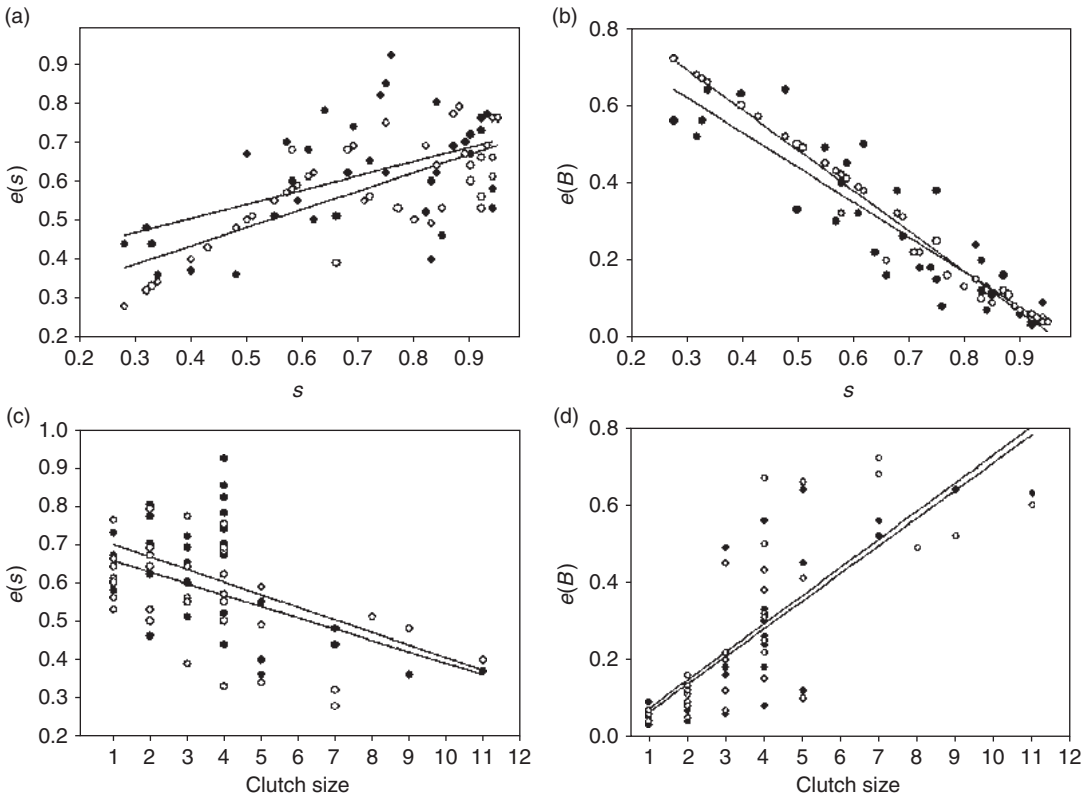


Fig. 2 The elasticity of adult survival rate $e(s)$ and fecundity rate $e(B)$ of birds in relation to adult survival rate s and clutch size, calculated for the actual asymptotic population growth rate λ (circles) and when the juvenile survival rate s_{juv} is chosen to give $\lambda = 1$ (squares).

demographic variance can then be estimated as the weighted mean across years of

$$\sigma_d^2(t) = E\frac{1}{a-1}\sum (R_i - \overline{R})^2$$

where $\bar{R}$ is the mean contribution of the females, a is the number of recorded contributions in year t, and E denotes the expectation. We can further partition σ_d^2 into components due to variation in fecundity, survival, and interaction between fecundity and survival. Writing B for the number of offspring produced, $I = 1$ if the mother survives and $I = 0$ if she dies, the demographic variance is the mean over years of the within-year variance of $R = B + I$. This can be split into its components $\mathrm{var}(R) = \mathrm{var}(B) + \mathrm{var}(I) + 2\mathrm{cov}(B, I)$ that can be estimated separately by simple sum of squares. In long-lived species, stochastic variation in age structure constitutes an important component of demographic stochasticity and for species with a mean age of maturity older than 3 years we must base our estimates on contributions (B_{jt}, I_{jt}) for the different age classes i in year t. We then calculate the demographic stochasticity from the projection matrix and separate this into components that are generated by demographic stochasticity in each vital rate.

It is obvious that such detailed demographic data are rarely available. However, in birds we can compute demographic variance of several species and relate the estimates to the position of the species along the `slow–fast' continuum of life-history variation. According to one hypothesis, σ_d^2 is expected to increase with adult survival rate (and hence to decrease with clutch size) because very few offspring recruit in short-lived species with a high first-year mortality. Alternatively, σ_d^2 can be expected to decrease with adult survival rate because life-history constraints (small reproductive rates, high life expectancy) generate small variability in fitness among individuals in long-lived species.

In birds, the two components of the demographic variance due to stochastic variation in fecundity and survival were positively correlated. As expected from this relationship, interspecific differences in demographic variance were closely related to the size of both the fecundity component and the survival component. Interspecific differences in demographic stochasticity were well explained by life-history variation. Larger values of σ_d^2 were found in species at the fast end of the avian life-history continuum, that is, in species with large clutch sizes (**Fig. 3a**), short life expectancy (**Fig. 3b**), early age at maturity, (**Fig. 3c**) and short generation times (**Fig. 3d**). This supports the hypothesis that the level of demographic stochasticity in avian population dynamics is subject to life-history constraints on the possible range of variation in fecundity or survival, resulting in small values of σ_d^2 in long-lived species with small reproductive rates.

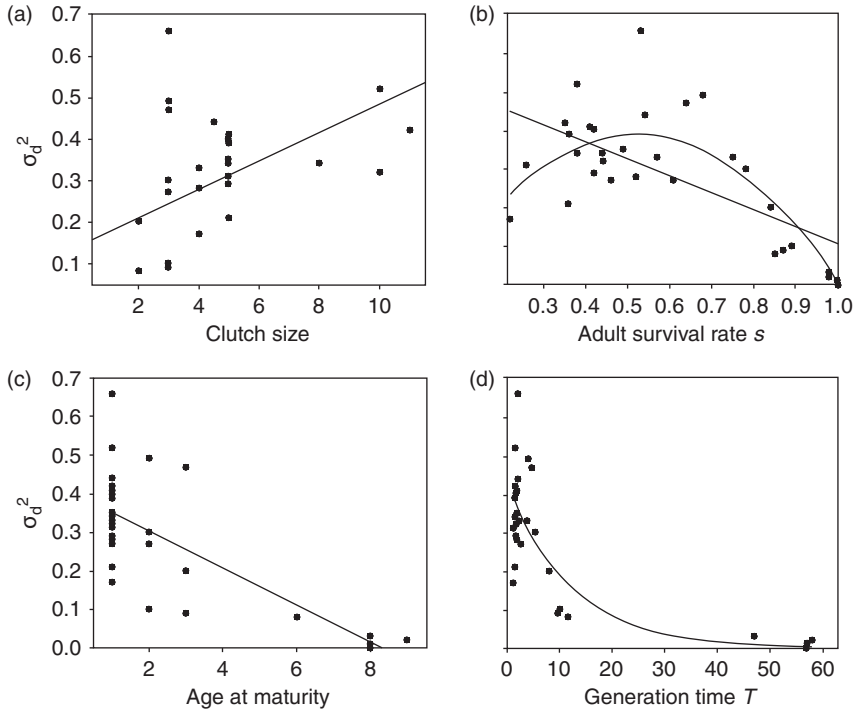


Fig. 3 The mean value across bird species of the demographic variance σ_d^2 in relation to (a) clutch size, (b) adult survival rate s, (c) age at maturity, and (d) generation time T.

A General Definition of Density Dependence of Age-Structured Populations

To analyze the effects of density dependence on the population dynamics, let us first consider a simplified life history in which individuals mature at age 1 year with adults having age-independent fecundity and survival. We can then write the dynamics $N(t) = \lambda[N(t-1)]N(t-1)$, where $N(t)$ is the population size in year t and $\lambda[N(t-1)]$ is the density-dependent finite rate of increase. The strength of density dependence γ in such a model can be defined as the negative elasticity of the population growth rate λ with respect to changes in population size N, evaluated at the carrying capacity $K : \gamma = -\left(\partial \ln\lambda / \partial \ln N\right)_K$. This approach can be extended to an age-structured density-dependent life history in which the total density dependence in the life history, D, should be defined as the negative elasticity of the population growth rate per generation, λ^T, with respect to the change in the size of the adult population when fluctuating around the carrying capacity, so that

$$D = -\left(T\frac{\partial\ \ln\lambda}{\partial\ \ln N}\right)K \qquad [1]$$

where T is the generation time. Thus, the annual rate of return to equilibrium then becomes $\gamma = D/T$.

Population Dynamics of Birds in Relation to Generation Time

This general definition of density dependence (eqn [1]) facilitates comparison of population dynamics across species with different types of life history. Here the author illustrates this approach by comparing the stochastic density-dependent population dynamics of different bird species. We assume that the expected adult annual survival and fecundity rates are independent of age, that density dependence is exerted by the adult fraction of the population on any combination of juvenile and adult vital rates, and, finally, that deviations of the adult population at time t from equilibrium $x(t) = N(t) - K$ are expected to be small or moderate. Based on these assumptions, we obtain a linearized autoregressive model with time delays from 1 to α years:

$$x(t) = \sum_{i=1}^{\alpha} b_i x(t-i) + \omega(t) \qquad [2]$$

where b_i is the autoregressive coefficient for time lag i, $\omega(t)$ is a noise term with mean zero and variance σ_ω^2, describing environmental stochasticity, including transient fluctuations in age structure and autocorrelations due to long-term fluctuations in the biotic or abiotic environment.

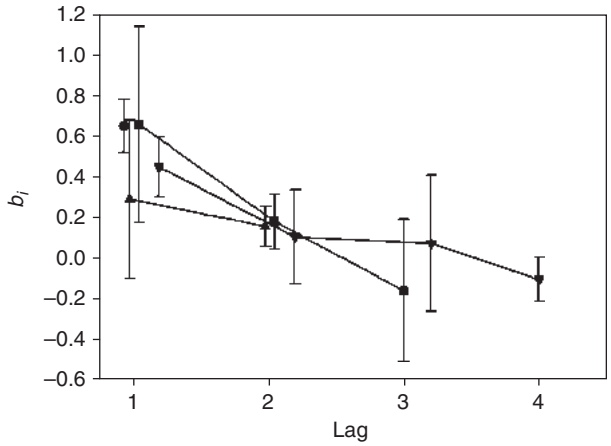


Fig. 4 Mean values (±SD) of the autoregression coefficients b_i for different lags in the population dynamics of birds in relation to variation in age at maturity of birds. Circles represent species that mature at 1 year, triangles age at maturity at 2 years, squares age at maturity at 3 years, and reversed triangles species that mature at 4 years or older.

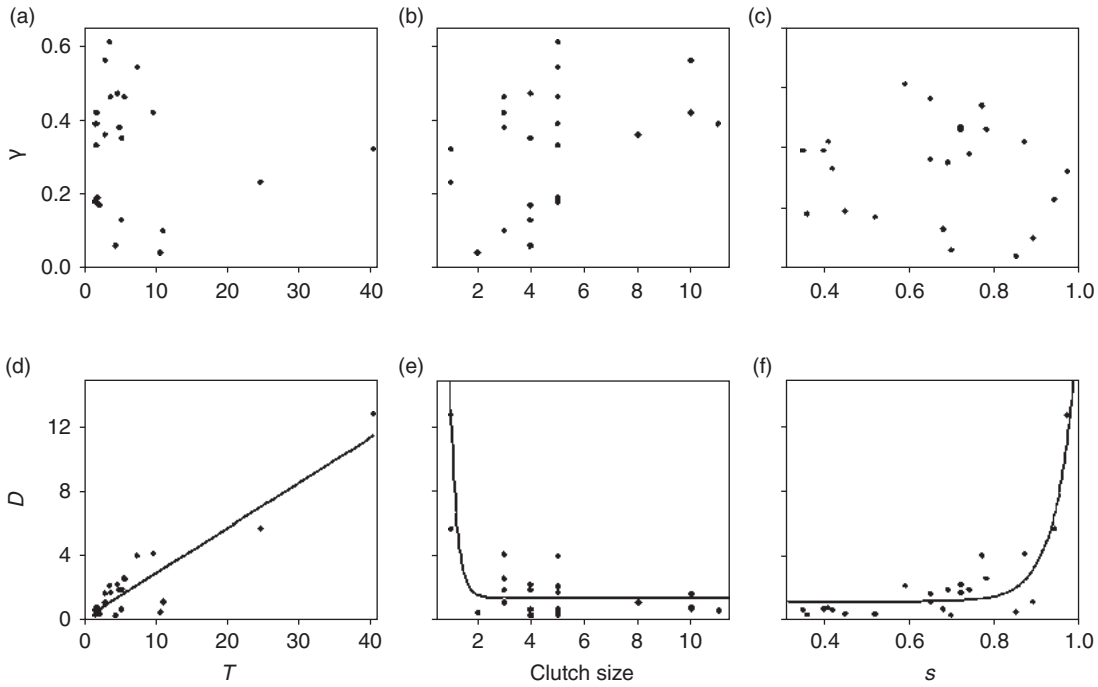


Fig. 5 The strength of density dependence in annual population fluctuations γ and in the total life history D in relation to (a), (d) generation time T; (b), (e) clutch size; and (c), (f) adult survival rate s of birds.

The autoregression coefficients b_i for species with age at maturity $\alpha \geq 2$ decreased with time lag (**Fig. 4**), indicating that the effects of the previous years' population sizes on current population size decreased with time. These autoregression coefficients b_i do not directly reveal the strength of delayed density dependence because they depend on life-history parameters as well as density dependence in the vital rates. In a species with $\alpha > 1$ with no density dependence in subadult or adult survival rates, b_1 equals the adult annual survival rate, s. Similarly, in a species that matures at 1 year, if $b_1 = 0$ the population autocorrelations for all time lags will be zero, corresponding to a white noise process for the population size $N(t)$ that indicates strong density dependence.

Again using the general definition of density dependence (eqn [1]), it is evident that there is a relationship between total density dependence in the life history and the autoregression coefficients

$$(1 - s)D = 1 - \sum_{i=1}^{\alpha} b_i \qquad [3]$$

In our data set from birds, the density-dependent effects at annual scale were independent of life history, including generation time (Fig. 5a) as well as clutch size (Fig. 5b) and adult survival rate (Fig. 5c). As a consequence, the stationary variance in the time series σ_N^2 was also independent of life history.

We have previously shown that many avian demographic traits such as clutch size and age at maturity scale closely with adult lifespan. Accordingly, we find that several features of population dynamics measured on a timescale of generations can be predicted from life-history characteristics. The strength of total density dependence in the life history D increased with generation time T (Fig. 5d) and adult survival rate (Fig. 5f) but decreased with clutch size (Fig. 5e). This implies that the effect on the population growth rate per generation of a change in population size was larger for long-lived than for short-lived species. Consequently, the rate of return to equilibrium measured in generations decreases with generation time T (correlation coefficient of $\log_{10}$-transformed values $= -0.73$, $P < 0.001$, $n = 23$).

To compare the residual variation in the population process, we must account for interspecific variation in age at maturity that will cause differences in the lag structure of the population dynamics. We first estimate the variance in the stationary distribution of the population sizes σ_N^2 in our model (eqn [2]) and then calculate the variance in the noise of a first-order process with a single time lag of 1 year, σ_ϵ^2, that will give the same stationary variance in population size as in the full model. Theoretical analyses show that variance for this white noise process for species with age at maturity larger than 1 year should be approximately equal to the environmental variance. In our bird data set, $\log_{10} \sigma_\varepsilon^2$ was independent of $\log_{10} T$ (Fig. 6a). In contrast, there was a highly significant linear increase in $\log_{10}(\sigma_\varepsilon^2 T)$ (Fig. 6b). Furthermore, the environmental variance for this process per generation ($\sigma_\varepsilon^2 T$) was closely related to life-history characters, that is, $\sigma_\varepsilon^2 T$ decreased with clutch size, but increased with adult survival rate. This suggests that environmental stochasticity per generation is greater for long-lived species than for species with short life expectancies. In contrast, interspecific differences in avian demographic stochasticity per generation are independent of life-history variation.

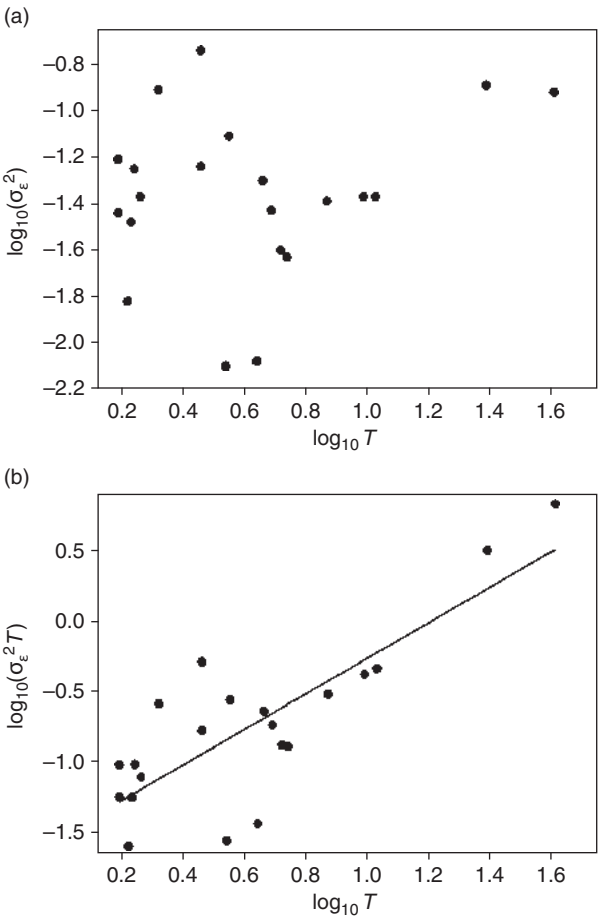


Fig. 6 (a) The residual variance in a first-order process, σ_ϵ^2, describing environmental stochasticity transient fluctuations in age structure as well as long-term autocorrelations in the environment and (b) in the total residual variance over a period of one generation $\sigma_\epsilon^2 T$, in relation to generation time T of birds.

Conclusion

Life-history traits show in most taxa a strong pattern of covariation. In many cases, species distribute them along a `slow–fast' continuum of life-history variation, for which generation time is a useful proxy. These analyses demonstrate clearly how many population dynamical patterns can be explained by the species' location along this continuum.

Acknowledgment

This work was supported by the Research Council of Norway (Strategic University Program in Conservation Biology).

See also: Evolutionary Ecology: Life-History Patterns. General Ecology: Age-Class Models. Terrestrial and Landscape Ecology: Plant Demography

Further Reading

Barker, R., Fletcher, D., Scofield, P., 2002. Measuring density dependence in survival from mark-recapture data. Journal of Applied Statistics 29, 305–313.
Caswell, H., 2001. Matrix Population Models, 2nd edn. Sunderland, MA: Sinauer.
Dennis, B., Ponciano, J.M., Lele, S.R., Taper, M.L., Staples, D.F., 2006. Estimating density dependence, process noise, and observation error. Ecological Monographs 76, 323–341.
Gaillard, J.M., Yoccoz, N.G., 2003. Temporal variation in survival of mammals: A case of environmental canalization? Ecology 84, 3294–3306.
Lande, R., Engen, S., Sæther, B.E., 2003. Stochastic Population Dynamics in Ecology and Conservation. Oxford: Oxford University Press.
Lande, R., Sæther, B.E., Engen, S., *et al*., 2002. Estimating density dependence from population time series using demographic theory and life-history data. American Naturalist 159, 321–332.
Sæther, B.-E., Bakke, Ø., 2000. Avian life history variation and contribution of demographic traits to the population growth rate. Ecology 81, 642–653.
Sæther, B.-E., Engen, S., Moller, A.P., *et al*., 2004. Life-history variation predicts the effects of demographic stochasticity on avian population dynamics. American Naturalist 164, 793–802.
Sæther, B.-E., Lande, R., Engen, S., *et al*., 2005. Generation time and temporal scaling of bird population dynamics. Nature 436, 99–102.

Detritus☆

Martin Zimmer, Leibniz Centre for Tropical Marine Research, Bremen, Germany and University of Bremen, Bremen, Germany

Glossary

Detritus Particulate dead organic matter of plant, animal or microbial origin.

Degradation Chemical breakdown of chemically diverse detrital compounds through, e.g., light (photo-degradation), microbial activity, or digestive processes of detritivores.

Decay Physico–chemical and microbial processing of detrital matter, such as leaching or microbial degradation.

Decomposition Entire process of detrital breakdown through decay, promotion by detritivores and direct contributions of detritivore, such as feeding and digestion, and mineralization.

Detritivore Animal that feeds on detritus (mostly used in aquatic ecology).

Decomposer Microbial contributor to decomposition.

Mineralization Final step of decomposition: transformation of organic matter into inorganic compounds.

Saprophage Animal that feeds on detritus (mostly used in terrestrial ecology for an animal that feeds on leaf litter).

Spatial subsidy (Organic) Matter, e.g., detritus, that is transported passively or actively from one (donor) habitat to another (recipient) habitat where it serves as resource to consumers in the recipient habitat; spatial subsidies are allochthonous material, contrasting autochthonous material that is derived from the same habitat.

Detritus

The term detritus summarizes nonliving organic matter, be it of animal, plant, or microbial (protist, fungal, bacterial, archeal) origin. For decades, it remained an ongoing debate as to whether or not "detritus" includes living microorganisms that colonize and nutritionally utilize the organic matter or inorganic components incorporated in detrital pellets. However, detritus sensu *strictu* consists of particulate organic matter (POM) that is to be degraded and does not include the agents of detritus degradation nor inorganic compounds that cannot be degraded. In aquatic ecology, dissolved organic matter (DOM) is sometimes considered a significant portion of detrital matter (together with POM). Part of DOM is derived from POM through initial steps of decay (see below), and remarkable amounts of DOM are exuded from living organisms. Hence, DOM is, strictly speaking, not to-be-degraded dead organic matter but either a product of detrital leaching (see below, for distinction between detritus and its leachate) or, similar to fecal matter (see below), derived from living organisms rather than the result of a death event. On the other hand, detrital matter is essentially always microbially colonized, and hence, a distinction of detritus and the living microbial biomass might not be reasonable from an ecological point of view.

Herein, the focus will be on dead plant material, not including microbial or inorganic components, since this is the most significant contributor to detritus in most terrestrial and aquatic systems. This view coincides largely with what is called "litter" in terrestrial ecology, whereas "POM" in aquatic ecology, and "wrack" in marine ecology, may also include organic matter of animal origin. Thus, it is obvious that a definition of the term detritus that is commonly agreed upon beyond ecological disciplines can hardly be provided.

Origin

As is obvious from the above, the origin of detritus is manifold. Because of its overwhelming importance in terms of both quantity and significance for nutrient cycling, mostly plant litter (i.e., detritus of vegetal origin) will be considered here. Detritus of animal origin (necromass) should be named carcass or carrion, and animals feeding on such organic matter are necrophagous (rather than saprophagous or detritivorous); this and other potential contributors to the total detrital mass will be briefly discussed below.

Two sources of plant detritus, namely vascular plants and macroalgae, that can be subdivided further, will be considered here. Besides photosynthetically active tissue, be it leaves, needles or stems, some vascular plants produce woody tissue (i.e., twigs, branches and logs). Dead roots contribute to the belowground pool of detrital matter and may decompose into mineral nutrients quickly (fine roots) or remain particulate over months and years (woody coarse roots; see woody detritus).

Although terrestrial by origin, vascular plants grow in, and produce detritus into, the terrestrial and both the freshwater (freshwater macrophytes and spatial subsidy from land plants) and the marine environment (seagrasses, saltmarsh vegetation, and mangroves). Macroalgae solely occur in the marine environment, but some of their detritus is exported into the semiterrestrial environment of the littoral and supralittoral coastal area, along with marine macrophyte detritus. Under particular conditions,

☆*Change History*: November 2017. M Zimmer is the sole author involved. The entire text has been slightly modified and updated. All figures where transferred in color figures.

This is an update of M. Zimmer, Detritus, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 903–911.

 doi:10.1016/B978-0-12-409548-9.10918-2

such as in floodplain forests, initially terrestrial detritus that has been deposited in freshwaters may, along with small amounts of detritus of aquatic origin (i.e., algae), occasionally be relocated into the terrestrial system (e.g., upon seasonal inundation or, as in the case of the marine-terrestrial transition zone, upon extreme tide events). On the other hand, inundation of flood plains may import additional terrestrial material from the terrestrial into the aquatic system.

Thus, detritus plays an important ecological role both as autochthonous source of nutrients (see **Table 1**) and as spatial subsidy (**Fig. 1**), i.e., as allochthonous resource that is passed from a donor in one habitat to a recipient in a second habitat. Detritus of terrestrial origin is also the major energy source to many lentic freshwaters and low-order streams in forested areas, making up several hundreds of grams per meter square per year. This subsidy of terrestrial detritus to freshwaters is mostly significant during winter (i.e., after autumnal leaf fall and under conditions of low productivity of planktonic or benthic aquatic plants). Huge amounts of up to several kilograms (dry mass) of detrital marine macrophyte wrack are deposited per km shoreline at the upper margin of the intertidal zone of rocky shores and soft sediment beaches with each incoming high tide, summing up to hundreds of tons per kilometer shoreline per year. Owing to the strong influence of tidal water movements and coastal winds as well as fast decomposition rates of some but not all marine detritus, the resulting high-intertidal wrack line is subject to enormous temporal variability and dynamics in both quantity and quality of detritus. Along the same line, tropical intertidal mangrove forests and temperate intertidal saltmarshes trap allochthonous detrital matter, originating from marine algae and macrophytes, as well as autochthonous detritus

Table 1 Estimated average production of detritus in selected generalized types of ecosystems (based on different sources)

	Production ($g\ m^{-2}\ a^{-1}$)
Freshwater macrophytes	300
Freshwater microalgae	
Phytoplankton	20
Phytobenthos	20
Marine microalgae	
Phytoplankton	20
Phytobenthos	30
Marine macrophytes	
Seagrass beds	500
Marine kelp forests	800
Macroalgae	300
Terrestrial macrophytes	
Saltmarshes	800
Mangrove forests	900
Tropical forests	300
Deciduous forests	100
Boreal forests	50
Grasslands	75
Tundras	25

Note that variation around these averages may be high when comparing, e.g., temperate and tropical systems, or seasonal patterns within a particular system.

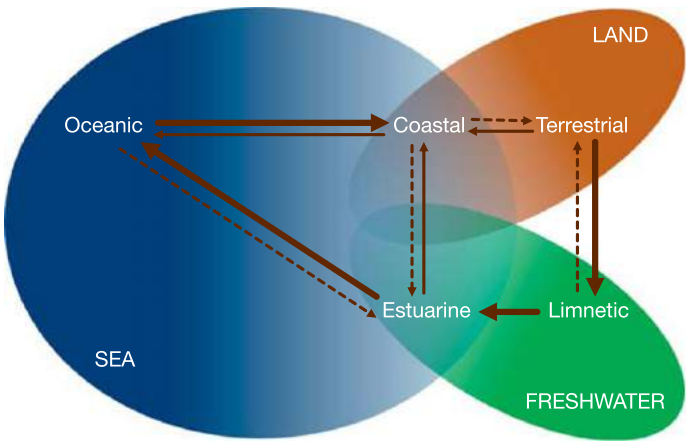


Fig. 1 Pathways of exchange of detritus among, and detrital connectivity of, ecological systems. The strength of an arrow indicates the significance and probability of the depicted connection and its direction.

among their aerial roots and stems. On the other hand, these intertidal ecosystems feed an annual amount of up to 1 t angiosperm detritus per kilometer coastline into estuaries, coastal waters and the open sea, and thus, contribute to the fertility of shelf zone.

In geographical zones with seasonal patterns of climatic conditions, leaf litter input into terrestrial and aquatic (mostly freshwater) systems (i.e., senesced leaves of broad-leaved deciduous trees) peaks during fall. However, some leaves are shed while still green and photosynthetically active, owing to physical disruption, be it through animal action or wind. The needle litter of coniferous trees is shed almost evenly over the year (but see *Larix decidua*, the European larch), but exhibits some rhythmicity of, e.g., 4 (Scots pine, *Pinus sylvestris*) or 7 (Norway spruce, *Picea abies*) years. It is mostly through physical disruption, but also senescence, that twigs and smaller branches of trees fall onto the forest floor or into forested water bodies. By contrast, bigger branches and logs become detritus through anthropogenic action (silviculture and logging) or upon death of the tree (e.g., through storm events or lighting strikes). All these events are more or less evenly distributed seasonally. Many aquatic vascular plants—freshwater macrophytes, marine seagrasses and intertidal saltmarsh vegetation—exhibit strong seasonality with increased release of detritus during fall, in some cases accompanied by a total die-off of above ground biomass. This seasonal pattern gets enforced by storm-driven wave action that predominates in fall and to a lesser extent in spring. Marine macroalgae, finally, show a pattern of detritus production that is similar to, but less marked than in, aquatic macrophytes. In tropical rainforests, the seasonality of litter fall seems to be correlated with rainfall seasonality, exhibiting peak litter fall after the dry season. However, species-specific litter fall dynamics result in a much more evenly distributed production of vegetal detritus. Two of the three predominant tree species in North Brazilian mangrove forests exhibit only little seasonality in litter fall, but *Avicennia germinans* sheds most of its leaves between the end of the rainy season and the start of the dry season.

Composition

The species composition of detritus in different ecosystems reflects the primary production of the respective autochthonous and allochthonous sources. Accordingly, it is necessary to distinguish between (terrestrial) angiosperms, aquatic macrophytes and algae (cf. **Fig. 1**). Terrestrial plant detritus has a substantially wider C:N (and C:nutrient) ratio than most aquatic organic matter, and is thus more difficult to decompose. An even narrower C:N ratio than in aquatic vegetal detritus is characteristic of detritus of animal origin (see below) the significance of which varies temporally and between distinct habitats. Algal detritus contains few and little recalcitrant compounds, such as cellulose, lignins or phenolics. Among terrestrial detrital sources, these parameters vary both among species and within a given species with environmental conditions (**Table 2**). Water-soluble compounds are readily leached off the detritus during early decomposition stages, particularly under moist or submerged conditions. Upon decay, the C:N ratio of the detritus-microbe conglomerate usually significantly drops, owing to the developing dense microbial biomass that accumulates nitrogen from the surrounding environment as well as leaching of carbonic but retention (immobilization) of nitrogenous detrital compounds.

Thus, wrack deposits in the high intertidal can be rich in nitrogen and poor in structural defense compounds, if they mainly consists of macroalgal material and animal necromass, or they are rich in cellulose and lignins but poor in nutrients, if mostly angiosperms, be it subtidal seagrasses, intertidal saltmarsh vegetation, or terrestrial vegetation, contributed to the wrack. Similarly, it is essential to freshwaters whether the detrital source is aquatic macrophytes and planktonic or benthic algae or terrestrial angiosperms. In the terrestrial realm, we distinguish between monocot detritus in grasslands and dicot detritus in forests, the latter, in turn, originating from either shrubs or woody plants, being either deciduous, evergreen or coniferous trees.

Decay and Decomposition

Detritus of whatever origin is decayed through leaching of water-soluble compounds of mostly low molecular mass and the action of microbial decomposers (decay: see Glossary). Feeding by detritivores (mostly in aquatic ecology; Latin: *detrítus*: scree, rock

Table 2 Comparison of terrestrial and aquatic sources of detritus with respect to estimated ranges of contents (%) of significant determinants of detrital quality to decomposing organisms (based on different sources)

	Limnetic		*Marine*		*Terrestrial*
	Angiosperms	*Algae*	*Angiosperms*	*Algae*	*Angiosperms*
Cellulose	10–30	0–25[a]	20–30	0–25[a]	20–50
Hemicellulose	10–20	–[b]	10–15	–[b]	10–30
Lignin	10–20	–[c]	15–20	–[c]	10–30
Phenolics	1–15	0–5	1–10	0–10	2–20
Lipids	1–5	1–5	1–5	0.5–5	1–10
C:N ratio	10–30	10–15	20–30	5–25	20–50
Nitrogen	1–5	1–5	2–5	1–5	0.1–5

[a]Among algae, it is only green algae (Chlorophyta) that contain cellulose in their cell walls.
[b]Brown (Phaeophyta) and red algae (Rhodophyta) contain alginate, carrageenan or agar, being structurally and functionally similar to hemicelluloses, in concentrations of up to 30 %, but hemicelluloses are restricted to angiosperms.
[c]Some green algae (i.e., Charaphyceae) contain lignin-like precursors, but lignins are restricted to angiosperms.

debris; *voráre*: to gobble, to gulp) and saprophages (mostly in terrestrial ecology; Greek: *saprós*: rotten, foul; *phageïn*: to feed) directly and indirectly contributes to the mass loss and chemical degradation of detrital matter and results in decomposition (see Glossary). Besides the purely physical process of leaching, biochemical—degradation by microbes and digestion by detritivores—and biological processes—dislocation, fragmentation and numerous animal-microbe interactions—play major roles in the mass loss of detritus (Fig. 2). Thus, decomposition dynamics can mathematically be described through exponential decay:

$$M_t = M_0 e^{-kt}$$

$$-k = \frac{\ln\left(\frac{M_t}{M_0}\right)}{t}$$

with M_t: mass of detritus at time t; M_0: initial mass of detritus (at time 0); t: time (in days); $-k$: decay rate. Decay rates are usually expressed per day (d^{-1}), but may also be given per year. Wood decay (see below) depends on the diameter of the twig, branch or stem, and is usually measured by means of tissue toughness or density.

Because of having received most of scientific interest in decomposition processes, leaf litter (i.e., annually shed leaves of deciduous trees) will be used as a model herein to exemplify decomposition processes. While the outline presented here can be generalized in first approximation, decomposition of other types of detritus may differ in detail to varying degree. Mass loss rates, usually used as a measure of decomposition rates, not only depend upon the type of detritus (see Table 3), but also on numerous environmental factors, such as moisture (in terrestrial habitats), temperature or the presence and activity of detritivorous animals. Accordingly, detritus can be classified as decomposing fast ($-k > 0.01$; $>1\%\ d^{-1}$), medium ($0.001 < -k < 0.005$; 0.1%–0.5% d^{-1}) or slowly ($-k < 0.005$; $< 0.1\%\ d^{-1}$) (see Table 3).

Ecological significance

It was only during the last couple decades that we began to understand or even appropriately acknowledge the essential role of detritus for within- and between-ecosystem cycles of matter and energy. Through the release of nutrients upon decomposition, detritus plays a major role in ecosystem processes by serving as the source for nutrient cycling and provision to primary producers in all biotopes. In many systems, as different as marine kelp forests, subtidal seagrass meadows, intertidal saltmarshes or mangroves, and deciduous or tropical forests, 80% or more of the annually produced plant biomass is recycled through the detritus pathway rather than consumed by herbivores. In terrestrial habitats, organisms involved in detritus decomposition are responsible for more than 95% of the total community metabolism. In freshwaters, detrital matter, be it dissolved or particulate (see above), greatly exceeds the organic matter present in living microorganisms, plants or animals.

Leaf litter

When freshly fallen plant litter is exposed to rain or dew, water-soluble compounds (e.g., amino acids, simple sugars or phenolics) are rapidly lost through leaching. As long as the leachate remains in contact with the litter as water film, this early-stage leaching accelerates litter degradation through the promotion of microbial activity and increased palatability of the litter to detritivorous

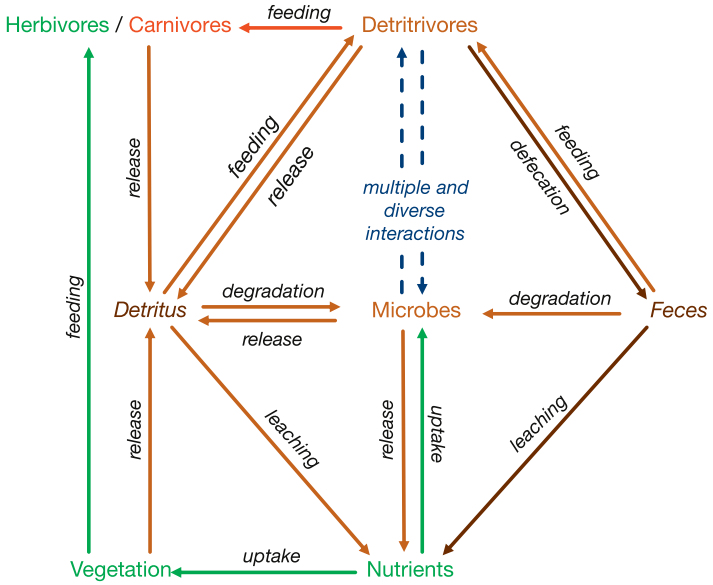


Fig. 2 Fluxes of matter among, and biotic interactions of, different compartments of the detrital subsystem in aquatic and terrestrial habitats. Broken arrows between detritivores and microbiota depict the diverse trophic, symbiotic and synergistic interactions of these central players in decomposition processes.

Table 3 Ranges (based on values reported in the literature) of decay rates ($-k$ d^{-1}) for selected types of detritus decomposing in different habitats

	$-$ k (d^{-1})	*Decomposition habitat*
Freshwater detritus		
Macrophytes	0.01–0.04	Aquatic
Marine detritus		
Macroalgae	0.0–0.07	Subtidal
	0.005-∞	Intertidal
Seagrass	0.007–0.02	Subtidal
	0.005–0.01	Intertidal
Terrestrial detritus		
Leaf and needle litter		
Narrow C:N	0.01–0.007	Aquatic
	0.01–0.2	Terrestrial
Wide C:N	0.001–0.01	Aquatic
	0.001–0.01	Terrestrial
Wood	0.001–0.007	Aquatic
	0.001–0.003	Terrestrial
Fine roots	0.001–0.02	Belowground

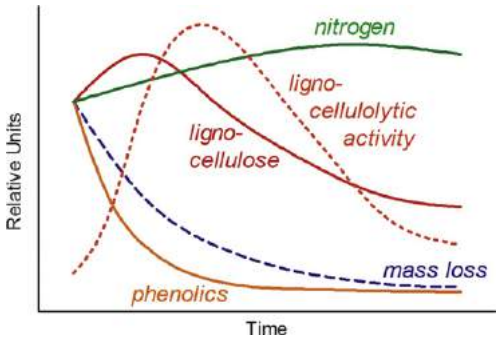


Fig. 3 Simplified scheme of changes upon decomposition in detrital mass (broken blue line) and contents of selected compounds (solid lines) [in % of initial value], as well as microbial activity against lignocellulose (dotted red line) [arbitrary units].

animals. The amount of mass lost through leaching varies according to both the leaf species and environmental conditions such as moisture. Physical disruption, e.g., through frost, drying and re-wetting cycles, sun light (UV), but also fragmentation upon feeding by detritivores, further promotes the loss of water-soluble compounds.

The remaining detritus itself, however, becomes enriched in larger and more recalcitrant compounds, as readily degradable substances are lost. Along the same line, leaching under water results in extensive release of dissolved organic matter (DOM), and differences in the chemical composition ("quality") of detrital sources are leveled out during early decay. Once dissolved in water, be it an aquatic environment or interstitial soil or pore water, DOM is prone to chemical degradation through microbial activity, eventually resulting in the formation of humic substances of varying chemical composition and the release of mineral nutrients. The other significant contributor to humic substances is the degradation of recalcitrant proteins and lignocellulose.

While leaching results in initial mass loss of detritus, the relative content of recalcitrant detrital compounds increases during the initial phase of decomposition (**Fig. 3**). At the same time, rapid microbial colonization leads to both a spatial accumulation and immobilization of nitrogen and increasing degrading activity on the detrital surface. Owing to a successive shift from microbes that utilize easily accessible low-molecular mass compounds to those that degrade lignocellulosic compounds, lignocellulolytic activity peaks after considerable mass loss through leaching and early degradation. Throughout late decomposition stages the formation of diverse humic substances due to the degradation of proteins, cellulose and phenolic compounds (including lignins) stabilizes the remaining detrital mass through inhibiting degradative microbial activity. In soils and sediments, this chemically stabilized organic matter is further decomposed at only very slow rates, particularly under anoxic conditions. Thus, huge stores of organic matter accumulate in sediments below marine seagrass beds or intertidal saltmarshes and mangroves and are essentially prevented from being transformed into climate-active gases, such as CO_2 or N_2O. Anaerobic decay, however, releases methane (CH_4), another potent climate gas, unless high sulfate concentrations inhibit methanogenesis.

The decay and decomposition rate, determining the release of nutrients to be available for growing plants and microorganisms, depends on both physical and chemical detritus characteristics (see **Tables 2** and **3**) and the activity of organisms involved in

decomposition. Accordingly, some types of leaf litter (e.g., alder, elder or chestnut) decompose within a couple months, while others (e.g., oak, beech or eucalypt) may last for years. Needle litter decomposes even more slowly. Microbial colonizers of detritus act in processing their substrate, rendering it an acceptable food source for detritivores. Most of the detrital biomass, particularly recalcitrant compounds such as phenolics and lignocellulosic cell wall compounds, requires (mostly aerobic) microbial activity for its degradation, which is, hence, slowed down under anoxic conditions, such as in (water-logged) soils and sediments. Thus, microbial litter colonizers serve detritivorous animals twofold. They precondition their detrital food sources and they are utilized as readily available source of (partly essential) nutrients. On the other hand, microorganisms are supported in degrading detrital matter through fragmentation by animals (see below), resulting in a significantly increased detrital surface. Further, microbial cells and propagules that survive the passage through a detritivore gut find themselves on a favorable substrate, detritivore feces, spread on a larger scale than microbial propagules themselves could achieve, and re-location of detrital particles by detritivores into favorable microhabitats facilitates microbial activity.

Consumption of detritus by detritivorous animals is the most obvious mode of mass loss. Detritivores of leaf litter belong to numerous diverse taxa, including molluscs, crustaceans, myriapods, insects, mites, nematodes and annelids. There is good evidence for general preferences of detritivores for some species of plant litter over others, and, as it is driven by detrital characteristics, such as the content of nitrogen and calcium, the C:N ratio or leaf thickness and toughness, this patterns seems to hold beyond biogeographical borders. Some consumers meet their decision based on the microbial community that inhabits the litter, but since microorganisms colonize different litter types at different species compositions (and densities), the outcome of this feeding behavior is similar and roughly reflects both microbial colonization and decay patterns. Typically litter of Betulaceae, Ulmaceae, Oleaceae, and Aceraceae is preferred over that of Fagaceae and gymnosperm trees, and litter of dicotyledonous plants over that of monocots. Toughness of plant tissue, and contents of nutrients and repellents, respectively, are probably major reasons for the observed feeding preferences. Generally, high nitrogen contents promote decomposition processes, while phenolics deter feeding by detritivores, impair enzymatic degradation processes, and thus, slow down decomposition. Thus, the loss of phenolics through leaching comprises an important initial step of decomposition that determines subsequent processes, and overwintered litter which is low in phenolics is a preferred food source to detritivores and decomposes faster than fresh litter. Along the same line of argument, terrestrial leaf litter loses deterrence upon submergence in water. These drivers of preferences interact with each other, and decomposition rates cannot be predicted by only one of them. Thus, the leaf litter of the mangrove *Avicennia germinans* in Northern Brazil is tougher than the litter of *Rhizophora mangle* but contains less phenolics—whereas detritivorous crabs prefer *Rhizophora* litter, mass loss of *Avicennia* litter proceeds much faster.

Similarly, most detritivores exhibit a marked preference for litter that carries high microbial biomass, exerting high microbial activity. This may proximately be due to phagostimulating compounds present in the microbial biomass. Further, detritivores obviously gain from feeding on microbially inoculated litter, but the ultimate reasons for this have not been unambiguously identified. Microbial biomass may significantly contribute to nutrition by providing essential nutrients, and some detritivores (e.g., terrestrial isopods) prefer feeding on those microbes that they can readily digest. Ingested microbiota may also assist in digestive processes by releasing digestive enzymes that can even break recalcitrant compounds such as lignocellulose. Alternatively, or in addition, microbial processing of the litter prior to detritivore consumption may be the determinant for this feeding preference.

Feeding by detritivores results in fragmentation of those detrital particles that are not ingested. Fragmentation of detritus increases leaching (see above) but in the midterm reduces nutrient release through immobilizing nutrients in increasingly produced microbial biomass; in the long term, however, fragmentation increases nutrient availability by stimulating microbial activity. Grazing on litter-colonizing microbes by detritivores may be expected to reduce microbial biomass. However, in part due to selective grazing on inactive or senescent cells and inducing nutrient release and compensatory re-growth, microbial activity and thus possibly even microbial biomass are increased in the long run. Since detritus has, on one hand, to be processed by microbes before detritivores feed on it, while, on the other hand, microbes are supported in decomposing the litter through fragmentation by soil animals, the interactions of the soil fauna and microbiota are, at least in part, considered truly mutualistic (**Fig. 2**).

Leaf litter decomposition on land

Early estimates suggested that the soil fauna contributes $<5\%$ to overall decomposition. Thus, indirect contributions to decomposition processes, promoting microbial decay, were assumed to prevail. These estimates, however, were based on respiration data and did not account for the large amounts of fragmented, ingested and digestively utilized litter. Currently, there seems to be common agreement that the contributions of microorganisms and animals are roughly equivalent. On average, soil animal activity increases decomposition by almost 25% through indirect promotion of microbial degradation of detrital compounds. In addition to those interactions of microorganisms and detritivores that are common to all habitats where decomposition is mediated by animals, the saprophagous soil animals control microbial activity during decomposition processes through the creation of favorable conditions for microbial decomposers and dislocation of leaf litter fragments into these conditions, as well as through dispersal of microbial propagules and inoculation of leaf litter.

Mechanical breakdown of lignocellulosic material, remaining in the leaf litter after leaching, by chewing detritivores (shredders in aquatic ecology, macrofauna in terrestrial ecology) facilitates its enzymatic degradation. Fragmentation of detritus positively affects the biomass and/or activity of the detritus-colonizing microbiota. Microbial biomass and activity increases when fragment size drops, but fragmentation may also inhibit fungal growth when particle size drops below a certain level of about 0.3 mm. Thus, the ratio of detritus-colonizing bacteria and fungi, and hence further decomposition processes, may in part be determined by particle size. In turn, both a detritivore's ability to fragment litter and the fragment size resulting from fragmentation strongly depend on the litter type.

Upon decomposition and subsequent mineralization brought about by microbial activity, nutrients derived from the decaying and disintegrating leaf litter material are eventually incorporated into the soil, where they are either taken up by the vegetation and soil microorganisms or may be incorporated as humic substances into clay-humic complexes that are recalcitrant towards further degradation and hence act as long-term nutrient store. Anoxic conditions contribute to the long-term stability of organic matter. In floodplains and water-logged sediments of the intertidal, leachates from litter remnants and nutrients released from detritivore feces or microbial cells that are dissolved in the porewater may eventually reach the hyporheic zone of the adjacent freshwater or the coastal waters upon submarine porewater discharge, providing input to the aquatic pool of dissolved organic matter (DOM). It is already during these fluxes that degradation of DOM proceeds.

Leaf litter decomposition in water

Except for the fact that leaching of water-soluble compounds proceeds much faster in water than on land, the basic processes of decay and decomposition are considered similar. Extensive leaching during early decay diminishes differences among detrital sources in terms of quality as resource for microbes and detritivores. Thus, decay rates of different detrital sources are more similar than on land.

While forested freshwaters, shaded by canopies, are fed by the energy input from terrestrial plant litter, open water bodies without (or with little) shoreline canopy exhibit a higher significance of autochthonous primary production by phytoplankton, macrophytes and their attached algae. In these systems, the significance of allochthonous detritus for nutrient cycles is mainly during winter, when in situ primary production is negligible. Overall, however, freshwaters must be considered heterotrophic (i.e., they are fueled from outside). By contrast, marine systems may be net producers or net consumers, depending on both adjacent habitat types (and the corresponding amounts and types of detritus) and the prevailing water currents. For intertidal saltmarshes and mangroves the question of whether they are sinks or sources of organic matter is still being discussed controversially. Common agreement seems to exist that being sink or source depends on the environmental setting and may change over time. Thus, the pneumatophores of the mangrove *Avicennia* seem to capture drifting leaf litter more efficiently than the prop roots of *Rhizophora*, but only when the tidal level does not exceed the length of the pneumatophores in which case floating leaves of high buoyancy would drift away above the roots.

Owing to water movements, in contrast to the situation in the terrestrial environment with mostly negligible effects of wind, detritus accumulates at sites with little or no flow, e.g., among stems of saltmarsh plants, roots of mangrove trees, or during slack high tide, rendering the in situ detritus distribution patchy (see above). While these trapped accumulations provide valuable shelters and refugia for benthic animals, detritus accumulation at large scales may result in temporal creation of anoxic conditions underneath the patches that, in turn, will reduce decomposition processes, because many of these are oxygen-dependent, especially those that involve animal activity.

In aquatic habitats, particularly in freshwaters, it is mostly fungi (aquatic hyphomycetes) that are involved in the breakdown of POM, but benthic shredders interact with them in similar manners to how terrestrial detritivores interact with microbes. Considering DOM a contributor to detrital matter (see above) that makes up 10–25-times the annual amount of POM, planktonic bacteria are just as important in decomposition as benthic microorganisms acting on POM are. Thus, the dominance of fungi that prevail the benthic decomposition of angiosperm detritus diminishes when suspended detritus of algal or microbial origin, being mostly decomposed by pelagic bacteria, is considered. Such suspended detrital matter in the water column has been coined "marine snow," and its dynamics have been increasingly studied recently, after we started appreciating its potential importance for matter fluxes in the marine realm. Upon eventual sedimentation of suspended POM, both pelagic bacteria, sedimenting along with the POM they colonize, and benthic bacteria and fungi act together.

In both cases, it is essentially only DOM (and some fine-POM) that can be transported to, and enter, adjacent habitats. Via this transport, up to 10^9 t C of terrestrial detritus reach the open sea annually worldwide. By contrast, POM sediments and is decomposed by benthic organisms. Upon seasonal flood events, part of the benthic detritus may be relocated into terrestrial flood plains or intertidal saltmarshes and mangroves, but the major portion of detritus is eventually incorporated into sediments and the pore water (the hyporheic interstitial), where oxygen consumption through aerobic degradation processes may result in anoxic conditions.

Vegetal detritus

World-wide, the annual plant biomass production exceeds animal biomass production by a factor of about 10. Thus, plant detritus is much more important in terms of amount and availability for nutrient release through the action of microorganisms and animals (decomposition: see above) than animal detritus. Accordingly, the ecological significance of vegetal matter decomposition has received a great deal of attention, but the decomposition of animal products (see below) has received relatively little. Among plant detritus, leaf litter may be the most prominent type of detritus in most ecosystems (see above), but depending on the biotope considered needle litter (in boreal forests), monocot litter (in grasslands and wetlands), or algal and angiosperm wrack (in coastal areas) may comprise the detrital pool predominantly or even entirely. The belowground compartment of terrestrial habitats and angiosperm-vegetated soft sediments is partly fueled by root detritus that decays in place unless being transferred to the oxic surface by bioturbation or human action.

Plant litter of whatever origin (see above) serves as substratum and food source for diverse microorganisms, be it bacteria, archeae, yeasts and fungi, microalgae or protists. Along with the plant litter itself, being of relatively low nutritive value (see above), these microorganisms are ingested by detritus-feeding animals and utilized as readily available and digestible

(supplementary) food source (see above). Thus, the digestion of both plant litter and microbial biomass eventually releases nutrients that can be utilized by the vegetation as soon as they are excreted, egested along with fecal masses, or set free upon the death and subsequent decomposition of the detritivore.

Macrophyte detritus

Macroalgae are considered nutrient sinks in coastal waters, but when they become detritus, they turn to being a nutrient source upon decomposition. Owing to reduced water flows in stands of macrophytes (including seagrass meadows and the intertidal vegetation of saltmarshes and mangroves), remarkable amounts of detritus may accumulate between blades, fronds, stems or aerial roots. Similarly, the stagnant water body at slack high tide deposits drifting wrack of little buoyancy along the high tide mark (until the wrack will be resuspended and removed from the wrack line by the next higher high tide). Once deposited, it is utilized as food by various invertebrates, such as sea urchins, molluscs and crustaceans, and fish that feed on large detrital particles, but also by filter feeders that ingested suspended remnants of detritus fragmentation. Once the wrack is worked into the soft sediment by wave action or bioturbation, the infauna, such as annelids and nematodes, contributes significantly to its decomposition, just like the micro- and mesofauna does in terrestrial habitats.

Primary production in near-shore macrophyte beds can also be exported to, and utilized by, other marine and nonmarine systems. Energy transfer from coastal to inland systems is brought about by terrestrial consumers that enter the intertidal area to feed upon, and fuel terrestrial food chains with, marine food sources that have been deposited ashore by tidal currents. The very same tidal currents may remove detritus of marine and intertidal origin from the intertidal and coastal subtidal zones into estuaries or the open sea (cf. **Fig. 1**) where it is subject to decomposition under submerged conditions. For instance, 15% of the CO_2 globally fixed in plant biomass, enters the open sea as seagrass detritus, while seagrasses make up less than 1% of the marine plant biomass. Thus, detrital spatial subsidies that are transported beyond habitat borders, where they serve as resources, play a significant role in fueling both terrestrial and aquatic ecosystems.

Woody detritus

Woody debris, be it small twigs, branches or whole stems, contribute significantly to the detrital biomass in forests, freshwaters and coastal marine areas. These inputs derive from both natural and anthropogenic activities including senescing trees, natural disturbance (e.g., windfall and slope failure), silviculture and forest management. In temperate forests, the input of woody debris amounts to roughly 1 t ha^{-1} a^{-1}. Generally, wood decomposition is slow, but small twigs in the litter layer may decompose at almost the same rate as leaf litter. By contrast, the decomposition of logs and trunks roughly equals the life time of the tree. Infestation of the wood by wood-boring and -inhabiting invertebrate, e.g., molluscan shipworms or crustaceans in the marine and intertidal environment, or beetle larvae and termites in the temperate and tropical, respectively, terrestrial realm, may accelerate wood decay by enlarging the surface area and promoting microbial activity, or may contribute to decomposition through digestive processes.

Particularly coarse woody debris adds long-lasting unique habitat structure and resources to both terrestrial and aquatic habitats, by being an important source of energy and nutrients for microorganisms and detritivores, trapping sediments, offering protection against harsh environmental conditions, and serving as refugium from predation to both consumers and inhabitants of decomposing wood. Coarse woody debris increases the diversity of biological communities in aquatic and terrestrial habitats. Certain species of fish depend on wood in streams to survive and to spawn. Logs, trunks and rootmats still attached to fallen trees are a key structural component in streams that deflects flows in ways that scour stream pools and induce meanders, but also anchors streambanks against high flows and floods by fixing the sediment, thus, reducing erosion and sedimentation downstream. Woody debris also serves as food source (albeit of low quality, owing to low nutrient and high lignocellulose contents) to xylophagous detritivores. Even more important, the surface of woody debris is densely colonized by microbial biofilms that, in turn, are grazed upon by various different animals. Along this line, the density and species composition of the microbial biofilm drives the settlement of potential animal colonizers on the wood surface. As the development of the biofilm, in turn, partly depends on wood exudates, chemical compounds that are leached off the woody detritus, the chemical composition of the wood indirectly mediates its colonization by invertebrates.

In contrast to most other types of detritus, dead wood in forests is not necessarily in contact with the forest floor—standing dead trees are also subject to decomposition, but the position of the trunk (along with the tree species, the composition of the forest community, soil characteristics and environmental conditions) determines the processes of decomposition. Even more than for angiosperm litter of photosynthetic tissue, the degradation, or at least modification, of lignin is a prerequisite for wood decomposition. Thus, it is essentially fungi (albeit of various taxonomic affiliations) that drive wood decay, being limited by the extremely wide C:N ratio of roughly 300–1000 (cf. **Table 2**). While soft- and brown-rotting fungi hydrolytically degrade cellulosic and hemicellulosic moieties of the cell walls, leaving brownish lignin remains, white-rot fungi oxidatively degrade lignins, but concomitantly also hydrolyze cellulose and hemicellulose—the main goal of lignin degradation is not the nutritive utilization of aromatic carbohydrates, but the removal of an embedding matrix to gain access to fiber polysaccharides. Fungal attack of freshly dead wood is facilitated by insects that actively intrude the wood by burrowing holes and galleries. Xylophagous insects are by more than 90% Coleoptera, at later stages of wood decomposition being accompanied by Collembola and noninsects, such as millipedes, isopods and lumbricids. In the tropics, it is mainly termites that destroy rotting wood. Most of these invertebrates are by now known to either bear endogenous (genes for) cellulases, or they rely on microbial symbionts to provide the enzymes needed to breakdown lignocellulose. Wood-boring shipworms (Bivalvia: Teredinidae), colonizing submerged woody debris,

harbor bacterial symbionts in their gills that produce lignocellulose-degrading enzymes, whereas crustaceans (isopods and amphipods, e.g., *Limnoria* and *Chelura*) that share the same lifestyle produce endogenous cellulases.

While the role of (coarse) woody debris is well-studied in forests and freshwaters, there is a dearth of information on its importance in marine coastal ecosystems. In high-intertidal tidepools with little algal cover, woody debris increases habitat complexity and heterogeneity and is readily colonized by motile invertebrate in search for shelter from harsh environmental conditions and predation. In low-intertidal pools and the open coast, the complex habitat structure of algal communities increases the relative significance of woody debris and its microbial biofilm as supplementary food source at the expense of its role as habitat structure.

While leaf and needle litter in grass- and woodlands is more or less evenly distributed, albeit with patchily scattered hot spots of high substrate quality and decomposing activity, the distribution of woody debris in a given habitat is rather heterogeneous, resulting in patches of refugia and supplementary food sources by means of biofilms. Similarly, coastal detritus, be it macroalgae or angiosperms, accumulate in patches, either at the upper margin of the intertidal zone (wrack line or drift line), in areas of little water movement beneath stems, roots, shoots or thalli of macrophytes, or aggregated in drift mats floating subtidally (see above). In either case, the patchy nature of these types of detritus provide habitat heterogeneity, although the effects on the community of animals making use of these patches may differ substantially, being either favorable (see above) or detrimental in terms of reducing the oxygen availability and producing hypoxic or even anoxic microhabitats underneath and within these patches.

Roots

Living roots shape their immediate environment, the rhizosphere, by exuding water-soluble organic compounds into the pore-water and by aerating the soil (or sediment in the case of seagrass, mangroves or saltmarsh plants). Upon death, root material is prone to belowground decomposition. This particular environment is characterized by relatively stable temperature and moisture conditions (as compared to conditions above ground). Except for very slow decay in boreal or permafrost soils, the climatic settings of the aboveground environment prove to be much less significant as drivers of root decay rates than root chemistry. However, the belowground environment may be prone to developing hypoxic or even anoxic conditions (especially when submerged, e.g., in floodplains or the intertidal), and even small-scale scarcity of oxygen may hamper root decomposition.

The detritus of coarse roots that are lignified resembles woody debris, and their decay is similar to the decay of wooden structures. Fine roots, however, essentially lack the structural compounds of wood and are relatively rich in nutrients and water-soluble compounds. Thus, fine roots can decay at rates similar to those of leaf litter, but, interestingly, difference in fine root decomposition rates do not seem to mirror those of leaf litter of different species. Fine root decay rates are mediated mostly by root chemistry, i.e., C:N ratio and calcium content. It is mostly the endogeic micro- and mesofauna (infauna of sediments) that contributes to fine root decomposition.

Animal detritus

Carrion

In comparison to plant litter, carrion (i.e., remnants of dead animals) is characterized by significantly higher nitrogen contents (i.e., a narrower C:N ratio) and fewer recalcitrant or deterrent compounds. Thus, unlike plant litter, animal necromass provides spatially and temporally rare detritus patches of high quality. Large carcasses are usually skeletonized after 1–3 years, but, depending on climate, complete decomposition of small carrion may be achieved within 1–2 months. However, relatively little is known about how carrion decomposition affects ecosystem functioning, how the size of the carrion patch controls decomposition and nutrient release, or how the huge amounts of carrion derived from catastrophic die-offs influence the respective ecosystem. From what is known, it can be concluded that the contribution to nutrient cycling from the decomposition of animal remnants can be substantial and, in such cases, should be considered in the formulation of nutrient budgets, even beyond habitat borders (e.g., in the case of terrestrial carnivores catching fish and leaving the carcass prone to decomposition in their terrestrial environment). For instance, dead whales (or other large marine vertebrates, such as sharks or sea elephants) that sink to the deep sea floor can provide allochthonous nutrient input to a variety of (scavenging) deep-sea organisms, such as crustaceans, sea cucumbers, annelids, slime eels or sharks, for decades. Owing to the particular environmental conditions of the deep sea (low temperature, high pressure) and the scarcity of scavengers, decay and decomposition of these carcasses is slowed down remarkably.

In contrast to detritivores, feeding on vegetal detritus, most of which are invertebrates, necrophagous consumers of carrion are represented by a high number of both invertebrate and vertebrate species. Necrophagous invertebrates mostly belong to the insect orders Diptera and Coleoptera. Invertebrate necrophages colonize and utilize animal necromass in waves of occurrence that are typical for a given habitat and climate. By contrast, the species of the dead animal has little influence on the process of carrion decomposition. After several stages that differ in the prevailing biochemical processes and the involved necrophages, the last steps result in mummification and subsequent sceletonization. The sequence of waves of invertebrate necrophages may, however, at any point be interrupted by scavenging vertebrates that feed upon carrion at any stage of decomposition.

Fecal detritus

Some authors consider feces of detritivores part of the detrital pool, but this point of view does not account for the particular role of feces as an intermediate step in decomposition processes. Detritivore feces—animal-processed material of initially vegetal origin, and hence considered animal detritus here—are usually characterized by a narrower C:N ratio than the detritus they derived from. Besides almost unchanged detrital fragments, detritivore feces contain considerable amounts of amorphous material, being

the result of digestive processes, with about 30%–50% of the ingested detritus complexed in humic substances. This amorphous feces fraction can subsequently be rapidly lost during feces disintegration. However, the peritrophic envelop that surrounds the feces of many detrivores may counteract disintegration and slow down fecal decay in some cases.

Fecal pellets of detritivores are often more readily decomposed through microbial action and coprophagous animals than the original detritus. Consequently, their C:N ratio rapidly drops during aging. Coprophagy, being promoted by the suitability of feces as environment for microbial decomposers, results in a further mechanical breakdown of already fragmented organic matter, and thus, directly contributes to decomposition processes. Owing to the digestion of litter-colonizing microorganisms during the gut passage, freshly dropped feces of detritivores usually contain less active microbiota than the leaf litter they are derived from. Further, possibly through both particle fragmentation and differential proliferation potential for bacteria and fungi in gut lumen of detritivores, the bacteria:fungi ratio is in favor of bacteria in the feces of many detritivores. Further changes in the microbial community occur during feces aging and decomposition. Reasons for fast microbial colonization of feces are the high initial content in viable microbial cells and propagules, the favorable surface:volume ratio, and the high content in easily accessible nutrients of a predigested substrate.

Most features of detritivore feces also apply to another particular type of detritus, namely dung (i.e., fecal masses derived from large herbivores, frugivores, carnivores and scavengers on carrion). Among these, fecal masses of herbivorous mammals (e.g., cattle) attract a particularly rich fauna of coprophages. Vegetal (algal) material that survives, and is still active after, gut passage is not detritus sensu *strictu*, as detritus is, *per definitionem*, dead. Other characteristics, such as distribution and dynamics in space and time, rather resemble carrion. Alike rotting fruits and dead animals, dung is characterized primarily by its spatial and temporal concentration in discrete and ephemeral patches of high energy and nutrient content. These patches provide home and food to diverse and highly dynamic communities of consumers making short-term use of this transient energy source. Coprophagous animals are found among dipteran, coelopteran and micro-lepidopteran insects, as well as oligochaete annelids. The sequence of coprophagous and coprophilous animals that appear on and in dung and the processes they initiate are highly predictable, but in detail depend on the habitat and climate under investigation. Specialized coprophilous fungi, similarly, exhibit a clear sequence of utilization of their habitat, spores of early stages already being present in the dung when it is deposited by the herbivore. Apparently, users of dung are highly specialized, and their community is significantly different from what usually is referred to as decomposers or detritivores. Depending on climatic conditions, dung patches may become completely decomposed in few months but may also last for several years. Thus, as for most types of detritus, nutrient release is not sudden but extended over a longer time interval, so that detritus provides a long-term store and source of nutrients at the basis of ecosystems.

See also: General Ecology: Soil Ecology. Global Change Ecology: Microbial Cycles

Further Reading

Chave, J., Navarrete, D., Almeida, S., Álvarez, E., Aragão, L.E.O.C., Bonal, D., Châtelet, P., Silva-Espejo, J.E., Goret, J.-Y., von Hildebrand, P., Jiménez, E., Patiño, S., Peñuela, M.C., Phillips, O.L., Stevenson, P., Malhi, Y., 2010. Regional and seasonal patterns of litterfall in tropical South America. Biogeosciences 7, 43–55.

Coleman, D.C., Crossley, D.A., Hendrix, P.F., 2004. Fundamentals in soil ecology. Amsterdam: Elsevier.

Cragg, S.M., Beckham, G.T., Bruce, N.C., Bugg, T.D.H., Distel, D.L., Dupree, P., Green Etxabe, A., Goodell, B.S., Jellison, J., McGeehan, J.E., McQueen-Mason, S.J., Schnorr, K., Walton, P.H., Watts, J.E.M., Zimmer, M., 2015. Lignocellulose degradation mechanisms across the tree of life. Current Opinion in Chemical Biology 29, 108–119.

Cross, W.F., Benstead, J.P., Frost, P.C., Thomas, S.A., 2005. Ecological stoichiometry in freshwater benthic systems: Recent progress and perspectives. Freshwater Biology 50, 1895–1912.

Findlay, S., Sinsabaugh, R. (Eds.), 2000. Dissolved organic matter in aquatic ecosystems. San Diego: Academic Press.

Graça, M.A.S., 2001. The role of invertebrates on leaf litter decomposition in streams—A review. International Review of Hydrobiology 86, 383–393.

Graça, M.A.S., Bärlocher, F., Gessner, M.O., 2005. Methods to study litter decomposition: A practical guide. Dordrecht: Springer.

Higgs, N.D., Gates, A.R., Jones, D.O.B., 2014. Fish food in the deep sea: Revisiting the role of large food-falls. PLoS One 9. e96016 doi:10.1371/journal.pone.0096016.

Hobbie, S.E., Oleksyn, J., Eissenstat, D.M., Reich, P.B., 2010. Fine root decomposition rates do not mirror those of leaf litter among temperate tree species. Oecologia 162, 505–513.

Ihnen, K., Zimmer, M., 2008. Selective consumption and digestion of litter microbes by *Porcellio scaber* (isopoda: Oniscidea). Pedobiologia 51, 335–342.

Mackensen, J., Bauhus, J., Webber, E., 2003. Decomposition rates of coarse woody debris—A review with particular emphasis on Australian tree species. Australian Journal of Botany 51, 27–37.

Moore, J.C., Berlow, E.L., Coleman, D.C., *et al.*, 2004. Detritus, trophic dynamics and biodiversity. Ecology Letters 7, 584–600.

Orr, M., Zimmer, M., Jelinski, D.E., Mews, M., 2005. Wrack deposition on different beach types: Spatial and temporal variation in the pattern of subsidy. Ecology 86, 1496–1507.

Quadros, A.F., Zimmer, M., Araujo, P.B., Kray, J.G., 2015. Litter traits and palatability to detritivores: A case study across biogeographical boundaries. Nauplius 22, 103–111.

Silver, W.L., Miya, R.K., 2001. Global patterns in root decomposition: Comparisons of climate and litter quality effects. Oecologia 129, 407–419.

Storry, K.A., Weldrick, C.K., Mews, M., Zimmer, M., Jelinski, D.E., 2006. Intertidal coarse woody debris: A spatial subsidy as shelter or feeding habitat for gastropods? Estuarine, Coastal and Shelf Science 66, 197–203.

Treplin, M., Zimmer, M., 2012. Drowned or dry: A cross-habitat comparison of detrital breakdown processes. Ecosystems 15, 477–491.

Zimmer, M., Auge, H., von Wühlisch, G., Schueler, S., Haase, J., 2014. Environment rather than genetic background explains intraspecific variation in the protein-precipitating capacity of phenolic compounds in beech litter. Plant Ecology and Diversity 7. doi:10.1080/17550874.2013.871655.

Dominance☆

Helmut Hillebrand, University of Cologne, Cologne, Germany

Introduction

The coexistence of species in ecological communities is almost always asymmetrical such that some species are dominant, but most are rare. This inequality of species contribution is reflected by the degree of dominance in species composition and the identity of the dominant species. The analysis of dominance has been a cornerstone in general ecology, tightly linked to questions of community structure and diversity. Understanding the degree of dominance and the identity of the dominant species is implicit in studies on community assembly and organization, on causes and consequences of diversity, and on evolutionary and macroecological constraints of regional species pools.

Dominance in a local community can be highly obvious as one species monopolizes the use of space and visually characterizes the structure of the assemblage (**Fig. 1**). Such striking dominance of single or few species can be seen in many terrestrial and aquatic communities. However, even in communities with seemingly lower dominance, only a few species contribute the majority of the biomass or the count of individuals.

The degree of dominance can be characterized by the maximum proportion of individuals (or biomass) contributed by a single species in an assemblage (dominance ratio). The dominance ratio can fluctuate between 0 and 1, but values >0.5 are very common, both for individuals and for biomass (**Fig. 2**). Another way of addressing dominance is represented by rank–abundance curves, where species are ordered in ranks according to their proportional contribution. All classical rank–abundance models—broken-stick, log series, or neutral models—predict the numerical dominance of just a few species out of all species present. Similar relationships can be derived between rank and biomass instead of abundance. The shape of the rank–abundance curve is a strong visualization of dominance (**Fig. 2**A) and a readily assessable way of analyzing the effects of ecological interactions and abiotic drivers on species dominance.

The complementary term to dominance is evenness, which describes the equality of the distribution of proportions across species. Communities with less well pronounced dominance have a higher evenness (a higher degree of equality in the distribution of abundance or biomass on different species). Consequently, the correlation between dominance (as maximum proportion of one species) and evenness is negative (**Fig. 2**B).

Patterns of Dominance

Across all types of realms and habitat types, the number of individuals and the total amount of biomass are almost inevitably dominated by one or few species, whereas most species are rare. However, the proportion of numbers or of biomass made up by the most dominant species differs strongly even in closely related assemblages or spatially adjacent communities (**Fig. 1**). Published dominance ratios range from 0.2 to almost 1 for abundance as well as biomass. The proportion contributed by the three most dominant species consistently exceeds 50% and very often 75%. The degree of dominance differs highly with the exact way dominance is measured, but also across spatial and temporal gradients. In order to understand the mechanisms driving dominance within communities and in regional scales, it is important to identify some of this variation in dominance ratios.

Measure-Based Patterns

The dominance ratios calculated for biomass often differ strongly from dominance ratios calculated for abundance, which is mainly due to the fact that the identity of the species dominating abundance differs from the species dominating biomass. As a consequence, the rank of any species within a community according to the proportion of abundance and the proportion of biomass tend to be negatively correlated. Therefore, the maximum dominance found for abundance is often uncorrelated to dominance assessed by biomass (**Fig. 2**C).

As a consequence of these different outcomes, ecologists have to define whether dominance in abundance or dominance in biomass has to be the main target of their study. The answer to this question is mainly driven by the ecological processes focused on. Number-driven processes such as dispersal and intra- or interspecific encounters are best reflected by abundance dominance, whereas other processes such as competition for resources or energy flux are mainly affected by dominance in biomass.

☆*Change History:* March 2018. I. Martins updated the references.

This is an update of H. Hillebrand, Dominance, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 938–944.

 doi:10.1016/B978-0-12-409548-9.11158-3

Fig. 1 Two adjacent rock pools at the Swedish west coast. The right rock pool is strongly dominated by ephemeral green algae (*Enteromorpha* sp.), the left rock pool is characterized by a large snail population (*Littorina* sp.) and crustose algae. Picture courtesy of Monika Feiling.

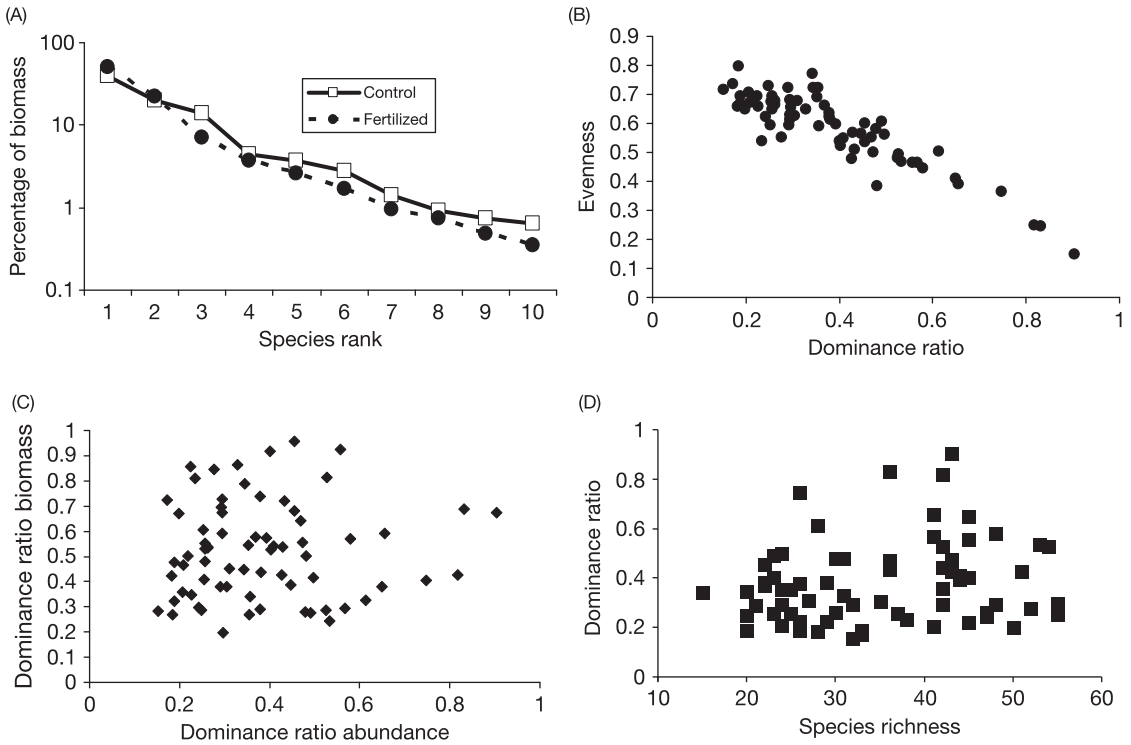


Fig. 2 Dominance patterns exemplified by data on benthic marine microalgae from the western Baltic Sea. (A) Relative biomass proportions of the ten most common microalgal species in control and fertilized treatments. (B) Correlation between evenness (measured as Pielous's index) and the dominance ratio in fertilization experiments from 1996 to 1998. (C) Correlation between dominance ratios calculated from abundances and dominance ratios calculated from biomass proportions. (D) Correlation between dominance ratios calculated from biomass proportions and species richness. All data are from Hillebrand, H. and Sommer, U. (2000). Diversity of benthic microalgae in response to colonization time and eutrophication. *Aquatic Botany* **67**, 221–236.

The degree of dominance is largely independent of the number of species present in a local assemblage, as becomes evident by the lack of correlation between taxonomic richness and dominance ratios (Fig. 2D). This pattern is corroborated by the repeatedly found lack of correlation between evenness and taxonomic richness in many community types. This uncoupling is mainly due to the fact that the overall species number is a consequence of the number of rare species, which have negligible influence on the share of dominant species in a community. In other words, a species can be highly dominant independent of the length of the tail of rare species in the assemblage.

Spatial and Temporal Patterns

The most ubiquitous spatial gradient in diversity is the latitudinal decline of species richness from the Tropics to the poles. This richness gradient is assumed to be accompanied by a gradient in evenness, where single species gain more dominance in temperate

or boreal areas compared to tropical ones. However, systematic evidence is rare for this pattern or similar changes along altitudinal gradients. Many macroecological studies on range–abundance relationships have shown that widespread species have higher local densities, but the mechanisms for these patterns are still widely discussed and comprise purely statistical reasons, population dynamics, or resource use.

In addition to geographical gradients, dominance also varies along environmental conditions. Dominance tends to be high under harsh environmental conditions, such that entire organism guilds are represented by single species. An example are saltmarshes, where the most stressful marine side of the marsh is often inhabited by few species (e.g., *Salicornia*, *Spartina*), whereas, more landward (and less affected by salt), the number of species increases and the dominance of a single species decreases. Also in freshwater–terrestrial contact zones, the aquatic side of the gradient is often dominated by single terrestrial species adapted to the stressful conditions such as reed. The reason for this pattern is the strong selection for certain adaptations in the harsh environments, which favors only few specialized species.

Also gradients in productivity are often related to dominance patterns. Dominance tends to be high under high productivity, whereas habitats of low to intermediate productivity seem to be less probable to be dominated strongly by a single species. Fertilization studies often enhance the dominance ratio (cf. **Fig. 2**A). Many studies concur on a decrease in evenness with increasing community biomass, which then translates into an increase of dominance with increasing biomass production.

Dominance varies not only over spatial gradients but also in time. This can be strongly seen not only in chronosequences of succession, but also in many paleoecological investigations. In early primary succession, the habitat tends to be harsh and few species dominate early seral stages. Later on, dominance ratios might decline when more species contribute more evenly to community biomass. When the chronosequence remains undisturbed, late successional stages may exhibit increasing dominance ratios again, as highly competitive species displace other organisms. Also in secondary succession, dominance is also most pronounced in early developing species guilds (colonizers) and in late stages (competitors), whereas mid-stages show lower dominance levels.

Paleoecological analyses of pollen or of diatom community composition nicely show the shifts in dominance ratios over time. Sequences of strong single species dominance are often interrupted by phases of more even abundance distribution. These shifts are often related to alterations of the abiotic environment on local (changes in nutrient input or water chemistry) or regional (climate) scales.

For all these temporal and spatial gradients, it should be noted that not only the degree of dominance changes but also the identity of the dominant species. Species dominating early and late successional stages differ as do species between the extremes of environmental or spatial gradients. Both the dominance ratio and the identity of species have strong consequences for the functioning of ecosystems.

Mechanisms of Dominance

The primary questions for the understanding of dominance is: what keeps most species rare and makes few species dominant? Even in neutral communities consisting of ecologically equivalent species, overall biomass or abundance are dominated by few species. Neutral theory predicts that even in the absence of any trait differences between species, stochastic events will create dominance for some and rarity of many species. The identity of the dominant species and the degree of dominance follows not from any ecological interactions or environmental conditions, but depends on ecological drift, dispersal success, and timing of colonization. However, most realized communities deviate from neutrality and the primary question posed above requires mechanistic answers. These answers will strongly differ with the scale of the analysis, for example, on the local scale of interacting populations (i.e., communities) and on the regional scale of species pools. Within communities, the main mechanisms creating dominance are interactions within and between species as well as the abiotic factors constraining these interactions. Between communities, dispersal and spatial niche differentiation are strong forces regulating how the regional species pool transfers into local community structure. Addressing the regional species pool itself, the evolution and descent of certain lineages is the major driver of regional dominance.

Local Communities

Competition

Interspecific competition generally leads to a reduction of the contribution of poor competitors to the community, whereas superior competitors are able to gain dominance. Thus, competition is a process strongly increasing dominance. The degree of competitive dominance depends on two factors: (1) on the asymmetry of the competition and (2) the time for superior species to develop their dominance (see further below). The former aspect depends on the distribution of traits in an assemblage of competing species. If traits are similar, and thus competitive advantage of a certain trait low, dominance will be low. On the other hand, strong dominance can be expected if traits are skewed and competitive advantage of a certain trait is large (**Fig. 3**). Competition of terrestrial plants for light is a lucid example for such an asymmetric competition for a unidirectional resource, as

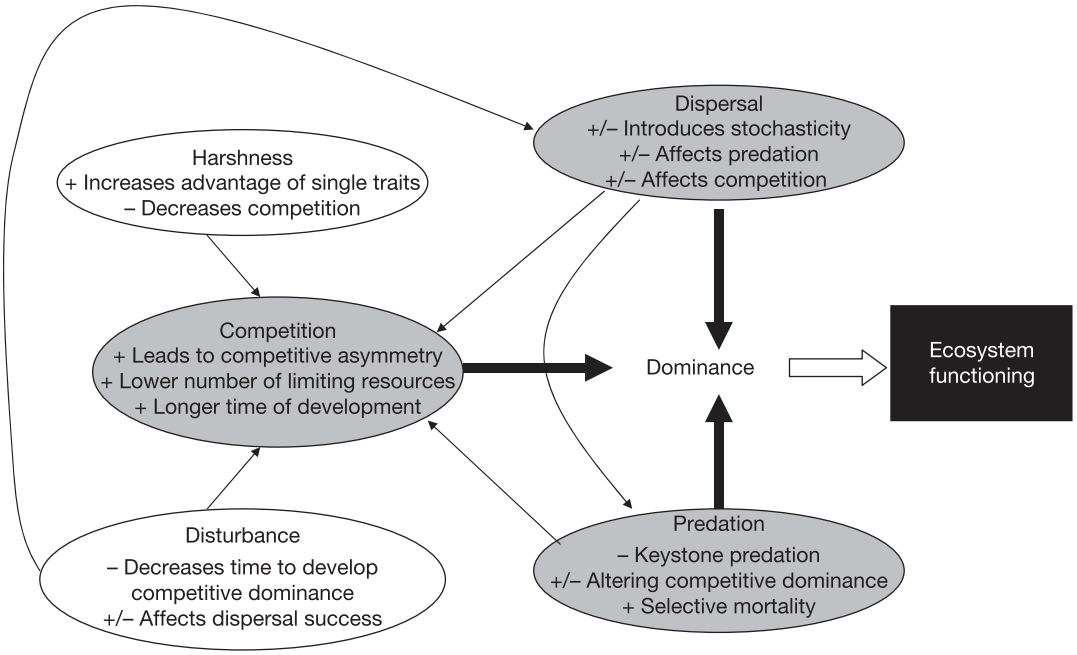


Fig. 3 Conceptual depiction of processes affecting dominance. *Gray ovals* comprise biotic processes, *white ovals* abiotic constraints. Direct effects are represented by *bold arrows*, indirect effects by *thin arrows*. '+' and '−' characterize processes or traits increasing or decreasing dominance, respectively.

competitive success mainly depends on one single trait (plant height), and species growing higher are able to strongly dominate communities. Similarly, filamentous species highly dominate many undisturbed assemblages of benthic algae (periphyton) as they are able to compete successfully for light and water column nutrients.

The skewed distribution of successful traits is also the reason why anthropogenic input of fertilizers often leads to enhance dominance (**Fig. 2**A). Fertilization of terrestrial or aquatic environments (eutrophication) generally enhances the availability of one or few resources, but not of others. Thereby, the array of traits leading to competitive success narrows. A typical example is represented by the addition of phosphorus (P) to lakes, which often leads to the dominance of cyanobacteria species which need a lot of P but are able to fix atmospheric nitrogen (N).

More generally speaking, dominance increases when the number of potentially limiting nutrients decreases (**Fig. 3**). It has been shown that the evenness of lake phytoplankton increases with increasing number of limiting nutrients. Similarly, the evenness of plant assemblages peaks at intermediate ratios of available N:P, but is low at low or high N:P ratios, respectively. Thus, a strong imbalance of N and P, which leads to limitation by either only N or only P, leads to greater dominance of single species. Similarly, gradients in productivity implicitly are also gradients in the number of limiting resources. Very unproductive terrestrial ecosystems often are systems with strong water shortage and dominated by species most successful in accessing water. Very productive terrestrial ecosystems tend to be limited only by light and again single species with certain traits will dominate. Therefore, dominance is least pronounced in mid-productive ecosystems.

Environmental harshness has often been supposed to reduce competition, which however has been falsified in recent years. Instead harsh conditions often require certain adaptations and the array of traits leading to survival and competitive success are also highly narrowed for organisms in extraordinarily harsh conditions. Thus dominance of single species is the general case in these environments.

The second aspect, time of development, describes whether the competitive advantage of a species can fully develop (**Fig. 3**). The gain of competitive dominance takes time, while the extent of time strongly depends on the generation time of the organisms and the already-mentioned asymmetry in competitive advantage. A highly superior fast-growing species will gain dominance more rapidly than a slow-growing species which has only a marginal competitive advantage compared to the cooccurring species. Competitive dominance in competing microalgae can arise within days or weeks, whereas it may last years in long-lived trees or in ecological similar moss species in a bog.

Because the gain of competitive dominance takes time, spatial as well as temporal heterogeneity may stop or even reverse this process. Spatial heterogeneity can be provided by patchiness in resource supply or by biotic or abiotic architecture affecting competitive success. Patchy resource supply creates a mosaic of resource ratios and competitively advantageous strategies. Architecture often opens up refuges or represents the basis for the success of other life-forms (such as epiphytes). Temporal heterogeneity is often investigated in the form of fluctuations in resource availability, which also reduces competitive dominance or reverses competitive ranks of the interacting species.

Disturbances, defined as an irregular mortality-inducing event, has been proposed as a univariate explanation of dominance and diversity patterns, most prominently by the intermediate disturbance hypothesis (IDH), which predicts highest species

richness at disturbance regimes of intermediate frequency and intensity. Similarly, it predicts that low-disturbance ecosystems are dominated by competitively superior species, whereas high disturbance creates dominance by species able to cope with the extreme mortality in these systems. Dominance is thus expected to be low in mid-disturbance regimes, where dominance opens niche opportunities for species by creating patchiness of environmental conditions allowing species to differentially express life-history tradeoffs. Strong evidence has accumulated that how disturbance (and also consumption as a biotic mortality agent) affects community structure is coupled to the productivity of the system, which affects the rate of biomass accrual and dispersal and thus the speed of competitive displacement in a community (see below).

Consumption

Consumers can act as a mortality agent keeping competitively superior species from gaining dominance. Consumption in this meaning comprises all kinds of consumer–prey interactions including herbivory, predation, and pathogens. Dominance can be enhanced or reduced by consumer presence, and the sign of the effect often is mediated by changing competitive interactions between prey species. Consumers can enhance dominance when they favor competitive success of well-defended or less-preferred prey species. More often, however, consumption reduces dominance if the superior competitor is especially prone to consumption.

The generality or selectivity of the consumer is highly important for their effect on dominance in prey communities. Selective consumers increase dominance if they favor species with a certain trait but they can reduce dominance if they are specialized on competitively superior species. The latter interaction is known in the ecological literature as keystone predation, which describes the preferential consumption of a competitively dominant species and thus the fostering of subdominant species. Keystone predation generally decreases dominance ratios and has been observed in many ecosystems strongly structured by competition.

Generalist consumers are also able to reduce dominance, mainly in situations where dominance is fostered by strong competition and fast competitive displacement. In highly productive ecosystems, dominance by single species is generally high, but these dominant species are often highly prone to consumption. Especially plants allocate less resources to defense (structural or chemical) with increasing ecosystem productivity (aka the growth rate hypothesis of defense). Therefore, plants adapted to dominate at high resource supply tend to be highly vulnerable to grazing and grazer presence reduces dominance. Thus, productivity is a key factor to understand the degree of consumer control on diversity and dominance. Several meta-analyses concurred on a dominance-reducing effect of consumers in highly productive ecosystems characterized by strong dominance of fast-growing species, which is strongly reduced by the presence of consumers. At low productivity, consumer effects on diversity are mainly negative, which translates to an increase in dominance.

Many trophic interactions are able to propagate through the food web and may change dominance indirectly on the next trophic level (trophic cascades). Predators can alter the dominance ratio and the identity of dominance species in plant assemblages via the reduction in herbivore density or activity. Other indirect effects in food webs have also been identified as altering and regulating the dominance in community structure. Apparent competition and associational resistance are two pathways mediating the contribution of different prey species to the entire prey assemblage. Also nutrient regeneration may change competitive advantages in communities where recycled nutrients play a major role in primary production. The presence of a P-rich consumer may enhance N-regeneration compared to P-regeneration and thus shift the available N:P elemental ratio and therewith competitive success of different prey taxa.

Facilitation

The role of positive interactions for community structure and dominance has only recently become a focus of ecological studies. The importance of mutualistic interactions is especially seen in the dominance of life-forms with long-evolved symbioses in certain ecosystems. The dependence of trees dominating boreal forests on mycorrhiza and the dependence of corals on symbiotic zooxanthellae are vivid examples highlighting the effects of positive interactions in community structure. Also nitrogen fixation via rhizobia can foster the dominance of certain plant groups such as the Fabaceae in terrestrial ecosystems with high light but poor N-availability.

Additionally, facilitation clearly plays a role in temporal gradients of dominance, as the presence of a certain pioneer species often is necessary in order to allow the dominance of late successional species. Facilitation can thus create dominance by the favored species, but it also can reduce dominance by allowing coexistence of species which would not be able to thrive without the presence of a facilitative species.

Dominance in Metacommunities

Ecological communities are not isolated, but connected via dispersal, which links local communities into metacommunities. Metacommunities provide a framework to understand the scale dependence of dominance and to highlight important controlling factors of dominance such as ecological interactions, dispersal, and species sorting. Regional and local dominance can strongly differ as do the processes leading to regional versus local dominance. Most locally dominant species are competitively superior species. However, depending on the frequency of disturbance, the overall probability of dispersal and the presence of source sink

dynamics, regionally dominant species may be either superior competitors or superior dispersers. The degree of regional (and local) dominance further depends on the spatial heterogeneity in environmental conditions across the local patches.

The effect of dispersal on dominance is clearly rate dependent (Fig. 3). At very low dispersal rates, neither competitive nor consumer–prey interactions are affected and the impact of dispersal is low. The identity of the competitive dominant species may strongly differ between the local communities, as each community will have a different species composition. At highest dispersal, all species are able to reach each of the local communities, and competitive dominance still plays the most important role, but now one species, the superior competitor in the regional species pool, will dominate all local communities. At intermediate dispersal frequencies, the dominance of the best competitor will be disturbed by frequent invasion of good colonizers as long as colonization–competition tradeoffs persist. The dominance can also be reduced in source–sink metacommunities, where local populations in harsh local (sink) patches are maintained by substantial dispersal from benign (source) patches.

The frequency of disturbances (or other sources of mortality) in the local patches determines how important different dispersal regimes can be. With no opening of space, dispersal cannot rescue colonizers from being excluded by competitors. However, with disturbance space is opened for colonizers to persist and the importance of colonization traits compared to competitive traits increases with disturbance frequency and intensity.

The identity of the dominant species can be highly dissimilar between local patches in a metacommunity, depending on dispersal rate (see above) and the environmental and geographical distance between the local patches. Higher geographical distance enhances the probability of dispersal limitation and higher environmental distance allows for species sorting. Beta-diversity, which measures the turnover in species composition between local patches, increases with increasing spatial distance and with decreasing environmental similarity.

Including metacommunity dynamics and community assembly into the analysis of dominance also allows addressing the effect of stochasticity and assembly sequence on dominance ratio and the identity of dominance species. As exemplified in Fig. 1, nearby habitats of similar abiotic conditions can have different species setup and can be highly dominated by one species, which is almost missing in the other system. These differences can occur due to founding effects, which stress the importance of colonization sequence. Overall dominance can be enhanced by a species colonizing an enemy-free space and growing into a size refuge (which is the mechanism probably leading to the dominance of green algae in the right rock pool Fig. 1) or by a species arriving able to monopolize and defend space (which is congruent with inhibition mechanisms in succession, see temporal patterns above).

Dominance in Regional Species Pools

From a global perspective, dominance became evident to the first exploring naturalists, who already described the changing vegetation on Earth by large units such as biomes. These biomes are not necessarily defined by the dominance of single species, but often by the dominance of certain phylogenetic lineages and life-forms. Thus, regional species pools tend to be dominated by higher taxonomic or evolutionary units. As the regional species pool is highly important for the species being able to colonize local communities (and metacommunities), the historical processes leading to regional dominance will also affect dominance patterns in local communities. Especially models focusing on neutrality of species traits or assuming strong stochastic dispersal effects but weak local interactions propose a strong regional imprint on local community structure. The importance of these processes including speciation–extinction dynamics or range expansions for local community patterns is still strongly debated and will probably be different in different biomes.

Consequences of Dominance

The degree and identity of dominance strongly differ between ecosystems, but does this variation matter for the functioning of ecosystems? Obviously the processes in ecosystems (such as productivity, respiration, energy transfer, resource retention) are mainly driven by the dominant species. The traits of dominant species thus have a higher importance for ecosystem functioning than the traits of rare species. The consequences of dominance are separate from the consequences of keystone species (sensu Payne) and foundation species (sensu Dayton). Whereas keystone species are species affecting ecosystem processes by having very high per capita impacts (e.g., top predators) and foundation species comprise species which have strong impact on ecosystem structure by enabling community structure (e.g., trees), the importance of dominance is based on the relative fraction of matter and energy flow canalized through one species. However, most foundation species are locally dominant, whereas most keystone species are not.

The question becomes highly important in the face of strong alterations of global diversity and of dominance patterns by human actions such as habitat destruction, species homogenization, and fertilization. Recent theory and experiments suggest that local species richness can affect local ecosystem processes. In many of these experiments, the identity of the species had a major effect on the outcome. The identity of the dominant species consistently affects primary production, decomposition rates, consumption rates, and overall community respiration. Species identity effects have in many studies been more important than the effects of species richness per se.

Few studies have manipulated not only the number of traits present but also their distribution. Studies looking for effects of evenness on ecosystem processes such as productivity are few to date, but one emerging pattern is that higher evenness in plants tends to reduce primary productivity. This can easily be explained by the fact that high evenness prevents the community from being dominated by the most productive species. Thus, important ecosystem processes tend to be best performed in communities characterized by strong dominance.

This primacy of dominant species in ecosystem functioning does not necessarily mean that rare species are unimportant. Many ecologists have proposed that the importance of subdominant species is an insurance against spatial or temporal fluctuations in abiotic conditions. Thus, rare species represent an insurance as they may be able to perform major ecosystems processes under different environmental constraints. More evenly structured communities may perform such shifts more easily than communities characterized by a strongly dominating species. However, this evenness insurance effect has not been analyzed theoretically or experimentally yet.

See also: Behavioral Ecology: Social Behavior and Interactions; Dominance Hierarchy. General Ecology: Communication

Further Reading

Cardinale, B.J., Hillebrand, H., Charles, D.F., 2006. Geographic patterns of diversity in streams are predicted by a multivariate model of disturbance and productivity. Journal of Ecology 94, 609–618.
Clark, G.F., Stark, J.S., Johnston, E.L., 2017. Tolerance rather than competition leads to spatial dominance of an Antarctic bryozoan. Journal of Experimental Marine Biology and Ecology 486, 222–229.
Gaston, K.J., Blackburn, T.M., 2000. Pattern and process in macroecology. Blackwell Scientific Publications: Oxford.
Hillebrand, H., Sommer, U., 2000. Diversity of benthic microalgae in response to colonization time and eutrophication. Aquatic Botany 67, 221–236.
Hohenbrink, S., Schaarschmidt, F., Bünemann, K., Gerberding, S., Zimmermann, E., Radespiel, U., 2016. Female dominance in two basal primates, Microcebus murinus and Microcebus lehilahytsara: Variation and determinants. Animal Behaviour 122, 145–156.
Holyoak, M., Leibold, M.A., Holt, R.D., 2005. Metacommunities—Spatial Dynamics and Ecological Communities. Chicago: University of Chicago Press, 513.
Huston, M.A., 1994. Biological diversity: The coexistence of species in changing landscapes. Cambridge University Press: Cambridge.
Magurran, A.E., 2004. Measuring Biological Diversity. Oxford: Blackwell Publishing.
Steiner, K.C., Stein, B.S., Finley, J.C., 2018. A test of the delayed oak dominance hypothesis at mid-rotation in developing upland stands. Forest Ecology and Management 408, 1–8.
Rosenzweig, M.L., 1995. Species diversity in space and time. Cambridge University Press: Cambridge.
Worm, B., Lotze, H.K., Hillebrand, H., Sommer, U., 2002. Consumer versus resource control of species diversity and ecosystem functioning. Nature 417, 848–851.
Yang, Q., Rogers, T., Dawes, J.H.P., 2017. Demographic noise slows down cycles of dominance. Journal of Theoretical Biology 432, 157–168.

Dormancy☆

Philip Withers, University of Western Australia, Crawley, WA, Australia
Christine E Cooper, Curtin University of Technology, Perth, WA, Australia

Introduction

Dormancy is a widely recognized behavioral and physiological state of both animals and plants that generally involves inactivity and reduced metabolic rate (Fig. 1). Torpor is a similar term to dormancy, meaning inactivity or lethargy. Dormancy or torpor can involve very different physiological states, in response to a variety of different stimuli, including low temperature, high temperature, lack of water, or lack of food. It can be a short-term event (<24 h), can occur for a few consecutive days, or may last an entire season or even many years. Dormancy can also involve a developmental arrest (diapause). Cryptobiosis, which literally means "hidden life," is a more extreme state than dormancy, with almost no detectable activity or metabolism. It is most prevalent in lower vertebrates, and is often a seasonal survival strategy to cold or desiccation.

Cryptobiosis

This state of "suspended animation" has been observed for a variety of invertebrate animals and plants during extreme environmental conditions. It was first described for invertebrate animals that survived an absence of water by becoming inactive and allowing their tissues to become desiccated (anhydrobiosis, e.g., rotifers). Two other forms of cryptobiosis also involve an altered state of cellular water, freezing temperatures (cryobiosis, e.g., a frozen insect), and high osmotic concentration (osmobiosis, e.g., brine shrimp eggs in a salt lake). Another form of cryptobiosis is survival of a lack of oxygen (anoxybiosis, e.g., killifish eggs sealed inside their egg capsule). The best-known example of cryptobiotic animals is probably the eggs of brine shrimp (*Artemia*), which can survive extended periods of complete desiccation, high salt concentration, or anoxia; their desiccated eggs are also remarkably resistant to extremes of temperature. Various "resurrection" plants are well-known examples of cryptobiotic plants, being able to recover from desiccation for extended periods. Seeds of some plants are also spectacularly resistant to desiccation, sometimes for very long periods of time (e.g., seeds more than 1000 years old of the Indian lotus from an ancient lake bed in China).

All of these forms of cryptobiosis involve complete inactivity. Ecological advantages of cryobiosis include survival of harsh environmental conditions, and dispersal of highly resistant life stages. However, the physiological adaptations required by these animals and plants to survive extreme conditions at no detectable metabolic rate are generally complex and specialized.

Diapause and Quiescence

Diapause is an ecological strategy for the avoidance of harsh conditions that involves the cessation of development of a subadult life stage. It is essentially a time-delaying tactic to synchronize further stages of the life cycle with appropriate environmental conditions. Diapause is especially common in insects but is also observed in a wide variety of other invertebrate animals (e.g., brine shrimp embryos) and vertebrate animals (e.g., annual killifish embryos), as well as many plants (e.g., buds, bulbs, rhizomes, and seeds). Some plant seeds require drying out before they can develop, ensuring that adverse dry seasons pass before the embryo starts to develop. Diapause is also a reproductive strategy in a variety of mammals for the delayed implantation and development of embryos (e.g., macropod marsupials, mustelids, and deer). Quiescence is a period of inactivity, similar to diapause, but is a facultative response to an immediate change in environmental conditions that is terminated simply by the resumption of more favorable environmental conditions, rather than a programmed and obligate response. It may be a response to harsh environmental conditions such as low or high temperature, or drought. Many invertebrates and plants (particularly their seeds) become quiescent.

Hibernation (Winter Dormancy)

Hibernation is when an organism spends the winter in a state of dormancy; it is long-term multiday torpor. Many plants survive extended periods of cold and desiccation, either as aboveground trees or shrubs, or as underground structures. Protective scales around stem tips allow buds of aboveground plants to endure winter conditions without damage. The aboveground structures of other plants die back in unfavorable conditions, leaving dormant underground bulbs, rhizomes, tubers or corms, for which the

☆*Change History:* February 2018. I. Martins made changes to the references.

This is an update of P.C. Withers and C.E. Cooper, Dormancy, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 952–957.

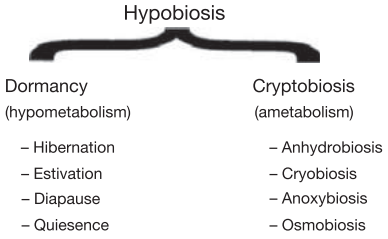


Fig. 1 Schematic summary of different hypobiotic (metabolism less than normal) states, including hypometabolism and ametabolism. Adapted from Keilin (1959) The problem of anabiosis or latent life: History and current concept. *Proceedings of the Royal Society of London.* **150**, 149–191.

Table 1 Summary of torpor patterns in monotreme, marsupial and placental mammals, and birds, for single-day torpor (T), hibernation (H), or estivation (E)

Taxon	*Torpor pattern*
Monotremata	
Tachyglossidae	H
Ornithorhynchidae	
Metatheria	
Didelphidae	T
Microbiotheriidae	H
Dasyuridae	T
Myrmecobiidae	T
Petauridae	T
Burramyidae	H
Acrobatidae	H
Tarsepididae	T
Eutheria	
Rodentia	T, H, E
Insectivora	T
Chiroptera	T, H
Carnivora	T, H?
Primates	T?, H/E
Macroscelidae	T
Aves	
Coliiformes	T
Trochiliformes	T
Strigiformes	T
Caprimulgiformes	T, H
Columbiformes	T
Coraciformes	T
Passeriformes	T

soil buffers environmental extremes. Many plants accumulate solutes in their fluids to prevent freezing during winter, while others can tolerate freezing of water in their xylem and other extracellular water pools. For ectothermic animals, hibernation is primarily a behavioral state with reduced body temperature, hence activity and metabolic rate. Some use supercooling or antifreeze solutes to avoid freezing, or tolerate freezing of their extracellular fluids (e.g., weta crickets and wood frogs). Many endothermic mammals also hibernate (Table 1). Mammalian hibernators typically use multiday torpor for weeks or even months (e.g., Fig. 2), and attain very low body temperatures (T_b's) (e.g., 0 to -5°C). Only one bird, the poorwill, is known to hibernate, although many other birds (and mammals) readily use single-day torpor during winter.

Estivation (Summer Dormancy)

Estivation is summer dormancy, that is, long-term torpor during summer for survival of hot and dry periods. Many desert plants survive extended periods of high temperature and low rainfall. Some survive as desiccated seeds (5%–10% water content), particularly annual species, but some survive desiccation as adults. These "resurrection" plants, such as the Rose of Jericho (*Selaginella*), can desiccate to about 5% water content during dry periods, but survive and "come back to life" after rain. Pincushion lilies similarly re-activate by regenerating from buds after rain.

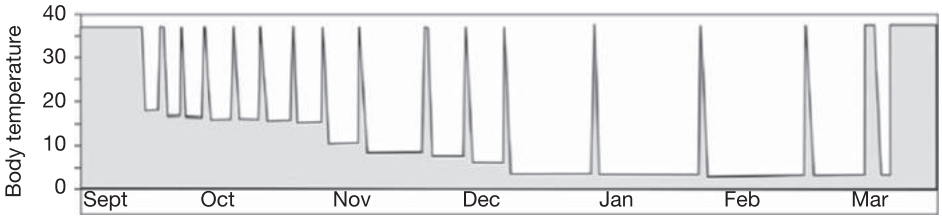


Fig. 2 Pattern of body temperature during a seasonal hibernation cycle for a ground squirrel. Modified from Wang, L.C.H. (1978) Energetic and field aspects of mammalian torpor: The Richardson's ground squirrel. In: Wang, L.C.H. and Hudson, J.W. (eds.) Strategies in Cold. Natural Torpidity and Thermogenesis, pp. 109–145. New York: Academic Press.

Fig. 3 Estivating frog (*Cyclorana cultripes*) in a cocoon of shed skin. Photograph by G. Thompson and P. Withers.

Amongst invertebrates (e.g., earthworms and insects) estivation usually involves an inactive stage with a water-resistant covering. For example, estivating earthworms form a mucus cocoon to resist desiccation, and many insect pupae are remarkably resistant to water loss. Amongst vertebrates, fishes, amphibians, and reptiles enter a similar estivation state. Fishes and amphibians often form a cocoon of dried mucus (e.g., African lungfishes) or shed epidermal layers (e.g., some desert frogs; **Fig. 3**) to resist epidermal water loss; the cocoon covers the entire body surface except for the nostrils. Reptiles have a relatively water-impermeable epidermis and do not need to form a cocoon to reduce evaporative water loss. Estivating ectotherms typically have an intrinsic metabolic depression for energy conservation.

Some mammals also estivate (**Table 1**). For example, desert ground squirrels enter a long-term estivation state that is physiologically similar to hibernation except for the higher ambient temperature (T_a) and T_b. Other mammals such as cactus mice and kangaroo mice use single-day torpor cycles during summer.

Thermal and Energetic Physiology of Torpor

Torpor involves a number of physiological changes, especially related to body temperature, metabolism, and water balance. These physiological changes are interrelated insofar as body temperature influences energetics, and water balance is related to both body temperature and metabolism. However, the detail of the physiological consequences of torpor differs between organisms.

Ectothermic Animals and Plants

For ectotherms, T_b is essentially equal to T_a during hibernation/estivation. This means that any decrease in T_a during hibernation or estivation is accompanied by a decrease in T_b, which in turn is accompanied by an exponential decline in metabolic rate (MR) as described by the Q_{10} relationship, that is,

$$Q_{10} = (\mathrm{MR}_{T_{b2}}/\mathrm{MR}_{T_{b1}})^{10/(T_{b2}-T_{b1})} \text{ or}$$
$$\mathrm{MR}_{T_{b2}} = \mathrm{MR}_{T_{b1}} Q_{10}^{(T_{b2}-T_{b1})/10} \quad (1)$$

where $\mathrm{MR}_{T_{b2}}$ is the metabolic rate at T_{b2} and $\mathrm{MR}_{T_{b1}}$ is MR at T_{b1}. For most physiological variables, Q_{10} is generally about 2.5. This decrease in MR results in substantial energy savings and thus a prolonged survival period in the cold.

For some ectotherms there is an unequivocal intrinsic metabolic depression during estivation that occurs without any change in T_b (e.g., snails, fishes, and amphibians). Some plant seeds during dormancy are also hypometabolic or even ametabolic. This

intrinsic metabolic depression, which is often a decrease in MR to about 20% of normal, occurs in the absence of any T_b, ionic, osmotic, or any other discernable physiological perturbation. The cue for intrinsic metabolic depression would appear to be a change in environmental conditions that indicates impending potential for desiccation. Intrinsic metabolic depression is not a short-term (e.g., daily) event; it often takes about 2–4 weeks for metabolic depression to become fully developed. It is probably more important for estivation, which has a lesser hypometabolism by lowered T_b than hibernation. The molecular or biochemical mechanisms for this intrinsic metabolic depression are not well understood; however its physiological significance is clearly extension of the hibernation/estivation period that can be survived by conserving energy.

Endothermic Animals

Endothermic vertebrate animals have a fundamentally different relationship between T_a and T_b than ectothermic animals and plants as a consequence of thermoregulatory thermogenesis. Thermal and energetic consequences of torpor are therefore more complex for endotherms because at low T_a their MR is normally increased above basal (BMR) by metabolic thermogenesis that maintains T_b constant (normothermia). During torpor, there is a profound decrease in MR, typically to 1% or even less of normothermic MR, and a concomitant decrease in T_b often close to T_a (**Fig. 4**). Entry into torpor appears to be a controlled physiological process, not simply an inability to thermoregulate. During torpor at moderate to high T_a's, T_b declines to nearly T_a and MR declines exponentially with T_b. This is the same pattern as for ectotherms, and indicates a state of nonthermoregulation. However, if T_b decreases below a species-specific set point at lowered T_a, then T_b is regulated at that set point by the onset of thermogenesis; this is the same as the normothermic thermoregulatory response except that the T_b set point is lower than for normothermia. For many single-day torpidators the torpor set point is about 20°C, but it is generally much lower for hibernators (about 0–5°C, but as low as −5°C for arctic ground squirrels).

Two mechanisms contribute to the marked decline in MR of endotherms during torpor. Defense of normal body temperature is relaxed and so the thermogenic increment in MR above BMR is eliminated. As a consequence, heat production is less than heat loss and T_b declines to close to T_a and so there is a further decline in MR due to the Q_{10} effect. However, if T_a decreases below the torpor set point where T_b is again defended, then MR increases for thermogenesis. For most endotherms, the decline in MR during torpor is accounted for by the elimination of the thermogenic MR increment and the decline in T_b and MR with a typical Q_{10} ($\approx$2.5).

Intrinsic metabolic depression is a third possible mechanism contributing to MR reduction during torpor. However, the contributions of the elimination of thermogenic MR increment and the decline in T_b and Q_{10} effect are so great that the contribution of intrinsic metabolic depression, should it occur in an endotherm, would be a relatively minor absolute energy saving.

Arousal from torpor is typically a physiologically driven event requiring considerable thermogenesis by shivering (skeletal muscle thermogenesis) or metabolism of specialized brown adipose tissue (in placental mammals but not marsupials or birds). There is also increasing evidence that many species use passive rewarming (e.g., basking) to arouse, since it greatly reduces the metabolic cost of arousal. Long-term hibernation by mammals is not necessarily a continuous period of prolonged inactivity. It is

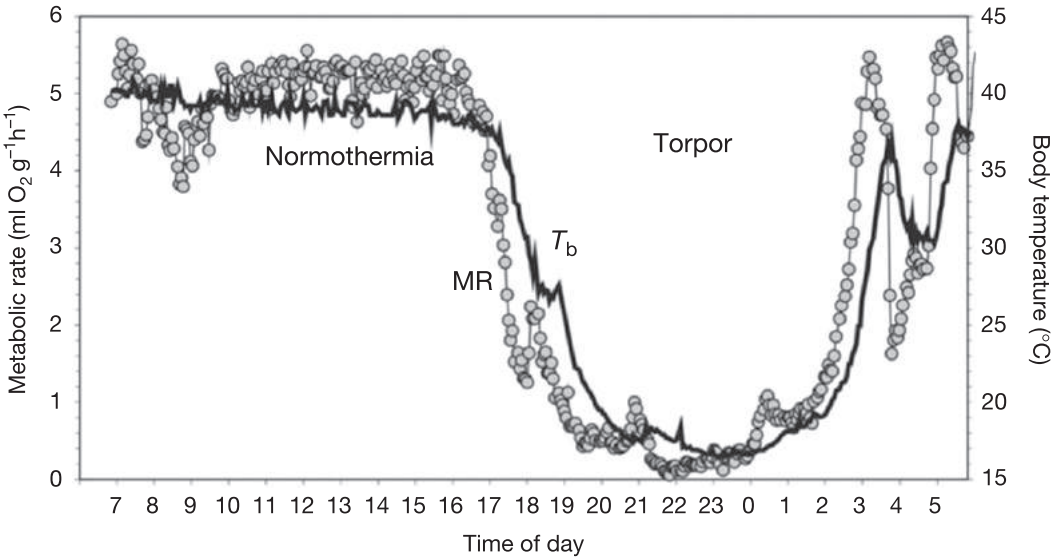


Fig. 4 A daily torpor cycle in a typical single-day torpidator, the dunnart *Sminthopsis macroura*, showing the decline in T_b and MR during entry into torpor, a short period of sustained torpor, and then the increase in T_b and MR during arousal from torpor. Data from F. Geiser, unpublished data.

periodically broken by a short period of arousal, then re-entry into hibernation (e.g., Fig. 2). The reason for these periods of arousal and re-entry is not clear. There appears to be some physiological "need" to periodically arouse. It has been suggested that perhaps some accumulated metabolite needs to be eliminated by urination, which only occurs if the animal is normothermic.

The beneficial energy savings of torpor are clearly evident from the difference between the high normothermic MR and the greatly reduced torpid MR, even after accounting for the metabolic cost of arousal. For daily torpor, the energy saving depends on the length of the torpor bout and the depth of torpor; for a dunnart, the daily energy saving is about 36% for 13 h of torpor (e.g., Fig. 4). For hibernation, the daily energy saving is greater because metabolic rate is low for typically 24 h per day; for a hibernating ground squirrel, the energy saving is about 85% over 6 months.

There is a complex pattern of single-day torpor, multiday hibernation, and multiday estivation amongst mammals and birds (Table 1) that partly reflects phylogeny but also body mass. Torpor is more advantageous for small than large species. Small species have a higher mass-specific metabolic rate and therefore benefit more from the energetic saving associated with torpor. The rate of entry into and arousal from torpor is strongly dependent on body mass. Small species enter torpor quicker because of their higher thermal conductance (higher surface-to-volume ratio) and they also arouse quicker because of their higher mass-specific MR and lower thermal inertia. In contrast, larger species cool and rewarm slower, so the energy savings are less, especially for daily torpor.

Cues for Dormancy

Many plants respond to the climatic cycle of their habitat. In particular, photoperiod, temperature, and rainfall are important cues for the commencement and also cessation of dormancy. Some species respond to long-term climatic cycles, while others undergo more immediate facultative responses to ambient temperature or water availability. For animals, single-day torpor can occur rapidly in response to short-term environmental changes, such as inclement weather. It generally occurs on a circadian cycle, corresponding to the normal period of activity/inactivity. Onset of hibernation or estivation, being seasonal long-term periods of dormancy, is a more prolonged and sometimes programmed response to an impending change in environmental conditions. For example, desert frogs initiate estivation if conditions become dry, by burrowing, forming a cocoon, and initiating intrinsic metabolic depression; this can take 3–4 weeks, but occurs independent of time of year. In contrast, hibernation by some mammals such as ground squirrels is obligate and only occurs at a specific time of the year after a period of preparation (e.g., seeking out or constructing suitable hibernation sites, increased activity and feeding, deposition of energy stores, and changes in body fluid solutes). This pattern of obligate hibernation is controlled on a circannual cycle by cues that include shortening photoperiod and decreasing air temperature. Reduced water availability and high T_a are primary cues for estivation.

Ecological Consequences of Dormancy

The general roles of torpor, hibernation, and estivation are avoidance of unfavorable short-term or long-term (seasonal) climatic conditions and conservation of energy during this period of inactivity. Seasonal dormancy also has obvious ecological benefits. It allows species to exploit ephemeral environments. Hibernation and estivation enable species to colonize habitats that would otherwise be unsuitable for growth or survival at certain times of the year due to harsh environmental conditions. Timing of active life stages or generations can be optimized. Seasonal dormancy therefore contributes to the fitness of individuals and species.

There would also appear to be costs associated with torpor. Many species do not use or survive torpor, and species capable of torpor do not necessarily use it on a routine basis. There is a fundamental physiological advantage (at least for endotherms and even for many ectotherms) of maintaining a high and stable body temperature, for example, growth, digestion, muscle contractility and immunological defense. There is also a physiological danger of thermal death or being unable to arouse if the T_a becomes too low (e.g., freezing), or death if energy reserves become insufficient for arousal. Ecological costs of torpor could include vulnerability to predation, competition from conspecifics that do manage to successfully forage, reduced reproductive success, and lower rates of essential activities such as cell division and digestion. There are also similar and additional costs of seasonal dormancy. It can delay reproduction and development, diminish posthibernal reproduction, require that short-lived species survive for longer, and result in sex-biased populations if there is differential survival based on gender. For multiday torpor by endotherms, there appears to be a necessity for periodic arousal, suggesting some physiological requirement for an occasional return to a high T_b (see above).

Further Reading

Dausmann, K.H., Glos, J., Ganzhorn, J.U., Heldmaier, G., 2005. Hibernation in the tropics: Lessons from a primate. Journal of Comparative Physiology B 175 (2005), 147–155.
Fadón, E., Rodrigo, J., 2017. Unveiling winter dormancy through empirical experiments. Environmental and Experimental Botany. doi:10.1016/j.envexpbot.2017.11.006.
Geiser, F., 1994. Hibernation and daily torpor in marsupials. A review Australian Journal of Zoology 42, 1–16.
Geiser, F., 2004. Metabolic rate and body temperature reduction during hibernation and daily torpor. Annual Review of Physiology 66, 239–274.
Guppy, M.G., Withers, P.C., 1999. Metabolic depression in animals: Physiological perspectives and biochemical generalizations. Biological Reviews 7 (1999), 1–40.
Gwinner, E., 1986. Circannual rhythms. Berlin: Springer.
Hochachka, P.W., Guppy, M., 1987. Metabolic arrest and the control of biological time. Cambridge: Harvard University Press.

Joergensen, R.G., Wichern, F., 2018. Alive and kicking: Why dormant soil microorganisms matter. Soil Biology and Biochemistry 116, 419–430.
Keilin, D., 1959. The problem of anabiosis or latent life: History and current concept. Proceedings of the Royal Society of London 150, 149–191.
Penfield, S., 2017. Seed dormancy and germination. Current Biology 27 (17), 874–878.
Storey, K.B., 2001. Molecular mechanisms of metabolic arrest. Life in Limbo. Oxford: Bios Scientific Publishers.
Tauber, M.J., Tauber, C.A., Masaki, S., 1986. Seasonal adaptations of insects. New York: Oxford University Press.
Wang, L.C.H., 1978. Energetic and field aspects of mammalian torpor: The Richardson's ground squirrel. In: Wang, L.C.H., Hudson, J.W. (Eds.), Strategies in cold. Natural torpidity and thermogenesis. New York: Academic Press, pp. 109–145.
Watanabe, Y., How, K.H., Zenke, K., Itoh, N., Yoshinaga, T., 2018. Dormancy induced by a hypoxic environment in tomonts of Cryptocaryon irritans, a parasitic ciliate of marine teleosts. Aquaculture 485, 131–139.
Withers, P.C., 1992. Comparative animal physiology. Philadelphia: Saunders College Publishing.

Ecological Effects of Acidic Deposition☆

Charles T Driscoll, Syracuse University, Syracuse, NY, United States
Irene Martins, University of Coimbra, Coimbra, Portugal

Introduction

Detailed studies by a large community of scientists for more than four decades have provided considerable insight into the ways in which atmospheric deposition alters ecosystems. When it was first identified, acidic deposition, or acid rain as it is commonly called, was viewed as a simple problem that was limited in scope. Scientists know now that acids and acidifying compounds enter ecosystems largely from atmospheric deposition and are transported through soil, vegetation, and surface waters, resulting in a range of adverse ecosystem effects. Controls on emissions of air pollutants, which were initiated in North America and Europe in the 1970s, have resulted in some recovery of ecosystems from the adverse effects of historical acidic deposition. In this article, information on patterns of acidic deposition, the effects of atmospheric deposition of sulfur and nitrogen on sensitive forest and freshwater resources, and the recovery that has resulted from controls on emissions of air pollutants is synthesized.

Acidic Deposition

Acidic deposition largely comprises sulfuric and nitric acid derived from sulfur dioxide and nitrogen oxides, respectively, and ammonium resulting from emissions of ammonia. Sulfur dioxide and nitrogen oxides, originating from human activities, are emitted into the atmosphere largely by the burning of fossil fuels, while ammonia is largely the result of agricultural activities. Once such compounds enter an ecosystem, they can acidify soil and surface waters, bringing about a series of ecological changes. The term " acidic deposition" encompasses all the forms of these compounds that are transported from the atmosphere to Earth, including gases, particles, rain, snow, clouds, and fog. Acidic deposition can occur as wet deposition, in the form of rain, snow, sleet, or hail; as dry deposition, in the form of particles or vapor; or as cloud or fog deposition, which is more common at high elevations and in coastal areas. Wet deposition is fairly well characterized by monitoring at approximately 200 National Atmospheric Deposition Programs (http://nadp.sws.uiuc.edu/) in the United States. In contrast, dry deposition is highly dependent on metrological conditions and vegetation characteristics, which can vary markedly over short distances in complex terrains. As a result, dry deposition is poorly characterized and highly uncertain. Dry deposition is characterized through the Clean Air Status and Trends Network (http://www.epa.gov/castnet/), which includes about 80 sites in the United States.

Sulfuric and nitric acids lower the pH of rain, snow, soil, lakes, and streams. Areas experiencing elevated acidic deposition include eastern North America (**Figs. 1** and **2**), the western United States, Europe, and Asia. In 2010–12, wet deposition (i.e., deposition from forms of precipitation such as rain, snow, sleet, and hail) in acid-sensitive regions of the eastern United States had average pH values of 4.1–5.0, which is about 2–10 times more acidic than background conditions.

Acidic deposition trends in the eastern United States and Europe mirror emission trends in the atmospheric source area or airshed for these regions. Long-term data from across the eastern United States and Europe show declining concentrations of sulfate in wet deposition since the mid-1970s (**Fig. 1**), coincident with decreases in sulfur dioxide emissions. In the United States, decreases in emissions of sulfur dioxide have occurred as a result of emission control programs largely for electric utilities associated with the 1970 and 1990 Amendments of the Clean Air Act. Based on these long-term data, a strong positive correlation exists between sulfur dioxide emissions in the source area and sulfate concentrations in wet deposition. It is expected that the sulfate concentration of wet deposition will decrease (or increase) in a direct linear response to the decrease (or increase) of sulfur dioxide emissions in the atmospheric source area. These observations strongly suggest a cause and effect relationship between emissions of sulfur dioxide and deposition of sulfate in sensitive regions. A similar relationship has started to become evident between emissions of nitrogen oxides and wet deposition of nitrate. In the United States, controls on nitrogen oxide emissions from electric utilities have occurred largely through the US Environmental Protection Agency Nitrogen Budget Program. This relationship for nitrate is not as strong as the relationship for sulfur because emissions of nitrogen oxides in the United States have substantially decreased only since the early 2000s. However, it appears that recent decreases in emissions of nitrogen oxides from electric utilities are starting to result in decreases in atmospheric deposition of nitrate (**Fig. 2**). Atmospheric ammonium deposition has not changed in recent decades, but is becoming an increasingly important consideration in air quality management and acidic and nitrogen deposition, with the decreases in nitrogen oxide emissions and nitrate deposition.

Effects of Acidic Deposition on Forest Ecosystems

In acid-sensitive regions, acidic deposition alters soils, stresses forest vegetation, acidifies lakes and streams, and harms fish and other aquatic life. These effects can interfere with important ecosystem functions and services such as forest diversity and

☆*Change History*: March 2018. CT Driscoll and I Martins updated all sections and added Figs. 1 and 2.

This is an update of C.T. Driscoll, Ecological Effects of Acidic Deposition, In Reference Module in Earth Systems and Environmental Sciences, Elsevier, 2013.

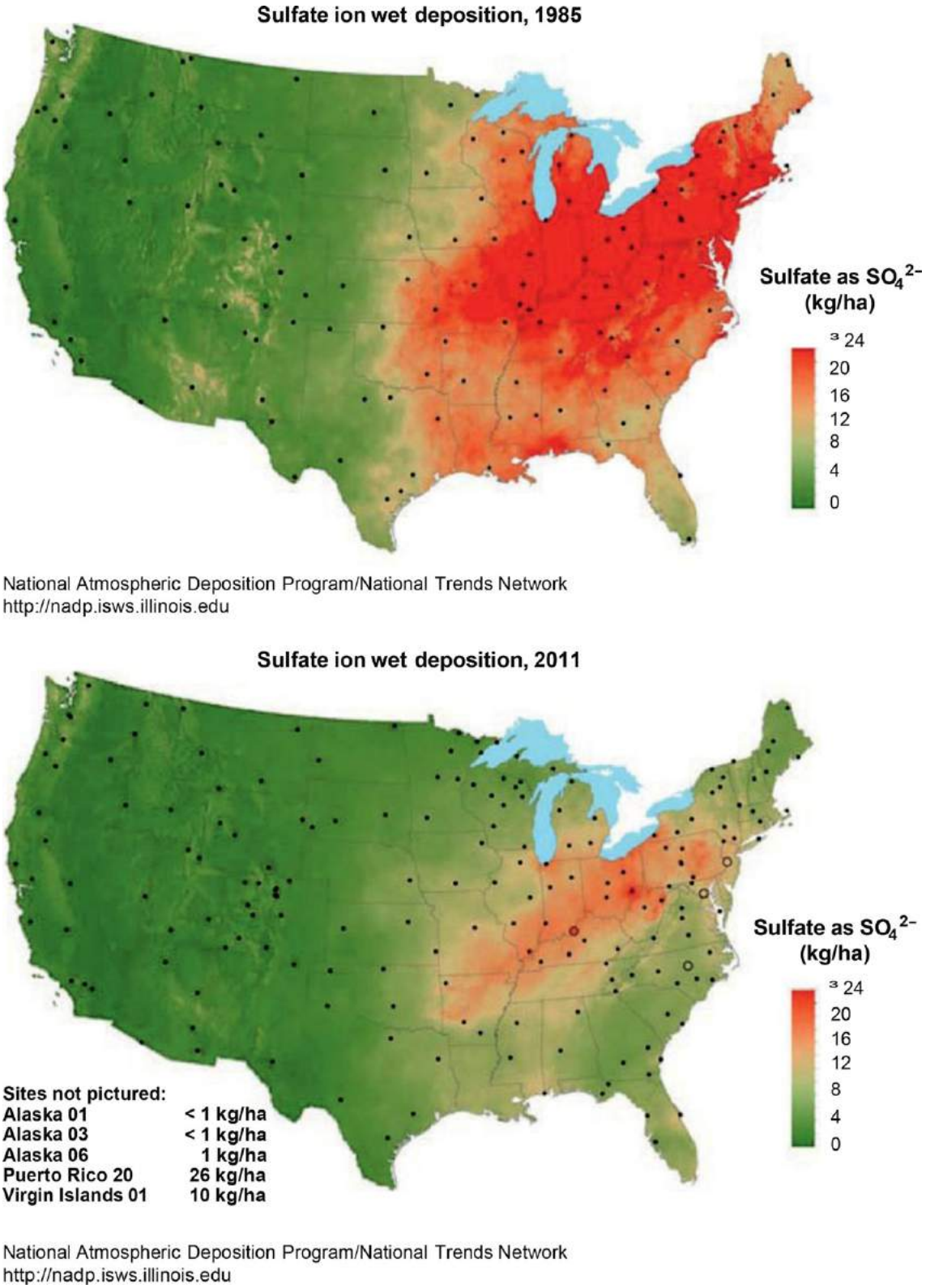


Fig. 1 Map of wet sulfate deposition for the United States for 1985 and 2011. Data are from the National Atmospheric Deposition Program (NADP) network.

productivity, and water quality. Years of acidic deposition have also made many ecosystems more sensitive to continuing air pollution. Moreover, the same pollutants that cause acidic deposition contribute to a wide array of other important environmental issues at local, regional, and global scales (see **Table 1**).

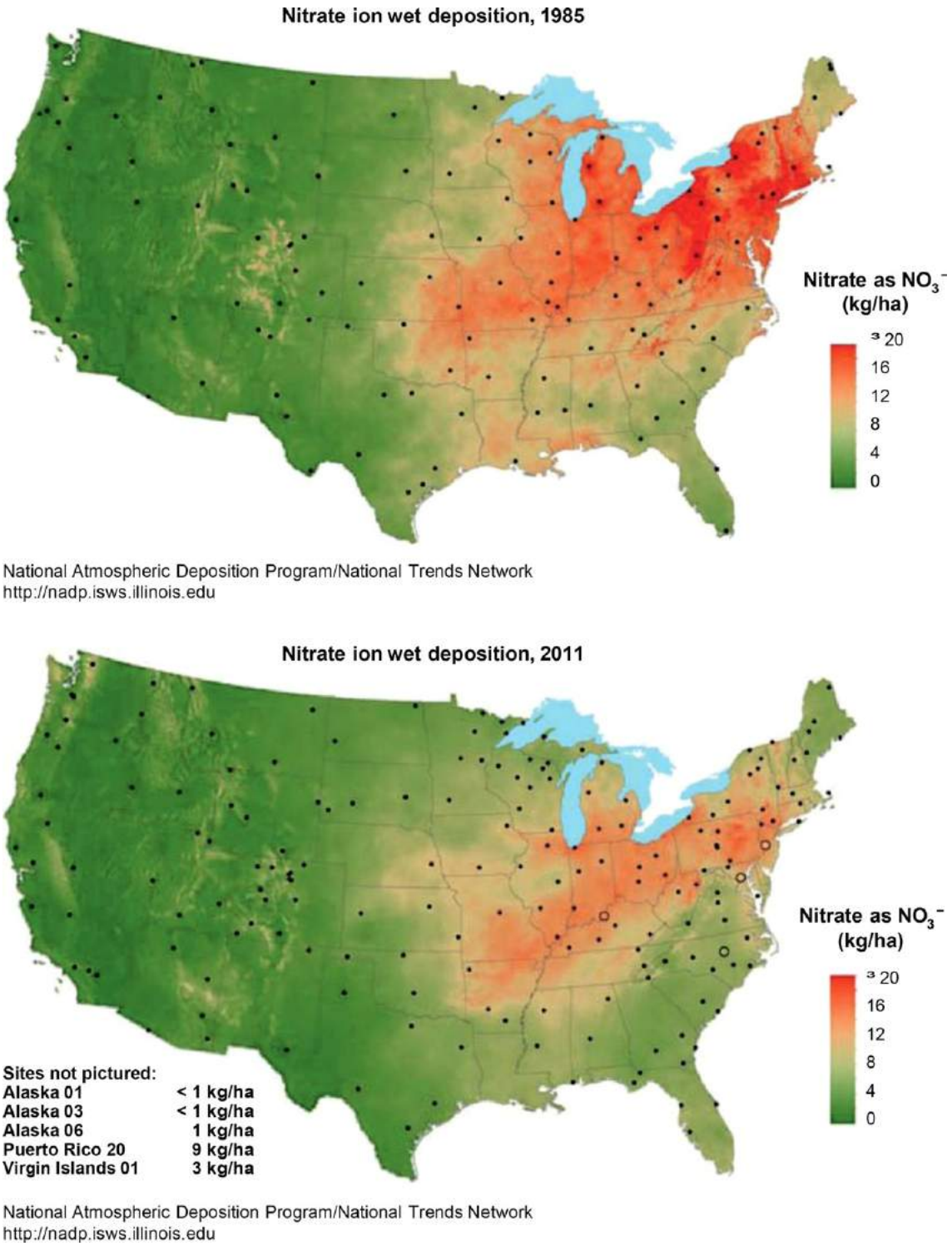


Fig. 2 Map of wet nitrate deposition for the United States for 1985 and 2011. Data are from the National Atmospheric Deposition Program (NADP) network.

Effects of Acidic Deposition on Forest Soils

Research has shown that acidic deposition has chemically altered soils with serious consequences for acid-sensitive ecosystems. Soils compromised by acidic deposition lose their ability to neutralize continuing inputs of strong acids, provide poorer growing conditions for plants, and extend the time needed for ecosystems to recover from acidic deposition. Acidic deposition has altered,

Table 1 The links between sulfur dioxide and nitrogen oxide emissions, acidic deposition, and other important environmental issues

Problem	*Linkage to acid deposition*
Coastal eutrophication	Atmospheric deposition adds nitrogen to coastal waters
Mercury	Surface water acidification increases mercury accumulation in fish
Visibility	Sulfate aerosols diminish visibility and view
Tropospheric ozone	Emissions of nitrogen oxides contribute to the formation of ozone

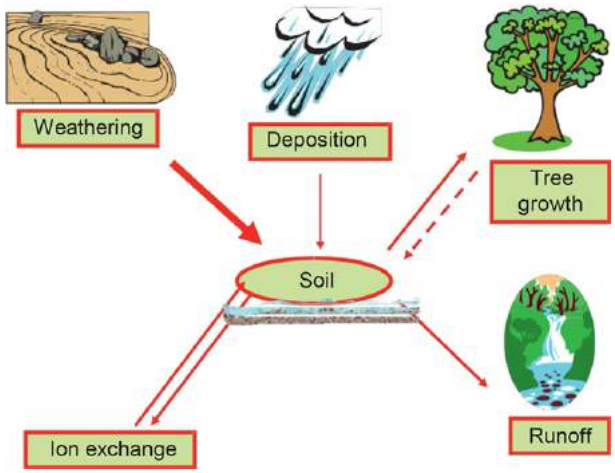


Fig. 3 Conceptual diagram illustrating the inputs and losses of calcium and other nutrient cations for forest ecosystems. Ecosystems with low weathering rates have low supplies of calcium and other nutrient cations available to soil and trees. High inputs of acidic deposition increase the leaching of calcium and other nutrient cations from soil. In acid-sensitive regions with low weathering rates, this can deplete soil pools of available calcium and other nutrient cations and limit the quantity available for tree growth.

and continues to alter, soils in sensitive regions in three important ways. Acidic deposition depletes available calcium and other nutrient cations from exchange sites in soil; facilitates the mobilization of dissolved inorganic aluminum into soil water; and increases the accumulation of sulfur and nitrogen in soil.

Loss of Calcium and Other Nutrient Cations

The cycling of calcium and other nutrient cations in forest ecosystems involves the inputs and losses of these materials (Fig. 3). For most forest ecosystems, the supply of calcium and other nutrient cations occurs largely by weathering (i.e., the breakdown of rocks and minerals in soil). Calcium and other nutrient cations may also enter forests by atmospheric deposition, although this pathway is generally much smaller than that of weathering. Losses occur largely by vegetation uptake and drainage waters. An important pool of ecosystem calcium and nutrient cations is the soil available pool or the soil cation exchange complex. Plants are generally able to utilize this source of nutrients. Forest ecosystems that are naturally sensitive to acidic deposition are generally characterized by low rates of weathering and generally low quantities of available nutrient cations. Under conditions of elevated inputs of acidic deposition and subsequent transport of sulfate and nitrate in drainage waters, nutrient cations will be displaced from available pools and leached from soil. This condition is not problematic for areas with high weathering rates and high pools of available nutrient cations. However, in acid-sensitive areas typically at higher elevation and with shallow soils, which contain minerals that are resistant to weathering, the enhanced loss of calcium and other nutrient cations can result in a depletion of available soil pools.

Over the last century, acidic deposition has accelerated the loss of large amounts of available calcium from acid-sensitive soil in acid-sensitive areas. Depletion occurs when nutrient cations are displaced from the soil by acidic deposition at a rate faster than they can be replenished by the slow breakdown of rocks or the deposition of nutrient cations from the atmosphere. This depletion of nutrient cations fundamentally alters soil processes, compromises the nutrition of some trees, and hinders the capacity for acid-sensitive soils to recover. For example, more than half of the available calcium has been lost from soil at the Hubbard Brook Experimental Forest, New Hampshire, over the past 60 years. Note that while acidic deposition to acid-sensitive areas is decreasing and there is some associated recovery of the acid-neutralizing capacity (ANC) of surface waters (see Surface Water Acidification section), in many regions the levels of acidic deposition remain high enough to continue to deplete exchangeable nutrient cations from soil. Moreover because soil weathering is a slow process, the depletion of cations from soil may limit the recovery of acid-impacted ecosystems over the long term.

Mobilization of Aluminum

Aluminum is often released from soil to soil water, lakes, and streams in forested regions receiving high acidic deposition, low stores of available calcium, and high soil acidity. One of the most significant ecological effects of acidic deposition is the mobilization of aluminum from soil and a shift in the form of aluminum in water from nontoxic organic forms to highly toxic inorganic forms. Concentrations of aluminum increase markedly with decreases in pH, particularly the toxic inorganic forms of aluminum. It is evident that concentrations of aluminum increase exponentially when surface water pH decreases below 6. Aluminum concentrations are thought to be ecologically significant when they increase to values above 2 mmol L^{-1}. This condition occurs below pH 6.0.

High concentrations of dissolved inorganic aluminum can be toxic to plants, fish, and other organisms. Concentrations of dissolved inorganic aluminum in streams in eastern North America and areas of Europe are often above levels considered toxic to fish and much greater than concentrations observed in forest watersheds that receive low inputs of acidic deposition.

Accumulation of Sulfur and Nitrogen in Soil

Elevated inputs of sulfur and nitrogen deposition also result in enhanced accumulation of these elements in soil. This response is generally not thought to have adverse effects on forest ecosystems. Indeed, elevated nitrogen deposition has the benefit of increasing the sequestration of carbon in soil, which offsets, to some extent, emissions of carbon dioxide to the atmosphere. Negative consequences of the accumulation of sulfur and nitrogen in soil will be realized only if these materials are mobilized to surface waters in response to decreases in inputs from atmospheric deposition. In the case of nitrogen, ecosystem accumulation of this growth-limiting nutrient can occur until it becomes no longer growth limiting. High inputs of nitrogen can result in a condition referred to as "nitrogen saturation," which coincides with increases in the leaching of nitrate to surface waters and the associated soil and water acidification. In the case of sulfate, it is unclear if soil sulfur is mobilized following decreases in atmospheric sulfur deposition or the associated eutrophication of ecosystems.

Effects of Acidic Deposition on Trees

Acidic deposition has contributed to the decline of red spruce and sugar maple trees in the eastern United States (Fig. 4). Symptoms of tree decline include poor condition of the canopy, reduced growth, and unusually high levels of mortality. The

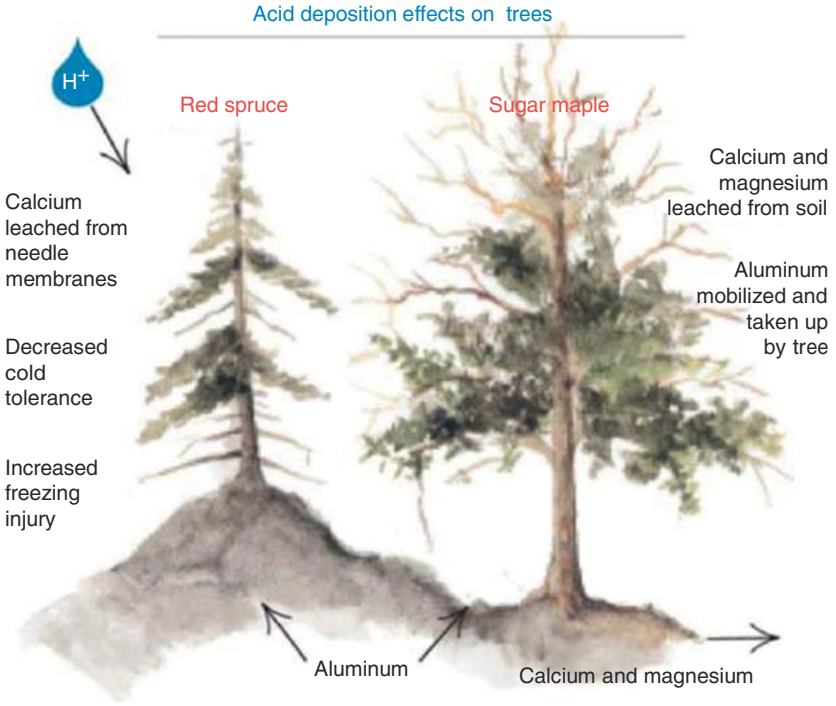


Fig. 4 Conceptual diagram illustrating the mechanisms by which acidic deposition impacts red spruce and sugar maple. Acidic deposition impacts red spruce through loss of membrane calcium due to direct leaching from foliage or reduced uptake of calcium from soil. The loss of membrane calcium makes red spruce more susceptible to winter injury. Acidic deposition results in loss of soil available calcium and magnesium and less uptake by sugar maple. This condition may make sugar maple more susceptible to insect or drought stress.

decline of red spruce and sugar maple in the northeastern United States has occurred over the last four decades. Factors associated with the decline of both species have been studied and include important links to acidic deposition.

Red Spruce

Acidic deposition is the major cause of red spruce decline at high elevations in the northeast. Many large canopy red spruce in the Adirondack Mountains of New York, the Green Mountains of Vermont, and the White Mountains of New Hampshire have died since the 1960s. Significant growth declines and winter injury to red spruce have been observed throughout its range.

Red spruce decline occurs by both direct and indirect effects of acidic deposition. Direct effects include the leaching of calcium from leaves and needles of trees (i.e., foliage), whereas indirect effects refer to acidification of the underlying soil chemistry.

The decline of red spruce is linked to the leaching of calcium from cell membranes in spruce needles by acid mist or fog. The loss of calcium renders the needles more susceptible to freezing damage, thereby reducing the tolerance of trees to low temperatures and increasing the occurrence of winter injury and subsequent tree damage or death. In addition, elevated aluminum concentrations in the soil may limit the ability of red spruce to take up water and nutrients through its roots. Water and nutrient deficiencies can lower the tolerance of trees to other environmental stresses and cause decline.

Sugar Maple

The decline of sugar maple has been studied in the eastern United States since the 1950s. Extensive mortality among sugar maples appears to have resulted from low levels of nutrient cations, coupled with other stresses such as insect defoliation or drought. The probability of decreases in the vigor of the sugar maple canopy or incidence of tree death increased on sites where the supply of calcium and magnesium to soil and foliage is lowest and stress from insect defoliation and/or drought is high. Low levels of nutrient cations can cause a nutrient imbalance and reduce the ability of a tree to respond to stresses such as insect infestation and drought.

Nitrogen Impacts

Inputs of nitrogen are generally considered to benefit forest ecosystems, because nitrogen is generally the growth-limiting element. However, recent research in the eastern United States has demonstrated a varied response to elevated nitrogen inputs, with many tree species apparently benefiting from this condition, but at the same time the health of some declining, such as red spruce, red pine, and white cedar. It appears that there are "winners and losers" as ecosystems respond to changes in the "chemical climate."

Effects of Acidic Deposition on Freshwater Aquatic Ecosystems

Surface Water Acidification

Acidic deposition degrades surface water quality by lowering pH (i.e., increasing acidity); decreasing ANC; and increasing dissolved inorganic aluminum concentrations. While sulfate concentrations in lakes and streams have decreased in eastern North America and Europe over the last 30 years, they remain high compared to background conditions (e.g., approximately 20 meq L^{-1}). Many lakes and streams in acid-impacted regions have shown some chemical recovery in response to emission controls, but have not achieved full recovery.

Acidification of surface waters due to elevated inputs of acidic deposition has been reported in many acid-sensitive areas receiving elevated inputs of acidic deposition, including Great Britain; Nordic countries; Northern, Central, and Eastern Europe; southwestern China; southeastern Canada; the northeastern United States; the Upper Midwest; and the Appalachian mountain region of the United States. Large portions of the high-elevation western United States are also potentially sensitive to acidic deposition; however, atmospheric deposition to this region is relatively low. Concern over effects of acidic deposition in the mountainous regions of the western United States may be overshadowed by the potential effects of elevated nitrogen deposition, including eutrophication of naturally nitrogen-limited lakes.

To illustrate the regional impacts of acidic deposition, a comprehensive survey of lakes greater than 0.2 ha in surface area in the Adirondack region of New York was conducted to obtain detailed information on the acid–base status of waters in this region. Of the 1469 lakes surveyed, 24% had summer pH values below 5.0. Also, 27% of the lakes surveyed were chronically acidic (i.e., ANC less than 0 meq L^{-1}) and an additional 21% were susceptible to episodic acidification (i.e., ANC between 0 and 50 meq L^{-1}). An analysis of the anion content of these lakes illustrates that these lakes have been acidified predominantly by atmospheric deposition of sulfate.

Seasonal acidification is the periodic increase in acidity and the corresponding decrease in pH and ANC in streams and lakes, which generally occurs during the higher flow fall, winter, and spring periods. Episodic acidification is caused by the sudden pulse of acids and a dilution of acid-neutralizing base cations (e.g., calcium, magnesium, sodium, potassium) during spring snowmelt and large rain events in the spring and fall. Increases in nitrate are often important to the occurrence of acid episodes. These conditions tend to occur when trees are dormant and, therefore, retaining less nitrogen. At some sites, short-term increases in

sulfate and organic acids can also contribute to episodic acidification. Episodic acidification often coincides with pulsed increases in concentrations of dissolved inorganic aluminum. Short-term increases in acid inputs to surface waters can reach levels that are lethal to fish and other aquatic organisms. All the acid-sensitive and acid-impacted regions discussed in this article have documented effects associated with episodic acidification.

Trends in surface water chemistry in Europe and eastern North America indicate that recovery of aquatic ecosystems impacted by acidic deposition has been occurring over a large geographic scale since the early 1980s. Some regions are showing rather marked recovery, while others exhibit low or nonexistent increases in ANC. Based on long-term monitoring, virtually all surface waters impacted by acidic deposition in Europe and eastern North America exhibit decreases in sulfate concentrations. This pattern is consistent with decreases in emissions of sulfur dioxide and atmospheric sulfate deposition. The exception to this pattern is streams in unglaciated regions, such as the central and southern Appalachian Mountains. Watersheds in this region and other portions of the southeastern United States exhibit strong adsorption of atmospheric sulfate deposition by highly weathered soils, and only some are exhibiting modest recovery from decreases in acidic deposition. In Europe, the most marked decreases in surface water sulfate have occurred in the Czech Republic and Slovakia, regions that historically experienced very high rates of atmospheric sulfate deposition. More than half of the surface waters monitored in Europe are showing increases in ANC. The rate of ANC increase in Europe is relatively high. This pattern is, in part, due to the relatively high rates of sulfate decreases. In the United States, some regions are showing statistically significant increases in ANC, including lakes in the Adirondacks, New England, and Upper Midwest, and streams in the Northern Appalachian Plateau.

Three factors limit the chemical recovery of the water quality of acid-impacted surface waters, despite the decreased deposition of sulfate. First, levels of acid-neutralizing base cations in streams have decreased markedly due to a loss of base cations from the soil and, to a lesser extent, a reduction in atmospheric inputs of base cations. Second, inputs of nitric acid have acidified surface waters and elevated their concentration of nitrate in many acid-impacted regions. Finally, sulfur that has accumulated in the soil, particularly in the southeastern United States, may be released to surface water as sulfate, even though sulfate deposition has decreased. It appears that the best approach to accelerate the recovery of acid-impacted lakes is to make additional cuts in emissions of sulfur dioxide and nitrogen oxides.

The decreases in sulfate and nitrate concentrations and increases in pH and ANC exhibited in some surface waters is an encouraging sign that impacted ecosystems are responding to emission controls and moving toward chemical recovery. Nevertheless, the magnitude of these changes is small compared to the magnitude of increases in sulfate and nitrate and decreases in ANC and pH that have occurred in acid-impacted areas following historical increases in acidic deposition. However, as discussed earlier, in many acid-sensitive regions, soils continue to acidify despite decreases in acidic deposition. Moreover, many acid-impacted watersheds in northern North America and Europe are showing increases in concentrations of dissolved organic carbon in response to decreases in acidic deposition. Dissolved organic carbon is indicative of the release of naturally occurring organic acids from soil. It is thought that acid deposition limited the transport of these naturally occurring organic acids to surface waters. Now with recovery, this process is being restored. However, as organic acids are acids, increased concentrations will limit the extent of pH and ANC increases in recovering ecosystems.

Response of Aquatic Biota to Acidification of Surface Waters by Acidic Deposition

Decreases in pH and elevated concentrations of dissolved inorganic aluminum have resulted in physiological changes to organisms and direct mortality of sensitive life-history stages, and reduced the species diversity and abundance of aquatic life in many streams and lakes in acid-impacted areas. Fish have received the most attention to date, but entire food webs are often adversely affected.

Decreases in pH and increases in aluminum concentrations have diminished the species diversity and abundance of plankton, invertebrates, and fish in acid-impacted surface waters. A detailed summary of the response of aquatic biota to the acidification of surface waters is provided in **Table 2**.

In the Adirondacks, a significant positive relationship exists between the pH and ANC levels in lakes and the number of fish species present in those lakes (**Fig. 5**). Surveys of 1469 Adirondack lakes conducted in 1984 and 1987 show that 24% of lakes (i.e., 346) in this region do not support fish. These lakes had consistently lower pH and ANC, and higher concentrations of aluminum than lakes that contained one or more species of fish. Experimental studies and field observations demonstrate that even acid-tolerant fish species such as brook trout have been eliminated from some waters in New York.

Similar relationships are evident in surface waters in acid-impacted regions throughout the world. Studies demonstrate the effects of acidic deposition on fish at three ecosystem levels:

- Effects on single organisms (condition factor—the relationship between the weight and the length of a fish). Fish condition factor is related to several chemical indicators of acid–base status, including minimum pH. This analysis suggests that fish in acidic streams use more energy to maintain internal chemistry than would otherwise be used for growth.
- Population-level effects (increased mortality). Bioassay experiments show greater mortality in chronically acidic streams than in high ANC streams. Egg and fry are sensitive life-history stages for fish.
- Community-level effects (reduced species richness). The species richness of fish and other aquatic organisms decreases with decreasing ANC and pH.

Table 2 Biological effects of surface water acidification in North America

pH decrease	*General biological effects*
6.5–6.0	Small decrease in species richness of phytoplankton, zooplankton, and benthic invertebrate communities resulting from the loss of a few highly acid-sensitive species, but no measurable change in total community abundance or production
	Some adverse effects (decreased reproductive success) may occur in highly acid-sensitive species (e.g., fathead minnow, striped bass)
6.0–5.5	Loss of sensitive species of minnow and dace, such as blacknose dace and fathead minnow; in some waters decreased reproductive success of lake trout and walleye, which are important sport fish species in some areas
	Visual accumulations of filamentous green algae in the littoral zone of many lakes, and in some streams
	Distinct decrease in the species richness and change in species composition of the phytoplankton, zooplankton, and benthic invertebrate communities, although little, if any, change in total community biomass or production
5.5–5.0	Loss of several important sport fish species, including lake trout, walleye, rainbow trout, and smallmouth bass; as well as additional nongame species such as creek chub
	Further increase in the extent and abundance of filamentous green algae in lake littoral areas and streams
	Continued shift in the species composition and decline in species richness of the phytoplankton, periphyton, zooplankton, and benthic invertebrate communities; decrease in the total abundance and biomass of benthic invertebrates and zooplankton may occur in some waters
	Loss of several additional invertebrate species common in oligotrophic waters, including *Daphnia galeata mendotae*, *Diaphanosoma leuchtenbergianum*, *Asplanchna priodonta*; all snails, most species of clams, and many species of mayflies, stoneflies, and other benthic invertebrates
	Inhibition of nitrification
5.04.5	Loss of most fish species, including the most important sport fish species such as brook trout and Atlantic salmon; few fish species able to survive and reproduce below pH 4.5 (e.g., central mudminnow, yellow perch, and in some waters, largemouth bass)
	Measurable decline in the whole-system rates of decomposition of some forms of organic matter, potentially resulting in decreased rates of nutrient cycling
	Substantial decrease in the number of species of zooplankton and benthic invertebrates and further decline in the species richness of the phytoplankton and periphyton communities; measurable decrease in the total community biomass of zooplankton and benthic invertebrates in most waters
	Loss of zooplankton species such as *Tropocyclops prasinus mexicanus*, *Leptodora kindtii*, and *Conochilus unicornis*; and benthic invertebrate species, including all clams and many insects and crustaceans
	Reproductive failure of some acid-sensitive species of amphibians such as spotted salamanders, Jefferson salamanders, and the leopard frog

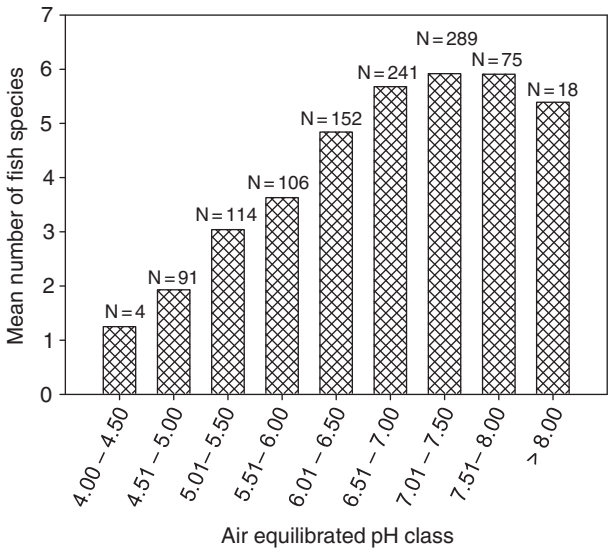


Fig. 5 Distribution of the mean number of fish species for ranges of pH from 4.0 to 8.0 in lakes of the Adirondack region of New York. *N* represents the number of lakes in each pH category.

Although chronically high acid levels stress aquatic life, acid episodes are particularly harmful because abrupt, large changes in water chemistry allow fish few areas of refuge. High concentrations of dissolved inorganic aluminum are directly toxic to fish and pulses of aluminum during acid episodes are a primary cause of fish mortality. High acidity and aluminum levels disrupt the salt and water balance of blood in a fish, causing red blood cells to rupture and blood viscosity to increase. Studies show that the viscous blood strains the heart of a fish, resulting in a lethal heart attack.

While most acid-impacted regions have monitoring programs that document the chemical recovery from acidic deposition, there are far fewer programs that track biological recovery. The limited studies that have been conducted generally show limited and delayed biological recovery in response to chemical recovery. Moreover, the endpoint of biological recovery is unlikely to be that which existed prior to the advent of acidic deposition, because of changes that have occurred to ecosystems in the intervening decades.

Effects of Acidic Deposition on Marine Ecosystems

Due to increased carbon dioxide emissions resulting from anthropogenic activities, hydrogen ion concentrations in surface oceans are predicted to increase 150% by 2100, which may pose threats for diverse marine organisms, particularly, shell-forming and calcifying organisms.

In the case of some commercially exploited filter-feeders, such as oysters, it is likely that the physiological impacts related to reductions in calcium carbonate availability affect ecosystem functioning and the provision of ecosystem services provided by these organisms. Nonetheless, the effects of other stressors (e.g. warming, eutrophication, metal contamination) cannot be discarded and are difficult to evaluate.

On the other hand, infaunal marine organisms, such as the clam *Mya arenaria*, which were considered to be relatively robust to changes in seawater pH and carbonate geochemistry, seem to suffer from several order impacts caused by sediment acidification, such as shell dissolution, behavioural alterations, more lesions, increased mortality.

Ecosystem-level analysis suggest that food web consequences of ocean acidification can extend beyond groups thought most vulnerable and affect fishery yield and ecosystem structure, yet, the indirect effects of ocean acidification on calcifying organisms are uncertain and can be counterintuitive. For example, primary productivity has been found higher at natural CO_2 vents with near-future CO_2 levels, compared to control sites with present-day CO_2 levels, which counterintuitively drove, via provision of more habitat and food, to a greater abundance of herbivorous gastropods.

Other works suggest that, while lower pH impairs the senses of reef fishes and reduces their survival, cephalopods and crustaceans will remain largely unscathed. In the same way, it is projected that habitat changes promoted by ocean acidification reduce seafood production from coral reefs, but increase production from seagrass and seaweed. Thus, the overall effects of ocean acidification on primary production and, hence, on food webs will result in hard-to-predict winners and losers. In fact, results from some experimental and numerical analyses propose that marine biota may be more resistant to ocean acidification than expected. In some cases, ocean acidification will enhance growth of marine autotrophs and reduce fertility and metabolic rates, but effects are likely to be minor along the range of pCO2 predicted for the 21st century, and feedbacks between positive responses of autotrophs and pH may further buffer the impacts.

Ecosystem Response to Air Quality Management

As discussed earlier, aquatic ecosystems have responded in North America and Europe to emission control programs. In Europe and Canada emission control programs have been specifically established to protect ecosystems from the effects of air pollution and acidic deposition. In contrast, in the United States, air quality management has been driven by human health concerns, and ecosystem recovery has been a cobenefit of these initiatives. However, as has historically been the case in Europe and Canada, the United States is starting to recognize the benefits of services associated with healthy ecosystems. This has led to interest in the concept of "Critical Loads" and "Target Loads," an approach that has been used in Europe for decades. A critical load is the loading of an air pollutant to an ecosystem below which harmful ecological effects have not been documented. A critical load is governed by the inherent biophysical characteristics of an ecosystem, such as rates of chemical weathering and soil characteristics. A target load is a political decision and maybe higher or lower than the critical load depending on political and economic considerations. In the United States, the US Forest Service, the National Park Service, and the US Environmental Protection Agency are starting to use critical loads to guide management in protecting ecosystems from air pollution effects. Calculations of critical loads of acidity and nitrogen have been conducted for regions of the US and the entire country. Those regions showing exceedances in critical loads are similar to those discussed above as experiencing surface water acidification (see Surface Water Acidification section).

In some regions, the recovery of ecosystems from acidic deposition has not been fast enough to protect impacted resources. Under these conditions, basic material, such as limestone or calcium carbonate, has been added as a management approach to mitigate the adverse effects of acidic deposition. Addition of basic materials has been successful in the protection of ecosystems from the impacts of acidic deposition. National programs have been implemented in Germany and Sweden. There have been no negative effects of liming in the treatment of acid-impacted watersheds or surface waters.

Further Reading

Aber, J.D., Goodale, C.L., Ollinger, S.V., *et al.*, 2003. Is nitrogen deposition altering the nitrogen status of northeastern forests? Bioscience 53, 375–389.
Branch, T.A., DeJoseph, B.M., Ray, L.J., Wagner, C.A., 2013. Impacts of ocean acidification on marine seafood. Trends in Ecology & Evolution 28 (3), 178–186.

Charles, D.F. (Ed.), 1991. Acidic deposition and aquatic ecosystems: Regional case studies. New York: Springer.
Clements, J.C., Hunt, H.L., 2017. Effects of CO_2-driven sediment acidification on infaunal marine bivalves: A synthesis. Marine Pollution Bulletin 117, 6–16.
Connell, S.D., Doubleday, Z.A., Hamlyn, S.B., Foster, N.R., Harley, C.D.G., Helmuth, B., Kelaher, B.P., Nagelkerken, I., Sarà, G., Russell, B.D., 2017. How ocean acidification can benefit calcifiers. Current Biology 27, R83–R102.
DeHayes, D.H., Schaberg, P.G., Hawley, G.J., Strimbeck, G.R., 1999. Acid rain impacts calcium nutrition and forest health. Bioscience 49, 789–800.
Driscoll, C.T., Lawrence, G.B., Bulger, A.J., *et al.*, 2001. Acidic deposition in the northeastern U.S.: Sources and inputs, ecosystems effects, and management strategies. Bioscience 51, 180–198.
Driscoll, C.T., Cowling, E.B., Grennfelt, P., Galloway, J., Dennis, R., 2010. Integrated assessment of ecosystem effects of atmospheric deposition: Lessons available to be learned. EM Magazine 11, 6–13.
Evans, C.D., Cullen, J.M., Alewell, C., *et al.*, 2001. Recovery from acidification in European surface waters. Hydrology and Earth System Sciences 5, 283–298.
Fay, G., Link, J.S., Hare, J.A., 2017. Assessing the effects of ocean acidification in the Northeast US using an end-to-end marine ecosystem model. Ecological Modelling 347, 1–10.
Fenn, M.E., Baron, J.S., Allen, E.B., *et al.*, 2003. Ecological effects of nitrogen deposition in the western United States. Bioscience 53, 404–420.
Gorham, E., 1989. Scientific understanding of ecosystem acidification: A historical review. Ambio 3, 150–154.
Greaver, T.L., Sullivan, T.J., Herrick, J.D., *et al.*, 2012. Ecological effects of nitrogen and sulfur air pollution in the US: What do we know? Frontiers in Ecology and the Environment 10, 365–372.
Hendriks, I.E., Duarte, C.M., Álvarez, M., 2010. Vulnerability of marine biodiversity to ocean acidification: A meta-analysis. Estuarine, Coastal and Shelf Science 86, 157–164.
Lehmann, C.M.B., Bowersox, V.C., Larson, S.M., 2005. Spatial and temporal trends of precipitation chemistry in the United States, 1985–2002. Environmental Pollution 135, 347–361.
Likens, G.E., Driscoll, C.T., Buso, D.C., 1996. Long-term effects of acid rain: Response and recovery of a forested ecosystem. Science 272, 244–246.
Lemasson, A.J., Fletcher, S., Hall-Spencer, J.M., Knights, A.M., 2017. Linking the biological impacts of ocean acidification on oysters to changes in ecosystem services: A review. Journalof Experimental Marine Biology and Ecology 492, 49–62.
Pardo, L.H., Fenn, M., Goodale, C.L., *et al.*, 2011. Effects of nitrogen deposition and empirical nitrogen critical loads for ecoregions of the United States. Ecological Applications 21, 3049–3082.
Schindler, D.W., Mills, K.H., Malley, D.F., *et al.*, 1985. Long-term ecosystem stress: Effects of years of experimental acidification. Canadian Journal of Fisheries and Aquatic Sciences 37, 342–354.
Stoddard, J.L., Jeffries, D.S., Lukewille, A., *et al.*, 1999. Regional trends in aquatic recovery from acidification in North America and Europe. Nature 401, 575–578.
Widdicombe, S., Dashfield, S.L., McNeill, C.L., Needham, H.R., Beesley, A., McEvoy, A., Øxnevad, S., Clarke, K.R., Berge, J.A., 2009. Effects of CO_2 induced seawater acidification in infaunaldiversity and sediment nutrient fluxes. Marine Ecology Progress Series 379, 59–75.
Wood, H.L., Widdicombe, S., Spicer, J.I., 2009. The influence of hypercapnia and macrofauna on sediment nutrient exchange? Biogeosciences 6, 2015–2024.

Ecological Efficiency☆

Lawrence B Slobodkin, SUNY, Stony Brook, NY, United States

Glossary

Ecosystem processes An intrinsic characteristic whereby an ecosystem maintains its integrity. Processes include decomposition, production, nutrient cycling, and fluxes of nutrients and energy.

Energy transfer The conversion of one form of energy into another, or the movement of energy from one place to another.

Entropy Is a measure of evenness of a distribution of energy between parts of a system.

Steady-state systems Matter entering the system is equivalent to the matter exiting the system.

Trophic level The position an organism occupies in the food chain.

Introduction

Historically, studies of multispecies assemblages have often focused on "energy." There were three general reasons for this:

1. Units of energy relate to all biological activity.
2. There was, for many years, the hope that a focus on energy would somehow bring ecology under the purview of thermodynamics.
3. Finally, once it had become customary to use energy units, there was no strong reason to change.

Concern with energy began as part of an attempt to extend the formal rigor of thermodynamics to biology. Lotka developed a diagram of an abstract ecological system at steady state. At the suggestion of Hutchinson, Lindeman used the notation of Lotka to describe the passage of energy through a lake.

In this trophic dynamic scheme, ecosystems are represented as a set of trophic levels. All the photosynthesizers (green plants, algae, and many kinds of colored bacteria) can be labeled as trophic level 0, the herbivores as level 1, carnivores feeding on herbivores as level 2, carnivores feeding on them as level 3, etc.

Ecological efficiency was rigorously defined by Lindeman as the fraction of the energy that is consumed by organisms on one trophic level that serves as nourishment for organisms on the next higher trophic level (**Fig. 1**). Any definable portion of a community has an ecological efficiency if, and only if, it is possible to measure the energy per unit time taken from it by predators and the energy that it consumed from its prey or food organisms or, in the case of photosynthetic organisms, from solar radiation.

The trophic dynamic diagram of Lindeman was not perfect. There are omnivores that consume both flesh and vegetation, thereby feeding on several trophic levels. There are multiple tropic levels among the decomposers. There are also cannibals and organisms that change their trophic level as they mature. These considerations complicate assignment of particular species to particular trophic levels.

With the exception of some deep sea and hot spring bacteria, that use chemical energy in the absence of light, photosynthesis directly, or indirectly, provides the energy for all organisms. Photosynthetic organisms gain radiant energy from sunlight and transform it into potential or chemical energy. Some of the chemical energy in plants is passed on to herbivores whence energy may be passed on either to higher-order carnivores or parasites or detritivores. The chemical energy consumed by herbivores is dissipated as heat, stored, or passed further up the food chain. In addition, all trophic levels contribute energy in the form of dead organisms and excreta to decomposer organisms, like bacteria, and molds.

The energy income to plants has been variously defined as the energy of the sunlight impinging on the plants or as the energy actually fixed by photosynthesis. This makes an enormous difference in the estimated value of ecological efficiency for the plants but does not matter for the ecological efficiency estimates of higher trophic levels.

Measuring Ecological Efficiency

To say that a particular population or trophic level has a particular ecological efficiency is an assertion about the ratio of energy received by that population or trophic level and the energy consumed from that population or trophic level by a predator or a next higher trophic level, during the same time interval. It is not measurable from a single population or a single trophic level.

☆*Change History*: March 2018. H Pethybridge included glossary, keywords, minor text edits, added Fig. 1, and updated references.

This is an update of L.B. Slobodkin, Ecological Efficiency, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 1019–1024.

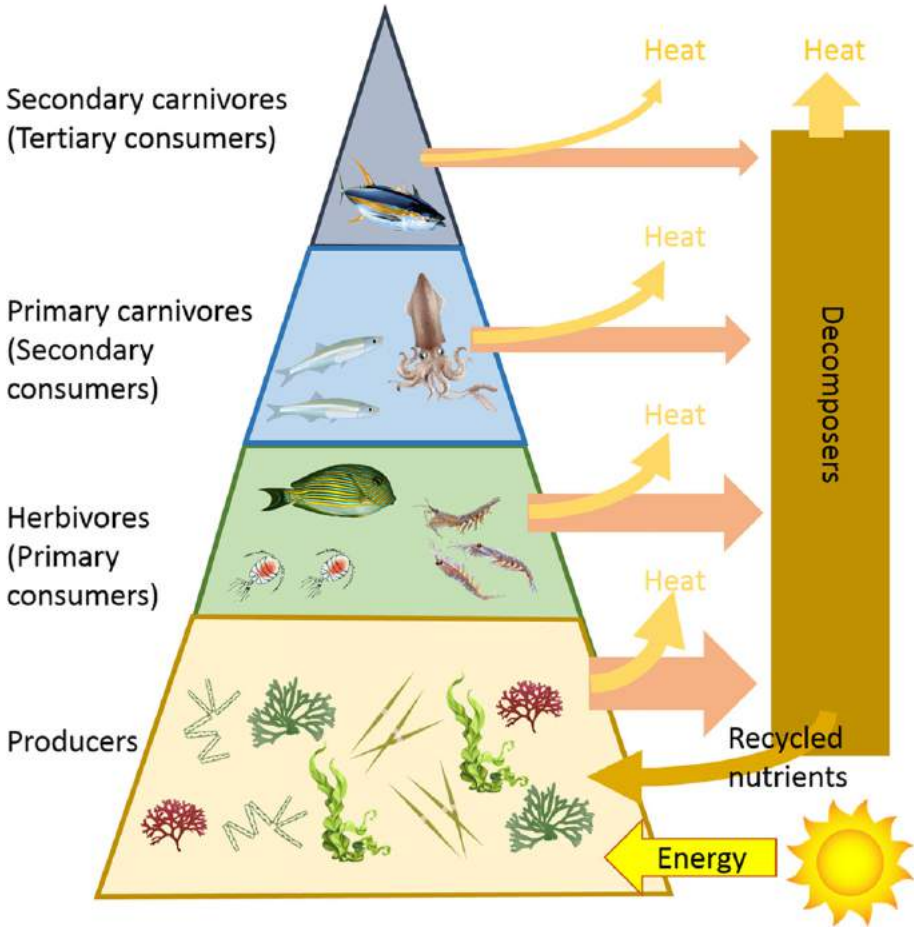


Fig. 1 The amount of energy transferred as biomass from one trophic level to the next (ecological efficiency).

Early efficiency estimates used surrogates for energy (biomass, population counts, total protein, etc.), but for the last 45 years direct energy measurements have been more common. Actual measurements of energy can be made of heat production by living organisms in a calorimeter or by combustion of whole organisms or their parts in a bomb calorimeter.

The caloric content per ash free gram of whole organisms varies between approximately 4500 calories per gram, found in carbohydrates and many whole plants, to 8000–9000 calories per gram found in pure fats. Most animals usually are around 5600, the value for protein mixed with a little fat. Efficiency of particular energy consumption processes can also be evaluated by considering biochemical equations directly, but ecologists do not often do this.

When circumstances permit, all organisms can reproduce and grow far more than is needed for their own replacement, however death rates will exceed the rates of growth and reproduction when circumstances change. Periods of low energy income and high mortality are interspersed with periods of high-energy availability and high growth and reproductive rate.

Ecological efficiency varies seasonally. In very small organisms, growth and multiplication can be very rapid—basically keeping up with energy income. In larger organisms, there are typically periods during which energy income exceeds the possibilities for enhanced growth and reproduction and periods of food shortage when organisms can only rely on stored energy. Higher values for ash-free total calorie concentration in animals are found either when food is particularly abundant or when energy is being stored in advance of some process that will make higher energy demands than can be met immediately by feeding. Large seeds, yolky eggs, prehibernating mammals, and premigratory birds have high caloric density.

In very small aquatic organisms, like in diatoms and copepods, oil droplets may serve for flotation as well as energy storage. In large animals like walrus, whales, and bears, fat may also serve as thermal insulation.

The Relation Between Population Maintenance and Ecological Efficiency

Generally, prey convert energy from an inedible form, which the predators cannot use as a food supply, to a more digestible and convenient form. Plants turn sunlight into food for animals. Also, a population of mice can transform inedible grain

into food for cats so that a steady-state population of mice can be considered to have an ecological efficiency from the standpoint of a steady-state population of cats. The ecological efficiency of the mice is defined as calories of mouse eaten per unit time by the cat population divided by calories of cat-inedible food, say grain, eaten per unit time by the mouse population.

How we choose to measure ecological efficiency of any population or trophic level will determine the value observed and will affect the usefulness of the measurement. We might measure the loss of the mouse population to all predators rather than to cats alone. This would also permit a measurement of the ecological efficiency of the mice but it would have a different, legitimate, meaning from that in which only cats were considered.

Both plants and animals that confront seasonality in growth and reproductive conditions have evolved energy storage mechanisms. In motile organisms and in plant seeds adapted for transportation or penetration of the soil, energy storage is usually in the form of fats and oils, which are maximally compact and least disturbing to the external shape of the organisms.

Lower caloric concentration in stored energy is found in organisms in which shape is not as important. Potatoes and other roots and some mollusks enclosed in shells store energy as starch or glycogen rather than fats or oils as do sedentary polychaetes, for example, lugworms. The transition of energy from one trophic layer to the next occurs when organisms are eaten but the process is always accompanied by production of heat, carbon dioxide, and indigestible material. At every trophic level, exudates, detritus, and dead organisms support detritivores, including bacteria and molds. This stepwise transformation of potential energy digestion results in increasingly recalcitrant residual material that finally may be added to sediments.

A steady-state stipulation may be important. If house-cats become excessively competent at catching mice, they may get a large temporary yield of mouse meat at the cost of destroying the mouse population completely. This is generally analogous to any problem of overexploitation of the kind that is encountered in fishing or hunting situations. Some predators may exploit their prey in a way that minimizes the danger to the persistence of the prey population. This is termed "prudent predation" and there was, at one time, a concern that this involved "group selection." The problem disappeared when it was shown that the focus of individual selection was on the escape and survival capacity of the prey and not on actual "prudence."

Unless the focus population (the one whose ecological efficiency is being measured) is a population of plants, it consumes energy from some other population, the food or prey population. Other organisms in the consumer, or predator, populations may eat energy-rich material from the focus population. The remainder of material produced by the focus population may be thought of as related to maintenance.

A population can have higher ecological efficiency if minimal energy is used for its own population maintenance. "Maintenance" costs are the activities of, and products from, the focus population that would occur even in the absence of predators or herbivores. Some of the products of maintenance are still energy rich and supply energy income to decomposer organisms—ranging from molds and bacteria to detritivorous animals. These materials include dead bodies or body parts and exudates, for example, dead leaves, fallen logs, molted skin, skin fragments, and mucous, fecal material, and loose scales. Maintenance requires oxygen consumption, produces carbon dioxide, and degrades some organic materials.

In moist and oxygenated situations, molds and bacteria and detritivores decompose organic wastes relatively quickly. Energy-rich compounds that are not consumed become humic stains in sediments, peat, coal, or oil and may eventually be burned as fossil fuels.

Farmers Are Not Predators

Focal populations may be "farmed" rather than preyed upon. Cows turn hay and feed into meat and milk, of interest to the farmer who maintains the animals in such a way as to minimize energy expenditure in exercise and aggression. Placid cows are more efficient but the entire situation in which humans do part of the work of maintenance of the population is highly unnatural.

In a farming situation, the consumption and the yield from a focal population can be measured in units of money and time. The ratio of price of produce to cost of husbandry is important, but since it is not defined in energetic terms it is not an ecological efficiency. Recall that part of the motivation of Lotka, Lindeman, and Hutchinson was to make theoretical contact with thermodynamics.

Ecological efficiencies in the sense of Lindeman, used in most ecological literature, do not have any meaningful definition for entire communities or entire ecosystems. There is a tendency to confuse the ideas of efficiency and effectiveness.

Geese can eat grass off golf courses, birds may eat noxious insects, and bacteria can help digest organic matter in sewage effluent. These valuable processes are only remotely connected to ecological efficiency. Advocates for vegetarianism quite correctly note that the amount of resources—sunlight, water, and fertilizers needed to grow animal flesh—is several times larger than that needed to produce an amount of vegetable food of equivalent or superior nutritional quality.

Ecosystems or communities can be managed for particular purposes, for example, maximum steady-state yields of fish, seaweed, grass, elephant tusks, or game. Management can increase these yields, but this is deeply different from ecological efficiency, which cannot be measured for an entire system. Animals and plants can be more or less effective at performing desired functions. Effectiveness at particular processes is of immediate intuitive value. It is less obvious why ecological efficiency matters for any practical purpose.

A high ecological efficiency does not imply a large amount of material being maintained at any trophic level.

Efficiency, Yield, and Stability

In many circumstances, ecological efficiency can be increased by more intensive predation, but only at the cost of reducing the availability of energy for population maintenance and probably lowering the total yield from that population. For example, if the fishing rate is excessive, a prey population may have a very high ecological efficiency, but also a dangerously low population size.

Whole organisms or parts of organisms in the focus population may be eaten. Feeding on another organism does not automatically infer the death of that organism. Parts of living plants may be eaten without killing the plants as in the feeding of bovines on grasses. Certainly, most macroscopic plants are partially eaten by herbivores as can be confirmed by examining almost any leaf. Almost all leaves have their outline broken by missing portions and their surfaces marked by tunnels, galls, or holes.

Some animals can be eaten in such a way as to not kill them. For example, parrotfish browse on the branches of coral without killing the entire coral. Stone crab claws are commercially harvested by breaking them off the living animals that are then released to presumably regenerate new claws. Clams are often fed upon by fishes in such a way as to leave the animals alive, removing only vulnerable parts of their anatomy—particularly the siphons.

Obviously, there are many procedures by which a focus population may avoid predators or at least minimize the harm from higher trophic levels. Mechanisms range from the flight of small birds before a hawk, the secretion of foul odors by many insects, and digging further into sediments by clams in the wake of crab attacks. Correspondingly, there are an endless variety of ways in which predators attack their prey. All parasites can be considered predators that have evolved ways of feeding that do not immediately kill their prey.

Most of the solar energy impinging on green plants dissipates as heat. The rest drives the photosynthetic process, converting light energy to chemical energy in organic molecules. The photosynthetic organisms usually use most of the sugars and carbohydrates that they produce for growth, reproduction, and repair.

Usually there is not a perfect local balance between photosynthetic energy fixation and energy losses to respiration. Residual material, particularly in oligotrophic situations, is a relatively small fraction of the organic material produced by the photosynthesizers in that ecosystem. This residual material may be washed away and used by some distant organisms, or buried, joining sediment or adding to fossil fuel or to the brown color of soil. The existence and condition of organic material in sediments depends broadly on the availability of water and oxygen for decomposer bacteria and molds.

The relative quantities of consumed energy that are dissipated as heat, used in growth and reproduction, stored as fats or oils, or passed on to higher trophic levels, vary among organisms, and vary with different times and circumstances. However, heat and buried or dissolved carbon compounds are the universal ends of all energy that enters an ecosystem.

Because some potential energy is converted to heat at each trophic level, the total energy per unit time that flows through the system must decrease with trophic level. If the maintenance cost per gram of live standing crop tissue is approximately independent of trophic level, a pyramid of standing crop abundance is generated, the classical "Eltonian pyramid."

When growth and metabolic rates (per calorie-day maintained) are very high in the lower trophic levels, and relatively low in organisms at higher trophic levels, the trophic pyramid can be inverted—with small standing crops of plants and herbivores turning over rapidly at the bottom while maintaining large populations at higher trophic levels. Examples include some aquatic systems that are based on phytoplankton.

Those organisms that are not food for any other organisms in the system have by definition ecological efficiency equal to zero. While the ecological efficiency of a top predator population is usually considered to be zero, but parasites on or in these predators can be taken as an even higher trophic.

In fact, there is probably no animal or plant population that actually has ecological efficiency of zero. Occasionally, even elephants, wolves, and tigers, if they are very young or very old, fall to predators or carry parasites, and even the driest and most toxic of vegetation usually shows obvious signs of having been partially consumed by some herbivore or mold.

In addition to feeding predators, energy is used in processes of growth and reproduction. Plants and animals both do work in moving internal fluids and intracellular parts. Animals are more mobile than plants, and a correspondingly larger portion of their ingested energy is expended as movement against external forces and friction.

In a growing population, the rate of energy consumption is greater than the rate at which energy is discharged from the population. At its upper limit, ecological efficiency is greatest when a population is rapidly growing and such a short time interval is considered that no deaths or losses of potential energy from individuals have occurred. At the lower limit, ecological efficiency of the focus population goes to zero as a starving population approaches extinction or if there are no predators.

In short, ecological efficiency will vary with conditions. Ecological efficiency will approach a steady state only if abundance, age distribution, and size distribution of the organisms in the focal population are not changing and neither are the predation rate and the distribution of age and size of the prey.

Ecological Efficiency and Evolution

The central mechanism of microevolutionary change involves selection among alleles in a single population. Measuring selection is quite different from measuring ecological efficiency. There are many ways of organizing an energy budget but ecological efficiency determination always requires measurements on at least three populations. This is extremely important in the context of how natural selection impinges on ecological efficiency.

There is no way for natural selection to act on ecological efficiency directly, but the interaction between populations of predators and the populations of their prey is inherently unstable in the absence of natural selection. It has long been known that the interaction of predators and prey can be unstable in laboratory experiments and in mathematical models. Typically, the last prey is eaten and the predators then starve to death unless special mechanisms and properties exist that prevent or at least delay this collapse of the system.

Obviously, predators and their prey are both found in nature, as if there were evolutionary mechanisms for stabilizing their coexistence. The actual food webs and food chains we find are the survivors of a massive group selective selection process—all the unstable webs and food chains are gone. In addition, the natural selection on the process of predation and on the escape of prey while not selecting for any particular value of ecological efficiency will tend toward stabilization of the particular predator–prey system involved.

Thermodynamics and Ecological Efficiency

There are many books and papers that use arguments about entropy in ecological systems. While we can focus on particular predators and consider the ecological efficiency of a particular focus population from the standpoint of particular predators, detritivores, etc., the boundary conditions for analyses of entropy and enthalpy in closed systems are certainly not met by ecological systems. 'Closing' an ecological system in the sense of elementary thermodynamics results in almost immediate death of all organisms in the system. They are not at equilibrium, nor is it generally possible to control temperature properly, and certainly they are not at chemical equilibrium. Ecological efficiencies are dimensionless ratios of rates defined in units of energy per time over some interesting area but they have not been measured within the boundary conditions for thermodynamics or steady-state thermodynamics.

Ecological efficiency involves the idea of a population or trophic level being fed upon by a predator or parasite population while feeding on some lower trophic level, but without the benefit of the farmer's attention. This seems to be a basically unstable arrangement. Why does not a higher trophic level completely eat out the lower one, or why cannot the lower trophic level evolve properties that make it immune to predation by the members of the higher tropic level?

Estimates of ecological efficiency vary widely. Some of this variance is due to actual biological differences but some is due to a variety of technical problems. We generally cannot directly measure transfers of energy in the field. Therefore, ecological efficiency measures require that more readily available measures, such as census data, stomach contents analyses, or some measure of biomass, are converted to calories by use of different conversion constants derived from measure of the caloric content of different categories of organisms in the system combined with field census data.

With the possible exception of relatively rare and large organisms in controlled situations, census evaluations in nature have relatively large and sometimes poorly known errors. Typically, there are not replicates available for the situation being examined.

Also, while steady-state conditions are not required for efficiency measurements, the meaning of non-steady-state evaluations is limited.

Ecological efficiency is not a constant across ecological situations, although this was, for a while, a tempting hypothesis. The same systems examined over different timescales may have different ecological efficiencies.

Attempts to circumvent the problem of sampling, repetition, and lack of a steady state can be made by using laboratory populations, in which food to the populations and yield from the populations can both be controlled. Of course, the relevance of these studies to field values remains uncertain.

When ecological efficiency was evaluated in laboratory populations of *Daphnia* and hydra, they approximately agreed with each other and were within the range of values reported from nature. It can be demonstrated that any apparent constancy of ecological efficiency cannot be the result of evolutionary selection. In fact, it is an epiphenomenon.

Ecological efficiency has a curious relation to natural selection, which may shed light on other evolutionary questions. Using the now classical metaphor of the spandrels of San Marco, Gould and Lewontin elegantly underscored the question of whether there exist biological properties that are not in themselves selected for or against by the evolutionary process but are carried along by selection in the process of selection for some other properties. Ever since, there have been arguments about whether or not such properties exist and whether or not there are any properties of the biological world that are ignored by selection.

An elementary tenet of modern evolutionary theory is the natural selection for differences in fitness among members of a population. Sometimes, the fitness advantages of the selected organisms are not obvious. This can generally be resolved by further study or suitable reexamination of details of the process. If necessary, the problem can often be resolved by considering the effects of selection not only on individuals but also on their kin. A mother whose probability of individual survival and longevity is materially reduced by reproduction may be producing a sufficient number and quality of young so that the total fitness of mother and young may be higher than that of some other mother and young, thereby providing raw material for selection.

In order for natural selection for ecological efficiency to achieve any particular value, the selective process must somehow work in a coordinated way on at least three populations within that community. There is no readily describable mechanism for that to occur. Possibly a massive group selection process acting on entire trophic systems might raise or lower the rates and quantities of energy transfer, but what is much more likely is that one or more of the populations in the trophic system will be eliminated or greatly reduced while ecological efficiency continues to be set by multispecies relationships that have no direct selective meaning.

The actual value of ecological efficiency seems to be an epiphenomenon of selection, operating on all of the components of energy transfer in a trophic system, rather than being in any way the direct target of selection. This implies that there is at least one property of ecological systems that is real and measurable but not related in any clear way to natural selection. Ecological efficiency values require coordination among parts of ecological systems, each of which are themselves subject to natural selection without particular coordination between the different selective forces.

See also: Conservation Ecology: Trophic Index and Efficiency; Turnover Time; Trophic Classification for Lakes. Ecological Data Analysis and Modelling: Conceptual Diagrams and Flow Diagrams; Climate Change Models. General Ecology: Ecological Stoichiometry: Overview. Global Change Ecology: Nitrogen Cycle

Further Reading

Barnes, C., Maxwell, D., Reuman, D.C., Jennings, S., 2010. Global patterns in predator–prey size relationships reveal size dependency of trophic transfer efficiency. Ecology 91 (1), 222–232.
Dickman, E.M., Newell, J.M., González, M.J., Vanni, M.J., 2008. Light, nutrients, and food-chain length constrain planktonic energy transfer efficiency across multiple trophic levels. Proceedings of the National Academy of Sciences 105 (47), 18408–18412.
Hairston, N., Smith, F., Slobodkin, L., 1960. Community structure, population control and competition. American Naturalist 94, 421–425.
Lindeman, R., 1942. The trophic dynamic aspect of ecology. Ecology 23, 399–418.
Lotka, A.J., 1925. Elements of physical biology. Baltimore: Williams and Wilkins.
Odum, E.P., Marshall, S.G., Marples, T.G., 1965. The caloric content of migrating birds. Ecology 46, 901–904.
Richman, S., 1958. The transformation of energy by Daphnia pulex. Ecological Monographs 28 (3), 273–291.
Sinsabaugh, R.L., Manzoni, S., Moorhead, D.L., Richter, A., 2013. Carbon use efficiency of microbial communities: Stoichiometry, methodology and modeling. Ecology Letters 16 (7), 930–939.
Slobodkin, L., Smith, F., Hairston, N., 1967. Regulation in terrestrial ecosystems and the implied balance of nature. American Naturalist 101, 109–124.
Vucic-Pestic, O.L., Ehnes, R.B., Rall, B.C., Brose, U., 2011. Warming up the system: Higher predator feeding rates but lower energetic efficiencies. Global Change Biology 17 (3), 1301–1310.

Relevant Websites

https://www.youtube.com/watch?v=R3Cr37a_9AU
http://astro.hopkinsschools.org/course_documents/earth_moon/earth/earth_science/biosphere/ecology/ecosystem_eco.htm
http://www.bio.utexas.edu/faculty/sjasper/Bio213/ecosystem.html

Ecological Stoichiometry: Overview

RW Sterner, University of Minnesota, St. Paul, MN, USA
JJ Elser, Arizona State University, Tempe, AZ, USA

Introduction

Most ecological analyses have been constructed from single-currency descriptions, ones based on single dimensions such as biomass, carbon, nitrogen, or energy. For example, the carbon budget of a forest or a lake would include the inputs and outputs and relevant interchanges of C within the ecosystem. This means of analysis has enabled a great deal of progress. However, a multivariate approach that looks simultaneously at multiple dimensions can provide additional insight and predictive power. For example, both carbon and nitrogen can be studied at once and such a description would have to include knowledge of the C:N ratio of different ecosystem components. These multivariate approaches considering multiple currencies are useful because each individual currency interacts with others. Sometimes these interactions are simple, sometimes they are complex. Some pairs of measures are tightly linked in certain locations. For example, the Ca:P ratio in different fish species is almost exactly equal to 2.3, which is the same ratio as Ca:P in bone. Linkages such as these result from the fact that energy and multiple substances do not flow independently through ecosystems. In other cases, element ratios can exhibit a great deal of flexibility with and across species. Understanding these linkages is the goal of ecological stoichiometry.

`Ecological stoichiometry' is a relatively new term, but stoichiometric approaches were used in some of ecology's classic studies. Considerations of limiting factors, as in Liebig's law of the minimum, are inherently stoichiometric. Resource ratio approaches to competition explicitly include ratios of nutrient supply and nutrient content of competitors, making these models stoichiometric. The classical view of the oceans as having N and P in balance with biotic demand, as first articulated by A. C. Redfield in the middle of the twentieth century, is a stoichiometric view. Ecological stoichiometric principles are involved in studies of food quality, nutrition, nutrient recycling, and others, which have long histories of investigation.

To be more explicit, ecological stoichiometry is defined as the study of the balance of energy and multiple chemical elements in ecological interactions. Most will be familiar with the concept of stoichiometry from introductory chemistry classes where it is used to analyze chemical reactions to identify, for example, the reactant that might limit the formation of some specific chemical product in a chemical reaction. In any chemical reaction when reactants combine to form products, mass must balance during these atomic rearrangements. Ecological stoichiometry seeks to apply this same line of reasoning to understand some of the factors regulating ecological processes such as trophic interactions (herbivory, predation, detritivory), competition, energy flow, and biogeochemical cycling. In considering the stoichiometry of a single chemical reaction, knowing the elemental composition of the reactants and products is essential. Similarly, in ecological stoichiometry, one must know the elemental composition of the organisms involved, and their abiotic world.

Consider the familiar example of the enzyme-catalyzed and light-driven reaction of photosynthesis:

$$6CO_2 + 12H_2O \rightarrow C_6H_{12}O_6 + 6H_2O + 6O_2$$

This reaction involving carbon dioxide and water as reactants, and glucose, water, and oxygen as products in actuality is the net outcome of dozens of individual reactions. It summarizes the overall requirements for CO_2 and H_2O needed for the formation of a glucose molecule by this complex, multistep biochemical process. Our concern here is that the fixed chemical structure of glucose firmly establishes the chemical bounds of the system's behavior. If only one carbon dioxide molecule is added to the above, CO_2 will be in excess, and no more glucose molecules can be produced due to limitation by the other reactant, water. Considering the stoichiometry of this reaction calls attention to the powerful ways in which chemistry imposes constraints on biology. All chemical elements on the left side of the reaction must be accounted for on the right, thus explaining a very important outcome of photosynthesis: the production of O_2. Also, the specific chemical formula of the desired product determines what mixture of reactants can be optimally used.

Ecological stoichiometry takes the same approach. However, instead of considering one to perhaps dozens of reaction steps, it summarizes the chemical balance of ecological transactions that are the net outcome not of dozens but perhaps of tens of thousands of reactions that comprise an organism's entire metabolism (its `metabolome'). Despite this leap in reaction number, the law of mass balance for all constituent elements must still be obeyed. Functional organisms cannot be constructed with arbitrary proportions of chemical elements. Thus, stoichiometric constraints impose order on ecological interactions same as they do on individual chemical reactions. Most ecological interactions involve some form of transfer of matter. These fluxes strongly control productivity and community structure both in terms of absolute magnitude (the fertility or richness of the habitat) and in terms of relative abundance (the ratios of limiting resources). Elements such as nitrogen and phosphorus act like limiting reagents in the highly complex set of chemical reactions that take place during organism assimilation, growth, and decay.

Two classical papers in ecological stoichiometry in particular provided groundwork for subsequent studies. The oceanographer A. C. Redfield noted the quantitative similarity in terms of N:P ratios between the mean chemical composition of deep oceanic waters and the chemical composition of active plankton in surface waters. He hypothesized that the similarity was not accidental. This `Redfield ratio' of 16 N:1 P atoms was considered by Redfield to be evidence that the biota exerted a large-scale influence on

the chemical composition of the sea, a new perspective relative to prevailing views that did not assign the biota with a major causative role in global chemical cycling. This stoichiometric process, operating over extremely large spatial and temporal scales, imposes a biotic fingerprint on the chemistry of the abiotic world. The influence of the Redfield ratio on oceanic biogeochemistry is difficult to overestimate but more recent developments have come to place this finding in a broader context.

The terrestrial biogeochemist W. A. Reiners proposed that the study of the chemical signatures of living things (such as their C:N:P ratios) and of their ecological coupling in nature provides a `complementary' perspective on ecosystem dynamics, supplementing understanding derived from the then-dominant single currency bioenergetics perspective. He argued that, at the core of living things, what he called `protoplasmic life', chemical composition was relatively constrained. However, around their protoplasmic core, living things deployed drastically different materials for structural support. These structural materials may have dramatically different elemental composition, for example, C-rich cellulose in plants, Ca-rich shells in mollusks, Si-rich frustules in diatoms, or P-rich bones in vertebrates. The evolution of these major structural adaptations, and the subsequent proliferation of biota bearing them, in turn had major impacts on ecological dynamics and ultimately on large-scale biogeochemical cycling. Reiners's argument highlights the importance of protoplasmic versus structural allocations in determining organismal elemental composition. While indeed this is important, it is also true that even `protoplasmic' life can vary in elemental composition in important ways due to differences in biochemical allocations connected to growth status and life-history strategy (as described below).

Homeostasis and the Constraints of Mass Balance and Chemical Proportions

A key aspect to ecological stoichiometry is the degree of variation of elemental composition of an organism or species. Is elemental composition a fixed trait characteristic of that species, or is it a flexible parameter that largely reflects local resource and environmental conditions? This contrast is captured in a quantitative way in the concept of `stoichiometric homeostasis', the degree to which the elemental composition of an organism maintains a strict chemical composition (like a glucose molecule) despite variation in the chemical composition of resources. Stoichiometric homeostasis can be parametrized in the form of a homeostasis coefficient, H (eta), as:

$$H = m^{-1}$$

where m is the slope of a plot of the elemental composition of the consumer under consideration against the elemental composition of the resource being consumed. Both the x- and y-variables should be log-transformed in this measure. In this approach, if the consumer shows no variation in its elemental composition despite wide variation in its resource supply, then the slope of such a line (m) approaches zero and H approaches infinity (strict homeostasis). On the other hand, if the consumer's elemental composition exactly tracks that of its resource supply, then m takes the value of 1 and H has a value of 1 (no homeostasis, complete plasticity).

Data for real organisms range between these extremes and can depend on the experimental conditions imposed. Strict homeostasis does not appear to be reached in many situations, although strong homeostasis appears to be the rule for many metazoans. For example, in experiments in which the P-content of the algae on which it is raised varies from 1% to 0.05%, the P-content of the crustacean herbivore *Daphnia* varies little (only from 1.5% to 0.9%, corresponding to H of ~ 7, strong but not strict homeostasis). In contrast, the green alga *Scenedesmus*, when grown on inorganic nutrient supplies with N:P ratios ranging from 5 to 80, shows almost a 1:1 correspondence between its biomass N:P ratio and that of its medium. In this case of weak homeostasis, H approaches a value of 1. These examples illustrate a general pattern in which photoautotrophic organisms (cyanobacteria, algae, vascular plants) are generally thought to exhibit great plasticity in elemental composition. In contrast, heterotrophic organisms (including bacteria but especially metazoans) regulate their elemental composition more strictly around particular values. The contrast of these physiological strategies has profound consequences for food web dynamics and energy flow and nutrient cycling in food webs.

The Biology of Elements

Each of the 100+ elements in the periodic table has characteristic chemical and physical features governing properties like the elements it will bond with, how easily it will ionize, etc. These properties arise mostly from the atomic size of each element and the arrangement of electrons within shells surrounding the nucleus. Metals such as Fe or Mo for instance can exist in different oxidation–reduction states, and thus these elements are involved in electron transport in membranes and in other biological locations. Other elements such as C or N are very useful for making highly complex three-dimensional (3-D) shapes. Thirty or more elements are thought to be essential to the growth of at least some organisms. This number and range of reactivities give biological systems a great range of chemical behaviors to choose from in order to organize their activities. The physical and chemical properties of elements have a great bearing on ecological stoichiometry.

As discussed above, differences in organism elemental content between species are very relevant in ecological stoichiometry and ecology in general. Nevertheless, some elements are consistently high in abundance in biological systems while others are consistently rare. For example, the chemical composition for a living human can be written:

$$H_{375000000}O_{132000000}C_{85700000}N_{6430000}Ca_{1500000}P_{1020000}S_{206000}Na_{183000}K_{177000}Cl_{127000}$$
$$Mg_{40000}Si_{38600}Fe_{2680}Zn_{2110}Cu_{76}I_{14}Mn_{13}F_{13}Cr_{7}Se_{4}Mo_{3}Co_{1}$$

This formula combines all the countless different compounds in a human being into a single abstract `molecule'. Ecological stoichiometry asks how far this analogy to a single complex molecule will take us.

Ecological stoichiometric principles should apply to any element, and possibly to some of the less-reactive biochemicals as well. However, stoichiometric analysis has focused most on several elements that make up moderate to large proportions of living biomass and that also may become limiting to organism growth. The four elements considered below relate strongly to ecological dynamics and evolutionary fitness.

Carbon

Carbon (C) is the third or fourth most abundant element in the universe. Though it is not especially abundant on the whole of the Earth, it is a major component element of life. C is a highly mobile element in ecosystems, with common forms in gaseous, liquid, as well as solid phases. C has a valence state of 4 and can form four strong bonds with four other atoms; this allows it to produce virtually limitless arrangements of chains, rings, and other highly complex 3-D shapes. Its importance in biological systems is indicated by the fact that the soft tissue of living organisms is generally 40–55% C by dry mass. This large amount and generally tightly bounded range reflects the constant, ubiquitous, structural role of C throughout living systems.

Most biochemicals fall within a narrow range of C contents. Carbohydrates are similar to many other types of molecules in being almost 50% C. Of all common biological molecules, fats have the highest carbon content, with about 75% of their mass contributed by C. Organic structural matter that may make up large amounts of organism biomass includes the cellulose and lignin of plants and the chitin of arthropod exoskeletons. Even most inorganic structural matter used in living things contains considerable carbon, for example, the calcium carbonate of mollusk shells and the apatite of vertebrate bones both have large amounts of C. An exception is the silicon-containing shells of diatoms; these are nearly carbon free.

Nitrogen

Nitrogen (N) neighbors carbon in the periodic table and has many similarities in terms of biological chemistry. It too is highly abundant in the universe and is similarly abundant as carbon on Earth. It too is a major element of life and is found in all three phases in ecosystems. Its redox reactions include N-fixation, denitrification, and nitrification. With a valence state of 3, it is not quite as capable as C in forming complex 3-D shapes. The soft tissue of living organisms is generally 2–10% N by dry mass. This relatively large value is due to the fact that molecules needed in large quantities in the cell, such as proteins, nucleic acids, and pigments, contain N.

Chief among the important N-contained biomolecules are proteins, which are about 15% N by mass. Proteins themselves can serve a structural role but, as enzymes, they also drive nearly all the biochemical activities of a cell. Nucleic acids also are about 15% N by mass. In the form of DNA, nucleic acids make up the genetic code, though DNA is not a large fraction of cell mass. Even more important to ecological stoichiometry is the more abundant RNA (especially rRNA), which is critically important to cell growth and thus evolutionary fitness. All organisms require abundant proteins and nucleic acids to live and grow. The lowest N-content in living things is generally found in very metabolically inactive creatures.

Phosphorus

Phosphorus (P) has a larger atomic mass than C or N, and it is less abundant in the universe than those two elements. However, it is more abundant on the whole of the Earth than is either C or N. P lacks a significant gaseous phase, imparting very different biogeochemical behaviors to this element compared to C or N. Its abundance in the aqueous phase is strongly controlled by the presence of other elements, most notably oxygen (O) and iron (Fe). Most of the P at Earth's surface is found in the oxidized form of PO_4^{3-} (phosphate). Phosphates can polymerize, and the reaction

$$ATP + H_2O \rightarrow ADP + H_2PO_4^-$$

has a high free energy change but occurs slowly without a catalyst; these characteristics make P a highly suitable element to be involved in transfer of chemical energy around a cell. The soft tissue of living organisms is generally 0.3–2% P by dry mass.

Phosphorus is widespread in biochemistry, being found in relatively high abundance in membranes and being involved in many biochemical reactions throughout a cell. However, P is thought to have two focal roles in whole-organism stoichiometry. First, P alternates with sugars in forming the backbone of nucleic acids. As rRNA, nucleic acid serves a key role in allowing a cell to make proteins. Eukaryotic ribosomes are $\sim$50% rRNA while prokaryotic ribosomes are $\sim$65% rRNA. In this way, we can say that phosphorus is the elemental engine for protein manufacture. Second, P in the form of the mineral apatite is a major component of bone and thus has a major structural role in vertebrates. In small, unicellular heterotrophs such as bacteria and in rapidly growing larvae of insects, generally around 50% and sometimes as much as 90% of total cellular phosphorus may be contained in nucleic acids, especially rRNA. In larger heterotrophs, structural support tissues such as bones take on greater importance in body P-pools and in accounting for differences between species. These interspecific differences can be ancient in evolutionary origin and represent major differences in body plans comparable to taxonomic levels such as family or above.

Iron

Iron is a considerably larger atom than even P. It is as much abundant in the universe as C and N, but it is more abundant than C and N on Earth. It is the first or second most abundant element in the whole of the earth. A majority of the Fe in the Earth's crust is in the form Fe^{2+} (ferrous) but this is quickly oxidized at the oxygen-rich surface to Fe^{3+} (ferric). Though making up 30% of the mass of the whole Earth (a very large fraction), concentrations of Fe in oxic waters can be extremely low due to the low solubility of ferric iron. Low iron concentrations have been shown to limit primary production rates, biomass accumulation, and ecosystem structure in a variety of open-ocean environments, including the equatorial Pacific, the subarctic Pacific, and the Southern Ocean, and even in some coastal areas. Binding of Fe with other molecules is very important to governing its overall solubility. Organically bound Fe may be the dominant chemical form in aqueous solutions.

Iron is chemically versatile. It can coordinate with O, N, or sulfur (S), and it can bind to small molecules. Iron-containing proteins are key features of many energy-transducing biological reactions in processes such as photosynthesis, respiration, and others. Iron serves no major structural role in biological systems. Instead, it is a renewable reaction center. Thus, the soft tissue of living organisms is generally $<1\%$ Fe by dry mass.

Molecules containing Fe include hemoglobin, cytochrome *c*, and ferridoxin.

Contrasting Homeostasis in Plants and Animals

Autotrophs rely on either light or chemical energy to turn CO_2 into organic carbon molecules. Photoautotrophs are photosynthesizing organisms such as algae and higher plants that use light for this process. Heterotrophs, in contrast, obtain their chemical energy from preexisting organic molecules. Examples of heterotrophs include bacteria, which absorb organic substances from their surroundings, and many different animals, which consume and digest other organisms. These two major contrasting nutritional strategies of autotrophy and heterotrophy also contrast in their stoichiometric flexibility. Autotrophs obtain carbon, energy, and nutrients from different, somewhat independent sources, whereas many heterotrophs obtain all of these at once from the same food parcels. This contrasting flexibility in turn has a great bearing on the specifics of how stoichiometry enters into ecology.

Photosynthesis relies on light energy to fix CO_2 into organic molecules such as sugars. From these building blocks many other biochemicals can be made. Carbon:nutrient stoichiometry (C:N or C:P ratios) in individual autotroph species can be quite variable. Biochemicals such as carbohydrates and many lipids, which contain only C, H, and O, are made without incorporation of nutrients such as N or P. An autotroph in the light and with adequate access to CO_2 can make a plentiful supply of these compounds (starches, oils, organic acids, etc.) without investment of other critical resources. It is often observed that autotrophs growing in high-light, low-nutrient environments will possess a great abundance of these molecules, so much so in fact that the C-content of the autotroph will be elevated under those types of conditions. Carbon:nutrient ratios within such plants can be exceedingly high (>1500 C:P, for example). When a slow-growing, nutrient-limited autotroph suddenly is exposed to high nutrient availability, it will take up those nutrients much faster than its growth rate. That is, nutrients are taken up in excess compared to growth requirements and in some extreme cases stored in specialized structures such as vacuoles or in specialized molecules such as polyphosphate. High carbon:nutrient ratios are also characteristic of large autotrophs such as trees, which require substantial investment in wood and ancillary tissues having high C:nutrient ratio. Ecological implications of these stoichiometric responses to light:nutrient ratios are discussed below.

Autotroph nutrient content is related to growth rate (μ, g g^{-1} d^{-1}). A quota (Q) is the mass or molar quantity of nutrients per cell (this discussion assumes a constant cell size). In unicellular autotrophs, the `cell quota' concept relates these two variables. The quota of the element that regulates growth rate will be very tightly related to growth rate by a relationship referred to as the Droop formula:

$$\mu = \mu'(1 - k/Q)$$

where μ' is a theoretical maximum growth, never attained, associated with infinite quota, and k is the minimum quota occurring at zero growth.

Under strongly nutrient limiting conditions where growth rate is low, quota of the limiting nutrient will be low, meaning a low nutrient:C or high C:nutrient ratio (see cellular C:P, **Fig. 1**, top panel). The minimum cell quota (k) is set by the level of nutrient-containing biochemicals necessary for basic metabolism, and nutrient requirements for growth are added to this basal level. A true upper level for nutrient content (less than μ') will be set by some combination of the composition of protoplasm at high growth rate or the ability of an autotroph to store excess quantities of any nutrient not currently needed for growth. In autotrophs, growth involves at least two specific major stoichiometric components, and probably more. The first is N for proteins involved in photosynthesis, especially the enzyme RUBISCO, which can be a major portion of cellular biomass. Metabolism in vascular plants relates more strongly and consistently to N than biomass or C. The second is P for ribosomes, which are needed to manufacture additional proteins.

In addition to these patterns relating content of the limiting nutrient to growth rate, the ratio of nutrient elements in an autotroph varies positively with the ratio of those nutrients in the environment. Soils or water of high N:P ratio will generally support plants or algae with high N:P ratio. This positive relationship derives in part from shifts in species across gradients such as these, with competition favoring species that have similar nutrient ratios as the supply ratio in the environment. It also derives

from intraspecific, physiological shifts associated with differing storage and utilization of the two nutrients similar to those described for quota above. Fig. 1 summarizes these different influences on autotroph nutrient content.

Samplings of whole assemblages of autotroph biomass have been examined in terrestrial, freshwater, and marine ecosystems, and have included microscopic as well as macroscopic species. Terrestrial ecosystems, with their larger, cellulose-rich, and woody plant species have higher and more variable C:P and C:N ratios than aquatic ecosystems. In the aquatic realm, offshore marine environments characteristically have low and less variable C:P and C:N ratios in their suspended matter, which contains a strong signal of autotroph biomass. We saw this relative constancy in the offshore marine realm when we discussed the Redfield ratio above. Redfield described the marine plankton to have a C:N:P ratio of 106:16:1. Today there is continuing interest in the Redfield ratio in the ocean, and it is known that it is not a true constant but rather varies with several factors, including climate. Freshwater ecosystems can be thought of as being intermediate in their stoichiometric patterns of C:N:P between terrestrial ecosystems and offshore marine ecosystems.

Animals and other heterotroph species also vary in their chemical content. Large shifts in C:N or C:P ratios in heterotrophs can follow from storage of large amounts of chemical energy in the form of lipids. Some invertebrates in seasonal environments, for instance, may assimilate and store lipids to the point where they are approximately half of organism mass. When those lipids are subsequently catabolized, dramatic shifts in C:N or C:P result. However, in contrast to the great stoichiometric flexibility often observed in autotrophs, unicellular and multicellular heterotrophs come closer to approaching an idealized, strictly homeostatic, abstract `molecule' of defined chemical composition. Reasons for this contrast between plants and animals are not well understood but might involve lack of specialized storage vacuoles in animal cells and the fact that animals obtain carbon, energy, and nutrients from living or recently living material, which is less chemically variable than the abiotic sources of carbon, energy, and nutrients used by plants.

Metazoan animal species exhibit a wide range of N:P ratios. Small, poorly skeletonized organisms such as tadpole stages of amphibians have N:P of ~ 20 whereas some fish species that are heavily endowed with calcium phosphate apatite mineral both in their internal skeleton and in their scales have N:P of ~ 5. Fish in fact are a highly stoichiometrically variable group. From that

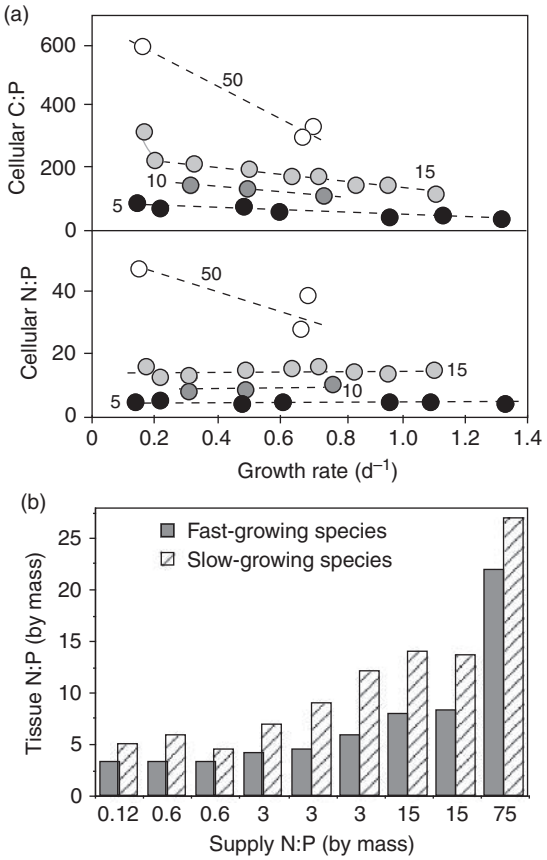


Fig. 1 Autotroph nutrient content as a function of both growth rate and nutrients in the external environment. (a) Experiments with the unicellular alga *Dunaliella tertiolecta*. Symbols refer to different N:P in the growth medium (5–50). (b) Experiments with two species of grasses, one (*Dactylis glomerata*) fast-growing and the other (*Brachypodium pinnatum*) slow-growing. In the upper part of (a), note that cellular C:P declines with increasing growth rate, and is highest at low growth rate and where the environmental N:P is greatest. Similarly, both panels of (a) show that environmental N:P has a positive effect on algal N:P at all growth rates. In panel (b), note again that environmental N:P has a positive influence on tissue N:P. Panel (b) also shows that for any given environmental N:P, the fast-growing species has lower N:P than the slow-growing species.

minimum N:P of about 5, different species of lower structural P content range upward to N:P of 15. Within fish, the Ca:P ratios are highly constrained, indicating that most of the stoichiometric differences in this group result from evolutionary pressures on structure and hardness of the integument.

These inter- and intraspecific patterns of elemental content combine in food webs of many species. Stoichiometric imbalance, where resource and consumer differ radically in their nutrient content, generates interesting ecological dynamics that we will consider next.

Stoichiometry of Limiting Elements

The earliest application of stoichiometric reasoning in ecology probably was that in the work by Justus von Liebig (1803–73), who was concerned with factors influencing crop growth yield. His studies led to what we now call `Liebig's law of the minimum'. One way to phrase Liebig's law is that growth is controlled not by the total of all resources available, but by the one scarcest resource. This is a direct analogy to the concept of a limiting reagent in a chemical reaction producing a stoichiometrically fixed product. `Scarcest' is a relative term and refers to the availability of a resource in the environment relative to biological demand. For example, chemical resources needed in large quantities such as C or N might be `scarce' in this context even if they are at higher concentration than elements such as Fe or Mo needed in much smaller quantities. Liebig's law does not apply to all possible resources. Ones that fit it most precisely are referred to as `nonsubstitutable' resources, ones which organisms cannot trade one for another. Individual elements are largely nonsubstituable with each other whereas many (though not all) organic molecules have substitutable alternatives.

If one knows the stoichiometry of resources and consumers, one can make predictions about which resource will run out and therefore become the `limiting reagent', regulating growth and production. We see this reasoning in such recently studied phenomena as the Southern Ocean, high-nutrient/low-chlorophyll (HNLC) region. This portion of the ocean is notable because of not only a simultaneous occurrence of low algal biomass but also high levels of the nutrient elements N and P. In most marine and freshwaters, one or both of N and P become exhausted and become limiting. However, in the HNLC regions, it is usually Fe that becomes exhausted first. Iron limitation in HNLC regions has been confirmed in small-scale bottle experiments, and in some spectacular, large-scale open-ocean fertilization experiments where Fe was introduced from a ship and the subsequent development and decay of a high algal patch was followed. Fe-limitation has also been described in near-shore oceans and even in lakes.

Liebig's law is most often thought of in terms of plant population dynamics, where individual elements form the set of possible resources. There are, however, also circumstances where Liebig's law has been applied to animal growth. Liebigian dynamics are strongly suggested, for instance, in the section below on stoichiometry and animal growth.

The stoichiometry of limiting factors varies considerably in space and time. Over long timescales associated with soil development during primary succession on a mineral soil, the N:P balance shifts because P is present in the original parent material, but N derives largely from biological processes including N-fixation. Thus, a new mineral soil will contain P but almost no N, and N will be limiting to plants growing on that site. N-fixers are favored. As N in the soil builds up over successive production/decay cycles, N comes into approximate balance with P. Over long timescales, highly weathered soils can have most of their P removed or bound into inaccessible forms, and P can become limiting. Soil development fitting this general pattern has been observed on the Hawaiian islands. Ancient tropical soils are often highly deficient in P, and there is a general pattern of an increased N:P in the foliage of coniferous trees, grasses, herbs, shrubs, and trees as one goes from the poles to the tropics.

When one seeks to apply Liebig's law in any individual case, complications often arise. For example, different species of an assemblage may be limited by different nutrients simultaneously, meaning multiple resources might limit a community. This form of multiple resource limitation is very well studied in theoretical models and has been the foundation of some recent elegant experimental studies as well; theory and experiment together suggest that the number of elements that are simultaneously limiting to different species has a large influence on the biodiversity of a community. Where multiple resources become limiting, biodiversity is higher. In this form of multiple-element limitation, one might potentially see the community as a whole respond to single additions of more than one resource. The largest response is expected where multiple resources are added, because this will stimulate the greatest number of species. In other cases, due to biochemical reasons, availability of one element might aid in the acquisition of another element. For example, the enzyme that allows cells to make use of organically bound P contains Zn; thus, potentially adding or subtracting Zn might be functionally similar to adding or subtracting P. Ecologists today do not have a comprehensive, widely accepted terminology for dealing with cases of multiple-nutrient limitation.

Liebig's law appears today in work referred to as resource competition theory (RCT). RCT provides a predictive, mechanistic framework for studying how community structure relates to resource availability. Good competitors for a resource are defined as those that require relatively small amounts of a resource in order to maintain themselves in a community in the face of mortality losses. Stochiometrically, good competitors often are those that themselves have low nutrient contents or high C:P or C:N ratios. However, both differential resistance to mortality and differing ability to acquire nutrients can also play strong roles in determining competitive ability.

Various tests of this theory have been performed and support for RCT predictions have come from the laboratory and field in experiments involving microorganisms (bacteria, cyanobacteria, algae) and vascular plants and even metazoan animals. For example, resource conditions with high Si:P ratios tend to favor diatoms (which have relatively low P-requirements but require silica), while other algal taxa (which have no Si-requirement) dominate over diatoms when Si:P ratios are low. Coexistence is predicted, and observed, when Si:P ratios are intermediate; in such a situation, the diatoms are limited by Si while its competitor would be limited by P. Similarly, the relative abundance of cyanobacteria (and especially N-fixing cyanobacteria) is often higher

when environmental nutrient supplies occur at low N:P ratios. Recent studies have also shown that the nonlinear resource utilization functions of species allow for multispecies coexistence due to the complex, and perhaps chaotic, dynamics that ensue when multiple resources (light, nutrient elements) can be limiting to multiple species (`coexistence by chaos').

Nutrient limitation by N or P interacts with the carbon cycle in interesting ways. Ecosystems that are strongly nutrient limited are often observed to have primary producers with elevated C:N or C:P ratios. Those nutrient-limited ecosystems can be said to have high nutrient-use efficiency. They make much biomass with each unit of nutrient acquired. As discussed above, in some systems it is the overall availability of light relative to nutrients that controls C:N or C:P ratios. No matter which terminology or thought process is used to study these relationships, the functional outcome is that growth rate and stoichiometry are related in nonhomeostatic organisms such as autotrophs.

Life Histories

Chemical composition is part of an organism's phenotype. It is therefore molded by natural selection and other evolutionary forces like any other aspect of an organism. Thus, the role of the organism in nutrient fluxes in ecosystems is shaped in evolutionary time. As we saw in our discussion of resource competition theory, ecological success or failure, and therefore evolutionary fitness, is tied directly to chemical content. Organism stoichiometry affects fitness. There are also more complex connections between stoichiometry and evolutionary pressures. Because of the distinct stoichiometry of structural matter, evolutionary pressures on body size or major body plans involving structure will have stoichiometric implications. In fact, anywhere organism function has some kind of distinct stoichiometric signature, evolution will be steered by and will in turn alter stoichiometry. Ecological stoichiometry can help us better understand some of the many selective forces on organism function. A good example is in life histories. A life history for a given species is its schedule of vital rates including growth, reproduction, and mortality. Life-history theory typically considers these schedules to be phenotypic traits exposed to selection and shaped by ecological forces. However, these vital schedules, though measurable, do not exist independent of the physical, inorganic world. Because certain vital rates have distinct chemical signature, ecological stoichiometry can illuminate key aspects of life-history evolution.

Growth Rate

The growth rate hypothesis (GRH) links elemental content to biochemistry and in turn links both of those to life history. It concerns the specific growth rate of body mass (μ, g g^{-1} d^{-1}). The GRH comes about because of the almost uniquely low N:P ratio of nucleic acids compared to all other major biochemicals. According to the central dogma of molecular biology, information flows from DNA to proteins by way of several forms of RNA. Protein is a large fraction of cell mass and to achieve high specific growth rates of cell mass, a great deal of ribosomal RNA is needed to manufacture large amounts of protein at high rates. In rapidly growing *Escherichia coli*, RNA makes up approximately 35% of total cell mass, and as much as 90% cellular P is in RNA. Such a large proportion of cell mass in a single biochemical of unusually low N:P generates a distinct stoichiometric signature for the whole organism. Due to this stoichiometric imprint of RNA, the GRH posits positive relationships among three variables: specific biomass growth rate, P-content, and RNA content. Support for the GRH comes from multiple groups of organisms ranging from unicells to small invertebrates such as *Drosophila* and *Daphnia* (Fig. 2).

Applicability of the GRH to larger organisms is still unclear. When organisms are examined over a range of many orders of magnitude in body size, RNA content is observed to decline greatly with body size, as does growth rate. Whole-organism P content, however, does not decline as strongly with body size in such a range. Hence, in considering a range of organisms from microbes to large multicellular animals, there must be a shift in P-containing pools along body-size gradients from RNA to other substances. As we will see in more detail below, in vertebrates bone replaces RNA as a major P-reservoir.

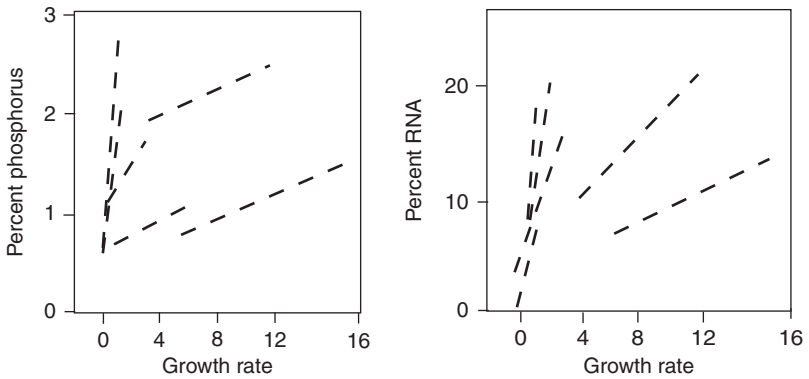


Fig. 2 Positive relationships among specific biomass growth rate (μ, d^{-1}) and both P-content and RNA content. Each dotted line represents a different species. Though species differ from one another, these variables are often linked in a way consistent with the GRH.

There is evidence too that the GRH applies to autotrophs as well as heterotrophs. In a study with the unicellular alga *Selenastrum*, high growth rate was associated with a shift in the boundaries between N- and P-limitation; the N:P of the boundary was lower at high growth rate. This is consistent with a need for larger quantities of low N:P RNA at high growth rate. Similarly, a model of unicellular algal growth was analyzed where cellular investment in two broad classes of biochemical machinery was examined: (1) assembly machinery, for example, ribosomes with low N:P ratios, and (2) resource-acquisition machinery, which is made up mainly of protein containing N but little or no P. Shifts in optimal N:P ratios in the model were associated with growth rate. Low N:P was favored at high growth, whereas high N:P was favored under low growth and strong resource scarcity. These studies of microscopic autotrophs are consistent with the GRH. In one study looking at foliar N and P in more than 100 vascular plants, N:P was lower in species that have higher growth rate. **Fig. 1b** showed one such contrast.

The GRH thus proposes a material basis to a key life-history parameter: growth rate. All else being equal, specific biomass growth rate will be closely coupled to such life-history parameters as age and size at first reproduction, and therefore is directly tied to fitness itself. Because it is especially P among all the elements that is disproportionately required for high growth, we can further hypothesize that high-growth phenotypes might be particularly sensitive to the presence or absence of phosphorus from the environment. A fascinating biomedical example of the GRH in action is in cancerous tumors, which in many cases are high in RNA. P-content of tumors has not often been studied, but evidence for high P-content in tumors has been seen.

Growth–Competition Tradeoffs

The GRH provides an explanation for why an organism might be selected to have high P content. Why then don't all organisms have high P content? As elsewhere in evolutionary biology, the answer lies in understanding key tradeoffs. What disadvantage might there be to having a high-growth, high-P lifestyle? Ecologists have long recognized a general syndrome of life-history characteristics, those that promote rapid-growth, colonist-type, life histories from those that promote success in the face of extreme competition due to presence of many competitors. The former are referred to as *r*-selected species and the latter are referred to as *K*-selected species. Stoichiometry has its own version of $r - K$ selection theory.

We have already seen that high-growth (*r*-selected) life histories require high P. However, it is equally the case that species of low nutrient content will outcompete species of high nutrient content during exploitative competition for resources. Stoichiometry predicts that this tradeoff will be particularly evidenced under conditions of P-limitation, because of the close tie between P and high growth rate. High-growth-rate species are anticipated to be poor competitors. According to stoichiometric theory, this tradeoff results from the material basis of a high-growth-rate lifestyle.

Nutrient Cycling, Flux, and Dynamics

The rate of cycling of potentially limiting elements is a critical feature of ecosystems, determining important aspects of their structure and function. This cycling includes movement of elements between abiotic and biotic pools as well as among living species. Autotrophs take up resources from their environment, and they may also leak some nutrients back into the surrounding soil or water media; they also produce litter which is broken down, and decays and releases nutrients into abiotic pools. Heterotrophic consumers participate in nutrient cycling when they ingest food and release wastes back to the abiotic pool. If we understand the controls on the rate that nutrients are made available or reavailable for living organisms to take up, we will have a powerful tool to understand and predict many features of ecosystems. Stoichiometric approaches to nutrient cycling are based on the conservation of matter and consider patterns among multiple elements.

Under strict homeostasis, we consider individual species to have fixed stoichiometric coefficients in their chemical makeup. Thus, the chemical flexibility needed to balance a reaction where resources are reactants and organism biomass is a product comes from the composition of waste products, another product in the reaction:

$$\text{Resources} \rightarrow \text{Organism biomass} + \text{Wastes}$$

If we know both the chemical content of organisms and the chemical content of their resources, we can use the conservation of matter to calculate the chemical composition of wastes. Note that variation in the chemical content of either resources or organism biomass can influence the chemical content of wastes.

One aspect of stoichiometry is merely the use of a set of tools for balancing chemical reactions involved in nutrient cycling and maintaining appropriate relationships among all the elements. However, stoichiometric homeostasis also has a characteristic and somewhat peculiar effect on nutrient cycling. A homeostatic consumer alters the fraction of ingested nutrients that are retained for growth in response to the chemical content of the food. A homeostatic consumer must generally retain most stringently the scarcest element. For example, when ingesting food of relatively low P-content, a homeostatic consumer must retain P with elevated efficiency. In other words, it must retain that substance which already is in scarce supply in the ecosystem while it resupplies the nonlimiting elements back to abiotic pools (**Fig. 3**). Nutrients that are found in relative excess tend to be recycled back to the environment while nutrients that are scarce in the food relative to a consumer's requirements are retained for growth. Thus, when put into the context of whole ecosystems, interesting dynamics result from homeostatic consumers.

These processes have been measured in many different situations. **Fig. 4** shows the effect of chemical variation of both resources and consumers in one set of aquatic organisms. Because natural ecosystems are composed of numerous species all with their own

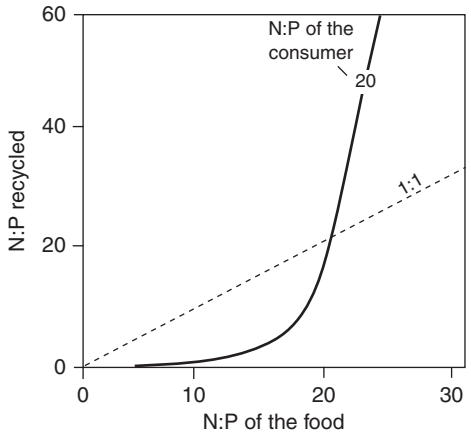


Fig. 3 The ratio of nutrient elements recycling by a feeding homeostatic consumer as a function of the ratios of elements in its food. At low N:P in the food, the homeostatic consumer retains N with high efficiency, recycling excess P, and hence the N:P recycled is very low. In contrast, at high N:P in the food, the homeostatic consumer retains P with high efficiency, releases relatively more N, and thus the N:P recycled is very high. Homeostasis in the consumer causes this relationship to curve and the greater the ability of the consumer to retain the element limiting its own growth, the tighter the bend in the function.

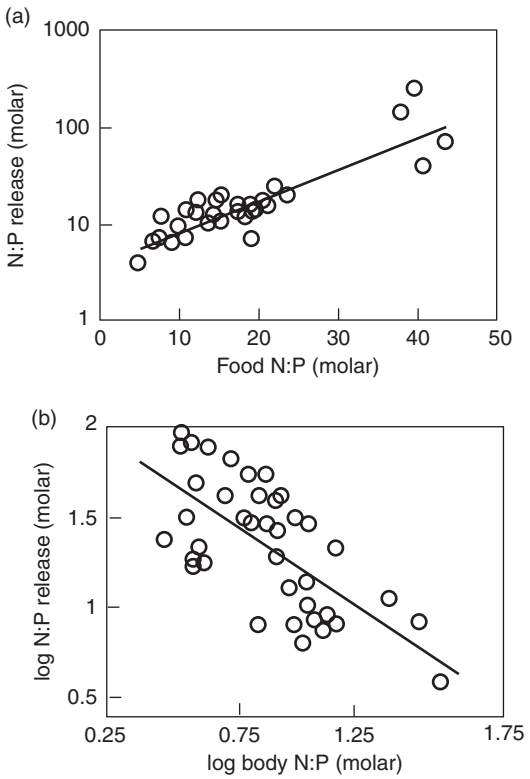


Fig. 4 Two aspects of stoichiometric determination of nutrient cycling, the nutrient content of the resources (a) and the nutrient content of the consumers (b). Both examples consider homeostatic animal consumers feeding upon chemically variable resources. In (a), the N:P ratio of nutrients released by freshwater zooplankton is positively related to the N:P ratio of the food they eat. This example shows that when consuming foods relatively high in either N or P, homeostatic consumers dispose of the excess nutrient and retain the nutrient most limiting to their own needs. Note the log scale of the *y*-axis, meaning a curvilinear function on linear axes. In (b), the stoichiometric variability of different aquatic vertebrate consumers generates a negative relationship between the N:P ratios of nutrients released and the N:P ratio of the consumers themselves. This example shows that consumers with bodies containing either relatively high N or P must retain that element high in their biomass and further that they will dispose of the element relatively scarce in their own biomass.

characteristic stoichiometry, natural food webs consist of many different resource–consumer pairs. Some of these will be stoichiometrically similar while others may be stoichiometrically dissimilar.

We also see a strong stoichiometric control on nutrient recycling in terrestrial systems. Leaves exhibit a wide range in C:N:P ratios, which is largely a function of the species or functional group of the plant, but as we have already seen it also depends on growth conditions and other factors. When leaves die, there is a corresponding wide range in nutrient content of the detritus they form. Due to nutrient resorption prior to leaf abscission, detritus often has a very low N- and P-content, a factor which emphasizes the stoichiometric dissimilarity between resources (detritus) and living consumers (microbial decomposers and detritivorous animals). Ecological stoichiometry is most easily revealed in systems where there is great chemical variability, like in this consideration of detritus and detritivore. Fig. 5 shows how strongly the nutrient content of detritus corresponds to the rate that detritus breaks down, at least at this large, cross-ecosystem scale.

This range in mineralization rate results from the stoichiometric match between the chemical composition of the litter and the needs of the organisms feeding on that litter (microorganisms, detritivores), so it combines stoichiometric food-quality effects, such as will be described below, with stoichiometric nutrient cycling rates. The highly biodiverse soil food web can grow and metabolize more rapidly when supported by high-nutrient litter instead of low-nutrient litter.

Another example of stoichiometric recycling effects involving plants, soils, and decomposers involves Ca-content. In a long-term (30-year) study where 14 different tree species were raised in monoculture, it was found that soil properties came to strongly depend on the Ca-content of the leaves of the tree species growing on that plot. Soil under tree species with high-Ca leaves, roots, and litter (e.g., maples, basswoods) was higher in pH, higher in C, lower in C:N, and higher in exchangeable Ca than was soil under tree species low in Ca (e.g., oaks, pines). Earthworm biomass also was higher under high-Ca tree species. The chemical parameter Ca was better related statistically to these and other similar relationships than was identity of the tree species themselves or the identity of functional groups like angiosperms versus gymnosperms, suggesting that it was indeed the stoichiometry of Ca in the litter that was the key factor involved.

Alteration of Community Composition and Ecosystem Dynamics through the Stoichiometry of Recycling

When we place these specific stoichiometric recycling effects into the context of complete, complex ecosystems, a variety of interesting dynamics result. For example, as we just saw, terrestrial plant species that produce low-nutrient litter will have a depressive effect on mineralization rates in their vicinity. A low mineralization rate will reduce primary productivity, lower plant biomass, and raise light:nutrient ratios. These conditions in turn will favor particular plant species that are good nutrient competitors. Because one factor that can improve a species' competitive ability for a given nutrient is a low content of that nutrient in its leaves, a cycle of positive feedback is favored: with good nutrient competitors altering nutrient cycling in their vicinity in such a way as to improve their chances of success against other species.

Shifts among consumer species differing in stoichiometry can also have ecosystem-level consequences. In freshwater plankton food chains, large-body-size, high-growth-rate *Daphnia* are favored under particular regimes of fish predation, for example, when there are many piscivores that deplete the planktivorous fish populations, releasing *Daphnia* from fish predation. Because *Daphnia* have a characteristic high-P stoichiometric signature, changes in fish predation regime can realign pelagic nutrient pools and fluxes. It has been observed that phytoplankton in the presence of abundant *Daphnia* grazers will be relatively more P- than N-limited and N-fixation will be reduced. These examples illustrate the close connections between community structure and ecosystem nutrient fluxes that are easily understood by consideration of stoichiometric principles.

Animal Growth: Stoichiometry and Food Quality

Ecological theory often incorporates the effects of food supply on consumer growth and reproduction by a hyperbolic functional response between animal growth and reproductive rates and the density of food. This is referred to as a `numerical response'. In

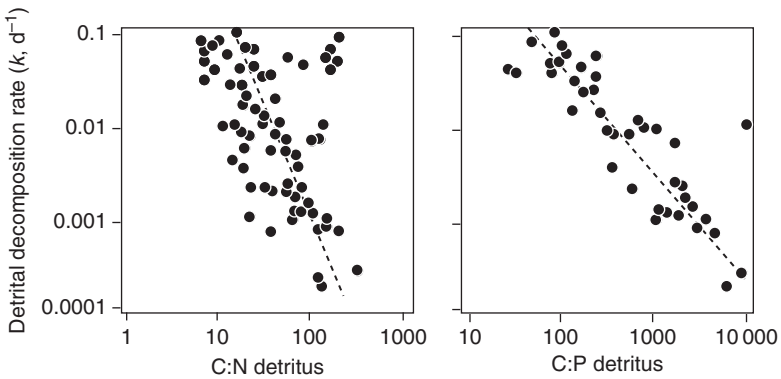


Fig. 5 The rate of breakdown of terrestrial detritus depends strongly on stoichiometry. The higher the C:N or C:P ratio, the slower the leaf litter breaks down (lower rate coefficient, *k*). As a consequence, the rate that nutrients are recycling back into the soil also depends on stoichiometry, with higher mineralization rates for detritus with low C:N or C:P.

addition, in the growth of an individual, a minimal quantity of food, alternatively known as the `threshold food concentration' or the `subsistence food level', is defined as that needed to offset respiratory losses. At the population level, a higher threshold is needed for growth to be sufficient to offset various sources of mortality. These effects of food quantity have been studied for a long time; however, incorporating the effects of the quality of food on consumer performance and dynamics has been much more difficult yet is done where ecological stoichiometry becomes very relevant.

While there are many known aspects of food quality (digestibility, palatability, toxicity, content of nutritive biochemicals such as essential amino acids and fatty acids), ecological stoichiometry measures food quality as the content of potentially important nutrient elements, especially N and P. This measure is in keeping with the extensive literature on the importance of N in affecting consumers in terrestrial and marine environments. It is also consistent with the increasing knowledge of the role of dietary P in freshwaters and perhaps in terrestrial ecosystems as well. In general, the quality of a food item is inversely related to its C:nutrient ratio, where the nutrient can be N or P. This effect can be understood to be an outcome of the fact that the nutrient element is increasingly diluted in the dominant C-biomass, making it difficult for the animal to extract sufficient limiting element from the food.

Stoichiometric theory gauges the onset of such limiting effects via calculation of the threshold elemental ratio (TER). The TER is defined as the food C:nutrient ratio above which the animal's growth becomes limited by nutrient X rather than C (carbon), where X might be N, P, or another element. In the simplest models, the TER is based on only three traits of the consumer: its own C:X ratio and its maximal gross growth efficiencies for C (E_C) and the nutrient (E_X). Gross growth efficiency is defined as the rate at which mass accumulates in production (or new growth) divided by the rate at which mass is ingested by the consumer. The maximum gross growth efficiency for an element is assumed to occur when that element is limiting growth. Assuming strict stoichiometric homeostasis of the consumer, the TER is calculated as

$$\mathrm{TER}_{\mathrm{C:X}} = (E_C/E_X)\mathrm{C} : \mathrm{X}_{\mathrm{consumer}}$$

Noting again that the TER is the C:X ratio in the food above which the consumer's growth should become limited by element X, a consumer with low TER is more likely to be limited by element X than a consumer with a high TER. Thus, an animal becomes less likely to be limited by nutrient X (its TER increases) as the consumer's C:X ratio increases (its nutrient content declines), as its ability to sequester element X increases (E_X increases), or as its carbon-assimilation efficiency decreases (E_C decreases).

The dependence of the TER on net C-growth efficiency implicitly means that the TER increases as food concentration decreases. This is because under such strong food-limited conditions, the respiratory demands of maintenance metabolism dominate the energy or C-balance, and C-growth efficiency is reduced. Thus, TER theory predicts that stoichiometric food quality should be of greatest importance under conditions of high food abundance. **Fig. 6** illustrates this in a general case, in which the line labeled `A' is the minimum TER for the consumer in question, asymptotically achieved at maximum $E_C$ at infinitely high food quantity. The line labeled `B' is the minimum amount of food that the consumer needs in order to maintain a constant body mass.

The effects of stoichiometric food quality have been extensively tested in the laboratory and the field, especially for freshwater zooplankton under conditions of potential P-limitation. For example, *Daphnia* growth commonly declines when food C:P ratio is above ~250. P-assimilation efficiency increases to high values as the TER is approached. Lab experiments involving short-term manipulation of food P-content or direct P-supply to the *Daphnia* have confirmed that this growth decline is at least partially due to a direct P-limitation of the animal. Furthermore, field studies have shown that *Daphnia* abundance is low under lake conditions in which seston C:P ratios are above 250. Further, short-term amendment of seston P-content increases *Daphnia* growth when seston C:P is above the TER, showing that the predictions of stoichiometric theory are confirmed not only in the lab but also under natural conditions. Also consistent with TER theory, animals with low body P-content, such as the crustacean *Bosmina*, appear to be relatively insensitive to food P-content, both in the lab and in lakes. Recent studies outside of the plankton have also provided

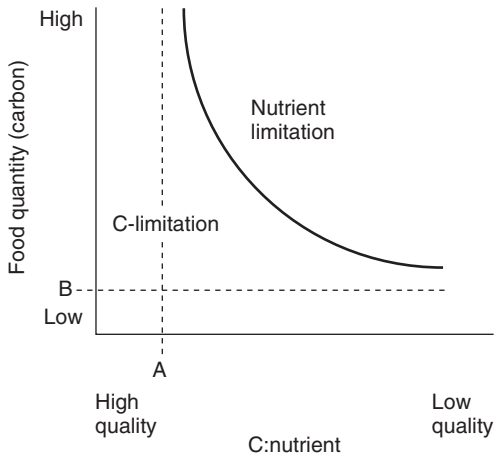


Fig. 6 Boundaries for limitation by food quantity (energy) vs. food quality (stoichiometry) for a homeostatic consumer as a function of food quantity. At high food abundance and high food C:nutrient ratios (low food quality), the consumer should be limited by the nutrients in its food. At low food abundance or low food C:nutrient, the consumer should be limited by total food quantity, or total energy content.

evidence of dietary P-limitation in benthic (stream insects, snails) and terrestrial animals (caterpillars, weevils), adding to previous evidence of dietary N-limitation from diverse habitats.

In sum, these studies show that stoichiometry is an important axis of food quality affecting basal consumers in diverse food webs. Relatively simple physiological processes and species traits can be used to predict their operation and impact.

Complex System Dynamics Driven by Stoichiometry

Theoretical ecology uses mathematical tools to place the nature of interactions into a formal, analytical framework. A goal is to understand features governing population dynamics and species diversity. Special emphasis has been placed on ecological competition and predator–prey dynamics. These studies often focus on the presence or absence of equilibrium points in model solutions, which indicate the possibility of species coexisting with each other over the long term. Moreover, a fundamental difference between stable equilibrium points and unstable equilibrium points is recognized. In the former, dynamics tend to restore the system to its equilibrium point after a small perturbation, whereas in the latter, like a ball rolling off the peak of a hill, small displacements from equilibrium can result in large changes and even be catastrophic to coexistence.

Herbivore Extinction Due to Poor Food Quality

Stoichiometrically implicit predator–prey and especially plant–herbivore interactions have been built and analyzed, providing some clues as to the role of stoichiometry in system dynamics and species diversity. In these models, consumers respond to food quality determined by element ratios, and nutrients are recycled according to stoichiometry and strict homeostasis. Recycling provides a feedback between a herbivore and its algal or plant prey. The herbivore's growth and nutrient release rates decline as the prey's C:nutrient ratio increases above its TER. These analyses involve development of versions of the classic Lotka–Volterra predator–prey models in which the homeostatic regulation of nutrient content varies between predator and prey. When the prey is modeled as a photoautotroph, its nutrient content will be variable and the predatory herbivore will have fixed nutrient.

The effects of introducing stoichiometric constraints to such models of trophic interaction are dramatic (Fig. 7). In classical Lotka–Volterra-type models, the predator can always grow whenever prey abundance is above some finite level. This is seen as a vertical, linear nullcline for the predator; the predator's nullcline is the set of points where the predator's net growth rate is zero. There is thus only one possible equilibrium point of plant and herbivore density for the system. However, in a stoichiometric version of these equations, both prey and predator nullclines are hump shaped. The prey's nullcline intersects the *x*- (prey-) axis due to nutrient limitation. The intersection point is determined by the maximal C:nutrient ratio the prey achieves under nutrient limitation. Stoichiometric models are unusual in that the predator's nullcline also is hump shaped, in this case due to the negative effects of poor food quality when the plant achieves high biomass on a fixed and limited amount of nutrient in the system. At high plant biomass, plants have high C:nutrient ratio, and further additions of biomass (C) can have a depressive effect on herbivore dynamics, a paradoxical result. The herbivore population is inhibited by the addition of bulk food. In fact, the herbivore can shift between positive growth and negative growth by the addition of plant biomass to the system! This counterintuitive aspect of the model has been termed the `paradox of energy enrichment'.

Stoichiometry alters nullclines so that both are nonlinear, meaning they can intersect each other at more than one point. Depending on the parameters used, there is potential for multiple equilibria, limit cycles, and potentially chaotic dynamics when one includes

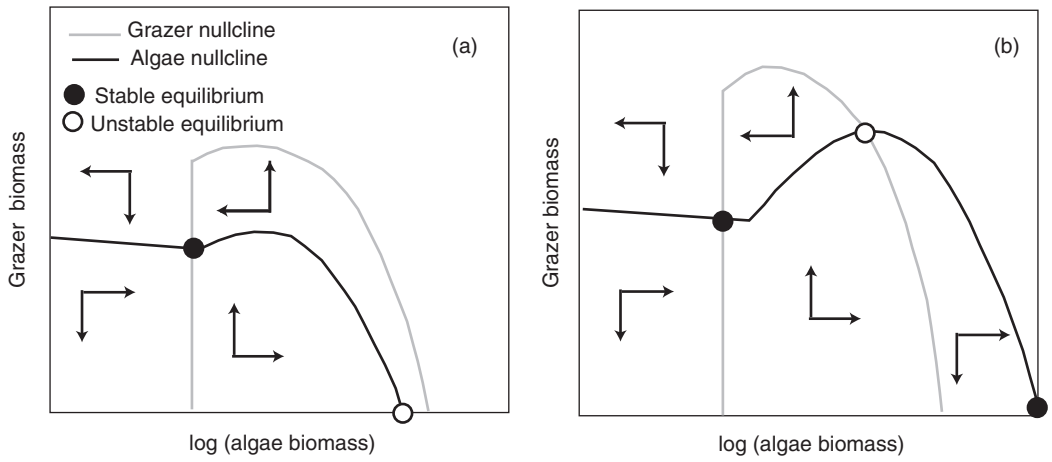


Fig. 7 Predator (grazer) and prey (algae) nullclines in a theoretical model. Panels (a) and (b) use different parameters for the same model. Stoichiometry makes the grazer nullcline hump shaped, which alters predicted dynamics and makes it possible even for a stable equilibrium point at zero grazer biomass.

stoichiometry in such a system. A particularly interesting equilibrium is illustrated in Fig. 7b, in which conditions are such that the prey nullcline intersects the prey axis at a high value, beyond the upper intersection of the predator's nullcline. In this situation, a stable equilibrium exists at the prey's intersection on the *x*-axis, a point where the grazer is extinct and cannot invade from low population levels. This is a grazer extinction point occurring in a `world' of very high food (plant) biomass! In ecophysiological terms, this situation is more likely to occur for grazers with high body nutrient content and for environmental conditions in which autotrophs develop very high biomass C:nutrient ratios. As described earlier, the latter occurs when nutrients are severely limiting and when light intensities and perhaps pCO_2 levels are high. Indeed, the theory predicts that local deterministic extinction of a herbivore can occur with increased light intensity or with any ecological change that induces high C:nutrient ratio in plant biomass.

A similar model has been built and analyzed where the prey is made to be stoichiometrically variable, but the predators are held to be homeostatic. An analogous situation to this would be the case of bacteria (homeostatic) being fed upon by Protozoa (stoichiometrically variable). When we change the point of strict homeostasis from the predator to the prey, the complex dynamics discussed for Fig. 7 disappear. It is both the presence of strict homeostasis and its location within a food web that influences system dynamics.

Increased Primary Production Can Reduce Food Chain Production

The idea that increased food abundance might have negative effects on consumers is a surprising stoichiometric prediction. Numerous lab and field studies have tested predictions of stoichiometric theory for trophic interactions with a special emphasis on examining how light and nutrients can jointly regulate herbivore production. For example, a laboratory study involving *Daphnia* and a green alga showed that growth of *Daphnia* was maximal at intermediate light. At low light, algal biomass was limiting to *Daphnia* growth. At very high light intensities, algal P-content was limiting (Fig. 8). Consistent with stoichiometric theory, the light intensity supporting maximal growth moved to higher light intensities as nutrient content in the system increased. In a longer-term study also involving *Daphnia* and a green alga (Fig. 9), increased light intensity was shown to inhibit *Daphnia* population growth and trophic efficiency but eventually nutrient recycling by the *Daphnia* was able to increase algal nutrient content so that high grazer densities eventually were achieved. In a result that supports the idea that increased light intensity can result in grazer extinction, in this study one replicate vessel received especially high light intensity, which resulted in unusually high algal C:P ratio and a *Daphnia* population that was never able to increase in abundance and was undergoing significant decline at the end of the experimental period.

Several studies involving field mesocosms have also shown that manipulation of light-nutrient balance has a strong effect on secondary production in nutrient-limited ecosystems. For example, a field experiment in a P-limited Canadian lake reduced light intensity to mescosms by more than tenfold. After 30 days, seston C:P ratio had decreased significantly and zooplankton production was increased nearly fivefold. Thus, there is strong evidence that secondary production in food webs is strongly influenced not only by overall rates of ecosystem productivity (`food quantity') but also by the quality of that production as indexed by its nutrient content.

Large-Scale Stoichiometry

Ecological stoichiometry also offers insights into factors regulating the multiple pathways by which energy and multiple chemical elements move or are stored at the scale of ecosystems and above. For example, whole-ecosystem manipulations of food web structure have shown how ecosystem nutrient levels and stoichiometric constraints influence the operation of cascading trophic

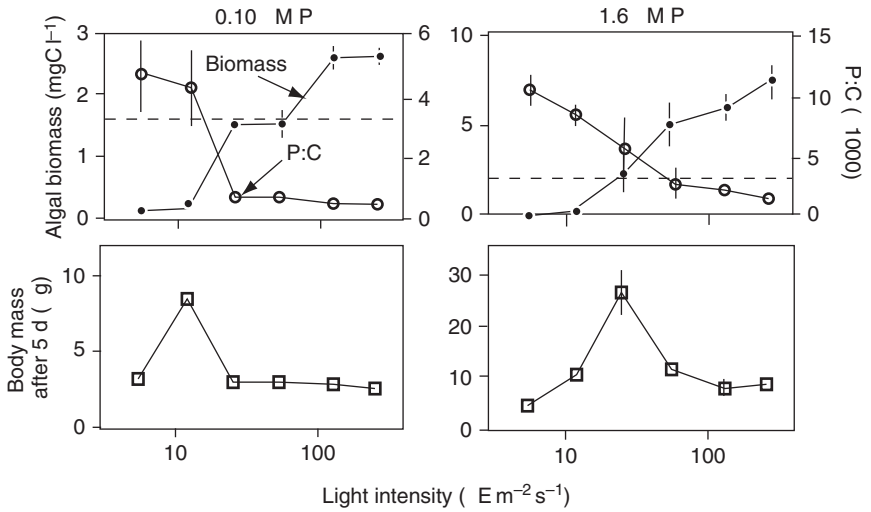


Fig. 8 Algae and grazers in light gradients. As light increases, algal biomass increases and algal P:C decreases, worsening food quality for grazers. Maximal grazer growth (highest mass) occurs at intermediate light levels.

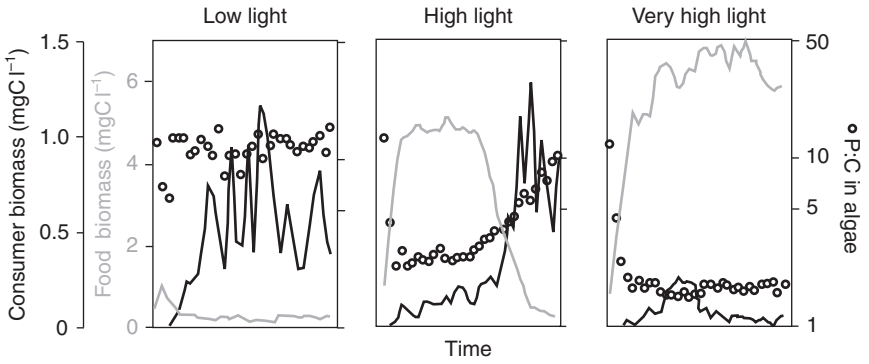


Fig. 9 Algal (food) and grazer (consumer) dynamics over time (>90 days) at three light levels. At low light, algal biomass remained low, P:C in algae was high, and grazers were variable but often abundant. At high light, algal abundance was initially high and P:C was low, but eventually the grazer became abundant and the system resembled the low-light condition. Finally, at very high light, algal abundance remained high, P:C remained low, and grazer abundance remained low.

interactions in which alterations in top trophic levels impinge on ecosystem productivity and nutrient cycling. In an experimentally fertilized, P-rich lake, introduction of a fourth trophic level (piscivorous northern pike, *Esox lucius*) led to strong reductions in planktivorous minnows and thus a major increase in zooplankton biomass and especially *Daphnia*. Consistent with stoichiometric nutrient recycling theory, increased *Daphnia* abundance was associated with increased N:P ratios in nutrient pools and a major reduction in the previously dominant N-fixing cyanobacteria. In a parallel experiment in which pike were introduced to a similar minnow-dominated but unfertilized lake, reduced minnow abundances did not result in increased *Daphnia* abundance. Instead, *Daphnia* abundances declined in parallel with changes in an unmanipulated control lake. Both *Daphnia* declines were associated with major increases in seston C:P ratios, suggesting that the *Daphnia* success in these lakes was strongly controlled by `bottom-up' food-quality constraints and relatively unaffected by `top-down' effects of food web interactions. These observations are consistent with an emerging view that strong trophic cascades may be confined to ecosystems having nutrient-rich autotrophic production at their base, as ecosystems with poor quality, nutrient-limited autotroph biomass may be unable to support high biomasses of herbivores capable of exerting significant grazing pressure.

Comparative studies across multiple ecosystems have suggested that stoichiometric constraints play an important role in regulating the fate of organic matter and the cycling of nutrients at the ecosystem scale. As mentioned earlier (**Fig. 5**), when considered across multiple studies in diverse terrestrial ecosystems, the rate constant of detrital breakdown correlates negatively with detritus C:nutrient ratio, although considerable variation can exist within particular studies. That is, nutrient-rich detritus (low C:nutrient ratio) breaks down rapidly, returning nutrients to available pools for reuptake by plants, while low-nutrient detritus (high C:nutrient ratio) breaks down slowly and may enhance immobilization of soil nutrients by microbiota, slowing the reuse by plants. Other large-scale patterns in the fate of C in ecosystems appear to be tied to autotroph nutrient content. When data for numerous ecosystems across diverse habitats (oceanic, limnetic, terrestrial) were compiled, strong positive correlations are seen between the percentage of primary production consumed by herbivores and plant nutrient content and turnover rate: ecosystems with fast-growing, nutrient–rich plant production support significant populations for herbivores which consume important quantities of that production. In contrast, in ecosystems with slow-growing, low-nutrient plant biomass, significant quantities of primary production escape consumption in the grazing food chain and enter detrital pathways and long-term C storage. Likewise, the release or retention of nitrogen or organic C appears to be a function of stoichiometric balance in watersheds. Several studies have shown that stream NO_3 concentrations or export rates are negatively correlated with average watershed soil C:N ratios, consistent with stoichiometric theory. Conversely, other studies have shown that concentrations or export rates of dissolved organic C from watersheds are positively correlated with watershed soil C:N ratios. Thus, ecophysiological limitations on the processing of organic C and nutrients by soil microorganisms ramify to affect the fluxes of materials at watershed and regional scales.

Stoichiometric constraints also play a role in the regulation of biosphere-scale processes governing oceanic and terrestrial C-cycling, atmospheric CO_2 concentrations, and thus global climate. In the open ocean, biogeochemical processes appear to be closely linked to multiple limitations associated with light intensity (as mediated by water column mixing processes), macronutrients (N, P), and micronutrients (especially iron). In the enormous central Pacific gyre, long-term studies have shown climatic variations associated with El Niño that reduce the intensity of vertical mixing processes and cross-thermocline nutrient transfers that enhance the success of light-limited, N-fixing cyanobacteria. Dominance of these algae, in turn, increases the C:P ratio of primary production and thus increases net carbon sequestration in deep waters. In other parts of the ocean, low iron supplies rather than light may limit primary production and N-fixation and thus stoichiometric coupling of C- and macronutrient cycles can ultimately depend on iron supply, itself regulated by long-term processes associated with delivery of continental dust. Since the production of continental dust is itself closely regulated by climatic conditions (rainfall patterns), a set of complex feedbacks that plays out over tens of thousands of years is established.

These feedbacks have been incorporated into global biogeochemical models that investigate the autoregulatory capability of the biosphere (`Gaia'). In these studies, it has been shown that a self-regulating system at the scale of the biosphere emerges if the organisms driving the Earth's biogeochemical systems operate under functional constraints on their processing of energy and matter. In other words, no higher-level selection at the scale of the biosphere (`Gaia' as organism) is required for planetary self-regulation to operate. Instead, `Gaia' is a complex interactive system with dynamics determined as an emergent property of selection operating primarily on the ecophysiological traits and constraints of individual organisms.

See also: Conservation Ecology: Trophic Index and Efficiency; Turnover Time; Trophic Classification for Lakes. Ecological Data Analysis and Modelling: Conceptual Diagrams and Flow Diagrams; Climate Change Models. General Ecology: Ecological Efficiency. Global Change Ecology: Nitrogen Cycle

Further Reading

Elser, J.J., Acharya, K., Kyle, M., *et al*., 2003. Growth rate – stoichiometry couplings in diverse biota. Ecology Letters 6, 936–943.
Elser, J.J., Urabe, J., 1999. The stoichiometry of consumer-driven nutrient recycling: Theory, observations, and consequences. Ecology 80, 735–751.
Redfield, A.C., 1958. The biological control of chemical factors in the environment. American Scientist 46, 205–221.
Reiners, W.A., 1986. Complementary models for ecosystems. American Naturalist 127, 59–73.
Sterner, R.W., Elser, J.J., 2002. Ecological Stoichiometry: The Biology of Elements from Molecules to the Biosphere. Princeton, NJ: Princeton University Press.
Sterner, R.W., Hessen, D.O., 1994. Algal nutrient limitation and the nutrition of aquatic herbivores. Annual Review of Ecology and Systematics 25, 1–29.
Williams, R.J.P., Fraústo Da Silva, J.J.R., 1996. The Natural Selection of the Chemical Elements: The Environment and Life's Chemistry. Oxford: Clarendon.

Ecophysiology

LA Ferry-Graham, California State University, Moss Landing Marine Labs Moss Landing, CA, USA
AC Gibb, Northern Arizona University, Flagstaff, AZ, USA

Introduction

This article contains an overview of the scientific discipline broadly defined as ecophysiology. Included is a definition of physiological ecology, a discussion of the relationship of this discipline to the fields of functional morphology and biomechanics, and a commentary regarding the use of the comparative method as a central aspect of the discipline.

What Is Ecophysiology?

Because the external environment is intrinsically variable, living organisms must maintain a separate and distinct internal environment. Physiology is, at its core, the study of the mechanisms organisms use to maintain the internal environment. The maintenance of a constant internal environment is termed homeostasis, and it is achieved through behavioral changes and internal regulation at the system, organ, cellular, and molecular level. The concept of homeostasis was first recognized by Claude Bernard (1813–78), who worked primarily in the field of human physiology. Humans demonstrate exceptionally stable internal environments, in contrast to many other animals and plants, where some aspects of the internal environment are maintained, while others are allowed to vary with changing conditions. When a given parameter of the internal environment is maintained at a constant level, or within a relatively small range, it is said to be regulated; an example of this is the internal body temperature of most mammals. However, other aspects of the internal environment may be allowed to vary with changes in the external environment, and are said to be conforming, such as the water content in certain intertidal algae. It is likely that organisms reach an evolutionary `compromise' between the potential physiological stress incurred by allowing a given parameter to vary with changes in the environment and the metabolic expense incurred by regulation.

Ecophysiology, or ecological physiology, is an area of research where physiological parameters are used to quantify the interaction between the external environment (i.e., ecology) and internal environment of the organism (Fig. 1). As such, this field is not really different from physiological ecology although some may choose one term over the other to reflect a relatively greater emphasis on ecological or physiological processes as part of their research program. One of the earliest areas of research within the ecophysiological realm was the field of `energetics', which was pioneered by H. T. Odum (1924–2002) and examines the energy flux between organisms and the environment. Energetics remains an important area of research, and the field of modern energetics includes studies of the energy required to perform a certain task (e.g., cost of transport) as well as of energy budgets – how energy is acquired and used by the organism (over various timescales) in performing all the tasks important to its survival, such as foraging, reproducing, growing, etc. Other current areas of research in ecophysiology include (but are not limited to): (1) energy acquisition in systems without light for primary production – how carbon is fixed to form the basis of food webs (e.g., hydrothermal vents); (2) modifications to the oxygen transport system – how animals respond to life in oxygen limited (e.g., elevation) or anaerobic environments (e.g., tidal marshes); (3) physiological responses to short-term, unpredictable, environmental stressors – how organisms respond to temporary, but life-threatening, changes in their environments, such a sudden, dramatic increase in ambient temperature (e.g., heat shock proteins). It is important to note that these research areas span a large range of biological organization: from whole-organism, or behavioral, responses to molecular responses. In addition, ecophysiologists often focus on environments that pose myriad and extreme physiological challenges to organisms that inhabit them. Two examples of environments of this type that have been well studied by ecophysiologists are the desert and the marine intertidal.

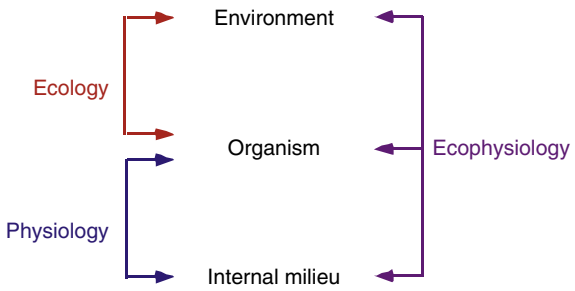


Fig. 1 Schematic diagram indicating the relative roles of ecology and physiology in creating the field of ecophysiology.

 doi:10.1016/B978-0-444-63768-0.00531-X

Example 1: The Desert

Organisms that live in the desert must maintain the internal environment in the face of high temperatures, large daily temperature fluctuations, and a perpetual scarcity of water. Whole-organism, or behavioral, responses are often used to mitigate these temperature extremes. Because muscles operate more effectively when they are warmer, behavioral thermoregulation is often employed to warm muscles after a cool desert night. By sunning on a warm rock, desert reptiles (ectotherms that produce little metabolic heat) warm themselves in the early morning and quickly reach a temperature that facilitates locomotion and feeding behaviors. However, mid-day temperatures frequently reach levels that present a thermal challenge. At high temperatures, animals encounter problems of evaporative water loss, which affects osmotic balance (maintenance of blood and intracellular volume), ionic balance (maintenance of chemical gradients), and acid–base balance (maintenance of internal pH) of the organism. At still higher temperatures, protein damage, including the deterioration of metabolic enzymes, occurs. Therefore, desert reptiles typically retreat to underground burrows or other shaded locations to avoid the heat of the day and are only active near dawn or dusk (crepuscular). Some mammals have taken this behavior to an extreme, and are active only during the cooler temperatures of the night (nocturnal). Here, mammals have an advantage because they are endothermic (i.e., produce substantial metabolic heat) and are able to maintain an internal temperature above that of the environment. Because muscles can be warmed by metabolic heat, nocturnal activity can be used to reduce water loss. Thus, evolutionary history may constrain the behavioral response of a given organism to a particular environmental stressor.

The stressors associated with the desert climate have also resulted in modifications, or adaptations, of the morphology and physiology of desert inhabitants. These adaptations often facilitate the retention of water. In plants, for example, leaves are greatly reduced or lost entirely thereby reducing evaporative water loss. Succulent for water retention stems are acquire. Interestingly, these modifications have occurred independently in several unrelated lineages, such as cacti in the Americas and euphorbs in Africa. In addition, some lineages of desert plants, notably CAM plants, have modified photosynthetic reactions with reduced CO_2 requirements. These plants keep their stomata (pores used for gas transfer) closed during the day, thereby reducing evaporative water loss. Animals also demonstrate physiological modifications to retain water in the desert. Kangaroo rats have highly modified kidneys that allow them to recover most of the water in their urine, and thus reduce excretory water loss. In fact, kangaroo rats are able to survive almost entirely on `metabolic water', that is, water produced as a by-product of the conversion of glucose into ATP. Consequently, they can survive without drinking water for extremely long periods of time.

The desert environment can also be used to illustrate research questions or ecological `problems' that are of interest to ecophysiologists. For example, how long does it take a basking lizard (an ectotherm) to achieve its preferred body temperature to warm its muscles for effective foraging and escape behaviors? To answer this question, a researcher might construct a simple mathematical model that includes the potential sources of heat gain and loss: total heat (H_T) = metabolic heat (H_M) ± conductive heat (H_C) ± convective heat (H_V) ± radiant heat (H_R) ± evaporative heat (H_E) (**Fig. 2**). The researcher might use this to make specific predictions, and then test this prediction by collecting empirical measurements from a model lizard (containing a thermocouple) placed on a basking rock. Alternately, measurements could be taken from a number of lizards of a given species in a certain size class sunning themselves in a desert environment to construct an equation that quantifies the rate of heat gain in lizards basking on rocks under a given set of environmental conditions.

Example 2: The Intertidal

The marine intertidal is the region of the shore that emerges with outgoing tides and submerges with incoming tides. Organisms in this habitat must tolerate breaking waves, which impose large and seasonally variable physical forces on the organisms. Exposed intertidal organisms, descendants of entirely aquatic ancestors, must cope with extreme desiccation stress. Tidepool organisms, or those that remain in small remnant pools of water, encounter fluctuations in water temperatures, water quality, and dissolved oxygen in the water. In addition, the longer tidepools remain separated from the ocean, the more conditions deteriorate.

As the tide recedes, organisms in the highest portion of the intertidal or attached to vertical surfaces are literally left `high and dry' – essentially thrust into a terrestrial environment. Exposure to heat and wind causes water loss, which in turn causes desiccation. Shelled invertebrates close their shells to protect their soft tissues from direct exposure. Small, mobile invertebrates and fishes hide within the intertidal algae, where they are sheltered from direct exposure, in a microenvironment with increased humidity. Many species of algae have evolved thick cuticles to reduce water loss by evaporation; other algal species have evolved the ability to tolerate large losses in water content for short periods of time. This latter group of algae serves as an example of species that do not attempt to maintain a constant internal milleu. Instead, they conform to surrounding conditions and are physiologically tolerant of the concomitant changes in tonicity and osmolarity. There are boundaries to every species' tolerance and zero percent water content would be fatal for these algae, but large changes in water content are tolerated. By either regulating or conforming, intertidal organisms can persist until they are submerged again by the incoming tide.

Cracks and crevices in the landscape often contain pools that persist between high tides, and many intertidal organisms can be found there. However, organisms living in these microhabitats face a different, but no less demanding, set of challenges between tides. These pools are typically small and shallow. During emersion, they are heated by the sun and experience large thermal fluctuations. These pools also have large surface areas, and water evaporates from the pools over time, which causes the salinity in the pools increase. Tidepool organisms also `pollute' their environment by releasing nitrogenous wastes into the water. Because these environmental challenges will vary depending on the elevation, size, and depth of a tidepool, intertidal fish species tend to

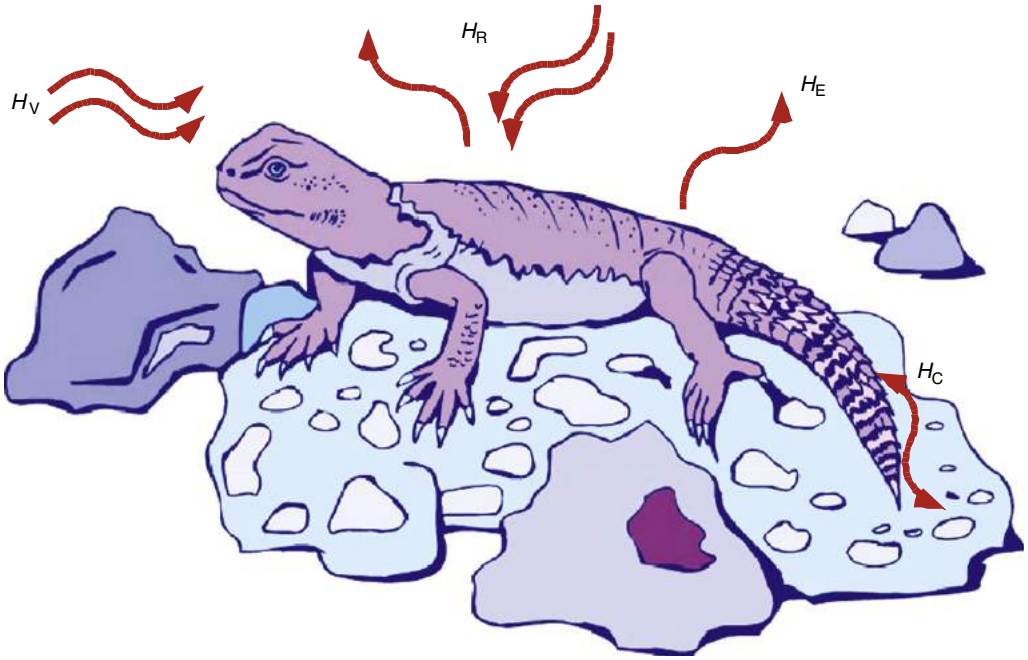


Fig. 2 Heat is transferred between an organism and its environment by radiation (H_R), convection (H_V), conduction (H_C), and evaporation (H_E). Information about ambient temperature, wind speed, color of the organism, etc., can be used to generate a model to predict the rate of heat transfer between an organism and the environment to answer specific ecophysiological questions: e.g., how fast can a lizard achieve its preferred body temperature?

distribute themselves according to their ability to either compensate internally or be tolerant of fluctuations in temperature, salinity, and nitrogen. Species that can accommodate wider fluctuations in these environmental variables tend to be found higher in the intertidal, where pools are exposed longer and conditions become more extreme.

Animals that remain in emergent tidepools also consume most of the dissolved oxygen in the water, which creates another physiological stressor. Tidepool algae or plants may exacerbate this problem by consuming oxygen at night, although they release oxygen into the pool during the day (sometimes to dangerously high levels). Because dissolved oxygen levels are typically low in tidepools, some tidepool fishes have evolved the ability to breathe air. An air-breathing fish that remains in a pool will take an `air breath' from the surface, and hold the air in its mouth, where oxygen is extracted. Some tidepool fishes also have rigid gills and gill filaments that remain erect when unsupported by water, which allows oxygen extraction to continue while air contacts the respiratory surfaces of the gill filaments. If tidepool conditions become too unfavorable, fishes with these gill modifications can exit the tidepool and potentially move to a new tidepool until the tides return and conditions improve. Emergent tidepool fishes may also extract oxygen through the skin, via a process termed cutaneous respiration. However, aerial respiration does pose a challenge for eliminating acidic and nitrogenous wastes, which normally diffuse readily through the gills into the surrounding water. In addition, because evaporative water loss occurs rapidly in a terrestrial environment, ionic and osmotic regulation are also problematic for these fishes. In fact, evaporative water loss may limit the distance and nature of the terrestrial excursions undertaken by marine intertidal fishes.

Ecophysiology, Ecomorphology, and the Ecomorphological Paradigm

The term ecophysiology, as interpreted broadly, also includes two related fields: functional morphology (or ecomorphology) and biomechanics. Studies in the field of functional morphology typically pertain to an animal's ability to move about in its environment (locomotion), to acquire energy (feeding), and to transmit its genes to the next generation (courtship and reproduction). Consequently, this field is particularly concerned with quantifying `performance' variables. Performance variables aim to quantify an animal's ability to survive and succeed in its habitat, such as maximum speed of locomotion, metabolic cost of transport, feeding rate, and maximum bite force. Biomechanics, in contrast, considers how the physical properties of the tissues that comprise an organism affect its ability to survive and succeed in its environment. This field includes studies that assess the strength and elasticity of animal and plant tissues (bone, tendon, cuticle, stems) to determine if environmental or behavioral stressors may contribute to failure of the organism to survive in its habitat, and modeling studies that atomize the organism into simple systems and consider the energy used and work being performed by that particular system.

A cornerstone of these and other related fields is the `ecomorphological paradigm', which states that structural or physical or even physiological differences among organisms are often reflections of differences in their ecology. For example, piscivorous

fishes might have long, sharp teeth for capturing prey, while planktivorous fishes may have lost teeth altogether and use modified gill arches to filter prey out of the water column. Therefore, diet can be predicted to some degree by tooth and gill morphology, and vice versa.

The Comparative Method

Modern ecophysiologists and related fields often rely upon the comparative method for studying patterns and processes. This method allows researchers to infer evolutionary processes from consistent correlations between specific modifications to organisms and a particular habitat. For example, several different plant lineages that live in desert habitats have lost their leaves (as described above). This morphological modification is hypothesized to minimize the absorption of solar radiation and loss of water. This change has occurred independently in several plant groups (i.e., the desert habitat was invaded multiple times by different plant groups), and the closest nondesert relatives of these plants do not show these morphological adaptations, which reveals that the plants did not have these traits prior to invading. Therefore, we infer that leaf loss is a trait that arose in direct response to living in the desert. By making these across and within-group comparisons, ecophysiologists can infer that leaf loss is an adaptive response to desert living that provides a fitness advantage to the plant species that possess it.

Conclusions

Ecophysiology attempts to understand the potential limits placed on organisms by their physiology, how organisms respond to particular environmental challenges, and how organisms have adapted to their ecological niches. In this context, it is important to note that while many environments are stressful and have the potential to disrupt homeostasis, the organisms that live in these environments are able to grow, mature, and reproduce under these challenging conditions. We can determine how a given species copes with an extreme environment through studies of specific physiological mechanisms. Behavior is the highest or broadest level of potential mechanism, followed by the system, organ, cellular, and molecular levels. Studies of animal behavior and functional morphology typically describe the highest levels of potential mechanisms. At lower levels, biomechanical and classical physiological mechanisms are examined. Ecophysiologists often hypothesize that these mechanisms represent adaptations to the ecological niche of the organism, and hypothesized adaptations are tested using the comparative method.

See also: Ecological Data Analysis and Modelling: Forest Models. Evolutionary Ecology: Body Size, Energetics, and Evolution. General Ecology: Metabolic Theories in Ecology: The Dynamic Energy Budget Theory and the Metabolic Theory of Ecology

Further Reading

Denny, M., 1988. Biology and the Mechanics of the Wave-Swept Environment. Princeton, NJ: Princeton University Press.
Feder, M.E., Bennett, A.F., Burggren, W.W., Huey, R.B. (Eds.), 1987. New Directions in Ecological Physiology. Cambridge: Cambridge University Press.
Harvey, P.H., Pagel, M.D., 1991. The Comparative Method in Evolutionary Biology. Oxford: Oxford University Press.
Hill, R.W., Wyse, G.A., Anderson, M., 2004. Animal Physiology. Sunderland. USA: Sinauer Associates.
Horn, M.H., Martin, K.L.M., Chotkowski, M.A. (Eds.), 1999. Intertidal Fishes: Life in Two Worlds. San Diego, CA: Academic Press.
Odum, T.P., 1969. The strategy of ecosystem development. Science 164, 262–270.
Sanford, G.M., Lutterschmidt, W.I., Hutchinson, V.H., 2002. The comparative method revisited. Bioscience 32, 830–836.
Wainwright, P.C., Reilly, S.M. (Eds.), 1993. Ecological Morphology: Integrative Organismal Biology. Chicago: University of Chicago Press.

Ecosystems

AK Salomon, University of California, Santa Barbara, CA, USA

What Is an Ecosystem?

Coined by A. G. Tansley in 1935, the term 'ecosystem' refers to an integrated system composed of a biotic community, its abiotic environment, and their dynamic interactions. A diversity of ecosystems exist through the world, from tropical mangroves to temperate alpine lakes, each with a unique set of components and dynamics (Fig. 1). Ecosystems can be classified according to their components and physical context yet their classification is highly dependent on the spatial scale of scrutiny. Typically, boundaries between ecosystems are diffuse. An 'ecotone' is a transition zone between two distinct ecosystems (i.e., the tundra–boreal forest ecotone).

History

Over 70 years ago, Sir Arthur Tansley (Fig. 2) presented the notion that ecologists needed to consider 'the whole system', including both organisms and physical factors, and that these components could not be separated or viewed in isolation. By suggesting that ecosystems are dynamic, interacting systems, Tansley's ecosystem concept transformed modern ecology. It led directly to considerations of energy flux through ecosystems and the pathbreaking, now classic work of R. L. Lindeman in 1942, one of the first formal investigations into the functioning of an ecosystem, in this case a senescent lake, Cedar Creek Bog, in Minnesota. Inspired by the work of C. Elton, Lindeman focused on the trophic (i.e., feeding) relationships within the lake, grouping together organisms of the lake according to their position in the food web. To study the cycling of nutrients and the efficiency of energy transfer among trophic levels over time, Lindeman considered the lake as an integrated system of biotic and abiotic components. He considered how the lake food web and processes driving nutrient flux affected the rate of succession of the whole lake ecosystem, a significant departure from traditional interpretations of succession.

By the late 1950s and early 1960s, system-wide energy fluxes were quantified in various ecosystems by E. P. Odum and J. M. Teal. In the late 1960s, Likens, Bormann, and others took an ecosystem approach to studying biogeochemical cycles by manipulating whole watersheds in the Hubbard Brook Experimental Forest to determine whether logging, burning, or pesticide and herbicide use had an appreciable effect on nutrient loss from the ecosystem. This research set an important precedent in demonstrating the value of conducting experiments at the scale of an entire ecosystem (see the section entitled 'Whole ecosystem experiments'), a significant advancement which continues to inform ecosystem studies today.

Ecosystem Components and Properties

Ecosystems can be thought of as energy transformers and nutrient processors composed of organisms within a food web that require continual input of energy to balance that lost during metabolism, growth, and reproduction. These organisms are either 'primary producers' (autotrophs), which derive their energy by using sunlight to convert inorganic carbon into organic carbon, or 'secondary producers' (heterotophs), which use organic carbon as their energy source. Organisms that perform similar types of ecosystem functions can be broadly categorized by their 'functional group'. For example, 'herbivores' are heterotophs that eat autotrophs, 'carnivores' are heterotrophs that eat other heterotrophs, while 'detritivores' are heterotrophs that eat nonliving organic material (detritus) derived from either autotrophs or heterotrophs (Fig. 3). Herbivores, carnivores, and dertitivores are collectively known as 'consumers'.

Classifying organisms according to their feeding relationships is the basis of defining an organism's 'trophic level'; the first trophic level includes autotrophs; the second trophic level includes herbivores and so on. Ecosystem components that make up a trophic level are quantified in terms of biomass (the weight or standing crop of organisms), while ecosystem dynamics, the flow of energy and materials among system components, are quantified in terms of rates.

Typically, ecologists quantifying ecosystem dynamics use carbon as their currency to describe material flow and energy to quantify energy flux. Material flow and energy flow differ in one important property, namely their ability to be recycled. Chemical materials within an ecosystem are recycled through an ecosystem's component. In contrast, energy moves through an ecosystem only once and is not recycled (Fig. 3). Most energy is transformed to heat and ultimately lost from the system. Consequently, the continual input of new solar energy is what keeps an ecosystem operational.

Solar energy is transformed into chemical energy by primary producers via photosynthesis, the process of converting inorganic carbon (CO_2) from the air into organic carbon ($C_6H_{12}O_2$) in the form of carbohydrates. Gross primary production is the energy or carbon fixed via photosynthesis over a specific period of time, while net primary production is the energy or carbon fixed in photosynthesis, minus energy or carbon which is lost via respiration, per unit time. Production by secondary producers is simply the amount of energy or material formed per unit term.

 doi:10.1016/B978-0-444-63768-0.00482-0

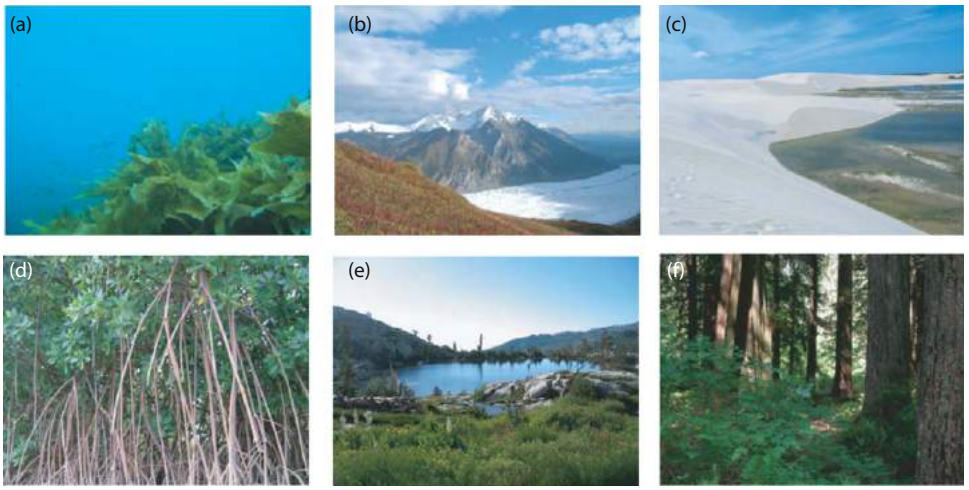


Fig. 1 (a) Kelp forest, (b) subarctic alpine tundra, (c) tropical coastal sand dune, (d) tropical mangrove, (e) alpine lake, and (f) temperate coastal rain forest. Photos by Anne Salomon, Tim Storr, and Tim Langlois.

Fig. 2 Sir Arthur G. Tansley coined the term ecosystem in 1935. From *New Phytologist* 55: 145, 1956.

A careful distinction needs to be made between production rates and static estimates of standing crop biomass, particularly because the two need not be related. For example, two populations at equilibrium, in which input equals output, might have the same standing stock biomass but drastically different production rates because turnover rates can vary (**Fig. 4**). For example, on surf swept shores from Alaska to California, two species of macroalgal primary producers grow in the low rocky intertidal zone of temperate coastal ecosystems (**Fig. 5**). The ribbon kelp, *Alaria marginata*, is an annual alga with high growth rates, whereas sea cabbage, *Hedophyllum sessile*, is a perennial alga with comparatively lower growth rates. Although they differ greatly in their production rates, in mid-July, during the peak of the growing season, these two species can have almost equivalent stand crop biomasses.

Ecosystem Efficiency

The efficiency of energy transfer within an ecosystem can be estimated as its 'trophic transfer efficiency', the fraction of production passing from one trophic level to the next. The energy not transferred is lost in respiration or to detritus. Knowing the trophic transfer efficiency of an ecosystem can allow researchers to estimate the primary production required to sustain a particular trophic level.

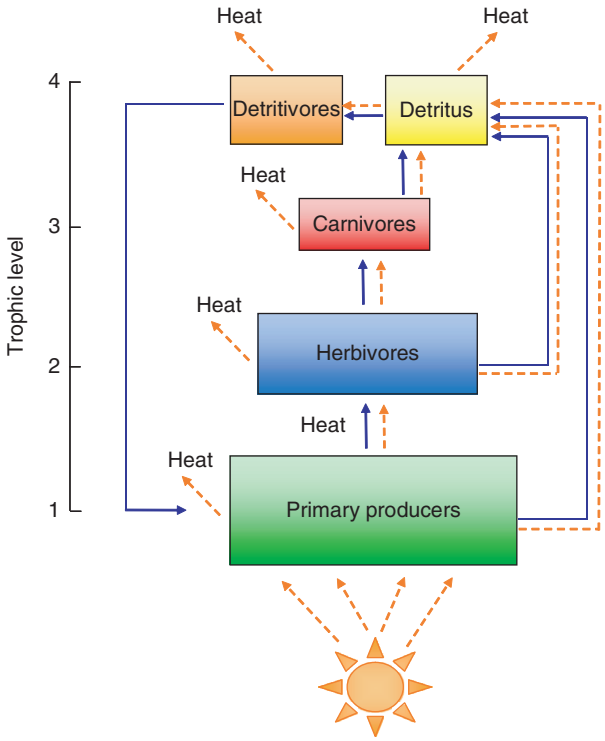


Fig. 3 Energy flows and material cycles in an ecosystem. Materials move through the trophic levels and eventually cycle back to the primary producers via the decomposition of detritus by microorganisms. Energy, originating as solar energy, is transferred through the trophic levels via chemical energy and is lost via the radiation of heat at each step. Adapted from DeAngelis DL (1992) *Dynamics of Nutrient Cycling and Food Webs*. New York, NY: Chapman and Hall.

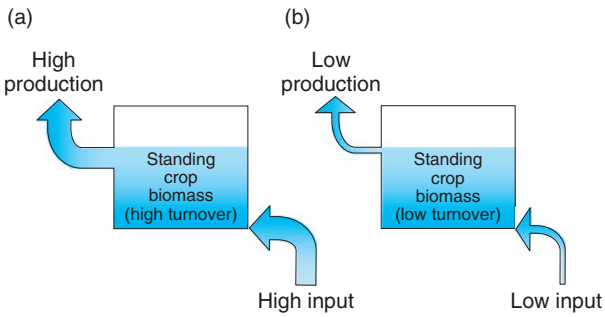


Fig. 4 Standing crop biomass is not always correlated to production rates. Here, two hypothetical species with populations at equilibrium, where input equals output, have an equivalent standing crop biomass but differ in their turnover rates. Population (a) has high input, high production, and high turnover rates, whereas population (b) has low input, low production, and low turnover rates. In reality, populations are rarely at equilibrium so standing crop biomass fluctuates depending on input rates and the amount of production consumed by higher trophic levels. Adapted from Krebs C (2001) *Ecology: The Experimental Analysis of Distribution and Abundance*, 5th edn. San Francisco: Addison-Wesley Educational Publishers, Inc.

For example, in aquatic ecosystems, trophic transfer efficiency can vary anywhere between 2% and 24%, and average 10%. Assuming a trophic efficiency of 10%, researchers can estimate how much phytoplankton production is required to support a particular fishery. Consider the open ocean fishery for tuna, bonitos, and billfish. These are all top predators, operating at the fourth trophic level. According to world catch statistics recorded by the Food and Agriculture Organization, in 1990, 2 975 000 t of these predators were caught, equivalent to 0.1 g of carbon per m^2 of open ocean per year. To support this yield of tuna, bonitos, and billfish, researchers can calculate the production rates of the trophic levels below, assuming a trophic efficiency of 10% and equilibrium conditions. Essentially, to produce of 0.1 gC m^{-2} yr^{-1} of harvested predators (tuna, bonitos, and billfish) requires 1 gC m^{-2} yr^{-1} of pelagic fish to have been consumed by the top predators, 10 gC m^{-2} yr^{-1} of zooplankton to be consumed by the pelagic fishes, and 100 gC m^{-2} yr^{-1} of phytoplankton. Note that these values represent the production that is transferred up

Fig. 5 (a) In the low intertidal zone of temperate coastal ecosystems, (b) the ribbon kelp, *Alaria marginata*, is an annual alga with high growth rates, whereas (c) the sea cabbage kelp, *Hedophyllum sessile*, is a perennial alga with lower growth rates. During the peak of the growing season, these two species can have a similar stand crop biomass but differ greatly in their production rates because one is an annual and the other is a perennial. Photo by Anne Salomon and Mandy Lindeberg.

trophic levels. They do not represent the standing stock of biomass at each trophic level. Knowing the net primary production of the photoplankton allows researchers to estimate the proportion of this production that is taken by the fishery.

It has been estimated that 8% of the world's aquatic primary production is required to sustain global fisheries. Considering continental shelf and upwelling areas specifically, these ecosystems provide one-fourth to one-third of the primary production required for fisheries. This high fraction leaves little margin for error in maintaining resilient ecosystems and sustainable fisheries.

Large-Scale Shifts in Ecosystems

A growing body of empirical evidence suggests that ecosystems may shift abruptly among alternative states. In fact, large-scale shifts in ecosystems have been observed in lakes, coral reefs, woodlands, desserts, and oceans. For example, a distinct shift occurred in the Pacific Ocean ecosystem around 1977 and 1989. Abrupt changes in the time series of fish catches, zooplankton abundance, oyster condition, and other marine ecosystem properties signified conspicuous shifts from one relatively stable condition to another (Fig. 6). Also termed 'regime shifts', the implications of these abrupt transitions for fisheries and oceanic CO_2 uptake are profound, yet the mechanisms driving these shifts remain poorly understood. It appears that changes in oceanic circulation driven by weather patterns can be evoked as the dominant causes of this state shift. However, competition and predation are becoming increasingly recognized as important drivers of change altering oceanic community dynamics. In fact, fisheries are well known to affect entire food webs and the trophic organization of ecosystems. Therefore, one could imagine that the sensitivity of a single keystone species to subtle environmental change could cause major shifts in community composition. Given this interplay between and within the biotic and abiotic components of an ecosystem, resolving the causes of regime shifts in oceanic ecosystems will likely require an understanding of the interactions between the effects of fisheries and the effects of physical climate change.

Studying Ecosystem Dynamics

Stable Isotopes

Important insights into ecosystem dynamics can be revealed through the use of naturally occurring 'stable isotopes'. These alternate forms of elements can reveal both the source of material flowing through an ecosystem and its consumer's trophic position. This is because different sources of organic matter can have unique isotopic signatures which are altered in a consistent manner as materials are transferred throughout an ecosystem, from trophic level to trophic level. Consequently, stable isotopes provide powerful tools for estimating material flux and trophic positions.

The elements C, N, S, H, and O all have more than one isotope. For example, carbon has several isotopes, two of which are ^{13}C and ^{12}C. In nature, only 1% of carbon is ^{13}C. Isotopic composition is typically expressed in δ values, which are parts per thousand differences from a standard. For carbon,

$$\delta^{13}C = \left[\left(\frac{^{13}C/^{12}\,C_{sample}}{^{13}C/^{12}\,C_{standard}}\right) - 1\right] \times 10^3$$

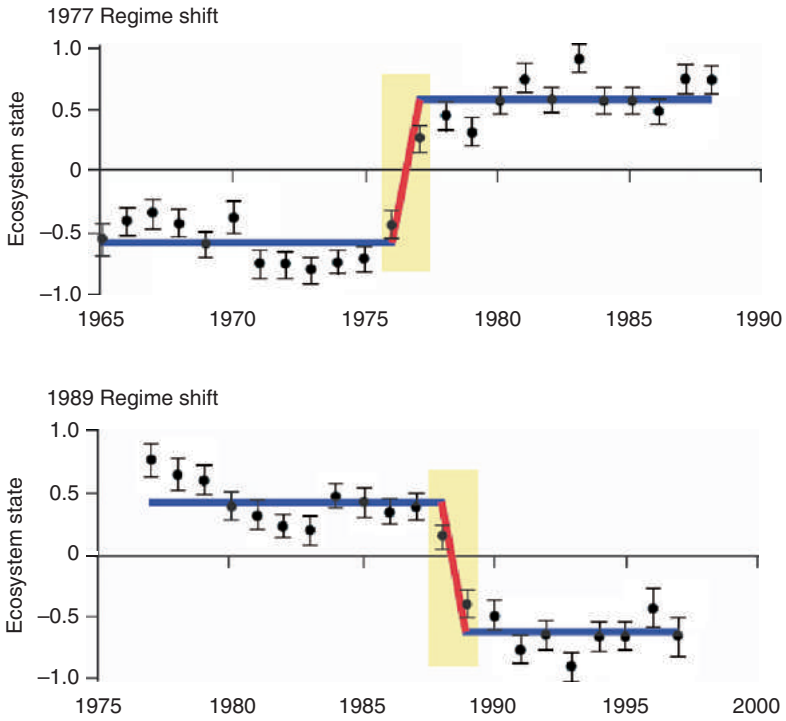


Fig. 6 Distinct shifts in ecosystem states, also referred to as 'regime shifts', occurred in the Pacific Ocean ecosystem around 1977 and 1989. The ecosystem state index shown here was calculated based on the average of climatic and biological time series. From Scheffer M, Carpenter S, Foley JA, Folke C, and Walker B (2001) Catastrophic shifts in ecosystems. *Nature* 413: 591–596.

Consequently, δ values express the ratio of heavy to light isotope in a sample. Increases in these values denote increases in the amount of the heavy isotope component. The standard reference material for carbon is PeeDee limestone, while the standard for nitrogen is nitrogen gas in the atmosphere. Natural variation in stable isotopic composition can be detected with great precision with a mass spectrometer.

Stable isotopes record two kinds of information. Process information is revealed by physical and chemical reactions which alter stable isotope ratios, while source information is revealed by the isotopic signatures of source materials. When organisms take up carbon and nitrogen, chemical reactions occur which discriminate among isotopes, thereby altering the ratio of heavy to light isotope. This is known as 'fractionation'. Although carbon fractionates very little (0.4‰, 1 SD = 1‰), the mean trophic fractionation of δ^{15}N is 3.4‰ (1 SD = 1‰), meaning that δ^{15}N increases on average by 3.4‰ with every trophic transfer. Because the δ^{15}N of a consumer is typically enriched by 3.4‰ relative to its diet, nitrogen isotopes can be used to estimate trophic position. Stable isotopes can provide a continuous measure of trophic position that integrates the assimilation of energy or material flow through all the different trophic pathways leading to an organism. In contrast, δ^{13}C can be used to evaluate the ultimate sources of carbon for an organism when the isotopic signatures of the sources are different.

Stable isotopes can track the fate of different sources of carbon through an ecosystem, because a consumer's isotopic signature reflects those of the key primary producers it consumes. For example, in both lake and coastal marine ecosystems, δ^{13}C is useful for differentiating between two major sources of available energy, benthic (nearshore) production from attached macroalgae, and pelagic (open water) production from phytoplankton. This is because macroalgae and macroalgal detritus (specifically kelp of the order Laminariales) is typically more enriched in δ^{13}C (less negative δ^{13}C) relative to phytoplankton due to boundary layer effects. Researchers have exploited this difference to answer many important ecosystem-level questions. Below are two examples.

During the late 1970s and early 1980s, in the western Aleutian Islands of Alaska, where sea otters had recovered from overexploitation and suppressed their herbivorous urchin prey, productive kelp beds dominated. There, transplanted filter feeders, barnacles and mussels, grew up to 5 times faster compared to islands devoid of kelp where sea otters were scarce and urchin densities high. Stable isotope analysis revealed that the fast-growing filter feeders were enriched in carbon suggesting that macroalgae was the carbon source responsible for this magnification of secondary production.

In four Wisconsin lakes, experimental manipulations of fish communities and nutrient loading rates were conducted to test the interactive effects of food web structure and nutrient availability on lake productivity and carbon exchange with the atmosphere. The presence of top predators determined whether the experimentally enriched lakes operated as net sinks or net sources of atmospheric carbon. Specifically, the removal of piscivorous fishes caused an increase in planktivorous fishes, a decrease in large-bodied zooplankton grazers, and enhanced primary production, thereby increasing influx rates of atmospheric carbon into the

lake. Atmospheric carbon was traced to upper trophic levels with $\delta^{13}C$. Here, naturally occurring stable isotopes and experimental manipulations conducted at the scale of whole ecosystems illustrated that top predators fundamentally alter biogeochemical processes that control a lake's ecosystem dynamics and interactions with the atmosphere.

Whole Ecosystem Experiments

Large-scale, whole ecosystem experiments have contributed considerably to our understanding of ecosystem dynamics. With its beginnings in wholesale watershed experiments in the 1960s, ecosystems are now being studied experimentally and analyzed as system of interacting species processing nutrients and energy within the context of changing abiotic conditions. This is particularly relevant these days given the effects of anthropogenic climate forcing and pollution in both terrestrial and oceanic ecosystems.

A classic series of whole-lake nutrient addition experiments conducted in northwestern Ontario by David Schindler and his research group illustrated the role of phosphorus in temperate lake eutrophication. To separate the effects of phosphorus and nitrate, the researchers split a lake with a curtain and fertilized one side with carbon and nitrogen and the other with phosphorus, carbon, and nitrogen. Within 2 months, a highly visible algal bloom had developed in the basin in which phosphorus had been added providing experimental evidence that phosphorus is the limiting nutrient for phytoplankton production in freshwater lakes. Certainly, algae may show signs of nitrogen or carbon limitation when phosphorus is added to a lake; however, other processes often compensate for these deficiencies. For instance, CO_2 is rarely limiting because physical factors such as water turbulence and gas exchange regulate its availiblity. Further, nitrogen can be fixed by blue-green algae. These species, which are favored when nitrogen is in short supply, increases the availability of nitrogen to algae, and the lake eventually returns to a state of phosphorus limitation. The practical significance of these results is that lake europhication can be prevented with management policies that control phosphorus input into lake and rivers.

Using Management Policies as Ecosystem Experiments

It has become increasingly common to use management policies as experiments and test their effects on ecosystem dynamics. An excellent example of this approach is the use of marine reserves to investigate the ecosystem-level consequences of fishing. Essentially, well-enforced marine reserves constitute large-scale human-exclusion experiments and provide controls by which to test the ecosystem effects of reducing consumer biomass via fishing at an ecologically relevant scale. Dramatic shifts in nearshore community structure have been documented in well-established and well-protected marine reserves in both Chile and New Zealand. In northeastern New Zealand's two oldest marine reserves, the Leigh Marine Reserve and Tawharanui Marine Park, previously fished predators, snapper (*Pagrus auratus*) and rock lobster (*Jasus edwardsii*), have increased in abundance by 14- and 3.8-fold, respectively, compared to adjacent fished waters. Increased predation leading to reduced survivorship and cryptic behavior of their herbivorous prey, the sea urchin (*Evechinus chloroticus*), has allowed the macroalga (*Ecklonia radiata*) to increase significantly within the reserves, a trend that has been developing in the Leigh reserve for the past 25 years (**Fig. 7**). Although this provides evidence that fishing can indirectly reduce ecosystem productivity, the trophic dynamics described above are context dependent and vary as a function of depth, wave exposure, and oceanographic circulation (**Fig. 8**). For example, both in the presence and absence of fishing, urchin densities decline to nearly 0 individuals per m^2 below depths greater than 10 m due to unfavorable conditions for recruitment, despite the presence or absence of snapper and lobster, while at depths above 3 m, wave surge can preclude urchin grazing both inside and outside the reserves. Furthermore, where oceanic conditions hinder urchin recruitment, the effects of fishing on macroalgae become less clear-cut. These physical constraints highlight the importance of abiotic context on biotic interactions. Ultimately, one can gain a lot of information by using management policies as experiments.

Although policy experiments have played an important role in elucidating ecosystem dynamics, in many cases, it is politically intractable or logistically impossible to experiment with whole ecosystems. Under such circumstances, researchers have used alternative techniques to explore ecosystem dynamics. Models in ecology have a venerable tradition for both teaching and understanding complex processes. Ecosystem models are now being used to gain insight into the ecosystem-level consequences of management policies, from fisheries to carbon emissions. For more information on ecosystem models and using management policies as experiments, see the section entitled 'Social–ecological systems, Humans as key ecosystem components'.

Ecosystem Function and Biodiversity

Accelerating rates of species extinction have prompted researchers to formally investigate the role of biodiversity in providing, maintaining, and even promoting 'ecosystem function'. Typically, studies experimentally modify species diversity and examine how this influences the fluxes of energy and matter that are fundamental to all ecological processes. In many cases, studies are designed to document the effects of species richness on the efficiency by which communities produce biomass, although the effects of species diversity on other ecosystem functions such as decomposition rates, nutrient retention, and CO_2 uptake rates have also been examined. Several seminal studies report a positive relationship between biodiversity and ecosystem function. Yet, the generality of the results, and the mechanisms driving them, have provoked considerable debate and several counterexamples exist.

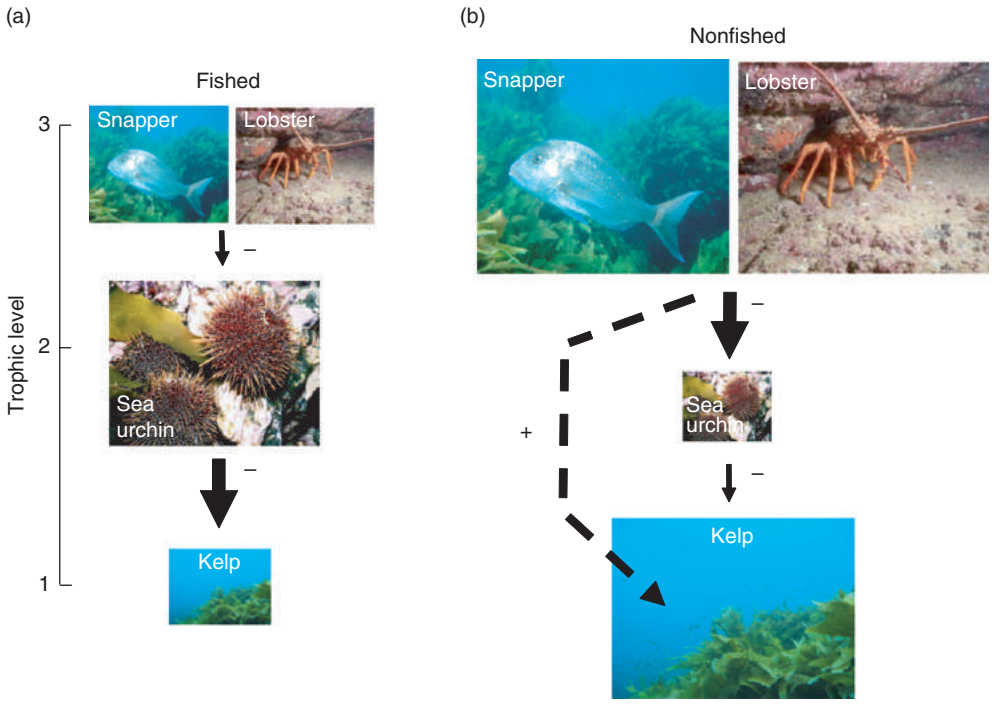


Fig. 7 (a) In nearshore fished ecosystems in northeastern New Zealand, snapper and lobster densities have been reduced due to fishing pressure resulting in high sea urchin densities, urchin barrens, and reduced kelp production. (b) In marine reserves, where previously fished snapper and lobster have recovered, sea urchins that have not been consumed by these predators behave cryptically, hiding in crevices. Consequently, kelp forests of *Ecklonia radiata* dominate. Photos by Nick Shears, Hernando Acosta, and Timothy Langlois.

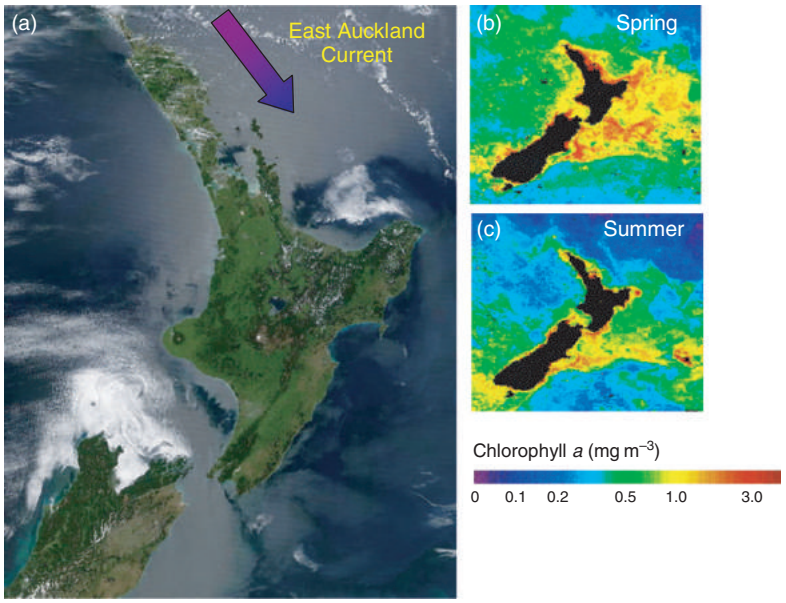


Fig. 8 The effects of fishing on nearshore ecosystems are influenced locally by wave exposure and regionally by oceanographic circulation. (a) In northeastern New Zealand, ocean circulation patterns influence nutrient delivery and thus (b) spring and (c) summer pelagic primary production. Satellite images: SeaWiFs Project, Ocean Color Web.

At the crux of the debate lies a question with deep historical roots: do some species exert stronger control over ecosystem processes than others? Imagine two distinct positive relationships between biodiversity and ecosystem function (**Fig. 9**). In type A communities, every single species contributes to the ecosystem function measured, even the rare species. By contrast, in type B communities, almost all of the ecosystem function measured can be provided by relatively few species, suggesting that many species are in fact redundant. Few empirical studies support type A relationships, rather, empirical evidence points to the

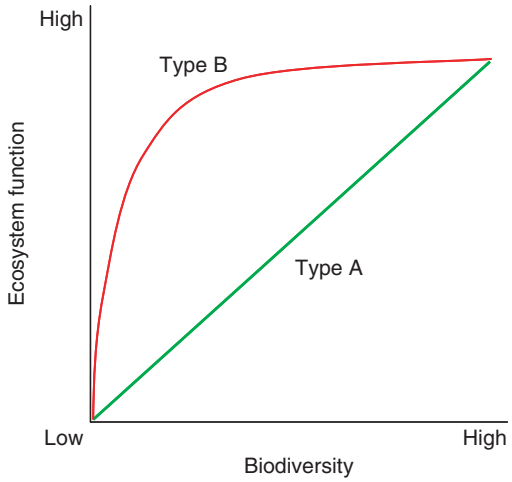


Fig. 9 Type A communities: every single species contributes equally to ecosystem functioning. Type B communities: ecosystem function is provided by only a few species.

prevalence of type B relationships. In fact, a recent meta-analysis of 111 such studies conducted in multiple ecosystems on numerous trophic groups found that the average effect of decreasing species richness is to decrease the biomass of the focal trophic group, leading to less complete depletion of resources used by that group. Further, the most species-rich polycultures performed no differently than the single most productive species used in the experiment. Consequently, these average effects of species diversity on ecosystem production are best explained by the loss of the most productive species from a diverse community. These results could be considered consistent with what has become known as the 'sampling effect'.

Critics argue that a positive relationship between species diversity and ecosystem function is a sampling artifact rather than a result of experimentally manipulated biodiversity *per se*. Such a 'sampling effect' can arise because communities comprising more species have a greater chance of being dominated by the most productive taxa. Yet, controversy surrounding the 'sampling effect' itself exists given the duality in its possible interpretation: is this a real biological mechanism that operates in nature or is it an experimental artifact of using random draws of species to assemble experimental communities? To add to the ecosystem function–biodiversity debate is the critical issue that many of these studies focus on a single trophic level and neglect or dismiss multiple trophic-level interactions, such as herbivory and other disturbances well known to alter ecosystem processes, calling into question the generality of these results.

Despite the controversy, these studies generally reinforce the notion that certain species exert much stronger control over ecological processes than others. However, identifying which species these are in advance of extinction remains a challenge. Nonetheless, identifying the mechanisms driving ecosystem functioning is an important conservation priority given that human well-being relies on a multitude of these functions.

Ecosystem Perspectives in Conservation Science

Ecosystem Services

Humans have always relied on nature for environmental assets like clean water and soil formation. Today, these assets are receiving global attention as 'ecosystem services', the conditions and processes by which natural ecosystems sustain and fulfill human life. Natural ecosystems perform a diversity of ecosystem services on which human civilization depends:

1. regulating services – purification of air and water, detoxification and decomposition of wastes, moderation of weather extremes, climate regulation, erosion control, flood control, mitigation of drought and floods, regulation of disease carrying organisms and agricultural pests;
2. provisioning services – provision of food, fuel, fiber, and freshwater;
3. supporting services – formation and preservation of soils, protection from ultraviolet rays, pollination of natural vegetation and agricultural crops, cycling of nutrients, seed dispersal, maintenance of biodiversity, primary production; and
4. cultural services – spiritual, esthetic, recreational.

Although critical to human existence, ecosystem services are often taken for granted or at best, greatly undervalued. This is ironic given that many ecosystem services are very difficult and expensive to duplicate, if they can be duplicated at all. Normally, ecosystem services are considered 'free' despite their obvious economic value. For example, over 100 000 species of animals

Fig. 10 Pollination services, provided by bees, bats, butterflies, and birds to name a few, have been valued at US$4–6 billion per year in the US alone. Consider the global value of this important ecosystem service. Photos by Steve Gaines, Heather Tallis.

provide free pollination services, including bats, bees, flies, moths, beetles, birds, and butterflies (Fig. 10). Based on the estimate that one-third of human food comes from plants pollinated by wild pollinators, pollination has been valued at US$4–6 billion per year in the US alone. Globally, the world's ecosystem services have been valued at US$33 trillion a year, nearly twice as much as the gross national product of all of the world's countries.

The idea of paying for ecosystem services has been gaining momentum. Yet, because ecosystem services are typically not sold in markets, they usually lack a market value. Given the value of natural capital, nonmarket valuation approaches are being developed by economists and ecologists to account for ecosystem services in decision-making processes. The notion being that economic valuation gives decision makers a common currency to assess the relative importance of ecosystem processes and other forms of capital.

Yet, assigning value to ecosystem services is tricky and some analysts object to nonmarket valuation, because it is a strictly anthropogenic measure and does not account for nonhuman values and needs. Yet, in democratic countries, environmental policy outcomes are determined by the desires of the majority of citizens, and voting on a preferred policy alternative is ultimately an anthropogenic activity. A second objection to nonmarket valuation is a disagreement with pricing the natural world and dissatisfaction with the capitalistic premise that everything is thought of in terms of commodities and money. The point of valuation, however, is to frame choices and clarify the tradeoffs between alternative outcomes (i.e., draining a wetland may increase the supply of developable land for housing but does so at the cost of decreased habitat and potential water quality degradation). Finally, a third objection to nonmarket valuation stems from the uncertainty in identifying and quantifying all ecosystem services. Advocates argue that economic valuation need not cover all values and that progress is made by capturing values that are presently overlooked.

Despite the uncertainties, valuing ecosystem services can sometimes pay off. When New York City compared the coast of an artificial water filtration plant valued at US$6–8 billion, plus an annual operating cost of US$300 million, the city chose to restore the natural capital of the Catskill Mountains for this watershed's inherent water filtration services and for a fraction of the cost (US $660 million). Ultimately, the valuation of ecosystem services, even if flawed, may get ecosystem processes on the decision-making table and lead to more sustainable policies in light of ever-expanding human populations.

Ecosystem services are threatened by growth in the scale of human enterprise (population size, per-capita consumption rates) and a mismatch between short-term needs and long-term societal well-being. With a global population soon to number 9 billion people, ecosystem services are becoming so degraded, some regions in the world risk ecological collapse. Many human activities alter, disrupt, impair, or reengineer ecosystem services such as overfishing, deforestation, introduction of invasive species, destruction of wetlands, erosion of soils, runoff of pesticides, fertilizers, and animal wastes, pollution of land, water, and air resources. The consequences of degrading ecosystem services on human well-being were examined in the Millennium Ecosystem Assessment (MA) 2005, which concluded that well over half of the world's ecosystems services are being degraded or used unsustainably. The MA developed global ecological scenarios as a process to inform future policy options. These scenarios were based on a suite of models that were designed to forecast future change. The MA based its scenario analyses on ecosystem services. Specifically, scenarios were developed to anticipate responses of ecosystem services to alternative futures driven by different sets of policy decisions. Following the completion of this ambitious ecological study, there is now a growing movement to make the value of ecosystem services an integral part of current policy initiatives.

Social–Ecological Systems, Humans as Key Ecosystem Components

Humans are a major force in global change and drive ecosystem dynamics, from local environments to the entire biosphere. At the same time, human societies and global economies rely on ecosystem services. As such, human and natural systems can no longer be treated independently because natural and social systems are strongly linked. Accumulating evidence suggests that effective environmental management and conservation strategies must take an integrated approach, one that considers the interactions and feedbacks between and within social, economic, and ecological systems. As a result, the concept of coupled 'social–ecological systems' has become an emerging focus in environmental and social science and ecosystem management. Social–ecological systems are considered as evolving, integrated systems that typically behave in nonlinear ways. The concept of resilience – the capacity to buffer change – has been increasingly used as an approach for understanding the dynamics of social–ecological systems. Two useful tools for building resilience in social–ecological systems are structured scenario modeling and active adaptive management.

Models of linked social–ecological systems have been developed to inform management conflicts over water quality, fisheries, and rangelands. These models represent ecosystems coupled to socioeconomic drivers and are explored with stakeholders to probe the management decision-making processes. Alternative scenarios force participants to be absolutely explicit about their assumptions and biases, thereby improving communication between stakeholders and exposing the ecological consequences of various management policies.

Adaptive management is an approach where management policies themselves are deliberately used as experimental treatments. As information is gained, policies are modified accordingly. This approach helps isolate anthropogenic effects from sources of natural variation and, most importantly, considers the consequences of a human perturbation on the whole ecosystem. In contrast, basic research on various parts of an ecosystem leads to the challenge of assembling all the data into a practical framework. Yet, biotic and abiotic ecosystem components are not additive, they interact. Due to these interactions, the dynamics of an ecosystem cannot be extrapolated from the simple addition of an ecosystem's components. Adaptive management examines the response of the system as a whole rather than a sum of its parts. Furthermore, this approach involves adaptive learning and adaptive institutions that acknowledge uncertainties and can respond to nonlinearities. In sum, structured scenario modeling and policy experimentation are tools that can be used to examine the resilience of social–ecological systems to alternative management policies and conservation strategies.

Ecosystem-Based Management

Recognizing the need to sustain the integrity and resilience of social–ecological systems has led to calls for 'ecosystem-based management', a management approach that considers all ecosystem components, including humans and the physical environment. With the overall goal of sustaining ecosystem structure and function, this management approach:

- focuses on key ecosystem processes and their responses to perturbations;
- integrates ecological, social, and economic goals and recognizes humans as key components of the ecosystem;
- defines management based on ecological boundaries rather than political ones;
- addresses the complexity of natural processes and social systems by identifying and confronting uncertainty;
- uses adaptive management where policies are used as experiments and are modified as information is gained;
- engages multiple stakeholders in a collaborative process to identify problems, understand the mechanisms driving them, and create and test solutions; and
- considers the interactions among ecosystems (terrestrial, freshwater, and marine).

Ecosystem-based management is driven by explicit goals, executed by policies and protocols, and made adaptable by using policies as experiments, monitoring their outcomes and altering them as knowledge is gained.

Traditionally, management practices have focused on maximizing short-term yield and economic gain over long-term sustainability. These practices were driven by inadequate information on ecosystem dynamics, ignorance of the space and timescales on which ecosystem processes operate, and a prevailing public perception that immediate economic and social value outweighed the risk of alternative management. Seeking to overcome these obstacles, ecosystem-based management relies on research at all levels of ecological organization, explicitly recognizes the dynamic character of ecosystems, acknowledge that ecological processes operate over a wide range of temporal and spatial scales and are context dependent, and presupposes that our current knowledge of ecosystem function is provisional and subject to change. Ultimately, ecosystem-based management recognizes the importance of human needs while addressing the reality that the capacity of our world to meet those needs in perpetuity has limits and depends on the functioning of resilient ecosystems.

See also: General Ecology: Community

Further Reading

Cardinale, B.J., Srivastava, D.S., Duffy, J.E., *et al.*, 2006. Effects of biodiversity on the functioning of trophic groups and ecosystems. Nature 443, 989–992.
Daily, G.C. (Ed.), 1997. Nature's Services: Societal Dependence on Natural Ecosystems. Washington, DC: Island Press.

DeAngelis, D.L., 1992. Dynamics of Nutrient Cycling and Food Webs. New York, NY: Chapman and Hall.
Krebs, C., 2001. Ecology: The Experimental Analysis of Distribution and Abundance, 5th edn. San Francisco: Addison-Wesley Educational Publishers, Inc.
Millennium Ecosystem Assessment, 2005. Ecosystems and Human Well-Being: Synthesis. Washington, DC: Island Press.
Pauly, D., Christensen, V., 1995. Primary production required to sustain global fisheries. Nature 374, 255–257.
Scheffer, M., Carpenter, S., Foley, J.A., Folke, C., Walker, B., 2001. Catastrophic shifts in ecosystems. Nature 413, 591–596.

Edaphic Factor☆

Nishanta Rajakaruna, College of the Atlantic, Bar Harbor, ME, United States
Robert S Boyd, Auburn University, Auburn, AL, United States

Introduction

This article describes the important roles geology and soil conditions play in the ecology and evolution of plant species and their associated biota. We seek to: (1) describe the edaphic factor as a life force responsible for generating and maintaining unique species assemblages and (2) emphasize the importance of conserving habitats with extreme edaphic conditions because of their biological diversity. First, we describe the edaphic factor: its definition and role in shaping the biotic world. Then we review our current knowledge of the ecology of unusual geologies, focusing on studies performed within and across biotic kingdoms. Further, we examine the process of plant evolution on extreme geologies, an area that has generated much interest among evolutionary biologists in the last few decades. Finally, we cover the applied ecology and conservation of plants and other biota restricted to unique geologies.

The Edaphic Factor: Its Role in Shaping the Biotic World

Ecologists have long noted the importance of geology in the global and regional distribution of organisms. Life, ranging from macro- to microscopic, exists on and within a mosaic of geologies that vary across both space and time. The contributions of geologic phenomena to maintaining and generating biotic diversity are twofold. First, large-scale geologic events (e.g., continental drift and rising of mountains) create discontinuous or patchy landscapes. Second, within this patchwork of landscapes, parental geologic materials such as igneous, metamorphic, or sedimentary rocks can become exposed, leading to the development of soils differing in chemical and physical characteristics. This creates opportunities for colonization and differentiation of species. The edaphic factor pertains to physical, chemical, and biological properties of soil resulting from these geologic phenomena. Discontinuities in the edaphic factor have contributed to the intriguing patterns of diversity we see in the biotic world. Edaphology is a branch of soil science that studies the influences of soils on organisms, especially plants. It includes agrology, the study of human uses of soils for agriculture, as well as how the features of soils affect human land use decisions.

According to soil ecologist Hans Jenny, soils owe their distinct characteristics to five interacting factors: climate, organisms, topography, parental rock, and time. If all but one factor (e.g., parental rock) remain unchanged, then variation in a soil body can be attributed to that one factor. Botanists have long recognized that the distribution, habit, and composition of vegetation are greatly influenced by the edaphic factor. The striking effects on vegetation of unusual and often extreme substrates (e.g., serpentine, limestone, dolomite, shale, gypsum) are apparent even to amateur naturalists. Whereas climate broadly defines major biomes (e.g., tropical rainforests, temperate deciduous forests, deserts, tundra), it is geology that enriches diversity within these zones. The role played by the edaphic factor in the distribution of plant species was keenly observed and recorded by many 18th- and 19th-century plant ecologists, who considered soils second only to climate as the major ecological determinant of plant distribution. It was in the twentieth century, however, that ecologists fully appreciated the role of the edaphic factor in generating habitats within which plants and their associated organisms live, interact, reproduce, and diverge over time.

Components of the Edaphic Factor

Plants generally obtain nutrient elements and water from soil. Thus, soil features that affect the availability and uptake of nutrients and water are of great importance to plants. Below is a brief overview of the soil features that most greatly affect plant growth.

Texture

Mineral particles in soils can be classified on the basis of their size (diameter). Clays are very small (<0.002 mm) particles, silts are larger (0.002–0.05 mm), and sands yet larger (0.05–2.0 mm). The percentage of each of these major particle classes soil determines the texture of a soil. Textures range from those predominantly containing one of the three major particle sizes and thus named for them (silt, sand, clay textures) or various intergradations (sandy clay, etc.). One major textural class (loam), which is ideal for plant growth, is not named for a predominant particle class because loam soils have similar amounts of all three particle

☆*Change History*: April 2018. Irene Martins has updated the text throughout the article.

This is an upadte of N. Rajakaruna and R.S. Boyd, Edaphic Factor, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 1201–1207.

classes. Texture is important for plant growth because it influences water availability and soil fertility, that is, the ability of a soil to supply nutrients to plant roots. Open spaces among soil particles represent the pore space of a soil. This pore space may be filled by water, air, or plant roots. Thus, the amount of pore space greatly influences the amount of water that a soil can contain. The tightness with which water is held in pore spaces determines whether or not the water and dissolved mineral ions will drain through the soil profile or remain in place and be available for uptake by plant roots. Coarse textured soils (e.g., sandy soils) have large pore spaces that do not hold water tightly enough to prevent gravity from pulling that water into deeper soil layers. Very fine textured soils (e.g., clay soils) have very small pores. Small pores hold water strongly and thus retain much water despite the pull of gravity. However, they also hold water against the pulling power of plant roots and so provide only small amounts of plant-available water. Soil texture is also an important determinant of the cation exchange capacity (CEC) of a soil, that is, the ability of a soil to adsorb and exchange mineral ions essential for plant growth. Soils with higher percentages of clay or silt particles generally have a greater CEC.

Structure

The three-dimensional arrangement of soil particles gives rise to soil structure. In many soils, groups of particles are held together to create lumps of soil materials termed peds. The space between peds can be important in allowing penetration of water and roots to deeper soil layers.

Depth

Soil depth can greatly influence the types of plants that can grow in them. Deeper soils generally can provide more water and nutrients to plants than more shallow soils. Furthermore, most plants rely on soil for mechanical support and this is especially true for tall woody plants (e.g., shrubs, trees). A classic example of the influence of soil depth on plant communities is seen on granite rock outcrops in the southeastern United States. As the granite weathers, it can form pools of soil that vary in depth from a few millimeters at the margin to tens of centimeters in the middle. The shallow marginal soils support certain annual plants, whereas deeper soils support herbaceous perennials and still deeper soils are colonized by woody plants. Plant zonation in these soil pools can be striking (**Fig. 1**). Some soils can develop special soil horizons (horizontal soil layers characterized by distinct chemical and physical features) that limit the soil depth available to support plants. These special soil horizons include claypans, zones of soil which contain large amounts of clay, and hardpans, layers of soil particles that have been cemented together by the deposition of mineral materials. Hardpans include calcic horizons (commonly called caliche), in which calcium carbonate cements the soil particles. The net effect of these dense horizons is to impede or prevent root growth and thus limit the effective depth of the soil. They also may affect soil oxygenation by restricting drainage at times in which large amounts of water are present.

Organic Matter

Organic matter in soils ranges from recognizable plant parts (roots, leaves, stems) to humus, which is partly decomposed plant material that is amorphous and spongy in nature. Organic matter contributes to a soil's ability to retain nutrients and water (i.e., soil's CEC). It aids in holding nutrients because negatively charged compounds in humus attract and hold positively charged plant nutrient ions. It helps provide water because humus can absorb 80%–90% of its weight in water and therefore contributes to a soil's ability to hold water under drought conditions. Organic matter in soils is also an important food source for decomposer and detritivore organisms. Organic matter also contributes to soil color, a factor that affects a soil's thermal properties.

pH

Soil pH is an extremely important ecological parameter. Its most important effect on plant growth is its influence on ion availability in the soil solution. Ions in soils are important for two major reasons. One is that many soil ions contain elements required for plant growth. These elements, called essential nutrients, are primarily obtained from the soil. The second reason is that plants obtain most of their water from soil and the amount of dissolved ions in soil water can influence a plant's ability to take up water (see the following section). The influence of pH on ion availability stems primarily from the influence of pH on the solubility of the various compounds present in the soil. In general, soil compounds containing some elements are more soluble at some pH values than others. For example, iron is relatively insoluble at pH values of 8 or greater and plants with a high iron requirement may perform poorly in soils with high pH values. Similarly, many heavy metals become increasingly available for plant uptake at pH values of 4–5. Thus, plants growing in low pH soils are more susceptible to heavy metal toxicities.

Ion Availability

Although we mentioned ion availability under pH (above), we should also mention that some ions are abundant in some soils primarily because they have been deposited in those soils in great amounts. Certain salts (often Na, Mg, or Ca salts) may be abundant in some soils in quantities that greatly affect plant growth. These salts include those from seawater (as in salt marshes) or

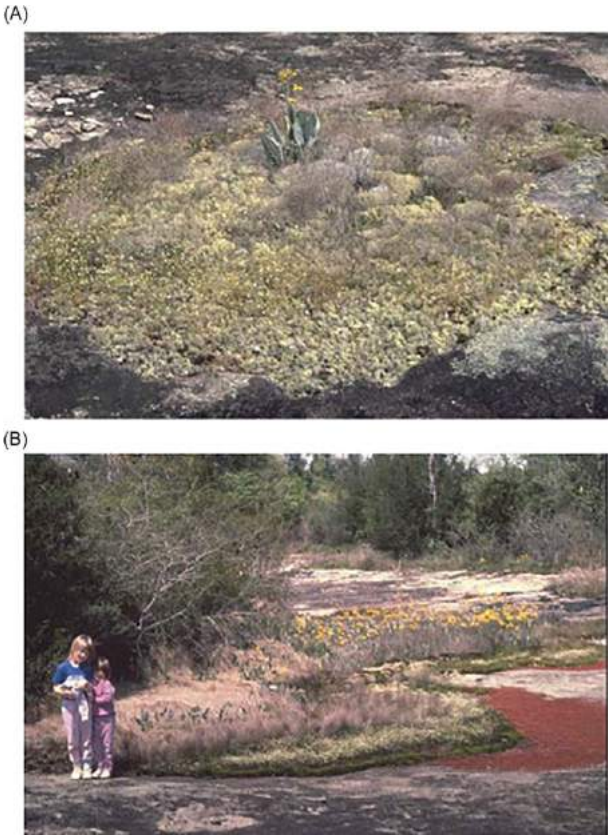


Fig. 1 (A) A small soil pool (about 2 m wide) on a granite outcrop in east-central Alabama. Shallow soil at the margins is dominated by lichens. The deepest soil in the center of the pool has been colonized by *Senecio tomentosus*, a yellow-flowered herbaceous perennial species. (B) A larger soil pool on the same granite outcrop shown in (A). Deep soil on the *left* (behind the children: Jenny and Kristina Boyd) is occupied by woody plants (shrubs and trees). The soil pool becomes more shallow to the *right*, where striking zonation of smaller plants can be observed. The most shallow soil on the extreme right is occupied by the small *red*-colored annual *Sedum smallii*. Slightly deeper soil to the left of the *Sedum* zone is dominated by moss (*Polytrichum commune*) and white-flowered annual *Arenaria* species. Still deeper soil between that zone and the woody plants is dominated by perennial grasses along with some *Senecio tomentosus*. Credit: R. S. Boyd.

those that build up in desert soils from evaporative concentration of relatively freshwater (e.g., the Great Salt Lake of Utah). Extensive irrigation of land in regions where there is high evapo-transpiration can also lead to accumulation of salts on the soil surface (i.e., secondary salinization). Salty soils (e.g., saline, sodic, saline-sodic soils) can impact plant growth by affecting water uptake, nutrient uptake or by causing toxicity due to specific ion effects. Water uptake can be slowed because the high ion concentration in the soil impedes water movement into plant roots. Nutrient uptake can be affected because ions can competitively inhibit the uptake of essential ions of similar size (e.g., Na^+ vs. K^+, Mg^{2+} vs. Ca^{2+}). Excess ions can also have specific toxic effects on plants by directly inhibiting essential physiological processes.

The Edaphic Factor in Ecology

Given the importance of soil features to plants, the edaphic factor's influence on plant ecology and evolution is unsurprising. In particular, soils with unusual features (extreme pH, nutrient imbalances, limited depth, etc.) may be a strong selective force shaping plant evolution. The floras of many unusual soils (serpentine soils, limestone soils, etc.) have at least some taxa that are found only on those soil types, whereas other species may evolve locally adapted populations (ecotypes, races, etc.). In many cases such taxa or populations have evolved in response to particular features of those soils. In other cases, unusual soils may be refugia for taxa that are unable to compete with species that dominate "normal" soils.

The ability of soils to affect ecology or evolution of organisms other than plants is less well known. It is also less likely for many animals, in part because their mobility and aboveground lifestyle render them less influenced by the various properties of soils. One soil feature that in specific cases has been shown to directly influence animal evolution is soil color. In habitats with little vegetation, such as deserts and beaches, the color of some animals has evolved to match the color of the soil. For example, white gypsum dunes, for which the White Sands National Monument in New Mexico is named, host a number of animals that are notably lighter-colored than those living on darker surrounding soils. These animals include

Fig. 2 The Ni tolerant insect *Melanotrichus boydi* (Heteroptera: Miridae) on a flower of its host plant, the California Ni hyperaccumulator *Streptanthus polygaloides* (Brassicaceae). The plant is found only on serpentine soils in California, and the insect is found only on *S. polygaloides*. The insect is tolerant of the high levels of Ni found in the plant tissues (usually >3000 μg Ni/g dry mass). The insects, about 5 mm long, contain about 800 μg Ni/g dry mass, enough to make them toxic to crab spiders that hunt for prey on flowers of *S. polygaloides*. Credit: R. S. Boyd.

insects, spiders, scorpions, lizards, amphibians, and mammals. The main selective advantage of this color matching is to provide camouflage that makes color-matched animals less likely to fall victim to predators. The evolution of burrowing and soil-dwelling animals is more likely to be influenced by soil properties due to the greater intimacy of their life histories with those soil features.

These direct effects of soils on biota are supplemented by a variety of indirect ways that soils may influence either animals or plants by affecting organism interactions. This is easily imagined when one considers the importance of plant communities in providing habitat for animals and other organisms. There are several intriguing cases of special plant–insect interactions under extreme edaphic conditions. Some plants endemic to heavy metal-rich serpentine soils harbor unique insect herbivores that are specialized to deal with the high metal concentrations found in the plant tissue (Fig. 2). It is rare to find cases in which the effects of soil on other organisms indirectly affect plants, but this does occur. For example, pocket gophers tunnel through soil and consume aboveground, and especially belowground, plant parts. In mountain meadows of Arizona, aspen trees suffer significant gopher-caused mortality on deep meadow soils but not on rocky outcrops where pocket gophers do not occur due to the lack of soil deep enough for them to make tunnels.

Soils and Biogeography

Landscape ecology is a subfield of ecology that examines the patterns and interactions between communities that make up relatively large areas. At this level of ecological scale, the pattern of soil types on a landscape may have important ecological consequences. One of these consequences is diversity (species richness, evenness). In a sense, patches of one soil type in a matrix of another are like islands in the sea (Fig. 3), and thus can be subjected to the ideas of Island biogeography. Island biogeography holds that the number of species present on an island is primarily determined by the size of the island and by its distance from sources of colonists. Thus, relatively small patches of unusual soils (i.e., edaphic islands) that are far from similar patches would be expected to have fewer species than large patches close to other areas of similar soils. This has been confirmed by recent studies of the flora on patches of serpentine soils in California.

Soils and Invasive Species

Invasive species are non-native species that become abundant enough to cause significant negative effects on some native species or the function of native ecosystems. Because soils are an important factor of the environment of organisms, it is no surprise that soil features can affect the ability of non-native species to become invasive. In many cases, disturbance of native communities (including changes in soils caused by disturbance) provides inroads for invasive species. Some studies have contrasted the susceptibility to invasion of unusual soils (such as serpentine soils) and more normal soils. The general conclusion is that the features of the unusual soils that make them challenging for plant growth often inhibit the invasiveness of non-native species. Anthropogenic activities can directly influence soil chemistry of some unusual edaphic habitats making such habitats conducive for colonization by invasive species. This appears to be the case for atmospheric nitrogen deposition on serpentine sites in California. Recent studies suggest that vehicle emissions along major highways in California may have increased the nitrogen content in serpentine soils. Non-native species, previously excluded from such soils due to nitrogen limitation, could potentially invade these unique habitats.

Fig. 3 Edaphic islands of serpentine outcrops in the Klamath Siskiyou Mountain range in northern California. The relatively barren serpentine outcrops (bare patches with reduced vegetation cover) are embedded in a mosaic of other geologies more favorable for plant growth. Credit: N. Rajakaruna.

It is common for invasive species (particularly plants) to impact soils and, through changes in soil characteristics, affect other organisms in those communities. For example, in northwestern US, diffuse knapweed (*Centaurea diffusa*; Asteraceae) has a direct soil-mediated impact on competing native plants. This invasive plant produces 8-hydroxyquinoline, a chemical that builds up in soils occupied by *C. diffusa* and poisons native plants growing in those soils. Invasive animal species have also been shown to alter soil features that then impact many other organisms in a habitat. For example, earthworms are not native to the forests of Minnesota but have been introduced in many locations. By consuming soil litter and accelerating its breakdown, these animals increase soil compaction, decrease water penetration, and change the nature of the litter layer habitat in ways that reduce its suitability for some native animals, herbaceous plants, and tree seedlings.

Plant Life on Selected Edaphic Conditions

Unusual edaphic conditions harbor unique plant associations often characterized by rarity and endemism. Such conditions also foster distinct morphological and physiological modifications leading to characteristic plant communities. One of the most remarkable edaphic habitats in which such unique plant communities are found is on serpentine soils derived from ultramafic and related rocks (i.e., rocks high in iron and magnesium silicates). Ultramafic rocks such as serpentinite and their associated serpentine soils are found throughout the world, concentrated, however, along continental margins and in regions of orogenesis (i.e., mountain building). Serpentine soils are unique in that they are often high in pH and heavy metals such as magnesium, nickel, and chromium, and generally low in essential nutrients, Ca/Mg ratio, and water-holding capacity. The rocks are often on open, steep slopes exposed to high light and heat conditions, and resulting soils are generally shallow and highly erodable (**Fig. 4**). The serpentine syndrome, the unique biological effects manifested by these extreme geoedaphic conditions, has led to research on plant physiology, ecology, and evolution in many parts of the world. Serpentine soils, although covering a mere few percent of the Earth's surface, host many endemic species. In the Californian Floristic Province, for example, 198 out of 2133 taxa endemic to that province are wholly or largely restricted to serpentine. Tropical islands of New Caledonia and Cuba provide even better examples of plant restriction to serpentine soils. In New Caledonia, 3178 taxa, roughly 50% of the native flora, are endemic to serpentine soils while in Cuba, 920 species, one-third of the taxa endemic to Cuba, have developed solely on serpentine soils. Similar restrictions and remarkable floristic associations are also found in serpentine areas of the Mediterranean, Africa, Australia–New Zealand, and Asia. Studies of metal hyperaccumulators (i.e., plants that accumulate at least 0.1% of their dry leaf weight in a heavy metal) of serpentine soils have not only led to the discovery of novel physiological pathways and their underlying genetic bases but have also laid the foundation for the development of innovative technologies such as phytoremediation (i.e., the use of hyperaccumulators to extract heavy metals from contaminated soils).

Gypsum ($CaSO_4 \cdots 2H_2O$), a substrate formed by the evaporation of saline waters, is also widely known for its distinctive indicator flora. The plant response to gypsum (gypsophily) manifests itself as unique communities consisting of gypsophilic endemics. While gypsum-associated plant communities are found in parts of Europe, deposits in xeric areas of the American southwest and adjacent Mexico are especially noted for their unique species composition.

Limestone forms by precipitation and lithification of $CaCO_3$, also leads to the formation of unique plant communities. In fact, some of the earliest observations on edaphic-plant relationships were made on landscapes overlying limestone and, by the late twentieth century, studies of limestone plant ecology had yielded a plethora of published work in both North America and Europe. Limestone and associated materials such as dolomite ($CaMgCO_3$) have exerted a profound influence on regional floras across the world resulting in unique vegetation compositions. Of interest are those temperate formations found on the White Mountains of

Fig. 4 Serpentine hills of Clear Creek Management Area, San Benito County, California. Serpentine exposures on these steep, open hills are prone to erosion. Credit: N. Rajakaruna.

eastern California, Mount Olympus of Greece, the European Alps, and tropical formations of Jamaica, Cuba, Turkey, and parts of Asia.

In addition to habitats formed on geologies with extreme chemical composition, other edaphically influenced habitats, such as savannas, barrens, guano-rich bird nesting rocks, coastal bluffs, alkaline flats, and vernal pools are also important sites that harbor unique communities of plants and animals.

Evolution Under Extreme Edaphic Conditions

Plant species or distinct populations belonging to certain species can often be distinguished by their faithfulness to particular edaphic conditions (Fig. 5). Plants that grow on chemically or physically extreme substrates are often derived from populations found off such substrates, suggesting the role extreme soil conditions can play in generating plant diversity. Influential work conducted during the mid-twentieth century on the grasses of heavy metal-contaminated mine tailings provides a classic demonstration of the role natural selection plays in maintaining diversity. This work, and subsequent work on many plant species, demonstrate that populations can evolve tolerance to extreme edaphic conditions and that this may lead to reduced gene flow between the ancestral population and the divergent, edaphically specialized, population. Such reproductive isolation, followed by further divergence, sets the foundation for the origin of new plant species. Plants that have either evolved in situ (i.e., neoendemics) or those that have had broader distribution but are currently restricted to extreme substrates (i.e., paleoendemics) are called edaphic endemics. While most plant species can be found under a range of edaphic habitats, it is these edaphically specialized taxa and their ancestral species that have attracted the attention of plant physiologists and evolutionary ecologists alike.

Edaphic endemics provide a model system to examine the process of plant evolution from adaptation and reproductive isolation to genetic divergence. Closely related species pairs are often distinguished by their distinct edaphic preferences. Such pairs can be found on adjacent yet contrasting soils formed naturally due to variation in parental rocks or by anthropogenic acts such as quarrying, mining, and even depositing of chemical waste in landfills. The process of divergence might proceed as follows: some individuals of a species have genetically determined traits that allow them to successfully survive in adjacent, chemically harsh soils. These individuals could become founders of a distinct population characterized by their tolerance to the extreme edaphic condition. Such a transition to a new habitat, if accompanied by a reduction in gene flow, can bring about full-fledged speciation. Evolution of tolerance to extreme conditions can occur quite rapidly, even within a few generations. Current phylogenetic analyses provide strong support for rapid evolution of edaphic specialists as recently illustrated for the species pair *Layia glandulosa–L. discoidea* (Asteraceae).

Conservation of the Biota of Extreme Geologies

Habitats on extreme geologies, from natural outcroppings of serpentine rocks to barrens resulting from anthropogenic activity, harbor unique species assemblages. Unfortunately, ever-expanding agriculture and forestry, mining activity, and urbanization have drastically affected the biota of many areas with unusual geologies. Plants associated with heavy-metal-rich geologies (i.e., metallophytes) are not merely biological novelties: they are the optimal choice for the restoration of metal-contaminated sites across the world. Phytoremediation is a growing field that uses metal-hyperaccumulating plants in the remediation of metal contaminated sites. The raw material for such endeavors comes from species found on extreme geologies such as serpentine

Fig. 5 An edaphically controlled vegetation boundary at Jasper Ridge Biological Preserve, San Mateo County, California. The yellow-flowered *Lasthenia californica* (Asteraceae) is restricted to serpentine soils. The sharply demarcated boundary between *L. californica* and grasses is defined by a serpentine-sandstone transition. Credit: Bruce A. Bohm.

outcrops, pointing to an immediate need for the conservation and detailed study of these habitats. Fortunately, recent years have seen the declaration of several preserves, set aside primarily due to their unique edaphic habitats and associated biota. Although they are spotty in their distribution and inadequate in number on a global scale, several preserves in the states of California, Oregon, and Washington, in the Province of Québec in eastern Canada, and in New Zealand and South Africa, have led the way in raising awareness of the immediate need for the conservation of these unique biotas. There has also been an urgent plea from those associated with research on metallophytes, advocating the prioritization of future research needs for the conservation of metallophyte diversity as well as the sustainable uses of metallophyte species in restoration and remediation of contaminated sites worldwide.

Further Reading

Alexander, E.B., Coleman, R.G., Keeler-Wolf, T., Harrison, S.P., 2007. Serpentine Geoecology of western North America: Geology, soils, and vegetation Oxford University Press, New York, NY

Anderson, R.C., Fralish, J.S., Baskin, J.M., 1999. Savannas, barrens, and rock outcrop plant communities of North America Cambridge University Press, New York, NY

Antonovics, J., Bradshaw, A.D., Turner, R.G., 1971. Heavy metal tolerance in plants. Advances in Ecological Research 7, 1–85.

Baldwin, B.G., 2005. Origin of the serpentine-endemic herb *Layia discoidea* from the widespread L. *glandulosa* (Compositae). Evolution 59, 2473–2479.

Boyd, R.S., 2004. Ecology of metal hyperaccumulation. New Phytologist 162, 563–567.

N.C. Brady, R.R. Weil. The nature and properties of soils 3rd edn. 1999 Prentice-Hall, Upper Saddle River, NJ

Brady, K.U., Kruckeberg, A.R., Bradshaw Jr., H.D., 2005. Evolutionary ecology of plant adaptation to serpentine soils. Annual Review of Ecology, Evolution, and Systematics 36, 243–266.

Chenchouni, H., 2017. Edaphic factors controlling the distribution of inland halophytes in an ephemeral salt lake "Sabkha ecosystem" at north African semi-arid lands. Science of the Total Environment 575, 660–671.

Guardia, G., Marsden, K.A., Vallejo, A., Jones, D.L., Chadwick, D.R., 2018. Determining the influence of environmental and edaphic factors on the fate of the nitrification inhibitors DCD and DMPP in soil. Science of the Total Environment 624, 1202–1212.

Jenny, H., 1980. The soil resource: Origin and behaviour McGraw Hill, New York, NY

Kruckeberg, A.R., 2002. Geology and plant life: The effects of landforms and rock types on plants University of Washington Press, Seattle, WA.

Lomolino, M.V., Riddle, B.R., Brown, J.H., 2005. Biogeography 3rd edn. Sinauer Associates, Sunderland, CT

Mabilde, L., De Neve, S., Sleutel, S., 2017. Regional analysis of groundwater phosphate concentrations under acidic sandy soils: Edaphic factors and water table strongly mediate the soil P-groundwater P relation. Journal of Environmental Management 203 (part 1), 429–438.

Macnair, M.R., Gardner, M., 1998. The evolution of edaphic endemics. In: Howard, D.J., Berlocher, S.H. (Eds.), Endless forms: Species and speciation. New York, NY: Oxford University Press, pp. 157–171.

Rajakaruna, N., 2004. The edaphic factor in the origin of plant species. International Geology Review 46, 471–478.

Singh B., Cattle S.R., Field D.J., 2014. Edaphic soil science, introduction to, editor(s): N. K. Van Alfen, Encyclopedia of Agriculture and Food Systems, Academic Press, Pages 35–58

Whiting, S.N., Reeves, R.D., Richards, D., *et al.*, 2004. Research priorities for conservation of metallophyte biodiversity and their potential for restoration and site remediation. Restoration Ecology 12, 106–116.

Endotherm☆

Marta K Labocha and Jack P Hayes, University of Nevada, Reno, NV, United States

What Is an Endotherm?

All organisms break complex molecules into simpler molecules, thereby obtaining energy to sustain life. This energy is used to transport ions, pump blood, move the body, and many other functions. Energy use ultimately leads to the production of internal heat. Endotherms are capable of producing sufficient internal heat to elevate their body temperature (or part of their body) above environmental temperature. In contrast, ectotherms can not produce sufficient internal heat to elevate their body temperature above environmental temperature. Endotherms have higher energy use than similar-sized ectotherms; consequently, endotherms consume more of the production by the ecosystem than do similar-sized ectotherms. The ability to be an endotherm depends on the rate of internal heat production, the size of the organism, the degree of insulation, and environmental circumstances.

Many endotherms, such as mammals and birds, elevate their entire body temperature above the environmental temperature. In other cases, endothermy is regional, such that only some parts of the organism (e.g., the brain or locomotor muscle of tuna or the flower of plant) are heated above environmental temperature. In some organisms, endothermy is not continuous but occurs only periodically or during particular activities. For example, brooding pythons heat their body only while incubating eggs, and insects elevate temperature of their active flight muscles only prior to and during flight.

In the older literature, the term endotherm was sometimes used interchangeably with the term homeotherm, but this usage is incorrect. Homeotherms are organisms that maintain relatively constant body temperature. In contrast, organisms whose body temperature varies substantially (generally tracking the environmental temperature) are called poikilotherms. Generally mammals freezing point of pure water). These fishes are homeotherms because their environmental temperature is relatively constant, not because they are thermoregulating. On the other hand, some, but not most, mammals and birds are heterothermic. That means their body temperature varies over a wider range than is typical for homeotherms (but it does not necessarily track environmental temperature as in poikilotherms). This variation in body temperature may occur over short timescales, such as nightly torpor in bats and hummingbirds or the rather variable body temperatures of some mammals, such as the egg-laying echidna. Variation in body temperature of birds and mammals may also occur over seasonal timescales, for example, as is seen in bears and ground squirrels. Animals that regulate their body temperature in the face of changing environmental temperature are called thermoregulators (**Fig. 1**). Thermoregulation and birds are homeothermic while most other animals are poikilothermic, but this is not always the case. For example, some Antarctic ice fish have body temperatures that show extremely little variation (and are below the can be accomplished physiologically by altering rates of heat production or heat loss. Animals may also thermoregulate behaviorally by selecting warmer or cooler microenvironments. Animals whose body temperature tracks the environmental temperature are called thermoconformers (**Fig. 1**).

Two other common terms in thermal physiology are cold blooded and warm blooded. These terms are no longer widely used by biologists because they convey little useful information beyond the temperature of an animal (or its blood). For example, animals may be cold blooded at one time and warm blooded at another depending on their thermal environment. Many reptiles have high body temperatures (i.e., are warm blooded during the day) when the thermal environment is hot, but they may be cold blooded when the thermal environment is cold (e.g., at night or during winter). Similarly heterothermic birds and mammals may be warm blooded during some parts of the diurnal or annual cycle and cold blooded at other times.

Other terms frequently used with relation to endothermy are standard metabolic rate (SMR), resting metabolic rate (RMR), basal metabolic rate (BMR), and field metabolic rate (FMR). SMR is metabolism measured at a particular environmental temperature while an animal is inactive and not digesting or absorbing food. RMR differs from SMR in that an animal may be digesting or absorbing food. BMR is a special case of SMR, when metabolic rate is measured within the thermoneutral zone (TNZ; see below). FMR is metabolic rate of animals in their natural environment. SMR, RMR, and FMR can be used to describe the metabolic rate of both endotherms and ectotherms, but the term BMR applies only to endotherms.

Which Organisms Are Endotherms?

The most familiar groups of endotherms are mammals and birds. Only mammals and birds are endothermic at rest. However, because some endothermic fish (e.g., tunas) must swim continuously to ventilate their gills, these fish might also be

☆*Change History:* April 2018. Editor, Irene Martins updated some references.

This is an update of M.K. Labocha and J.P. Hayes, Endotherm, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 1270–1276.

 doi:10.1016/B978-0-12-409548-9.11221-7

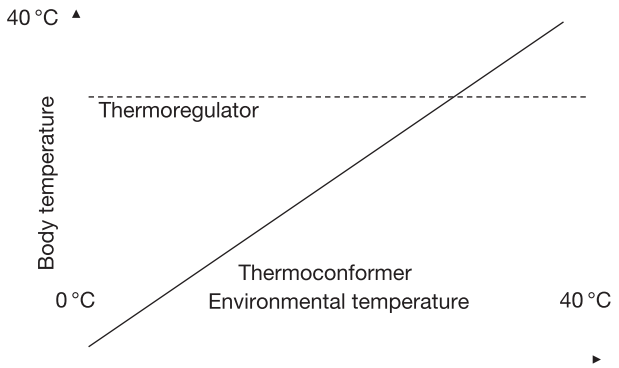


Fig. 1 Change in body temperature of thermoregulators and thermoconformers with environmental temperature change.

considered to be endotherms even at rest or as close to rest as their biology allows. Besides birds and mammals, other endotherms include some species of reptiles (e.g., large sea turtles, brooding pythons), ray-finned fishes (e.g., tunas and billfishes), lamnid sharks, insects (e.g., beetles, dragonflies, cicadas, moths, and bees), gymnosperm plants (i.e., cycads), and flowering plants (e.g., some species from the families Annonaceae, Aracacea, Araceae, Cyclanthaceae, and Rafflesiaceae).

Source of Heat

Mammals and birds differ from other endotherms because they can maintain high body temperature at rest. A high basal metabolism is necessary for mammals and birds to elevate their body temperature above environmental temperature even if the animal is resting. Birds and mammals have much higher standard metabolic rates than ectotherms, and the differences in standard metabolism are associated with different molecular, cellular, and organ characteristics. Compared to ectotherms, birds and mammals have higher mitochondrial volume, greater membrane surface per tissue volume, and higher aerobic enzyme activity. Their cellular membranes have greater ion fluxes, such that physiologists describe the cellular membranes of endotherms as being 'leakier' than those of ectotherms. Indeed, the greater leakiness of cellular membranes has been hypothesized to be one of the major processes responsible for the higher standard metabolic rate of endotherms.

Most endothermic reptiles, fishes, and insects support an elevated body temperature with heat produced by working muscle, and they do not maintain endothermy at rest. So their molecular, cellular, and organ characteristics are more similar to ectotherms than they are to endotherms. Besides producing heat by muscle, endothermic fishes developed vascular countercurrent heat exchangers to reduce heat loss and thereby promote endothermy. Some sea turtles also have countercurrent heat exchangers (in their flippers). In fishes that elevate their brain or eye temperature, the method of heat production varies among species, and the method of heat production is known not for all species. In lamnid sharks, heat is transferred from muscle, whereas billfishes and mackerel produce heat with specialized heater tissues (evolved from ocular muscles). In plants, endothermy is achieved by decoupling ATP generation from heat production via use of an alternative electron-transport mechanism (i.e., cyanide-insensitive respiration).

Endothermy, Metabolic Rate, and Body Size

The metabolic rate necessary to achieve endothermy is influenced by size because heat loss to the environment is roughly proportional to an animal's surface area. Surface area depends on body mass (i.e., for similarly shaped organisms, surface area increases with body mass raised to the power 2/3). Because surface area increases more slowly than body mass, larger animals generally can achieve endothermy with lower metabolic intensities (i.e., metabolic rates per unit mass) (**Table 1**). Indeed, even with low metabolic intensities, animals with very large body masses (e.g., sauropod dinosaurs) may have been endothermic. A lower limit on size of vertebrate endotherms also exists. The smallest vertebrate endotherms are roughly 2 g in mass but the smallest vertebrate ectotherms can be much smaller.

Endothermic Response to Temperature

Endothermic homeotherms alter their metabolic rate in response to environmental temperature (**Fig. 2**). The response to temperature depends in part on how precisely or not endotherms maintain body temperature. Over a range of intermediate temperatures, called the thermal neutral zone (TNZ), metabolic rate remains constant. If body temperature and metabolic rate are constant, but the thermal gradient between the animal and the environment is changed, then the insulation (or its reciprocal conductance) of the animal must change. A typical endothermic homeotherm would have maximal conductance at the upper end

Table 1 Basal metabolic rate (BMR) of some mammals over a large range of body mass

Species	*Common name*	*Body mass (g)*	*BMR (mL O_2 h^{-1})*	*Mass-specific BMR (mL O_2 g^{-1} h^{-1})*
Sorex minutus (Sparti, 1992)	Pygmy shrew	3.3	28.4	8.60
Peromyscus maniculatus (Hinds and Rice-Warner, 1992)	Deer mouse	14.9	44.4	3.00
Clethrionomys glareolus (Labocha *et al.*, 2004)	Bank vole	20.9	52.5	2.50
Petrodromus tetradactylus (Downs and Perrin, 1995)	Four-toed elephant shrew	206.1	179.5	0.87
Erinaceus europaeus (Shkolnik and Schmidt-Nielsen, 1976)	European hedgehog	750.0	337.5	0.45
Fossa fossa (McNab, 1989)	Fanaloka	2260.0	904.0	0.40
Felis pardalis (McNab, 1989)	Ocelot	10,416.0	3229.0	0.31
Antilocapra americana (Wesley *et al.*, 1973)	Pronghorn	37,800.0	9318.0	0.25
Panthera onca (McNab, 1989)	Jaguar	68,900.0	12,402.0	0.18
Alces alces (Renecker and Hudson, 1986)	Moose	325,000.0	51,419.0	0.16

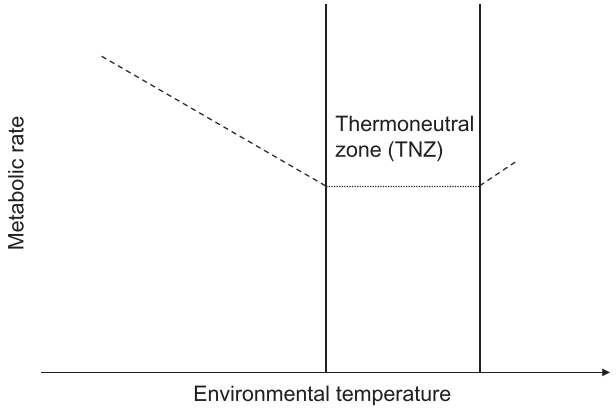


Fig. 2 Change in metabolic rate of endotherms in response to environmental temperature. Other response patterns are possible, but this pattern is roughly typical for endotherms that are strict thermoregulators.

of its TNZ and minimal conductance at the lower end of its TNZ. Above the upper end of the TNZ, metabolic rate and body temperature increase, and the animal may undertake such behaviors as panting to increase evaporative heat loss. Below the lower end of the TNZ, endotherms increase heat production so that they can offset increased heat loss in colder environments; thereby, they balance heat production and heat loss to maintain a constant body temperature (**Fig. 2**).

Endotherms that live in cold environments (e.g., aquatic environments, where the thermal conductivity of the water is many times greater than that of air) tend to be well insulated, and aquatic endotherms tend to be large. Insulation in vertebrates can take the form of fur, feathers, or fat layers. As insulation increases, the TNZ widens and the lower end of the TNZ occurs at lower temperatures. Animals that are very well insulated (the arctic fox is a classic example) have thermal neutral zones that may extend to extremely cold temperatures (e.g., they may experience temperatures as low as $-40^{\circ}C$ or perhaps even lower without having to increase their basal metabolic rate).

High temperatures can present problems for endotherms. When the thermal environment is hot, endotherms must increase evaporation to avoid overheating. Hence, hot environments can be challenging, particularly if ample water is not available.

Benefits and Costs of Endothermy

The ability to control body temperature enables endotherms to live in cold or thermally variable environments yet remain active. On a geographical scale, endotherms can live in regions too cold for most ectotherms. On a temporal scale, endotherms can be active year-round in regions with cold seasons, as well as be active during cold parts of the day or night. Endotherms not only have high SMR, RMR, and FMR, they also have high aerobic capacities (maximal aerobic metabolism). These high aerobic capacities

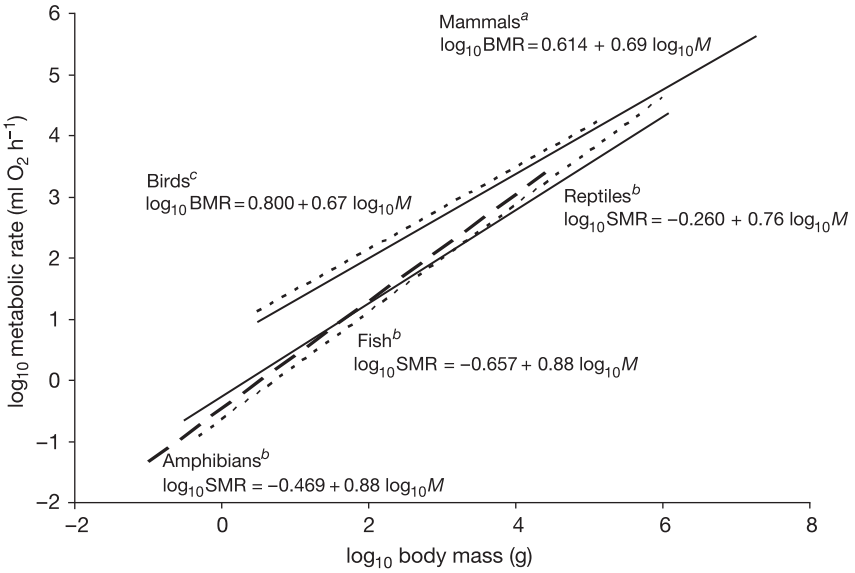


Fig. 3 Relationship between metabolic rate and body mass in vertebrates. The data for amphibians, fish, and reptiles were normalized to 38°C as described in White et al. Equation for birds was recalculated from Watts (used by McKechnie and Wolf) to mL O_2 h^{-1} (Lovegrove, 2000; White et al., 2006; McKechnie and Wolf, 2004).

Table 2 Field metabolic rate (FMR) of endotherms and ectotherms of similar body mass

Species	*Common name*	*Endotherm (ENDO)/ectotherm (ECTO)*	*Body mass (g)*	*FMR (kJ d^{-1})*
Peromyscus maniculatus (Hayes, 1989)	Deer mouse	ENDO	18.0	57.3
Erithacus rubecula (Tatner and Bryant, 1993)	Robin	ENDO	18.7	71.3
Mabuya striata (Nagy and Knight, 1989)	Striped skink	ECTO	19.5	2.9
Thamnophis sirtalis (Peterson *et al.*, 1998)	Common garter snake	ECTO	22.0	5.2
Spermophilus parryi (Nagy, 1994)	Arctic ground squirrel	ENDO	630.0	817.0
Alectoris chukar (Kam *et al.*, 1987)	Chukar	ENDO	440.0	306.2
Tiliqua scincoides (Christian *et al.*, 2003)	Bluetongue lizard	ECTO	574.0	45.7
Aptenodytes patagonicus (Kooyman *et al.*, 1992)	King penguin	ENDO	12,900.0	7410.0
Chelonia mydas (Southwood *et al.*, 2006)	Green sea turtle	ECTO	16,000.0	1867.0

make it possible to sustain vigorous activity for extended time periods. As a result, endotherms can maintain large territories and migrate for long distances.

While there are diverse benefits of endothermy, endothermy also has costs. In particular, endotherms have higher metabolic rates than ectotherms (**Fig. 3**), so they must locate and process more food to provide the energy to sustain higher metabolism. Typically, an endotherm requires much more energy and food than a similar-sized ectotherm (**Table 2**). The cost of endothermy relative to ectothermy changes with environmental temperature. At warm environmental temperatures, the energetic cost of endothermy relative to ectothermy is not as great as at cold environmental temperatures (**Table 3**).

Energetic Influence of Endotherms on Ecosystems

Endothermy is more energetically costly than ectothermy. Because endotherms use more energy than ectotherms, the same amount of food can maintain a larger population of similar-sized ectotherms than endotherms. Moreover, 90% or more of the energy assimilated by endotherms is converted to heat, so only a small percentage of the food energy drawn from the ecosystem by endotherms is converted to biomass (i.e., to grow tissue or produce offspring). In other words, endotherms have lower production efficiency than ectotherms.

Because of the high energetic cost of endothermy, endothermic carnivores require higher prey densities than ectothermic carnivores. In systems with low primary productivity they will be absent or rare. Even folivorous endotherms may be absent in habitats with extremely low productivity.

Table 3 A comparison of metabolic rates of endotherms and ectotherms of similar body mass

Species	*Common name*	*Endotherm (ENDO)/ ectotherm (ECTO)*	*Temperature (°C)*	*Body mass (g)*	*Metabolic rate* ($mL\ O_2\ h^{-1}$)
Dipsosaurus dorsalis (Bennett and Dawson, 1972)	Desert iguana	ECTO	25	35	1.68
Dipsosaurus dorsalis (Bennett and Dawson, 1972)	Desert iguana	ECTO	30	35	2.45
Dipsosaurus dorsalis (Bennett and Dawson, 1972)	Desert iguana	ECTO	35	35	5.25
Dipsosaurus dorsalis (Bennett and Dawson, 1972)	Desert iguana	ECTO	40	35	6.30
Notoryctes caurinus (Withers *et al.*, 2000)	North-western marsupial mole	ENDO	30.8	34	21.4
Phyllostomus discolor (McNab, 1969)	Pale spear-nosed bat	ENDO	34.6	33.5	11.1
Gerbillus allenbyi (Haim, 1984)	Allenby's gerbil	ENDO	36.3	35.3	38.8
Acanthodactylus erythrurus (Pough and Busack, 1978)	Fringe-toed lizard	ECTO	20	9.0	1.17
Acanthodactylus erythrurus (Pough and Busack, 1978)	Fringe-toed lizard	ECTO	25	9.0	1.62
Acanthodactylus erythrurus (Pough and Busack, 1978)	Fringe-toed lizard	ECTO	30	9.0	2.25
Acanthodactylus erythrurus (Pough and Busack, 1978)	Fringe-toed lizard	ECTO	35	9.0	3.15
Crocidura crossei (Sparti, 1990)	Crosse's shrew	ENDO	34.3	10.2	22.4
Perognathus longimembris (Chew *et al.*, 1967)	Little pocket mouse	ENDO	34.7	8.9	9.5
Pteronotus davyi (Bonaccorso *et al.*, 1992)	Davy's naked-backed bat	ENDO	38.8	9.4	24.4

For endotherms, basal metabolic rate is shown. For ectotherms, standard metabolic rate over a range of temperatures is shown.

Evolution of Endothermy

Because endothermy is energetically expensive and evolved more than 100 million years ago (at least in birds and mammals), the selective forces leading to the evolution of endothermy are unclear. The earliest hypotheses to explain the evolution of endothermy postulated that selection for higher resting metabolism led to an expanded thermal niche or increased thermal stability. Later, the aerobic capacity model posited that endothermy evolved as by-product of selection for high aerobic capacity (i.e., maximal oxygen consumption capacity during exercise). Based on data for extant organisms, aerobic capacity was argued to be inescapably correlated with resting metabolism. The newest models suggest that endothermy evolved as consequence of selection for intense parental care. Which of these models for the evolution of endothermy is best is unresolved.

Acknowledgment

This contribution was supported in part by US National Science Foundation award IOB 0344994 to J. Hayes.

See also: General Ecology: Homeotherms

References

Bennett, A.F., Dawson, W.R., 1972. Aerobic and anaerobic metabolism during activity in the lizard Dipsosaurus dorsalis. Journal of Comparative Physiology 81, 289–299.
Bonaccorso, F.J., Arends, A., Genoud, M., Cantoni, D., Morton, T., 1992. Thermal ecology of moustached and ghost-faced bats (Mormoopidae) in Venezuela. Journal of Mammalogy 73, 365–378.
Chew, R.M., Lindberg, R.G., Hayden, P., 1967. Temperature regulation in the little pocket mouse, Perognathus longimembris. Comparative Biochemistry and Physiology 21, 487–505.
Christian, K.A., Webb, J.K., Schultz, T.J., 2003. Energetics of bluetongue lizard (Tiligua scincoides) in a seasonal tropical environment. Oecologia 136, 515–523.
Downs, C.T., Perrin, M.R., 1995. The thermal biology of three southern African elephant-shrews. Journal of Thermal Biology 20, 445–450.
Haim, A., 1984. Adaptive variations in heat production within gerbils (genus Gerbillus) from different habitats. Oecologia 61, 49–52.
Hayes, J.P., 1989. Field and maximal metabolic rates of deer mice (Peromyscus maniculatus) at low and high altitudes. Physiological Zoology 62, 732–744.

Hinds, D.S., Rice-Warner, C.N., 1992. Maximum metabolism and aerobic capacity in heteromyid and other rodents. Physiological Zoology 65, 188–214.
Kam, M., Degen, A.A., Nagy, K.A., 1987. Seasonal energy, water, and food consumption of Negev chukars and sand partridges. Ecology 68, 1029–1037.
Kooyman, G.L., Cherel, Y., Le Maho, Y., *et al*., 1992. Diving behavior and energetics during foraging cycles in king penguins. Ecological Monographs 62, 143–163.
Labocha, M.K., Sadowska, E.T., Baliga, K., Semer, A.K., Koteja, P., 2004. Individual variation and repeatability of basal metabolism in the bank vole, Clethrionomys glareolus. Proceedings of the Royal Society B 271, 367–372.
Lovegrove, B.G., 2000. The zoogeography of mammalian basal metabolic rate. American Naturalist 156, 201–219.
McKechnie, A.E., Wolf, B.O., 2004. The allometry of avian basal metabolic rate: Good predictions need good data. Physiological and Biochemical Zoology 77, 502–521.
McNab, B.K., 1969. The economics of temperature regulation in neotropical bats. Comparative Biochemistry and Physiology 31, 227–268.
McNab, B.K., 1989. Basal rate of metabolism, body size and food habits in order Carnivora. In: Gittleman, J.L. (Ed.), Carnivore behavior, ecology, and evolution. Ithaca, NY: Cornell University Press, pp. 335–354.
Nagy, K.A., 1994. Field bioenergetics of mammals: What determines field metabolic rates. Australian Journal of Zoology 42, 43–53.
Nagy, K.A., Knight, M.H., 1989. Comparative field energetics of a Kalahari skink (*Mabuya striata*) and gecko (*Pachydactylus bibroni*). Copeia. 1), 13–17.
Peterson, C.C., Walton, B.M., Bennett, A.F., 1998. Intrapopulation variation in ecological energetics of the garter snake Thamnophis sirtalis, with analysis of the precision of doubly labeled water measurements. Physiological Zoology 71, 333–349.
Pough, F.H., Busack, S.D., 1978. Metabolism and activity of the Spanish fringe-toed lizard (Lacertidae: Acanthodactylus erythurus). Journal of Thermal Biology 3, 203–205.
Renecker, L.A., Hudson, R.J., 1986. Seasonal energy expenditures and thermoregulatory responses of moose. Canadian Journal of Zoology 64, 322–327.
Shkolnik, A., Schmidt-Nielsen, K., 1976. Temperature regulation in hedgehogs from temperate and desert environments. Physiological Zoology 49, 56–64.
Southwood, A.L., Reina, R.D., Jones, V.S., Speakman, J.R., Jones, D.R., 2006. Sesonal metabolism of juvenile green turtles (Chaledonia mydas) at Heron Island, Australia. Canadian Journal of Zoology 84, 125–135.
Sparti, A., 1990. Comparative temperature regulaton of African and European shrews. Comparative Biochemistry and Physiology A 97, 391–397.
Sparti, A., 1992. Thermogenic capacity of shrews (Mammalia, Soricidae) and its relationship with basal rate of metabolism. Physiological Zoology 65, 77–96.
Tatner, P., Bryant, D.M., 1993. Interspecific variation in daily energy expenditure during avian incubation. Journal of Zoology 231, 215–232.
Wesley, D.E., Knox, K.L., Nagy, J.G., 1973. Energy metabolism of pronghorn antelopes. Journal of Wildlife Management 37, 563–573.
White, C.R., Phillips, N.F., Seymour, R.S., 2006. The scaling and temperature dependence of vertebrate metabolism. Biology Letters 2, 125–127.
Withers, P.C., Thompson, G.G., Seymour, R.S., 2000. Metabolic physiology of the north-western marsupial mole, Notoryctes caurinus (Marsupialia: Notoryctidae). Australian Journal of Zoology 48, 241–258.

Further Reading

Alagaili, A.N., Bennett, N.C., Mohammed, O.B., Zalmout, I.S., Boyles, J.G., 2017. Body temperature patterns of a small endotherm in an extreme desert environment. Journal of Arid Environments 137, 16–20.
Bennett, A.F., 1991. The evolution of activity capacity. Journal of Experimental Biology 160, 1–23.
Bennett, A.F., Dawson, W.R., 1972. Aerobic and anaerobic metabolism during activity in the lizard Dipsosaurus dorsalis. Journal of Comparative Physiology 81, 289–299.
Bennett, A.F., Ruben, J.A., 1979. Endothermy and activity in vertebrates. Science 206, 649–654.
Block, B.A., Finnerty, J.R., Stewart, A.F.R., Kidd, J., 1993. Evolution of endothermy in fish: Mapping physiological traits on a molecular phylogeny. Science 260, 210–214.
Bonaccorso, F.J., Arends, A., Genoud, M., Cantoni, D., Morton, T., 1992. Thermal ecology of mustached and ghost-faced bats (Mormoopidae) in Venezuela. Journal of Mammalogy 73, 365–378.
Boyles, J.G., McKechnie, A.E., 2010. Energy conservation in hibernating endotherms: Why "suboptimal" temperatures are optimal. Ecological Modeling 221 (12), 1644–1647.
Boyles, J.G., Warne, R.W., 2013. A novel framework for predicting the use of facultative heterothermy by endotherms. Journal of Theoretical Biology 336, 242–245.
Chew, R.M., Lindberg, R.G., Hayden, P., 1967. Temperature regulation in the little pocket mouse, Perognathus longimembris. Comparative Biochemistry and Physiology 21, 487–505.
Christian, K.A., Webb, J.K., Schultz, T.J., 2003. Energetics of bluetongue lizard (Tiligua scincoides) in a seasonal tropical environment. Oecologia 136, 515–523.
Dickson, K.A., Graham, J.B., 2004. Evolution and consequences of endothermy in fishes. Physiological and Biochemical Zoology 77, 998–1018.
Downs, C.T., Perrin, M.R., 1995. The thermal biology of three southern African elephant-shrews. Journal of Thermal Biology 20, 445–450.
Haim, A., 1984. Adaptive variations in heat production within gerbils (genus Gerbillus) from different habitats. Oecologia 61, 49–52.
Hayes, J.P., 1989. Field and maximal metabolic rates of deer mice (Peromyscus maniculatus) at low and high altitudes. Physiological Zoology 62, 732–744.
Hayes, J.P., Garland Jr., T., 1995. The evolution of endothermy: Testing the aerobic capacity model. Evolution 49, 836–847.
Heinrih, B., 1993. The hot-blooded insects: Strategies and mechanisms of thermoregulation. Cambridge, MA: Harvard University Press.
Hinds, D.S., Rice-Warner, C.N., 1992. Maximum metabolism and aerobic capacity in heteromyid and other rodents. Physiological Zoology 65, 188–214.
Hulbert, A.J., Else, P.L., 2000. Mechanisms underlying the cost of living in animals. Annual Review of Physiology 62, 207–235.
Kam, M., Degen, A.A., Nagy, K.A., 1987. Seasonal energy, water, and food consumption of Negev chukars and sand partridges. Ecology 68, 1029–1037.
Kemp, T.S., 2006. The origin of mammalian endothermy: A paradigm for the evolution of complex biological structure. Zoological Journal of the Linnean Society 147, 473–488.
Kooyman, G.L., Cherel, Y., Le Maho, Y., *et al*., 1992. Diving behavior and energetics during foraging cycles in king penguins. Ecological Monographs 62, 143–163.
Koteja, P., 2000. Energy assimilation, parental care and the evolution of endothermy. Proceedings of the Royal Society—Biological Sciences (Series B) 267, 479–484.
Labocha, M.K., Sadowska, E.T., Baliga, K., Semer, A.K., Koteja, P., 2004. Individual variation and repeatability of basal metabolism in the bank vole, Clethrionomys glareolus. Proceedings of the Royal Society B 271, 367–372.
McNab, B.K., 1969. The economics of temperature regulation in neotropical bats. Comparative Biochemistry and Physiology 31, 227–268.
McNab, B.K., 1989. Basal rate of metabolism, body size and food habits in order Carnivora. In: Gittleman, J.L. (Ed.), Carnivore behavior, ecology, and evolution. Ithaca, NY: Cornell University Press, pp. 335–354.
McNab, B.K., 2002. The physiological ecology of vertebrates: A view from energetics, chs. 4 and 5. Ithaca, NY: Cornell University Press.
Nagy, K.A., 1994. Field bioenergetics of mammals: What determines field metabolic rates? Australian Journal of Zoology 42, 43–53.
Nagy, K.A., Knight, M.H., 1989. Comparative field energetics of a Kalahari skink (Mabuya striata) and gecko (Pachydactylus bibroni). Copeia. 1), 13–17.
Peterson, C.C., Walton, B.M., Bennett, A.F., 1998. Intrapopulation variation in ecological energetics of the garter snake Thamnophis sirtalis, with analysis of the precision of doubly labeled water measurements. Physiological Zoology 71, 333–349.
Pough, F.H., Busack, S.D., 1978. Metabolism and activity of the Spanish fringe-toed lizard (Lacertidae: Acanthodactylus erythurus). Journal of Thermal Biology 3, 203–205.
Renecker, L.A., Hudson, R.J., 1986. Seasonal energy expenditures and thermoregulatory responses of moose. Canadian Journal of Zoology 64, 322–327.
Ruben, J., 1995. The evolution of endothermy in mammals and birds: From physiology to fossils. Annual Review of Physiology 57, 69–95.
Schweitzer, M.H., Marshall, C.L., 2001. A molecular model for the evolution of endothermy in the theropod-bird lineage. Journal of Experimental Zoology 291, 317–338.
Seymour, R.S., Schultze-Motel, P., 1997. Heat-producing flowers. Endeavor 21, 125–129.
Shkolnik, A., Schmidt-Nielsen, K., 1976. Temperature regulation in hedgehogs from temperate and desert environments. Physiological Zoology 49, 56–64.

Southwood, A.L., Reina, R.D., Jones, V.S., Speakman, J.R., Jones, D.R., 2006. Sesonal metabolism of juvenile green turtles (Chaledonia mydas) at Heron Island, Australia. Canadian Journal of Zoology 84, 125–135.
Sparti, A., 1990. Comparative temperature regulaton of African and European shrews. Comparative Biochemistry and Physiology A 97, 391–397.
Sparti, A., 1992. Thermogenic capacity of shrews (Mammalia, Soricidae) and its relationship with basal rate of metabolism. Physiological Zoology 65, 77–96.
Tatner, P., Bryant, D.M., 1993. Interspecific variation in daily energy expenditure during avian incubation. Journal of Zoology 231, 215–232.
Werner, J., Sfakianakis, N., Rendall, A.D., Griebeler, E.M., 2018. Energy intake functions and energy budgets of ectotherms and endotherms derived from their ontogenetic growth in body mass and timing of sexual maturation. Journal of Theoretical Biology 444, 83–92.
Wesley, D.E., Knox, K.L., Nagy, J.G., 1973. Energy metabolism of pronghorn antelopes. Journal of Wildlife Management 37, 563–573.
Withers, P.C., Thompson, G.G., Seymour, R.S., 2000. Metabolic physiology of the north-western marsupial mole, Notoryctes caurinus (Marsupialia: Notoryctidae). Australian Journal of Zoology 48, 241–258.

Epiflora and Epifauna☆

Richard B Taylor, University of Auckland, Auckland, New Zealand

Introduction and Definitions

The terms epiflora and epifauna (collectively "epibiota") are usually used to describe nonparasitic organisms living on the surfaces of other organisms (see examples in **Fig. 1**). Some epibiota have an intimate relationship with their host, but for many or perhaps even most species the host is just a surface to which they can attach, shelter from consumers, or gain better access to light or nutrients. These species typically show little host-specificity and may also occur on nonliving surfaces.

Numerous terms have been used to describe surface-dwelling organisms and their hosts (Box 1). Epiphyte, the most familiar of these, usually denotes a terrestrial plant living on another plant, but has also been used in the aquatic literature to describe plants on plants, and animals on plants. Here epiphyte and epiflora are used interchangeably to refer to plants living on other organisms. For aquatic epifauna the term has been broadened to include animals inhabiting nonliving surfaces (but not those forming large-scale habitat like coral reefs). Microorganisms (with the exception of periphyton) and terrestrial animals are not considered here.

Epiflora

Terrestrial epiphytes comprise a taxonomically and morphologically diverse range of relatively large vascular plants (notably bromeliads, orchids, and ferns), and small nonvascular bryophytes (mosses and liverworts), lichens, and free-living algae. Growth forms include creepers, mats, brackets, nests, pendants, and shrubs. Most terrestrial epiphytes are self-contained on the bark or branches of their host, except for hemiepiphytes, which have roots reaching down to the ground. Not considered here are the mistletoes, which are parasitic, and the vines or lianas, which although using other plants for physical support do not have a crown attached to the host.

Aquatic epiflora are composed of red, green, and brown algae, and various other photosynthetic microorganisms. In freshwaters, microscopic, mostly unicellular, periphyton (or "aufwuchs") dominate, while in the ocean macroalgae (seaweeds) are also important. Periphyton grows in thin films, while macroalgal epiphytes take encrusting, sheetlike, or branching forms, the latter two usually attached to their host by a small discoid holdfast or penetrative rhizoids.

Epifauna

Aquatic epifauna can be categorized by the sieve mesh size they are retained on (meiofauna 0.063–0.5 or 1 mm; macrofauna >0.5 or 1 mm), and by whether they are mobile or sessile (attached). Freshwater macrophytes are inhabited by meiofaunal nematodes and macrofaunal gastropods, insect larvae and oligochaete worms (all mobile). Estuarine and marine macrophytes are inhabited by meiofaunal nematodes and harpacticoid copepods, and macrofaunal gastropods, polychaete worms, and amphipod and isopod crustaceans (all mobile), along with sessile hydroids, bryozoans and tube-building polychaetes. Larger marine sessile animals are host to crustaceans (shrimps, crabs, amphipods, isopods), gastropods, pycnogonids, and brittlestars (all mobile). Sessile animals such as sponges and bryozoans may also be present. Animals living on lake and river beds include most of the taxa listed above for freshwater macrophytes, along with crayfish, mysid shrimps, and fishes (all mobile) and bivalves (sessile). The mobile fauna of nonliving marine surfaces includes gastropods, echinoderms (starfishes, brittlestars, sea urchins, sea cucumbers), crustaceans (crabs, lobsters, shrimps, amphipods, isopods, cumaceans), and fishes. The sessile fauna of hard substrata includes ascidians, brachiopods, bryozoans, crustaceans (barnacles), cnidarians (hard and soft corals, sea anemones, gorgonians, hydroids), echinoderms (brittlestars, crinoids, sea cucumbers), tube-building polychaetes, and sponges. On soft sediments common sessile epifaunal taxa include bivalves and sponges, many of which actually grow attached to shells (alive or dead) or cobbles. There is a general trend for the maximum body size of epifaunal individuals to increase from aquatic macrophytes (<1 cm) to sessile invertebrates (several cm) to nonliving surfaces (10s of cm).

Types of Hosts

On land the major hosts for epiphytes are woody vascular plants: the trees and shrubs, with animals (e.g., sloths) a very minor habitat. In freshwaters the main living hosts for both epiflora and epifauna are vascular plants, with bryophytes (mosses and

☆*Change History:* October 2017. Richard B Taylor updated the Further Reading section and Fig. 1, and added "Life Histories" and "Threats" sections (including Fig. 3).

This is an update of R.B. Taylor, Epifauna and Epiflora, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 1389–1393.

Fig. 1 Representative epibiota, using New Zealand examples. (A) *Notheia anomala*, a brown alga occurring only on the intertidal brown seaweed *Hormosira banksii*, (B) *Collospermum hastatum*, a lily growing on the tree *Metrosideros excelsa*, (C) *Luidia australiae*, an asteroid echinoderm (starfish), on a sandy seafloor, (D) *Protohyale rubra*, an amphipod crustacean, and filamentous red algae on the green seaweed *Codium fragile*. White bars represent 10 mm (A,D), 100 mm (C) or 1 m (B). Photos copyright Richard Taylor.

Box 1 Definitions

Aufwuchs. Microscopic organisms forming a film on aquatic surfaces
Epi-. Living on the surface of
Epibenthos. Organisms living on the surface of a sea-, lake- or river-bed
Epibiota. Organisms living on any surface
Epifauna. Animals living on an aquatic surface
Epiflora. Plants or algae living nonparasitically on surfaces of other organisms
Epilithic. Growing on rock
Epiphyte. Plant or algae living nonparasitically on the surface of another organism
Epizoon. Organism living on the external surface of an animal
Hemiepiphyte. Epiphyte with roots reaching down to the ground
Holoepiphyte. Epiphyte living its entire life cycle as an epiphyte
Liana. Vine rooted in the ground
Macrophyte. Macroscopic aquatic plant or alga
Periphyton. Microscopic organisms forming a film on aquatic surfaces
Phorophyte. Terrestrial plant hosting epiflora

liverworts) and macroalgae less important. In the oceans, the hosts are seaweeds (macroalgae), and seagrasses, mangroves and saltmarshes (all vascular plants), the latter two found only along sheltered coastal and estuarine margins. Marine epiflora also occur on many aquatic animals, both sessile (e.g., sponges, corals) and mobile (e.g., gastropods, turtles). In addition to macrophytes, aquatic epifauna inhabit numerous sessile animals (sponges, bivalves, tunicates, corals, gorgonians, hydroids, bryozoans), and some mobile animals. As noted in the introduction, animals living on the surfaces of marine and estuarine soft sediments (mud, sand, and gravel) and hard bottoms (rocks, wood, and artificial structures such as jetties and shipwrecks) are often termed epifauna, and those inhabiting similar freshwater habitats are also included here for completeness.

Host Attributes

Total abundances, total biomasses and maximum sizes of epibiota individuals tend to increase with host age, size, and structural complexity, for both plant and animal hosts. Epiphyte distributions are often strongly stratified within individual hosts.

Terrestrial host plants tend to be large, rigid, and may live hundreds or even thousands of years, while most aquatic host plants are small, flexible, and live for a decade at most. These factors, along with inherent phylogenetic constraints and the high drag resulting from the greater density of water as a medium, limit the maximum age and size attained by aquatic epiphytes and sessile epifauna. While terrestrial epiphytes can weigh hundreds of kilograms, aquatic epiphytes usually weigh less than a few grams.

Host-Specificity

Restriction to a single host species is rare among epibiota, with many species found on multiple host species or even nonliving surfaces. For instance, many terrestrial epiphytes also grow at ground level on soil or rocks, and most mobile epifaunal taxa found on aquatic macrophytes will also colonize artificial plants, suggesting that for those epibiota there is nothing intrinsically special about the host. Obligate host-specificity is probably commonest among epibiota having a tight mutualistic relationship with their host, e.g., coral-dwelling shrimps and crabs that derive shelter and eat mucous released by their host in exchange for protecting it from crown-of-thorns starfish and other predators. There is little evidence for coevolution of such relationships.

Environmental Gradients

Terrestrial epiphytes are most diverse and abundant in neotropical forests that experience year-round humidity and no frosts. Diversity is lower in the palaeotropics (possibly due to historical aridity), and decreases at higher altitudes and latitudes, but epiphytes can still be important in temperate forests with oceanic climates. Aquatic epiflora are found virtually everywhere their host macrophytes exist, i.e., on sunlit surfaces free of sessile animals or destructive grazers. Epiflora are often abundant in eutrophic waters.

Epifauna occur in all aquatic environments: in lakes and rivers, from the intertidal to ocean trenches, and from tropical to polar latitudes. Densities are usually highest where there is strong net water movement supplying food for filter-feeders. Strong light supports algal growth, which benefits mobile epifauna that eat periphyton, but not sessile epifauna, which are often susceptible to being overgrown. On soft sediments, epifaunal densities tend to decline with increasing depth due to the reduction in food input from the sunlit surface waters, though densities can be locally high where detritus accumulates. Sessile epifauna are often delicate and long-lived, so are vulnerable to physical disturbances like dredging and trawling.

Acquisition of Resources

Solid Substratum

Structures vary greatly in their suitability as a surface to grasp or stick to, with solidity, shape, inclination, aspect, texture, and sediment- and water-holding ability among the important characteristics for epibiota.

Light

In dense forests light at ground level is $< 2\%$ of that falling on the canopy, so epiphytism is the only way that small plants can gain access to strong light for photosynthesis. However, not all terrestrial epiphytes require full sun and some are tolerant of deep shade. Light levels are also greatly reduced under dense aquatic macrophyte beds, but for epiphytes there tends to be less vertical light stratification within hosts than that experienced by their terrestrial counterparts, due to the flexibility of macrophytes subject to water motion.

Water

In the canopies of many forests, precipitation is sporadic and exposure to the wind and sun creates drier conditions than on the forest floor, so water availability strongly affects the growth and survivorship of epiphytes lacking a root connection to the ground. Epiphytes maximize water uptake and minimize loss by (1) rapidly absorbing rain, dew or mist using specialized leaf and root structures, (2) storing water when it is plentiful (in their own tissues, root-mass humus, or "tanks" in the case of bromeliads), and (3) using CAM photosynthesis to reduce losses through transpiration. Many can tolerate severe desiccation. Water availability is obviously not a factor for aquatic epiphytes, except those exposed by low tide or drought.

Nutrients/Food

Nutrients are often in short supply for terrestrial epiphytes, since most have no root connection to the ground and must instead obtain their nutrients from solutes in rain and stemflow water, wind-borne particles, litterfall, carnivory, or animal excrement (especially the waste products of ants living in the root mass or within specialized plant organs). For aquatic organisms, living on surfaces can increase their access to nutrients/food in a number of ways. (1) Organisms obtaining nutrients from the water column, either as ions in solution (epiphytes) or suspended particles (many epifauna), should profit from the increased access to

water flow achieved by living on other organisms that project into the water column, or in the case of epifauna living directly on the seafloor, inhabiting the surface of sediments or rocks rather than living in burrows. Aquatic epiphytes can take up nutrients from the surrounding water across their entire surfaces, so they generally have better access to nutrients than do terrestrial epiphytes. They usually have higher surface: area volume ratios than their host macrophytes, enabling them to take up nutrients more rapidly, but like their hosts they are still potentially limited by C, N, or P. Other food sources available to epifauna are (2) periphyton, for which host surface area is important, (3) trapped detritus, for which host structure is important, and (4) host macrophyte tissue, for which intrinsic host properties such as nutritional value and chemical defenses are important.

Refuge From Consumers

Living in a tree or aquatic macrophyte reduces the threat to epibiota from ground- or seafloor-dwelling consumers. Predation is intense in many aquatic habitats, but epibiota can acquire critical refugia on hosts possessing (1) structural complexity, which reduces the foraging efficiency of predators, and/or (2) chemical defenses against consumers, which provide epibiota with "enemy-free space" where they are safe from incidental consumption or destruction of their host (an "associational defense"). For herbivorous seaweed epifauna, shelter from predators often appears to outweigh food quality as a factor influencing host selection or subsequent survivorship.

Contribution to Ecosystem Diversity, Biomass and Productivity

Epiflora

Worldwide about 27,000 species or 9% of vascular plants are epiphytes, with the proportion up to 50% in tropical forests. About two-thirds of vascular epiphyte species are orchids. Similar figures are unavailable for nonvascular epiphytes. The green biomass of epiphytes may match that of their host trees, though epiphyte productivity is likely to be lower due to resource limitation (see above). In moist tropical forests the habitat complexity, water and food provided by epiphytes is a major contributor to the high diversity of insects and other animals.

The productivity of aquatic epiphytes often exceeds that of their host macrophytes, though in some habitats (e.g., seagrass beds) the dominance of epiphytes may be a recent artifact of anthropogenic eutrophication or the overharvesting of vertebrate herbivores (sirenia, turtles) that once cropped macrophytes before they could be colonized by epiphytes. In aquatic systems, epiflora are usually more suited to direct consumption by herbivores than their vascular plant hosts, which tend to be tougher, and contain less nitrogen than epiflora but more indigestible fiber and secondary metabolites like phenolics.

Epifauna

It is difficult to compare epifaunal diversities among habitats because of variation among studies in the size range of organisms examined and in the number of hosts or area of seafloor sampled. However, some generalizations are possible. For instance, marine epifauna are far more taxonomically diverse than freshwater epifauna, and complex structures host more species than simple ones. In the oceans, species of macrophytes and sessile animals host up to several hundred epifaunal species.

Small organisms are more metabolically active than larger ones (they have higher mass-specific metabolic rates and shorter generation times), and will likely play increasingly greater roles in aquatic ecosystems as larger organisms are progressively removed by fishing. On New Zealand rocky reefs, small (0.5–10 mm) mobile epifauna contribute about 80% of total secondary (animal) production, and probably make similar contributions to total food consumption and nutrient regeneration by animals.

Small mobile epifauna are important trophic links between fishes and primary producers, while suspension-feeding epifauna link pelagic and benthic ecosystems by accelerating the flow of organic matter from the water column to the seafloor.

Interactions

As shown by the aquatic example in Fig. 2, there are numerous ways in which the host, its epiflora and epifauna can potentially interact. Some of the more important interactions are as follows. Sessile epibiota reduce their host's access to light and nutrients, and their additional mass can cause tree branches to break off or aquatic macrophytes to be torn loose by water movement. In an apparent response, some hosts actively discourage epibiota by allelopathy (the release of settlement- or growth-inhibiting chemicals) or by the shedding/sloughing of heavily fouled tissues. There is ample scope for indirect effects, the most important of which is probably the favor that many herbivorous epifauna do their host macrophyte by removing fouling epiphytes – some seaweeds are rapidly smothered by epiphytes in the absence of certain epifaunal grazers. On the other hand, direct grazing by epifauna can seriously damage the host macrophyte, especially when vulnerable tissues or microscopic life stages are targeted.

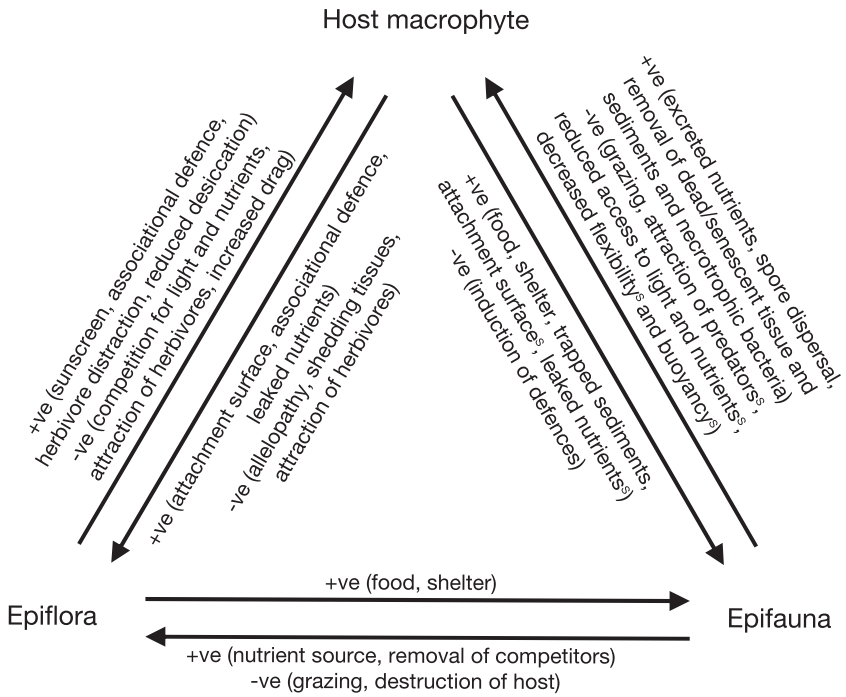


Fig. 2 Summary of potential interactions between aquatic macrophytes, epiflora, and epifauna. A positive effect of one group of organisms on another is denoted by "+ve", and a negative effect by "-ve". Superscripted "S" refers to sessile epifauna.

Trophic Cascades: The Importance of Predators

In aquatic ecosystems, the impact of grazing epifauna on both host macrophytes and their epiflora is often moderated by small predatory fishes (a so-called "trophic cascade").

Several mesocosm experiments have found that fishes prevent epifaunal overgrazing of seaweeds. Interestingly, for seaweed epifaunal taxa feeding on periphyton the main effect of fish predation is to remove the larger epifaunal individuals, which actually results in higher total numbers of epifauna due to the freeing up of the limiting periphyton resource for many more small individuals.

In freshwater macrophyte beds, epifauna appear to have a much greater impact on periphyton than on the host, so by reducing epifaunal densities fish promote periphyton growth and thus decrease host vigor. Fish predation on epifauna may ultimately shift shallow lakes from macrophyte- to phytoplankton-dominated ecosystems.

Light and nutrients also strongly affect primary productivity in aquatic ecosystems, but it is not yet possible to generalize about the relative importances of bottom-up (resource) and top-down (consumer) controls on community structure.

Life Histories

The taxonomic diversity of epibiota is reflected in their wide range of life histories. Some species have complex life cycles with phases that look like they belong to different species (Fig. 3). Epibiota face particular life history challenges. Suitable host species may be short-lived or patchily distributed, making it difficult for individual epibiota to find hosts, grow to maturity and locate mates. If the host has a strongly seasonal life cycle then specialized obligate epibiota need to synchronize their own life cycle accordingly. In common with nonepibiota, tradeoffs in life history traits are to be expected. For example, should offspring be dispersed widely in the hope of finding a vacant settlement site, or settle next to the parent in what may be an overcrowded or deteriorating habitat (such as a rotten tree branch)? Although traits such as self-compatibility may be more frequent in certain epibiota species than related nonepibiota, life histories of epibiota generally reflect their phylogeny much more strongly than their lifestyle, i.e., there is no particular life history mode that is characteristic of epibiota as a whole.

Threats

Human activities affect epibiota in a variety of ways. Like other organisms, epibiota are and will be affected by climate change directly through changing temperature, precipitation (on land) and acidification (in water), and indirectly through altered species interactions. The diverse epiphyte communities in tropical cloud forests are particularly vulnerable to climate change due to their

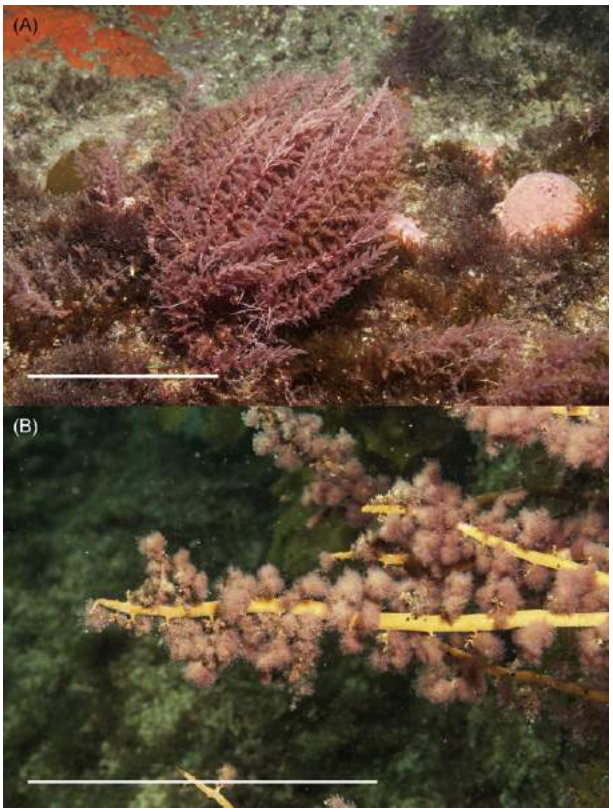


Fig. 3 Alternating life stages of the red alga *Asparagopsis armata*. (A) gametophyte, growing on turfing algae, (B) tetrasporophyte, growing on the brown seaweed *Carpophyllum plumosum*. Both life stages occur mainly as epiphytes. The gametophyte is haploid and dioecious. Sperm from males fertilizes eggs within a carposporangium on the female. The resultant diploid carpospores are released, and germinate into tetrasporophytes. When mature these develop haploid tetraspores via meiosis, which are released and germinate into gametophytes to complete the life cycle. The gametophyte is present in spring and early summer, while the tetrasporophyte is present year-round. Before the life cycle was worked out the tetrasporophyte was assigned to a separate genus and species. White bars represent 100 mm. Photos copyright Richard Taylor.

dependence on atmospheric moisture. Habitat modification, such as the destruction or fragmentation of natural forests, has varying effects on epiphytes. Some species are able to establish in plantation forests, while opening up the forest to wind through fragmentation favors epiphytes that tolerate a drier microclimate, at the expense of species that cannot. In the sea, the effects of dredging and trawling on seafloor epifauna have been compared to the clear-felling of forests on land. Air pollution affects many terrestrial epiphytes, and lichens are even used as indicator species in this regard. In aquatic systems, pollutants such as fine sediments and nutrients have greatly altered many ecosystems. Interestingly, eutrophication (excess nutrients) sometimes favors algal epiphytes over macrophytes, as the high surface area: volume ratio of microalgae and filamentous macroalgae enables them to take up nutrients more rapidly. Many terrestrial epiphytes are heavily harvested for medicinal or ornamental purposes.

Further Reading

Benzing, D.H., 1990. Vascular epiphytes: General biology and related biota. Cambridge: Cambridge University Press.
Brönmark, C., 1989. Interactions between epiphytes, macrophytes and freshwater snails: A review. Journal of Molluscan Studies 55, 299–311.
Dickson, L.G., Waaland, J.R., 1985. *Porphyra nereocystis*: A dual-daylength seaweed. Planta 165, 548–553.
Edgar, G.J., Moore, P.G., 1986. Macro-algae as habitats for motile macrofauna. Monographiae Biologicae 4, 255–277.
Gili, J.-M., Coma, R., 1998. Benthic suspension feeders: Their paramount role in littoral marine foodwebs. Trends in Ecology and Evolution 13, 316–321.
Hay, M.E., Parker, J.D., Burkepile, D.E., Caudill, C.C., Wilson, A.E., Hallinan, Z.P., Chequer, A.D., 2004. Mutualisms and aquatic community structure: The enemy of my enemy is my friend. Annual Review of Ecology and Systematics 35, 175–197.
Hayward, P.J., 1988. Animals on seaweed. Richmond: Richmond Publishing Co. Ltd.
John, D.M., Hawkins, S.J., Price, J.H. (Eds.), 1992. Plant-animal interactions in the marine benthos. Oxford: Clarendon Press.
Lowman, M.D., Nadkarni, N.M. (Eds.), 1995. Forest canopies. San Diego: Academic Press.
Lüttge, U. (Ed.), 1989. Vascular plants as epiphytes: Evolution and ecophysiology. Berlin: Springer-Verlag.
Mondragón, D., Valverde, T., Hernández-Apolinar, M., 2015. Population ecology of epiphytic angiosperms: A review. Tropical Ecology 56, 1–39.
Nadkarni, N.M., Merwin, M.C., Nieder, J., 2001. Forest canopies: Plant diversity. In: Levin, S. (Ed.), Encyclopedia of biodiversity. San Diego: Academic Press, pp. 27–40.
van Montfrans, J., Wetzel, R.L., Orth, R.J., 1984. Epiphyte-grazer relationships in seagrass meadows: Consequences for seagrass growth and production. Estuaries 7, 289–309.
Zotz, G., 2016. Plants on plants—The biology of vascular epiphytes. Switzerland: Springer.

Generation Time☆

Enric Cortés, National Oceanographic and Atmospheric Administration, Panama City, FL, United States
Gregor M Cailliet, Moss Landing Marine Laboratories, Moss Landing, CA, United States

Glossary

Cohort In general a collection of individuals that share some trait(s). When referring to generation time, a group of individuals born in the same year (or other unit of time).

Intrinsic rate of population growth The maximum rate at which a population can increase with no resource limitation. It is a constant expressed in year^{-1}. Also known as maximum rate of population increase (r_{max}), Malthusian parameter, or per capita rate of growth.

Life table A schedule of survival probabilities and fecundity, typically at age, that is used to calculate the intrinsic rate of population growth and other demographic parameters, such as generation time.

Projection matrix A mathematical model—a matrix—that projects the population from time t to time $t+1$ assuming constant schedules of survival and reproduction at age (or stage or size). Analogous to a life table.

Stable age distribution A principle of ecology by which the proportions of individuals belonging to different age groups will remain constant generation after generation when the population reproduces in a constant environment with no resource limitation.

Definition of Generation Time

The estimation of generation time, or generation length, is an important part of the demographic analysis of populations because it is a measure of the time required for a newborn to become a parent and this influences a population's potential for growth. The word "generation" is generally defined as the average age at which a female from any given population gives birth to her first female offspring, or the average age of mothers in a population. It is therefore equivalent to the time that it takes a population to increase by a factor equal to the net reproductive output of those females.

There are several definitions of the term generation and ways of calculating its length (generation time) for age-structured populations. The first is to calculate the time T required for a population to increase by a factor of R_0 (its net reproductive rate). Two other definitions include (1) the mean age (μ_1) of the females bearing offspring produced by a cohort over its lifetime (not requiring a stable age distribution) and (2) the mean age ($\overline{A}$) of the females bearing pups produced by a population at the stable age distribution—for this last one, in a stationary population with $\lambda = 1.0$, $\mu_1 = \overline{A}$, where λ is the finite rate of population growth.

Calculation of Generation Time

While the concept of generation and generation time (or length) has been an important, but varying, concept over the years, ways of calculating it have also varied. Generation times can be calculated either through the use of life tables or projection matrices. Early authors defined generation time as the mean length of a generation (or the average time from the birth of a mother to the birth of her first female offspring), and used it as T in the exponential growth equation $N_t = N_0 e^{rT}$. Then, from their definition of the net reproduction rate as $N_t/N_0 = R_0$, they solved for $r = \ln R_0/T$ and $T = \ln R_0/r$. But, it was then pointed out that the mean length of a generation cannot be obtained until the best value of r, the intrinsic rate of population growth, is obtained, which is only possible after iteratively solving the Euler–Lotka equation, which in its discrete form is expressed as $\sum l_x m_x e^{-rx} = 1$,where l_x is the probability of a female being alive at the beginning of age x and m_x is the number of female offspring produced by a female of age x. This ultimately led to the other two expressions of generation time, μ_1 and $\overline{A}$. The first, μ_1, is equal to $\sum x l_x m_x / \sum l_x m_x$, whereas the second, $\overline{A}$, is expressed as $\sum x l_x m_x e^{-rx}$. Both of these equations have commonly been used in many papers and textbooks to explain how to calculate generation times. The above calculations of generation time (and other parameters) involve the creation of life tables—a more classic approach to demography. A relatively more contemporary approach to demography involves the use of Leslie (age-based) or Lefkovitch (stage-based) matrix models.

Demographic analyses are useful for understanding the population dynamics of many organisms and their likely response to exploitation. The role of generation time in these analyses is crucial in understanding the interaction between age- or stage-specific survivorship and reproductive-output functions. Such approaches have been used to produce estimates of generation time for many types of organisms including algae, flowering plants, invertebrates, fishes, reptiles, birds, and mammals.

☆*Change History:* February 2018. E Cortés and GM Cailliet updated Abstract, updated, extended, and added sections to the text, extended the Further Reading, and modified Figure 1.

This is an update of G.M. Cailliet, Generation Time, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 1734–1736.

Ecological Considerations

Generation time is positively correlated with body size (see Fig. 1), leading some authors to categorize this as a "generation-time law." The mathematical relationship is that the generation time increases with body size at a power of approximately one-fourth of the body mass, a pattern which appears to hold across taxa. Body size (measured as weight or length) in a variety of organisms is also positively related to reproductive output (e.g., clutch weight or volume, number of eggs per female parent), hatchling weight (egg or other product), gestation (or brood) time, maturation time, and maximum life span. The intrinsic rate of population growth appears to decrease with body size across taxa following the same allometric relationship as generation time, in other words it decreases with body size at a power of approximately one-fourth of the body mass. However, in some taxa it can be either negatively or not related to body size and generation time.

The relationship between generation time and the intrinsic rate of increase with body mass led some authors to define the intrinsic rate per generation (rT), which is a dimensionless quantity because r is measured in years^{-1} and T is measured in years, such that: $rT = a_r M^{-0.25} a_t M^{0.25} = a_r a_t M^0 = a_r a_t$, where M is body mass and a_r and a_t are the constants in the allometric equations for intrinsic population growth rate and generation time, respectively. The intrinsic rate per generation is related to an important reference point in population dynamics of exploited populations, the position of the inflection point of population growth curves, which is a measure of where "maximum sustainable yield" (MSY) or "Catch" can be attained. The relationship between the intrinsic rate per generation and the position of the inflection point of population growth curves is independent of body size. These relationships, often employing size at maturation or first reproduction rather than actual generation times, have also been employed to better understand how the life-history traits of many organisms have evolved.

The intrinsic rate per generation, or product of the allometric constants, has been shown to be approximately equal to 1 in several taxa. This led to two simplified definitions of generation time under ideal conditions and assuming constant adult survival probability (s) and fecundity after the age at first reproduction (α): $\overline{T}_{op} = \alpha + {}^{s}/_{e^r} - s$ and $\overline{T}_{op} \approx {}^{1}/_{e^r} - 1$.

The Use of Generation Time in Conservation and Fisheries

The concept of "generation time" has been used extensively by the International Union for the Conservation of Nature (IUCN) as part of their process of categorizing the status of species and their populations worldwide. They state "Generation length is the

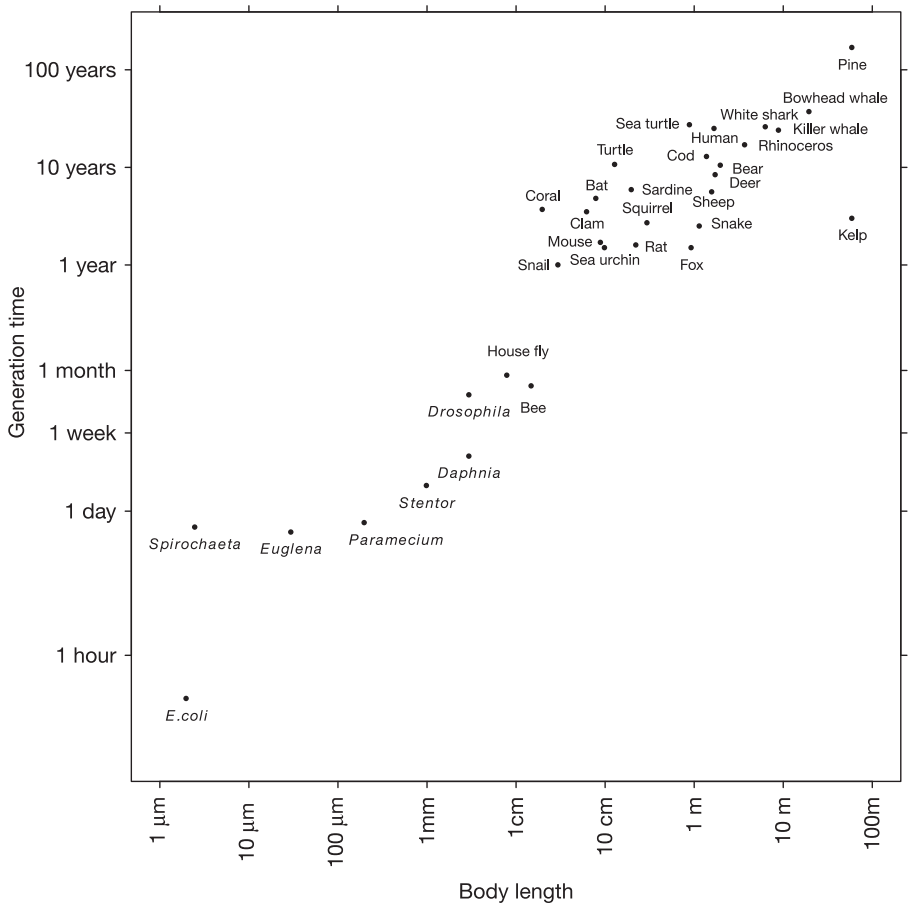


Fig. 1 The relationship between body size and generation time for several groups of living organisms.

average age of parents of the current cohort (i.e., newborn individuals in the population). Generation length therefore reflects the turnover rate of breeding individuals in a population. Generation length is greater than the age at first breeding and less than the age of the oldest breeding individual, except in taxa that breed only once." Generation length is used to categorize the conservation status of species when determining whether they are declining by a certain percentage in abundance "over the last 10 years or three generations, whichever is longer." This enables the IUCN to decide whether a species should be placed on the Red List and categorized as "critically endangered" (CR $\geq$ 90% reduction), "endangered" (EN $\geq$ 70% reduction), or "vulnerable" (VU $\geq$ 50% reduction). Therefore, in this process, the accuracy of both the estimates of population abundance and generation time estimates is extremely important.

Because estimates of age-specific survival and fecundity are often difficult to obtain, the IUCN recommends using two proxies—which are simplifications like the equations for T_{op} we saw in the previous section—to calculate generation time. The first equation assumes constant adult survival and fecundity and infinite lifespan and simplifies to: $T = \alpha + {}^{1}/_{M}$ where M is the instantaneous rate of natural mortality. The second equation is: $\hat{T}_z = \alpha + z(\omega - \alpha)$ where ω is lifespan and therefore $\omega - \alpha$ is a measure of reproductive lifespan, and z is a constant that "depends on survivorship and the relative fecundity of young vs. old individuals in the population." However, it has been found that these proxies can introduce bias and it is preferable to use the traditional definitions of generation time, highlighting the importance of collecting the detailed life history information needed for their computation.

Generation time also comes into play when calculating rebuilding times for overexploited species in fisheries, both in the United States and worldwide. Under precautionary approaches, scientists at the National Marine Fisheries Service (NMFS) of the United States Department of Commerce first calculate the rebuilding time (time it takes for the stock to reach the biomass that produces MSY under zero fishing mortality) and if that rebuilding time exceeds 10 years, then they must add the generation time to it to estimate a new rebuilding period.

See also: General Ecology: Growth Models

Further Reading

Bonner, J.T., 1965. Size and cycles: An essay on the structure of biology. Princeton, NJ: Princeton University Press.
Caswell, H., 2001. Matrix population models. Construction, analysis, and interpretation, 2nd ed. Sunderland, MA: Sinauer Associates.
Caughley, G., 1977. Analysis of vertebrate populations. London: Wiley Press.
Coale, A.J., 1972. The growth and structure of human populations: A mathematical investigation. Princeton, NJ: Princeton University Press.
Ebert, T.A., 1999. Plant and animal populations: Methods in demography. San Diego, CA: Academic Press.
Fowler, C.W., 1988. Population dynamics as related to rate of increase per generation. Evolutionary Ecology 2, 197–204.
Fung, H.C., Waples, R.S., 2017. Performance of IUCN proxies for generation length. Conservation Biology. doi:10.1111/cobi.12901.
IUCN. (2012). IUCN Red List Categories and Criteria: Version 3.1 (2nd Ed.). Gland, Switzerland/Cambridge, UK: IUCN. iv + 32 pp.
Heppell, S.S., Crowder, L.B., Menzel, T.R., 1999. Life table analysis of long-lived marine species with implications for conservation and management. American Fisheries Society Symposium 23, 137–148.
Mertz, D.B., 1970. Notes on methods used in life-history studies. In: Connell, J.H., Mertz, D.B., Murdoch, W.W. (Eds.), Readings in ecology and ecological genetics. New York: Harper and Row, pp. 4–17.
Niel, C., Lebreton, J.D., 2005. Using demographic invariants to detect overharvested bird populations from incomplete data. Conservation Biology 19, 826–835.
Restrepo, V. R., Thompson G. G. and Mace, P. M. *et al.* (1998). Technical guidance on the use of precautionary approaches to implementing National Standard 1 of the Magnuson-Stevens fishery conservation and management act. NOAA technical memo. NMFSF/SPO-31. Springfield, VA: National Technical Information Center.
Stearns, S.C., 1976. Life-history tactics: A review of the ideas. Quarterly Review of Biology 51, 3–47.
Wilson, E.O., Bossert, H.W., 1981. A primer of population biology. Sunderland, MA: Sinauer Associates.

Relevant Websites

http://www.esf.edu/species/—International Institute for Species Exploration.
http://www.iucnredlist.org/—IUCN Red List.
http://www.nmfs.noaa.gov/sfa/laws_policies/msa/—Magnuson-Stevens Fisheries Conservation and Management Act.

Growth Constraints: Michaelis–Menten Equation and Liebig's Law☆

Sven E Jørgensen, Copenhagen University, Copenhagen, Denmark

Glossary

Enzyme kinetics The study of the enzyme-catalyzed reactions that depends on the concentration of the compounds directly interacting with the enzyme and the highest rate achievable by the enzyme that are catalyzed by enzymes.

Rate equation Rate equation is a mathematical formula that shows the effect of changing the concentrations of the reactants on the rate of the reaction.

Reaction rate constant (or coefficient) Reaction rate constant quantifies the rate of a chemical reaction. The value depends on changes on conditions such as temperature or the catalyst.

Substrate The chemical species of interest that is being modified.

Threshold agents The agents that, in a suitable concentrations, promote growth. Nonthreshold agents, in contrast have a negative or no effect on the growth.

What Stops Growth? Liebig's Law

Considering the generalized growth and development phenomenon, we have to ask the question: Why does it not continue indefinitely? What stops growth? Why do the trees not grow to the sky, bacteria not fill the oceans, etc.? The answer is that the elements needed to build the biomass of various species are present only in a limiting amount in the environment/the ecosystems. Liebig's law considers these interactions between the organisms and their growth on the one side and the concentrations of nutrients on the other side. The elementary composition varies from plant to plant and from organism to organism and even the same species may have a different composition in different environments and at different times of the year. There are, however, some basic biochemical processes that require a combination of elements in a certain ratio.

The biochemistry and the biochemical reactions for all organisms on Earth are very similar, although they vary slightly from organism to organism. About 20 elements are necessary to build an organism, sometimes with the addition of one to five more elements. **Table 1** shows an example of the elementary composition of freshwater plants. The relative quantities of the 19 essential elements in plant tissue are shown. The composition is not strictly stoichiometric, as is that of a chemical compound, but may vary between lower and upper values following the composition of different biochemical compounds in the cells. For instance, phytoplankton can contain approximately 0.5%–2.5% of phosphorus and 5%–12% of nitrogen by dry weight.

The required elementary composition of an organism reflects the chemical constraints for growth. Growth implies that the environment delivers the elements—for plants as suitable compounds dissolved in water in contact with the plants, and for animals by the composition of available food. The environment rarely has exactly the same composition as required for growth, which implies that the element less abundant in the environment compared with the need determines the limits to growth. This is expressed in Liebig's law of the minimum (see **Fig. 1**); if a nutrient is at a minimum relative to its use for growth, there is a linear relation between growth and the concentration of the nutrient. If the supply of other factors is at a minimum, further addition of the nutrient in question will not, as shown, influence growth.

Growth and the Limiting Factor(s)

The relationship between growth $v = dw/dt$ (w is weight) and the concentration of the limiting factor is described by the Michaelis–Menten equation, which is also used to give the relation between the concentration of a substrate and the rate of a biochemical reaction:

$$dw/dt = v = ks/(ks + s) \quad (1)$$

where v is the rate (e.g., growth rate), k a rate constant, s the concentration of a substrate, and ks the half-saturation constant. The dimensions of ks and s are the same: kg m^2 or kg m^3. Notice that Eq. **(1)** corresponds to a zero-order reaction when $s > > ks$ and a first-order reaction when $ks > > s$. Growth follows a zero-order reaction when the resources are abundant, while Eq. **(1)** describes the influence of limiting resources on the growth rate ($s < < ks$).

The situation in nature is often not so simple, as two or more resources may be limiting simultaneously. This can be described by

☆*Change History:* March 2018. H. R. Pethybridge included glossary, keywords, abstract, rearranged existing text, and updated references.

This is an update of S.E. Jørgensen, Growth Constraints: Michaelis–Menten Equation and Liebig's Law, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 1797–1799.

 doi:10.1016/B978-0-12-409548-9.11151-0

Table 1 Average freshwater plant composition on wet basis

Element	*Plant content (%)*
Oxygen	80.5
Hydrogen	9.7
Carbon	6.5
Silicon	1.3
Nitrogen	0.7
Calcium	0.4
Potassium	0.3
Phosphorus	0.08
Magnesium	0.07
Sulfur	0.06
Chlorine	0.06
Sodium	0.04
Iron	0.02
Boron	0.001
Manganese	0.000,7
Zinc	0.000,3
Copper	0.000,1
Molybdenum	0.000,05
Cobalt	0.000,002

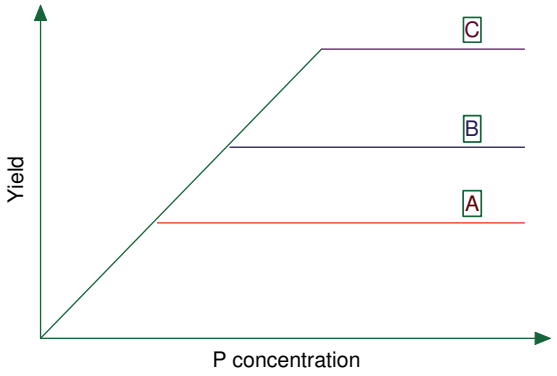


Fig. 1 Illustration of Liebig's law. The phosphorus concentration is plotted against the yield. At a certain concentration another component will be limiting, and a higher P concentration will not increase the yield. The three levels A, B, and C correspond in this case study to three different potassium concentrations.

$$v = K^r(N_1/(ks_1 + N_1))(N_2/(ks_2 + N_2)) \tag{2}$$

where N_1 and N_2 are the nutrient concentrations, K^r is a rate constant, and ks_1 and ks_2 are the half-saturation constants related to N_1 and N_2.

This equation will often limit the description of growth too much and is in disagreement with many observations. Eq. (3) seems to overcome these difficulties:

$$v = K^r \min\,(N_1/(ks_1 + N_1)), (N_2/(ks_2 + N_2)) \tag{3}$$

This expression is in accordance with Liebig's minimum law. Another possibility would be to apply the average of two or more limiting factors, for instance, for two limiting factors (nutrients):

$$v = K^r[(N_1/(ks_1 + N_1)) + (N_2/(ks_2 + N_2))]/2 \tag{4}$$

Finally, it is also possible to use the following expression:

$$v = 2K^r/[((ks_1 + N_1)/N_1) + ((ks_2 + N_2)/N_2)] \tag{5}$$

If one element (nutrient) is limiting, the growth Eq. (1) can be used. It is often the case for lakes, where phosphorus is the limiting nutrient. In coastal areas both phosphorus and nitrogen may be limiting at different parts of the year and it is necessary to apply Eq. (2), (3), (4), or (5). Experience has shown that Eq. (3) and (4) have given the most promising results in the sense that it has been possible to obtain better calibration and validation for models that are based on these two equations.

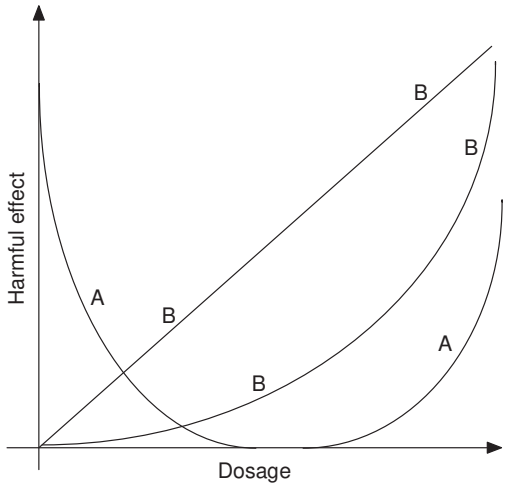


Fig. 2 A indicates threshold agent; B indicates nonthreshold or gradual agent. To have a threshold agent it is sufficient that one of the two A-plots is valid. The two B-plots represent two different dose–response curves.

The Michaelis–Menten equation is extensively applied in ecological models for nutrient-limiting plants' growth, for grazing, and for predation. The equation takes into consideration the needs for growth in form of nutrients or food and the availability of the resources. It is obvious that interactions between growth of organisms and the resources determining the growth are frequently describing the actual situation in an ecosystem, which explains the frequent application of the Michaelis–Menten equations. The same equation is applied frequently in biochemistry to describe the kinetics of biochemical enzymatic processes. The enzymes correspond to the organisms and the substrates correspond to the nutrients or food.

Threshold and Nonthreshold Agents

Following the general terminology, we distinguish between threshold agents, that in a suitable concentration promote growth, and nonthreshold agents, that in all concentrations have a negative or no effect on the growth.

Threshold agents include various nutrients, such as phosphorus, nitrogen, silica, carbon, vitamins, and minerals (calcium, iron, zinc, etc.). When they are added or taken in excess, the organism or the ecosystem can be overstimulated, and the ecological balance may be damaged. Examples are the eutrophication of lakes, streams, and estuaries from fertilizer runoff or municipal wastewater. The threshold level and the type and extent of damage vary widely with different organisms and stresses. The thresholds for some pollutants may be quite high, while for others they may be as low as 1 ppm or even 1 ppb.

The threshold level is often closely related to the concentration found in nature under normal environmental conditions. This shows that the organisms have become adapted to the chemical composition of the environment.

Nonthreshold agents are a number of heavy metals, pesticides, and a number of toxic organic compounds. It is important to determine the noneffect level or noneffect concentration for these compounds to determine at which level or concentration a negative effect on organisms and ecosystems can be expected.

Fig. 2 shows the different effects of threshold and nonthreshold agents. Threshold agents have often harmful effects when they are not present in sufficient quantities to support growth. The range where there chemical composition of organism to the chemical composition of the environment takes place. The biochemistry developed reflects the available elements and compounds. For instance, the role of magnesium and manganese ions and later zinc ions as cofactors for many enzymatic processes could be due to the presence of these ions in a suitable concentration, while other ions could have been used had they been present in an appropriate concentration. The use of calcium in skeletons can be explained more or less in this way.

See also: Ecological Processes: Gross and Net Production in Different Environments. Evolutionary Ecology: Limiting Factors and Liebig's Principle. General Ecology: Ecological Stoichiometry: Overview; Biomass; Growth Models. Global Change Ecology: Energy Flows in the Biosphere; Nitrogen Cycle. Human Ecology and Sustainability: Limits to Growth

Further Reading

Cao, J., 2011. Michaelis–Menten equation and detailed balance in enzymatic networks. The Journal of Physical Chemistry B 115 (18), 5493–5498.
Danger, M., Daufresne, T., Lucas, F., Pissard, S., Lacroix, G., 2008. Does Liebig's law of the minimum scale up from species to communities? Oikos 117 (11), 1741–1751.

Degryse, F., Shahbazi, A., Verheyen, L., Smolders, E., 2012. Diffusion limitations in root uptake of cadmium and zinc, but not nickel, and resulting bias in the Michaelis constant. Plant Physiology 160 (2), 1097–1109.
Hadeler, K.P., Jukić, D., Sabo, K., 2007. Least-squares problems for Michaelis–Menten kinetics. Mathematical Methods in the Applied Sciences 30 (11), 1231–1241.
Johnson, K.A., Goody, R.S., 2011. The original Michaelis constant: Translation of the 1913 Michaelis–Menten paper. Biochemistry 50 (39), 8264–8269.
Jørgensen, S.E., Bendoricchio, G., 2001. Fundamentals of ecological modelling, 3rd edn. Amsterdam, The Netherlands: Elsevier, 530 pp.
Liebig, J., 1840. Chemistry in its application to agriculture and physicology. London: Taylor and Walton.
López-Urrutia, Á., San Martin, E., Harris, R.P., Irigoien, X., 2006. Scaling the metabolic balance of the oceans. Proceedings of the National Academy of Sciences 103 (23), 8739–8744.
Min, W., Gopich, I.V., English, B.P., Kou, S.C., Xie, X.S., Szabo, A., 2006. When does the Michaelis – Menten equation hold for fluctuating enzymes? The Journal of Physical Chemistry B 110 (41), 20093–20097.
Odum, E.P., Odum, H.T., Andrews, J., 1971. Fundamentals of ecology. vol. 3. Philadelphia: Saunders.
Shelford, V.E., 1913. Animal communities in temperate America, as illustrated in the Chicago region: A study in animal ecology. Bulletin No. 5 University of Chicago Press.
Sterner, R.W., Elser, J.J., 2002. Ecological stoichiometry: The biology of elements from molecules to the biosphere. Princeton University Press.
Tzafriri, A.R., Edelman, E.R., 2007. Quasi-steady-state kinetics at enzyme and substrate concentrations in excess of the Michaelis–Menten constant. Journal of Theoretical Biology 245 (4), 737–748.

Relevant Websites

https://www.khanacademy.org/test-prep/mcat/biomolecules/enzyme-kinetics/v/steady-states-and-the-michaelis-menten-equation–KHANAACADEMY.
https://chem.libretexts.org/Core/Biological_Chemistry/Catalysts/Enzymatic_Kinetics/Michaelis-Menten_Kinetics–LibreTexts Project.
http://slideplayer.com/slide/6355981/–SlidePlayer.

Growth Models☆

Todd M Swannack, US Army Engineer Research and Development Center, Vicksburg, MS, United States

Introduction

This article presents an overview of the mathematical models of population growth for single-species populations. The first section discusses the models used to model population growth for species with nonoverlapping generations. These models are represented mathematically by linear or nonlinear difference equations. The second section focuses on the models used for species with overlapping generations—these range in complexity from the classic exponential and logistic growth models to models incorporating time lags, or population projection matrices with age (or class)-specific natality and mortality rates. The third section discusses how population growth is affected by the inclusion of stochastic variables.

Discrete Generations

Nonoverlapping generations are common biological phenomenon and population growth is modeled using difference equations. These models all assume that the population size at time $t + 1$ depends only on the conditions at the time t (usually the beginning of each growing or breeding season). The simplest model can be expressed as

$$N_{t+1} = R_0 N_t \tag{1}$$

where N_{t+1} is the population size at time $t + 1$ and N_t is the population size at time t. R_0 is a constant per capita reproductive rate, which is independent of population size. There are three possible outcomes of this model. If $R_0 > 1$ the population will increase geometrically (Fig. 1A), if $R_0 = 1$ the population is at equilibrium (Fig. 1B), and if $R_0 < 1$ the population size will decrease (Fig. 1C).

The above model assumes that R_0 does not change with population size; however, populations do not generally grow at a constant rate. The growth rate of a population should be different for different densities. The simplest assumption regarding this change in reproductive rate based on density is to assume that the R_0 decreases linearly as population size increases. R_0 can be calculated by using a linear equation:

$$R_0 = 1 - B\left(N_t - N_{eq}\right) \tag{2}$$

where $(-)B$ is the slope of the line and N_{eq} is the equilibrium size of the population ($R_0 = 1$ at N_{eq}). Therefore, Eq. (1) can be rewritten as

$$N_{t+1} = \left(1 - B\left(N_t - N_{eq}\right)\right)N_t \tag{3}$$

This model can generate several different behaviors, depending on the value of BN_{eq}. If BN_{eq} is between 0 and 2, the model will converge on N_{eq} either without oscillations ($1 > BN_{eq} > 0$) or with damped oscillations ($2 > BN_{eq} > 1$) (Fig. 2a). If BN_{eq} is between 2 and 2.57, the growth form is stable limit cycles continuing indefinitely (Fig. 2b), and if $BN_{eq} > 2.57$, the population size fluctuates in a nonrepeatable (chaotic) pattern (Fig. 2c).

☆*Change History*: March 2018. Todd M. Swannack updated Reference section.

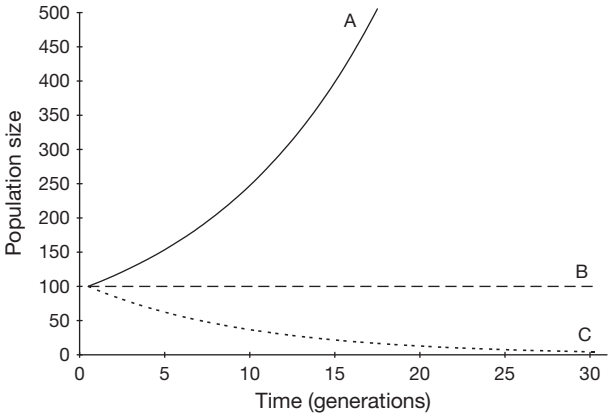


Fig. 1 Population growth for discrete generations using a difference Eq. (1) with $N_0 = 100$. A represents $R_0 = 1.1$; B represents $R_0 = 1.0$, and C represents $R_0 = 0.9$.

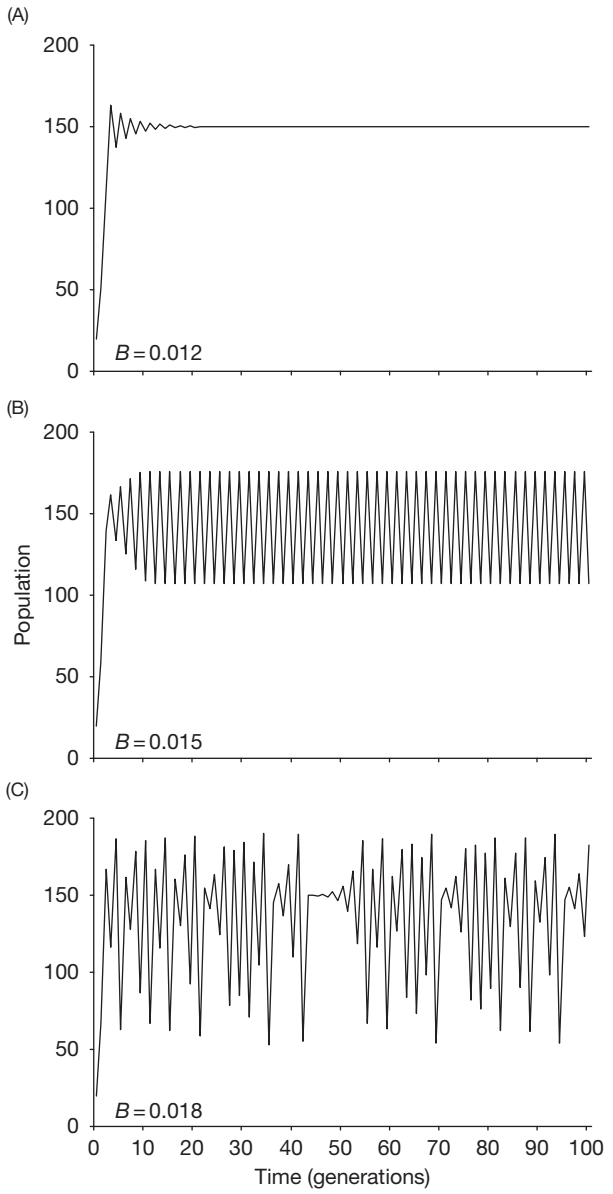


Fig. 2 Population growth with discrete generations and R_0 as a linear function of population density (Eq. 3). The model parameters were $N_0 = 20$, $N_{eq} = 150$, $B = 0.012$ (A), $B = 0.015$ (B), and $B = 0.018$ (C). Modified from Krebs, C. J. (2001). *Ecology*, 5th edn. San Francisco, CA: Benjamin Cummings.

Beverton–Holt Model

The Beverton–Holt model describes discrete-time populations:

$$N_{t+1} = \frac{\lambda N_t}{(1 + bN_t)} \tag{4}$$

where λ represents the finite rate of increase and b is a positive constant. This model approaches a stable equilibrium point for all parameter values and the growth form is analogous to the continuous-time logistic model (Eq. 10).

May Model

Models which assume a nonlinear relationship between the growth rate and the population size can also generate realistic growth forms. One of the simplest nonlinear models (commonly referred to as the May model) is expressed as

$$N_{t+1} = N_t e^{[r(1-N_t/k)]} \tag{5}$$

where e is the base of the natural logarithm, r is the intrinsic rate of increase, and K is the carrying capacity of the environment. This model assumes the relationship between r and N is nonlinear and the growth form exhibited by this model depends on the value of r, K, and the initial value of N. The growth forms of this model are analogous to that of Eq. (3). The model will reach a stable equilibrium at K if $1.99 \geq r \geq 0$, stable limit cycles (with 2, 4, 8, or 16 endpoints) if $2.692 \geq r \geq 2.0$, or become chaotic (nonrepeatable pattern) if $r \geq 2.693$.

Ricker Model

The Ricker model is another nonlinear model which has a wide range of behaviors:

$$N_{t+1} = \lambda N_t e^{bN_t} \tag{6}$$

This model, like Eqs. (4) and (5), can generate growth forms reaching a stable equilibrium point, can oscillate in a repeatable stable limit cycle, and can also generate chaotic, nonrepeating behavior, depending on the initial values of the parameters.

Continuous Generations

Many species have evolved life-history strategies that favor longer lives with multiple reproductive events occurring during the life span. Population growth of these iteroparous species has been commonly modeled using either continuous-time differential equations or population projection matrices.

Birth–Death Model

The simplest model of population growth for species with continuous generations assumes that birth and immigration are combined and death and emigration are combined into a general addition and loss rate, respectively. These rates are constants and the model is represented as

$$\frac{\mathrm{d}N}{\mathrm{d}t} = bN - \mathrm{d}N \tag{7}$$

where b is the instantaneous birth–immigration rate, d is the instantaneous mortality–emigration rate, N is the population size, and $\mathrm{d}N/\mathrm{d}t$ represents the rate of change in N.

Exponential Growth Model

Assuming that in any time interval ($\mathrm{d}t$) an individual can be added to the population either through birth or immigration and in the same interval there is a probability of dying or emigrating, then the instantaneous rate of per capita growth will be

$$r = b - d \tag{8}$$

where r represents the intrinsic rate of growth for the population at time t. The differential equation representing the increase in population size during successive time intervals is

$$\frac{dN}{dt} = rN \tag{9}$$

where N represents the population size and r represents the population's intrinsic capacity for increase (also referred to as the Malthusian parameter). This model assumes that environmental resources are unlimited and r is a constant. The population will increase as long as r is positive (i.e., as long as $b > d$) (Fig. 3A).

Logistic Growth Models

Population growth cannot continue to infinitely large sizes because resources within any environment are finite. The most common population growth model representing growth in a limited environment is a logistic equation (commonly called the Verhulst–Pearl logistic equation) that has a maximum limit at the carrying capacity (K) of the environment:

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right) \tag{10}$$

where r is the intrinsic rate of increase and the $(1 - N/K)$ term represents the effect of individuals in the same population competing for resources (i.e., the effect of intraspecific competition). The growth curve is sigmoid shaped (Fig. 3B). The rate of change in population size is highest at low population sizes and has a value of zero once the population reaches carrying capacity. The Verhulst–Pearl model assumes the population size at time t only depends on the conditions at t, K is a constant, and every new individual decreases the rate of population increase by the fraction $1/K$, causing the decrease in r to be a linear function of N. At any initial value of N, the model will approach K monotonically.

Theta Logistic Model

The Verhulst–Pearl model can be modified by adding an exponent (θ) to the $(1 - N/K)$ term allowing a nonlinear relationship between r and N. The θ-logistic model is as follows:

$$\frac{dN}{dt} = rN\left[1 - \left(\frac{N}{K}\right)^{\theta}\right] \tag{11}$$

where θ controls the shape of the relationship between r and N. The value of θ depends on how individuals of a given population interact at different values of N. The growth form of this model is sigmoid for all values of θ, yet different values of θ have different ecological interpretations. If $\theta > 1$ there is a convex relationship between r and N, inferring that a population grows rapidly until it approaches K then growth slows rapidly. If $\theta = 1$, then the model behaves exactly like Eq. (7) (i.e., the negative effect of each individual is the same, regardless of population size). If $\theta < 1$ then there is a concave relationship between r and N, implying that the per capita reduction in population growth is greater at lower values of N.

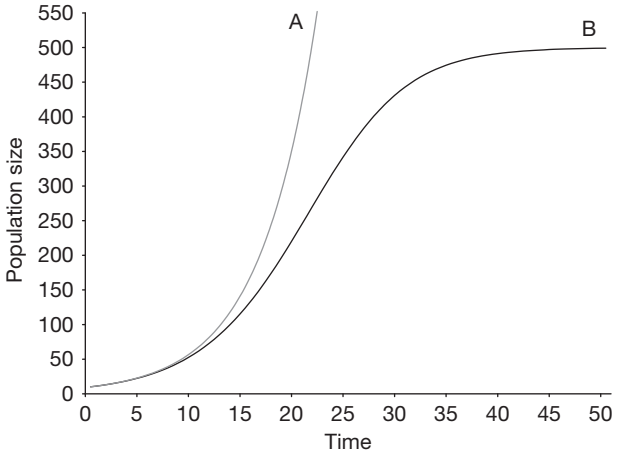


Fig. 3 Population growth forms for species with continuous generations. (A) Exponential (geometric) population growth in an environment with unlimited resources (Eq. 6, $r = 0.085$, $N_0 = 10$). (B) Logistic (sigmoid) population growth in an environment with limited resources (Eq. 7, $r = 0.085$, $K = 500$, $N_0 = 10$).

Time-Lag Models

Time–lag models attempt to create population growth forms which better reflect natural conditions because organisms rarely respond instantaneously to a change in the system. Introducing a time lag into a model can destabilize the model and cause dramatic fluctuations in population size, depending on the length of the time delay. The simplest time-lag model is a modification of the birth and death rate model (Eq. 7):

$$\frac{dN(t)}{dt} = b(N(t-\tau)) - d(N(t)) \tag{12}$$

where b is the instantaneous birth rate, t is the current time, τ represents the time delay (the $N(t-\tau)$ term is the population size at time $t-\tau$), and d is the instantaneous death rate. This model assumes that d is related to the current population size, but that b depends on the population size at time τ.

The Verhulst–Pearl logistic equation can be modified to incorporate time lags:

$$\frac{dN(t)}{dt} = rN_t\left[1 - \frac{N(t-\tau)}{K}\right] \tag{13}$$

where t represents the current time and τ represents the time delay that individuals experience before they are capable of contributing new individuals to the population. The dynamics of this model are controlled by values of the combination $r\tau$. The model will reach a stable equilibrium for $0 < r\tau < e-1$ ($e-1 \approx 0.36788$), damped oscillations for $e-1 < r\tau < \pi/2$, and a stable limit cycle for $r\tau > \pi/2$.

Eq. (13) can be further modified to incorporate more than one time lag:

$$\frac{dN(t)}{dt} = rN(t-\tau_g)\left[1 - \frac{N(t-\tau)}{K}\right] \tag{14}$$

where τ_g is a second time lag. The second time lag will further slow the rate of population increase, but will create patterns of growth similar to those in Fig. 4, that is, the larger the time lag, the larger the amplitude of the oscillations.

Matrix Models

As organisms age, their realized fecundity as well as their survival rates can change. Age-based matrix models separate populations into different age classes and each age class can possess different fecundity and survival rates. Matrix models are useful when hypothesizing about the importance of specific population parameters for certain age classes. Identifying the most sensitive age class is important for management strategies, especially for endangered species.

The most common matrix model is an age-based matrix, commonly referred to as the Leslie matrix. The individuals in a population are separated into classes based on their age (N_x) and each class is assigned a probability of survival (P_x) and a realized fecundity (F_x). F_x and P_x are put into a transition matrix **M** with F_x along the top row of the matrix and P_x along the subdiagonal, and all other elements in **M** are assigned a value of zero (Fig. 5A). Population projections can be made by placing the current number of individuals in each age class into a vector $\boldsymbol{n}$ and obtaining the product of the **M***n* matrices:

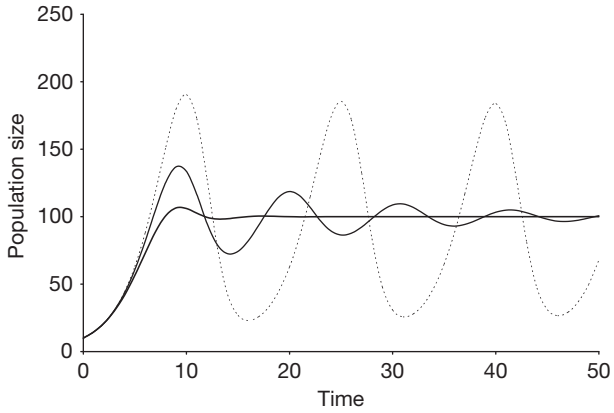


Fig. 4 Population growth of a model with time lags. The *heavy black line* was modeled without a time lag ($\tau = 0$) and reaches a stable equilibrium point, the *light black line* was created with a $\tau = 1$ and exhibits damped oscillations, and the dotted line was created using $\tau = 2$ and exhibits a two-endpoint stable limit cycle.

(A)

$$\mathbf{M}_A = \begin{bmatrix} F_0 & F_1 & F_2 & \dots & F_{m-1} & F_m \\ P_0 & 0 & 0 & \dots & 0 & 0 \\ 0 & P_1 & 0 & \dots & 0 & 0 \\ 0 & 0 & P_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & P_{m-1} & 0 \end{bmatrix} \quad \mathbf{n}_A = \begin{bmatrix} N_0 \\ N_1 \\ N_2 \\ \vdots \\ N_{m-1} \\ N_m \end{bmatrix}$$

(B)

$$\mathbf{M}_B = \begin{bmatrix} G_1 & F_2 & F_3 & \dots & F_{m-1} & F_m \\ P_1 & G_2 & 0 & \dots & 0 & 0 \\ 0 & P_2 & G_3 & \dots & 0 & 0 \\ 0 & 0 & P_3 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & G_{m-1} & 0 \\ 0 & 0 & 0 & \dots & P_{m-1} & G_m \end{bmatrix} \quad \mathbf{n}_B = \begin{bmatrix} N_1 \\ N_2 \\ N_3 \\ \vdots \\ N_{m-1} \\ N_m \end{bmatrix}$$

Fig. 5 Population projection matrices. (A) An age-structured Leslie matrix ($\mathbf{M}_A$) and a vector ($\mathbf{n}_A$) of the number of the individuals in each age-class. (B) A stage-based matrix ($\mathbf{M}_B$) and a vector ($\mathbf{n}_B$) of the number of the individuals in each age class. Subscripts represent the age or size class with m being the oldest or largest class, respectively.

$$\mathbf{M}\mathbf{n}_t = \mathbf{n}_{t+1} \tag{15}$$

where $\mathbf{n}_t$ is the vector whose elements are the number of individuals in each age class, and $\mathbf{n}_{t+1}$ is the vector containing the number of individuals in each age class at time $t + 1$.

For some species, size or stage of development may be better indicators of an individual's fecundity and survival rate. These models do not assume that every individual leaves the class at every time step. The probability an individual will survive and remain in class x during the next time step is G_x. Stage-based matrices incorporate this term along the diagonal in the transition matrix $\mathbf{M}$. P_x is defined as the probability that an individual will move into the succeeding class $x + 1$. The P_x elements are placed along the subdiagonal in the transition matrix $\mathbf{M}$ (Fig. 5b). F_x is defined the same as in the Leslie matrix and population projections are calculated the same as above. The standard stage-based matrix model assumes that the first size class is not reproductively mature and any the remaining classes each have a specific realized fecundity.

Stochastic Models

Deterministic models of population growth allow the population size at time $t + 1$ to be calculated based on the knowledge of N_t plus the other factors that are included in the models. Stochastic models attempt to include the probabilistic nature of biological systems by representing one (or more than one) variable in the given model as a random variable. Natural populations can start at the same size but due to chance increase or decrease at different rates based on the value of the random variable at each time step. Any variable in a model that has any degree of uncertainty can be parametrized as a random variable. Stochastic variables can be based on a probability of occurrence (i.e., probability of breeding at any given time step, or probability of having a certain number

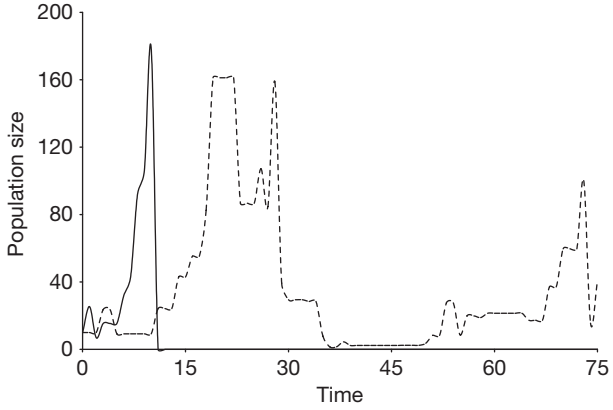


Fig. 6 Two simulations of a hypothetical population using a stochastic birth–death model. The probability of birth or death occurring at each time step was 0.5. The rest of the parameters were: $N_0 = 1.2$, $b = 1.2$, $d = 0.8$.

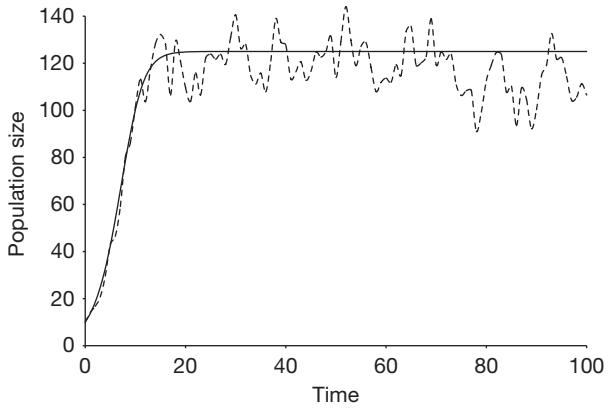


Fig. 7 Population growth for a deterministic model (*heavy line*) and model with the birth rate was a random variable (*dotted line*) chosen from a normal distribution with a mean of 1.2 and a standard deviation of 0.1.

of offspring during a given reproductive bout). These probabilities are generally assigned based on previous knowledge of the system. Stochastic models where both births and deaths occur randomly allow for a chance of extinction. Fig. 6 depicts two simulations of a random birth–death model (Eq. 7) where the probability of births and deaths occurring during a given time step was 0.5 (probabilities of occurrence were independent of each other). This representation exemplifies the nature of stochastic population models for small population sizes. The first simulation went extinct after 10 time steps while the second simulation did not go extinct during the simulation. Small populations are at a greater risk of extinction due to their inability to rebound after a catastrophic event. Stochastic models are often used to estimate the probability of extinction for a given population. This probability is based on several repetitions of the same stochastic model.

Random variables can also be chosen from a frequency distribution of historical data or from a statistical distribution fitted to historical data—Fig. 7 was created using the latter case, depicting growth curves for a hypothetical population created with both deterministic and random models. The same parameters were used except for the random model the birth rate was chosen at the beginning of each time step from a normal distribution with a mean of 1.2 and standard deviation of 0.1 while the birth rate for the deterministic model was 1.2. The population size of the stochastic model oscillates around K but does not converge on a single equilibrium point.

Summary

Population growth for a single-species population can be modeled in a variety of ways. Populations will grow exponentially if the per capita reproductive rate is constant and independent of population size. Simple difference equations used to model populations with discrete generations can generate complex behaviors including converging on an equilibrium population size, limit cycles, and chaos. The most common model used to estimate growth in populations with overlapping generations is the Verhulst–Pearl logistic equation where population growth stops at the carrying capacity. The Verhulst–Pearl model has restrictive assumptions, and was modified into the θ-logistic model which relaxes the assumptions of the relationship between the intrinsic rate of increase and population size. Time–lag models allow for the future population size to be dependent on a population size in the past. Adding a time lag to a model can destabilize it and cause fluctuations in population size, the amplitude of the fluctuations depending on the length of the time lag. Age- or size-based matrices group individuals into classes with each class having its own natality and mortality rates, and these models can project the future number of individuals for each class. Stochasticity can be parametrized into any model and the outcome will no longer be deterministic and will better reflect the probabilistic nature of biological populations.

Further Reading

Brauer F, Castillo-Chavez C, and Castillo-Chavez C (2012) *Mathematical models in population biology and epidemiology*. vol. 1. New York: Springer.
Caswell H (2001) *Matrix population models: Construction, analysis, and interpretation*, 2nd edn. Sunderland, MA: Sinauer Associates.
Gadgil M and Bossert WH (1970) Life historical consequences of natural selection. *American Naturalist* 104: 1–24.
Krebs CJ (2001) *Ecology*, 5th edn San Francisco, CA: Benjamin Cummings.
Leslie PH (1945) On the use of matrices in certain population mathematics. *Biometrika* 33: 183–212.
MacDonald N (1989) *Delays in biological systems: Linear stability theory*. Cambridge: Cambridge University Press.
Matis JH, Kiffe TR, Matis TI, Jackman JA, and Singh H (2007) Population size models based on cumulative size, with application to aphids. *Ecological Modelling* 205(1–2): 81–92.
May RM (1973) *Stability and complexity in model ecosystems*. Princeton, NJ: Princeton University Press.
May RM (1974) Biological populations with nonoverlapping generations: Stable points, stable cycles, and chaos. *Science* 186: 645–647.
McCallum H (2000) *Population parameters: Estimation for ecological models*. Oxford: Blackwell Science.

Pearl R and Reed LJ (1920) On the rate of growth of the population of the United States since 1790 and its mathematical representation. *Proceedings of the National Academy of Sciences of the United States of America* 6: 275–288.
Pianka ER (1972) *r* and *K* or *b* and *d* selection? *American Naturalist* 106: 581–588.
Renshaw E (1991) *Modelling biological populations in space and time*. Cambridge: Cambridge University Press.
Reynolds JD and Freckleton RP (2005) Population dynamics: Growing to extremes. *Science* 309: 567–568.
Sibley RM, Barker D, Denham MC, Hone J, and Pagel M (2005) On the regulation of populations of mammals, birds, fish, and insects. *Science* 309: 607–610.
Turchin P (2003) *Complex population dynamics: A theoretical/empirical synthesis*. Princeton, NJ: Princeton University Press p. 2003.
Verhulst PF (1838) Notice sur la loi que la population suit dans son accroissement. *Correspondence Mathématique et Physique* 10: 113–121.

Habitat

J Stamps, University of California, Davis, CA, USA

Introduction

'Habitat' can be defined as a location in which a particular organism is able to conduct activities which contribute to survival and/or reproduction. This definition emphasizes the notion that the term habitat is organism-specific; that is, it focuses on the biotic and abiotic factors that affect the survival or reproduction of a particular type of organism, and on the areas that contain these factors. In addition, the term 'habitat' focuses on particular areas within a larger landscape. In contrast, the related term 'niche' considers the range of environmental conditions which allow the members of a species to live and reproduce, but does not specify the areas where these conditions exist.

The term habitat can refer to locations at several different spatial scales. At the largest spatial scale, habitat refers to large geographic areas in which the members of a population are able to survive and reproduce. At a medium spatial scale, habitat refers to areas in which a single member of a population is able to survive and reproduce. And finally, at a small spatial scale, habitat refers to areas in which a single member of a population can conduct particular activities (e.g., foraging, shelter, offspring production). Often, the term 'microhabitat' is used to refer to areas in which organisms conduct particular activities, while the term 'macrohabitat' refers to areas used by populations or species.

Implied in the concept of habitat at the intermediate (individual) level is that it contains all of the biotic and abiotic factors which affect survival and reproduction, throughout the life of that individual. However, it is frequently difficult (if not impossible) for humans to identify and measure all of the factors that affect survival and reproduction for taxa of interest. As a result, researchers interested in defining habitat for a particular species usually take 'shortcuts', trying to find factors that humans can readily measure which might be related to the factors which determine whether or not an organism is able to live and reproduce in a given area. Some examples of the shortcuts used by humans to describe habitat include using vegetation species or structural features to describe habitat for birds, using type of soil or incline to describe habitat for trees, or using salinity or distance from the sea to describe habitat for organisms that live in estuaries. In these and comparable cases, researchers assume that the factors they are using to describe habitat either have strong direct effects on survival and reproduction, or that they are strongly correlated with other biotic and abiotic factors that affect survival and reproduction.

If these assumptions are valid, then surrogate measures of habitat can be reasonably successful. Thus, if a particular plant community provides a particular species of bird with nest-sites, food, water, protection from predators and inclement weather, and locations appropriate for social activities, then the space-use patterns of that species might map relatively closely onto the spatial distribution of that plant community. The assumption that some features of a habitat are more important than others is implicit in the term 'habitat-forming species', which refers to organisms that create habitat that can be used by other organisms. For instance, certain species of sea anemones provide habitat to anemonefish, which are able to live, grow, and reproduce within the protection of their tentacles. At a broader spatial scale, 'foundation species' create areas that are capable of supporting an entire community of other organisms. An obvious example are the corals which provide food, shelter from predators, protection from wave action, and other benefits to the large array of organisms that are only able to make a living on coral reefs.

However, while it is tempting to assume that one can determine habitat for one species by mapping the distributions of other organisms, this practice can sometimes lead one astray. Even if the members of a species only occur in association with another organism, this does not mean that every area that contains that organism will support members of that species. For instance, the presence or absence of avian brood parasites (e.g., cowbirds) in a landscape can have a major impact on the reproduction and sustainability of avian populations in that area, but vegetation maps alone are unlikely to reveal whether or not avian brood parasites are present in a particular region.

Given the considerable uncertainty that exists when humans attempt to describe habitats, there can be quite a gap between our notion of what constitutes a habitat and the constellation of biotic and abiotic factors which determine whether the members of a species would be able to sustain themselves in a given area. This uncertainty is one reason for use of the phrase 'suitable habitat', which ecologists often use to refer to areas which actually do contain the factors which determine whether an area can be used by the members of a population.

The Importance of 'Empty' Suitable Habitat

As a practical matter, empiricists begin studies of the habitat of a given organism by focusing on the areas in which it currently lives and reproduces. However, an area can provide suitable habitat for an organism without currently being occupied by any members of that species. The assumption that a habitat can be suitable but empty lies at the heart of many important concepts in basic and applied ecology. Metapopulation biology is based on the assumption that populations live in patches linked by dispersal, and that

at any given time, some of the patches in a given landscape may be unoccupied. Similarly, restoration ecology is based on the premise that humans can change previously unusable habitat into habitat that is suitable for a given species, but that newly suitable habitat will remain empty until it is colonized (either artificially, via translocations, or naturally, via dispersal) by the members of that species.

The concept of empty, suitable habitat is also relevant to range shifts and range expansions, topics of particular interest for applied ecologists in this era of global warming. Many areas which currently sustain the members of a given species may become unsuitable in the future, as a result of changes in temperature, rainfall, and all of the changes in biotic factors that can occur as an indirect result of climate change. Conversely, other regions which currently lack any members of a given species may become suitable for them within a few decades, as a result of the direct and indirect effects of global warming. While it is difficult to predict how climate change will affect the ranges of particular species, the geographic locations of suitable habitat are already beginning to shift for many organisms, and they are likely to shift even more in the coming years. As a result, ecological models which are based on the assumption that habitats are fixed in space across long periods of ecological and evolutionary time are gradually giving way to new approaches which acknowledge temporal shifts of habitat across the landscape.

Habitat Quality

Besides simply asking whether or not a habitat can support an organism, we can further categorize suitable habitats based on their 'quality', where the quality of a given habitat can be defined as the fitness an individual can expect if it lives in that habitat, after controlling for any positive or negative effects of conspecifics on fitness. In practice, habitat quality can be estimated by measuring fitness components of organisms living in different types of habitats at the same population density, or by controlling statistically for positive and negative relationships between population density and fitness.

Habitat quality measured at the spatial scales that are relevant to individuals has important implications for habitat quality at the larger spatial scales that are relevant to populations and species. For instance, if habitat type A supports higher individual survivorship and reproduction than habitat type B at any given population density, then habitat A will contribute more new biomass and more new recruits to the population than habitat type B per unit area. Hence, accurately estimating habitat quality is a major concern both for basic ecologists interested in the effects of habitat features on population dynamics, and for applied ecologists interested in identifying the locations with the largest potential impact on endangered or pest species.

As should be apparent from the previous section, population density should not be used as a surrogate for habitat quality. Although organisms sometimes accumulate in high-quality habitats, there are plenty of situations in which this does not occur. Indeed, the best habitats from the perspective of individual survival and reproduction may be areas which currently contain moderate to low population densities, or even areas which currently do not contain any members of the species. Every decade or so, ecologists write influential articles explaining the various reasons why population density should not be equated with habitat quality. Unfortunately, however, this practice is still fairly common, perhaps because it is easier for humans to estimate the number of organisms living at a given location than to measure the survivorship and reproductive success of those organisms.

Habitat Selection

Another interesting implication of the notion that habitats vary in quality is that we would expect organisms to evolve physiological or behavioral processes which increase their chances of finding, recognizing, and using higher-quality habitats. In fact, mechanisms for habitat selection are nearly ubiquitous in the natural world. Even sessile organisms exhibit habitat selection, albeit at the smaller spatial scales. Examples of microhabitat selection by sessile organisms include plants growing toward the light, or cnidarian colonies growing in directions that enhance the flow of nutrient-rich water past their foraging polyps. However, the champions of habitat selection are mobile animals. In fact, one of the major advantages of mobility in animals is that it allows them to select different microhabitats or habitats in which to conduct different activities, instead of attempting to maximize every component of fitness at a single location. Animals are able to forage in some microhabitats which contain food and safety from predators, rest in other microhabitats which feature favorable abiotic features (temperature, moisture, light, etc.) and biotic features (protection from predators and ectoparasites), advertise for potential mates in microhabitats in which communication signals can be perceived over long distances, etc. At larger spatial scales, individuals can select some types of habitat for long-term use during reproductive periods (breeding home ranges or territories), other types of habitat for long-term use when they are not reproductive (e.g., juvenile home ranges, wintering areas), and still other types of habitat to use when traveling over long distances (migration or dispersal).

Just as habitat can be defined at different spatial scales, habitat selection can occur at the level of microhabitats, individuals, or populations. In each situation, habitat selection refers to the behavioral processes by which individuals choose particular locations for conducting particular types of activities. At the smaller spatial scales, habitat selection refers to the processes by which animals choose microhabitats to use for foraging, offspring production (e.g., oviposition or nest-site selection), courtship or mating, sleeping, or other activities. At medium spatial scales, habitat selection refers to the choice of areas in which individuals will spend extended periods of time, for example, the choice by natal dispersers of a new patch of habitat in which to establish a home range or territory. Finally, at larger (population level) spatial scales, habitat selection refers to the processes by which different

individuals in the same population select a place in which to live. At this spatial scale, interactions between individuals within the population often play a major role in habitat selection. For instance, in some species newcomers are attracted to areas which already contain conspecifics (conspecific attraction), so that even if different areas contain the same biotic and abiotic features, newcomers are more likely to settle in areas that already contain members of their own species.

Habitat Selection versus Habitat Use

As a practical matter, it is much easier to study the behavioral processes responsible for habitat selection at smaller than at larger spatial scales. Many species will express reasonably natural microhabitat selection behavior under controlled conditions in the laboratory or in field enclosures, so we know quite a bit about how animals select foraging sites, oviposition sites, shelter sites, etc. However, it is much more challenging to design and conduct elegant experiments of habitat selection at larger spatial scales. As a result, researchers interested in habitat selection at the level of individuals and populations often estimate habitat selection indirectly, by comparing the habitat features of areas actually used by individuals or groups of individuals with habitat features at points in the same region that have been randomly chosen by the researcher. The implicit assumption here is that any difference between the two distributions reflects active habitat choice by the individuals in that population. However, this assumption is often problematical. For instance, the habitat-use patterns of new recruits in marine fish and invertebrates are strongly affected by habitat-specific mortality immediately after settlement. As a result, nonrandom distributions of new recruits across the landscape often have more to do with differential mortality than they do with differential settlement, or movement by new arrivals to preferred habitats.

Several authors have suggested using the term 'habitat use' rather than 'habitat selection' to refer to nonrandom spatial distributions of animals and other mobile organisms in heterogeneous landscapes. Habitat use carries no implicit assumptions about the processes which are responsible for associations between organisms and particular habitat features, and instead simply indicates habitat features which are associated with the space-use patterns of the organism of interest. As was indicated earlier in this article, identifying the factors which are associated with the habitat-use patterns of a given organism is an important first step toward identifying the factors that are required to support individuals and populations of any given organism.

Further Reading

Armstrong, D.P., 2005. Integrating the metapopulation and habitat paradigms for understanding broad-scale declines of species. Conservation Biology 19, 1402–1410.
Dennis, R.L., Shreeve, T.G., Van Dyck, H., 2006. Habitats and resources: The need for a resource-based definition to conserve butterflies. Biodiversity and Conservation 15, 1943–1966.
Jones, J., 2001. Habitat selection studies in avian ecology: A critical review. Auk 118, 557–562.
Robertson, G.A., Hutto, R.L., 2006. A framework for understanding ecological traps and an evaluation of existing evidence. Ecology 87, 1075–1085.
Stamps, J.A., 2001. Habitat selection by dispersers: Integrating proximate and ultimate approaches. In: Clobert, J., Danchin, E., Dhondt, A.A., Nichols, J.D. (Eds.), Dispersal. Oxford, UK: Oxford University Press, pp. 230–242.
Stamps, J., Krishnan, V.V., 2005. Nonintuitive cue use in habitat selection. Ecology 86, 2860–2867.
Sutherland, W.J., 1996. From Individual Behaviour to Population Ecology. Oxford, UK: Oxford University Press.
Van Horne, B., 1983. Density as a misleading indicator of habitat quality. Journal of Wildlife Management 47, 893–901.

History of Ecology☆

Frank N Egerton, University of Wisconsin-Parkside, Kenosha, WI, United States
Nathalie Niquil, Université de Caen, Caen, France
Irene Martins*, University of Coimbra, Coimbra, Portugal

Introduction

Ernst Haeckel coined the word "oecology" in 1866 for a new science, but relevant observations and ideas had already been accumulating since the ancient Greeks. The balance of nature was the first ecological idea; Carl Linnaeus expanded it beyond animals to include plants and named it *Oeconomia Naturae*. Specialized sciences began to emerge in the early 1800s; among the earliest was phytogeography, founded by Alexander von Humboldt. Evolutionary theories by Lamarck and Charles Darwin were relevant to ecological ideas, since Lamarck thought species evolve rather than become extinct; Darwin saw competition as a cause of extinction. The roots of the main ecological specializations—plant ecology, animal ecology, limnology, and marine ecology—emerged in the 1800s, and limnology and plant ecology became organized by the 1890s. These four specializations were developed throughout the 1900s, as were new ones—primarily population ecology and ecosystem ecology. Ecological societies and journals came to the fore in the 1900s, as did institutions and specialized schools in various universities. Biogeochemistry arose in Russia (the USSR) in the early 1900s, and the Gaia theory arose in 1972. After the Second World War, environmentalism became important in all countries, and ecologists were needed as consultants. The International Biological Program (1964–74) produced many publications on ecosystems throughout the world.

Natural History

Antiquity and Middle Ages

The earliest ecological concept was the balance of nature, which arose from observations by Herodotus that predatory animals have fewer offspring than their prey and by Plato that each species has means to survive. Later, the sciences of zoology and botany were founded at Aristotle's school, the Lyceum. Supposedly, Aristotle wrote the zoological treatises and Theophrastos the botanical treatises; more likely, Theophrastos organized and compiled all of them. In Roman times, these sciences were abstracted in Pliny's *Natural History* and then remained part of a science of natural history.

During the Middle ages, ancient natural history knowledge was recovered and expanded. Emperor Friedrich II recorded careful observations on raptors and their prey in his treatise on falconry. Scholastic Albertus Magnus wrote two lengthy encyclopedias on the natural history of plants and on animals, synthesizing ancient Arabic and contemporary European knowledge (including his own observations).

Scientific Revolution in the 1500s and the 1600s

During the 1500s, Albertus' encyclopedic tradition was continued by Italian, French, and German herbalists, and the Swiss scholar Conrad Gessner (Fig. 1) wrote natural histories of both animals and plants but only published the former before he died of plague in 1565. Two Italian physicians made important contributions to contagion theory: Girolamo Fracastoro (Fig. 2) published *De contagione* (1546) and Girolamo Gambuccini published the first book on parasitic worms (1547). Aldrovandi and Penny wrote encyclopedias on insects in the late 1500s which were published in the 1600s. Scientific societies arose in mid-1600s and published both books and periodicals which included natural history. Francesco Redi's, John Ray's, and Antoni van Leeuwenhoek's publications were especially important for natural history (Figs. 3–5). All three, for example, studied animal parasites. John Graunt, William Petty, and Matthew Hale advanced human demography in the late 1600s.

The 1700s

Luigi Ferdinando Marsigli published the first treatise on what we call "oceanography" in 1725, *Histoire physique de la mer*, which included information on temperature, salinity, tides, currents, depth contours, and marine plants and animals. He realized that further studies required organized efforts. That began with Captain James Cook's three world exploration voyages, 1768–79, with accompanying naturalists. René Antoine de Réaumur (Fig. 6) made detailed and precise studies of the life histories of insects and

☆*Change History*: April 2018. Frank Egerton updated to include female ecologists.

This is an update of F.N. Egerton, History of Ecology, In Reference Module in Earth Systems and Environmental Sciences, Elsevier, 2013.

Present address: CIIMAR—Interdisciplinary Centre of Marine and Environmental Research of the University of Porto, 4450-208 Matosinhos, Portugal.

 doi:10.1016/B978-0-12-409548-9.00864-2

Fig. 1 Conrad Gessner (1516–65).

Fig. 2 Girolamo Fracastoro (*c.* 1478–1553).

Fig. 3 Francesco Redi (*c.* 1626–98).

Fig. 4 John Ray (1627–1705).

other invertebrates in his *Histoire des insects* (6 vols., 1734–42). The balance of nature concept was originally limited to animals. Carl Linnaeus (**Fig. 7**) named the concept "economy of nature" and broadened it to include plants—in dissertations which his students defended for their degrees, but which were essentially written by him. His *Oeconomia Naturae* (1749) was the first attempt to organize an ecological science. Linnaeus believed in the stability of species, and his economy of nature concept was not as dynamic as post-Darwinian ecology became, but he believed that members of each species flourish or not within a cycle of plant succession. In *Philosophia Botanica* (1751), he itemized 25 different stations (habitats) of plants. His contemporary, Georges Louis

Fig. 5 Antoni van Leeuwenhoek (1632–1723).

Fig. 6 René Antoine de Réaumur (1683–1757).

Fig. 7 Carl Linnaeus (1707–78).

Fig. 8 Georges Louis Leclerc, Comte de Buffon (1707–88).

Leclerc, Comte de Buffon (Fig. 8), with collaborators, wrote a large *Histoire naturelle* (44 vols., 1749–1804, though his contributions ended with his death in 1788). Linnaeus and Buffon received many plants and animals from explorers in distant places, and they struggled to explain the similarities and differences among species. Buffon developed a speculative theory that a limited number of species arose initially in a small area, and as their populations increased, they spread out into the rest of the world and were modified into new species by their new surroundings. Clergyman Gilbert White (Fig. 9) published *The Natural History and Antiquities of Selborne* (1789) that contains accurate observations on locations and habits of local plants and animals; it became the most widely read work on natural history. Investigators of different kinds of "airs" tested them on plants and animals, beginning

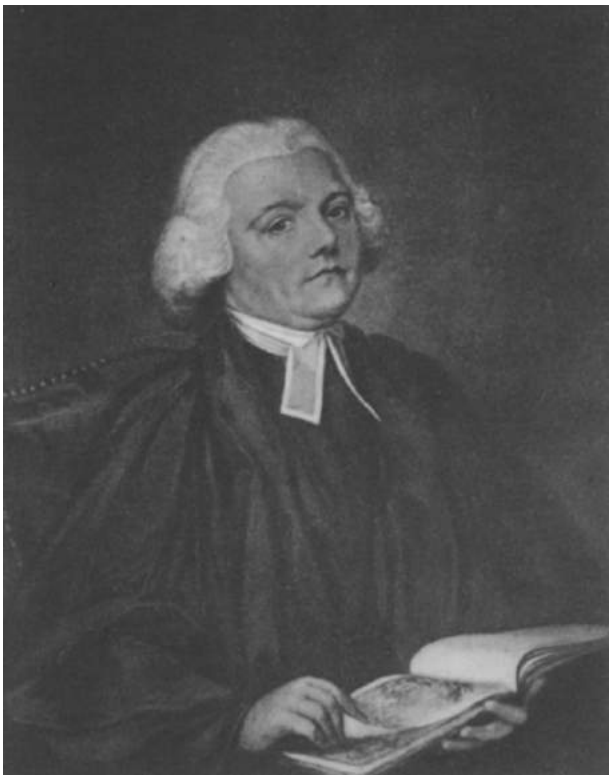


Fig. 9 Gilbert White (1720–93).

with Stephen Hales, the founder of plant physiology, and culminating with Joseph Priestley and Jan Ingen-Housz, who discovered that plants in sunlight produce a gas (named "oxygen" by Antoine Laurent Lavoisier) that animals need for respiration. This was new evidence of the balance of nature.

Specialized Sciences

The 1800s

Early in the century, Alexander von Humboldt (**Fig. 10**) became famous for his investigations and publications. He used inherited wealth to fund 5 years of exploration in Latin America, and then spent two decades in Paris, publishing his findings. "Humboldtian science" involved correlations between plant associations (e.g., grasslands, rain forests, tundra) and environmental factors (e.g., temperature, precipitation, topography), which he carefully measured to understand the distribution and abundance of species of plants and animals. He was the effective founder of phytogeography, emphasizing distribution of vegetation rather than floristics.

In 1809, Lamarck (**Fig. 11**) published his theory of evolution, arguing that species do not become extinct; they become different species. August de Candolle (**Fig. 12**), who rejected Lamarck's theory, emphasized (1820) competition between species as important for determining distributions and extinctions. His writings were used by geologist Charles Lyell (1832), who believed in the stability of species and that fossils do represent extinct species.

Hewett C. Watson (**Fig. 13**), founder of British phytogeography, was strongly influenced by Humboldt and Lamarck but not by de Candolle. He investigated the variability of British plant species in different parts of their ranges, and also compared the British and Azores Island floras.

Charles Darwin (**Fig. 14**), who read widely, was already familiar with the argument for the importance of competition in nature before he read in 1837 Thomas R. Malthus' *Essay on the Principle of Population* (1798, 5th edn. 1825). Darwin's theory of evolution by natural selection focused attention on the struggle of organisms against biotic and abiotic environmental forces. In such a revolutionary book as *On the Origin of Species* (1859), one might expect Darwin to reject the balance of nature concept, since the struggle for existence can cause extinctions. However, his famous cats–mice–bees–clover story seemed to support the balance concept. It was Alfred Russel Wallace who, thinking of extinctions, asked "Where is the balance?"

A Darwin disciple, Ernst Haeckel (**Fig. 15**), in 1866 expressed the need for a new science, "oecology." In 1870, he explained, "By oecology we mean the body of knowledge concerning the economy of nature—the investigation of the total relations of the animal both to its inorganic and to its organic environment." It took about three more decades for ecological sciences to begin to

Fig. 10 Alexander von Humboldt (1769–1859).

Fig. 11 Jean Baptiste de Lamarck (1744–1829).

organize, but meanwhile, various ecological ideas were elaborated. Darwin's theory of evolution included positive relations between individuals and between species, but the emphasis seemed to be on the negative relations of competition, predation, and parasitism. Several authors explored positive relations: Anton de Bary coined the word "symbiosis" and wrote a book on it (1869); P. J. van Beneden defined mutalism in his book *Les commensaux et les parasites* (1875); Alfred Espinas described diverse forms of mutualism in *Des sociétés animals* (1878); and Peter Kropotkin wrote on both animals and humans in *Mutual Aid* (1902).

Fig. 12 August-Pyramus de Candolle (1778–1841).

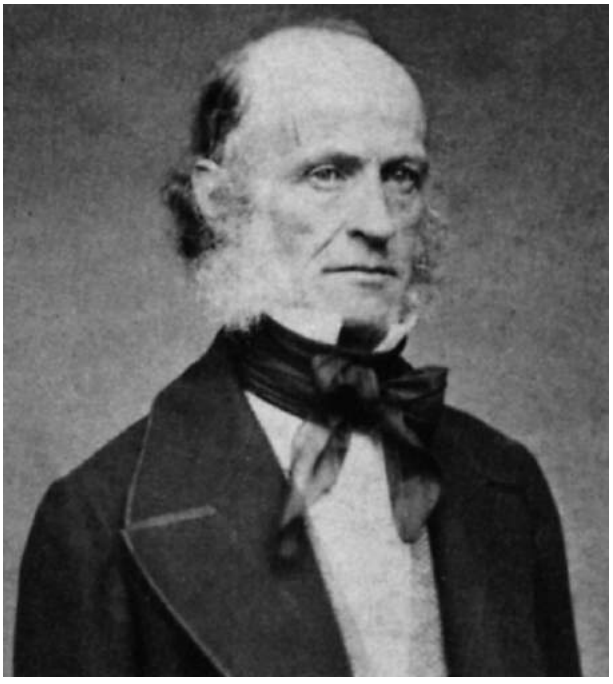


Fig. 13 Hewett Cottrell Watson (1804–81).

Parasitology and the germ theory of disease were important developments in the late 1800s, but their relevance for ecology was little developed.

Edward Forbes (Fig. 16) began studying marine animals in the 1830s and published many works concerning the distribution of species in relation to depth and other factors. He postulated the existence of an "azotic" zone in the depths of the sea. J. J. Coste established a marine zoology laboratory at Concarneau, France, in 1859; more influential was Anton Dohrn's Stazione Zoologica,

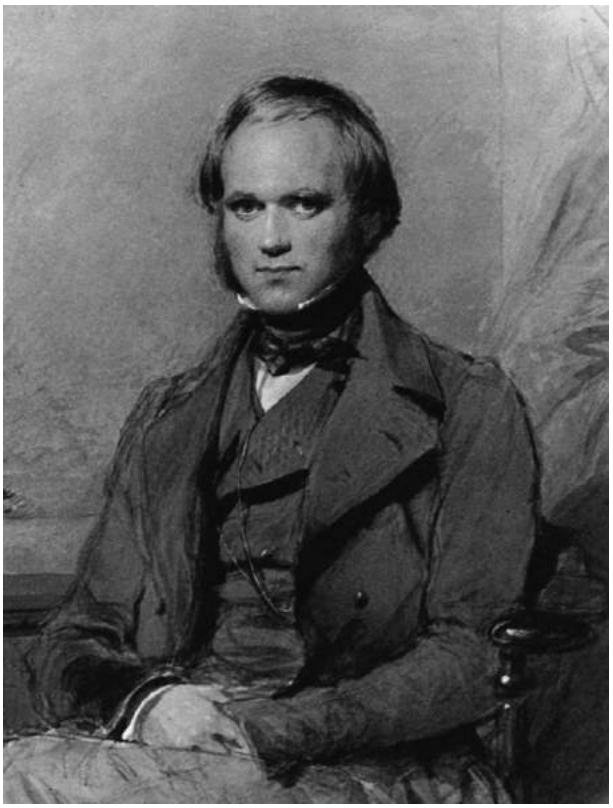


Fig. 14 Charles Robert Darwin (1809–82).

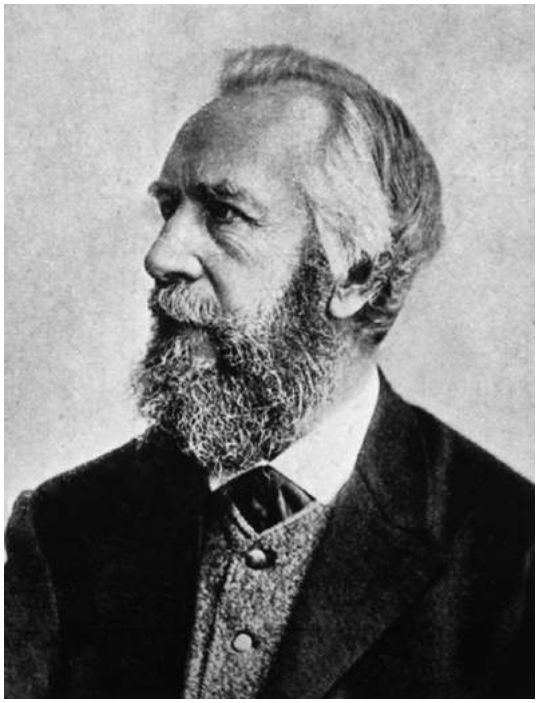


Fig. 15 Ernst Heinrich Haeckel (1834–1919).

established at Naples in 1872. The latter influenced Spencer Fullerton Baird to obtain funds for a US Fish Commission Laboratory at Woods Hole, Massachusetts, in 1884. In 1888, the independent Woods Hole Oceanographic Institution opened nearby. Charles

Fig. 16 Edward Forbes, Jr. (1815–54).

Wyville Thomson (**Fig. 17**) was the first student of deep oceanic life, and his book, *The Depths of the Sea* (1872), helped him become head naturalist on the worldwide voyage of the HMS Challenger (1872–76), and later he directed publication of 50 large volumes of its scientific reports (1885–95). Impressive discoveries of this expedition inspired other countries to sponsor oceanographic voyages. Karl Möbius (German 1877, English 1883) studied oyster beds to assist in their management. He argued that an oyster bed is a "community of living beings, a collection of species and a massing of individuals, which find here everything necessary for their growth and continuance." His study was influential for developing the concept of a biotic community.

American zoologist Stephen A. Forbes (**Fig. 18**) wrote an influential essay, "The Lake as a Microcosm" (1887), which may have been influenced by Möbius' essay, and which used the balance of nature concept to explain the apparent stability of species in a lake despite competition and predation. Meanwhile, since the 1860s, the Swiss zoologist François Alphonse Forel (**Fig. 19**) devoted his research to the life in Lake Geneva and the environmental conditions. His studies culminated in a three-volume paradigm treatise on this research (1892–1904), which he named as "limnology," and he published the first textbook on it in 1901.

Although Swiss botanist Carl Schröter first introduced the terms "autecology" and "synecology" in 1910, those subjects already existed by the 1890s. A strong German tradition in plant physiology and phytogeography led to the development of plant autecology, and when a botanical congress in Madison, Wisconsin, in 1893 adopted the term "ecology," "autecology" was meant. However, the first plant ecology treatise, by Eugenius Warming (**Fig. 20**), *Plantesamfund* (1895), was on synecology, based on his lectures in the first course ever taught in ecology, at the University of Copenhagen. Also in 1895 S. A. Forbes claimed that "economic entomology is simply applied ecology."

Due to strong gender discrimination, women attempting to make a career in natural sciences faced strong antagonism. This was the case of Emilie Snethlage (1868–1929) (**Fig. 21**), a pioneer in fieldwork and ornithology in the Amazonian region. Initially, Emilie worked as a governess in Germany, Ireland, and England, until 1899, when she received a small inheritance and decided to enter the University of Berlin to fulfill her youthful dream: study natural history. However, as a woman, she could only attend lectures hidden behind a screen, she could not speak, and she had to enter and leave lecture rooms 15 min before and after male students. In spite all this, Emilie earned a doctorate in 1904, under supervision of August Weismann. She worked as zoological assistant at the Berlin Natural History Museum until hired by Emilio Goeldi at the Natural History Museum in Belém, Brazil. In 1914, she became the museum director. Between 1908 and 1928, she undertook numerous field expeditions in the Amazon forest and other remote regions in Brazil to study natural history, zoogeography, and ethnology, and describing numerous new species of tropical birds. In some expeditions, only Indians from local tribes accompanied her, through risky regions. Between 1905 and 1928, she published more than 35 scientific papers and books, one of which was the first Catalogue of Amazonian Birds (1914). The Madeira Parakeet (*Pyrrhura snethlageae*), described in 2002, was named in her honor.

Libbie Henrietta Hyman's (1888–1969) (**Fig. 22**) graduated from Fort Dodge High School in 1905, and one of her teachers obtained a scholarship for her at the University of Chicago, where she majored in zoology, under the direction of Professor Charles

Fig. 17 Charles Wyville Thomson (1830–82).

Fig. 18 Stephen Alfred Forbes (1844–1930).

M. Child (B.S. 1910). Child's focus and hers was on invertebrates (PhD 1915). Her dissertation was published a year later. In graduate school, she worked as a laboratory assistant in several zoology courses, and she continued to be Child's assistant after receiving her PhD. She was dissatisfied with available manuals for introductory courses and published "A Laboratory Manual for Elementary Zoology" (University of Chicago Press, 1919, edn. 2, 1929) and "A Laboratory Manual for Comparative Anatomy" (University of Chicago Press, 1922, edn. 2, 1942). In 1940, she began publishing her magnum opus, "The Invertebrates" (6 vols., 1940–67, over 4000 pages). She became the leading invertebrate zoologist of North America and received numerous honors from scientific societies.

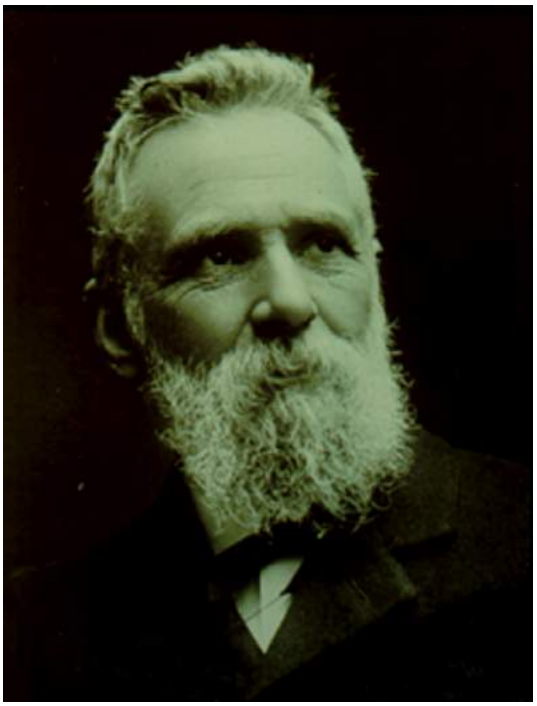


Fig. 19 François Alphonse Forel (1841–1912).

The 1900s

Specializations

Four specializations of ecology that emerged during the late 1800s—plant ecology, animal ecology, limnology, and marine ecology—persisted throughout the 1900s, but other specializations also arose, most notably population ecology and ecosystem ecology (**Fig. 23**). Population ecology focused primarily on animals, but both plant and animal ecologists developed ecosystem ecology.

American plant ecologists Frederic E. Clements and Henry Chandler Cowles (**Figs. 24** and **25**) led the investigation of plant communities. Clements wrote the first ecology textbook in English (1905), which was used in the US and in the British Empire. Clements believed plant communities were united into a "superorganism," but in 1917H. A. Gleason questioned the existence of biotic communities, and in 1926 he developed "The Individualistic Concept of the Plant Association." Arthur G. Tansley (**Fig. 26**) at Oxford University challenged the concept of a community being a superorganism in "The Use and Abuse of Vegetational Concepts and Terms' in 1935. Gleason's perspective was adopted in the later 1940s by John T. Curtis (**Fig. 27**) in his school of plant ecology at the University of Wisconsin, where they collected quantitative data for gradient analysis to document a "continuum" of plant species distributions.

Emma Lucy Braun (1889–1971) (**Fig. 28**) was one of 50 botanists awarded a Certificate of Merit by a jubilee committee of the Botanical Society of America "for her contribution to our knowledge of the origin and structure of the Eastern American deciduous forest. Her critical evaluation of the works of others, her capacity to observe correctly in the field and to interpret forcefully have given biogeographers a new point of departure." One of her papers, "The Development of Association and Climax Concepts, Their Use in Interpretation of the Deciduous Forest" is in 50 years of Botany: Golden Jubilee Volume of the Botanical Society of America (1958). For a 1994 symposium on "The Eastern Deciduous Forest since E. Lucy Braun 1950," Ronald Stuckey compiled an 83-page booklet of biographical accounts, maps, and photographs. Lucy published four books and 180 articles. She was vice president of ESA in 1938 and president in 1950—the first woman to hold either office.

She was also the first woman president of the Ohio Academy of Science (1933–34).

In plant ecology, Verona Margaret Conway (1910–1986) (**Fig. 29**) was one of the initiator of studies linking field observations and experimentation, and applying mathematical analysis to environmental physics. She studied processes changing sedge-dominated vegetation into grassland. She studied submerged parts of sedges and the air-spaces through which gases diffuse, thus going from anatomy to ecology. She explained influence of annual temperature changes on waterlogged peat, leading to blanket peats.

Margaret Bryan Davis (1931–2014) (**Fig. 30**) study, "Climatic Changes in Southern Connecticut Recorded by Pollen Deposition at Rogers Lake" (1969) was the only paper by a woman included in Foundations of Ecology: Classic Papers with Commentaries (Real and Brown 1991:650–663). Her social skills were evident in popularity as an advisor to graduate students and her service on national and international committees, leading to her election as president of the American Quaternary

Fig. 20 Johannes Eugenius Warming (1841–1924). Courtesy of Hunt Institute for Botanical Documentation.

Association (1978), National Academy of Sciences (1982) and ESA (1987–88). She was the third woman to serve as ESA president, and she also received its Eminent Ecologist Award (1993). She was also a mentor of women students and faculty members, having spent a quarter of her scientific efforts on these issues.

In limnology there was a strong interest in classifying lakes according to their biological productivity and in measuring their physical and chemical qualities. American leadership was by Edward A. Birge and Chauncey Juday at the University of Wisconsin; for many Wisconsin lakes, they studied temperature stratification, light penetration, dissolved minerals, and hydrogen ion concentration. Arthur D. Hasler (**Fig. 31**), who succeeded them, emphasized experimental limnology and fish ecology. G. Evelyn Hutchinson (**Fig. 32**) of Yale University published an encyclopedic *Treatise on Limnology* (1957–67). In Europe, the leading limnologist was August Thienemann (**Fig. 33**), at the Hydrobiologische Anstalt at Lake Plön and at the University of Kiel. He studied the physical qualities of north German lakes and their invertebrate fauna, and in 1928–29, to make comparisons, he led an expedition to Indonesia to study life in tropical lakes.

Emmaline Moore (1872–1963) (**Fig. 34**) led a survey of the physical, chemical, and biological aspects of Lake George, to determine how to increase fish productivity. That survey was successful and next came a survey of the 60,000 square miles of the New York watershed. In 1926 she became Director of the Biological Survey and chief aquatic biologist. The Survey performed biological surveys of lakes and rivers (1927–40), all published, which in 1996 were still the most comprehensive surveys of any state's water and aquatic resources. In 1928 she became the first woman president of the American Fisheries Society.

The first attempts to organize animal ecology was by Charles C. Adams, who taught animal ecology at the University of Illinois and published a *Guide to the Study of Animal Ecology* (1913), and by Victor E. Shelford (**Fig. 35**), who taught animal ecology at the University of Chicago and published *Animal Communities in Temperate America as Illustrated in the Chicago Region* (1913). A more sweeping synthesis was by Charles Elton (**Fig. 36**) of Oxford University, whose *Animal Ecology* (1927) explained the dynamics of food chains and transformed Joseph Grinnell's descriptive concept (1917) of the niche of California thrashers into a functional concept.

In 1939, Clements and Shelford collaborated on a textbook, *Bio-Ecology*, which sought to bridge the gaps between their specializations. It helped, though the specializations persisted. In 1949, five Chicago animal ecologists—Warder C. Allee, Alfred E.

Fig. 21 Emilie Snethlage (1868–1929).

Fig. 22 Libbie Henrietta Hyman's (1888–1969).

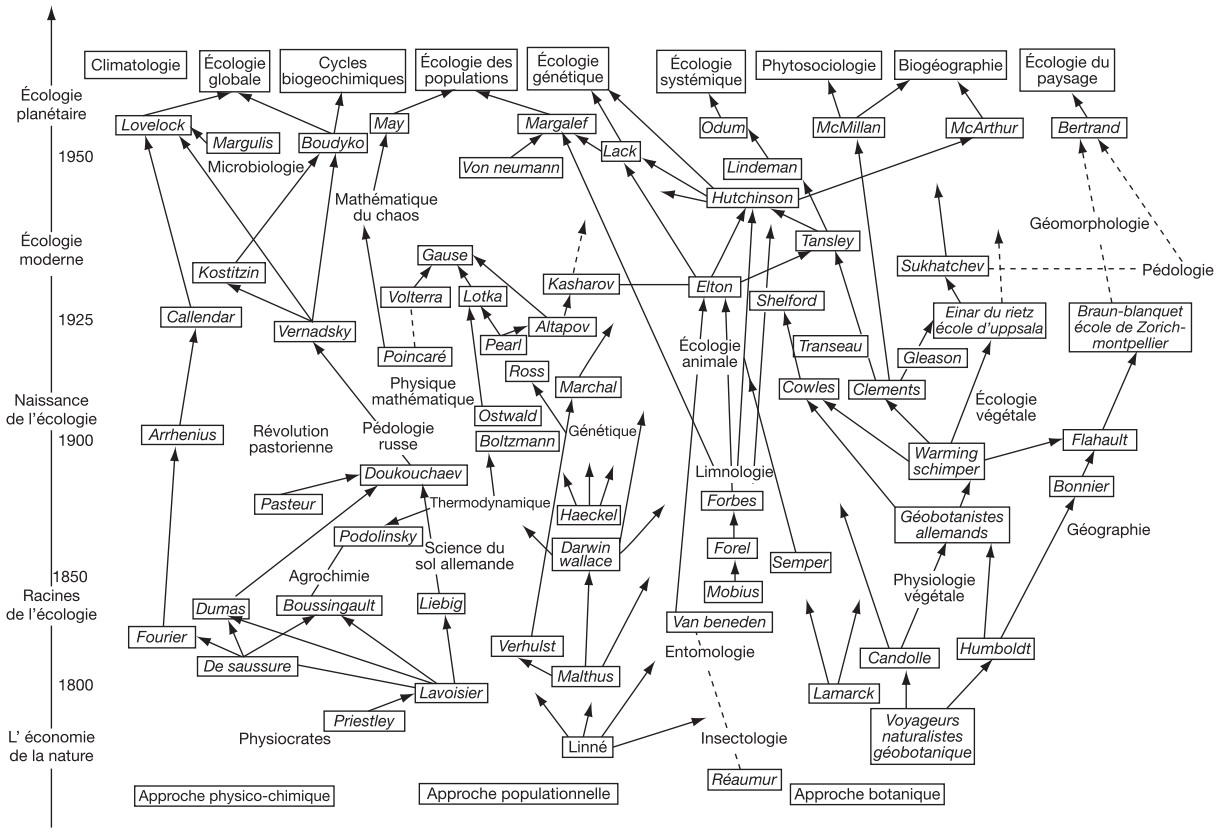


Fig. 23 Some roots and branches of the tree of ecological knowledge (Deléage 1992). It includes names of some ecologists not mentioned in this article. Courtesy of Éditions la Découverte.

Emerson, Orlando Park, Thomas Park, and Karl P. Schmidt—collaborated on the most detailed and comprehensive synthesis ever written (including two historical chapters), *Principles of Animal Ecology*, which emphasized cooperation in nature.

Marine ecology focused on estimating the abundance of plankton, at the bottom of the food chain, and fish at the top, seeking in both cases causes of fluctuation in abundance.

Marie Lebour (1876–1971) (**Fig. 37**) was an expert on planktonic stages of marine animals, pioneering the use of a newly invented plunger jar for studying eggs and larvae of krill in the North Atlantic, Bermuda, and Antarctica. She specialized on dinoflagellates and diatoms, publishing the first two comprehensive books in English on them: The dinoflagellates of Northern Seas (1925) and The Planktonic Diatoms of Northern Seas (1930). She joined the staff of the Marine Biological Association's Laboratory at Plymouth in 1915, where she remained until 1946, then an honorary staff member until 1964. Marie Lebour discovered at least 28 new marine species and published more than 175 scientific papers.

Ruth Dixon Turner (1914–2000) (**Fig. 38**) received her PhD from Harvard University in 1954, with a dissertation on Teredinidae (shipworms, marine bivalve molluscs), which remains a standard work to this day. In 1971, Ruth Turner became the first woman to dive in the deep-sea submersible ALVIN. This was only the first of many deep-sea dives to study long-term biodeterioration and species diversity in the deep-ocean. In 1976 she became Professor of Biology—being one of Harvard's first tenured women professors-, Curator of Malacology at the Museum of Comparative Zoology and joint editor of the journal *Johnsonia*. In 1992, Ruth Turner—who continued to SCUBA dive until her 70s—received the Diver of the year award from the Boston Sea Rovers. She was also named Woman Pioneer in Oceanography by the Woods Hole Oceanographic Institution, an award given only once before—to Turner's long-time hero, Mary Sears (1905–97), also of Harvard University. Ruth Turner has over 200 scientific publications.

Eugenie Clark (1922–2015) (**Fig. 39**) was an expert on shark behavior and tropical fishes. She was a pioneer in the field of SCUBA diving for research purposes. Her first memoir, Lady with a Spear (1953) described her adventures collecting fish in the South Pacific and in the Red Sea. It was very popular and translated into seven languages. Her second book, The Lady and the Sharks (1969) described her experiences there (1955–66). She then spent the rest of her career on the faculty of the University of Maryland.

In France Katharina Mangold-Wirz (1922–2003) (**Fig. 40**) was a pioneer in the domain of embryonic development and life cycle in relation to environmental conditions, that she applied to cephalopods. Her work on cephalopod ecology led the way for other marine eco-physiologists. After her death, two new cephalopod species were named after her: *Microeledone mangoldi* (2004) and *Asperoteuthis mangoldae* (2007).

Fig. 24 Frederic Edward Clements (1874–1945). Courtesy of Hunt Institute for Botanical Documentation.

Fig. 25 Henry Chandler Cowles (1869–1939).

Fig. 26 Arthur George Tansley (1871–1955). Courtesy of Hunt Institute for Botanical Documentation.

Fig. 27 John Thomas Curtis (1913–61). Courtesy of Hunt Institute for Botanical Documentation.

Advances in commercial fishing technology after the Second World War created an industry that could deplete oceans and great lakes of their resources, and ecologists and fishery biologists were needed to explain the level of harvesting that can be sustained. The same is true for shellfish, which must also be protected from pollution. Hasler's study of homing instinct in salmon facilitated development of "salmon farming."

Tansley named and defined the "ecosystem" concept in 1935 but did not lay its foundation. (Instead, he completed his large monograph, *The British Islands and Their Vegetation*, in 1939.) Raymond L. Lindeman, a Midwestern postdoctoral student under Hutchison, developed the ecosystem concept in the most important ecological article ever published, "The Trophic-Dynamic Aspect of Ecology" (1942). It includes both original research and a synthesis of previous literature. Lindeman died, but Hutchinson got it published. After the Second World War, the brothers Eugene and Howard Thomas Odum (Fig. 41) led in studies on

Fig. 28 Emma Lucy Braun (1889–1971).

Fig. 29 Verona Margaret Conway (1910–86).

Fig. 30 Margaret Bryan Davis (1931–2014).

Fig. 31 Edward Asahel Birge (1851–1950), Chauncey Juday (1871–1944), and Arthur Davis Hasler (1908–2001).

the productivity of ecosystems. Eugene Odum also published a very popular textbook, *Fundamentals of Ecology* (1953, 3rd edn. 1971) that publicized ecosystem concepts. His ecology school at the University of Georgia studied the ecosystem within the vast US Atomic Energy Commission's Savannah River Installation Area. H. T. Odum pioneered ecosystem modeling, and today modeling is fully integrated into ecology and applied ecology due to important contributions from Denmark, the Netherlands, Sweden, Switzerland, and the US. Recent integrated ecosystem theory is indebted to H. T. Odum (maximum power and energy), R. Ulanowicz (ascendency), B. C. Patten (network theory), and thermodynamics (S. E. Jørgensen). R. O'Neill and Tim Allen developed hierarchy theory.

Evelyn Christine Pielou (1924–2016) (Fig. 42) pioneered mathematical ecology, with methods, indices, and hypothesis tests and applied her methods to organisms and ecosystems from boreal forests to intertidal algae. She linked mathematical ecology, biogeography, and paleoecology. Her awards included George Lawson Medal (Canadian Botanical Association, 1984), Eminent

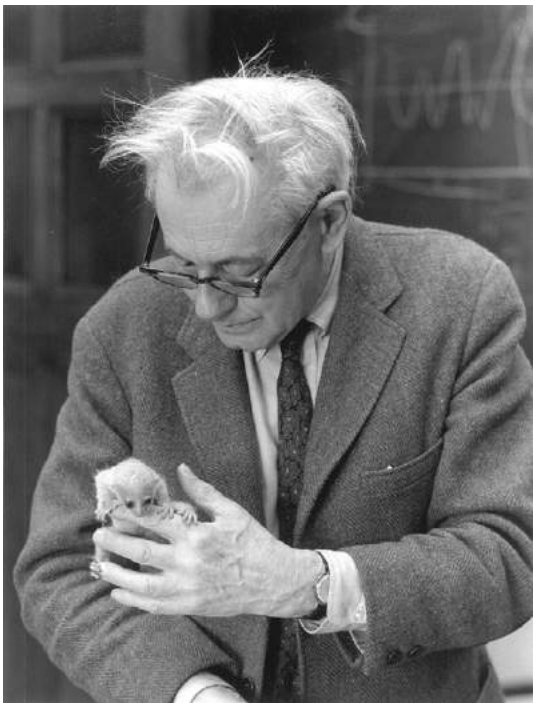


Fig. 32 George Evelyn Hutchinson (1903–91). Courtesy of Yale Peabody Museum of Natural History.

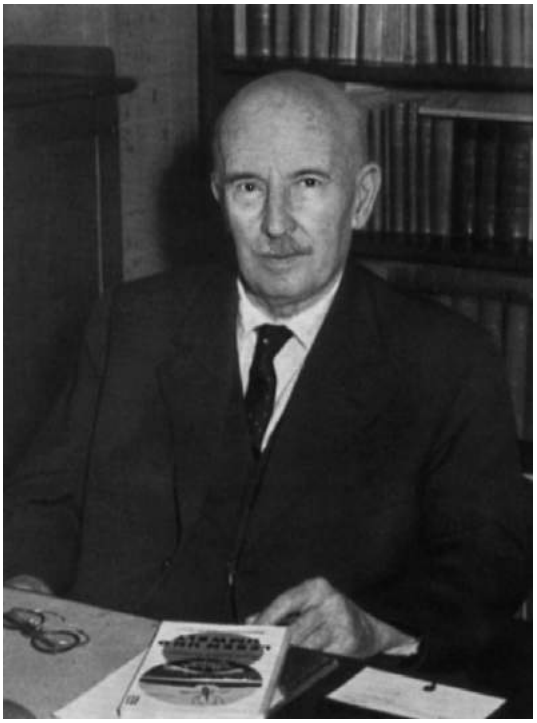


Fig. 33 August Friedrich Thienemann (1882–1960).

Ecologist Award (Ecological Society of America, 1986), and Distinguished Statistical Ecologist (International Congress of Ecology, 1990). The books she published e.g. Introduction to Mathematical Ecology (1969), Population and community ecology: principles and methods (1974), Ecological diversity (1975) served as a basis for learning mathematical ecology for generations of ecology researchers.

Fig. 34 Emmaline Moore (1872–1963).

Fig. 35 Victor Ernest Shelford (1877–1968).

Population studies focused primarily on causes of fluctuations in abundance and on developing accurate mathematical models to describe and predict these fluctuations. The studies were mainly on mammals, birds, fish, and insects. Charles Elton led members of his Bureau of Animal Population (1932–67) into field work and theoretical debate at Oxford University. He discredited the notion of a balance of nature and replaced it with the idea of constant change. Aspects of ethology, the study of

Fig. 36 Charles Elton (1900–91).

Fig. 37 Marie Lebour (1876–1971).

Fig. 38 Ruth Dixon Turner (1914–2000).

Fig. 39 Eugenie Clark (1922–2015).

Fig. 40 Katharina Mangold-Wirz (1922–2003).

animal behavior, are relevant to animal ecology. The three founders of ethology in the 1930s (Konrad Lorenz, Niko Tinbergen, and Karl von Frisch) received a Nobel Prize in 1973.

The Russian geochemist Vladímir Ivanovich Vernadsky (Fig. 43) studied organisms as the source of atmospheric gases and founded the science of biogeochemistry, explained in *La biosphère* (Russian 1926, French 1929, English 1997). Biogeochemistry has remained a dominant aspect of Soviet/Russian ecology. A similar (but more radical) idea to Vernadsky's is James Lovelock's Gaia theory (1972), which postulates that the Earth is a superorganism that regulates its life forms to maintain an environment that supports life. Gaia theory has attracted popular support and continues to intrigue some Earth scientists, as seen in *Scientists Debate Gaia* (2004).

Fig. 41 Gene Elden Likens (1935–), Howard Thomas Odum (1924–2002), Eugene Plesants Odum (1913–2002), and William Eugene Odum (1942–91). Courtesy of Betty Jean Craige and University of Georgia Press.

Fig. 42 Evelyn Christine Pielou (1924–2016).

Ramon Margalef became the leading ecologist of Spain and the Spanish-speaking world with his 950-page *Ecologia* (1974, 4th edn. 1986), though he also published papers and *Perspectives in Ecological Theory* (1968) in English. Michael H. Graham, Paul K. Dayton, and Robert T. Paine have debated in *Ecology* (June 2002) whether ecological concepts are advanced according to a Popperian or Kuhnian model of science or by evolution of a paradigm—showing that a definitive answer probably requires a book-length study.

In coastal ecology, Deborah Rabinowitz (1947–87) (**Fig. 44**) focused her research in mangrove vegetation. Her experiments challenged a previous paradigm that mangrove swamps were zoned according to different tolerances of species for physical conditions. She made important findings on coexistence, species rarity, and the role of dispersal in maintaining species coexistence. She earned her PhD at the University of Chicago (1975) and became assistant professor in the Department of Ecology and Evolutionary Biology, University of Michigan. In 1982 she became a tenured associate professor at Cornell University, until she died of cancer at the age of 40.

Fig. 43 Vladimir Ivanovich Vernadsky (1863–1945).

Fig. 44 Deborah Rabinowitz (1947–87).

Societies and Journals

The founding of societies and journals is an important indicator of scientific progress. The International Council for the Exploration of the Sea, founded in 1902 with headquarters in Copenhagen, pioneered in relating oceanography to fisheries. Its publications were *Journal du Conseil* and *Rapports et Procés Verbaux*. The first ecological society arose in Britain. Tansley and William Smith had formed the British Vegetation Committee in 1904, which was very active but had difficulty finding funds to publish its memoirs and maps. They decided that an ecological society would be more successful in attracting funds. The British Ecological Society held its first meeting on 12 April 1913, and the first issue of its *Journal of Ecology* was printed for that meeting. Tansley edited it, beginning from 1916. British animal ecology developed more slowly, and only in 1932 did the British Ecological Society begin publishing its *Journal of Animal Ecology*, edited by Elton. The Ecological Society of America (encompassing United States and Canada) was founded in 1915; in 1917, it began publishing its *Bulletin* and in 1920 its main journal, *Ecology*.

In 1922, Thienemann and Einar Naumann, a phytoplankton specialist at the University of Lund, founded the Internationale Vereinigung für Limnologie (later called Societas Internationalis Limnologiae), which meets biannually and publishes its proceedings. In 1936, Americans and Canadians founded the Limnological Society of America, which in 1948 became the American Society of Limnology and Oceanography.

It began publishing *Limnology and Oceanography* in 1956. Since ecological papers were published in botanical, zoological, and general science journals, not all countries formed ecological societies, and some countries did so only later. Germany's Gesellschaft für Okologie was founded in 1970 and Italy's Società Italiana di Ecologia in 1976. More than a dozen specialized ecological journals in English have begun since 1960.

The development of ecosystem models by H. T. Odum and his associates encouraged Sven Erik Jørgensen to lay the foundation for the journal *Ecological Modeling* in 1975. Volume 1 had 320 pages; current volumes contain some 4000 pages per year. This increase is illustrative of the explosion in all aspects of applied ecology since the last decades of the 1900s. Subjects initially published in *Ecological Modeling* soon acquired even more specialized journals: *Ecological Economics* (1988), *Ecological Engineering* (1992), *Ecosystem Health* (2001), *Ecological Complexity* (2004), *Ecological Informatics* (2006), and others.

Institutions

Ecology has flourished in both research institutions and university departments. The *World Directory of Hydrobiological and Fisheries Institutions* which Robert W. Hiatt compiled in 1963 provides information on hundreds of institutions, and their numbers have increased considerably since then. Recent lists of aquatic and terrestrial institutions are in Europa Publications' annuals, entitled *The World of Learning*. Here are just two American examples: the Marine Biological Association of San Diego established its laboratory in 1903, and it became the Scripps Institution of Oceanography in 1925; the Carnegie Institution of Washington in 1903 established the Carnegie Desert Botanical Laboratory outside Tucson; its funding declined during the Depression and in 1940 it was given to the US Forest Service, which sold it in 1960 to the University of Arizona. There are histories of several ecological schools: the plant ecology schools at the Universities of Nebraska and Chicago (Tobey); the Chicago animal ecology school (Mitman); each of the three schools at Oxford University: Tansley's plant ecology school (Anker), Elton's Population Bureau (Crowcroft), and Tinbergen's ethology school (Kruuk, Thorpe); the schools of plant ecology (Fraelish et al.) and limnology (Becker and Egerton) at the University of Wisconsin-Madison; and Eugene Odum's ecosystem school at the University of Georgia (Barrett and Barrett; Craige).

Conservation and Environmentalism

The role of natural history museums for conservation and understanding of ecosystems has been preponderant throughout time.

Maria Corinta Ferreira (1922–2003) (**Fig. 45**) was an entomologist and simultaneously the head of a museum. After graduating in Lisbon (1945), she moved to Mozambique in 1949. In Africa she proceeded her research on entomology, which included field work for the collection of specimens, targeting the enrichment and diversification of the museum's insect collections and increasing the knowledge on the entomological fauna of Mozambique. As the full responsible of Museum of Natural History of Mozambique (MAC) she introduced the dual organizational criteria, which combined an exhibition for the general public (public exhibition), with one aimed at a specialized audience (reserve collection). Until 1974, she also planned and implemented a network of individual and institutional contacts, which allowed to disseminated collections throughout international museums, particularly, in South Africa and Europe. She published more than 120 published papers in English and Portuguese, monographies on Coleoptera, catalogues of new species, among other types of scientific and outreach documents. She was also a brilliant scientific illustrator with over 3000 morphological drawings.

Fig. 45 Maria Corinta Ferreira (1922–2003).

The botanist Ada Hayden (1884–1950) (Fig. 46) specialized in prairies and especially in anatomical adaptions and diversity of prairie plants. She used her research to engage in actions for prairie conservation. Her teaching was remarkable. She was first an Instructor (1910–18), and then Assistant Professor (1919–50) of Botany. She was deeply involved with the Iowa State College Herbarium, from the 1930s until her death. She added to its collection more than 40,000 specimens. She also worked to establish criteria for choosing which prairies should be preserved. From more than 100 tracts, she identified 32 as potential preserves. The College Herbarium and a prairie preserve in Howard County, Iowa, were named in her honor.

Ruth Myrtle Patrick (1907–2013) (Fig. 47) was an aquatic biologist and educator widely regarded as one of the early pioneers of the science of limnology. She was a pioneer on translating her knowledge of river biotic communities into a diagnostic tool for detecting river pollution and on deploying multidisciplinary teams of researchers to assess and monitor aquatic systems. She joined the National Academy of Sciences in 1970 and in 1973 she became chairman of the board at the Academy in Philadelphia. In 1975 she won the John and Ann Tyler Ecology Award, which, at $150,000, was the most generous award in science.

After World War II, both conservationists and ecologists became concerned with negative aspects of the rapid human transformation of the environment. This was an opportunity for ecologists to advise on how to minimize environmental damage and preserve natural areas. This opportunity was first realized in the United Kingdom with the establishment of the Nature Conservancy in 1949. It both preserves natural areas and provides for ecological research in them. In the United States, the triple threats of atomic radiation, phosphate detergents, and dichlorodiphenyltrichloroethane (DDT) insecticide transformed pre-war conservationists into post-war environmentalists. Barry Commoner led the fight against atomic radiation and phosphate detergents, and Rachel Carson (1907–64) (Fig. 48) led the fight against DDT. The outcome was passage of the National Environmental Policy Act (NEPA, 1969), requiring environmental impact assessments before major projects modify natural areas, and presidential establishment of the US Environmental Protection Agency (EPA, 1970), which monitors compliance with environmental laws. Both of these require expert advice from ecologists. Other countries have more or less followed the British and American examples.

Dian Fossey (1932–85) (Fig. 49) earned a B.A. degree (1954) in occupational therapy at San José State University. In 1963, she spent 7 weeks in Africa, where she met Louis and Mary Leakey, who introduced her to the first contact with wild gorillas. Back in USA, she wrote three articles on her trip. Later, she showed her articles to Louis Leakey who lectured in Louisville. He suggested she come to Africa, with his assistance and study gorillas. She did, and began by visiting Jane Goodall at Gombe and observed her studies on chimpanzees. She founded Karisoke Research Center, a remote rainforest camp for studying mountain gorillas. She became the world authority on their physiology and behavior. In doing so, she imitated their behavior and drew pictures of their "noseprints" for identification. She traveled to Cambridge University, UK, and earned a PhD in 1980, and lectured at Cornell University in 1981–83. She achieved considerable fame with her book, Gorillas in the Mist (1983). She not only studied gorillas but also worked to protect them, which may have led to her murder in her camp on December 26, 1985.

Fig. 46 Ada Hayden (1884–1950).

Fig. 47 Ruth Myrtle Patrick (1907–2013).

Fig. 48 Rachel Louise Carson (1907–64).

Disease Control

Ecology research is crucial to understand the biological vectors associated with infectious diseases and, thus, crucial for disease and plague control.

Ann Bishop (1899–1990) (Fig. 50) was a pioneer in malaria research. She specialized in protozoology and parasitology and studied Plasmodium, the malaria parasite and its vector, the mosquito *Aedes aegypti*. She tested various anti-malarial chemicals, then studied drug resistance in both parasites and vectors. In 1932 she earned D.Sc. from Manchester University for research on

Fig. 49 Dian Fossey (1932–85).

Fig. 50 Ann Bishop (1899–1990).

blackhead parasites. She was elected to membership in the Royal Society in 1959 and later joined the Malaria Committee of the World Health Organization. In 1992, the British Society for Parasitology created Ann Bishop Traveling Award, to aid young parasitologists to study parasites where they were endemic.

IBP

The success of research conducted during the International Geophysical year (1957–58) inspired three prominent scientists in 1959 to begin planning the International Biological Program, organized by the International Union of Biological Sciences. Barton Worthington (Fig. 51), head of Britain's Nature Conservancy, became its scientific head. Nationally supported IBP research lasted from 1964 to 1974, and some projects continued beyond 1974. The projects were mainly ecological. In the United States, significant new funding enabled ecology to evolve into "big science." George M. Van Dyne and others used a method of monitoring ecosystems with computers called "systems ecology." Without new funds, the USSR contributed about as much IBP research and publications as did the USA. Worldwide, IBP held over 200 meetings, and its research cost over $40 million annually. The resulting publications were numerous and generally of high caliber.

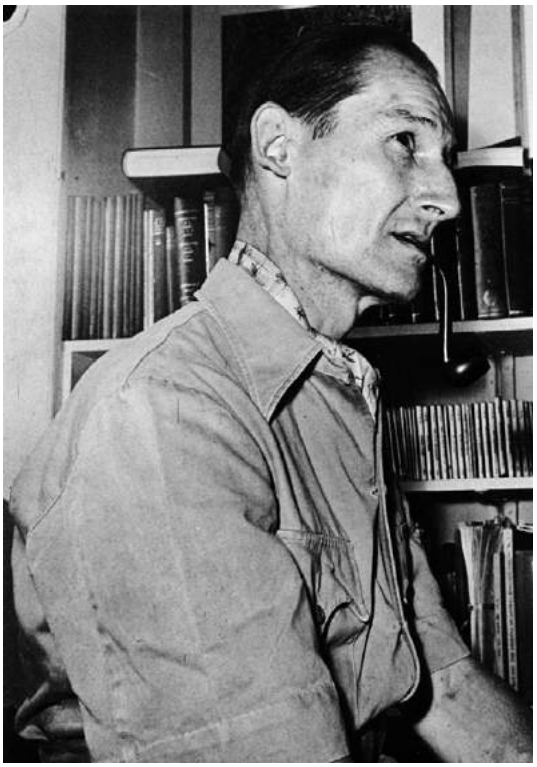


Fig. 51 Edgar Barton Worthington (1905–2001).

Conclusion

Ecology began with natural history observations and a balance of nature concept. It expanded cumulatively through the ages until Linnaeus and Buffon began to expand the theoretical aspects of natural history in mid-1700s. During 1800s ecological disciplines began to emerge, beginning with Humboldt's plant geography. Natural history was revolutionized by Darwin's theory on the origin of species, leading to Haeckel's coining of the word *oecologie* in 1866 and a gradual elaboration of limnology, marine biology, plant ecology, and animal ecology by the early 1900s. During the 1900s, additional specializations emerged (including population ecology and ecosystem ecology), ecological organizations and journals were founded, and ecological professorships were added by universities. A sophisticated science developed well before 2000. Ecology now has the broadest scope of all the sciences, and applied ecology specialties flourish.

Further Reading

Acot, P., 1994. Histoire de l'écologie, 2nd edn Paris: Presses Universitaires de France.

Acot, P., 1998. The European origins of scientific ecology (1800–1901). Amsterdam: Editions des Archives Contemporaines/Gordon and Breach.

Allcock, A.L., von Boletzky, S., Bonnaud-Ponticelli, L., Brunetti, N.E., Cazzaniga, N.J., Hochberg, E., *et al.*, 2015. The role of female cephalopod researchers: Past and present. Journal of Natural History 49 (21–24), 1235–1266.

Anker, P., 2001. Imperial ecology: Environmental order in the British Empire, 1895–1945. Cambridge, MA: Harvard University Press.

Anonymous, 1994. Citations for Honorary Members: E. C. ("Chris") Pielou. ESA Bulletin 25, 68–69.

Antunes, L.P., 2016. Maria Corinta Ferreira (1922–2003?): Naturalist at the Museu Dr. Álvaro de Castro, Lourenço Marques [now Maputo], Mozambique, 1949–1974. HoST—Journal of History of Science and Technology 10, 103–124. doi:10.1515/host-2016-0005.

Barrett, G.W., Barrett, T.L. (Eds.), 2001. Holistic science: The evolution of the Georgia Institute of Ecology, 1940–2000. New York: Taylor and Francis.

Beckel, A., Egerton, F.N., 1987. Breaking new waters—A century of limnology at the University of Wisconsin. Madison, WI: Wisconsin Academy of Sciences, Arts and Letters.

Bentley, B., 1987. Eminent ecologist: E. C. Pielou. ESA Bulletin 68, 30–31.

Bishop, A., Gilchrist, B.M., 1946. Experiments upon the feeding of *Aëdes aegypti* through animal membranes with a view to applying this method to the chemotherapy of malaria. Parasitology 37 (1–2), 85. doi:10.1017/S0031182000013202.

Blair, W.F., 1977. Big biology: The US/IBP. Stroudsburg: Dowden, Hutchinson and Ross.

Bocking, S., 1997. Ecologists and environmental politics: A history of contemporary ecology. New Haven: Yale University Press.

Brown, M.T., Hall, C.A.S., 2004. Through the macroscope: The legacy of H. T. Odum. Ecological Modelling 178, 1–294.

Burgess, R.L., 1996. American ecologists: A biographical bibliography. Huntia 10, 5–116.

Burkhardt Jr., R.W., 2005. Patterns of behavior: Konrad Lorenz, Niko Tinbergen, and the Founding of Ethology. Chicago: University of Chicago Press.

Cittadino, E., 1990. Nature as the laboratory: Darwinian plant ecology in the German Empire, 1880–1900. Cambridge: Cambridge University Press.

Craige, B.J., 2001. Eugene Odum: Ecosystem ecologist and environmentalist. Athens, GA: University of Georgia Press.

Cramer, J., 1987. Mission-orientation in ecology: The case of Dutch fresh-water ecology. Amsterdam: Rodopi.

Croker, R.A., 1991. Pioneer ecologist: The life and work of victor Ernest Shelford, 1877–1968. Washington: Smithsonian Institution Press.
Croker, R.A., 2001. Stephen Forbes and the rise of American ecology. Washington: Smithsonian Institution Press.
Crowcroft, P., 1991. Elton's ecologists: A history of the bureau of animal population. Chicago: University of Chicago Press.
Dajoz, R., 1984. Eléments pour une histoire de l'écologie: La naissance de l'écologie moderne au XIXe siècle. Histoire et Nature 24–25, 5–111. 6 plates.
Decacon, M., 1997. Scientists and the sea, 1650–1900: A study of marine science. UK: Ashgate Aldershot.
Deléage, J.-P., 1993. Histoire de l'écologie, 2nd edn. Paris: Seuil.
Drouin, J.-M., 1993. L'écologie et son histoire, 2nd edn. Paris: Flammarion.
Egerton, F.N., 1962. The scientific contributions of François Alphonse Forel, the founder of limnology. Schweizerische Zeitschrift für Hydrologie 24, 181–199.
Egerton, F.N., 1973. Changing concepts of the balance of nature. Quarterly Review of Biology 48, 322–350.
Egerton, F.N., 1977. History of American Ecology. New York: Arno Press.
Egerton, F.N., 1983. The history of ecology: Achievements and opportunities (Part 1). Journal of the History of Biology 16, 259–311.
Egerton, F.N., 1985. The history of ecology: Achievements and opportunities (Part 2). Journal of the History of Biology 18, 103–143.
Egerton FN (2001 A history of the ecological sciences. *Bulletin of the Ecological Society of America*, quarterly, part 61A; http://esapubs.org/esapubs/journals/bulletin.htm. (accessed January 2007).
Egerton, F.N., 2003. Hewett Cottrell Watson: Victorian Plant Ecologist and Evolutionist. Aldershot: Ashgate.
Egerton, F.N., 2007. Understanding food chains and food webs, 1700–1900. Bulletin of the Ecological Society of America 88, 50–69.
Egerton FN (2012) *Roots of Ecology: Antiquity to Haeckel*. Berkeley, University of California Press.
Egerton FN (2015) *A Centennial History of the Ecological Society of America*. Boca Raton, CRC Press/Taylor and Francis.
Ferreira M.C. 1954. Biological reconnaissance of the territories around the Indian Ocean. Proceedings of the Pan Indian Ocean Science Congress. (2nd), [Perth, W.A.]
Fossey, D., Harcourt, A.H., 1977. Feeding ecology of free-ranging mountain gorilla (Gorilla gorilla beringei). In: Clutton-Brock, T. (Ed.), Primate ecology: Studies of feeding and ranging behaviour in lemurs, monkeys and apes. London: Academic Press, pp. 415–447.
Fossey, D., 1983. Gorillas in the mist. New York: Houghton Mifflin Company.
Fralish, J.S., McIntosh, R.P., Loucks, O.L., John, T., 1993. Curtis: Fifty years of Wisconsin plant ecology. Wisconsin Academy of Sciences, Arts and Letters Madison, WI.
Frey, D.G., 1963. Limnology in North America. Madison, WI: University of Wisconsin Press.
Glaser, J.R., Zenetou, A.A. (Eds.), 1994. Gender perspectives: Essays on women in museums. Washington, D.C.: Smithsonian Institution press.
Golley, F.B., 1993. A history of the ecosystem concept in ecology. Yale University Press, New Haven, CT.
Hagen, J.B., 1992. An entangled Bank: The origins of ecosystem ecology. Rutgers University Press, New Brunswick, NJ.
Haines, C.M., 2001. International women in science: A biographical dictionary to 1950. Santa Barbara, CA: Abc-clio.
Höxtermann, E., Kaasch, J., Kaasch, M., 2001. Verhandlungen zur Geschichte und Theorie der Biologie, 7: Berichte zur Geschichte und Theorie der Ökologie. Deutsche Gesellschaft für Geschichte und Theorie der Biologie Neuburg an der Donau. Neuburg an Donau: Deutsche Gesselschaft fur Geschichte und Theorie der Biologie.
Hutchinson, G.E., 1979. The kindly fruits of the earth. New Haven: Yale University Press.
Joseph, L.E., 1990. Gaia: The growth of an idea. New York: St. Martin's Press.
Kingsland, S.E., 1995. Modeling nature: Episodes in the history of population ecology, 2nd edn. Chicago: University of Chicago Press.
Kingsland, S.E., 2005. The evolution of American ecology, 1890–2000. Baltimore: Johns Hopkins University Press.
Kormondy, E.J., McCormick, J.F., 1981. Handbook of contemporary developments in world ecology. Westport, CT: Greenwood Press.
Kruuk, H., 2003. Niko's nature: The life of Niko Tinbergen and his science of animal behaviour. Oxford: Oxford University Press.
Kwa, C.L., 1989. Mimicking nature: The development of systems ecology in the United States, 1950–1975. Amsterdam: University of Amsterdam.
Langenheim, J.H., 1996. Early history and progress of women ecologists: Emphasis upon research contributions. Annual Review of Ecology and Systematics 27, 1–53.
Lebour, M., 1925. The dinoflagellates of Northern seas. Plymouth: Marine Biological Association of the United Kingdom, p. p. 250.
Lebour, M., 1930. The Planktonic diatoms of Northern Seas. London: Ray Society, p. p. 244.
Nicolson, M., 1990. Henry Allan Gleason and the individualistic hypothesis: The structure of a botanist's career. Botanical Review 56, 91–161.
McGinnies, W.G., 1981. Discovering the desert: Legacy of the Carnegie Desert botanical laboratory. Tucson, AZ: University of Arizona Press.
McIntosh, R.P., 1985. Background of ecology: Concept and Theory. Cambridge: Cambridge University Press.
Mills, E.L., 1989. Biological oceanography: An early history, 1870–1960. Ithaca, NY: Cornell University Press.
Mitman, G.A., 1992. The state of nature: Ecology, Community and American Social Thought. Chicago: University of Chicago Press.
Nicolson, M., McIntosh, R.P., 2002. H.A. Gleason and the Individualistic Hypothesis Revisited. Ecological Society of America Bulletin 83, 133–142.
Ogilvie, M. 2000. The biographical dictionary of women in science: Pioneering lives from ancient times to the mid-20th century, Vol. 1, 1, Taylor & Francis US, pp. 129–130, ISBN 9780415920384.
Palladino, P., 1996. Entomology, ecology and agriculture: The making of scientific careers in North America, 1885–1985. Amsterdam: Harwood Academic Publishers.
Pigott, D., 1988. Verona Margaret Conway. Journal of Ecology 76, 288–291.
Real, L.A., Brown, J.H., 1991. Foundations of ecology: Classic papers with commentaries. Chicago: University of Chicago Press.
Russell, F.S., 1972. Dr Marie V. Lebour. Journal of the Marine Biological Association of the United Kingdom 52 (3), 777–788. doi:10.1017/S0025315400021718.
Schlee, S., 1973. The edge of an unfamiliar world: A history of oceanography. New York: Dutton.
Sears, M., Merriman, D., 1980. Oceanography: The past. New York: Springer.
Sheail, J., 1987. Seventy-five years in ecology: The British ecological society. Oxford: Blackwell.
Shortland, M., 1993. Science and nature: Essays in the history of the environmental sciences. British Society for the History of Science Oxford.
Simberloff, D., 1988. Conservation biology. The Journal of the Society for Conservation Biology 2 (1), ISSN: 0888–8892; online ISSN: 1523–1739.
Söderqvist, T., 1986. The ecologists: From merry naturalists to saviors of the nation: A sociologically informed narrative survey of the Ecologization of Sweden, 1895–1975. Stockholm: Almqvist and Wiksell.
Snethlage, E., 1912. A travessia entre o Xingú e o Tapajoz. (in Portuguese). vol.7. Belém: Boletim do Museu Goeldi, pp. pp. 49–92.
Snethlage, E., 1914. Catálogo das Aves Amazônicas (in Portuguese). 8. Belém: Boletim do Museu Paraense Emílio Goeldi de Historia Natural e Etnografia, pp. pp. 1–530.
Snethlage, E., 1917. Nature and man in Eastern Pará, Brazil. Geographical Review 4 (1), 41–50.
Steleanu, A., 1989. Geschichte der limnologie und ihrer grundlagen. Frankfurt am Main: Haag and Herchen.
Thorpe, W.H., 1979. The origins and rise of ethology: The science of the natural behaviour of animals. London: Heinemann.
Tobey, R.C., 1981. Saving the prairies: The life cycle of the founding School of American Plant Ecology, 1895–1955. Berkeley, CA: University of California Press.
Trepl, L., 1987. Geschichte der Ökologie vom 17. Jahrhundert bis zur Gegenwart Zehn Vorlesungen. Athenäum: Frankfurt am Main.
Turner, R.D., 1981. "Wood islands" and "thermal vents" as centers of diverse communities in the deep sea. Biologiya Morya 7 (I), 3–10.
Turner, R.D., Lutz, R.A., 1984. Growth and distribution of mollusks at deep-sea vents and seeps. Oceanus 27 (3), 55–62.
Weiner, D.R., 1988. Models of nature: Ecology, conservation, and cultural revolution in Soviet Russia. Bloomington, IN: Indiana University Press.
Worster, D., 1994. Nature's economy: A history of ecological ideas, 2nd edn. Cambridge: Cambridge University Press.
Worthington, E.B., 1975. The Evolution of IBP. Cambridge: Cambridge University Press.
Worthington, E.B., 1984. The ecological century: A personal appraisal. Oxford: Clarendon.

Homeotherms

P Frappell and K Cummings, La Trobe University, Melbourne, VIC, Australia

Introduction

Living organisms experience daily, seasonal, and long-term climatic variations in temperature. Many animals respond to changes in body temperature (T_b) with behavioral adjustments, migration, or changes in metabolism (e.g., torpor). However, for some animals and almost all plants such responses are impossible. Animals that effectively regulate their T_b to a thermal set-point despite larger variations in ambient temperature are referred to as homeotherms (in Greek *homoios* means `same, identical', and *therm* means `heat'). Stenotherms (in Greek *Stenos* means `narrow', and *therm* means `heat') are animals that live in a thermally stable environment and can tolerate only narrow temperature changes. Excellent examples of stenotherms are the Antarctic marine notothenioid fishes and invertebrates that have evolved in waters having an annual temperature variation of less than 2 °C. However, despite the narrow range in T_b of stenotherms, they are not homeotherms as they do not actively regulate T_b. The regulation of a T_b in homeotherms, independent of ambient temperature, is determined by the rates of heat gain and/or loss through a combination of physical and physiological means and, in some cases, by heat production.

Although not necessary for homeothermy, many homeotherms regulate T_b through the retention of metabolically generated heat, that is, endothermy (in Greek *endo* means `within' and *therm* means `heat'). Examples of endothermy have arisen in some plants, insects, fishes, and reptiles, and in all mammals and birds (**Fig. 1**). Many endotherms (e.g., hummingbirds, echidnas, billfishes) show substantial daily or seasonal variation in T_b, which in many cases is associated with variation in the thermal set-point. Among ectotherms (in Greek *ecto* means `outside' and *therm* means `heat') there are examples where homeothermy is achieved through heat gained from the environment and where behavioral thermoregulation coupled with improved heat retention enables thermal stability. What then are the benefits and costs of homeothermy? To appreciate the importance of homeothermy in deciding the metabolic niche space of an animal one needs to consider how metabolic rate and other physical factors such as temperature and body size determine an animal's T_b and subsequent life functions.

Temperature

Temperature is an important environmental factor. It influences the biology of an animal through changes in the rates of biochemical and physiological processes and in the stability of biomolecules. An animal's T_b is the result of thermal balance between the rates of heat gain and loss.

Fourier Equation for Heat Balance

Heat balance of an animal is determined by the net exchange of heat. Heat balance is decided through avenues of heat gain or loss and metabolic heat production. The general equation describing heat balance is

$$H_{prod} = \pm H_{cond} \pm H_{conv} \pm H_{rad} \pm E_{vap} \pm H_{store}$$

H_{prod} is metabolic heat production, H_{cond} is heat transfer by conduction through physical contact, H_{conv} is heat transfer across the fluid boundary layer at the surface of the animal, and H_{rad} is radiative heat transfer between the animal and its surrounds; all depend on temperature gradients, either between the animal's body or surface temperature and the ambient temperature or the temperature of the surrounding objects. E_{vap} is the heat that is lost during evaporation (or heat gained during condensation) and depends on the difference in vapor pressure between the animal and the surrounding air, assuming the air at the animal's surface is saturated. H_{store} is the heat that is gained or lost by a change in T_b, and must equal zero for an organism to be in heat balance and thus maintain a constant T_b.

All avenues of heat transfer depend on the surface area of the animal, being greater the larger the surface area. Therefore, by reducing surface area, thermal conductivity, or the coefficient of radiation, an animal can reduce heat transfer to the environment. For a homeotherm to maintain constant T_b the equilibrium between heat loss and gain or production must be maintained at all ambient temperatures (**Fig. 2**). For an ectotherm at a given T_b, if the properties of heat transfer remain unaltered, there can only be a single ambient temperature at which the rates of heat loss and gain with the environment are in equilibrium. An endotherm, without altering the heat transfer properties, at the same ambient temperature, would have a higher T_b; the exact T_b would be determined when equilibrium is established between the rates of heat production and transfer. Under these conditions, T_b will remain constant over time. At ambient temperatures lower than the point of equilibrium the animal must increase heat gain or production to match heat loss, reduce heat loss or a combination of the two to prevent a decline in T_b. At warmer temperatures, the animal must increase heat loss and reduce heat gain or production to maintain T_b. Many endothermic homeotherms extend the single ambient temperature at which equilibrium between heat production and loss occurs into a zone, the thermoneutral

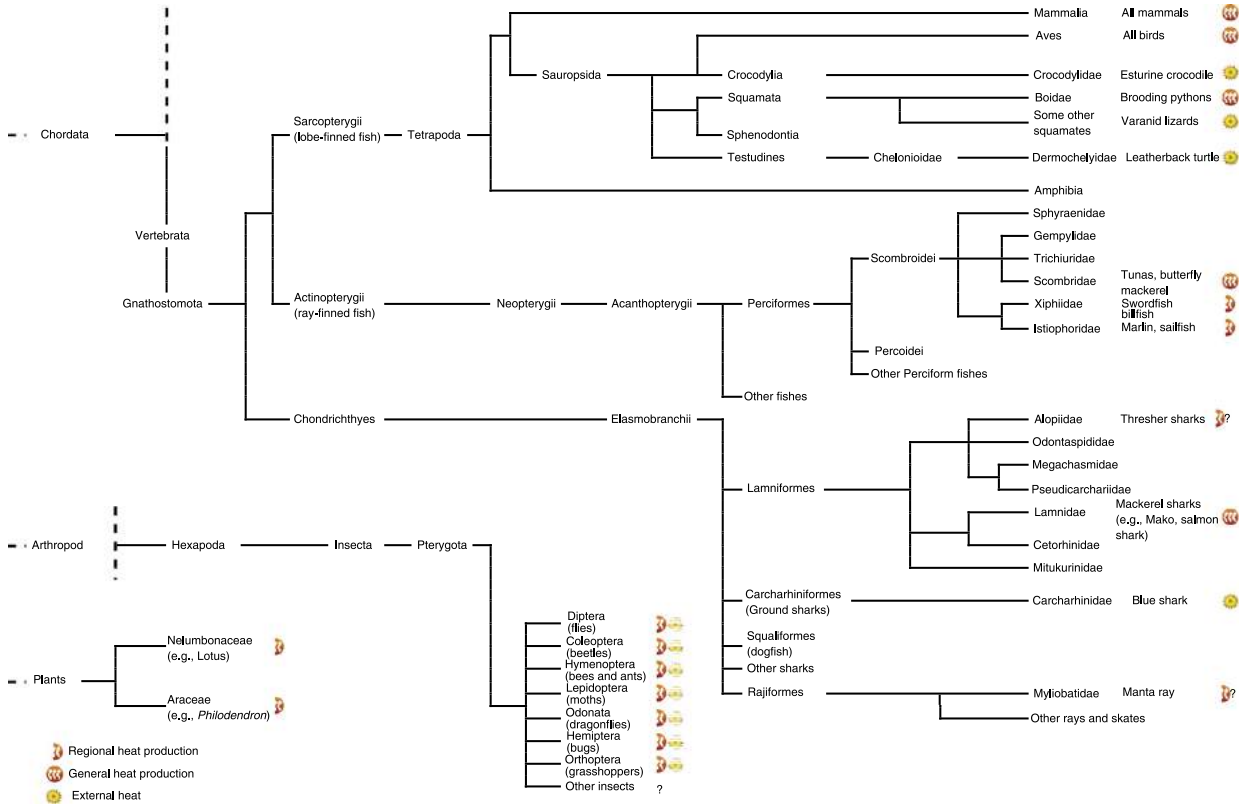


Fig. 1 Cladograms showing proposed phylogenetic relationships among homeothermic organisms. Homeothermy may be achieved in organisms that rely predominantly on external heat obtained from the environment (ectotherms) or in organisms in which the majority of heat is metabolically generated (endotherms). The metabolically generated heat can be either confined to specific areas of the body (e.g., thorax in insects, cranial regions in fish) or more generally distributed and may be purposely generated for thermoregulation (thermogenesis) or as a by-product of other activities.

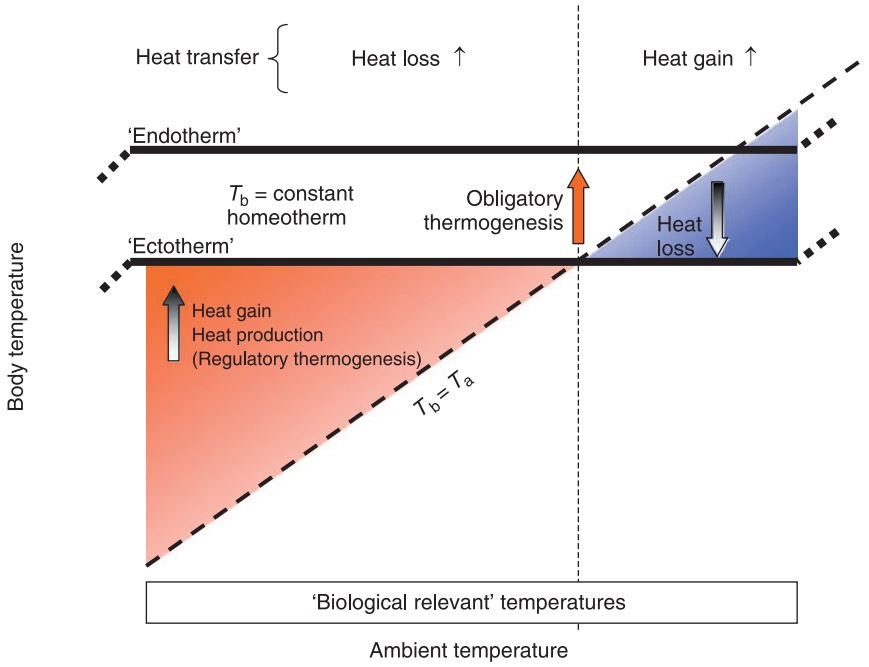


Fig. 2 Body temperature is established as a balance between heat input and heat loss. Heat input occurs through heat transfer or from obligatory or regulatory thermogenesis. Heat transfer, either loss or gain, between the animal and its environment can occur via conduction, convection, radiation, or evaporation/condensation. The rate of heat transfer for each mode is proportional to the surface area and, except for evaporation/conduction, is proportional to the temperature gradient between the animal and the environment. For an `ectotherm' at an established body temperature (T_b) in equilibrium with a given ambient temperature (T_a) heat gain equals heat loss (intersection with $T_b = T_a$). At temperatures on either side of this single T_a, for T_b to remain constant changes in heat transfer need to occur; at lower T_a heat loss will increase and heat needs to be added (or heat loss reduced) and for higher T_a heat gain will increase and heat loss needs to increase (and/or heat input reduced). The increase in cellular metabolism associated with leakier membranes in `endotherms' is coupled with increased heat production (obligatory thermogenesis). The production of heat will result in an increase in T_b to a temperature where heat loss from the increasing T_b–T_a gradient will reestablish equilibrium at the single T_a in which equilibrium was originally established in the `ectotherm'. To maintain $T_b$ with changing $T_a$ again requires adjustments in heat transfer. In the case of an `endotherm', a decrease in T_a is initially met over a narrow temperature range through changes in heat transfer to offset the increasing heat loss (known as the `thermal neutral zone', TNZ), after which further decline in T_a is countered with increased heat production (regulatory thermogenesis).

zone, without activating regulatory heat production. This is achieved by altering heat transfer (e.g., conductance through changes in surface area achieved through posture or peripheral adjustments to the circulation) or by utilizing heat from obligatory heat-generating processes (activity, postprandial metabolism).

Sources of Heat

External heat. Generally, metabolic heat production by ectotherms is negligible. If T_b equals ambient temperature (T_a) then the animal must thermoconform. However, T_b can be elevated above ambient temperature through behavioral adjustments (e.g., basking). In addition, T_b can be elevated by increasing radiative heat gain (e.g., black bodies absorb more radiant heat than white) and/or conductive heat gain (e.g., ventral contact with solar-heated substrate) and can be maintained upon removal of the heat source by minimizing conductive heat loss (e.g., reduced peripheral circulation, addition of insulative fat layers).

Internal heat. Obvious examples of endotherms are mammals and birds but other examples can be found among the reptiles, fish, insects, and plants. In all these examples sufficient heat can be metabolically produced to elevate tissue temperatures above that of the ambient temperature. In many of these organisms the increase in temperature is reflected as a general increase in overall T_b (mammals, birds, tuna, brooding pythons), whereas in others the temperature increase is regional (lamnid sharks, swordfish, many insects, the flowers of some plants). Heat generated from the activity of exercising aerobic skeletal muscle can be retained; this can be an important source of heat in small endotherms and in some endothermic fish. In tuna and lamnid sharks the red muscle is located in a more medial and anterior body position and the perfusion of this muscle enables countercurrent heat entrapment of metabolic heat that accompanies continual locomotion and the maintenance of a stable T_b despite excursions by these fish into cold water (**Fig. 3a**). In lamnid sharks (e.g., Salmon shark) that inhabit cold waters the red muscle has specialized to function within an elevated temperature range (20–30 °C).

In other cases, the skeletal muscles can be activated for nonlocomotory isometric high-speed contractions, that is, shivering, to generate heat. An example is provided by the sphinx moth that requires a thoracic temperature higher than 35 °C for flight,

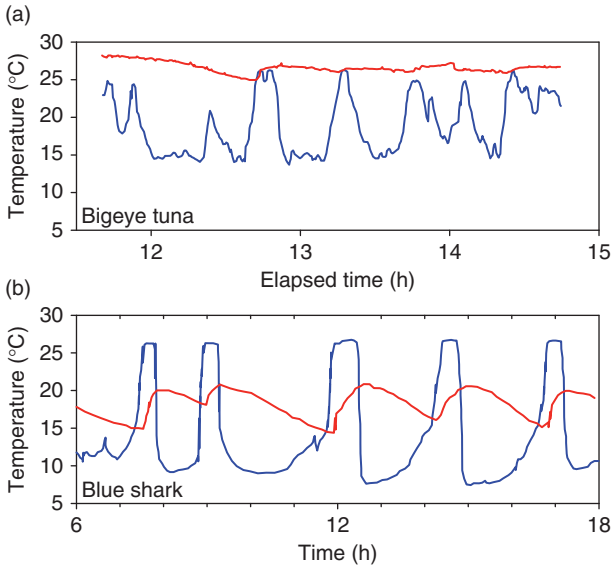


Fig. 3 Body (red line) and water (blue line) temperatures as a function of time in (a) an endothermic fish, the bigeye tuna and (b) a large ectothermic shark. Being an endotherm the tuna has the capacity to elevate T_b above that of the environment, whereas the ectothermic shark adopts a temperature closer to the average water temperature that it is exposed to while diving through the water column. In the case of the shark regulation is dependent on the diving behavior to contain T_b within limits. In the tuna T_b is largely independent of the temperatures encountered. (a) Redrawn from Holland KN and Sibert JR (1994) Physiological thermoregulation in bigeye tuna, *Thunnus obesus*. *Environmental Biology of Fishes*. 40: 319–327. (b) Redrawn from Carey EG and Scharold JV (1990) Movements of blue sharks (*Prionace glauca*) in depth and course. *Marine Biology* 106: 329–342.

achieved by shivering of wing muscles prior to flight. During flight, heat is produced as a by-product of flight metabolism and is independent of ambient temperature across the range of 15–35 °C, yet thoracic temperature is held reasonably constant by adjusting blood flow and hence heat transfer from the pubescent abdomen. At warm temperatures a similar situation occurs in the dragonfly. At cooler T_a the dragonfly and orchid bee, which are both poorly insulated, appear to modulate heat production. In the orchid bee this is achieved by beating their wings at elevated frequencies, reducing flight muscle efficiency, and producing more heat. Either strategy achieves the goal of regulating T_b, particularly the thorax, throughout the flight (**Fig. 4**).

Another mechanism for generating heat is nonshivering thermogenesis (NST). Only two animal tissues are specialized for NST: brown adipose tissue in small eutherian mammals and cranial heater tissue in billfishes and the butterfly mackerel. Brown adipose tissue contains uncoupling protein 1 (UCP 1) that permits futile cycling of the mitochondrial electron transport chain to produce heat without ATP synthesis and degradation. Fish cranial heater tissue have lost their myofibrillar contractile apparatus and participate in futile cycling of Ca^{2+} between the cytoplasm and the sarcoplasmic reticulum which is mediated in the ryanodine receptor by Ca^{2+}-ATPase. In birds it would appear that NST also occurs via Ca^{2+} release uncoupled from ATP synthesis. However, this occurs in the skeletal muscle and is owing to isoforms of the sarcoplasmic/endoplasmic reticulum Ca^{2+}-ATPase not associated with muscle fiber contraction. Interestingly, UCPs have also been identified in plants. However, in these cases most of the heat is generated via an alternative, cyanide-insensitive oxidase pathway that reduces the electrochemical proton gradient across the mitochondrial membrane. These later cases are examples of regulatory thermogenesis.

A further source of obligatory heat can arise from postprandial metabolism following ingestion of a meal (often referred to as specific dynamic action). The mechanical processing of food contributes little, whereas the high energetic cost associated with the synthesis of macromolecules (e.g., proteins) can be a substantial contributor owing to its high energetic cost. The resulting heat can warm both the viscera and the body core (e.g., albacore tuna).

Thermal Set-Point, Load Error, and Temporal Variation

Body temperature is regulated through multiple temperature sensors (core and periphery) that fall into two groups, one responding with increasing activity to rising temperature and the other responding with increasing activity to falling temperature (**Fig. 5**). Centrally, the coordinated control of T_b takes place mainly in the pre-optic/anterior hypothalamus. The load error is the minimal deviation of T_b from the thermal set-point that is tolerated by the system. Deviation from the thermal set-point triggers a response from multiple effectors, adjusting heat gain or loss, as appropriate, to restore equilibrium. When the signal rates from the warm- and cold-responding sensors balance each other, T_b is at its thermal set-point and load error is zero. Load error is determined in part by the delay of the temperature sensors and the gain of the response by the various effectors. Any regulatory change in heat gain or loss is proportional to the load error and is generated by comparing the signals from both groups of sensors.

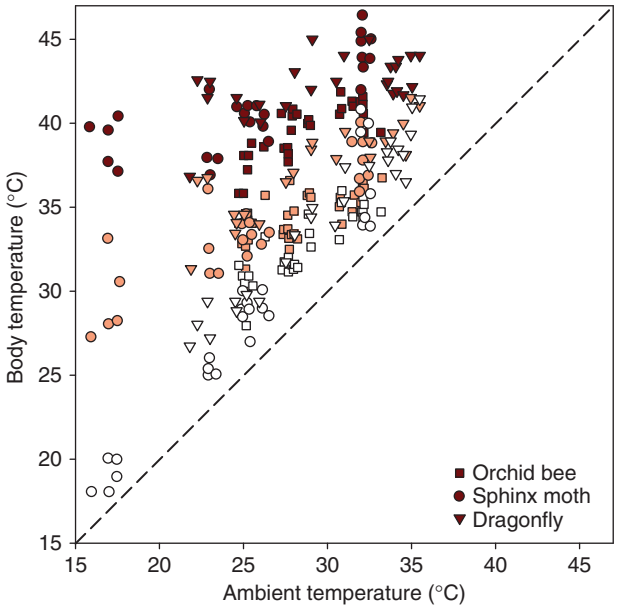


Fig. 4 Body temperatures as a function of ambient temperature during flight in three representative insects. Note that the temperature of the thorax (dark symbols) and, to a lesser extent, the head (light shading) are relatively independent of ambient temperature while that of the abdomen (open symbols) is not. Dashed line $T_b = T_a$. Drawn from data taken from May ML (1995) Simultaneous control of head and thoracic temperature by the green darner dragonfly *Anax Junius* (Odonata: Aeshnidae). *The Journal of Experimental Biology* 198: 2373–2384; Borrell BJ and Medeiros MJ (2004) Thermal stability and muscle efficiency in hovering orchid bees (Apidae: Euglossini). *Journal of Experimental Biology* 207: 2925–2933; and Hegel JR and Casey TM (1982) Thermoregulation and control of head temperature in the sphinx moth, *Manduca Sexta*. *Journal of Experimental Biology* 101: 1–15.

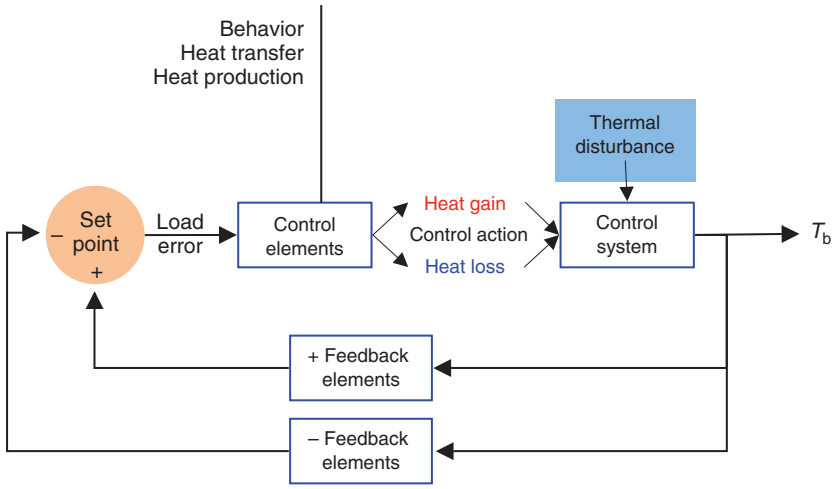


Fig. 5 Dual controller model for the regulation of body temperature with a thermal set-point. One group of feedback elements (sensors) responds with increased activity to rising temperature while the other group responds with increased activity to falling temperature. By comparing the activity of both groups of sensors the load error is generated (the difference between set-point and T_b) and the control elements are activated in proportion to the load error. An increase in T_b results in dominance of warm sensor activity and the control elements (e.g., changes in behavior, heat transfer properties, or heat production) restore equilibrium by increasing heat loss. When the activity from warm and cold sensors equals each other the load error is zero and T_b is at its set-point. For T_b to be regulated the system requires T_b to be displaced from set-point. How far the system can be displaced from equilibrium establishes the load error that is tolerated for the system.

In endothermic homeotherms the load error is generally small; deviations in T_b of a fraction of a degree elicit responses to return T_b toward set-point. Ectothermic homeotherms, on the other hand, often tolerate much larger variations but are still capable of good regulatory control of T_b. For example, varanid lizards, a highly active group of reptiles, are able to behaviorally regulate their T_b between 30 and 36 °C on cold days when provided with an appropriate radiant source of heat (**Fig. 6**).

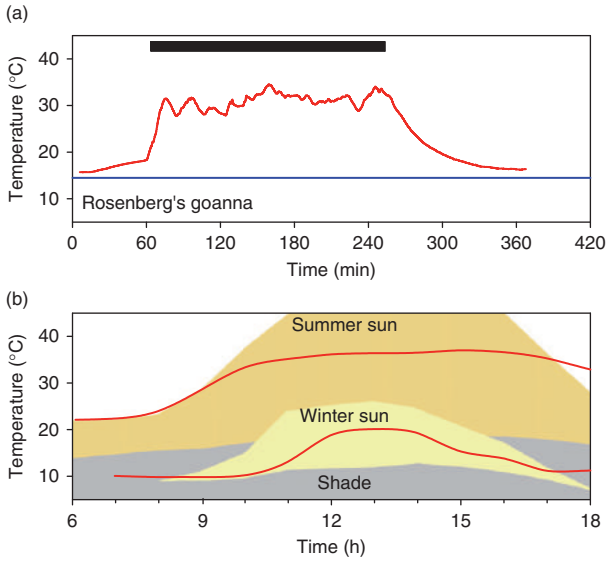


Fig. 6 Body temperature (red line) in relation to sources of heat input as a function of time in the varanid lizard, Rosenberg's goanna. In (a) under constant cool conditions (T_a = blue line) but the animal was free to shuttle in and out from underneath a radiant heat source (black bar) and in (b) under natural conditions that occur in summer and winter. When provided with an appropriate radiant source of heat from the sun Rosenberg's goanna can quickly elevate T_b above T_a in the shade, achieved through dark skin coloration and adjustments in peripheral circulation that favor heat gain. Once the desired (set-point) temperature is achieved the lizard can regulate T_b through behavioral adjustments that affect heat transfer. (a) From Clark TD, Butler PJ, and Frappell PB (2006) Factors influencing the prediction of metabolic rate in a reptile. *Functional Ecology* 20: 105–113. (b) Drawn from data obtained and modified from Christian KA and Weavers BW (1996) Thermoregulation of monitor lizards in Australia – An evaluation of methods in thermal biology. *Ecological Monographs* 66: 139–157.

Some ectothermic homeotherms have a smaller range in T_b. For example, ectothermic fish are not usually considered homeothermic, but many species spend 75% of their time within $\pm 2\,^{\circ}C$ of their preferred temperature. It has also been demonstrated that ectothermic sharks possess thermoreceptors that are capable of temperature resolution to $0.001\,^{\circ}C$ and can occupy narrow thermal niches.

Load error should not be confused with temporal adjustments that may occur in thermal set-point. Even in endotherms, T_b oscillates by as much as several degrees over nycthemeral and circadian cycles (**Fig. 7**). The circadian rhythm of activity can accentuate the daily oscillations of T_b (**Fig. 8**), though elevation of T_b by activity or exercise is not always indicative of a change in set-point. The thermal set-point is also known to change in accord with other cycles. During the human ovarian cycle, for example, the thermal set-point is about $0.5\,^{\circ}C$ higher in the luteal phase. In circannual cycles (e.g., seasons) the set-point is lowered and some animals display a marked reduction in metabolic rate and an associated regulated decline in T_b (e.g., torpor, **Fig. 8**). Finally, stressful situations such as infection result in an increase in thermal set-point (fever), whereas injury, hypoxia, starvation, and other situations can result in a decrease in the thermal set-point (anapyrexia).

Why Elevate and Regulate T_b above Ambient Temperature?

Catalytic or enzymatic rates are temperature dependent according to the Boltzmann's constant (Q_{10} is used to describe the effects of a $10\,^{\circ}C$ temperature difference on reaction rate). A constant T_b allows enzymes to always be at their optimal temperature for catalysis and overcomes the need for having isoenzymes adapted to operate at different temperatures. Interestingly, across the taxa, homeotherms have tended toward T_b's of 30–40 °C. This may be an attempt to minimize the effects of temperature on enthalpy and entropy required for activation energy for biochemical reactions.

A warm constant T_b means that rate processes such as metabolism, digestion, neural function, muscle force, and speed of contraction, growth, and other processes dependent on temperature can be maximized and maintained independent of ambient temperature. In turn, greater and more sustainable levels of activity and various processes enable expansion of both the temporal niche (i.e., animals can forage for longer) and thermal niche (i.e., animals are freed from fluctuations in ambient temperature).

Warm Is Not Always Better

In endothermic homeotherms the major downside to maintaining a warm T_b in the cold is the high energy expenditure for thermogenesis. Obviously, energetic savings can be made through a reduction in thermal set-point and the associated hypometabolism. For all animals a disadvantage of a higher T_b is that the Q_{10} effect will increase demand for energy. Some animals optimize the costs of a

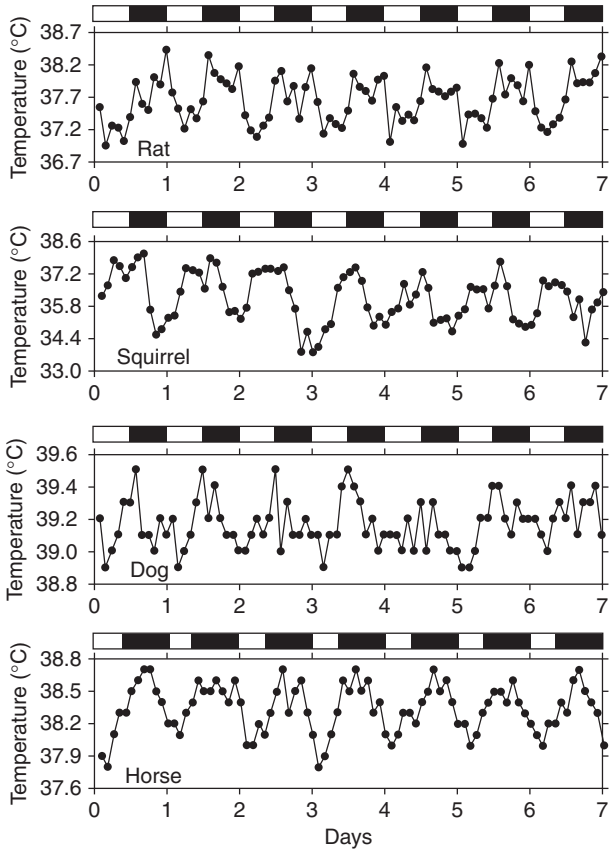


Fig. 7 Daily variation in body temperature in four mammalian species reveals a range of average T_b and variation about the mean. The acrophase (time of peak) for the nocturnal rat was much later than that for the diurnal species. Light and dark phases are indicated. Reproduced from Refinetti R and Piccione G (2005) Intra- and inter-individual variability in the circadian rhythm of body temperature of rats, squirrels, dogs, and horses. *Journal of Thermal Biology* 30: 139–146.

higher set-point through behavioral adjustments. In the case of the dogfish, a squaliform shark, it forages in warm water but lowers energy costs by resting and digesting in cooler water, thereby increasing bioenergetic efficiency and growth rate.

Body Size

The influence of environmental factors is size dependent. Small animals have relatively more surface area and relatively less metabolically active tissue with which to generate heat. Therefore, we would expect a small animal, when challenged with cold, to lose heat more rapidly than a larger animal. If such an animal attempted to generate heat to offset the heat loss and maintain T_b, it would require a relatively higher rate of energy utilization. Small endothermic homeotherms, by virtue of their higher metabolic requirements, are therefore dependent on a reliable supply of energy-yielding foods. Exposure to extreme cold and/or a food shortage would place a small endothermic homeotherm under profound metabolic stress. Obvious energetic savings are to be made under such conditions by lowering the thermal set-point for T_b regulation and depressing metabolism (i.e., torpor, **Fig. 8**). Examples of torpor are found in all three infraclasses of mammals and some groups of birds.

On the other hand, large animals have relatively small surface areas and cool slowly if placed in a cold environment. Once warmed larger insects take longer to cool than small organisms of equivalent taxa. Subsequently, they can successfully forage in a cool microhabitat provided that the foraging bout was less than the equilibration time. Large body size and the resulting thermal inertia together with improved ability for heat retention (insulative layers of fat, counter-current heat exchangers, or physiological adjustments to the circulation that limit peripheral heat loss) and behavioral thermoregulation (basking) enable the largest reptiles, the leatherback turtle and estuarine crocodile (both up to 1000 kg), to achieve a high degree of thermal stability (**Fig. 9**). Large estuarine crocodiles, whose large size is also associated with an increased T_b, actually risk overheating from solar radiation during the day. To minimize this risk they spend much of the day in the water and at night are frequently observed on land. Leatherback turtles also have elevated T_b's, from a few degrees above the warm oceanic waters at low latitudes, to as much as 18 °C above environmental temperatures when diving into high-latitude subsurface waters as cold as 0.4 °C. In the case of leatherback turtles it would appear that metabolic rate is higher than predicted and, in association with relatively warm surface waters and improved ability for heat retention, this is sufficient to maintain warm and fairly stable T_b's when

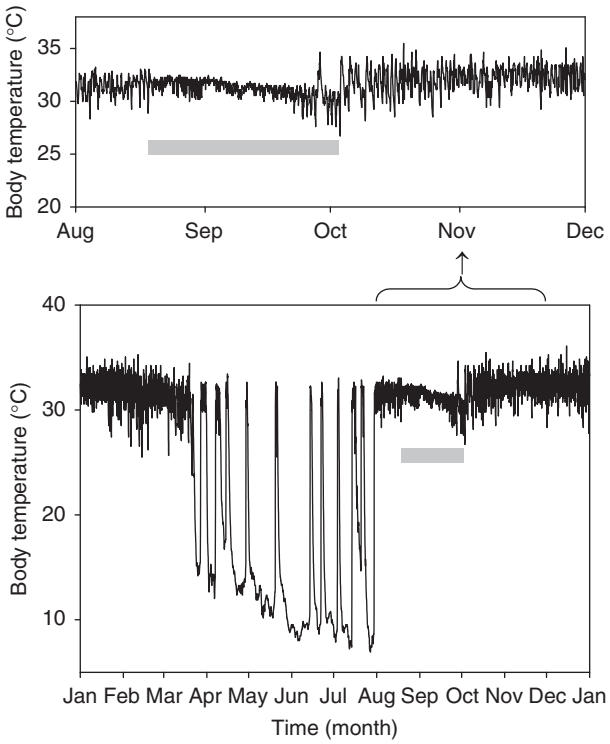


Fig. 8 Body temperature in a monotreme, the echidna, collected over 1 year in Tasmania. During the winter months (April–August) the animal enters torpor, characterized by a lowering of the thermal set-point and subsequent marked decline in T_b together with characteristic test arousals. The characteristic large circadian variation in T_b of echidnas (more easily seen in the upper panel) is a function of the activity cycle. Departures from this daily T_b cycle are associated with a decrease in the activity pattern as a result of the animal spending time in the burrow following egg laying (indicated by shaded bar). Data provided by SC Nicol and modified from Nicol SC and Andersen NA (2006) Body temperature as an indicator of egg-laying in the echidna, *Tachyglossus aculeatus*. *Journal of Thermal Biology* 31: 483–490.

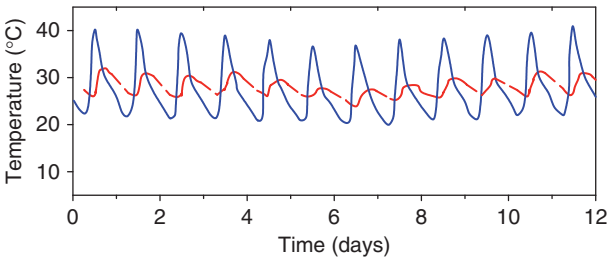


Fig. 9 Body temperature (red line) and operative temperature (blue line, temperature calculated at the body surface and based on heat inputs) for a 520 kg estuarine crocodile. The large body size of the crocodile together with behavioral adjustments permits a high degree of thermal stability in comparison with daily changes in operative temperature. Modified from Grigg GC, Seebacher F, Beard LA, and Morris D (1998) Thermal relations of very large crocodiles, *Crocodylus porosus*, free-ranging in a naturalistic situation. *Proceedings of the Royal Society of London B* 265: 1793–1799.

foraging in cold waters. Interestingly, it appears that muscle tissue metabolism is independent of temperature in leatherback turtles, thereby freeing muscle activity somewhat from thermal constraints in the environment.

In large, non-lamnid sharks, such as the blue shark, a thermal hysteresis permits warming more quickly than cooling. Coupled with regular, vertical excursions in the water column that return the fish to warmer surface water, this enables these sharks to maintain their T_b within a reasonably narrow range (14–21 °C) while diving through water from 26 to 7 °C (**Fig. 3b**).

Homeothermy and Thermal Niche Expansion

Many species encounter wide ambient temperature ranges, the result of moving through their milieu, migrating or changing climatic conditions that occur over various temporal timescales (days, seasons, or evolutionary). Homeothermy permits

exploitation of a range of thermal environs by freeing T_b from ambient temperature. In turn, thermal niche expansion, coupled with elevated T_b which improves temperature sensitive processes (e.g., locomotory performance, digestion, neural function, and metabolic rate) and the advantages of an improved metabolic rate provide the potential to optimize foraging, growth, and reproduction.

The thermal niche of pelagic tuna, billfish, lamnid, and large sharks has expanded because of homeothermy associated with elevated T_b. Improved temporal resolution resulting from retinal warming and temperature stability in endothermic fish such as the swordfish provides benefits over their prey which will have eyes equilibrated with the temperature of the water in which they swim and thus lower temporal resolution, diminishing the ability to avoid predation. Leatherback turtles by their enormity are homeothermic and consequently, like the warm pelagic fish, can travel vast distances across ocean basins to forage in high latitudes or dive to depths in cold waters. At the other extreme, small insects are capable of homeothermy and elevated T_b in order to enable flight for foraging, defense of territory, and mate selection during cool periods. Many insects, together with a host of reptiles rely on behavioral approaches to improve heat transfer between themselves and their thermal environment in an attempt to regulate T_b.

Decreasing T_a is associated with the increased metabolic costs of thermogenesis in endothermic homeotherms. Continual increases in metabolic rate in response to cold can become unsustainable, particularly if resources are limiting. In response to extrinsic constraints many endotherms respond by lowering their thermal set-point and entering a regulated hypometabolic state (e.g., torpor). Others choose thermally buffered microenvironments (e.g., burrows) or behaviors such as huddling or communal roosting to reduce heat loss, conserve energy, and preserve T_b. A worthy example of the latter approach is found in the red squirrel that remains active during winter ($T_a = -26\ ^\circ C$). The red squirrel minimizes energy expenditure during winter by ensuring easy access to a food hoard and a well-insulated nest in which it remains inactive for most of the time, venturing outside only for short periods on the warmest winter days.

Summary

Homeothermy associated with a T_b elevated above ambient temperature has enabled expansion of thermal niches through exploitation of colder habitats and improved rates of biological activity. On the other hand, an organism's metabolism is influenced by ecological determinants such as temperature and the resources available to it. Adaptation to change in environmental temperature across short-term and evolutionary timescales is largely dependent on an organism's ability for adjustments in a multitude of temperature-dependent processes. Homeotherms are able to minimize these changes by stabilizing body temperature.

See also: Aquatic Ecology: Microbial Communities

Further Reading

Borrell, B.J., Medeiros, M.J., 2004. Thermal stability and muscle efficiency in hovering orchid bees (Apidae: Euglossini). Journal of Experimental Biology 207, 2925–2933.
Brown, J.H., Gillooly, J.F., Allen, A.P., Savage, V.M., West, G.B., 2004. Toward a metabolic theory of ecology. Ecology 85, 1771–1789.
Carey, E.G., Scharold, J.V., 1990. Movements of blue sharks (*Prionace glauca)* in depth and course. Marine Biology 106, 329–342.
Christian, K.A., Weavers, B.W., 1996. Thermoregulation of monitor lizards in Australia – An evaluation of methods in thermal biology. Ecological Monographs 66, 139–157.
Clark, T.D., Butler, P.J., Frappell, P.B., 2006. Factors influencing the prediction of metabolic rate in a reptile. Functional Ecology 20, 105–113.
Dickson, K.A., Graham, J.B., 2004. Evolution and consequences of endothermy in fishes. Physiological and Biochemical Zoology 77, 998–1018.
Frappell, P.B., Butler, P.J., 2004. Minimal metabolic rate, what it is, its usefulness, and its relationship to the evolution of endothermy: A brief synopsis. Physiological and Biochemical Zoology 77, 865–868.
Grigg, G.C., Seebacher, F., Beard, L.A., Morris, D., 1998. Thermal relations of very large crocodiles, *Crocodylus porosus*, free-ranging in a naturalistic situation. Proceedings of the Royal Society of London B 265, 1793–1799.
Hegel, J.R., Casey, T.M., 1982. Thermoregulation and control of head temperature in the sphinx moth, *Manduca Sexta*. Journal of Experimental Biology 101, 1–15.
Heinrich, B., 1993. The Hot Blooded Insects: Strategies and Mechanisms of Thermoregulation. Cambridge, MA: Harvard University Press.
Holland, K.N., Sibert, J.R., 1994. Physiological thermoregulation in bigeye tuna, *Thunnus obesus*. Environmental Biology of Fishes 40, 319–327.
Ivanov, K.P., 2006. The development of the concepts of homeothermy and thermoregulation. Journal of Thermal Biology 31, 24–29.
Jessen, C., 2001. Temperature Regulation in Humans and Other Mammals. Berlin: Springer.
May, M.L., 1995. Simultaneous control of head and thoracic temperature by the green darner dragonfly *Anax Junius* (Odonata: Aeshnidae). The Journal of Experimental Biology 198, 2373–2384.
Nicol, S.C., Andersen, N.A., 2006. Body temperature as an indicator of egg-laying in the echidna, *Tachyglossus aculeatus*. Journal of Thermal Biology 31, 483–490.
Peters, R.H., 1983. The Ecological Implications of Body Size. Cambridge: Cambridge University Press.
Portner, H.O., Bennett, A.F., Bozinovic, F., *et al*., 2006. Trade-offs in thermal adaptation: The need for a molecular to ecological integration. Physiological & Biochemical Zoology 79, 295–313.
Refinetti, R., Piccione, G., 2005. Intra- and inter-individual variability in the circadian rhythm of body temperature of rats, squirrels, dogs, and horses. Journal of Thermal Biology 30, 139–146.
Seymour, R.S., 2002. Temperature regulation by thermogenic flowers. In: Taiz, L., Zeiger, E. (Eds.), Plant Physiology, 3rd edn. Sunderland: Sinauer Associates. Essay 11.3.
Somero, G.N., 2005. Linking biogeography to physiology: Evolutionary and acclimatory adjustments of thermal limits. Frontiers in Zoology 2, 1. doi:10.1186/1742-9994-2-1.

Hunting☆

M Nils Peterson, North Carolina State University, Raleigh, NC, United States

Introduction

Hunting is the practice of pursuing, capturing, or killing wildlife. This broad definition can be divided into subsistence, commercial, and recreational hunting. Subsistence hunting provided the primary source of protein for most humans prior to widespread domestication of animals and evolution of agricultural societies (2,500,000–10,000 BCE). When agriculture and animal husbandry emerged (10,000 BCE), hunting began moving from a subsistence practice to a cultural practice. This transition is largely complete in developed nations where hunting is an insignificant source of protein. Subsistence hunting, however, still persists in economically depressed areas (e.g., much of rural Africa), and in relatively isolated cultural groups (e.g., Arctic Inuit).

In areas where survival was not the primary objective of hunting, the practice evolved into extermination, commercial, and recreational varieties. Extermination hunting involves attempts to eliminate wildlife which prey on or compete with domestic livestock or crops or threaten human safety. Commercial hunting for meat, hides, plumage, or other tissues (e.g., tusks, antlers, horns, claws, and skulls) often involves special training and utilizes efficient, rather than traditional or stylized, weapons. Recreational hunting evolved as a luxury sport of higher social classes in early agricultural societies (e.g., ancient Egypt and Mesopotamia). In Europe, sport hunting remained in the realm of royalty or aristocracy through the colonial period, and that tradition spread to some colonies (e.g., India). Social and ecological contexts in colonial North America, however, led to a modern hunting culture without strong ties to social class. While modern recreational hunting is often associated with sport hunting (e.g., trophy hunting), many hunters participate primarily to experience the outdoors and associate with other hunters. This hunting culture plays a major role in advertising, economics, and social practices in many rural areas throughout the world. Formal "cultural hunting" regulations also recognize the central role of hunting for some tribal groups.

In the United States, self-sufficient and conservation-minded sportsmen hunters embodied by Theodore Roosevelt represented a shared view of hunting prior to the 1960s. The civil and women's rights movements, Vietnam War and the peace movement, and environmental movement, however, led to a more critical public in late 1960s and early 1970s. Since then, animal welfare and animal rights groups have successfully lobbied against several forms of sport hunting (e.g., "canned" hunts of penned wildlife, hound hunting of fox and bear) and commercial whaling. These successes combined with steady decline in hunter recruitment, retention, and numbers during the same period suggest future changes in hunting culture and practice.

Ecological Effects of Hunting

The ecological effects of hunting are no less diverse than its history. While hunters act as apex predators in ecological terms, human hunters rarely conform to the assumptions of the Lotka–Volterra equations (predator–prey equations). Human hunters can and often have operated as keystone predators, exerting a disproportionate effect on prey populations relative to human numbers. The ecological effects of modern hunting, however, are mediated by governmental regulations. Throughout much of the twentieth century, the maximum sustainable yield paradigm guided hunting regulations. During the latter half of the 20th century fixed harvest quotas based on maximum sustainable yield decimated many global fisheries and whale populations. This failure led to regulation of hunting effort (season timing and length, bag limits, means/methods (archery vs. firearms)). Such regulations control harvest indirectly by altering hunting effort and hunter efficiency. Because hunter effort required to bag an animal increases as prey abundance declines, effort-based harvest regulations are less risky than quota-based systems should initial prey populations be overestimated. This allows regulators time to adaptively implement changes designed to decrease take should this be required.

Idealized narratives of indigenous cultures suggest early humans evolved sustainable hunting systems in a delicate balance with local ecosystems. While some indigenous cultures clearly exhibit sustainable hunting practices (e.g., the Amazonian Korubo, Artic Inuit, South African Bushmen), the apparent stability may reflect historic extinctions of most species vulnerable to indigenous hunting techniques. Many indigenous hunters actively manipulated ecosystems to facilitate hunting. Early hunters in every continent transformed ecosystems by setting fires to drive animals into traps, facilitate tracking animals, and create favorable habitat for prey species. The landscapes greeting European colonists in North America and Australia reflected the fires aboriginals used to create habitat for hunted species.

The overkill hypothesis for the Pleistocene megafauna extinction event represents the paradigm case of unsustainable hunting (Fig. 1). Radiation of hunting cultures into North and South America around 10,000 BCE coincided with extinction

☆*Change History:* February 2018. I. Martins made minor changes to the references.

This is an update of M.N. Peterson, Hunting, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 1912–1915.

 doi:10.1016/B978-0-12-409548-9.11168-6

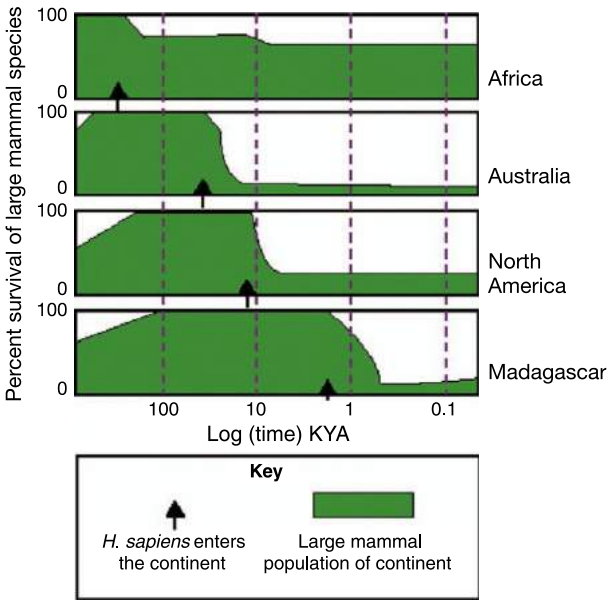


Fig. 1 Timeline of large mammal extinctions and entry of humans in Africa, Australia, North America, and Madagascar. Adapted from Martin 1984 by Elin Whitney-Smith.

of 33 of 45 and 46 of 58 large mammal genera in each continent, respectively. Since 1600, hunting caused 23% of all known animal extinctions. The list of species driven to extinction (e.g., passenger pigeon (*Ectopistes migratorius*), heath hen (*Tympanuchus cupido cupido*)), extirpated from most of their range (e.g., bison (*Bos bison*)), or threatened (e.g., most whale species) by commercial hunting is staggering. Such changes wrought by hunting often have ripple effects throughout ecosystems. For instance, as domestic livestock replaced extirpated bison, brown-headed cowbirds (*Molothrus ater*) adopted the sedentary lifestyle of their domestic symbionts and became North America's most notorious avian brood parasite. Recreational hunting has had more mixed ecological impacts. In some contexts (e.g., Medieval and Colonial Europe), hunting motivated preservation of forest ecosystems as game preserves. Aristocratic sport hunting, however, nearly led to extinction of the Bengal tiger (*Panthera tigris tigris*).

Hunting became an almost universally positive ecological force during the conservation movement of the late 1800s and early 1900s (to which prominent hunters including Theodore Roosevelt contributed heavily). In the United States, federal excise taxes on hunter purchases ($>$\$200 million annually) support most state-level wildlife management programs in the United States, and the mandatory purchase of Federal Duck Stamps by migratory waterfowl hunters since 1934 helped purchase more than 20,000 km^2 of wildlife habitat in the National Wildlife Refuge System. Recreational hunters also contributed to restoration of North America's decimated game species including white-tailed deer (*Odocoileus virginianus*), pronghorn (*Antilocapra americana*), black bear (*Ursus americanus*), wild turkey (*Meleagris gallopavo*), and wood duck (*Aix sponsa*). Hunters also contributed to establishment of wildlife reserves and conservation-hunting programs throughout the world. The conservation–hunting programs epitomized by Zimbabwe's Communal Areas Management Programme for Indigenous Resources (CAMPFIRE) provide locals in developing nations a sustainable supply of money and meat and preserve local wildlife species. Finally, in areas dominated by private land ownership, hunting provides economic incentives for protecting natural and agricultural lands from suburban or commercial developments.

Some ecologically questionable practices, however, have evolved to capitalize on the growing economic value of hunted species, including high fencing, supplemental feeding, and landscape manipulation to support economically valuable species. While these practices may render keeping private lands in a somewhat natural state more economically viable, they are ecologically problematic. High fences effectively fragment landscapes for many species, supplemental feeding increases the risk of disease transmission, and landscape manipulation intended to help commercially valuable wildlife can threaten other species.

Modern hunting also provides an important tool for managing overabundant wildlife species. Hunting allows natural resource managers to control populations of species before they exceed carrying capacity of their habitat, damage vegetation, threaten other species, or threaten human health and safety. Managers also can use hunting to control a wildlife population's density in efforts to minimize risks associated with wildlife-related diseases (e.g., Lyme disease, bovine brucellosis). In 2005, hunting became a major tool to fight degradation of saltwater marshes caused by growing populations of lesser snow geese (*Chen caerulescens caerulescens*) along the Hudson Bay. Some groups assert reintroducing predators would achieve the same benefits as hunting. Such reintroductions, however, prove politically problematic. Further, considerable evidence suggests prey species typically control predator abundance rather than vice versa.

Ecological Effects on Hunting

Because successful hunting requires a clear understanding of the relationship between the prey species and its biotic and abiotic environment, hunters were probably the first ecologists. In early human history, hunting made most people ecologists by necessity. Now millions of recreational ecologists study the relationships between wildlife and their environments with hopes of increasing the likelihood of successful hunts. Because hunters were among the first ecologists, and continue to study relationships among game species, other organisms, and the abiotic environment, ecosystems shaped and shape the practice of hunting. Hunting generally occurs in areas where prey species predictably occur. Like nonhuman predators, hunters have always focused their efforts in areas and at times where prey species meet critical needs (e.g., food, cover, rest, reproduction). In arctic areas hunters target seals at breathing holes in ice, in arid areas hunters wait near watering holes, and salt and other mineral deposits (natural and artificial) provide a common hunting location throughout the world. Deer hunters often position themselves between cover and foraging areas during dawn and dusk when deer predictably move between these areas. Waterfowl hunters position themselves on or near small water bodies where waterfowl rest along migration routes during fall migrations.

Biodiversity also influenced the persistence of subsistence hunting. In areas lacking domesticatable species, hunting remained essential to human survival until domestic plants and animals were imported from other areas. Changing landscape patterns have also influenced hunting. As agriculture and urban sprawl created a fragmented landscape in many areas, popular game species—including white-tailed deer and Canada geese (*Branta canadensis*)—became nuisances in suburban areas. In many such cases, hunting in nearby areas became an important tool for controlling those nuisance species.

While technology (e.g., firearms, motor vehicles) has allowed modern hunting to develop independent of ecological relationships in some ways, ecosystems have shaped the social nature of hunting. In relatively open landscapes, persistence hunting (using teamwork to run down prey) evolved. In densely forested areas, hunting evolved to be less of a group activity. Attributes of prey also influenced the social nature of hunting. Large prey species required larger groups of hunters, processors, and eaters. Even with modern technology, however, ecosystems influence the social dynamics of hunting. Emerging zoonotic diseases are shaping perceptions of hunting risk and influencing hunting participation. When chronic wasting disease (CWD) was discovered in Wisconsin (USA) deer herds in 2002, hunter numbers began declining and more than 50% of deer hunters using firearms that hunted in 2001, but not 2002, cited CWD as their reason for not hunting.

Further Reading

Angula, H.N., Stuart-Hill, G., Ward, D., Matongo, G., Diggle, R.W., Naidoo, R., 2018. Local perceptions of trophy hunting on communal lands in Namibia. Biological Conservation 218, 26–31.
Diamond, J.M., 1997. Guns, germs, and steel: The fates of human societies. NY: W.W. Norton.
Dizard, J.E., 2003. Mortal stakes: Hunters and hunting in contemporary America. Boston: University of Massachusetts Press.
Groombridge, B., 1992. Global biodiversity: Status of the Earth's living resources. Chapman and Hall London.
Leopold, A., 1933. Game management. Madison: The University of Wisconsin Press.
Lund, J.F., Jensen, F.S., 2017. Is recreational hunting important for landscape multi-functionality? Evidence from Denmark. Land Use Policy 61, 389–397.
Martin, P.S., 1984. Prehistoric overkill: A global model. In: Martin, P.S., Klein, R.G. (Eds.), Quaternary extinctions: A prehistoric revolution. Tucson: University of Arizona Press, pp. 354–404.
Needham, M.D., Vaske, J.J., Manfredo, M.J., 2004. Hunters' behavior and acceptance of management actions related to chronic wasting disease in eight states. Human Dimensions of Wildlife 9, 211–231.
Pang, A., 2017. Incorporating the effect of successfully bagging big game into recreational hunting: An examination of deer, moose and elk hunting. Journal of Forest Economics 28, 12–17.
Peterson, M.J., 2001. Northern bobwhite and scaled quail abundance and hunting regulation: A Texas example. Journal of Wildlife Management 65, 828–837.
Peterson, M.N., 2004. An approach for demonstrating the social legitimacy of hunting. Wildlife Society Bulletin 32, 310–321.
Whitney, G.G., 1994. From coastal wilderness to fruited plain: A history of environmental change in temperate north America, 1500 to the present. Cambridge: Cambridge University Press.
Whytock, R.C., Morgan, B.J., Awa, T., Bekokon, Z., Abwe, E.A., Buij, R., Virani, M., Vickery, J.A., Bunnefeld, N., 2018. Quantifying the scale and socioeconomic drivers of bird hunting in central African forest communities. Biological Conservation 218, 18–25.

The Intermediate Disturbance Hypothesis☆

RW Osman, Smithsonian Environmental Research Center, Edgewater, MD, USA

Introduction

The intermediate disturbance hypothesis (Connell, 1978; Grime, 1973a,b) postulates that for any ecological system the diversity of species will be highest when or where disturbance is at some intermediate level. The hypothesis simply overlays environmental disturbances, which negatively impact one or more species, onto the temporal process of community development or succession. Starting with a new, unoccupied patch of habitat the number of species in it will increase as it is colonized. As population densities and species diversity continue to increase resources will eventually become limiting and the species will compete for these resources. With increased competition, inferior competitors will be displaced and eventually lost. At some point species losses should be greater than gains and the number of species will decline and, in the extreme, end with a single dominant species remaining. Thus, the temporal pattern of diversity can be represented by a unimodal curve in which diversity initially increases and then declines over time (**Fig. 1**). Any disruption or disturbance of this process will likely reset the community to some earlier state. High rates of disturbance will keep the community in an early stage with few species and low rates of disturbance will allow competitive exclusion to reduce diversity. Therefore, some rate or magnitude of disturbance should exist that forestalls competitive exclusion (Connell, 1978; Grime, 1973a,b) and maintains the community in an intermediate developmental state of highest diversity.

At its core the hypothesis makes four principal assumptions: (1) that populations and the communities that they form are constantly changing, (2) that these changes can result in the gain or loss of species, (3) that the process of community change can be disturbed or disrupted by any number of physical and biological events that remove individuals, populations, species, or the whole community, and (4) the change in the number or diversity of species is nonlinear as a consequence of the interaction between species recruitment and the limits placed on continued increases by finite available resources. The first three of the assumptions are straightforward. First, normal processes of individual birth, growth, and mortality will produce change. These changes may be limited to replacement of individuals in stable populations with little measurable change at the community level. It is more likely that populations will fluctuate in size and produce changes in species' relative abundances including complete loss of some species. Secondly, for any community, there is always some probability of new species immigrating or existing species emigrating or going extinct locally. Thirdly, disturbances that produce losses certainly can occur in any system, but their effects will vary based on their magnitude, frequency, and spatial extent. The fourth assumption is probably the most contested. The increase in species as a habitat is colonized clearly must occur and the rate of increase will decline as both available resources and the pool of available new species decrease with the addition of each new species. Thus the critical aspect is whether after some period of time competition for limiting resources or some other process results in sufficient species losses to cause a decline in diversity.

Although the intermediate disturbance hypothesis is most often tied to changes in species associated with succession, it is not really dependent on any particular process of community development. Whether one perceives the species composition of a community changing as a fairly systematic successional sequence of species or as a process of random immigrations and extinctions, the process can be interrupted or perturbed by weather, fire, flood, drought, herbivores, predators, disease, tree falls, waves, etc. These disturbances can affect any or all species, randomly or selectively. If diversity increases and then decreases as the community develops through time, then some level of disturbance is likely to maintain the community at a state of maximum diversity.

Because of its apparent generality the intermediate disturbance hypothesis has been studied and tested in a vast array of empirical studies of most types of natural communities and habitats, in experimental communities in small laboratory chambers, and in larger ecosystems and landscapes that can include multiple habitats and communities. There have also been a substantial number of theoretical and modeling studies that have tried to develop and test constraints on the hypothesis and its applicability. As a consequence, there is a diversity of opinions about when, where, and to what systems the hypothesis should be applied. Before limiting the hypothesis to one extreme of catastrophic disturbances that remove all species in fragments or patches within a landscape or to the other of non-catastrophic disturbances across homogeneous habitat that maintain high diversity by reducing populations to non-competitive levels, it is important to examine the historical context of the hypothesis and how it has evolved.

Hypothesis History and Range

The idea that an intermediate level of disturbance can lead to the highest diversity within a system has been proposed independently by numerous ecologists based on their observations of a variety of communities. Although observations of the

☆*Change History:* December 2014. RW Osman has made the following changes: References were added to the text and were updated to include studies up to 2014. Some minor modification of the text was made to incorporate the references. Additional discussion of both recent criticisms of the intermediate disturbance hypothesis and responses to these criticisms. Discussion of some alternatives was modified to accommodate the discussion of criticism. A short section on the links to the intermediate productivity hypothesis and the dynamic equilibrium model of Huston (1979) was included.

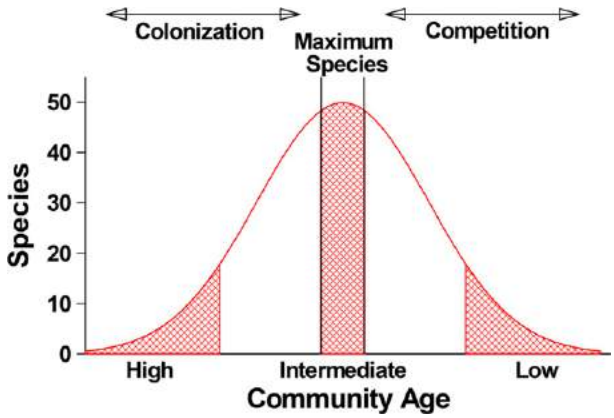


Fig. 1 General concept of the intermediate disturbance hypothesis which overlays disturbance at different times on the diversity changes as a community develops from early stages of colonization with few species to maximum diversity prior to the reduction in diversity through competition after resources become limiting.

phenomenon can be found in studies from the early to mid-1900s (e.g. Gleason, 1917; Hutchinson, 1951, 1953), the link between intermediate levels of disturbance and high diversity was more fully explored in the 1970s (e.g. Connell, 1978; Grime, 1973a,b; Horn, 1974; Huston, 1979; Osman, 1977; Osman and Whitlatch, 1978; Sousa, 1979). Connell (1978) is usually credited with giving the first formal definition and description of this idea as the intermediate disturbance hypothesis. Connell (1978) found the intermediate disturbance hypothesis to be the best of six explanations for the high diversity of species found in the tropical coral reefs and rain forests that he was studying. Alternate explanations included: 'equal chance' in which all species are equal and diversity is a function of the environment and species available in a region, 'gradual change' in which environmental changes prevent competitive exclusion, 'niche diversification' or the finer division of resources among competing species, 'circular networks' in which competition is not hierarchical, and 'compensatory mortality' in which noncompetitive mortality is greatest for dominant species. Connell (1978) was careful to limit the scope of the intermediate disturbance hypothesis to the main structural species of two tropical systems, corals and trees. He saw them as representative of communities of sessile, often long-lived, species with dispersal largely limited to the larval and seed life-stages produced through reproduction. He specifically excluded more motile species which have the potential to avoid or quickly adjust to the impacts of disturbance.

In this context, Connell (1978) described these diverse systems as dynamic, non-equilibrium communities that did not approach maximal diversity by the coexistence of species which have divided available resources on finer and finer scales. Rather than a balance or equilibrium among coexisting species each with their unique way of using some portion of the available resources, species contested limiting resources with those superior competitors ultimately pushing out the inferior. Competitive exclusion and the ability of disturbances to forestall the loss of inferior competitors are critical elements of the hypothesis. Within long-lived coral reef and forest communities competitive exclusion is not instantaneous and may take decades to centuries. In these long time periods any number of events can disrupt the competitive process, remove or reduce the abundance of dominants, and make resources more available. Although one can imagine how disturbances such as disease, wind, or waves might exclusively or disproportionately affect large dominant coral or tree species, such types of disturbance are not necessary for disturbance to contribute to higher diversity. Disturbances that randomly remove individuals or parts of the habitat will create areas available for colonization with available resources and reduced competitor abundance. New or previously lost species have the opportunity to colonize these areas with the potential of increasing local diversity.

At its core the intermediate disturbance hypothesis is recognition that environmental change can alter ecological processes and patterns, particularly when the change is of sufficient magnitude, frequency, or spatial extent to cause mortality and the local loss of populations, species, or whole communities. How disturbance disrupts processes, the nature of the processes being disrupted, and ultimately, the consequences for diversity have had their own diversity of interpretations. The most dominant view has been that of physical processes operating at variable frequencies causing interruptions (e.g. tree falls, rock turnovers) of a successional sequence that progresses from good colonizers to competitive dominants. Even with these constraints intermediate disturbance can be viewed differently. At one end of the spectrum intermediate disturbance can be seen as some frequency of catastrophic or complete removal of the community in patches within a fragmented landscape (e.g. Cushman and McGarigal, 2003; McGuinness, 1984; Osman and Whitlatch, 1978; Paine and Levin, 1981; Roxburgh *et al.*, 2004; Turner, 1989). If this produces patches of different ages, disturbed at different times then some intermediate frequency of disturbance will maximize the cumulative diversity of all patches (**Fig. 2**). At the other end of the spectrum intermediate disturbance can be seen as part of secondary succession in which non-catastrophic disturbances operating in a homogeneous system keep abundances sufficiently low and resources sufficiently available to prevent competitive exclusion from occurring (Horn, 1974). In this context intermediate disturbances provide the environmental variability and temporal diversity in resource availability that can produce opportunities for a greater diversity of species life-histories or adaptations.

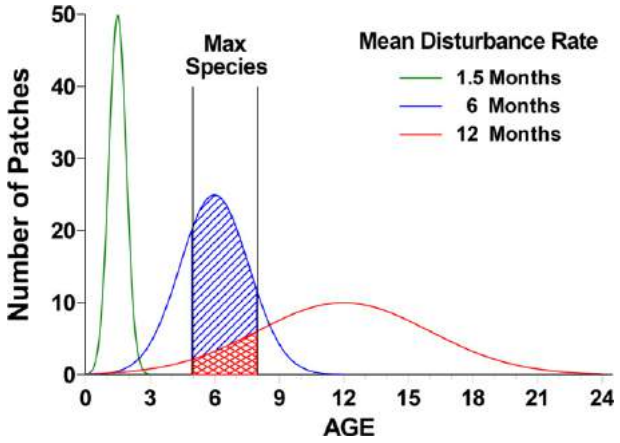


Fig. 2 Distributions of patch ages at different frequencies of disturbance showing the proportion of patches at maximum diversity assuming this is attained in communities 5–8 months old.

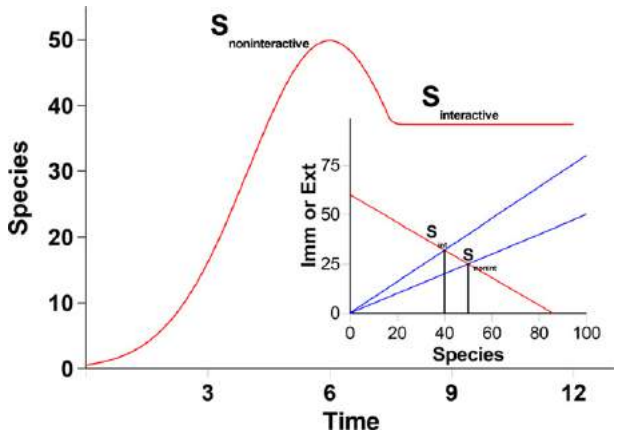


Fig. 3 Diversity (S) changes during colonization as a function of species gains and losses (immigration and extinction) (based on island biogeographic theory). Inset: When species interact or compete for resources the extinction rate increases resulting in a lower number of species. Thus, disturbance at intermediate frequencies can result in highest diversity.

Alternately, the link between high diversity and intermediate levels of disturbance can be maintained independent of any of the specific interspecific relationships that are part of succession. For example, island biogeographic theory (MacArthur and Wilson, 1963, 1967) presents an alternate scenario for the colonization of an island or patch of habitat that is neutral to species identities. The number of species present will be a function of the number of species immigrating minus the number of species lost or going extinct locally. Immigration rate will be a declining function of the number of species present, reaching zero when all species in the available pool of species are present. With all species having some probability of being lost, the cumulative extinction rate will increase as the number of species present increases. Therefore, some balance or equilibrium in species number should exist when immigration and extinction rates are equal. Species extinction rates should also be higher when they begin interacting or when population sizes are limited by available resources. Thus there should be a higher non-interactive equilibrium species number that might be expected to precede a lower interactive equilibrium species number (**Fig. 3**). In patchy environments, the catastrophic random disturbance of individual patches will produce a distribution of patches with a mean age equal to the mean disturbance rate. At some intermediate rate of disturbance, the majority of patches will be at an age when they are at or near the non-interactive equilibrium number of species creating the overall highest diversity for the system (**Fig. 2**). Likewise, non-catastrophic disturbance can lower extinction rate by opening resources and keep individual patches at the highest diversity. This approach is not limited to specific types of species, makes no assumptions about competitive dominance or the relative strengths or directions of any interactions. It relies only on the system reaching some dynamic balance between species losses and gains that will change once resources become limiting. Regardless of whether the probabilities of both immigration and extinction are equal or variable among species, some intermediate level of disturbance should produce the highest number of species.

There have been a number of studies that have questioned the applicability or validity of the hypothesis. These have been either metaanalyses of published papers evaluating the empirical evidence (Hughes *et al.*, 2007; Mackey and Currie, 2001) or studies that have criticized the hypothesis on various theoretical grounds (Chesson and Huntly, 1997; Fox, 2013; Shea *et al.*, 2004). These

criticisms have been addressed (Huston, 2014; Sheil and Burslem, 2013) with Huston (2014) providing a broader context in terms of his dynamic equilibrium model (Huston, 1979) that includes a similar relationship of highest diversity at intermediate levels of productivity (Grime, 1973a,b). The lack of support in many empirical studies points to the importance of defining disturbances in terms of the life-histories of the organisms being studied (see below) as well as the multiple mechanisms contributing to diversity patterns (Sheil and Burslem, 2013). For example, Huston (2014) in linking productivity and diversity found that the level of disturbance necessary for highest diversity can vary with productivity. As Huston (2014) has shown, much of the theoretical criticism has resulted from refuting the ability of intermediate levels of disturbance to maintain stable communities of highest diversity. However, the intermediate disturbance hypothesis is defined in terms of a dynamic process of community development and changing diversity and makes no claims for maintaining diversity at a stable level (Connell, 1978; Huston, 1979, 2014). This can be seen more clearly by examining the elements of the hypothesis.

Elements of the Hypothesis

There are two principal elements of the intermediate disturbance hypothesis, a temporal sequence of community development or change and a set of one or more environmental changes that are seen as disrupting the temporal sequence.

The Developmental Process

A nonlinear change in the number of species in a community or any other ecological system is a necessary component of the intermediate disturbance hypothesis. This change is usually seen as unimodal with a peak in diversity at some intermediate point in time. This is in contrast with two other potential patterns, a continuous increase in species number over time and an asymptotic increase with species number reaching and stabilizing at some maximum (**Fig. 4**). Assuming that there is some finite pool of available species, then stabilization at this number of species represents the upper limit to diversity. It follows that if over time diversity increases or increases to an asymptote without any decrease, then no frequency or magnitude of disturbance can result in higher diversity and the intermediate disturbance hypothesis would be rejected.

Beyond a pattern of increasing and then decreasing diversity over time, the process by which a community develops is unimportant. If development proceeds in an orderly successional sequence of species or through random colonization, the aspect critical to the intermediate disturbance hypothesis is that diversity peaks at some intermediate stage of the process. On the other hand, the characteristics of the system, especially the taxa or species included, the variability in their life-histories, and their trophic relationships, have important consequences to the application and testing of the hypothesis.

Connell (1978) was careful to restrict the intermediate disturbance hypothesis to particular types of species when he applied it to the structural tree and coral species in forest and reef systems. It is logical to assume that within these taxa there will be a limited range of life-histories such that high to low rates of disturbance can be defined. However, if we included all species from bacteria to vertebrates that might be considered to be part of tropical reef or forest systems, then imposing a single meaningful scale of disturbance is difficult. Disturbance is relative to the taxa being examined and their life-histories or generation time (Shea *et al.*, 2004). An intermediate rate of disturbance for a bacterial community may be on the order of hours to days while it may be decades or more for corals or trees. Even for communities of taxa with similar generation times, the intermediate disturbance frequency can also vary with the size of the habitat given that in smaller patches resources will become limiting faster (e.g. Miller, 1982). The size of a disturbance will also vary among taxa with different generation times. Large disturbances to a bacterial community will be too small to be disturbances to trees. In such complex systems it may be necessary to classify disturbances in

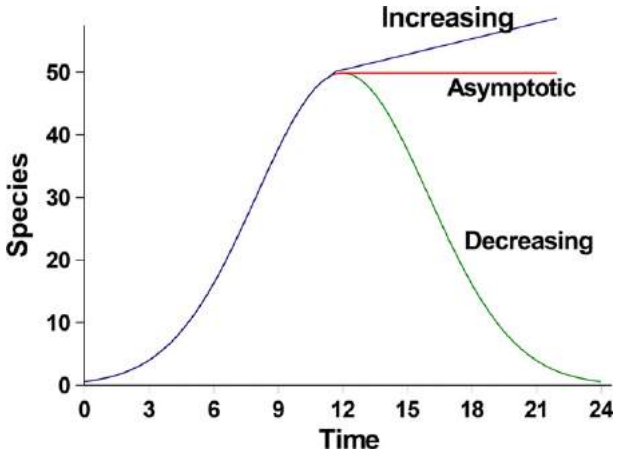


Fig. 4 Alternate patterns of temporal changes in diversity. If diversity is asymptotic or increasing intermediate disturbance would not maximize diversity.

terms of frequency, magnitude, and spatial extent simultaneously and identify a combination that maximizes overall diversity. For example, if an intermediate disturbance frequency for trees is on the order of decades, smaller magnitude daily intermediate disturbances for bacteria will occur multiple times within the period of highest tree diversity. This would produce a set of multiple intermediate disturbances of specific frequencies and magnitudes that result in several diversity maxima.

Including multiple trophic levels within the system being considered produces a similar set of problems (e.g. Gallet *et al.*, 2007; Jeppesen *et al.*, 2000; Wootton, 1998). Modeling has demonstrated that the degree to which disturbances affect each trophic level can influence the overall impact on diversity (Wootton, 1998). If a disturbance affects all levels equally then an intermediate level producing the highest diversity will exist. Alternatively, if the impact of a disturbance is disproportional there may not be an intermediate level producing the highest diversity. For example, if a disturbance principally causes mortality in a lower trophic level this would open resources at this level and some intermediate disturbance rate should lead to higher diversity at this trophic level. At the same time herbivores or predators in the next trophic level would experience reduced resources because of the lowered overall abundances of prey. This should lead to greater competition, lower diversity in the higher trophic level, and an unknown change in overall diversity. If species at higher trophic levels are more selective in their prey then the changes in their resources with disturbance and the effect on overall diversity are even more difficult to predict.

In a similar manner, the degree or number of positive symbiotic interactions within any system may affect directly the relationship between disturbance and diversity (Brooker *et al.*, 2008; Hacker and Gaines, 1997). At its core the intermediate disturbance hypothesis depends on resources becoming limiting and resulting negative effects on abundances and the survival of species. If a significant number of interspecific interactions within a system are positive, then disturbances may actually reduce available resources or the ability of species to utilize them. For such a system no level of disturbance may increase diversity.

Types of Disturbances

Within the context of the intermediate disturbance hypothesis, disturbances are any events or processes that produce mortality. Disturbances will vary in frequency, magnitude, and spatial extent (e.g. Huston, 1979, 2014; Miller *et al.*, 2011; Shea *et al.*, 2004) and they are usually viewed as unpredictable. Frequency is the most common attribute associated with intermediate disturbances (Gaedeke and Sommer, 1986; Lenz *et al.*, 2004; Sousa, 1979; Svensson *et al*, 2007) and the effects of changes in disturbance frequency are, perhaps, those most easily seen. In part, this results because the rate at which disturbances occur can be imposed directly onto the temporal change in a system's diversity. Disturbance rate translates directly into a mean age of the community and its diversity.

The magnitude of a disturbance (Armesto and Pickett, 1985; Bertocci *et al.*, 2005; Biswas and Mallik, 2011) is measured in terms of the mortality it causes which can vary between catastrophic with complete loss of all individuals of all species and non-catastrophic with the loss of only a few individuals and possibly no loss of species. Secondary succession is often seen as resulting from disturbances of low to intermediate magnitude (Horn, 1974). Disturbances of intermediate magnitude and frequency may reduce abundances without causing the loss of any species. The lowered abundances of some or all of the species can free resources, reduce or eliminate competition, and allow coexistence and high diversity.

Finally, disturbances can vary in spatial extent (Altman and Whitlatch, 2007; Angelini and Silliman, 2012; Cadotte, 2007; Grau, 2002; Roxburgh *et al.*, 2004), affecting all or only certain parts of a system. Because species are usually not distributed evenly, the spatial extent and location of a disturbance can result in the complete loss of some species regardless of the magnitude or frequency of the disturbance. Spatial extent of disturbances is particularly important in patchy environments. In a patchy system, the community within each patch may develop independently and the age and diversity of communities can differ among patches as a function of their disturbance histories. Within such systems there may be two or more disturbance levels that promote the highest diversity, one for each patch which may vary with size and a mean disturbance rate for the whole system that maximizes the differences in species composition among patches and thus overall diversity. The degree to which the patches are in phase or disturbed at the same time and to the same degree will affect the impact of patchiness on intermediate disturbance (Abugov, 1982).

Disturbances can result from physical or biological processes and the source of a disturbance will determine its impact on diversity. Physical disturbances are the most common ones associated with intermediate disturbance. A long list of physical phenomena including wind, flooding, waves, fire, exposure to extreme heat or cold, increased UV radiation, low dissolved oxygen or anoxia, drought, extreme pH, etc. can cause various degrees of mortality and disturbance. Although the probability of any of these occurring may vary predictably with location or season, we usually see physical disturbances as unpredictable in the degree to which they affect populations and as being fairly unselective in terms of which species are affected. Species will differ in their tolerances to any of these and other physical phenomena, but these disturbances do not target particular species. Biological disturbances most commonly result from disease, herbivory, or predation all of which may target particular species and result in selective disturbance. If the targeted species are dominants then an intermediate level of disturbance is likely to produce higher diversities by making resources available to other species. If biological disturbances cause a disproportionate loss of rare or competitively inferior species, diversity could be lowered by exacerbating the rate at which dominants monopolize resources. This is not to say that biological disturbances are all selective. For example, a grazing limpet might remove all species from all or part of rock, regardless of whether they are food items. This disturbance would not differ from a similar removal through physical scraping.

The Effect of Intermediate Disturbance on Different Communities

Marine communities dominated by attached or sessile species have been seen as systems in which intermediate levels of disturbance facilitate increases in overall diversity. These communities include not only the coral reefs that led Connell (1978) to hypothesize the effects of intermediate disturbance, but also rocky intertidal invertebrate and algal communities subject to disturbances by waves, predators and herbivores, ice, and desiccation (Sousa, 1979; Svensson *et al.*, 2007), communities on subtidal cobbles and boulders that are turned over by waves or covered by sediment (Osman, 1977), salt marshes with differences in stress associated with tidal zone, sessile communities experimentally disturbed (Lenz *et al.*, 2004) or disturbed by territorial fish (Hixon and Brostoff, 1983), disturbed sandflat communities (Lee *et al.*, 2011), and phytoplankton communities (Gaedeke and Sommer, 1986; Grover and Chrzanowski, 2004). Although most systems have been found to conform to the predictions of the hypothesis, mixed results were found for at least one rocky intertidal habitat subjected to multiple types of disturbances including rock turnover by waves and sand burial, and gradients of temperature and desiccation stress as well as predation. In this habitat rock turnover had little effect on the diversity of algal community on top of rocks but intermediate disturbance by sand burial did increase the diversity of the fauna attached to the undersides of rocks (McGuinness, 1984). Intermediate disturbance has also been found to have mixed effects on the diversity of many infaunal invertebrate communities found in sedimentary environments (Lee *et al.*, 2011; Schratzberger and Warwick, 1999; Villnäs *et al.*, 2012; Zajac *et al.*, 2013). In these complex habitats with multiple resources, competition among infaunal species may be weak with no clear dominants. If resources are not limiting and competition is weak, then disturbance may have little effect on resource availability and, therefore, on diversity.

The effect of intermediate disturbance has been studied on a variety of scales within terrestrial plant communities. Examples include disturbances from storms to tree falls creating light gaps in tropical, temperate, and boreal forests (Molino and Sabatier, 2001; Woods, 2004), fire disturbance and herbivory in grasslands and prairies (Collins *et al.*, 1995; Seastedt and Knapp, 1993), disturbance from flooding, waves, and ice on riparian vegetation (Naiman and Decamps, 1997; Pollock *et al.*, 1998), agricultural disturbance and recovery of old fields (Armesto and Pickett, 1985), and bryophyte communities disturbed by tree falls (Jonsson and Esseen, 1990). In most of these communities patterns predicted by the intermediate disturbance hypothesis have been observed, but there have also been inconsistencies leading to alternative interpretations. In many ways the inconsistencies result from applying a deceptively simple yet somewhat vague hypothesis to very complex communities. For example, tree fall disturbances in tropical forest that create light gaps have been used as a prime example of diversity being maintained by intermediate disturbance (Denslow, 1987; Grau, 2002). The recolonization of these disturbed areas can be viewed as a lottery with all species having an equal chance of recruiting and vying for the newly-released resources and competition eventually causing the loss of some species. This view ignores differences in species not only in their competitive ability but in their ability to recruit. Species differ in their proximity to the disturbed area or their production of seeds that will cause differences in their probabilities of recruiting. Seed banks and degrees of seed dependence on disturbance for germination can also alter recruitment probabilities and herbivory as well as competition can determine survival. Recruitment limitation in which some species have little probability of recruiting into a particular gap can reduce the ability of disturbance to act as a mechanism for maintaining diversity (Hubbell *et al.*, 1999). That is, if recruitment is limited to only those species in mature communities surrounding the gap then increasing diversity by disturbance opening resources for early succession species will not occur. Nevertheless, it remains that the inclusion of early succession species or poor competitors as part of the forest increases overall diversity and resources must be renewed for them by disturbance or some other process. Alternatives such as the clustering of tree fall gaps that create larger gaps that increase the distance of central areas of the gap from the surrounding community may increase the recruitment probability of more distant or early succession species. Although it is clear that adaptations, differences in natural history of species, and chance patterns of distribution will influence the degree to which intermediate disturbance plays a role in maintaining diversity, the frequency, magnitude, and spatial pattern of disturbances must all be considered when evaluating its overall effect. Small tree fall gaps may simply be too small, while other disturbances may happen too frequently to contribute to the maintenance of diversity.

Similar mixed results have also been seen for grasslands and prairies as well as riparian vegetation (Pollock *et al.*, 1998). In prairies the variability among species in their adaptations and tolerances to fire as well as resources that may shift on different time scales can create differential, non-random responses to fire disturbance (Seastedt and Knapp, 1993). The heterogeneity or diversity of these systems appears dependent on the sizes of disturbed areas relative to the size of the whole system. In some experiments, lower diversities were generated at frequencies of one and four years with highest diversity seen after10 years (Collins *et al.*, 1995). This suggests that disturbance frequencies of 10 years or more could maintain the highest diversity in this system. For riparian systems there appear to be multiple sources of disturbance including flooding, erosion from waves, and ice damage and these vary with stream or river size, making it difficult to define an intermediate level of disturbance (Naiman and Decamps, 1997). Nevertheless, when a system is largely affected by one source such as flooding, diversity patterns largely agree with those predicted by the intermediate disturbance hypothesis (Pollock *et al.*, 1998).

As originally proposed by Connell, the intermediate disturbance hypothesis is less likely to be applicable to more motile species and there have been few investigations of its applicability to terrestrial animal communities. When examined, the effects of productivity and multiple trophic levels have been seen limiting the influence of intermediate disturbance in maintaining higher diversity. In some systems with rodent herbivores and carnivore predators, rodent diversity can be linked to variability in productivity with higher diversities at intermediate levels of productivity while carnivore diversity is greater at higher productivity levels (Owen, 1988). Unfortunately, productivity is related to rainfall and its variability can also be seen as a disturbance regime affecting the resources for rodents. The absence of a similar peak in carnivore diversity as seen with the rodents may be less a

refutation of the application of the intermediate disturbance hypothesis than identifying the level of rodent abundance that is necessary for maintaining the highest carnivore diversity (Owen, 1988). This may not coincide with highest rodent diversity. Investigations of less motile animals such as spiders indicate their diversity is greatest at intermediate levels of physical disturbance (Spiller and Schoener, 1988). On the other hand overlaying an additional disturbance from lizard predation has no effect; possibly because spider densities are already held below levels at which competitive exclusion would reduce diversity.

Aquatic systems exhibit mixed responses to intermediate disturbance. The invertebrate fauna of streams and rivers is highly motile, habitats are subject to a high degree of disturbance, the importance of competition in structuring these systems is not clear, and intermediate levels of disturbance are difficult to define. Nevertheless some studies have found patterns consistent with the intermediate disturbance hypothesis (Butler, 1989; Death and Winterbourn, 1995). Disturbance resulting from dragonfly predation produces highest species richness at intermediate levels and fish diversity in some, but not all rivers investigated is highest where disturbance is intermediate (Thorp and Cothran, 1984). Sessile stream species such as bryophyte communities that can be disturbed by both boulder turnover and ice also conform (Virtanen *et al.*, 2001). Dynamics in lakes in which fish predation is the source of disturbance seem to conform to the predictions of the hypothesis. In shallow areas with high predation and deep areas with few predators invertebrate prey diversity remains low, but with different species dominating the two extremes (Butler, 1989). In between these two zones predation and diversity are intermediate. Studies of plankton communities also have exhibited mixed results with diversity in some lakes being highest at intermediate levels of disturbance while in others diversity is most closely linked to the availability of resources (Grover and Chrzanowski, 2004).

The intermediate disturbance hypothesis has also been tested in laboratory mesocosm studies using experimental communities of microorganisms or bacteria. A number of experiments in these systems have produced evidence refuting the intermediate disturbance hypothesis (Huston, 2014). Although the communities used in these types of studies are usually limited to a fairly small number of species, the combinations of species as well as the disturbance regime can be manipulated. Some studies have simultaneously manipulated both frequency and magnitude of disturbance and have found that these characteristics of disturbance can interact such that frequent low magnitude and infrequent high magnitude disturbances both result in low diversity, but for different reasons (Polishchuk, 1999). Frequent low magnitude disturbances do little to disrupt the community or make resources available for new species. Infrequent high magnitude disturbances do free up resources but the long time period between disturbances allow competitively superior species to dominate and lower diversity. In a study of bacterial communities it was found that without strong competition disturbance regime had little effect on diversity (Wohl *et al.*, 2004).

Gradients and Landscapes

The intermediate disturbance hypothesis has been applied and tested in larger systems, in particular to spatial patterns of diversity along gradients and within landscapes. Gradients and zonation patterns often have communities dominated by different species at opposite ends and a mixture of both at intermediate locations, often with higher diversity. The diversity pattern is consistent with the effects of intermediate disturbance, but other factors can contribute to this pattern. Gradients can result not only from unidirectional variations in disturbance but from variations in one or more environmental variables. Gradients in factors such as temperature, light, pH, nutrients, soil type, rainfall, or salinity can result in different communities at the extremes based on species tolerances with the potential for a mixture of species or an ecotone between the two. Diversity will likely be higher at intermediate sites, producing a pattern consistent with the intermediate disturbance hypothesis but not resulting from an actual gradient in disturbance. Gradients in diversity with high intermediate values are not by themselves proof of the intermediate disturbance hypothesis.

Ecological landscapes represent complex spatial variation in habitats or communities and disturbance can be an element in producing these patterns. The relationship of landscape pattern to disturbance is clearest when disturbance acts to fragment a mature homogenous system into a terrain of patches of different ages or states of recovery from disturbances of different magnitudes (Paine and Levin, 1981; Roxburgh *et al.*, 2004). Some intermediate level of disturbance should produce a landscape of greatest diversity (Osman and Whitlatch, 1978). Given the potentially large spatial extent of landscapes, they can be a mixture of habitat types in which overall diversity results from habitat diversity and not disturbance or the landscape can be influenced by multiple types of disturbances that affect only some subset of habitats or fragments (Turner, 1989). If the landscape consists of habitable fragments or patches imbedded within an unsuitable matrix, such as ecological reserves within an urban area, then the intermediate disturbance hypothesis may not apply (Cushman and McGarigal, 2003).

Variations and Alternatives

The attractiveness of the intermediate disturbance hypothesis is its deceptive simplicity of applying some form of disruption to the observable process of community development. Clearly as a barren patch of habitat is colonized species richness should increase regardless of whether one envisions the community developing as a defined succession or as a random accumulation of species. Eventually resources should become scarce with competition causing some decline in diversity. It follows that if this process of change can be halted at the right time diversity will be maintained at the highest level. Reality, however, is not so tractable. Even within a homogeneous habitat, community development will vary as a consequence of random recruitment as well as differences

in competitive ability or susceptibility to predators. Recruitment, itself, may vary with season, distance from source, interspecific differences in productivity, etc. and not be random. Together these can result in delaying any competition for resources and a decline in diversity. Likewise, the simultaneous variability in the magnitude, frequency, and spatial scale of disturbances, whether they operate catastrophically or non-catastrophically, whether they are seen as operating within or among patches, or whether one or more types occur, all create a complexity that makes testing the hypothesis difficult.

This has resulted in both empirical and theoretical challenges to the hypothesis (Fox, 2013; Hughes *et al.*, 2007; Huston, 1979; Mackey and Currie, 2001). In two metaanalyses of published studies (Hughes *et al.*, 2007; Mackey and Currie, 2001) support for the intermediate disturbance hypothesis was found to be weak with less than 50% of the studies supporting the hypothesis. However, Sheil and Burslem (2013) and Huston (2014) have argued that other factors affecting diversity such as variation in productivity (Huston, 2014), the difficulty in sampling the complete temporal sequence of community development, the degree to which the level of disturbance examined was matched to the life histories of the organisms studied, and the specific exclusion of some types of communities (e.g. motile organisms; Connell, 1978) contribute to the expected patterns not being observed.

Fox (2013) presented a set of criticisms based on theory. He refuted the ability of intermediate levels of disturbance to increase diversity by (1) keeping species' densities low and reducing competitive exclusion, (2) interrupting competitive exclusion and preventing the system from reaching an equilibrium, and (3) creating variability in dominance. As Huston (2014) demonstrated, these arguments are focused on the ability of intermediate disturbance to maintain continually high diversity in a community at equilibrium, which is not part of the hypothesis as originally developed by Connell (1978). The hypothesis recognizes the dynamic nature of communities and how they develop and identifies a time during this development at which diversity may be higher before reductions from competitive exclusion. This point is neither stable or in equilibrium, but disturbance can keep the community more in this state than earlier or later states of lower diversity. Other theoretical and modeling studies have led to variations and modifications of the intermediate disturbance hypothesis, including variations in the community or ecological system, variations in disturbance characteristics, inclusion of positive interactions, and linking or contrasting other environmental parameters with disturbance (Cadotte, 2007; Gallet *et al.*, 2007; Hacker and Gaines, 1997; Sugden *et al.*, 2008). For example, when the complexity of the community is increased by the inclusion of multiple trophic levels or groups of species with very different life-histories, intermediate disturbance becomes harder to define (Jeppesen *et al.*, 2000; Lee *et al.*, 2011; Wootton, 1998). A single frequency or magnitude of disturbance is unlikely to affect species of vastly different sizes and generation times in the same way or necessarily be intermediate for all. Likewise, disturbances that maintain higher diversity and presumably lower population levels of prey may have the opposite effect on predators. Characteristics of disturbance such as how the timing or phasing of disturbances among patches can also modify the overall effect of disturbance (Abugov, 1982). Even at the same magnitude and frequency, whether a disturbance affects one or all patches at the same or different times will determine its impact and what level of disturbance produces the highest diversity. Other parameters such as recruitment (Hubbell *et al.*, 1999) or productivity (Grime, 1973a,b; Huston, 1979, 2014; Waide *et al.*, 1999) can also have a nonlinear impact on diversity and change the effect of disturbance. Low and high levels of recruitment may minimize the effects of disturbance. At low recruitment resources may not be limiting and disturbance would have little effect while at high recruitment levels any effect of disturbance would be quickly overwhelmed. Only at intermediate levels would differences in disturbance affect diversity.

There is a strong link between the intermediate disturbance hypothesis and productivity. Huston (1979, 2014) developed a more general, dynamic equilibrium model of diversity that incorporated specific links between the intermediate disturbance hypothesis and a similar intermediate productivity hypothesis that he traces back to Grime (1973a,b). Both hypotheses project a unimodal distribution of diversity with a peak at either an intermediate level of disturbance or productivity. Huston (2014) argued for much stronger empirical support for the intermediate productivity hypothesis and further suggested that intermediate levels of disturbance maximized diversity only at intermediate levels productivity. He suggested that a low levels of productivity; growth and reproductive rates would also be low, making it difficult for communities to take advantage of any resources made available after a disturbance. Likewise, at high levels of productivity high growth and reproductive rates would quickly limit any resources opened by disturbance. However, it could be argued that the effective rate or magnitude of a disturbance would shift with productivity making the 'ideal' intermediate rate a direct function of productivity.

In summary, the intermediate disturbance hypothesis remains a compelling concept that provides a non-equilibrium view of how communities can gain or maintain a richness in species without an intricacy of adaptations. Strong and weak competitors, good and poor colonizers, can all coexist in a system where some process such as disturbance prevents the monopolization of resources at different times and places. Intermediate disturbance remains an hypothesis that is far from proven or universally applicable to all communities. It needs to be tested in each community or habitat with an open mind as to what constitutes a community, a disturbance, and an intermediate frequency, magnitude, or spatial scale of the latter.

References

Abugov, R., 1982. Species diversity and phasing of disturbance. Ecology 63, 289–293.
Altman, S., Whitlatch, R.B., 2007. Effects of small-scale disturbance on invasion success in marine communities. Journal of Experimental Marine Biology and Ecology 342, 15–29.
Angelini, C., Silliman, B.R., 2012. Patch size-dependent community recovery after massive disturbance. Ecology 93, 101–110.
Armesto, J.J., Pickett, S.T.A., 1985. Experiments on disturbance in old-field plant communities: impact on species richness and abundance. Ecology 66, 230–240.

Bertocci, I., Maggi, E., Vaselli, S., Benedetti-Cecchi, L., 2005. Contrasting effects of mean intensity and temporal variation of disturbance on a rocky seashore. Ecology 86, 2061–2067.
Biswas, S.R., Mallik, A.U., 2011. Species diversity and functional diversity relationship varies with disturbance intensity. Ecosphere 2.art52.
Brooker, R.W., Maestre, F.T., Callaway, R.M., Lortie, C.L., Cavieres, L.A., Kunstler, G., Liancourt, P., Tielborger, K., Travis, J.M.J., Anthelme, F., Armas, C., Coll, L., Corcket, E., Delzon, S., Forey, E., Kikvidze, Z., Olofsson, J., Pugnaire, F., Quiroz, C.L., Saccone, P., Schiffers, K., Seifan, M., Touzard, B., Michalet, R., 2008. Facilitation in plant communities: the past, the present, and the future. Journal of Ecology 96, 18–34.
Butler IV, M.J., 1989. Community responses to variable predation: field studies with sunfish and freshwater microinvertebrates. Ecological Monographs 59, 311–328.
Cadotte, M.W., 2007. Competition-colonization trade-offs and disturbance effects at multiple scales. Ecology 88, 823–829.
Chesson, P., Huntly, N., 1997. The roles of harsh and fluctuating conditions in the dynamics of ecological communities. American Naturalist 150, 519–553.
Collins, S.L., Glenn, S.M., Gibson, D.J., 1995. Experimental analysis of intermediate disturbance and initial floristic composition: decoupling cause and effect. Ecology 76, 486–492.
Connell, J.H., 1978. Diversity in tropical rain forests and coral reefs. Science 199, 1302–1310.
Cushman, S.A., McGarigal, K., 2003. Landscape-level patterns of avian diversity in the Oregon coast range. Ecological Monographs 73, 259–281.
Death, R.G., Winterbourn, M.J., 1995. Diversity patterns in stream benthic invertebrate communities: the influence of habitat stability. Ecology 76, 1446–1460.
Denslow, J.S., 1987. Tropical rainforest gaps and tree species diversity. Annual Review of Ecology and Systematics 18, 431–451.
Fox, J.W., 2013. The intermediate disturbance hypothesis should be abandoned. Trends in Ecology and Evolution 28, 86–92.
Gaedeke, A., Sommer, U., 1986. The influence of the frequency of periodic disturbances on the maintenance of phytoplankton diversity. Oecologia 71, 25–28.
Gallet, R., Alizon, S., Comte, P.A., Gutierrez, A., Depaulis, F., van Baalen, M., Michel, E., Müller-Graf, C.D.M., 2007. Predation and disturbance interact to shape prey species diversity. American Naturalist 170, 143–154.
Gleason, H.A., 1917. The structure and development of the plant association. Bulletin of the Torrey Botanical Club 44, 463–481.
Grau, H.R., 2002. Scale-dependent relationships between treefalls and species richness in a neotropical montane forest. Ecology 83, 2591–2601.
Grime, J.P., 1973a. Competitive exclusion in herbaceous vegetation. Nature 242, 344–347.
Grime, J.P., 1973b. Control of species density in herbaceous vegetation. Journal of Environmental Management 1, 151–167.
Grover, J.P., Chrzanowski, T.H., 2004. Limiting resources, disturbance, and diversity in phytoplankton communities. Ecological Monographs 74, 533–551.
Hacker, S.D., Gaines, S.D., 1997. Some implications of direct positive interactions for community species diversity. Ecology 78, 1990–2003.
Hixon, M.A., Brostoff, W.N., 1983. Damselfish as keystone species in reverse: intermediate disturbance and diversity of reef algae. Science 220, 511–513.
Horn, H.S., 1974. The ecology of secondary succession. Annual Reviews of Ecology and Systematics 5, 25–37.
Hubbell, S.P., Foster, R.B., O'Brien, S.T., Harms, K.E., Condit, R., Wechsler, B., Wright, S.J., Loo de Lao, S., 1999. Light-gap disturbances, recruitment limitation, and tree diversity in a neotropical forest. Science 283, 554–557.
Hughes, A.R., Byrnes, J.E., Kimbro, D.L., Stachowicz, J.J., 2007. Reciprocal relationships and potential feedbacks between biodiversity and disturbance. Ecology Letters 10, 849–864.
Huston, M.A., 1979. A general hypothesis of species diversity. American Naturalist 113, 81–101.
Huston, M.A., 2014. Disturbance, productivity, and species diversity: empiricism vs. logic in ecological theory. Ecology 95, 2382–2396.
Hutchinson, G.E., 1951. Copepodology for the onithologist. Ecology 32, 571–577.
Hutchinson, G.E., 1953. The concept of pattern in ecology. Proceedings of the Academy of Natural Sciences of Philadelphia 105, 1–12.
Jeppesen, E., Jensen, J.P., SØndergaard, M., Lauridsen, T., Landkildehus, F., 2000. Trophic structure, species richness and biodiversity in Danish lakes: changes along a phosphorus gradient. Freshwater Biology 45, 201–218.
Jonsson, B.G., Esseen, P.-A., 1990. Treefall disturbance maintains high bryophyte diversity in a boreal spruce forest. Journal of Ecology 78, 924–936.
Lee, K.-M., Lee, S.Y., Connolly, R.M., 2011. Short-term response of estuarine sandflat trophodynamics to pulse anthropogenic physical disturbance: Support for the Intermediate Disturbance Hypothesis. Estuarine, Coast and Shelf Science 92, 639–648.
Lenz, M., Molis, M., Wahl, M., 2004. Experimental test of the intermediate disturbance hypothesis: frequency effects of emersion on fouling communities. Journal of Experimental Marine Biology and Ecology 305, 247–266.
MacArthur, R.H., Wilson, E.O., 1963. An equilibrium theory of insular zoogeography. Evolution 17, 373–387.
MacArthur, R.H., Wilson, E.O., 1967. The theory of island biogeography. Princeton, NJ: Princeton University Press.
Mackey, R.L., Currie, D.J., 2001. The diversity–disturbance relationship: is it generally strong and peaked? Ecology 82, 3479–3492.
McGuinness, K.A., 1984. Species-area relations of communities on intertidal boulders: testing the null hypothesis. Journal of Biogeography 11, 439–456.
Miller, T.E., 1982. Community diversity and interactions between the size and frequency of disturbance. American Naturalist 120, 533–536.
Miller, A.D., Roxburgh, S.H., Shea, K., 2011. How frequency and intensity shape diversity–disturbance relationships. Proceedings of the National Academy of Sciences 108, 5643–5648.
Molino, J.-F., Sabatier, D., 2001. Tree diversity in tropical rain forests: a validation of the intermediate disturbance hypothesis. Science 294, 1702–1704.
Naiman, R.J., Decamps, H., 1997. The ecology of interfaces: riparian zones. Annual Review of Ecology and Systematics 28, 621–658.
Osman, R.W., 1977. The establishment and development of a marine epifaunal community. Ecological Monographs 47, 37–63.
Osman, R.W., Whitlatch, R.B., 1978. Patterns of species diversity: fact or artifact? Paleobiology 4, 41–54.
Owen, J.G., 1988. On productivity as a predictor of rodent and carnivore diversity. Ecology 69, 1161–1165.
Paine, R.T., Levin, S.A., 1981. Intertidal landscapes: disturbance and the dynamics of pattern. Ecological Monographs 51, 145–178.
Polishchuk, L.V., 1999. Contribution analysis of disturbance-caused changes in phytoplankton diversity. Ecology 80, 721–725.
Pollock, M.M., Naiman, R.J., Hanley, T.A., 1998. Plant species richness in riparian wetlands – a test of biodiversity theory. Ecology 79, 94–105.
Roxburgh, S.H., Shea, K., Wilson, J.B., 2004. The intermediate disturbance hypothesis: patch dynamics and mechanisms of species coexistence. Ecology 85, 359–371.
Schratzberger, M., Warwick, R.M., 1999. Impact of predation and sediment disturbance by *Carcinus maenas* (L.) on free-living nematode community structure. Journal of Experimental Marine Biology and Ecology 235, 255–271.
Seastedt, T.R., Knapp, A.K., 1993. Consequences of nonequilibrium resource availability across multiple time scales: the transient maxima hypothesis. American Naturalist 141, 621–633.
Shea, K., Roxburgh, S.H., Rauschert, E.S.J., 2004. Moving from pattern to process: coexistence mechanisms under intermediate disturbance regimes. Ecology Letters 7, 491–508.
Sheil, D., Burslem, D.F.R.P., 2013. Defining and defending Connell's intermediate disturbance hypothesis: a response to Fox. Trends in Ecology and Evolution 28, 571–572.
Sousa, W.P., 1979. Disturbance in marine intertidal boulder fields: the nonequilibrium maintenance of species diversity. Ecology 60, 1225–1239.
Spiller, D.A., Schoener, T.W., 1988. An experimental study of the effect of lizards on web-spider communities. Ecological Monographs 58, 57–77.
Sugden, H., Lenz, M., Molis, M., Wahl, M., Thomason, J.C., 2008. The interaction between nutrient availability and disturbance frequency on the diversity of benthic marine communities on the north-east coast of England. Journal of Animal Ecology 77, 24–31.
Svensson, J.R., Lindegarth, M., Siccha, M., Lenz, M., Molis, M., Wahl, M., Pavia, H., 2007. Maximum species richness at intermediate frequencies of disturbance: consistency among levels of productivity. Ecology 88, 830–838.
Thorp, J.H., Cothran, M.L., 1984. Regulation of freshwater community structure at multiple intensities of dragonfly predation. Ecology 65, 1546–1555.
Turner, M.G., 1989. Landscape ecology: the effect of pattern on process. Annual Review of Ecology and Systematics 20, 171–197.

Villnäs, A., Norkko, J., Lukkari, K., Hewitt, J., Norkko, A., 2012. Consequences of increasing hypoxic disturbance on benthic communities and ecosystem functioning. PLoS One 7, e44920.
Virtanen, R., Muotka, T., Saksa, M., 2001. Species richness-standing crop relationship in stream bryophyte communities: patterns across multiple scales. Journal of Ecology 89, 14–20.
Waide, R.B., Willig, M.R., Steiner, C.F., Mittelbach, G., Gough, L., Dodson, S.I., Juday, G.P., Parmenter, R., 1999. The relationship between productivity and species richness. Annual Review of Ecology and Systematics 30, 257–300.
Wohl, D.L., Arora, S., Gladstone, J.R., 2004. Functional redundancy supports biodiversity and ecosystem function in a closed and constant environment. Ecology 85, 1534–1540.
Woods, K.D., 2004. Intermeidate disturbance in a late-successional hemlock-northern hardwood forest. Journal of Ecology 92, 464–476.
Wootton, J.T., 1998. Effects of disturbance on species diversity: a multitrophic perspective. American Naturalist 152, 803–825.
Zajac, R.N., Vozarik, J.M., Gibbons, B.R., 2013. Spatial and temporal patterns in macrofaunal diversity components relative to sea floor landscape structure. PLoS One 8, e65823.

Keystone Species and Keystoneness

Simone Libralato, Istituto Nazionale di Oceanografia e di Geofisica Sperimentale—OGS, Trieste, Italy

Glossary

Ecological network analysis Is a methodology to analyze a system of interactions to identify general properties.
Food web model A conceptual or quantitative representation of trophic interactions in an ecosystem.
Interaction strength Is a quantification of the magnitude of effect of one species on another.
Species functional role Main activity characterizing a species or groups of species as a contribution to the dynamics of the system.
Trophic level The position of an organism or species in the food chain.

Keystone Species

Keystones are species whose modifications in abundance cause large effects on the ecological communities they are part (Paine, 1969). These effects are disproportionate with respect to their biomass proportion in the system (Power *et al.*, 1996).

History of the Keystone Species Concept

In the 1960s, Professor Robert T. Paine, while conducting experiments on manipulation of rocky shore communities, noticed that removing the sea star predator *Pisaster ochraceus* resulted in overgrowth of the mussel *Mytilus californicus* (Paine, 1966). The experiment was conducted in the field on rocky shores of the west coast of United States, by manually removing *P. ochraceus* from areas and then protect with metallic cages the plots from sea star predation activity on species in the rocky shores. Under the absence of predator, the mussel become dominant in abundance with consequent reduction of biodiversity. In his later seminal work, Paine defined the sea star as a keystone species (Paine, 1969), that is, a consumer whose intense predation keep at low density potentially dominant species and thus it maintains high the biodiversity. Such work put the attention to the fact that not all predators' alteration result in large community changes, that is, species have not the same functional role. Some species are called keystones because removing them or altering their density result in large perturbations of the communities around them (Paine, 1969).

The large influences of keystones on the communities were connected with the presence of a set of trophic interactions between the keystone species and other surrounding species, not necessarily strong but critical for the prey itself to keep it under control. This original keystone species concept was mainly focused around the approach based on food webs but the introduction of the interaction strength concept and quantification of direct and indirect measurement of interactions introduced new avenues for identification of keystones. These approaches for determining the keystone species resulted in a flourishing of new meanings on the concepts of keystones species (Mills *et al.*, 1993) creating confusion with other ecologically relevant species (see Piraino *et al.*, 2002).

The need to clarify the keystone species concept and avoid confusion resulted in new practical definition that include the accounting of effects per unit of biomass (Power *et al.*, 1996). The peculiarities of keystone species, to distinguish them from dominant species, are that they exert influences on communities through intense predation, thus their direct and indirect effects on other species are disproportionally large compared to their biomass in the system (Power *et al.*, 1996; see **Fig. 1**). Further additional characteristics to determine keystone species included the prevalence of top–down effects and dominance within a functional group (Davic, 2003).

Examples of Keystone Species

Species that exert important impact on many others are routinely observed through trophic cascades analyses from both controlled manipulations and natural experiments evaluated ex-post (Estes *et al.*, 1978; Myers *et al.*, 2007). For example large shark species in the Atlantic Ocean, which includes requiem (Family Carcharhinidae) and hammerhead sharks (Family Sphyrnidae), are ecologically important because they consume mesopredators, which in turn consume estuarine shellfish, such as scallops (*Argopecten irradians*) and hard clams (*Mercenaria mercenaria*). Long-term exploitation resulted in strong reduction of sharks populations which resulted beneficial for rays, and eventually lead to the collapse of estuarine shellfish populations in Core Sound, North Carolina, United States (Myers *et al.*, 2007).

The reduction in the number of sea otter (*Enhydra lutris*) on Alaskan coasts, also revealed it to be a keystone. In normal conditions, sea otter keep low the abundance of the sea urchin through intensive predation, but as the number of sea otters decline

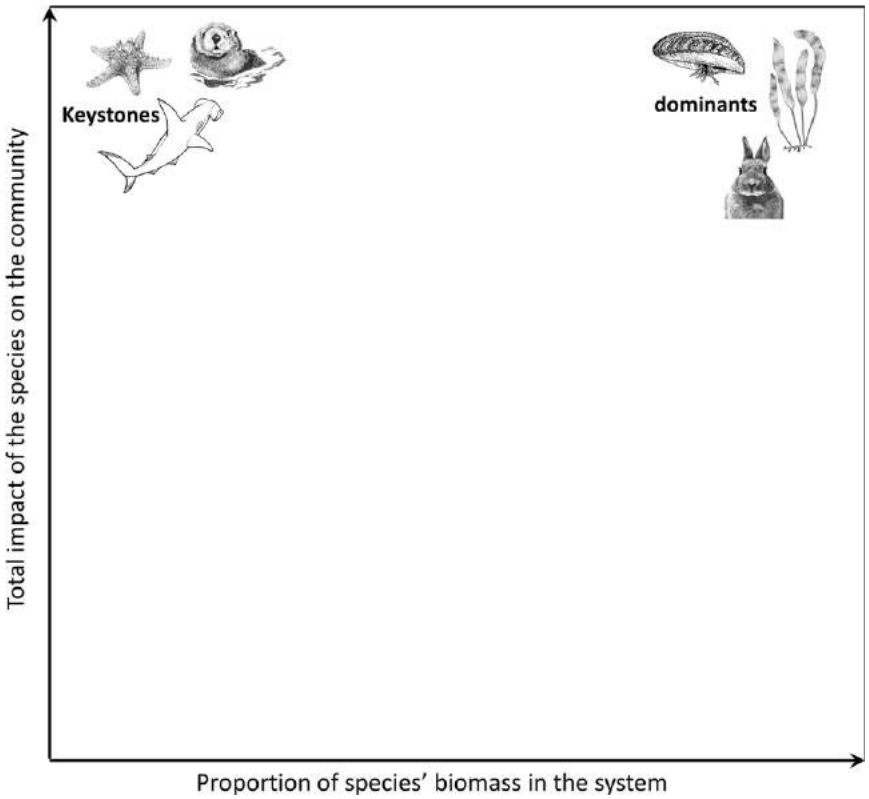


Fig. 1 The schematic representation of the species role on the basis of definition provided by Power *et al.*, 1996. In that essay both keystone and dominants species are considered having high impacts on the community, but they are distinguished in terms of biomass proportion in the system. Redrawn from Power, M.E., Tilman, D., Estes, J.A., Menge, B.A., Bond, W.J., Mills, L.S., Daily, G., Castilla, J.C., Lubchenco, J., Paine, R. T., 1996. Challenges in the quest for keystones. Bioscience 46, 609–620.

the flourishing on sea urchin population result in great reduction of kelp forest (*Laminaria* spp) with relevant consequences and changes for the whole community (Estes *et al.*, 1978).

On terrestrial ecosystem, for example, the Iberian lynx was defined as keystone, as its predation keep rabbit number under control otherwise overgrowing and changing completely the community.

Importance for Conservation and Management

Since the first introduction of the concept, the large interest for the keystone was related to the possibility to identify target priority species for protection and maximize effects of conservation (Paine, 1969; Mills *et al.*, 1993). Although likelihood of extinction is a good reason often used for guiding conservation efforts, rare species are not necessarily the most important species ecologically. Conversely, the ecological role that species play in the ecosystem may be a better measure of the species' need for protection and the keystone species concept allows to place attention on priority species defined on the basis of the role in the ecosystem. Conserving keystone species, in fact, would preserve ecosystem functioning and ecosystem services that are at the very basis of biodiversity maintenance but would also have a large impact on many other species of the food web, including the more rare ones (Chapin *et al.*, 1997). Therefore, keystone concept facilitate meaningful prioritization of conservation interventions and their ecological efficacy. Preserving keystone species, moreover, would allow maintaining high the biodiversity and thanks to control of high competitive species, would allow preserving also rare species and favor ecological redundancy in ecosystems. Overall the keystone species would support maintenance of ecosystem resilience an attribute fundamental for contrasting effects of perturbations, such as those induced by climate change.

Difficulties in Determining Keystone Species

Empirical determination of keystone species is, however, difficult. Experimental manipulation has been the most fundamental method for identify species role and their contribution to ecosystem functioning. It consists in imposing controlled changes on species density to detect community effects: it is a very relevant approach to determine local effects, but empirical determination has relevant limitations. Experiments limited to microscopic communities, in fact, can be successfully carried out in laboratory but direct field experiments on larger invertebrates or vertebrates are not always possible. Field experiments, in fact, are usually

restricted to species with no or limited movement ability such as plants and macrobenthic species, mainly sessile or bottom-moving species and can be carried out on very small plots, where it is very difficult to document a population level response and to describe the space-temporal heterogeneity of physical conditions and biological densities of natural systems. Moreover, direct and indirect effects of biological and anthropic activities can propagate along the components of the ecosystem reaching species not directly impacted, and produce important changes on the ecosystem functioning. These "cascade effects" are usually identified on the basis of observations on few taxa, while models can represent them in their completeness (Libralato *et al.*, 2006).

Experimental approach, moreover, need to be conceived to last in proportion of the life span of the species involved: in case of "natural experiments," therefore, medium term observations are often required, thus increasing their cost, a factor which considerable limit the application of experimental methods. Moreover, since the environment is characterized by stochastic variability and fluctuations, one further limitation is related to the need of distinguishing the results of ecosystem interactions from those of environmental noise: excluding the influence of all other possible factors other than those imposed is not easy and it is not always practicable. Mathematical models allow one to overcome some of the limits outlined for the experimental approach, and represent a valuable shortcut and a priori analysis for subsequent ad hoc experimental measurements (Jordán, 2009).

Keystoneness

Determining Keystone Species From Food Web Models

Several metrics applied to ecological network models such as regular equivalence and other topological measures demonstrated to be useful to quantify the trophic importance of the species and the identification of keystone species (Jordán *et al.*, 2008). The effects of removal in terms of secondary extinctions also revealed useful for the identification of important species. Food web and ecosystem model perturbation, moreover, can also help predicting the ecosystem effects of removal and thus were used to determine keystones. However, the keystoneness derived on the basis of mixed trophic impact analysis on quantitative food web networks demonstrated to be a simple, efficient and useful concept to determine keystone species from a wide range of models. Keystoneness is conceived as a characteristic of any species, and can be used to rank the species or group of species. Only species ranking high in keystoneness would be classified as keystone, but all can have a degree of keystoneness.

This continuous property allows accounting for the fact that species have a potential for explicit their keystone role that also vary from local conditions and can change in space and in time.

Keystoneness is determined for all nodes of an ecological network on the basis of the mixed trophic impact analysis. The approach allows determining not only the total effects of a taxa on the food web, but also to distinguish them into top–down and bottom–up effects. The mixed trophic analysis allows for quantify both keystoneness and top–down contribution of each node of a food web.

Mixed Trophic Impact to Quantify Effects of a Species on the Food Web

The mixed trophic impact analysis results from the application of input output analysis on quantitative weighted network of trophic interactions and permits to account for the direct and indirect impacts that any single node have on any other node of the network. For each couple of nodes (i and j) of network with n nodes it is possible to calculate the net direct impact q_{ij} as the difference between the direct positive impact that a prey flow has for the predator (g_{ij}) and the direct negative impact that a predator has on the prey (f_{ji}). These direct impacts are calculated on the basis of the flux of predation from prey i to predator j, T_{ij} (Ulanowicz and Puccia, 1990). The positive direct impact g_{ij} is calculated as the proportion of the flow T_{ij} with respect to the total consumptions of the predator:

$$g_{ij} = \frac{T_{ij}}{\sum_{k}^{n} T_{kj}}$$

And the negative direct impact as the proportion of the weighted link T_{ij} over all possible consumptions on prey i:

$$f_{ij} = \frac{T_{ij}}{\sum_{k}^{n} T_{ik}}$$

The net direct impact is thus calculated as:

$$q_{ij} = g_{ij} - f_{ji}$$

The matrix of direct net impacts Q includes all the q_{ij} elements for the network of n nodes. It can be demonstrated that the propagation of each net direct impact to the food web can be calculated by multiplying q_{ij} elements of any distinct pathway in the food web. Multiplying Q by itself (Q at the power of 2) allows looking for second order impacts; multiplying Q three times (Q at the power of 3) effects of third order, etc.

Mathematically the complete set of direct and indirect impacts can be obtained by the sum of all integer powers of the matrix Q resulting in a matrix of mixed trophic impacts M. The mixed trophic impact matrix of a network with n nodes, is a squared matrix

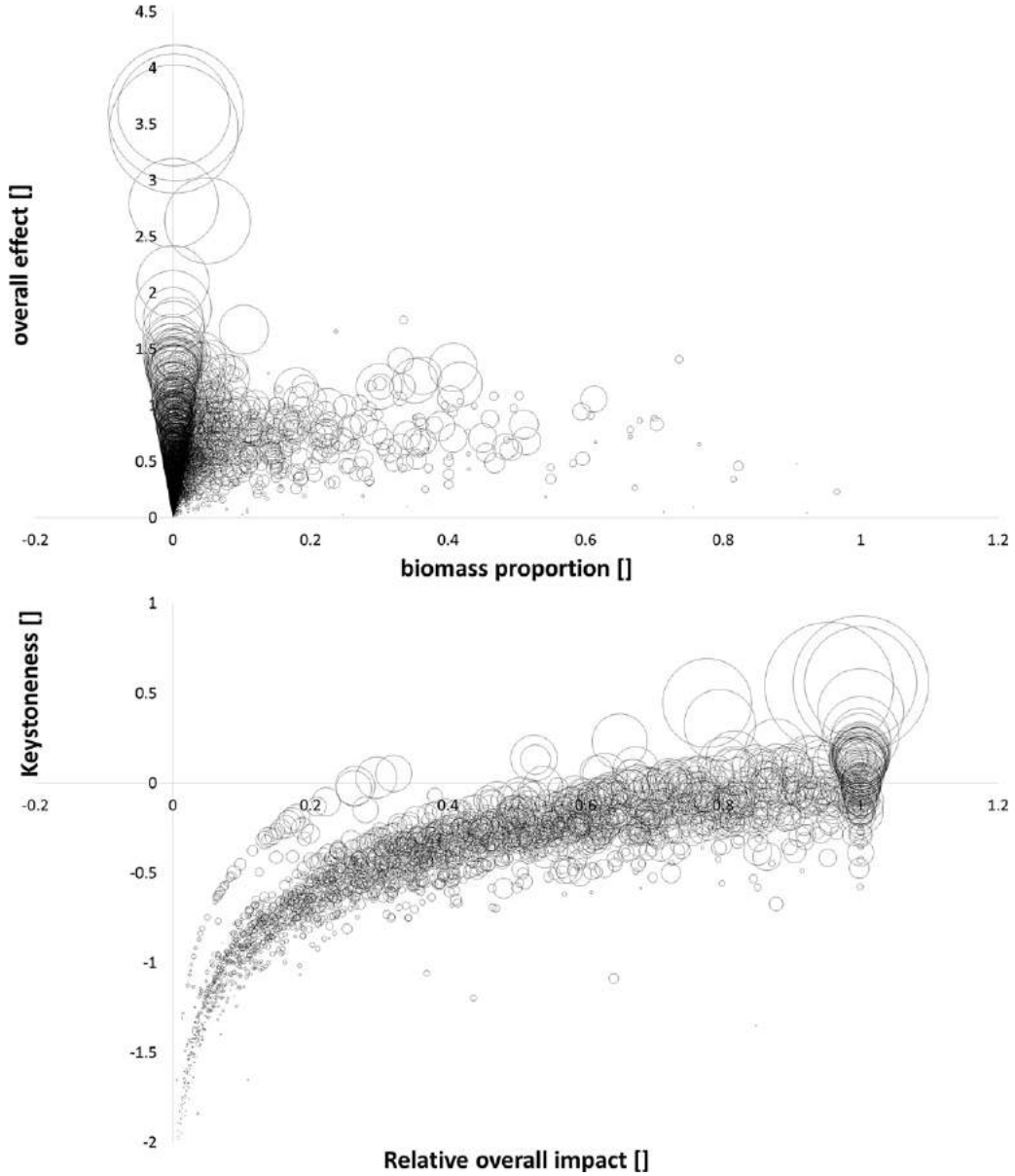


Fig. 2 Overall effect, biomass proportion and keystoneness for the 2635 functional groups of the 107 food webs of marine ecosystems. Area of the bubbles is proportional to the top–down impact measured as the sum of negative MTI elements. Upper panel is built in conformity with **Fig. 1**. Lower panel represent the keystoneness in a common graphics, showing keystones as the groups with keystoneness close to 0 or above.

with $n \times n$ elements, m_{ij}. Each element of the matrix can be positive or negative and represent the impact of the node i on the node j of the network through all possible direct and indirect pathways propagating through the food web (Ulanowicz and Puccia, 1990).

It has been demonstrated that the elements m_{ij} of the matrix M are corresponding to the sensitivity of the node j to perturbations of the node i at the equilibrium. This equivalence was demonstrated on the basis of results from a set of simulations obtained by changing density, biomass or productivity of the group i and evaluating the long term biomass changes on the other groups of the web (Libralato *et al.*, 2006). This equivalence set the basis for defining overall effects (direct and indirect) of a group on all the others of the web, on the basis of the matrix M. It was thus proposed to consider all possible direct and indirect effects produced by a species, as the sum of absolute values of perturbation induced in the food web by a change in one node. Namely, the overall effect $\mathscr{E}_i$ of the node i is calculated as the of all the effects of i over the whole food web. In order to avoid that positive and negative factors eliding each other, for accounting the total overall effect the squared elements of m_{ij} are summed by row as in the following:

$$\mathscr{E}_i = \sqrt{\sum_{j=1}^{n} m_{ij}^2 - m_{jj}^2}$$

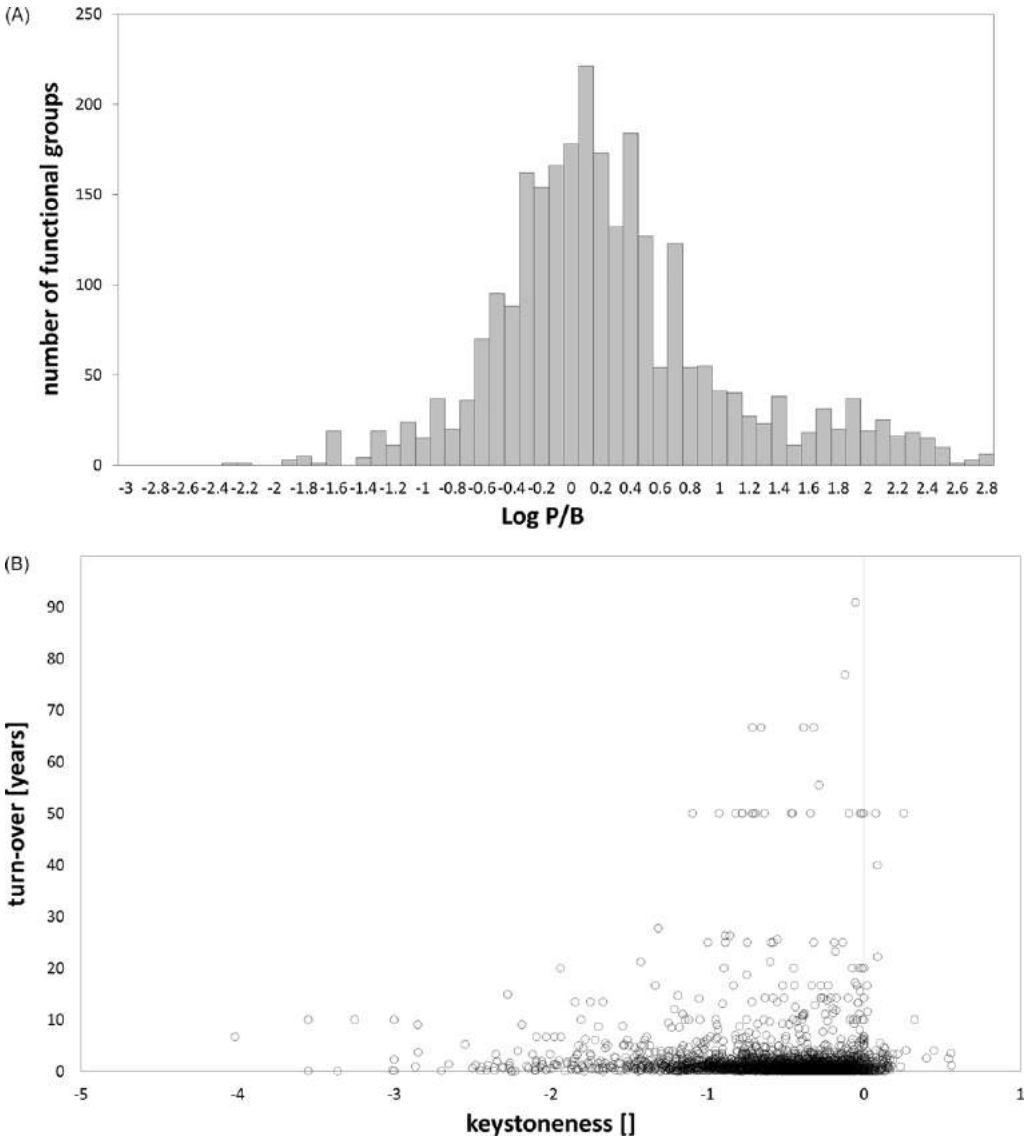


Fig. 3 (A) Turn-over (in years) expressing average longevity of the population as a characteristics of species for the 2635 functional groups of the 107 food webs. Data showed to be normally distributed. (B) Results in terms of keystoneness for the 2635 functional groups with respect to turn-over as a measure of longevity. Results highlight a tendency for high keystoneness to be related with high turnover, that is, to species characterized by high longevity.

where the self-effects are not considered. The overall effect $\mathscr{E}_i$ is thus a synthetic measure of the impact that a species has on all the others and can be considered for evaluating species role.

Notably, it is also possible to consider the contribution of positive and negative m_{ij} elements on the overall effect. Considering that negative m_{ij} indicates a species with direct and indirect negative effects on another, negative m_{ij} is indicative of a top–down effect (Libralato *et al.*, 2006). Exception to this is represented by the beneficial predator (Ulanowicz and Puccia, 1990): a species whose positive indirect effects on the prey are larger than negative impact on a prey due to predation removal (negative impact). Since beneficial predation is quite exceptional feature, considering negative m_{ij} as top–down effects is quite robust (Libralato *et al.*, 2006). Conversely positive m_{ij} is more related to bottom–up effects.

Keystoneness as a Combination of Overall Effect and Biomass Proportion

In line with per capita interaction strength as a measure of keystone role, the impact of a species ($\mathscr{E}_i$) contrasted with its biomass proportion in the system (p_i) was used as a way to distinguish keystone species from dominants (Power *et al.*, 1996). Using the same approach it is possible to contrast overall effect determined from ecological networks on the basis of mixed trophic impact

analysis and the biomass proportion (Libralato *et al.*, 2006). Using these considerations several measures of species importance were evaluated and the keystoneness was chosen as:

$$KS_i = \log\left[\mathcal{E}_i \bullet (1 - p_i)\right]$$

Each node of an ecological network is thus characterized by a keystoneness KS_i which can assume any value in the range from 1 to negative infinity (**Fig. 2**). The keystoneness allows to rank species by keystone role played and by convention species or functional groups with *KS* close or above 0 are to be considered with high keystoneness and thus keystone species (Libralato *et al.*, 2006). Other ways of combining overall effects and biomass proportion in the system have been formally tested highlighting that new possible weighting and/or ranking could be used (Valls *et al.*, 2015), but confirming the potential of these two factors determined from food web models as a basis for determining keystone species.

Applications of Keystoneness

Broad meta-analyses of *KS* application on several marine food webs highlighted that species with high *KS* have also high contribution of negative mixed trophic impacts (**Fig. 2**), i.e., a large proportion of their large impacts are due to top–down effects (Heymans *et al.*, 2014). Among the 2635 functional groups analyzed from a set of 107 ecological networks representing marine ecosystem all over the world highlighted the presence of several high trophic level species having high keystoneness. Therefore keystone species identified included in particular top predators such as large pelagics and sharks (Libralato *et al.*, 2006; Heymans *et al.*, 2014).

Although in the 107 food web models there are many groups long living, such as seals, sharks, whales, turtles, the majority of taxa represented have short generation time and the average turn-over is 2.5 years (**Fig. 3**A) . Species or groups of species with high keystoneness (high impact on the food web) were long living taxa (**Fig. 3**B). This implies that we need longer time series to help our models (and modelers) to be more accurate in their ecosystem predictions.

See also: Ecological Data Analysis and Modelling: Climate Change Models

References

Chapin, F.S., Walker, B.H., Hobbs, R.J., Hooper, D.U., Lawton, J.H., Sala, O.E., Tilman, D., 1997. Biotic control over the functioning of ecosystems. Science 277 (5325), 500–504.

Davic, R.D., 2003. Linking keystone species and functional groups: A new operational definition of the keystone species concept. Conservation Ecology 7 (1), r11. http://www.consecol.org/vol7/iss1/resp11/ (online).

Estes, J.E., Smith, N.S., Palmisano, J.F., 1978. Sea otter predation and community organization in the western Aleutian Islands, Alaska. Ecology 59 (4), 822–833.

Heymans, J.J., Coll, M., Libralato, S., Morissette, L., Christensen, V., 2014. Global patterns in ecological indicators of marine food webs: A modelling approach. PLoS One 9. e95845

Jordán, F., 2009. Keystone species and food webs. Philosophical Transactions of the Royal Society, B: Biological Sciences 364 (1524), 1733–1741.

Jordán, F., Okey, T.A., Bauer, B., Libralato, S., 2008. Identifying important species: Linking structure and function in ecological networks. Ecological Modelling 216 (1), 75–80.

Libralato, S., Christensen, V., Pauly, D., 2006. A method for identifying keystone species in food web models. Ecological Modelling 195, 153–171.

Mills, L.S., Soulé, M.E., Doak, D.F., 1993. The keystone-species concept in ecology and conservation. Bioscience 43 (4), 219–224.

Myers, R.A., Baum, J.K., Shepherd, T.D., Powers, S.P., Peterson, C.H., 2007. Cascading effects of the loss of apex predatory sharks from a coastal ocean. Science 315 (5820), 1846–1850.

Paine, R.T., 1966. Food web complexity and species diversity. The American Naturalist 100 (910), 65–75.

Paine, R.T., 1969. A note on trophic complexity and community stability. The American Naturalist 103 (929), 91–93.

Piraino, S., Fanelli, G., Boero, F., 2002. Variability of species' roles in marine communities: Change of paradigms for conservation priorities. Marine Biology 140, 1067–1074.

Power, M.E., Tilman, D., Estes, J.A., Menge, B.A., Bond, W.J., Mills, L.S., Daily, G., Castilla, J.C., Lubchenco, J., Paine, R.T., 1996. Challenges in the quest for keystones. Bioscience 46, 609–620.

Ulanowicz, R.E., Puccia, C.J., 1990. Mixed trophic impacts in ecosystems. Coenoses 5 (1), 7–16.

Valls, A., Coll, M., Christensen, V., 2015. Keystone species: Toward an operational concept for marine biodiversity conservation. Ecological Monographs 85 (1), 29–47.

Relevant Websites

http://ecopath.org/—Ecopath with Ecosim ecosystem modeling.
https://www.britannica.com/science/keystone-species—Encyclopedia Britannica.
https://www.iucn.org/—IUCN, International Union for Conservation of Nature.
https://www.nationalgeographic.org/encyclopedia/keystone-species/—National Geographic.

Leaf Area Index

NJJ Bréda, National Institute for Agricultural Research (INRA), Champenoux, France

Introduction

The plant canopy is a site of physical and biochemical processes associated with the terrestrial biosphere. The functional and structural attributes of plant canopies are dependant on species composition, microclimatic conditions, nutrient dynamics, herbivore activities, and many other activities like management. The amount of foliage in a plant canopy is one of the basic ecological characteristics reflecting the integrated effects of these factors in an ecosystem. In turn, canopy leaf area is the dominant driving force of primary production, water and nutrient use, energy exchange, and other physiological functions of a range of ecosystem processes. Understanding the organization and function of plant canopies is of central importance when conducting many types of comparative ecological studies or when developing biophysical Earth system models involving water and carbon balances. Yet, characterizing plant canopies presents many challenges, largely because of their complex geometry, and because of the difficulties of obtaining meaningful quantitative indices that relate back to fundamental processes such as light interception, transpiration, and photosynthesis. Ecophysiologists, managers (farmers and foresters), ecologists, climate and weather forecast modelers, ecosystem modelers, and atmosphere–ecosystem interaction modelers, request information about canopy leaf area index (LAI), one of the most widely used descriptors of the canopy.

LAI is a measure of canopy foliage content commonly used in studies of vegetation and ecosystems. LAI is the total area of one side of the leaf tissue per unit area of ground surface. According to this definition, LAI is a dimensionless quantity characterizing the canopy of an ecosystem. One unit of LAI is equivalent to 10 000 m^2 of leaves per hectare. LAI has been recognized as the most important attribute of vegetation structure for characterizing canopies from the stand to large areas at broad spatial scales. In defining ecology as the study of the structure and function of ecosystems, LAI is one of the core parameters in ecology, as it links canopy structure and ecosystem function.

Magnitude of LAI across the World

Biomes are major biogeographic regions consisting of distinctive plant life forms (e.g., forest, grasslands, desert, etc.). LAI ranges from 1.3 $\pm$ 0.9 for deserts to 8.7 $\pm$ 4.3 for tree plantations, and up to 20 depending on the biome. Temperate evergreen forest (needle leaf and broadleaf) displays the highest average LAI (5.1–6.7) out of the natural terrestrial vegetation classes (Fig. 1). Biomes with the highest LAI values are tree plantations, temperate evergreen forests, and wetlands. Exceptionally high values have been reported for hybrid poplars grown under intensive culture which could develop LAI values of 16–45, depending on the tree spacing. Those with the lowest LAI values are deserts, grasslands, and tundra.

Environmental Controls of LAI

As climate (especially the mean and variations in annual precipitation and temperature) is the primary force shaping the major biomes of the world, much of the observed pattern of LAI distribution is initially driven by similar climatic factors. Second, biome distribution is controlled by edaphic conditions (water supply and soil fertility), which also control LAI. According to the resources optimization theory, LAI may adjust to climate and site potential. Reviews of plant science literature have computed the response of LAI to variations in soil moisture, soil fertility, and atmospheric CO_2. LAI is probably co-limited by a number of resources, including water, nitrogen, and light. A linear response of LAI to N was reported for some crops or coniferous species, but is not uniform for all plant species, soil nutrients, or fertilization rates. Fertilization (nitrogen, phosphorus, or potassium) strongly increases LAI but the response could be a short-term one, after which an acclimatization of the canopy occurs. A new steady state is adjusted in the following growing seasons, especially in terms of leaf area versus root ratio. LAI sensitivity decreases (i.e., LAI saturates) indicating that something other than soil fertility is a limiting factor for canopy development. In many cases, water supply acts as a strong limitation resulting from LAI increase and related water uptake needs. The response to increasing soil water content is close to that of soil fertility. Increasing soil water availability in soils suffering from severe drought causes a significant increase in LAI. Interestingly, the response to increases in atmospheric CO_2 is nonlinear. LAI curves for crops and plant communities indicate a strong response of LAI to increases in atmospheric CO_2 up to the current ambient content. Afterward, the impact is more limited. At some resource thresholds, the addition of fertilizers or water will have no further influence on LAI. The saturation of LAI is likely to be indicative of light limitation, due to self-shading of leaves and negative carbon balances in low canopy layers. A large proportion of natural ecosystems lies below a threshold of optimal resource availability. LAI saturation is a threshold beyond which any further increase in LAI is compensated for by a negative carbon balance in shaded lower canopy leaves.

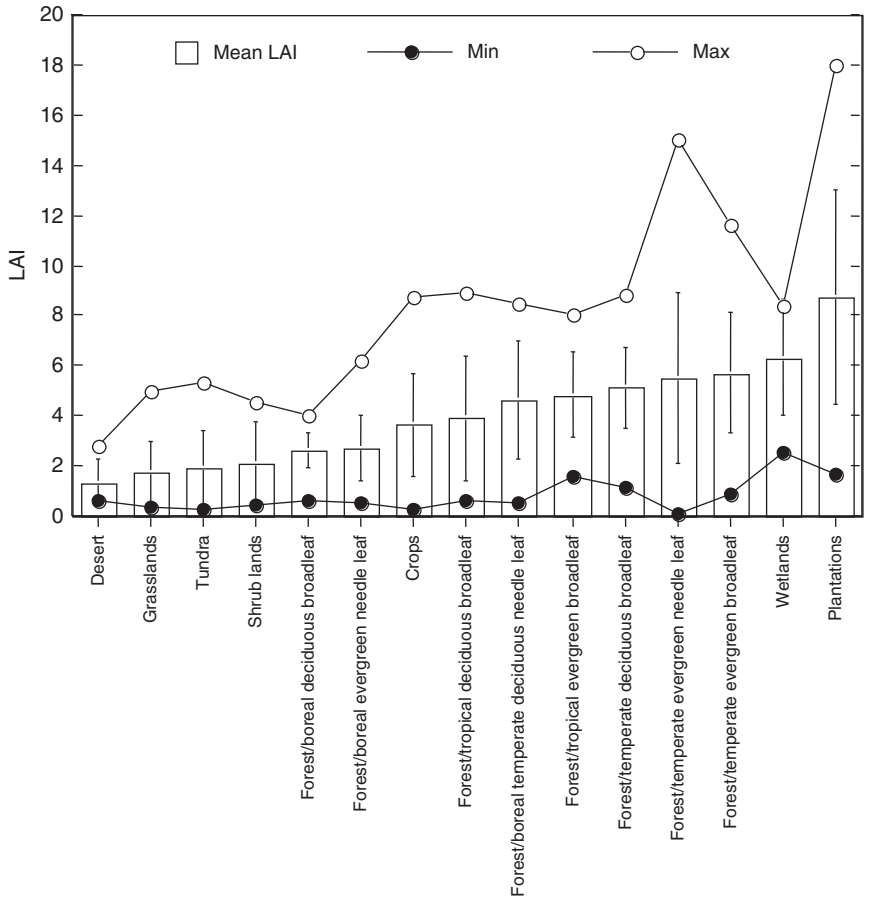


Fig. 1 Mean LAI (+/− standard deviation) of biomes and cover types. Data from Scurlock, J.M.O., Asner, G.P., Gower, S.T., 2001. Worldwide Historical Estimates and Bibliography of Leaf Area Index, 1932–2000. ORNL Technical Memorandum TM-2001/268. Oak Ridge, TN: Oak Ridge National Laboratory. Global Leaf Area Index Data from Field Measurements, 1932–2000. Data set available online (http://daac.ornl.gov) from the Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA.

Natural Temporal Variation in LAI

Both seasonal and interannual dynamics of LAI are characteristic of ecosystems, and seasonal dynamics of LAI is one of the aspects of phenology and life form. Evergreen versus broadleaved forests exhibit contrasting seasonal progression of LAI, with seasonal variation of less than 10% of LAI for evergreen. LAI expansion in deciduous species occurs within 1 month from budburst to maximum LAI which remains quite constant over the growing season and then decreases at leaf fall.

On an interannual time basis, adjusting LAI is a response of the ecosystem to cope with drought. Over a large range of climates, changes in LAI have been studied at both the individual and ecosystem scales along gradients from higher to lower rainfall amounts or from moister to drier habitats in broad- or needle-leaved tree or shrub communities. The control of LAI and morphology is often the most powerful means that a mesophytic plant has to influence its fate when subjected to long-term water stress in the field. The main response of the shrubs to different precipitation regimes in the chaparral range is to change LAI, and not physiological parameters like stomatal regulation. This adjustment is largely species dependent and both leaf size and number are affected. For example, mature eucalyptus trees are tall and produce large leaves at moist sites, whereas at drier sites, trees are shorter and tend to produce smaller leaves.

Aging of forest stands also leads to temporal changes in LAI. LAI firstly increases to a maximum at ages ranging from 16 to 50 (depending on species and site index) and subsequently stabilises or declines slightly, up to 20% lower than peak value. Interestingly, aboveground net primary production follows a similar trend with aging.

Natural canopy disturbances like fire or windstorm also induce abrupt changes in LAI. The severity of disturbance determines the regeneration options (by growth of suppressed seedlings and saplings or from the seed bank), as well as the time for stand LAI recovery. Biotic attacks by herbivores and leaf-eating insects, combined with drought, severe early spring freeze–thaw events, and fungal pathogens cause substantial reduction in LAI and hence in productivity. The loss of foliage due to insect defoliation could result in spectacular changes to LAI. Such LAI reduction could be large enough to be detected by indirect measurements, in a quantified way as compared to visual assessment of the severity of defoliation. Monitoring of LAI decrease in this way may be of

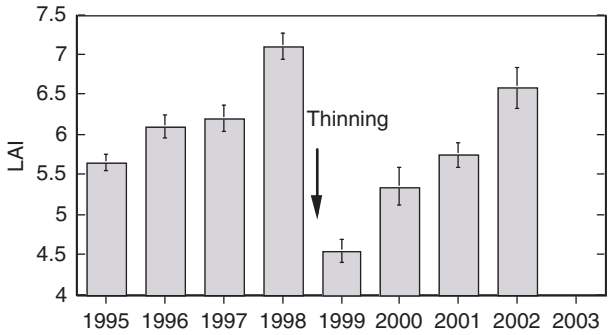


Fig. 2 Stand LAI recovery after thinning in a beech stand (RENECOFOR Network, Plot HET88). LAI was measured using a plant canopy analyzer (LAI 2000, Li-Cor, Nebraska, USA). Data from N. Bréda.

importance in mapping the spatial extension of the attack, may help to predict stand dieback and eventually plant mortality in the following years. Whatever the kind of natural disturbance or extreme event, the time needed to recover pre-event LAI could be used as an index of ecosystem resilience.

Management of LAI

Except in natural ecosystems, the canopy is periodically managed by farmers, foresters, grazing animals, and agriculturists. All management operations including cutting, grazing pasture, thinning, fertilization, liming, mowing, pruning, species sawing, and high herbage use, affect LAI. As stocking rate increases in grazed pasture, the total consumption per hectare goes up, while the net primary productivity decreases. The LAI is usually 2–3 with lower stocking densities and 1–2 at higher densities. In both crops and young forests, fertilizer effects on LAI are significant, and should increase LAI up to 3 units. An example of LAI management in forests is thinning, which reduces stand basal area, density, and LAI. In some cases, especially in even-aged mono-species stands like coniferous plantations, the LAI reduction is proportional to the basal area removed. Nevertheless, in most cases, the percentage of basal area removed is not proportional to those of LAI. Thinning improves water balance, radiation penetration within the canopy of the remaining trees, soil biology, and organic matter mineralization as a result of microclimatic changes. Canopy recovery occurs over several years depending on the intensity of the thinning, tree age, and site fertility (**Fig. 2**).

Managers control LAI to control productivity and water uptake, but in fact agricultural and forest managers should be interested in using estimates of LAI to gauge the vigor of cultures or plantations (crop or forest decline, pathogen attacks), to adjust management practices and thus produce optimum LAI (**Fig. 3**).

LAI as a Descriptor of Canopy Structure

Canopy structure means: (1) the whole of the vegetation community including species, number, leaf area, and leaf history; (2) its spatial organization, horizontal and vertical arrangement; and (3) its time progression (season, year, decade, and more). Then geometric complexities of different canopies are reduced to a simple quantification of the sum of all leaf layers as LAI.

The vertical distribution of LAI in mixed canopies reflects the functional abilities of species or leaves (shade, air humidity, and temperature tolerance). LAI controls both within- and below-canopy microclimate, determines and controls rainfall, snow and deposition interception, radiation extinction, wind velocity slackening, light quality and quantity below the canopy, and hence influences the living conditions of fungi, plants, insects, macro- and micro-fauna communities.

LAI as the Driving Force of Canopy Exchanges

LAI describes a fundamental property of the plant canopy in its interaction with the atmosphere, especially concerning radiation, energy, momentum, and gas exchange. Stand function includes: (1) the rate of biological energy flux through the ecosystem, that is, rates of production and respiration; (2) the rate of material and nutrient cycling, that is, the biogeochemical cycles; and (3) biological and ecological regulation, including prevention of soil erosion and regulation of water uptake, or radiation interception and conversion. LAI is the favored canopy variable because it is required for estimating many process rates, from canopy gas exchange to nutrient return in litterfall, including understorey microclimate control and competition for light, water, and mineral nutrients.

LAI acts as the canopy–atmosphere interface where water and carbon gas exchange occurs and is, therefore, a core parameter of biogeochemical cycles in ecosystems. Any change in canopy LAI, as a result of frost, storm, defoliation, grazing, drought, or management practice, is accompanied by modifications in stand productivity. Process-based ecosystem simulations require LAI as

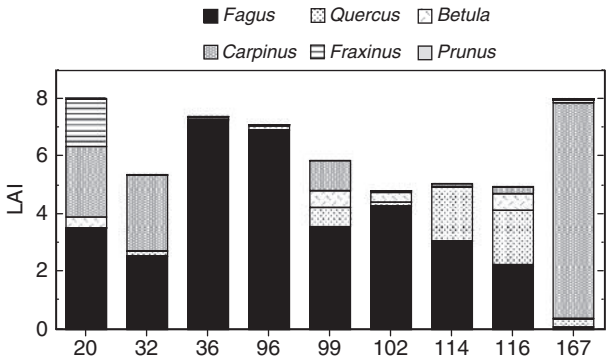


Fig. 3 Spatial variability of LAI in 60 ha of a managed beech forest (Hesse, France). Ground-based LAI measurements were distributed according to a systematic network (50 m x 50 m) using two cross-calibrated plant canopy analysers (LAI 2000, Li-Cor, Nebraska, USA). The scale ranges from 2 to 6, i.e., a similar range to that of the biomes presented in Bouriaud, O., Soudani, K., Bréda, N., 2003. Leaf area index from litter collection: Impact of specific leaf area variability within a beech stand. Canadian Journal of Remote Sensing 29, 371–380.

a key input parameter to produce quantitative analyses of productivity. When LAI of a community is low (<4), which is usually in arid environments or during the establishment of a crop, the transpiration rate (T/PET) is linearly related to LAI. Beyond this point, transpiration rate increases more slowly due to: (1) the saturation of canopy radiation interception and (2) soil water availability limitation (Fig. 4). Growth rates are also dependent on LAI, but as LAI increases, the growth rate reaches a maximum value. Thereafter, it may decline. The existence of an optimum LAI was first observed for herbaceous plants. At the slope inflexion, called critical LAI, an increase in LAI and its associated CO_2 uptake will not counterbalance the reduction of CO_2 uptake in the existing leaf area because of self-shading. The community might still continue to gain in biomass, but at a lower rate. Plant growth and life form strongly affects optimum LAI depending on leaf angle, clustering, and vertical distribution of leaves resulting in differences in the self-shading and greater or lesser depths of penetration of light into the canopy.

Assessment of LAI

Despite its functional importance, the measurement of LAI is not easy, due to its spatial (horizontal and vertical) and temporal heterogeneity. A plethora of ground-based optical, allometric, or litter collection methods and remote sensing approaches to estimate LAI are available.

Direct or Semidirect Methods

These involve a leaf area measurement. The main advantage of direct methods is that they are the only ones giving real LAI without any other plant organs. For that reason, they are considered as reference methods for indirect and remote sensing calibration. However, these direct methods are tedious, time consuming, and some of them are destructive.

Harvesting

The harvest method is one of the oldest methods, used for various vegetation types from crops to forests. A sample to be treated destructively has to be collected 4–5 times during the crop life cycle (i.e., sampling is done at 10–15 day intervals from seedling emergence) or relative to some of the developmental stages of the crop (emergence, flowering, physiological maturity). At each collection date, leaves are separated from the other parts and subsamples are selected for leaf area measurements. Then leaf area models have to be calibrated as: leaf area = (leaf length $\times$ leaf breadth) $\times k$, where k is a species specific coefficient depending on leaf shape and indentation ($k=0.5$ for triangle, 0.75 for grasses such as sorghum and maize, near 2/3 for many dicots). Leaves of the subsample are dried, weighed, and the dry leaf weight ratio is computed. Finally, LAI is computed as the dry weight of leaves $\times$ dry leaf weight ratio.

Litter collection

Using traps, in nets, or sampled on the ground is useful for LAI calculation of broadleaved plants. Species composition in the canopy is not distinguished when using canopy LAI. By sorting leaves from different species in mixed broadleaved forest, the contribution of each species from the community may be quantified (Fig. 5). Few data exist on leaf biomass of temperate deciduous forest communities which include undergrowth. If biomass is nearly negligible as compared to total stand biomass, the contribution of leaf area of undergrowth has been demonstrated to be a significant part (from 15% to 60%) of LAI of the whole community (tree or shrub layer + ground vegetation) within a narrow range of 7. Finally, the measured leaf properties (individual leaf area, specific leaf area, number of leaves, etc.) are of key importance for a comprehensive analysis of LAI changes between stands, species, or dates. Indeed, differences in LAI for a given stand during certain years could be the result of any change in number of plants, number of leaves, or individual leaf area.

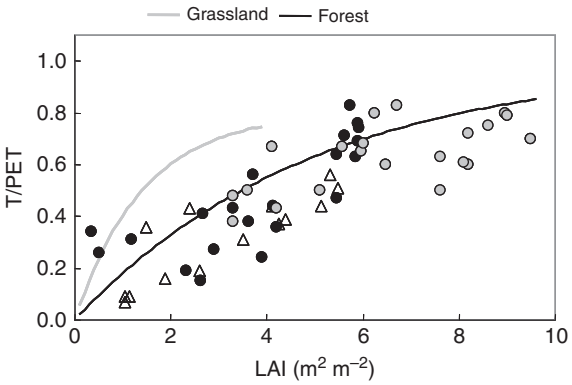


Fig. 4 Transpiration ratio (T/PET) as a function of LAI for forest and grasslands. For forest, symbols are for phenological periods: triangles: LAI increase during leaf expansion; black circles: LAI decrease during autumn; gray circles: maximum LAI for contrasting stands. Data from Granier, A., Bréda, N., Biron, P., Villette, S., 1999. A lumped water balance model to evaluate duration and intensity of drought constraints in forest stands. Ecological Modelling 116, 269–283 and Saugier, B., 1996. Evapotranspiration des prairies et des cultures. Comptes-rendus de l'Académie d'Agriculture de France 82, 133–153, for forests and grasslands, respectively.

Allometry

A relationship exists between sapwood or tree basal area and leaf area. This destructive approach is site-, species-, age-, and management dependent. Allometric relationships give the state at a given date.

Indirect Methods

These use the consequences of LAI on radiation interception or reflectance and are nondestructive. As any measure of radiation, these methods are sensitive to sky condition (direct vs. diffuse radiation or a clear and stable sky).

Ground-based approaches using optical instruments

At plot or stand level, the most common method of estimating LAI and its seasonal variation is from measurements of the fraction of light transmitted through the canopy to the ground. Nondestructive optical methods have been developed to estimate LAI periodically. The simplest approach, using Beer's law inversion with an extinction coefficient depending on the crop or tree properties, is useful and efficient for crops and broadleaved forests. However, for many evergreen species, the procedure requires some corrections to take the clumping of needles and branches into account. In any case, accurate equipment and methods for ground estimates of LAI are now available. Optical methods of estimating LAI use the inversion of gap fraction data. One fruitful approach involves measurement of the gap fraction, the proportion of unobscured sky in a set of sky directions as seen from beneath a plant canopy. Recent advances in the theory make it possible to calculate a useful array of canopy properties from gap fraction measurements, including light extinction coefficients, LAI, and leaf angle distribution. A variety of techniques can be employed to obtain gap fraction measurements, such as linear arrays of light sensors (SunSCAN, Delta-T Devices Ltd., Cambridge, UK and AccuPAR, Decagon Devices, Pullman, USA). Two other devices measure gap fraction for different zenith angles. The LAI-2000 (Li-Cor, Lincoln, Nebraska, USA) measures 5 zenith angles simultaneously, through a fisheye light sensor, while the DEMON instrument (CSIRO, Canberra, Australia) measures direct beam radiation from the Sun through a directional narrow angle of view (0.302 sr). Measurements with the DEMON instrument have to be repeated several times from early morning until noon to collect data over a range of zenith angles. Hemispherical photography and imaging hemispherical sensors (e.g., the CI-100 Canopy Analyzer, CID, Vancouver, USA) are also widely used but these frequently underestimate LAI.

Remote-sensed approaches

As opposed to ground-based methods, remote sensing deciphers the reflected, instead of the transmitted radiation. The plant communities are full of chlorophylls, a set of pigments which absorb part of the solar radiation for photosynthesis. As a result, reflectance of radiation in different spectral bands, especially the two widely used infrared and red ones, is changed proportionally to the amount of green vegetation. Since large-area maps of LAI are needed for global land-surface modeling, plenty of empirical relationships (i.e., statistical correlations) have been proposed between satellite or airborne image reflectance and ground-based (*in situ*) estimations of LAI. There are many vegetation indices developed from radiances in a wide range of channels corresponding to spectral bands. LAI estimation from satellite data requires ground data for validation and testing for bias. Satellite data must be corrected for atmospheric effects, thus requiring additional information on the state of the atmosphere (especially water vapor, aerosols, and ozone). Nondestructive (optical) measurements are the preferred approach for obtaining ground measurements. Classical values of LAI derivation from remote sensed vegetation indices (like Normalized Difference Vegetation Index, NDVI) range from 0 to 4–5, fitting an empirical exponential function with a plateau indicating a saturated signal for higher LAI. Such a relationship was first established for wheat and maize, at various states of growth. Unfortunately, LAI often reaches values above 5 and up to 15 in temperate mixed broadleaved forests or coniferous plantations. Then

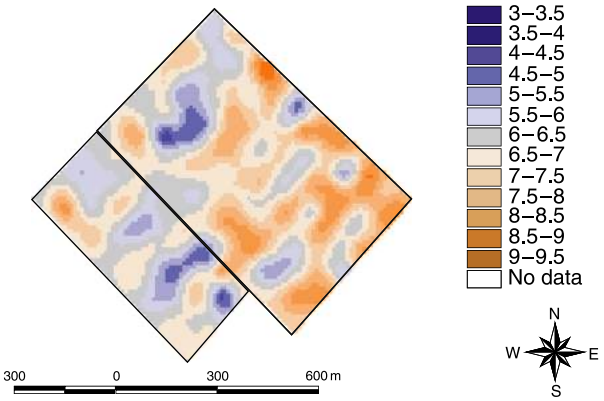


Fig. 5 Species contribution to total stand LAI in beech-dominated forest as estimated by litter collection and species sorting. Each plot number is related to spatial variability within the stand. For further details, see Bouriaud, O., Soudani, K., Bréda, N., 2003. Leaf area index from litter collection: Impact of specific leaf area variability within a beech stand. Canadian Journal of Remote Sensing 29, 371–380.

remote LAI derivation is not well adapted for such kinds of vegetation. New algorithms using the intrapixel variability of signals have been proposed but remain to be tested using ground-based data sets of forest LAI differing from the calibration data set.

Worldwide Maps of LAI and Controversies About Scaling in Global Modeling

Because LAI is a dimensionless quantity and an extensive surface parameter, it can be measured, analyzed, and modeled across a range of spatial scales, from individual tree crowns to whole regions or continents. Estimation of LAI across a landscape is needed for regional ecosystem analysis or modeling, but recent meta-analysis of worldwide LAI measurements highlights weaknesses in the ecological and geographical coverage of LAI measurements on a global basis. LAI is a canopy parameter in several models of growth or net primary production, but also in interactive models of land surface and atmospheric processes. Across wide landscapes, remote sensing is used to derive LAI maps to initiate regional or global modeling and automatic mapping of LAI at 8 km spatial resolution (NASA MODIS satellite data) offers global coverage of the biosphere. The reason for such a broad global ecological research interest in LAI is due to its emergent properties but it is still the subject of controversy in terms of scaling.

Further Reading

Asner, G.P., Scurlock, J.M.O., Hicke, J.A., 2003. Global synthesis of leaf area index observations: Implications for ecological and remote sensing studies. Global Ecology and Biogeography 12, 191–205.
Bouriaud, O., Soudani, K., Bréda, N., 2003. Leaf area index from litter collection: Impact of specific leaf area variability within a beech stand. Canadian Journal of Remote Sensing 29, 371–380.
Bréda, N.J.J., 2006. Ground-based measurements of leaf area index: A review of methods, instruments and current controversies. Journal of Experimental Botany 54, 2403–2417.
Cowling, S.A., Field, C.B., 2003. Environmental control of leaf area production: Implications for vegetation and land-surface modeling. Global Biogeochemical Cycles 17 (1), 1007. doi:10.1029/2002GB001915.
Gower, S.T., Norman, J.M., 1991. Rapid estimation of leaf area index in conifer and broad leaf plantations. Ecology 72, 1896–1900.
Granier, A., Bréda, N., Biron, P., Villette, S., 1999. A lumped water balance model to evaluate duration and intensity of drought constraints in forest stands. Ecological Modelling 116, 269–283.
Norman, J.M., Campbell, G.S., 1991. Canopy structure. In: Pearcy, R.W., Elheringer, J., Mooney, H.A., Rundel, P.W. (Eds.), Plant Physiological Ecology. London: Chapman and Hall, pp. 301–323.
Odum, E.P., 1962. Relationships between structure and function in ecosystems. Japanese Journal of Ecology 12, 108–118.
Pierce, L.L., Running, S.W., 1988. Rapid estimation of coniferous forest leaf area using a portable integrating radiometer. Ecology 67, 1762–1767.
Prentice, K.C., 1990. Bioclimatic distribution of vegetation for general circulation model studies. Journal of Geophysical Research 95, 11811–11830.
Saugier, B., 1996. Evapotranspiration des prairies et des cultures. Comptes-rendus de l'Académie d'Agriculture de France 82, 133–153.
Scurlock, J.M.O., Asner, G.P., Gower, S.T., 2001. Worldwide Historical Estimates and Bibliography of Leaf Area Index, 1932–2000. ORNL Technical Memorandum TM-2001/268. Oak Ridge, TN: Oak Ridge National Laboratory.
Waring, R.H., Running, S.W., 1998. Forest Ecosystems. Analysis at Multiple Scales, 2nd edn. San Diego, CA: Academic Press, 370pp.
Watson, D.J., 1947. Comparative physiological studies in the growth of field crops. Part I. Variation in net assimilation rate and leaf area between species and varieties, and within and between years. Annals of Botany 11, 41–76.

Relevant Website

http://daac.ornl.gov
DAAC-Distributed Active Archive Centre, for biogeochemical dynamics, Oak Ridge National Laboratory.

Metabolic Theories in Ecology: The Dynamic Energy Budget Theory and the Metabolic Theory of Ecology

Jaap van der Meer, NIOZ Royal Institute for Sea Research, Texel, The Netherlands; Utrecht University, Utrecht, The Netherlands; and VU University Amsterdam, Amsterdam, The Netherlands

Introduction

Animal metabolism is the processing of material that individual animals take up from the environment. Some fraction of this material is unusable and leaves the body as feces. This fraction is sometimes called the non-metabolic waste and not considered as part of metabolism, because it does not pass the gut wall and therefore not really enters the animals body. The remaining usable fraction of the food uptake is converted into other chemical forms and subsequently used for various life processes such as body maintenance, somatic growth and reproduction. Part of the usable material is thus stored in the body, maybe only temporarily as reproductive output, but a substantial amount of material also leaves the body as metabolic waste, in the form of gas such as carbon dioxide, liquid such as water or urine, or even as solid waste in the form of hairs, feathers, scales and skin. If the focus is on energy, a similar division can be made. The free chemical energy of the food is partly transformed into other forms of chemical energy, while a fraction is used for all kinds of internal work and finally dissipates as heat. Metabolic rate is usually defined as the heat production of an animal. So despite the term, metabolic rate is not sufficient to describe metabolism.

A complete description should also include food intake rate and defecation rate (or at least assimilation rate, which is here defined as the difference between these two rates), somatic growth rate, reproductive rate and the rate at which metabolic waste is produced. The description of these rates can be in terms of power measured as Watt, or in terms of material flows measured for example by C-moles per unit of time.

Ecology is the study of organisms in relation to their environment, and because organisms, as we have seen above, can be considered as processors of material, it is obvious that metabolism and how it is affected by the environment should play a pivotal role in ecology. Influential scholars as Lotka, Elton, Von Bertalanffy, Hutchinson, and Lindeman already long ago emphasized the transfer of material and energy from one component of the ecosystem to another. Despite the important work done by these early scientists, by the end of the previous century the focus in ecology had shifted away from energy and material flows and the subject of metabolism was receiving little attention. But two decades ago, papers by James Brown and colleagues put metabolism again strongly on the research agenda of ecology. Brown was, as many others before him, puzzled by Kleiber's observation that metabolic rate of adult animals of different species is proportional to their mass raised to the 3/4 power. Many scientists tried to find laws that could explain this regularity, but none had convinced the scientific community as a whole. The zoologist Brown and plant biologist Enquist joined forces with the physicist West, and they worked out the idea that whole-organism metabolic rate is limited by the internal delivery of resources to cells. Resources have to be distributed through branching networks, and the fractal-like design of these networks cause the supply rate and therefore the metabolic rate, to scale as a 3/4 power of body volume (West *et al.*, 1997). In later publications they worked out this idea and included other aspects of metabolism, e.g. somatic growth (West *et al.*, 2001; Hou *et al.*, 2008), reproduction (West *et al.*, 2001) and feeding (Hou *et al.*, 2011). The same research group was also interested in the effect of temperature on metabolic rate and claimed that it could be understood from basic cellular processes. Combining the two ideas led to the famous, but much questioned core equation of what has become known as the metabolic theory of ecology (MTE). The equation relates the metabolic rate B measured in Watt, to body mass m measured in kg and temperature T measured in K, and is given by

$$B = B_0 m^{3/4} e^{-E_a/(kT)} \tag{1}$$

where the parameter B_0 in W $kg^{-3/4}$ is a normalization parameter, E_a in eV is supposed to be the "average activation energy of metabolism" which should be about 0.6 eV, and k is the Boltzmann constant which equals 8.617×10^{-5} eV K^{-1}. Of course, very similar equations were already in use for a long time, but the claim for novelty was that a mechanistic explanation was offered.

More than a decade earlier the Dutch theoretical biologist Kooijman laid the basis of what has become known as the dynamic energy budget (DEB) theory in two papers, which were hardly recognized in those days (Kooijman, 1986a,b). Kooijman, like Brown, was interested in general patterns in biology, not only Kleiber's law but also the observation that the growth of most heterotrophic organisms, in size differing between a yeast cell and an elephant, follows Von Bertalanffy's equation quite accurately. Many more patterns had his interest, and a complete list will follow below. But Kooijman proceeded in a different way than Brown and colleagues. He realized that metabolic rate is related to many fundamental life processes, such as somatic maintenance, growth and reproduction, which each have their overhead costs, and feeding. In his view a description of these fundamental processes should form the core of a metabolic theory, and metabolic rate will follow automatically from the overall energy balance. He further made the assumption that most of these fundamental processes are simply dependent on either surface area or body volume. He also realized that many of the patterns that he was interested in could not be explained by using a one-compartment model animal. At least two compartments are needed, and they were labeled structural body and reserve, where the main

difference is that structure has maintenance costs, but reserve has not. DEB theory claims that it provides the simplest model that is able to explain all the patterns Kooijman had in mind. Simpler alternatives just do not exist. These early 1986 papers already explained a whole series of empirical scaling relationships, and many more predictions would follow during the development of the theory which led to several monographs (Kooijman, 2010).

Below I discuss and compare the two theories, DEB and MTE, in more detail. A theory in science should be based on assumptions that are in accordance with broadly accepted and more fundamental theories. If measurable quantities are involved in these assumptions, they should be in line with empirical observations. So I will list and discuss where needed the assumptions of the two theories. A theory should further be internally consistent. This will also be discussed. Predictions should be in line with empirical data and finally the theory should not be needlessly complicated. I will comment on the apparent complexity of the two theories. For reasons of convenience I restrict the overview to animals, although both MTE and DEB have something to say about other forms of life.

DEB, Development and Criticisms

The simplest DEB model, which is the standard DEB model, aims to predict a series of empirical patterns observed in living animals. A list of these patterns, which are all related to feeding, assimilation, respiration, growth, development, reproduction and changes in the chemical composition of animals, is provided in **Table 1**. Subsequently DEB theory provides a set of model assumptions that completely define the mathematical formulation of the standard DEB model. The underlying idea is that this set is as limited and as simple as possible, but the model that follows from it is capable of reproducing the observed patterns. A summary of the model assumptions is provided in **Table 2**. A graphical representation of the energy fluxes in the standard DEB model is provided in **Fig. 1**. The model was initially developed by Kooijman (1986a,b), but has been modified in details since then, which basically means that the list of assumptions has been refined, see for example Sousa *et al.* (2008). These assumptions automatically lead to the mathematical description of the model, in terms of a set of coupled differential equations, one for each of the three state variables. State variables and environmental factors are explained in **Table 3**, and the primary parameters of the model in **Table 4**. For completeness the differential equation will be given. For reserve density it is

$$\frac{\mathrm{d}[E]}{\mathrm{d}t} = V^{-1/3}(\{\dot{p}_{Am}\}f - \dot{v}[E]) \tag{2}$$

where f refers to the dimensionless scaled functional response, which is a function of food density X, and can vary between 0 and 1. The scaled functional response relation is basically Holling type II, and is given by $f = X/(X_K + X)$, where X_K is the half-saturation coefficient. This coefficient is directly related to various primary parameters of the DEB model (see **Table 4**) and to μ_X, which is the chemical potential of the food. It is given by $X_K = \{\dot{p}_{Am}\}/(\mu_X \kappa_X \{\dot{F}_m\})$. The differential equation for structural volume is

$$\frac{\mathrm{d}V}{\mathrm{d}t} = \frac{(\kappa\dot{v}[E] - \{\dot{p}_T\})\ V^{2/3} - [\dot{p}_M]V}{\kappa[E] + [E_G]} \tag{3}$$

which reduces to the well-known Bertalanffy equation when the reserve density is in equilibrium. This happens at constant food density. Finally, the equation for maturity is

$$\frac{\mathrm{d}E_H}{\mathrm{d}t} = (1-\kappa)\dot{p}_C - \dot{k}_J E_H \tag{4}$$

for $E_H < E_H^p$. Else, that is when the animals have become mature and $E_H = E_H^p$, maturity does not change anymore and $\mathrm{d}E_H/\mathrm{d}t = 0$.

Table 1 Empirical patterns on feeding, respiration, growth and reproduction that DEB theory aims to explain

Process	*Pattern*
Feeding	During starvation, organisms are able to survive for some time, to reproduce and to grow
	At abundant food feeding rate is at some maximum
Growth	Growth of animals at abundant food is well described by the von Bertalanffy growth curve
	Many animals do not stop growing after reproduction has started, i.e. they exhibit indeterminate growth
	Fetuses increase in weight proportional to cubed time
	The inverse of the von Bertalanffy growth rate of the same species at different food availabilities corrected for a common body temperature decreases linearly with ultimate length
	The logarithm of the von Bertalanffy growth rate of different species corrected for a common body temperature decreases almost linearly with the logarithm of the species maximum size
Respiration	Freshly laid eggs do not use oxygen in significant amounts
	The use of oxygen increases with decreasing mass in embryos, but increases with mass in juveniles and adults
	The use of oxygen scales with body mass raised to a power close to 0.75
Stoichiometry	Well-fed animals have a different body chemical composition than poorly-fed organisms
	Animals growing with constant food density converge to a constant chemical composition
Reproduction	Reproduction increases with size intra-specifically, but decreases with size inter-specifically

Source: Sousa, T., Domingos, T. and Kooijman, S. A. L. M. (2008). From empirical patterns to theory: A formal metabolic theory of life. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**(1502), 2453–2464.

Table 2 The basic assumptions of the standard DEB model

1. An organism is characterized by a structural body, reserves, and maturity. The chemical composition of both structural body and reserves is constant, which is called the assumption of strong homeostasis. Maturity indicates the level of development and represents information, but has no energy or mass capacity
2. Each organism starts its life as an embryo (which does not feed and does not reproduce). When the embryo has reached a certain level of maturity, it changes into a juvenile (which feeds, but does not reproduce). Similarly, a juvenile changes into an adult (which feeds and reproduces) when it has reached the maximum maturity value
3. Ingestion depends upon food density by Holling's disc equation (but recall that embryos do not feed). Both the searching rate and the maximum ingestion rate, the latter similar to the inverse of the handling time, are proportional to the surface area of the organism
4. A fixed fraction of the ingested food is assimilated and enters the reserves
5. Reserve density, which is the amount of reserves per amount of structural body, reaches an equilibrium under constant food conditions. This assumption is called the weak homeostasis assumption. The use of reserves only depends upon the amount of reserves itself and on body volume, which in combination with the weak homeostasis assumption implies that the mobilization of the reserves occurs at a rate proportional to the reserve density
6. A fixed fraction κ of the mobilized reserves goes to somatic maintenance and growth of the structural body, with a priority for maintenance. The rest goes to maturity maintenance and either to maturity (for embryos and juveniles) or to reproduction (for adults)
7. Somatic maintenance rate is basically proportional to structural volume, but in specific cases part of the maintenance costs is proportional to surface area of the organism (e.g. heating rate in endotherms). Maturity maintenance costs are proportional to maturity
8. Energetic costs of growth are proportional to body volume and energetic costs per egg are such that the newborn juvenile has the same energy density as its mother

Source: Kooijman, S. A. L. M. (2010).*Dynamic energy budget theory for metabolic organisation*, 3rd edn. Cambridge: Cambridge University Press.

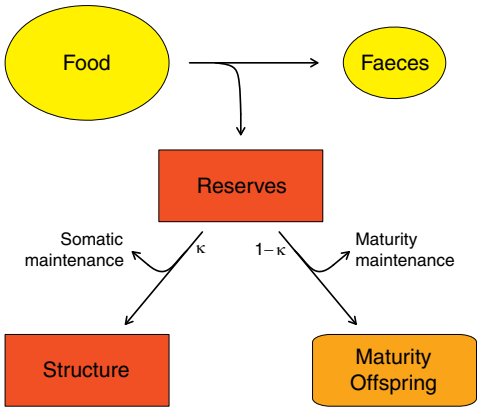


Fig. 1 Conceptual scheme of the main energy fluxes in the standard DEB model.

Table 3 State variables of the standard DEB model and environmental variables

Symbol	*Dimension*	*Interpretation*
V	L^3	Structural body volume
$[E]$	eL^{-3}	Reserve density
E_H	e	Maturity
T	T	Temperature
X	$\#l^{-2}$ or $\#l^{-3}$	Food density in the environment

L stands for the dimension length of the structural body, e for energy, # for mass measured in terms of C-moles, and l for the dimension length of the environment.

For adults the energy flow is channeled to reproduction and the rate of cumulative reproduction is given by

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \frac{\kappa_R}{E_0}\left((1-\kappa)\dot{p}_C - \dot{k}_J E_H^p\right) \tag{5}$$

where κ_R is the reproduction efficiency and E_0 the energy content of an egg. The term $\dot{p}_C$ is the rate at which the reserves are mobilized. This rate is given by

$$\dot{p}_C = \frac{[E]}{\kappa[E] + [E_G]}\left((\dot{v}[E_G] + \{\dot{p}_T\})\ V^{2/3} + [\dot{p}_M]V\right) \tag{6}$$

Table 4 Primary parameters of the standard DEB model

Symbol	*Dimension*	*Interpretation*	*Process*
$\{\dot{p}_{Am}\}$	$eL^{-2}t^{-1}$	Surface-area-specific maximum assimilation rate	Assimilation
$\left\{\dot{F}_m\right\}$	$l^*L^{-2}t^{-1}$	Surface-area-specific searching rate	Feeding
κ_X	–	Digestion efficiency	Digestion
$\dot{v}$	Lt^{-1}	Energy conductance	Mobilization
κ	–	Fraction of mobilization rate spent on maintenance plus growth	Allocation
$[\dot{p}_M]$	$eL^{-3}t^{-1}$	Volume-specific maintenance rate	Turnover/activity
$\{\dot{p}_T\}$	$eL^{-2}t^{-1}$	Surface-area-specific maintenance rate	Heating/osmosis
$[E_G]$	eL^{-3}	Volume-specific costs of growth	Growth
$\dot{k}_J$	–	Specific maturity maintenance	Regulation/defense
κ_R	–	Reproduction efficiency	Egg formation
E_H^b	e	Maturity at birth	Life history
E_H^p	e	Maturity at puberty	Life history

Source: The "wild card" * stands for two when food density is expressed per area or three when expressed per volume, t stands for time. See further **Table 3**.

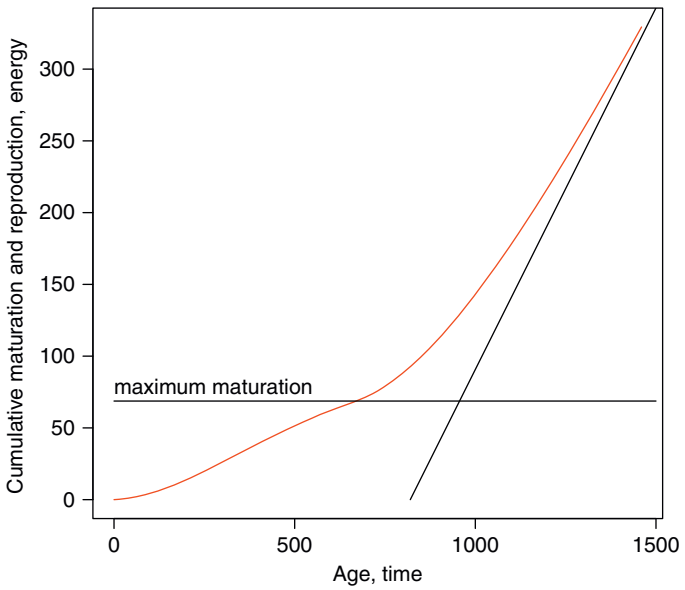


Fig. 2 Cumulative maturation and reproduction as a function of age. Scaled functional response f is 1. When the total investment in maturation has reached its maximum, reproduction starts and the cumulative reproduction steadily approaches a diagonal asymptote.

Combining Eqs. **6** and **5** will give an explicit expression for the reproduction rate as a function of size and reserves (**Fig. 2**).

DEB theory has been mainly criticized for its perceived complexity (Brown *et al.*, 2004; Zuo *et al.*, 2009; Marquet *et al.*, 2014). Brown *et al.* (2004) state that DEB models are complex, using many variables and functions. They argue that how much complexity in a model is desirable is partly a matter of the purpose for which the model is used, and that there is room for a complementary and even more general approach than DEB, something that MTE offers. Zuo *et al.* (2009) consider DEB as a very detailed model and they (erroneously) state that the model requires the measurement of 17 variables and 18 parameters. Marquet *et al.* (2014) even call DEB theory an inefficient theory as it contains too many species specific parameters, which hinders the generation of general predictions.

MTE, Development and Criticisms

The development of the MTE started with a rather technical paper by West *et al.* (1997), using principles of fluid dynamics to show that optimal transport networks yield a supply rate to the cells that scales with the 3/4 power of body mass. The network design, which could for example represent the cardiovascular network of a mammal, was based on certain specific assumptions, such as that the capillaries have to reach all cells. Optimal was defined as having minimal energy loss through dissipation and wave reflections. The approach was applauded by many, but also heavily criticized, both by what I for the sake of simplicity will call

biologists and physicists. The biologists questioned the relevance of the assumed network for animal life. A first problem is that closed branching networks are rarely found in animals. Apart for vertebrates and a few invertebrate groups, such as cephalopods, they do not exist in the animal kingdom. A second problem is that cardiovascular networks are probably designed for peak exercise during short periods, and it seems rather unlikely that the much lower resting metabolic rate that Kleiber was talking about, or even the average or field metabolic rate, has the same proportion to the maximum rate for all animal species. The biologists further thought that the predictions on the type of network that resulted from the optimization procedure were far away from reality. The predicted area preservation of the network implies that the flow rate in the capillaries is as large as that in the aorta. If that would be the case for you and me, we would almost immediately die if we would cut ourselves in one of our fingers. But most criticisms pointed to the fact that there is actually no need at all for predicting a 3/4 relationship, because such universal relation simply does not exist in nature. The physicists merely disagreed on the reliability of some more detailed assumptions on the network itself, e.g. it was originally assumed that the network is infinitely large, and they questioned how critical such assumption is for the outcome. They came to the conclusion that more realistic assumptions on the size of the network would have led to a non-allometric relationship that can be approximated by an allometric one with a scaling coefficient of 0.81. Others had major problems with the optimization procedure that had been used. So Price *et al.* (2012) had to point out in a recent review that "the community at large has not reached a consensus as to whether the theory is or is not logically consistent." This community at large is in fact rather small, as most biologists are not well equipped to follow the detailed arguments of the physicists in this part of the debate.

The development of MTE went on with the paper by West *et al.* (2001), later followed by Hou *et al.* (2008). West and co-workers presented the so-called ontogenetic growth model (OGM). The OGM assumed that organism growth rate is proportional to the difference between the supply rate at which resources are delivered to the cells, which was set equivalent to the average resting metabolic rate B, and the maintenance rate B_{maint}, defined as the power needed to sustain the organism in all its activities. The fractal-like design of the network through which the resources are supplied to the cells causes the supply rate to scale as the 3/4 power of body mass, as earlier proposed by West *et al.* (1997). The maintenance rate was thought to be proportional to body mass $B_{maint} = B_m m$. Hence, the model can be expressed as

$$\frac{dm}{dt} = \frac{B_0 m^{3/4} - B_m m}{E_m} \tag{7}$$

where m is total body mass, t is time, and E_m is the energy required to create a unit of mass of new tissue. Though the OGM was highly influential and widely cited, it was also severely criticized (Makarieva *et al.*, 2004; Van der Meer, 2006). Criticisms pointed among other things to lack of consistency and incompleteness. The criticism that the OGM lacked consistency referred to the ambiguous definition of metabolic rate. The term was both used for the supply rate of energy and for the maintenance rate. At the same time the difference between these two rates was supposed to be used for the build-up of new body tissue. The proportionality parameter E_m which converts energy into biomass, and which thus converts the difference between the supply and maintenance rate into the rate of biosynthesis, was derived from the empirical combustion energy per unit mass of mammalian tissue, for which a value of 7 kJ/g was used. The use of such parameter is in accordance to the idea that the difference between supply and maintenance is entirely stored into new tissue, but the fact that thus a part of the supply rate is not dissipated causes a problem, as you cannot, almost literally, have your cake and eat it. The problem could only be resolved if not the supply rate, but the maintenance rate would be equated to the resting metabolic rate (Van der Meer, 2006). Holding on to the notion that the resting metabolic rate equals the supply rate would violate the energy conservation law (Makarieva *et al.*, 2004). The logic consequence of equating resting metabolic rate to maintenance rate is that the intraspecific scaling power of resting metabolic rate equals 1 (Van der Meer, 2006). Ironically, this disputes the strong claim that the MTE explains Kleiber's law of a scaling power of 3/4, at least as far as intraspecific comparisons are concerned. The supposed incompleteness of the OGM referred to the fact that the metabolic rate of active animals can be much higher than the resting metabolic rate and to the ignorance of all overhead costs of growth (Makarieva *et al.*, 2004; Van der Meer, 2006).

A new model by Hou *et al.* (2008), called the extended ontogenetic growth model (EOGM) and mainly meant to be applied to birds and mammals, aimed to repair these flaws. The EOGM distinguishes between total metabolic rate B_{tot} and resting metabolic rate B_{rest}. The difference between these rates is used for activity. Total metabolic rate is assumed to be proportional to resting metabolic rate, $B_{tot} = fB_{rest}$. Following West *et al.* (1997) total metabolic rate and thus also resting metabolic rate scale as the 3/4 power of body mass. The new model also takes the overhead costs of growth into account and distinguishes between the energy content stored in newly synthesized biomass and the energy expended in synthesizing this biomass from the constituent materials. The difference between resting metabolic rate and maintenance rate is now entirely used for the overhead costs of growth. This implies that the resting metabolic rate is completely dissipated to heat. The new growth model still looks like the old OGM, but the parameter E_m is now interpreted as the overhead costs of growth. So the problems of violating the energy conservation law or contrasting Kleiber's law were solved. One new problem arose and that is, how are the building materials that need to be stored in newly synthesized biomass transported to the cells? Hou *et al.* added the rate S at which the energy content of new biomass accumulates, on top of the total metabolic rate B_{tot} in order to arrive at the assimilation rate of food $A = B_{tot} + S$. They further supposed a fixed ratio between the supply of building materials and the overhead costs of growth (both in terms of power): $S = \gamma(B_{rest} - B_{maint})$. This implies that the assimilation rate equals $A = (f + \gamma)B_0 m^{3/4} - \gamma B_m m$. But Hou *et al.* did not tell how these assimilated products are delivered to the cells and what kind of network is required to do so. The core of MTE is that within organisms the supply rate to the cells is constrained by some specific fractal-like design of branching networks. Consequently, the supply rate scales as the 3/4 power of body mass. Now two options with respect to the design of the network are possible. The

Table 5 Empirical patterns on feeding, respiration, growth and reproduction that MTE theory aims to explain

Process	*Pattern*
Growth	Growth of animals is well described by a von Bertalanffy type growth curve
Respiration	The use of oxygen scales with body mass raised to a power close to 0.75

Source: West, G. B., Brown, J. H. and Enquist, B. J. (1997). A general model for the origin of allometric scaling laws in biology. *Science* **276**, 122–126; West, G. B., Brown, J. H. and Enquist, B. J. (2001). A general model for ontogenetic growth. *Nature* **413**, 628–631; Hou, C.; Zuo, W.; Moses, M. E., Woodruff, W. H., Brown, J. H. and West, G. B. (2008). Energy uptake and allocation during ontogeny. *Science* **322**(5902), 736–739.

Table 6 The basic assumptions of the metabolic theory of ecology, and more specifically of the extended ontogenetic growth model (EOGM)

1. An organism is characterized by its body mass
2. Nine detailed assumptions on the design of the distribution network (not spelled out here, but see assumptions A1–A9 in Price *et al.* (2012)) result in the prediction that the delivery of oxygen to the cells is proportional to body mass raised to the power 3/4
3. The oxygen delivered to the cells is used for maintenance, overhead costs of growth, and locomotion and other activities
4. Oxygen delivery needed for maintenance is proportional to body mass
5. Oxygen delivery needed for locomotion and other activities is proportional to the oxygen needed for maintenance and overhead costs of growth
6. The material needed as building blocks for new body tissue is proportional to overhead costs of growth. The delivery of this material to the cells is not restricted by the capacity of the distribution network

Source: Interpreted from West *et al.* (1997) and Hou *et al.* (2008). See text for further explanation.

network is either designed such that its capacity is just sufficient to transport the matter that is dissipated to heat. This means that the supply rate through the network equals the total metabolic rate $fB_0m^{3/4}$. This option thus implies that the EOGM requires additional and unknown ways of transport to deliver the building materials for biosynthesis to the cells. The other option is that the network is able to transport the building materials as well. It can easily be shown that the network should be designed such that the capacity equals $(f + \gamma)B_0m^{3/4}$. Smaller 3/4 scaling networks cannot transport the assimilated matter when the animals are small. In this case the supply rate is no longer constrained by the network when the animal gets larger. For larger animals the network has a huge overcapacity and it remains unclear by what mechanism the total metabolic rate is constrained. In a personal communication the authors stated that the first option is basically the right one, but that "it is primarily the uptake and delivery of oxygen that has selected for the fractal-like design of the vascular system, the matching design of the respiratory system and the 3/4 power scaling of metabolic rate." The processing of food is not constrained by the network as food "can be stored in the body and utilized for metabolism on widely varying time scales." This is an important and essential addition to the original contribution, as it for example suggests that reproduction, which requires very little overhead costs, will hardly be limited by the supply network. Apparently, the transport of food is not hampering the transport of oxygen.

To simplify a comparison with the DEB approach discussed earlier, I also list the type of patterns MTE aims to explain on feeding, respiration, growth and reproduction (**Table 5**). West *et al.* (2001) made a small remark on reproduction, but this was further ignored in the EOGM. Feeding is more recently discussed by Hou *et al.* (2011), but this paper contains some serious problems, about which a discussion is beyond the scope of the present overview. Hence, a single pattern for respiration and one for growth remain. A list of the underlying assumptions of MTE (**Table 6**) also illustrates the more limited scope compared to that of DEB (**Table 2**).

Predicting Scaling Relationships

For aerobic animals metabolic rate, which is the rate at which chemical energy is transformed into heat, is usually indirectly measured by the rate of oxygen consumption. DEB theory is clear about the intraspecific scaling relationship of metabolic rate. Maintenance costs are proportional to structural body volume, but there are other processes contributing to the oxygen consumption. For example, only part of the energy that is allocated to growth is fixed in new body tissue, the rest is dissipated as overhead costs. These overhead costs are proportional to the difference between a surface area-related term and a volume-related term. For endotherms, the heating costs, which scale with surface area, also contribute to the total oxygen use. Taking all these other processes into account, DEB theory predicts that the total metabolic rate will scale (if individuals of different size within the same species are compared) with body volume with a power somewhere between 2/3 and 1. The EOGM assumes that the supply rate of oxygen to the cells sets the rate at which chemical energy is transformed into heat. As the supply rate scales with 3/4 power of body mass, so does the metabolic rate. But, only when animals of different size of the same species are compared.

But what about Kleiber's law of the 3/4 power scaling of metabolic rate, which was not based on an intraspecific but on an interspecific comparison? Assuming that such a comparison concerns full-grown adult organisms, implies that growth has ceased. According to MTE all supplied energy is then used for maintenance. This can be seen from setting the growth Eq. **7** equal to zero. It follows that the ultimate size M_∞ equals $(B_0/B_m)^4$. So it seems logical either to assume that the supply parameter B_0 scales with the double square root of ultimate body mass, that is $B_0 = m_\infty^{1/4}$, or to assume that the maintenance parameter scales inversely with

m_∞, i.e. $B_m = m_\infty{}^{-1/4}$. MTE makes the second assumption, that is cellular maintenance costs scale interspecifically with a power of $-1/4$. The maintenance costs of a lizard are thus much higher than those of a baby crocodile of the same size. It is this assumption, which comes without any mechanistic biological reasoning, that in fact reveals Kleiber's law. So despite several claims, MTE's main idea of transport of resources to the cells through a fractal-like hierarchy of branching vessels does not suffice to predict Kleiber's law.

DEB theory predicts (for convenience restricted here to well-fed ectotherms) a maximum volume that equals $(\kappa\{\dot{p}_{Am}\}/[\dot{p}_M])^3$. Maximum volumetric length (the cubic root of the structural body volume) is thus proportional to the ratio between the maximum area-specific assimilation parameter $\{\dot{p}_{Am}\}$ and the volume specific maintenance parameter $[\dot{p}_M]$. Contrary to MTE, DEB assumes that animals that reach a large ultimate size do not differ in their volume specific maintenance rate from animals that remain small. Animals grow larger because they have higher assimilation capacities. This implies that metabolic rate at ultimate size is proportional to ultimate structural body volume, which seems to contradict Kleiber's law. However, DEB also predicts that species with a large ultimate structural volume have a higher maximum energy density than species that remain small. And because the body mass that is measured in practice is the sum of the structural body mass and the reserve mass, metabolic rate is not proportional to measured body mass. In fact, DEB predicts a non-allometric relationship between metabolic rate and measured body mass, which can be approximated by an allometric relationship with a scaling power very close to 3/4. For endotherms, heating costs, which are related to surface area, have to be added when animals are outside the thermoneutral zone and the approximate scaling power will then be somewhat lower.

DEB distinguishes between parameters that depend upon the local biochemical environment and those that depend upon physical design. The latter parameters are related to the ultimate size of the organism, the first group of parameters do not depend upon size. The maximum area-specific assimilation parameter is one of the few parameters that belong to the second group. Others are the E_H^b and E_H^p, the maturity at birth and at puberty. A variety of body size scaling relationships follow directly from this parameter classification. One example is the Bertalanffy growth rate, whose relation with ultimate body size can be approximated by a scaling relation with a co-efficient of -1. One of the first DEB papers already came up with predicted scaling relations for about 20 physiological variables, as diverse as maximum starvation time or minimum incubation time (Kooijman, 1986b).

Complexity

The standard DEB model contains three state variables (**Table 3**) and 12 primary parameters (**Table 4**). The extended ontogenetic growth model (EOGM) only considers one state variable and has four parameters. So at first sight DEB seems more complex than MTE. One should, however, realize that DEB is able to make predictions about many more processes than MTE, e.g. embryonic growth in eggs and wombs, hatching, maturation, reproduction, food depression, interspecific scaling etc. MTE has nothing to say about these issues. Of course, this difference comes at a cost. But DEB theory allows a reduction of the standard DEB model under certain restrictions. If food conditions are constant, reserve density is also constant after hatching, and growth can simply be described by a Von Bertalanffy growth equation with only three parameters and one state variable. Note that although the equation is the same, it is based on a different biological rationale than Von Bertalanffy had in mind. He did not apply the energy conservation law to the overall organism, but defined growth as the difference between anabolism (synthesis) and catabolism (breakdown).

Beyond the Organismic Level: Populations and Ecosystems

The standard DEB model provides a detailed description of the flows of energy in and out of the individual organism, including reproduction. It is therefore usable as a building block in either physiologically-structured populations models or agent-based models. Several applications are available, e.g. Martin *et al.* (2013). Ecosystem models have also been (partly) based on DEB theory (Saraiva *et al.*, 2017). MTE is less suited for these goals, as it neither incorporates feeding (but see Hou *et al.*, 2011) nor reproduction. Predictions of MTE at for example the population level are restricted to applying the so-called energy-equivalence rule, which says that every population of animals receives the same amount of energy as food. Linking MTE and this rule predict that population abundance scales with body mass with a scaling coefficient of minus $-1/4$. One should however notice that this prediction entirely relies on the 3/4-scaling relationship of metabolic rate, which is not just a result of MTE but also of DEB. In fact, any other explanation of Kleiber's law could have been used. Marquet *et al.* (2014) define an efficient theory as one that generates many predictions. It is a bit ironic that MTE sensu stricto, that is the idea of a fractal-like distribution network, has not generated any further hypothesis above the level of the individual. All predictions could have been made as a consequence of Kleiber's law. And several of the few detailed predictions at the level of the individual, such as that of an area-preserving network, are clearly flawed. One therefore could argue that Marquet *et al.* should have categorized the MTE theory as inefficient.

Temperature Effects

Temperature is, apart from food availability, another important environmental variable that affects metabolism. Both DEB and MTE use the Van't Hoff–Arrhenius equation to describe the dependency of physiological rates on temperature. This equation has

its origin in statistical thermodynamics, where the behavior of a system containing a very large number of a single type of molecules is predicted from statistical considerations of the behavior of individual molecules. In its basic form the Van't Hoff–Arrhenius equation looks like

$$\dot{k}(T) = \dot{k}_\infty e^{E_a/(kT)} \tag{8}$$

where $\dot{k}(T)$ is a reaction rate that depends upon the absolute temperature T (in Kelvin), $\dot{k}_\infty$ is a (theoretical) maximum reaction rate, which is the reaction rate when all molecules would react. The term $\exp(E_a/(kT))$ is the Boltzmann factor, which gives the fraction of the molecules that obtain the critical activation energy E_a (in joules per molecule) to react. This fraction increases with increasing temperature. The constant k (not to be confused with the reaction rate $\dot{k}$) is the Boltzmann constant and equals 1.38×10^{23} J/K. The Van't Hoff–Arrhenius equation can also be re-written in the form

$$\dot{k}(T) = \dot{k}_1 \qquad \exp\left(\frac{T_A}{T_1} - \frac{T_A}{T}\right) \tag{9}$$

where $\dot{k}_1$ is the reaction rate at a reference temperature T_1, and T_A the so-called Arrhenius temperature (which equals E_a/k). The Van't Hoff–Arrhenius equation is approximate for bi-molecular reactions in the gas phase, and Kooijman (2010) emphasizes the enormous step from a single reaction between two types of particles in the gas phase to physiological rates where many compounds are involved and gas kinetics do not apply. He therefore regards the application of the Van't Hoff–Arrhenius relation to physiological rates as an approximation only, for which the parameters have to be determined empirically. For this reason, Kooijman prefers the use of an Arrhenius temperature instead of the use of an activation energy, which would give a false impression of mechanistic understanding. Initially, MTE expressed the more candid view that the Van't Hoff–Arrhenius equation must be applied because it links whole-organism metabolism directly to the kinetics of the underlying biochemical reactions (Gillooly *et al.*, 2001). They furthermore stated that for all aerobic species a single parameter value for E_a can be used, because such species have the same biochemistry. This point of view has been criticized by, among many others, e.g. Glazier (2015). Similar to Kooijman, they stress that the Van't Hoff–Arrhenius equation is just a statistical generalization, and they too conclude that at present we still lack a clear understanding of the relationship between temperature and metabolism at the organismal scale. Glazier also pointed to the huge variation in estimates of the Arrhenius parameter. Indeed the claim of mechanistic understanding has never been substantiated.

Within MTE E_a and k are expressed in eV instead of joules and for E_a usually a value of 0.6 or 0.65 eV is used. Dividing by the Boltzmann coefficient gives an Arrhenius temperature of about 7000 or 7500 K, which is not far from the value of 8000 K used as the default value in DEB theory (Kooijman, 2010). But, as mentioned earlier, in DEB theory the Arrhenius parameter is considered as an estimable parameter, and estimates vary in practice mostly between 6000 and 10,000 K, which is, expressed in eV, equivalent to the range 0.52–0.86. Glazier (2015) mentions even an observed range of 0.2–1.2 eV for various metabolic processes.

Summary

Metabolic ecology describes the uptake of energy and matter by individual organisms and the subsequent allocation to various life processes such as body maintenance, somatic growth and reproduction. The ecological role that organisms play is strongly related to their metabolism and understanding the determinants of metabolism is essential to understanding this ecological role. Two alternative and influential theories on metabolic ecology, the dynamic energy budget theory (DEB) and the metabolic theory of ecology (MTE) are described in detail, with a focus on the underlying assumptions and predictive power. Core of the DEB theory concerns the assimilation of food, limited by the surface area of the organism, the transfer into reserves and the allocation of mobilized reserves, also limited by the reserve-structural body interface, to the various life processes. Metabolic rate, i.e. heat production, follows consequentially from making the energy balance. MTE, on the other hand, considers the supply rate of oxygen, limited by a branching fractal network, as setting the pace of life. Fuel will automatically be available. The theories are further compared in terms of complexity and internal consistency. At first sight, DEB seems to be the more complex theory, but if the scope and predictive power are also considered, it appears that DEB cannot be considered as more complex than MTE. The limited scope of MTE, exemplified for example by the ignorance of reproduction and the possibility of variable food intake rate, the lack of internal consistency and the low predictive power, makes it the least efficient theory of the two.

Further Reading

A contrasting view is expressed by Glazier (2015), who believes that in both DEB theory and MTE the importance of informational control, as mediated by various genetic, cellular, and neuro-endocrine regulatory systems, is underestimated (but see Kooijman, 2010, pp. 15–16).

Recently, several papers suggested ways to test the DEB and MTE theories against each other (Kearney and White, 2012; Maino *et al.*, 2014). Kearney and White (2012) came up with a list of 10 possible tests for which DEB predictions and MTE predictions differed. One example is a test on limb regeneration, in for example lizards. DEB predicts a regeneration rate up to threefold faster than ontogenetic growth rate, whereas these rates should be equal according to MTE. Maino *et al.* (2014) point to the relationship between inter- and intraspecific scaling of biological rates (see also discussion above) that should point to tests that could lead to a further refinement of the current metabolic theories.

See also: General Ecology: Ecophysiology

References

Brown, J.H., Gillooly, J.F., Allen, A.P., Savage, V.M., West, G.B., 2004. Response to forum commentary on 'toward a metabolic theory of ecology'. Ecology 85, 1818–1821.
Gillooly, J.F., Brown, J.H., West, G.B., Savage, V.M., Charnov, E.L., 2001. Effects of size and temperature on metabolic rate. Science 293 (5538), 2248–2251.
Glazier, D.S., 2015. Is metabolic rate a universal 'pacemaker' for biological processes? Biological Reviews 90 (2), 377–407.
Hou, C., Zuo, W., Moses, M.E., Woodruff, W.H., Brown, J.H., West, G.B., 2008. Energy uptake and allocation during ontogeny. Science 322 (5902), 736–739.
Hou, C., Bolt, K.M., Bergman, A., 2011. A general model for ontogenetic growth under food restriction. Proceedings of the Royal Society B: Biological Sciences 278 (1720), 2881–2890.
Kearney, M.R., White, C.R., 2012. Testing metabolic theories. American Naturalist 180 (5), 546–565.
Kooijman, S.A.L.M., 1986a. Energy budgets can explain body size relations. Journal of Theoretical Biology 121, 269–282.
Kooijman, S.A.L.M., 1986b. Population dynamics on basis of budgets. In: Metz, J.A.J., Diekmann, O. (Eds.), The dynamics of physiologically structured populations. Berlin: Springer-Verlag, pp. 266–297.
Kooijman, S.A.L.M., 2010. Dynamic energy budget theory for metabolic organisation, 3rd edn. Cambridge: Cambridge University Press.
Maino, J.L., Kearney, M.R., Nisbet, R.M., Kooijman, S.A.L.M., 2014. Reconciling theories for metabolic scaling. Journal of Animal Ecology 83 (1), 20–29.
Makarieva, A.M., Gorshkov, V.G., Li, B., 2004. Ontogenetic growth: Models and theory. Ecological Modelling 176, 15–26.
Marquet, P.A., Allen, A.P., Brown, J.H., Dunne, J.A., Enquist, B.J., Gillooly, J.F., Gowaty, P.A., Green, J.L., Harte, J., Hubbell, S.P., O'Dwyer, J., Okie, J.G., Ostling, A., Ritchie, M., Storch, D., West, G.B., 2014. On theory in ecology. Bioscience 64 (8), 701–710.
Martin, B.T., Jager, T., Nisbet, R.M., Preuss, T.G., Grimm, V., 2013. Predicting population dynamics from the properties of individuals: A cross-level test of dynamic energy budget theory. American Naturalist 181 (4), 506–519.
Price, C.A., Weitz, J.S., Savage, V.M., Stegen, J., Clarke, A., Coomes, D.A., Dodds, P.S., Etienne, R.S., Kerkhoff, A.J., McCulloh, K., Niklas, K.J., Olff, H., Swenson, N.G., 2012. Testing the metabolic theory of ecology. Ecology Letters 15 (12), 1465–1474.
Saraiva, S., Fernandes, L., van der Meer, J., Neves, R., Kooijman, S., 2017. The role of bivalves in the balgzand: First steps on an integrated modelling approach. Ecological Modelling 359, 3448.
Sousa, T., Domingos, T., Kooijman, S.A.L.M., 2008. From empirical patterns to theory: A formal metabolic theory of life. Philosophical Transactions of the Royal Society B: Biological Sciences 363 (1502), 2453–2464.
Van der Meer, J., 2006. Metabolic theories in ecology. Trends in Ecology & Evolution 21 (3), 136–140.
West, G.B., Brown, J.H., Enquist, B.J., 1997. A general model for the origin of allometric scaling laws in biology. Science 276, 122–126.
West, G.B., Brown, J.H., Enquist, B.J., 2001. A general model for ontogenetic growth. Nature 413, 628–631.
Zuo, W., Moses, M.E., Hou, C., Woodruff, W.H., West, G.B., Brown, J.H., 2009. Response to comments on 'energy uptake and allocation during ontogeny'. Science 325 (5945).

Microclimate

KAS Mislan and B Helmuth, University of South Carolina, Columbia, SC, USA

Introduction

The microclimatological conditions surrounding an organism (the local conditions of climate as determined by aspects of the organism's microhabitat) control the exchange of factors such as heat, water, and nutrients between each organism and the surrounding environment. As a result, microclimates drive many aspects of organismal physiology and ecology. In this article, we first describe the physical characteristics of microclimates and the mechanisms by which organisms interact with microclimates through their behavior and morphology. We then discuss the physiological and ecological implications of microclimates over a range of temporal and spatial scales. Using these principles, we conclude with a discussion of mathematical modeling approaches that can be used to quantitatively predict organism distributions using microclimate characteristics.

Physical Characteristics

The ecological concept of microclimate depends largely on the question being addressed. To a landscape ecologist, a microclimate may comprise the side of a mountain, or a section of desert tens or perhaps even hundreds of kilometers in extent. To a reproductive ecologist interested in the life history of a mosquito breeding in a pitcher plant, the term microclimate refers to the inside of a plant only a few centimeters in diameter. Thus, from an ecological viewpoint, microclimate is very much defined by the organism or community in question. As a result, cohabitating species may have different responses to heterogeneity in their local microhabitat, and what constitutes ecologically and physiologically important variability for one species may comprise environmental noise for another. This complex interaction between each organism and its local environment is one of the foundations of population and community dynamics.

The dimensions of any particular microclimate are highly dependent on the mobility and dispersal capabilities of the organism in question. To an organism with limited movement and dispersal capabilities, the world may be restricted to a very small space; to an organism capable of foraging over large areas, such microclimates may appear as mere noise as it shuttles from place to place in search of food and shelter. Consider, for example, the section of rocky shore shown in **Fig. 1**. To a mussel tightly adhered to the substratum, living on the north versus south face of a rock may make an enormous difference in terms of the amount of solar radiation that the animal receives, and thus the maximum body temperature that the animal experiences. Differences in body temperature of 10 °C or more are not uncommon between sessile animals separated by a few centimeters due to differences in solar radiation created by substratum angle. In contrast, a gull foraging in the same intertidal zone may not be able to distinguish this small-scale thermal variability from the larger microclimate of the entire intertidal area. Furthermore, if microclimate conditions become intolerable in the intertidal area, the gull can spread its wings and fly to a more favorable location, perhaps a ledge on a nearby cliff. Measuring an organism's microclimatic conditions thus mandates that we first have an understanding of how far the organism is likely to `sample' its local environment.

The ecological consequences of microclimate heterogeneity are directly dependent on the relative differences in scale between predators (e.g., gulls, crabs, seastars) and prey (mussels, barnacles), and between competitors. Moreover, this scale determines the scale that scientists must measure or model microclimatic parameters. Equally important, however, is the fact that organisms themselves affect the flux of materials between themselves and their local environments. In the next section, we discuss the role of organism morphology and size in driving the transfer of heat to and from the organism's body, and how this affects the means by which we measure microclimates in the field.

Organism–Microclimate Interactions

All organisms exchange heat, water, and nutrients with their surrounding environment, and microclimate can have a significant influence on these rates of exchange. While the exchange of each of these factors is equally important, throughout the rest of this article, heat exchange will be used to explain the intricate relationship between organisms and their local microclimates.

In general, heat flux between an individual organism and the environment is divided into six categories: short-wave (visible) solar radiation, long-wave (infrared, IR) radiation to and from the sky and from the organism's surroundings, conduction to and from the ground, heat convected between the animal and the air, heat lost through the evaporation of water, and (for endotherms) heat generated through metabolism. Environmental variables such as habitat type, substrate orientation, as well as characteristics of the organism itself such as mass, morphology, and color, also drive rates of heat exchange. For example, organisms that have large proportions of their total surface area in contact with the underlying substratum (e.g., barnacles, lichens) may have body

 doi:10.1016/B978-0-444-63768-0.00520-5

Fig. 1 Rocky intertidal zone at Tatoosh Island, Washington State. The importance of microhabitat variability, and thus the concept of microhabitat, varies with movement and dispersal capability of the organism.

temperatures that are tightly coupled with ground temperature. In addition, some organisms use behavioral mechanisms to moderate the amount of heat flux. Snakes bask on hot rocks during the day and take refuge underneath rocks at night when surface conditions are too cold for them to survive. Due to the many variables involved in heat exchange, microclimate (air or surface) temperature and organism temperature are often dissimilar. Moreover, two organisms, even ones of the same species, can experience markedly different rates of heat transfer when exposed to identical microclimates, and can thus experience very different body temperatures.

Endotherms are able to maintain a relatively constant body temperature (i.e., are homeothermic) through the production of metabolic heat (during periods of increased heat loss) and cooling. However, the ability to produce heat is energetically costly for endothermic organisms, particularly when they are in microclimates with temperatures above or below their ideal environmental range. On cold winter nights, elk can reduce loss of radiant heat, and thus save metabolic energy that may be used for breeding in the spring, by seeking refuge beneath trees with needles. The elk can further reduce heat loss by laying down, but a prone position may make them easier prey for wolf predators. For an endotherm, finding the ideal physiological microclimates may be a trade-off between ecological costs such as avoiding predators and the physiological benefits such as gestating young.

In contrast, ectotherms, with negligible metabolic heat production, have body temperatures that change with the rate of heat transfer in and out of their bodies (i.e., are poikilothermic). As a result, most ectotherms have body temperatures that change rapidly as microclimatic conditions change; in some cases, these fluctuations can be rapid and quite large.

Therefore, when making measurements of microclimatic factors such as air temperature, wind speed, and surface temperature, it is important to consider how these factors are translated into factors such as body temperature, as well as to consider both the direct and indirect effects of body temperature on organismal physiology and ecology.

In most cases, the temperature of plants and animals does not track any single microclimatic parameter (such as air temperature; Fig. 2). Moreover, because organisms respond to the same environmental parameters in different ways because of their size and morphology, patterns in the same environmental parameter may not translate into the same pattern of organism stress. Importantly, it is the relative impacts of microclimate on different species that determine the importance of microclimate on ecological interactions.

Physiological Impacts

Weather extremes can cause the conditions of a microclimate to quickly exceed normal limits; these extremes can have catastrophic effects on organisms, particularly ectotherms. A heat wave is an example of a weather extreme that can cause harmful increases in microclimate temperature. In order to quantify the catastrophic effects of extremes on organisms, a physiological index, lethal temperature 50 (LT50), which describes the conditions when 50% of a population dies, has been used. However, the LT50 only provides a way of quantifying lethality at the population level, and there are more subtle effects that may precede these threshold events. Therefore, recent studies have emphasized the importance of quantifying the sublethal effects of weather extremes on organisms.

One of the new ways for quantifying damage from heat stress is to make measurements at the subcellular level. Extreme high temperatures can cause proteins to denature (unfold), but this damage can be prevented by the activity of heat shock proteins (HSPs), which prevent inappropriate interactions between the ends of damaged proteins that have begun to unfold. By measuring levels of HSPs, scientists can determine the relative heat stress an organism experiences. If organisms are responding to acute heat

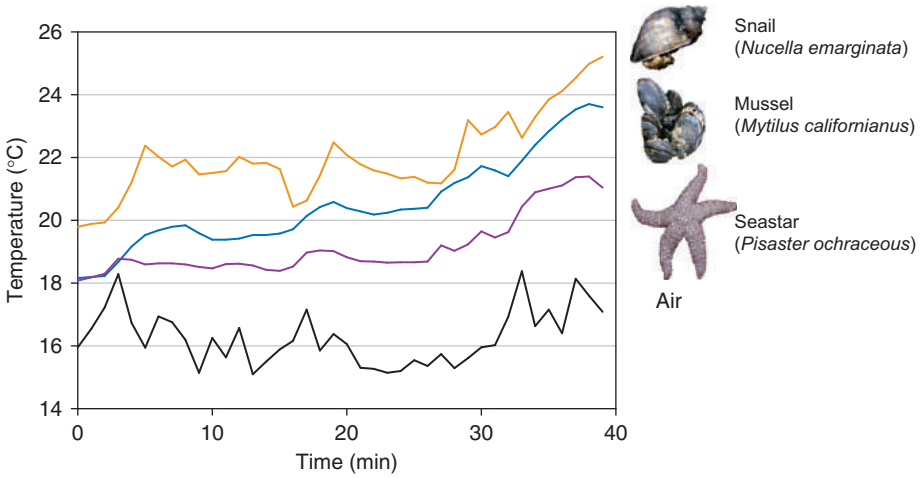


Fig. 2 Comparison between air temperature and body temperatures. Note that body temperatures do not track air temperature and vary for different organisms located in the same microhabitat. Data collected at Strawberry Hill, OR, USA, on 15 July 2006 by L. Yamane and L. VanThiel.

stress at the subcellular level, it is likely that there are other sublethal effects on the physiology of an organism. Making links between extreme temperatures, sublethal effects, and the ability of an organism to grow, survive, and reproduce is an area of active research.

Biogeographic Impacts

Climate plays an important role in determining the limits of species ranges. A species range is usually viewed as a north-to-south continuum with population densities slowly decreasing toward each end. However, variability in microclimate over a range of scales may cause breaks within this continuum, so that the abundance of organisms and levels of physiological stress wax and wane along each species distribution. Therefore, identifying the scale over which climate drives patterns of species distributions is the key to predicting the effects of climate change on ecological communities.

Climates are dynamic and warming and cooling trends occur both slowly, over the course of millennia, and more quickly over a period of years (e.g., El Niño Southern Oscillation and related La Niña events). Microclimates are ultimately derived from the overall climate of an area, so significant changes at the large-scale level will usually be reflected at the small-scale level. However, microclimates include multiple factors, and it can be difficult to predict the long-term direction of change. By observing the changes in particular microhabitats more vulnerable to climate change, it may be possible to make better predictions. For example, a community living on the south face of a mountain in the Northern Hemisphere may exhibit responses to climatic change earlier than a community on a nearby north face. Conditions observed in these vulnerable areas can be used in mathematical models to explore the possible changes in species distributions.

Mathematical Models

Recent approaches have used mathematical models to quantify heat flux between organisms and the environment using microclimate and climate data. These models play an increasingly vital role in improving our understanding of the biogeography and life history of many organisms, from plants to snails to elk. Climate data that must be collected for the models include air temperature, wind speed, solar radiation, cloud cover, and humidity in the vicinity of the organism in question.

The accuracy of the output from these models can be within 1–2 °C when compared to body temperature data collected at the same time from actual organisms. Because collecting data at the scale of the organism is often not practical, new ways have been developed to take factors such as substratum angle into account so that large-scale models of climate may be scaled to microclimatic levels.

There are many applications for mathematical models and an even greater number of ecological questions that can be explored. Some of the applications include: (1) generate and test hypothesis about where and when organism distributions are limited by aspects of their microhabitat (vs. the presence of predators or competitors); (2) hindcast past environmental conditions using historical climate data and make comparisons against historical changes in species distributions; (3) make short-term body temperature and microclimate forecasts to predict where physiological stress is most likely to occur; and (4) explore future climate conditions under a context of global climate change to determine the possible impacts on the physiological state and biogeography of different species.

Summary

Ecologists have long realized the crucial role of climate in determining the distribution and abundance of organisms and have benefited from recent technological innovations improving their ability to study microclimates. Clearly, scale is the most important element when considering microclimates, and it must be considered in the context of the organism of interest. Once the organism and microclimate scale are identified, it is possible to account for sources of heat, water, and nutrient flux and processes that mediate them. Microclimates have pervasive effects from the subcellular level to the biogeographic level making mathematical models, which integrate microclimate with environmental factors and organism physiology, the best way to further test ecological concepts.

Further Reading

Franco, A.C., Nobel, P.S., 1989. Effect of nurse plants on the microhabitat and growth of cacti. The Journal of Ecology 77, 866–870.
Helmuth, B., 2002. How do we measure the environment? Linking intertidal thermal physiology and ecology through biophysics. Integrative and Comparative Biology 42, 837–845.
Hofmann, G.E., 2005. Patterns of Hsp gene expression in ectothermic marine organisms on small to large biogeographic scales. Integrative and Comparative Biology 45, 247–255.
Holtmeier, F.-K., Broll, G., 2005. Sensitivity and response of Northern Hemisphere altitudinal and polar treelines to environmental change at landscape and local scales. Global Ecology and Biogeography 14, 395–410.
Huey, R.B., Peterson, C.R., Arnold, S.J., Porter, W.P., 1989. Hot rocks and not-so-hot rocks: Retreat-site selection by garter snakes and its thermal consequences. Ecology 70, 931–944.
Kingsolver, J.G., 1983. Thermoregulation and flight in *Colias* butterflies – Elevational patterns and mechanistic limitations. Ecology 64, 534–545.
Lausen, C.L., Barclay, R.M.R., 2006. Benefits of living in a building: Big brown bats (*Eptesicus fuscus*) in rocks versus buildings. Journal of Mammalogy 87, 362–370.
Nobel, P.S., 1988. Environmental Biology of Agaves and Cacti. Cambridge: Cambridge University Press.
Pincebourde, S., Casas, J., 2006. Multitrophic biophysical budgets: Thermal ecology of an intimate herbivore insect–plant interaction. Ecological Monographs 76, 175–194.
Porter, W.P., Sabo, J.L., Tracy, C.R., Reichman, O.J., Ramankutty, N., 2002. Physiology on a landscape scale: Plant–animal interactions. Integrative and Comparative Biology 42, 431–453.
Rosenberg, N.J., Blad, B.L., Verma, S.B., 1983. Microclimate: The Biological Environment. New York: Wiley.
Wethey, D.S., 2002. Biogeography, competition, and microclimate: The barnacle *Chthamalus fragilis* in New England. Integrative and Comparative Biology 42, 872–880.

Migration and Movement

Lars-Anders Hansson, Lund University, Lund, Sweden

Glossary

Dispersal Individuals or populations (or parts of populations) that move to reach new areas, but do not return.
Invertebrate Animal that has no backbone.
Migration Individuals or populations (or parts of populations) that move between two well-defined habitats on a temporally predictable basis.
Movement Individuals or populations (or parts of populations) that change position at any temporal or spatial scale. Movement includes all other ways of displacement.
Partial migration When some, but not all, individuals of a population migrate.

The Overall Question: To Move or Not to Move, and Why?

Movement is a fundamental process of life for most animals, and the ability to move fast and for a long time is often important both for escaping threats and for catching food. Although movement and migration are impressive phenomena, with continent-wide bird migrations and enormous biomasses of large animals walking long distances on African savannas, movement and migration are by no means the rule in nature. Instead, a stationary lifestyle seems to be as successful as costly movements. For example, any tree you see has been there for decades; it has been on the same spot, during bad as well as good times, irrespective of whether the place the seed once fell on was good or bad. On the other hand, the little bird searching for insects in the tree has chosen a completely different lifestyle by spending all day endlessly searching for the best food patch, the best life companion, the best protection against predators; it is always moving and when times get worse, it migrates to another continent. Hence, the moving animal searches for patches providing well-being, may it be food, a partner, or simply protection, and in places resembling paradise it stays longer, whereas it quickly leaves places of less quality (Brown, 1988). Hence the animal, be it a bird or a snail or a wolf, instantly judges the environment in order to optimize its well-being, which, in simple terms can be said to be the overall driving force of movements and migrations. This driving force towards well-being can be divided into measurable entities; the most common ones are summarized in **Fig. 1** and include components such as food, partner, temperature, predation, and infection risk.

Based on the composition of such drivers in the local environment, each taxon, or individual, may adopt a "movement strategy" aiming at optimizing its fitness (Bastille-Rousseau *et al.*, 2017). Such strategies may include dispersal, that is, leaving the area for potentially better conditions, migration to other places during rough times, such as during winter, but returning when conditions improve, or simply to stay in the same area year around (Mueller and Fagan, 2008). To stay always in the same area requires that moving has higher costs than benefits, whereas the decision to migrate or disperse is based on the benefits exceeding the costs of movement. Such strategies may, of course, change if conditions in the home area change. For example, locusts (e.g., *Locusta migratoria*) are driven to perform mass movements when food becomes scarce and then they become pests, eating anything that comes in their way. Similarly, many rodents, such as lemmings (*Lemmus lemmus*) are often relatively stationary, but perform mass movements (dispersals) away from barren habitats towards some distant paradise, although few of them reach it. However, the ones who happen to disperse to a suitable spot can successfully start exploiting and reproducing.

Partial Migration

In some species, for example, many insect-eating birds, all individuals are always taking the decision to migrate, whereas in other species, none migrates; they all stay in the same area year-round. However, in many species some individuals, but not all, decide to migrate. Such partial migration is actually more the rule than the exception and in most species there are often some individuals that migrate and others that stay, or at least perform only short migrations (Chapman *et al.*, 2011). This suggests that there is an individual dimension involved in the decision to migrate or stay. A possible logic behind this is that if all others migrate, then an individual staying year-round, and surviving, will benefit from low competition and an excess of food. Similarly, if all stay, the one who migrates and survives will have the advantages of less competition and lots of food during the migration period. This means that the advantages depend on which decisions the other individuals in the population are taking (Sinclair, 1983). Hence individuals in a population may take decisions based on expected costs and benefits, for example, they weigh the risk of predation against possibilities for food, and thereby growth.

An example of this dilemma is the partial migration in the fish roach (*Rutilus rutilus*) from the lake where predation is high and food level low in winter, to surrounding streams and wetlands where food is even lower, but where predators, such as pike (*Esox esox*) rarely follow them (Brönmark *et al.*, 2008). Hence a large portion (up to 60%) of the population may leave the lake during

 doi:10.1016/B978-0-12-409548-9.10596-2

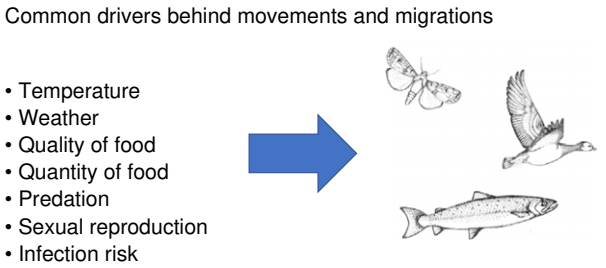


Fig. 1 Common drivers for organisms to undertake movements, dispersal, or migrations.

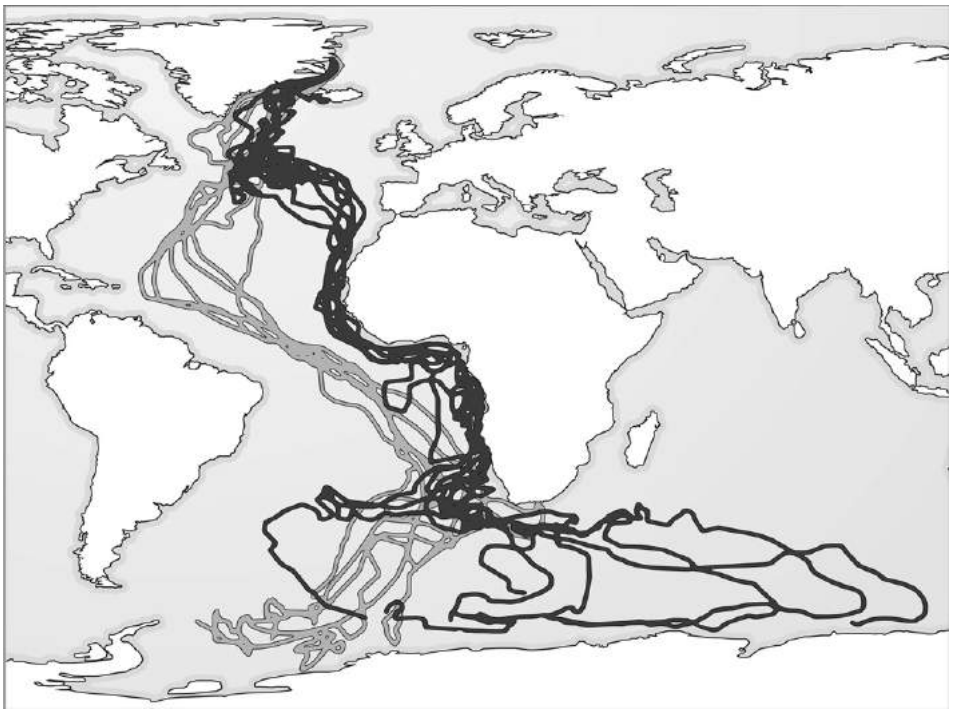


Fig. 2 Yearly migratory journey of Arctic terns (*Sterna paradisaea*) between Arctic regions and Antarctica. The tracks show the route of terns from breeding grounds in fall (*black lines*) and the return in spring (*gray lines*). Redrawn from Egevang, C., Stenhouse, I. J., Phillips, R. A., Petersen, A., Fox, J. W. and Silk, J. R. D. (2010). Tracking of arctic terns Sterna paradisaea reveals longest animal migration. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 2078-20781.

winter, reducing the ratio between predation risk and growth (food), which may be a way to endure, at least for the strongest individuals. Interestingly, less fit individuals are more prone to stay in the lake despite the predation risk, whereas individuals that are more well fed are, to a higher extent, taking on the winter migration to the predator refuge in streams and wetlands (Brodersen *et al.*, 2008). In spring the migratory part of the population returns to the lake and joins the residents. Thus, the decision to migrate or stay is context-dependent; that is, it differs from year to year according to the ratio between risk and benefit. It may, as suggested by the roach migrations, also be condition-dependent; that is, each individual asks themselves the question: "Am I fat enough to endure the whole winter in the stream?"

Generalist or Specialist?

The movements of organisms are a profound phenomenon in nature which has received considerable attention in science as well as in media. For example, the mass migration of large herds of ungulates on the African savannah (Bartlam-Brooks *et al.*, 2013), the migration of birds following similar routes every year, and, of course, all the extreme movements, such as the non-stop flying by wader birds (Gill *et al.*, 2009), the global-scale migrations in Arctic terns (*Sterna paradisaea*) (Egevang *et al.*, 2010) (**Fig. 2**), and the apparently almost endless duration of flight in swifts (Hedenström *et al.*, 2016). It is often the extremes that catch our attention.

To manage extreme journeys, an organism has to specialize and become the best flyer, runner, or swimmer. However, this has high costs and requires investments in organs and skeletons that leads to translocation of energy from other skills. This is why no organism on Earth is the champion in all locomotion modes: running, swimming, and flying. Instead the ones that are specialists in fast running are generally bad swimmers and do not fly at all, and the same is true for excellent flyers, which generally do badly as runners or swimmers; similarly, fast swimmers can generally not walk or fly. Since a normal life often requires that an organism has at least some skills in mastering more than one medium, most animals are more generalists; they are reasonably good at both walking and swimming, for instance, but would never win a competition against a champion. Actually, and maybe somewhat surprisingly, an organism that handles all movement skills reasonably well, that is, is a perfect generalist, may be something like a duck, which can fly, may reach impressive speed when running, and is also a good swimmer. Hence, although a generalist may not be the champion in either running, flying, or swimming, it manages all those locomotion modes pretty well. However, the generalist, such as the duck, could never take up a fight successfully with a specialist in swimming, running, or flying, and below we shall look more closely at some record-winning specialists.

Records in Moving

From an energy point of view, it is cheapest to move in water and most expensive to fly, whereas running on land is in between (Schmidt-Nielsen, 1972). Generally, the more muscle an organism has, the faster it can move, be it in water, air, or on land. However, this is true only up to a certain weight, and when an organism is too heavy—for example, an elephant—it becomes limited in its ability to provide quick energy to the muscles (Hirt *et al.*, 2017). Thus it is medium-sized animals, and not the largest with most muscles, that are generally among the fastest, such as the cheetah (*Acinonyx jubatus*) on land, the tuna fish (*Thunnus thynnus*) in water, and the peregrine falcon (*Falco peregrinus*) in air. In addition to high speed, endurance is also important and some organisms are very good at moving over long distances and for a long time. One of the champions is the wader bird godwit (*Limosa lapponica baueri*), which each year travels non-stop over the Pacific Ocean from breeding grounds in Alaska to New Zealand and Australia. This is a distance of about 10,000 km and the godwit manages it in only 5–9 days (Gill *et al.*, 2009). Since there is no place to rest or eat along the flight, this is an extraordinary extreme of avian flight performance.

Another extreme is the common swift (*Apus apus*). To investigate their position, speed, and altitude, researchers equipped some swifts with micro data loggers and accelerometers. The birds were then recaptured when returning back to their nests 10 months later with their data loggers, which showed that the swifts had been flying for 10 months without touching the ground (Hedenström *et al.*, 2016). This means that they eat, mate, and sleep airborne, and thus live the majority of their life in the air. These are just a few examples of fascinating record stories among moving animals, and we switch now to an even more fascinating question: how are organisms finding their way?

Finding Their Way

In addition to being able to move, an organism has to find its way; this skill is performed using many different senses and cues. The most obvious is to recognize and remember landmarks, such as rivers, lakes, or mountains. More advanced cues include the sun and the stars, despite them appearing to change position in the sky over time. This requires more advanced senses including a time-compensated sun compass (Schmidtkoenig, 1990; Åkesson *et al.*, 2014). Such a "biological compass" compares the position of the sun with the organism's internal clock, allowing them to determine their movement direction. Although the understanding of how such biological compasses function is still not complete, it has been shown to be used by many insects, such as honey bees, ants, and butterflies, as well as by birds, but it is likely that many more moving animals are able to use it (Åkesson *et al.*, 2014). In addition to the sun and stars, the Earth's magnetic field is also used by many organisms to find their way, and this is a very reliable cue especially when the weather is cloudy and the sun, stars, and landmarks cannot be used (Åkesson *et al.*, 2014; Muheim *et al.*, 2014). Many organisms also utilize smell to find their way, including migratory salmon (e.g., *Salmo salar*) which find their way from the ocean and back to the river where they were born by using smell (Stabell, 2012). Interestingly, most organisms utilize a combination of several of these cues and senses to find their way.

Skewed Knowledge on Animal Movements and Migration

Much of the fascination about animal movement and migration lies in the fact that we can see and track their movements, which has led to that the majority of studies performed, and thereby our understanding, is thus focused on larger animals, such as some large insects, birds, and mammals. However, there are probably about 7.7 million eukaryotic animals on Earth (Mora *et al.*, 2011) and most of those are smaller than 10 mm. As a thought experiment, we may combine the approximate number of animals in different size classes with the number of studies performed on their movements and migrations (**Fig. 3**). This thought-provoking analysis provides an interesting illustration of our skewed understanding of animal movements, especially when examining the knowledge per size class of existing species on Earth. This clearly shows that we know a lot about a few large organisms, whereas the movements and migrations of smaller organisms are still a black box (**Fig. 3**). However, this thought experiment may also

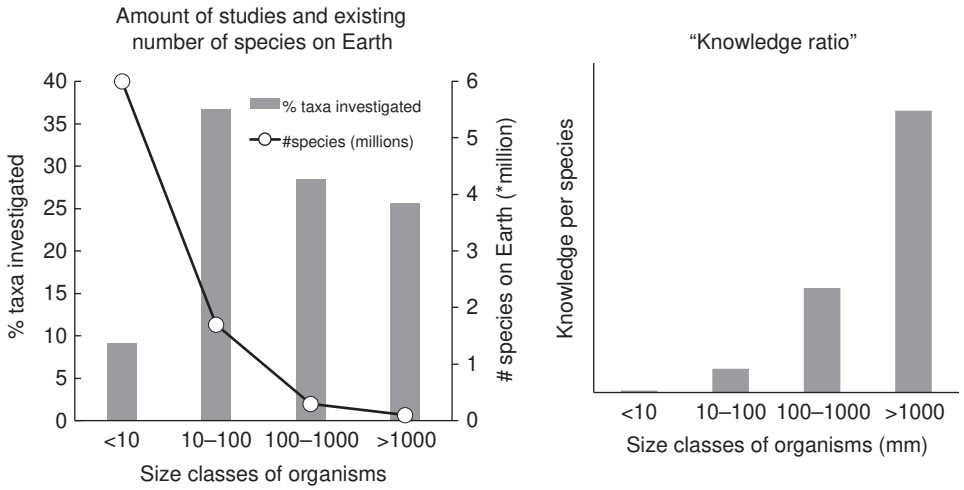


Fig. 3 Analysis of our knowledge of movements and migration of different-sized classes of animals, estimated as number of references (*left panel*; bars) in recent books (Hansson and Åkesson, 2014; Milner-Gulland *et al.*, 2011; Cheshire and Uberti, 2016). The line (*left panel*) shows the estimated number of eukaryotic species on Earth (Mora *et al.*, 2011). In the right panel this data is expressed as a ratio between amount of studies and number of species, showing that our knowledge is highly skewed towards organisms larger than 100 mm, whereas our knowledge on the movement and migration on smaller organisms is indeed scarce.

provide an incitement for researchers not only to study the movement of larger animals, but also to explore the fascinating variety of movement strategies adopted by smaller creatures. A major problem, and reason for the skewed knowledge, for studying small organisms is that they are unable to carry traditional tracking equipment, such as tags, radio, or GPS transmitters. A considerable technical development is therefore required, and below we shall look at a few examples on how such technical advancements can open up the study of the fascinating movements and migrations of small animals: a field where our knowledge is negligible in comparison to that of larger organisms.

Examples of Small Movers

Enormous amounts of insects are moving high up in the atmosphere and specialized radar techniques are required to study these high-flying insect migrants, as they are too small to carry transmitters or to be observed by any other means (Hu *et al.*, 2016). Because many migrant insects are extremely abundant, seasonal migration transfers enormous amounts of energy, nutrients, propagules, pathogens, and parasites between regions. For example, it has been estimated that such "bioflows" contain ~3.5 trillion insects (3200 tons of biomass) passing over the southern United Kingdom every year (Hu *et al.*, 2016). Although some of these insects perform proper migrations, most of them do not return. Instead they disperse, reproduce, and then the offspring makes the return flight. Below we shall look more deeply at some fascinating trans-generational migrations of insects, but also at the diel vertical migrations of tiny animals under the water surface, constituting enormous amounts of biomass migrating up and down through oceans and lakes.

Trans-generational migration: The monarch butterfly

Short-lived organisms, such as insects, do not live long enough to perform long-distance migrations. However, some, such as many lepidopterans, have solved this by spreading out the journey over several generations. An iconic example of such transgenerational migrations is the Monarch butterfly (*Danaus plexippus*), which during fall migrates southwards from the northern United States and Canada to Mexico, where it spends the winter. In spring, these butterflies fly north and lay eggs in the southern United States. The new generation then continues northwards to the northern United States and Canada and thereby closes the transgenerational migratory cycle (Reppert *et al.*, 2010). In all parts of this cycle the individual animal is naïve to the route—that is, it has never flown the route before. Therefore, it is puzzling how they manage to find their way, but it seems as though the Monarch butterflies are using several types of compasses, including both sun- and magnetic compasses (Guerra *et al.*, 2014).

The Bogong moth's nocturnal travels

Another transgenerational migrant is the Australian Bogong moth (*Agrotis infusa*), which make yearly migrations of more than 1000 km between the inland of the Australian continent and the south-eastern coast (Warrant *et al.*, 2016). In high altitude caves (up to 1800 m above sea level) the moths hibernate for several months and then migrate back to warmer regions, where they lay eggs and die. The next generation performs the same long journey without having any experience of how to find their way to the alpine caves. Since they are flying during night-time, they cannot use the sun to navigate, and although the moon and stars are possible cues when finding their way, these are less reliable. Instead a likely cue is the Earth's magnetic field (Åkesson *et al.*, 2014, Muheim *et al.*, 2014), although the successful navigation of Bogong moths is still an open question (Warrant *et al.*, 2016).

Aquatic zooplankton move away from predators and ultraviolet radiation

One of the largest migrations on Earth is the enormous numbers of mm-sized crustacean zooplankton moving downwards from sunlit surface waters of oceans and lakes during day and returning to surface waters every night to feed (Hays *et al.*, 1995). The drivers of this diel (or diurnal) vertical mass-migration (DVM) seem to be a combination of avoiding predation from visually hunting predators (fish), but also to escape from harmful solar ultraviolet radiation (UVR) (Hansson and Hylander, 2009a; Williamson *et al.*, 2011). Hence, zooplankton respond to each of these threats separately, but more strongly to a combination of the threats (Hansson and Hylander, 2009b). Although the DVM phenomenon has been known for decades, little is known about the individual dimension of the migration, mainly due to lack of suitable techniques. Recent developments in nanotechnology, however, have provided opportunities to utilize tiny "nano-lamps," that is, fluorescent particles that can be attached to the animal, allowing cameras to see and track their migrations and movements (Ekvall *et al.*, 2013) (**Fig. 4**). A short movie showing how two individuals of *Daphnia* move from harmful ultraviolet radiation down to a depth refuge is shown here: https://youtu.be/aqOGNziAf-0.

The technological development within nanotechnology has advanced our understanding of how, when, and why zooplankton move. For example, despite their small size and simple neurological systems, it seems that individuals of the zooplankter genus *Daphnia* may have some kind of "personality"; that is, different individuals show repeatability in their behavior, for example, some show a strong and some a weak response when exposed to a threat (Heuschele *et al.*, 2017). Although this is a very weak expression of personality, the study shows that the evolution of personality, and thereby individual decisions regarding, for example, migration, may not be restricted to more complex organisms, but may actually have evolved also in invertebrates. Such individual decisions may also partly explain why DVM is not performed by all individuals in a population, that is, it is a partial migration.

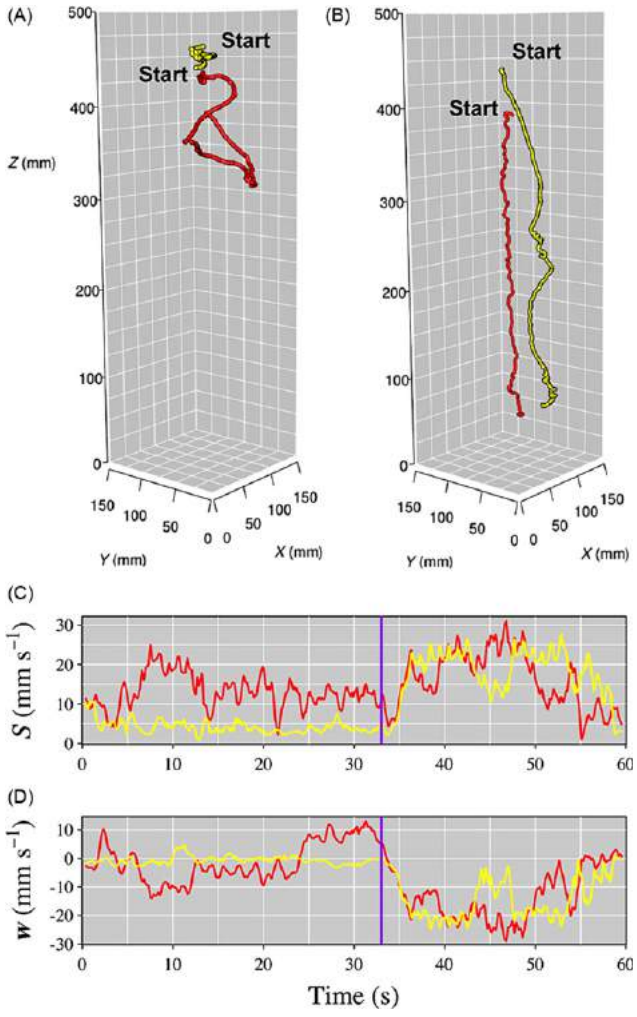


Fig. 4 Tracking of the position, speed, and vertical displacement of two *Daphnia magna* individuals in the absence (A) and presence (B) of UV radiation. The animals were marked with yellow and red nanoparticles (quantum dots), and monitored simultaneously to obtain 3D trajectories. Swimming speed (*S*) for both animals is shown in (C) and vertical speed (*w*) in (D). From Ekvall, T. M., Bianco, G., Linse, S., Linke, H., Bäckman, J. and Hansson, L.-A. (2013). Three-dimensional tracking of small aquatic organisms using fluorescent nanoparticles. *PloS One* 10.1371/journal.pone.0078498.

Parasites on the Move

It is likely that all moving animals are hosts to one or several parasites and pathogens, such as viruses and bacteria, which will hitch-hike on the host migrator to wherever it flies, walks, or swims. Hence, in addition to being forced to have an immune-defense against parasites and pathogens in its resident place, a migrant also has to handle the menagerie of parasites at the site to which it is migrating (Westerdahl *et al.*, 2005). Moreover, the diversity and abundance of parasites and pathogens are greater at lower (tropical) latitudes than at higher (temperate and polar regions) ones (Guernier *et al.*, 2004). Many birds migrate from temperate/polar regions, where they reproduce and raise their offspring, and the tropics, where they stay during winter when the temperature is low and food is scarce at higher latitudes. A logical question here is why do they not stay in the tropics all year round and reproduce where temperatures are high and food is in excess? It is probable that one of the drivers here is the high diversity of pathogens in the tropics, which has proven more beneficial to performing long migrations and raising offspring where the infection risk is lower, than to staying in the tropics year-round (Westerdahl *et al.*, 2014). This means that breeding at high latitudes may provide not only long days and a lot of food for the offspring, but also a pause for the immune system, suggesting that pathogens and parasites may actually be an important piece in the jigsaw puzzle of the evolution of migratory strategies in many organisms.

Migration and Climate Change

The climate is constantly changing, but during recent decades the change has been much faster and more pronounced in many regions, and since temperature is a major driver of movement, considerable changes in the migratory systems of organisms are also to be expected. There are already examples of how timing in migration has changed due to higher temperatures, and many songbirds have advanced their arrival by more than 2 weeks during the latest 50 years (Stervander *et al.*, 2005). Similar observations have been made on the migration of pink salmon (*Oncorhynchus gorbuscha*) up through the Alaskan rivers (Taylor, 2008). Climate-induced changes in migratory patterns may not have any major effects given that other variables, such as food availability, change in a similar way. However, different organisms respond differently to altered temperature, which has led to mismatches between return migration times of, for example, pied flycatchers (*Ficedula hypoleuca*) and their insect prey in the breeding areas in Europe. In the Netherlands, for example, spring, and thereby the burst of insect larvae, occurs earlier, whereas the arrival of migrating flycatchers has not changed much. This has meant that the flycatcher nestlings miss the period when insect larvae peak in abundance, that is, a climate-induced mismatch between predator and prey (Both *et al.*, 2006).

Human Movements

It is not only animals that are moving; humans have also moved and migrated throughout history. It is currently believed that man's first movement is likely to have started in eastern Africa, and that our ancestors then dispersed to other continents. Ever since, humans have moved on and conquered new land. However, pure migrations have also occurred, for example, when whole populations of Stone Age people went north to the edge of the ice-shelf to hunt in spring, and then migrated back south to warmer regions in fall. Similarly, many nomad tribes, for example, in North Africa, migrate back and forth depending on temporal differences in opportunities. Moreover, only 150 years ago a considerable part of the North European population left the prospect of starvation in Scandinavia, for example, and moved westward to settle in North America. Some of them returned to Europe, that is, they were migrants; others dispersed and stayed on the new continent and are now assimilated in the North American population. The main reason for this large-scale movement was lack of food (**Fig. 1**), and rumors of the "promised land" on the other side of the ocean, where food and land were available, which made millions of starving farmers take on the risky journey westwards. Similar movements have been common on all continents throughout the history of man, leading to permanent human settlings on all continents except Antarctica. Sadly, the reasons for human movements nowadays are still shortage of food or disasters, such as war. Hence, we may finally conclude that the drivers of movement, dispersal, and migration (**Fig. 1**) of humans are by no means different from similar phenomena among other animals.

See also: Behavioral Ecology: Dispersal–Migration

References

Åkesson, S., Boström, J., Liedvogel, M., Muheim, R., 2014. Animal navigation. In: Hansson, L.-A., Åkesson, S. (Eds.), Animal movement across scales. Oxford: Oxford University Press.

Bartlam-Brooks, H.L.A., Beck, P.S.A., Bohrer, G., Harris, S., 2013. In search of greener pastures: Using satellite images to predict the effects of environmental change on zebra migration. Journal of Geophysical Research: Biogeosciences 118 (4), 1427–1437.

Bastille-Rousseau, G., Gibbs, J.P., Yackulic, C.B., Frair, J.L., Cabrera, F., Rousseau, L.P., Wikelski, M., Kummeth, F., Blake, S., 2017. Animal movement in the absence of predation: Environmental drivers of movement strategies in a partial migration system. Oikos 126, 1004–1019.
Both, C., Bouwhuis, S., Lessells, C.M., Visser, M.E., 2006. Climate change and population declines in a long-distance migratory bird. Nature 441, 81–83.
Brodersen, J., Nilsson, P.A., Hansson, L.A., Skov, C., Brönmark, C., 2008. Condition-dependent individual decision-making determines cyprinid partial migration. Ecology 89, 1195–1200.
Brönmark, C., Skov, C., Brodersen, J., Nilsson, P.A., Hansson, L.A., 2008. Seasonal migration determined by a trade-off between predator avoidance and growth. PLoS One 3. e1957
Brown, J.S., 1988. Patch use as an indicator of habitat preference, predation risk, and competition. Behavioral Ecology and Sociobiology 22, 37–47.
Chapman, B.B., Brönmark, C., Nilsson, J.A., Hansson, L.A., 2011. The ecology and evolution of partial migration. Oikos 120, 1764–1775.
Cheshire, J., Uberti, O., 2016. Where the Animals Go. Oxford: Particular Books.
Egevang, C., Stenhouse, I.J., Phillips, R.A., Petersen, A., Fox, J.W., Silk, J.R.D., 2010. Tracking of arctic terns sterna paradisaea reveals longest animal migration. Proceedings of the National Academy of Sciences of the United States of America 107, 2078–2081.
Ekvall, T.M., Bianco, G., Linse, S., Linke, H., Bäckman, J., Hansson, L.-A., 2013. Three-dimensional tracking of small aquatic organisms using fluorescent nanoparticles. PLoS One. doi:10.1371/journal.pone.0078498.
Gill, R.E., Tibbitts, T.L., Douglas, D.C., Handel, C.M., Mulcahy, D.M., Gottschalck, J.C., Warnock, N., Mccaffery, B.J., Battley, P.F., Piersma, T., 2009. Extreme endurance flights by landbirds crossing the pacific ocean: Ecological corridor rather than barrier? Proceedings of the Royal Society B: Biological Sciences 276, 447–458.
Guernier, V., Hochberg, M.E., Guegan, J.F., 2004. Ecology drives the worldwide distribution of human diseases. Plos Biology 2, 740–776.
Guerra, P.A., Gegear, R.J., Reppert, S.M., 2014. A magnetic compass aids monarch butterfly migration. Nature Communications 5, 4164.
Hansson, L.-A., Åkesson, S. (Eds.), 2014. Animal movement across scales. Oxford: Oxford University Press.
Hansson, L.A., Hylander, S., 2009a. Effects of ultraviolet radiation on pigmentation, photoenzymatic repair, behavior, and community ecology of zooplankton. Photochemical & Photobiological Sciences 8, 1266–1275.
Hansson, L.A., Hylander, S., 2009b. Size-structured risk assessments govern daphnia migration. Proceedings of the Royal Society B: Biological Sciences 276, 331–336.
Hays, G.C., Warner, A.J., Proctor, C.A., 1995. Spatio-temporal patterns in the diel vertical migration of the copepod *M*etridia lucens in the Northeast Atlantic derived from the continuous plankton recorder survey. Limnology and Oceanography 40, 469–475.
Hedenström, A., Norevik, G., Warfvinge, K., Andersson, A., Backman, J., Akesson, S., 2016. Annual 10-month aerial life phase in the common swift apus apus. Current Biology 26, 3066–3070.
Heuschele, J., Ekvall, M.T., Bianco, G., Hylander, S., Hansson, L.A., 2017. Context-dependent individual behavioral consistency in daphnia. Ecosphere 8 (2), e01679. doi:10.1002/ecs2.1679.
Hirt, M., Jetz, W., Rall, B., Brose, U., 2017. A general scaling law reveals why the largest animals are not the fastest. Nature Ecology and Evolution 1, 1116–1122.
Hu, G., Lim, K.S., Horvitz, N., Clark, S.J., Reynolds, D.R., Sapir, N., Chapman, J.W., 2016. Mass seasonal bioflows of high-flying insect migrants. Science 354, 1584–1587.
Milner-Gulland, E.J., Fryxell, J.M., Sinclair, A.R.E., 2011. Animal Migration—A Synthesis. Oxford: Oxford University Press.
Mora, C., Tittensor, D.P., Adl, S., Simpson, A.G.B., Worm, B., 2011. How many species are there on earth and in the ocean? PLoS Biology 9.e1001127
Mueller, T., Fagan, W.F., 2008. Search and navigation in dynamic environments—From individual behaviors to population distributions. Oikos 117, 654–664.
Muheim, R., Boström, J., Åkesson, S., Liedvogel, M., 2014. Sensory mechanisms of animal orientation and navigation. In: Hansson, L.-A., Åkesson, S. (Eds.), Animal movement across scales. Oxford: Oxford University Press.
Reppert, S.M., Gegear, R.J., Merlin, C., 2010. Navigational mechanisms of migrating monarch butterflies. Trends in Neurosciences 33, 399–406.
Schmidtkoenig, K., 1990. The sun compass. Experientia 46, 336–342.
Schmidt-Nielsen, K., 1972. Locomotion—energy cost of swimming, flying, and running. Science 177, 222–228.
Sinclair, A.R.E., 1983. The function of distance movement in vertebrates. In: Swingland, I., Greenwood, P.J. (Eds.), The Ecology of Animal Movement. Oxford: Clarendon Press.
Stabell, O., 2012. Migration and navigation. In: Brönmark, C., Hansson, L.-A. (Eds.), Chemical ecology in aquatic systems. Oxford: Oxford University Press.
Stervander, M., Lindstrom, K., Jonzen, N., Andersson, A., 2005. Timing of spring migration in birds: Long-term trends, north Atlantic oscillation and the significance of different migration routes. Journal of Avian Biology 36, 210–221.
Taylor, S.G., 2008. Climate warming causes phenological shift in pink salmon, *Oncorhynchus gorbuscha*, behavior at Auke Creek, Alaska. Global Change Biology 14, 229–235.
Warrant, E., Frost, B., Green, K., Mouritsen, H., Dreyer, D., Adden, A., Brauburger, K., Heinze, S., 2016. The Australian Bogong moth *agrotis infusa*: A long-distance nocturnal navigator. Frontiers in Behavioral Neuroscience 10, 77.
Westerdahl, H., Waldenstrom, J., Hansson, B., Hasselquist, D., Von Schantz, T., Bensch, S., 2005. Associations between malaria and MHC genes in a migratory songbird. Proceedings of the Royal Society B: Biological Sciences 272, 1511–1518.
Westerdahl, H., Bensch, S., Nilsson, J.-Å., O'Connor, E., Sehgal, R., Tesson, S., Hasselquist, D., 2014. Pathogens and hosts on the move. In: Hansson, L.-A., Åkesson, S. (Eds.), Animal movement across scales. Oxford: Oxford University Press.
Williamson, C.E., Fischer, J.M., Bollens, S.M., Overholt, E.P., Breckenridge, J.K., 2011. Towards a more comprehensive theory of zooplankton diel vertical migration: Integrating ultraviolet radiation and water transparency into the biotic paradigm. Limnology and Oceanography 56, 1603–1623.

Monocultures Versus Polycultures[☆]

Matthew ES Bracken, Northeastern University, Boston, MA, United States

Introduction

In evaluating the interactions between organisms, ecologists are often interested in whether the performance of an individual species by itself (in a monoculture) is different from the performance of that species when other species are present (in a polyculture). Experiments comparing the growth, fecundity, or physiological rates of species in monocultures versus polycultures are used to assess competitive versus facilitative interactions between species and to evaluate the degree to which species are partitioning limiting resources.

These approaches have their origins in studies conducted to understand and improve the yield of agricultural crops, but they have recently been applied to more basic ecological questions, including evaluations of resource use, competition, and complementarity among species. Comparisons of monoculture versus polyculture performance have been especially useful in understanding the mechanistic links between the number of organisms and the rates of ecosystem-level processes in a given area.

Below, we briefly present the agricultural roots and history of experiments evaluating monocultures versus polycultures, describe some of the statistical methodology used in these comparisions, and illustrate how they have been used to evaluate both competitive interactions and the relationship between biodiversity and ecosystem functioning.

Agricultural Origins

The use of polycultures in agriculture, usually referred to as intercropping, is based on the traditional knowledge that carefully selected mixtures of crops are characterized by higher overall yields. This occurs because of more thorough use of limiting resources (complementarity), lower fertilizer requirements, greater resistance to herbivorous pests, and greater soil stability in polycultures when compared to monocultures. Additionally, growing multiple crops in a field provides farmers with a form of insurance: there is still something to harvest if one crop fails.

Because of these benefits, intercropping was the primary method of agriculture worldwide throughout most of history. Intercropping remains widespread in developing countries, though it has been largely abandoned in developed countries (e.g., the United States and in Europe) in the latter half of the 20th century due to the industrialization of agriculture. More recently, interest in organic farming techniques and sustainable agriculture has prompted First World farmers to return to this time-tested technique for increasing crop yield without applying chemical fertilizers and pesticides. Because of this agricultural legacy, many of the earliest experimental comparisons of monocultures and polycultures were conducted to evaluate the effects of mixed cropping on crop yields and to understand mechanisms of competition and coexistence between different agricultural species.

Competition and Coexistence

As ecologists began to experimentally evaluate the mechanisms underlying competitive interactions between species, it became clear that insights into these interactions could be obtained by comparing the performance of individuals in monocultures versus polycultures. For example, in Georgyi Gause's classic experiments on competition, he tested simple theoretical models of competition by measuring the relative population densities of two species of *Paramecium* (microscopic heterotrophs) competing for bacteria in experimental microcosms. Gause specifically compared the abundances of the two species in monoculture and in polyculture and found that both species perform well in monoculture, but when both species are cultured together, one of them (*Paramecium aurelia*) outcompetes the other (*Paramecium caudatum*), driving it locally extinct. Similarly, in another classic ecological experiment, Thomas Park evaluated competition between two species of *Tribolium* flour beetles by comparing their performance separately versus together and found that one species always competitively displaces the other. In the case of *Tribolium*, one of the species (*T. castaneum*) is susceptible to infection by a parasite. When the parasite is present, *T. castaneum* is outcompeted, whereas when the parasite is absent, *T. castaneum* outcompetes its congener.

[☆]*Change History:* February 2018. I. Martins made minor changes to the references.

However, the diversity of coexisting organisms on Earth suggests that competitive exclusion is not the rule. For example, 200–300 species of trees can coexist in a 100 × 100 m region of tropical rainforest, despite the fact that all of those species have similar basic requirements for potentially limiting resources such as light, nutrients, and water. How does this occur? Theory suggests that species coexist by carving out unique niches in resource space, so that they minimize their competition for resources.

What evidence is there for this phenomenon of resource complementarity? Again, comparisons of monocultures and polycultures have been used to quantify the degree to which species compete versus coexist. For example, plant assemblages are often composed of monocots, which are characterized by tall, erect shoots and shallow, laterally spreading roots, and dicots, which occupy lower aboveground strata and have deeper, less lateral rooting systems. When both monocots and dicots are mixed together in a polyculture, the yield (harvested dry mass) is higher than when either monocots or dicots are grown alone. This suggests that intraguild competition (monocots vs. monocots or dicots vs. dicots) is more intense than interguild competition (monocots vs. dicots), because the plants use different spatial niches. Mixtures of shallow- and deep-rooted species are characterized by more efficient use of limiting nutrients (e.g., nitrogen), resulting in higher rates of total nitrogen use, and leading to higher productivity in polycultures.

Several techniques have been used to statistically evaluate the relative importance of intra- versus interspecific competition in structuring assemblages of organisms. One of the most widely used metrics is the "relative yield total" (RYT, which is similar to the land equivalent ratio often used in agricultural intercropping studies). The relative yield of a species in a polyculture is defined as

$$\mathrm{RY}_i = \frac{P_i}{M_i} \tag{1}$$

where P_i is the observed yield of species i in polyculture and M_i is its yield in monoculture. Summing these values across all species gives the RYT:

$$\mathrm{RYT} = \sum \mathrm{RY}_i \tag{2}$$

This metric allows researchers to test the null hypothesis that observed yields are associated with proportional changes in the contributions of each component species (RYT = 1). When RYT = 1, the component species achieve their "expected yield." There is either no evidence of interactions between species, or the intensities of intra- and interspecific competitive interactions are equivalent. Deviations from RYT = 1 can provide some information on the interactions between the component species. For example, when RYT < 1, the component species may be competing for limiting resources, whereas when RYT > 1, the species may be complementary in their use of those resources.

It is important to realize, however, that the RYT simply reflects the combined interactions of the species in the mixture, which may include dominance, complementarity, competition, and facilitation. The mathematical properties of the RYT can also make it somewhat misleading as a summary of competitive versus complementary interactions. Consider, for example, a case where two species are grown in monoculture and in mixture. When the most productive species in monoculture has a lower than expected yield in polyculture, but the other species performs equally well in monoculture and in mixture, RYT can be >1, despite the fact that the polyculture yield is lower than that predicted by a weighted average of the component species' yields.

A conceptually simpler way to evaluate resource use by an assemblage of species is to compare the expected performance of a polyculture with the performances of its component species. When a polyculture performs better than predicted based on a weighted average of the performance of its component monocultures, it is said to overyield, which is evidence for either facilitation or resource-use complementarity. Overyielding can take two forms, nontransgressive overyielding, which occurs when a mixture performs better than the weighted average of its component monocultures, but does not perform better than the best-performing monoculture, and transgressive overyielding, which occurs when a mixture performs better than the best-performing monoculture

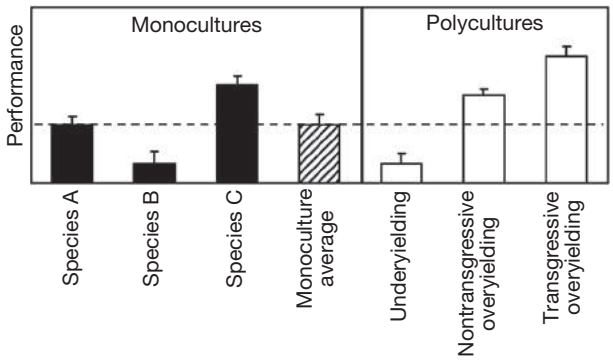


Fig. 1 Performance of monocultures and polycultures. On the left side of the figure the performances (e.g., yield, fecundity, or physiological rates) of three species grown in monoculture (*solid filled bars*) and the average performance of those three monocultures (*hatched bar* and *dashed horizontal line*) are illustrated. On the right side of the figure are three possible outcomes of an experiment containing all three species in mixture: underyielding, when the mixture does not perform as well as the monoculture average; nontransgressive overyielding, when the mixture performs better than the monoculture average but not better than the best-performing monoculture; and transgressive overyielding, when the polyculture outperforms the best-performing monoculture.

(Fig. 1). Distinguishing between mechanisms responsible for overyielding (i.e., facilitation vs. complementarity) requires a solid knowledge of the natural history of the system being investigated.

This method of comparing monoculture and polyculture performance is summarized by the $D_{\max}$ metric, which quantifies the degree of transgressive overyielding:

$$D_{max} = \frac{\sum P_i - \text{Max}\,(M_i)}{\text{Max}\,(M_i)} \tag{3}$$

where $\sum P_i$ is the total performance of the mixture (the sum of the yields of each component species in polyculture) and $\text{Max}(M_i)$ is the performance of the best-performing monoculture. When $D_{\max} > 0$, either complementarity or facilitation must be occurring.

Biodiversity and Ecosystem Functioning

More recently, these metrics have been formalized by researchers seeking to evaluate the effects of declining biodiversity on the transformation and flux of energy and matter in ecosystems. In the mid-1990s, scientists began to realize that more diverse assemblages were characterized by higher rates of productivity, growth, and resource utilization. However, this early work was criticized because researchers had not explicitly differentiated between two possible mechanisms for enhanced ecosystem functioning in more diverse assemblages: the "sampling effect," which occurs because a more diverse assemblage is more likely to contain one or more species with a dominant effect on the process being measured, and "complementarity," which occurs when organisms successfully partition limiting resources, reducing competitive overlap between them and leading to higher rates of collective resource use. Statistical comparisons of monocultures and polycultures were invoked as effective methods of differentiating between these possible mechanisms which link diversity and ecosystem functioning.

In particular, when the performance of a diverse assemblage is greater than predicted, it is possible to partition this positive influence of diversity into the portion attributable to complementarity (or facilitation) and the portion associated with the sampling effect. The sampling effect occurs when the best-performing species dominate the multispecies assemblages, and it is mathematically calculated as the covariance between the species yields in monoculture and their yields in the mixture. The net biodiversity effect is expressed as

$$\Delta Y = \sum \Delta \text{RY}_i M_i = N\overline{\Delta \text{RY} M} + N \,\text{cov}\,(\Delta \text{RY}, M) \tag{4}$$

where ΔY is the deviation of the polyculture from the yield predicted by its component monocultures, ΔRY_i is the deviation of species i from its expected relative yield, M_i is the yield of species i in monoculture, N is the number of species in the polyculture, and $\overline{\Delta \text{RY}}$ and $\bar{M}$ are the average deviation of the N species from their expected yields and the average yield of the N species in monoculture, respectively. The net biodiversity effect (ΔY) can then be separated into the complementarity effect $N\overline{\Delta \text{RY} M}$ and the sampling effect (N cov $(\Delta \text{RY}, M)$).

One real benefit of using these comparisons of monoculture and polyculture performance in evaluating the relationship between diversity and ecosystem functioning is the fact that important insights into the relationship can be gained using a relatively small number of experimental units. For example, imagine a (rather small) local pool of ten species. In order to generate a diversity gradient containing all combinations of those species, with richness levels of 1, 2, 3, ..., 10 species, a researcher would need to assemble over 1000 different experimental treatments. Replicating each of those treatments multiple times would create an incredibly unwieldy (though statistically powerful) array of experimental units. In contrast, a researcher could compare monocultures of each individual species with a polyculture of all ten species using only 11 unique treatments, allowing sufficient replication with a manageable sample size.

Approaches using the metrics and comparisons described above have advanced our knowledge of the relationship between biodiversity and ecosystem functioning in many different systems, including grasslands, soils, benthic marine habitats, streams, and estuaries. Other experiments using these techniques to evaluate the consequences of changing biodiversity have included comparisons of the invasibility of plots containing monocultures versus polycultures and evaluations of the effects of genetic (intraspecies) diversity on the recovery of species following disturbance events. Comparisons of monocultures and polycultures, originally used to evaluate crop yields and competitive interactions, have proven to be robust techniques for evaluating the effects of biodiversity on many different community and ecosystem functions, including both long-term processes (e.g., recruitment and/or growth of organisms over time) and physiological rates (e.g., uptake of nitrogen or utilization of carbon).

Further Reading

Bruno JF, Boyer KE, Duffy JE, Lee SC, and Kertesz JS (2005) Effects of macroalgal species identity and richness on primary production in benthic marine communities. *Ecology Letters* 8: 1165–1174.

Cardinale BJ, Palmer MA, and Collins SL (2002) Species diversity enhances ecosystem functioning through interspecific facilitation. *Nature* 415: 426–429.

Connell JH (1983) On the prevalence and relative importance of interspecific competition: Evidence from field experiments. *American Naturalist* 122: 661–696.

Emmerson MC and Raffaelli DG (2000) Detecting the effects of diversity on measures of ecosystem function: Experimental design, null models and empirical observations. *Oikos* 91: 195–203.

Federer WT (1993–1999) *Statistical design and analysis for intercropping experiments*. vol. 2. New York: Springer.
Feng L, Wang G, Han Y, Li Y, Zhu Y, Zhou Z, and Cao W (2017) Effects of planting pattern on growth and yield and economic benefits of cotton in a wheat-cotton double cropping system versus monoculture cotton. *Field Crops Research* 213: 100–108.
Fridley JD (2001) The influence of species diversity on ecosystem productivity: How, where, and why? *Oikos* 93: 514–526.
Groc S, Delabie JHC, Fernandez F, Petitclerc F, Corbara B, Leponce M, Céréghino R, and Dejean A (2017) Litter-dwelling ants as bioindicators to gauge the sustainability of small arboreal monocultures embedded in the Amazonian rainforest. *Ecological Indicators* 82: 43–49.
Hector A (1998) The effect of productivity on diversity: Detecting the role of species complementarity. *Oikos* 82: 597–599.
Horwith B (1985) A role for intercropping in modern agriculture. *Bioscience* 35: 286–291.
Jolliffe PA (1997) Are mixed populations of plant species more productive than pure stands? *Oikos* 80: 595–602.
Jolliffe PA (2000) The replacement series. *Journal of Ecology* 88: 371–385.
Loreau M (1998) Separating sampling and other effects in biodiversity experiments. *Oikos* 82: 600–602.
Loreau M and Hector A (2001) Partitioning selection and complementarity in biodiversity experiments. *Nature* 412: 72–76.
Pradana YS, Sudibyo H, Suyono EA, and Indarto BA (2017) Oil algae extraction of selected microalgae species grown in monoculture and mixed cultures for biodiesel production. *Energy Procedia* 105: 277–282.
Stachowicz JJ, Fried H, Osman RW, and Whitlach RB (2002) Biodiversity, invasion resistance, and marine ecosystem function: Reconciling pattern and process. *Ecology* 83(9): 2575–2590.
Syafiq M, Rahman A, Atiqah N, Ghazali A, Asmah S, Yahya MS, Aziz N, Puan CL, and Azhar B (2016) Responses of tropical fruit bats to monoculture and polyculture farming in oil palm smallholdings. *Acta Oecologica* 74: 11–18.
Tilman D (1999) The ecological consequences of changes in biodiversity: A search for general principles. *Ecology* 80: 1455–1474.
Vandermeer JH (1989) *The ecology of intercropping*. Cambridge: Cambridge University Press.

Numerical Ecology

Pierre Legendre, Université de Montréal, Montréal, QC, Canada

Introduction

Numerical ecology is the field of quantitative ecology devoted to the numerical analysis of [mostly multivariate, but also time series] ecological data, with emphasis on community composition data. Community ecologists, whose data are multivariate by nature (many species, several environmental variables), are the primary users of these methods. Hence, population dynamics, single-species distribution models or the analysis of single species spatial patterns, which are powerful applications of mathematical ecology, are not considered parts of numerical ecology sensu stricto.

Numerical ecology is a sub-discipline of ecology, not of statistics or other mathematical discipline. In numerical ecology, the analysis starts with consideration of an ecological question and the data available to answer it. Numerical analysis methods are chosen to answer the question at hand and test ecological hypotheses about the data. When tests of statistical significance are in order, ecologists use permutation tests in most cases. These tests are applicable to non-normal univariate or multivariate data, in particular multivariate community composition data.

Many of the methods used in numerical ecology have been developed by ecologists, specialists of classification methods, geneticists and other researchers who were facing questions about multivariate data in their fields of study. Their training often combined statistics and some field of ecology or biology, in different proportions. So, historically, it is the people who needed to analyze intricate data sets to address high-level scientific questions in their fields of research, and had statistical or numerical training, who often developed statistical or numerical methods of data analysis. The first statisticians, people like Galton, Pearson and Spearman, who created the bases of modern-day statistics, had not been trained as statisticians either: that field did not exist before their work.

The field was reviewed and synthesized by Legendre and Legendre in five editions of a book published in French ("*Écologie numérique*") and in English ("*Numerical Ecology*") from 1979 to 2012. Because of the successive editions of this successful book over more than three decades, people often associate the field to the names of these authors. This short article will show how these books were part of a trend in the ecological literature that started before the 1960s and involved many researchers.

When Legendre and Legendre published the first editions of their book, they called it "*Écologie numérique*" and "*Numerical Ecology*" to emphasize the lineage with the field of *numerical taxonomy*, founded in 1963 by microbiologist Peter H. A. Sneath and population geneticist Robert R. Sokal. Numerical taxonomy aimed at testing hypotheses about biological systematics, population biology, and phenetic, phyletic and phylogenetic relationships, using multivariate data analysis. The approach included explicit steps to create dendrograms and cladograms using numerical methods, instead of the subjective syntheses of data that were generally favored until then. Likewise, numerical ecology includes steps to test ecological hypotheses using data and explicit methods of numerical analysis.

A Brief History

Pioneer Researchers

Numerical ecology developed thanks to the work of a numerous researchers. Until about 1970, community ecology had been mostly a descriptive science, although some ecologists had ventured into mathematical analyses. Pioneer researchers who developed key concepts and numerical methods of great importance for multivariate data analysis include the following well-known scientists:

- Vegetation scientist Paul Jaccard, working in the Alps, developed the first similarity coefficient used to analyze vegetation survey data (Jaccard, 1900). His coefficient is still in wide use nowadays in all fields where scientists analyze multivariate presence-absence observational data.
- In 1954, the vegetation ecologist David Goodall was the first to use factor analysis in community ecology. Goodall proposed the term "ordination" to designate this type of analysis, a term now widely used in textbooks and publications in community ecology and many other fields (Goodall, 1954). At that before-computer time, several other ecologists had experimented with numerical methods to address ecological questions.
- Robert R. Sokal (State University of New York, Stony Brook, United States) developed *numerical taxonomy* with Peter H. A. Sneath (University of Leicester, England; their foundation textbooks were published in 1963 and 1973; see Further Reading) and promoted the use of multivariate data analysis in biology and ecology. These two researchers proposed several methodological developments, including similarity coefficients and clustering methods, and experimented with the use of computers.
- John C. Gower spent his career (1955 to present) developing numerical methods of analysis for numerical taxonomy, numerical ecology and agricultural experimentation. He was also a pioneer in the use of computers at the Rothamsted

Experimental Station in England. The author of this article had the privilege to work with him on the properties of dissimilarity coefficients (Gower and Legendre, 1986).

- Robert H. Whittaker (Cornell University, United States) proposed the five-kingdom taxonomic classification of the world's biota into the Animalia, Plantae, Fungi, Protista, and Monera (Whittaker, 1969) and developed the key ecological concepts of alpha, beta, and gamma diversity (Whittaker, 1972). He hired young collaborators who wrote and distributed important computer software for community ecology.
- Cajo J. F. ter Braak (Wageningen University, The Netherlands) developed canonical correspondence analysis and many other key methods related to canonical ordination. He also wrote the Canoco program (the first version was developed in 1985), which was the first generally available software for community ordination, simple and canonical. Successive versions were released to researchers from 1988 (version 2.1) and in the following years up to now. The history of the Canoco software is recounted in ter Braak (1988) and in Section 1.2, entitled "Canoco for Windows", of the successive versions of the Canoco manual, for example, ter Braak and Šmilauer (2002).

Numerical ecology is the result of many years of collaborative work among many dedicated researchers in the fields of numerical classification and quantitative ecology, too many to be listed here. Many of them are cited in the *References* sections of the "*Numerical Ecology*" and "*Numerical Ecology with R*" books. These collaborations are illustrated in **Fig. 1**.

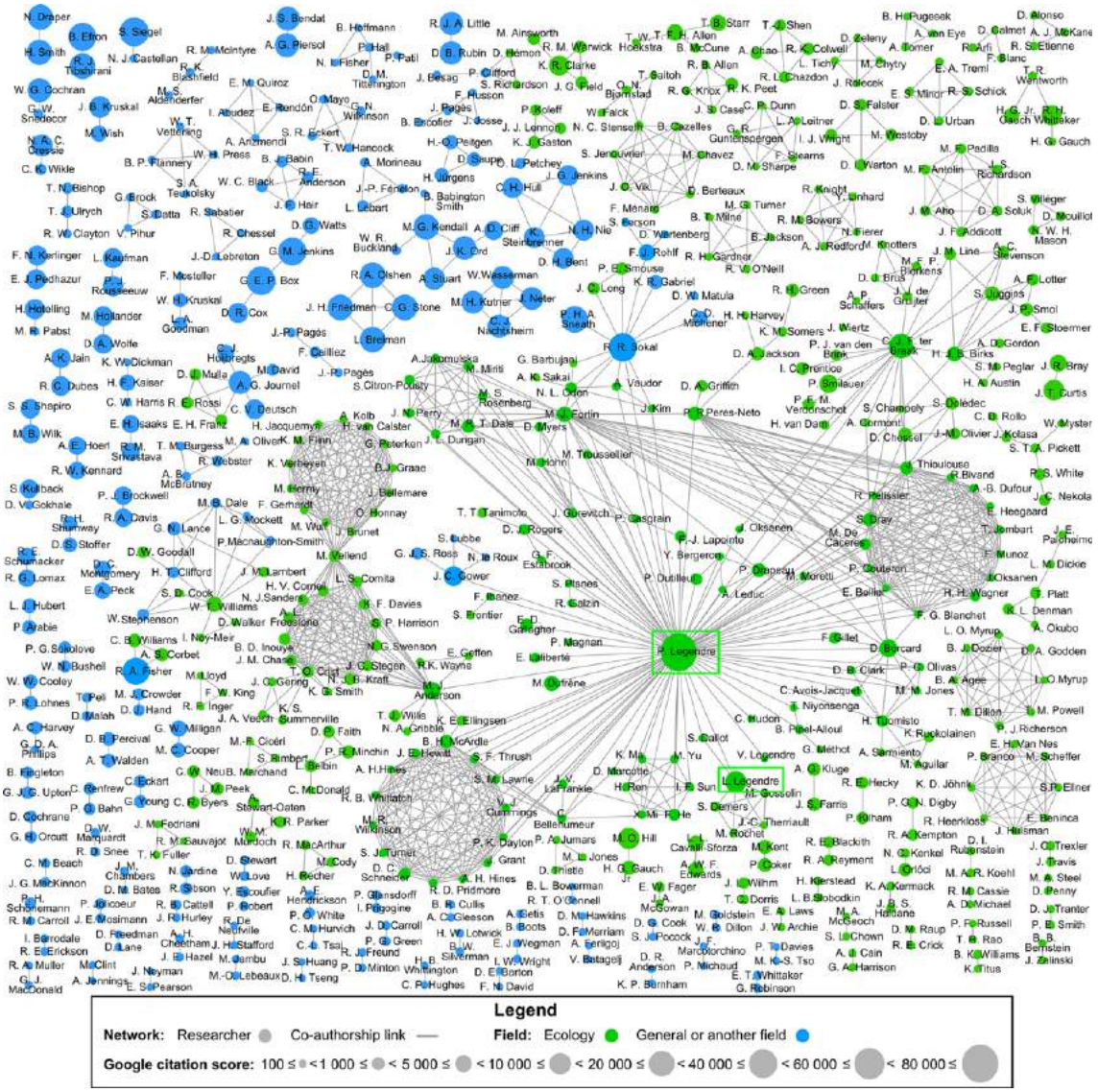


Fig. 1 This figure describes the network of collaborators who produced the references in the 2012 edition of the *Numerical Ecology* book (Legendre and Legendre, 2012). Single-author references were excluded. Network computed and kindly provided for use in this article by Prof. V. Makarenkov, Department of Computer Sciences, Université du Québec à Montréal.

Numerical ecology has been able to make great progress in the computer age thanks to the dedication of many developers of statistical packages, especially in the R language, who wrote software designed to analyze ecological data. Cited here, in alphabetic order, are some of the packages, available on the Comprehensive R Archive Network (CRAN) Web site, that have been developed by and for ecologists: ade4, adespatial, BiodiversityR, cocorresp, codep, ecodist, FactoMineR, FD, labdsv, lmodel2, mvpart, pastecs, picante, princurve, rioja, vegan, vegclust. That list is not exhaustive. Many other packages and functions are available on researchers personal web pages or in appendices of published papers describing statistical methods for ecological analysis.

Textbooks

From 1969 to 1979, the contributions of the previous decades were synthesized in four textbooks that marked the foundation of the field of numerical ecology:

- Statistical ecologist Evelyn Christine Pielou, Professor of mathematical biology at Queens' University (Kingston, Canada), formally introduced the field in 1969 by publishing a textbook entitled "*An Introduction to Mathematical ecology*" (Pielou, 1969). The Preface opened with the following sentences:

  "The fact that ecology is essentially a mathematical subject is becoming ever more widely accepted. Ecologists everywhere are attempting to formulate and solve their problems by mathematical reasoning, using whatever mathematical knowledge they have acquired, usually in undergraduate courses or private study. The purpose of this book is to serve as a text for these students and to demonstrate the wide array of ecological problems that invite continued investigation." (Pielou 1969, p. v.)
- László Orlóci, University of Western Ontario, London, Canada, published in 1975 "*Multivariate Analysis in Vegetation Research*" with a clear orientation towards community ecology (Orlóci, 1975). The main articles described ways of computing resemblance functions as well as methods of ordination and classification.
- Roger Green, who was Orlóci's colleague at the University of Western Ontario, London, Canada, published in 1979 "*Sampling Design and Statistical Methods for Environmental Biologists*" oriented towards animal ecology and sampling designs (Green, 1979). The book is a comprehensive guide to the principles of sampling design and methods of statistical analysis. It reviews the principles of inference, sampling and statistical design, and hypothesis formulation, with reference to ecological data.
- The first French and English editions of the Legendre and Legendre (1979, 1983) numerical ecology textbook provided a differently oriented synthesis of statistical methods aimed at all fields of ecology. The authors presented the mathematical bases of the methods of data analysis, and illustrated these methods with easy-to-compute numerical examples and real-data ecological applications drawn from the published literature.

Contribution of the Legendre Brothers

In May 1975, a dozen or so ecologists, mostly marine, sat during 3 days in a classroom on the second floor of a historical building of the *Station marine de Villefranche-sur-Mer* (Université Paris 6, France), a few meters away from the Mediterranean shore, to discuss developments concerning a new trend in the ecological literature: the statistical analysis of multivariate ecological data. The meeting was called "Séminaire de mathématiques appliquées à l'océanographie biologiques" and had a marine ecology orientation.

Because they had both worked in data analysis, Louis Legendre (oceanographer, Université Laval, Canada) and Pierre Legendre (community ecologist, Université du Québec à Montréal, Canada) had been independently invited to participate in the seminar, where they contributed several presentations. On the evening of the closing day of the meeting, sitting at the terrace of a restaurant with view on the harbor, Louis and Pierre Legendre wrote, on a paper place mat, a list of subjects, which was to become the table of contents of a book about the new subdiscipline of ecology that had been discussed during the seminar. They published the first edition of the book in 1979, in French, under the title "*Écologie numérique*."

History of Publication of the Numerical Ecology Textbook

- The first edition of "*Écologie numérique*" (in French) was published in 1979 by Masson, in Paris, and Presses de l'Université du Québec in Québec City (two volumes, 473 pages in total).
- The work was translated into English, under the supervision of the two authors, and published in 1983 by Elsevier Scientific Co. in Amsterdam under the title "*Numerical Ecology*" (435 pages).
- A second French edition, revised and augmented, was published in 1984 by the two original publishing houses (2 volumes, 618 pages in total) (Legendre and Legendre, 1984a).
- During the 1980s, community ecologists started to study species-environment relationships thanks to the computer package Canoco made available by Cajo ter Braak. During the 1990s, they became aware of the importance of spatial structures to understand the spatial variation of community composition (Levin, 1992; Legendre, 1993). The second English edition of "*Numerical Ecology*" was published in 1998 by Elsevier (868 pages) (Legendre and Legendre, 1998). It mostly focused on modeling the multivariate structure of community composition data. It incorporated a whole chapter on canonical ordination

and one on spatial analysis, and described the partitioning of the variation of community composition data into spatial end environmental components, a method, now very popular, that had been proposed by Borcard *et al.* (1992).

- The years 2000 were marked by the development of multiscale variation partitioning, a development initiated by Borcard and Legendre (2002), and the progress of methods for the analysis of beta diversity. The third English edition of "*Numerical Ecology*," published in 2012 (1006 pages), featured a new chapter on multiscale spatial eigenfunction analysis, as well as substantial additions to most other chapters.

In the meantime, a companion book, "*Numerical Ecology with R*", had been written by Daniel Borcard, François Gillet and Pierre Legendre and published by Springer Science in 2011 in the *Use R!* book series. The book contained detailed accounts of the computation of the numerical ecology methods of analysis using R packages. It was based on the major developments of R packages for ecologists since the year 2000, produced by various groups and their collaborators around the world, including the packages vegan (2001) and ade4 (2002). The list of R packages used in the various articles occupies several pages at the end of the book. A second edition of the R book was published by Springer in 2018. The R book was translated to Chinese by Jiangshan Lai (Institute of Botany, Chinese Academy of Science) and published in 2014 by Higher Education Press (Beijing). The data sets used in the R books (Borcard *et al.*, 2011, 2014, 2018) and the scripts of all analyses are freely available on a Web page cited at the end of this article.

In February 2018, the various editions of the "*Numerical ecology*" textbook had been cited more than 18,000 times in the scientific literature and the "*Numerical Ecology with R*" book more than 1600 times.

Important Papers Across the Years

Users of numerical methods and graduate students often wonder where the basic ideas of the methods we are routinely using come from and how they were developed. Here is a selection of papers that have changed the way ecologists analyze multivariate data during the past 50 years and the teaching of numerical ecology to graduate students in universities. The following list is by no means exhaustive.

The years 1960 and 1970—Development of redundancy analysis (RDA); Rao (1964) called the method "principal components of instrumental variables"; van den Wollenberg (1977) called it "redundancy analysis." Principal coordinate analysis (PCoA): Gower (1966). The concepts of alpha-beta-gamma diversity: Whittaker (1972). Time-constrained clustering: Gordon and Birks (1972, 1974).

1980–89—Spatially-constrained clustering: Lefkovitch (1978); Legendre and Legendre (1984b). Metric and Euclidean properties of dissimilarity coefficients: Gower and Legendre (1986). Canonical correspondence analysis (CCA): ter Braak (1986, 1987a,b). Spatial analysis as a tool for community ecologists: Legendre and Fortin (1989).

1990–99—The method of variation partitioning: Borcard *et al.* (1992). Spatial autocorrelation, a new paradigm for ecology: Levin (1992), Legendre (1993). Co-inertia analysis (CoIA): Dolédec and Chessel (1994). Indicator species analysis: Dufrêne and Legendre (1997). RLQ analysis: Dolédec *et al.* (1996). Fourth-corner analysis: Legendre *et al.* (1997); Dray and Legendre (2008); Dray *et al.* (2014). Distance-based redundancy analysis (dbRDA): Legendre and Anderson (1999).

2000–09—Transformations for community composition data prior to linear ordination, Legendre and Gallagher (2001), leading to transformation-based PCA (tbPCA) and transformation-based RDA (tbRDA). Spatial eigenfunction analysis—Moran's eigenvector maps (MEM): Borcard and Legendre (2002); Dray *et al.* (2006); asymmetric eigenvector maps (AEM): Blanchet *et al.* (2008). Concordance analysis of species associations: Legendre (2005). The rationale for estimation of beta diversity by the variance of the community composition data table, Var(**Y**): Legendre *et al.* (2005). Improving indicator species analysis: De Cáceres and Legendre (2009); De Cáceres *et al.* (2010).

2010 to present—Should the Mantel test be used in spatial analysis? Legendre and Fortin (2010), Legendre *et al.* (2015). Testing the space-time interaction in community surveys: Legendre *et al.* (2010). Test of significance of the canonical axes in RDA: Legendre *et al.* (2011). Partitioning beta diversity: Legendre and De Cáceres (2013), Legendre (2014). Temporal and space-time analysis of beta diversity: Legendre and Gauthier (2014). Study of temporal beta diversity: Legendre and Salvat (2015). Multiscale codependence analysis (MCA), which quantifies the joint spatial distribution of a pair of variables at different spatial scales (Guénard *et al.*, 2010), was generalized to handle multivariate response data (Guénard and Legendre, 2018).

Workshops

On 3–11 June 1986, Pierre and Louis Legendre, assisted by Marie-Josée Fortin (now Professor at University of Toronto), organized a NATO Advanced Study Workshop on Numerical Ecology at the *Station biologique de Roscoff* in France. Methods of data analysis were presented by statisticians and methodologists, followed by discussions of their application to ecological problems by working groups of ecologists. A book of proceedings was published after the workshop (Legendre and Legendre, 1987).

On 26–28 May 2008, a workshop entitled Spatial Ecological Data Analysis with R (SEDAR) was held at *Université Claude Bernard* in Lyon. It had been organized by Stéphane Dray to coordinate efforts among researchers developing the spatial analysis of ecological data and make plans for the future. One of the results of this workshop was a new R package, adespatial, for spatial and time-series analysis of community data. Written under the direction of Stéphane Dray, adespatial appeared on CRAN on 6 June 2016. New functions are still being added to this package.

On 6–7 October 2016, a workshop organized by Pedro Peres-Neto (Concordia University) and Marie-Josée Fortin (University of Toronto) was convened at Concordia University in Montreal. Twenty participants discussed future developments of the field. Following the meeting, one of the participants, Prof. Vladimir Makarenkov, computed a network describing the scientific collaborations that produced the wealth of references to numerical methods included in the 2012 edition of the "*Numerical Ecology*" book. Although the list of references at the end of a textbook is admittedly biased in favor of its authors, this network (**Fig. 1**) illustrates the fact that the development of data analysis methods for ecologists is the result of a broad and fruitful collaboration among many scientists.

Developments in Progress

Community Ecology

One of the primary concerns of community ecology nowadays is to test hypotheses about the processes that generate and maintain biodiversity in ecosystems, in particular beta diversity (i.e., the spatial variation in community composition among sites) in a region, through neutral processes, abiotic environmental species filters and biotic interaction filters. Numerical ecology develops and provides the statistical methods to test such ecological hypotheses.

Methodological developments on which researchers are presently working include the following topics:

- Beta diversity analysis of spatially distributed genetic, molecular and trait data. This work extends the concept of beta diversity, which was originally defined as the spatial variation of community composition data, to other types of biodiversity data.
- Beta diversity analyses across temporal and space-time surveys. Comparison of two and multiple surveys across time. Identification of the processes that cause changes in community composition for species, genetic, molecular, and trait data.
- New advances in spatial modeling by spatial eigenfunction analysis. Translation of landscape resistance networks into spatial eigenfunctions.
- Paleoecological analysis: new advances in time-constrained clustering and other methods for modeling abrupt changes in multi-species paleoecological data series.
- Three-table analysis. The basic methods, called RLQ analysis (Dolédec *et al.*, 1996) and fourth-corner analysis (Legendre *et al.*, 1997), test hypotheses of relationships between species traits and environmental characteristics of the sites mediated by the observed site-by-species data matrix. The two methods were unified by Dray and Legendre (2008) and Dray *et al.* (2014). Future developments, recently published or under discussion, involve other characteristics of the species, for example their phylogeny, and other characteristics of the sites, for example their spatial structure. Examples of such extensions are given in Dray *et al.* (2014). Mathematical extensions of the method are also considered.
- Analysis of multi-species community data using multivariate generalized linear mixed models (GLMM) or a latent variable model (LVM) that combines GLM with Markov Chain Monte Carlo (MCMC) methods (Warton *et al.*, 2015), as in the boral R package (Hui, 2017).
- Analysis of ecological networks.

The methods developed for community ecology can be transferred to other fields where empirical research is also concerned with frequency data, namely: gene frequencies, molecular data (including those used in microbiology nowadays), as well as biological and behavioral trait analysis.

Software Development

In most cases nowadays, new methods of analysis are published accompanied by software. For decades, Fortran, then C dominated the programming environment. An important program, which implemented a variety of simple and canonical ordination methods and allowed researchers to apply them to data, is Canoco; version 2.1 became available in 1988 (ter Braak, 1988).

R is a free software environment for statistical computing and graphics distributed on the CRAN Web site. The first stable version, R 1.0, appeared on CRAN on 29 February 2000. The R language is in fashion at the moment, and it is likely to be around for quite some time, given that the research community has produced thousands of packages, each containing from a few to hundreds of functions for data analysis. More than 12,000 packages are presently distributed on the CRAN site, in addition to the many other packages and functions available on individual researchers' Web pages or found in appendices of methodological papers. R may eventually be replaced by other software development environments, or complemented by other more specialized programming and computing environments. The future will tell, but for sure, ecologists will keep computing.

Conclusion

As pressing new ecological questions emerge in the world, ecologists and methodologists will keep developing methods of data analysis to answer these questions using multivariate data and enrich the methodological framework of numerical ecology.

Acknowledgments

Many thanks to Louis Legendre (Université Paris 6, France) for comments on a first draft of this article, and to Vladimir Makarenkov (Université du Québec à Montréal, Canada) who computed the network of scientific collaborations shown in **Fig. 1**.

References

Blanchet, F.G., Legendre, P., Borcard, D., 2008. Modelling directional spatial processes in ecological data. Ecological Modelling 215, 325–336.
Borcard, D., Legendre, P., 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecological Modelling 153, 51–68.
Borcard, D., Legendre, P., Drapeau, P., 1992. Partialling out the spatial component of ecological variation. Ecology 73, 1045–1055.
Borcard, D., Gillet, F., Legendre, P., 2011. Numerical ecology with R. Use R! series. New York: Springer Science.
Borcard, D., Gillet, F., Legendre, P., 2014. Numerical ecology with R, Chinese edition (translation: J. Lai, Institute of Botany, Chinese Academy of Sciences). Beijing: Higher Education Press.
Borcard, D., Gillet, F., Legendre, P., 2018. Numerical ecology with R. Use R! series, 2nd edn. New York: Springer Science.
De Cáceres, M., Legendre, P., 2009. Associations between species and groups of sites: Indices and statistical inference. Ecology 90, 3566–3574.
De Cáceres, M., Legendre, P., Moretti, M., 2010. Improving indicator species analysis by combining groups of sites. Oikos 119, 1674–1684.
Dolédec, S., Chessel, D., 1994. Co-inertia analysis: An alternative method for studying species environment relationships. Freshwater Biology 31, 277–294.
Dolédec, S., Chessel, D., ter Braak, C.J.F., Champely, S., 1996. Matching species traits to environmental variables: A new three-table ordination method. Environmental and Ecological Statistics 3, 143–166.
Dray, S., Legendre, P., 2008. Testing the species traits-environment relationships: The fourth-corner problem revisited. Ecology 89, 3400–3412.
Dray, S., Legendre, P., Peres-Neto, P.R., 2006. Spatial modelling: A comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecological Modelling 196, 483–493.
Dray, S., Choler, P., Dolédec, S., Peres-Neto, P.R., Thuillier, W., Pavoine, S., ter Braak, C.J.F., 2014. Combining the fourth-corner and the RLQ methods for assessing trait responses to environmental variation. Ecology 95, 14–21.
Dufrêne, M., Legendre, P., 1997. Species assemblages and indicator species: The need for a flexible asymmetrical approach. Ecological Monographs 67, 345–366.
Goodall, D.W., 1954. Objective methods for the classification of vegetation. III. An essay in the use of factor analysis. Australian Journal of Botany 2, 304–324.
Gordon, A.D., Birks, H.J.B., 1972. Numerical methods in quaternary palaeoecology. I. Zonation of pollen diagrams. New Phytologist 71, 961–979.
Gordon, A.D., Birks, H.J.B., 1974. Numerical methods in quaternary palaeoecology. II. Comparison of pollen diagrams. New Phytologist 73, 221–249.
Gower, J.C., 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53, 325–338.
Gower, J.C., Legendre, P., 1986. Metric and Euclidean properties of dissimilarity coefficients. Journal of Classification 3, 5–48.
Green, R.H., 1979. Sampling design and statistical methods for environmental biologists. New York: John Wiley & Sons.
Guénard, G., Legendre, P., 2018. Bringing multivariate support to multiscale codependence analysis: Assessing the drivers of community structure across spatial scales. In: Methods in Ecology and Evolution, 9. pp. 292–304.
Guénard, G., Legendre, P., Boisclair, D., Bilodeau, M., 2010. Multiscale codependence analysis: An integrated approach to analyze relationships across scales. Ecology 91, 2952–2964.
Hui, F.K.C., 2017. Boral: Bayesian ordination and regression analysis. R package version 1.3.1. https://CRAN.R-project.org/package=boral.
Jaccard, P., 1900. Contribution au problème de l'immigration post-glaciaire de la flore alpine. Bulletin de la Societe Vaudoise des Sciences Naturelles 36, 87–130.
Lefkovitch, L.P., 1978. Cluster generation and grouping using mathematical programming. Mathematical Biosciences 41, 91–110.
Legendre, L., Legendre, P., 1979. Écologie numérique. Tome 1: Le traitement multiple des données écologiques. Tome 2: La structure des données écologiques. Paris: Masson and Québec: Presses de l'Université du Québec.
Legendre, L., Legendre, P., 1984a. Écologie numérique, deuxième édition. Tome 1: Le traitement multiple des données écologiques. Tome 2: La structure des données écologiques. Paris: Masson and Québec: Presses de l'Université du Québec.
Legendre, L., Legendre, P., 1983. Numerical ecology. In: Developments in environmental modelling, vol. 3. Amsterdam: Elsevier scientific Publ. Co.
Legendre, P., 1993. Spatial autocorrelation: Trouble or new paradigm? Ecology 74, 1659–1673.
Legendre, P., 2005. Species associations: The Kendall coefficient of concordance revisited. Journal of Agricultural, Biological, and Environmental Statistics 10, 226–245.
Legendre, P., 2014. Interpreting the replacement and richness difference components of beta diversity. Global Ecology and Biogeography 23, 1324–1334.
Legendre, P., Anderson, M.J., 1999. Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. Ecological Monographs 69, 1–24.
Legendre, P., De Cáceres, M., 2013. Beta diversity as the variance of community data: Dissimilarity coefficients and partitioning. Ecology Letters 16, 951–963.
Legendre, P., Fortin, M.-J., 1989. Spatial pattern and ecological analysis. Vegetatio 80, 107–138.
Legendre, P., Fortin, M.-J., 2010. Comparison of the mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. Molecular Ecology Resources 10, 831–844.
Legendre, P., Gallagher, E.D., 2001. Ecologically meaningful transformations for ordination of species data. Oecologia 129, 271–280.
Legendre, P., Gauthier, O., 2014. Statistical methods for temporal and space-time analysis of community composition data. Proceedings of the Royal Society B 281.20132728
Legendre, P., Legendre, L., 1998. Numerical ecology. In: Developments in environmental modelling, 2nd English edn, vol. 20. Amsterdam: Elsevier Science BV.
Legendre, P., Legendre, L., 2012. Numerical ecology. In: Developments in environmental modelling, 3rd English edn, vol. 24. Amsterdam: Elsevier Science BV.
Legendre, P., Legendre, L. (Eds.), 1987. Developments in numerical ecology,NATO ASI series, vol. G-14. Berlin: Springer-Verlag.
Legendre, P., Legendre, V., 1984b. Postglacial dispersal of freshwater fishes in the Québec peninsula. Canadian Journal of Fisheries and Aquatic Sciences 41, 1781–1802.
Legendre, P., Salvat, B., 2015. Thirty-year recovery of mollusc communities after nuclear experimentations on Fangataufa atoll (Tuamotu, French Polynesia). Proceedings of the Royal Society B 282.20150750
Legendre, P., Galzin, R., Harmelin-Vivien, M.L., 1997. Relating behavior to habitat: Solutions to the fourth-corner problem. Ecology 78, 547–562.
Legendre, P., Borcard, D., Peres-Neto, P.R., 2005. Analyzing beta diversity: Partitioning the spatial variation of community composition data. Ecological Monographs 75, 435–450.
Legendre, P., De Cáceres, M., Borcard, D., 2010. Community surveys through space and time: Testing the space-time interaction in the absence of replication. Ecology 91, 262–272.
Legendre, P., Oksanen, J., ter Braak, C.J.F., 2011. Testing the significance of canonical axes in redundancy analysis. Methods in Ecology and Evolution 2, 269–277.
Legendre, P., Fortin, M.-J., Borcard, D., 2015. Should the mantel test be used in spatial analysis? Methods in Ecology and Evolution 6, 1239–1247.
Levin, S.A., 1992. The problem of pattern and scale in ecology. Ecology 73, 1943–1967.
Orlóci, L., 1975. Multivariate analysis in vegetation research. The Hague: Dr. W. Junk B. V.
Pielou, E.C., 1969. An introduction to mathematical ecology. New York: John Wiley & Sons.

Rao, C.R., 1964. The use and interpretation of principal component analysis in applied research. Sankhyā: The Indian Journal of Statistics, Series A 26, 329–358.
ter Braak, C.J.F., 1986. Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. Ecology 67, 1167–1179.
ter Braak, C.J.F., 1987a. The analysis of vegetation-environment relationships by canonical correspondence analysis. Vegetatio 69, 69–77.
ter Braak, C.J.F., 1987b. Ordination. In: Jongman, R.H.G., Ter Braak, C.J.F., van Tongeren, O.F.R. (Eds.), Data analysis in community and landscape ecology. Wageningen, The Netherlands: Pudoc, pp. 91–173. Reissued in 1995 by Cambridge, England: Cambridge University Press.
ter Braak, C.J.F., 1988. CANOCO—An extension of DECORANA to analyze species-environment relationships. Vegetatio 75, 159–160.
ter Braak, C.J.F., Šmilauer, P., 2002. CANOCO reference manual and CanoDraw for windows user's guide—Software for canonical community ordination (version 4.5). Ithaca: Microcomputer Power.
van den Wollenberg, A.L., 1977. Redundancy analysis. An alternative for canonical correlation analysis. Psychometrika 42, 207–219.
Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C., Hui, F.K.C., 2015. So many variables: Joint modeling in community ecology. Trends in Ecology & Evolution 30, 766–779.
Whittaker, R.H., 1969. New concepts of kingdoms or organisms. Science 163, 150–160.
Whittaker, R.H., 1972. Evolution and measurement of species diversity. Taxon 21, 213–251.

Further Reading

Sneath, P.H.A., Sokal, R.R., 1973. Numerical taxonomy—The principles and practice of numerical classification. San Francisco: W. H. Freeman.
Sokal, R.R., Sneath, P.H.A., 1963. Principles of numerical taxonomy. San Francisco: W. H. Freeman.

Relevant Websites

Numerical ecology, n.d., www.numericalecology.com—Legendre Numerical ecology page.
Numerical Ecology with R (NEwR) books, n.d., http://adn.biol.umontreal.ca/∼numericalecology/numecolR/—NEwR book.
Numerical taxonomy, n.d., https://en.wikipedia.org/wiki/Numerical_taxonomy—Numerical taxonomy.
Paul Jaccard, n.d., https://en.wikipedia.org/wiki/Paul_Jaccard—Paul Jaccard.
David Goodall, n.d., https://en.wikipedia.org/wiki/David_W._Goodall—David Goodall.
John C. Gower, n.d., http://onlinelibrary.wiley.com/doi/10.1111/insr.12094/pdf—A conversation with John C. Gower.
Robert H. Whittaker, n.d., https://en.wikipedia.org/wiki/Robert_Whittaker—Robert H. Whittaker.

Paleoecology☆

H John B Birks, University of Bergen, Bergen, Norway

Glossary

Chronology Obtaining absolute (or relative) dating of events in the archive (such as a fossil or sediment sequence).
Fossil Any preserved remains, impression, or trace of any once-living thing from a past geological age.
Geological time scale A system of chronological dating that relates geological strata to time.
Proxies Parts of an archive that provide the evidence of the biota and the physical environment.
Uniformitarianism An essential assumption and philosophical principle in paleoecology that "the present is the key to the past." That is processes that took place in the geologic past are the same as the ones that are observed today.

Research Approaches

Just as there are many approaches to ecological research (e.g., population, community, landscape, ecosystem, global ecology; descriptive, deductive, experimental ecology), there are several approaches to paleoecological research, resulting in several types of paleoecology. One division is based on the biological scales of study ranging from the paleoecology of individuals (adaptation, evolution) to population, community, landscape, ecosystem, and global paleoecology. Other types of paleoecological research can be taxonomically based or habitat based as individual paleoecologists often study particular organisms (e.g., vertebrates, insects, and diatoms) or work in a particular habitat (e.g., wetlands, lakes, oceans, and deserts). The major division within paleoecology, however, concerns timescales and time periods.

Ecologists are primarily interested in timescales of hours, days, weeks, months, years, or decades (so-called real time as these timescales lie within the realm of direct human experience and observation). In contrast, paleo-ecologists are interested in timescales of hundreds, thousands, or millions of years. The major division within paleoecology is between deep-time paleoecology and Quaternary-time paleoecology.

Deep-time researchers use fossils from preQuaternary sediments to study the distribution, evolution, and dynamics of past biota over timescales from thousands to millions of years. Research emphases are on adaptation, evolution, extinction, and biogeography. Quaternary-time (Q-time) researchers use techniques from paleontology, sediment geology, geochemistry, and isotope analysis to reconstruct past biota and environments and to study biotic responses to environmental change at Quaternary timescales (decades, hundreds, or thousands of years) during the last 2 million years of Earth's history. In practice, much of Q-time paleoecology is centered on the past 50,000 years, the period over which radiocarbon dating can be used to provide a chronology. The Quaternary has witnessed many major climatic oscillations between temperate interglacial and cool glacial stages. It has also witnessed the evolution, cultural diversification, and global spread of humans. The last few hundred years have seen the increasing role of humans in altering Earth's biota and environment, resulting in the so-called Anthropocene in which we live today where there is detectable human impact on the atmospheric composition and climate.

Deep-time biologists are usually called paleontologists or paleobiologists, whereas Quaternary biologists are often called paleoecologists.

A real-time ecologist may ask why study paleoecology and what can paleoecologists contribute to contemporary ecology? There are many reasons for studying paleoecology and integrating paleoecology into ecology. These are given as follows:

1. Present-day ecology benefits from a long-term perspective. Paleoecology provides direct evidence for ecological dynamics over long timescales that supplements ecological observations and tests ecological theories about succession and community dynamics.
2. Paleoecology can provide valuable insights into ecological legacies from human activity and environmental change. Ecological legacies are properties of an ecological system today that can only be explained by events or conditions absent from the system today.
3. Paleoecology provides a long-term context for current ecological phenomena and landscape patterns. Many ecological processes occur over decades to millennia (e.g., succession, migration, and pedogenesis). A long temporal perspective is essential to understand factors that determine the rates and causal mechanisms of these processes.
4. Reconstructing past environments is important to evaluate the extent of natural environmental variability, to place current environmental changes, particularly climate, into a long-term context, and to detect if current changes lie within the range of natural variability.

☆*Change History*: March 2018. H R Pethybridge included glossary, keywords, and updated references.

This is an update of H.J.B. Birks, Paleoecology, In Reference Module in Earth Systems and Environmental Sciences, Elsevier, 2013.

 doi:10.1016/B978-0-12-409548-9.00884-8

5. Understanding past climatic change and studying the response of organisms to those changes can contribute to predicting biotic changes in the future.

Philosophy of Paleoecology

Paleoecology is primarily a descriptive historical science involving inductive reasoning and research approaches and techniques drawn from the earth and biological sciences. Its language is therefore derived from both sciences. As fossils are central to paleoecology, careful identification, sound taxonomy, and unambiguous nomenclature are essential.

The method of multiple working hypotheses, presented by Thomas Chamberlain in the mid-nineteenth century, is essential in paleoecology as several explanations are often possible for an observed biotic change. The principle of simplicity (Occam's razor), proposed by William of Occam (1280–1349) is also essential. It proposes that given a set of competing explanations, all of which offer an adequate explanation for a given data set, the simplest explanation is preferable.

An essential assumption and philosophical principle in paleoecology is uniformitarianism, namely "the present is the key to the past"." Since James Hutton (Fig. 1) in the late eighteenth century and Charles Lyell (Fig. 1) in the 19th century, earth scientists have debated this assumption. Stephen Jay Gould resolved the debate by emphasizing the fundamental distinction between substantive uniformitarianism where rates of geological processes are thought to be constant in time and methodological uniformitarianism or actualism where the nature of the processes and their underlying laws are assumed to be the same through time but the rates may be very different at different times. Catastrophes (e.g., floods and volcanic eruptions) do occur so the rates of change can vary greatly but they all follow the basic laws of nature because the properties of matter and energy are invariant with time. Methodological uniformitarianism is an untestable methodological assumption common to all sciences. It represents the simplest approach to paleoecology and is thus an application of the principles of simplicity and induction.

Paleoecological Evidence

The geological record for Q-time and deep-time paleoecology can be rich in biological and environmental information.

Biological information comes from fossils preserved in sediments. The most common types of fossils in Quaternary paleoecology have much original material preserved, such as "hard parts" of shells, insect exoskeletons, diatom frustules, leaf cuticles, bones, pollen, seeds, and wood. Other types of fossils include impressions and films, petrifications and replacements, molds and casts, and trace fossils. These can provide valuable biological evidence in deep-time paleoecology.

Preservation of biotic remains usually requires deposition in anaerobic environments such as lake bottoms, ocean floors, or wetlands, or more rarely, freezing or desiccation. Plant microfossils (e.g., pollen) and macrofossils (e.g., seeds, leaves, wood) provide information about past flora and vegetation. Remains of animals (e.g., ostracods, beetles, cladocerans, testate amebae, and vertebrates) give insights into the past fauna.

In Q-time paleoecology, fossils are identified, as far as possible, by comparison with living taxa and are given names of the living taxa that the fossils most closely resemble. In deep-time paleoecology, the fossils are regarded as representing extinct taxa and taxonomies are developed for particular fossil groups based entirely on fossil remains.

Environmental information (e.g., climate, lake levels, and water temperatures) can sometimes be obtained from the sediments in which fossils are preserved. Such information is obtained by physical, inorganic and organic chemical, biogeochemical, and stable-

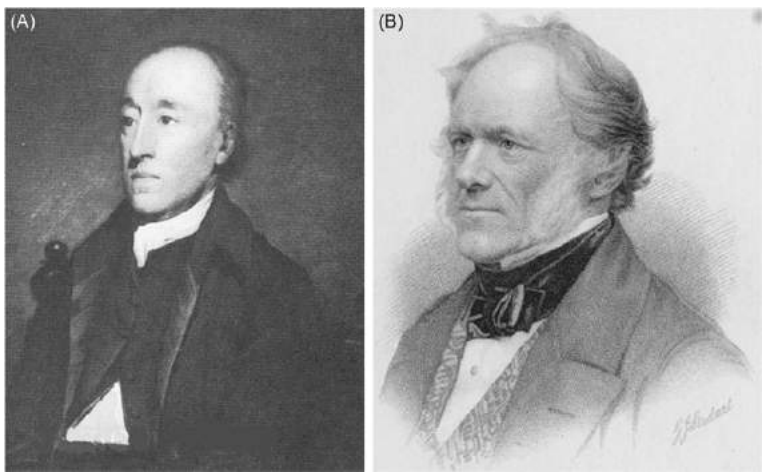


Fig. 1 (A) James Hutton (1726–97) and (B) Charles Lyell (1797–1875), the founders of modern earth science and of the concept of uniformitarianism, "the present is the key to the past," the basic principle of paleoecology. (A) From http://www.science.siu.edu/geology/people/ferre/hutton.html (accessed January 2008). (B) From http://kentsimmons.uwinnipeg.ca/16cm05/1116/16evolut.htm (accessed January 2008).

isotope analyses of sediments. Components of the sediments such as charcoal provide direct evidence for the occurrence of fires. Environmental information is often derived indirectly from biological evidence. For example, changes in lake-water pH can be inferred from changes in the composition of fossil diatom assemblages, under the assumption that the ecological preferences of the modern taxa are the same as they were in the past. Indirect sources of environmental information are called environmental "proxies."

A sound chronology is essential in paleoecological investigations to determine timings of past events, estimate rates of change, and permit correlations. Various dating techniques provide chronologies. In Q-time research, radiocarbon dating and other radiometric techniques (e.g., ^{210}Pb-dating) are the major geochronological tools. Other techniques include tree rings, annually laminated sediments, volcanic-ash layers, and paleomagnetics.

Paleoecological data are frequently quantitative and consist of many variables (e.g., diatom taxa and chemical elements) and many samples. Numerical methods such as principal components analysis are valuable tools in summarizing major patterns of variation in complex, multivariate paleoecological data. Statistical techniques (e.g., regression) are important in testing hypotheses about possible causes for observed changes in the paleo-ecological record.

Selected examples of different types of Quaternary paleoecological evidence are listed in **Table 1**, along with the type of paleoecological inferences possible from such evidence. If the inferences are quantitative, this is indicated along with the applicability of these types of evidence in deep-time research. Evidence types unique to deep-time studies are not given.

The major techniques in Quaternary paleoecology are pollen analysis, paleolimnology (e.g., analyses of diatoms, chironomids, and ostracods), paleoceanography (e.g., analyses of foraminifers, radiolarians, and cocco-lithophorids), tree-ring analysis, paleoentomology, peat stratigraphy, plant–macrofossil analysis, inorganic and organic geochemistry, and geochronological dating. Important reference works discussing many of these techniques are given in **Table 2**.

Table 1 Selected examples of Quaternary paleoecological evidence and their paleoecological inferences

Evidence	D	Q	Inference
1. Individual fossils			
Number and variation in growth rings(e.g., tree rings)	+	+	Age, seasonality, growth rates, growth conditions
Stomatal density in leaves	+	+	Atmospheric CO_2
Body size	+	+	Size, biomass, size frequency distribution
Density	+	+	Biomass, spatial patterns
Geographical location	+	+	Geographical distribution, habitat, dispersal, climate
Tooth wear	+		Diet in mammals
Fossil DNA			Species presence, evolutionary relationships and rates
2. Fossil assemblages			
Taxonomic composition	+	+	Taxa present, environmental conditions
Taxonomic richness, evenness, and diversity	+	+	Taxonomic richness, evenness, diversity
Relative abundance of taxa	+	+	Abundance frequency distribution, dominance, rarity
Geographical location	+	+	Geographical distribution, climate
Presence/absence and relative abundance of plant functional types	+	+	Vegetation structure and ecological properties (e.g., albedo)
3. Geochemical and isotopic composition of fossil hard parts			
Oxygen isotopes	+	+	Temperature, ice volume, salinity, moisture sources
Carbon isotopes	+	+	Food sources, vegetation type, productivity, atmospheric CO_2, moisture, temperature
Nitrogen isotopes	+	+	Food source, trophic level
Hydrogen isotopes	+	+	Temperature, moisture source
Mg/Ca ratios	+	+	Temperature
Sr/Ca ratios	+	+	Temperature range
4. Geochemical evidence from sediments			
Heavy metals and organic toxins		+	Human disturbance, atmospheric contamination
Mn, Fe, Mo, Cr	+	+	Redox conditions
Sulfur isotopes	+	+	Redox conditions, microbial activity
K, Ca, Mg		+	Erosion
5. Biogeochemical evidence from sediments			
Biomarkers and other molecular markers	+		Species presence, trophic structure
Photosynthetic and other pigments	+	+	Productivity, trophic status
N and C isotopes in organic compounds	+	+	Productivity, organic sources
Organic compounds (e.g., alkenones)	+	+	Temperature, organic sources
6. Lithological evidence			
Sediment grain-size	+	+	Sedimentary environment (e.g., glacial, aolian, alluvial)
Peat and coal stratigraphy	+	+	Wetland hydrology, moisture changes
Charcoal	+		Fires

Plus signs in the column labeled D, Deep-time; and Q, quantitative indicate that this evidence is also applicable in Deep-time research and that the paleoecological inferences are quantitative.

Table 2 Major reference sources for paleoecological techniques

Techniques	*Sources*
Coring and chronological techniques	Last, W. M. and Smol, J. P. (eds.) (2001). Tracking environmental change using lake sediments. Vol. 1: Basin analysis, coring, and chronological techniques, Dordrecht: Kluwer
Physical and geochemical methods	Last, W. M. and Smol, J. P. (eds.) (2001). Tracking environmental change using lake sediments. Vol. 2: Physical and geochemical methods, Dordrecht: Kluwer
Biological fossils from terrestrial environments, algae, and other siliceous fossils	Smol, J. P., Birks, H. J. B., and Last, W. M. (eds.) (2001). Tracking environmental change using lake sediments. Vol. 3: Terrestrial, algal, and siliceous indicators. Using lake sediments. Vol. 3: Terrestrial, algal, and siliceous indicators. Dordrecht: Kluwer
Zoological fossils	Smol, J. P., Birks, H. J. B. and Last, W. M. (eds.) (2001). Tracking environmental change using lake sediments, Vol. 4: Zoological indicators. Dordrecht: Kluwer
Plant micro- and macrofossils	Jones, T. P. and Rowe, N. P. (eds.) (1999) Fossil plants and spores. London: The Geological Society
Numerical methods	Birks, H. J. B. and Gordon, A. D. (1985). Numerical methods in quaternary pollen analysis. London: Wiley Birks, H. J. B. (1998) Numerical tools in palaeolimnology—progress, potentialities, and problems. *Journal of Paleolimnology* 20, 301–332

Paleoecology can be studied in any deposits containing fossils that provide evidence for past life. In Quaternary paleoecology, especially the last 11,500 years of the Holocene (postglacial) and the recent past (last 100–250 years), lakes and their sediments provide the most diverse records of past biota and environment because lake sediments integrate a range of regional and local biological and environmental signals. Lake sediments consist of material from several sources. The material can be divided into allochthonous (derived from outside the lake from the catchment or atmosphere) and autochthonous (derived from the lake itself) material.

Autochthonous material includes fossils of limnic organisms such as diatoms, chrysophytes, other algae and their pigments, bacteria, aquatic macrophytes, cladocerans, ostracods, chironomids, fish, etc. Input of allochthonous material to a lake has three main sources—groundwater, catchment or watershed, and atmospheric. Groundwater inputs include solutes (e.g., Ca, Mg, Na, K, Cl, SO_4, and HCO_3), nutrients (e.g., P and N), and toxins (e.g., pesticides). Catchment inputs include fossils of terrestrial biota (e.g., seeds, leaves, charcoal, and insects), inorganic material, nutrients, organic detritus, and toxins. Atmospheric inputs include radioactive nuclides (^{210}Pb, ^{137}Cs, ^{241}Am, ^{14}C, etc.), charcoal, spheroidal carbonaceous particles, pollen, trace metals (e.g., Pb, Cu), volcanic ash, toxins (e.g., persistent organic pollutants), sulfates, nitrates, and dust.

Stages in a Paleoecological Study

Although each paleoecological study is unique, depending on the research problems, geographical area of study, site type, expertise of the paleoecologists, age of the sediments, etc., there are several stages that are common to many, if not all, Quaternary paleoecological studies.

1. *Definition of research problem*. Careful definition of the research problem and hypotheses to be tested is important at the outset, as paleoecology is a very labor-intensive and time-consuming activity. A poorly designed project results in a considerable waste of time and effort.
2. *Selection of site to be sampled*. Careful site selection is essential if the research questions are to be answered. Site selection requires not only knowledge of the study area and its geology, topography, and hydrology but also knowledge of the ecology and land use of the possible sites to be selected (**Fig. 2**). Exploratory studies are invaluable.
3. *Selection of coring site*. Once a site has been selected, the next stage is to select where to take a core of the sediments for paleoecological study. In general, the aim is to maximize between-site variability and hence to minimize within-site variability. Experience has shown that the deepest point in a basin is often the place where within-site variations are minimal and is thus the preferred place for sediment coring. Again exploratory studies are invaluable.
4. *Collection of sediment cores*. There are several different types of coring devices. The choice depends on many factors, including water depth, nature of the sediments to be sampled, amount of sediment needed for study (pollen analysis only needs 1 cm^3 of sediment, whereas paleoentomology requires large volumes), temporal resolution required, and remoteness of the study site. Details of suitable coring devices can be found in the works given in **Table 2**. The most common type is illustrated in **Fig. 3**, along with a sediment core.
5. *Sampling and describing sediments*. This involves describing the sediments in terms of physical properties (e.g., color, stratification, and water content), humification (degree of decay), and composition (clay, silt, sand, organic detritus, mud, etc.). The subsampling resolution depends on the research questions and the rate of sediment accumulation. In some arctic lakes where the last 10,000 years may be represented by only 1 m of sediment, sampling at 10 years intervals requires a sampling resolution

Fig. 2 Lille Kjelavatn, a small lake at 1000 m in southern Norway. This is an ideal site for a paleoecological study as the lake is small and relatively deep ensuring that the pollen record is from the catchment (watershed) and the sediments are undisturbed; the surrounding slopes are gentle and there is no sedge-swamp around the lake ensuring that terrestrial plant macrofossils are washed into the lake; and the bedrock is acid, thereby minimizing errors in radiocarbon dating.

Fig. 3 (A) Coring lake sediments using a Livingstone piston corer at Haugtjern, southeast Norway and (B) a 1 m long, 2.5 cm diameter core of lake sediment extruded from a Livingstone piston corer at Mangrove Lake, Bermuda.

of every 1 mm. Such sampling must be done in the laboratory using specially constructed equipment. Time spent on careful sampling is time well spent, because contaminated samples are worthless.

6. *Dating*. An absolute chronology is essential in almost all Quaternary studies. For the last 150 years, radio-metric-dating techniques involving ^{210}Pb, ^{137}Cs, and ^{241}Am are invaluable. For the last 15,000 years, radio'carbon dating is the major chronological tool.
7. *Collecting paleoecological data*. A very wide range of data types can be collected, including physical, chemical, and biological data (see **Table 1**).
8. *Presentation of paleoecological data*. The resulting data may be complex and it is a challenge to present the results of several paleoecological analyses in a clear and effective way. Much thought is required and many critical questions need to be considered—should all variables be presented when there may be 200 diatom taxa present, should the data be presented as relative percentages or "absolute" accumulation rates, should the data be plotted on a sediment age scale or on an estimated age scale? Numerical techniques can be valuable in summarizing the patterns of variation within and between different data sets from the same sediment sequence.

9. *Interpretation*. There are two major approaches to the interpretation of paleoecological data: (1) paleoecological reconstructions, a primarily descriptive approach where the emphasis is on reconstructing past biota, populations, communities, landscapes, and ecosystems; (2) ecological paleoecology, a hypothesis-testing approach where the emphasis is on interpreting the observed changes in terms of underlying causes.

Paleoecological Reconstructions

In this approach, paleoecologists use the available evidence for reconstruction purposes at a range of ecological scales.

1. Past biota—what taxa were present in the past?
2. Past populations—what were the population sizes in the past?
3. Past communities—what communities or "life assemblages" were present in the past?
4. Past landscapes—what was the past landscape and how did it vary in space and time?
5. Past environments—what was the environment (e.g., climate and lake-water pH) at particular times in the past?
6. Past ecosystems—what was the ecosystem at particular times in the past?

Reconstructions can be based on a few "indicator species" or on the fossil assemblage as a whole. The reconstructions may be qualitative or quantitative. An important approach for quantitative environmental reconstructions in Quaternary paleoecology involves modern organism–environment transfer or calibration functions to transform fossil assemblages into estimates of the past environment (Fig. 4).

Transfer functions are mathematical regression-type models that express the relationship between modern assemblages of organisms (e.g., pollen) preserved in surface sediments (Y_m) and the contemporary environment (e.g., mean July temperature—X_m):

$$X_m = \hat{U}_m Y_m$$

where $\hat{U}_m$ is the modern transfer function estimated by U_m inverse regression or calibration.

The transfer function is assumed to be invariant in time and space and is applied to fossil assemblages, Y_f, to derive estimates of the past environment, X_f:

$$X_f = \hat{U}_m Y_f$$

This general approach has been used to estimate sea-surface temperatures from fossil foraminifers, radiolarians, coccolithophorids, diatoms, and dinoflagellate-cyst assemblages, terrestrial climate from fossil pollen, chironomid, cladoceran, diatom, mollusk, and insect assemblages and from tree-rings, lake conditions (e.g., pH, salinity, total P, and anoxia) from fossil diatom, cladoceran, chironomid, chrysophyte, and ostracod assemblages, bog moisture from fossil moss and testate amebae assemblages, and atmospheric CO_2 from stomatal density of fossil leaves.

As all the biotic evidence is generally used for reconstruction purposes, it is not possible to use the reconstruction as a basis for interpreting the observed biotic changes, except in the rare cases where the environmental reconstruction is based on independent sources of evidence such as stable isotopes, sediment geochemistry, or one biological proxy that is used solely for reconstruction purposes.

Ecological Paleoecology

In this approach, paleoecologists are interested in the causal underlying processes or "forcing functions" for the observed stratigraphical patterns. For example, what factors caused the observed changes in pollen stratigraphy at a site over the last 11,000 years? Are the biotic changes responses to changes in climate, soils, biotic interactions, pathogens, disturbance regimes, land-use, etc.? It is essential to interpret paleoecological data in terms of underlying causal factors if the paleoecological record is to be used as a long-term record of ecological dynamics that can help understand present-day systems.

In causal interpretations, there may be two or more competing hypotheses to explain the observed patterns. At least three independent proxies are needed to test two hypotheses. A major development in Quaternary paleoecology has been multiproxy studies where several proxies are studied on the same sediment core. To test competing hypotheses, one or more proxy is used to reconstruct the past environment. These reconstructions and other independent paleoenvironmental variables are then used as predictor variables to test hypotheses about the causes of change in the other proxies when they are considered as response variables. A range of statistical regression-modeling techniques can be used to test different hypotheses about the causes of the observed changes in the response variable in relation to the predictor variables. Statistical significance can be assessed by permutation tests that take into account the numerical properties of time-ordered paleoecological data.

Contributions of Paleoecology to Ecology

Paleoecological research has in the last 30 years made many important contributions to our understanding of present-day ecological systems. Some examples are summarized in Table 3 in terms of the ecological questions considered and the paleoecological techniques employed. Further examples can be found in the "Further reading" section.

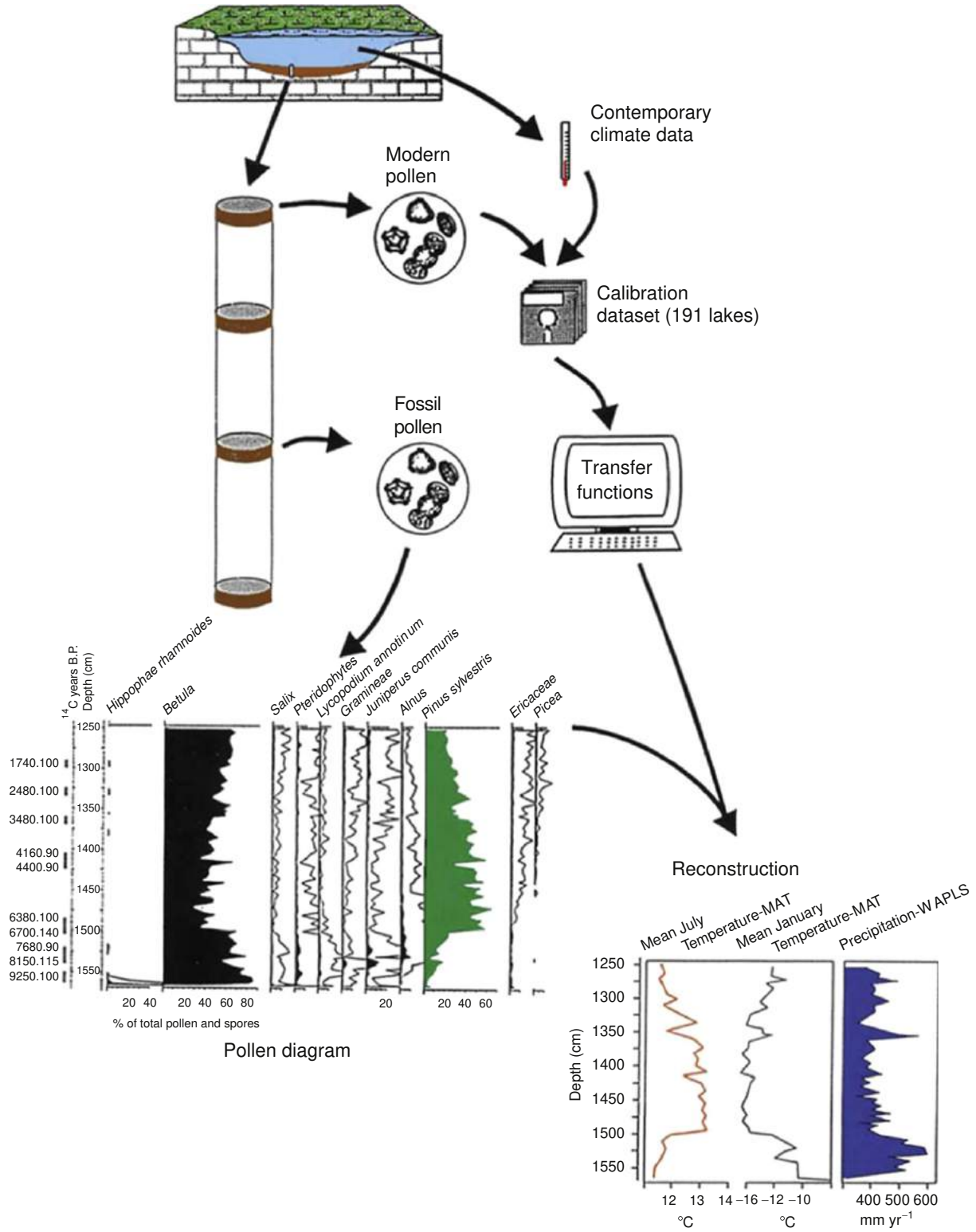


Fig. 4 A schematic representation of the stages involved in deriving a quantitative reconstruction of past environment from pollen-stratigraphical data using a modern calibration training set and transfer or calibration functions. Based on an unpublished diagram by Steve Juggins.

One of the major contributions (Table 3) that paleoecology has made to our understanding of present-day ecosystems concerns the cause of recent surface-water acidification in Europe and North America. Competing hypotheses for recent pH decreases are given as follows:

Table 3 Examples of Quaternary paleoecological studies that have contributed to modern ecological understandings

Ecological question	*Paleoecological techniques used*
What is the cause of recent surface-water acidification?	Paleolimnology (especially diatoms), transfer functions, chronology
What role does a lake's catchment play in lake development?	Paleolimnology, pollen analysis, plant macrofossils, statistics and hypothesis testing, chronology
What is the origin of landscape mosaics (e.g., forest and grassland)?	Pollen analysis of pond sediments, statistics, chronology
What is the role of human activity in landscape change and species extinction?	Vertebrates, pollen analysis, plant macrofossils, sediment geochemistry and magnetism, chronology
What is the role of soil development in long-term vegetation dynamics?	Pollen analysis, sediment inorganic geochemistry, chronology
Is there an orderly and predictable succession of community types in hydroseral successions?	Sediment composition, pollen and macrofossil analysis, chronology
Does the "regeneration complex" hypothesis explain the mechanisms of bog growth and development of pools and hummocks?	Peat composition, plant macrofossils, chronology
Has the frequency and extent of wild-fires changed in time?	Charcoal, pollen, sediment geochemical analyses, statistics, chronology
Have pathogens been important in influencing long-term forest dynamics?	Pollen analysis, statistics, chronology
Have forest trees spread since the last glacial stage as assemblages or individualistically?	Pollen analysis, chronology, mapping

1. the recent changes result from land-use, vegetational, and soil changes (e.g., land dereliction, secondary forest succession, and soil deterioration);
2. the recent changes are the result of natural long-term vegetational and soil acidification that has occurred over the last 10,000–11,000 years; and
3. the recent changes are a result of atmospheric deposition of strong acids following the combustion of fossil fuels, the so-called "acid-rain" hypothesis.

As a result of international research programs in Europe and North America in the 1980s and 1990s, detailed paleoecological studies of lake sediments showed that the recent land-use and natural long-term acidification hypotheses could be falsified, whereas despite several attempts to reject the acid-rain hypothesis it could not be falsified. The paleoecological studies (**Fig. 5**) primarily involved fossil algal (diatom and chrysophyte) assemblages preserved in lake sediments, sediment inorganic geochemistry, counting of spheroidal carbonaceous particles formed by the combustion of fossil fuels, ^{210}Pb-dating, transfer functions linking modern algal assemblages to lake-water pH, and fine-resolution sampling (about every 5–10 years for the last 250 years).

Paleoecology, particularly for the last 100 years, is making major contributions to assessing the "health" of ecosystems in terms of contamination by heavy metals (e.g., Cu, Cd, Zn, Pb, Ni, and As) and persistent organic pollutants (e.g., polycyclic aromatic hydrocarbons), of eutrophication by nutrients (e.g., N and P), and of recovery following decreases in the atmospheric deposition of "acid rain."

A current research activity for many paleoecologists is detecting impacts of recent climate change on biological systems. A synthesis of paleoecological data for Arctic lakes has shown that major changes in biotic composition have occurred in the last 150 years in over 80% of the lakes examined above the Arctic Circle, whereas further south, less than 60% of the lakes showed any major changes. All available evidence suggests that arctic lakes have changed dramatically and directionally within the last 150 years. "Regime-shifts" have occurred rapidly, characterized by changes to taxonomically more diverse and increasingly productive aquatic ecosystems, with more complex community and trophic structures and enhanced plankton development.

A new application of paleoecology, conservation paleoecology, is developing where results of paleoecological studies are used to provide historical perspectives relevant to nature conservation and ecosystem management. These perspectives include insights into biotic responses to environmental change, rates and mechanisms of biological invasions, extent of naturalness in ecological systems, and the frequency of disturbance, especially wild-fires. It is in nature management in relation to fire that conservation paleoecology is making major contributions. It is essential in management to know the natural variability of wild-fires so that this can be used as a reference against which contemporary conditions and future alternatives can be evaluated. Such assessments are often based on short-term records (<50 years) only. Although climate change and human activity are recognized as the major drivers of fire frequency and extent, paleoecological studies show that these relationships are often surprisingly complex. Results from studies involving pollen, plant macrofossils, charcoal, sediment geochemical, and time-series analyses show that warmer/drier conditions do not necessarily result in a high fire frequency. Fires often occur more frequently under wetter conditions, because in moist intervals biomass production and hence fuel loads increase. Fire frequency and extent thus oscillate with climate through time. Paleoecological data suggest that in a given region, there may be several possible ecosystem or "regime" states corresponding to different fire frequencies which, in turn, are a function of climate–fuel relationships. Such insights can be used to define the context of current forest conditions and potential future changes.

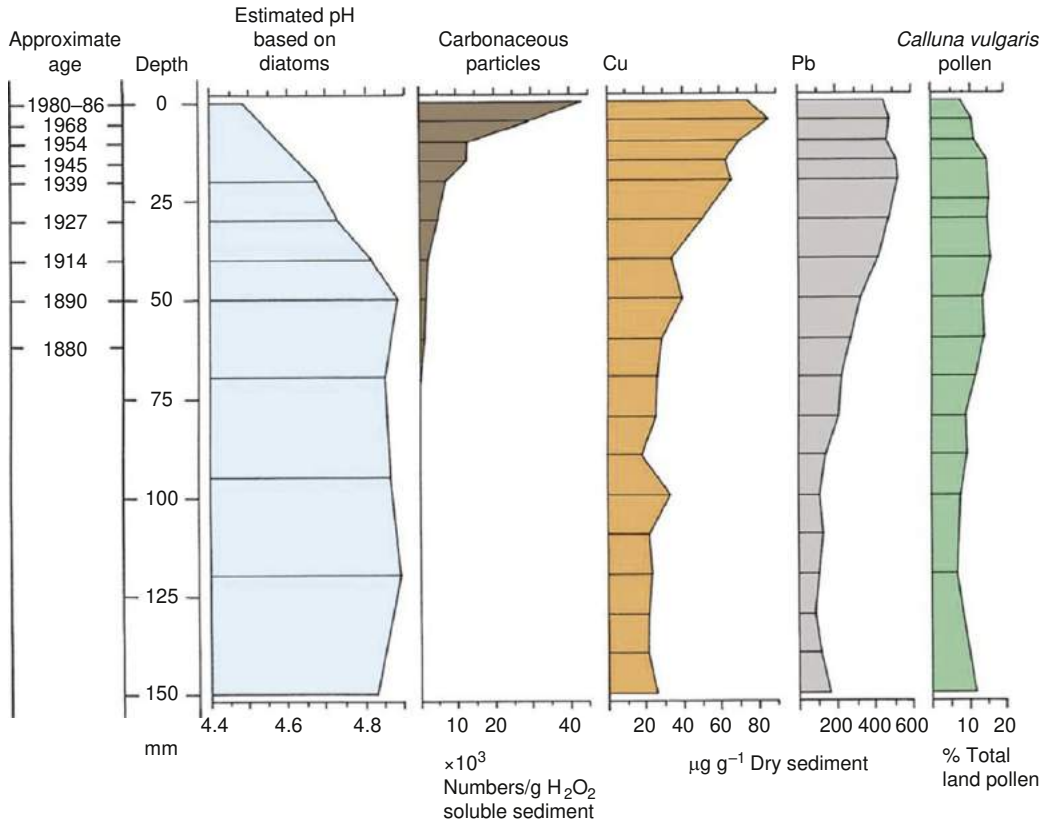


Fig. 5 A stratigraphical diagram from Holetjørn, a small hill-top lake in southwest Norway that covers the last 250–300 years. The diagram shows the decrease in lake-water pH beginning at about 1890. This acidification coincides with the beginning of the deposition of carbonaceous particles formed by the high temperature combustion of fossil fuels. The concentrations of the heavy metals Cu and Pb also increase at about 1914, reflecting atmospheric contamination due to industrialization and the use of lead in petrol. There is a small decline in *Calluna vulgaris* (heather, lyng) pollen since 1939, reflecting the increase in upland grazing in recent years. The age scale is based on ^{210}Pb-dating. Modified from an original diagram by H. J. B. Birks.

Future Directions and Potentialities

There have been major advances in Quaternary paleoecology since the pioneering pollen-analytical studies by Lennart von Post (Fig. 6) in Sweden in the early 20th century. After the sophistication of pollen analysis by Johs Iversen, Knut Faegri (Fig. 7), and others in the 1940s–1970s, important advances in paleoecology have come from the developments of paleoceanography in the 1960s–1980s, paleolimnology in the 1980s–1990s, and a wealth of isotope and geochemical techniques in the last decade.

A major future development in paleoecology will come from recent advances in earth sciences. These involve stable-isotope analysis, organic geochemistry, and detection of molecular markers that will allow past environmental conditions to be inferred independent of biological proxies. To date, there have been few paleo-ecological studies that exploit these developments to explore links between independently derived records of environmental change, particularly climate change, and observed biotic changes. When this is done more fully, it will allow paleoecological research to focus directly on the nature of biotic response to environmental changes over a range of temporal and spatial scales, namely to study the ecology of the past. This will be a major break-through, as the geological record of fossils can then be used as an ecological observatory or laboratory for studying long-term ecological changes over timescales beyond direct ecological observations. As the paleoecological record is a unique record of biotic responses to a wide range of environmental changes, it can help predict biotic responses to future environmental change, a major concern of much ecological and conservation research today.

Ecology, conservation, and nature management are primarily concerned with the present and increasingly with the future. Paleoecology considers the past but can provide a historical perspective to the present. With the ever-increasing quantity and quality of paleoecological data at a fine spatial and/or temporal resolution, there is great potential for close interactions between real-time ecologists and Quaternary paleoecologists, as there is an enormous contribution that paleoecology can make to ecology. The paleoecological record contains many "lessons from the past" about biotic responses to environmental change. With recent developments in paleoecology and earth sciences, it will soon be possible to read this record ecologically and to learn from these "lessons" so as to improve our understanding of ecological dynamics over long timescales and to contribute to nature conservation and management policies for the future.

Fig. 6 Lennart von Post (1884–1950), a Swedish geologist who invented pollen analysis, the dominant technique in Quaternary terrestrial paleoecology.

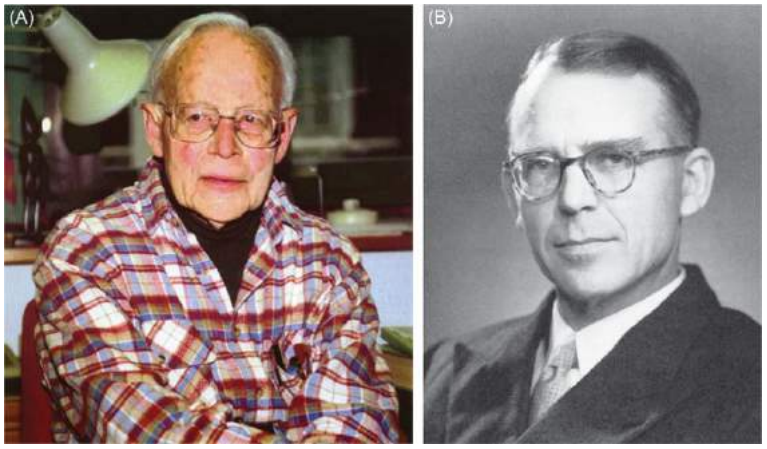


Fig. 7 (A) Knut Faegri (1909–2001), a Norwegian botanist and (B) Johs Iversen (1904–71), a Danish botanist. In the 1950s–1970s, Faegri and Iversen developed pollen analysis into a powerful paleoecological technique that is now being used to resolve critical questions in community, landscape, conservation, ecosystem, and global ecology.

See also: Evolutionary Ecology: Genetic Drift; Macroevolution; Evolutionary Ecology; Metacommunities. General Ecology: History of Ecology

Further Reading

Birks, H.J.B., Birks, H.H., 2004. Quaternary palaeoecology. Caldwell, NJ: Blackburn Press.
Birks, H.J.B., 1995. Quantitative palaeoenvironmental reconstructions. In: Maddy, D., Brew, J.S. (Eds.), Statistical modelling of Quaternary science data Technical Guide5. , pp. 161–254.
Dodd, J.R., Stanton Jr., R.J., 1981. Paleoecology, concepts and applications. New York: Wiley.
Hammar, Ø., Harper, D.A.T., 2006. Paleontological data analysis. Oxford: Blackwell Publishing.
Hu, F.S., Hampe, A., Petit, R.J., 2009. Paleoecology meets genetics: Deciphering past vegetational dynamics. Frontiers in Ecology and the Environment 7 (7), 371–379.

Jackson, J.B.C., Erwin, D.H., 2006. What can we learn about ecology and evolution from the fossil record? Trends in Ecology and Evolution 21, 322–328.
Jackson, S.T., Overpeck, J.T., 2000. Responses of plant populations and communities to environmental changes of the late Quaternary. Paleobiology 26 (supplement), 194–200.
Kohn, M.J., 2010. Carbon isotope compositions of terrestrial C3 plants as indicators of (paleo) ecology and (paleo) climate. Proceedings of the National Academy of Sciences 107 (46), 19691–19695.
McElwain, J.C., Steinthorsdottir, M., 2017. Paleoecology, ploidy, paleoatmospheric composition, and developmental biology: A review of the multiple uses of fossil stomata. Plant Physiology 174 (2), 650–664.
Pandolfi, J.M., 2011. The paleoecology of coral reefs. In: Coral reefs: An ecosystem in transition. Netherlands: Springer, pp. 13–24.
Seddon, A.W., Mackay, A.W., Baker, A.G., Birks, H.J.B., Breman, E., Buck, C.E., Ellis, E.C., Froyd, C.A., Gill, J.L., Gillson, L., Johnson, E.A., 2014. Looking forward through the past: Identification of 50 priority research questions in palaeoecology. Journal of Ecology 102 (1), 256–267.
Smith, C.R., Glover, A.G., Treude, T., Higgs, N.D., Amon, D.J., 2015. Whale-fall ecosystems: Recent insights into ecology, paleoecology, and evolution. Annual Review of Marine Science 7, 571–596.
Smol, J.P., Wolfe, A.P., Birks, H.J.B., Douglas, M.S., Jones, V.J., Korhola, A., Pienitz, R., Rühland, K., Sorvari, S., Antoniades, D., Brooks, S.J., 2005. Climate-driven regime shifts in the biological communities of arctic lakes. Proceedings of the National Academy of Sciences of the United States of America 102 (12), 4397–4402.

Relevant Websites

https://www.journals.elsevier.com/palaeogeography-palaeoclimatology-palaeoecology/
http://www.angelfire.com/az3/mohgameil/paleoecology.html
http://personal.colby.edu/∼ragastal/Pecology.htm

Parasites

KD Lafferty, University of California, Santa Barbara, CA, USA

Introduction

Parasites Are Everywhere

Parasitism is the most popular lifestyle on Earth. Roughly, half of plants and animal species are parasitic at some stage of their life cycle. Few species, if any, lack any parasites and most species have at least one host-specific parasite species. Even parasites have parasites. We know the most about humans, who host 342 parasite species, not including viruses and bacteria. Many of these only parasitize humans. Parasites tend to be more abundant in places where hosts are abundant and more diverse in places where hosts are diverse. For instance, more parasites species are found in the tropics than at high latitudes. While parasite diversity follows host diversity, parasitism may also select for diversity in host communities.

What Is a Parasite?

Parasite comes from the Greek *parasitos*. Originally, *parasitos* referred to a dinner guest. However, around 400 BC, Greek comedies began featuring a stereotyped character: a hard to get rid of guest whose obnoxious nature becomes tiresome. Hereafter, *parasitos* took on the meaning of a freeloader.

Tapeworms, digenetic trematodes, and acanthocephalans are familiar parasitic taxa. Among the animals, there are nine entirely parasitic phyla and 22 predominantly ($>99\%$) predatory phyla. For the remaining 11 phyla that have a mix of parasitic and free-living species (e.g., Nematoda, Platyhelminthes, and Arthropoda), clades within a phylum are often either entirely parasitic or predatory. Nonetheless, parasitism lacks a single evolutionary origin and is spread among phyla.

The field of parasitology traditionally limits its focus to animal groups that one can see clearly with a microscope. These include protozoans (amebas, flagellates, ciliates, apicomplexans, myxozoans, and mesozoans) and metazoans. Of the metazoans, several parasitic groups are familiar: platyhelminthes (monogenes, trematodes, cestodes), nematodes, acanthocephalans, pentastomids, and arthropods (crustaceans and insects). Some of these groups (particularly the dipterans) are best defined as micropredators. Small and nonanimal parasitic taxa (viruses, bacteria, fungi, and plants) are not included in the field of parasitology. These groups are often termed pathogens. For these reasons, an ecological/evolutionary concept of parasitism does not necessarily conform to the taxa treated by the field of parasitology.

Natural enemies of all taxa take nourishment from a victim (host) or victims (prey) using a variety of trophic strategies. The particular strategy an individual natural enemy uses can vary from one victim to the next. How do parasites differ from predators? At first glance, it appears that relative body size is the key factor. Parasites are generally much smaller than their hosts and predators are larger than their prey. But a wolf is not a parasite of the larger moose and body size is not a particularly useful means of distinguishing among trophic strategies, despite the fact that it appears to be a key correlate and evolutionary driver of trophic strategies.

The durability of the enemy–victim interaction has been argued as a useful distinction between predators and parasites (i.e., parasitic interactions are durable). While parasites are generally intimately associated with their hosts, intimacy is difficult to apply as a dichotomous operational definition. There is no obvious durability criterion that would clearly allow us to distinguish predator from parasite. For instance, a lion may stay for several days feeding on a zebra carcass while the nymph of a tick takes one blood meal from a deer and drops off within a few hours.

If intimacy is difficult to apply as a categorical definition of parasitism, what other criteria are available? Population models provide a possible perspective of trophic strategies. In infectious disease models, a parasite exploits only one host during a particular phase of its ontogeny. In contrast, most population models in ecology incorporate a functional response that considers how predators eat many prey. In other words, population modelers have established a well-accepted dichotomy between parasites and predators based on the number of hosts or prey attacked during a particular life-history stage. Attacking one versus more than one victim is easier to apply as a means to distinguish parasites from predators than durability of an attack, though the two views are not in conflict (exploiting only one victim during a life-history stage is necessarily a durable attack).

Micropredators

Mosquitoes are included in parasitology texts, yet because an individual feeds on more than one victim during the adult life stage, adult mosquitoes fail to meet the criterion for a parasite. Instead, they represent a type of predator that does not reduce the fitness of prey to zero. For this reason, `micropredators', like mosquitoes, have similarities with parasites. Many natural enemies are obligate micropredators, others such as vampire bats, lampreys, cookie-cutter sharks, and many herbivores (e.g., deer) switch between micropredation and predation, depending on the relative size of the predator and its prey.

The Diversity of Parasite Life Histories

Whereas the criterion `Does the enemy attack more than one victim?' separates predators from parasites, criteria that help distinguish among parasitic life-history strategies are (1) does the enemy eliminate victim fitness, (2) does the enemy require a victim's death? and (3) does the enemy cause intensity-dependent pathology? Predators can also be subdivided into useful categories by applying the first and third criteria. The next sections describe common types of parasite life histories.

Parasitoids

Parasitoids are a type of parasite that requires their victim's death, thereby reducing host fitness to zero. Entomologists use the term parasitoid to describe wasps and flies that lay eggs in or on insect hosts. The larvae then consume the host from the inside. When the carcass is consumed, the parasitoid wasps or flies metamorphose into free-living adults. For this reason, parasitoids are relatively large parasites. Other taxonomic groups (turbellarians, nematodes, crustaceans) also use a parasitoid life-history strategy. For instance, an intertidal turbellarian infects young crabs, grows to a large size, and, on adulthood, bursts through the crabs' exoskeleton to become a free-living adult worm. Because a diversity of taxonomic groups must kill their hosts as a normal aspect of their development, the term parasitoid is best used as a description of a parasite life-history strategy instead of a taxonomic category for certain flies and wasps.

Microparasites and Macroparasites

Intensity-dependent pathology is an important criterion for distinguishing among parasites, and parasite body size is somewhat correlated with this criterion. The protozoans, bacteria, and viruses have short generation times, rapid reproduction inside the host, a tendency to induce immunity in surviving hosts, and a short duration of infection. Most parasites and pathogens that fit these assumptions are relatively small. The population dynamics of such parasites are well described by `intensity-independent' (SEIR) models. Because such parasites are small, these models have gained the nickname of microparasite models.

The population dynamics of most parasitic worms are not well described by SEIR models. Because pathology increases with the number of worms in an infection (or intensity), and worms typically aggregate among hosts, pathology varies considerably from host to host. To model such worms more appropriately requires accounting for parasite intensity and aggregation, and more specifically keeping track of the number of individual parasites in the parasite population, the number of hosts, and the number of parasite free-living stages. Intensity-dependent models are better able to accommodate the biology of many species of adult parasitic worms. Because helminths are much larger than are protozoans, bacteria, and viruses, intensity-dependent models are termed macroparasite models.

While the terms macroparasite and microparasite have utility for modeling purposes, they have been inappropriately used as a coarse taxonomy, presumably because the prefixes in the terms focus attention to the body size of the consumer. Protozoa and smaller microbes are `microparasites' and helminths and arthropods are `macroparasites' even though some small parasites may be better modeled as macroparasites and many large parasites may be better modeled as microparasites.

Parasitic Castrators

Parasitic castrators take the bulk of their energy from their host's reproductive tissues, often reducing the host's fitness to zero. While the host is alive, and appears well, it is dead from an evolutionary perspective. This strategy is not commonly recognized in the veterinary and medical fields, because large vertebrates generally lack parasitic castrators. In many other systems, however, parasitic castrators may be common: cestodes can castrate their fish and invertebrate hosts, larval trematodes usually castrate their molluskan first intermediate hosts, and parasitic barnacles castrate their crustacean hosts. Parasitic castrators are large with respect to their hosts, but ironically, they are generally best modeled as microparasites because their effects on the host are intensity independent.

Complex Life Cycles

Within a complex life cycle, a parasite may use more than one life-history strategy, exploiting several host species in succession. Most complex life cycles can be considered predator–prey (e.g., acanthocephalans), vector (e.g., malaria), or free-living stage transmitted (e.g., trematode cercariae). For instance, trematodes most likely began as parasites of mollusks and later added vertebrate definitive hosts, while many parasitic nematodes whose adults live in vertebrate guts later added intermediate hosts. There are a variety of ways parasites can add new hosts. Biting arthropods clearly provided a convenient means for blood and tissue parasites to contact new hosts. In addition, there should be selection for parasites to survive the predation of their hosts by parasitizing their host's predator. The latter case provides a new category of parasite, the trophically transmitted parasite. Such parasites do not kill their intermediate host, but require its death for transmission.

Ecological Effects of Parasites

Parasite ecology is a developing tradition and already has a rich set of its own jargon (Table 1). Parasites are small but their numbers and biomass can add up so that they can have impacts at the level of host individuals, populations, and communities. Parasites also have important evolutionary effects in that they promote the evolution of host defense and sexual reproduction.

Competition

Density-dependent transmission allows parasites to disproportionately affect common species. This helps maintain rarer competitors, thereby promoting coexistence and stability. In addition, when hosts share parasites, parasites can be competitive weapons. For example, two competing species of amphipod may coexist in nature because a trematode reverses their relative population growth rates. Competitive weapons, such as parasites, can also reduce biodiversity if subordinate species are more susceptible. For instance, a larval tapeworm shared by two flour beetle species increases the rate at which the dominant beetle excludes the subordinate.

Parasites could tip the balance in competitive interactions between native and introduced species. Invasive species typically bring only a small fraction of their parasites to invaded regions, and what they pick up from the native community rarely makes up for the difference. Alternatively, an invader has no coevolved history with the few new parasites it acquires, and these could limit the invasion. Parasites can cause two species to interact indirectly even if these species do not compete for resources. Such apparent competition occurs because one host (the more tolerant or resistant) helps maintain the abundance of a natural enemy that then differentially affects the second species.

Predation

Disease can affect top predators. For instance, a Scandinavian outbreak of sarcoptic mange (caused by mites) in the late 1970s through the 1980s reduced the density of red fox. Prey (rodents, rabbits, ground birds, deer) increased as a result and then declined after the epidemic waned and fox populations recovered.

The fate of most parasites is tied to that of their hosts. If their hosts die, this is usually a bad thing for host and parasite alike. In nature, parasites are part of nearly every meal. This puts tremendous evolutionary pressure on parasites to survive the ingestion process, particularly if they can relocate in the predator. Perhaps as a result, many parasitic species have complex life cycles where an intermediate host must be eaten by a final host. In such life cycles, the parasite must wait for the consumption of the intermediate host by an appropriate final host. However, not all parasites are patient. Some parasites manipulate the behavior or appearance of the intermediate host to increase the rate at which a predator host will catch and eat it. For instance, in southern California estuaries, the most common trematode, *Euhaplorchis californiensis*, encysts on the brain of killifish; the worms alter the fish's behavior, making it shimmy and swim to the surface. These fish are 10–30 times more likely to be eaten by birds, the final host of the worm. In this system, the worms essentially dictate which fish live and die. They also provide an easy snack for egrets and herons that otherwise might have to work harder for a living. Some mathematical models indicate that such parasite-increased trophic transmission can reduce prey density; it can also increase predator density so long as the energetic costs of parasitism for the predator are not too severe. Other mathematical models suggest that some predators may depend on parasites to supply them with easy prey.

Parasitism

Parasites can interact with each other. Some parasites have parasites (i.e., hyperparasites), while other parasites compete with each other for host resources. For larval trematodes, competition for resources within the snail is intense and trematodes have special

Table 1 Ecological measures of parasitism

Parasite term	*Definition*
Abundance	Parasites per host (whether infected or not)
Aggregation	A statistical measure of the distribution of parasite abundance among hosts
Colonization	The infection of an uninfected host
Component population	All individuals of a specified life history stage at a particular place and time
Density	Parasites per sampling unit (per host, per gram tissue, etc).
Incidence	Rate at which uninfected hosts become infected
Infrapopulation	All individuals of a species of parasites in a single host
Intensity	Parasites per infected host
Prevalence	Proportion or % of hosts infected by a parasite
Suprapopulation	All developmental phases of a species at a particular place and time

Modified from Bush AO, Lafferty KD, Font JM, and Shostak AW (1997) Parasitology meets ecology: Definitions, clarifications, examples and Margolis *et al.* revisited. *The Journal of Parasitology* 83: 575–583.

morphological and behavioral adaptations for interspecific interactions. For example, adding dominant trematode species to ponds can exclude subordinate trematode species. Parasites can interact with the host, often via the immune system, to displace other parasites or alter their pathogenic effects on the host. Despite all the potential for parasite–parasite interactions, few studies have considered what this means at the community level.

Mutualism/Facilitation

Sometimes, by altering their host, parasites can alter communities dependent on these hosts or their actions. In one case, such manipulations can have dramatic and unexpected consequences for communities. The trematode *Curtuteria australis* reduces the ability of cockles to bury into New Zealand mudflats (perhaps this increases an infected clam's vulnerability to predation by final host birds). The shells of infected clams stick up out of the mud and provide a hard substrate for sessile invertebrates, such as limpets, that otherwise could not persist in the soft sediment. Parasites can affect substrate-forming species as well, shifting communities in the opposite direction. For instance, trematodes reduce populations of a tube-building corophiid amphipod, thereby destabilizing the sediment and altering the faunal composition of a Danish mudflat.

Food Web Topology

Many ecologists acknowledge the potential importance of parasites in food webs and advocate their inclusion. Parasites add links and species to food webs (Fig. 1). Like any consumer, this has the potential to change the chain length, linkage density, or connectance of a food web (which may alter stability). But as mentioned above, parasites differ from predators in several ways, the most notable being their intimate association with their prey and their relatively low individual biomass. Although the individual body size of a parasite is very small, parasite abundance can be high, leading to comparable total biomass of parasites and top predators. Relatively little is known about parasites in food webs, but the studies published to date indicate that parasites are likely to be worth including. They may comprise most links in a food web and, at least for generalist species, be more densely linked in webs than predators. It would seem that no food web is complete without parasites.

Ecological Effects on Parasites

Host Diversity

Parasitism has been posited as a factor that promotes biodiversity, but it is equally logical to expect that high host diversity and abundance should promote parasitism. Hosts serve as both habitat and dispersal agents for parasites and an abundance of hosts should lead to an abundance of parasites. Further, because parasites tend to be host specific, increased species heterogeneity of

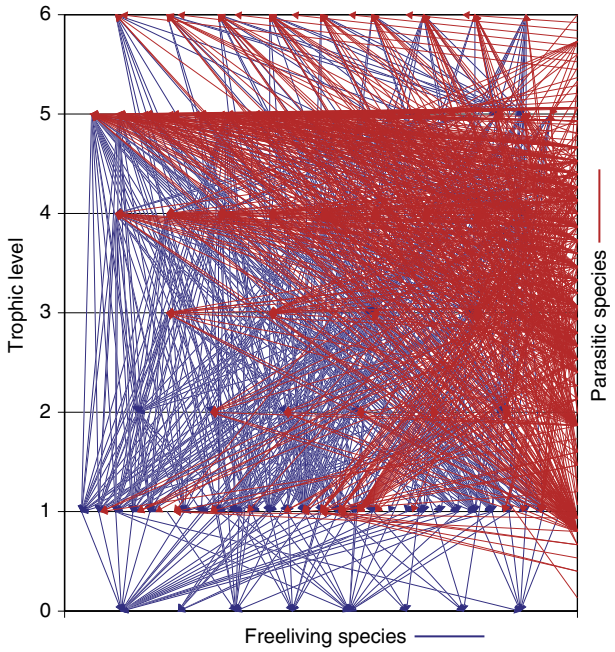


Fig. 1 A food web of Carpinteria Salt Marsh showing predator–prey links in blue and parasite–host links in red.

host communities can facilitate increased species heterogeneity of parasite communities. As a result, a high richness of hosts should contribute to a high richness of parasites. In some cases, parasites may make costeffective bio-indicators of ecosystem health, but here the prediction is counter intuitive – ecosystems with abundant and diverse hosts should support abundant and diverse parasites. There are two exceptions to this prediction. If host abundance decreases because of high diversity, few parasite species may be able to sustain transmission. In addition, if a parasite's vector attacks hosts the parasite cannot develop in, a high diversity of hosts can mean a high probability of contact with non-hosts, thereby diluting transmission.

Environmental Change

Pollutants, malnutrition, and thermal stress due to climate change are all examples of stressors hypothesized to increase individual susceptibility to infectious diseases. This line of thought suggests that parasites should increase in response to environmental stress. For example, intensities or prevalences of ciliates on fish gills increase with oil pollution, pulp mill effluent, industrial effluent, and thermal effluent. This appears to be due to an increase in host susceptibility because toxic conditions impair mucus production which is a fish's main defense against gill parasites.

An opposing prediction generally emerges from considering the population dynamic context of infection. Outside stressors that depress host population density should reduce the chance of an epidemic, or even the ability of a parasite to persist at all, because factors that reduce host density also reduce contact rates between infected and uninfected individuals. Threats to biodiversity, which are generally mediated through reductions in abundance, should indirectly reduce risk to host-specific parasites. By this same reasoning, direct reduction of host density should reduce disease. Culling of seal populations reduces intestinal nematode parasites by reducing host density below transmission thresholds. Fishing can similarly reduce parasites in fish populations and may be responsible for long-term declines in fish parasites in the ocean. For instance, a species of swim bladder nematode was apparently extirpated from native trout in the Great Lakes after a variety of stressors reduced trout populations to very low levels. Alternatively, some stressors may increase parasitism by increasing host density. In particular, the addition of nutrients to aquatic systems increases primary productivity that indirectly increases some grazers and predators. This is probably why the stress most commonly observed to be associated with increased parasitism in fishes and invertebrates is eutrophication.

Stressors may more negatively influence parasites than their hosts. Toxic chemicals and metals have a relatively consistent negative effect across studies of intestinal helminths. Selenium, for example, is more toxic to tapeworms than to their fish hosts. A pollutant may also kill sensitive free-living stages of the parasite. For example, trace metals in sewage-sludge reduce the survival of free-living cercariae and miracidia, leading to a lower trematode prevalence in intermediate-host snails. It is also possible for parasitic infection to make the host more susceptible to toxins. For instance, cadmium is much more toxic to amphipods infected with larval acanthocephalans than uninfected amphipods. While this latter effect decreases the spread of an epidemic through a population, it also increases the impact of disease on infected individuals.

This heterogeneous array of potential effects of stress on infectious disease makes it unclear how a particular stressor should affect the overall course of an epidemic in a host population, or endemic levels of a disease. Although stressed individuals should be more susceptible to infection if exposed, the stressor could simultaneously reduce opportunities for infection because the contact rate between infected and uninfected individuals will decline with the extent that the stressor reduces host density. In addition, populations of some parasites that are directly susceptible to the stressor may not be able to persist at all. It would also appear, *a priori*, that stress can either aggravate or diminish the population-level impact of a host-specific infectious disease organism upon its host.

Further Reading

Bush, A.O., Fernandez, J.C., Esch, G.W., Seed, R.J., 2001. Parasitism: The Diversity and Ecology of Animal Parasites. Cambridge, UK: Cambridge University Press.
Bush, A.O., Lafferty, K.D., Font, J.M., Shostak, A.W., 1997. Parasitology meets ecology: Definitions, clarifications, examples and Margolis *et al.* revisited. The Journal of Parasitology 83, 575–583.
Combes, C., 2001. Parasitism: The Ecology and Evolution of Intimate Interactions. Chicago: University of Chicago Press.
Hudson, P.J., Rizzoli, A., Grenfell, B.T., Heesterbeek, H., Dobson, A.P., 2002. The Ecology of Wildlife Diseases. Oxford: Oxford University Press.
Lafferty, K.D., 2003. Is disease increasing or decreasing, and does it impact or maintain biodiversity. The Journal of Parasitology 89, S101–S105.
Lafferty, K.D., Kuris, A.M., 2002. Trophic strategies, animal diversity and body size. Trends in Ecology and Evolution 17, 507–513.
Torchin, M.E., Lafferty, K.D., Dobson, A.P., McKenzie, V.J., Kuris, A.M., 2003. Introduced species and their missing parasites. Nature 421, 628–630.
Zimmer, C., 2002. Parasite Rex. New York: The Free Press.

Philosophy of Ecology: Overview

K deLaplante, Iowa State University, Ames, IA, USA

Introduction

At its most general level, the philosophy of ecology is the philosophical study of (1) ecological phenomena and (2) those disciplines that study ecological phenomena.

This definition has certain virtues, but it lacks content until we specify what we mean by `ecological phenomena' and what sorts of disciplines study such phenomena. The task is complicated by the fact that the term `ecology' is used in different ways in different contexts.

Ecology is of course a science, but ecology is also identified with a broader philosophical and ethical worldview that in various respects predates modern ecological science. In the `romantic ecology' of the nineteenth century associated with writers like Wordsworth, Thoreau, and Emerson, it was associated with a rejection of mechanistic, atomistic, and reductionistic science and philosophy that was believed to be responsible for a variety of human and natural ills. This conception carried over into the `ecology movement' of the 1960s, an environmental movement tied to broader sociocultural movements of that decade (women's liberation, civil rights, and a range of anticonsumerist, anticapitalist, and antimilitarist movements). In recent decades the term has been appropriated by a number of sociopolitical movements and philosophies that seek to diagnose and ameliorate humanity's dysfunctional relationship with nature (deep ecology, social ecology, socialist ecology, ecofeminism, etc.). Do all of these philosophies count as philosophies of ecology? Are they all branches of the philosophy of ecology?

Within academic philosophy, the most common approach to this question tries to draw a distinction between ecological science and the ethical, social, and broader philosophical uses of the term. Proponents of this approach reserve the term `philosophy of ecology' for the philosophical study of ecological science qua science, with a focus on conceptual issues in fields like behavioral ecology, population ecology, community ecology, evolutionary ecology, and ecosystem ecology. On this view, ecology is conceived as a branch of the natural, biological sciences, and the philosophy of ecology as a specialization within the philosophy of science. Though they may occasionally appeal to the ecological sciences for intellectual support for their various philosophical positions, deep ecology, social ecology, and other radical ecophilosophies are regarded as branches of social theory or environmental philosophy, not the philosophy of ecology.

The approach just described has much to recommend it, but the conception of the philosophy of ecology that will be developed in this article takes a somewhat different tack, one that endorses a broader conception of both the domain of ecology and the philosophy of ecology than is commonly found in the literature. This approach views ecology as a discipline that spans both the natural and social/behavioral sciences. Once this broader conception of ecological science is acknowledged, it becomes increasingly difficult to draw sharp lines between philosophical issues raised by ecological science and philosophical issues raised by a broader ecological worldview.

Ecology: The Study of Ecological Phenomena

Ecology is, at its most general level, the study of `ecological phenomena'. This simple definition is more useful than it appears.

First, though it may initially seem vacuous, the definition acquires content when we specify what it is for a phenomenon to be `ecological'. It is a useful exercise because it forces us to consider the `object' of ecological theorizing rather than the specific techniques, theories, or methodologies that characterize particular forms of ecological science. We want to know what it is about a given subject matter that suggests to the investigator that ecological concepts may be appropriately or usefully applied to it in the first place.

Second, the definition allows us to distinguish ecological `science' from other forms of ecological inquiry simply by defining ecological science as `the scientific study of ecological phenomena'. It is important to keep the issue of the scientific status of different forms of ecological inquiry separate from the question of what it is about some phenomenon that motivates an ecological inquiry in the first place.

Third, we can now give a correspondingly straightforward definition of a `philosophy of ecology'. If ecology is the study of ecological phenomena, then a philosophy of ecology is `the ``philosophical'' study of ecological phenomena'. However, this definition fails to capture the second-order character of much philosophical theorizing (thinking about thinking about X); we will also want to talk about the philosophical study `of the study' of ecological phenomena, that is, the philosophical study of `ecology'.

For the remainder of this article we will interpret ecology as ecological science and the philosophy of ecology as the philosophy of ecological science.

 doi:10.1016/B978-0-444-63768-0.00247-X

Two Conceptions of the Domain of Ecology

What Exactly Is the Domain of Ecology?

Among ecologists and philosophers we can distinguish two schools of thought on this question, each motivated by a variety of methodological, institutional, and social factors. There are those who defend a more `restrictive' view of the domain of ecology, and those who endorse a more `expansive' view. The distinction is important as it imposes a corresponding distinction on ways of understanding the philosophy of ecology.

The Restrictive View: Ecology as Natural Biological Science

The more restrictive view of ecological science is the one encountered today in most standard textbooks used to teach ecology, and currently has the status of orthodoxy within academic ecology. This view was strongly influenced by the rise of evolutionary and population ecology in the 1960s and 1970s, which reinforced an organism-centered conception of ecology that focused on demographic properties of populations and communities. It also developed in response to the appropriation of the term `ecology' by the environmental movement during the same period, which pressured ecologists to clarify how ecology differs from a general concern for environmental welfare.

Supporters of the restrictive conception of ecology are inclined to agree with the following claims:

- Ecology is a pluralistic discipline with many subfields, but ultimately it should be understood as a natural (as opposed to social), biological (as opposed to physical) science.
- The ultimate aim of ecology is to explain and predict patterns and changes in the distribution and abundance of organisms. Ecology is, fundamentally, a science of demographic processes. Ecosystem processes acquire their ecological relevance indirectly, in virtue of their impact on demographic properties of ecological systems.
- Ecology focuses on the natural world of plants and animals. Ecology does not study the root causes of human impacts on the environment, or the social ramifications of such impacts. That is the job of the human social sciences and the humanities, and interdisciplinary fields like environmental studies, which should be distinguished from the natural science of ecology.

The Expansive View: Ecology as Synthetic Systems Science

More expansive conceptions of ecology have also been popular, both within and outside academic ecology. Expansive conceptions of ecology flourished during the heyday of ecosystem and systems ecology under the influence of Eugene and Howard Odum (roughly 1950 to late 1970s). Though currently a minority view in mainstream academic ecology, expansionism has seen a resurgence in certain branches of applied ecology (e.g., systems approaches in conservation ecology and ecosystem management), and it has always been a foundational premise of ecological economics and those traditions of human ecology that claim a strong kinship to scientific ecology.

Supporters of a more expansive conception of ecology are likely to agree with the following claims:

- Ecology is a pluralistic discipline with many subfields, and should be understood both as an interdisciplinary science that spans the physical, biological, and social sciences, and as a synthetic science that has as one of its aims to integrate ecologically relevant information from various different spatial and temporal scales and levels of organization, including human social organization.
- The ultimate aim of ecology is to explain and predict properties of living systems (individuals, populations, communities) as functions of their relationships to their various biotic and abiotic environments. These properties include, but are not restricted to, demographic processes concerning abundance and distribution of organisms.
- Human beings are the most ecologically influential species on the planet and human ecology – the study of the ecological dimensions of human nature and human behavior, including the root causes of environmental attitudes and practices – is an important and legitimate branch of ecology.

Though there exists a set of research traditions in sociology that bear the name `human ecology', this expression is better understood as a general umbrella term for a wide range of scientific disciplines that address different aspects of human–environment relations. Human ecology would thus include fields like ecological economics, ecological anthropology, ecological history, and ecological psychology.

The distinction between the restrictive and expansive conceptions of ecology outlined above induces a corresponding distinction between restrictive and expansive conceptions of the philosophy of ecology.

Issues in the Philosophy of Ecology: Restrictive Mode

Philosophy of ecology in its more restrictive mode focuses on philosophical issues in population, community, evolutionary, and ecosystem ecology as these fields are ordinarily represented in the standard textbooks and journals. In this mode, the philosophy

of ecology is generally understood as a specialization within the philosophy of the natural, biological sciences, alongside (and overlapping with) other such specializations, such as the philosophy of evolutionary theory. For the most part, this is how the subject matter of the philosophy of ecology is conceived within the tradition of Anglo-American philosophy of science.

Many of the philosophical issues studied within this mode are better understood by situating them within the context of the broader intellectual debate that has served to structure much of the foundational discourse of ecology in the twentieth century. This is the debate between `holistic' and `reductionistic' research traditions in ecology.

As the story is usually told, holists believe that ecological systems exhibit order, structure, and regularity at population, community, and ecosystem levels of organization, with higher-level properties and regularities both emerging out of and constraining lower-level properties and regularities. Hence, holists believe the search for law-like generalizations governing the behavior of populations, communities, and ecosystems is a reasonable and desirable goal of ecological research, and formal investigations of community and ecosystem structure are a worthwhile – indeed, indispensable – activity. The ecosystem concept has its home within this broadly holistic picture of ecological systems.

Reductionists, on the other hand (as the story goes), believe that ecological systems are nothing more than assemblages of individual species populations whose behavior is determined largely by response to local environmental conditions (both biotic and abiotic). There are no such things as `communities' or `ecosystems' with emergent causal properties of their own; any properties they have are, at best, epiphenomenal statistical properties of the collection of species populations that compose them. The ecological properties of species populations are best understood in evolutionary terms, as products of natural selection and other evolutionary mechanisms. Consequently, reductionists eschew the search for general laws governing large classes of ecological systems, for it is assumed there are none to discover; rather, their focus is on local, historically contingent, site-specific investigations of population behaviors and environmental conditions.

This dualistic narrative, or some variant of it, has provided the motivating context for most of the writings on foundational issues in ecology, from the early decades of the twentieth century through to the present (e.g., the Clements–Gleason debate over the nature of communities and ecological succession). In this context, to engage in the philosophy of ecology is to take up and defend a position on foundational issues that place one somewhere along the spectrum between extreme holism and extreme reductionism.

One can characterize the core issues in the philosophy of ecology in terms of a set of metaphysical and epistemological questions on a handful of key topics:

The metaphysical status of ecological entities. What is a population, a community, an ecosystem? Do ecological entities have emergent properties that play a causal role in determining how ecological systems change over time? Is the concept of a community or an ecosystem even operationally meaningful?

Law-like regularity versus historical contingency. Does ecology have general laws? If so, what are the causal properties of ecological systems that ground these regularities? Is the existence of such laws consistent with neo-Darwinian selection theory operating at the level of individual organisms? At what levels of organization should we expect to see such laws?

The epistemology of modeling. What is the proper role of theoretical models and model-building in ecological science? If models can only give approximate descriptions of real-world ecological systems, how should their predictions be tested and assessed? Should we interpret theoretical models realistically or as mere tools for organizing, explaining, and predicting observable patterns in ecological data?

Model-driven versus data-driven research traditions. Should ecological research focus on empirical case studies of particular ecological systems rather than general model-building? How should we compare the results of controlled ecological experiments with the results of comparative field studies of natural systems? What are the weaknesses and advantages of each approach?

Evolution and ecology. Is natural selection acting at the level of individual organisms sufficient to explain the organization and structure of communities? Do ecosystems coevolve with their component species populations? How, in general, do evolutionary and ecological mechanisms interact?

This list is incomplete, but the majority of philosophers of science who specialize in the philosophy of ecology have research programs that bear directly on some subset of these questions.

Have any consensus views emerged with respect to any of these questions? As with any branch of philosophy it would be overly optimistic to expect consensus on foundational questions. One can, however, identify historical and recent trends in how ecologists and philosophers have viewed these issues. One can say, for example, that the 1950s and 1960s were dominated by holistic approaches in ecology, and there was a high degree of optimism about the prospect of a mature ecological science that could compare favorably with law-governed fields like physics. Opinion swung the other way in the 1970s and 1980s with the rise to dominance of evolutionary and population approaches in ecology, during which period greater emphasis was placed on historically contingent, site-specific features of ecological systems, along with an attending skepticism about laws in ecology and criticism of holistic approaches in ecology generally. Professional philosophers of science have really only started looking at these questions within the last 15 years, but recent work indicates that the pendulum is swinging back to a more intermediate position between the holistic optimism of the 1950s and the reductionistic pessimism of the 1980s.

Issues in the Philosophy of Ecology: Expansive Mode

Those who endorse a more expansive conception of ecological science believe that ecological concepts and theories may be usefully applied to a broader range of phenomena than do defenders of the more restrictive conception. In this section we review two sources of motivation for the expansive conception, and introduce a set of issues for the philosophy of ecology pertaining to each source.

Systems Ecology

One of the sources of motivation for the expansive conception of ecology is reflection on the domain of systems ecology. Systems ecologists use a variety of formal techniques – network theory, information theory, dynamical systems theory, etc. – to describe the structural and dynamical properties of whole ecosystems. But why should systems ecology be associated with the expansive conception of ecology? Is systems ecology not commonly viewed as a branch of traditional ecosystem ecology?

To answer this question it may be useful to distinguish three related types of ecological properties or phenomena:

1. properties of `biological entities' that depend on or make essential reference to relations to environments;
2. properties of `environments' that depend on or make essential reference to relations to biological entities; and
3. properties of the `relations' that obtain `between' biological entities and their environments.

Different research traditions in ecology can be distinguished in part by which of these three categories of ecological phenomena are the main focus of study. Population and community ecology, for example, are organism-centered branches of ecological science that focus on phenomena of type (1). The magnitude and rate of change of properties like population size and density and community composition all depend in various ways on the relationships that the component populations have to their biotic and abiotic environments.

Empirically oriented forms of ecosystem ecology (i.e., biogeochemistry and ecological stoichiometry) tend to focus on phenomena of type (2) involving stocks and flows of biologically relevant elements, nutrients, and minerals. For example, the standing stock of phosphorus in a lake ecosystem is a property of the environment of the lake's biotic community, but its properties depend in part on the biotic activities of this community.

Systems ecology, by contrast, focuses on phenomena of type (3). The nodes of an abstractly defined ecological network are meant to correspond to functionally defined ecological types (predators, filter feeders, deposited detritus, microbiota, etc.), but for the most part the phenomena of interest to systems ecologists are the `network' or `organizational' properties of such systems (e.g., connectance, cycling indices, throughputs, and other measures of network structure and function). Systems ecologists are perhaps better described as `complex systems' ecologists; they seek to describe and explain macro-level patterns in the structure and behavior of ecosystems – patterns, for example, associated with self-organizing processes – that may be characteristic of certain generic classes of complex systems.

A striking feature of such complex systems patterns is that they can often be realized in systems of different kinds (physical, chemical, biological, ecological, etc.). To give just two examples: (1) the same critical point phenomena observed in phase transitions in gases and fluids can be observed in the transition from ferromagnetic to paramagnetic state in magnetic materials; and (2) the same `period-doubling route' to chaotic dynamics has been observed in systems as diverse as fluids, chemical clocks, electrical circuits, lasers, and acoustic systems. These and other complex systems behaviors have the following generic features:

1. The details of the system (those details that would feature in a complete causal-mechanical explanation of the system's behavior) are largely irrelevant for describing the behavior of interest.
2. Many different systems with completely different `micro' details will exhibit identical behavior.

What do fluids, chemical systems, electrical circuits, lasers, and acoustic systems have in common that would explain their common period-doubling route to chaotic dynamics? Whatever it is, it cannot have much to do with the specific material properties of the components that make up these systems. Any explanation must refer to the relational or structural features that the systems have in common – in short, it must abstract away from the `matter' to identify the underlying `form' that is common to all the systems in question.

Theoretical systems ecology looks to discover properties of this type that describe ecological systems, but due to their formal character, they may equally describe properties of neural networks in the brain or human socioeconomic networks. It should not be surprising, then, that systems ecologists have a tendency to speculate on the implications of their work for phenomena in other branches of natural and social science.

Philosophers of ecology have largely ignored systems ecology, an unfortunate situation given the number of interesting philosophical questions that the field raises. For example,

- What precisely does it mean to say that a real-world ecosystem instantiates or exemplifies the organizational properties of a formal ecosystem model?
- How do we know when a real-world system actually instantiates a particular formal model? How can such claims be tested? What evidence would bear on them?

- How do the formal properties of ecological systems (the properties that might be instantiated in many different kinds of systems) interact with the material properties of ecological systems (the properties that are particular to the material constitution of the system in question), to generate observed structural and behavioral patterns?
- Does the existence of formal properties of ecosystems demand a holistic view of ecosystems, or is it compatible with a reductionistic view whereby the properties of the whole are determined by the properties of the component parts?

Human Ecological Sciences

We have already noted a second motivation for an expansive conception of ecological science, namely, the fact that there already exists a variety of ecological sciences that deal with human–environment relations.

There are subdisciplines within traditional ecology, such as human paleoecology and human paleobiology, that focus on human–environment relations in the evolutionary past. These disciplines are components of human origins research, an interdisciplinary field that draws on expertise in anthropology, archeology, and linguistics as well as traditional ecology and biology. Philosophers have taken great interest in human origins research. The field is foundational for human sociobiology and evolutionary psychology, which in turn are foundational for naturalistic theories of cultural evolution, and for a variety of positions in the philosophy of mind, language, and ethics.

There also exist a variety of ecological disciplines that study human–environment relations in the present, such as ecological economics, ecological psychology, ecological anthropology, and ecological sociology. Those who work in these fields usually have disciplinary affiliations in economics, psychology, anthropology, or sociology, rather than biology or ecology, yet it is typical for workers in these nontraditional ecological disciplines to view their field as continuous with a general ecological science of organism–environment relationships.

Note that acknowledging these disciplines as ecological sciences does not imply that they all employ the same scientific methods nor that they are all equally successful as sciences. It implies only that at some level they are part of a common scientific enterprise.

It should be obvious that a philosophy of ecology that includes all these human ecological sciences within its scope will have a correspondingly broader sweep than its more restrictive counterpart, since it self-consciously includes philosophical issues relating to the ecological dimensions of human cognition, human social organization, and human–environment relations more broadly. In this mode, the philosophy of ecology naturally spans both the natural and social sciences, and reaches deeper into the domain of sociopolitical philosophy and ethics than it does in its more restrictive mode.

Consider, for example, philosophical issues in ecological economics. One goal of ecological economics is to devise methods of economic valuation and organization that promote the goals of long-term ecological and economic sustainability. A fundamental challenge of this goal is to provide a meaningful definition of `sustainable' that applies to ecological and economic systems. Research on this question has shown that the concept of sustainability is inherently value-laden, and that one cannot properly address the issue without considering the ethical and sociopolitical consequences of public policies that would operationalize the concept. Evaluating these consequences is, naturally, a task for ethics and sociopolitical philosophy. But if ecological economics is a branch of ecology, and the foundational issues of ecological economics belong to the philosophy of ecology, then the challenge of evaluating the ethical and sociopolitical dimensions of the concept of sustainability also belongs to the philosophy of ecology. Similar reasoning applies to the foundational problems of all the human ecological sciences noted above.

Ecology-the-Science and Ecology-the-Worldview

In the introduction we noted two senses of the term `ecology', one associated with ecology as a natural biological science and the other associated with ecology as a philosophical worldview concerned with human–environment relations in the broadest sense. The dominant tradition in the philosophy of ecology tries to separate these senses as much as possible, restricting the philosophy of ecology to the investigation of foundational issues in ecological science and relegating the ethical, political, and more speculative metaphysical dimensions of the broader ecological worldview to other branches of philosophy. This approach has its merits; it is consistent with the way most professional ecologists understand ecology and it makes more efficient use of the professional division of labor among philosophers.

We also noted, however, that there has been disagreement among ecologists over how to understand the domain of ecology. Some argue for a more restrictive conception of ecology that identifies it with the traditional ecological disciplines taught in natural science departments. Others argue for a more expansive conception that includes the study of human–environment relations. We saw how this expansive conception of ecology draws support from two sources: first, a consideration of the distinctive character of systems ecology; and second, the existence of a variety of human ecological sciences.

If we accept an expansive conception of ecology, do we lose the sharp distinction between ecology-the-science and ecology-the-worldview that was such an attractive feature of the restrictive conception? Yes and no. On the one hand, the philosophy of ecology in its expansive mode will inevitably include questions that address metaphysical, epistemological, and normative issues that are also addressed in more speculative ecological and environmental philosophies. The domains of the philosophy of ecology and environmental philosophy will necessarily overlap.

On the other hand, in demanding that ecology operates as a science that is beholden to the epistemological standards of the scientific disciplines that it encompasses, and not to the presuppositions of any particular philosophical worldview, then ecological science will retain its autonomy and identity as a science. Though their domains may overlap, the methods of empirical science distinguish ecology-the-science from ecology-the-worldview.

Further Reading

Allen, T.F.H., Hoekstra, T.W., 1992. Toward a Unified Ecology. New York: Columbia University Press.
Brennan, A., 1988. Thinking About Nature: An Investigation of Nature, Value and Ecology. Athens: University of Georgia Press.
Cooper, G.J., 2003. The Science of the Struggle for Existence: On the Foundations of Ecology. New York: Cambridge University Press.
Cuddington, K., Beisner, B. (Eds.), 2005. Ecological Paradigms Lost: Routes of Theory Change. London: Elsevier Academic Press.
deLaplante, K., 2004. Toward a more expansive conception of ecological science. Biology and Philosophy 19, 263–281.
Ginzburg, L., Colyvan, M., 2004. Ecological Orbits: How Planets Move and Populations Grow. New York: Oxford University Press.
Gunderson, L.H., Holling, C.S. (Eds.), 2002. Panarchy: Understanding Transformations in Human and Natural Systems. Washington: Island Press.
Keller, D.R., Golley, F.B. (Eds.), 2000. The Philosophy of Ecology: From Science to Synthesis. Athens: University of Georgia Press.
Mikkelson, G.M., 2003. Ecological kinds and ecological laws. Philosophy of Science 70, 1390–1400.
Odenbaugh, J., 2003. Complex systems, trade-offs and mathematical modeling: A response to Sober and Orzack. Philosophy of Science 70, 1496–1507.
Peters, R., 1991. A Critique for Ecology. Cambridge: Cambridge University Press.
Real, L.A., Brown, J.H. (Eds.), 1991. Foundations of Ecology: Classic Papers with Commentaries. Chicago: University of Chicago Press.
Sarkar, S., 2005. Biodiversity and Environmental Philosophy: An Introduction. New York: Cambridge University Press.
Schrader-Frechette, K.S., McCoy, E.D., 1993. Method in Ecology: Strategies for Conservation. New York: Cambridge University Press.
Taylor, P.J., 2005. Unruly Complexity: Ecology, Interpretation, Engagement. Chicago: University of Chicago Press.

Phytosociology

J Dengler, University of Hamburg, Hamburg, Germany
M Chytrý, Masaryk University, Brno, Czech Republic
J Ewald, University of Applied Sciences Weihenstephan, Freising, Germany

Introduction

Phytosociology is a subset of vegetation science, in which it stands out by focusing on extant (vs. fossil), taxonomic (vs. physiognomic or functional) plant assemblages at the scale of vegetation stands (vs. landscapes or biomes). Its principal goal is the definition and functional characterization of vegetation types based on the total floristic composition of stands. Phytosociology distinguishes between concrete vegetation stand (phytocoenosis), which can be represented by a plot record (relevÅ), and abstract vegetation type (syntaxon), representing a group of all stands sharing certain attributes. The classification framework (syntaxonomy) is designed in close analogy to plant taxonomy, with association as the basic unit.

The fundamental concepts of phytosociology were developed by Josias Braun-Blanquet in the 1920s. He combined a standardized protocol for plot sampling, sorting of species-by-plot matrices, demarcation of community types, and their hierarchical ordering into a practical and efficient framework for the study of vegetation. In this article, we use the term phytosociology for the Braun-Blanquet approach and its modern extensions.

Phytosociology is the mainstream vegetation classification scheme in Europe, as well as in several countries outside Europe, and has become increasingly popular worldwide from the 1990s onward. Within modern ecology, phytosociology represents the most comprehensive and consistent methodology for vegetation classification. RelevÅs are the most widely used standardized protocol for sampling plant species co-occurrences at the stand scale. Being derived from the vast body of relevÅ data, syntaxonomy provides a comprehensive yet open system of vegetation types, which are indispensable in land-use management and nature conservation. Consisting of abundance data on individual plant species, relevÅs and vegetation types organized in large phytosociological databases are an enormous source of fine-scale biodiversity information. If linked to the growing body of plant trait or indicator value data or environmental information in geographical information systems (GISs), phytosociological data open new avenues for exploring large-scale ecological patterns and processes, and provide spatially explicit information necessary for environmental management.

Phytosociological Data

Data Records

In phytosociology, the data of a single plot are called a relevÅ (French for record, see **Table 1**), which consists of `header' and species data. The `header' comprises plot identification, methodological information, and metric, ordinal, or categorical data on geographic position, environmental conditions, and overall vegetation structure. Some of these data are essential, others optional, depending on the purpose and resources of a project (**Table 2**).

The species data are composed of a list of plant taxa (species and infraspecific taxa; further referred to as `species') and their attributes. A full relevÅ lists all plant species occurring in the plot and growing on soil, including bryophytes, lichens, and macroalgae. Additional recording of species growing on substrata other than soil, such as on living plants (epiphytes), rocks (saxicolous plants), or dead wood (lignicolous plants), is desirable, but not standard in phytosociology. Every species observation is assigned to a vertical stratum (e.g., tree layer, shrub layer, herb layer, and cryptogam layer). Woody species occurring in different layers are recorded separately for each layer. For each species observation in a layer, an importance value is estimated and usually expressed on a simplified scale of abundance (number of individuals/ramets) and/or cover (area of the vertical projection of all aerial parts of a species relative to the total plot area) (**Table 3**). As mixed cover-abundance scales pose problems in data analysis, pure cover scales are preferred when precise quantitative estimates are required, for example, in studies of vegetation change in permanent plots. Sometimes, additional characteristics of the species – such as sociability (degree of clustering of the individuals), vitality, fertility, age class (e.g., seedling or juvenile), and phenological status – are recorded, but these are of little or no importance for standard analyses.

Selection and Size of Plots

Plot sites in the field are positioned in vegetation stands that are relatively homogeneous in terms of structure, species composition, and environment, so that variation is minimized within and maximized between plots.

The traditional sampling strategy in phytosociology, preferential sampling, in which the researcher selects stands that are considered as representative of some vegetation units, has several disadvantages: it is not repeatable by other researchers, tends to neglect some vegetation types and oversample others, and produces a nonrepresentative sample of vegetation diversity in the study area. In spite of these disadvantages, probabilistic sampling strategies, such as random or systematic sampling, have never received

 doi:10.1016/B978-0-444-63768-0.00533-3

Table 1 Example of a forest relevÅ with five vegetation layers distinguished: upper tree layer (T1), lower tree layer (T2), shrub layer (S), herb layer (H), and cryptogam layer (C)

Plot ID/methodology	
Field number	291
Author	J Ewald
Plot size (m^2)	144
Plot shape	square
Sampling date	3 June 1997
Preliminary syntaxon	Galio-Fagetum adenostyletosum
Geographic data	
UTM coordinates	32 U 4434393 E – 5272800 N
Locality	Ettaler Manndl, Höllenstein, 3 km W from Eschenlohe,Garmisch-Partenkirchen, Bavaria, Germany
Environmental data	
Elevation (m a.s.l.)	1300
Slope aspect (°)	35
Slope inclination (°)	32
Soil type	Cambisol
Parent material	Cretaceous sandstone
Management	Protective forest
Stand age (year)	140
Structural data	
Height upper tree layer (m)	30
Height lower tree layer (m)	6
Height shrub layer (m)	3
Cover upper tree layer (%)	75
Cover lower tree layer (%)	3
Cover shrub layer (%)	1
Cover herb layer (%)	20
Cover cryptogam layer (%)	3

Layer	*Species*	*Importance*	*Layer*	*Species*	*Importance*
T1	*Fagus sylvatica*	3	H	*Oxalis acetosella*	2
	Picea abies	3		*Paris quadrifolia*	+
				Polypodium vulgare	+
T2	*Picea abies*	1		*Prenanthes purpurea*	+
				Primula elatior	+
S	*Picea abies*	1		*Ranunculus lanuginosus*	1
				Rumex alpestris	+
H	*Acer pseudoplatanus*	+		*Salvia glutinosa*	1
	Aconitum vulparia	+		*Sanicula europaea*	+
	Adenostyles alliariae	1		*Saxifraga rotundifolia*	1
	Adoxa moschatellina	+		*Senecio fuchsii*	1
	Athyrium filix-femina	+		*Stellaria nemorum*	2
	Cardamine flexuosa	+		*Thelypteris limbosperma*	+
	Chaerophyllum hirsutum	+		*Veronica urticifolia*	+
	Chrysosplenium alternifolium	+		*Viola biflora*	+
	Cicerbita alpina	+			
	Deschampsia cespitosa	+	C	*Atrichum undulatum*	1
	Dryopteris dilatata	+		*Brachythecium rutabulum*	+
	Dryopteris filix-mas	+		*Conocephalum conicum*	+
	Epilobium montanum	+		*Ctenidium molluscum*	+
	Galeopsis tetrahit	+		*Dicranella heteromalla*	+
	Galium odoratum	+		*Dicranum scoparium*	+
	Geranium robertianum	+		*Fissidens taxifolius*	+
	Gymnocarpium dryopteris	+		*Mnium spinosum*	+
	Impatiens noli-tangere	+		*Plagiochila porelloides*	+
	Lamiastrum montanum	1		*Plagiomnium undulatum*	+
	Luzula sylvatica subsp. *sieberi*	+		*Plagiothecium curvifolium*	+
	Lysimachia nemorum	1		*Polytrichum formosum*	+
	Mercurialis perennis	+		*Rhizomnium punctatum*	+
	Mycelis muralis	1		*Thuidium tamariscinum*	+
	Myosotis sylvatica	+			

Table 2 Essential (*) and selected optional data to be included in the `header' of a phytosociological relevÅ

Group	*Data*	*Comment*
ID/methodology	Field number*	
	Author(s)*	
	Plot size*	
	Plot shape	
	Sampling date*	
	Preliminary assignment to a syntaxon	
Geographic data	Geographic coordinates*Locality in textual form*	For example, Greenwich coordinates, UTM including political and/or natural geographic units
Environmental data	Elevation (m a.s.l.)*	
	Slope aspect*	
	Inclination*	
	Soil	For example, type, texture, depth, pH, humus form, humuscontent, C/N ratio
	Geology (parent material)	
	Management	
Structural data	Height of vegetation layers (m)	For example, tree layer, shrub layer, herb layer, cryptogam layer
	Cover of vegetation layers (%)*	Cover of each layer and total cover
	Cover of other surfaces (%)	For example, bare soil, litter, woody debris, rocks, open water

Table 3 Customary version of an extended Braun-Blanquet cover-abundance scale with ordinal values, which are often used for numerical interpretation. In the original Braun-Blanquet scale, 2m, 2a, and 2b were joined under the symbol `2´

Symbol	*Abundance (number of individuals/ramets)*	*Cover interval (%)*	*Ordinal value*
r	1	0–5	1
+	2–5	0–5	2
1	6–50	0–5	3
2m	More than 50	0–5	4
2a	Any	5–12.5	5
2b	Any	12.5–25	6
3	Any	25–50	7
4	Any	50–75	8
5	Any	75–100	9

wider acceptance in phytosociology. While providing reliable estimates of vegetation attributes, probabilistic sampling is less suited to phytosociology's goal of representing maximum variation in vegetation diversity across a study area, as it tends to undersample or even miss rare types. GIS and global positioning system (GPS) technology have made stratified-random sampling schemes increasingly popular in phytosociology. Based on the overlay of digital maps in a GIS, the study area can be stratified into patches with certain combinations of land-cover types and environmental variables that are supposed to correlate with plant distribution. Within each of these strata, plot positions are randomly placed and subsequently found in the field with a GPS receiver. A related sampling strategy is a gradient-oriented transect or gradsect, which establishes plot sites along a landscape transect that runs parallel to an important environmental gradient.

Phytosociological plots are usually squares or rectangles, which, as a rule of thumb, are roughly as large in square meters as the vegetation is high in decimeters (e.g., 200 m^2 for a forest of 20 m height). Despite this rule and other suggestions in textbooks, actual plot sizes used may span more than one order of magnitude within the same vegetation type. Standardization of plot sizes is hindered by the vague and misleading concept of `minimal area', which is thought to be a certain plot size specific for each vegetation type, beyond which any further enlargement has negligible effects on species richness and composition. However, plot size strongly influences estimates of species richness and other vegetation parameters. Joint use of differently sized relevÅs in a single analysis may thus produce artifacts in classification, ordination, and calculation of fidelity of species to vegetation units. To safeguard data compatibility, standard plot sizes have been proposed for use within certain structural formations, for example, 200 m^2 in forest vegetation; 50 m^2 in scrub vegetation; 16 m^2 in grassland, heathland, and other herbaceous vegetation; and 4 m^2 in aquatic and low-growing herbaceous vegetation.

Vegetation Databanks

Phytosociology has a long tradition of publishing, archiving, and re-analyzing relevÅs as its basic primary data. Many phytosociological journals print full tables including all relevant relevÅs, thus making data accessible for future compilation and analysis, which was

traditionally performed as synoptic tables on paper. The limitations of manual data management were overcome by using table editing and databank software, which allows seizing, storing, managing, filtering, and analyzing relevÅ data in multiple ways.

Compilation in a databank requires that all information obeys stringent formal and technical rules laid down in reference lists, meta-data and data models. Databanks of different formats and complexity were established, ranging from simple spreadsheets to relational and object-based data models that allow flexible definitions and comprehensive documentation of meta-data. Simple databanks are able to exchange data freely if the same standards, database formats, definitions, and reference lists are used. The success of phytosociological databanks is so far due to rather simple management software packages such as TURBOVEG, which is currently the most widespread program in Europe and beyond, distributed free of charge or at small cost along with taxonomic reference lists and tools to create, edit, and analyze phytosociological tables.

While early databank development revolved around fixing standards for data types and references for plant taxon concepts and names, modern ecoinformatics provides tools to exchange data of different formats and taxonomic reference and, ultimately, link up databanks of any format in networks. Rather than enforcing standard formats, these systems require that data are recovered and stored with as much original information as possible, including meta-data on sampling design and methods, cover-abundance scales, definition of layers, taxonomic references, and original data sources.

Classification of Vegetation

Aims and Criteria

Vegetation classifications are performed with three fundamental goals: (1) delimiting and naming parts of the vegetation continuum to enable communication about them; (2) predicting a multitude of ecosystem attributes (e.g., species composition, site conditions, and ecological processes) from the assignment of a particular stand to a vegetation unit; and (3) making multi-species co-occurrence patterns representable by verbal descriptions, tables, diagrams, and maps. Floristically defined vegetation types are thus suitable reference entities for ecological research, bioindication, and nature conservation.

Reaching these aims requires of the classification approach:

1. coherence of units with respect to major ecosystem properties;
2. simple and clear discernability of units;
3. completeness of the system (i.e., coverage of all vegetation types of the given area);
4. robustness (i.e., minor changes of the data should not considerably change the classification);
5. tolerance against varying data quality;
6. supra-regional applicability;
7. applicability for a range of different purposes;
8. hierarchical structure, allowing for different degrees of generalization;
9. equivalence of units of the same hierarchical level; and
10. adequate number of units with respect to practical use.

As no single classification can ideally meet all of these criteria at the same time, and their relative importance depends on the purposes, competing classifications of the same objects and data are a reality. Thus, the interpretation of local data will change with scaling up from local to regional and supra-regional context. However, there is also a practical requirement to have a unified supra-regional classification to enable communication among scientists, managers, and authorities between regions.

Braun-Blanquet Approach

The `Braun-Blanquet approach' provides a methodological framework for vegetation classification that seeks an optimal combination of the above criteria and that reconciles conflicting requirements of different scales and purposes. However, it is not an unambiguous and uniform set of recipes, and it has been subject to diverse modifications. Despite the variety of different versions, practitioners agree on certain fundamentals, which distinguish the Braun-Blanquet approach from most other ways of vegetation classification: (1) The classification is based on the (total) species composition of the sample plots (floristic–sociological method), whereas structural or environmental criteria play a subordinate role. (2) The classification units called syntaxa (singular: syntaxon) are arranged into a hierarchical system according to their floristic similarity. The principal ranks of this system are, from bottom up, association, alliance, order, and class. (3) There are generally accepted rules for the scientific naming of syntaxa (see the section entitled `Phytosociological ranks and nomenclature').

Within the Braun-Blanquet approach, the concept of character and differential species is important for the recognition of previously defined syntaxa. Differential species are those that positively differentiate, by their occurrence, the target syntaxon from other syntaxa. Character species are a special case of differential species: they positively differentiate the target syntaxon from all other syntaxa. The differential and character species combined are called diagnostic species. The validity of diagnostic species may be restricted to comparisons within the syntaxon of the next higher rank or within a physiognomic vegetation type. Diagnostic species are based on the concept of fidelity, that is, concentration of their occurrence or abundance within the given syntaxon. Traditionally, arbitrary measures of fidelity were used, such as constancy in the target syntaxon had to be at least twice as high as in

any other syntaxon. Nowadays, statistical fidelity measures are increasingly used (see the section entitled `Numerical approaches'). However, in spite of several attempts at a formal definition of differential and character species, no widely accepted agreement in this respect has been reached so far.

Phytosociology faces difficulties in the classification of vegetation types that lack species of narrow ecological amplitude which could be used as character species of the respective syntaxa. This problem led early practitioners to avoid stands without specialist species as `atypical' and `fragmentary' and oversample those containing presumed character species. Even when sampled and recognized, such poorly characterized vegetation types were often excluded from the syntaxonomic system. Vegetation types poor in diagnostic species may be incorporated into the system in several ways, for example: (1) Deductive classification affiliates such units as so-called basal or derivative communities to higher syntaxa of the system, from which their formal names are derived (e.g., *Elymus repens* [*Artemisietea vulgaris*] derivative community). (2) According to the concept of central syntaxon, there can be one negatively differentiated syntaxon within the next superior syntaxon of the hierarchy; central syntaxa have the same ranks (e.g., association) and nomenclature as normal syntaxa.

This diversity of approaches within the Braun-Blanquet system must be unified where all vegetation types of a large area are to be placed in a single coherent system, such as in modern projects of national vegetation classifications. These projects have usually developed consistent systems of standardized and operational methodology of vegetation classification based on the Braun-Blanquet approach.

Numerical Approaches

Traditional phytosociological work was based on the subjective delimitation of vegetation units, made either already during the field reconnaissance and sampling or in the process of manual sorting of relevÅs and species within tables. The need for more formal, transparent, efficient, and repeatable classification procedures led to the introduction of numerical classification methods in phytosociology since the 1960s. They can be either agglomerative or divisive. Agglomerative methods start with linking individual relevÅs based on the similarity of their species composition, forming relevÅ clusters and subsequently linking these clusters to form a hierarchical classification, usually presented as a dendrogram. Divisive methods start with dividing the set of relevÅs into subsets, which are further divided into subsets on a lower hierarchical level, thus eventually proceeding to the single relevÅs. The most popular divisive method is two-way indicator species analysis (TWINSPAN), which uses the ordination method of correspondence analysis to divide the relevÅs into subsets. Simultaneously with the classification of relevÅs, TWINSPAN classifies species, and produces an ordered species-by-relevÅ table similar to that used in traditional phytosociology (see the section entitled `Phytosociological tables'). The classifications of the same data sets produced by agglomarative clustering and TWINSPAN usually roughly correspond but differ in details. Agglomerative clustering is the method of choice when cluster homogeneity is the principal goal, while TWINSPAN better reflects the main gradients in species composition of the input data set. An important choice in any numerical procedure is the transformation of cover-abundance data, which determines to what degree species cover-abundance will be accounted for in the analysis.

In addition to numerical classification, phytosociology frequently uses various ordination methods, such as correspondence analysis (CA), detrended correspondence analysis (DCA), or principal components analysis (PCA). Sometimes, ordination and classification are perceived as antagonististic approaches, representing the Gleasonian continuum concept and the Clementsian concept of superorganism, respectively. However, phytosociologists never engaged in that ideological debate, and nowadays both approaches seem to be reconciled: classification studies often use ordination to visualize the position of vegetation units along gradients, and ordination patterns are used to propose the delimitation of relevÅ groups for certain purposes.

Applicability of an established classification crucially depends on finding those species that are typical of relevÅ groups (vegetation units) and make them recognizable by simple floristic criteria. Such species may include the most frequent species, dominant species, or diagnostic species. The former two groups of species can be easily defined by setting some threshold of constancy or cover-abundance values that a species must exceed to be considered as frequent (constant) or dominant, respectively. Diagnostic species are determined based on the concept of fidelity, which quantifies the degree of concentration of a species' occurrence or abundance in the relevÅs of the target vegetation unit. If a species occurs mainly in the relevÅs of the target vegetation unit while it is largely absent elsewhere, it is considered as faithful to this vegetation unit. Fidelity can be quantified by various statistical measures. If it is based on species presence/absence, various measures of association between categorical variables can be used, for example, chi-square, G statistic, or phi coefficient of association. Some fidelity measures have also been proposed to deal with cover-abundances, for example, the Dufrêne–Legendre indicator value. The properties of different fidelity measures vary slightly, for example, with respect to the weight given to rare or common species. Statistical significance of fidelity can be either derived directly from the values of some of these measures or determined by a separate procedure such as permutation test. Apart from the selection of the appropriate fidelity measure, fidelity can be measured in two different ways. First, species occurrence in the target group of relevÅs can be compared with all the relevÅs in the data set that do not belong to the target group, irrespective of the divisions of the rest of the data set. Second, species frequency in that group of relevÅs where it is most common is compared with its frequency in the group where this species is the second most common. In both cases, some arbitrary threshold fidelity value is selected and species that exceed this value are considered as diagnostic. The first approach is not affected by the divisions of the data set outside the target vegetation unit, thus yielding a more general result, whereas the latter approach is only valid in the context of a given table or classification, but it provides a clearer separation of vegetation units through diagnostic species within this table or classification. The results of both approaches depend on the geographical extent, sampling design, and delimitation of the available set of relevÅs, the `universe of investigation'.

Integrating the Different Approaches

While having the basic aims in common, the traditional Braun-Blanquet approach and numerical approaches differ in some respects. Indeed, no approach produces an objective or `the correct' classification. In spite of the high degree of formalization involved in numerical classification, the numerous choices concerning the data set composition, cover-abundance transformation, numerical coefficients, classification algorithms, or number of vegetation units to be accepted result in the fact that numerical methods, like the traditional expert-based approaches, may suggest many different partitions of the same data.

Unlike the expert-based classifications, which often use unclear classification criteria, numerical classification methods consistently use explicit information on species occurrence and cover-abundance and apply it consistently across vegetation types. However, while experts often implicitly incorporate in the classification process knowledge of species behavior in a broad geographical and environmental range, numerical methods only use information contained in the particular data set, which often results in rather idiosyncratic classifications. It is therefore difficult to combine different numerical classifications into a single system of syntaxa, which would be valid over large areas and different habitats, without relying on expert judgment.

To avoid these problems, supervised classification methods have gained importance recently. They take traditional syntaxa that are widely recognized by phytosociologists as given and assign new relevÅs to these syntaxa by numerical procedures. Such an approach supports both the stability of the traditional phytosociological system, which has already received wide acceptance, and the application of formal, unequivocal classification procedures. A simple approach is to calculate an index of similarity of species composition between new relevÅs and constancy columns of synoptic tables (see the section entitled `Phytosociological tables') that summarize the traditional classification and subsequent matching of each new relevÅ to the vegetation unit to which it has the highest similarity. More sophisticated methods of supervised classification include quadratic discriminant analysis, multinomial log-linear regression, classification trees, and artificial neural networks. The latter, for example, can establish a classifier based on the previous knowledge of what the relevÅs belonging to a certain vegetation unit look like. When new relevÅs are submitted, the classifier assigns them, with some degree of uncertainty, to the correct vegetation unit.

Another method of supervised classification is COCKTAIL, which was specifically designed to imitate traditional Braun-Blanquet classification. It uses the external information on species behavior, extracted from large phytosociological databases, and forms sociological groups of species with statistical tendency of co-occurrence in the relevÅs of the database. Then, unequivocal definitions of syntaxa are created that involve decision rules, postulating which of the sociological species groups must be present or absent for a particular relevÅ to be assigned to the target syntaxon. COCKTAIL definitions can be created to fit the meaning of the syntaxa of traditional phytosociology. In such a way, traditional syntaxa can be defined formally and applied in the computer expert systems, which automatically assign newly encountered relevÅs to syntaxa.

Phytosociological Tables

In phytosociology, original data and classification results are presented as tables of species by relevÅs or community types. There are two types of phytosociological tables, relevÅ tables (**Table 4(a)**) and synoptic tables (**Table 4(b)**). In both cases, species are listed in the lines and relevÅs (in relevÅ tables) or combined groups of relevÅs (in synoptic tables) in the columns.

Both types of tables are normally presented in a structured manner. Lines and columns are arranged in such a way that `species blocks' (i.e., groups of nonempty table cells) form more or less a diagonal from the top left to the bottom right. Therefore, the diagnostic species corresponding to the syntaxa ordered from left to right are to be found from top downward (except for the negatively differentiated syntaxa). In tables representing multi-layered woody vegetation, plant species of upper layers are normally listed at the top of the table to give an impression of stand structure. At the bottom of the table, those species are listed that have no diagnostic value within the respective table. These may be diagnostic species of superior syntaxa or `companions', that is, species that have no diagnostic value for any syntaxon included in the table. Within blocks, species are sorted by decreasing constancy or decreasing fidelity. Species blocks or individual diagnostic species can be highlighted by frames or shadings in the tables; the criteria for doing so are related to species fidelity to syntaxa and should be clearly defined in particular studies.

In synoptic tables, all relevÅs assigned to the same vegetation unit are represented by a single column with constancy values (i.e., the percentage proportion of relevÅs in which the species is present). Constancy values are often presented as classes indicated by Roman numerals (I: 1–20%; II: 21–40%; …; V: 81–100%), but the use of percentages has several advantages, for example, it does allow the application of modern fidelity concepts and merging of different synoptic tables without loss of accuracy. In addition to the constancy values, medians or ranges of the cover-abundance values or fidelity levels may be indicated. It is important to note (though long-neglected in phytosociology) that the calculation and comparison of constancy values does only make sense for plots of the same or similar size, because constancy values are strongly influenced by plot size.

Phytosociological Ranks and Nomenclature

Abstract vegetation units defined by floristic–sociological criteria are termed syntaxa. They are positioned in a hierarchy of different ranks (**Table 5**), which is meant to make the multitude of units manageable and offers the opportunity to vary the conceptual resolution of analysis, maps, and graphs. The association is considered as the basic unit, comparable to species in taxonomy. Ranks below the association level are often used to express edaphic (subassociations and variants), climatic (altitudinal forms), geographic (vicariants or races), structural (facies of dominant species), and successional variation (phases).

Table 4 (a) Worked example (I): RelevÅ table containing three associations of three alliances and two classes of the subalpine heathland and grassland vegetation of the Czech Republic (see **Table 6** for their position in the syntaxonomic hierarchy). Species of the cryptogam layer are marked with `C', the other species belong to the herb layer. Blocks of diagnostic species are shaded. Within blocks, diagnostic species are ranked by decreasing fidelity to the given syntaxon. Fidelity was measured with the phi coefficient of association and was based on the comparison of species occurrences within the syntaxa of this table only; species with $\phi > 0.25$ were considered as diagnostic. As each association belongs to a different alliance, diagnostic species of the associations can be partly considered as diagnostic of the alliances. Companion species are ranked by decreasing constancy within the entire table. Data were taken from the Czech National Phytosociological Database. Species occurring in a single relevÅ are not shown. (b) Worked example (II): Synoptic table based on the same data as **Table 4a**. The numbers in the table are percentage constancies

	(a) RelevÅ table																															*(b) Synoptic table*		
Association	*Junco trifidi-Empetretum hermaphroditi*												*Cetrario-Festucetum supinae*									*Carici-Nardetum*									*J-E*	*C-F*	*C-N*
RelevÅ number	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*	*26*	*27*	*28*	*29*	n = *13*	n = *10*	n = *6*
Diagnostic species of the association *Junco trifidi-Empetretum hermaphroditi*																																
Empetrum hermaphroditum	5	3	3	3	5	4	3	3	3	3	4	4	4	.	.	.	.	+	.	.	.	.	.	.	.	.	.	.	.	100	10	.
Hylocomium splendens (C)	.	.	1	r	.	2	.	.	1	.	.	.	r	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	38	.	.
Vaccinium myrtillus	1	1	2	2	1	2	1	1	2	2	1	2	2	1	+	.	.	+	1	.	.	.	+	.	.	.	+	+	+	100	50	50
Melampyrum sylvaticum	.	.	+	.	.	+	.	.	.	+	.	.	+	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	31	.	.
Pleurozium schreberi (C)	.	.	1	r	.	2	.	.	3	r	.	.	r	.	.	.	.	.	.	.	.	.	.	.	.	.	.	+	.	46	.	17
Polytrichum piliferum (C)	2	+	.	.	.	.	2	+	.	.	2	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	38	10	.
Diagnostic species of the association *Cetrario-Festucetum supinae*																																
Cladonia bellidiflora (C)	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	+	+	.	.	+	.	.	.	.	.	.	30	.
Thamnolia vermicularis (C)	.	.	.	.	.	.	.	+	.	.	.	.	.	.	.	.	+	.	.	+	+	.	.	.	+	.	.	.	.	8	40	.
Diagnostic species of the association *Carici bigelowii-Nardetum strictae*																																
Nardus stricta	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	2	.	+	1	3	4	4	4	4	5	.	40	100
Galium saxatile	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	+	.	2	.	.	.	50
Anthoxanthum alpinum	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	r	.	1	2	.	.	10	50
Deschampsia cespitosa	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	+	.	1	.	.	.	.	33
Festuca rubra agg.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	+	.	.	.	33
Luzula campestris agg.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	+	.	.	.	33
Potentilla erecta	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	+	.	.	.	33
Diagnostic species of the class *Juncetea trifidi*																																
Calluna vulgaris	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2	1	+	1	1	+	2	.	+	2	.	+	1	r	.	.	.	90	50
Bistorta major	.	.	.	.	.	.	.	+	.	.	.	.	.	.	.	1	+	1	+	+	+	1	.	1	+	+	.	+	+	.	.	8	90	50
Agrostis rupestris	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	+	.	1	1	.	.	+	.	.	.	.	.	40	17
Carex bigelowii	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	4	+	.	3	+	.	1	1	1	.	1	+	.	+	.	+	.	70	67
Hieracium alpinum agg.	.	.	.	.	.	.	.	1	.	.	.	.	+	.	.	.	+	2	.	+	+	1	2	1	+	+	1	+	+	+	.	15	80	83
Companion species																																
Avenella flexuosa	+	+	+	1	1	1	2	1	1	+	1	1	1	1	+	1	3	+	.	3	1	2	2	4	+	3	2	.	2	100	90	83
Vaccinium vitis-idaea	.	2	1	2	.	2	.	1	2	1	+	1	2	1	1	.	.	.	+	.	.	.	.	.	+	+	+	2	.	77	30	67
Cetraria islandica (C)	.	+	1	r	.	1	.	+	2	r	2	.	r	.	+	2	.	+	.	.	2	3	+	.	.	.	.	.	.	69	60	.
Festuca supina	.	.	.	1	.	.	.	.	1	.	1	.	.	2	2	3	1	.	3	.	2	+	.	.	.	+	.	1	+	23	70	50

Solidago virgaurea	.	+	+	.	.	.	.	.	.	+	+	.	.	.	.	.	.	+	r	+	.	.	1	+	.	.	+	.	r	31	40	50
Homogyne alpina	.	.	r	.	.	.	.	.	.	r	.	.	.	.	.	+	.	+	.	.	.	+	1	+	.	.	2	1	.	15	40	50
Huperzia selago	.	+	.	.	.	.	1	+	.	.	.	.	.	.	.	+	.	+	+	.	.	1	+	.	.	.	.	.	.	23	50	.
Calamagrostis villosa	.	.	.	.	+	.	.	.	.	.	+	+	.	.	.	.	+	.	.	.	.	.	.	.	.	+	.	r	+	23	10	50
Juncus trifidus	1	2	.	.	.	.	.	3	.	.	.	.	.	.	2	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	23	20	.
Trientalis europaea	.	.	+	.	.	.	.	.	.	+	+	+	.	.	.	.	.	.	.	.	.	.	.	.	.	.	+	.	.	31	.	17
Vaccinium uliginosum	.	.	.	+	.	+	.	.	+	.	.	.	+	.	.	.	.	+	.	.	.	.	.	.	.	.	.	.	.	31	10	.
Racomitrium lanuginosum (C)	.	2	.	.	.	.	.	2	.	.	.	.	.	.	+	.	.	+	.	.	.	.	.	.	.	.	.	.	.	15	20	.
Cladonia arbuscula (C)	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	+	.	.	.	+	.	.	.	.	.	.	.	8	30	.
Cladonia rangiferina (C)	.	+	.	.	.	.	.	.	.	.	.	.	.	.	+	1	.	+	.	.	.	.	.	.	.	.	.	.	.	8	30	.
Dicranum fuscescens (C)	.	1	.	.	.	.	2	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	23	.	.
Polytrichastrum alpinum (C)	.	+	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	15	10	.
Alectoria ochroleuca (C)	.	.	.	.	.	.	.	+	.	.	.	.	.	.	.	.	.	.	.	.	2	.	+	.	.	.	.	.	.	8	20	.
Campanula bohemica	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	+	.	.	.	.	.	+	.	.	.	.	1	.	.	.	20	17
Pulsatilla alpina ssp. *austriaca*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	+	.	.	.	.	+	.	.	.	.	1	.	.	20	17
Pohlia nutans (C)	.	+	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	+	.	.	.	.	.	8	.	17
Polytrichum strictum (C)	.	.	2	.	.	.	.	.	.	r	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	15	.	.
Silene vulgaris	.	.	.	.	.	.	.	.	.	.	.	.	.	+	.	.	.	.	.	.	.	.	.	.	.	.	+	.	.	.	10	17
Cetraria cucullata (C)	.	.	.	.	.	.	.	.	.	.	.	.	.	.	+	.	.	+	.	.	.	.	.	.	.	.	.	.	.	.	20	.
Diphasiastrum alpinum	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	+	.	.	.	.	+	.	.	.	.	.	.	.	20	.
Polytrichum juniperinum (C)	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	+	.	.	.	.	+	.	.	.	.	.	.	.	20	.
Cladonia pyxidata (C)	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	+	.	.	.	.	.	+	.	.	.	.	.	10	17
Veratrum album subsp. *lobelianum*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	r	.	.	.	.	+	.	.	.	.	.	10	17

Table 5 Syntaxonomic ranks whose names are regulated by the International Code of Phytosociological Nomenclature (ICPN)

Rank	*Termination*	*Example (without author citation)*
Class[a]	*-etea*	*Koelerio-Corynephoretea*
Subclass[b]	*-enea*	*Koelerio-Corynephorenea*
Order[a]	*-etalia*	*Phragmitetalia australis*
Suborder[b]	*-enalia*	*Oenanthenalia aquaticae*
Alliance[a]	*-ion*	*Fagion sylvaticae*
Suballiance[b]	*-enion*	*Cephalanthero-Fagenion*
Association[a]	*-etum*	*Cornicularion aculeatae-Corynephoretum canescentis*
Subassociation[b]	*-etosum* or `*typicum*' or `*inops*´	*Cornicularion aculeatae-Corynephoretum canescentis cladonietosum*

[a]Principal rank (obligatory).
[b]Supplementary rank (optional).

Table 6 Worked example (III): Syntaxonomic hierarchy including all principal ranks and full syntaxon names with author citations for the syntaxa presented in Table 4

Class: *Loiseleurio-Vaccinietea* Eggler ex Schubert 1960
  Order: *Rhododendro-Vaccinietalia* Braun-Blanquet in Braun-Blanquet et Jenny 1926
    Alliance: *Loiseleurio procumbentis-Vaccinion* Braun-Blanquet in Braun-Blanquet et Jenny 1926
      Association: *Junco trifidi-Empetretum hermaphroditi* Šmarda 1950

Class: *Juncetea trifidi* Hadač in Klika et Hadač 1944
  Order: *Caricetalia curvulae* Braun-Blanquet in Braun-Blanquet et Jenny 1926
    Alliance: *Juncion trifidi* Krajina 1933
      Association: *Cetrario-Festucetum supinae* Jeník 1961
    Alliance: *Nardo strictae-Caricion bigelowii* Nordhagen 1943
      Association: *Carici bigelowii-Nardetum strictae* (Zlatník 1928) Jeník 1961

Like other fields of biological systematics, syntaxonomy is an open-ended process that is carried out by a large community of independent researchers and requires unequivocal rules for naming classification units. Therefore, the Nomenclature Commission of the International Association for Vegetation Science (IAVS) and the FÅderation Internationale de Phytosociologie (FIP) have established the International Code of Phytosociological Nomenclature (ICPN), similar to the nomenclature codes used in botanical and zoological taxonomy.

The ICPN regulates the scientific nomenclature of four principal and four supplementary ranks of syntaxa. Neither synusial nor symphytosociological units (see the section entitled `Symphytosociological approaches'), nor informally named syntaxa (e.g., *Elymus repens* community) fall under the ICPN. The ICPN provides precise instructions for the formation of syntaxon names, their valid publication, and the decision about which of several available names from the earlier literature to apply. According to the ICPN, every syntaxon of a certain circumscription and rank has only one correct name. However, the ICPN only regulates the nomenclature and does not define rules for proper delimitation and classification of syntaxa. Aiming to provide unambiguity and stability of syntaxon names, the ICPN is based on two major principles: (1) among several names for a syntaxon, the oldest valid (published) name is the correct one (priority); (2) each syntaxon name is connected to a nomenclatural type (a single relevÅ for associations, a validly described lower-rank syntaxon for higher syntaxa), which determines the usage of the name when this syntaxon is split off, merged with others, or otherwise changed in its delimitation.

Syntaxon names are formed of the scientific names of one or two (in the case of subassociations, up to three) plant species or infraspecific taxa, which usually are, but need not be, characteristic in the respective vegetation type. The formation of the scientific syntaxon names involves connecting vowels, the declination of the taxon epithets, and addition of terminations indicating syntaxonomic rank (Table 5). An `author citation' (i.e., the author(s) and year of the first valid publication) also forms part of the complete syntaxon name (see Table 6).

Other Levels of Classification

Synusial approaches

While phytosociological classifications are usually based on all plant species occurring in vegetation stands, for some purposes sampling may be restricted to certain taxonomic, functional, or structural parts of these. Abstract types of such partial communities are called synusiae (singular: synusia) in order to differentiate them from normal community types (syntaxa) (Table 7). Synusiae include plant assemblages of horizontally differentiated microhabitats within larger vegetation stands, of vertical vegetation layers, and of seasonally separated phenological phases. Epiphytic cryptogams inhabiting tree bark are a typical example of a synusia, which is recorded in small plots with cover projection estimated perpendicular to the substrate surface. Synusiae should be placed in a separate hierarchical system with ranks of their own and the union as its basic unit. However, many studies of partial

Table 7 Levels of classification from synusial phytosociology to sigmasociology

Concrete object	*Elements recorded in relevés*	*Abstract type*
Partial vegetation stand (e.g., layer)	Species	Synusia
Vegetation stand (phytocoenosis)	Species	Syntaxon
Vegetation stand (phytocoenosis)	Synusiae	Coenotaxon
Vegetation mosaic (tesela)	Syntaxa or coenotaxa	Sigmataxon
Landscape mosaic (catena)	Sigmataxa	Geosigmataxon

communities place their units in the system of syntaxa, leading to the ambiguous situation that the same name can refer to both a synusia and a syntaxon.

Symphytosociological approaches

While plant communities and partial communities are assemblages of plant species and their individuals, symphytosociological units are assembled of synusiae or syntaxa (Table 7) and represent a coarser view of community diversity. Sampling and classification basically follow the phytosociological method, but use synusiae or syntaxa (fine-scale vegetation types) instead of plant species as objects of observation. Two major concepts fall in this category and may be combined: (1) `Integrated synusial phytosociology' of some French authors classifies separate `associations' for tree, shrub, herb, and cryptogam layers, which in the normal terminology would be synusiae. These `associations' are recorded in relevÅs of entire stands, analyzed like species in normal phytosociological tables, and such relevÅs are then classified to form so-called coenotaxa. (2) Sigmasociology records syntaxa (or coenotaxa) in large relevÅs of uniform macrotopography, substrate, and climate (tesela), which are tabulated and classified to form sigmataxa, which at a yet coarser scale (catena) become the elements of landscape units called geosigmataxa.

Applied Phytosociology

Ecological Assessment

The study of species–environment and community–environment relations is the key to the functional interpretation of plant communities and to applications of phytosociology in bioindication and predictive modeling. Since the time of Braun-Blanquet, many relevÅs have been made in conjunction with measurements of soil, topographic, and climate variables. If relevÅ coordinates are known, environmental variables can also be *post hoc* read from maps or modeled from geodata. Environmental data are of multivariate nature, which requires condensing their information content and choosing the most meaningful variables. As in species-by-relevÅ matrices, the dimensions of environmental variation can be reduced by extracting continuous gradients (ecological factors) or by forming clusters (site types). Relationships with the environment can be established for community types or species.

While often restricted to verbal descriptions and simple outlines of schematic correspondences (e.g., vegetation type–soil type) in early phytosociology, vegetation type–environment relationships are nowadays studied based on measured variables. These data enable to establish environmental envelopes, which define the possible occurrence of each vegetation type in ecological space. The overall significance of environmental differentiation between types can be tested, for example, by nonparametric permutation procedures (MRPPs).

There is a long tradition in phytosociology of defining ecological groups of species that exhibit similar behavior along gradients and represent species of similar realized niche (`ecological amplitude') rather than fundamental niche (`physiological amplitude'). Such groups are mainly based on expert knowledge and are only partly calibrated on independent measurements of environmental variables. Also, the derivation of ecological indicator values of plant species strongly relies on phytosociological descriptions of vegetation patterns, from which the principal ecological gradients are extracted. While separate species group systems and indicator value systems have been devised for vegetation of arable fields, grasslands, and forests, Heinz Ellenberg created a general, semi-quantitative system of indicator values for the central European flora, in which most species of vascular plants, bryophytes, and lichens are assigned a value on an ordinal scale, ranging from 1 to 9 and representing the estimated ecological optimum with respect to the principal factors light, temperature, continentality, moisture, soil reaction, nutrient availability, and salinity. Being unique in summarizing the niches of an entire flora, Ellenberg values are widely used for calibrating ecological conditions based on plant communities. The concept of plant indicator values has been recently adapted for use beyond central Europe.

Vegetation Maps

Information on spatial distribution of syntaxa is often summarized in vegetation maps. Maps of actual vegetation show the current distribution of vegetation types in a given area, usually in small areas of particular interest, such as nature reserves. Fine-scale mapping of actual vegetation requires operational definitions of syntaxon boundaries and their differential floristic and structural features, which are laid down in detailed mapping keys.

For mapping larger areas, the concept of potential natural vegetation (PNV) is often used. PNV is hypothetical vegetation that would exist at certain sites under current site conditions and current climate, provided the vegetation is not disturbed by humans

and is allowed to develop into equilibrium with the prevailing site conditions. Being based on the knowledge of the relationship between habitat and natural vegetation, PNV maps implicitly or explicitly rely on models, which can take different forms. Traditional phytosociology establishes the correspondence between actual (e.g., certain meadow or weed communities) and natural vegetation (e.g., certain forest types), and maps PNV units by interpreting actual vegetation. More modern PNV models are calibrated from joint descriptions of vegetation and site conditions of remnant natural stands, and use combinations of site conditions to extrapolate natural vegetation for any point in the landscape. Process-based models predict the outcome of competition between the dominant plant species, but have so far rarely been used to construct PNV maps.

While maps of actual or potential vegetation provide full coverage of a study area and its vegetation units, selective maps show the distribution of certain syntaxa, based on the available relevÅs. They can be presented as dot maps of exact plot positions or as grid maps, indicating presence or absence of the syntaxon in grid cells. As, however, information on distribution of syntaxa is often less comprehensive than on plant species, the potential range of a syntaxon can be modeled by superimposing distribution maps of its diagnostic species. The more the number of these co-occur in a certain area, the higher the probability to find the respective community type there. Models of potential syntaxon ranges can be based on outline or grid maps and on simple or weighted sums of species, but the prediction value is best for high-resolution grid maps where the contribution of diagnostic species to the prediction of a syntaxon is weighted by their fidelity to the latter.

Spatial models of syntaxon distribution can also be based on the knowledge of the relationships between environmental variables (including land use) and syntaxon occurrence. If digital maps of environmental factors and landscape structures relevant to plant distribution are available, the model can be made with the probability of syntaxon occurrence as a response variable and a set of landscape variables as predictors. The relevant environmental maps are then overlaid in a GIS and the probabilities of syntaxon occurrence predicted by the model are mapped.

Monitoring Temporal Change

As spatially and temporally explicit, detailed representations of vegetation, phytosociological relevÅs and maps are appropriate tools for monitoring change in plant species composition and the underlying environmental conditions. Thus, fine-scale monitoring systems in agriculture, forestry, nature conservation, and civil engineering have used repeated phytosociological relevÅs at permanently marked locations over many decades, which allow us to analyze trends in diversity of species and species groups (such as Ellenberg indicators or plant functional types). Besides detecting gradual changes, phytosociology expresses succession as a change of community types. Where many permanent plots conform to the same rules, succession can be generalized into temporal gradients and/or sequences of community types (seres). However, many phytosociological succession models have been based on comparative observation (space-for-time substitution) rather than real time series. Larger groups of old relevÅs without permanent marking are sometimes used to detect succesional trends by making new relevÅs in the supposed old positions (`quasi-permanent plots') and by detecting systematic differences between old and new data.

Repeated mapping may reveal changes in the spatial delimitation of vegetation units and allow representation of succession in a transition matrix. However, its validity crucially depends on fully operational mapping keys that unequivocally define the criteria for drawing boundaries between types.

Nature Conservation

The conservation of species depends on the maintenance of their habitats. Habitat classifications can be founded on structural or abiotic features, but they are often based on syntaxa, conveying a summary of ecosystem properties that are difficult to measure or model. Preserving the diversity of extant plant communities is thought to safeguard the survival of typical species not only of plants, but also of animals, fungi, and microorganisms, and the maintenance of current ecosystem processes.

In Europe, phytosociological units were important in defining habitats (biotopes) in the CORINE and EUNIS systems, which contain a comprehensive classification of European habitats. The CORINE classification provided the basis for inclusion of habitat types under the Habitats Directive of the European Union, the most powerful legislative instrument for nature conservation in Europe. In the Union-wide conservation network Natura 2000, phytosociologically defined habitat types are crucial for the delimitation, inventory, monitoring, and management of protected areas.

In landscape planning and policy making, phytosociological units are used to underpin normative judgments and set conservation priorities by evaluating their naturalness and endangerment. Naturalness, or its reciprocal concept, hemeroby, ranks communities by the strength of human influence and consequent alterations of species composition, structure, and ecological processes. Methodologies range from assigning community types to classes of naturalness to complex evaluation schemes taking detailed account of community features.

Reporting the degree of threat to the habitats of a region, red lists of plant communities are another potentially powerful policy tool in nature conservation. Compilation of red lists presupposes a comprehensive and well-established phytosociological classification for the target region, including detailed knowledge about distribution, commonness, and temporal trends of syntaxa. With the advent of phytosociological databanks and GIS, red list compilation is moving from pure expert judgment to a process driven by relevÅ data and rule-based decisions on the vulnerability and conservation value of plant communities. While vulnerability considers current distribution, quantitative development in the past, and foreseeable threats in the future, conservation

value may be based on the frequency and status of component red-listed plant species, naturalness of the inhabited sites, and responsibility of the target region for the global preservation of a syntaxon. The combination of vulnerability and conservation value may be used to set reasonable priorities for conservation measures.

Further Reading

Barkman, J.J., 1990. Controversies and perspectives in plant ecology and vegetation science. Phytocoenologia 18, 565–589.
Berg, C., Dengler, J., Abdank, A., Isermann, M. (Eds.), 2001–04. Die Pflanzengesellschaften Mecklenburg-Vorpommerns und ihre Gefährdung, 2 vols. Jena: Weissdorn.
Braun-Blanquet, J., 1964. Pflanzensoziologie – Grundzüge der Vegetationskunde, 3rd edn. Vienna: Springer.
Bruelheide, H., 2000. A new measure of fidelity and its application to defining species groups. Journal of Vegetation Science 11, 167–178.
Chytrý, M., Otýpková, Z., 2003. Plot sizes used for phytosociological sampling of European vegetation. Journal of Vegetation Science 14, 563–570.
Chytrý, M., Tichý, L., Holt, J., Botta-Dukát, Z., 2002. Determination of diagnostic species with statistical fidelity measures. Journal of Vegetation Science 13, 79–90.
Dengler, J., 2003. Archiv naturwissenschaftlicher Dissertationen 14: Entwicklung und Bewertung neuer Ansätze in der Pflanzensoziologie unter besonderer Berücksichtigung der Vegetationsklassifikation. Nümbrecht: Galunder.
Dierschke, H., 1994. Pflanzensoziologie – Grundlagen und Methoden. Stuttgart: Ulmer.
Ellenberg, H., Weber, H.E., Düll, R., *et al.*, 1992. Scripta Geobotanica 18: Zeigerwerte von Pflanzen in Mitteleuropa, 2nd edn. Göttingen: Goltze.
Ewald, J., 2001. Der Beitrag pflanzensoziologischer Datenbanken zur vegetationsökologischen Forschung. Berichte der Reinhold-Tüxen-Gesellschaft 13, 53–69.
Gillet, F., Gallandat, J.-D., 1996. Integrated synusial phytosociology: Some notes on a new, multiscalar approach to vegetation analysis. Journal of Vegetation Science 7, 13–18.
Mucina, L., SchaminÅe, J.H.J., Rodwell, J.S., 2000. Common data standards for recording relevÅs in field survey for vegetation classification. Journal of Vegetation Science 11, 769–772.
Rodwell, J.S., SchaminÅe, J.H.J., Mucina, L., *et al.*, 2002. Rapport EC-LNV 2002/054: The Diversity of European Vegetation – An Overview of Phytosociological Alliances and Their Relationships to EUNIS Habitats. Wageningen: National Reference Centre for Agriculture, Nature and Fisheries.
Weber, H.E., Moravec, J., Theurillat, J.-P., 2000. International code of phytosociological nomenclature, 3rd edn. Journal of Vegetation Science 11, 739–768.
Whittaker, R.H. (Ed.), 1973. Handbook of Vegetation Science 5: Ordination and Classification of Communities. The Hague: Junk.

Plant Ecology☆

James C Hull, Towson University, Towson, MD, United States
Howard S Neufeld, Appalachian State University, Boone, NC, United States
Frank S Gilliam, University of West Florida, Pensacola, FL, United States

Introduction

Three fundamental properties differentiate the ecology of plants and animals: resource use, mobility, and patterns of growth. Plants are photoautotrophic (except for parasitic species), gaining energy solely from light, and take up essential nutrients (e.g., nitrogen) mostly in inorganic forms. Plants are sessile, lacking mobility. Finally, plants exhibit indeterminate, modular growth, meaning that they continue to grow their entire life span, and do so via expansion of modules (e.g., roots, stems, leaves) that often operate independent of each other. Individual plant success is dependent upon its arrival as a seed (or some vegetative propagule, such as a rhizome) at a suitable location and upon its ability to occupy that site and tolerate local conditions of the environment. Avoidance of predation (herbivory) by plants is dependent upon mechanical, chemical, and life cycle characteristics. In contrast, animals get their energy from other organisms, move within and between habitats to maintain a suitable habitat and, in addition to mechanical and chemical deterrence, depend upon behavior for predator avoidance as well as for mating. In addition, being sessile means plants must obtain light energy for photosynthesis by producing organs with a large surface area to volume ratio, such as occurs with leaves, to harvest light energy, and extensive root systems with which to withdraw water and nutrients from the soil. In contrast, animals have a relatively small surface area to volume ratio, and internalize their absorptive tissues within their bodies through extensive folding (i.e., intestinal villi, alveoli in lungs), where they are protected against the vicissitudes of the environment, such as desiccation and wind.

Plants have specific environmental requirements. For example, they require light, nutrients, water, carbon dioxide, and oxygen. Each of these is a resource needed for growth, survival (maintenance), and reproduction. Acquisition of resources, along with variation in availability of these resources determine the presence or absence of plants at specific sites. Additionally, tolerance ranges of plants to environmental variables such as temperature, wind, physical soil properties, soil pH, fire, salinity, atmospheric humidity, and the presence or absence of herbivores and dispersers all affect the ability of plants to colonize and utilize the resources at a site. It is often the extremes of these factors, such as record low or high temperatures or early and late frosts, that determine the persistence of a plant in a particular habitat, and not the average or range of values. It may only take one cold snap to extirpate a species from an area, whereas a species might persist for many years even if the optimal level of resources is reduced, but not by enough to cause death. Finally, resources are dispersed over time and space, which provide additional niche dimensions in which plants are found.

This entry will examine the properties of the principal resources of plants, the factors affecting the availability of resources to plants, and the response of plants to these variations. Finally, how these resources influence the interactions of plants to form dynamic biotic communities will be discussed.

☆*Change History*: March 2018. H Neufeld updated all text, references, and new figures. Copyright © 2014 Elsevier Inc. All rights reserved.

Plant Life History Traits and the Principle of Allocation

Virtually all organisms are limited in their access to essential resources, especially energy, regardless of form, and plants are no exception. The principle of allocation states that organisms have limited resources to allocate to three main functions required for success of a species: growth, survival (maintenance), and reproduction. The wide variety of life history patterns among plants (e.g., annuals, biennials, perennials) is the result of evolution selecting for optimal allocation of essential resources. Annual plants (e.g., crabgrass) initially allocate resources to growth, but only long enough to produce flowers and seeds, after which essentially all resources are allocated toward reproduction. In sharp contrast, long-lived perennials (e.g., oak trees) allocate more resources toward maintenance (consider their large woody stem), with allocation toward reproduction confined mostly to growing seasons with high resource availability (Fig. 1).

Light

Light is the energy resource of plants captured through the process of photosynthesis, but it also affects plants in myriad other ways. The light used in photosynthesis is roughly comparable to the visible portion of the spectrum for humans (i.e., from 400 nm, the blue wavelengths, up to 700 nm, the red wavelengths) and is expressed as the *photosynthetic photon flux density* (*PPFD*) in units of μmol photons m^{-2} s^{-1}. PPFD at sea level on a bright summer day is approximately 2000 μmol photons m^{-2} s^{-1}. Alternatively, beneath a forest canopy the PPFD can range from a low of only 5 up to 200 μmol photons m^{-2} s^{-1} (0.3%–10% of full sunlight) although in most forests it tends to be less than 140 μmol photons m^{-2} s^{-1} (7% of full sunlight). Most plants reach their maximum photosynthetic rates at PPFDs between 150 and 600 μmol photons m^{-2} s^{-1} although some plants, especially certain grasses like *Zea mays* (corn) do not saturate even at full sunlight. Shade plants tend to saturate at lower PPFD than do sun plants, which is one factor contributing to their ability to persist in deep shade.

At lower irradiances, light is limiting to photosynthesis and plant growth. Plant species with high irradiance requirements are called *sun plants*, while those with lower light requirements are shade plants. Sun plants photosynthesize at greater rates in high light, but are often unable to maintain themselves under low-light environments. Conversely, *shade plants* do well in low light, but can be

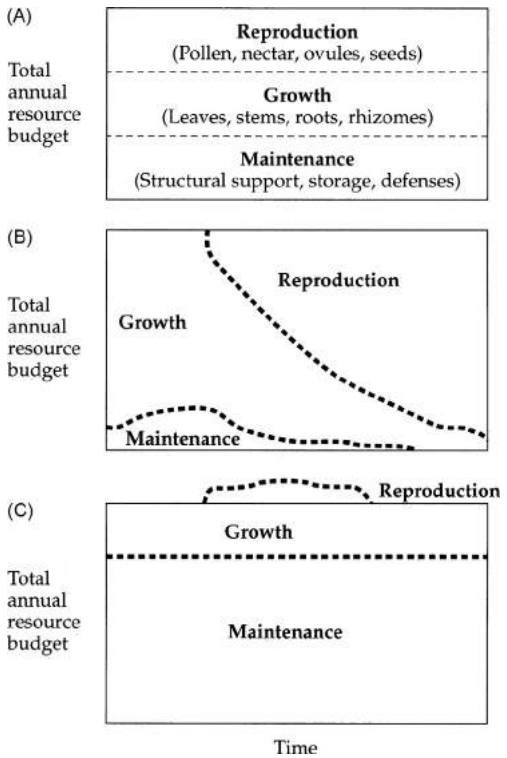


Fig. 1 Three patterns of life history among flowering plants illustrating variation in allocation of resources toward basic plant functions of reproduction, growth, and maintenance. (A) A hypothetical situation wherein allocation is equal to all three; (B) an annual plant with allocation initially toward growth, followed by reproduction; (C) long-lived perennial (i.e., tree) with disproportionate allocation of resources chronically toward maintenance, with allocation toward reproduction confined to periods of high resource availability. Redrawn from Barbour, M.G., Burk, J.H., Pitts, W.D., Gilliam, F.S., and Schwartz, M.S. (1999). *Terrestrial Plant Ecology*, 3rd edn., Menlo Park, CA: Addison Wesley Longman, Inc.

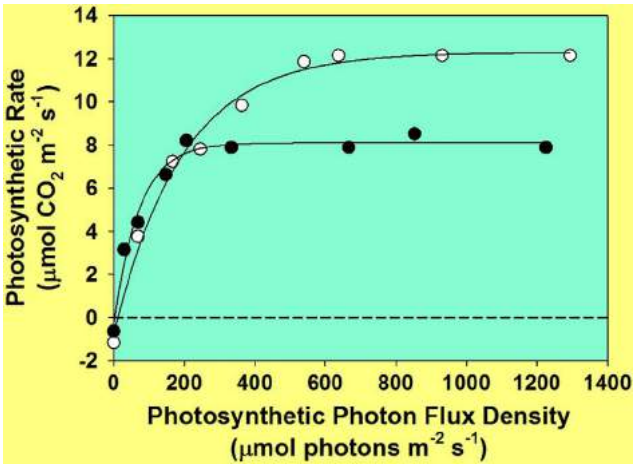


Fig. 2 Curves of net photosynthesis (PS_{net}) as a function of light intensity (PPFD) for a sun plant (*Helianthus annuus*, sunflower, open symbols) and a shade plant (*Oxalis rubra*, wood sorrel, filled symbols). Positive values for PS_{net} indicate uptake of CO_2 into the leaf, whereas negative values indicate release of CO_2 from the leaf, such as would occur if the leaf was only respiring (e.g., in the dark). Net photosynthesis is simply the difference between uptake (gross photosynthesis, PS_{gross}) and respiration (R): $PS_{net} = PS_{gross} - R$. The dashed line at zero photosynthesis indicates when $PS_{gross} = R$. Sun plants have higher maximum rates of PS_{net} in high PPFD, and greater rates of R in the dark than shade plants. However, shade plants often have higher quantum efficiencies at low light as indicated by a steeper slope of the photosynthetic curve below 200 μmol photons m^{-2} s^{-1}. This means that shade plants achieve a greater increase in PS_{net} per photon absorbed than sun plants, which is mainly a function of the fact that shade leaves are thinner and have less self-shading. The light compensation point is the PPFD where $PS_{net} = 0$ (where PS_{net} crosses the dashed line) and is lower in shade plants than sun plants. This, coupled with lower R, may contribute to the ability of shade plants to persist in low light habitats. Conversely, sun plants can achieve higher rates of photosynthesis and outcompete shade plants in full sun habitats. Adapted from data in Figures 3 and 10 in Böhning, R.H. and Burnside, C.A. (1956). The effect of light on rate of apparent photosynthesis in leaves of sun and shade plants. *American Journal of Botany* **43**, 557–561.

inhibited by high light (Fig. 2). Shade plants have thinner, broader leaves which increase the efficiency of light capture, while sun leaves are thicker and often contain an extra layer of cells with which to capture light for photosynthesis (Fig. 3).

On cloudless days, light is largely directional while on overcast days, it is mostly diffuse and appears to come from all directions. Competition for light, and in particular, direct light, is mediated by differences in vertical plant size and angular distribution of leaves. Shade plants frequently display their leaves in ways that minimize overlap (i.e., a mono-layer canopy such as that shown by flowering dogwoods (*Cornus florida*), while sun plants have an architecture composed of multiple leaf layers, which maximizes light interception (e.g., pines, *Pinus* sp., and birch trees, *Betula* sp.). The upper leaves of sun plants are often more vertically oriented while the lower ones are more horizontal. This allows more light to penetrate farther into the crown, maximizing light interception by the plant. These physiological and structural differences restrict shade plants to understory habitats and sun plants to more open habitats.

Not only do plants respond to the quantity of light, but they can also detect the quality of that light. In full sun, for example, the ratio of red to far-red light is approximately 1:1, whereas in the understory, it is substantially reduced, often to as low as 0.1:1. This is because the green leaves of the canopy primarily absorb the red light and allow more of the far-red light to penetrate to the understory. Plants have special pigments, such as *phytochromes* and *cryptochromes*, which respond to the lower ratio of red to far-red light. This can stimulate height growth in shade-intolerant plants and serve as a mechanism to avoid prolonged shading. Some shade-intolerant species, such as black cherry (*Prunus serotina*) have seeds that germinate only when the *red: far-red ratio* is near one, because this indicates an absence of an overstory canopy, and thus conditions more suitable for germination and subsequent survival.

Plants also differ with respect to their tolerance to the periodicity of light. In virtually any canopy environment, some light passes through the canopy relatively unaffected by leaves and branches. This creates a *light fleck* (or *sunfleck*) on the surface below it. These light flecks differ considerably in their duration and energy content. In one study in Great Smoky Mountains National Park, 80% of the sunflecks lasted less than 1 min each, but contributed nearly 80% of the daily PPFD even though their cumulative duration totaled only 6 h. Most sunflecks last less than 120 s, yet the 5% that are longer can contribute up to 75% of the total radiation input to the forest understory. The intensity of light received in these sunflecks is proportional to their size and duration: larger and longer sunflecks have higher PPFDs.

The photosynthetic apparatus of plants requires the prior presence of light to react quickly to a sunfleck, a phenomenon called photosynthetic induction. Understory plants have a low irradiance requirement to remain induced, while sun plants require higher light to maintain induction. Even though sunflecks are present for only a small fraction of the time, they contribute toward a substantial amount of photosynthetic carbon gain by shade plants. In tropical forests, sunflecks may contribute from 30% to 60% of the daily carbon gain, whereas it is a lower percentage (10%–20%) in temperate forests.

Another response to periodicity reflects changes in day length accompanying seasonal changes in temperate regions. For many plants the length of the photoperiod controls flowering. Short-day plants will initiate flowering when day length gets shorter than

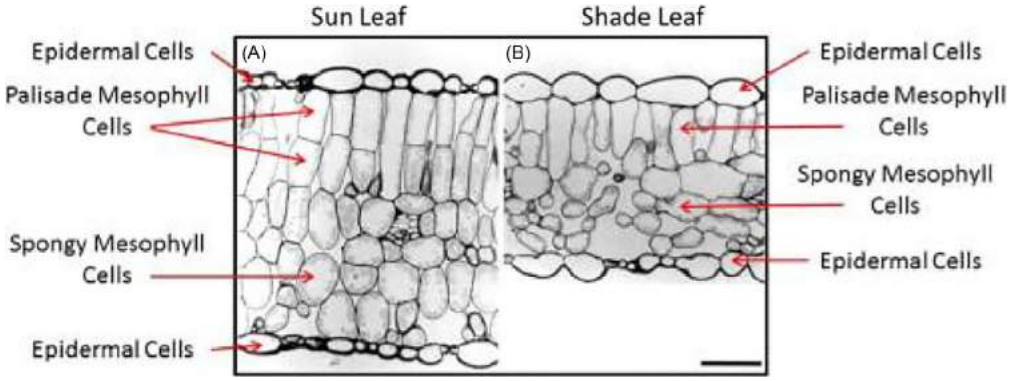


Fig. 3 Light micrographs of cross-sections of (A) sun and (B) shade leaves of *Chenopodium album*. Sun leaves were grown in PPFD of 360 μmol photons m^{-2} s^{-1}, and shade leaves in 60 μmol photons m^{-2} s^{-1}. The bar in the lower right is 50 μm in length. Note the double layer of palisade mesophyll cells in the sun leaves, and the greater thickness of the sun leaf compared to the shade leaf, which has only one layer of palisade mesophyll cells. The *light gray* objects arrayed within and on the periphery of the mesophyll cells are chloroplasts. From: Yano, S. and Terashima, I. (2001). Separate localization of light signal perception for sun or shade type chloroplast and palisade tissue differentiation in *Chenopodium album*. *Plant Cell Physiology* **42**, 1303–1310.

some critical period (or more properly, as the night gets longer, since plants respond more to what happens at night than what occurs during the day). Alternatively, long-day plants (i.e., short-night plants) will flower when day length exceeds a critical period. A new hypothesis suggests photoperiodic control of flowering may be linked to internal circadian rhythms in plants. The ability to properly time flowering enables plants to complete their reproductive cycle before adverse environmental changes limit their activity. In the prairies of North America grasses must initiate flowering early enough to complete their reproductive cycle before the first killing frost. For the same species in the north, flower initiation occurs in June, while in the south flowers are initiated in October. The two populations are ecotypes (genetically different populations of the same species), and the differences in flowering time are maintained when they are grown together in the same environment.

Water

The availability of water controls plant distribution on scales ranging from microhabitats to continents. Precipitation, temperature and nutrient availability are the primary factors that distinguish the various biomes of the world and which exert controlling influences on their annual productivity. Terrestrial *annual net primary productivity* (*ANPP*) is measured as g C produced m^{-2} $year^{-1}$ and ANPP is the difference between *annual gross primary productivity* (*AGPP*), or how much C is assimilated by plants per m^2 $year^{-1}$, and how much is annually respired (R_a) away: ANPP = AGPP − R_a. Annual evapotranspiration (AE) is a measure that integrates precipitation and temperature and is positively correlated with ANPP across terrestrial biomes (Fig. 4).

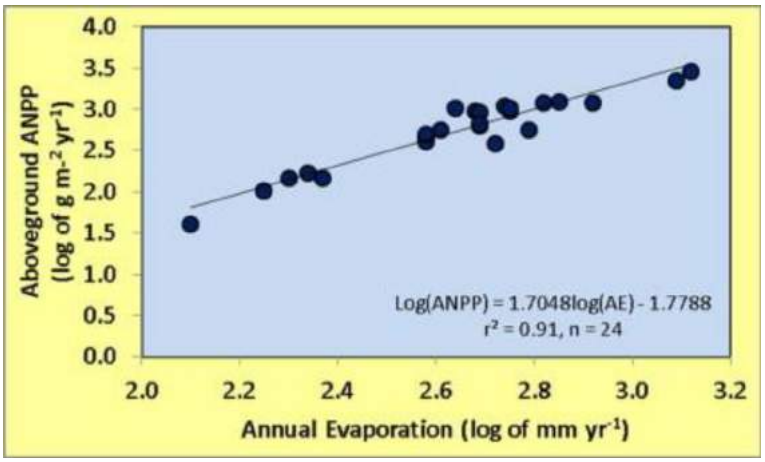


Fig. 4 Relationship between the log of annual evapotranspiration (AE) and annual above ground net primary productivity (ANPP). The original units for ANPP and AE are g biomass m^{-2} $year^{-1}$ and mm H_2O m^{-2} $year^{-1}$, respectively. AE is the amount of water that enters the atmosphere from the combined activities of evaporation from plant and soil surfaces and that which is transpired by plants. AE integrates the availability of both water and solar energy, which is why it is so highly correlated with ANPP. The more available water, when combined with abundant solar energy to drive photosynthesis, leads to greater ANPP in ecosystems. Figure Drawn from data in: Rosenzweig, M.L. (1968). Net primary productivity of terrestrial communities: Prediction from climatological data. *American Naturalist* **102**, 67–74.

Water in plant systems is best described using concepts of thermodynamics as exhibited by the concept of water potential (Ψ_w). This value, derived from the natural gas laws, expresses the energy contained in water in units of pressure (e.g., MegaPascals, MPa; 0.1 MPa = 1 bar = ~0.9869 atmospheres = 750 mmHg = 14.5 psi at sea level). The Ψ_w of pure water is defined as 0 MPa. Numerous physical and chemical factors affect Ψ_w, including solutes (Ψ_s), matric (i.e., surface) attributes (Ψ_m), pressure (Ψ_p), and gravity (Ψ_g). The addition of solutes to water lowers the free energy of water in that system by altering hydrogen bonding between water molecules and by lowering their kinetic energy. Both of these changes result in a lowered capacity for that solution to do work compared to pure water, which is why water containing solutes always has a negative Ψ_s. For example, a 1.0 m solution (g solute/kg water) of a nondissociating solute will create a $\Psi_s = -2.5$ MPa. If this solution is separated from pure water by a semipermeable cell membrane, then a potential energy gradient will exist from the water to the solution.

This energy gradient means that water will move from areas of high potential energy (pure water) to areas of low potential energy (solution) following basic principles of thermodynamics. Since cells contain numerous solutes, water will move down the potential gradient into the cell via osmosis resulting in an increase in cell volume. This increase in volume is quickly constrained by the cell wall, which creates a positive pressure on the water in the cell. The end result is an increase in the Ψ_p of the cell. Ultimately the Ψ_p (which is positive) would partially or totally offset the Ψ_s (which is negative), and the cell would reach an equilibrium Ψ_w of 0 MPa. For example, a plant rooted in a wet, saturated soil, which would have a Ψ_{soil} near zero, would have an equilibrium Ψ_{plant} also near zero. Total plant water potential is simply the sum of all the components of water potential described above: $\Psi_{plant} = \Psi_s + \Psi_p + \Psi_m + \Psi_g$.

Matric potential (Ψ_m) is most important component in determining soil water availability, but is only a minor component of Ψ_{plant}. The surfaces of soil particles (sand, silt, and clay) provide a strong attraction to water and limit the ability of plants to absorb this water. As soils dry down, the Ψ_{soil} begins to decline and becomes more negative due to matric effects that result from the strong adsorption of water (via hydrogen bonds)to the soil particles.

Water usually moves from the soil, then to the root, stem and leaf before eventually evaporating into the atmosphere. It does so by moving passively down a Ψ gradient from soil to atmosphere, which is known as the Soil-Plant-Atmosphere Continuum, or *SPAC*. With few exceptions, this is a one way gradient, which is why water moves primarily from soil to plant to atmosphere. This occurs because the atmosphere has a very negative Ψ (e.g., at 50% relative humidity and 20°C, $\Psi_{air} = -93.6$ MPa) while a dry soil might be only −0.5 MPa. Plants effectively act as giant wicks, moving water from the relatively higher Ψ in the soil, to the very low Ψ in the atmosphere (Fig. 5).

A temperate forest tree, transpiring at its maximum rate of about 10–15 mol H_2O m^{-2} s^{-1} might have a Ψ_{leaf} of −1.5 MPa. In contrast, a creosotebush (*Larrea tridentata*) growing in the dry Chihuahuan Desert, may have a much lower $\Psi_{leaf} \sim -7$ MPa. This shows the great range in Ψ_{plant} possible among species adapted to habitats with different availabilities of water. Friction in the xylem conduits, through which the transpirational water moves, and the inability of the root system to provide water at the same rate at which it is lost from the leaves, leads to this drop in Ψ_{plant} from the much higher Ψ_{soil}. It is this drop in Ψ that induces movement of water from the soil into the roots, while the large gradient from the leaf to the atmosphere insures that water evaporates out of and not into, the leaf.

The upward movement of water in the xylem of plants is widely accepted as occurring by cohesion-tension forces, a theory first proposed in 1895 by the Irish botanist Henry Horatio Dixon and the physicist John Joly. Water lost by evaporation in leaves is replaced by water pulled up by the strong cohesive forces between water molecules. Because more water is lost than can be replaced by root uptake, and because of frictional resistances in the xylem, the water column is placed under tension (negative pressure) and becomes metastable.

Pressure potentials within living cells are typically positive; however, the pressure potential of the water moving up the plant through its xylem is negative, and frequently referred to as xylem potential (Ψ_x). It is easily measured using a pressure chamber and serves as a proxy for plant water status. The potential energy of water in the xylem column is also affected by gravity and Ψ_x decreases with height by 0.01 MPa m^{-1}. Thus a fully hydrated tree, 50 m in height, would have a maximum Ψ_x of −0.5 MPa instead of 0 MPa due to the impact of Ψ_g. Maximum tree height (*c.* 130 m) may even be a function of the decreased Ψ_{plant} with height because this decreases turgor, increases the chances of embolisms, and ultimately reduces photosynthetic carbon gain at the tops of tall tree crowns.

Plants have a physiological limitation for water absorption based upon their ability to lower osmotic potentials of root cells. This limitation is described as the permanent wilting point (PWP), which is an expression of the water content of a soil when a plant is no longer capable of extracting water from it. The Ψ_{plant} at PWP varies with species, but for mesophytic, agricultural plants is approximately −1.5 MPa. At PWP in sandy soil the remaining water may be less than 1% by dry weight, but in clay soil the water content may be nearly 20% by weight. Small clay particles attract water more tightly than larger sand particles making it more difficult for roots to extract.

Water use by individual plants varies considerably depending on plant size, availability of soil water, and atmospheric temperature and humidity. A large tree may transpire as much as 1000 L H_2O day^{-1} whereas a desert cactus may lose less than 1 g day^{-1}. Transpiration ratio (TR) is a measure of water loss to carbon gain by plants, typically expressed as mass of water lost to mass of plant weight gained (mol H_2O/mol CO_2). This value will vary between plants and the photosynthetic pathways they possess (discussed below). Plants with the C3 photosynthetic pathway (most trees, wheat, for example) have a TR of 666–2000 and have the least efficient TR; those with the C4 pathway (maize, sorghum) have an intermediate TR of 500–1000; while those with the CAM pathway (cacti, pineapple) have the most efficient TR of 100–250. If calculated in terms of the demand for water

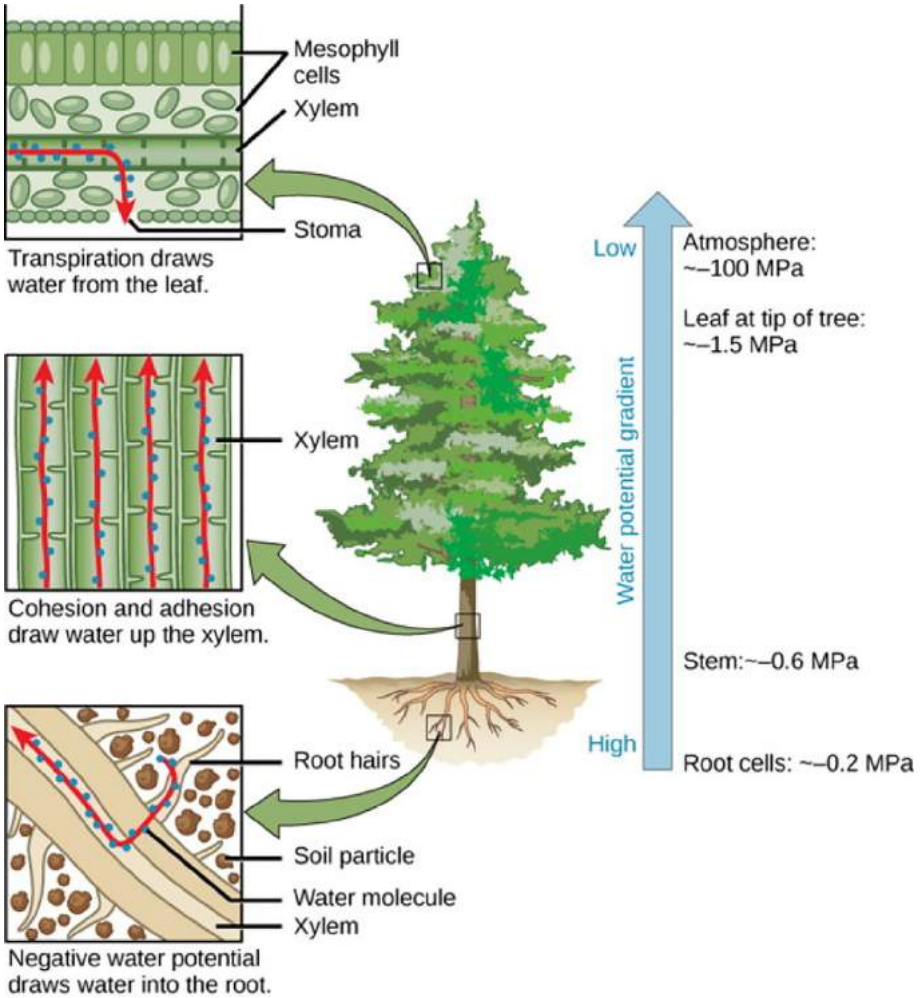


Fig. 5 The Soil–Plant–Atmosphere-Continuum (SPAC). Water moves passively down a free energy gradient (i.e., down the Ψ_w gradient) as it moves from the soil into roots, then stems and leaves, before evaporating into the atmosphere. Because the atmosphere has such a low Ψ_w, the gradient is unidirectional. In essence, plants act as giant wicks, removing water from the soil and transpiring it back to the atmosphere. Frictional resistances in the xylem, and an evaporative demand that exceeds the capacity of the roots to take up water, result in a lowering of the Ψ_w along the pathway through the tree. Figure courtesy of Pearson publishing.

(Mg H_2O ha^{-1} $year^{-1}$), which includes both the fraction taken up by the plant and that evaporated from the soil back to the atmosphere, per unit biomass produced (Mg biomass (tonnes) ha^{-1} $year^{-1}$), the differences in the ratios become quite dramatic. For C3 plants, the ratio ranges from 400 to 1200, for C4 from 286 to 571, while for CAM it is 60–150. These differences account, in part, for the relatively higher proportion of C4 and CAM plants in hot, arid regions where water is frequently limiting to plant growth. The very high water use efficiencies of C4 and especially CAM plants is one reason why scientists are attempting to incorporate these alternative photosynthetic pathways into C3 crop plants, so future agricultural water use can be reduced.

The water status of plants varies both seasonally and diurnally. Stomata usually open early in the morning and close at dusk. Some plants may close their stomata in the afternoon because of drought stress resulting from an inability to match absorption to transpiration rates. This results in a decrease in mid-day photosynthetic capacity, often termed the mid-day depression (Fig. 6). Similarly, on a seasonal basis, when evapotranspiration rates have depleted soil water, low Ψ_{soil} makes water uptake difficult. As a consequence Ψ_{plant} decreases, resulting in stomatal closure during much of the day, and preventing the occurrence of catastrophic drought stress.

The velocity of water movement in *xylem vessels* and *tracheids* is proportional to the fourth power of the radius of the conduits. Thus a simple doubling of conduit diameter can result in a 16-fold increase in flow, which means that larger xylem elements will move most of the water in a plant. The first vascular plants evolved tracheids to conduct water. These relatively primitive xylem cells have end walls and narrow conduits, and for water to move from one cell to another must pass through circular structures known as pits, which have thinner cell walls than the rest of the cell surface. The most recently evolved plants, the angiosperms, or flowering plants, have evolved shorter, but wider xylem cells known as vessel elements. These cells are joined end to end to form a conducting tube, with end walls only occurring occasionally. Because of the lack of end walls, and their much wider diameters, water flows more easily through vessels than tracheids, and angiosperms have a higher hydraulic conductivity than more primitive vascular plants,

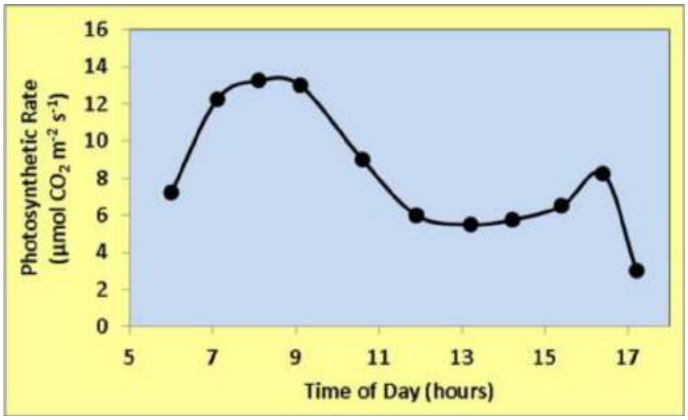


Fig. 6 Mid-day depression in photosynthesis for sun leaves of the understory herb *Arisaema heterophyllum*. The mid-day depression starts just after 10 am and is maximal at about 1 pm (13 h) in the afternoon. The depression is initiated by the closing of the stomata due to excessive transpiration, which in turn reduces the CO_2 concentration inside the leaf. Low internal CO_2 directly leads to lower photosynthetic rates, but it also favors relatively more oxygenation of the RUBISCO molecule and higher rates of photorespiration, further contributing to lower photosynthetic rates. Later in the afternoon, when transpirational demands lessen, stomata open and photosynthetic rates rise temporarily before dropping again, this time due to low light late in the day. Figure redrawn from: Muraoka et al. (2000). Contributions of diffusional limitation, photoinhibition and photorespiration to midday depression of photosynthesis in *Arisaema heterophyllum* in natural high light. *Plant, Cell and Environment* **23**, 235–250.

like gymnosperms, as a result. Hydraulic conductivity is simply the ease with which a plant can move water through its xylem. Tracheid diameters in some conifers, such as redwood trees may only be 20 μm, whereas in angiosperm lianas, such as kiwi (*Actinidia deliciosa*) and grapevine (*Vitis vinifera*), which have very large vessels up to 500 μmin diameter), their hydraulic conductivities are among the highest for any plants.

Because the water in the xylem of a transpiring plant is under tension and metastable, large negative pressures developing within the xylem can result in *cavitation*, which is the formation of air bubbles in the xylem column. Cavitation can lead to *embolisms*, which are breaks in the water column that result in the loss of conduction from that point on in the xylem.

Embolisms form when air is drawn into a xylem element because of a pressure difference across the cell wall (Fig. 7). Xylem under large tension can aspirate an air bubble from outside the cell and pull it through pit pores that separate the xylem cell from a neighboring air space. Once inside the cell, the negative pressure allows the bubble to rapidly expand, resulting in the embolism and cessation of flow. The larger the pit pores, the more easily a bubble can be aspirated (a consequence of surface tension), and susceptibility to embolisms is partially correlated with pit pore sizes.

Freezing can also result in the formation of embolisms due to the release of gases as the temperature of the xylem decreases (note how gas bubbles form in ice cubes as the water freezes). Subsequent thawing under tension results in the expansion of these trapped air bubbles and the formation of an embolism. The larger the xylem conduit, the more susceptible the plant is to such freezing-induced embolisms. As a consequence, lianas, which have the largest diameter vessels, are effectively absent from colder habitats, such as occur at high elevations and latitudes. Instead, they reach their peak abundance in tropical habitats where freezing events are absent. Conversely, the small conduits of tracheids found in gymnosperms afford protection from this type of embolism and may account in part for their prevalence in such habitats. Most angiosperms contain a mixture of large diameter, rapidly conducting vessels, and narrow diameter, slower conducting, tracheids, and vessels. The combination of these xylem elements provides a measure of hydraulic safety for plants.

Nutrients

Plants generally require the same nutrients, although their absolute and relative requirements may vary. While nutrients have traditionally been categorized into two classes (macro- and micronutrients) depending upon their proportion in the plant body, this division has been difficult to justify from a physiological standpoint. A recent proposal divides nutrients into four classes based on their biochemical and physiological functions in the plant. These are nutrients such as (1) N and S, used in organic compounds, (2) P, Si, B, which are used for structural integrity and energy storage, (3) K, Ca, Mg, and others which are used in ionic form and, (4) Fe, Zn, Cu, and other metals, which are involved in redox reactions.

Some nutrients are used in high amounts, and can constitute more than 1% of the weight of plants. These include C, H, O, N, P, S, Ca, Mg, and K. Carbon, H, and O are incorporated into plants through photosynthesis and respiration while the remaining nutrients are absorbed from the soil, although some, such as N, can be absorbed in gaseous form on occasion. Most of the other nutrients make up less than 1% of a plant's weight, such as Fe, Mo, Cu, Zn, Mn, and B. Although plants may contain Na, Cl, Se, Ni, and Si, it is not always clear whether these elements are required. Nonetheless, some of these may contribute to the survival of

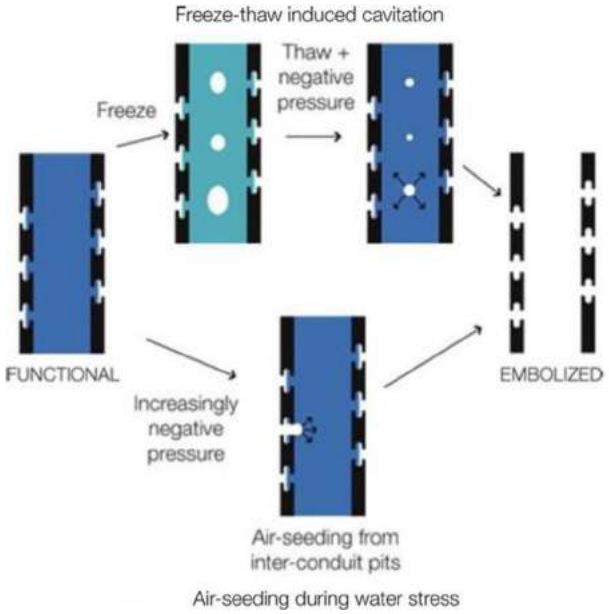


Fig. 7 Cavitations can arise from either air-seeding or freezing in xylem conduits. The functional conduit on the left is a xylem element containing water under tension. The conduit on the right is an embolized xylem element, with no water inside. The upper two xylem elements illustrate freezing-induced cavitation while the bottom one shows cavitation caused by air-seeding. With air seeding, the pressure differential across the xylem cell wall is so great that an air bubble is aspirated from an intercellular airspace outside the xylem, or from an adjacent, embolized xylem element, and pulled into the functional conduit through the pit pores. Pit pores are spaces between the cellulose microfibrils where the secondary cell walls are absent and can facilitate intercellular movement of water. Because functional xylem is under negative pressure (tension), and the airspace outside is at atmospheric pressure, the air bubble is directed inwards toward the functional conduit. If the pressure differential becomes large enough, it overcomes the surface tension of the air bubble, enabling it to be pulled into the conduit. This action constitutes the cavitation event. Once inside the xylem cell, the low xylem pressure allows the bubble to rapidly expand and form an embolism. Once embolized, the xylem conduit can no longer conduct water to the leaves. With freezing-induced embolisms, bubbles form as the sap freezes. When the conduit thaws under tension, the bubbles rapidly expand and embolize the xylem element. Figure taken from: Cain, M.L., Bowman, W.D. and Hacker, S.D. (2011). *Ecology*, 2nd edn, Sunderland, MA: Companion website by Sinauer Associates.

certain plants. Silicon is found in high amounts in grasses and horsetails (*Equisetum* sp.) while other plants accumulate nickel to very high amounts (*hyperaccumulators*), both of which may be strategies to deter herbivores. Sodium is thought to be required for plants that perform C4 photosynthesis.

Soil nutrient availability may be divided into two groups: (1) nutrients whose availability is largely controlled biologically, for example, N, and (2) those whose availability is largely controlled abiotically, for example, P. The nitrogen cycle consists of biological processes involving incorporation of atmospheric nitrogen into ammonia or ammonium nitrogen (fixation), release of organic nitrogen into the soil solution (mineralization), and conversion of ammonia into nitrate (nitrification), which is the form of nitrogen most commonly absorbed by plants. Each of these steps is mediated by soil microbes that control the rate of availability of N to plants. Soil factors such as oxygen content and pH have important effects on the availability of N because of effects on soil microbes and root functioning.

Successional processes (the changes in plant communities through time) can also affect the partitioning between nitrate and ammonium nitrogen. Apparent nitrate uptake decreases during succession while ammonium uptake increases, which are mechanisms by which late successional communities conserve nitrogen and energy. Since nitrate is a negatively charged anion it is repulsed by the negatively charged soil particles and is lost from ecosystems after disturbances. Ammonium, on the other hand, is a positively charged cation and is retained in the soil. Furthermore, less energy is required by plants to build nitrogen containing organic compounds from ammonium than from nitrate.

Phosphorus is most commonly available to plants as phosphate and is ultimately derived from the substrate. The availability of phosphorus is controlled by pH and the abundance of other ions that immobilize it in the soil. Calcium, silicates, and oxides of aluminum and iron bind phosphorus at different soil pH resulting in low soil availability. Phosphorus availability to many plants is enhanced by mutualistic relationships between plants and mycorrhizal fungi. These fungi derive carbon energy from the host plant and have hyphae that extend through the mineral soil and absorb large quantities of phosphorus (as well as other nutrients such as N, also C and even water), which are shared with the host plant. These mutualistic associations arose as soon as plants began colonizing terrestrial habitats, perhaps in response to the extremely limiting amounts of available phosphorus on land before the formation of deep, nutrient rich soils. So, despite the abiotic control of phosphorus availability in the soil, the absorption by plants of phosphorus at low concentrations is often dependent upon an association with mycorrhizal fungi.

Mycorrhizae can even transfer nutrients and carbon among individual trees whose roots are connected by the same hyphae. Suzanne Simard and colleagues used ^{13}C and ^{14}C isotopes to follow the movement of carbon from one tree to another through

interconnected hyphae for individuals of Douglas fir (*Pseudotsuga menziesii*) and across species boundaries to paper birch (*Betula papyrifera*). Trees in full sun served as "donor" trees and provided C to the trees in the shade ("receiver" trees). This illustrates that there are complex biological interactions taking place belowground among plants, fungi and soil fauna that have strong influences on aboveground processes.

Nutrients present in excess amounts in the soil can result in toxicity symptoms in plants. In theory, any element may be present in excess, including nitrogen, but the most common forms of nutrient toxicity arise from salinity and heavy metals. Salinization of soils is a worldwide problem that results from high evaporation rates after irrigation in arid regions. Evaporation concentrates salts in the upper soil layers, where it can reach toxic levels. There also biological processes that concentrate salt. The invasive saltcedar (*Tamarix* spp.) transpires copious amounts of salt-laden water that it absorbs from deep within the soil (Fig. 8). The result is the accumulation of salts in its leaf tissues, which it then exudes from special salt glands. The salt then falls off the tree and accumulates in the upper layers of the soil to the detriment of surrounding plants, since the salt is both toxic and makes water uptake difficult because it decreases ψ_{soil}.

Excess N deposition may cause plant diversity to decrease in various ecosystems. Preindustrial N deposition rates may have been as low as ~2 kg N ha^{-1} $year^{-1}$, whereas today in Europe and the eastern United States, this represents <10% of rates in the eastern United States and <1% of the maximum rates in Europe. Mechanisms for the decline in diversity include altering interspecific competition to favor nitrophilic species over a diverse N-efficient flora, increasing degree of herbivory, decreasing success of mycorrhizal infections, increasing severity of pathogenic fungal infections, and enhancing success of invasive species. Nitrogen deposition may also reduce *carnivory* in pitcher plants by shifting production away from pitchers, which are the leaves that capture prey, to *phyllodes*, which are leaves that only do photosynthesis.

Lastly, acid deposition can cause the leaching of cations from the soil, creating nutrient deficiencies for certain plants. Acid deposition (as either rain, fog, dew or snow) is the result of nitrogen and sulfur compounds being added to the atmosphere through combustion processes such as coal burning, and their transformation to nitric and sulfuric acids. These acids compete for binding sites on soil particles and cause leaching of Ca^{+2}, K^{+}, Mg^{+2}, and other cations. The loss of these ions, and in particular, Ca^{2+}, can result in an inability of some trees, particularly red spruce (*Picea rubens*) to withstand extreme cold periods, and may be a cause of forest decline in northern areas. Soil acidification also increases the solubility of heavy metals, such as Al^{3+}, which is extremely toxic to roots. Acidic deposition may have leached so much Ca^{2+} out of the soils at the Hubbard Brook site in New Hampshire, United States, where nutrient cycling patterns have been studied since the 1960s, that forests there are unable to increase their biomass from year to year.

Adaptive Plant Strategies

One of the goals of any science is to develop predictive theories on which future research can be based. This is how science progresses, which distinguishes it from faith based initiatives, where the conclusions are reached without reference to either data or experimentation. In plant ecology, scientists continue to search for particular adaptive (i.e., evolutionary) plant strategies that may be used conceptually to increase our knowledge about how plants interact to form communities and ecosystems. If a set of

Fig. 8 Salt Cedar (*Tamarix* sp.) is an exotic invasive tree that currently occupies over 1.6 million acres of land in North America alone. It is highly competitive with native species, often displacing them from riverine habitats. It can extrude salt from glands in its leaves and raise the salinity level of soil so much so that native plants are excluded. It also has very high transpiration rates which deplete soil water and further enhance its competitive abilities. Photo courtesy of Dr. Jonathan Horton.

functional traits can be found that have broad predictive capabilities, this would eliminate the necessity of studying each of the ~225,000 species of plants, which would be an essentially impossible task. In this section, various adaptive plant strategies are examined, but it should be kept in mind that this list is by no means complete, and that the very concept of functional traits is still a matter of much discussion among ecologists. Also, the reader should rest assured that in applying the concept of "strategy" to plants, we are not implying that they have either intent or foresight to prepare for the future. Rather, adaptive plant strategies are the means by which plants come to be competitively successful through the exploitation of various physiological and structural attributes, each of which have evolved as the result of natural selection.

Plant Functional Groups—Alternative Photosynthetic Pathways

All higher plants assimilate CO_2 through the process known as C3 photosynthesis, named for the fact that the first stable product is a 3C sugar, 3-phosphoglycerate, or 3-PGA. In addition, however, two alternative photosynthetic pathways have evolved; the C4 pathway, named for the fact that the first stable product of fixation is a 4C organic acid and the crassulacean acid metabolism pathway, or *CAM* as it is known (named after plants in the Crassulaceae family). Both alternative pathways incorporate a new primary fixation event involving another enzyme, phosphoenolpyruvate carboxylase (or PEPc) before shuttling the fixed carbon to the C3 apparatus. These new pathways can be viewed as "add-ons" to the more evolutionarily primitive C3 pathway. The C4 pathway involves a spatial separation, whereas CAM involves a temporal separation, of additions steps.

The evolution of CAM preceded that of C4, and is restricted to vascular plants, whereas C4 has evolved only in the angiosperms. CAM plants make up about 6%–7% of the flora, whereas C4 plants comprise just 3%. Both groups, however, are ecologically important in ways that belie their taxonomic rarity; CAM plants predominate in desert and epiphytic habitats, whereas C4 plants thrive in warm, dry areas, but with notable exceptions in both cases.

During the process of C3 photosynthesis, light energy is used to fix CO_2 in the chloroplasts to create a stable 3C compound, known as a triglyceride, which gives this pathway its name. All plants are derived ultimately from the green algae, which also perform C3 photosynthesis, and it is the evolutionarily most primitive pathway in plants. The fixation process is catalyzed by an enzyme known as *RUBISCO*, which stands for ribulose-1,5-bisphosphate carboxylase oxygenase. Having evolved long before global oxygen levels reached their current levels, it is probably an accident of evolution that it can also fix O_2 in addition to CO_2. Under ambient conditions, about one in four fixation events is an oxygenation instead of a carboxylation, resulting in an efficiency of carboxylation of ~75%. Oxygenation adds no new C to the plant, but instead results in the loss of CO_2 through a process known as photorespiration.

Beginning ~100 MYa, atmospheric CO_2 levels began declining and during Pleistocene glaciations (~2 MYa) may have ranged between 180 and 200 ppm. For comparison, preindustrial concentrations were 270–290 ppm, whereas current levels (2018) now hover around 409 ppm, an increase almost solely due to human activities especially the burning of fossil fuels. These low prehistoric atmospheric CO_2 concentrations would have reduced the gradient for diffusion into the leaf, making photosynthesis difficult for C3 plants. Since CO_2 and O_2 competitively inhibit each other for binding sites on RUBISCO, the low CO_2 levels would have also favored a relative increase in the rate of oxygenation. Beginning around 35 MYa fossil leaves of plants structurally different from typical C3 plants are found, constituting the first evidence for the evolution of a new photosynthetic pathway known as C4 photosynthesis. These plants are categorized by their ability to concentrate intercellular CO_2 around RUBISCO to enhance photosynthesis when atmospheric concentrations are low.

C4 photosynthesis evolved independently at least 66 times in 19 families of angiosperm plants (of which four, Poaceae, Cyperaceae, Chenopodiaceae, and Amaranthaceae account for most lineages). While C4 plants constitute only about 3% of all plant species (~7500 species), they account for nearly 25% of all annual net primary productivity worldwide (mainly in the large grasslands such as the savannas and prairies) making them extremely important ecologically. In addition, some of our most important crops are C4, such as maize (*Zea mays*), sugarcane (*Saccharum officinarum*), and sorghum (*Sorghum bicolor*).

C4 leaves differ structurally from those in C3 plants (Fig. 9). C3 plants have an upper layer of long columnar cells (palisade) and a lower layer of irregularly spaced and shaped cells (spongy) where photosynthesis occurs. Leaves of C4 plants, on the other hand, are divided into mesophyll cells and another set of large cells surrounding the vascular bundles known as bundle sheath cells. This type of leaf anatomy is called *Kranz anatomy* ("Kranz" means "wreath" in German), and was known to plant scientists in the late 19th century. However, the link with C4 photosynthesis was not realized until the mid-20th century.

PEPc is kinetically faster than RUBISCO and insensitive to O_2. It fixes CO_2 onto a 3C pyruvate molecule in the mesophyll cells to form a stable 4C compound, usually oxalic or malic acid. The malic acid is then shuttled to the bundle sheath cells where the recently fixed CO_2 is then removed from the organic acid and subsequently fixed via RUBISCO and the C3 pathway (Fig. 10). To put it simply, C4 is an "add-on" to the C3 pathway. The end result of this process is that it functions to "pump" CO_2 to the bundle sheath cells, where it can build up to concentrations nearly $10\times$ those in ambient air. This shifts the competitive equilibrium between carboxylation and oxygenation on RUBISCO in favor of the former, essentially eliminating photorespiration. As a consequence, rates of photosynthesis can be substantially higher than those found in C3 plants.

Although C4 photosynthesis may have originally evolved as a way to cope with declining atmospheric CO_2 concentrations, any ecological situation that resulted in a substantial loss of carbon from photorespiration probably served as a necessary selection pressure for this photosynthetic pathway. In habitats where high salinity, temperature, or drought would cause stomatal closure and a lowering of the internal CO_2 concentration, there would be more photorespiration and evolution of the C4 syndrome would be favored.

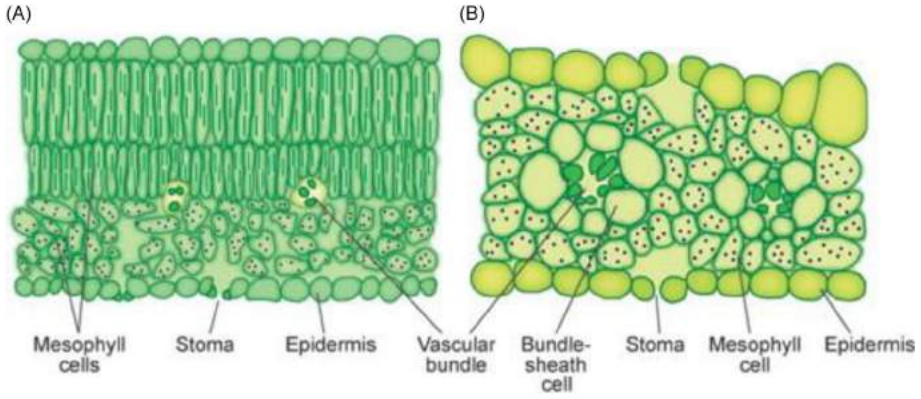


Fig. 9 Idealized cross-sections of (A) C3 and (B) C4 leaves. The enlarged bundle sheath cells surrounding the vascular bundle in C4 leaves give rise to the anatomical formation known as Kranz anatomy. The term "Kranz" refers to the wreath-like arrangement of the cells around the vascular strand. In C3 plants, CO_2 is assimilated in both the palisade and spongy mesophyll cells via the C3 pathway, using the enzyme RUBISCO to perform the fixation reaction. In C4 plants, CO_2 is initially fixed in the mesophyll cells by phospho-enol pyruvate carboxylase (PEPc), and a four carbon organic acid is then transported to the bundle sheath cells, subsequently decarboxylated, and the resultant CO_2 fixed by RUBISCO via the C3 pathway. Any photorespired CO_2 is re-assimilated by the mesophyll cells before it can escape the leaf, which is one reason for the absence of any apparent photorespiration in C4 plants. Figure courtesy of: Tipple, B.J. and Pagani, M. (2007). The early origins of terrestrial C4 photosynthesis. *Annual Review of Earth and Planetary Sciences* **35**, 435–461.

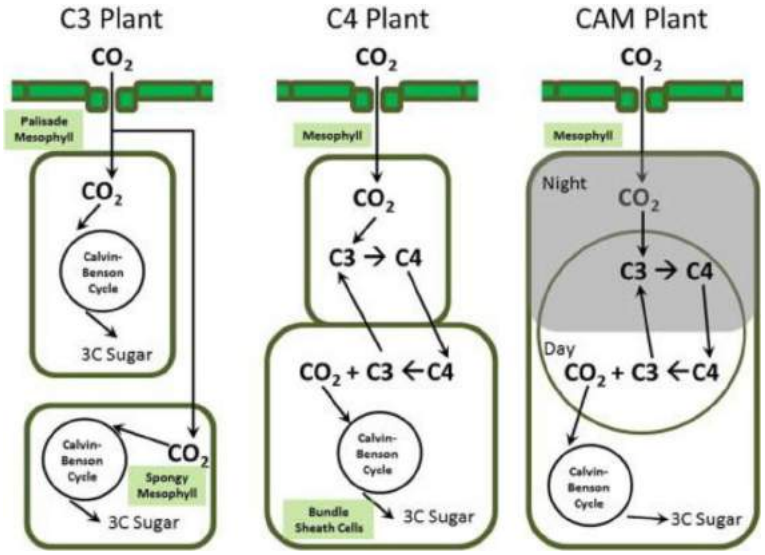


Fig. 10 Schematic comparisons of the three major photosynthetic pathways in plants. C3 plants fix CO_2 onto a 5C sugar (ribulose-1,5-bisphosphate) in both palisade and spongy mesophyll cells using the enzyme RUBISCO. The resultant 6C compound splits into two 3C sugars, which constitute the first stable products of photosynthesis and which give the pathway its name. C4 plants perform an initial fixation of CO_2 onto a 3C compound (phosphoenolpyruvate) by PEPc that is found only in mesophyll cells, to form a 4C organic acid. This acid is then shuttled to the bundle sheath cells where the CO_2 is removed. The 3C pyruvate is recycled back to the mesophyll cells while the CO_2 is fixed by RUBISCO via the C3 pathway, which is localized only to these specific cells. The CAM pathway is very similar to the C4 pathway, but both fixation reactions occur in the same cell, although at different times. The initial fixation by PEPc occurs at night, whereas the decarboxylation and re-fixation by RUBISCO occur during the day using the energy of the sun to power these reactions. By opening their stomata at night, CAM plants achieve the highest water use efficiency of any of the photosynthetic pathways. NOTE: Original artwork by authors.

The lack of photorespiration shifts the temperature optimum for photosynthesis higher for C4 compared to C3 plants. C3 plants reach maximum photosynthesis in a broad plateau centered around 25°C, whereas in C4 plants it is 35°C or higher. The higher photosynthetic rates of C4 plants allow them to assimilate more C per unit of water taken up, which is why their water use efficiencies are so much higher than for C3 plants (see earlier discussion). This is one reason why C4 plants are common in warmer, drier areas. In more mesic habitats they often reach their peak activity during the drier portion of the year. In the prairies of central North America, C3 species dominate in the cool, wet spring, while C4 species become more common in late summer when it is hotter and drier. Biogeographically, the proportion of a flora that is C4 declines with both latitude and altitude to the point that there are no native C4 plants in Alaska, or in most arctic or alpine areas (Fig. 11).

Lastly, C4 plants discriminate less than C3 plants against heavy stable isotopes of carbon, such as ^{13}C (Fig. 12). C4 plants have a discrimination of between −10 and −18 parts per mil, whereas C3 plants range between −23 and −36 parts per mil, with no overlap. This means that stable isotopes analyses can be used to distinguish which plants are C3 or C4. Stable isotope analysis can be used to analyze annual average water use efficiencies of plants, dietary preferences (you are what you eat), agricultural practices, and the relative productivity of C3 and C4 plants worldwide. Scientists have analyzed the stable ^{13}C isotope ratios of human skeletons from South America and used the change in $^{13}C/^{12}C$ ratios to estimate that after 1000 AD maize (a C4 plant) became a prominent staple food in the diet.

A third photosynthetic pathway, first noted in the family Crassulaceae, and now called crassulacean acid metabolism (or CAM) bears strong resemblance to C4 photosynthesis in that it concentrates CO_2 around RUBISCO, but the separation of the two fixation reactions is temporal instead of spatial and occurs in a single cell. These plants include all the cacti, many succulents, and even pineapples (*Ananas comosus*) which grow in the most xeric environments, such as deserts, as well as a variety of epiphytes like Spanish moss (*Tillandsia usneoides*), an iconic CAM plant that grows on trees in the southeastern United States. CAM plants, for reasons discussed below, can, paradoxically, also occur in aquatic habitats.

The unique distinguishing feature of CAM plants is that they open their stomata at night, at which time CO_2 diffuses in and is fixed by PEPc. The resultant malic acid is stored in the large central vacuole overnight. This requires a large volume of water, which is why many CAM plants have thick, succulent leaves. During the day, the malic acid is decarboxylated and the sun's energy is used to fix the CO_2 into triglycerides using RUBISCO and the C3 pathway. By restricting stomatal opening to the cooler night time hours, these plants greatly reduce transpirational water losses, and as a result, have the highest water use efficiencies of any plants, which in turn, enables them to grow in the driest habitats.

CAM plants are also found in aquatic habitats, particularly small pools of water where the CO_2 concentrations during the day are depleted by the photosynthetic activities of other plants and algae. Plant such as *Isoëtes* sp. take up CO_2 at night when respiratory processes have raised the CO_2 concentrations in the water, and then during the day, they decarboxylate the malic acid and fix the CO_2 into 3-PGA via the C3 pathway.

Plant Functional Traits

Currently, 225,000 species of plants are catalogued, but scientists estimate that there may actually be 298,000 species throughout the world. If we are to understand global patterns in functioning, and use these inputs for predicting community successional patterns and ecosystem functioning, we will need to develop proxies for various species traits, such as photosynthesis, reproduction and growth, since it will be impossible to measure these for each and every species. Plant functional groups, such as the alternative photosynthetic pathway traits discussed earlier, or *Raunkiær's lifeform classifications* based on the location of the perennating bud, are

Fig. 11 The percentage of the grass flora that is C4 plants in North America. The percentage decreases with increases in latitude. There are no native C4 grasses in Alaska, although crabgrass (*Digitaria sanguinalis*), which is C4, has been introduced there now. Similar biogeographical patterns hold for C4 forbs (herbaceous plants that are not grasses). Figure adapted from Teeri, J.A. and Stowe, L.G. (1976). Climatic patterns and distribution of C3 and C4 grasses in North America. *Oecologia* **23**, 1–12 as presented on the web by Forseth, I.N. (2010). The Ecology of Photosynthetic Pathways. *Nature Education Knowledge* **3**(10), 4. http://www.nature.com/scitable/knowledge/library/the-ecology-of-photosynthetic-pathways-15785165.

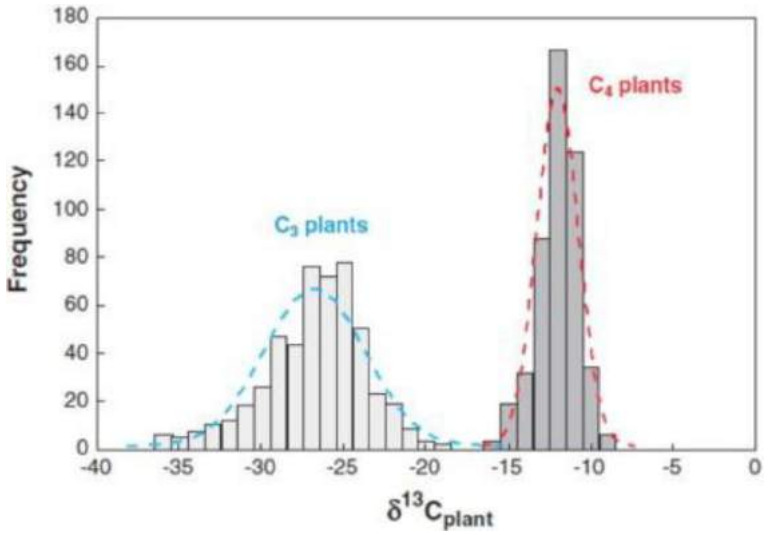


Fig. 12 C3 plants discriminate more against the uptake of ^{13}C than do C4 plants. The *light gray* bars show the discrimination in C3 plants in $\delta^{13}C_{plant}$ units and the *dark gray* bars that by C4 plants. $\delta^{13}C_{plant}$ units are a way of expressing how much less ^{13}C is in a plant relative to a fixed standard, such as Pee Dee belemnite, a clay obtained from South Carolina and which is used to calibrate such measurements. Two things are clear from this graph: (1) C3 plants have a wider range of discrimination than C4 plants and (2) the mean discrimination is much greater in C3 than C4 plants. Also, since there is no known overlap between the two groups of plants, $\delta^{13}C_{plant}$ analysis can be used to distinguish the C3 and C4 photosynthetic pathways of a plant without having to resort to measuring gas exchange or leaf biochemistry. Figure from: Tipple, B.J. and Pagani, M. (2007). The early origins of terrestrial C4 photosynthesis. *Annual Review of Earth and Planetary Sciences* ***35***, 435–461, and the original figure was obtained from data in Cerling, T.E. and Harris, J.M. (1999). Carbon isotope fractionation between diet and bioapatite in ungulate mammals and implications for ecological and paleoecological studies. *Oecologia* **120**, 347–363.

attempts at just such an endeavor. However, these classifications, while useful for distinguishing major plant types across biogeographical distances, ultimately fail when it comes to predicting growth because they contain no physiological mechanisms, and also because structural traits tend to cross over these particular groupings, making the distinctions less useful for modeling purposes. For example, some genera of trees and herbs may have species with thin leaves and high photosynthetic rates, as well as species with thick, evergreen leaves and low photosynthetic rates. Thus, the anatomical classifications are of limited utility for determining those factors constraining physiological functioning, productivity and reproduction.

An alternative strategy is find a suite of traits that is quantitative, continuous, easily measured, and has broad explanatory power without reference to a particular species. Such traits would represent broad-based evolutionary patterns that serve as proxies for how plants respond to environmental selection pressures, and could be incorporated into predictive models, such as, for example, how plants will respond to future global climate change.

A variety of traits have been proposed for just such an effort (Fig. 13). The most successful attempts to date include three major trait dimensions: (1) a leaf economics spectrum, running from cheap (on a carbon basis), short-lived, leaves with high nitrogen content (and consequently high photosynthetic rates), to thicker and denser (i.e., higher mass per unit area), longer-lived leaves with lower nitrogen content; herbs, grasses, and deciduous trees cluster with the former group, while evergreen plants are representative of the latter; (2) reproductive output, whereby species with larger seeds have a lower output per m^2 of canopy, but higher survival rates upon germination and throughout the plant's early life, and (3) final height at maturity, where final height trades off with rapid, early height growth or tolerance to shade.

Newly formed leaves become "profitable" after photosynthesis accrues enough new carbon to pay back the construction costs for that leaf, that is, the carbon that was expended to make new cell walls and membranes, as well as the photosynthetic apparatus. Once these costs are paid back, any additional assimilated carbon can be considered profit, and is exported to other parts of the plant to build new roots, stems, more leaves, and of course, reproductive structures such as flowers and seeds. Thin, less dense leaves require less time to become profitable because of the smaller amount of carbon invested in them, whereas thicker and denser leaves take longer. As a consequence, leaf life spans are shorter where the pay back times are shorter, such as for thin leaves, those with high photosynthetic rates, and those in full sun. Spring vernal herbs, such as trout lily (*Erythronium americanum*) make such leaves. These plants take advantage of a small window of opportunity for photosynthesis that is limited to the period before the overstory canopy leafs out, when most of the carbon assimilation for these plants is accomplished. They have fairly high photosynthetic rates and can assimilate most of their annual carbon in just a few weeks. Once the canopy closes their leaves senesce rapidly and their total lifespan may be only a few weeks. In contrast, evergreen plants, which invest a considerable amount of carbon to make durable leaves, have lifespans that exceed 1 year, and some pines can maintain needles for over a decade. When scientists combine climate parameters with leaf trait variables their models are able to explain nearly 75% of the variation in photosynthetic rates, nitrogen concentrations, respiration rates and specific leaf area ($g\ m^{-2}$ of leaf), all without reference to any one particular species.

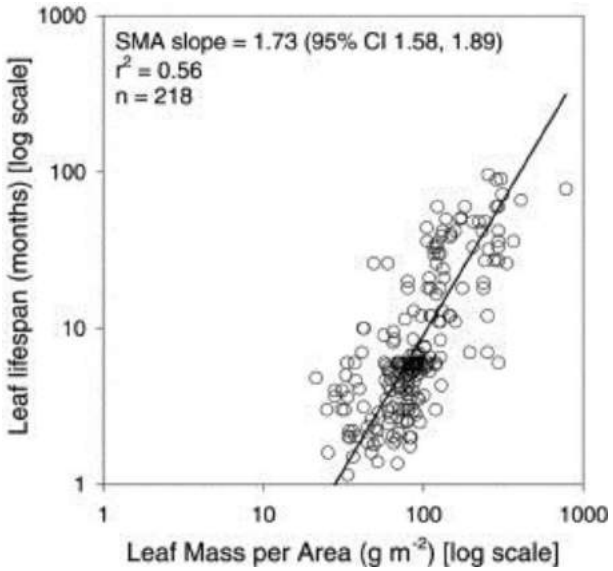


Fig. 13 The lifespan of a leaf is highly correlated with the mass of leaf tissue per unit area of leaf (g m^{-2}). Leaves with a higher mass per area have longer leaf lifespans because they require a longer period of time to pay back their construction costs (i.e., the amount of carbon and nutrients invested in building the new leaf). Once these costs are met, the leaf can then provide a net benefit in terms of carbon to the plant. This is but one of many correlations among various leaf traits that make up what is known as the worldwide leaf economics spectrum. These correlations cross over species and may be useful in developing models of plant functioning without having to make detailed physiological measurements on all known species. Figure from: Westoby, M., Falster, D.S., Moles, A.T., Vesk, P.A. and Wright, I.J. (2002). Plant ecological strategies: Some leading dimensions of variation between species. *Annual Review of Ecology and Systematics* **33**, 125–159; and is redrawn from original data in: Reich, P.B., Walters, M.B., and Ellsworth, D.S. (1997). From tropics to tundra: Global convergence in plant functioning. *Proceedings of the National Academy of Sciences of the United States of America* **94**(13), 13730–13734.

Functional traits other than leaves are also being investigated. They include a worldwide wood economics spectrum, whereby woody plant strategies can be estimated using easily measured traits such as density, which integrates both structural and functional attributes. For example, wood density is an excellent predictor of mid-day water stress and whole tree transpiration rates, independent of any particular species designation. A recent metaanalysis (a metaanalysis is a study of general patterns obtained from a suite of previous independently done studies) showed that the relationship between density and water use was quadratic in form: trees with either low or high wood density had lower transpiration rates than those with an intermediate density. The explanation proposed for this pattern is that trees with low wood density have wider diameter conducting xylem elements that are more susceptible to cavitation. Under stressful conditions these trees close their stomata to restrict water loss, while conversely, trees with high wood density have narrower conduits that limit the maximum rate of water movement up the trunks due to the constraints imposed according to the Hagen–Poiseuille law, where flow is proportional to the fourth power of the radius.

The evolution of potential plant strategies could be restricted by limited genetic variation in particular traits, which would constrain their ability to evolve in the face of strong pressures from natural selection. However, recent analyses of the amount of genetic variation in populations for various functional traits suggests that genetic constraints are less important than natural selection in determining the suite of traits that make up the worldwide leaf and wood economics spectra. The interesting opportunities for future research are to more clearly understand the tradeoffs among adaptive traits that result in the plant forms and functions specific to particular habitats and biomes, and why there is selection against alternative trait syndromes.

Plant Interactions

Interactions between plants are one of the driving forces controlling plant community structure. The nature, importance, and demonstration of these interactions generate considerable discussion among ecologists. Such interactions can result in the elimination of species from communities through the various processes of competition, whereas even very small niche differences may be enough to maintain a high biodiversity of perennial plants, sometimes for several hundreds of years.

Interaction between plants resulting in increased plant growth is called facilitation, which is most frequently viewed as one species making the environment less harsh to which the second species responds favorably. An example of facilitation is hydraulic lift. In arid environments surface soils become too dry to support shallow-rooted species, but deep-rooted species persist. In the cold deserts of North America, big sage (*Artemisia tridentata*) brings water to the surface from deeper layers. Some of this water then diffuses into the soil and supports herbs beneath the shrubs. Another example involves the fixation of atmospheric nitrogen by legumes that colonize young, infertile soils recently exposed by retreating glaciers. The addition of nitrogen to the soils then allows other plants with higher nutrient requirements to establish.

Negative interactions between plants occur when one or both plants grow poorly in the presence of the other. The term "*competition*" is frequently used to describe any negative interaction; however, this use does not fully describe the mechanisms of the interaction. If the interaction is a result of two plants utilizing the same limiting resource such as water, light, or nutrients, then that interaction is called "*resource competition*" or "*scramble competition.*" Typically the species differ in their ability to utilize the resource or to tolerate the result of reduction of the resource. Competition for water was demonstrated in a study of the interaction between the perennial bluebunch wheatgrass (*Agropyron spicatum*) and the introduced annual cheatgrass (*Bromus tectorum*) in the Palouse prairie grasslands of Northwestern United States, where most of the limited precipitation falls as winter snow. Cheatgrass grows in the winter, exhausts the water supply, and survives the summer drought as seed, while the bluebunch wheatgrass grows slowly in winter, experiences low water availability for its summer growth, and dies. This exemplifies the preemptive use of a resource and the tolerance of the result by cheatgrass.

Negative interactions may result if a substance is added to the environment by one plant with a detrimental effect on a second plant. This may be referred to as "*interference competition*" or "*contest competition*." Plants add a wide range of chemicals to the environment that have negative effects and include salts, phenolics, terpenes, alkaloids, mustard oils, and cyanides. If the added material has a negative effect on neighboring plants, then the term "*allelopathy*" is used to describe the interaction. However, such chemicals may also have additional effects in the ecosystem, such as deterring herbivores, altering litter decomposition, and affecting the nitrogen cycle. Chemical interactions among plants is now a vibrant field of study, and may help shed light on such topics as why invasive species often do so much better in their nonnative environment than the one they evolved in.

An early classic study of allelopathy by Cornelius H. Muller in the California chaparral described interaction of an aromatic shrub (*Salvia leucophylla*) with the surrounding annual grassland in which a zone immediately surrounding the shrubs was devoid of grasses, a second zone stretching out an additional 2 m had stunted grasses of a limited number of species, and a third zone was composed of uninhibited grasses, including wild oat (*Avena fatua*) and ripgut brome (*Bromus diandrus*) (Fig. 14). Muller demonstrated that the shrubs produced a variety of volatile terpenes (e.g., cineole and camphor) with the potential to inhibit growth of seedlings. However, later studies showed that the situation was much more complicated and could not be attributed solely to the production of allelochemicals from the *Salvia*. When herbivore exclosures were erected in the bare zones, grasses and other plants began growing there, suggesting that animals were also involved in the maintenance of the bare zones about the shrubs.

More commonly, allelopathic interactions result from water-soluble materials leaching from the canopy or litter or which are actively exuded from roots. In some cases, production of allelochemicals is stimulated by another stress, such as herbivory. When sagebrush is eaten by herbivores, it responds by producing allelochemicals that inhibit the germination of nearby plants. The Novel Weapons Hypothesis was proposed recently to explain biogeographical patterns in the success of the invasive species *Centaurea diffusa*. It states that when species are introduced to a new ecosystem, they may produce novel allelochemicals that inhibit native plants and facilitate their competitive abilities. Despite a long history of research, the demonstration of allelopathy in the field is difficult, since there are so many competing hypotheses to deal with. As such, the phenomenon remains controversial. As stated by Inderjit et al. (2011), "Mere production of chemicals by a plant is not sufficient to ensure their allelopathic potential." To show that a negative interaction is allelopathic, an investigator must demonstrate no other biotic interaction (resource competition or herbivory) is causative, a chemical is produced and released in sufficient quantity and at time to be effective, that the target plant is susceptible to the chemicals, and finally, that microbial transformations in the soil do not modify the allelochemicals in ways that reduce their toxicity.

Fig. 14 Allelopathy purportedly demonstrated in shrubland–grassland interface in Santa Ynez Valley, CA, United States. At left is a stand of purple sage (*Salvia leucophylla*) surrounded by a bare zone with few or no grass species. An inhibition zone of 1–2 m with foxtail fescue (*Festuca megalura* (= *Vulpia myuros*)) and soft chess (*Bromus mollis*) grading into uninhibited grassland of wild oat (*Avena fatua*) and ripgut brome (*B. diandrus*). The aromatic shrubs produce volatile compounds including cineole and camphor, which can be toxic to some of the grasses.

Plant Community Dynamics

Plant communities are assemblages of species co-occurring in a given time and place (see McGill et al., 2006). The organization, development, and repeatability of plant communities in similar habitats and through time entailed considerable discussion and controversy among plant ecologists during the 20th century. Disagreements regarding the structure and functioning of plant communities resulted from attempts to define the composition, stability, and boundaries of communities and as a result, numerous competing theories have been proposed over the years to explain these patterns. The modern synthesis states that the plant community is an assemblage of species' populations aggregated in a region resulting from dispersal mechanisms, physiological tolerances to local site characteristics, and responses to disturbance. Others suggest that new approaches are now necessary to understand the rules governing community structure and how communities may respond to disturbance and climate change.

Plant succession is described as the change in composition of vegetation in one place over time. Succession is viewed as one of two related processes: primary succession or secondary succession. Primary succession is the development of vegetation on substrate previously lacking plants (Fig. 15). Examples include the development of vegetation on sand dunes, glacially exposed substrate, filling in of lakes, etc. The earliest ecological descriptions of primary succession were by H. C. Cowles on the sand dunes of Lake Michigan and by W. S. Cooper on the development of vegetation following retreat of glaciers in Glacier Bay, Alaska. In both, the development of vegetation was viewed as a chronosequence characterized by the predominance of conspicuous plants. Conspicuous early invaders were replaced over time by species that were inconspicuous at the onset or arrived later. An example of such a chronosequence at Glacier Bay is the presence of avens (*Dryas drummondii*), followed by alder (*Alnus tenuifolia*), and subsequently by Sitka spruce (*Picea sitchensis*), a process estimated to take several hundred years.

Secondary succession is the development of vegetation following a disturbance to the original plant community. Disturbances leading to secondary succession include fire, windstorms, and anthropogenic disturbances such as logging and farming (old field succession). These successions occur on substrate that is wholly or partially intact, and consequently they occur more rapidly. The pattern of secondary succession is greatly dependent upon the type of the disturbance. For example, in a study of vegetation development at one location over several hundred years, the outcome was dependent upon whether the disturbance initiating the succession was fire or windstorm. In a classic study of old field succession, Catherine Keever found that the time of year the stand was released from farming affected the sequence of development of early plants. From this arose a body of evidence supporting the view that species' characteristics such as seed dispersal, photosynthetic light requirements, growth rate, and longevity control much of succession. Accordingly, early-successional plants are often rapidly dispersed, require high light, have high photosynthetic rates and growth rates, but are relatively short lived. Species possessing these traits are classified as r-selected. These plants give way in time to species with slower dispersal rates, greater shade tolerance, slower growth rates, but are long lived (K-selected). This concept of tradeoffs among suites of traits between early- and late-successional species frames much of the current discussion of the mechanisms of succession.

A complementary theory of plant strategies was proposed in the mid-1970s by J.P. Grime. He suggested that the dichotomous categorization of plants as early and late successional was too simplistic, and that a third axis, stress tolerance, was necessary to encompass the range of possible adaptive plant strategies (Fig. 16). Grime employed a three axis strategy, often referred to as C–S–R theory: Competitive species (C, comparable to K-selected or late successional species) are found where productivity is high and competition severe; and stress tolerators (S) occur where environmental stress, such as shade or low nutrients, is high and consequently competition is low. Ruderal species (R, comparable to r-selected, or early successional species) inhabit sites with high disturbance rates but where competition is low. The last possible adaptive strategy, which combines all three factors (high disturbance, competition, and stress), was viewed by Grime as an impossible evolutionary outcome, so no plants fell into that portion of the schema.

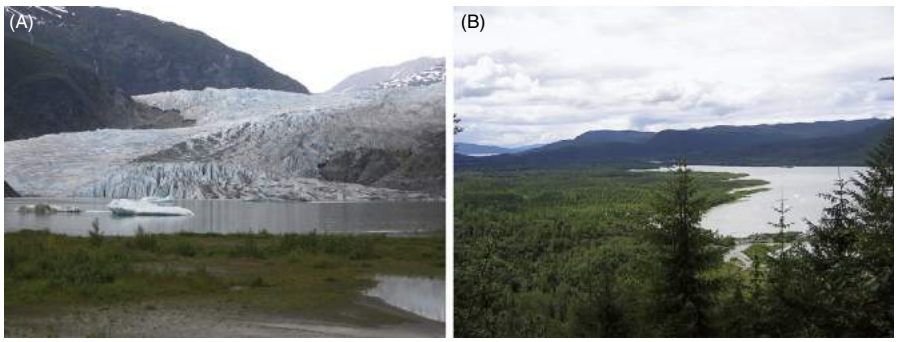


Fig. 15 Primary succession of forest development following retreat of the Mendenhall Glacier, Alaska, United Sytates. (A) Initial stages of vegetation on newly exposed glacial till in the foreground. The site was exposed within the last 20 years, and the vegetation consists of willows (*Salix* sp.), fireweed (*Epilobium* sp.), lupines (*Lupinus* sp.), alders (*Alnus* sp.), scattered Sitka spruces (*Picea sitchensis*), and western hemlocks (*Tsuga heterophyla*). (B) Forest development on glacial till adjacent to Mendenhall Lake. Ice occupied these sites in the mid-1700 s and is currently retreating at a rate of approximately 40 m year^{-1}. The vegetation consists largely of alders, Sitka spruces, and western hemlocks.

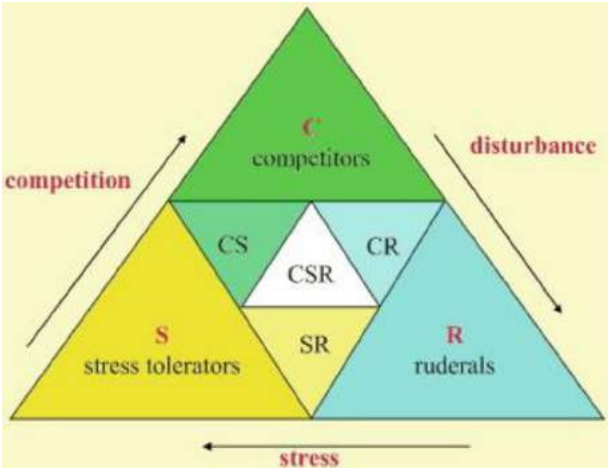


Fig. 16 Philip Grime's C–S–R triangle of plant strategies. Grime proposes that plants must adapt to three major environmental pressures: competition with other plants, habitat disturbance, and stress, caused by either excesses or deficiencies in water, light, nutrients, or herbivory. Early successional plants (often called pioneer or Ruderal plants) would be found primarily in highly disturbed, low stress environments, where competition is low (*right hand corner*). Plants in late successional environments are subject to high amounts of competition but low rates of disturbance (*upper corner*) and are known as competitors. Plants in habitats with high rates of disturbance and stress are most likely subject to low rates of competition, and are termed stress tolerators (*lower left corner*). There are no plants known that are able to withstand high levels of all three pressures, and so there is no fourth plant strategy. Of course, many habitats fall between these extremes, and may contain plants with a mixture of strategies, depending on the relative amounts of each of the three main driving factors (CS, CSR, CR, and SR). Figure from Shiro Tsuyuzaki, Hokkaido University: http://hosho.ees.hokudai.ac.jp/~tsuyu/top/dct/lc.html.

In 1988, David Tilman proposed a competing theory, known as the resource-ratio theory, to address perceived deficiencies in Grime's C–S–R theory. The resource-ratio theory predicts that plants require different ratios of resources, rather than simply different amounts of resources, and it is the ratios that place species into different niche spaces, thus allowing them to co-exist and to not violate the dictum that species with identical niches cannot co-exist indefinitely. However, extensive research over the past 25 years has failed to find much evidence in support of this theory, and it is now no longer considered as viable an explanation of plant strategies. Rather, plants compete and co-exist by modifying their physiological tolerances and their responses to environmental variation in ways that allow them to compete successfully against neighboring plants.

A rigorous review of Grime's C–S–R theory by J. Wilson and W. Lee found that there were many deficiencies in its assumptions: many of its predictions were either impossible to make, or had never been tested. Where they had been tested, the C–S–R theory proved useful in identifying various adaptive strategies of plants in particular habitats. Its main contribution appears to be that it includes the extra dimension of stress tolerance and Wilson and Lee suggest combining it with the Leaf Amortization theory, which says that the more carbon and nutrients invested in a leaf, the longer it takes for the plant to payback those construction costs before the leaf begins yielding a net profit (e.g., C) to the plant. Stress tolerant plants often display a common suite of adaptations, such as long leaf lifespans (i.e., evergreen leaves), tough leaves to withstand harsh winter conditions, low photosynthetic rates and consequently low growth rates coupled with low nitrogen content and demand from the soil.

Historically, communities were thought to progress from early successional states to a self-sustaining endpoint known as the climax community. In the early 20th century, Frederick Clements envisioned such communities as highly repeatable through time (i.e., climax communities were co-evolved sets of species) and implied, but never explicitly, that such communities might even be viewed as super organisms to an extent. This view was sharply challenged in the 1920s by H.A. Gleason, a botanist from the New York Botanical Gardens, who promoted the theory that species segregated individually along environmental gradients according to their own physiological tolerances, an idea known as the individualistic hypothesis. Most ecologists today concur with Gleason's hypothesis and paleoecological studies of community composition changes that occurred as the glaciers receded in North America and other places document the presence of unique communities unlike any we see today.

In the current view a climax community represents an oscillation of community organization around a dynamic equilibrium stage, that is, there are small, mostly random changes in species composition, but with little or no direction (Fig. 17). The difficulty of describing the climax state is one of scale. Vegetation change is observed differently as the size of the measurement unit increases from a single stand of a few individuals to regional scales incorporating hundreds of stands. Small-scale disturbance on the order of loss of single individuals may cause a microsuccessional sequence, if that individual is predictably replaced by another species with other ecological attributes. This may create a patchwork of communities, which on one scale appear to be in different stages of succession while on a larger scale appearing to represent a dynamic mosaic of regional vegetation.

The maintenance of biodiversity in plant communities has been the subject of much discussion for nearly a century now. The classic explanation has been that niche differences allow species to coexist because resources are partitioned slightly differently among species and that species with similar niches cannot co-occur indefinitely. But a competing theory, the neutral theory of biodiversity, states that niches are not important, and that all competitors are equally competent, which is what allows species to

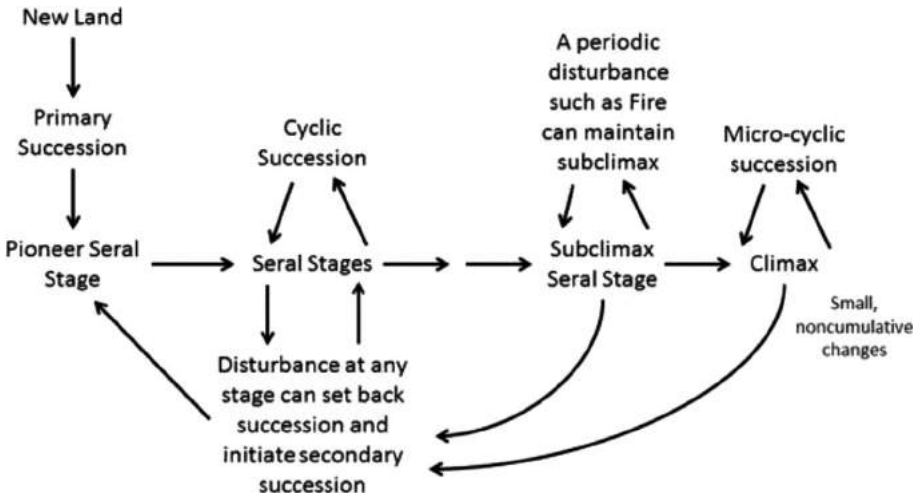


Fig. 17 A diagram of various successional pathways in plant communities. Primary succession occurs on newly exposed land, such as land exposed after glaciers retreat, or is newly produced by lava flows from volcanoes. Various plant species establish over time leading to a series of successional stages that culminate in a dynamic climax community. Some successional stages may oscillate or revert back to earlier stages (cyclic succession), while periodic disturbances may set succession back to earlier several stages. At the climax stage, there are small-scale changes in the plant communities (micro-cyclic succession), dependent on localized plant death (disease) and disturbance (wind throw), and which occur in a mosaic of patches across the landscape. Redrawn from: Barbour, M.G., Burk, J.H., Pitts, W.D., Gilliam F.S. and Schwartz, M.S. (1999). *Terrestrial Plant Ecology*, 3rd edn, Menlo Park, CA: Addison Wesley Longman, Inc.

co-exist. Recent research suggests that niche diversity allows the maintenance of competing species for the following reasons. When a species is rare in the community, it is more likely to compete with other individuals of that same species, who also share essentially the same niche. They are less likely to compete with another different species, so most of the competition is with members of its own species. Only when that species becomes more abundant will its effects on competing species become significant. The per capita growth rate of the population then, is higher when that species is rare and its competitors common, which results in a stabilizing effect, because it prevents the competitors from displacing them from the community (Fig. 18).

Disturbance is characteristic of most plant communities. Natural disturbance includes fire, wind damage, grazing, insect damage, frost heaving, flooding, disease, and other causes. The spatial extent of disturbance may differ from an individual to large regional effects. Disturbance has two attributes affecting vegetation: frequency (or return interval) and intensity. Fire is one disturbance that illustrates these attributes (Fig. 19).

Fire has an important influence on community structure in most semiarid regions including biomes such as grasslands, savanna, and Mediterranean scrub. Fire also is suggested to be important in boreal forests, coniferous and deciduous forests, and even in bogs. Fires may reoccur at frequencies of 1–5 years in grasslands, 25–100 years in Mediterranean scrub communities and eucalypt forests, and 100–500 years in coniferous and deciduous forest communities, although these return frequencies can vary depending on edaphic conditions, such as soil depth and susceptibility to drought. For example, in the Linville Gorge Wilderness Area in the mountains of western North Carolina, the lower slopes are dominated by deciduous forest while on the thin-soiled ridges, fire-adapted species such as Table Mountain Pine (*Pinus pungens*) predominate. Historical records suggest that the return frequency for ground fires ranges between 5 and 12 years, while catastrophic crown fires occur about every 75 years. As a consequence, many species have unique adaptions to fire. Table Mountain Pine (and its close relative pitch pine, *Pinus rigida*) both have *serotinous pine cones*, which only open and disperse their seeds after being heated in a fire.

Each fire frequency results in different modifications to the species and communities that persist in the region. In the absence of frequent fires communities may change in composition and structure. For example, as a result of anthropogenic decreases in fire frequencies, the vegetation of the prairie peninsula of Midwestern United States was converted from grasslands to deciduous forest. In the eastern United States, fire intolerant species, such as red maple (*Acer rubrum*) have become more common. Furthermore, the absence of fire results in a buildup of litter on the forest floor, creating a large fuel load that when ignited, leads to catastrophic crown fires that can devastate an area and destroy the forest. Once the forest is lost, rain events can lead to uncontrolled soil erosion, which may delay successional recovery for decades. In bogs, which can occasionally become dry during years of low rainfall, the absence of fires allows woody shrubs to invade, which then shade out various light-demanding herbaceous species, such as the charismatic and carnivorous Venus fly-trap (*Dionaea muscipula*).

The intensity of a fire can be gauged by the level of its destruction. Fires may be classified as high, moderate, or low intensity. Highly destructive fires kill all the aboveground living matter, and only seeds, rootstocks, and rhizomes that are far enough underground to avoid the extremely high temperatures remain to repopulate the community. Such is the case in grassland fires, Mediterranean scrub, and some coniferous fires, especially those with large fuel loads (see above).

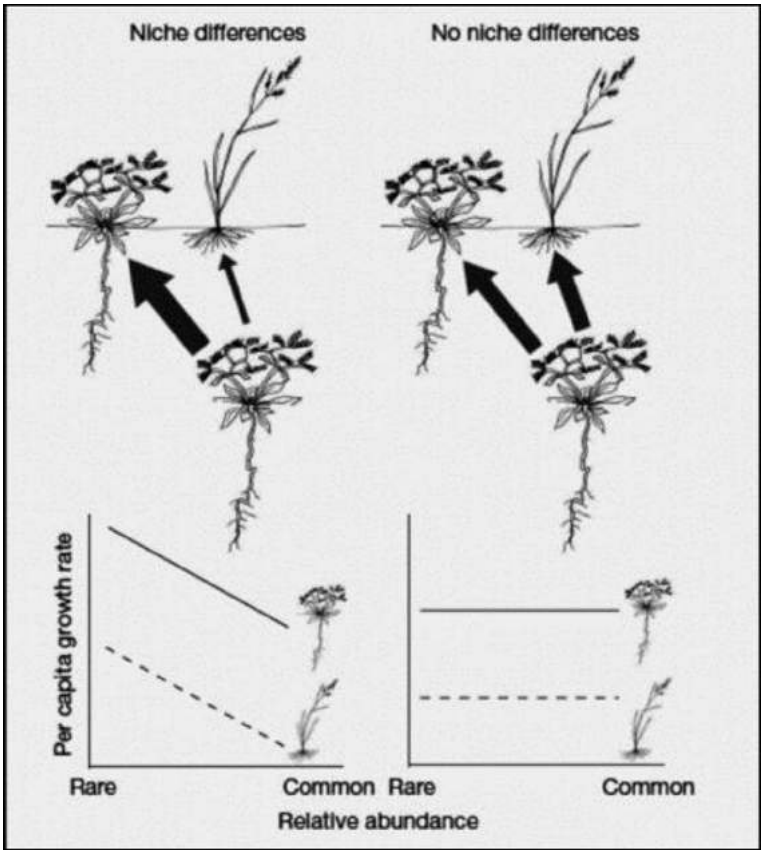


Fig. 18 An example of how differences in niches may maintain species diversity in plant communities. In this case, two species differ in rooting depth and most likely withdraw water and nutrients from different soil compartments. The larger the width of the arrow, the greater is the competitive influence of that species. When a species is rare (see graph and figures on *left*), it is more likely to compete against another species which is common. Since the two species differ in their belowground niche spaces, interspecific competition is minimal and the rare species can have high per capita growth rates. When rare species increase in abundance, they begin to compete more among themselves and less against their competitors, which reduces their per capita growth rates. If there were no niche differences (see graph and figures on *right*), then species would limit themselves as much as their competitors, whether rare or common, and per capita growth rates would remain unchanged. These dynamics then allow the maintenance of a high diversity of plant species in the community, despite the apparent similarity in their resource requirements. Figure courtesy: Levine, J.M. and HilleRisLambers, J. (2009). The importance of niches for the maintenance of species diversity. *Nature* **461**, 254–258.

Moderate-intensity fires kill photosynthetic structures and some meristematic tissues. Standing vegetation may resprout and produce new photosynthetic structures, for example, the Pine Barrens of New Jersey, United States, where pitch pine produces new branches from epicormic buds beneath the bark and Mediterranean ecosystems, wherein shrub species sprout prolifically after fires. The germination of some Mediterranean species occurs only after exposure to smoke, insuring that seedlings germinate when the probability of competition from vegetation is low and environmental resources such as light and nutrients are high.

Finally, low-intensity fires can affect the ground layer of vegetation without a dramatic effect on the canopy. The effect of this type of fire is to control the composition and reproduction of the community. Examples of these communities include the ponderosa pine forests of western North America and the longleaf pine–wire grass communities of the southeastern Coastal Plain of North America. In both cases, frequent fires maintain the dominant overstory trees (*Pinus ponderosa* or *Pinus palustris*, respectively) by reducing the successful establishment of broad-leaved competitors as well as eliminating diseases that attack the pines. Fire interacts with an additional disturbance regime—tropical storms/hurricanes—to maintain the structure and function of longleaf pine ecosystems.

Each type of disturbance produces different ecological and evolutionary constraints on the vegetation. Frequent, low-intensity disturbance results in a distinctively different vegetation than infrequent, high-intensity disturbance. Each of these vegetation types will have a different response to the disturbance. With frequent, low-intensity disturbance there may be little compositional change following the disturbance. The plants of this vegetation have adaptations that allow rapid recovery following the disturbance. Conversely, with infrequent, high-intensity disturbance a successional sequence may occur because the adaptations of initial invaders that favor response to disturbance may not convey adaptive advantage in competition over time.

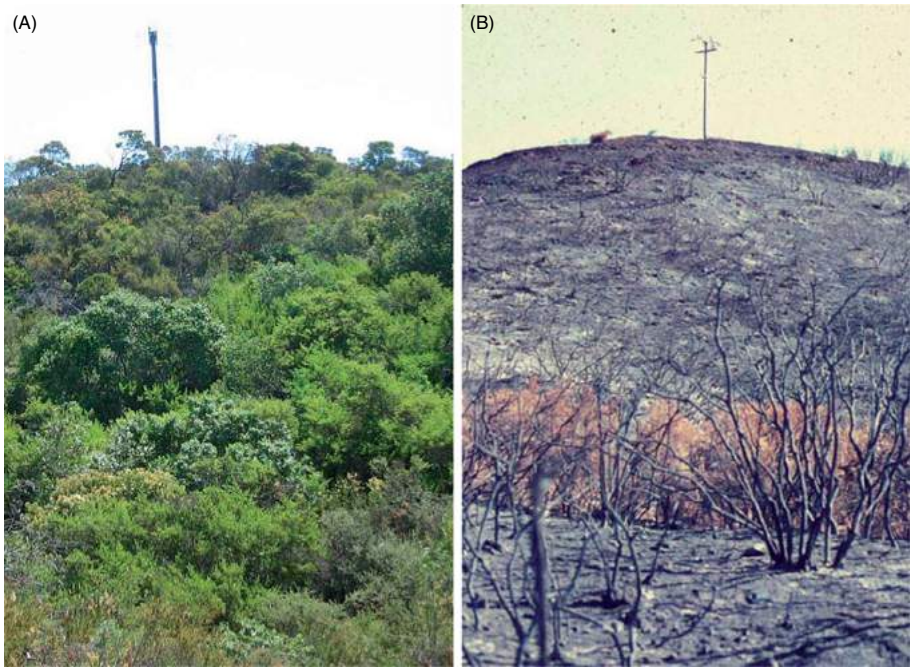


Fig. 19 Unburned and burned chaparral in Santa Barbara County, CA, United States. Both photographs were taken of the same area but separated by 32 years. (A) Typical unburned chaparral in 2003 consisting of chamise (*Adenostoma fasciculatum*), California lilacs (*Ceanothus* spp.), and toyon (*Heteromeles arbutifolia*). (B) Burned chaparral following an intense fire in 1971. The utility pole serves as a reference point. The photographs indicate the rapid recovery of the vegetation following fire.

Summary

The sessile nature of plants requires they integrate the immediate environment for the duration of their lives. The presence of a plant suggests that, for the period of occupying that site, the environment has met all minimal requirements for light, water, and nutrients while not surpassing the tolerances of the plant to abiotic and biotic factors. Plant interactions alter the physical and biotic effects of the environment and may lead to alteration in community composition. When plants are no longer able to alter the environment beyond the tolerance of the occupants, a dynamic stability in community composition is established. This composition is often changed as a result of disturbance in the form of herbivory, wind, flooding, or fire. Depending upon the frequency and intensity of disturbance, an altered and dynamic community may occupy the site.

Acknowledgments

The authors would like to thank Drs. Lara Souza and Catherine Cole for helpful improvements to the manuscript.

Reference

Inderjit, Wardle DA, Karban R, and Callaway RM (2011) The ecosystem and evolutionary contexts of allelopathy. *Trends in Ecology & Evolution* 26: 655–662.

Further Reading

Adler PB, Ellner SP, and Levine JM (2010) Coexistence of perennial plants: An embarrassment of niches. *Ecology Letters* 13: 1019–1029.

Anderson RC, Fralish JS, and Baskin JM (1999) Deep-soil savannas and barrens of the Midwestern United States. In: *Savanna, Barrens, and rock outcrop plant communities of North America*. New York: Cambridge University Press.

Bais HP, Vepachedu R, Gilroy S, Callaway RM, and Vivanco JM (2003) Allelopathy and exotic plant invasion: From molecules and genes to species interactions. *Science* 301: 1377–1380.

Barbour MG, Burk JH, Gilliam FG, and Schwartz MW (1999) *Terrestrial plant ecology*, 3rd edn Menlo Park, CA: Benjamin/Cummings, an imprint of Addison Wesley Longman.

Bobbink R, Hicks K, Galloway J, Spranger T, Alkemade R, Ashmore M, Bustamante M, Cinderby S, Davidson E, Dentener F, Emmett B, Erisman J-W, Fenn M, Gilliam F, Nordin A, Pardo L, and De Vries W (2010) Global assessment of nitrogen deposition effects on terrestrial plant diversity: A synthesis. *Ecological Applications* 20: 30–59.

Borland AM, Griffiths H, Hartwell J, and Smith JAC (2009) Exploiting the potential of plants with Crassulacean acid metabolism for bioenergy production on marginal lands. *Journal of Experimental Botany* 60: 2879–2896.

Boyd RS (2007) The defense hypothesis of elemental hyperaccumulation: Status, challenges, and new directions. *Plant and Soil* 293: 153–176.
Brady NC and Weil RR (2002) *The nature and properties of soils*, 13th edn. Upper Saddle River, NJ: Prentice-Hall.
Bucci SJ, Goldstein G, Meinzer FC, Scholz FG, Franco AC, and Bustamante M (2004) Functional convergence in hydraulic architecture and water relations of tropical savanna trees: From leaf to whole plant. *Tree Physiology* 24: 891–899.
Caldwell MM, Dawson TE, and Richards JH (1998) Hydraulic lift: Consequences of water efflux from the roots of plants. *Oecologia* 113: 151–161.
Callaway RM (1995) Positive interactions among plants. *Botanical Review* 61: 306–349.
Casper BB and Jackson RB (1997) Plant competition underground. *Annual Review of Ecology and Systematics* 28: 545–570.
Chave J, Coomes D, Jansen S, Lewis SL, Swenson NG, and Zane AE (2009) Toward a worldwide wood economics spectrum. *Ecology Letters* 12: 351–366.
Chazdon RL and Pearcy RW (1991) The importance of sunflecks for forest understory plants. *Bioscience* 41: 760–766.
Darley WM (1990) The essence of "plantness". *American Biology Teacher* 52: 354–357.
Donovan LA, Maherall H, Caruso CM, Huber H, and de Kroon H (2011) The evolution of the worldwide leaf economics spectrum. *Trends in Ecology & Evolution* 26: 88–95.
Driscoll CT, Lawrence GB, Bulger AJ, Butler TJ, Cronan CS, Eagar C, Lambert KF, Likens GE, Stoddard JL, and Weathers KC (2001) Acid rain revisited: Advances in scientific understanding since the passage of the 1970 and 1990 Clean Air Act Amendments. In: *Hubbard Brook Research Foundation*. vol. 1. Hanover, NH: Science Links™ Publication. no.1.
Gilbert IR, Jarvis PG, and Smith H (2001) Proximity signal and shade avoidance differences between early and late successional trees. *Nature* 411: 792–795.
Gilliam FS (2006) Response of the herbaceous layer of forest ecosystems to excess nitrogen deposition. *Journal of Ecology* 94: 1176–1191.
Gilliam FS, Platt WJ, and Peet RK (2006) Natural disturbances and the physiognomy of pine savannas: A phenomenological model. *Applied Vegetation Science* 9: 83–96.
Gurevitch J, Scheiner SM, and Fox GA (2006) *The ecology of plants*, 2nd edn. Sunderland, MA: Sinauer Associates.
Haines BL (1977) Nitrogen uptake—Apparent pattern during old field succession in southeastern United States. *Oecologia* 26: 295–303.
Halsey RW (2004) In search of allelopathy: An eco-historical view of the investigation of chemical inhibition in California coastal sage scrub and chamise chaparral. *The Journal of the Torrey Botanical Society* 131: 343–367.
Horn HS (1971) *The adaptive geometry of trees, monographs in population biology*. vol. 3, Princeton, NJ: Princeton University Press p. 144.
Horton JL and Neufeld HS (1998) Photosynthetic responses of *Microstegium vimineum* (Trin.) A. Camus, a shade-tolerant, C-4 grass. *Oecologia* 114: 11–19.
Hull JC (2002) Photosynthetic induction dynamics to sunflecks of four deciduous forest understory herbs with different phenologies. *International Journal of Plant Sciences* 163: 913–924.
Kallarackal J, Otieno DO, Reineking B, Jung E-Y, Schmidt MWT, Granier A, and Tenhunen JD (2013) Functional convergence in water use of trees from different geographical regions: A meta-analysis. *Trees* 27: 787–799.
Koch GW, Sillett SC, Jennings GM, and Davis SD (2004) The limits to tree height. *Nature* 428: 851–854.
Larcher W (2003) *Physiological plant ecology, ecophysiology and stress physiology of functional groups*, 4th edn New York: Springer.
Levine JM and HillerisLambers J (2009) The importance of niches for the maintenance of species diversity. *Nature* 461: 254–257.
MacArthur RH (1972) *Geographical ecology: Patterns in the distribution of species*. New York, NY: Harper & Row.
McCook LJ (1994) Understanding ecological community succession: Casual models and theories, a review. *Vegetatio* 110: 115–147.
McGill BJ, Enquist BJ, Weiher E, and Westoby M (2006) Rebuilding community ecology from functional traits. *Trends in Ecology and Evolution* 21: 178–185.
Neufeld HS and Young DR (2014) Ecophysiology of the herbaceous layer in temperate deciduous forests. Chapter 3, In: Gilliam F (ed.) *The herbaceous layer in forests of Eastern North America*, 2nd edn., pp. 34–95. New York, NY: Oxford University Press, Inc.
Onoda Y, et al. (2011) Global patterns of leaf mechanical properties. *Ecology Letters* 14: 301–312.
Osnas JLD, Lichstein JW, Reich PB, and Pacala SW (2013) Global leaf trait relationships: Mass, area and the leaf economics spectrum. *Science* 340: 741–744.
Paul GS and Yavitt JB (2011) Tropical vine growth and the effects of forest succession: A review of the ecology and management of tropical climbing plants. *Botanical Review* 77: 11–30.
Raunkiær C (1934) *The life forms of plants and statistical plant geography*. Being the collected papers of C. Raunkiær. Oxford: Clarendon Press.
Reich PB, Wright IJ, and Lusk CH (2007) Predicting leaf physiology from simple plant and climate attributes. *Ecological Applications* 17: 1982–1988.
Rosenzweig M (1968) Net primary productivity of terrestrial communities: Prediction from climatological data. *American Naturalist* 109: 87–94.
Sage RF (2003) The evolution of C4 photosynthesis. *New Phytologist* 161: 341–370.
Salisbury FB and Ross CW (1992) *Plant physiology*, 4th edn. Belmont, CA: Wadsworth Publishing.
Schenk HJ (2006) Root competition: Beyond resource depletion. *Journal of Ecology* 94: 725–739.
Simard SW and Durall DM (2004) Mycorrhizal networks: A review of their extent, function, and importance. *Canadian Journal of Botany* 82: 1140–1165.
Simard SW, Perry DA, Jones MD, Myrold DD, Durall DM, and Molina R (1997) Net transfer of carbon between ectomycorrhizal tree species in the field. *Nature* 388: 579–582.
Taiz L and Zeiger E (2010) *Plant physiology*, 5th edn. Sunderland, MA: Sinauer Associates, Inc.
Tilman D (1990) Constraints and tradeoffs: Towards a predictive theory of competition and succession. *Oikos* 58: 3–15.
Violle C, Navas M-L, Vile D, Kazakou E, Fortunel C, Hummel I, and Garnier E (2007) Let the concept of trait be functional. *Oikos* 116: 882–892.
Vitousek PM, Aber JD, Howarth RW, Likens GE, Matson PA, Schindler DW, Schlesinger WH, and Tilman DG (1997) Human alteration of the global nitrogen cycle: Sources and consequences. *Ecological Applications* 7: 737–750.
Way DA and Pearcy RW (2012) Sunflecks in trees and forests: From photosynthetic physiology to global change biology. *Tree Physiology* 32: 1066–1081.
Westoby M and Wright IJ (2006) Land-plant ecology on the basis of functional traits. *Trends in Ecology & Evolution* 21: 261–268.
Westoby M, Falster DS, Moles AT, Vesk PA, and Wright IJ (2002) Plant ecological strategies: Some leading dimensions of variation between species. *Annual Review of Ecology and Systematics* 33: 125–159.
Wilson JB and Lee WG (2000) C–S–R theory: Community-level predictions, tests, evaluation of criticisms, and relation to other theories. *Oikos* 91: 77–96.
Wright IJ, et al. (2004) The worldwide leaf economics spectrum. *Nature* 428: 821–827.
Yeang H-Y (2013) Solar rhythm in the regulation of photoperiodic flowering of long-day and short-day plants. *Journal of Experimental Botany* 64: 2643–2652.

Plant Physiology☆

Ulrich Lüttge, Technical University of Darmstadt, Darmstadt, Germany

What Is Physiology? What Is Ecology?

A well-known dictionary of the English language says that physiology is "the science of the function of living organisms and their parts" and that ecology is "the branch of biology dealing with the relations of organisms to one another and to their physiological surroundings" and if we follow more closely Ernst Haeckel's original definition when he first introduced the term ecology in 1866 we have it as "the entire science of the relations of the organism to its surrounding environment, comprising in a broader sense all conditions of its existence." Hence, if we want to bridge the interface between general ecology and plant physiology we must consider functions of parts and the whole of organisms on the one hand and all conditions of their existence on the other hand and integrate both. It shall be done by looking first at plant physiology and then at ecology and finally by combining both, choosing physiological ecology of photosynthesis as a case study because photosynthesis with its primary production of new biomass from inorganic precursors is of paramount importance for all life on Earth.

Plant Physiology: Function of Plants and Their Parts

The Parts

In a hierarchical order the parts of plants are molecules, membranes, organelles, cells, tissues, and organs. Macromolecules such as polynucleic acids, proteins, polysaccharides (carbohydrates), and lipids may have both structural and functional roles. Biological membranes composed of lipids and proteins (lipoprotein membranes) border the living cells at their surface (the plasma membrane) and separate and conceal various compartments inside the cells, for example, the central cell sap vacuole (the tonoplast) typical of plant cells. Important organelles within cells are the mitochondria and the chloroplasts. The plant cells are surrounded by cell walls composed of polysaccharides, most importantly cellulose. Individual cells can already be independent autotrophic organisms, such as prokaryotic photosynthetically active bacteria and cyanobacteria, which as endosymbionts also have become the evolutionary precursors of chloroplasts, and eukaryotic unicellular algae. In the pluricellular algae, bryophytes, and vascular plants, many cells build up tissues, different tissues form organs and various organs, such as roots, stems, leaves, and flowers make up the whole vascular plant. Eukaryotic plant cells have a nucleus with chromosomes where the central genome is located, but they have two additional genomes in the two organelles, the mitochondria and the chloroplasts, which as original endosymbionts in the phylogenetic history of the eukaryotic cells have retained their own deoxyribonucleic acid (DNA) carrying genetic information.

The Functions

The main concern of plant physiology is the causality of functions. A basic property of life is metabolism. Therefore, we may distinguish functions of biochemistry and functions of development.

☆*Change History:* March 2018. Irene Martins made minor changes to the text and references.

Biochemical Functions

A major distinction of biochemical functions is dissimilation and assimilation. Dissimilation breaks down substrates for energy metabolism and also for the formation of monomeric building blocks for the synthesis of macromolecules. Assimilation in photosynthetically active plants is the most noble character of plants because the strictly autotrophic plants can build up their entire organic biomass from inorganic precursors, not only carbon compounds from carbon dioxide and water in photosynthesis, but also nitrogen- and sulfur-containing organic molecules from inorganic ammonia and nitrate, sulfate and sulfide. This is of eminent basic ecological importance as it is the only pathway for the entry of inorganic matter into the organic biomass on Earth.

Dissimilative processes are fermentation in the cytoplasm and respiration in the mitochondria breaking down carbohydrates and serving energy metabolism as well as providing building stones for the synthesis of proteins and lipids. Similarly the breakdown of proteins and lipids (fatty acids) to their monomeric building stones are dissimilative processes which also can considerably contribute to energy metabolism in the case of lipids. The major assimilative process is photosynthesis. The biosynthesis of structural and functional macromolecules uses the monomeric building stones of organic bases and pentoses (five carbon sugars) for polynucleic acids, amino acids for proteins, various monomeric sugars for polysaccharides, and activated acetic acid (acetyl-coenzyme A) for lipids. A list of important compounds in the biochemistry of plants and their functions is given in Table 1. The major role of polynucleic acids is in storage, transmittance, and active use of genetic information. Proteins serve structural functions but are particularly important as biocatalyzers in enzymatic reactions. Lipids together with proteins build up the lipoprotein biomembranes. Carbohydrates have major structural functions in the cell walls of plants and, in this way, also are the basis of the formation of wood in the stems of plants and the trunks of large trees.

In addition, we must note that plants are the most inventive biochemists we can imagine. They can produce a vast diversity of natural products some of which may be classified as terpenoids, phenolic compounds and organic bases, and alkaloids (Table 1). Among the latter the stimulating compounds of coffee, tea, and tobacco as well as drugs such as atropine, chinine, curare, opium, cocaine, mescaline, and coniine, are well known.

Having mentioned the separate biochemical functions of cytoplasm, mitochondria, and chloroplasts above, it should be evident that metabolism in plant cells is compartmentalized. This is possible by the separating functions of membranes concealing the compartments. However, like any good border such membrane borders must not block off the various compartments hermetically from each other. On the contrary, they must allow a controlled and regulated cooperation of compartments. This is provided by transport proteins crossing the membranes, that is, carriers and channels for substrates mediating the exchange between the compartments. Separate biochemical and physiological functions of the organs leaves, roots, and flowers need translocation in the long-distance transport pathways of the phloem for assimilates and the xylem for water plus mineral nutrients.

Analytical techniques have now advanced to the extent that one can collect information about entire complements of various compounds. We call this "omics," such as genomics deciphering entire genomes, transcriptomics covering gene transcripts, proteomics analyzing sets of proteins, and metabolomics assessing occurrence of metabolites. This is very descriptive of the functional system of organisms and provides huge sets of data. New theoretical approaches are developed to digest this information for the quest of understanding causality, for example, network dynamics (Fig. 1).

Table 1 Important chemical compounds and their major functions in plants

Compound class	*Monomers*	*Polymers*	*Functions*
Primary metabolism and functions			
Carbohydrates	Monosaccharides (pentoses, C_5; hexoses, C_6)	Polysaccharides	Structure Storage (e.g., starch)
Proteins	Amino acids	Proteins	Biocatalyzers Structure (membranes) Storage
Lipids	Activated acetyl Glycerol	Fatty acids Lipids	Structure (membranes) Storage (fats)
Nucleic acids	Organic bases, pentoses (ribose) (deoxyribose)	Polynucleic acids: ribonucleic acid (RNA), deoxyribonucleic acid (DNA)	Genetic information
Natural products and their functions			
Terpenoids	Isopentenyl-pyrophosphate (C_5)	Mono- (C_{10}), di- (C_{20}), tri- (C_{30}), tetra- (C_{40}), poly- ($C_{500-5000}$) terpenes	Phytohormones, scent compounds, pigments, defense, resins
Phenolic compounds	Phenol	Flavones Flavonols	Electron transport Pigments Structure (the lignin of wood)
Anthocyanidines			
Alkaloids	Amino acids Organic bases		Defense Nitrogen storage

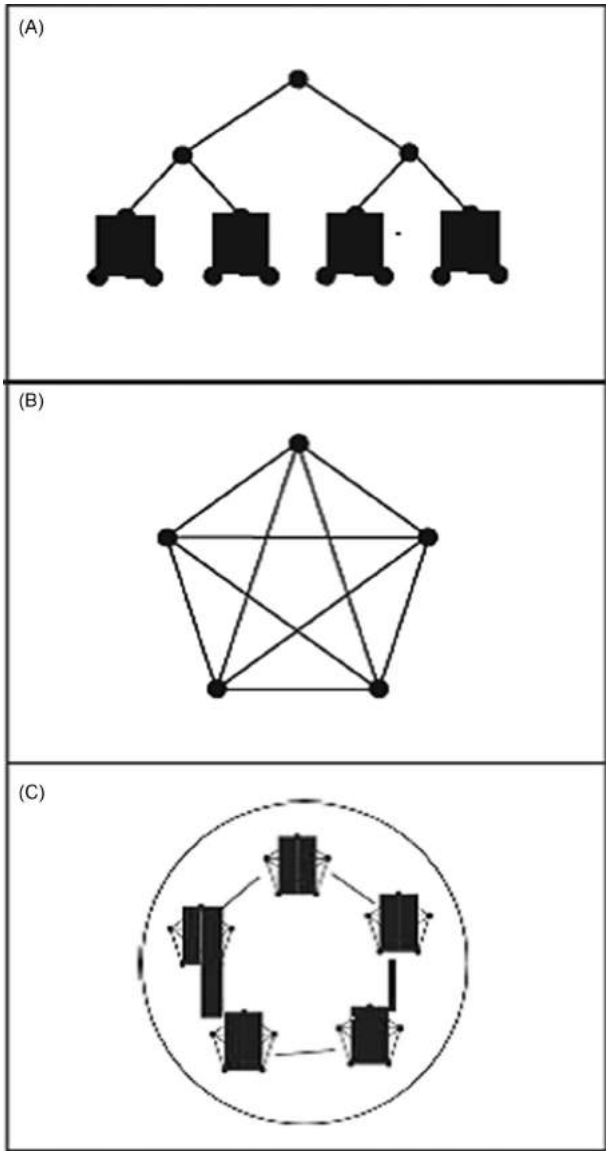


Fig. 1 Schemes of different types of networks: (A) open network with a hierarchy of levels, (B) closed network with (knots) and edges (connecting lines) where the hierarchy of levels is lost, (C) complex network composed of different closed networks, where each of the subnetworks composing it can be considered as a knot and where the supernetwork itself may become a knot in a yet more highly integrated supranetwork.

Developmental Functions

The life cycle of higher plants comprises germination of seeds, growth, flowering, fertilization, and sexual reproduction forming seeds. Since plants normally cannot move around and are bound to a given location their orientations in space also are largely given by developmental processes. These processes involve increases in plant size and mass by assimilation. However, life cycles also require the most complex differentiations. These are regulated by effectors which can be external or internal control parameters and which evoke differential activation/inactivation of genes. External control parameters are environmental cues linking developmental functions to ecological responses. Internal control parameters among others are various phytohormones controlling different developmental functions. External and internal control parameters interact. There are primary and secondary messengers. All of this is interwoven in signaling networks of an extraordinarily high degree of complexity, an example of which is shown in Fig. 2 where the signal transduction pathways from an effector to the phosphorylation of proteins are depicted. Protein phosphorylation is an essential element in differential gene regulation as well as directly in cellular metabolic functions.

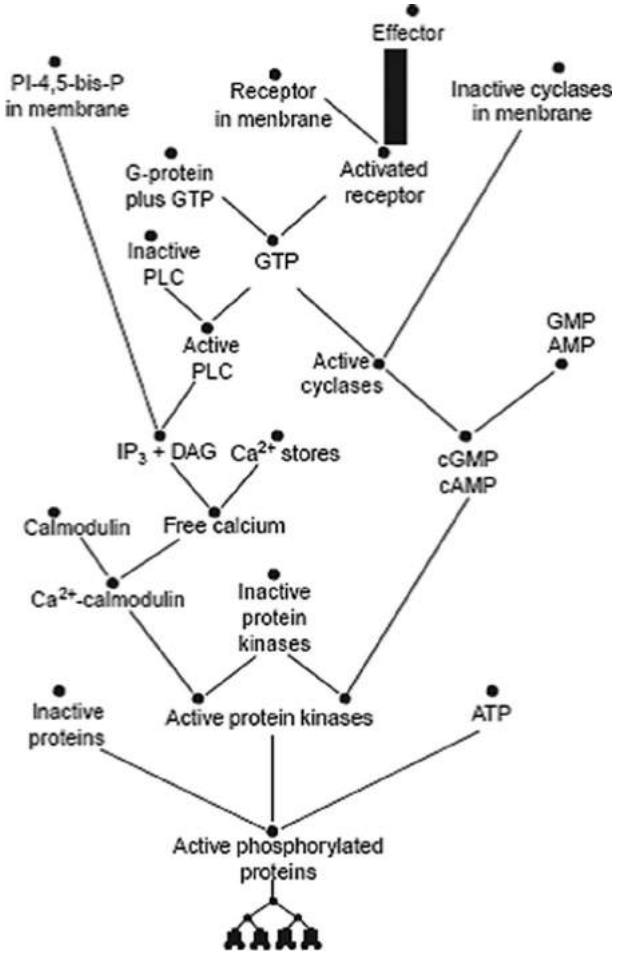


Fig. 2 Molecular regulation network in the signaling cascade from effector to active phosphorylated proteins. An effector which can be any external factor including light, that is, mainly blue and red light, or any internal factor including metabolites and phytohormones reacts with an appropriate receptor (mainly in a membrane, e.g., the plasma membrane). The activated receptor releases guanosine-triphosphate (GTP) from a GTB-binding so-called G-protein, which binds to and activates phospholipase C (PLC), active PLC hydrolyzes membrane-bound phosphatidyl-inositol-4,5-bisphosphate (PI-4,5-bis-P) to inositol-1,4,5-tris-phosphate (IP_3) and diacyl glycerol (DAG) which elicit the release of the second messenger calcium from membrane-bound Ca^{2p}-storing compartments, free Ca^{2p} ions bind to the small peptide calmodulin, and Ca^{2p}-calmodulin activates protein kinases. GTP also can activate membrane-bound cyclases which from guanosine- and adenosine-monophosphates (GMP, AMP) form the respective cyclic derivatives cGMP and cAMP which then activate protein kinases. Activated protein kinases phosphorylate proteins, and the activated proteins can either directly affect metabolic processes or move into the nucleus and function as gene regulation factors.

Plant Ecology: All Conditions of Its Existence

The conditions of the existence of organisms in their environment are given by abiotic and biotic factors. In or near their optimum these factors or external control parameters fulfill basic requirements. However, according to the biological stress concept any of a vast multitude of factors can become a stress factor or "stressor" when its dosage is either too low or too high.

Abiotic Factors

The most important abiotic factors for plants are light, carbon dioxide, water, temperature, nutrients, and salinity. Their actions are interrelated in a functional network involving all major processes of plant physiology in the ecological performance of plants (Fig. 3).

Light drives photosynthetic CO_2 fixation and also photorespiration and excess light can lead to overenergization of the photosynthetic apparatus and formation of reactive oxygen species (ROS) and oxidative stress. Light is heating up the leaves. Light may also have a signaling function. Day length (photoperiod) may have developmental consequences. Particularly, red and blue light can function as effectors in signaling networks (Fig. 2) and in this way, light can also lead to the developmental formation of sun and shade leaves. Light affects the movements of stomatal guard cells, opening and closing of stomatal pores in leaves which regulate gas exchange, that is, CO_2 uptake and loss of water vapor by transpiration. This affects CO_2 assimilation and also transpirational cooling of leaves. Temperature has important effects on metabolism. Heat, cold, and freezing are important

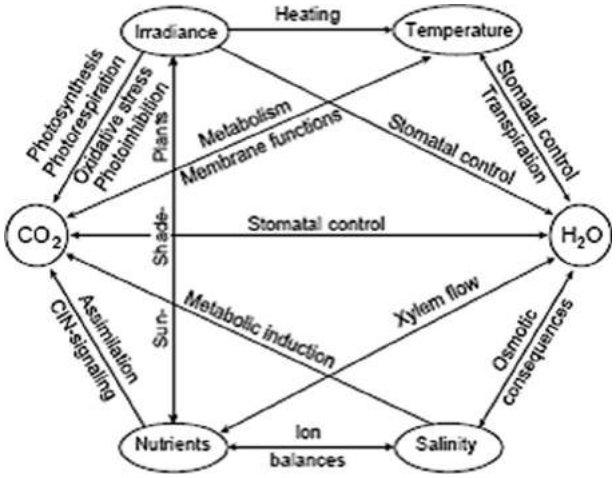


Fig. 3 Factor/function network of physiological ecology.

stressors. Transpiration and the xylem flow of water affect uptake and distribution of nutrients. Nutrients and light interact, for example, in the formation of sun and shade plants, where the shade plants generally have a higher demand of nitrogen. CO_2 assimilation and nutrients interact in the assimilation of inorganic nitrogen and sulfur and in carbon/nitrogen signaling functions in the whole plant. Salinity is one of the outstanding ecological challenges worldwide and, in particular, is a great problem in irrigation agriculture. Salinity affects plant–water relations due to osmotic consequences and has adverse effects on ion nutrient balances, and especially the sodium ions of NaCl have adverse effects if accumulated in plant cells and not sequestered by transport across the tonoplast into the central cell vacuole. These are but a few of the possible interactions and the reader may discover others by moving around in a scheme like that of Fig. 3.

Biotic Factors

Important biotic factors are competitors and nurse plants, predators, parasites, and symbionts eliciting antagonisms and mutualisms. Competition between plants for light often is given by shading and also by filtering certain wavelengths out of the solar spectrum, especially in the red region, where they are important for photosynthesis and for signaling functions. In the soil, competition is for water and nutrients. Plants develop chemical defense mechanisms, so-called allelopathic reactions. Conversely, plants can also nurse other plant species allowing seedling establishment and early growth under their canopy, a phenomenon which like in ecology is also very important in agroforestry. Animal predators and microbial and animal parasites, of course, cause manifold stress situations which need not be enumerated here. Plants often develop characteristic responses against such parasites where, in a so-called hypersensitive reaction, they produce aggressive defense compounds, often ROS, sacrificing small parts of their own tissue and killing the intruders with it. Important plant/plant parasites are the hemipara-sites which are green and photosynthetically active but divert the xylem stream from their hosts parasitizing for water and nutrients, such as mistletoes and the agricultural weed Striga in the tropics. In plant/plant holoparasitism, the parasites are not active photosynthetically and tap both the phloem and the xylem of their hosts, obtaining assimilates together with water and nutrients. Mutualisms and important symbioses of plants are the mycorrhiza with fungi for obtaining water and nutrients in exchange against photosynthetic assimilates and the root nodules with bacteria fixing atmospheric nitrogen.

Photosynthesis: A Special Case Story

Green photosynthesizing prokaryotes, algae, and plants are the primary producers of ecosystems. Hence, it appears most appropriate to choose photosynthesis as a particular case to illustrate the close relations of plant physiology and ecology, physiological ecology as it were, and cast a bridge between the two disciplines.

Scalar Levels in Space and Time

Starting the hierarchical levels of the parts of plants above at the bottom of the scale with molecules was prudent but not quite right. When we look at photosynthesis we realize that we must extend the scale to a still very much finer level, that is, that of the elementary particles of physics, the photons and electrons. When the appropriate photons of the blue and the red part of the solar spectrum of wavelengths are absorbed by the molecules of chlorophyll, electrons are excited and moved along an electron transport chain in the internal membranes, the thylakoid membranes, of the chloroplasts to generate the reduction equivalents and chemical energy required for the assimilation of CO_2. At the other end of the scale, we move from leaves to canopies entire ecosystems,

biomes, and even the whole planet. Thus, the spatial scale taken in one dimension (meters) from molecules to the entire globe covers about 15–16 orders of magnitude. For the timescale, we note that the half-life of excited states of chlorophyll are 10^{-15} to 10^{-13} s for the so-called second singlet state attained after the absorption of a blue photon and 10^{-11} to 10^{-9} s for the first singlet state after absorption of a red photon. Time constants for electron transport in the membranes of chloroplasts are up to the range of seconds. Activation of the CO_2-reducing Calvin cycle of photosynthesis takes minutes. Stomatal responses in photosynthetic gas exchange may reach several tens of minutes. Time constants of photosynthetic productivity in whole plants, trees, forests, ecosystems, and biomes are weeks, months, and years, decades, and hundreds of years, and the evolution of life on Earth took about $(3–4) \times 10^9$ years. Thus, the temporal scale (seconds) covers at least 32 orders of magnitude (Fig. 4). Another couple of numbers illustrating the vast range of scales covered is that the surface of the globe receives a solar irradiance energy of about 7×10^{16} Js^{-1}, while the energy content of one exciton stable for 10^{-15} to 10^{-9} s after the absorption of a photon by a chlorophyll molecule is $(30–45) \times 10^{-20}$ J.

Environmental Control Parameters

Water, irradiance, and carbon dioxide (CO_2) are the dominating environmental control parameters in the physiological ecology of photosynthesis and frequently stress limitations are due to water supply and high irradiance. The interaction of these factors can be assessed using Fig. 3 and will be discussed below.

Hydraulic Limitation

The most important regulatory response of photosynthesizing plant leaves to problematic water supply is stomatal closure by guard cell movements. This prevents or at least highly reduces the loss of water in the gaseous form by transpiration. However, it must be considered a compromise and also has negative effects. The most obvious one is that as long as stomata are closed CO_2 cannot be taken up from the atmosphere and production of assimilates stalls. Stomatal closure also amplifies adverse effects of high irradiance. It prevents transpirational cooling when high insulation heats up the leaves. Blocking CO_2 uptake, stomatal closure contributes much to overenergization of the photosynthetic apparatus because the remaining CO_2 in the internal airspaces of leaves is rapidly fixed in photosynthesis and then CO_2 is lacking as an acceptor for the reduction and energy equivalents of reduced nicotinamide adenine dinucleotide phosphate (NADPH) and adenosine triphosphate (ATP) (Fig. 5) produced in photosynthetic electron transport.

High Irradiance

When the production of NADH and ATP is faster than their consumption by the energy-demanding photochemical work of CO_2 reduction and assimilation the molecular elements of the photosynthetic electron transport chain rapidly get reduced and closed for

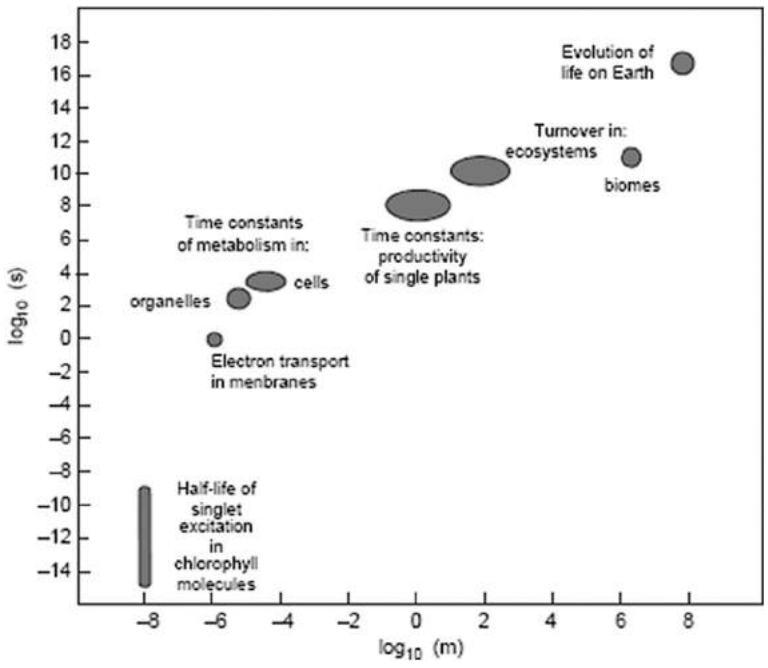


Fig. 4 Scalar levels of structures (in a one-dimensional notation of meters) and time constants of functions (in seconds) in relation to photosynthesis and dependent processes.

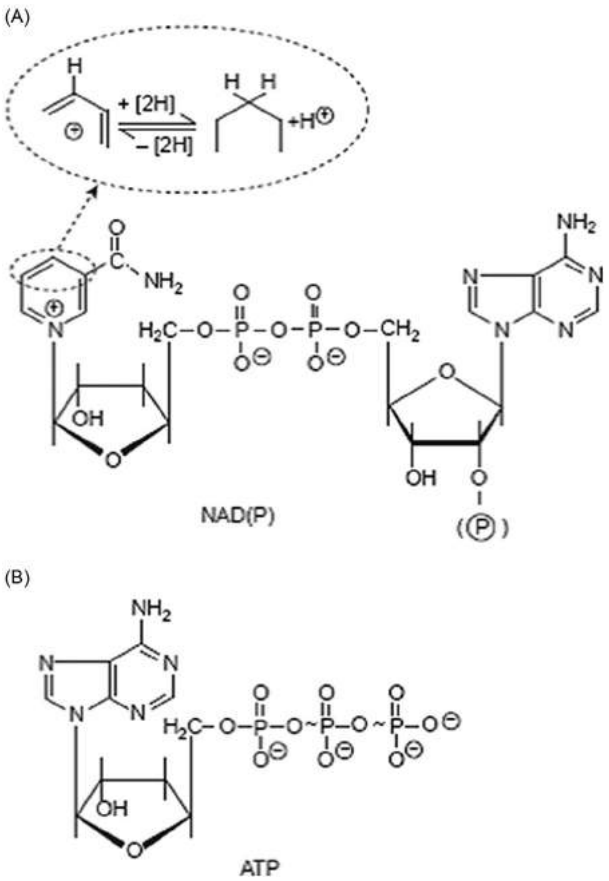


Fig. 5 Chemical structures of (A) nicotine adenine dinucleotide (phosphate) (NAD(P)) and (B) adenosine triphosphate (ATP), where the insert at the upper left corner in (A) shows the transfer of a reduction equivalent [2H] and the squiggled bonds in (B) indicate energy-rich phosphate bonds of energy equivalents.

further electrons. This occurs at high irradiance when not all of the excitation energy of chlorophyll can be used in photochemical work and the excited chlorophyll cannot relax in this way. The plant has evolved several protection mechanisms for this situation.

First, there is photorespiration. This is based on the dual affinity of the CO_2-fixing enzyme ribulose-bis-phos-phate carboxylase/oxygenase (RubisCO) for both CO_2 and oxygen (O_2). When it reacts with O_2, at low CO_2/O_2 concentration ratios in the leaves, carbon and metabolite flow occurs through the cycle of photorespiration. This is also photochemical work, but it is futile and dissipates energy. Second, the energy absorbing and transferring photosystems of the thylakoid membranes have a very complex sophisticated structure, and there are pigments absorbing light energy in the light-harvesting antennas of photosystems and diverting the excitation away from the central reaction centers of the photosystems as well as futile cycles of oxidation (forming epoxides)/reduction of xanthophyll-type pigments dissipating energy in the form of heat. Third, partial and reversible destruction of structural components reduces the light-absorbing activity of photosystems. This involves turnover of proteins as well as the arrangement of antennas and cores of photosystems in the thylakoid membranes. Fourth, the ROS formed from excited electrons of chlorophyll molecules notwithstanding the first three protective mechanisms can be deactivated by superoxide dismutases, and biochemical redox active mechanisms, like the glutathione/ascorbate cycle.

We realize that there are cascades of measures to overcome high irradiance stress. They are related to the phenomenon called photoinhibition which may be protective even if partially destructive, when destruction is reversible. Finally, when the intensity of stress overrules the protective measures parts of the plant or the whole plants may die: they are literally burnt by the ROS.

Carbon Dioxide (CO_2)

Currently anxieties are nourished by dramatic anthropogenic increases of atmospheric CO_2 concentrations. However, we may note that during the geological history of our planet and the evolution of photosynthesis, CO_2 concentrations have fluctuated very much, for example, 230 million years ago they were about 6 times the present levels, 270–310 million years ago they were similar, and 460–535 million years ago they were much higher, that is, about 20 times more than right now. Thus, the affinity to its substrate which the key enzyme of primary photosynthetic CO_2 fixation, RubisCO, has evolved is not so high that it could operate at

saturation under the current atmospheric CO_2 concentration. This problem is amplified for water plants by the low solubility of CO_2 which dissolves in the form of bicarbonate (HCO_3^-) while the actual substrate of RubisCO is CO_2. Hence, many photosynthesizing water plants, cyanobacteria, and algae, have developed carbon-concentrating mechanisms in which carbonic anhydrase, the enzyme catalyzing the $CO_2 = HCO_3^-$ equilibrium, plays a central role.

Among the terrestrial higher plants two different mechanisms of carbon concentrating have evolved which are of eminent ecophysiological relevance, especially in relation to hydraulic and irradiance stress. These are C4 photosynthesis and Crassulacean acid metabolism (CAM). The standard photosynthetic pathway is C3 photosynthesis named after the first stable compound produced after fixation of CO_2 by RubisCO, the three'carbon organic acid 3-phosphoglyceric acid (3PGS), which is then reduced to carbohydrate in the Calvin cycle. C4 photosynthesis and CAM are modifications of this pathway, where, in both cases, the primary fixation of CO_2 is not by RubisCO but by phosphoenolpyruvate carboxylase (PEPC) and the first stable fixation product is the four carbon organic acid malic acid (or malate the anion of this acid). This gave C4 photosynthesis its name. Subsequently, malate is decarboxylated again and the CO_2 regenerated is refixed and assimilated via RubisCO. Why this detour via PEPC and what is the difference between C4 photosynthesis and CAM? The CO_2 affinity of primary fixation by PEPC is 60 times higher than that of RubisCO. This enables PEPC to operate at substrate saturation and drive CO_2-concentrating mechanisms.

In C4 photosynthesis, primary CO_2 fixation by PEPC occurs in an outer tissue (the so-called mesophyll tissue), the malate produced is transported into an inner tissue around the veins or bundles of the leaves (the so-called bundle sheath tissue) where decarboxylation and CO_2 concentrating to about six times that of the ambient atmosphere for subsequent fixation by RubisCO takes place. At this elevated CO_2, RubisCO works close to its saturation and can use a higher proportion of excitation energy for photochemical work. Even if stomata are partially closed to reduce transpiration, still enough CO_2 can diffuse into the leaves for high-affinity fixation by PEPC. We can see that compared to C3 photosynthesis C4 photosynthesis is better suited to deal with stress due to hydraulic limitation and high irradiance.

CAM occurs in many different families of vascular plants, but the name is derived from the family of Crassulaceae (Crassulacean acid metabolism, CAM) where it was first discovered. In C3 and C4 plants, CO_2 fixation and the light-dependent reactions of photosynthetic electron transport must run simultaneously which includes the dilemma that during CO_2 uptake water vapor is lost by transpiration through open stomata. In CAM, primary CO_2 fixation occurs in darkness during the night when the driving force for transpiration, that is, warm and dry ambient air, is highly reduced and loss of water vapor is minimized. The malic acid produced is stored overnight in the central cell sap vacuoles and is remobilized and decarboxylated during the day. In this phase, stomata are closed and very high internal CO_2 concentrations between 2-fold and 60-fold atmospheres are built up so that RubisCO works at substrate saturation while transpiratory loss of water is prevented. Again, this is an effective mechanism of dealing with hydraulic limitation and high irradiance. The principal difference between C4 photosynthesis and CAM is that the two carboxylation processes are separated in space (mesophyll and bundle sheath) in the former and in time (night and day) in the latter, that is, that CO_2 concentrating is restricted in space (bundle sheath in C4 photosynthesis) and time (light period in CAM), respectively.

The three modes of photosynthesis are constitutive in many plant species. However, there are also C3/CAM intermediate plants. Constitutive CAM itself is already quite plastic because when water stress is not severe, CAM plants can open the stomata in the later afternoon, when the nocturnally stored malic acid is consumed, and take up and fix atmospheric CO_2 like C3 plants. In the truly C3/CAM intermediate species, the expression of either of the two photosynthetic phenotypes is flexible and, in many cases, reversible depending on the environmental conditions. For various intrinsic physiological and biochemical reasons the productivity of CAM is lower than that of C3 photosynthesis, and therefore it is only profitable for the plant to perform CAM when the environmental situation is stressful and it is preferable to perform C3 photosynthesis when water availability is sufficient.

Impact at the Community Level: From Physiological Autecology to Physiological Synecology

Primarily, physiological ecology is autecology assessing the physiological traits of individual plants or species in relation to ecological performance. Conversely, synecology covers larger vegetation units, ecosystems, or biomes and is mainly based on floristic, phytosociological, and phytogeographical approaches. However, physiological ecology can also develop to synecology when larger data sets on physiological reactions for different plants composing the vegetation can be obtained. This is increasingly facilitated in the physiological ecology of photosynthesis due to miniaturization of equipment for measuring gas exchange (CO_2 and water vapor) and parameters of photosynthetic electron transport. The former is assessed by infrared gas analysis (IRGA). The latter is particularly easy to achieve by measurements of chlorophyll fluorescence. When discussing the above the various ways of energy dissipation of activated chlorophyll, fluorescence has not been mentioned yet. In fact, a small part of the energy not used for photochemical work can be released again by emission of light, that is, fluorescence, which is readily recorded using miniaturized pulse amplitude-modulated fluorometers. In addition, appropriate sampling for analyses of metabolic compounds and also stable isotopes (mainly ^2H or deuterium, ^{13}C, ^{18}O, ^{15}N) provides a wealth of information. There are several isotope effects in metabolism in general and in photosynthesis in particular. For example, the discrimination of RubisCO against the rare ^{13}C isotope in the substrate CO_2 as compared to the abundant ^{12}C isotope is much larger than that of PEPC, and thus, with the appropriate precautions mass spectrographic analyses of carbon isotope ratios in dried plant material allows to distinguish C3, C4, and CAM plants.

General Importance of Physiological Processes for Fitness at the Community Level

Fitness is frequently defined and quantified as the number of seeds produced ("Darwinian fitness"). However, this may cause difficulties, because it is too limited a view. This is readily recognized when we consider, for example, *K*- and *r*-strategies of plants with small and high seed numbers, respectively, or clonal growth, where particular plant species may dominate entire ecosystems without any generative propagation. Germination of seeds and establishment and survival of seedlings are essential. Thus, we see that the entire complement of physiological processes contributes to fitness, for example, in addition to the capacity of photosynthesis treated above in some detail as a special case story, and similarly important functions such as dormancy of seeds with germination at the ecologically appropriate time and the physiology of development, competition, defense, and mutualism as alluded to above.

Further Reading

Atwell B, Kriedemann P, and Turnbull C (eds.) (1999) *Plants in action*. South Yarra, Australia: MacMillan Education Pty Ltd.
Ezer D and Wigge PA (2017) Out in the midday sun, plants keep their cool. *Current Biology* 27(1): 28–30.
Fitter A and Hay R (2001) *Environmental physiology*. Amsterdam: Elsevier.
Hasegawa Y, Murohashi F, and Uchida H (2016) Plant physiological activity sensing by bioelectric potential measurement. *Procedia Engineering* 168: 630–633.
Heldt H-W (2005) *Plant biochemistry*, 3rd edn, Amsterdam: Elsevier.
Hemsley A and Poole I (2004) *The evolution of plant physiology*. Amsterdam: Elsevier.
Hütt M-T and Lüttge U (2004) Network dynamics in plant biology: Current progress in historical perspective. *Progress in Botany* 66: 277–310.
Lüttge U (1997) *Physiological ecology of tropical plants*. Berlin, Heidelberg, New York: Springer.
Lüttge U (2004) Ecophysiology of Crassulacean acid metabolism (CAM). *Annals of Botany* 93: 629–652.
Lüttge U and Scarano FR (2004) Ecophysiology. *Revista Brasileira de Botânica* 27: 1–10.
Nobel P (2005) *Physicochemical and environmental plant physiology*, 3rd edn, Amsterdam: Elsevier.
Ntagkas N, Woltering EJ, and Marcelis LFM (2017) Light regulates ascorbate in plants: An integrated view on physiology and biochemistry. *Environmental and Experimental Botany* 147: 271–280. https://doi.org/10.1016/j.envexpbot.2017.10.009.
Raghavendra AS (ed.) (1998) *Photosynthesis. A comprehensive treatise*. Cambridge: University Press.
Raven PH, Evert RF, and Eichhorn SE (1999) *Biology of plants*, 6th edn, New York: W.H. Freeman.
Truernit E (2017) Unveiling the dark side of phloem translocation. *Current Biology* 27(9): 348–350.

Pollination

E Pacini, Università di Siena, Siena, Italy

Introduction and Definitions

Pollen is the male gametophyte of gymnosperms and angiosperms. Its size ranges from 15 to 200 μm. Pollination is transport of pollen from its site of production to the female landing site. If successful, it is followed by fertilization and seed development. Irrespective of systematic group, pollination is always affected by biotic and abiotic factors (**Table 1**).

Pollen presentation is the manner in which pollen is presented for dispersal. Pollen may be dispersed as single grains, as in many plants relying on wind dispersal, or in groups of grains, as in almost all zoophilous species. Pollen grains may be held together by:

1. common walls, as in tetrads where grains derived from the same meiotic division stay together, or multiple tetrads which may number up to several thousand, as in orchids;
2. threads on the pollen surface or derived from the anther; and
3. viscous fluids, of which pollenkitt is the most common, which also have other functions, such as to keep pollen in the anther until dispersal, to stick pollen to pollinators, to make pollen attractive through scent or color, and to hide or expose pollen to insect sight.

Pollen dispersal in clumps is typical of angiosperms.

Gymnosperm pollen is produced by pollen sacs in male cones. It is transported by air currents to the ovule micropyle. Angiosperm pollen is produced by anthers of flowers. It is carried to stigmas by animals, air currents, and sometimes water (**Table 2**).

At the end of its flight, pollen may land on female parts of the same or another species, giving rise to legitimate and illegitimate pollination, respectively (**Fig.1**). Pollen has two walls and when accepted by the female part, emits a tube leading the male gametes toward the female ones inside the ovule. Possible crosses in angiosperms depend on the sexual expression of the plant and on pollen vectors (**Fig.1**).

Pollination From an Evolutionary Point of View

Gymnosperm pollination is invariably anemophilous (primary); only recently evolved genera as *Ephedra* and *Welwitschia* are pollinated by insects. There is general agreement that early angiosperms were pollinated by Coleoptera and Diptera. Woody and herbaceous secondary anemophilous angiosperms may descend from zoophilous species. Hydrophily is probably derived from anemophily. Hydrophilous pollen of seagrasses characterized by submarine pollination is 2–3 mm long and a few dozen microns wide. The genus *Callitriche* has terrestrial, amphibious, and submerged freshwater species; their pollen is spherical and that of submerged species is devoid of exine, as in all species with submarine pollination.

Competition to attract pollinators is high when many entomophilous species bloom at the same time and pollinators are few. This is avoided by different blooming periods and anemophily. Examples of entomophilous families with anemophilous members are: Ranunculaceae, *Thalictrum*; Euphorbiaceae, *Mercurialis* and castor bean *(Ricinus)*; Asteraceae, ragweed (*Ambrosia*) and *Artemisia*. Few entomophilous species bloom at the same time in January and February in Northern Hemisphere Mediterranean

Table 1 Effects of the main climatic parameters on pollination

Temperature	High *T* damages flowers and pollen duringpresentation or dispersal
	Moderate *T* facilitates flower and antheropeningLow *T* slows pollen ripening and flower openingand reduces pollinator activity
Rain and mist	Purge pollen from air, especially at low *T*
	Slow flower opening and anther-pollendehydration
	May inappropriately rehydrate and reactivatepollen
	Hinder small animal movements
Sky brightness	Brightness facilitates diurnal pollinator flight Dullness hinders diurnal pollinator flight
Air currents	Facilitate pollen removal and dispersal inanemophilous species
	Facilitate flower opening and antherdehydration
	High wind speeds hinder pollinator flight
Pressure variations	Ascending air currents facilitate long-distancepollen dispersal
	Descending air currents facilitate pollen fallout

 doi:10.1016/B978-0-444-63768-0.00859-3

Table 2 Features of major pollen vectors. Honeybees are the best pollinators because they visit flowers of a given species for as long as they are available, and store pollen for their progeny, unlike nonsocial insects

Pollen vectors			*Specificity (with respect to a flower attractant)*	*Efficiency*[a]	*Distances*[a] *(minimum and maximum)*
Animals	Insects	Honeybees	Very high	Very high	Several cm → few hundred m
		Solitary bees and wasps	High	High	Several cm → few hundred m
		Flies	Low	Moderate	Several cm → few hundred m
		Butterflies and moths	Low[d]	Low[d]	Several cm → few hundred m
		Coleoptera[b]	Low	Low	Several cm → few hundred m
	Birds		Moderate	Moderate–high	Few m → few hundred m
	Flying animals and small marsupials		Low	Low	Several cm → few hundred m
Air currents	Breezes		Low[e]	Low	Few cm → few hundred m
	Strong winds		Very low	Very low	Few hundred m → several km
Water	Salt[c]		Very low	Very low	Some m → several hundred m
	Fresh		Very low	Very low	Few dm → several hundred m

[a]Biotic and abiotic environmental parameters may radically affect efficiency and distance traveled.
[b]Coleoptera are not good pollinators because they have few hairs and their chewing mouthparts damage flowers.
[c]Underwater pollination occurs in all seaweeds.
[d]Efficiency is low because these animals feed only on nectar and must visit different species in order to have a balanced diet.
[e]Specificity is low but species with this pollination syndrome often grow close to each other, as in the case of grasses and tress.

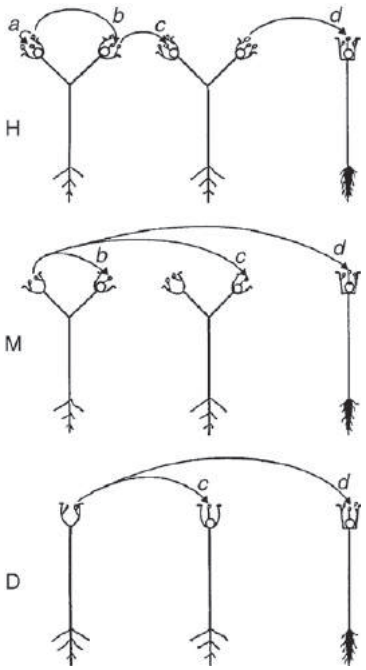


Fig. 1 Scheme of hermaphrodite and unisexual flowers and possible crosses in angiosperms according to plant sexual expression H = plant with hermaphrodite flowers; M = monoecious plant with flowers of both sexes; D = dioecious plant with male and female flowers on different plants. Two plants of the same and one of another species are shown for each type of sexual expression. *a* = autogamy, *b* = geitonogamy, and *c* = heterogamy are legitimate pollination styles though they imply different genetic reassortment; *d* = xenogamy is illegitimate pollination and gives rise to hybrids in the absence of barriers. Plants with hermaphrodite flowers have all possible crosses; these reduce progressively in monoecious and dioecious plants. Pollen vectors determine pollen cross types. Air and water currents are agents of all types of crosses. Social bees are commonly responsible for *a*, *b*, and *c*, rarely *d*. Butterflies and moths are most often agents of *d*, because they only feed on nectar and must visit different flowers to have a balanced diet.

environments, when few insects are active. *Helleborus bocconei* and *H. foetidus* grow in similar environments and share the same pollinator, but pollen attaches to different parts of the pollinator body, so that useless pollination is avoided. Anemophilous species of *Juniperus* growing in the same environments disperse pollen in different periods.

Preparing for Dispersal

A fluid fills the anther cavity in which pollen develops. It disappears by evaporation and/or resorption when pollen is ripe, prior to pollen dispersal. Pollen sacs, anthers, and pollen also lose some water. Pollen could be damaged if it were released with a high water content and high metabolic activity. In order to avoid this possibility, pollen is dispersed in a quiescent state, in which cell division and metabolism are arrested. Two classes of pollen, with different levels of developmental arrest, are determined on the basis of their water content at dispersal: partially dehydrated pollen (PDP) has a water content of less than 30% and partially hydrated pollen (PHP) has a water content of more than 30%. The former has metabolic devices to keep its low water content constant; the latter does not have devices and loses water quickly, especially in dry environments. PDP and PHP may be transported by animals or air currents. Both have advantages and disadvantages as well as devices to ensure successful pollination. PDP survives longer at low relative humidity, whereas PHP dries out readily and dies. However, the latter germinates quickly, taking only a few minutes after landing on a stigma. PDP is more common in dry temperate environments and is dispersed during the dry hours of the day. PHP is more common in the Tropics and wet environments; in temperate regions, it is presented for dispersal when RH is high, such as at night or in winter and autumn. Species belonging to these two groups have physiological and ecological strategies to ensure safe pollination journeys, which are quick and short in the case of PHP.

Pollen Presentation and Dispersal Mechanisms

When the anther opens, pollen is presented for dispersal and may (1) be launched from the anther by ballistic movements of the anther or stamen and dispersed by air currents, as in *Parietaria* and castor bean (*Ricinus communis*); (2) be scattered from the anther by flower movements caused by an insect in search of nectar, and loaded on insect hairs as in *Spartium junceum* and other Phaseolaceae; (3) be dropped by the anther for lack of any forces to keep it attached, as in grasses and many herbaceous and woody anemophilous species; (4) remain stuck to the anther by pollenkitt or other sticky fluids or threads, pending removal by animals or air currents; (5) be kept in anthers having only small apertures, being released in small doses when the flower is shaken by animals or breezes, as in tomato (*Solanum lycopersicum)* and Ericaceae; and (6) be dislodged from the anther and presented in another part of the flower (secondary presentation), when flowers are small and disposed in inflorescences where there is no space to expose pollen in the anther, as in daisies such as *Bellis*.

Pollen may be presented for different lengths of time: for zero time, when it is launched, leaving the anther when it opens, as in many anemophilous species having PHP; for a few hours to a month, as in many entomophilous species having PDP; for longer, in the case of orchids.

Pollen presentation ceases when the flower closes, as in some species with PHP, probably to avoid dispersal of unviable pollen, or when the anthers are discarded.

Animals are attracted to flowers by the prospect of rewards or shelter, or by misleading messages (deception). Common rewards include pollen and nectar, both rich in nutrients: the former contains more proteins whereas the latter contains more carbohydrates. Pollen is collected actively by animals that feed on it and/or passively by those collecting nectar. Each visit to a flower is associated with pollen uptake and discharge.

Factors Affecting Pollen Exposure, Dispersal, and Success

The timing for anther opening and pollen release varies with geographical area and season. There are few reports of anthers closing when the weather is wet or during rain. Many flowers open when insects start to fly. In an anemophilous temperate-zone species such as *Mercurialis annua,* that blooms all year around, anthers open around 7 a.m. in summer and 11–12 a.m. in winter. In temperate zones, anthers of anemophilous and entomophilous flowers generally open from 8 a.m. to 2 p.m., night pollination is restricted to dry summer periods, and night mists purge the air of pollen. In tropical countries, pollination occurs around the clock, but night pollination is always zoophilous. In temperate countries, pollen vectors vary with the seasons, anemophilous trees pollinating in late autumn and late winter–early spring, when many have shed their leaves. The period when entomophilous pollination may occur progressively reduces from the tropics to the poles. **Table 1** shows the main environmental parameters affecting pollination.

Pollen may have different probabilities of effecting legitimate pollination (**Fig.1**), depending on the dispersing vector, and the distances it may travel vary (**Table 2**). Pollen vectors have different specificities, that is, possibility of transporting pollen to the right landing site. **Table 2** shows the specificity, efficiency, and distances reached with different pollination vectors. The main features of anemophilous and entomophilous pollination are shown in **Table 3**.

Table 3 Features of common pollination syndromes

Features	*Entomophily*	*Anemophily*
Habitus	Isolated herbs, shrubs, and trees	Trees and social herbs such as grasses
Environment and season	Tropical, all year aroundTemperate, only spring and summer	Tropical, only dry periodsTemperate, mainly late autumn and late winter
Inflorescence	Different types or solitary flowers	Often pendulous and monoecious
Flowers	Hemaphrodite, rarely monoecious or dioecious	Monoecious, dioecious, rarely hermaphrodite
	Large	Inconspicuous, not attractive
	Stamen and stigma often inside corolla	Pendulous stamen with long filaments and stigma often outside corolla
Pollen	With abundant ornamentation	With reduced ornamentation
	With pollenkitt or devices for mass transport	Free, independent grains
Ovules/ovary	Generally many	Generally one

Abiotic pollination appears random since there are no mechanisms to ensure cross-pollination. Pollen is released into air or water, the movements of which disperse and transport it. When air or water speed is high, pollen may be dispersed long distances, to places where no individuals of the species grow. The further the pollen is dispersed, the less its chance of finding the right female counterpart. Pollen flight is quick and short in farmed anemophilous species, such as wheat, oats, rice, and corn, all of which have PHP.

Although anemophily may seem random, in some gymnosperms and angiosperms at least, anemphilous pollen dispersal ceases being random when pollen grains approach the tip of a cone or flower. By virtue of the shape of the grains and female parts, pollen is conveyed by air currents to the right landing site. Pollen lands on the stigma by gravity (anemophily), or because a pollen-dusted insect incidentally touches the stigma. Electrostatic forces are invoked to explain pollen uptake by insects and release on the stigma.

Biotic pollination first occurred when an animal touched an anther and incidentally delivered pollen to the female counterpart. Interactions between partners began in this way, sometimes leading to species-specific relationships, which are dangerous because if the pollinator becomes extinct or rarefied, the plant can no longer reproduce sexually and will die out if unable to reproduce vegetatively. While pollination is important for plants, being one of the first steps of plant sexual reproduction, for the animal counterpart it represents a source of food: pollen and/or nectar. Abiotic pollination is less expensive for plants because investment in rewards is not necessary; however, investment in an excess of pollen is necessary for random dispersal.

Further Reading

Ackerman, J.D., 2000. Abiotic pollen and pollination: Ecological, functional, and evolutionary perspectives. Plant Systematics and Evolution 222, 167–185.
Cooper, R.L., Osborn, J.M., Philbrick, C.T., 2000. Comparative morphology and ultrastructure of the Callitrichaceae. American Journal of Botany 87, 161–175.
Dafni, A., Kevan, P.G., Husband, C., 2005. Practical Pollination Biology. Cambridge, ON: Enviroquest.
Gelbart, G., von Aderkas, P., 2002. Ovular secretions as part of pollination mechanism in conifers. Annals of Forest Science 59, 345–357.
Linder, H.P., Midgley, J., 1996. Anemophilous plants select pollen from their own species from the air. Oecologia 108, 85–87.
Lindgren, D., Paule, L., Xihuan, S., *et al*., 1995. Can viable pollen carry Scots pine genes over long distances? Grana 34, 64–69.
Nepi, M., Franchi, G.G., Pacini, E., 2001. Pollen hydration status at dispersal: Cytophysiological features and strategies. Protoplasma 216, 171–180.
Pacini, E., 2000. From anther and pollen ripening to pollen presentation. Plant Systematics and Evolution 222, 19–43.
Pacini, E., Hesse, M., 2005. Pollenkitt – Its composition, forms and function. Flora 200, 399–415.
Thien, L.B., Azuma, H., Kawano, S., 2000. New perspective on the pollination biology of basal angiosperms. International Journal of Plant Sciences 161 (supplement 6), S225–S235.
Vaknin, Y., Gan-mor, S., Bechar, A., Ronen, B., Eisikowithch, D., 2001. Are flowers morphologically adapted to take advantage of electrostatic forces in pollination? New Phytologist 152, 301–306.

Principal Components Analysis☆

Craig Syms, James Cook University, Townsville, QLD, Australia

Glossary

Biplot An ordination diagram that simultaneously presents dependent variables.
Centroid The (weighted) mean of a multivariate data set. Can be represented by a vector.
Correlation coefficient A measure of strength of the relationship between two variables.
Eigenanalysis The process of finding eigenvectors and eigenvalues of a matrix. Sample scores are often eigenvectors while eigenvalues are usually ranked from highest to lowest, and termed the first, second, third, etc. eigenvalues.
Matrix A set of numbers arranged in rows and columns. A correlation matrix consists of correlation coefficients and is an example of a square symmetric mtrix. The rows and the columns represent the variables. A Covariance matrix consists of covariant entries, where the diagonal elements will equal the variances. A similarity matrix reflects the degree of similarity or dissimilarity between groups and are easily produced from, or converted into, a distance matrix.
Ordination The ordering of a set of data points with respect to one or more axes so that relationships among the points in any number of dimensions can be visible on inspection.
Principal Components The axes of a principal components analysis. The first principal component will, ideally, represent the dominant gradient. The second component will be orthogonal to the first, and will explain some of the residual variation. The third will be orthogonal to the first and second components, and so on.

Introduction

Principal components analysis (PCA) is a distance-based ordination technique used primarily to display patterns in multivariate data. It aims to display the relative positions of data points in fewer dimensions while retaining as much information as possible, and explore relationships between dependent variables. In general, it is a hypothesis-generating technique that is intended to describe patterns, rather than test formal statistical hypotheses. Although PCA was originally developed to analyze continuous variables, it can also be used on ordinal and presence–absence data.

PCA is carried out on the response of dependent variables in a multivariate data set. Consequently it is an unconstrained ordination, in which hypothetical causal independent variables are not explicitly included in the analysis. For example, if the abundance of several species of fish (the response or dependent variables) were measured at a range of different sites with different characteristics such as wave exposure (causal or independent variables), the exposure information would not be explicitly included in the analysis. Patterns recovered in PCA are solely a function of relationships between the dependent variables. For this reason, PCA can also be classified as an indirect gradient analysis, in which hypothetical causal processes such as exposure, moisture, etc., are inferred from patterns in the dependent variables. PCA assumes that the relationships between dependent variables are linear. This implies that PCA should be applied when most dependent variables have nonzero values across most of the samples, and that bivariate scatterplots of each variable against each other variable should be linear or at least monotonically increasing or decreasing. PCA is a very useful analytical tool, and is one of the most widely used ordination methods in ecology.

☆*Change History*: March 2018. H R Pethybridge included the addition of a glossary and keywords, updates to the abstract, and a revised references list.

An Informal Explanation

Given a multivariate data set consisting of a number of samples in which many variables have been recorded intuitively, PCA is a process in which the original variable axes are aligned along lines of variation in the data and the values for each sample on those new axes are calculated. For example, consider a data set containing 10 samples of abundances of two species. The relative position of each sample in two dimensions can be displayed with a scatterplot of species A versus species B (Fig. 1A). A new set of ordination axes can be generated by moving the old axes to the center of the data set by subtracting the mean of each variable from the sample values—a process known as centering—then rotating these axes so that they lie along the major lines of variation (Fig. 1B). In PCA, the first axis (principal component 1) lies along the greatest line of variation, the second axis lies along the next greatest line of variation on the condition that it lies at right angles to the first, and so on for subsequent axes. This guarantees a property known as orthogonality; which means that each axis is independent of each other. The next step is to project the sample positions onto these new axes—these axes are called principal components (PCs) (Fig. 1C).

In this two-species example, it is possible to display all the variation in a two-dimensional (2D) scatterplot of the original variables. However, if the aim was to explain as much variation as possible in only one dimension (i.e., a line) then the PCs have an important advantage over the variable values. In this example the species explain similar amounts of variation because their variance is approximately the same. At best, about 50% of the entire variation in the data could be displayed in one dimension by plotting values for a single species. In contrast, a plot of the first PCs in this example would explain 85.5% of the total variation in the data set. PCA partitions the variation so that the first PCs will explain more variation than any single variable, assuming there is some correlation between variables. The importance of this is apparent when there are more than two species in a sample. More variation can be presented in a plot of two PCs than can be presented by plotting any pair of species. The PCs can also be interpreted in terms of the original species abundances. A projection of the original centered species axes into the reduced space plot can be used to derive a measure of association of that species with the PC axis (Fig. 1D). In this example, samples that lie on the left of the axis had low abundances of both species A and species B, whereas samples that lie to the right of the axis had high numbers of both species.

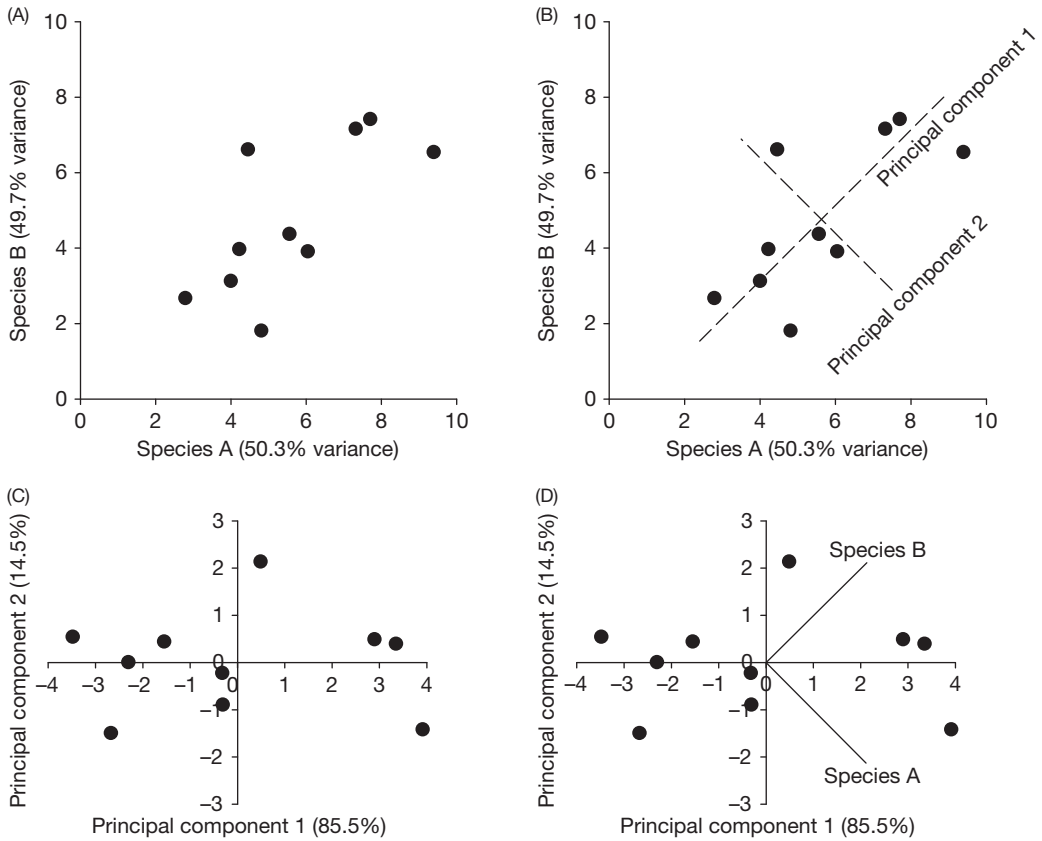


Fig. 1 Deriving principal component (PC) axes of a two-species data set. (A) The original data points can be displayed as a scatterplot of the two species. (B) A new set of axes (PCs) can be derived by placing axes at the center of the data mass, rotating the first axis along the main line of variation in the data set, and rotating the second axis along the next line of variation, conditional on independence with the first axis. (C) The position of the data points on the PCs can be plotted as a reduced space plot. (D) The direction of the original centered species axes can also be projected into the space to generate a biplot.

Calculation of PCA

Mathematically, PCA can be calculated from a mean-centered (i.e., the mean of each variable is subtracted from all values of that variable) data matrix, **Y**. From **Y**, the covariance matrix is calculated using the formula: $1/(n-1)\mathbf{Y}'\mathbf{Y}$ (i.e., the sums of squares and cross-products matrix, rescaled by the n – 1 degrees of freedom). This square, symmetric matrix can be decomposed by an eigenanalysis or "singular value decomposition" into eigenvalues (**L**) and eigenvectors (**U**), which are normalized or scaled to a length of 1. The eigenvalues represent the amount of variation explained by each axis and are usually expressed as a proportion or percentage of the sum of all the eigenvalues. The PCs (**F**) are calculated by projecting the mean-centered data into the ordination space by postmultiplying the centered data by the eigenvectors: $\mathbf{F} = \mathbf{YU}$. An important point to note is that the value of a sample on the PC is a linear combination of the values of the variables in the sample, multiplied by their corresponding eigenvectors. The eigenvectors represent the projection of the old species axes into the new ordination space.

An alternative method of calculating PCA is to use an iterative method such as the two-way weighted summation (TWWS) algorithm. This method starts with a mean-centered data matrix, and arbitrary initial scores on the first PC axis are assigned. The eigenvectors on the initial PC scores are calculated, and then the sample PC scores on these eigenvectors are calculated and rescaled to a length of 1. An estimate of the eigenvalue is obtained from the standard deviation divided by the number of samples, and the procedure is rerun until the eigenvalue does not change with further iterations. Upon convergence, the eigenvectors are scaled to a length of 1, and the PCs are scaled to the eigenvalue. Subsequent axes are calculated in a similar way, except that the PC score estimates at each iteration stage are made uncorrelated with previous ones using the Gram–Schmidt orthogonalization procedure. Both methods yield the same result (within iterative tolerance limits). The eigenanalysis method is easier to program in languages that support matrix operations, whereas the TWWS algorithm can be more efficient for very large data sets because each PC axis is calculated sequentially.

Presentation and Interpretation of PCA Results

A plot of the samples on the PC axes is the primary output of PCA. This reduced space plot displays the relative positions of samples in multivariate space in fewer dimensions. Although a simple scatterplot of samples on the PC axes might provide some useful information on data structure—for example, whether samples are clustered together or occupy a gradient—additional information is usually included on the plot to assist interpretation. This can be illustrated by a PCA of triplefin (Pisces: Tripterygiidae) fish abundance at a range of sites in northeastern New Zealand. The data were collected from sites with different exposure and location characteristics and so graph symbols could be used to reflect these characteristics of the samples (Fig. 2A). This contributes to the interpretation of patterns in the samples based on additional information and is an informal exploration that can identify hypotheses about causal processes. There appears to be a gradient in triplefin assemblages across exposure gradients from sheltered to exposed sites, but assemblages on offshore exposed and sheltered mainland sites are distinct from the semiexposed and exposed mainland sites (Fig. 2A). Information about the value of individual variables can also be included in the reduced space plot to identify which variables are responsible for the observed patterns. If plot symbols are scaled to reflect the value of a single variable in the analysis we see that the triplefin *Forsterygion varium* was more abundant in semiexposed and exposed mainland sites, but relatively uncommon on sheltered mainland sites and practically absent from exposed offshore sites (Fig. 2B). Presenting multiple bubble plots of species abundances is often not a suitable option due to the large number of graphs required; so a more compact and formal presentation of the dependent variables can be generated by plotting the eigenvectors into the reduced space plot (Fig. 2C). This presentation is known as a biplot, and follows from the mathematics of PCA in which the samples are projected into the space by premultiplication of the eigenvectors. In this example the importance of *F. varium* in characterizing mainland exposed/semiexposed sites is clear from the length and direction of its eigenvector. *Notoclinops segmentatus* is characteristic of exposed sites, regardless of mainland or offshore status, and *F. lapillum* is characteristic of sheltered sites. Examination of bubble plots and species–frequency histograms at each site support this interpretation.

Numerical Scale: Transformation and Standardization

Like most (if not all) multivariate methods PCA can be sensitive to data transformation and standardization. Because PCA is an eigenanalysis of a variance–covariance matrix, which is dependent on the numerical scale of the data, variables with large absolute values will dominate the data structure. If the data table consisted of variables measured on different scales (e.g., abundance, kilograms, milliliters, pH) then this scale dependency could exert unwanted effects on the analysis. In addition, a quantity such as a volume if measured in milliliters in one sample, for example, would exert more effect than a volume measured in liters in another sample, even if both samples contained the same volume. In the triplefin example, the PCA of the covariance of untransformed triplefin abundance data was dominated by the numerically dominant species, *N. segmentatus* and *F. varium* (Fig. 3A) and largely insensitive to less-common species. This may be a problem if the intent of the analysis is to retain information on less-abundant species or, more generally, variables with small but biologically important values.

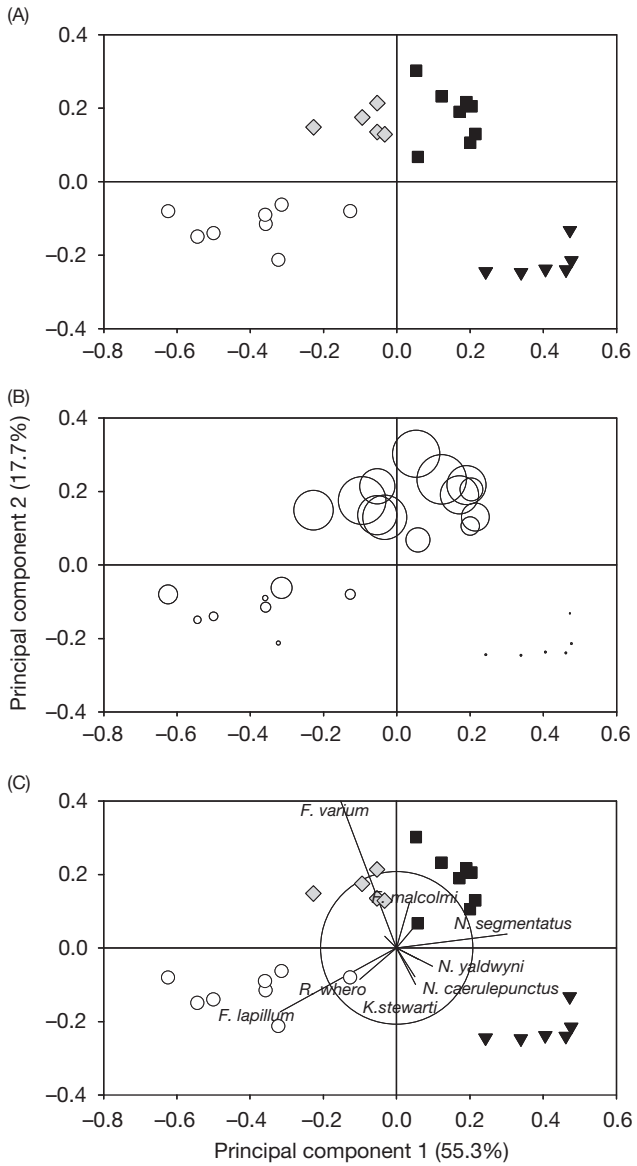


Fig. 2 Reduced space plots of a principal components analysis (PCA) of triplefin fish (Family: Tripterygiidae) abundances. The PCA was calculated using the covariance matrix of the square root of the proportional species abundance in a sample. Percent variance explained is derived from the eigenvalues. Note the equal scaling of the *x* and *y* axes—this ensures the ordination space is not distorted in the plot. (A) Information about wave exposure and location of sites is added by changing *plot* symbols: *circle* indicates sheltered mainland; diamond indicates semiexposed mainland; square indicates exposed mainland; downward triangle indicates exposed offshore. (B) Symbol size can be scaled in proportion to the value of the variables, in this case triplefin species abundance. *Forsterygion varium* is characteristic of exposed and semiexposed mainland sites, uncommon in sheltered sites, and practically absent from offshore exposed sites. (C) A joint presentation of the reduced space and eigenvectors forms a biplot. Distances between sites approximate the Euclidean distance of the transformed data and the eigenvectors are the projection of the original species axes into the space. Eigenvectors have been rescaled to half of their original value for clarity in the plot. The circle is the equilibrium contribution of the eigenvectors. Species outside this circle are influential in defining the ordination space. In this example, *Notoclinops segmentatus* is characteristic of exposed offshore and exposed mainland sites, *F. varium* is characteristic of exposed and semiexposed mainland sites, and *F. lapillum* is characteristic of sheltered sites.

There are two ways to reduce the effect of variables with large absolute values, and increase the effect or weight of variables with small values. First, the data can be centered and transformed to standard deviation units. This process is called standardization, and is implicit in many software implementations of PCA. If data are standardized to unit variance prior to the analysis, then PCA becomes an eigenanalysis of a correlation rather than a covariance matrix. The effect of this standardization is to give all variables equal weight in the analysis and is commonly used and recommended in ecological applications. In contrast with the covariance matrix PCA, an analysis of the correlation matrix of untransformed triplefin abundance data yields an ordination in which both common and uncommon species are important in defining the ordination space (Fig. 3B). Second, data can be numerically transformed using functions such as a square-root, fourth-root, and log transform. Transformations are often used to improve linearity between variables or to reduce the effect of variables with large values in the analysis. However different transformations

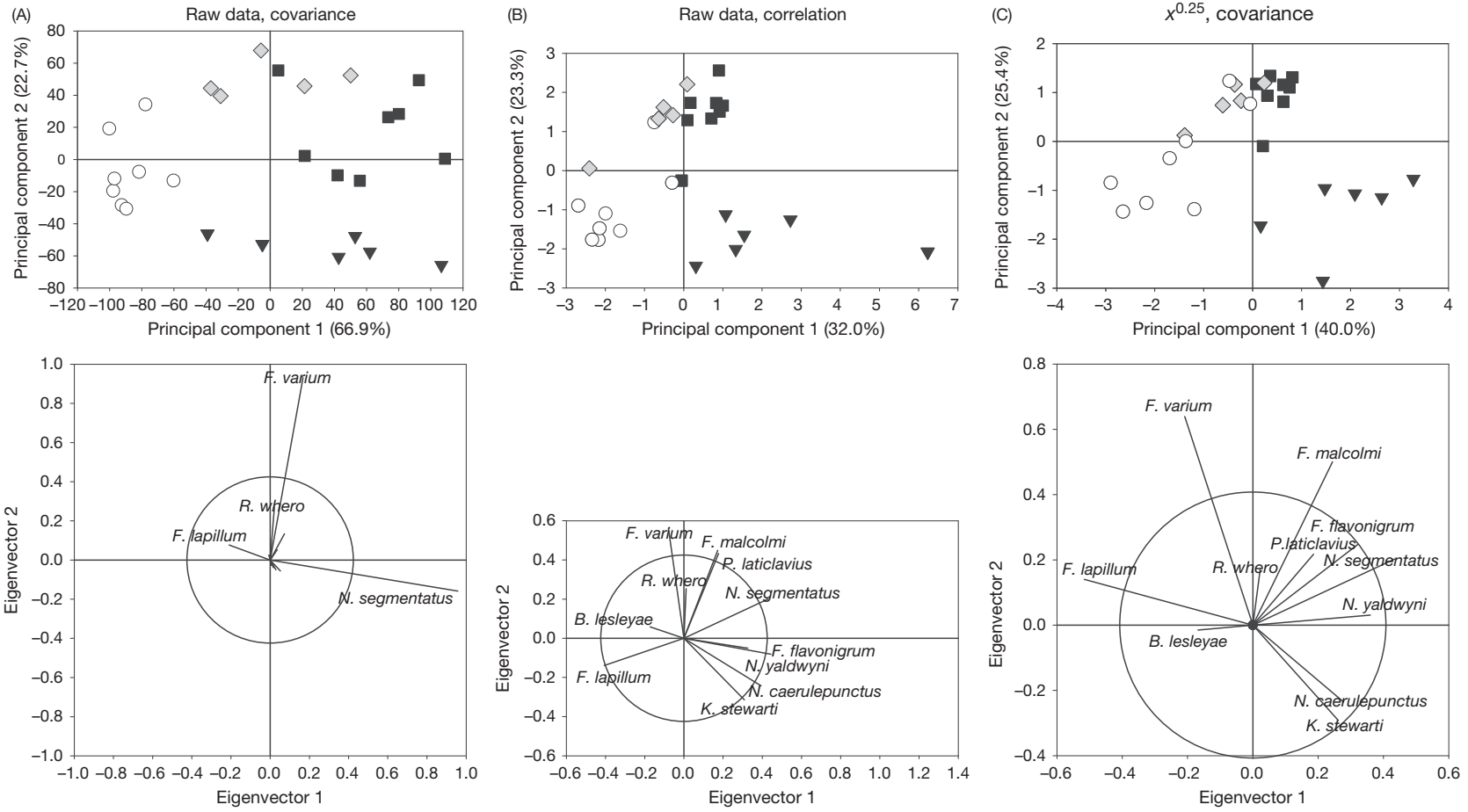


Fig. 3 Effects of data transformation and standardization on PCA of triplefin assemblages. Top row of graphs are the reduced space *p* corresponding eigenvector plots with equilibrium contribution circles. (A) Untransformed covariance matrix analysis is strongly influenced species, *Notoclinops segmentatus* and *Forsterygion varium*. (B) Untransformed correlation matrix analysis reduces the influence of the nu increases the weight of the rarer species. (C) Covariance analysis of fourth-root transformed data also reduces the influence of the numerically dominant species, and increases the weight of the rarer species.

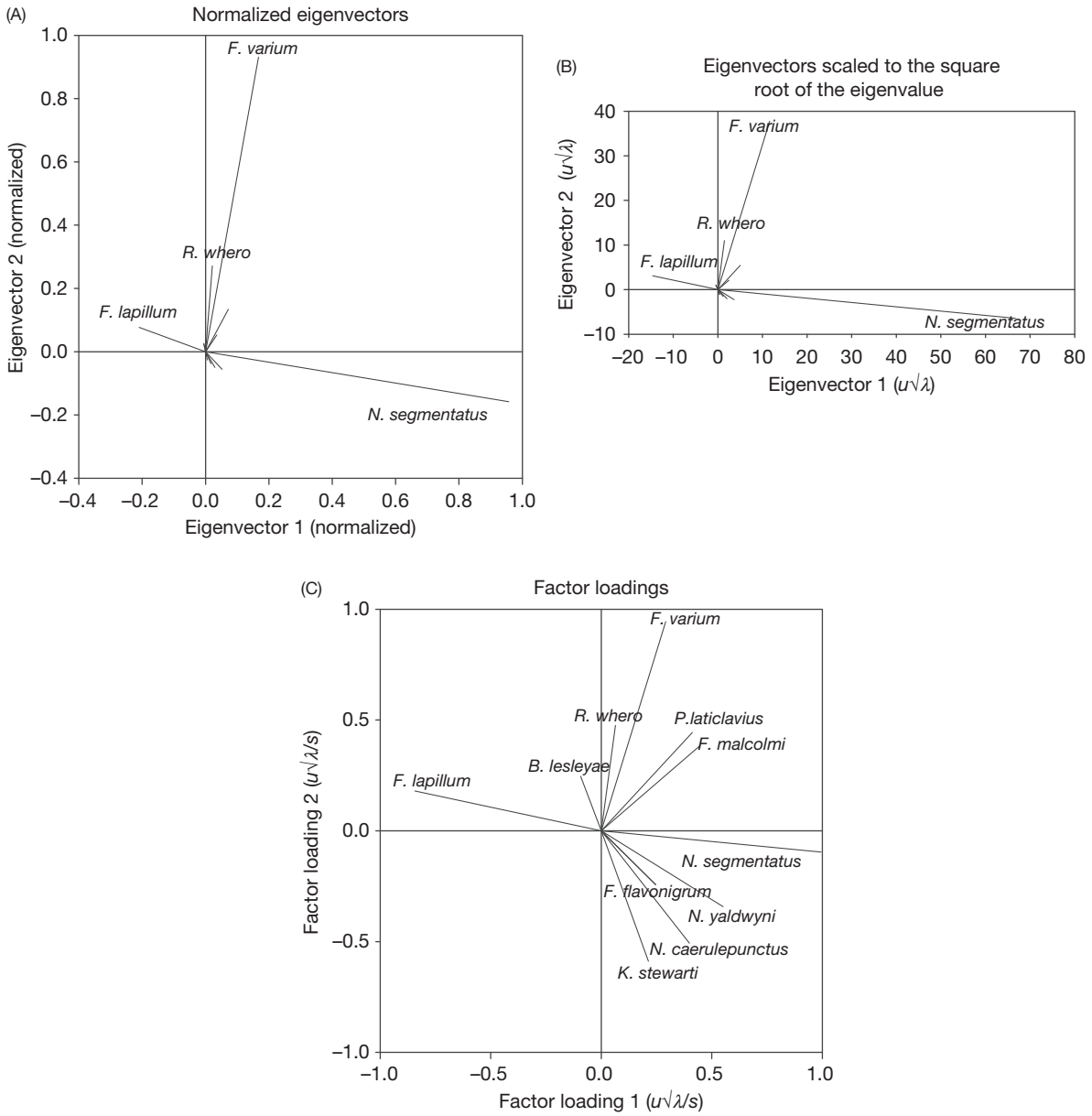


Fig. 4 Effects of eigenvector scaling on a PCA of the raw covariance matrix. (A) Normalized eigenvectors associated with distance biplots are scaled to a length of 1. (B) Covariance biplots scale the eigenvectors so that the length of the vector approximates their standard deviation, and the cosine of the angle between variables approximates their covariance. (C) Factor loadings rescale the covariance biplot scaling by the standard deviation of the variable. This gives an estimate of the importance of the axis in explaining the variance of the variable—it does not represent the importance of the variable in explaining the axis. Note that if the PCA was carried on the correlation matrix the factor loadings would be equal to the covariance biplot because each variable's standard deviation is made equal to 1. In this example the ordination space is defined primarily by two species—*Notoclinops segmentatus* and *Forsterygion varium*.

will alter the importance of different variables in defining the ordination space, and hence may alter the ecological interpretation. In general, increasing levels of transformation (e.g., $x^{0.5}$, $x^{0.25}$, $\log(x)$) will progressively shift analytical emphasis from abundance to compositional aspects of the data. For example, less-abundant triplefin species assume more importance in a covariance matrix PCA with a fourth-root transform ($x^{0.25}$), although not to the same extent as a PCA on the correlation matrix (Fig. 3C).

Biplot Scaling

In addition to data-scaling considerations, biplots can also be scaled differently, which may in turn alter their interpretation. Two common biplot scalings are used to display PCA. A superimposed plot of the PCs (**F**) and the normalized eigenvectors (**U**) is known as a distance or Euclidean biplot. In this biplot, the PC scores are scaled so that their sums of squares equals the eigenvalue (λ) for a given axis, the positions of samples in ordination space approximate their distance in Euclidean space, and the eigenvectors represent the projection of the dependent variable axes into the ordination space (Fig. 4A). The length of the eigenvector indicates

the contribution of the variable to the space—an eigenvector approaching a length of 1 indicates that the variable contributes strongly to defining the ordination space. In addition, the approximate values of the dependent variables in each sample can be reconstructed by projection at right angles of the sample values onto the eigenvector axes. Another common scaling is to scale the eigenvectors to equal their standard deviation ($\mathbf{UL}^{0.5}$) and standardize the PC scores to unit sum of squares ($\mathbf{G} = \mathbf{FL}^{-0.5}$). This is the covariance (or correlation if the data have been standardized) biplot. Unlike the Euclidean biplot, the distances between samples in the reduced space do not approximate their Euclidean distances—they have been standardized by a variance measure. In the covariance biplot the eigenvectors are rescaled to equal the square root of the eigenvalue (cf. normalized in the Euclidean biplot). This projection effectively rescales the eigenvectors to standard deviation units, and has some interesting properties. The length of the vector approximates the standard deviation of the variable, not its contribution to the ordination space. The angle between dependent variable vectors provides a measure of their covariance: covariance $\approx \cos\theta$, where θ is the angle between dependent variable vectors (Fig. 4B). If the PCA has been carried out on standardized data, then this angle will represent the correlation. These angles will only provide a good covariance or correlation estimate if the number of samples is large, the vectors are well represented in the analysis, and the variation explained by the axes is large. Both biplots have the property that the centered data can be reconstructed from the sample scores and the variable vectors: $\mathbf{FU}' = \mathbf{G}(\mathrm{UL}^{0.5})' = \mathbf{Y}$.

The correlations between the original variables and the values of the samples on the Euclidean PC axes may also be used to project dependent variables into a reduced space plot. These values are often termed factor loadings or factor patterns, but their usage should be treated with caution. If the PCA has been carried out on standardized data (i.e., the correlation matrix) then the covariance biplot eigenvector scaling ($\mathbf{UL}^{0.5}$) is equal to the factor unimportant. Conversely variables with large variance might appear to be strongly associated with an axis, when in fact they contribute nothing to its construction. Factor loadings describe how important an axis is to a variable, not how important a variable is to an axis (Fig. 4C). The rationale for this approach comes from a related method—"factor analysis"—which considers measured variables as a function of a hypothesized causal process represented by the PCs, rather than the variables defining a reduced ordination space.

Adequacy of the PCA Solution

PCA generates as many PC axes as there are variables. The axes with the larger eigenvalues hopefully describe trends in the data, whereas the axes with smaller eigenvalues simply represent random variation. There are no authoritative rules for deciding how many PCs are interpretable. Initial recommendations were based on the cumulative percentage of variation explained by the eigenvalues. However ecological data sets differ in their correlation structure, so defining an arbitrary level of variation (e.g., 75%–95%) is not a biologically relevant criterion and its use has been widely disregarded. The plot of the eigenvalues against the axis order (a scree plot) can guide the identification of "important" PCs (Fig. 5). Scree plots can be used to visually identify breaks between PCs that potentially explain trends, and those that represent statistical noise. Typically the trivial eigenvectors on the right of the scree plot will form a linear series, and major magnitude changes on the left may represent trends. The efficacy of this visual determination of breaks is dependent on the underlying data structure. The Kaiser–Guttman criterion requires that eigenvalues that exceed the average expectation should be retained. In a PCA of the correlation matrix, all variables have equal variances and hence the sum of eigenvalues is equal to the number of variables. Consequently the Kaiser–Guttman criterion on a PCA on the

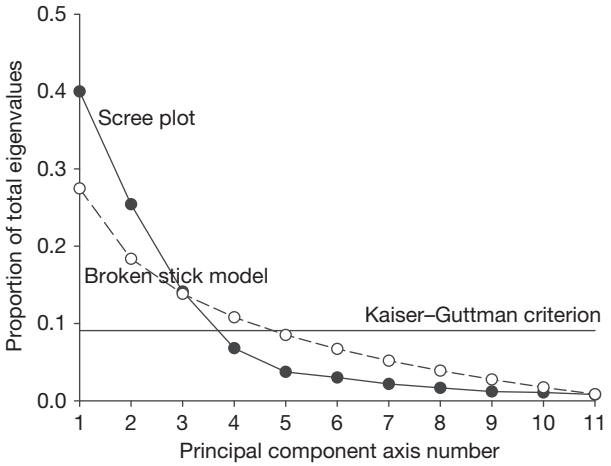


Fig. 5 Scree plot of the proportion of variation explained by each successive principal component (PC) of an analysis on the covariance matrix of $x^{0.25}$-transformed triplefin data (*filled circle*). The Kaiser–Guttman criterion (average proportion explained by eigenvalues) suggests that three axes should be retained for analysis. The intersection of the broken stick model (*open circle*) with the scree plot suggests that two axes should be retained for analysis. The inflection point of the scree plot also suggests that three axes should be retained. While it is clear that at least two axes should be retained, most users of PCA would also examine the third axis to ascertain if it described some ecologically interpretable pattern.

correlation matrix is that eigenvalues greater than 1 should be interpreted. While intuitively the Kaiser–Guttman criterion seems reasonable, there is sampling variability in ecological data sets and so the average expectation may not be a suitable null model. An alternative approach is to use the "broken stick model" to identify the null distribution of eigenvalues, if there was no structure in the data (Fig. 5). Expected eigenvalues for a given axis under the broken stick model can be calculated as $b_k = 1/p \sum_{i=k}^{p} 1/i$, where p is the number of variables, and b_k is the expected proportional eigenvalue for the kth component. Computationally intensive randomization tests such as bootstrap confidence intervals can also be used to identify which eigenvalues are nontrivial. Formal statistical tests such as Bartlett's test of sphericity, and both Bartlett's and Lawley's test of homogeneity of the correlation have generally fared poorly in simulations. A general recommendation would be to use the broken stick model to identify nontrivial PC axes if bootstrapping was not available.

Similar issues exist for interpreting which eigenvectors are important in a PCA. When eigenvectors are normalized, their total length is 1. Consequently if an eigenvector on a particular PC axis has a value close to 1 then that variable is well represented on that axis and less well represented on other axes. However if a variable is not strongly associated with any PC, the eigenvectors for that variable should be equal across axes. The expected eigenvector for a variable that is not associated with any PCs is known as the equilibrium contribution, p and is given by $\sqrt{d/p}$, where d is the number of dimensions of interest, and p is the number of variables. Eigenvectors with values larger than the equilibrium contribution for a single axis can be considered to be associated with that axis. Similarly eigenvectors with values larger than the equilibrium contribution for two axes could be considered to be associated with forming a 2D space. If the eigenvectors are not normalized then the equilibrium contribution must be calculated separately p for each variable and is given by $s_j \sqrt{d/p}$, where s_j is the standard deviation of the jth variable. If the eigenvectors are normalized, the equilibrium contribution can be presented on a graph as a circle (e.g., Fig. 2C).

Assumptions, Limitations, and Other Considerations

PCA was originally developed to describe patterns in multivariate normal (MVN) data. However, deviations from MVN are generally not as critical to the success of unless there are fewer samples in the data set than there are variables. Most software packages will still calculate the analysis under this condition, with the restriction that the number of PCs will equal the number of samples rather than the number of variables. In general, the first PC axes will still be interpretable but minor axes may not be because of overfitting of the model. This is analogous to multiple regression analysis in which fitting too many variables with too few data points will yield a degenerate solution. Several guidelines have been suggested for establishing appropriate sample sizes to generate robust PCA solutions. Some researchers have suggested that studies should aim to achieve absolute sample sizes ranging from 50 (very poor) through 200 (fair), 300 (good) up to 1000 or (excellent). Others have suggested that the ratio of samples to variables is of more importance, with minimum values of 5:1–10:1. It is important to note that these recommendations have generally been recommended by users of "factor analysis," in which robust covariance estimates are key to identifying stable analytical solutions. In addition, many of these suggestions stem from the social sciences in which raw data, such as questionnaire responses, are often "indicators" of variables, rather than direct measures of the variable itself. In most ecological applications these sample sizes are unrealistic and the focus of the analysis is on description of an assemblage, rather than the "factor analysis" objective of recovering underlying causal factors. For most purposes, a rule of thumb for PCA would be to ensure that there are more replicates than variables. In general, the greater the ratio of replicates to variables, the better. PCA is a data exploration and display tool—not a hypothesis-testing method subject to strict distributional assumptions—so it should yield useful insight into data set structure even if the replication is not as great as desired. Robustness of the PCA solution can always be evaluated by bootstrapping, as described above for eigenvalues and eigenvectors.

It is important to be aware of software idiosyncrasies when calculating PCA. Many software implementations calculate PCA on the correlation matrix by default. In addition, PCA is mathematically related to "factor analysis" (FA), a method used widely in the social sciences. The main conceptual difference between PCA and FA is that PCA considers the ordination axes as a product of the variables—they are a linear combination of the eigenvectors. FA considers the variables as a product of the axes themselves. In this interpretation, the variable values are "caused" by the hypothetical ordination axes rather than the axes simply reflecting patterns in the data. The mathematical similarity of the two approaches has led to many software packages combining PCA routines into FA routines. Implementations that incorporate FA and PCA may, by default, yield a covariance biplot of the correlation matrix, with the sample scores scaled to 1 and the eigenvectors scaled to their standard deviation (the factor loading or pattern). FA uses the correlation matrix by default, so as outlined above an analysis of the covariance matrix may substantially change the interpretation of the resulting biplot. FA software also offers the option of axis rotations. Rotations are intended to align axes so that factor loadings are maximized, that is, to make variables associated with single axes. These procedures should not be used for PCA, unless the intent is to use PCA in an exploratory FA and not as a descriptor of ecological data. As with all ecological data analysis, it is important to ensure the correct technique is being employed.

PCA is a very flexible procedure. In its basic form, it is an eigenanalysis of a covariance or correlation matrix. Consequently, it is possible to calculate PCA on a nonparametric correlation matrix such as Spearman's rank correlation. This approach can be useful to deal with nonlinearity of variables. It also follows that PCA can also be calculated on a partial correlation or covariance matrix. A partial correlation coefficient is one that has been statistically adjusted for another variable, essentially correcting for a covariate. Ecological applications of partial PCA are rare, but morphometric studies frequently use partial PCA to assess relationships between morphometric variables after correcting for size of the organism. The ecological application is clear. Dominant variables such as

wave exposure, moisture gradients, etc., could be effectively removed from the analysis prior to PCA to yield an ordination that statistically 'corrects' for dominant variables.

Although PCA was developed as an ordination method to summarize patterns in multivariate data, it also has a range of other uses. PCA can be used in multiple regression to detect collinearity of predictor variables, and as a variable-reduction tool. For example, collinear variables could be replaced in a multiple regression with their first PC. This reduction in number of regression coefficients will increase the power and stability of a multiple regression by reducing the number of variables and improving independence of the coefficients. PCs can themselves be used in data presentations and analyses. For example, a contour plot of PCs of spatially structured data could provide information on a range of variables in a single graph.

Further Reading

Abdi H and Williams LJ (2010) Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* 2(4): 433–459.

Abdi H, Williams LJ, and Valentin D (2013) Multiple factor analysis: Principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Statistics* 5(2): 149–179.

Budaev SV (2010) Using principal components and factor analysis in animal behaviour research: Caveats and guidelines. *Ethology* 116(5): 472–480.

Candès EJ, Li X, Ma Y, and Wright J (2011) Robust principal component analysis? *Journal of the ACM (JACM)* 58(3): 11.

Demšar U, Harris P, Brunsdon C, Fotheringham AS, and McLoone S (2013) Principal component analysis on spatial data: An overview. *Annals of the Association of American Geographers* 103(1): 106–128.

Jackson DA (1993) Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology* 74: 2204–2214.

Jolliffe IT (1986) *Principal components analysis*. New York: Springer.

Legendre P and Gallagher ED (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia* 129: 271–280.

Jolliffe IT and Cadima J (2016) Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A* 374(2065): p. 20150202.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, and Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38(8): 904.

Shlens, J., 2014. A tutorial on principal component analysis *arXiv preprint arXiv:1404.1100*.

ter Braak CJF, ter Braak CJF, and van Tongeren OFR (1995) Ordination. In: Jongman RHG (ed.) *Data analysis in community and landscape ecology*, pp. 91–173. Cambridge: Cambridge University Press. ISBN: 0-521-47574-0.

Witten DM, Tibshirani R, and Hastie T (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10(3): 515–534.

Relevant Websites

http://ordination.okstate.edu/—Ordination methods for ecologist.

https://commons.wikimedia.org/wiki/Category:Principal_component_analysis—Wikimedia commons.

Rhizosphere Ecology☆

Corey D Broeckling, Mark W Paschke, and Jorge M Vivanco, Colorado State University, Fort Collins, CO, United States
Daniel Manter, USDA-ARS, Fort Collins, CO, United States

Introduction

The rhizosphere is defined as the region of soil surrounding plant roots which is under the influence of the root. This region is centered around the root, and is best defined by the biotic response to the influence of the root (Fig. 1). Practically, this region is measured using biological indicators such as microbial density, enzymatic activity, or mapping root-derived chemical gradients. Thus, the spatial limits of the rhizosphere are determined by the soil biotic community under the direct or indirect influence of plant roots. The composition and dynamics of this biotic community is dependent on plant species, root architecture, plant carbon allocation, soil physical and chemical properties, microbial population diversity, among a host of other factors.

The plant root system, though comprising approximately half of a given plant's biomass, is relatively poorly studied compared to aboveground tissues. Consequently, examination of the ecological interactions in the rhizosphere lags behind aerial studies. This is primarily due to the technical challenges of working in a complex soil matrix. More problematic still is that much of the rhizosphere community is microbial, and only a fraction of microbe species are amenable to laboratory culture (see below). Though root biology and ecology are more challenging than aboveground studies, the biological and ecological importance of the root system and surrounding rhizosphere has prompted many detailed studies of root biology and rhizosphere interactions.

Roots are highly branched organs which aid the plant in uptake of water and nutrients from the soil. This branched nature results in a vast surface area available for colonization by soil organisms. Due to the challenges associated with quantifying root characteristics in a natural soil matrix, estimates of root surface area vary by orders of magnitude. One study reported that 1 m^2 of soil in temperate grassland ecosystems contains an estimated 80 m^2 of root surface area. Another study reported that a single 1-month-old rye plant can generate 620 km of root length, with over a billion root hairs, and over 600 m^2 of total surface area. Regardless of the precise values (which vary by plant species, soil type, nutrient status, etc.), this incredible surface area generates an abundant and heterogeneous matrix for soil biota to thrive.

The rhizosphere contains a complex food web with the plant as the primary source of carbon. Aerial plant parts convert carbon dioxide to carbohydrates through the process of photosynthesis. Fixed carbon is transferred through the plant vasculature to the root system, generally in the form of carbohydrates, amino acids, and other primary plant metabolites. These compounds serve as carbon and nitrogen substrates to support root system growth and this growth subsequently impacts the rhizosphere by modulating interactions with rhizosphere organisms through the secretion of organic compounds into the soil (root exudation), regulation of border cell release to the soil, and the alteration of the physical properties of individual root cells, the root system as a whole, and the physical properties of the soil. Plant-derived contributions to the rhizosphere subsequently influence the physiology, behavior, fitness, and interactions of the organisms inhabiting that area. These interactions can be positive, negative, or neutral to plant fitness, and each specific contribution to the rhizosphere may induce a specific change in community activity or structure.

Rhizosphere Food Web

The rhizosphere is an exceptionally nutrient-dense region compared to bulk soil, with energy derived from root exudates, sloughed root cells, dead and decaying root tissue, cellular leakage derived from herbivory and pathogen attack, proteinaceous secretions (mucilage), and symbiotic relationships between plants and microorganisms which shuttle carbon from the root to surrounding soil. This rich nutrient source supports a dense and diverse population of primary consumers and an elaborate trophic web with the root as the primary carbon source. The length of a trophic chain depends on the amount of input from the primary producer: the more nutrient input, the longer the theoretical upper limit of chain length. In the rhizosphere, plant roots are the primary source of carbon, with roots serving as a sink for aboveground photosynthetically fixed atmospheric carbon. Detrital food webs are able to

☆*Change History:* March 2018. Irene Martins made minor changes to the text and references.

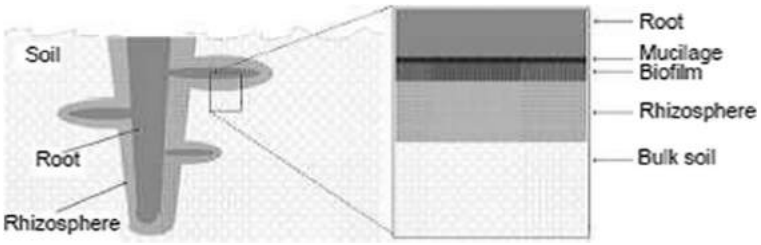


Fig. 1 Schematic representation of a plant root and surrounding rhizosphere. Though drawn here as a discrete boundary, the precise delineation separating bulk and rhizosphere soil is typically vague. Mucilage is secreted by plant roots and typically is composed of polymers such as polysaccharides and polypeptides. Biofilm is a dense microbial layer immediately adhering to the root surface. Shading is approximately proportional to nutrient density in the soil, with a decreasing gradient from the root surface toward the bulk soil.

support anywhere from three to eight trophic nodes, and the rhizosphere food web is predicted to support a web of similar length and complexity.

The primary consumers of plant exudates are microbes including bacteria and fungi. The nutrient source for these organisms includes small metabolites, which may be either actively or passively released by the plant into the surrounding soil. Root exudates include metabolites such as amino acids, organic acids such as citric and malic acids, and secondary plant compounds such as flavonoids and terpenoids. The microbial community is able to utilize these various compounds with some specificity, with individual microbial species more effectively utilizing a given carbon substrate than another species. Further, the root exudates from two different varieties of the same plant can select for different genotypes of the same bacterial species, in part through a differing transcriptional (gene expression) response by the bacteria to the exudates. Plant secondary metabolites, which often contain phenolic ring structures, are particularly resistant to general microbial degradation. However, certain microbial species are capable of utilizing these compounds, despite the fact that these compounds often display general antimicrobial activity. Larger metabolites and biopolymers such as polysaccharides and polypeptides are also actively secreted from the root, which can be degraded and utilized by rhizosphere microbes. Additionally, soil microbes can be pathogenic to the plant, invading and killing the root or root system. Death of a root will rapidly generate a large detrital nutrient pool, containing both intracellular contents and cell wall and membrane components, which can then be utilized by saprophytic fungi or bacteria.

Invertebrates including insects, noninsect arthropods, and nematodes may also be primary consumers, feeding on live, dead, or decaying plant material. The most common root herbivores are soil nematodes, which feed on living root tissue via piercing/sucking mouthparts. The digestive system of herbivorous nematodes serves as a conduit for plant nutrient to be passed from the root to the soil via defecation. Likewise, nematode feeding can induce the death of the root in that region, feeding the saprophytic food web. Alternatively, arthropods such as collembola and soil mites may feed on dead or decaying vegetation. These arthropods will also commonly feed on rhizosphere microbes, thus as a class they may be considered either primary or secondary consumers. Predatory amoeba will feed on soil bacteria and can be consumed by nematodes. Likewise, certain nematode species feed on fungi, bacteria, or even other nematodes, so can be considered primary, secondary, or even higher-order consumers. However, feeding preference will generally be species-specific, with a given species either herbivorous or predatory. Many insect species spend part of their life cycle beneath the soil, and many are specifically adapted for root herbivory. Consumption of the root by an insect (commonly the immature larval stages of beetles) will result in wounds that expose plant cellular contents to the microbial community. Often the larval stage will complete development below ground, and much of the material consumed by the insect passes through the digestive system, again increasing nutrient availability for microbes. Predatory arthropods including insects and mites may also be tightly associated with the rhizosphere, feeding on nematodes, collembola, and mites. So a theoretical food chain might proceed from plant root exudate through bacteria, amoeba, nematode, predatory nematode, predatory mite, predatory insect larvae. These insect larvae can then serve as a food source for larger vertebrates such as mice or birds—extending the food chain above ground.

Rhizosphere trophic cascades are documented in which predatory nematodes cause a decrease in root damage inflicted by an herbivorous moth species. When the predatory nematodes are present, the host plant demonstrates increased growth and seed set over the course of a single growing season. The logical extension of the observed effects (increased growth of the host plant, in this case a member of the leguminous *Lupinus* genus) is that with increased growth and seed production may come increased *Lupinus* biomass and hence increased levels of symbiotic nitrogen fixation by rhizobia.

Biotic and Abiotic Influences on Rhizosphere Properties

The rhizosphere is a highly dynamic region, the properties of which are directly influenced by abiotic factors such as mineral composition and physical properties of the soil. Physical properties such as water permeability, soil texture, abrasiveness, and mineral composition and distribution can determine which plant species survive. Mineral concentrations in even a small region of soil can vary in concentration by 100–10,000-fold. Additionally, minerals can bind to plant-derived organic compounds (a process known as chelating), potentially altering their availability to soil microorganisms. These and similar soil characteristics can affect both plant growth and the microbial community. In this way, abiotic factors can influence root growth and can dictate the biotic effects on the rhizosphere physical and chemical properties (Fig. 2).

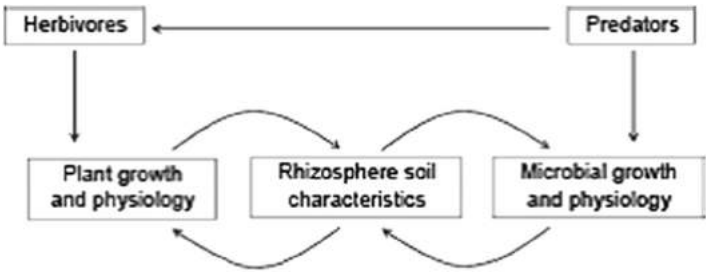


Fig. 2 Simple model of the biotic interactions influencing the rhizosphere.

Plant Modification of Soil Characteristics

After a given plant is established at a site, the soil characteristics will influence rhizosphere properties through effects on plant growth and physiology. Roots function in the uptake of nutrients, and the relative immobility of certain nutrients (particularly phosphate, potassium, and ammonium) in the soil results in local depletion of essential nutrients, which demands new root growth to probe for nutrient-rich regions. Individual plant roots are relatively short-lived. For many agricultural crops, approximately 20%–50% of the individual roots die within a week, an indicator of the dynamic nature of the root system. This implies that any abiotic factor that a plant can detect and respond to may affect the rhizosphere through root death or new growth.

Soil physical characteristics will directly impact plant growth and physiology through a variety of mechanisms. For example, physical abrasion in sandy soil may result in elevated rates of carbon transfer from root to rhizosphere, likely through increased sloughing of cells and secretion of polysaccharide mucilage to prevent root damage. Likewise, a highly compacted soil becomes difficult for plant roots to penetrate, and thereby restricts root mass and surface area, and hence rhizosphere volume.

Plant root architecture is also influenced by nutrient availability. Soils that are low in phosphate, for example, may induce increased production of fine root hairs with a decrease in secondary roots. Likewise, nutrient depletion often results in an increased root: shoot ratio, often with decreased absolute biomass of the root system. Nutrient deficiency or mineral toxicity (such as aluminum) will often result in an increased secretion of organic materials into the rhizosphere. These compounds may be organic acids, which regulate rhizosphere pH and thus reduce the solubility of aluminum in soil water, or higher molecular weight proteins, which are thought to bind and sequester aluminum. Likewise, plants can alter exudation in response to phosphorus deficiency. These secretions may increase availability of sparse minerals and decrease the toxicity of overly abundant minerals. Though a plant may secrete such compounds into the soil for the purpose of detoxification of minerals, such secretion will simultaneously increase carbon availability in the rhizosphere, as many microbes can metabolize these compounds.

Biotic Influences on Rhizosphere Properties

The highly dynamic nature of the rhizosphere is governed in part simply by plant growth and death. However, this view is simplistic in that the inhabitants of the rhizosphere impact soil nutrient status and plant physiology. The plant response to these stresses will further affect rhizosphere characteristics. The rhizosphere is a continuously evolving habitat, the characteristics of which are impacted by interacting biotic components.

Both bacterial and fungal species produce secondary metabolites that are of biological and ecological importance to rhizosphere dynamics. Both clades are capable of producing metabolites that mimic plant hormones such as auxin and gibberellins. As plant growth is governed, at least in part, by hormone signaling, these metabolic products can directly impact root growth and therefore rhizosphere dynamics. Additionally, expression of the genes responsible for biosynthesis of these secondary metabolites can be regulated by environmental factors such as carbon status of the plant, nitrogen status of the soil, and rhizosphere pH—all of which are impacted by plant physiological mechanisms. Further, some of the secondary metabolites of fungi have antimicrobial properties, and bacterial products can inhibit growth of other bacteria or fungi. Thus microbial competition in the rhizosphere is partially mediated by secondary metabolites, in addition to direct competition for organic and inorganic resources. In addition, plant secondary metabolites can influence this competition, favoring those microbes that can metabolize the specific molecules in the root exudates. Further, plant exudates have been demonstrated to regulate the virulence of the soil microbes—that is, whether a particular microbial species is pathogenic (virulent) to the plant or is simply a rhizosphere inhabitant (avirulent).

Microbes may form symbiotic relationships with plants which benefit the plant through increased growth and/or seed production. This relationship can be formed by certain taxonomic groups of fungi and bacteria. In the case of mycorrhizal fungi, the fungus transfers primarily mineral nutrients and water to the plant in exchange for photosynthetically derived carbon. The carbon supports extensive growth of the fungus outside of the rhizosphere, which increases the area available for nutrient uptake by the fungus. This relationship then benefits both partners. Fungal hyphae can extend well beyond the reach of the root system, and this network of fungal influence is called the mycorhizosphere. Dinitrogen-fixing bacteria (diazotrophs) can either be contained within specialized root organs called nodules or living in the soil matrix surrounding the root. Diazotrophs such as *Rhizobia* spp. and *Frankia* spp. are contained within root nodules, and thus contribute little to the rhizosphere nitrogen pool directly. They directly transfer fixed nitrogen to host plants. Nitrogen is a limiting nutrient in many ecosystems, and plants that are able to form symbioses

with nitrogenfixing bacteria often display increased growth following nodulation, which indirectly increases the size and complexity of the rhizosphere.

Plant-growth-promoting rhizobacteria constitute another functionally (not necessarily taxonomically) related group that influences plant growth. However, this class does so without developing an endosymbiosis. In this class, the presence of specific bacterial species in the rhizosphere promotes the growth of the plant through associative dinitrogen fixation, nutrient mineralization and chelation, and protection from pathogens. Many diazotrophs are rhizobacteria and they can directly contribute available nitrogen to the rhizosphere. One study revealed that a possible mode of communication between rhizobacteria and plants is through volatile metabolites produced by the rhizobacteria. When the plant was exposed to these compounds, it responded with increased growth rate.

The properties of the rhizosphere are also dependent on the properties of the aboveground portion of the plant. Herbivory on leaf tissue can alter gene expression in the root system, often resulting in altered susceptibility to soil pathogens. Aboveground wounding can increase the rates of symbiosis between plants and arbuscular mycorrhizal colonization. This wounding has also been demonstrated to result in increased quantities of bacterial-feeding nematodes in the rhizosphere. Conversely, root herbivory will effect aboveground physiology and can reduce fitness parameters, such as seed production. Nutrient status can also affect the ability of a plant to mount an induced defense upon exposure to aboveground herbivory. The major wound hormone of plants, jasmonic acid, can be transported from shoot to root, allowing leaf herbivory to elicit defense responses in roots.

Soil nutrient status will also affect the interactions between rhizosphere inhabitants. Symbiosis between diazotrophs and leguminous or actinorhizal plants is more likely to be established in soil of low-nitrogen availability than soil with abundant nitrogen. Under low-nitrogen conditions, leguminous plants will increase production of flavonoid secondary metabolites which, when released into the soil, serve a communicative role to nodule-forming rhizobia. Likewise, when the rhizobia recognize the presence of a legume root (via detection of the flavonoid signal), they release lipooligosaccharides into the soil which the plant recognizes to initiate nodule formation. These compounds clearly serve a role in communication between two highly coevolved species. Following the establishment of nodules, plants respond with increased growth and the rhizosphere microbial community becomes more active, presumably due to increased carbon availability in the rhizosphere.

Plant roots have the ability to not only contribute carbon to the rhizosphere, but to take organic metabolites from the rhizosphere. The sum of the rate of efflux and influx provides a measure of net contribution to the rhizosphere. Microbial secondary metabolites have been demonstrated to increase the efflux of plant-derived amino acids from the root into the rhizosphere by 200%–2000%. Bacterial *Pseudomonas* spp. produce a metabolite, 2,4-diacetylphloroglucinol, which was found to block amino acid uptake by the plant, while fungal *Fusarium* spp. produce a metabolite, zearalenone, which increases amino acid efflux from the roots of alfalfa. In this way microbes are thought to play an active role in the plants' ability to modify the rhizosphere.

Methods of Studying Rhizosphere Ecology

Understanding rhizosphere ecology and the interactions between plants and soil-borne organisms often requires determination of the identity and distribution of the vast diversity of organism(s) (bacteria, fungi, arthropods, etc.). However, it is this diversity that makes the rhizosphere one of the most difficult communities to study, often dictating the use of a variety of methodologies dependent upon the research objectives and organisms of interest. Traditional methods based on the isolation and growth of live organisms in culture may significantly limit the population being evaluated. For example, it has been suggested that only 1% of a bacterial population can be cultured by common laboratory techniques. It is unknown if this limited sample is representative of the entire population and is unculturable due to a physiological state, or a highly selective sample that is phenotypically and/or genetically suited for laboratory growth on artificial media. To overcome these problems, a variety of methods have been developed including direct observation, fatty acid analysis, chemical, and molecular techniques (DNA and RNA), etc. A detailed discussion of each of these techniques is beyond the scope of this article and the reader should consult one of the many reviews that discuss the advantages and disadvantages in more detail. A brief introduction to some of the available techniques is shown in Table 1. As

Table 1 Summary of some methods for studying the rhizosphere

General methodology	*Example techniques*	*Target organism*
Direct observation	Serial plating counts, trapping	Fungi, bacteria, arthropods
	Rhizotron	Macroinvertebrates, roots
	Funnels, trapping	Arthropods
	Ergosterol, chitin, fatty acid markers	Fungi, bacteria
	Fluorescence microscopy	Fungi, bacteria
Molecular markers	Non-PCR techniques (GC content, reassociation, and hybridization, microarrays)	All
	PCR techniques (DGGE, TGGE, SSCP, RFLP, T-RFLP, RISA)	All—primer specific
Functional measures	Carbon utilization (e.g., Biolog plates)	Bacteria
	Nitrogen fixation rate	Diazotrophs
	Respiration rate	Bacteria, fungi, roots

should be obvious from Table 1, the methodology chosen may significantly influence the type and diversity of organisms identified in rhizosphere ecology studies. In addition, the research must also consider a number of other factors in the design and analysis of rhizosphere ecology studies. For example, considerable temporal and spatial heterogeneity may be observed associated with factors such as plant species and distribution, microclimate, soil physical and chemical properties, and the life stage or physiological state of rhizosphere microorganisms.

Detection and/or isolation by any of the above methods may not lead to a definitive identification of an organism. For many taxonomic groups, there is no official definition of species. For example, the genetic plasticity of bacteria allowing DNA transfer through plasmids, bacteriophages, and transposons complicates the concept of species. Fungal taxonomy has similar problems in identifying vegetative structures, as most taxonomy is based on sexual structures. Species-level arthropod identification is time consuming and typically conducted by systematists, especially when examining immature (larval) specimen. Molecular techniques may alleviate some of these problems but are still limited due to incomplete databases, genetic polymorphyisms, multiple gene copies, and intraspecies variation. The utility of some molecular approaches for studying rhizosphere ecology may also be limited by their inability to separate organisms that are dormant and/or not participating in rhizosphere processes from those organisms that play key roles in the rhizosphere.

Larger-scale techniques such as microscopy and rhizotron-based observation have also been used to examine higher-order rhizosphere structure and organization. Fluorescent microscopy of bacteria, either naturally fluorescing (such as *Pseudomonas* spp.) or strains expressing a fluorescent protein, allows for detailed spatial organization of microbial communities in a relatively intact setting. Rhizotrons are essentially buried glass-walled containers that allow viewing of intact rhizosphere as it expands to the glass surface. Such devices are of limited use for microbial spatial dynamics, but can be valuable as tools for studying larger arthropods and root–root interactions.

Summary

The rhizosphere is a nutrient-rich region of the soil immediately surrounding the plant root. This region is highly dynamic and supports a dense and diverse fauna. Despite the challenges associated with studying ecological interactions in a soil matrix, researchers are beginning to understand the complex ecological interactions occurring in the rhizosphere. Chemical communication plays an integral role in the ecology of the rhizosphere, and new functions for intra- and interspecific signals continue to surface. Much of the knowledge on rhizosphere biology has been revealed by agricultural researchers, who have studied many of the positive and negative relationships between plants and soil microbes. Though detailed descriptions exist for many rhizosphere interactions, the complex and cryptic nature of the rhizosphere will continue to challenge scientists interested in the ecology of the plant–soil interface and its associate biota.

Further Reading

Ahkami AH, White RA, Handakumbura PP, and Jansson C (2017) Rhizosphere engineering: Enhancing sustainable plant ecosystem productivity. *Rhizosphere* 3(2): 233–243.

Allen MF, Swenson W, Querejeta JI, Egerton-Warburton LM, and Treseder KK (2003) Ecology of mycorrhizae: A conceptual framework for complex interactions among plants and fungi. *Annual Review of Phytopathology* 41: 271–303.

Bais HP, Park S-W, Weir TL, Callaway RM, and Vivanco JM (2004) How plants communicate using the underground information superhighway. *Trends in Plant Science* 9: 26–32.

Coleman DC, Crossley DA Jr., and Hendriz PF (2004) *Fundamentals of soil ecology*, 2nd edn., New York: Elsevier Academic Press.

Dazzo FB and Garoutte A (2017) *Rhizosphere, Reference Module in Life Sciences*. Elsevier.

Dessaux Y, Grandclément C, and Faure D (2016) Engineering the rhizosphere. *Trends in Plant Science* 21(3): 266–278.

Garbeva P, van Veen JA, and van Elsas JD (2004) Microbial diversity in soil: Selection of microbial populations by plant and soil type and implications for disease suppressiveness. *Annual Review of Phytopathology* 42: 243–270.

Hawes MC, Brigham LA, Wen F, Woo HH, and Zhu Z (1998) Function of root border cells in plant health: Pioneers in the rhizosphere. *Annual Review of Phytopathology* 36: 311–327.

Hey J (2001) The mind of the species problem. *Trends in Ecology and Evolution* 16: 326–329.

Kent AD and Triplett EW (2002) Microbial communities and their interactions in soil and rhizosphere ecosystems. *Annual Review of Microbiology* 56: 211–236.

Kilronomos JN, Rillig MC, and Allen MF (1999) Designing belowground field experiments with the help of semi-variance and power analyses. *Applied Soil Ecology* 12: 227–238.

Kirk JL, Beaudette LA, Hart M, et al. (2004) Methods of studying soil microbial diversity. *Journal of Microbiology Methods* 58: 169–188.

Klein DA and Paschke MW (2004) Filamentous fungi: The indeterminate lifestyle and microbial ecology. *Microbial Ecology* 47: 224–235.

Mark GL, Dow JM, Kiely PD, et al. (2005) Transcriptome profiling of bacterial responses to root exudates identifies genes involved in microbe-plant interactions. *Proceedings of the National Academy of Sciences of the United States of America* 102: 17454–17459.

Morris CE and Monier JM (2003) The ecological significance of biofilm formation by plant-associated bacteria. *Annual Review of Phytopathology* 41: 429–453.

Phillips DA, Fox TC, King MD, Bhuvaneswari TV, and Teuber LR (2004) Microbial products trigger amino acid exudation from plant roots. *Plant Physiology* 136: 2887–2894.

Torsvil V, Daae FL, Sandaa R-A, and Ovreas L (1998) Review article: Novel techniques for analyzing microbial diversity in natural and perturbed environments. *Journal of Biotechnology* 64: 53–62.

Waisel Y, Amram E, and Kafkafi U (2002) *Plant roots: The hidden half*, 3rd edn., New York: Dekker.

Wintzingerod FV, Govel UB, and Stackebrandt E (1997) Determination of microbial diversity in environmental samples: Pitfalls of PCR-based rRNA analysis. *FEMS Microbiology Reviews* 21: 213–229.

Salinity

DM Talley, San Francisco Bay National Estuarine Research Reserve, Tiburon, CA, USA
TS Talley, University of California, Davis, Davis, CA, USA

Introduction

Salt has played an important role in human history, serving as a currency, figuring prominently in fables and expressions, and being an important nutritional and culinary element. Salt also plays a crucial role in the structuring of organismal assemblages and ecosystems. Salt is an essential element for all living organisms, and is critical to many metabolic functions, including nerve and muscle action, blood pH and electrolyte balance, and cell regulation. Salt is, however, only salubrious within a certain range, with too little or too much leading to metabolic failure. Because of this, salt is a common and often dominant abiotic factor structuring ecological communities. For example, transitions from fields of grasses, herbs, and succulents to tall stands of reeds and cattails as one heads upstream from an ocean is in large part due to the effects of decreasing salinity.

Definitions and Measures

While common `table salt' is sodium chloride (NaCl), salts can more broadly be defined as the product formed from neutralizing an acid, where a metal atom (or positively charged radical) replaces one or more of the acid's hydrogen atoms. Salts are thus neutral ionic compounds composed of positively charged ions (`cations'; e.g., calcium, magnesium, potassium, sodium) and negatively charged ions (`anions'; e.g., bicarbonate, carbonate, chloride, nitrate, sulfate).

Salinity is a measure of the `saltiness' or concentration of salt in water or soil. Salinity is synonymous with halinity, which derives from the word halides meaning chloride, and means the total concentration of soluble salts. An oft-ignored convention is for oceanographers to use the term `halinity' when referring to oceanic conditions, due to the dominance of NaCl in seawater, and `salinity' being used for soil or freshwater systems. Salinity historically has been measured as the ratio of the mass of dissolved salts to the solution in which it is dissolved (e.g., parts per thousand or ppt). More recently, salinity has been measured in terms of practical salinity units (psu) – the ratio of the conductivity of the sample water to a potassium chloride standard (32.435 6 g KCl/kg water); psu is a ratio, and thus is a dimensionless measure of salinity.

Salinity is often measured using conductivity meters, which measure the conductance of electricity through solution. Since salt ions conduct electricity, conductivity is proportional to the concentration of salts in a solution. While measures of conductivity can be converted to salinity, and vice versa, the algorithms can be complex as they also depend upon temperature and pressure. Measurements of salinity can also be inferred through refractance, water density, and sound speed. Refractometers measure refractance, or the bending of light waves as they pass through a solution. Both refractometers and salinity meters require a fluid solution, while conductivity meters can measure conductance of moist sediment slurry.

There are a number of systems for classifying aquatic environments based on salinity. For marine waters, the `Venice system' (or some variation thereof) is commonly used, with freshwater (<0.5), oligohaline (0.5–5), mesohaline (5–18), polyhaline (18–30), euhaline (30–40), and hyperhaline (>40). Other commonly used categories include brackish (0.5–30) and brine (>50). Seawater salinity usually ranges from 32 to 38 and is an average of 35.

Organisms can also be classified with regard to their salt tolerance. Those that can tolerate a wide range of salinities are called euryhaline (e.g., many intertidal fishes, mussels), while those with narrow salinity tolerances are `stenohaline' (e.g., most ascidians, freshwater fishes). Some organisms complete different stages of their life cycle in different salinity regimes. For example, fish that live most of their lives in the sea but that breed in freshwater are called anadromous (e.g., salmon). Conversely, catadromous organisms are those that live in freshwater but breed in saltwater (e.g., eels, Chinese mitten crabs).

Sources of Salt

The seas originally received their salt when a young Earth's atmosphere, filled with hydrogen chloride and other materials, dissolved into the primitive ocean. The majority of the salt, however, gets added to the ocean through the gradual weathering of terrestrial rocks by water (and a smaller percentage through hydrothermal or volcanic inputs). Precipitation is slightly acidic due to carbonic acid that forms when water interacts with atmospheric carbon dioxide. Rain thus not only erodes the rock but its acidity also dissolves minerals and salts, carrying them in solution downstream to the ocean. Additionally, the sodium and chloride ions that are present in freshwater sources are mostly not used by organisms and are transported to the ocean, where evaporation concentrates in flowing water leading to saline conditions. There are also mechanisms by which salt gets removed from the ocean,

Fig. 1 Halophytic plants, such as the pickleweed (*Salicornia bigelovii*) depicted here, have adaptations to allow them to deal with the excess salts in their environment. Pickleweed actively transports excess salt to the tips of the terminal jointed leaves, which are then shed. Photo credit: D. Talley.

keeping the system in steady state, such as corals using calcium to build reefs, ions attaching to clay particles, and salts precipitating out of solution.

Inland water bodies receive salts in a similar manner. In arid regions and water bodies that have no outflow, evaporation can lead to salinization and the formation of inland seas. Examples include the Great Salt Lake in Utah, and the Salton Sea and Mono Lake in California (USA), the Dead Sea in Israel, and the Caspian Sea in Eurasia.

Terrestrial systems may also be saline. Salt may be carried inland from the ocean in prevailing winds and deposited in rainfall and dust, creating salt deposits in sediments. Erosion and release of salts from parent rocks, as well as the isolation and evaporation of ancient seas, may also contribute to the formation of saline soils. Often salts remain buried within the sediment profile, but agricultural practices such as irrigation and the removal and replacement of deep-rooted woody plants with shallow-rooted crops tend to increase salinities. Woody plants with roots of varying depths draw up fresh groundwater and shade the soil surface, thereby reducing evaporation and salinization. Shallow-rooted crops, on the other hand, do little to shade surface soils and do not tap into the groundwater, thereby allowing it to rise upward, pushing hovering salt layers to the surface. Furthermore, irrigation water is often high in minerals and salts, further exacerbating salinization. This has caused a shift in some arid regions from native flora and fauna and valuable agricultural species to salt-tolerant natives and exotics and barren areas, such as in the wheat belt in Australia.

Adaptations to Salinity

Salt affects organisms through alteration of the water balance of cells and through salt toxicity. Since osmotic pressure pushes water toward tissues with higher solute concentrations, and seawater is often more concentrated than the cells of organisms, water tends to flow out of cells in aquatic organisms in the presence of seawater. Similarly, water uptake into organisms and cells is impaired when in saline solutions, resulting in water limitation similar to what occurs in arid systems. Salt toxicity results when salt affects enzymatic activities and energy processes, such as reductions in photosynthesis and respiration. Organisms living in saline environments have two general adaptive strategies – tolerating and avoiding salt stress. Organisms adapted to live in saline environments are called halophiles; in particular, such plants are called halophytes (**Fig. 1**).

Stress tolerance is generally achieved through metabolic or physiological adaptations. Both plants and animals have a diversity of adaptations to various levels of salinity in the environment, as either hyper- or hyposalinity can cause physiological stress. Osmoconformers are organisms that keep their internal fluids isotonic to their environment, that is, they maintain an internal salinity similar to their ambient conditions (e.g., most marine invertebrates, seagrass). Osmoregulators, conversely, maintain a constant osmotic pressure within their bodies by balancing water uptake and loss through the controlled movement of solutes across membranes between internal fluids and the external environment (e.g., most aquatic vertebrates, some marine invertebrates such as fiddler crabs). Organisms may have selective cellular uptake of particular salts, for example, preventing sodium but allowing potassium uptake (salt grass, cordgrass). Plants and animals may take saline water into their tissues but then accumulate organic compounds to increase cell osmotic potential to prevent cellular explosion (e.g., cordgrass). Other organisms have glands through which salt is excreted (e.g., gulls, some salt marsh plants), tissues through which they take up salts to maintain osmotic balance (e.g., freshwater fishes), or move salt to cells which are eventually lost, such as the specialized hairs or leaf tips of halophytes (e.g., salt marsh plants such as pickleweed and black grass) (**Fig. 2**). Succulent plants dilute salts by taking up more water, but still have to regulate salts by sequestering them in cell vacuoles, isolated from the cytoplasm and organelles of the cells.

Fig. 2 This species of side-blotched lizard, *Uta tumidarostra*, lives on islands in the Gulf of California where *in situ* terrestrial production is quite low, and thus feeds extensively on intertidal invertebrates. This species has evolved large nasal salt glands that allow it to excrete the excess salts consumed with its prey. Reproduced by permission of L. Grismer.

Stress avoidance includes regulation or direct avoidance of salt, either through structural adaptations or behavioral responses. In halophytes, for example, structural adaptations may include specialized root cells that may filter out salt and result in the sap consisting of pure water. Other organisms have behavioral adaptations, such as timing of reproduction, emergence, or dispersal to avoid undesirable conditions (e.g., insects, crab larvae). Mobile organisms, such as fish or crustaceans, may be able to avoid hypo- or hypersaline conditions by moving out of an area.

All of these adaptations come at a cost – the energy used to perform these functions is thus unavailable for other physiological demands, such as growth and reproduction. Therefore, organisms dealing with salt stress, like other physical stresses, are usually faced with a tradeoff between coping with salt and facing competition or predation in less saline areas. Most vascular plants, for example, are salt tolerant but would perform better in fresher conditions if not for competition with taller brackish and freshwater plants.

Many of these factors broadly apply to terrestrial organisms as well. Here, osmoregulation is often handled through specialized organs (e.g., kidneys, Malpighian tubules), and a common physiological stress is a lack of salt, as opposed to an overabundance. Nonetheless, similar issues of adaptation and tolerance for high or low salinity environments apply.

Scales of Salinity Variations

Salinity varies over a vast range of temporal and spatial scales, with profound effects on ecological processes at each scale.

Fine Spatial Scales

Patchy salinity patterns. Even within a very localized system, there can be fine-scale variations in salinity influenced by other abiotic variables, such as substrate type and microtopography, and biotic variables, such as substrate organic content and the presence or activity of local species. Substrates of fine particle size and high organic matter, for example, may maintain lower salinities due to lower evaporation rates than coarser, more mineral soils. The presence of microtopography (depressions, peaks, slopes) may also vary evaporation rates and therefore salinity. The presence of organisms that directly or indirectly alter salinity and, in turn, the associated community (i.e., ecosystem engineers) may contribute to fine-scale salinity variability. The presence of shade-casting species, such as plants with dense canopy like trees, grasses, or mat-forming species would likely maintain lower salinities than areas without such (or any) plants. Activities of organisms present could also alter salinity – bioturbators might turn surface soils minimizing evaporation and salinity accumulation, and salt-excreting species, such as tamarisk, may increase local salinities.

Vertical gradients. Salinity gradients exist across intertidal zones, with distance from the sea into the upland. Whether tidal flats, marsh, rocky intertidal, or sandy beach, salinities generally increase in arid regions and decrease in nonarid regions as one heads from lower to upper intertidal. Freshwater upland runoff and evaporation rates of tidal water, wave swash, and sea spray determine the extent and severity of the gradient. Organisms are thought to be limited by abiotic factors, such as high salinity and desiccation or freshwater runoff, in the upper intertidal and biotic factors, such as predation or competition, in the lower intertidal where stresses are not limiting.

Salinity gradients can also be found in water columns. In estuaries where wave energy and turbulence is low, salinity can have complex vertical patterns. Fresher river water often flows over the denser seawater in estuaries, forming a salinity wedge that migrates horizontally with the tides. The duration and thickness of the wedge varies with rainfall and runoff, as well as amount of mixing due to wind and water turbulence. Some organisms, such as crab larvae, may use these salinity gradients and boundaries as

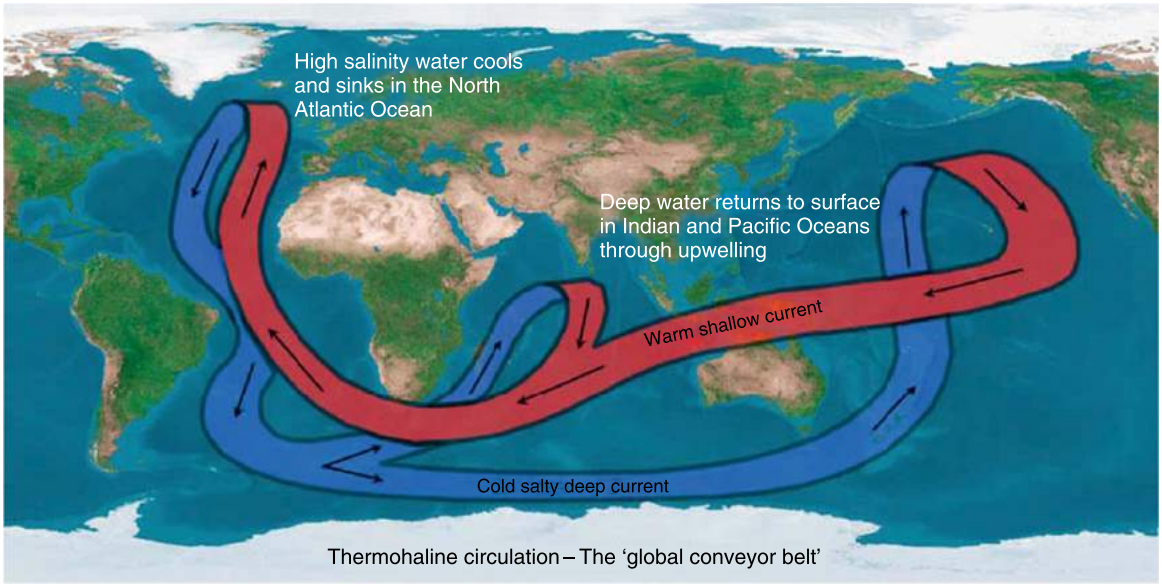


Fig. 3 global thermohaline circulation. Density patterns caused by differences in salinity and temperature drive global oceanic circulation patterns.

a cue for migration direction or settlement. While mobile species may be able to avoid undesirable salinities, less mobile species need to be able to tolerate the daily and seasonal fluctuations associated with tidal cycles and weather patterns.

Broad Spatial Scales

Whole ecosystem. When we think of saline systems, we generally think of marine systems, or those influenced by seawater. Estuaries often form where rivers flow into the sea, resulting in a gradient of salinity that decreases with distance upstream from the ocean. Reverse estuaries form where gently sloped, relatively arid coastal areas are flooded by the sea, resulting in salinity gradients that increase with distance from the ocean due to evaporation and lack of fresh water inputs toward the head of the estuary. The structure of associated communities corresponds to changes in salinity, with shifts from higher proportions of marine taxa, such as polychaete worms, large clams, oceanic fish and crustaceans and seaweeds, to higher proportions of brackish and freshwater taxa, such as insects, aquatic vegetation, small clams and mussels, estuarine fish and crustaceans.

Latitudinal variation. Latitudinal variation in climate may influence the levels of salt stress, which in turn affect the structure and controls on associated communities. For example, coastal marshes of the eastern United States generally have higher salinities in the southern, hotter climate than in the northern, cooler climate. The southern marshes are dominated by succulents, with high salinity restricting the lower tidal limits of non-salt-tolerant plants. In comparison, grasses and herbs dominated northern marshes where flooding determined lower limits.

Global patterns. There are global-scale patterns of salinity, both horizontally (regional high sea-surface salinities; e.g., the Pacific Ocean is fresher than the Atlantic) and vertically (e.g., the North Atlantic Deep Water (NADW) is a mass of dense, salty water, part of which flows at a depth of 2000–4000 m on the Atlantic coast of North America). Both precipitation and evaporation affect salinity making the relatively shallow, semienclosed seas of arid climates hypersaline. These include the Red Sea, the Mediterranean Sea, and the Caspian Sea, among others. In fact, it is the large-scale vertical and horizontal distribution of salinity that helps drive global ocean circulation patterns (the thermohaline circulation of the `global conveyor belt'; **Fig. 3**). These patterns of circulation have profound effects on the ecology of the world, driving climate patterns, propagule distribution, and numerous other large-scale ecological patterns.

Fine Temporal Scales

Transient fluctuations. The salinity of a particular location may be temporarily influenced by transient abiotic or biotic conditions. Temporary reductions in local salinity could be caused by the short-term inputs of freshwater or shading, such as during rain or flood events and additions of wrack mats, detritus, short-lived plants, or algal blooms. Salinity increases could result from removal of freshwater or structure, or addition of saline water, such as during high tide events or tidal surges that scour ground cover and increase seawater inundation, or during short periods of dry, hot temperature (**Fig. 4**). The effects of such events may persist for a time after the actual event ends, for example, salt residue may remain in the upper intertidal following a high tide series.

Cyclic fluctuations. Frequent events such as daily and monthly tidal cycles can cause concomitant fluctuations in salinity of soils and water. Organisms within the tidal zone are either adapted to the salinity shifts or move in and out with the tides.

Fig. 4 Intertidal pools in salt marshes have salinities that are strongly affected by meteorological events (e.g., rainfall, wind), timing of inundation, and ambient air temperature. These pools can change in salinity from more than 125 psu to virtually freshwater in the course of just hours. Photo credit: D. Talley.

Seasonal fluctuation in climate and tides have similar but longer-term effects on the salinity of systems. In many arid and temperate regions, winter and early spring bring precipitation and long pulses of fresh and brackish conditions, while summer and early fall bring drier and therefore more saline conditions. Tropic regions may experience the highest salinities in the winter dry season and the lowest during summer tropical storm events. Organisms not adapted to these conditions die off or move to more favorable environments (downstream to more saline waters if marine, or upstream to fresher conditions if not marine).

Broad Temporal Scales

Succession. Shifts in salinity in systems over longer timescales may in part be driven by succession and also play a part in succession. For example, in disturbed or newly created salt marshes, the first colonizers, consisting only of salt-tolerant species, may shade (e.g., annual succulents) or rework (e.g., rove beetles, burrowing crabs) the substrate surface, reducing evaporation and moderating salinity. Subsequent species may need to be less salt tolerant. As cover of the substrate increases, salinity is further reduced and buffered from dramatic fluctuations.

Decadal and longer. Longer-term fluctuations in the environment and climate may lead to alteration of salinity patterns. For example, the El Niño/Southern Oscillation (ENSO) events can alter patterns of precipitation and current flow, with cascading effects on salinity patterns worldwide. Over geologic timescales, alterations in aquatic salinity patterns have been profound, coupled with changes in oceanic circulation, glaciation and deglaciation, etc., with all of the potential ecological consequences of altered salinity on a global scale.

Summary

Salinity is one of the dominant physical factors structuring terrestrial and aquatic ecosystems, with both overabundance and paucity creating physiological stress. Organisms have evolved a number of adaptations to deal with these stresses, which can affect successional patterns, competitive interactions, and species diversity and distribution. The salinity of any given environment is controlled by both physical and biological processes, and varies over spatial scales ranging from centimeters (e.g., vertical distribution in sediments) to thousands of kilometers (oceanic), and over temporal scales from minutes (e.g., ocean waves cresting a berm) to geologic (the evolution of saline conditions in the oceans).

Further Reading

Bertness, M.D., Hacker, S.D., 1994. Physical stress and positive associations among marsh plants. American Naturalist 144, 363–372.
Broecker, W.S., 1987. The biggest chill. Natural History 96, 74–82.
Kurlansky, M., 2003. Salt: A World History. New York: Penguin.
Maetz, J., 1974. Aspects of adaptation to hypo-osmotic and hyperosmotic environments. In: Malins, D.C., Sargent, J.R. (Eds.), Biochemical and Biophysical Perspectives in Marine Biology. London: Academic Press, pp. 1–167.
Open University, 1995. Seawater: Its Composition, Properties and Behaviour, 2nd edn. Oxford: Pergamon Press.
Open University, 2001. Ocean Circulation, 2nd edn. Oxford: Pergamon Press.
Osmond, C.B., Austin, M.P., Berry, J.A., *et al*., 1987. Stress physiology and the distribution of plants: The survival of plants in any ecosystem depends on their physiological reactions to various stresses of the environment. BioScience 37, 38–48.

Scavengers

OJ Schmitz, HP Jones, and BT Barton, Yale University, New Haven, CT, USA

Introduction

Ecological communities can be envisioned as collections of species that are organized into food chains and webs in which each species is a consumer of resources and is itself a resource for other consumers. Ecologists call these consumptive interactions trophic interactions. And so, species engaging in a particular kind of trophic interaction are said to belong to a distinct trophic group. Ecologists routinely idealize food chains and webs as being comprised of four trophic groups. Species that consume mineralized nutrients and CO_2 in order to photosynthesize carbohydrates belong to the plant trophic group, species that consume living plant tissue belong to the herbivore trophic group, species that prey on herbivores belong to the carnivore trophic group, and species that recycle dead organic material back into the nutrient pool belong to the decomposer trophic group.

Such classic idealization of ecological systems typically ignores another trophic group, scavengers. After all, mobs of bloody headed vultures vying for their share of a carcass or insect larvae crawling in stinking, rotting meat do not engender the same sense of awe and natural wonder as do grazing antelope and lions coexisting on the Serengeti plains of Africa. Scavengers get short shrift in ecological thinking because their role is typically viewed as being a repulsive behavioral curiosity or the ecological equivalent of garbage men that sustain themselves on nature's offals. Scavengers are sometimes viewed merely as parasites that steal food – called kleptoparasitism – from the more noble carnivores. These are, however, unfortunate and inaccurate characterizations of scavenging. As we will show below, scavenging serves an important role to the welfare of many species and it can be an important determinant of the structure and functioning of ecological communities. Moreover, scavenging involves many more species than those few specialists that are routinely highlighted as serving this role.

Scavenging as a Trophic Interaction

Scavenging, like carnivory, involves the act of consuming the flesh of dead animals – carrion. Technically, however, scavenging differs from carnivory in that it does not actively involve killing animals. Scavengers also differ from decomposers. Decomposers like bacteria and fungi break down the protein of dead animals into its constituent carbon-, nitrogen-, and hydrogen-based elements (**Figure 1**). Those elements are then broken down further into mineralized form to be taken up later by plants. Scavengers, on the other hand, consume organismal protein and convert it into their own body tissue (**Fig. 1**).

Research has shown that many organisms die from sources other than predation. Although the exact value varies among species and sizes of prey, predation accounts for between 2% and 75% of organism losses annually, thus leaving 25%–98% to be scavenged. In the Serengeti alone, the annual amount available to scavengers is estimated to be on the order of 26 million kg. Clearly, neither the Serengeti plains, nor any other location globally, is littered with dead animal carcasses, testimony to the magnitude of this trophic interaction. Depending on the size and species of carrion, a carcass can be despatched within hours to days. Research has also shown that scavenging efficiency, defined as the proportion of a carcass that was consumed within this time frame, averages 75%, a value that rivals the efficiencies of carnivores consuming their hunted prey.

Scavenging can be temperature dependent because of interplay between microbial decomposition and chemical detection of carcasses. This interplay leads to intermediate, optimal temperatures for scavenging, especially within temperature regions of the globe. Decomposers alone are rarely able to utilize entire carcasses. So, to avoid competition with scavengers, decomposers have evolved capacities to produce noxious and odorous chemicals that can make the entire carcass distasteful or even toxic. At moderate temperatures (e.g., 10–15 °C) microbial decomposition is at a level that produces modest concentration of chemicals leading to putrid odors that signal the location of edible carrion to scavengers. Indeed, experimental studies have demonstrated that under such conditions scavengers can find and begin to remove carrion within minutes to hours after becoming available. Higher temperatures and associated higher rates of decomposition lead to higher concentrations of toxic amines and sulfur compounds that signal to scavengers that the item is inedible. Lower temperatures are less favorable for microbial activity and accordingly there is little or no production of chemical odors. Scavenging may thus be limited at low temperatures because, without the chemical cues, scavengers may have difficulty finding a potentially edible carcass.

Who Scavenges?

Carrion tends to be an unreliable resource in any one location or point in time. This has hindered the evolution of strict or obligate scavenging behavior, except in a few notable species like vultures or certain flying insects. Vultures in particular have several traits that facilitate specialization as scavengers. First, they have large broad wings that enable them to expend minimal energy by soaring

 doi:10.1016/B978-0-444-63768-0.00544-8

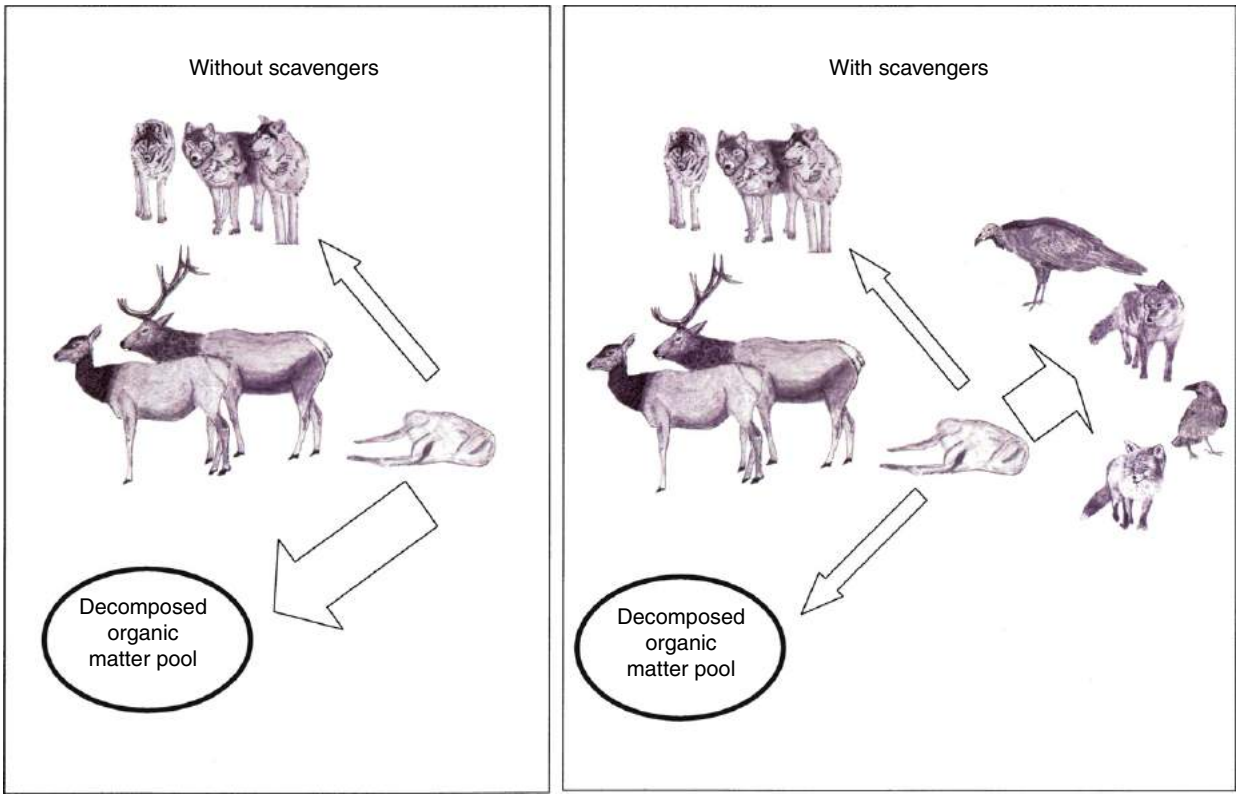


Fig. 1 In many systems, including the example of wolves preying on elk, predation accounts for 2%–75% of losses of individual prey annually (denoted by thin arrow from the elk carcass to wolves). The remaining 25%–98% of all individuals that die annual succumb to nonpredation causes. Much of this abundant carrion can be broken down into constituent chemical elements by decomposers (denoted by thick arrow from elk carcass to the decomposed organic matter pool) if scavengers are absent. In many cases, however, it is redirected to bolster populations of a diversity of scavenger species, including turkey vultures, coyotes, ravens, and foxes (denoted by thick arrow from elk carcass to scavengers and thin arrows to decomposed organic pool). Hence, scavengers compete with decomposers for carrion. Scavengers, however, are often more efficient than decomposers at despatching carrion.

over vast areas to locate carrion. They have sharp eyesight and a keen sense of smell. They also are able to consume carcasses very rapidly once they have been discovered. Even so, they represent a very small faction of the range of species that scavenge.

Most scavenging is facultative; in essence a dietary supplement. It is undertaken by a broad range of species. Most notably carnivores of all stripes, including the majestic birds of prey such as eagles, hawks and falcons but also other birds such as ravens and magpies; canid, felid, ursid, and hyenid mammal predators; snakes, lizards, and spiders all consume fresh carrion when it is found. After all, it does not make evolutionary sense (in terms of improving individual survival and reproduction) to pass up a free meal, one that effectively differs little from a hunted prey item, whenever it is encountered. The proportion of the diet that comes from carrion, however, can vary widely among carnivore species, making some species like hyenas and ravens seem close to being obligate in their scavenging.

Many seemingly unlikely species such as herbivores also scavenge. For example, on islands in Lake Michigan, white-tailed deer consume large quantities of dead alewives, herring-like freshwater fish that undergo annual mass die offs and wash onto shore in spring. It has been estimated that alewives comprise 30%–54% of the daily diet of individual deer during the spring period. Such purposeful scavenging appears to provide the deer with an important dietary supplement during a period when terrestrial food resources are in critically low supply after lengthy winter browsing. Alewives have higher protein, fat, energy, and mineral (especially salt) contents and are more easily digested than the heavily browsed plants on the islands. Other herbivores including grasshoppers and hippopotamus also readily engage in scavenging.

Ecological Effects of Scavengers

Behavior Modification of Carnivores

Scavengers do not actively hunt and kill prey. Instead, they must seek out carrion across broad distances on landscapes. But carrion is highly ephemeral in space and time and so it can be quite difficult to find it unless one can search wide distances quickly and

efficiently. Most scavengers do not have this searching ability. So they beat the odds against finding carcasses by associating themselves with species that actively hunt and kill prey.

A classic example of such association is between wolves and ravens. Ravens are typically present at wolf-killed carcasses and in some locations such as on Isle Royale in Lake Superior they are omnipresent. There are even cases in which ravens are rarely found on the landscape except in close association with wolves. Ravens can derive a very good livelihood from scavenging carcasses. An individual can ingest and hoard between 0.5 and 2 kg of wolf-killed prey per day. Thus, wolves may routinely lose between 2 and 20 kg of food per day to flocks of ravens. There are notable cases in which flocks of ravens devour up to half of the moose carcasses.

Such a high level of scavenging imposes strong competitive pressure on wolves to the extent that it may alter wolf grouping dynamics. A classically held belief is that groups of wolves comprise related kin in which altruistic behavior of the kin contribute toward overall family welfare (survival and reproduction). But recent research shows that wolf packs contain unrelated individuals. Moreover, pack sizes are often larger than one would expect if individual wolves were attempting to maximize their foraging returns. Such behavior is not expected to be favored by natural selection. This counterintuitive behavior can, however, be reconciled if we add in the costs of food loss to scavengers. In the absence of scavenging, wolves maximize their foraging returns by associating in groups of two or three individuals because one individual alone is inefficient at killing prey and beyond two or three individuals competition for access to a carcass increases with group size leading to diminishing per individual foraging return rate. Loss of food to scavengers may change this structure because it forces wolves to hunt more frequently. Larger packs tend to be more efficient at killing prey frequently. Also, individual foraging return varies little with group size under conditions of scavenging and frequent hunting. Thus, the foraging cost of living in large groups may be offset by the benefit of frequent prey capture in wolves and perhaps in other social carnivores like lions that also face competition with scavengers.

Behavior Modification of Herbivores

Because scavengers associate closely with predators, their vocalizations and movement behavior may signal imminent predation risk to prey. Moose in boreal forest ecosystems that face high predation risk have been shown to respond dramatically to this signal. In boreal forests, ravens associate closely with wolves, especially during wolf hunting forays. In these regions, the probability of survival, especially of young individuals, can often be as low as 30%. Research using playback calls of ravens has shown that moose in such high risk areas decrease their foraging rates and become increasingly vigilant by being watchful of imminent danger. This contrasts sharply with a lack of a behavioral response to playback of raven calls in geographic locations where wolves and other predators of moose have long been extirpated. In such areas, survival probability is at least three times higher than in the high risk areas. Differences in foraging rates between high and low risk areas are known to have differential effects on ecosystems because they lead to differences in the abundance of plant species that comprise the herbivore's forage.

Carnivore–Scavenger Interactions

Because carnivores hunt year-round, they often provide a steady supply of carrion. The exact supply rate of such a resource is known to change the seasonal behavior of scavenger species as well as be an important determinant of the spatial composition of scavenging species within landscapes.

Grizzly bears are important scavengers throughout most of their geographic range. In most cases, however, they hibernate during the winter months as a means to survive periods of chronic food shortages. However, Grizzly bears are known to forego hibernation in conditions when the supply of carrion is high. This may often happen in winters with high snow depth because moose and elk species that comprise the prey base for wolves are encumbered by deep snow and thus are especially vulnerable to being captured. Under such conditions, wolves frequently abandon partially eaten carcasses in favor of capturing new prey, leaving a continuous and plentiful supply of left-over meat, bone and hide to be scavenged.

Scavenging is undertaken by many generalist species that opportunistically use carrion when it is available while sustaining themselves on other resources when carrion is unavailable. These species do not live in isolation of one another on landscapes. So the availability of carrion within the landscape can lead to strong interactions among species as they vie for their share of the resource. Moreover, the nature of carrion supply in space and time can have an important bearing on the kinds of scavenger species found within a location.

If the amount of carrion provided by carnivores is small and much localized, then this resource will attract scavenger species with small foraging radii – those species that forage largely within a small local area. This highly limited resource will be most likely consumed by species that are competitively dominant. These are typically the more fearsome species like coyotes or hyenas that are able to usurp the food by scaring away other species. If the local supply of carrion is large, then it will saturate the ability of the local, competitively dominant scavengers to consume the carrion in its entirety. In such cases, wandering species – those with large foraging radii – will also be attracted to the resource leading to a high diversity of scavenger species at a carcass. The plentiful supply of the resource also diminishes the intensity of competitive interactions among the scavenger species. Because many of these scavengers are also generalist carnivores, such a high, local resource supply represents an important survival subsidy that maintains the multiple trophic level structure of ecological food chains and webs. Predator species that temporarily resorted to scavenging can resume their normal carnivore role once the pulse of carrion supply subsides.

In addition, carnivores, by adding to natural mortality of prey, can add to the spatial and temporal supply of carrion. In the absence of carnivores, herbivore species often die in high numbers during parts of the year in which resources are in short supply or poor in quality such as drought periods in savanna grasslands or late winter in northern temperate regions. Scavengers take advantage of these short pulses of resources to sustain their populations. Nevertheless, their population dynamics are influenced by the vagaries of this carrion supply because it can fluctuate widely with weather conditions from year to year. Large hunting carnivores can change the temporal dynamics of carrion supply from a short seasonal pulse to one that is more even and protracted throughout the year. This subsidy in turn can help to stabilize the long-term population dynamics of carnivore species that scavenge opportunistically, leading to a higher diversity of species on the landscape.

Energetic subsidies in the form of carrion can also undergird food chain structure in locations where long food chains are unlikely to be sustained by local levels of resource production. Arid oceanic island ecosystems off Baja Mexico normally provide an inhospitable environment: they are covered with *Opuntia* cactus and myriad species of flying insects and their web-building spider predators. Curiously, however, the islands support extraordinarily high densities of spider predators. This occurs because a considerable abundance of nutrient-rich resources in the form of drowned animal carcasses washes up onto the shore from oceanic drift. This resource input sustains insect species that scavenge the decomposing carcasses, thereby creating a highly abundant resource for carnivore species, especially on islands where there is little plant production and hence limited production of herbivore prey. The marine-island food energy conduit thus bolsters the structure of the island food web. In turn, the abnormally high abundance of spiders led to an unusually high capacity to control the abundance of the island's herbivorous insects, thereby lessening the insect damage to plants. Thus, the effects of the subsidy, mediated by scavenging, reverberate through the whole island system. Shut off the supply of carrion and the island ecosystem could collapse to a comparatively barren desert.

Summary

The view that scavengers are repugnant, behavioral oddities is an unfortunate and inaccurate representation of an ecological role. Scavenging provides an important means to bolster the structure and functioning of ecological systems because it mediates the ebb and flow of a major resource in space and time. It seems evident that in its absence, many ecological systems would have considerably lower species diversity because of reduced productivity and longevity of the myriad species that avail themselves of carrion when it is in supply. In addition, scavenging may serve an important and integral role in the functioning of ecosystems in that it provides a source of energy to top carnivores that that can rival or exceed in magnitude levels of energy supply provided by the supply chain from plants through herbivores to carnivores.

Further Reading

Berger, J., 1999. Anthropogenic extinction of top carnivores and interspecific animal behaviour: Implications of the rapid decoupling of a web involving wolves, bears, moose and ravens. Proceedings of the Royal Society of London Series B – Biological Science 266, 2261–2267.

Case, D.J., McCullogh, D.R., 1987. White-tailed deer forage on Alewives. Journal of Mammalogy 68, 195–197.

DeVault, T.L., Rhodes, O.E., Shivik, J.A., 2003. Scavenging by vertebrates: Behavioral, ecological, and evolutionary perspectives on an important energy transfer pathway in terrestrial ecosystems. Oikos 102, 225–234.

Polis, G.A., Hurd, S.D., 1995. Extraordinarily high spider densities on islands: Flow of energy form the marine to terrestrial food webs and the absence of predation. Proceedings of the National Academy of Sciences of the United States of America 92, 4382–4386.

Schmitz, O.J., Krivan, V., Ovadia, O., 2004. Trophic cascades: The primacy of trait-mediated indirect interactions. Ecology Letters 7, 153–163.

Vucetich, J.A., Peterson, R.O., Waite, T.A., 2004. Raven scavenging favours group foraging in wolves. Animal Behaviour 67, 1117–1126.

Wilmers, C.C., Crabtree, R.L., Smith, D.W., Murphy, K.M., Getz, W.M., 2003. Trophic facilitation by introduced top-predators: Grey wolf subsidies to scavengers in Yellowstone National Park. Journal of Animal Ecology 72, 909–916.

Wilmers, C.C., Stahler, D.R., Crabtree, R.L., Smith, D.W., Getz, W.M., 2003. Resource dispersion and consumer dominance: Scavenging at wolf- and hunter-killed carcasses in Greater Yellowstone, USA. Ecology Letters 6, 996–1003.

Seasonality

GH Dayton, Moss Landing Marine Laboratories, Moss Landing, CA, USA

Introduction

Seasonal shifts in environmental conditions play a fundamental role in influencing the abundance and distribution of organisms throughout space and time. These seasonal factors ultimately influence key biological parameters of organisms (Table 1). Changes in climatic conditions directly impact primary producers, which in turn influence secondary consumers throughout the food web. Fluctuations of vegetative growth between the growing and nongrowing season result in patchy concentrations of essential resources such as food, water, and energy. Thus, energetic resources for primary and secondary consumers are largely driven by changes in seasonality. These shifts in resources have direct and indirect effects on the ecology of all organisms. Beyond energetic constraints, seasonal variation in climatic conditions influence when and where particular organisms can persist. As a result, many species exhibit highly adaptive traits and life histories that enable them to persist in changing environments. This is especially true in higher latitudes where climatic conditions change more drastically throughout the year in comparison with low latitudes where climatic fluctuations are less variable. In this article, the author outlines some of the seasonal factors that affect organisms, as well as highlights examples of how these factors drive seasonal shifts in the abundance and distribution of various species. It is important to note that many, if not all, of the seasonal factors discussed below are correlated with one another and in many cases it is difficult to disentangle the factors individually. For example, seasonal shifts in photoperiod and temperature (tightly correlated with one another) play important roles in the ecology of many organisms; however, these factors are determined by the tilt of the axis of rotation of the Earth.

Rainfall

Seasonal shifts in rainfall have a pronounced impact on aquatic organisms; this is particularly true in ephemeral water bodies. The highest diversity of aquatic organisms occurs in temporary water bodies that are seasonal in nature – filling after cyclic rains and drying during nonrainy periods. Species that are dependent upon these habitats exhibit life history strategies that enable them to exploit seasonal aquatic habitats as well as to persist during dry periods. Many of these organisms have stages in their life cycles in which they enter diapause (a period of quiescence characterized by the cessation of growth and reduction of metabolic activity) or remain dormant during the dry season. For example, several species of aquatic invertebrates lay eggs in temporary pools, after which the eggs settle into sediments where they remain dormant during the dry period until seasonal rains once again flood the site and the aquatic larvae life stages emerge. Eggs of some aquatic invertebrates can remain dormant for over 125 years. Plant species associated with vernal pools follow a predictable flowering phenology after seasonal rains; these species are primarily annual and reproduce as pools draw down during the spring and summer months. Spring wildflowers in vernal pool ecosystems provide important resources for pollinating insects, some of which only collect pollen from vernal pool plant species.

Seasonal rains also play a vital role in sustaining organisms that live in permanent water bodies. Most large riverine systems throughout the world experience annual floods during wet seasons. These floods expand the spatial extent of rivers into habitats that for most of the year remain dry. Water spills over the riverbanks, flooding forest habitats and connecting water bodies that are usually isolated from one another. As a result, aquatic organisms are able to exploit habitats that are inaccessible throughout much of the year. As swelling rivers move into upland habitat, nutrients leach into the water, increasing primary production. Aquatic plants assimilate these nutrients that are eventually recycled back into the environment via decomposition. This pulse of nutrients plays a significant role in supporting the base of the food web and in turn sustains an increased number of herbivores and predators. Seasonal floods are common in large tropical riverine systems, which support a diverse assemblage of fish. One of the factors thought to have led to the diversity of fishes in these systems is the accessibility to a wide breadth of feeding niches that are available during seasonal flood events. Fish are able to exploit inundated upland habitats and thus are able to gain access to a wide

Table 1 Organismal responses to seasonally varying factors

Seasonal forcing factor	*Organismal response*
Ice	Dormancy, recruitment
Rainfall	Growth, reproduction, migration, germination
Photoperiod	Hibernation, reproduction, migration, diapause, dormancy, food caching, molting, dormancy, recruitment, growth
Storms	Dispersal, reproduction, growth
Temperature	Germination, migration, reproduction, hibernation, recruitment, growth, dormancy
Wind	Dispersal, reproduction

 doi:10.1016/B978-0-444-63768-0.00545-X

variety of resources. As a result, fish assemblages in large tropical rivers exhibit a wide variety of feeding strategies and have a disproportionately greater number of herbivorous, detritivorous, and omnivorous feeding behaviors.

Desert regions, by definition, have very little rainfall throughout the year, with most of the precipitation occurring in seasonal storms during a 2–3 month period. Similar to environments that experience freezing temperatures during winter months, organisms that inhabit deserts are adapted to exploit patchily distributed resources throughout the year. The tight correlation between rainfall and seed production in desert regions has played an adaptive role in selecting for life history strategies that favor individuals that breed during wet periods of the year when resources are abundant. Many species of granivorous rodents, ants, and birds show predictable fluctuations in abundance throughout the year; peaks are associated during the winter and/or summer rain periods when seed production is high (**Fig. 1**). Organisms that inhabit desert environments throughout the year have adaptations that enable them to persist during times of low resource availability and extreme temperatures. Several desert species exhibit behavioral modifications that enable them to persist through drastic seasonal shifts in the environment. Amphibians, for example, bury themselves underground or take refuge beneath cover during dry months. These behaviors prevent desiccation and reduce their energetic requirements during times of low resource availability. Other organisms, such as the kangaroo rat which has highly specialized kidneys that are extremely efficient at conserving water, have physiological modifications that enable them to persist during times of low resource availability.

Oceanic Upwelling

Seasonal and periodic upwelling of deep nutrient-rich water to shallow depths plays a major role in supporting marine organisms worldwide. Upwelling occurs when cold nutrient-rich deep water replaces warmer nutrient-poor surface water. Increases in nutrients promote growth of marine algae and phytoplankton, which in turn provide the basis of the food web for many fish, marine mammals, and marine birds. These shifts in resources influence the abundance of marine organisms throughout the world. Upwelling events provide animals such as humpback whales, shearwaters (a largely pelagic bird), many tuna, and other primary and secondary consumers, with a significant portion of their annual energy requirements.

Photoperiod

Photoperiod (day length) plays an important role in regulating the timing of migration as changes in photoperiod influence temperature, rainfall, and ultimately primary productivity. Most plants respond to changes in photoperiod by producing seeds, new growth, and/or fruit at specific times of the year. A common animal response to shifts in vegetative production is migration. A

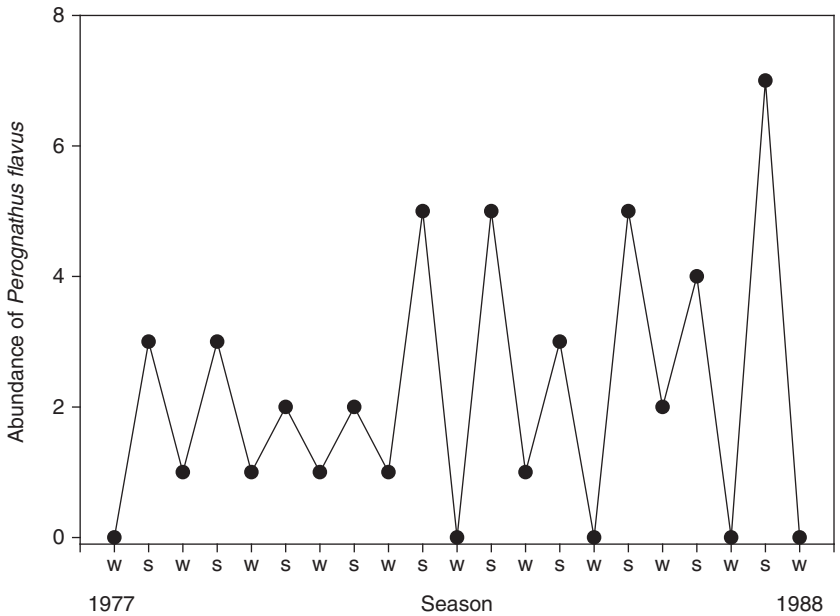


Fig. 1 Changes in abundance of the Silky Pocket Mouse (*Perognathus flavus*) over a 10 year period in the Chihuahuan Desert (w = winter; s = summer). Abundance was calculated as the 6 month average. Seasonal peaks in abundance are correlated with summer precipitation events and are likely largely due to increased primary productivity associated with rain events. Adapted from Brown JH and Henske EJ (1990) Temporal changes in a Chihuahuan Desert rodent community. *Oikos:* 59: 290–302.

classic example of a migrating species is the Serengeti wildebeest which migrates from areas of low resource production to areas of high resource production on an annual basis. Wildebeests occur throughout the Serengeti ecosystem which extends along the border region of Tanzania and Kenya. Forty percent of the Serengeti is comprised of grassland habitat which is the primary food source for wildebeests. Seasonality in rainfall throughout this region results in a dichotomous climate in which most of the rainfall occurs in the wet season from January–May, with very little rainfall occurring during the dry season (June–August). Grass production is high during the wet periods and virtually nonexistent during much of the dry season. Throughout the year, wildebeest will move from areas of low grassland productivity to areas of high grassland productivity.

Migratory species often modify the habitat in the area they migrate to, themselves having a seasonal impact on the environment. For example, many waterfowl congregate at breeding grounds where they reproduce and feed prior to migrating back to their winter habitat. It is not uncommon for breeding grounds to have tens of thousands of birds concentrated in relatively small areas for 2–3 months at a time. During these events, foraging birds have direct impacts on the plants they feed upon. Grubbing (digging up plant roots) by large numbers of waterfowl can have detrimental consequences on the environment as these areas are stripped of vegetation and plant community composition is significantly altered. However, for some plant species, heavy grazing (foraging on aboveground plant material) actually increases the overall net aboveground primary production, and plays a critical role in maintaining aboveground biomass and species composition of the vegetation. This occurs because although birds are eating plants, they are not killing them; and while foraging they are defecating on the ground, essentially adding fertilizer for plant growth. In wintering grounds, migratory waterfowl primarily roost in wetlands and spend most of the day foraging in upland habitats. Daily foraging forays to uplands followed by returns to roosting sites creates an agent for nutrient transport between habitats, which in turn can influence local ecosystem processes. The transfer of nutrients into wetlands have positive impacts on many aquatic plant and animal species; however, large numbers of migratory waterfowl roosting in small wetlands can result in high levels of nutrient loading which can be detrimental to water quality and ultimately the aquatic ecosystem.

Seasonal shifts in primary productivity driven by changes in photoperiod and/or rainfall have consequently played an important role in selecting for behavior of many species such as wildebeests and waterfowl, to move between hospitable habitats on an annual basis. This in turn influences the habitat in the areas they migrate to. Furthermore, top predators, such as lions and foxes, gain a substantial amount of their annual energy budget from migratory animals as they migrate through, or into, their ranges. Thus, these migratory behaviors in turn have direct and indirect impacts on the environment, which can result in large ecological effects that ripple throughout the food web.

Temperature

While many animals exhibit life history strategies that take advantage of patchy resources (both spatial and temporal) by moving between areas where resources are available throughout the year, other species have life histories that enable them to remain dormant during periods of low resource availability. Species that inhabit extreme environments, such as deserts and regions that experience frequent freezing temperatures, have evolved mechanisms to deal with the consequences of seasonal variation in environmental conditions. Nonmigratory species that inhabit areas remaining under snow for periods of the year have to cope with seasonal changes in resources by significantly reducing activity or by lowering their metabolic rate during cold periods. Reduction of metabolic processes saves energy during times when food resources are not available.

Several species cope with seasonal shifts in resources by caching food during times when resources are plentiful in order to survive periods when food is scarce or nonexistent. Many members of the avian family *Corvidae* harvest seeds throughout the spring, summer, and fall months and cache them in various locations in order to have a food source during the winter and early spring when seeds are not being produced. A classic example of such a species is Clark's nutcracker, which is known to cache over 30 000 seeds at greater than 7000 individual cache sites over a spatial extent of greater than 20 linear km. The Clark's nutcracker will bury seeds beneath soil and plant material throughout the summer and fall. During the winter and early spring, nutcrackers will return to their cache sites by utilizing landmarks and will then unbury their seeds. Clearly, seasonality has played a large role in the selective pressures that have led to the adaptations that facilitate behavior and cognitive abilities of the Clark's nutcracker. Indeed, when a comparative approach is taken, researchers have shown that Corvid species that are less reliant upon stored food for survival are not as diligent at caching food nor are they as good at relocating caches compared to Corvid species that inhabit harsher environments and thus largely rely upon cached seeds for survival.

Summary

Seasonality plays a critical role in influencing the persistence of all living organisms. Seasonal shifts in climatic conditions influence the availability of resources, which in turn influences the presence or absence of species throughout the environment at both the temporal and spatial scale. Over time, natural selection has favored individuals that display behaviors, phenotypes, and physiological adaptations that enable them to maximize seasonally patchy resources and cope with extreme environmental conditions. Individuals that persist in seasonal environments are better suited to adapt to shifts in environmental conditions, and

thus are able to exploit a vacant niche. When examining factors that effect the distribution and abundance of organisms, it is imperative to consider how seasonality has, and continues to, influence species persistence.

Further Reading

Aidley, D.J., 1981. Animal migration. New York, NY: Cambridge University Press.

Bakum, A., 1990. Coastal ocean upwelling. Science 4439, 198–201.

Balda, R., Kamil, A., 2006. The ecology and life history of seed caching corvids. In: Brown, M.F., Cook, R.G. (Eds.), Animal Spatial Cognition: Comparative, Neural, and Computational Approaches. Available online. http://www.pigeon.psy.tufts.edu/asc/balda/(accessed on 10 October 2007).

Boyce, M.S., 1979. Seasonality and patterns of natural selection for life histories. American Naturalist 114, 569–583.

Brown, J.H., Reichman, O.J., Davidson, D.W., 1979. Granivory in desert ecosystems. Annual Review of Ecology and Systematics 10, 201–227.

Brown, J.H., Heske, E.J., 1990. Temporal changes in a Chihuahuan Desert rodent community. Oikos 59, 290–302.

Cáceres, C.E., 1998. Interspecific variation in the abundance, production, and emergence of *Daphnia* diapausing eggs. Ecology 79, 1699–1710.

Colburn, E.A., 2004. Vernal pools: Natural History and Conservation. Blacksburg, VA: The McDonald & Woodward Publishing Company.

Kerbes, R.H., Kotanen, P.M., Jefferies, R.L., 1990. Destruction of wetland habitats by Lesser Sow Geese: A keystone species on the west coast of Hudson Bay. The Journal of Applied Ecology 27, 242–258.

Kitchell, J.F., Schindler, D.E., Herwig, B.R., Post, D.M., Olson, M.H., 1999. Nutrient cycling at the landscape level: The role of diel foraging migrations by geese at the Bosque del Apache National Wildlife Refuge, New Mexico. Limnology and Oceanography 44, 828–836.

Knapp, R.A., Matthews, K.R., Orlando, S., 2001. Resistence and resilience of alpine lake fauna to fish introductions. Ecological Monographs 71, 401–421.

Wilmshurst, J.F., Fryxell, J.M., Farm, B.P., Sinclair, A.R.E., Henschel, C.P., 1999. Spatial distribution of Serengeti wildebeest in relation to resources. Canadian Journal of Zoology 77, 1223–1232.

Winemiller, K.O., Jepsen, D.B., 1998. Effects of seasonality and fish movement on tropical river food webs. Journal of Fish Biology 53, 267–296.

Seed Dispersal☆

Anna Traveset, Institut Mediterrani d'Estudis Avançats (CSIC-UIB), Mallorca, Balearic Islands, Spain
Javier Rodríguez-Pérez, University of Évora, Évora, Portugal; Aranzadi—Society of Sciences, Donostia/San Sebastian, Spain

Glossary

Dispersal effectiveness The contribution that each frugivore/seed disperser species within a community makes to plant fitness; in other words, the number of new adult plants produced by the disperser's activity relative to the rest of dispersers in the community.

Dispersal kernel Function summarizing the frequency of seed density after arrival from dispersal sources. The "dispersal kernel" is a population-based characteristic defined by a specific shape of an unimodal leptokurtic statistical function with a peak at or close to the origin, followed by a rapid decline and a long, more or less fat, tail.

Dispersal syndrome Particular combination of fruit and seed traits that are associated to a dispersal vector (e.g., water, wind, animal). The morphological characteristics of a dispersal syndrome are shared by phylogenetically unrelated species and have emerged as a consequence of parallel evolution forces.

Janzen-Connell hypothesis It proposes a trade-off between seed dispersal distance, which is maximum around mother plant, and the pressure of herbivores, pathogens and natural enemies which are maximum at high seed densities. This hypothesis predicts that the maximum probability of seed and seedling survival occurs at medium distances from origin. Despite it was developed to explain the high diversity in tropical forests, it has been demonstrated in other ecosystems as well.

Secondary seed dispersal Multi-step process of seed dispersal, that is, involving at least one more stage of seed dispersal after a primary seed dispersal. It may include either biotic or abiotic vectors not directly related to primary seed dispersal process.

Seed rain The spatial distribution of dispersed seeds within a landscape as a consequence of multiple dispersal events, sources and agents. Seed rain provides important population-based processes related to the density and richness of seeds.

Seed shadow The spatial location of seeds around a source point after being dispersed. The spatial location of each seed will determine the location where it will be established within the landscape.

Background

The ecology of seed dispersal is a topic of much interest to naturalists, although it has not been until the last three decades that has received considerable attention by scientists. Seed dispersal is one of the key phases in the process of plant regeneration, as it determines the potential area of recruitment at the same time that acts as a template for the rest of phases in such process. Dispersal can be defined as the process by which individuals move from the immediate environment of their parents to establish in an area more or less distant from them. In contrast to animal dispersal, plant dispersal is always passive in the sense that seeds have no control of where they will end up; moreover, seed dispersal is more determined by the traits of the parents than by the traits of the seeds themselves. Two widely used terms in the study of seed dispersal "seed shadow" and "seed rain" represent individual and population-based perspectives, respectively, which are needed if we are to understand the shaping of the spatial seed distribution and, ultimately, the evolutionary and demographic processes affecting plant recruitment.

☆*Change History*: October 2017. Javier Rodríguez-Pérez extended sections, included glossary and updated Abstract, included new references and added Figures 2, 4 and 5. A Traveset revised text.

Traits of Propagules and Dispersal Vectors

Seeds are dispersed in a great variety of ways. The morphological devices that enhance dispersal are usually quite evident and interpretable. Thus, for instance, we find wind-borne diaspores bearing wings, hairs or plumes that increase air resistance and slow the rate of fall (a dispersal syndrome named anemochory), seeds that float in the water by means of a buoy (hydrochory), seeds with hooks or barbs that adhere to the exteriors of animal vectors (exozoochory), seeds with elaiosomes for ant dispersal (myrmechochory), or diaspores with flesh appendages or coverings that are consumed by animals which later eject the seeds (i.e., endozoochory). Some plants disperse their offspring ballistically, by the explosive opening of the fruits or the springing of a trip-lever. Other species lack any evident dispersal device, which makes us wonder whether dispersal is less advantageous in these species or how they achieve effective dispersal.

The dispersal mode of seeds has commonly been associated with seed size; thus, for instance, species dispersed ballistically have significantly larger seeds than those dispersed by anemochory. This occurs even within a genus, such as in *Pinus*, in which seeds weighing less than c. 100 mg are wind dispersed whereas heavier seeds tend to have adaptations for either ballistic or animal dispersal. The relationship between particular seed traits and vectors has been called "dispersal syndrome." However, the usefulness of dispersal syndromes has often been questioned, especially for vertebrate seed dispersal, as fruit traits often are more influenced by plant phylogeny than by vertebrates. In fact, after accounting for phylogeny, fruit size seems to be the only trait—out of a large number of fruit traits considered in a review—significantly associated to dispersal (Fig. 1).

Seeds may have more than one opportunity of being dispersed. After a first phase consisting on the initial movement of seeds away from mother plant (primary dispersal), there may be a second phase (secondary dispersal) in which seeds are further dispersed, usually by another mechanism or agent. This is common in plants which are first dispersed by endozoochory or ballistically and are subsequently moved by ants, dung beetles, rodents, birds or even predators of such frugivores that carry seeds in their digestive tracts. There are cases where seeds have reward structures to attract ants, such as elaiosomes or flesh appendages which are removed once the seeds are dispersed. In other species, however, primarily dispersed seeds are removed by seed predators (e.g., rodents, granivorous ants) that store a fraction of them in sites—like holes in trees or in soil, ant nests, etc.—, and that later "forget" to collect them.

Plant and Animal Adaptations

The dispersal mode of any plant species is the result of many different pressures and constraints. Phylogenetic constraints are responsible for the fact that entire families or genera usually exhibit only slight variations on a single mode of dispersal. Nonetheless, large variation in some families or genera evidences that these constraints are not universal but contingent on the selective pressures and environmental variability which constrains the seed dispersal process.

From the plant's viewpoint, the timing of fruit maturation is key to enhance the seed dispersal process. A few general patterns in dispersal phenology have been described. Wind-dispersed neotropical trees, for instance, mature their fruits during the dry season, when trade winds are strong and trees are leafless, contrasting with the more or less constant throughout the year production of fleshy or dry fruits. By contrast, in the north temperate zones, mature fruits are produced in late summer and autumn, when avian frugivores are usually abundant, whereas further south, ripe fruits are also found through the winter, when flocks of wintering migrant birds are foraging. Nevertheless, such fruiting patterns do not need to be interpreted as adaptations to dispersal, as constraints to such timing may derive from selection to avoid pathogens or predators, to shift the flowering time or to modify the length required for fruit maturation. It is widely accepted that seasonality in temperature and water availability set limits on the time of fruit and seed development and maturation.

Fruit consumption by frugivores, and in particular vertebrates, has selected for fruit traits that enhance their detectability such as advertisement and/or visual chemical signals. Fruits, thus, tend to have a conspicuous coloration, distinctive odor or a combination of both. A common pattern found both in the tropics and in the temperate zones is that bird-dispersed plants usually have red or black-colored fruits. In some species, a bicolored fruit advertisement, contrasting the ripe fruits with the surrounding foliage is what presumably gives visual conspicuousness (what has been termed the "foliar flag" hypothesis). Also, some ripe fruits reflect ultraviolet light which enhances the detectability by birds, as their color vision extends to the near UV. The fruits dispersed by vertebrates also tend to have a pulp rich in water and carbohydrates whilst are poor in protein and lipids; however, there is much interspecific variability in nutrient composition, and fruit pulp quality does not show to be a trait reflecting plants' adaptations to dispersers. Fruit pulp usually contains also secondary metabolites (phenolics, alkaloids, etc.), sometimes to the point of being lethal to animals, which require an adaptive explanation not yet found. One possibility is that such compounds serve as defense against microbial pathogens and invertebrate pests that preclude the consumption of the fruits by legitimate dispersers.

Regarding animal adaptations, the majority of frugivores do not require important morphological and physiological adaptations, especially those that feed occasionally on fruits; in temperate climates, fruit resources are seasonally available and thus frugivores need to cope with periods of fruit scarcity or high fruit abundance. In the tropics, by contrast, there are frugivore species highly adapted to fruit consumption due to high availability of such resource there. Frugivores there can have the following distinctive traits: birds tend to have shorter, broader and flatter bills, and wider gapes than those not consuming fruits; some birds also have smaller and less muscular gizzards, larger livers and shorter intestines; frugivorous bats have shorter canines and broader palates than insectivorous ones; frugivorous lizards have longer intestines than those consuming mostly animal material. This is probably the reason why there are not frugivores specializing on only one or a few plant species.

Fig. 1 Examples of seed dispersal syndromes (a) European ash (*Fraxinus excelsior*), wind dispersal in Poland (Author: Pleple 2000; License: Creative Commons); (b) Coconut (*Cocos nucifera*), Sea dispersal in the shore of India (Author: Dheeraj Madala; License: Creative Commons); (c) Tree spurge (*Euphorbia dendroides*), ballistic and ant dispersal in Israel (Author: Gideon Pisanty; License: Creative Commons); (d) Red squirrel (*Sciurus vulgaris*) removing accorns (*Quercus robur*) in Estonia (Author: Hannu; License: Creative Commons); (e) Lilford's wall lizard (*Podarcis lilfordi*) consuming Joint pine (*Ephedra fragilis*) fruits in Balearic Islands (Author: Javier Rodríguez-Pérez, use under permision); (f) Elephant apple (*Dillenia indica*), seed dispersal by mammals in India (Author: Shijan Kaakkara; License: Creative Commons).

Evolutionary Advantages

From the broad range of dispersal strategies, evolutionary or selective forces could shape the rich diversity of strategies found in nature. In general, seed dispersal strategies (or the selective forces that shape them) tend to increase seed dispersal rates from source point, providing individuals and/or species more capable to generate offspring. Hence, the two major benefits of seed dispersal are: (1) departure from the mother plant, which usually avoids sibling competition and reduces offspring mortality by predators or pathogens; and (2) colonization of new sites. Seed density usually is maximum at short distances from the mother plant (Fig. 2). Deviations from this conventional seed shadow shape can result from patchiness of habitat structure or from other ecological factors such as the frugivores' behavior, which can promote nucleation process (for instance, by depositing seeds under particular trees used for resting).

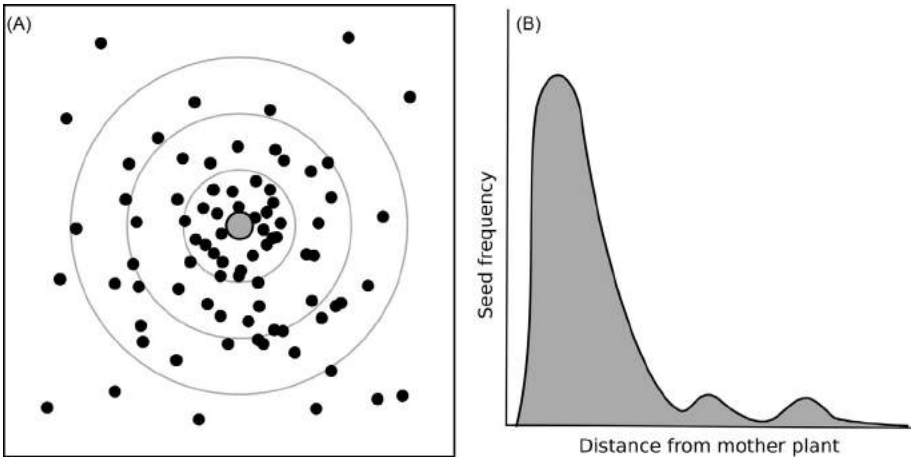


Fig. 2 Example of seed shadow and relationship between dispersal distance and probability of seed dispersal. In the left panel, *black* points represent the individual location of dispersed seed at increasing distances from the origin (mother plant in *gray*). In the right panel dispersal kernel of seed frequency at increasing distances from origin.

Frequently, the benefits of leaving the immediate vicinity of the mother plant depend upon the advantages obtained by (1) increasing the distance to it and (2) avoiding a highly intense sibling competition; the effects of both factors are not easy discernible without field experimentation. Regarding the advantage of colonizing new sites, seeds that leave mother plants have the capacity of occupying vacant habitats and suitable microhabitats for either germination and growth. This allows, for instance, the latitudinal or altitudinal migration of many plants in front of climate changes, the recolonization of a land after a volcanic eruption, the colonization of rapid-growth species of an abandoned field, enhancing thus the ecological succession, etc. There are also species that have "directed seed dispersal," benefiting from it as seeds are deposited in sites or microsites that are especially suitable for germination and seedling establishment. Seed mistletoes, for instance, are usually defecated by birds on host twigs, which are required for germination and seedling recruitment. From the evolutionary point of view, processes against dispersal could additionally act as selective forces upon propagule traits, at the expenses to loss propagule viability (i.e., seed germination and seedling growth). Under latter circumstances, the mechanisms against seed dispersal may occur, especially in plants inhabiting extreme habitats (i.e., cliffs or mountains) with very small distribution ranges, and that usually have promotes evolutionary processes of local adaptation and low genetic variability in plant populations.

Patterns of Dispersal from Entire Plant

Once released from mother plant, the movement of seeds may be governed by the laws of the probability, with different chances to stay close or far away from origin (mother plant). As discussed in previous sections, seed traits could determine, at first glance, the origin and end points of each dispersal event. However, seed traits do not necessarily guarantee such dispersal properties but depend on stochastic processes affecting the spatial and temporal dynamics of each locality. For wind-dispersed species, for instance, the processes that influence dispersal distances are either atmospheric (the spatial and temporal statistics of the wind velocity field—vertical, longitudinal, and latitudinal—their covariance structure and their integral time scale properties) or biological (terminal velocity of the dispersal unit, release height, and timing of release) factors. In the case of seed dispersal by animals, seed shadows could be additionally modified by environment or the attraction to animals to suitable habitats. For such reason, the magnitude and direction of each dispersal event is random and we thus need to approach statistical functions to better predict the probability that each dispersal event arrives to a given end. Borrowing concepts and methodologies from mathematics, the statistical distribution of dispersal distances in a plant population is called the "dispersal kernel" (Fig. 2). Thus, the shape of the dispersal kernel impacts on many plant population processes at local or population scales, such as the recruitment patterns, genetic structure and community diversity processes.

Despite the majority of seed dispersal events occur in the vicinity of mother plants, recent population and community studies show that the low-likely events associated to extreme dispersal distances are critical for rates of range expansion, speciation and, genetic connectivity between distant populations. Such long-dispersal events are extremely difficult to capture in nature and, thus, the use of simulation models based on assuming that a complex system can be subdivided into discrete processes (i.e., abiotic and biotic) are a promising tool because it provides reliable predictions of standard (local dispersal) as well as nonstandard (long-distance dispersal LDD) dispersal events. In the case of wind-dispersed species, knowledge of the average wind velocities appears to be sufficient to predict local dispersal and thus, to predict LDD, we need additional information on extreme events of updrafts and strong gusts is needed. By contrast, seed dispersal distance in endozoochorous species is mostly a function of seed retention time in

the frugivore's digestive tract and of frugivore's movement patterns (home range, habitat use, daily activity patterns) and thus migration movements of frugivores could give clues about LDD. These simulation models are predicting that dispersal could be up to two orders of magnitude higher than those previously obtained by empirical methods.

Quantitative and Qualitative Components of Seed Dispersal Effectiveness

The seed dispersal's effectiveness, or the contribution that frugivores make to plant fitness, can be considered at a variety of scales (from individuals to species communities) and decomposed in quantitative and qualitative components in terms of the effect of disperser to plant regeneration. The seed dispersal concept is more developed from the viewpoint of biotic seed dispersal of fleshy-fruited plants and thus more extensive modifications should be necessary for abiotic seed dispersal. In the case of biotic dispersal, the quantitative component is dependent upon the number of visits made on fruiting plants, upon the number of seeds dispersed in each visit, and is influenced by factors that are either intrinsic to the mother plant (e.g., size, fruit crop size, pulp/seed ratio) or extrinsic to it (e.g., fruit crops of neighbors, surrounding vegetation). In addition, the biology of frugivores is related to the species traits and activity (gape size, fruit-handling methods, degree of trophic generalism, etc.). By contrast, the quality of seed dispersal, usually more difficult to evaluate, is a function of (1) the quality of the dispersed seed (often associated with fruit and/or seed size and rather variable within an individual plant, and influenced by factors such as number of seeds/fruit); (2) the quality of seed treatment in each digestive tract of the frugivores (in turn dependent upon traits like seed coat thickness, chemical composition of pulp, gut passage time, morphology and physiology of the digestive tract, type of food ingested along with seeds, etc.); and (3) the quality of the microhabitat where the seed is deposited, which will ultimately determine the probability of germination and establishment; the sites where the seeds are deposited will be determined by factors such as frugivore movements after fruit removal, frugivore habitat preferences, etc. whilst the quality of the microsite will depend on abiotic (light levels, soil texture and humidity, etc.) and biotic conditions (levels of predation, competition, herbivory, etc.).

A successful plant recruitment depends on frugivore activity, on what type of fruit is selected, how it is processed, and the movements of the dispersers after seed ingestion, and is further determined by the biotic and abiotic factors prevailing in the recipient microhabitat where the seed is dropped. All together, this is crucial if we are to assess the demographic and evolutionary consequences of frugivore activity. With the available information, we know that the effects of the qualitative components of dispersal may erase the initial differences among frugivores in their quantity component, but more studies are needed to know which is more important determining the final pattern of plant recruitment.

Linking Seed Dispersal Patterns to Plant Establishment

Dispersal is the first of a series of stages that affects the subsequent plant regeneration process (Fig. 3), determining the spatial arrangement and, consequently, the population dynamics of most plant species. In the case of fleshy-fruited plants, the facilitation effect of nurse or shelter plants is in turn enhanced by the fact that most seeds are deposited under shrubs that act as feeding source and shelter for frugivores (the so-called perching effect). The effect of nurse plants seems to be crucial in many ecosystems because they ameliorate the negative effect of summer drought in seedling survival, at least when resources are limiting, as it occurs in arid ecosystems. However, in most cases the recruitment dynamics is usually very complex, and thus the seed template produced by seed dispersal can be subsequently modified by other regeneration stages (from seed dispersal to seedling establishment). This is because processes acting at different stages are usually independent and "uncoupled," and promote subsequently seed-seedling conflicts. One of the major hypothesis to explain the maintenance of tree species biodiversity is the "Janzen-Connell hypothesis" which coined the importance of seed dispersal mechanisms to avoid the negative effect of herbivores, host–pathogens and natural enemies in the vicinity of mother plants (Fig. 4). Hence, the deposition of large amounts of seeds by frugivores in the close vicinity of mother plants (despite microhabitat could be adequate there) may be subjected to intense post-dispersal seed predation and/or be unsuitable for emergence or recruitment of seedlings. Moreover, such uncoupling among regeneration stages may be also variable

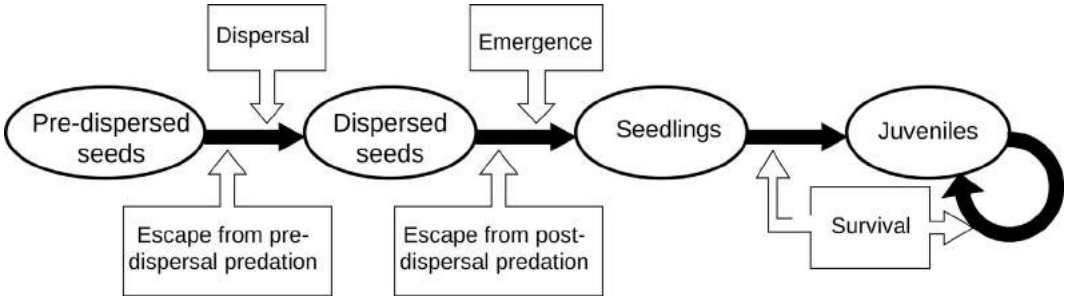


Fig. 3 Diagram representing the different stages (in *circles*) and processes (in *squares*) affecting them along the plant regeneration cycle. The overall probability of recruitment is obtained from the product of the partial probabilities of recruitment of each stage.

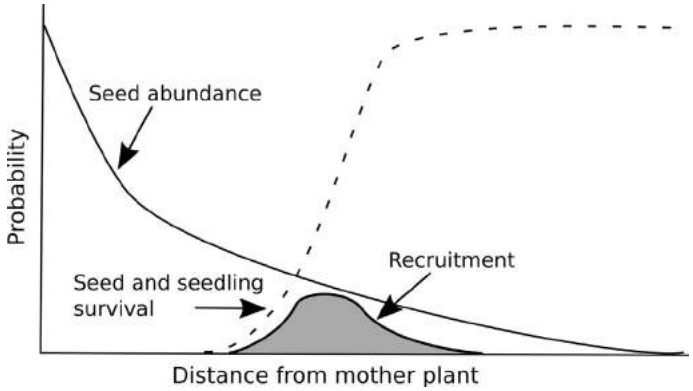


Fig. 4 Janzen-Connell hypothesis represented as the probability of distance from mother plant and the probability of survival to reproduction. The hypothesis predict that the expected maximum probability of seed survival occurs at intermediate distances from mother plants.

across spatial and temporal scales producing unpredictability of the plant regeneration process. Still, another possible source of conflict in the plant regeneration process is seed size. In general, larger seeds tend to have a higher chance of seedling emergence or establishment but may also have a lower probability of being ingested by frugivores or escaping from post-dispersal seed predators. These conflictive pressures will consequently affect the optimal value of seed size in many species to maximize the fitness through the overall regeneration process.

Consequences to Population Genetic Structure

In contrast to the great effort dedicated to figuring out demographic consequences, the effects of seed dispersal on the genetic structure of populations have received less attention, although information on this aspect is growing rapidly. Plant genes are dispersed either through haploid pollen or diploid seeds, and inheritance may be maternal (chloroplast DNA in angiosperms), paternal (chloroplast DNA in conifers) or biparental (nuclear DNA). Recent studies have revealed that plant species dispersed by animals have characteristically high levels of within-population genetic variation compared to other seed-dispersal vectors, and that variation is associated with extensive gene flow via seed dispersal in addition to outbreeding via pollen flow. Moreover, when plant populations dispersed by animals are structured in space (e.g., fragmented populations, metapopulations), frugivores have shown to strongly influence the among-population gene flow via seeds. The tools provided by microsatellites have also shown unequivocal genetic fingerprints of source mother plants in the population, revealing a marked heterogeneity in the genetic composition of the seed rain in different microhabitats, and also making it possible to know the fraction of seeds that come from other populations. Therefore, despite their low occurrence in nature, long-distance dispersal (LDD) events can now be tracked at different scales (i.e., landscape, regional or continental) by these genetic markers.

Plant–Frugivore Networks: Interpreting the Biodiversity of Interactions

In many ecological communities, plant–frugivore interactions are a key ecological process for maintaining biodiversity. At the community level, species interact with each other in a variety of ways and thus the interaction between plants and frugivores could be integrated in complex webs which help to contextualize the complexity of natural histories of partner species. Graph theory offers an analytical methodology with tremendous implications in the study of ecology of food webs, and has also offered an ideal conceptual framework for the study of such mutualistic networks integrated within species communities. This approach allows the description of the macroscopic structure of the entire web at the time that allows determining how fragile such species interactions are in front of different types of disturbances (e.g., introduction of alien species, changes in the abundance, or extinctions of particular species) (Fig. 5).

In the particular case of plant–frugivore interactions, common patterns emerging from studies of ecological networks are: (1) a low number of strong dependencies, (2) a high level of asymmetry in the interactions; thus if a plant depends strongly on a frugivore species, the animal depends weakly on the plant, and (3) a great heterogeneity in the strength of interactions among species. The three characteristics contribute to the maintenance of species coexistence in the community. As it occurs in many mutualisms between species there exist a high abundance of interactions not observed (forbidden links) which are usually prevented by either species traits and/or habitat specificity. For plant–frugivore interactions, the majority of frugivores cannot consume large-sided fruits (i.e., gape-size limitations), whereas non-overlapping phenologies of fleshy-fruited plants impede the temporal encounter between potentially interacting species. Thus, forbidden interactions are of major importance to understand the architecture of ecological

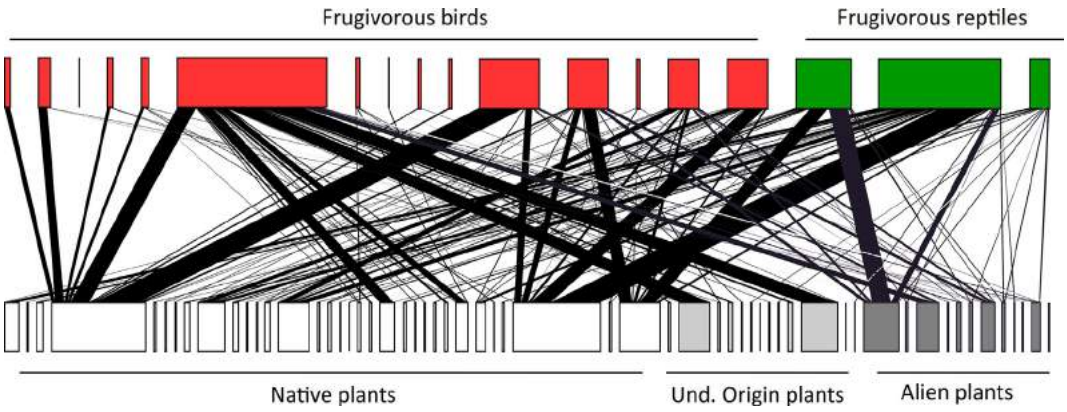


Fig. 5 Mutualistic network between fleshy-fruited plants and frugivore vertebrates in Galapagos archipelago. Nodes represent frugivore species (*upper boxes*) and plant species (*lower boxes*) whereas lines links between plant species and frugivores, that is, fruit consumption. Node size represents the relative occurrence of each interaction, in terms of the number of seeds of plant species consumed by frugivore species.

networks, and they allow us to weight the relative contribution of neutral mechanisms (i.e., abundant species have higher change to interact each other) or trait-matching (i.e., frugivores with large gape-sizes can consume the majority of fleshy-fruit species).

Seed Dispersal as Key Ecosystem Function and Service

Seed dispersal may drive plant gene flow, plant population dynamics and functional connectivity along landscapes and affect to key ecosystem functions related to (a) revegetation, recolonization and population dynamics of vegetation, and (b) the connectance and connectivity of information (species and genetic diversity), and it intimately depends on the scale of landscape structure of habitat. Due to these roles, seed dispersal is now considered a key ecosystem function as it has major implication for the colonization and recovery of fragmented and altered landscapes and the conservation and resilience of native ecosystems. Seed dispersal outcome may even be quantified in economic terms, when linking the process of vegetation recovery to recreation or carbon-sink uses as seed dispersal enhances the ecological succession. As suggested for other provisioning ecosystem services, at least three relevant components may be distinguished in seed dispersal function: the magnitude of seed delivery (abundance of seeds), the composition of seed input (richness of species dispersed), and the spatial pattern of seed rain which cascades into those processes structuring species communities.

In the case of seed dispersal by animals, frugivores are considered to be mobile links as they are able to connect habitat patches across landscapes. A large proportion of plants in tropical and temperate ecosystems bear fleshy fruits and thus the plant species richness there should be highly dependent on the relationship between the plant spatial pattern within the landscape and the activity of frugivorous animals. Frugivores may drop many seeds under plant canopies, and this deposition may be highly contingent to the individual plant location relative to co-fruiting partners, which strongly vary yearly. In this sense, seed dispersal by frugivores could additionally has properties related to spatially-explicit ecological networks, defined by nodes (represented by individual habitat patches or individual trees) and links (represented by the probability of dispersal among individual habitat patches or trees).

Loss of Seed Dispersal Function to Plant Communities

Seed dispersal is universally considered important for biodiversity conservation. The structure of the landscape has strong effects on the distances traveled by seeds, regardless whether they are dispersed by abiotic factors (wind) or by animals. Therefore, any type of disturbance, such as habitat fragmentation or habitat modification by an invasive plant species for instance, is likely to change the patterns of seed movement, the patterns of seed recruitment as well as the genetic structure of the plant populations. For wind-dispersed species, it is known that seeds travel much further distances in open landscapes than in dense forest, due to differences in the shape of the wind profile. On the other hand, plants depending on animals for seed movement are susceptible to dispersal failure when their seed vectors become rare or extinct. Disruption of the seed dispersal mutualism can have serious consequences for the maintenance of the plant populations. In tropical areas in particular, the widespread decimation of frugivores by overhunting and habitat loss has devastating demographic, genetic and evolutionary consequences for the maintenance of tree species diversity. In island ecosystems, such defaunation processes have even worse consequences due to the more simple communities found in them and thus due to the lower probability of species redundancy. A number of examples from islands has shown that the extinction or decimation of the main disperser of a plant has lead it to the near or even total extinction. An excessive long distance

dispersal of elements alien to ecosystems represents also a threat to biodiversity, especially if it goes along an insufficient dispersal of native species. In any case, long-lived plant populations could still reproduce and persist for decades without apparent seed dispersal processes, producing an "extinction debt" for their seed dispersal function after the vector is lost.

Seed Dispersal Under Global Change

It is expected that the current biodiversity crisis associated to anthropogenic alterations would promote functional consequences of ecological functions and services at multiple scales. Habitat destruction and fragmentation, overexploitation, biological invasions and climate change may likely cascade into deterministic species extinctions, loss of seed dispersal function and loss of ecosystem resilience. In the case of fragmentation of native habitats, decline of habitat areas and increasing isolation of native habitats have been found to reduce both the quantity of fruit removal and seed dispersal distances and the fruit removal by frugivores within and between fragments. An increasing number of studies are showing how the populations of frugivores are being decimated, both in the tropics and in the temperate zones, and how this translates into a lower dispersal success of the plants depending upon them. Biological invasions, or the introduction of novel species within species communities, create profound changes in the abundance, distribution and behavior of native species, most of them negative. After the introduction of alien plant species, for instance, native frugivores consume either alien and native plant species and thus potentially reduce the function provided to native plant species (see Fig. 5 for an example about this). The introduction of alien frugivores, by contrast, may provide either beneficial or negative effects on seed dispersal, depending if the alien frugivore is more or less effective dispersing plant species. Still, climate change is a factor producing loss of seed dispersal function, but there is still few evidence of its relevance in the long-term. Most evidence predicting changes in plant distribution ranges under climate change uses the "climate envelope" models relating future climate scenarios with changes in species distribution. However, such geographic models paradoxically usually ignore the majority of processes affecting vegetation changes, in particular species interactions and seed dispersal among others. Finally, global change drivers frequently act in combination with others, which is key to evaluate the synergistic, additive and antagonistic effect on the plant population resilience.

Further Reading

Bascompte J and Jordano P (2007) Plant-animal mutualistic networks: The architecture of biodiversity. *Annual Review of Ecology, Evolution, and Systematics* 38: 567–593.
Cousens R, Dytham C, and Law R (2008) *Dispersal in plants: A population perspective*. Oxford University Press.
Comita LS, Queenborough SA, Murphy SJ, Eck JL, Xu K, Krishnadas M, et al. (2014) Testing predictions of the Janzen–Connell hypothesis: A meta-analysis of experimental evidence for distance- and density-dependent seed and seedling survival. *Journal of Ecology* 102(4): 845–856.
Dennis A, Schupp EW, Green R, and Westcott DA (2007) *Seed dispersal—theory and its implications in a changing world*. Wallingford & New York: CABI International.
Forget P-M, Lambert JE, Hulme PE, and Van der Wall SB (2005) Seed fate: Predation. In: *Dispersal and seedling establishment*. Wallingford & Cambridge: CABI publishing.
Herrera CM and Pellmyr O (2002) *Plant animal interactions: An evolutionary approach*. Oxford: Blackwell publishing.
Jordano P (2017) What is long-distance dispersal? And a taxonomy of dispersal events. *Journal of Ecology* 105(1): 75–84.
Levey DJ, Silva WR, and Galetti M (2002) *Seed dispersal and frugivory: Ecology, evolution and conservation*. Wallingford & New York: CABI publishing.
Levin SA, Muller-Landau HC, Nathan R, and Chave J (2003) The ecology and evolution of seed dispersal: A theoretical perspective. *Annual Review of Ecology Evolution and Systematics* 34: 575–604.
McConkey KR, Prasad S, Corlett RT, Campos-Arceiz A, Brodie JF, Rogers H, and Santamaria L (2012) Seed dispersal in changing landscapes. *Biological Conservation* 146(1): 1–13.
Nathan R and Muller-Landau HC (2000) Spatial patterns of seed dispersal, their determinants and consequences for recruitment. *Trends in Ecology and Evolution* 15: 278–285.
Schupp EW, Jordano P, and Gómez JM (2010) Seed dispersal effectiveness revisited: A conceptual review. *New Phytologist* 188(2): 333–353.
Schupp EW, Jordano P, and Gómez JM (2017) A general framework for effectiveness concepts in mutualisms. *Ecology Letters* 20(5): 577–590.
Traveset A and Richardson DM (2006) Biological invasions as disruptors of plant reproductive mutualisms. *Trends in Ecology and Evolution* 21: 208–216.

Poikilotherms☆

Inna Sokolova, University of North Carolina at Charlotte, Charlotte, NC, United States

Etymology and Definitions

Poikilotherms (Greek poikilos—"various, spotted," and therme—"heat") are defined as organisms with variable body temperature (T_b) (Fig. 1). Typically, T_b in poikilotherms changes as a function of the temperature of their surroundings under normal physiological and environmental conditions. The opposite of this term is homeotherm (Greek homo—"the same," and therme —"heat"), which refers to organisms maintaining constant or nearly constant body temperature. The concept of poikilothermy originated in and was predominantly used for animal studies, but is in principle applicable to any organism. Poikilothermy is determined by the combination of extrinsic factors (i.e., variation of environmental temperature) with intrinsic (physiological) constraints of the organism such as low levels of metabolic heat production, rapid heat dissipation due to the small body size, high heat conductivity of the external medium (for soil dwellers and aquatic organisms), or excessive energy costs associated with the maintenance of the constant T_b. Although there is no explicit consensus about how broad the variations in T_b should be for an organism to be considered a poikilotherm, it is generally accepted that if the body temperature changes by more than 1.5°C–2°C under physiological conditions, the organism is a poikilotherm. Of cours
e, this rule of thumb excludes pathological conditions such as fever or specialized adaptations such as the regulated drop in T_b during entrance into hibernation or torpor in mammals and small birds, when a change in body temperature can considerably exceed 2°C.

Poikilotherm is one of the oldest terms in ecological physiology, and probably also one with the longest history of confusion. It would be impossible to understand the place of the concept of poikilothermy in thermal biology without clearing away some old but very persistent fallacies. In many texts as well as in vernacular language, poikilotherms are often incorrectly called "cold-blooded" organisms, in contrast to the "warm-blooded" homeotherms. In addition to muddling the meaning of poikilothermy, this statement is simply not true. In fact, poikilotherms can have body temperature similar to or even higher than in homeotherms. For example, some desert reptiles and insects can have T_b up to 40°C–44°C so that there is nothing "cold-blooded" about these poikilotherms. Overall, the cold-blooded and warm-blooded terminology has no real value for understanding of biological mechanisms or physiological consequences of thermoregulation and is best avoided. Similarly, poikilothermy should not be confused with ectothermy, which indicates the predominant reliance of an organism on external heat sources for thermoregulation. Although ectothermic thermoregulation is rarely efficient enough to provide constant T_b in thermally variable environments and most ectothermic organisms are in practice poikilotherms, there are several important exceptions from this rule, and poikilotherms can be found among endotherms as well as ectotherms.

Evolutionary Adaptations to Poikilothermy and Its Ecological Implications

Evolutionary adaptations of poikilotherms are dictated by the necessity to withstand a substantial variation in body temperature. Across the animal kingdom, different species of poikilotherms have evolved to operate at body temperatures from −1.86°C (e.g., some polar fish and invertebrates) to up to 44°C–45°C in certain tropical fish, desert insects, and reptiles, while dormant or quiescent life stages of some animals (such as some rotifers and tardigrades) can survive temperatures spanning from nearly −273 to over 100°C. Within each species of poikilotherms, the range of tolerated body temperatures is smaller, but can still be very appreciable. Thus, in temperate and subpolar poikilotherms, seasonal temperature changes may lead to a gradual change in T_b by 15°C–30°C. On a short-term basis, some land insects and reptiles from temperate climates and marine intertidal invertebrates may experience rapid variations of T_b in excess of 20°C–30°C during diurnal or tidal cycles. Behavioral escape mechanisms (such as migration or habitat choice) may reduce thermal stress but are rarely sufficient to completely prevent a change in T_b. As a result, physiological and biochemical functions of poikilotherms have evolved to withstand a wide range of fluctuations in T_b which would be immediately lethal for most active homeotherms.

Temperature change directly affects the rates of all biological processes as well as stability of macromolecules and membrane structures. At high temperatures, increasing molecular motion may lead to structural destabilization and eventually damage. At low temperatures, a decrease in kinetic energy of the molecules results in low rates of biochemical reactions and the loss of membrane fluidity incompatible with sustaining active life. If the temperature drops further, below the freezing point of intracellular fluids, water crystallization and resulting mechanical damage to the cells becomes a problem. Therefore, a major challenge of poikilothermy is to maintain cellular and systemic homeostasis in the face of temperature-induced functional and structural

☆*Change History:* February 2018. Irene Martins made minor changes to the text and references.

This is an update of I. Sokolova, Poikilotherms, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 2851–2854.

 doi:10.1016/B978-0-12-409548-9.11131-5

Fig. 1 Poikilothermic animals: (A) a reef fish Anthias squamipinnis; (B) a nudibranch mollusk Coryphella on a bryozoan colony; (C) a lion mane jellyfish *Cyanea capillata*; (D) a freshwater crustacean Daphnia (water flea); (E) a spring frog Rana; (F) a tiger swallowtail *Papilio glaucus* on honeysuckle. Photos: (A–E) Mikhail Fedyuk, (F) Inna Sokolova.

alterations in their cells. Poikilotherms have evolved multiple ways to achieve this homeostasis, which include profound alterations of intracellular milieu, membrane composition and properties, enzyme activities, and concentrations of molecular chaperones and cryoprotectants.

Biological membranes are among the most temperature-sensitive cellular sites in poikilotherms. Changes in T_b strongly affect membrane fluidity, which in turn may affect its integrity and permeability, as well as signal transduction and function of membrane-associated proteins and cytoskeleton. A suite of biochemical mechanisms known as homeoviscous adaptation allows poikilotherms to maintain optimal levels of membrane fluidity in the face of temperature change. These mechanisms involve adaptive changes in the degree of acyl chain saturation of the membrane phospholipids, changes in the cholesterol content and ratio of different phospholipid classes (phosphatidyl choline to phosphatidyl ethanolamine) in the membrane. In different poikilotherms, homeoviscous adaptation may be brought about by the de novo synthesis of certain lipid classes, biochemical modification of existing membrane lipids, cholesterol synthesis or breakdown, as well as by seasonal changes in the diet. Some mammalian hibernators selectively feed on plants rich in polyunsaturated fatty acids before entering into hibernation. This leads to an increase of the unsaturated lipid content in their membranes and fat depots, lower temperature set points during hibernation, and improved winter survival rates. Interestingly, diet can also affect temperature preference of

an organism resulting in modified behavior. For example, Australian shingleback skinks select cooler environments when fed diets artificially enriched in polyunsaturated fatty acids, and this diet-induced shift in the preferred body temperature may reach 5°C.

Another key aspect of the variable T_b in poikilotherms is variation in the rates of enzymatic reactions, which has profound "ripple" effects on the rates of all integrative processes, from metabolism and growth to neurotransmission and behavior. Decrease in body temperature results in slowing down the rates of enzymatic reactions, which may in turn result in reduced rates of growth and reproduction, as well as impaired locomotion and ability to escape predators or to find food. On the short-term scale, homeostasis of enzymatic reaction rates may be achieved by changing concentrations of reaction substrates and products, or variation in intracellular levels of allosteric regulators of enzyme activity. During a prolonged decrease in T_b (e.g., during seasonal cold acclimatization), decreasing reaction rates can be compensated by elevated enzyme concentrations, expression of less-temperature-sensitive isoforms of enzymes, or both. However, this compensation is often incomplete, and in most poikilotherms a decrease in body temperature is associated with a decreased activity and growth rate.

Although elevated temperatures enhance rates of enzymatic processes (and thus, "the rate of living") in poikilotherms, an excessive increase in T_b is damaging and potentially lethal due to the destabilization and eventual denaturation of cellular proteins. In order to protect against such denaturation, poikilotherms may express molecular chaperones (particularly so-called heat shock proteins, or HSPs), which assist in proper folding of partially denatured proteins and stabilization of their native conformation. Expression of HSPs is almost universal response to heat stress in the animal kingdom and found in all poikilotherms, as well as most homeotherms. The only known exception is some extremely stenothermal and cold-adapted Antarctic fish species which have lost the ability to induce HSPs in response to heat stress. Increasing T_b also results in a decline in intracellular pH in poikilotherms, which helps to support normal folding and function of intracellular proteins through the maintenance of constant levels of protonation of their critical o:-imidazol groups. Taken together, these changes in intracellular milieu help to maintain structural integrity and cellular homeostasis in poikilotherms facing a change in T_b.

Preventing ice formation is a significant challenge for poikilotherms living in habitats where environmental temperatures fall below the freezing point of intracellular fluids. Many poikilothermic species such as Arctic and Antarctic fishes, terrestrial arthropods and amphibians, plants and fungi are known to seasonally synthesize and accumulate antifreeze agents such as glycerol, sorbitol (and other polyols), trimethylamine-*N*-oxide (TMAO), as well as specialized antifreeze proteins and glycoproteins. These compounds decrease the freezing point of intracellular fluids and some of them also provide thermal hysteresis (lowering of the temperature required for crystal growth beyond that needed for crystal melting), thus preventing formation and growth of intracellular ice crystals. Owing to these mechanisms, some glycerol-rich insects may supercool to -60°C without freezing. Caterpillars of the butterflies *Aporia crataegi* can survive several months with body temperature as low as -50°C; to achieve such remarkable hardiness, 14% of their body weight is composed of cryoprotectants. In hibernating land frogs, high tissue levels of glucose serve as cryoprotectants. Synthesis of the cryoprotectants in poikilotherms is regulated by hormonal systems, which in turn are typically activated by photoperiod rather than temperature. This allows animals to accumulate sufficient levels of cryoprotectants in their tissues before the environmental temperature actually drops below freezing.

It is worth noting that most of the above-described adaptive changes to maintain homeostasis in the face of changing T_b require considerable times to be accomplished (e.g., days to weeks) and are typically associated with long-term acclimation or acclimatization of poikilotherms to the changed thermal environment, for example, during seasonal temperature changes or evolutionary adaptation to different climates. During short-term temperature fluctuations, poikilotherms have to put up with temporary disturbances of cellular homeostasis and must depend on the robustness of their intracellular systems to survive those disturbances. Due to the inevitable constraints on structure and function of macromolecules (and thus on the range of the temperatures to which the organisms may be successfully adapted), there is no species that "could take it all" and could survive the changes of T_b spanning over the whole range of temperatures consistent with active life. Due to the varying T_b as a function of ambient temperature and the high temperature sensitivity of their physiology, distribution patterns of poikilotherms often closely follow gradients or discontinuities in environmental temperature. It is perhaps no wonder that the most striking examples of the temperature-induced shifts in species distribution come from poikilotherm species. The threshold effects of temperature (i.e., the minimum amount of the temperature change which is sufficient to result in a significant shift of the species distribution limits) may be quite sublime, and a change of the mean temperature by 1°C–2°C can strongly shift the geographical distribution of poikilotherms. This high temperature dependence of poikilotherm biogeography is evidenced not only by paleontological record but also by the recent observations of major distribution shifts in aquatic and terrestrial poikilotherms which are correlated with (and likely caused by) increases in ambient temperature of about 1.2°C–2.2°C in the last century. A less-than-exhaustive list of recent climate-driven changes in poikilotherm distribution include major faunal shifts in shallow-water marine habitats, local extinctions of poikilotherm populations at the southern limits of the distribution range, declines in zooplankton abundance, extensive bleaching of coral reefs, increases in mosquito-borne diseases in highlands, and the northward shift of ranges of nonmigratory insects. With the global climate change, more research will be needed to improve our understanding of physiological and biochemical mechanisms underlying the distribution shifts of poikilotherm populations and to analyze the profound ecosystem-level effects of these shifts.

Acknowledgments

The author was supported by National Science Foundation (IBN-0347238) and the University of North Carolina at Charlotte Faculty Research Grant during work on this manuscript. The author also wishes to thank Mikhail Fedyuk for providing photographs.

Further Reading

Culos, G.J., Tyson, R.C., 2014. Response of poikilotherms to thermal aspects of climate change. Ecological Complexity 20, 293–306.
Guschina, I.A., Harwood, J.L., 2006. Mechanisms of temperature adaptation in poikilotherms. FEBS Letters 580 (23), 5477–5483.
Hazel, J.R., 1995. Thermal adaptation in biological membranes: Is homeoviscous adaptation the explanation? Annual Review of Physiology 57, 19–42.
Heinrich, B., 1993. The hot-blooded insects. Cambridge, MA: Harvard University Press.
Hochachka, P.W., Somero, G.N., 2002. Biochemical adaptation. Mechanism and process in physiological evolution. Oxford: Oxford University Press.
Prosser, C.L. (Ed.), 1991. Environmental and metabolic animal physiology. New York: Wiley-Liss.
Somero, G.N., 2005. Linking biogeography to physiology: Evolutionary and acclimatory adjustments of thermal limits. Frontiers in Zoology 17, 1–9.
Willmer, P., Stone, G., Johnston, I., 2000. Environmental physiology of animals. Oxford: Blackwell Science.

Soil Ecology

MA Pavao-Zuckerman, University of Arizona, Tucson, AZ, USA

Introduction

Soils are an essential component of the world's ecosystems, providing rooting medium for plants and serving as the habitat for the saprophytic organisms that recycle energy, matter, and nutrients through the decomposition process. Soils have been an essential provider of ecosystem services throughout human history, as most food comes ultimately from plants that grow in soils and sediments. Ancient societies dating back to the Mesopotamians and Egyptians have recognized the importance of soils. We can see the deep importance of soils in the root and origins of many words, for example, the shared root of the words humus and human. While the relationship between soil organisms and soil health has been linked for centuries, the discipline of soil ecology has its origins in microbial ecology of the nineteenth and twentieth centuries, soil science of the early twentieth century, and soil management (particularly agricultural systems) in the mid-twentieth century. Recent efforts in soil ecology have focused on developing a mechanistic understanding of how organisms and soils interact to yield patterns of soil biodiversity, nutrient cycling function within ecosystems, and feedbacks to global change mechanisms.

Soil ecology is concerned with interactions, be it between organisms or between organisms and the soil environment. Soil ecology has its origins in soil biology and soil zoology, the study of organisms in the soil habitat. Soil ecology and soil science are related, yet different disciplines, with soil science focusing more on physical processes, the classification and genesis of soils, soil chemistry, and soil physics. When we ask ecological questions about soils from an ecosystems perspective, the inclusion of physical and chemical interactions with organisms extends this overlap between soil ecology and soil science. The conceptual and analytical tools between the two disciplines differ because they have different epistemological and metaphysical foci. The breadth of the conceptual domain of the discipline of soil ecology can be seen in **Fig. 1**.

Components and Characteristics of Soils

Soils are a mixture of weathered mineral rock particles, organic matter (i.e., both living, and dead and decaying), water, and air. Soils can be thought of as functional entities that are the resulting products of the interaction of physical, chemical, and biological processes. Soil mineral particles are usually described based upon their relationship to soil texture, and include sand (0.05–2.0 mm), silt (0.002–0.05 mm), and clay (<0.002 mm). The relative distribution of these particles is used to describe different soils, and affects soil properties such as bulk density, pore space, and particle density. Variation in these physical properties structure habitats

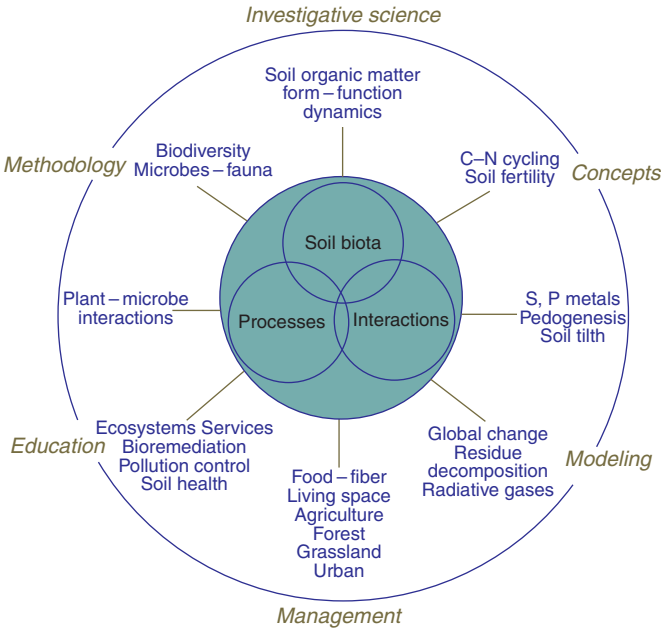


Fig. 1 The conceptual domain of soil ecology includes organisms, processes, and the societal context for research. From Paul, E.A. (Ed.), 2006. Soil Microbiology, Ecology, and Biogeochemistry, 3rd edn. Burlington, MA: Academic Press.

 doi:10.1016/B978-0-444-63768-0.00850-7

for different soil microbes and fauna. Soils are also composed of organic materials, which include the partially fragmented and decomposed remains of plant material, animals, microbes, and feces. Organisms are also an important organic constituent of soils, and their activity can influence soil properties (e.g., porosity), the binding of inorganic materials into soil aggregates, and nutrient cycling in soils. Soil air is the gases that are present in soil pores that are not filled with water. Oxygen and carbon dioxide (CO_2) are important constituents, and their concentration in the soil affects many processes (e.g., nitrification and denitrification). Soil water can contribute up to 30% of soil volume, and is essential for the activity and physiological functioning of organisms in the soil.

Soil structure is a description of the spatial arrangement of soil particles into aggregates. Several hierarchical schemes for classifying soil structure have been proposed that link soil aggregates across several orders of magnitude of scale, from the submicron level (microaggregates) to several square meters (the pedon). The aggregate structure of soils has important implications for the distribution of pores in soils, which affects the distribution and availability of water and gasses. While soil structure generally refers to this aggregate structure, soils also have a spatial structure with depth. Soils in general are composed of different horizons, if one views them in profile. On the surface is a layer of recently fallen plant litter, with smaller amounts of excreta and dead bodies. Below this litter layer is a horizon enriched in organic material from the breakdown of the litter layer, and further below are mineral horizons of various textures and thicknesses. This vertical structure derives from the varying influence of soil formation processes with depth. In the 1940s, Jenny described the formation of soils as an ecological process that results from the interactions of factors, including climate, organisms, parent material, and relief. More recent theoretical advances resulting from agroecological and urban ecosystem research have extended the `organisms' category to explicitly include the actions of humans in the formation of soils and soil structure. These physical, chemical, and biotic formation factors lead to variations in the distribution to depth of various mineral, chemical, and organic components of soil, giving soils a depth profile (**Fig. 2**).

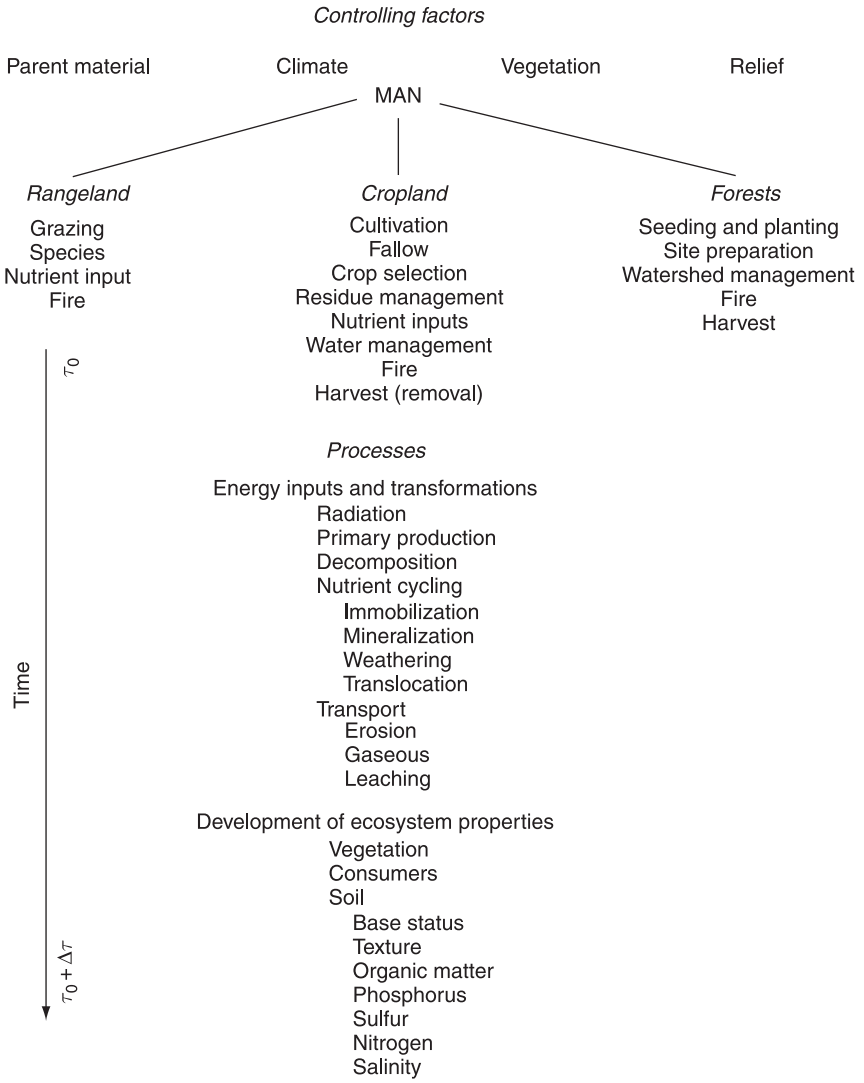


Fig. 2 Factors that control soil formation. Soil formation is an ecological process that involves the interactions of climate, organisms, parent material, and relief. From Coleman, D.C., Crossley Jr., D.A., Hendrix, P.F., 2004. Fundamentals of Soil Ecology, 2nd edn. Burlington, MA: Academic Press. Modified from Jenny, H., 1980. The Soil Resource: Origin and Behavior. In Ecological Studies 37. New York: Springer.

Soil Biota

Soil ecology has part of its roots in soil zoology, and places heavy emphasis on the study of organisms within the context of the unique soil habitat. While historically emphasis had been placed upon the description and classification of soil organisms, contemporary focus links biodiversity and interactions of soil organisms to broader-scale ecological phenomena, including carbon fluxes, nutrient cycling, and biogeography. It should be noted that general understanding of biodiversity in soils has been historically undeveloped in part because *in situ* identification of soil microbes has been problematic; only around 10% of soil microbes have been successfully identified via *in vitro* culturing, and the taxonomic resolution of soil fauna is incomplete. Complicating matters further, soils are a complex and opaque medium, rendering even the *in situ* observation of macrofauna problematic. However, recent applications of genomic techniques to soil microbes and biogeographic perspectives on soil taxonomy have helped to increase our basic understanding of the diversity, abundance, and distribution of soil organisms.

Diversity of Soil Organisms

Soils are a highly diverse habitat and harbor such high levels of biodiversity that they are often referred to as `the poor person's rainforest'. Microbes, including bacteria, fungi, and cyanobacteria are the most numerous and diverse taxonomic groups in the soil. There are also many species of animals that live in the soil, including single-celled amoebae, free-living and parasitic nematode worms, Acari, oligochetes, and insects. In addition to these `permanent' and `periodic' soil residents, animals can also be `temporary' or `transient' residents of soil as life cycle stages may be completed below ground (i.e., egg and larval stages of Diptera) or behaviors may drive organisms below ground (i.e., nesting of many small birds, reptiles, and mammals). Plant roots are also important biotic components of soil that both influence soil physical structure and contribute organic detritus to soil food webs.

Conceptual Organization of Soil Organisms

Soil ecologists rely upon several classification schemes to conceptually link the diversity of soil organisms to ecological functions and processes. One of the more useful classifications is based upon size and uses body diameter to distinguish soil biota into microflora (<2 μm), microfauna (2–100 μm), mesofauna (0.1–2 mm), macrofauna (>2 mm), and megafauna (>20 mm). Organisms in different size classes have different spatial effects in soils and interact with different soil processes. Individually, microflora impact relatively small spatial scales and are important in the breakdown of organic residues and the production of exudates that form soil aggregates; microfauna are important in regulating these processes. Meso- and macrofauna operate at larger spatial scales, and physically breakdown plant residues, create pore structure, and mix and redistribute mineral and organic soil components (**Fig. 3**).

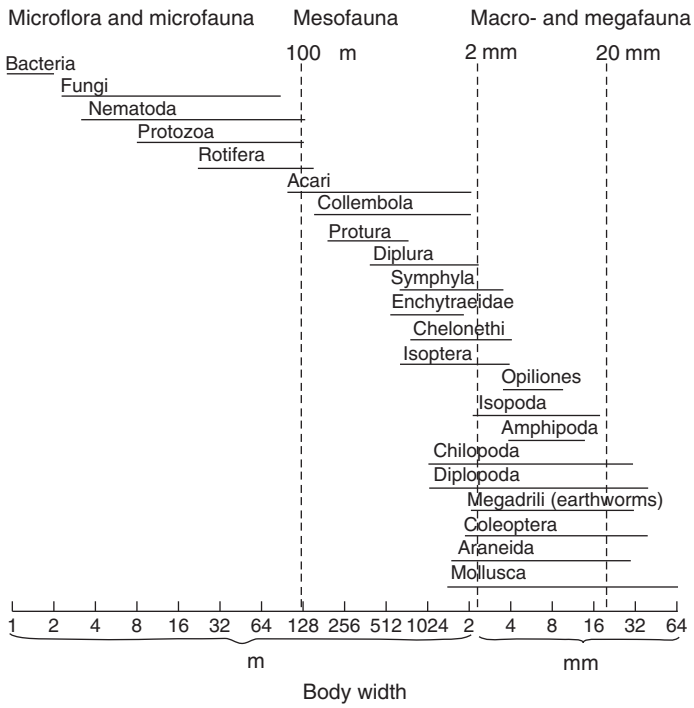
Much success is also found from viewing soil organisms from a food web or trophic perspective. Problems associated with *in situ* identification of microbes and the degree of unnamed species of soil fauna makes the use of broader taxonomic groups necessary, including functional groupings of organisms. Soil organisms are commonly grouped based upon feeding habits, including microbe feeders (microbivores), those that feed on litter and organic matter (detritivores), those that feed off or parasitize plant roots (herbivores), or those that feed on other soil animals (carnivores). Soil food web dynamics are elucidated through a mixture of lab, field, and modeling exercises. The use of radioactive and stable (both pulses additions of labels and natural abundances) isotopic tracers and statistical techniques (such as interaction strength and path analysis) have been important in determining the spatial and temporal dynamics of soil food webs. Generally, it is thought that the interaction of bottom-up (productivity based) and top-down (predatory) controls lend to the stability of soil food webs. Soil food web analyses have been important in demonstrating the roles of microbe feeders (e.g., nematodes and protozoa) on regulating both microbial populations and decomposition and nutrient dynamics in soil ecosystems. Experimental research suggests that there are two distinct pathways in the soil food web: (1) a `fast' bacterial pathway, and (2) a `slow' fungal pathway. A third pathway via the activity of root grazers (e.g., plant parasitic nematodes) is also possible, but is theoretically less developed. Observations in experiments and field studies of nematode-trapping fungi and ectomycorrhizal fungi that prey upon collembolans complicate the trophic structure traditionally conceptualized in soil food webs, and while they have not yet been integrated into soil food web models, they point to the degree of trophic linkages possible in soils (**Fig. 4**).

A third approach integrates spatial scale and trophic perspectives, linking specific spatial scales to soil processes. Ecosystem engineers, such as earthworms and termites, can indirectly influence the cycling of nutrients through direct impacts on soil structure. A second group is considered the litter transformers (micro- and macroarthropods) and fragments or comminutes litter into smaller pieces, thereby increasing the surface area available to microbial decomposition. Members of a third group are part of a `micro-food web', and include microbes and microfaunal predators (primarily nematodes and protozoa). Each level in this conceptual scheme influences ecosystem properties through actions at different size, space, and timescales.

The Role of Organisms in Soil Functions and Processes

Soil organisms play key roles in ecosystems through their effects on physical properties and processes, and the biological contributions to carbon and energy fluxes and cycling of nutrients. The importance of soil fauna for soil physical properties generally

(a)



(b)

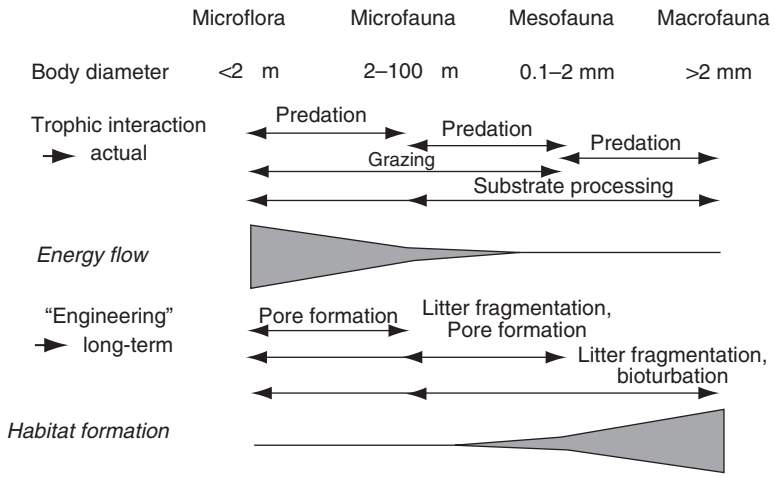


Fig. 3 (a) Size classification of soil organisms by body width. (b) The interactions of soil organisms are dependent on size of the organism. Depicted here are trophic interactions and ecosystem engineering effects. Note that with increasing size the relative effects on energy flow decrease, while the effects on habitat formation increase. (a) Swift, M.J., Heal, O.W., Anderson, J.M., 1979. Decomposition in Terrestrial Ecosystems. Oxford: Blackwell Scientific Publications. (b) From Scheau, S., Setälä, H., 2002. The soil food web: Structure and perspectives. In: Tscharntke, B., Hawkins, B.A. (Eds.), Multitrophic Level Interactions. Cambridge, UK: Cambridge University Press, pp. 223–264.

increases with larger body sizes. Soil macrofauna, such as earthworms, ants, and termites, can have dramatic effects on soil porosity, creating macropores and tunnels that allow for preferential flow of water into the soil profile. The movement of macrofauna through the soil profile (such as some species of earthworms) can mix mineral particles from one horizon into another, and can bring fragments of leaf litter from the surface to mineral soil horizons, thus affecting soil texture, bulk density, and organic matter contents. A notable exception to this size relationship is the role of microbes in the formation of soil aggregates. The activity of microbes, particularly myccorhizal fungi, produces exudates that help to bind together soil particles into aggregates.

Roughly 80–90% of net primary production enters the soil as dead plant material (e.g., leaf litter, stems, roots) or organic exudates from roots. This is decomposed primarily by bacteria and fungi in the soil and surface litter layer. It is the physiological

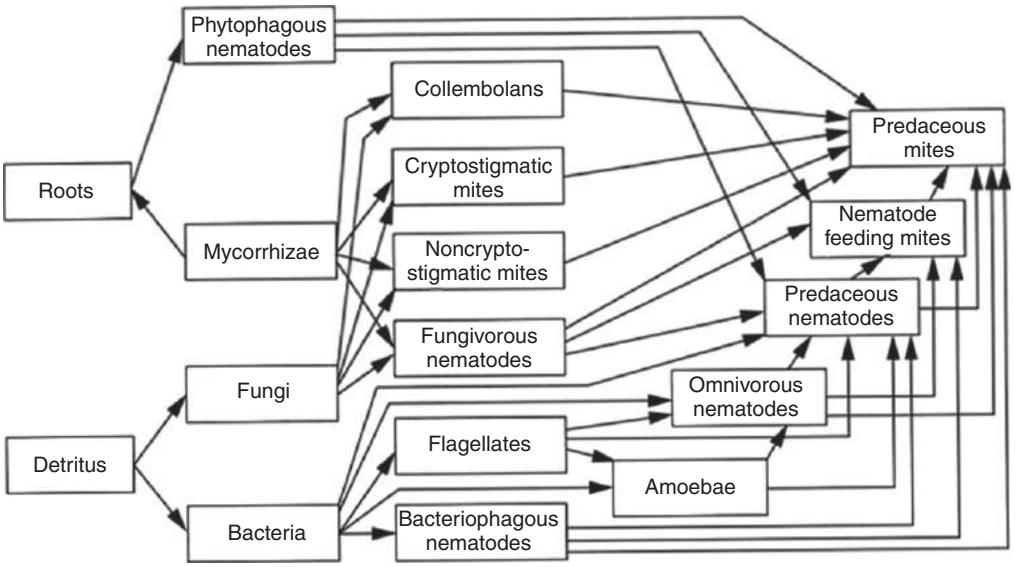


Fig. 4 A conceptual model of a shortgrass prairie soil food web. From Bardgett, R.D., 2005. The Biology of Soil: A Community and Ecosystem Approach (Biology of Habitats). Oxford: Oxford University Press. Adapted from deRuiter, P.C., Neutel, A.-M., Moore, J.C., 1995. Energetics, patterns of interaction strengths, and stability in real ecosystems. Science 269 (5228), 1257–1260.

processes of, and enzymatic digestion by the microbes, as well as the trophic interactions in the soil food web that control the cycling of nutrients in the soil. Decomposition is the transformation of an organic substrate from one form to another, and is the source of CO_2 that is respired from soils, and the recycling of nutrients.

Decomposition is controlled by the interactions of the physical environment, the quality of the substrate, and the community decomposer organisms. Temperature and soil moisture are the dominant controls on decomposition, with warmer, wetter soils generally having faster rates of organic matter decomposition. Substrate quality is a function of the stoichiometry of the plant material and the composition of the litter in terms of structural components (lignin, cellulose) and secondary metabolites (tannins, phenolics), which can decrease the palatability of litter and make it more recalcitrant. The composition of the community of soil organisms mediates decomposition because different organisms have different functional roles in soils. For example, the fungal community will display successional patterns as leaf chemistry is changed during the process of decomposition, matching the suite of active fungi to substrate quality. Exclusion studies have demonstrated that soil fauna also play a role in controlling the decomposition process. The microbial loop refers to the stimulation of nutrient availability via the trophic interactions between bacteria and their consumers (protozoa and nematodes), where predation on microbes stimulates their population growth and enhances rates of nutrient cycling. Earthworm casts provide a physical and chemical environment that is more suited to microbial activity than the surrounding bulk soil. So, while decomposition processes are driven by the extracellular enzymes of microbes, the production of enzymes and the environment in which they function is regulated by soil fauna.

Soils are notoriously heterogeneous habitats when viewed from the perspective of the soil flora and fauna. Over the scale of micro- and millimeters organisms may encounter films of water clinging to soil particles, air-filled soil pores, concentrations of organic material, aggregates of soil, and plant root hairs. One conceptual construct for managing the complexity of soils in ecological studies is to think of `hot spots' of activity, where because of concentrations of available resources and/or microclimatic conditions, organism growth and activity is concentrated. These hot spots in soils make up a small portion of the total volume of soil (up to 10%) but will contain the majority of biological activity in soils (over 90%). The hot spots are located in (1) the drillosphere (area influenced by earthworm burrows), (2) the rhizosphere (area influenced by plant roots), (3) the aggregatusphere (soil aggregates where soil mineral particles are bound with organic materials), (4) the detritusphere (larger concentrations of leaf litter and organic matter), and (5) the porosphere (the spaces between soil aggregates). It is within these hot spots that habitats and resources are suitable for soil organisms, and where most decomposition and nutrient cycling happens in soils. Because of their effect on the spatial distribution of these hot spots, the distribution, composition, and physiology of the plant community thus can have very strong controls on organism activities in soils. Through their controls on nutrient cycling, the availability of nutrients in the soil, and soil physical characteristics, the activity and composition of the soil microbes and fauna, in turn, has a strong influence on the plant community. The nature and degree of such above- and belowground linkages is a major research question in contemporary soil ecology.

While specific soil organisms or functional types are linked to decomposition and nutrient-cycling processes, the explicit link between levels of biodiversity and soil ecosystem function is still in question. While this is likely the result of the high functional redundancy in the community of soil organism, this trend is also partly due to problems of taxonomic resolution mentioned above. However, recent advances in molecular techniques hold the promise of being able to move past problems of *in situ* identification of soil microbes. Techniques can now identify the presence of genes that code for specific biogeochemical processes

(such as, lignin decomposition, nitrification, and phosphorus utilization), and some techniques are roughly quantifiable, so it is possible to see how much of a particular gene is in the soil, actively working on an enzymatic process. The integration of these molecular techniques into studies of biogeochemistry is in its infancy, but holds strong promise for allowing researchers to link specific organisms and physiological processes at the level of microbes to a specific nutrient cycling pathway in ways that were previously unthinkable.

Conclusions

Soils are key sites of ecosystem functions, providing catabolic analogs to the aboveground processes of photosynthesis, assimilation, and primary production. Decomposition of plant material drives the cycling of nutrients, and it is the activity of microbes that controls the availability of nutrients in the soil. The role of soil biota in ecosystem processes and functions has long been recognized as important, yet the specific nature of these interactions and effects is still under investigation. Particularly of interest are linkages across broad spatial and temporal scales. There is much to be learned about how processes that occur at the micro- and millimeter-scale in soils (and is observed in meter-scale plot research) can be scaled up to the level of ecosystems and the biosphere. This has implications that extend beyond basic scientific research. The majority of carbon that is stored in terrestrial ecosystems exists in soil organic matter. The response of soils to global change processes is an important control and feedback point in global biogeochemical cycles. For example, an important aspect unknown is how soils will respond to potential shifts in detritus quantity and quality that comes about from plant responses to global climate change. Further complicating matters is that global change factors such as elevated atmospheric CO_2 levels, global climate change, chronic N-deposition, and landscape transformations will have interactive effects on soils, yet we do not know if they will be additive or multiplicative interactions. The implications of these interactions for the management of soils within the context of global change, ecological restoration, and landscape transformations remain an important component of understanding and managing the resilience and sustainability of ecological systems.

Further Reading

Bardgett, R.D., 2005. The Biology of Soil: A Community and Ecosystem Approach (Biology of Habitats). Oxford: Oxford University Press.
Bardgett, R.D., Usher, M.B., Hopkins, D.W. (Eds.), 2005. Biological Diversity and Function in Soils. Cambridge: Cambridge University Press.
Beare, M.H., Coleman, D.C., Crossley Jr., D.A., Hendrix, P.F., Odum, E.P., 1995. A hierarchical approach to evaluating the significance of soil biodiversity to biogeochemical cycling. Plant and Soil 170, 5–22.
Brussard, L., 1998. Soil fauna, guilds, functional groups and ecosystem processes. Applied Soil Ecology 9, 123–135.
Cardon, Z.G., Whitbeck, J.L., 2007. The Rhizosphere: An Ecological Perspective. Burlington, MA: Academic Press.
Coleman, D.C., Crossley Jr., D.A., Hendrix, P.F., 2004. Fundamentals of Soil Ecology, 2nd edn. Burlington, MA: Academic Press.
deRuiter, P.C., Neutel, A.-M., Moore, J.C., 1995. Energetics, patterns of interaction strengths, and stability in real ecosystems. Science 269 (5228), 1257–1260.
Special feature: New directions in microbial ecology. Jackson, R.B., Fierer, N., Schimel, J.P. (Eds.), Ecology 88, 1343–1400.
Jenny, H., 1980. The Soil Resource: Origin and Behavior. In Ecological Studies 37. New York: Springer.
Moore, J.C., McCann, K., Setala, H., de Ruiter, P.C., 2003. Top-down is bottom-up: Does predation in the rhizosphere regulate aboveground dynamics? Ecology 84, 846–857.
Paul, E.A. (Ed.), 2006. Soil Microbiology, Ecology, and Biogeochemistry, 3rd edn. Burlington, MA: Academic Press.
Scheau, S., Setälä, H., 2002. The soil food web: Structure and perspectives. In: Tscharntke, B., Hawkins, B.A. (Eds.), Multitrophic Level Interactions. Cambridge, UK: Cambridge University Press, pp. 223–264.
Swift, M.J., Heal, O.W., Anderson, J.M., 1979. Decomposition in Terrestrial Ecosystems. Oxford: Blackwell Scientific Publications.
Wall, D.H. (Ed.), 2004. Sustaining Biodiversity and Ecosystem Services in Soils and Sediments. Washington, DC: Island Press.
Wardle, D.A., 2002. Communities and Ecosystems: Linking the Aboveground and Belowground Components. Princeton, NJ: Princeton University Press.
Wolters, V., 2000. Invertebrate control of soil organic matter stability. Biology and Fertility of Soils 31, 1–19.
Zak, D.R., Blackwood, C.B., Waldrop, M.P., 2006. A molecular dawn for biogeochemistry. Trends in Ecology and Evolutionary Biology 21, 288–295.

Stable Isotope Ecology

Alexandra Baeta, MARE - Marine and Environmental Sciences Centre, University of Coimbra, Portugal

Isotope Basics

What Are Isotopes?

All atoms of a given element have the same number of protons and electrons, but some atoms have more neutrons than other atoms of the same element and therefore weigh more (Fig. 1). These particular forms of an element defined by a specific number of neutrons are referred to as isotopes of the element. Isotopes are denoted by an atomic "formula." The atomic number symbolizes the number of protons in the nucleus of each atom of an element, and is used to identify the position of the element on the periodic table. The mass number is the total number of neutrons and protons present in its nucleus.

In nature, an element occurs as a mixture of its isotopes. The element hydrogen, for example, has three naturally occurring isotopes. One, simply known as hydrogen, ^{1}H (usually written as simply H; also called protium), has one proton and no neutrons; the deuterium isotope (or D), ^{2}H, contains one proton and one neutron; and tritium (or T), ^{3}H, has one proton and two neutrons (Fig. 1). Hydrogen is the only element whose isotopes are given different names. As another example, consider the three isotopes of the element carbon, which has the atomic number 6. The most common isotope is carbon-12, ^{12}C, which accounts for about 99% of the carbon in nature, has six neutrons. Most of the remaining 1% of carbon consists of atoms of the isotope ^{13}C, with seven neutrons. A third isotope, ^{14}C, has eight neutrons; it is present in the environment in minute quantities. Notice that all three isotopes of carbon have six protons—otherwise, they would not be carbon. Because isotopes differ in mass, they have slightly different physical properties (=characteristics), but they always have similar chemical properties. The similarity occurs because only the electrons are used in chemical reactions, not the neutrons or protons.

Word Origin and History

The term "isotope" is formed from the Greek roots *isos* ("equal") and topos ("place"), meaning "at the same place," thus the word "isotope" comes from consideration of the periodic table of the elements, and means that different isotopes of a single element occupy the same position on the periodic table of the elements.

The existence of isotopes was first suggested by the British chemist Frederick Soddy, in 1913, based on studies of radioactive decay chains that indicated about 40 different species referred to as *radioelements* (i.e., radioactive elements) between uranium and lead, although the periodic table only allowed for 11 elements from uranium to lead. Soddy proposed that several types of atoms (differing in radioactive properties) could occupy the same place in the table. The term "isotope" was suggested to Soddy by Dr. Margaret Todd, a Scottish physician and family friend, in 1913, during a conversation in which he explained his ideas to her. This term was accepted and used by Soddy, and has become standard scientific nomenclature. In 1921, he received the Nobel Prize in Chemistry "for his contributions to our knowledge of the chemistry of radioactive substances, and his investigations into the origin and nature of isotopes" (see http://nobelprize.org/chemistry/laureates/1921).

Measurement, Notation and Terminology

Mass spectrometry is the analytical technique used to make precise determinations of the isotopic ratios. The complete process involves the conversion of the sample into gaseous ions, with or without fragmentation (usually by combustion), which are then characterized by their mass to charge ratios (m/z) and relative abundances. The isotope ratio mass spectrometer (IRMS) is a specialization of mass spectrometry, and is the most common method to quantify stable isotopes: after conversion of a sample to

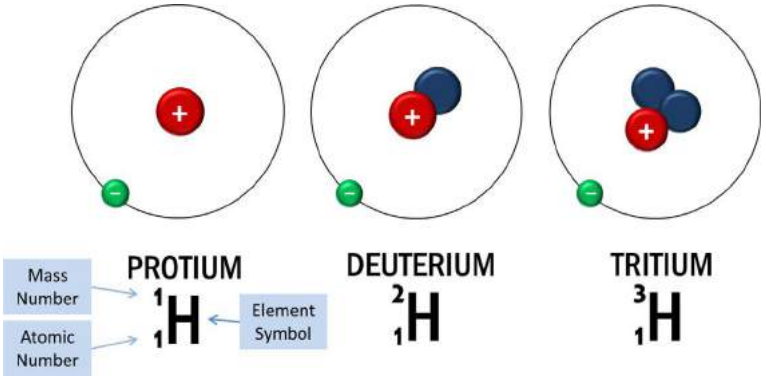


Fig. 1 The three naturally occurring isotopes of hydrogen, showing their protons (marked in *red*, with positive charge), neutrons (*blue*, with no charge), and electrons (*green*, with negative charge). Based on http://nuclearconnect.org/know-nuclear/science/isotopes and https://www.giss.nasa.gov/research/briefs/schmidt_03/.

gaseous form (e.g., for C, it would be converted to CO_2), it is introduced into an ionizing source and the results are then compared to a standard. Thus, the IRMS is generally coupled to an elemental analyzer, in which the sample is converted to a gas. The isotopic composition of the pure gas is measured using IRMS.

For quality assurance and quality control purposes, the two most important factors to consider when working with an IRMS laboratory is the calibration, that should occur regularly, and standards used. Calibration should occur regularly and the laboratory should be able to demonstrate that their current measures match either international or laboratory standards (or both) to within ±1‰. To demonstrate this, laboratories generally incorporate standards at the start or finish (or both) of each "run" (i.e., a discrete set of samples analyzed continuously). The instrument is calibrated with samples of a standard mass; when samples are prepared, an effort should be made to conform all sample masses to the amount indicated by the laboratory or the results can be biased. When analyzing samples, the analytical facility should introduce replicate standards at regular intervals (e.g., every 10–12 samples in the run) for quality assurance. For acceptable precision, replicate reference material should have an error ≤ 0.2‰. Rapid technological advances over the past decade have greatly stimulated the use of isotope analyses, and this analytical approach is now among the most popular in ecology because of the insights provided by isotope ratios at natural abundance levels. The main stable isotopes used in ecology are ^{13}C, ^{15}N, ^{18}O, ^{2}H, ^{34}S, and sample preparation will depend on the physical form of the sample and the element to be measured.

Isotopic composition is reported relative to an internationally accepted standard and expressed in "parts per mil" (‰) deviation from that standard following the formula: $\delta(‰) = [(R_{sample}/R_{standard}) - 1] \times 10^3$, where R is the ratio of heavy-to-light isotope, R_{sample} is that in the standard. For example, the defined reference standard for ^{13}C is the Pee Dee Belemnite (PDB) limestone, while for ^{15}N the atmospheric nitrogen. These analytical reference materials can be obtained from the IAEA (International Atomic Energy Agency). Isotopic values were expressed in the δ unit notation. The Greek letter delta, Δ or δ (capital and small letter delta, respectively), symbolizes change, and expresses the relative differences in isotopic ratios between samples and standards that are measured by isotope ratio mass spectrometry. A positive δ value means that the ratio of the heavy to the light isotope (e.g., ^{2}H:^{1}H, ^{13}C:^{12}C, ^{15}N:^{14}N, ^{34}S:^{32}S) is higher in the sample than it is in the standard, whereas a negative delta value has the opposite meaning. Heavy isotopes are usually much rarer than light isotopes so the ratio can be very small. Because the ratio is so small, it is multiplied by 1000 for a more useable value, and as a result, the notation is parts per mil (‰). For example, the ^{13}C:^{12}C (R) for the PDB standard is 0.0112372, or about 1.12 ^{13}C atoms for every 100 ^{12}C atoms. If a sample has 1.11 ^{13}C atoms for every 100 ^{12}C, then the $\delta^{13}C$ value would be −21.1‰. In ecological studies, for example, the difference in $\delta^{13}C$ values between two locations is often 5‰ to 15‰, which is a small difference in the natural abundance of ^{13}C.

There are several common means to compare the isotopic composition of two materials (including, e.g., "heavy" vs. "light," "high" vs. "low" values, "enriched" vs. "depleted"), and it is therefore essential to use correct, clear and consistent terminology, and phraseology, in studies using isotope analysis. For example, the recommended terms for expression of the stable carbon isotope ratio are "$\delta^{13}C$ value," "carbon isotope composition," "^{13}C-depleted (or enriched) sample," and "high (low) $\delta^{13}C$ values."

Types of Isotopes

There are two types of isotopes, stable and radioactive isotopes. For example, both ^{12}C and ^{13}C are stable isotopes, meaning that their nuclei do not have a tendency to lose particles (stable over time). The isotopes ^{1}H and ^{2}H are also examples of stable isotopes. An isotope tends to be stable when the number of neutrons (N) and the number of protons (Z) are quite similar ($N/Z \leq 1.5$). The isotopes ^{14}C and ^{3}H, however, are unstable, or radioactive. A radioactive isotope has an unstable nucleus that decays spontaneously, giving off particles and energy (emitting alpha, beta, or gamma rays), until stability is reached. When they decay, they release particles that may be harmful. This is why radioactive isotopes are dangerous. A loss of nuclear particles transforms the atom to an atom of a different element. For example, radioactive carbon decays to form nitrogen-14, while tritium decays into helium-3.

Eighty out of the first 82 elements in the periodic table have stable isotopes. In the total, there are ~300 stable isotopes, over 1200 radioactive isotopes. The distribution between light and heavy isotope in nature varies among isotopes, but for most cases the light isotope is the most abundant. As such the natural abundance of the lightest stable isotope of the elements hydrogen (H), carbon (C), nitrogen (N), oxygen (O), and sulfur (S) is >95% (Table 1). On the contrary, the heavy stable isotopes of some elements such as boron (B) and lithium (Li) are the abundant isotopes, >80% of the total. Only a few elements such as bromine (Br), silver (Ag), and europium (Eu) show approximately equal distribution between light and heavy stable isotopes. By definition, natural abundance is the measure of the average amount of a given isotope naturally occurring on Earth. Most elements have more than one stable isotope; only 21 elements that are known to have only one isotope. For example, hydrogen has two ^{1}H and ^{2}H, with the lighter isotope being much more abundant (Table 1).

Uses of Radioisotopes

Radioactive isotopes have many useful applications. In biology, for example, researchers use the ^{14}C, a radioactive isotope of carbon, to measure the age of ancient biological materials, a method known as radiocarbon dating (also referred to as carbon dating or carbon-14 dating). In agriculture radioisotopes also show wide applications to develop better fertilizers, control insects and improve plant breeds, as well as in food processing and preservation.

Radioisotopes are also useful as tracers to follow atoms through metabolism, with various uses in medicine (medical therapy, diagnostics and research), since cells use the radioactive atoms as they would nonradioactive isotopes of the same element, but the radioactive tracers can be detected. Nuclear medicine, for example, is a field of medicine that uses a trace amount of radioactive substances for the diagnosis and treatment of many health conditions such as certain types of cancer, and neurological and heart diseases. In medicine, two of the most commonly used radioisotopes are technetium-99m and iodine-131. Radioactive isotopes are also used for medical research to study normal and abnormal functioning of organs and systems.

Although radioactive isotopes are very useful in biological research and medicine, radiation from these decaying isotopes also poses a hazard to life by damaging cellular molecules. The severity of these damages depends on the type and amount of radiation an organism absorbs. One of the most serious environmental threats is radioactive fallout from nuclear accidents.

Application of Stable Isotopes

Today, stable isotopes play an important role in science. Their unique properties enable them to be used in a broad variety of applications across diverse scientific disciplines including hydrology, oceanography, geology, (paleo)climatology, biogeochemistry, archeology, biology, ecology, physiology, astronomy, medicine, forensics, among many others (Fig. 2).

In hydrology, for example, stable isotopes are useful tools for characterizing several different water dynamics within a watershed. Isotope hydrology is based on the notion of tracing a water molecule through the hydrological cycle. Water is the most fundamental component of the Earth's climate. Its presence in all three phases—liquid, solid and gas—defines our planet in a very profound way. Tracking the movements of water through the system—in oceans, air, clouds, rain, snow, ice, lakes, rivers, and back to the oceans—is therefore a primary concern of climatologists. Where does the water come from? Where does it go? What feedbacks are involved? Common stable isotopes used in hydrology as conservative tracers are ^{2}H and ^{18}O. These isotopes occur naturally in the environment, but their natural abundance differs with different environmental conditions. These environmental isotopes (applied through meteoric processes) can be used to trace and identify different air and water masses contributing precipitation to a watershed since the stable isotope composition of water changes only through mixing and well-known fractionation processes

Table 1 Average abundances of stable isotopes that are important for understanding ecological systems

Element	*Isotope*	*Average abundance (%)*
Hydrogen	^{1}H	99.985
	^{2}H	0.015
Carbon	^{12}C	98.89
	^{13}C	1.11
Nitrogen	^{14}N	99.63
	^{15}N	0.37
Oxygen	^{16}O	99.759
	^{17}O	0.037
	^{18}O	0.204
Sulfur	^{32}S	95.00
	^{33}S	0.76
	^{34}S	4.22
	^{35}S	0.014

From West, J.B., Bowen, G.J., Cerling, T.E., Ehleringer, J.R. (2006). Stable isotopes as one of nature's ecological recorders. *Trends in Ecology and Evolution* **21**, 408–414.

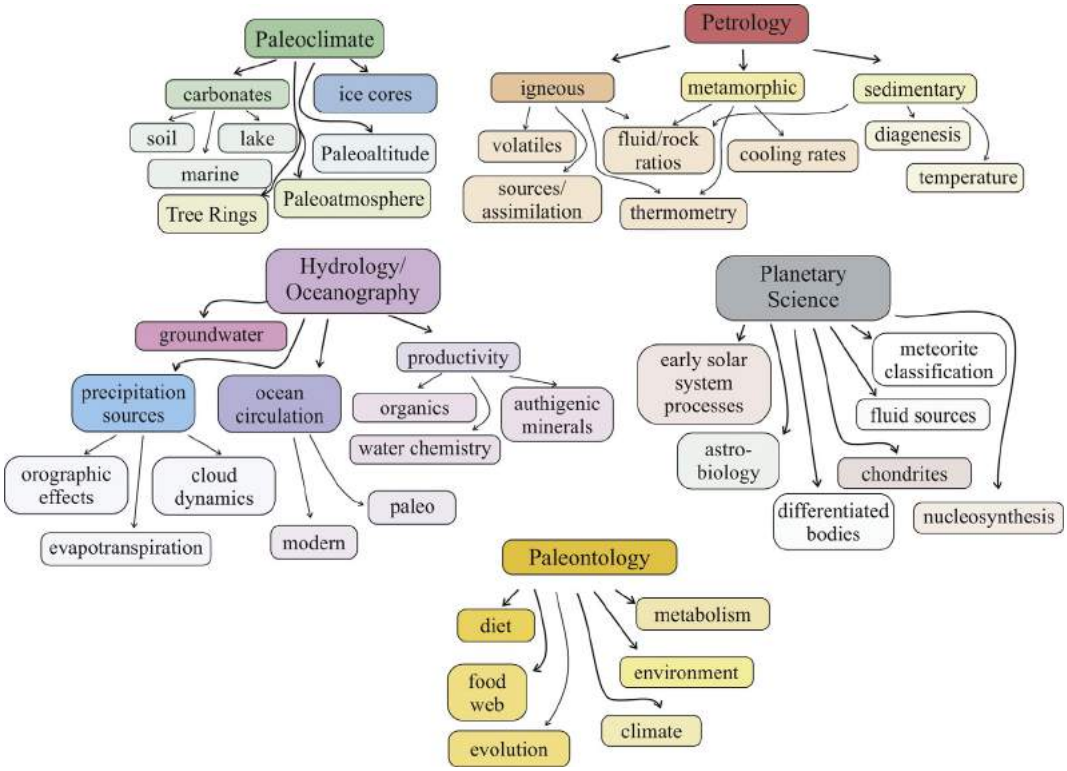


Fig. 2 Examples of the various fields and types of applications of stable isotopes. From Sharp, Z. (2017). *Principles of stableisotope geochemistry* (2nd edn.). https://doi.org/10.5072/FK2GB24S9F, reprinted with permission.

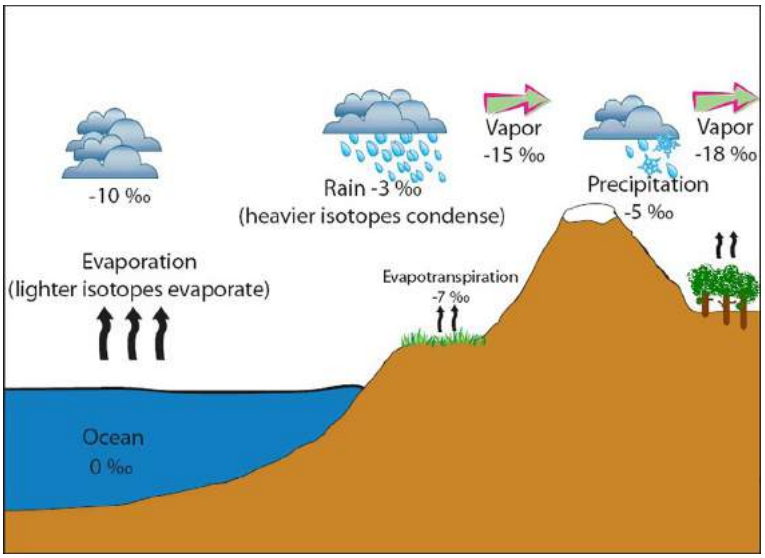


Fig. 3 A schematic diagram of the isotope fractionation process via evaporation, condensation, and evapotranspiration (combination of evaporation and transpiration). Notice that waters are lighter when they evaporate and are relatively heavier when condensed in the form of precipitation. From Bruckner, M. Z. (2007) Stable isotope primer and some hydrological applications, microbial life educational resources. Retrieved with permission from https://serc.carleton.edu/microbelife/research_methods/environ_sampling/stableisotopes.html (30th April, 2018).

that occur during evaporation and condensation (Fig. 3). For water in the climate system, the most important fractionations occur when water evaporates from the ocean (the evaporating water is slightly lighter, or more depleted, than the ocean water it came from) and when it condenses in clouds (the condensate is a littler heavier, or more enriched, than the water vapor). This pattern leads to a very clear pattern of progressively more depleted isotopes as you move toward the poles from the equator (Fig. 4). Thus, water from different places or processes can have a different isotope signal, a characteristic fingerprint of its origin and therefore can

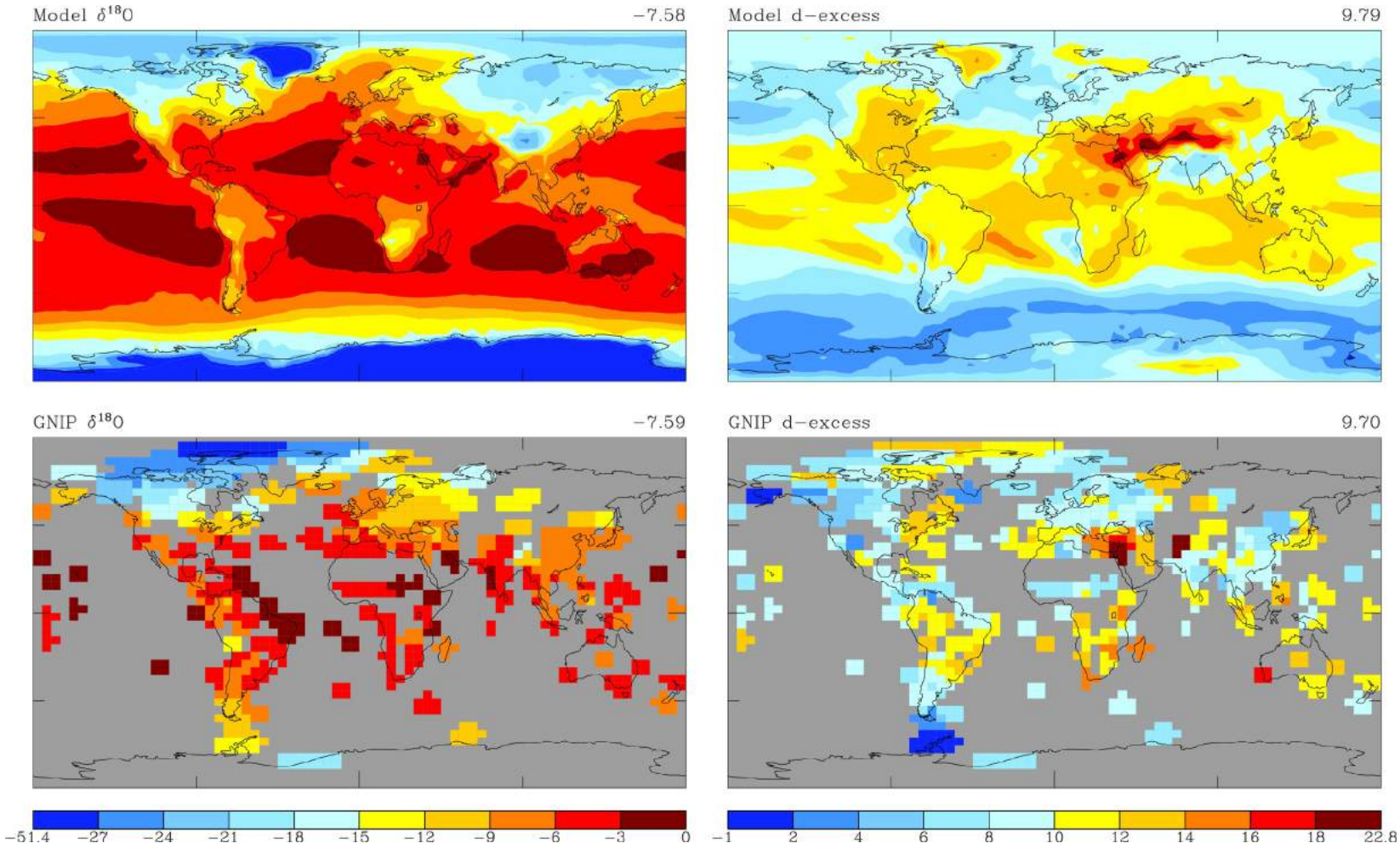


Fig. 4 Annual mean $\delta^{18}O‰$ (the isotopic ratio in precipitation) in a control simulation (averaged over last 5 years; *top*) compared to observations (*bottom*) in the GNIP database. Ratios are measured as a deviation from a standard, with more negative values implying less heavy isotopes than normal, positive values indicating more. The general pattern is for more depletion as you go toward the pole or away from the coast. *Reproduced with permission from* Aleinov, I. and Schmidt, G.A. (2006). Water isotopes in the GISS ModelE land surface scheme. *Global and Planetary Change* **51**, 108–120.

help identify where the water in the stream comes from. This happens in polar studies where the proportion of ice melt can be tracked through the ocean, and in hydrology where seasonal variations in isotopes can be tracked through a watershed. In paleoclimate studies, for example, the ratio of ^{18}O to ^{16}O in ice cores and fossil remains of microorganisms is commonly used to identify colder versus warmer periods in the Earth's past; higher abundances of ^{16}O indicate warmer periods (increased evaporation), whereas higher abundances of ^{18}O indicate cooler periods (decreased evaporation).

Natural stable isotope abundance techniques are also used by ecological researchers globally as an important tool to understand origins, migration and other movements (e.g., birds and fishes), resource portioning, host–parasite interactions, plant water use and nutrient status, ecophysiological processes, ecosystem fluxes of carbon, nitrogen and water, pollution, food webs, and sources of organic matter. For this purpose, stable isotopes of carbon and nitrogen are of specific interest.

In fact, the versatility of stable isotope applications has greatly expanded the variety of studies possible. Therefore, this article will focus on the application of stable isotopes in food web research.

Stable Isotopes in Food Web Research

At present, the major goal for ecology is to understand how communities and ecosystems will change following pollution, habitat destruction, overexploitation, invasion, and climate change. Such changes have triggered and accelerated the decrease in biodiversity and modified the structure and functioning of ecosystems, thereby jeopardizing the maintenance of goods and services provided to humans. A vision of community and ecosystem changes related to trophic interactions may allow a more holistic understanding of how communities and functioning may change in response to global. Under this framework, food web studies should take into account the dynamic nature of the trophic relationships, which vary following species behaviors and population dynamics as well as for spatial and temporal variability of the habitat considered. Such an approach may contribute to the preservation and management of ecosystems in view of global change.

It is not surprisingly that the benefit of using stable isotope analysis (SIA) approach to elucidate trophic interactions is high, with significant advantages over traditional methods (e.g., stomach content analysis). For example, the stable isotope approach, particularly those of nitrogen and carbon, provides a number of potential advantages over dietary methods, and has enhanced our understanding of trophic structure and dynamics of ecological communities, as well as ontogenetic shifts in consumer diet. Stable isotopes offer three potential advantages in terms of food web analysis; firstly, the $\delta^{13}C$ and $\delta^{15}N$ ratios of animal tissue represent the integration of carbon and nitrogen over a prolonged period (conservative markers; integration of different functions and environmental events over time periods); secondly, they are based on assimilation rather than ingestion (measurement of the effective use of food); and third they can be measured from comparatively small samples. In addition to time-integrated trophic information, isotope signatures have the potential to simultaneously capture complex interactions, including trophic omnivory, and to track energy or mass flow through ecological communities.

Stable Isotopes of Nitrogen and Carbon

The central assumption of SIA in trophic ecology is perhaps best represented by the observation first pointed out by DeNiro and Epstein (1976) "You are what you eat (plus a few per mil)". Thus, the isotopic structure of the prey is roughly assumed by the predator, based on that assumption. Indeed, predator metabolism operate a selection between the isotopic forms, preferentially respiring the light C isotope (^{12}C) and excreting the light N isotope (^{14}N). As a result consumers tend to be isotopically heavier than its food source (since lighter isotopes are preferentially used in metabolism) (Figs. 5 and 6), a process called discrimination factor (also fractionation factor or trophic enrichment) ($\Delta^{13}C$ and $\Delta^{15}N$ for carbon and nitrogen, respectively). This effect is more pronounced in nitrogen and leads to increases in $\delta^{15}N$ of approximately 3‰–4‰ at each trophic level. Thus, "you are what you eat (or more correctly assimilate) plus a few per mil." Because of the discrimination that occurs with trophic transfers, $\delta^{15}N$ can be used as a proxy for trophic positions.

Stable carbon isotope ratios vary substantially among primary producers with different photosynthetic pathways, but change little with trophic transfers, generally from 1‰ to 2‰ from prey to predator. Therefore, $\delta^{13}C$ are typically used to determine which primary producer components are the ultimate carbon source. For example, comparing marine and freshwater systems, marine phytoplankton have a $\delta^{13}C$ value of ca. −24‰ to −19‰ because the $\delta^{13}C$ of total dissolved inorganic CO_2 (DIC) in the ocean is about 0‰ and the incorporation of carbon by C3 plants proceeds with a fractionation of about −21‰. Riverine sources, such as matter derived from terrestrial vegetation (C3 plants, ca. −27‰ $\delta^{13}C$), soils (ca. −26‰ $\delta^{13}C$), and phytoplankton (< -30‰ $\delta^{13}C$), generally are depleted in ^{13}C compared to the estuary and marine systems (Figs. 6 and 7). Thus, freshwater organisms may be ^{15}N-depleted compared to marine organisms because terrestrial derived organic matter and freshwater algae often have a $\delta^{15}N$ value of −2‰ to 7‰, which is generally less than estuarine and coastal marine phytoplankton (7‰ to 10‰). These different isotopic compositions of primary producers are reflected in the tissues of their consumers, and again in the next level up the food web.

Natural-abundance SIA provides ecological information not only on food web structure, trophic interactions, and nutrient flux, but offers also a method for retrospective geolocation. The correct interpretation of stable isotope data requires an understanding of spatial and temporal variation in the isotopic compositions at the base of the food web. Incomplete knowledge of the likely spatio-temporal variation in baseline isotope values over an animal's foraging range, and/or over seasonal, annual, or multiannual cycles, can lead to poor sampling design and inaccurate interpretation of results.

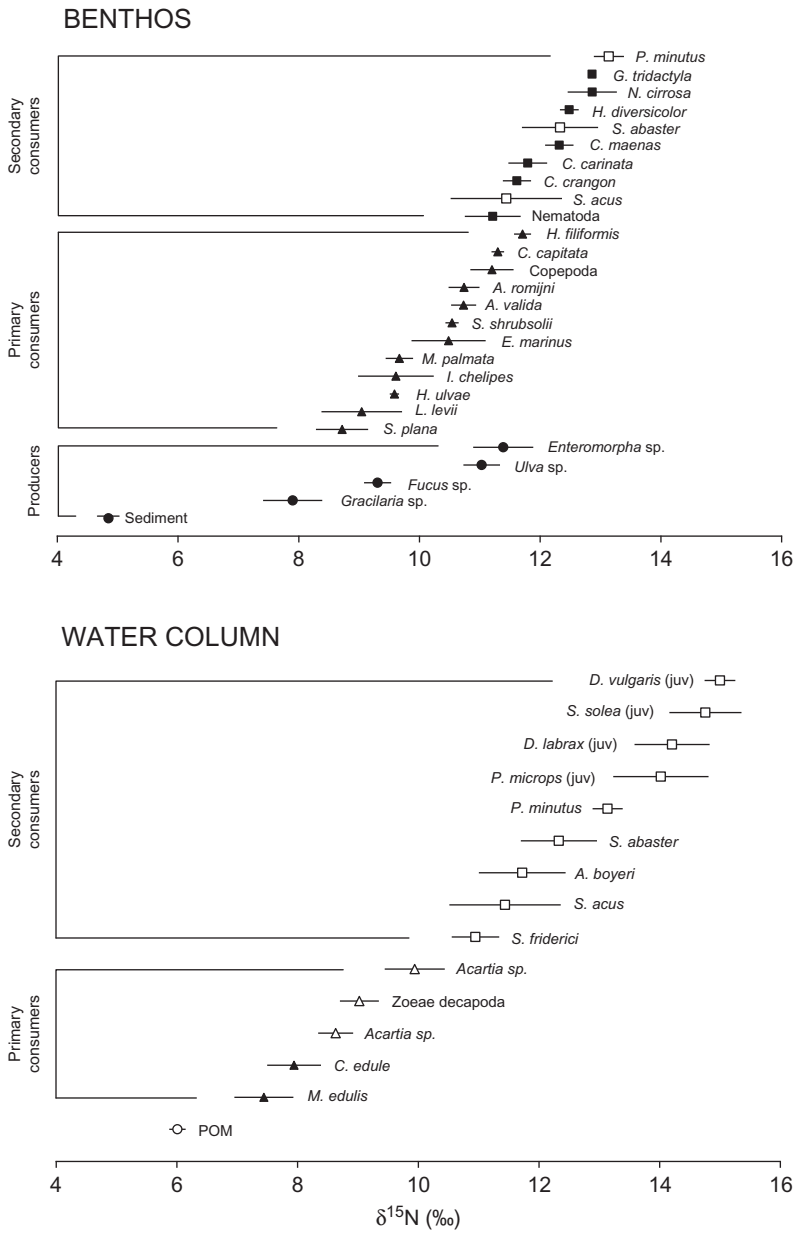


Fig. 5 Nitrogen stable isotope signatures of producers and consumers for the benthic (*top*) and water column (*bottom*) food webs from the *Zostera* site, in the Mondego estuary, Portugal. From Baeta, A., Valiela, I., Rossi, F., Pinto, R., Richard, P., Niquil, N., and Marques, J.C. (2009). Eutrophication and trophic structure in response to the presence of the eelgrass *Zostera noltii*. *Marine Biology* **156**, 2107–2120.

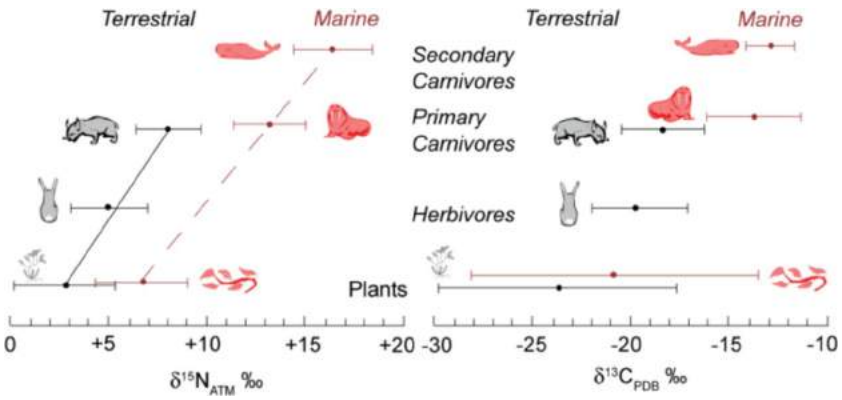


Fig. 6 Values of $\delta^{13}C$ and $\delta^{15}N$ in various marine and terrestrial organisms. From White, W.M. (2014). *Isotope geochemistry*. Wiley-Blackwell. https://doi.org/10.5072/FK2GB24S9F.

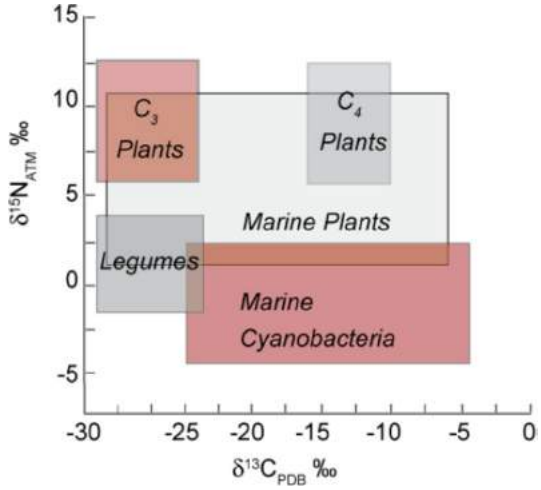


Fig. 7 Relationship between $\delta^{13}C$ and $\delta^{15}N$ among the principal classes of autotrophs. From White, W.M. (2014). *Isotope geochemistry*. Wiley-Blackwell. https://doi.org/10.5072/FK2GB24S9F.

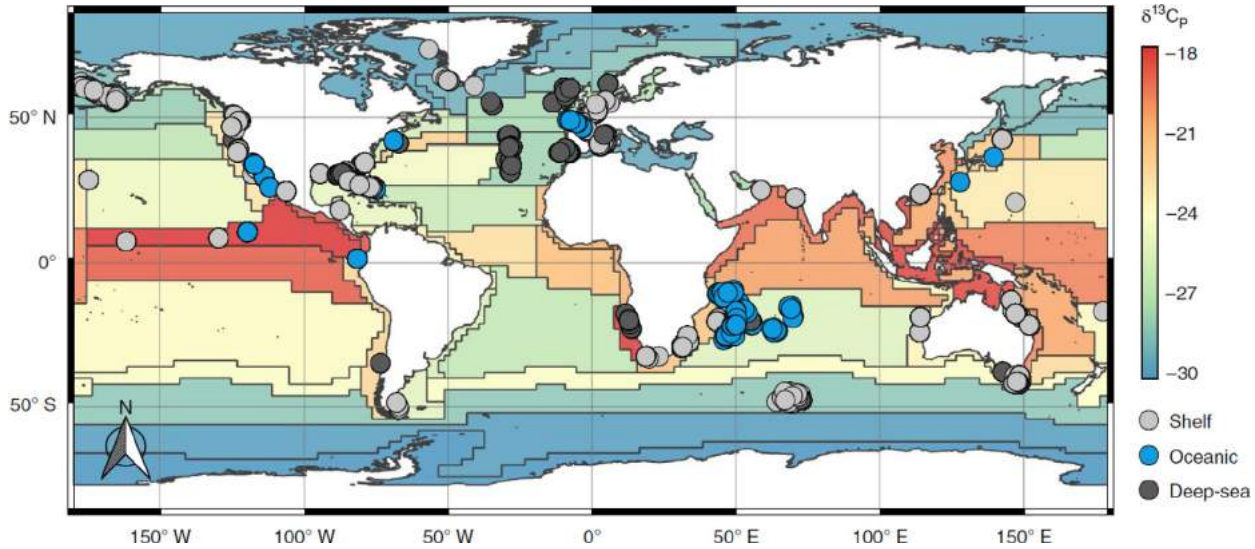


Fig. 8 Distribution of compiled shark data overlaid on a spatial model of annual average biomass weighted $\delta^{13}C_P$ within Longhurst biogeographic provinces from the median sampling year (2009). The colored points signify the habitat classification of those samples. Most studies provided one location for multiple samples. From Bird, C.S., Veríssimo, A., and Trueman, C.N. (2018). A global perspective on the trophic geography of sharks. *Nature Ecology & Evolution* **2**, 299–305.

Bird et al. (2018), for example, studied the global feeding habits of sharks by using stable isotopes of carbon in shark muscle as a natural tracer. The relative abundances of the two stable isotopes of carbon ^{12}C and ^{13}C in marine phytoplankton vary systematically with latitude, and among different coastal primary producers such as algae, seagrasses and mangroves. Differences in carbon isotope compositions are then passed from phytoplankton to consumers throughout food webs during feeding. This works of the premise of "you are what you eat" and also "you are *where* you ate." In this study, by tracking the original site of photosynthetic fixation of carbon atoms that were ultimately assimilated into muscle tissues of sharks, it was possible to identify globally consistent biogeographic traits in trophic interactions between sharks found in different habitats. Sharks that live near to the coast or in shelf seas appear to get most of their food from local phytoplankton based food webs (regional pelagic sources), but contain individuals that populations of these shelf sharks appear to be made up of individuals that forage within additional isotopically diverse local food webs, such as those supported by terrestrial plant sources, benthic production and macrophytes. In contrast, oceanic sharks that are found throughout the world's oceans seem to get most of their food from phytoplankton based food webs in areas of cooler water in the northern and southern hemispheres, between 30° and 50° of latitude from the Equator. Accordingly, increased understanding of species feeding ecology is essential to design more effective protection measures to conserve biodiversity and ecosystems (Fig. 8).

$\delta^{15}N$ as an Indicator of Nitrogen Pollution to Ecosystems

Isotopes are also useful tracers in ecological systems. For example, $\delta^{15}N$ is an effective tool to track natural and anthropogenic nitrate pollution sources to marine, freshwater, and terrestrial systems. Due to the wide variety of sources, nitrogen can enter in these systems in the form of nitrate (NO_3^-), nitrite (NO_2^-), and ammonium (NH_4^+). Moreover, nitrogen can be converted from one form into another through various reactions and processes.

In marine systems, for example, high levels of $\delta^{15}N$ in basal sources can be used as an alternative method to establish the level of human eutrophication, since anthropogenic sources of nitrogen are generally enriched in the heavy isotope when compared to natural ones. Actually, several studies have shown the utility of stable isotopes in trophic interactions studies, highlighting the relevance of the stable nitrogen isotope a reliable bioindicator to detect patterns of anthropogenic nitrogen contamination in several ecosystems. This is because different N sources have different relative abundances of $\delta^{15}N$, it is possible to evaluate the contribution of nitrogen sources to the local N pool, at least qualitative. For example, nitrate (NO_3^-) from human and animal waste is enriched in $\delta^{15}N$, with isotopic compositions of 10‰ to 22‰, whereas nitrate in synthetic fertilizer, which is fixed from atmospheric N, generally has much lower $\delta^{15}N$, with values ranging from −3‰ to 3‰.

Baeta et al. (2017), for example, correlated changes in the natural abundance of $\delta^{15}N$ in fish larvae to anthropogenic N inputs from both waste water and agricultural sources. Inferring, however, which specific nitrogen sources contribute to elevate $\delta^{15}N$ might be difficult in ecosystems that contain a variety of different pollution sources that contribute to the nitrogen load to these systems, and additional studies would be necessary. Stable isotopes used in conjunction with other techniques, such as compound-specific stable isotope analysis (CSIA) of natural isotopic abundance, may greatly enhance the evaluation of sources and transformation processes of pollutants, relevant to environmental decision makers. By increasing greatly the resolution of SIA, advances in CSIA will provide new opportunities and applications for stable isotopes studies throughout ecology.

Applying Stable Isotopes to Examine Food-Web Structure: Mixing Models, Ecological Networks, and Limitations

Stable isotopes have been also used to quantify the contribution of different food sources in the diet of an organism. Stable isotope mixing models (SIMM) have been proposed to reconstruct consumer diets, by assessing the proportions of food sources assimilated by a consumer (not just ingested by the consumer). Ecological researchers use mixing models to achieve as accurate as possible depiction of the trophic links in a food web. However, a good understanding of the study system is essential before using these models to determine the diet of consumer, because researchers need to know some aspects of an animal's diet a priori such as the identification of food sources that might be in the diet of the study species. Once the dietary sources have been identified, SIMM offer an excellent way to quantify diet, by estimating the proportional contribution of each dietary source to the diet of a consumer. Over the last two decades, a number of isotope mixing models have been proposed. These techniques range from simple, qualitative inferences based on the isotopic niche, to Bayesian mixing models that can be used to characterize food web structure at multiple hierarchical levels. In fact, current models have become more sophisticated, estimating probability distributions of source contributions, and incorporating complexities such as variability in isotope signatures, discrimination factors, hierarchical variance structure, covariates, and concentration dependence.

Another promising approach is the use of stable isotopes in the modeling of food web network structure. By combining stable isotope techniques with quantitative modeling approaches, it will be possible to gain additional insight in the structure of food webs (Fig. 9). Understanding how ecosystems react to and recover from perturbations is a fundamental goal of ecology. The increasing interest toward ecosystem status and performance, and the need to approach complex environmental problems, stimulate the application of tools for whole-system assessment. With the advent of quantitative ecosystem modeling tools, food web analysis is leading toward a better understanding of food web structure and the design of better management strategies for conservation. Accordingly, incorporation of diet information derived from SIMM to mass-balance models (that represent a "snapshot" of the trophic flows in the ecosystem and can be used to describe at least part of the reality) results in more constrained food web models.

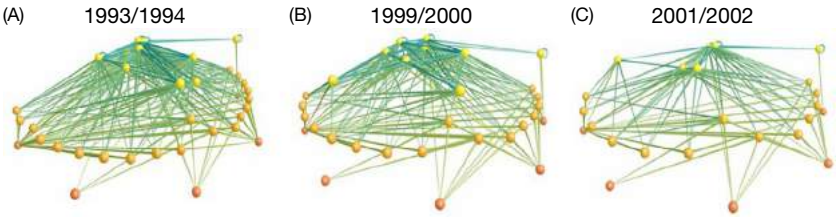


Fig. 9 3D representation of three food webs from the *Zostera* site in the Mondego estuary, Portugal. (A–C) *Zostera* site in 1993/1994, 1999/2000 and 2001/2002 respectively. Images were produced with FoodWeb3D written by R.J. Williams, Pacific Ecoinformatics and Computational Ecology Laboratory. The different *colored dots* represent functional groups from different trophic levels: *red* = primary producers, *orange* = primary consumers, and *yellow* = secondary consumers. The *light* and *dark gray* links represent feeding links. Adapted from Baeta, A., Niquil, N., Marques, J.C., and Patrício, J. (2011). Modelling the effects of eutrophication, mitigation measures and an extreme flood event on estuarine benthic food webs. *Ecological Modeling* **222**, 1209–1221.

Despite the multiple advantages of the natural isotope tracing approach, this technique, as with all tools, bears some notable limitations. In food web studies, the use of stable isotopes is based on an understanding of how the isotopic composition of animal diet and tissues are related. There are two main aspects to this. First, there is substantial unexplained variation in isotope composition of a given tissue, across taxa, and across tissue types. Therefore, the use of appropriate consumer-diet discrimination is crucial. The use of SIMM for diet reconstruction is characterized by an overreliance on literature values for key parameters. The use of constant discrimination factors declines the predictive power of models and inferences become limited—the models may illustrate general patterns, but not reliable and precise diets reconstruction and trophic positions. The second relates to tissue turnover time: an animal tissue does not immediately reflect the isotopic composition of its diet, but rather integrates over some period of time. Isotope turnover is a function of somatic growth and metabolism. Somatic growth-based isotope turnover occurs as the animal adds new tissue, diluting the pool of tissue derived from the diet in its previous location. Metabolic isotope turnover occurs as tissue is broken down and new tissue is synthesized. In general, tissues with fast growth and high metabolism have quicker turnover rates. For example, isotopic signatures in bone may integrate diets over the course of a year, hair over a few months, muscle over several weeks, and blood plasma over several days. Understanding the dynamics of isotopic turnover may allow ecologists to detect seasonal, or even episodic, changes in an animal's diet. Accordingly, further experimental investigation (laboratory and field) to quantify accurate values of consumer-resource discrimination and turnover rates of tissues is essential to inform scientifically sound food web studies using SIA.

References

Baeta A, Vieira LR, Lírio AV, Canhoto C, Marques JC, and Guilhermino L (2017) Use of stable isotope ratios of fish larvae as indicator to assess diets and patterns of anthropogenic nitrogen pollution in estuarine ecosystems. *Ecological Indicators* 83: 112–121.https://doi.org/10.1016/j.ecolind.2017.07.062.

Bird CS, Veríssimo A, and Trueman CN (2018) A global perspective on the trophic geography of sharks. *Nature Ecology & Evolution* 2: 299–305. https://doi.org/10.1038/s41559-017-0432-z.

DeNiro M and Epstein S (1976) You are what you eat (plus a few per mil): The carbon isotope cycle in food chains. *Geological Society of America* 8: 834–835. Abstracts with Programs.

Further Reading

Aleinov I and Schmidt GA (2006) Water isotopes in the GISS ModelE land surface scheme. *Global and Planetary Change* 51: 108–120. https://doi.org/10.1016/j.gloplacha.2005.12.010.

Baeta A, Valiela I, Rossi F, Pinto R, Richard P, Niquil N, and Marques JC (2009) Eutrophication and trophic structure in response to the presence of the eelgrass *Zostera noltii*. *Marine Biology* 156: 2107–2120. https://doi.org/10.1007/s00227-009-1241-y.

Baeta A, Niquil N, Marques JC, and Patrício J (2011) Modelling the effects of eutrophication, mitigation measures and an extreme flood event on estuarine benthic food webs. *Ecological Modelling* 222: 1209–1221. https://doi.org/10.1016/j.ecolmodel.2010.12.010.

Boecklen WJ, Yarnes CT, Cook BA, and James AC (2011) On the use of stable isotopes in trophic ecology. *Annual Review of Ecology, Evolution, and Systematics* 42: 411–440. https://doi.org/10.1146/annurev-ecolsys-102209-144726.

Bruckner MZ (2007) *Stable isotope primer and some hydrological applications, microbial life educational resources*. Retrieved with permission from https://serc.carleton.edu/microbelife/research_methods/environ_sampling/stableisotopes.html (30th April, 2018).

Campbell L, Venkiteswaran J, Bond AL (2014) Common mistakes in stable isotope terminology and phraseology. https://figshare.com/articles/Common_Mistakes_in_Stable_Isotope_Terminology_and_Phraseology/1150337.

Caut S, Angulo E, and Courchamp F (2009) Variation in discrimination factors ($\Delta^{15}N$ and $\Delta^{13}C$): The effect of diet isotopic values and applications for diet reconstruction. *Journal of Applied Ecology* 46: 443–453. https://doi.org/10.1111/j.1365-2664.2009.01620.x.

Chang R and Goldsby K (2014) *Chemistry*, 11th edn. New York: McGraw-Hill Education.

Eggers T and Jones TH (2000) You are what you eat. . .or are you? *Trends in Ecology and Evolution* 15(2): 65–266.

Fry B (2006) *Stable isotope ecology*. New York: Springer.

Hoffman JC (2016) Tracing the origins, migrations, and other movements of fishes using stable isotopes. In: Morais P and Daverat F (eds.) *An introduction to fish migration*, pp. 169–196. Boca Raton, FL: CRC Press.

Kendall C, Elliott EM, and Wankel SD (2007) Tracing anthropogenic inputs of nitrogen to ecosystems, chapter 12. In: Michener RH and Lajtha K (eds.) *Stable isotopes in ecology and environmental science*, 2nd edn, pp. 375–449. Pittsburg, PN: Blackwell Publishing.

Layman CA, Araujo MS, Boucek R, Hammerschlag-Peyer CM, Harrison E, Zachary RJ, Matich P, Rosenblatt AE, Vaudo JJ, Yeager LA, Post DM, and Bearhop S (2012) Applying stable isotopes to examine food-web structure: An overview of analytical tools. *Biological Reviews* 87: 545–562. https://doi.org/10.1111/j.1469-185X.2011.00208.x.

Lotze HK, Lenihan HS, Bourque BJ, Bradbury RH, Cooke RG, Kay MC, Kidwell SM, Kirby MX, Peterson CH, and Jackson JBC (2006) Depletion, degradation, and recovery potential of estuaries and coastal seas. *Science* 312: 1806–1809.

Michener R and Lajtha K (eds.) (2007) *Frontmatter, in stable isotopes in ecology and environmental science*, 2nd edn Oxford, UK: Blackwell Publishing Ltd. https://doi.org/10.1002/9780470691854.fmatter.

Neubauer P and Jensen OP (2015) Bayesian estimation of predator diet composition from fatty acids and stable isotopes. *PeerJ* 3: e920, https://doi.org/10.7717/peerj.920.

Post DM (2002) Using stable isotopes to estimate trophic position: models, methods, and assumptions. *Ecology* 83: 703–718. https://doi.org/10.1890/0012-9658(2002)083[0703,USITET]2.0.CO;2.

Sharp Z (2017) *Principles of stableisotope geochemistry*, 2nd edn. https://doi.org/10.5072/FK2GB24S9F.

Smith JA, Mazumder D, Suthers IM, and Taylor MD (2013) To fit or not to fit: Evaluating stable isotope mixing models using simulated mixing polygons. *Methods in Ecology and Evolution* 4: 612–618. https://doi.org/10.1111/2041-210X.12048.

Vander Zanden MJ, Clayton MK, Moody EK, Solomon CT, and Weidel BC (2015) Stable isotope turnover and half-life in animal tissues: A literature synthesis. *PLoS One* 10(1): e0116182. https://doi.org/10.1371/journal.pone.0116182.

West JB, Bowen GJ, Cerling TE, and Ehleringer JR (2006) Stable isotopes as one of nature's ecological recorders. *Trends in Ecology and Evolution* 21: 408–414. https://doi.org/10.1016/j.tree.2006.04.002.

West JB, Bowen GJ, Dawson TE, and Tu KP (2010) *Isoscapes: Understanding movement, pattern, and process on earth through isotope mapping*. The Netherlands: Springer.
White WM (2014) *Isotope geochemistry*. Wiley-Blackwell.
Yoon I, Williams RJ, Levine E, Yoon S, Dunne JA, and Martinez ND (2004) In: *Webs on the Web (WoW): 3D visualization of ecological networks on the WWW for collaborative research and education*, vol. 5295, *Proceedings of the IS&T/SPIE Symposium on Electronic Imaging, Visualization and Data Analysis,* 124–132.

Relevant Websites

http://nuclearconnect.org/know-nuclear/science/isotopes.
https://www.giss.nasa.gov/research/briefs/schmidt_03/.
http://nobelprize.org/chemistry/laureates/1921.

Succession

JM Pandolfi, University of Queensland, Brisbane, QLD, Australia

Introduction

This article explores the notion of ecological succession, one of ecology's oldest and most controversial concepts. Simply stated, succession is the orderly changes in communities of a specific place over a period of time. More formally, it is the orderly progression of directional community development that results from modification of the physical environment by the community and culminates in an ecosystem in which maximum biomass and interactions among component organisms are maintained. Many workers relate succession to disturbance (fires, hurricanes, drought, etc.), such that succession refers to changes in a community after a perturbation opens up ecospace. Most of the earlier work on succession (dating back to 1899!) described the sequential invasion of species invading a particular site; later studies deal with a whole range of community characteristics from both empirical and theoretical standpoints (Table 1). Most of the work on succession has been dominated by forest communities, but successional studies also exist for a number of other ecosystems including benthic and pelagic marine ecosystems.

Different Kinds of Succession

Two kinds of succession have been proposed. `Autogenic' succession refers to sequential changes in the community that are brought about by the influence and activities of the organisms themselves upon the environment or habitat. `Allogenic' succession occurs to communities through the influence of factors external to the organisms themselves. The utility of allogenic succession may be limited because it may be too broadly applicable to any temporally repetitive change. Most of the debates surrounding succession, and studies which provide detailed examples are concerned with autogenic succession.

`Primary' succession refers to the colonization of sites where a severe disturbance has left no trace of the preexisting community, or where entirely new ecospace has been created. These might include a terrestrial or submarine lava flow, glacial recession, or the movement of submarine or subaerial sand dunes. Harsh environments usually characterize the earliest phases of primary succession, and communities tend to develop more slowly than with more minor disturbances to the ecosystem. `Secondary' succession refers to colonization of sites which did have a previous community established, but a perturbation has removed a portion of it or all of it over a limited area. Here again, much of the debate surrounding succession and most of its examples come from secondary succession. However, most unassailable instances of succession are examples of primary succession.

Expected Community Trends of Succession

The expected community trends of succession are explicit and detailed (Table 1). They rely on an early-successional community that ultimately leads to the `climax' community, or the most stable state for the successional sequence. This climax community is maintained until an extrinsic disturbance resets the successional clock. The attributes are grouped according to community energetics, community structure, life-history characteristics, cycling of nutrients, selection pressure, and overall homeostasis. These trends follow from the notion that the climax community is a steady state maintained by internal feedback control mechanisms. But they break down if it is not, and there is little evidence so far that they are.

In forests there are a variety of aspects of production ecology which vary over the successional sequence including forest floor depth, chemistry and content of decaying wood, nutrient availability, leaf area, carbon allocation, species composition and their nutritional adaptations, the relative roles of trees and minor vegetation, the relative importance of geochemical, biogeochemical and internal cycles, rate of soil organic matter accumulation, and total ecosystem organic matter. Early-successional forest species generally are shade intolerant, require a mineral seedbed, may be fast-growing, can generally utilize nutrients that occur in newly formed gaps, may be nitrogen-fixing, are tolerant of microhabitats in newly disturbed settings, and are superior competitors in such settings. Late-successional forest species generally are shade tolerant where they are able to grow slowly, may require a more organic seedbed, and are generally poorly adapted to newly disturbed settings where they are poor competitors. Nutrients are generally gained from litter decomposition and mycorrhizae.

Models of Succession

Three alternative models have been proposed for determining the sequence of species that might occur after a perturbation. The first is the facilitation model in which only certain species (`early successional') are able to arrive immediately following a

 doi:10.1016/B978-0-444-63768-0.00547-3

Table 1 Predicted changes that occur in major structural and functional characteristics of a developing ecosystem during ecological succession

Ecosystem attributes	*Developmental stages*	*Mature stages*
Community energetics		
1. Gross production/community respiration (*P*/*R* ratio)	Greater or less than 1	Approaches 1
2. Gross production/standing crop biomass (*P*/*B* ratio)	High	Low
3. Biomass supported/unit energy flow (*B*/*E* ratio)	Low	High
4. Net community production (yield)	High	Low
5. Food chains	Linear, predominantly grazing	Weblike, predominantly detritus
Community structure		
6. Total organic matter	Small	Large
7. Inorganic nutrients	Extrabiotic	Intrabiotic
8. Species diversity – variety component	Low	High
9. Species diversity – equitability component	Low	High
10. Biochemical diversity	Low	High
11. Stratification and spatial heterogeneity (pattern diversity)	Poorly organized	Well-organized
Life history		
12. Niche specialization	Broad	Narrow
13. Size of organism	Small	Large
14. Life cycles	Short, simple	Long, complex
Nutrient cycling		
15. Mineral cycles	Open	Closed
16. Nutrient exchange rate, between organisms and environment	Rapid	Slow
17. Role of detritus in nutrient regeneration	Unimportant	Important
Selection pressure		
18. Growth form	For rapid growth (`r-selection')	For feedback control (`*K*-selection')
19. Production	Quantity	Quality
Overall homeostasis		
20. Internal symbiosis	Undeveloped	Developed
21. Nutrient conservation	Poor	Good
22. Stability (resistance to external perturbations)	Poor	Good
23. Entropy	High	Low
24. Information	Low	High

Analysis of these parameters provides empirical tests of the expected development of ecosystems during succession.

disturbance. The most important factor governing succession in the facilitation model is the degree to which early-successional species alter the environment for later-successional species. Later species can only colonize after these early-successional species have sufficiently modified the environmental conditions. For example, in forest communities the amount of net primary production of the plant community will influence its ability to alter the site. This model implies a high degree of organization in ecological communities. However, it may still operate in communities where such organization is not apparent, for example, in heterotrophic successions or primary successions.

The second is the tolerance model in which the sequence of species to inhabit an ecosystem after a disturbance is determined exclusively by their life-history characteristics. The species that occur are simply those that are most efficient in exploiting resources in a given time or place. Thus, the success of later species is unrelated to the species composition of the earlier community. Two circumstances have been forwarded where this model of succession might be most prevalent. The first is with certain groups of animals, such as large vertebrates or insects that display a high degree of social integration, that have evolved a high degree of independence from both biotic and physical environments. The second is in situations where populations are limited more by resources than natural enemies or environmental stress. In the tolerance model, life-history characteristics such as the relative longevity, growth form, reproductive mode, and size of the early versus later species becomes paramount. A further elaboration of this model is given below in the section titled `The role of competition and life history'.

The third is the inhibition model in which early-successional species inhibit the success and growth of later-successional species. Those species that happen to occupy the site first maintain their membership in the community until their populations are damaged or die, only then enabling later species to invade. As such, the species composition gradually shifts to species with greater longevities. No special mechanism need be invoked for the inhibition model of succession.

Most models of succession have been criticized. Some authors treat them as purely descriptive and, like the entire successional process, a particular manifestation of the outcome of interacting individual life histories. In fact, one of the original authors of these three models has recently acknowledged that directional species replacement is only one change that might occur after a disturbance, and is often not the most likely change.

Some authors have taken these three models and related them to the search for ecological assembly rules during succession and compared them with a fourth alternative, the random colonization model, where chance survival of different species initiates the successional sequence; subsequent associations occur as a result of subsequent random colonization by new species. The search for assembly rules in ecology has a rich and controversial history that parallels that in succession. In fact, successional changes in community composition might be viewed as a special case of an overall search for deterministic processes in the maintenance and diversity of ecological communities.

Factors Influencing Succession

There are a number of factors that can influence the way in which community succession proceeds, if at all. Probably the most important of all are differences in life-history traits such as dispersal, ability to establish in open sites, shade tolerance, etc. The tradeoff between the availability of seeds and propagules of later-successional species versus the degree to which the early-successional community is resistant to invasion will influence the degree to which the former can increase at the successional site. Invasion success itself is dependent upon a large number of factors including, but not limited to, changes in microclimate or microhabitats, the composition of the seedbed, diseases, predators, and competition. Superimposed upon all of these are factors external to the organisms themselves, such as precipitation, temperature, salinity, and disturbance regime which can serve to reset the successional sequence.

The Role of Competition and Life History

Patterns from succession are numerous and varied, and have been mainly reported from the population and ecosystem level. However, a generalized theory of succession has been proposed for plants based on individual organisms. It is based on a model with three features: (1) the demographic history (birth, growth, death) of each individual is tracked through time; (2) every individual is assigned a suite of species-specific life-history traits (e.g., size, maximum age, and shade tolerance); and (3) competition and resource availability (light) is modeled explicitly through varying each individual's growth and mortality based on its degree of shade tolerance and light availability. The model successfully produces a significant portion of the variability in successional patterns using the mechanism of competition. Again, the importance of life-history traits among interacting plants is manifested.

Two-species simulations from the model were classified into five groupings based on the temporal patterns of species divergence (**Fig. 1**). The first was classical successional replacement where one species was able to dominate early in succession and the other species later, and is due to the correlation among life-history parameters of the two species. Divergence occurs when competitive ability starts out the same, but then one species outcompetes the other whereas convergence is the opposite: species are unequal in competitive ability early in succession but this equals out later in succession. Total suppression occurs when one species has the competitive advantage throughout the entire successional development. Finally, pseudocyclic replacement occurs when a temporary period of dominance or co-dominance by the initial cohort of species 2 results during the period between the senescence of the initial cohort of species 1 and the maximum lifespan of species 2. Thus early succession is characterized by the establishment of both species, but the initial cohort of species 2 persists longer than that of species 1. Species 2 is later replaced by the competitively superior species 1. The outcomes of the simulations depend more upon the equality of the life-history characteristics among species as opposed to the absolute values of the characteristics themselves; thus outcomes are governed by the qualitative relationship between the species with respect to their competitive abilities through time. Results of the simulations were similar to the two-species models when more than two species were considered and when multiple resources were considered. To summarize this important model to the concept of succession, the summary of the authors is quoted:

> An approach based on competition among individual plants is presented as an explanation for species replacements during plant succession. Inverse correlations among life-history and physiological traits that confer competitive ability under different environmental conditions are shown to be sufficient to produce successional replacements but not sufficient for understanding the complex variety of successional patterns unless they are applied at the individual rather than at the population level or higher. With models based on competition among individual plants, various combinations of life-history and physiological traits can produce the great variety of population dynamics found in natural successions. The classic successional pattern of species replacement results from a particular structure of correlations among life-history sand physiological characteristics. Atypical patterns of succession result when this correlation structure is altered. Both primary and secondary succession are modeled as nonequilibrium processes capable of interacting with disturbances to produce steady-state communities whose properties depend on abiotic conditions, such as temperature and resource levels, and on the type and frequency of disturbances. (Huston and Smith, 1987, p. 193)

Before leaving the subject of competition, it must be noted that there may well be a role for positive interactions among species comprising a community. There is increasing evidence that direct positive interaction, or facilitations, among species can play a vital role in ecological communities. While relatively unexplored in successional studies, the role of facilitations might provide an important context for testing the importance of successional patterns and processes in the maintenance of communities.

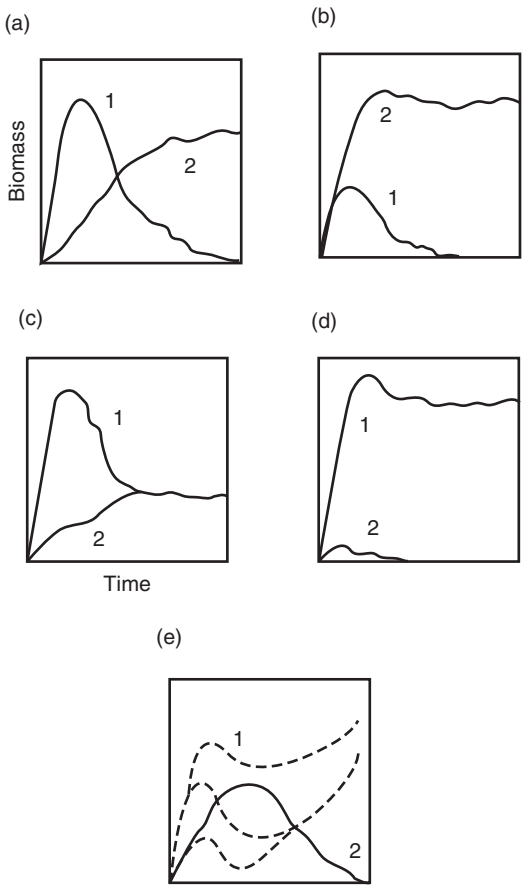


Fig. 1 General patterns of species interactions in a two-species succession. Competitive dominance occurs as a result of differing life-history traits. (a) Classical successional replacement that results from an inverse relationship between the two species in the attributes that confer early and late competitive advantage; (b) divergence that results when competitive abilities are equal early on, followed by a competitive advantage by one of the species; (c) convergence, the inverse of (b), that results when one species has a competitive advantage that equals out during the succession; (d) total suppression that results from one species having a competitive advantage over the other throughout succession; and (e) pseudocyclic replacement that results from alternating periods of dominance or co-dominance between the two species. From Huston MH and Smith T (1987) Plant succession: Life history and competition. *American Naturalist* 130: 168–198.

Is Succession Real?

Many models of succession are based on the idea that species that occur early in the successional sequence facilitate their own replacement by altering the environment to be more favorable to later species. Of course, this is anathema to natural selection – no species would `intentionally' create more favorable circumstances for any other species than itself (the `evolutionary paradox' of succession). For this reason the whole idea of succession has been questioned, and some authors consider it to be no more than the result of the passage of time, with no real mechanistic basis.

Many of the factors that might influence succession have been called into question, such that many authors believe that it is not a real phenomenon, or if it is, is unimportant in the maintenance of species diversity. Of critical importance is the degree to which communities act in changing the environment. Some reviews of the idea of succession conclude that most purported examples of succession do not fit the predicted trends: changes in structural and functional properties are not consistently associated with changes in species composition; successional stages are not consistently unidirectional; and effects of earlier species might just as easily delay or prevent as facilitate later-successional replacement.

Workers critical of successional theory have also put forward alternatives. One of these relates purported successional changes to consequences of differential growth, survival, and colonizing ability of species adapted to growth at different points on environmental gradients. Differences in taxonomic composition among successional stages are dependent upon interspecific competition while structural and functional changes in the community results in correlations in plants among size, longevity, and growth rate. In fact, successional changes may be exclusively governed by the presence of slower-growing later-successional species during the early phases of succession.

Much of the debate stems from the classic divergence in opinion about the influence of factors that result in the way in which species are maintained within ecological communities. On the one hand there may only be a loose association of species who

happen to inhabit the same habitat at the same time. In this model, species do not `feel' each other or interact much, or if they do, the effects of these interactions are minor in comparison with their shared environmental tolerances. In the extreme view, communities are just haphazard associations of species and any spatial or temporal continuity in the composition of a community is either minimal, or determined solely by environmental parameters.

A completely different view holds that biological communities are tightly integrated units composed of species that not only respond together to environmental changes, but also maintain linkages among each other that result in continuity of occurrence over broad spatial and temporal scales. Some caricatures of this model have even referred to the community as a `super-organism' with emergent properties that extend beyond individual species. While this view is probably overstated, there is a large amount of evidence that species interactions within communities result in spatial and temporal continuity of species assemblages within habitats. It is these interactions and other biological properties that have led many authors to believe that succession is a very plausible model for understanding changes in species composition over time.

Some Examples of Succession

Regardless of the continuing debates on whether there is any mechanistic basis for explaining directional patterns in species occurrences, excellent examples of the latter occur in nature. One of the best studies of primary succession was a test of the mechanisms involved in succession following deglaciation at Glacier Bay, Alaska. The main conclusion from this work was that no single factor or mechanism can totally account for primary succession. Plant life-history traits, such as seed size, growth rate, age at first reproduction, maximum height, and longevity played a fundamental role, but competition among species, and facilitation by addition of soil N and organic matter were also important. Although the relative importance of these factors was influenced by resource availability and environmental stress, they varied predictably through the successional sequence. The authors concluded that (1) life-history traits and availability of propagules determine the pattern of succession, (2) the mechanism for changes in species abundance patterns that accompanies succession is competition, and (3) the rate of change and nature of the climax community depends upon initial site conditions and the role of facilitation.

Perhaps one of the most famous and long-standing successional stories comes from the Krakatau Islands of Indonesia. The volcanic eruptions between May and August 1883 resulted in the almost complete `sterilization' of the islands. A classic primary successional sequence has been documented for the main volcanic island of Rakata which continues uninterrupted, whereas the other islands display characteristics of secondary succession (**Fig. 2**). On Rakata, a number of different successional pathways are being expressed (**Fig. 2a**). A similar diversity of succession characterizes the other islands, but eruptive activity of Anak Krakatau – a new volcano that was established in the center of the group between 1927 and 1930 – disrupted the vegetation of some of the other islands such as Serung and R. Kecil. There, varying disturbance events have reset the successional sequence (**Fig. 2b**).

Differences in vegetation and in the curves of species colonization among the Krakatau Islands have been attributed mainly to intermittent disturbances of varying intensities from Anak Krakatau. The importance of succession in viewing the history of terrestrial ecosystems of the Krakatau Islands is summed up by Whittaker *et al.*

> Neither the pattern of change in the flora, nor that for some groups of fauna, e.g., the birds and butterflies… can be understood without reference to vegetation succession and the key period in the 1920s when the savanna vegetation of the interiors gave way to forest. … a successional model of island recolonization is required, involving evaluation of habitat and dispersal mechanisms. (Whittaker *et al.*, 1989, p. 103)

Succession in the Fossil Record

One of the great limitations of assessing the frequency or validity of the processes of succession is that the complete successional sequence might take longer than the period of time over which most ecosystems are studied within scientific research programs. Since succession is by its very nature a process that occurs over time, it is not surprising that paleontologists have investigated the degree to which it can be understood from evidence in the fossil record. In contrast to modern ecologists working on mainly terrestrial forest communities, paleoecologists have focussed more on marine systems, studying succession in fossil coral reefs, marine hardgrounds, and some sandy substrates. A critical debate for paleoecologists is the degree to which species replacements in communities are ecological versus evolutionary phenomena.

Palaeontologists have used ancient reefs as model ecosystems in the study of succession. Four major phases of succession in eight ancient reefs ranging in age from the Early Ordovician (488 Ma) to the Late Cretaceous (65.5 Ma) have been established: stabilization, colonization, diversification, and domination. The stabilization stage involves the initial colonization of the seafloor that results in the establishment of a firm substrate. The colonization stage is characterized by encrusters and frame-builders, those organisms capable of colonizing a hard substrate and that begin to build three-dimensional structures above the sediment–water interface. Some authors consider the first two of these stages to be the same, equivalent to the pioneering stage of ecologists studying succession in living forest communities. The third stage in the succession of fossil reefs is referred to as the diversification stage, where maximum diversity of the reef community is developed. Diversity then decreases as a `domination' stage sets in whereby a single functional entity characterizes the reef community. As the communities ascend through the first three stages of succession, species diversity, degree of stratification and pattern diversity, niche specialization, and the complexity of food chains all increase. These successional sequences typically occur through significant intervals of time and are the result of the gradual

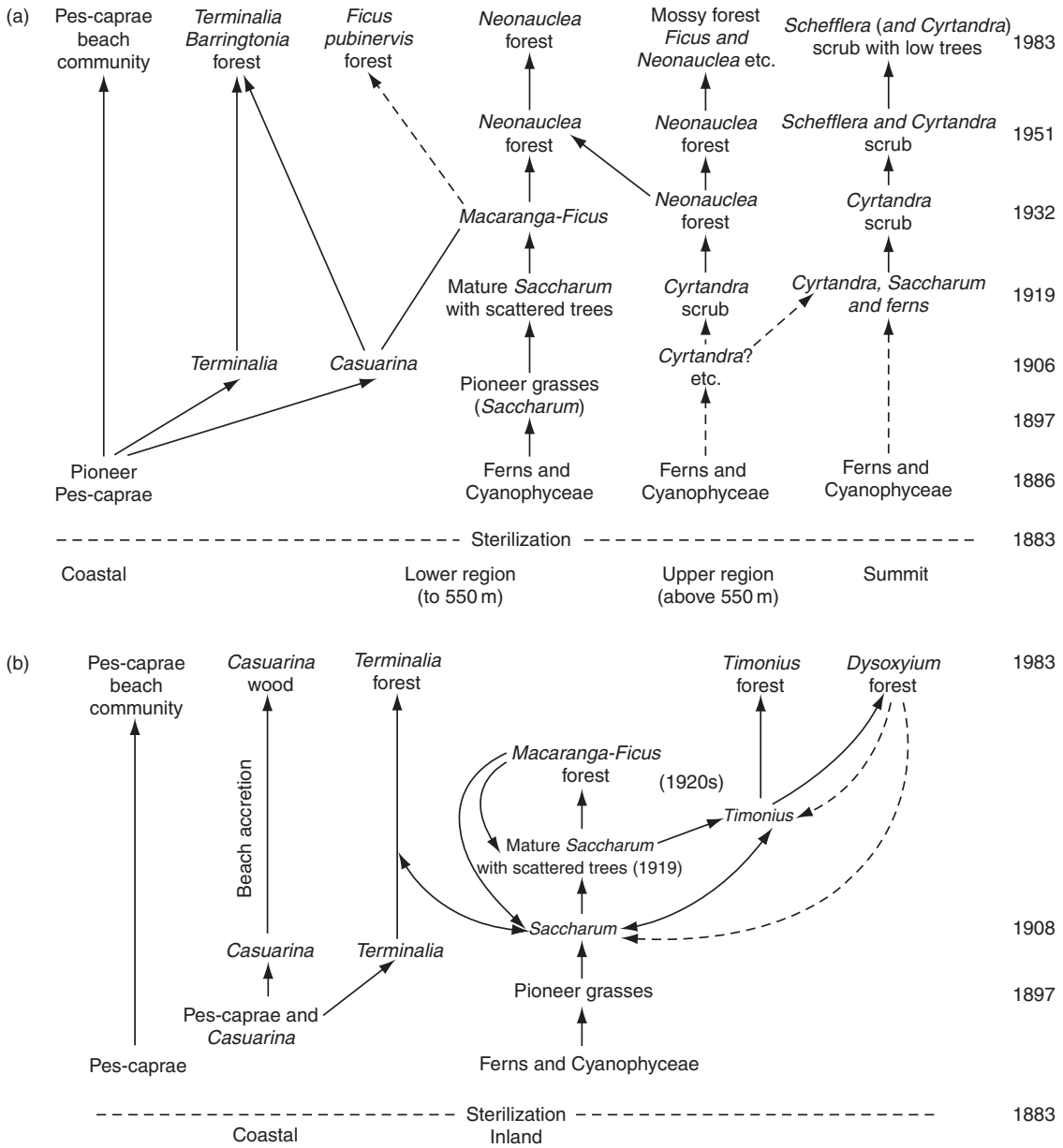


Fig. 2 Principal successional pathways on (a) Rakata and (b) Sertung and R. Kecil plant communities. Rakata has not experienced any volcanic disturbance for over 100 years, whereas forest development on Sertung and R. Kecil has been interrupted by volcanic activity several times since 1930 by Anak Krakatoa. Eruptive activity changed late-successional forest communities to early-successional or `pioneer' communities. From Whittaker RJ, Bush MB, and Richards K (1989) Plant recolonization and vegetation succession on the Krakatau Islands, Indonesia. *Ecological Monographs* 59: 59–123.

alteration of the submarine substratum by individual species, and the elaboration of energy-flow pathways as the community proceeds through time. Shorter-term successional episodes have also been observed where occasional environmental perturbations destroy the preexisting community. Here, subsequent community development involves a rapid biologically induced sequence of communities that culminate in the climax community that was established prior to the perturbation. Just as in the longer-term successional sequences, increases in major functional and structural attributes of the community accompany these shorter-term successional sequences.

There are, of course, a number of potential problems associated with the identification of succession in the fossil record. In the model above, an increase in diversity may come about just as easily from an increase in the zonation of the reef as to intrinsic biological changes. In the domination phase, the accompanying decrease in diversity may reflect a community that caps a reef after it dies, as opposed to a decrease during the life of the reef. However, in the fossil record of the Quaternary period (the past 1.8 Myr) these difficulties can often be overcome due to the precise age dating and sea-level history of the reefs.

In Kenya, `obligatory succession' in the Pleistocene consisted of an early assemblage of sediment-tolerant corals (dominated by massive or doming corals) that was replaced by predominantly branching and platy/encrusting corals. This kind of succession was mainly confined to the earliest portions of reef development on a bare substrate. Patterns through longer successional phases were varied, but under certain conditions massive or domed-shape corals might replace the branching assemblages. However, clearly defined zones were rare – most temporal changes in reef species associations were random or unstructured.

Applications to Restoration Ecology

The concepts of succession and community assembly have recently been taken up by the field of `restoration ecology'. The nature of community assembly and the degree to which communities can be expected to return on a path generally directed to their original states (though return to their exact previous state is probably not possible nor desirable) is of primary interest to restoration ecologists interested in mitigating and reversing the effects of human-induced habitat degradation. Because of this central concern and the fundamental questions which each theory provides, it is an important exercise to understand the similarities and differences between community succession and community assembly. Both concepts recognize the importance of historical events in shaping community composition, which develops over time toward a final state, and both recognize competition as a major driving force.

Single or multiple stable states are predicted for both concepts, but the way in which species are maintained in such states is very different. We have already discussed the idea of a climax community – the final stable state of a successional sequence. However, where succession occurs, multiple stable states can be maintained across an ecosystem by disturbances that keep the community in a variety of successional states, so the `climax' is never achieved at any one site or time. Where these disturbances are generated by the characteristics of the community itself (herbivory or fire) the community may be kept at an `arrested' successional state. Arrested succession might also arise where early-successional species maintain their advantage over later-successional species for extremely long periods of time. `Cyclic' succession occurs when there is no one stable successional state and all of the stages yield to others within the successional cycle. Stochastic differences in the arrival and colonization of late-successional species might also result in variability among ecosystems in their community composition.

Community assembly refers to the process by which species colonize, interact with other species and establish a community. This assembly may or may not be affected by `assembly rules' which predict community composition on the basis of a few key attributes such as size of the species pool, the abiotic environment, or interspecific interactions. Community assembly may also result in either a single community type or multiple stable states. Where there is a strong match between environment and community, then community composition should converge towards a single state where similar environmental conditions hold, regardless of the historical order in which the species invade. But multiple stable equilibria have also been observed with varying historical sequence of species invasions, so that different timing of introduction of species colonization leads to different species composition.

Regardless of whether species are maintained through a successional sequence, or the product of their sequence of invasion into an ecosystem, an understanding of the processes which bring about various community states is fundamental to those wishing to move the system in one direction or another. Even where such processes are particularly opaque, an understanding of the presence of particular states may help in providing goals for management and restoration. Some predictions about the likelihood of the existence of a single or multiple stable states have been put forward. While these need further testing, a single stable community state has been predicted to arise in ecosystems with small species pools, high levels of dispersal, low productivity, and high rates of disturbance; whereas multiple stable states should be more apparent in ecosystems with large species pools, low levels of dispersal, high productivity, and low rates of disturbance. Thus, an integrated approach to understanding the processes underlying the historical trajectory of ecosystems, coupled with an in-depth knowledge of life-history variability, and a few key attributes of the ecosystem can go far in our understanding of how communities are likely to respond in the future to environmental changes today.

Further Reading

Chapin, F.S., Walker, L.R., Fastie, C.L., Sharman, L.C., 1994. Mechanisms of primary succession following deglaciation at Glacier Bay, Alaska. Ecological Monographs 64, 149–175.

Clements, F.R., 1916. Plant Succession: An Analysis of the Development of Vegetation. Washington, DC: Carnegie Institute.

Connell, J.H., Slayter, R.O., 1977. Mechanisms of succession in natural communities and their role in community stability and organization. American Naturalist 111, 1119–1144.

Cowles, H.C., 1899. The ecological relations of the vegetation on the sand dunes of Lake Michigan. Botanical Gazette 27, 95–117. 167–202; 281–308; 361–391.

Crame, J.A., 1980. Succession and diversity in the Pleistocene coral reefs of the Kenya coast. Palaeontology 23, 1–37.

Drury, W.H., Nisbet, I.C.T., 1973. Succession. Journal of the Arnold Arboretum 54, 331–368.

Egler, F.E., 1954. Vegetation science concepts. I. Initial floristic composition, a factor in old-field development. Vegetatio 4, 412–417.

Gleason, H.A., 1926. The individualistic concept of the plant association. Bulletin of the Torrey Botanical Club 53, 7–26.

Colonization, Succession and Stability. In: Gray, A.J., Crawley, M.J., Edwards, P.J. (Eds.), The 26th British Ecological Society Symposium held jointly with the Linnean Society of London. Oxford: Blackwell.

Huston, M.H., Smith, T., 1987. Plant succession: Life history and competition. American Naturalist 130, 168–198.

Lawton, J.H., 1987. Are there assembly rules for successional communities? In: Gray, A.J., Crawley, M.J., Edwards, P.J. (Eds.), Colonization, Succession and Stability: The 26th British Ecological Society Symposium Held Jointly with the Linnean Society of London. Oxford: Blackwell, pp. 225–244.
McCook, L.J., 1994. Understanding ecological community succession: Causal models and theories, a review. Vegetatio 110, 115–147.
Odum, E.P., 1969. The strategy of ecosystem development. Science 164, 262–270.
Walker, K.R., Alberstadt, L.P., 1975. Ecological succession as an aspect of structure in fossil communities. Paleobiology 1, 238–257.
Whittaker, R.J., Bush, M.B., Richards, K., 1989. Plant recolonization and vegetation succession on the Krakatau Islands, Indonesia. Ecological Monographs 59, 59–123.
Young, T.P., Chase, J.M., Huddleston, R.T., 2001. Community succession and assembly. Ecological Restoration 19, 5–18.

Suspension Feeders☆

Brian T Hentschel, San Diego State University, San Diego, CA, United States
Jeff Shimeta, University of Melbourne, Parkville, VIC, Australia

What Is Suspension Feeding?

Suspension feeding is the capture and ingestion of food particles that are suspended in water. These particles can include phytoplankton, zooplankton, bacteria, and detritus. Some suspension feeders are primarily grazers of planktonic algae, while others are carnivores, and some that feed at the sediment–water interface are primarily detritivores. Some suspension feeders are largely nonselective omnivores, whereas others display strong preferences for certain particles according to size or chemical properties.

Suspension feeders are often described as employing passive or active means to capture particles. Passive suspension feeders depend entirely on ambient water flow to supply particles to their feeding structures (e.g., foraminiferans, corals, and brittle stars). In contrast, active suspension feeders usually create their own feeding current to enhance the local supply of food particles or actively swim or engage in other feeding-related behaviors when they sense the presence of nutritious particles (e.g., ciliates, sponges, crustaceans, and bivalves). Some animals can feed either passively or actively, for example, some barnacles, which wave their feeding appendages in weak flow but hold them steady in stronger flow. Many active suspension feeders are often referred to as "filter feeders" because they pump water through a structure that functions as a filter, removing particles from suspension (e.g., sea squirts, certain worms that secrete mucus nets, and baleen whales).

Organisms That Suspension Feed

All of the major animal clades include species that suspension feed (Table 1). Most small animals and protozoans that inhabit the plankton employ some form of suspension feeding, as do some larger drifters such as jellies and salps. Some nekton such as clupeiform fishes (herrings, sardines, anchovies, menhaden), manta rays, whale sharks, and baleen whales are suspension feeders. Numerous benthic invertebrates also remove particles from near-bottom waters. Many of these taxa obtain their nutrition almost exclusively by suspension feeding, for example, most species of sponges, hydroids, anemones, bivalve mollusks, bryozoans, phoronids, brachiopods, crinoids, sea squirts, and some polychaetes (such as fan worms) and crustaceans (such as barnacles and mole crabs). There are, however, several species of bivalves, polychaetes, and crustaceans that are known to switch between suspension feeding and deposit feeding depending on the supply of suspended particles in the near-bottom water.

A great variety of morphological structures, as well as secreted mucus structures, are used by suspension feeders to capture particles (Table 1 and Fig. 1). The taxonomic and morphological diversity of suspension-feeding organisms makes it difficult to draw many ecological generalities about suspension feeding as a process. Nonetheless, much has been learned by focusing attention on the small-scale mechanisms by which suspension feeders capture particles. Capture is a two-step process that involves contacting and retaining particles.

Mechanisms of Particle Contact

The simplest models of particle contact have considered nonturbulent flows that have Reynolds numbers much less than one, where the Reynolds number is a dimensionless ratio of inertial to viscous forces on a fluid. Four general mechanisms (Fig. 2) have been described and related to the form and function of animals' particle-collecting structures such as tentacles or setae, referred to here as particle collectors.

"Direct interception" occurs when particles follow the streamlines of the flow field and the center of a particle comes within one particle radius of the organism's collector. The area of the fluid that is sampled by direct interception at low Reynolds numbers is approximately twice the radius of the particle. The fluid velocity near the collector and the size of the particles determine the rate at which particles are contacted by direct interception. Direct interception is a common contact mechanism for animals that suspension feed with tentacles or setae, for example, polychaetes, echinoderms, crustaceans, etc. (Table 1).

"Inertial impaction" can bring more distant particles in contact with a collector when the specific gravity of the particle exceeds the specific gravity of the fluid. The momentum of a relatively heavy particle can transport it to the collector even if the streamlines

☆*Change History:* March 2018. Irene Martins made minor changes to the text and references.

This is an update of B.T. Hentschel and J. Shimeta, Suspension Feeders, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 3437–3442.

 doi:10.1016/B978-0-12-409548-9.11133-9

Table 1 Examples of marine suspension feeders and the structures used by adults to capture particles[a]

Phylum	*Examples*	*Particle-collecting structures*	*Habitat*
Protists	Flagellates, ciliates, foraminiferans, radiolarians, heliozoans	Flagella, microvilli, cell surface, cilia, pseudopodia, spines	Planktonic, benthic
Porifera	Sponges	Microvilli of choanocytes	Benthic
Cnidaria	Hydroids, hydromedusae, siphonophores, anemones, zoanthids, corals, sea pens, jellies	Tentacles, mucus nets	Planktonic, benthic
Ctenophora	Comb jellies	Tentacles	Planktonic
Rotifera	Rotifers	Ciliated corona	Planktonic, benthic
Entoprocta	Entoprocts	Ciliated tentacles	Benthic
Cycliophora	Cycliophorans	Cilia	Benthic
Sipuncula	Peanut worms	Ciliated tentacles	Benthic
Echiura	Inn-keeper worm	Mucus net	Benthic
Annelida	Polychaetes	Ciliated tentacles, mucus threads or nets	Benthic
Mollusca	Pteropods, snails, limpets, vermetids, clams, mussels, oysters, scallops	Ciliated gill filaments, ciliated parapodia, mucus threads	Planktonic, benthic
Arthropoda (subphylum Crustacea)	Copepods, krill, crabs, shrimps, cephalocarids, branchiopods, leptostracans, mysids, cumaceans, tanaids, barnacles, amphipods, ostracods	Setae, cirri	Planktonic, benthic
Bryozoa	Bryozoans	Ciliated tentacles	Benthic
Phoronida	Phoronids	Ciliated tentacles	Benthic
Brachiopoda	Lamp shells	Ciliated tentacles	Benthic
Echinodermata	Sea stars, brittles stars, basket stars, sea cucumbers, sea urchins, sand dollars, crinoids	Tube feet, spines, pinnules, tentacles, pedicellariae, mucus threads	Benthic
Hemichordata	Acorn worms, pterobranchs	Ciliated proboscis, ciliated tentacles, mucus nets	Benthic
Chordata	Sea squirts, salps, larvaceans, lancelets, fishes, baleen whales	Mucus nets, gill rakers, filter plates, baleen plates	Planktonic, nektonic, benthic

[a]Most of these animals also have a suspension-feeding larval stage in which particles are captured by setae (arthropods) or cilia (other taxa).

of the flow will not bring the particle to the collector via direct interception. The rate at which particles are contacted by inertial impaction is affected by the local velocity, the size of the particles, and the specific gravity of the particles relative to that of the fluid. Inertial impaction is especially important for animals that suspension feed on organic-mineral aggregates near the sediment–water interface, such as various worms, crustaceans, and echinoderms.

"Gravitational deposition" can also cause particles that have a relatively high specific gravity to contact the collector from above. Unlike direct interception and inertial impaction, the rate at which settling particles are contacted by gravitational deposition is not affected by the local fluid velocity at low Reynolds number. Gravitational deposition is especially important for drifting protozoans, pteropods, and benthic cnidarians.

"Diffusional deposition" occurs when the random motion of a particle causes it to cross fluid streamlines that would otherwise prevent it from contacting the collector by direct interception or inertial impaction. The rate at which particles are contacted is affected by the surface area of the collector, the concentration gradient of the particles, and the diffusivity constant of the particle which describes its rate of random motion. This mechanism can be especially important for contacting living, motile particles, for example, protozoa feeding on bacteria and cnidarians feeding on zooplankton.

These four mechanisms of particle contact can, and usually do, act in combination when suspension-feeding organisms live in natural mixtures of particles that have different sizes, shapes, concentrations, and specific gravities. The relative contributions of each contact mechanism can also vary due to fluid velocity. Together, these contact mechanisms can account for selective feeding due to differential contact rates among particle types.

When the Reynolds number of the collector approaches unity, which is the case if the diameter of the collector is approximately 0.01 cm and the local velocity is roughly 1 cm s^{-1}, streamlines become compressed near the sides of the collector and separate in the collector's lee (**Fig. 3**). The streamline compression along the sides of the collector results in sampling particles from a greater area of the fluid than when it occurs without streamline compression at low Reynolds number. Streamline compression also enhances particle contact by inertial impaction, gravitational deposition, and diffusional deposition.

When the Reynolds number of the collector is greater than 10, which can occur if the diameter of the collector is approximately 0.1 cm and the local velocity is roughly 1 cm s^{-1}, streamlines around the collector can separate and form downstream vortices (**Fig. 3**). These vortices allow for particle contact on the downstream side of the collector. Organisms themselves usually obstruct the flow and create downstream vortices. Many benthic suspension feeders, especially colonial

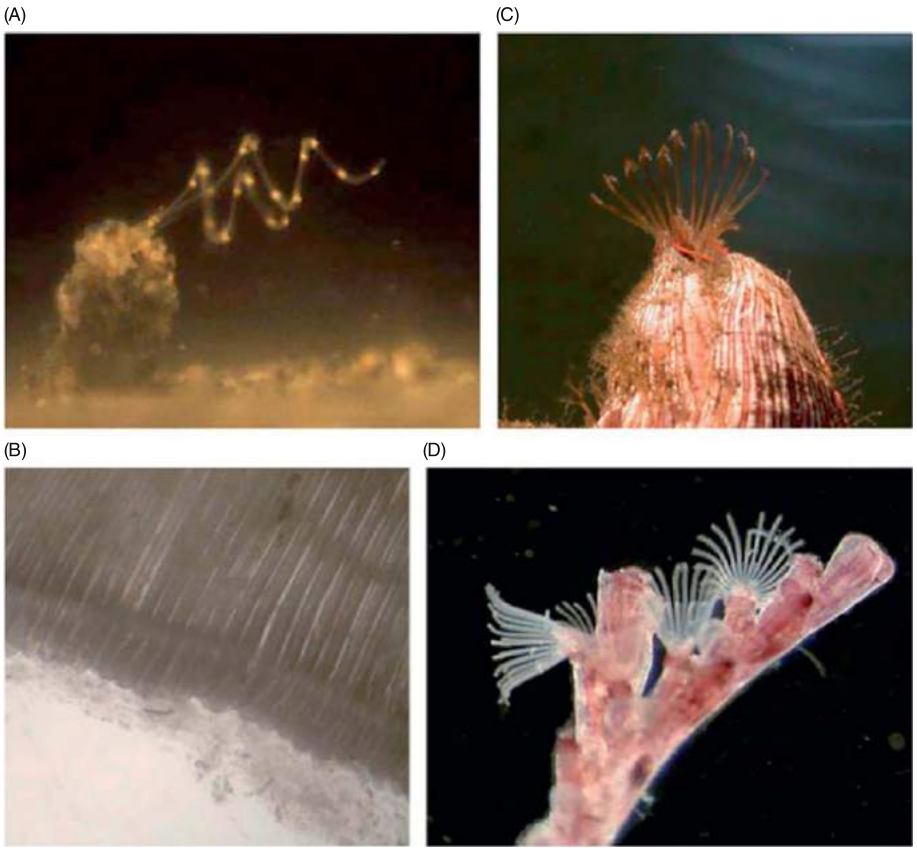


Fig. 1 Photographs of some of the more common feeding structures of benthic suspension feeders. (A) The two tentacles of a spionid polychaete extending from the worm's tube into the benthic boundary layer in a coiled pattern downstream. (B) Magnified image of a bivalve gill that shows the stacks of lamellae that filter particles; a mucus string that aids in transporting captured particles to the mouth is visible at the bottom of the image. (C) The cirri appendages of a barnacle extending above its shell. (D) Magnified image of the lophophore tentacles of a colonial bryozoan. (A) Photograph by J. Shimeta; (B–D) Photograph by B. T. Hentschel.

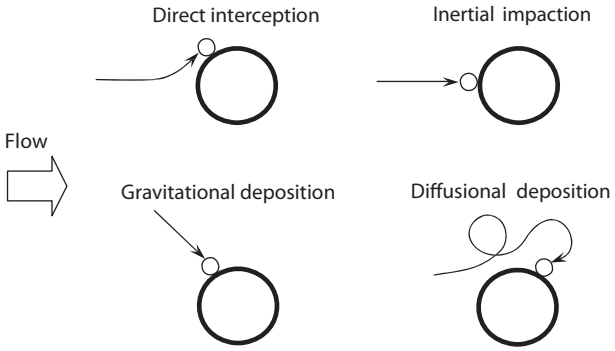


Fig. 2 Mechanisms of particle contact for a cylindrical collector (large circle, shown in cross section), such as a tentacle or seta.

ones like bryozoans and corals, capture particles in the wake of upstream neighbors. Capture rates can be highest on the downstream edge of such colonies.

Retaining Contacted Particles

Particles contacted by the collector(s) of a suspension feeder are not necessarily captured and ingested. They must be retained during transport to the organism's mouth. To successfully retain a particle after its initial contact, some type of adhesive force is necessary to overcome the force of drag and the particle's inertia. Drag and inertia increase with particle size and with increasing

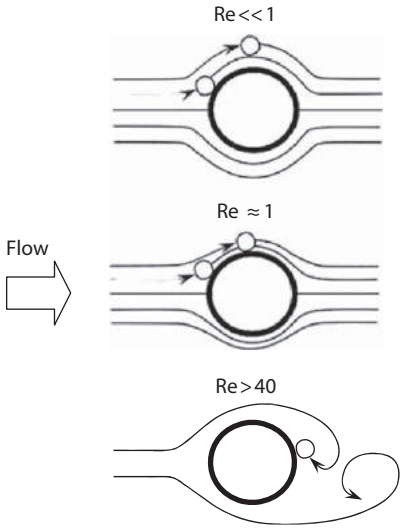


Fig. 3 Flow patterns and particle contact by direct interception for a cylindrical collector at various Reynolds numbers.

velocity. The inertia of a particle also increases with its specific gravity. Therefore, retention is greater at slower velocities and for smaller particles that have a low specific gravity. Because of this, most passive suspension feeders experience a maximal capture rate at an intermediate flow strength.

A variety of mechanisms serve in retaining particles. Mucus and other organic coatings secreted on the surface of feeding structures can enhance retention efficiency. For example, many benthic suspension feeders retain particles on strings or sheets of mucus that cover their tentacles, gills, or pharynx and are then transported to the mouth (e.g., various worms, bivalves, echinoderms, hemichordates, and sea squirts, **Table 1**). Similarly, particles that have sticky organic coatings are more likely to be retained than are relatively clean particles. The electrostatic charge or hydrophobicity of particles also influences their retention. In cnidarians, the nematocysts retain zooplankton prey with barbs and toxins. In filter feeders, networks of collectors form a sieve that retains all particles larger than the sieve's pore size (e.g., the overlapping setae on crustacean appendages, and the mucus nets of worms, sea squirts, and other invertebrate chordates, **Table 1**).

Suspension Feeding in More Complicated Flow Regimes

Although much has been learned from relatively simple modeling of the mechanisms underlying particle contact and retention, suspension feeders often live in flow environments that are much more complicated than steady, nonturbulent conditions. In some active suspension feeders, particle capture depends on strong velocity gradients produced by the feeding currents. For example, ciliary capture occurs in ciliates, polychaetes, entoprocts, bivalves, bryozoans, phoronids, brachiopods, and many invertebrate larvae (**Table 1**). In this capture mechanism, cilia or cirri redirect approaching particles onto a ciliary tract, often by a reversal of the ciliary beat direction. The particles are retained in mucus strings or within the currents of the ciliary tract, without necessarily contacting the cilia. As another example, copepods intercept and sieve particles with complex appendage motions that isolate and trap desired particles from the suspension.

Complexities in the ambient flow also have strong impacts on suspension feeders. The feeding rates of passive suspension feeders depend entirely on variability in the surrounding flow regime. Even the feeding rates of some active suspension feeders such as sponges can be enhanced by ambient flow. Strong flow can deform an animal's feeding structures or otherwise interfere with the animal's ability to create an effective feeding current. Bivalves and other active suspension feeders are known to alter their pumping rates in response to ambient velocities and the concentration of food particles. The growth form or orientation of some benthic suspension feeders is adjusted to maximize exposure to flow, for example, gorgonian corals, crinoids, brachiopods, and scallops.

Most benthic and planktonic suspension feeders experience fluid turbulence. Turbulence can affect the local velocities and the concentration gradients of food particles near suspension feeders. Turbulent pulses of increased velocity reduce particle retention due to greater drag on contacted particles. Under nonturbulent conditions, colonial or aggregated suspension feeders can deplete particle concentrations before the fluid reaches downstream regions of the colony. Under turbulent conditions, however, local depletion of particles is less likely to occur. Many planktonic protozoans were once thought to be smaller than turbulent eddies and, therefore, not affected by turbulent variability in fluid motions. Recent studies have, however, found that the feeding rates of some protozoans can increase or decrease significantly in response to moderate levels of turbulence.

Suspension feeders living in the benthic boundary layer face strong vertical gradients in velocity, turbulence, and particle concentrations. Many passive suspension feeders have a stalked morphology or build tubes that elevate their feeding structures to regions of enhanced particle supply (e.g., foraminiferans, sponges, hydroids, corals, polychaetes, crinoids, sea squirts). If the

concentration and horizontal flux of food particles reach a local maximum at some height above the bottom, many passive suspension feeders such as tube-building polychaetes can optimize the height at which they feed by varying the height of their tube or the extension of their feeding tentacles.

Many suspension feeders inhabiting shallow, coastal areas experience flow that oscillates in time due to wave motion. The behaviors of many benthic suspension feeders have been observed to differ between steady, unidirectional flows and oscillatory flows. Quantitative measures of particle contact, retention, and capture in oscillatory flows are, however, poorly understood relative to those in steady, unidirectional flows.

Ecological Interactions Related to Suspension Feeding

Like all trophic processes, suspension feeding is integral to many ecological interactions. For example, bacterivory by suspension-feeding protozoans and grazing on those protozoans by larger zooplankton are major linkages in pelagic food webs. Unlike many other predator–prey interactions, however, the activities of most suspension feeders extend beyond biotic interactions to affect a wide range of biogeochemical processes.

The vast majority of sessile invertebrates that colonize hard substrata are suspension feeders. These organisms include bryozoans, ascidians, hydroids, encrusting sponges, mussels, and barnacles. Dense assemblages of these sessile suspension feeders often form what are termed "fouling communities" that create unique microhabitats for other organisms. Suspension-feeding corals create an even more extensive habitat that supports diverse communities.

Another obvious impact that suspension feeders have on the environment involves the aggregation and removal of many small particles from suspension. Pelagic grazers such as copepods and ciliates process thousands of microalgal and bacterial cells every hour. The capture and ingestion of these small, dilute food items usually results in aggregation in the form of fecal pellets that sink more rapidly out of the water column and increase the export of organic material from the photic zone to deeper depths.

Benthic suspension feeders can also remove vast quantities of phytoplankton and other particles from suspension. Bivalve mollusks typically pump on the order of 10 L day^{-1}, but some large mussels and oysters have filtration rates as high as 1000 L day^{-1}. The unintentional introduction of zebra mussels (*Dreissena polymorpha*) into North American lakes and rivers has greatly altered the ecosystem because dense populations can effectively clear the entire bodies of water they inhabit of phytoplankton and other particles every few days

In soft-sediment habitats, some infauna suspension feed by pumping water through their burrows or tubes and capturing food particles with a mucus net. The echiuran "inn-keeper worm," Urechis campo, typically irrigates its burrow with 50–70 L day^{-1}. Chaetopterid polychaetes typically pump 10–20 L day^{-1} through their tubes. In addition to removing particles from suspension, this type of infaunal pumping irrigates subsurface, anoxic layers of sediment, creating suitable habitat for many small metazoans.

Suspension feeders that inhabit soft sediments tend to be more common in sandy substrates than in mud. Mud and silt particles accumulate only in regions of reduced water flow, which is not conducive to passive suspension feeding. In addition, fine-grained mud and silt can clog the filter elements of some suspension feeders.

Many benthic suspension feeders live in dense aggregations. In fact, many suspension feeders such as corals, bryozoans, and ascidians are colonial. Nearby neighbors can alter local flow fields and particle concentrations. Downstream members of a colony often experience reduced velocities and particle concentrations due to the "current shading" of upstream neighbors. When roughness elements that obstruct flow (e.g., whole organisms or their feeding appendages) have diverse sizes and shapes, as typically occurs in a mixed-species assemblage, the topographic complexity can lead to enhanced local velocities and feeding rates. When the density of roughness elements exceeds approximately 8% of the bottom area, the wakes surrounding individual organisms interact to create what is termed "skimming flow" around the entire aggregation. This can lead to reduced local velocities and depleted particle concentrations within the aggregation, for example, over stretches of coral reefs or mussel beds.

Further Reading

Cardinale, B.J., Palmer, M.A., Collins, S.L., 2002. Species diversity enhances ecosystem functioning through interspecific facilitation. Nature 415, 426–429.
Chandravanshi, M.L., Mukhopadhyay, A.K., 2017. Dynamic analysis of vibratory feeder and their effect on feed particle speed on conveying surface. Measurement 101, 145–156.
Cresson, P., Ruitton, S., Harmelin-Vivien, M., 2016. Feeding strategies of co-occurring suspension feeders in an oligotrophic environment. Food Webs 6, 19–28.
Eckman, J.E., Duggins, D.O., 1993. Effects of flow speed on growth of benthic suspension feeders. Biological Bulletin 185, 28–41.
Frechette, M., Butman, C.A., Geyer, W.R., 1989. The importance of boundary-layer flows in supplying phytoplankton to the benthic suspension feeder *Mytilus edulis* L. Limnology and Oceanography 34, 19–36.
Hentschel, B.T., Larson, A.A., 2005. Growth rates of interface-feeding polychaetes: Combined effects of flow speed and suspended food concentration. Marine Ecology Progress Series 293, 119–129.
Johnson, A.S., 1990. Flow around phoronids: Consequences of a neighbor to suspension feeders. Limnology and Oceanography 35, 1395–1401.
Koehl, M.A.R., Strickler, J.R., 1981. Copepod feeding currents: Food capture at low Reynolds number. Limnology and Oceanography 26, 1062–1073.
Miller, D.C., Bock, M.J., Turner, E.J., 1992. Deposit and suspension feeding in oscillatory flows and sediment fluxes. Journal of Marine Research 50, 489–520.
Muschenheim, D.K., 1987. The dynamics of near-bed seston flux and suspension-feeding benthos. Journal of Marine Research 45, 473–496.
Naji, A., Nuri, M., Vethaak, A.D., 2018. Microplastics contamination in molluscs from the northern part of the Persian Gulf. Environmental Pollution 235, 113–120.

Okamura, B., 1985. The effects of ambient flow velocity, colony size, and upstream colonies on the feeding success of bryozoa. II. *Conopeum reticulum* (Linnaeus), an encrusting species. Journal of Experimental Marine Biology and Ecology 89, 69–80.
Patterson, M.R., 1991. The effects of flow on polyp-level prey capture in an octocoral *Alcyonium siderium*. Biological Bulletin 180, 93–102.
Riisgard, H.U., Larsen, P.S., 2001. Mini review: Ciliary filter feeding and bio-fluid mechanics—Present understanding and unsolved problems. Limnology and Oceanography 46, 882–891.
Sebens, K.P., Witting, J., Helmuth, B., 1997. Effects of water flow and branch spacing on particle capture by the reef coral *Madracis mirabilis* (Duchassaing and Michelotti). Journal of Experimental Marine Biology and Ecology 211, 1–28.
Shimeta, J., Jumars, P.A., 1991. Physical mechanisms and rates of particle capture by suspension feeders. Oceanography and Marine Biology: Annual Review 29, 191–257.
Shimeta, J., Koehl, M.A.R., 1997. Mechanisms of particle selection by tentaculate suspension feeders during encounter, retention, and handling. Journal of Experimental Marine Biology and Ecology 209, 47–73.
Smith, B.R., Aldridge, D.C., Tanentzap, A.J., 2018. Mussels can both outweigh and interact with the effects of terrestrial to freshwater resource subsidies on littoral benthic communities. Science of the Total Environment 622–623, 49–56.
Wildish, D., Kristmanson, D., 1997. Benthic suspension feeders and flow. Cambridge: Cambridge University Press.

Synecology

ER Pianka, University of Texas, Austin, TX, USA

The *Encyclopædia Britannica* defines synecology as the study of a group or community of organisms and their relationships to each other and to their common environment. Whereas autecology is the study of interrelationships between organisms and their environments at the level of an individual, a population, or an entire species, synecology is concerned with the highest level of biological organization: entire systems of interacting populations in a complex and dynamic physical environmental setting. Synecology is the study of ecosystems, which include both the abiotic nonliving physical environment as well as a biotic component, the community (or biome) of living microbes, fungi, plants, and animals that occur together at any given spot. The concept of a community is itself an abstraction; communities are seldom clear cut and distinct but almost always grade into one another. By considering ecological systems as `open' rather than `closed', and by allowing for continual inflow and outflow of materials, energy, and organisms, this difficulty can be partially overcome and the community concept can be quite useful. Although communities change both in space and in time, they can be examined from an instantaneous view of a fairly localized portion of a larger community. Community ecology is the study of the distribution, abundance, demography, and interactions between populations of coexisting species. Synecology is the most abstract and most difficult kind of ecology, but it is also exceedingly tantalizing and vitally important, as well as extremely urgent as humans complete their domination and usurpation of planet Earth.

As in almost any academic endeavor, two, diametrically opposed, approaches to ecosystem ecology exist: one approach views ecological systems in terms of their component parts, nutrient pools coupled to complex networks of interacting populations. The other approach is more holistic, coming at ecosystems from the top down, rather than from the bottom up. These two perspectives each have their own advantages and limitations, but both are useful.

Community structure concerns all the various ways in which members of communities relate to and interact with one another, as well as any community-level properties that emerge from these interactions. Just as populations have properties that transcend those of the individuals comprising them, communities have both structure and properties that are not possessed by their component populations.

Community ecologists are still in the process of developing a vocabulary. Identification of appropriate aggregate variables or macrodescriptors is essential, but constitutes a double-edged sword; macrodescriptors allow progress but simultaneously constrain direction(s) that can be pursued. To be most useful, such macrodescriptors must simplify population-level processes while retaining their essence without fatal oversimplification. Examples include trophic structure (food webs), connectance, rates of energy fixation and flow, efficiency, diversity, stability, distributions of relative importance among species, niche relationships, resource partitioning, guild structure, assembly `rules', successional stages, and so on. At this early stage in community ecology, it seems prudent not to become overly `locked in' by words and concepts. Even the trophic level concept (producers, consumers (herbivores, omnivores, and carnivores), and decomposers) should not be inviolate (**Fig. 1**).

Clements saw communities as organized superorganisms, whereas Gleason envisioned them as merely statistical ensembles of individualistic species. This debate lives on today, albeit in a somewhat different form. Some putative system-level properties are undoubtedly simply epiphenomena that arise from pooling components; examples would presumably include trophic levels, subwebs, nutrient cycles, and ecological pyramids. But, do communities also possess truly emergent properties that transcend those of mere statistical collections of populations? For example, do patterns of resource utilization among coexisting species become coadjusted so that they mesh together in meaningful ways (see **Fig. 2**)?

If such resource partitioning occurs, truly emergent community-level properties arise as a result of orderly interactions among component populations. The null model approach has been exploited to search for such patterns by Winemiller and Pianka, who constructed randomized replicates of real communities termed `pseudo-communities' and compared these with their prototypes to detect guild structure and resource partitioning.

Such transcendent phenomena or epiphenomena simply cannot be studied at individual or population levels, but must be approached at the level of an entire assemblage or community.

A major problem for community and ecosystem ecologists is that communities are not acted upon directly by natural selection (as individual organisms are). We must keep clearly in mind that natural selection operates by differential reproductive success of individual organisms. It is tempting but dangerously misleading to view organisms or ecosystems as having been `designed' for orderly and efficient function. Antagonistic interactions at the level of individuals and populations (competition, predation, parasitism) must frequently impair certain aspects of ecosystem performance. Effective studies of community organization thus require a pluralism of approaches, including all of the following levels: individuals, family groups, populations, trophic levels, and community networks, as well as historical and biogeographic studies. All these approaches have something useful to offer. The approach taken must be fitted to the questions asked as well as to the peculiarities of the system under study. Much more effort needs to be devoted to connecting community-level attributes and phenomena to how natural selection operates on the behavior and ecology of individual organisms.

 doi:10.1016/B978-0-444-63768-0.00550-3

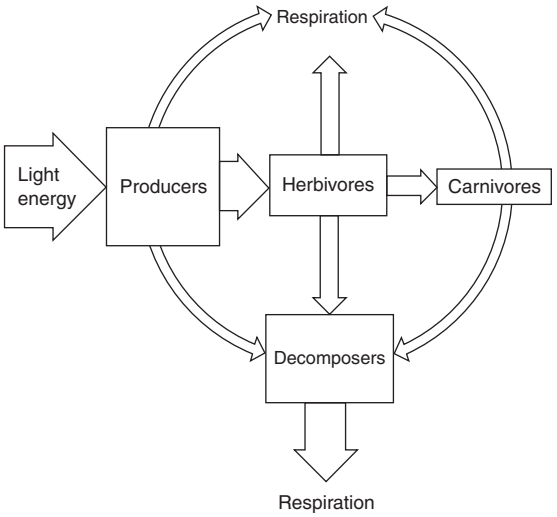


Fig. 1 A simple trophic level `compartment' model of a community, with arrows indicating flow of energy through the system. Widths of arrows reflect rate of flow of energy between particular parts of the system.

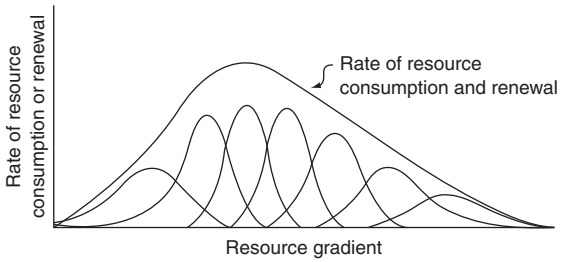


Fig. 2 Niche relationships among members of competitive communities can be represented with bell-shaped utilization curves along a resource spectrum such as height above ground or prey size. Among the seven hypothetical species shown, those toward the tails have broader utilization curves because their resources renew more slowly. In such a competitive community, consumers are at equilibrium with their resources and the rate of resource consumption is equal to the rate of renewal along the entire resource gradient (uppermost curve).

For example, efficiencies of transfer of energy from one trophic level to the next have been estimated to average ~10%–15%. Natural selection operating on individual prey organisms favors escape ability, which in turn reduces the rate of flow of matter and energy through that trophic level, decreasing ecological efficiency but increasing community stability. In contrast, predators evolve so as to be better able to capture their prey, which increases the efficiency of flow of energy through trophic levels but reduces a system's stability. In the coevolution of a predator and its prey, to avoid extinction, the prey must remain a step ahead of its predator. As a corollary, community-level properties of ecological efficiency and community stability may in fact be inversely related because natural selection operates at the level of individual predators and prey. Moreover, the apparent constancy and low level (10%–15%) of ecological efficiency could be a result of the `compromise' that must be reached between prey and their predators.

Ecosystem-level studies are also plagued by difficult problems of scale in both space and time. Patch size and dynamics, climatic events and climate change, nutrient cycles, disturbance frequency, and dispersal ability are just a few of many factors that vary widely within and among systems, as well as over space from local to geographic areas and through time from the short term to the long term.

A plethora of interesting questions can be asked about communities: What structure do they have that transcends population-level processes? What are the effects of community-level attributes on the component organisms living in a given community? What are the roles of parasitism, predation, mutualism, and interspecific competition in shaping community structure? How important are indirect interactions among species and to what extent do such interactions balance out direct effects? How many niche dimensions separate species, and which ones? To what extent are species spread out evenly in niche/resource space? (Such an overdispersion in niche space might be predicted under a competitive null hypothesis, with each species minimizing its interactions with all others.) Do clusters of functionally similar species (`guilds') exist? If so, how can such guild structure be detected and measured? What are its components? Are such guilds merely a result of built-in design constraints on consumer species, and/or do guilds simply reflect natural gaps in resource space? Can guild structure evolve even when resources are continuously distributed as a means of reducing diffuse competition? (A community without guild structure would presumably have greater

diffuse competition than one with guild structure.) Do more diverse communities have more guild structure than simpler communities? What factors determine the diversity and stability of communities and what is their relationship to one another?

Ultimately, we must be able to answer such fundamental questions about `how' natural systems are put together before we will even begin to be able to ask more interesting questions about `why' ecosystems have any particular observed properties, such as ``What are the effects of indirect interactions among populations and/or guild structure on the assembly, structure, stability, and diversity of communities?''

The extreme complexity of most ecosystems makes their study quite difficult but at the same time extremely challenging. Humans now dominate all of Earth's biomes – pristine natural ecosystems no longer exist. Unfortunately, we still know very little about how natural communities function, much to our peril as we continue to usurp whatever limited resources this planet has to offer. Community ecology has thus become exceedingly urgent: humans sorely need to understand how natural ecosystems function and evolve, if only so that we can manage our artificial human-engineered ecosystems more sensibly than we have so far.

Further Reading

Clements, F.E., 1916. Publication No. 242: Plant Succession: Analysis of the Development of Vegetation. Washington, DC: Carnegie Institute.
Cody, M., Diamond, J.M. (Eds.), 1975. Ecology and Evolution of Communities. Boston: Harvard University Press.
Diamond, J., Case, T. (Eds.), 1986. Community Ecology. New York: Harper & Row.
Gleason, H.A., 1926. The individualistic concept of the plant association. Bulletin of the Torrey Botanical Club 53, 7–26.
Gotelli, N.J., Graves, G.R., 1996. Null Models in Ecology. Washington, DC: Smithsonian Institution Press.
Kikkawa, J., Anderson, D.J. (Eds.), 1986. Community Ecology: Pattern and Process. London: Blackwell Scientific Publications.
Morin, P.J., 1999. Community Ecology. Oxford: Blackwell.
Odum, E.P., 1959. Fundamentals of Ecology. Philadelphia: W. B. Saunders.
Pianka, E.R., 1980. Guild structure in desert lizards. Oikos 35, 194–201.
Pianka, E.R., 1981. Competition and niche theory. In: May, R.M. (Ed.), Theoretical Ecology, 2nd edn. Oxford: Blackwell, pp. 167–196. ch. 8.
Pianka, E.R., 1987. The subtlety, complexity, and importance of population interactions when more than two species are involved. Revista Chilena de Historia Natural 60, 351–362.
Pianka, E.R., 1992. The state of the art in community ecology. In: Adler, K. (Ed.), Proceedings of the First World Congress of Herpetology at Canterbury. Contributions to Herpetology, No. 9: Herpetology. Current Research on the Biology of Amphibians and Reptiles., pp. 141–162. Society for the Study of Amphibians and Reptiles.
Pianka, E.R., 2000. Evolutionary Ecology, 6th edn. San Francisco: Benjamin-Cummings, Addison-Wesley-Longman.
Putman, R.J., 1994. Community Ecology. London: Chapman and Hall.
Weiher, E., Keddy, P., 1999. Ecological Assembly Rules. Cambridge: Cambridge University Press.
Winemiller, K.O., Pianka, E.R., 1990. Organization in natural assemblages of desert lizards and tropical fishes. Ecological Monographs 60, 27–55.

Temperature Regulation☆

Inna Sokolova, University of Rostock, Rostock, Germany

Physiological Function of Thermoregulation (The "Why")

Regulation of body temperature (T_b) plays a key role in organism's physiology, ecology, and behavior and encompasses physiological and/or behavioral adaptations whose function is to alter the amount of body heat and thus T_b of the organism. The crucial role of thermoregulation is based on the fact that body temperature directly affects the rates of all biochemical and physiological processes and structure of biological macromolecules. This temperature-dependency of life is rooted in the basic laws of physics and chemistry, particularly in the direct relationship between the temperature and the amount of molecular motion. According to the collision theory of reaction rates, a chemical reaction occurs when the centers of mass of colliding reactants come close enough for a new bond to be formed or an existing bond to be broken. For this to happen, potential energy of the reactants must be equal or greater than a certain threshold energy level called activation energy (E_a) (Fig. 1A). Elevated temperature increases mean velocity of the reactant molecules and the fraction of the molecules with kinetic and potential energy above E_a (Fig. 1B) so that both frequency of molecular collisions and energy of these collisions increase leading to higher reaction rates. Additionally, in enzyme-catalyzed biological processes elevated temperature may convey conformational flexibility to enzymes favoring more rapid binding of their substrates and/or release of the products. Since ligand binding and release are often rate-limiting steps in enzymatic reactions, this temperature-conferred enzyme flexibility also increases reaction velocity. Thus, all other things being equal, higher body temperature (T_b) leads to higher rates of biochemical reactions and thus higher rates of integrative processes such as metabolism, neurotransmission, locomotion, reproduction, and growth.

Different species have evolved to normally function in different ranges of body temperatures. In the Earth's biosphere, the lowest recorded temperature is −89.2°C in continental Antarctica, and the highest is around 100°C in geothermal hot springs and 350°C in certain deep-sea hydrothermal vents. While prokaryotic life can be found over much of this temperature range, the limits of body temperatures compatible with active eukaryotic life are much narrower—from −1.86°C (freezing point of the ocean water) to +45–50°C (large polar mammals and birds which can tolerate external temperatures down to −60°C are no exception to this rule, as their body temperatures are maintained around 35–40°C). However, within each species the range of tolerated body temperatures is narrower still and centered on so called optimal T_b, which provides the highest level of physiological performance. At the molecular level, thermal optimum is determined by trade-off between the rate-enhancing effects of moderate increase in temperature on biochemical or physiological processes, on one hand, and the destructive effects of heat causing structural damage to proteins and excessive lability of biological membranes, on the other. At the whole-organism level, optimal body temperature is determined by integration of the thermal optima of different physiological and biochemical processes. The range of tolerated body temperatures is usually limited by the most temperature-sensitive process, which in many metazoans is neural function.

The proximate function of thermoregulation is to maintain T_b as close to the thermal optimum as possible within given physiological constraints of the organism and physical constraints of its environment. Physiological constraints which may limit the ability to maintain the optimum T_b are set by the rates at which heat is generated and retained by the organism. For example, low metabolic rates may limit the amount of metabolically generated heat available for thermoregulation, whereas the ability to retain body heat may be constrained by high surface-to-volume ratios in small animals or heat loss across the respiratory surfaces in aquatic organisms. On the other hand, physical constraints of the environment can make perfect thermoregulation energetically unfeasible even if physiologically possible due to excessive energy costs or limited access to food resources. Therefore, the degree of constancy of T_b as well as the nature of sources that are used for thermoregulation may greatly differ between the organisms. Based on the predominant sources of the heat used to maintain T_b, an organism can be classified as an ectotherm (i.e., predominantly using external heat sources to thermoregulate) or an endotherm (mostly relying on internal (metabolically generated) heat to maintain T_b). Irrespective of the nature of the heat sources used for thermoregulation, the end-result of it may be either (nearly)

☆*Change History*: March 2018. The author Sokolova updated section 'Metabolic heat generation' and 'Further reading.' Fig. 4 has been added.

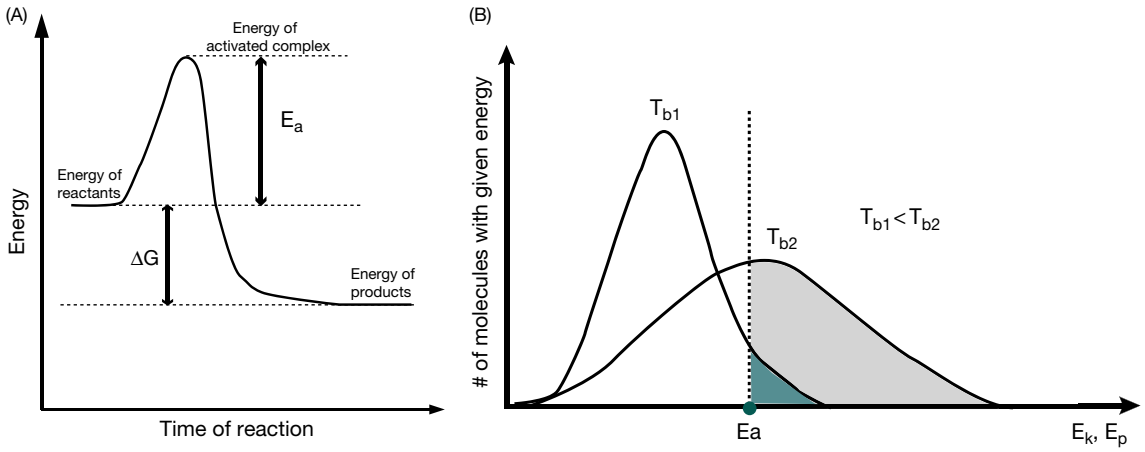


Fig. 1 (A) Activation energy (E_a) of a chemical reaction. ΔG—Gibbs free energy change of the chemical reaction. For the reaction to occur, the energy of reactants must be increased by the value of E_a, after which the reaction proceeds spontaneously. A key function of enzymes as biological catalysts is to reduce the activation energy barrier of activated complex and thus to facilitate biochemical reactions. (B) The rate of biochemical reactions increases with increasing body temperature (T_b) due to increase in the fraction of molecules with energy levels exceeding E_a.

constant T_b, in which case the organism is called a homeotherm, or T_b varying as a result of seasonal and/or diurnal changes in the habitat temperature, in which case the organism is classified as a poikilotherm. These two classification systems are closely related and complementary but not synonymous, as each provides different information about thermal physiology of an organism: the poikilotherm–homeotherm axis refers to the end result of thermoregulation, and the ectotherm–endotherm one—to the means by which it is achieved.

In practice, ectothermic thermoregulation is rarely efficient enough to provide the constant T_b, and most ectotherms are indeed poikilotherms (and vice versa, most endotherms are homeotherms). However, there are several important exceptions to this rule which emphasize that these two classification systems are best kept apart (Fig. 1). For example, a large group of aquatic ectotherms which live in thermally stable environments (such as some species of deep sea invertebrates, pelagic fishes and invertebrates of the open ocean, and Antarctic fish and invertebrates) are essentially homeotherms by the virtue of the constant temperature of their environments. Although lacking active physiological thermoregulation, these organisms have body temperatures that vary by $<0.5^{\circ}$C and are unable to withstand fluctuations of T_b exceeding a few degrees centigrade, similar to many "true" endothermic homeotherms such as mammals or birds. On the other end of the spectrum, several groups of endothermic animals are poikilotherms and experience large fluctuation of T_b under normal physiological and environmental conditions. In fact, most mammals and birds, which are endothermic homeotherms and maintain core T_b close to their physiological optimum (30–45°C for different species) as adults, are temporarily poikilothermic as neonates. Body temperature of newly born mammals and birds may significantly drop when parents leave the nest (by up to 8–10°C in some species) and return back to normal within minutes after the parents return to the nest to warm up the newborns. Some adult birds and mammals from temperate and polar climates can switch to temporary poikilothermy during hibernation or torpor. The degree of poikilothermy during hibernation varies; some species maintain their body temperature at a low but fairly constant set point whereas others allow T_b to follow that of the environment. Some small mammalian hibernators such as Arctic ground squirrels may allow their T_b to fall as low as -1.9°C. Finally, some endotherms (such as African mole rats and golden moles) retain true poikilothermy throughout their adult life. These highly specialized groups of mammals live in arid savannahs in a vast net of narrow burrows which they dig underground. Their bodies are permanently in a close contact with the walls of the burrows which results in a rapid conductive heat transfer between their bodies and the surrounding ground. Maintaining constant T_b would be energetically very expensive under these conditions, and the mole rats have evolved secondary poikilothermy permitting their body temperature to closely follow the temperature of the surrounding ground thereby reducing heat loss. In mammals, such poikilothermy is only possible in a thermally stable environment such as these underground burrows where variations of the temperature do not exceed a few degrees centigrade.

Biological Mechanisms of Thermoregulation (The "How")

Biological mechanisms of thermoregulation are grounded in the First Law of Thermodynamics, which creates physical basis for thermoregulation. The First Law states that the change of the internal energy of a system is equal to the heat added to the system minus the work done by the system. It implies that energy cannot be created or destroyed but it may be transformed or transferred between different systems (e.g., between an organism and its environment). As applied to thermoregulation, the consequence of this Law is that for any organism at any time, net rate of heat gain (ΔH) equals the rate of heat storage (ΔH_s).

$$\Delta H_s = \Delta H \tag{1}$$

For biological purposes, the net rate of heat gain (ΔH) can be broken down to construct a less rigorous but more physiologically meaningful equation of heat balance of an organism:

$$\Delta H_s = \Delta H_m \pm \Delta H_r \pm \Delta H_{cv} \pm \Delta H_{cd} - \Delta H_e \quad (2)$$

where H_s is total heat stored in the body, H_m—metabolically generated heat, H_r—heat gain or loss due to radiation, H_{cv}—heat gain or loss due to convection, H_{cd}—heat gain or loss due to conduction, and H_e—heat loss due to evaporation. If sum of all terms in the Eq. (1) equals to zero ($\Delta H_s = 0$), there is no net heat gain or loss in the organism, and its T_b is constant. In contrast, if ΔH_s is positive or negative, T_b of the organism will increase or decrease, respectively.

While the above outlined equation is in principle applicable to all living organisms, active thermoregulation is mostly restricted to animals and a few species of thermogenic plants, which will be focus of the following discussion. From the above equation it is also clear that thermoregulation can be achieved by physiological as well as by behavioral means using both external and external heat sinks and sources. There is no single mechanism of thermoregulation which works best in all environments, and different strategies can be equally efficient in providing regulation of T_b thus representing the many evolutionary ways to climb the same adaptive mountain.

Thermoregulation Using External Heat Sources

Conduction

Conductive heat transfer is transfer of heat down a temperature gradient between two bodies in close physical contact. The rate of conductive heat transfer depends on temperature gradient between the two bodies, the area of contact and the conductive properties of these bodies. In conductive heat transfer, heat always flows from the warmer body to the cooler one.

All organisms can manipulate conductive heat transfer to some extent, mostly through changing areas of contact between the body and the substrate (e.g., by changing posture) or selecting conductive properties of the substrate upon which they rest through habitat selection or use of insulating nesting materials. For example, penguins can reduce heat conduction from their feet to snow by standing on their heels and using tail feathers as a third point of a tripod to stabilize the posture. Because tail feathers have much lower heat conductivity than the feet, reducing of the area of contact between the feet and the snow greatly reduces heat loss to the environment. Many reptiles can press their body against a warm rock in order to gain heat or lift themselves from the cold substrate thereby reducing heat loss. In fact, if you have ever pressed your hands against a warm oven when you came back from a skiing trip or leaned with your cheek on a cool window pane on a hot summer day, you were using conduction to thermoregulate.

In endothermic mammals and birds, one of the most important long-term adaptations to minimize conductive heat loss is development of insulation, which reduces heat conductivity of the external body surfaces. These insulators can be internal (blubber, fat layer, internal air sacs) or external (feathers, furs, cuticular bristles). Typically, external insulation is much more efficient in preventing heat loss: for example, two centimeters of mammalian fur have about the same insulating properties as 60 cm thick layer of blubber.

Regional heterothermy is another adaptation to manipulate conductive heat flow, which involves regulation of the temperatures of exposed body parts in order to change the thermal gradient between the body surface and the environment in a desirable direction. In cold environments and especially in water where rapid conductive heat loss presents a danger, many species maintain the temperature of their extremities at a lower set point than the core T_b. For example, in otters, seals and arctic birds the temperature of feet, tails and flippers can be maintained at 15–20°C below the core T_b due to vasoconstriction and countercurrent heat exchange (Fig. 2). On the other hand, diving mammals can increase body flow to their flippers during extensive exercise to exhaust excess metabolic heat and to prevent overheating.

Convection

Convective heat transfer is the transfer of heat between two bodies by currents of moving gas or fluid. In free convection, air or water moves away from the heated body as the warm air or water rises and is replaced by a cooler parcel of air or water. In forced convection, air or water is forcibly moved across the body surface (such as in wind or wind-generated water currents) and efficiently removes heat from the body. Convection is a very efficient way of heat transfer because it maintains a steep temperature gradient between the body and surrounding air or water.

Evolutionary adaptations that prevent or enhance heat convection are important constituents of the mechanisms of thermoregulation. In mammals and birds from warm climates, forced convection plays a key role in preventing overheating of the body as heated blood moves from the core of the body towards the peripheral blood vessels where heat is dissipated to the environment. In cold climates, convective heat loss may be reduced by vasoconstriction of the peripheral blood vessels when exposed to cold, and especially by development of dense fur or down which creates a layer of still air next to the body and reduces convection currents. However, one of the most elegant examples how the convection principle was put to use to conserve metabolic heat, is represented by countercurrent blood circulation systems that have evolved independently in several groups of aquatic and terrestrial vertebrates. In countercurrent circulation, peripheral blood vessels are arranged in a spatially organized fashion in close proximity to each other forming nets called rete mirabile. In rete mirabile (the "wonderful net"), blood in arteries and veins flows in opposite direction, and the outward flowing arterial blood loses most of the heat to the colder inward-flowing venous blood (Fig. 2). The heat transfer between arterial and venous blood is very efficient due to the forced convection which maintains temperature gradient between arteries and veins. Therefore, when arterial blood reaches body surface, it has cooled down so that the temperature gradient between the blood and the environment is much lower than between the core body temperature and the environment, and the heat loss is reduced. Metabolic heat received by the venous blood from arterial blood is carried by the venous blood back to the body core.

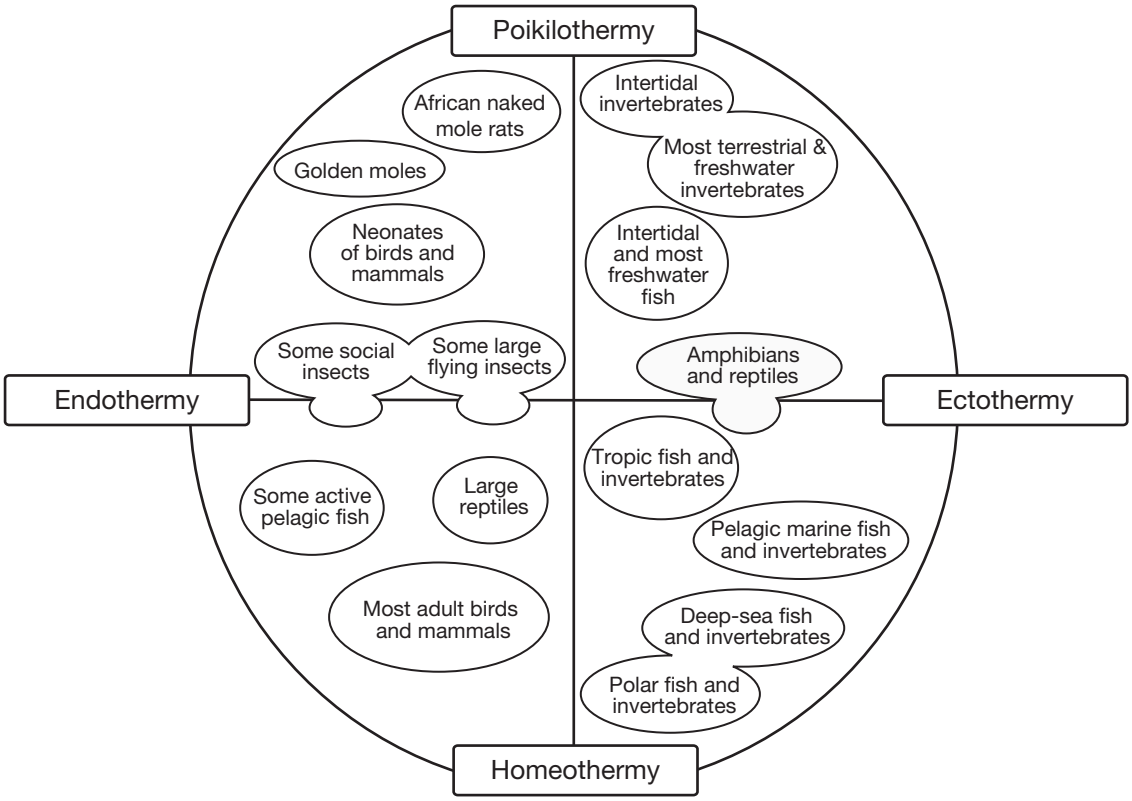


Fig. 2 Examples of organisms' classification in thermal biology based on the degree of constancy of their body temperature (poikilotherms–homeotherms) or on the heat sources used for thermoregulation (ectotherms–endotherms).

These countercurrent systems are used in extremities (feet) of many polar birds, in flippers of aquatic mammals and penguins, and in skin and peripheral muscles of tuna fish, allowing these animals to minimize heat loss despite high thermal conductivity and low temperature of their environments (Fig. 3).

Radiation

All physical bodies with temperature over absolute zero (0 K or −273°C) radiate energy in the form of electromagnetic waves. The wavelength of the radiated energy is directly dependent on the temperature: the hotter is the body, the shorter is the wavelength of the radiation, and the more energy can be transferred from that body to another one. Living creatures with the body temperatures between 0°C and 45°C radiate long-wave radiation in the infrared part of the spectrum thereby losing some of their body heat. They may also acquire heat from more heated objects radiating energy—the best example is sun radiation, which is widely used by both endotherms and ectotherms as an energetically "cheap" way to thermoregulate in cold environments.

In addition to behavioral adaptations to such as basking in the sun, physical properties of an organism may evolve to allow regulation of radiative heat exchange—in particularly, body color and reflective properties. Typically, darker bodies absorb more visible and infrared radiation than the light-colored ones, therefore increasing the rate of heat gain from the environment. For example, some insects and reptiles show latitudinal clines in body coloration with increasing frequency of dark morphs towards high latitudes. In some organisms variation in body coloration is associated with different microclimates on a small spatial scale. Thus, in intertidal snails light-colored morphs are often found on open, sun-lit rock surfaces whereas darker morphs are found in shady crevices, under stones and seaweeds. In some white morphs from open-rock habitats the shell surface consists of a highly reflective calcium carbonate layer which reflects >95% of radiation. This helps snails to prevent overheating during low tides when solar radiation is intense and heat loss is reduced in less conductive aerial environment. However, the relationship between body color and climate is by no means straightforward and should be treated with caution. Firstly, reflectance of the body surface is not always a direct function of color; for example, white and dark human skins have approximately same reflective properties. Similarly, shiny black insect cuticles have nearly the same reflectance as that of the pale matte counterparts. Most importantly, body coloration has many other important functions such as crypsis, mate choice and visual signals to conspecifics or potential predators which may conflict with thermoregulation—hence there are as many exceptions to the "darker in the cold" pattern as there are instances of it (consider for example white polar bears, or snowshoe hares which have dark fur in summer and white—in winter).

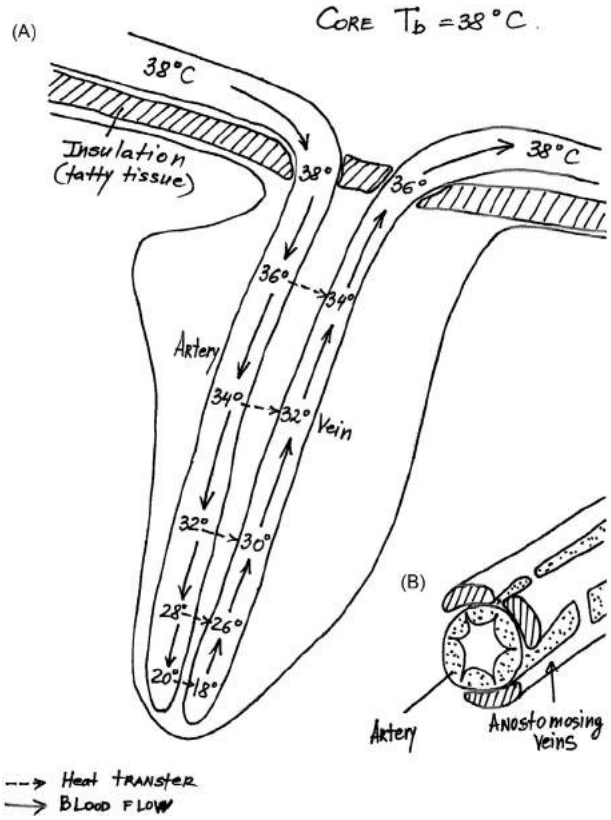


Fig. 3 A schematic representation of blood countercurrent in a penguin flipper (A) and an example of spatial arrangement of blood vessels (B) in rete mirabile.

Evaporation

Evaporation is an efficient means to get rid of the excessive body heat, which can be used by terrestrial and intertidal organisms. Due to high latent heat of vaporization of water (2.5 kJ g^{-1} at 0°C and 2.4 kJ g^{-1} at 40°C), evaporation carries off a significant amount of heat energy from the surface of the body, thereby effectively cooling it. However, using evaporation for cooling has an associated cost of high water loss. Along the latitudinal climatic gradient, evaporation is mostly used by organisms from cold or temperate climates, where water losses can be easily replenished. In arid climates where access to water is limited, evaporation is minimized and its role in thermoregulation is low due to the high risk of dehydration (although some level of evaporation is inevitably present due to gas exchange at the respiratory surfaces). This gradient in use of evaporation for thermoregulation is found not only along the large-scale climatic gradients but also on small scale—for example, along the vertical shore gradient in the intertidal zone of seas and oceans. In the low intertidal horizons where animals have regular access to water, they have developed behavioral adaptations preventing overheating by evaporation from the body surface that allow them to remain active during the low tide. In contrast, high-shore species which may be covered by water only once or twice in a fortnight during spring high tides, avoid evaporation by becoming inactive and isolating themselves from the environment in shelters or within the shells, which often leads to extremely high body temperatures, often in excess of 35–40°C during low tide.

Thermoregulation Using Internal Heat Sources

Metabolic heat generation

Due to the constraints on energy transformations imposed by the laws of thermodynamics, dissipation of energy in the form of heat is an inevitable by-product of metabolism in all organisms. Background metabolic heat production due to the intrinsic "inefficiency" of metabolic systems can be quite high and is similar in ectotherms and endotherms. For example, it is estimated that only about 40% of energy released during catabolism of glucose is conserved in the form of ATP, and 60% is released as heat. However, the amount of metabolic heat generated by ectotherms or endotherms is different due to the differences in the overall rates of metabolism. Most ectotherms belong to so called bradymetabolic organisms (Greek *brady-*, slow), which have low metabolic rates. In ectotherms, metabolic heat gain is insufficient to significantly increase the T_b in the face of the constant heat loss to the environment. In contrast, endotherms are typically tachymetabolic (Greek *tachy-*, fast), that is, characterized by high rates of metabolism, which allows them to use metabolic heat to maintain T_b above that of the surroundings. These differences in the rate of living are due to the intrinsic properties of metabolic machinery of ectotherms and endotherms and are not a simple reflection of

the difference in the body temperature. Even at the same T_b and size, metabolic rates of endotherms are 4–5 times higher than in ectotherms.

Endotherm adaptations to thermoregulation essentially represent evolutionary modifications of the biochemical systems, whose original functions were ATP synthesis, locomotion, ion transport or antioxidant defense. Typically, these modifications involve futile cycles or short circuits in energy production and ATP turnover, which allow these systems to "waste" energy in the form of heat for thermoregulation. In small mammals (such as rodents), all newborn mammals and mammalian hibernators during arousal hibernation, a major site of metabolic heat production is a specialized thermogenic tissue called brown adipose tissue (BAT). BAT is a highly vascularized tissue, rich in mitochondria (which give the tissue its characteristic brown color) and fat depots. The main function of BAT mitochondria is not ATP production, but generation of heat for thermoregulation through futile proton cycling, in a process called non-shivering thermogenesis. BAT mitochondria contain high levels of so called uncoupling protein 1 (UCP1), or thermogenin, which greatly facilitates proton conductance (so called proton leak) in BAT mitochondria. Due to the mitochondrial proton leak, the electrochemical gradient created by the electron transport chain is dissipated by UCP1 bypassing the mitochondrial ATP synthase; thus, besides the heat production, no other useful work is done. Heat generated during this process is carried out of the BAT by a well developed vascular net and distributed through the rest of the body.

Although BAT seems to be a strictly mammalian evolutionary invention, mitochondrial heat generation is used by other organisms for endothermic thermoregulation. In some thermogenic plants such as lotus and arum lilies, large amount of heat is produced in flowering parts to enhance production and volatilization of floral scents to attract pollinators and/or to prevent freezing of the plant reproductive organs. This elevated heat production is also largely due to the futile mitochondrial proton cycling, which is achieved by a mechanism very different from that of BAT—namely, through branching of the electron transport system (ETS) in the inner membrane of mitochondria to an enzyme called alternative oxidase (AOX) instead of cytochrome *c* oxidase. This alternative electron pathway bypasses energy-conserving sites of the ETS thus reducing the electrochemical proton gradient across the membrane and releasing chemical energy as heat. Uncoupling proteins homologous to UCP1 are also found in plants and may play a role in heat production, although their thermogenic function has not been unequivocally confirmed in plants. The arum lilies and sacred lotus can use the mitochondrial thermogenesis to increase temperature in their flowering parts up to 25–35°C above their surroundings, and to maintain it at a fairly constant temperature despite environmental fluctuations. In fact, metabolic rates and heat output in some of those plants can exceed those of the most active endotherm animals such as hummingbirds.

A significant proportion of metabolic heat generation is produced during locomotion and other types of muscular activity. In vertebrates, muscles represent 1/3 of the total body mass and have among the highest rates of ATP turnover of all tissues. Heat generated during muscle contraction provides a significant contribution to the overall metabolic heat production in endotherms. In fact, several species from predominantly ectotherm lineages (such as fishes) have developed regional or complete endothermy solely based on the muscle-generated heat. For example, in some highly active pelagic fish such as tunas and lamnid sharks, metabolic heat generation in white muscle is sufficient to maintain core T_b at a fairly constant level 10–15°C above the temperature of the surrounding water—a very impressive physiological feat given limitations of ectotherm metabolic heat production and high thermal conductivity of water. This locomotion-based regional endothermy is supported by morphological and anatomical adaptations such as the countercurrent retia mirabilia vasculature in the muscles and the gills which prevents excessive heat loss, and the fact that white muscles in tuna fish are located deep within the body and covered by a layer of the red muscles providing insulation to the heat-generating white muscle. A countercurrent heat retention system also helps maintaining the brain and eye temperatures of tunas at 5–6°C above the ambient. Despite these adaptations that keep the muscle and the brain warm, the rest of the sharks' and tunas' body remains at near the ambient temperature. Notably, the same mechanisms that result in regional endothermy in tunas and lamnid sharks (i.e., heat gain by locomotion combined with the heat conservation by insulation and counter-current retia mirabilia) led to the whole-body endothermy in at least one large species of pelagic fish—the moonfish, or opah (Fig. 4). In the opah, large amounts of metabolic heat are produced by the contraction of the massive pectoral fin muscles

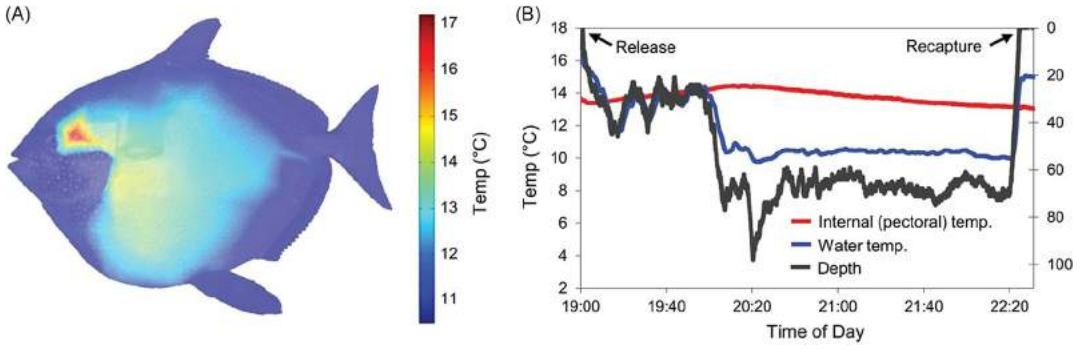


Fig. 4 Whole-body endothermy in the opah, *Lampris guttatus*. (A) In situ internal temperature profile (~4–5 cm below the skin) of an opah at an ambient temperature of 10.5°C. (B) In vivo pectoral muscle temperature for an opah swimming at depth. In both examples, large fish (39–40 kg body mass) were measured. Figure reproduced from Wegner N.C., Snodgrass O.E., Dewar H. and Hyde J.R. (2015). Whole-body endothermy in a mesopelagic fish, the opah, *Lampris guttatus*. *Science* **348**, 786–789 with permission from the American Association for the Advancement of Science.

(constituting 16% of the total body mass and nearly 40% of the total muscle mass) propelling the fish forward. The heat-generating muscle is insulated by a thick layer of fatty connective tissue, and heat is further conserved by the unusually extensive and tightly coupled retia mirabilia inside the gill that warm the oxygenated blood from the gills as it enters the body. To date, the opah is the only known fully endothermic fish species.

Heat generation due to the muscle contraction is also used by birds and mammals facing cold stress in so called shivering thermogenesis. Here an old mechanism of locomotion has been evolutionary modified to perform a new function of thermogenesis. Similar to a decoupling of ETC function from ATP synthesis in non-shivering thermogenesis, the mechanisms of increased heat production in shivering thermogenesis involve decoupling of ATP turnover from locomotion in contracting muscles. During shivering thermogenesis, sympathetic neural system activates actomyosin ATPases in locomotory muscles, but this activation results in uncoordinated small muscle contractions rather than locomotion. Interestingly, in some social insects (such as bees) elevated temperature of the colony can be maintained by a mechanism analogous to collective shivering thermogenesis brought about by rapid uncoordinated contractions of flying muscles of numerous workers in the hive. As a result of shivering thermogenesis, there is significant stationary heat generation but no net movement. The drawback of this mechanism is that it prevents locomotion, which may conflict with other functions such as search for food or escape from predators.

Muscle ATPases are also used as heat-generating engines in cranial heater organs which provide local endothermy to eyes and brain of some fish species such as marlins, swordfish and the butterfly mackerel. However, the futile ATP turnover which generates heat in these organs uses Ca^{2+} ATPases of endoplasmic reticulum (ER) rather than actomyosin ATPases. Neural stimulation results in massive release of Ca^{2+} from the ER of the cranial organ cells which is then pumped back using ATP. ATP is replenished by mitochondrial oxidative phosphorylation. The cycle is repeated as long as the neural stimulation is present using up ATP and resulting in heat generation. The cranial heat organs allow maintaining the brain and eye temperature fairly constant and 10–15°C higher than the temperature of the surrounding water thus providing sustained neural function in the cold and allowing these fish to hunt deep-living prey such as squid. Similar futile Ca^{2+} cycling can also lead to a pathology if runs uncontrolled—for example as in malignant hyperthermia, a genetic disorder in humans and pigs caused by a mutation of the Ca^{2+} release channel in the endoplasmic reticulum (so called ryanodine receptor). Under certain conditions (such as during stress or anesthesia) massive abnormal Ca^{2+} release may occur in the mutation-bearers resulting in intense thermogenesis and hyperthermia.

Endothermic thermoregulation offers many ecological advantages including a greater degree of autonomy from the thermal environment, ability to maintain nocturnal activity, high locomotory performance and stable levels of neural activity. However, it also comes at a considerable cost. High rates of heat loss in small animals due to high surface-to-volume ratio require high metabolic rate and access to plentiful food resources in order to maintain endothermy. In aquatic habitats, high heat loss is inevitably associated with respiratory gas exchange—the heat transfer across the gill surfaces occurs at 50 times higher rate than gas exchange. Some endothermic adaptations for thermoregulation (especially those using futile ATP or proton cycling) may conflict with other fitness-related functions of an organism diverting energy from growth, reproduction or locomotion to energetically "wasteful" heat generation. In contrast, ectothermy is associated with much lower costs of living and allows for a smaller body size, which provides access to ecological niches unavailable to endotherms. In fact, ectothermy appears to be a superior evolutionary strategy in many environments, and 99.9% of species on this planet are ectotherms.

Further Reading

Bennett AF (1987) Evolution of the control of body temperature: Is warmer better? In: Dejours P, Bolis L, Taylor CR, and Weibel ER (eds.) *Comparative physiology: Life in water and on land*. Padova: Livian Press.

Bicudo JEPW, Bianco AC, and Vianna CR (2002) Adaptive thermogenesis in hummingbirds. *Journal of Experimental Biology* 205: 2267–2273.

Heinrich B (1993) *The hot-blooded insects*. Cambridge MA: Harvard University Press.

Jarmuszkiewicz W, Sluse-Goffart CM, Vercesi A, and Sluse FE (2001) Alternative oxidase and uncoupling protein: Thermogenesis versus cell energy balance. *Bioscience Reports* 21: 213–222.

Lowell BB and Spiegelman BM (2000) Towards a molecular understanding of adaptive thermogenesis. *Nature* 494: 652–660.

Montanari T, Poscic N, and Colitti M (2017) Factors involved in white-to-brown adipose tissue conversion and in thermogenesis: A review. *Obesity Reviews* 18: 495–513.

Seymour RS (2001) Biophysics and physiology of temperature regulation in thermogenic flowers. *Bioscience Reports* 21: 223–236.

Solmonson A and Mills EM (2016) Uncoupling proteins and the molecular mechanisms of thyroid thermogenesis. *Endocrinology* 157: 455–462.

Somero GN, Lockwood BL, and Tomanek L (2017) *Biochemical adaptation: Response to environmental challenges, from Life's Origins to the Anthropocene*. Oxford: Sinauer Associates Inc. and Oxford University Press.

Wegner NC, Snodgrass OE, Dewar H, and Hyde JR (2015) Whole-body endothermy in a mesopelagic fish, the opah, *Lampris guttatus*. *Science* 348: 786–789.

Willmer P, Stone G, and Johnston I (2004) *Environmental physiology of animals*. New York: John Wiley & Sons, Inc.

Tolerance Range

AJ Cullum, Creighton University, Omaha, NE, USA

Introduction

The ability of an organism to survive and reproduce in a particular habitat depends on both the biotic and abiotic components of its environment. While the combination of all these components define the organism's ecological niche, the more specific term tolerance range can be used to define the maximum and minimum values of a particular abiotic variable that the organism can withstand. While it may seem obvious to the modern student of ecology that any number of abiotic factors may limit the fitness and distribution of species, this idea was not well developed until the publication of F. F. Blackman's law of limiting factors in 1908 and Victor Shelford's law of toleration in 1913. These laws included for the first time the concept of a minimum and maximum limit of toleration for different aspects of the environment.

The Physical Environment

The tolerance range of an organism can be defined for any of the abiotic – that is, nonliving, or physical or chemical – components of its environment. Abiotic variables commonly examined in this context include temperature, osmolarity or salinity, pH, oxygen and carbon dioxide levels, relative humidity, light levels, wind and water currents, hydrostatic pressure, and the levels of various specific ions, compounds, and toxins. These factors have been of interest because they tend to have significant effects on biological processes and hence on fitness. Some of these variables are relevant to all types of habitats (e.g., oxygen levels), while others obviously apply only to a subset (e.g., relative humidity to terrestrial environments). Additionally, some factors may be very stable in one type of habitat but highly variable in others (e.g., temperature in the ocean benthos versus temperature at the soil surface in a desert). One characteristic of all these variables, however, is that they can reach levels that will begin to have a significant impact on the fitness of organisms (see the section titled `Fitness effects'). When abiotic variables reach these levels, they are sometimes referred to as stressors. But note that a condition that might be stressful for one species may be benign for others.

Although all of the factors listed above have been investigated to some degree in the context of tolerance, by far the most widely studied aspect of the abiotic environment in this regard is temperature. There are a number of reasons for this. One is that temperature is a factor that is relevant to environments of all types and that has profound effects on living things. Another reason is that it varies tremendously across the inhabited environments of the earth, with organisms metabolically active at temperatures ranging from -70 to $+130$ °C. Third, temperature has the interesting property that, below about 0 °C, it causes a phase change in the primary component of both living things and many habitats (i.e., liquid water changes to ice). Finally, temperature can be measured easily in almost any habitat and manipulated easily for experimental purposes. As a result, the most complete descriptive and explicative data on tolerance ranges are for temperature, and thus many of the examples in this article, though not all, will relate to this variable.

Fitness Effects

In general, the tolerance range of an organism or population for a particular variable can be defined by the upper and lower limits at which some minimum level of fitness is achieved. (For the purposes of this article, the term `fitness' will be used in the broad sense (i.e., including both viability and fecundity) rather than the stricter Darwinian sense (i.e., inclusive fitness).) The relationship between fitness and an abiotic factor is normally one in which organisms exhibit maximal fitness over a limited range, with fitness decreasing the further outside this optimal range the abiotic factor varies. This relationship between an abiotic variable and fitness can be termed a tolerance or fitness curve; a hypothetical example is shown in **Fig. 1**. Typically, the term tolerance range is used to indicate the critical minimum and maximum values of the abiotic variable that allow long-term (i.e., effectively indefinite) survival (**Fig. 1**). In some cases, however, the term may reflect alternate levels of tolerance, and the limits for the range will almost always be different as a result. In the context of population sustainability or distribution, for example, the tolerance range would be defined by the limits at which individuals can reproduce (**Fig. 1**). Consider the brown alga *Tinocladia falklandica*, which can survive water temperatures up to 28 °C, but is restricted to the southernmost coasts of South America because reproduction does not appear to occur above about 15 °C. In other contexts, tolerance may be defined by short-term rather than long-term survival. Many insects, for example, can survive temperatures below 0 °C for no more than 24 h; however, if temperatures drop below freezing only at night and are warmer during the day, then long-term survival at such temperatures would not be necessary for long-term survival in this habitat.

Although this article focuses on the concept of tolerance range, it is worth considering a few points about tolerance curves. First, a more complete description of an individual or population's pattern of fitness across an environmental gradient can be provided by a complete tolerance curve, such as those shown in **Fig. 1**, rather than a simple tolerance range. But while tolerance curves offer

 doi:10.1016/B978-0-444-63768-0.00556-4

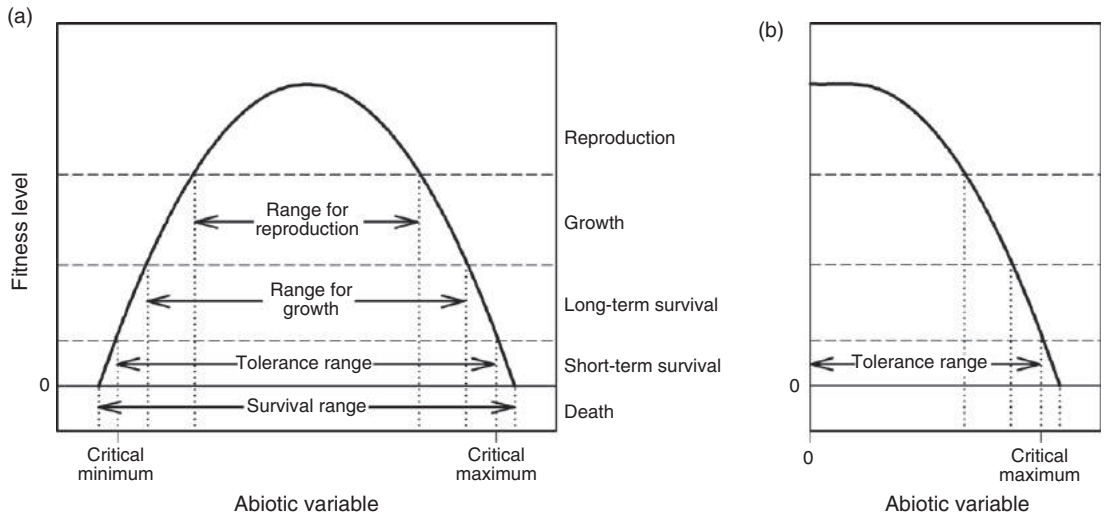


Fig. 1 Tolerance curves showing organismal fitness as determined by the value of an abiotic variable. The tolerance range for a variable is normally defined by the critical minimum and critical maximum values that allow long-term survival by the organism. Other fitness ranges of ecological significance include the survival range, representing the limits at which the organism can survive for brief periods, and the ranges over which reproduction or growth can occur. (a) A bilateral tolerance curve, with reduced fitness on either side of an optimum abiotic range. Such a curve might be seen for temperature or pH. (b) A unilateral tolerance curve, with optimal fitness when the abiotic variable is at or near zero, and declining as the variable increases. Such a curve might be seen for various toxic compounds.

the advantage of providing a specific fitness level for any value of an abiotic variable, tolerance ranges offer two more practical advantages. One is that they are relatively easy to determine, especially at the level of survival. It is much easier to conduct experiments to determine simple lethal limits than to derive detailed measures of fitness across the full range of an abiotic variable. The second is that they are much more concise to define, which can be useful for providing summary information.

A second point about tolerance curves concerns their shape. Note that Fig. 1a shows a symmetrical relationship between fitness and the abiotic variable on either side of an optimal value. However, for many variables the shape may be considerably skewed. In the case of temperature, for example, fitness may decrease much more rapidly on the right side of the curve (high temperatures) than the left (low temperatures). In the case of some variables, such as environmental toxins, the curve may be truncated; that is, maximal fitness is at a toxin concentration of zero, with decreasing fitness as concentrations increase (Fig. 1b).

Tolerance Ranges: Examples for Temperature

A complete review of tolerance ranges for all significant abiotic factors is beyond the scope of this entry. A brief discussion of some limits associated with temperature, however, will help to illustrate some of the patterns that may be seen for many abiotic variables.

Eurythermal species

The prefix eury- is used to indicate species with a wide tolerance range, in this case for temperature. Because organisms must be able to survive in an environment in order to persist in it, most species from habitats with moderately to highly variable temperatures are eurytherms. For aquatic environments, these habitats include shoreline areas and all but the larger freshwater bodies. Animal from these habitats can typically survive temperature ranges of 20 °C or more, and some can survive freezing temperatures. Terrestrial habitats can show an even greater temperature range, in some locations more than 75 °C, and there are birds and mammals that can tolerate this range (but the body temperatures of these endotherms do not, of course, vary by this magnitude). Among ectotherms, a more typical eurythemal range would be about 30 °C for long-term survival.

Stenothermal species

Among the most invariant thermal environments are those in the open ocean, particularly at depth or in polar regions. Species from these habitats typically have narrow thermotolerance ranges, with perhaps the narrowest being the notothenioid fishes of the Antarctic that survive only over a range from about -2 to $+4$ °C. Some freshwater and terrestrial microhabitats also have relatively constant temperatures, and in general tropical regions show less temperature variation than temperate regions. Species from these relatively stable habitats thus tend to have narrower tolerance ranges than those from more variable environments. Many coral species, for example, are having their upper tolerance limits exceeded by increases in ocean temperatures of as little as 1 °C.

Extreme limits

Organisms that thrive at high temperatures are known as thermophiles. Among terrestrial animals, sustained temperatures of 45 °C in vertebrates (the desert pupfish *Cyprinodon*) and 54 °C in arthropods (the desert ant *Cataglyphis*) represent known upper tolerance limits. A few invertebrates from deep-sea hydrothermal vent habitats appear to regularly tolerate temperatures of 60 °C or more. Some desert plants can also tolerate tissue temperatures up to 70 °C for part of the day. Thermophilic algae and fungi can tolerate about 60 °C, and eubacteria about 70 °C. But by far the most extreme thermophiles are among the archaea; the current record is held by a species known as `strain 121', also a hydrothermal vent species, which grows at 121 °C and survives 131 °C.

Cold-adapted organisms, or psychrophiles, are somewhat less variable in the range of limits seen, in large part because of the tendency of water to solidify below about 0 °C. Many species from all kingdoms do well at down to about −2 °C or so. Below this, species may survive but most become inactive, although some organisms remain metabolically active below −10 °C. Cold-adapted endotherms can, of course, stay active at low temperatures by maintaining much higher body temperatures; some mammals survive temperatures as low as −70 °C. In the inactive state, a variety of invertebrates can survive temperatures of −30 °C or lower, and some plants as low as −200 °C. Many single-celled organisms, as well as nematodes and tardigrades, can survive temperatures of −269 °C, or just above absolute zero.

Functional Determinants of Tolerance Limits

One of the most obvious questions to ask about the tolerance range of an organism or species is what determines these particular limits. This type of question had traditionally been addressed by physiologists or biochemists, but as biological inquiry has become more integrated across fields of study, many ecologists have taken a more direct interest in the underlying reasons for tolerance limits.

Physiological Mechanisms

Cells that live independently or as part of very simple multicellular organisms must be able to tolerate directly the range of abiotic environments they encounter if they are to survive (see below). More complex multicellular organisms, however, have the potential to provide an internal environment (a term coined by the famous nineteenth-century physiologist Claude Bernard) that differs from the external environment. Thus the osmolarity, solute composition, pH and even temperature of interstitial fluid, blood, and/or hemolymph of an organism can be maintained at levels different from those of the organism's habitat. This differentiation of the internal state from the external conditions occurs through active physiological regulation of the variable or variables in question, and organisms that maintain such differentiation are referred to as osmoregulators, thermoregulators, etc., as appropriate to the regulated variable. Not all variables need be regulated, however, and the term conformer is applied to organisms that allow an internal state to be at equilibrium with the environment.

For any abiotic factor for which an organism shows conformation, survival in a range of environments requires that cellular mechanisms (discussed below) be tolerant of this range, because internal conditions are similar to external ones. When a variable is regulated, on the other hand, this offers the advantage that individual cells need be tolerant of only a limited range of conditions for the variable in question. Organisms with strong regulation of a variable therefore tend to be able to tolerate a greater range of conditions than a conformer would, and increased regulatory ability has generally meant an increased ability to colonize habitats, such as dry terrestrial environments, that are unfavorable for simpler eukaryotic life.

Given the ability to regulate a variable, then, why do organisms still show some limit to their tolerance range? The general reason is that any regulatory system has a finite capacity to adjust rates of influx and efflux, and if the difference between the organism's internal and external conditions becomes too great, the regulatory system is overwhelmed and regulation fails. Once this occurs, an organism may experience a very rapid decline in fitness, because individual cells may be relatively intolerant of changing conditions.

Even in cases where regulation is possible over the short term, it may be unsustainable energetically. Regulation is an expensive process, and organisms in unfavorable habitats may spend more energy on regulation than they are taking in from the environment. Under these conditions, an individual may be able to live for a limited period of time, but the energy imbalance would make long-term survival impossible.

Biochemical and Cellular Factors

In organisms that either do not regulate a particular abiotic variable (see above) or in which regulation has failed, processes occurring at the cellular level must be tolerant of the changing values of this environmental factor. In cases where tolerance is not possible, injury or death is likely to occur. The particular reasons that cells reach their tolerance limits are, not surprisingly, quite variable, depending on both the environmental factor and the species involved. However, there are two general aspects of cellular biology that are influenced by many abiotic factors and thus merit a brief examination: enzyme activity and membrane biology.

Enzymes

Enzymes are the crucial catalysts in biochemical pathways, and hence are involved to some degree in just about all cellular activities. Changes in the amount of activity exhibited by a particular type of enzyme can cause a pathway to show increased or decreased throughput due to the catalytic effect on reaction rates. While in some cases such biochemical changes may be part of normal cellular control mechanisms, in other cases they can cause significant problems by limiting or bringing to a stop vital cellular processes.

The rate of an enzymatic reaction depends to a significant degree on the enzyme's molecular flexibility. To be effective in catalyzing a reaction, enzymes must have a relatively stable structure in order to present appropriate binding sites to reactants, but must also be flexible enough to rapidly undergo the changes in conformation necessary to allow the reaction to occur. Too much or too little flexibility causes an enzyme to show decreased activity. Temperature is well-known to affect protein flexibility, but pH, ion concentrations, and hydrostatic pressure also influence enzymes in this way. If changes in the organism's environment bring about intracellular changes in these variables, the result can be that enzyme stability and hence activity is affected. As these effects begin to disrupt cellular biochemistry, fitness will decrease.

Membranes

Biological membranes serve two key roles in cellular biology. One is to allow compartmentalization, both of cytosol for the cell as a whole and for the contents of individual organelles. The other is to support membrane-bound proteins and other molecules that allow generation of ATP, cell signaling, and the controlled transport of materials in and out of the cell, among other functions. Just as enzymes must exhibit an appropriate level of flexibility in order to function, biological membranes must exhibit appropriate levels of fluidity and phase structure to function. Even moderate changes in membrane fluidity can influence the permeability of the membrane to different solutes and the function of membrane-bound proteins. This in turn can lead to disruption in the ability of cells to maintain cytostolic composition, carry out some biochemical reactions, and gather information from the extracellular space.

The fluidity of a membrane is determined in part by its composition and in part by abiotic variables. Again, it is temperature that has the most well-characterized influence, with increasing temperature causing membranes to become more fluid. But as before, other environmental factors also affect the membrane's state, with pH, hydrostatic pressure, and the concentration of various solutes all influencing fluidity. When changes in fluidity begin to affect the viability of the cell, the organism may have reached its tolerance limit.

Behavioral Tricks

Tolerance limits that exist at the physiological or biochemical level may often be avoided by behavioral responses that can effectively alter the abiotic environment. Even when constrained to a particular macrohabitat, motile organisms may find more favorable conditions by the selection of an appropriate microhabitat. For example, many terrestrial animals make use of burrows to avoid high or low temperature extremes that may occur above the soil surface. Other types of behavioral strategies can also effectively increase tolerance limits, such as when some fish gulp air and force it across their gills if their aquatic environment becomes too hypoxic. Even relatively or completely sessile organisms can use behavior to effectively increase the range of environmental conditions they can withstand. For example, some desert plants change the orientation of their leaves to the sun during the day to avoid cellular temperatures that would damage living tissue.

Changes in Tolerance Range

The tolerance range or tolerance curve of an individual or a population may change over time. In the case of individuals, such changes are normally strictly phenotypic, with no alteration of the genome; the general term for this type of change is phenotypic plasticity. Shifts in tolerance ranges of populations or species over multiple generations, on the other hand, can include a genetic component and thus represent evolutionary changes.

Changes in tolerance range, whether genetically based or strictly phenotypic, could occur in a number of potential ways. Consider an example in which an organism or population experiences a change in its environment, with the level of an abiotic variable increasing as a result. Some of the possible changes in the pattern of tolerance are shown in Fig. 2. The result might be an increase in the optimum level of the variable as well as in increase in the upper and lower tolerance limits (Fig. 2a), or an increase in the optimal level with an increase in just the upper tolerance limit (Fig. 2b), or an increase in just the optimal level with no change in tolerance limits (Fig. 2c). Other patterns, including intermediates between some of these examples, are also of course possible.

Phenotypic Plasticity

Many organisms have the ability to alter some aspects of their biochemistry, physiology, and/or morphology in response to changes in the abiotic environment, or in anticipation of such changes, thus modifying their tolerance range. These modifications may take as little as a few minutes or as long as several months or even years (but always within the lifetime of the individual), and may be reversible or may be

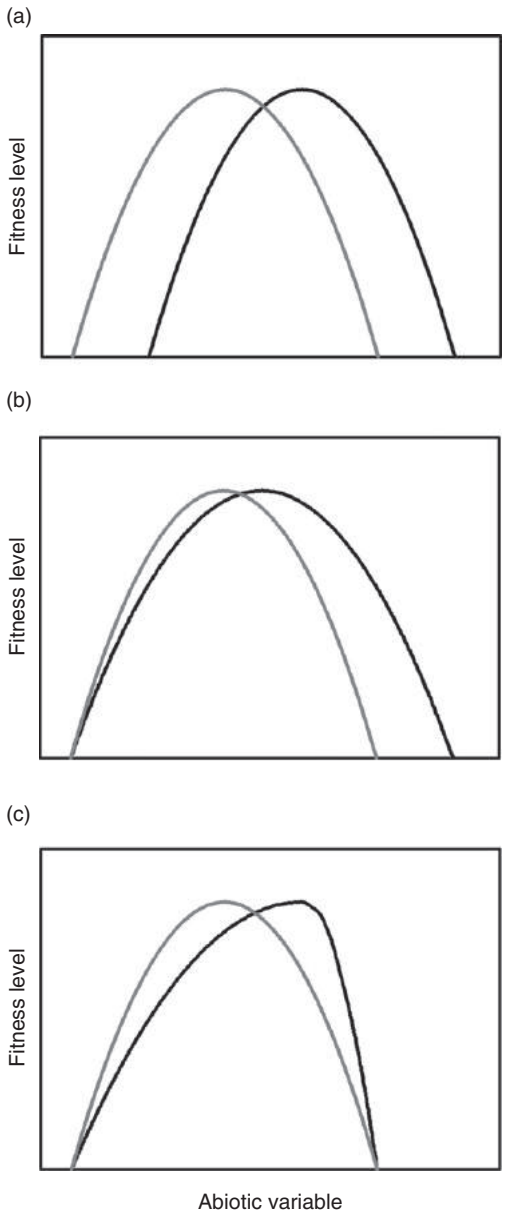


Fig. 2 Potential changes in tolerance curves due to phenotypic plasticity or evolution. Gray curves represent the initial curves, black curves the shifted curves. (a) A shift in the tolerance curve that results in an increase in both the critical minimum and maximum as well as optimal fitness value. (b) A shift in the tolerance curve that results in an increase in just the critical maximum and optimum. (c) A shift in the tolerance curve that results in an increase in the optimum but not in the critical minimum or maximum. These patterns represent just a few of the possible changes that could occur.

permanent. Although the general term phenotypic plasticity can be applied to any such changes, specific classes of change are recognized. Two of the more common such classes are acclimation and developmentally related changes.

Acclimation

Some responses to environmental change occur very quickly; these are normally considered part of cellular or physiological regulatory mechanisms and can be classified as acute responses. The term acclimation, on the other hand, is applied to organismal changes that may take longer to occur. Formally, acclimation refers to a response to changes in a single abiotic variable such as would occur in a controlled laboratory environment, while acclimatization refers to a response to changes in multiple variables such as occurs in nature during, for example, seasonal or altitudinal changes. Often, however, the term acclimation is used to cover both meanings, and that will be the convention here. (To add to the confusion, the term adaptation is sometimes used to refer to these sorts of phenotypic changes, especially by organismal biologists such as physiologists. Additionally, the terms heat-hardening and cold-hardening are also sometimes used to indicate temperature acclimation.)

Acclimatory responses tend to be especially common in organisms that experience large changes in environmental conditions on a regular basis. Although these responses are normally associated with physiological or morphological changes that may take days or weeks in response to seasonal changes, they may also occur in the space of an hour in response to daily environmental fluctuations. Other changes in organisms may take months or more and involve longer-term environmental changes. Whatever their timecourse, acclimatory changes are generally reversible, but there are some exceptions to this trend.

In the context of tolerance ranges, an acclimatory response to changed conditions by an organism would mean a shift in the tolerance curve (**Fig. 2**). The use of the term acclimation typically implies that this shift is favorable to the organism, by improving its fitness under the new or even more extreme conditions. Thus conditions that previously would have been lethal may now be tolerable. For example, sheepshead minnows that have been living at 21 °C for an extended period lose neurological function at 40.1 °C. But if allowed to acclimate for 30 days to 38 °C, a temperature near but still below their original upper thermal limit, this thermal limit increases to 44 °C. Thus a period of pre-exposure to an environmental change can provide a favorable change in tolerance limits, allowing an organism to survive or thrive in conditions that previously would have been detrimental or fatal. Note however that the magnitude of change in tolerance limits may be significantly smaller than the magnitude of change in environmental conditions.

The mechanisms by which acclimatory changes in tolerance occur are variable, but generally depend at least in part on factors discussed in the section titled `Functional determinants of tolerance limits'. In response to a change in an abiotic factor (or in some other variable that helps to predict a change in this factor) organisms can make partially or fully compensatory changes in enzyme activity by activating or deactivating existing enzyme molecules, synthesizing or degrading more molecules, or expressing different isoforms of an enzyme. They may also make changes in membrane composition to return them to appropriate levels of fluidity (a process known as homeoviscous adaptation). In addition, some short-term expansions of tolerance ranges can be achieved through increased synthesis of stress proteins (sometimes called heat shock proteins), which help to stabilize other molecules under conditions that would otherwise tend to cause them to denature or otherwise lose functionality.

The benefit provided by an acclimatory shift in tolerance range may not be without potential tradeoffs, however. Such tradeoffs may occur in several ways. First, any improvement in tolerance at one end of the range often involves a decrement to the tolerance limit at the opposite end of the range (**Fig. 2a**). In the case of the temperature acclimation by minnows discussed above, the lower critical limits for normal function also increased with increased acclimation temperature, from 6.9 °C in the 21 °C-acclimated fish to 11.3 °C in the 38 °C-acclimated fish. Thus colder temperatures that would not have been effectively lethal originally become so as the tolerance range shifts upward. A second cost can result from the expression of stress proteins. Although they allow survival in what might otherwise be lethal conditions, their expression in large amounts can have effects on growth and fecundity that persist even after the organism has returned to more benign conditions.

Note that acclimatory responses may not occur at all in many species, especially those that do not normally experience the sort of variation in the abiotic environment that could select for phenotypic plasticity. Thus these organisms have never evolved, or have secondarily lost, the cellular and physiological mechanisms that underlie the ability to change tolerance ranges. Also, although the term acclimation typically designates a beneficial change in a tolerance range or curve, in some cases the opposite pattern of change can result; that is, pre-exposure to an environmental change may actually decrease an organism's ability to survive further change. For example, while mayfly larvae (*Zephlebia dentata*) kept at 20 °C had higher survival rates at 26 °C than those kept at 15 °C, the opposite pattern was seen for stonefly larvae (*Zelandobius furcillatus*), with lower survival rates at 26 °C in the group kept at 20 °C than those kept at 15 °C. Again the reason for such a loss of tolerance may vary from case to case, but two likely causes are stress-based injury of cells and tissues, or the additional metabolic cost of compensating for the acclimation conditions, with a resulting reduction in energy reserves for surviving the more extreme conditions.

Development

Organisms may show substantial changes in their tolerance range for one or more abiotic factors during ontogeny. Some of these changes may be due to acclimatory responses, as discussed above; that is, different life stages may show different tolerances simply because each stage is acclimating to the different conditions in which it finds itself. If the environmental conditions were held constant, then the tolerance range would not change. One important aspect of acclimation that takes place during development, however, is that physiological or morphological changes that may occur in response to different conditions are more likely to be irreversible than are acclimatory changes in mature organisms.

Other changes in tolerance during development, however, are genetically programmed; that is, they occur to at least some degree regardless of the environment in which the organism actually finds itself. One major reason that such canalized changes in tolerance evolve is that they help prepare individuals for the conditions they will likely encounter at different life stages; that is, they are anticipatory rather than reactive. An example of this is the increase in salinity tolerance that is seen in some salmon species, such as the chum salmon, that occur just prior to the migration of young to marine environments.

Evolutionary Changes

Evolutionary changes in tolerance range (i.e., those having a genetic basis) will generally occur at the level of populations or species rather than individuals. When changes in the abiotic environment occur, populations must be able to survive and reproduce under these new conditions; those that fail to adapt will not persist. Thus a combination of successful adaptation and

extinction means that, over long timescales, species generally show tolerance curves that are suitable, if not optimal, for their habitats; this pattern was clear in the discussion of thermal tolerance ranges above.

But while this general, macroevolutionary relationship between tolerance and environment may be relatively predictable, the dynamics of the evolutionary process for tolerance ranges over shorter timescales are less clear. A change in the abiotic environment might lead to evolutionary changes in tolerance curves in any of the patterns shown in **Fig. 2**, depending in part on the available genetic diversity and appearance of new mutations in the population, and in part on the nature of the changes in the environment. All three of the patterns in **Fig. 2** are seen in nature, both among populations within species and among closely related species, when reproductively isolated groups experience different abiotic conditions. In addition, laboratory-based evolution experiments have shown the same variety of alterations to tolerance ranges under changing conditions. An illustrative example involves lines of *Escherichia coli* allowed to evolve at 20 °C or 42 °C; these temperatures are within a degree of the initial genotype's lower and upper thermal limits (37 °C is the historically optimal temperature for this strain). After 2000 generations, the 20 °C lines showed a decrease in both their upper and lower tolerance limits by 2 °C (the general patterns shown in **Fig. 2a**) compared to the ancestral genotype. Some of the 42 °C lines, on the other hand, showed an increase in their upper thermal limit by about 1 °C with no increase in their lower thermal limit (**Fig. 2b**), and a second group of 42 °C lines showed no change in either thermal limit despite a shift in the thermal optimum (**Fig. 2c**). Thus over microevolutionary timescales, it may be difficult to predict *a priori* how tolerance ranges may be altered in response to environmental changes.

Tolerance to Multiple Stressors

All of the discussions above have focused on tolerance when only a single abiotic variable is changing; that is, in the study on temperature acclimation in sheepshead minnows, for example, water salinity or pH was not changed at the same time. In nature, however, multiple abiotic variables may reach stressful values simultaneously. Such a relationship may occur for two primary reasons. One is that changes in one variable may directly cause changes in another. For example, as soils become more acidic, the change in pH often mobilizes toxic heavy metals that would otherwise be biologically inactive. The other reason for such simultaneous changes is that variables may covary in association with some other change. Consider an increase in altitude, which results in both low oxygen partial pressures and low water content in air.

Organisms exposed to nonoptimal levels of two or more abiotic variables simultaneously often have smaller tolerance ranges for each than they do when exposed to only one stressor at a time. This reduction in tolerance ranges may be due to a number of factors, including additive or even synergistic effects of the variables on cellular biochemistry, and the additional metabolic cost of regulation when multiple environmental challenges occur. An example of interacting effects between two abiotic variables is seen in the early larval stages of the wharf crab, *Sesarma cinereum*, which survive best in seawater of normal salinity and a 25 °C temperature. The largest salinity tolerance range for the species occurs at this 25 °C optimum, with a higher or lower temperature causing a decrease in the salinity tolerance range. Likewise, the maximum temperature tolerance range occurs at normal seawater salinity, and decreases as salinity increases or decreases relative to this optimum. Similar interacting effects are often observed for other combinations of variables.

Further Reading

Cossins, A.R., Bowler, K., 1987. Temperature Biology of Animals. New York: Chapman and Hall.
Feder, M.E., Hofmann, G.E., 1999. Heat-shock proteins, molecular chaperones, and the stress response: Evolutionary and ecological physiology. Annual Review of Physiology 61, 243–282.
Hochachka, P.W., Somero, G.N., 2002. Biochemical Adaptation: Mechanism and Process in Physiological Evolution. New York, NY: Oxford University Press.
Hoffman, A.A., Parsons, P.A., 1997. Extreme Environmental Change and Evolution. Cambridge, UK: Cambridge University Press.
Huey, R.B., Bennett, A.F., 1990. Physiological adjustments to fluctuating thermal environments: An ecological and evolutionary perspective. In: Morimoto, R.I., Tissieres, A., Georgopoulos, C. (Eds.), Stress Proteins in Biology and Medicine, vol. 19. Cold Springs Harbor, NY: Cold Springs Harbor Laboratory Press, pp. 37–59.
Jenks, M.A., Hasegawa, P.M. (Eds.), 2005. Plant Abiotic Stress. Oxford, UK: Blackwell Science Ltd.
Johnston, I.A., Bennett, A.F. (Eds.), 1996. Society for Experimental Biology Seminar Series, Vol. 59: Animals and Temperature: Phenotypic and Evolutionary Adaptation. Cambridge, UK: Cambridge University Press.
Spicer, J.I., Gaston, K.J., 1999. Physiological Diversity and Its Ecological Implications. Oxford, UK: Blackwell Science.
Wharton, D.A., 2002. Life at the Limits: Organisms in Extreme Environments. Cambridge, UK: Cambridge University Press.
Willmer, P., Stone, G., Johnston, I., 2005. Environmental Physiology of Animals, 2nd edn. Oxford, UK: Blackwell Science Ltd.
Wright, D.A., Welbourn, D., Campbell, P.G.C., Harrison, R.M., 2002. Environmental Toxicology. Cambridge, UK: Cambridge University Press.

Trophic Structure

E Preisser, University of Rhode Island, Kingston, RI, USA

Introduction

Trophic structure is defined as the partitioning of biomass between trophic levels (subsets of an ecological community that gather energy and nutrients in similar ways, that is, producers, carnivores). The forces controlling biomass accumulation at each trophic level have been a central concern of ecology dating from the early twentieth-century work of Elton and Lindeman. While interspecific interactions such as omnivory and intraguild predation can make it difficult to assign many organisms to a single trophic level, several broadly defined trophic levels are nonetheless clearly distinguishable. Primary producers, autotrophic organisms (primarily plants and algae) that convert light or chemical energy into biomass, make up the basal trophic level. Primary consumers, generally referred to as herbivores, feed on primary producers. Their consumption of individual producers can range anywhere from a small fraction of the total producer biomass (caterpillars feeding on trees) to the entire organism (fish feeding on algae). Secondary consumers kill and feed on heterotrophic organisms such as herbivores and/or detritivores. Although secondary consumers are referred to generally as predators, this trophic level includes parasites, parasitoids, and pathogens in addition to carnivores. Secondary carnivores, or top predators, are organisms that eat carnivores. The most common examples of this trophic level occur in aquatic systems, where piscivorous fish such as tuna or pike eat smaller fish that feed on zooplankton (which, in turn, feed on phytoplankton). Finally, detritivores derive sustenance from dead organic matter emerging from each of the above trophic levels. While relatively little attention is paid to this trophic level, decomposers process a large fraction of net primary productivity (`NPP') and are integral to nutrient cycling and ecosystem-level processes.

Control of Trophic Structure

Factors affecting the partitioning of biomass between trophic levels can be divided into two broad categories. The first of these categories, bottom-up control, emphasizes the role(s) played by nutrient limitation and energetic inputs to producers and the subsequent efficiency of energy transfer between trophic levels in determining the biomass accumulation at each trophic level. The second category, top-down control, emphasizes the importance of predation in producing patterns of biomass accumulation that are often at odds with those predicted by energy inputs alone. While recognizing the differences between these two factors, it is also important to emphasize that both bottom-up and top-down factors represent extremes along a continuum of importance for regulatory control. While ecologists debate the extent to which bottom-up versus top-down control influence trophic structure in particular ecosystems, there is a broad consensus that both need to be considered when considering community dynamics.

Bottom-Up Control

Bottom-up control of trophic structure means that the production of biomass at each trophic level is a function of energy input into the primary producer trophic level. Biomass accumulated by producers then passes to higher trophic levels as a function of the between-level transfer efficiency. The resulting biomass pyramids are generally characterized by abundant producer biomass and sharp reductions in each higher trophic level. An important exception to this pattern occurs in aquatic food webs, where `inverted biomass pyramids' can occur as a consequence of the extremely short generation time of unicellular producers relative to resident herbivores and predators (see the section titled `Aquatic versus terrestrial ecosystems').

Energy transfer to producers

Although a large amount of light energy is potentially available to producers, only a small fraction of the total is actually converted to producer biomass. Net photosynthetic efficiency (the percentage of available light energy that becomes biomass) in naturally occurring terrestrial and aquatic communities falls between 0.01% and 3%, with values approaching 10% for intensively managed agricultural systems. The resulting NPP is critically dependent on temperature and affected by water availability in terrestrial systems and nutrient levels in aquatic systems.

Energy transfer from producers to higher trophic levels

The overall transfer efficiency of energy between trophic levels is a function of three separate processes. The first of these, consumption efficiency, is the percentage of available productivity at a lower trophic level that is eaten by a higher trophic level. Grazers in temperate lakes, for example, remove nearly four times the fraction of primary productivity eaten by terrestrial grazers. The second process, the consumer's assimilation efficiency, determines what fraction of the biomass ingested by the consumer is

converted to energy. Finally, the consumer's production efficiency determines the percentage of assimilated energy that yields new biomass. Taking all three processes into account, the overall between-level transfer efficiency ranges from 2% to 24%.

System-wide patterns of consumer–resource transfer efficiency are also affected by ecological stoichiometry, the `match' between the nutrient needs of consumers and the nutrient supply of their resources. Transfer efficiencies are highest when consumers feed on resources whose nutrient ratios are similar to their own, and decrease sharply when they feed on resources with dissimilar ratios. Consumer–resource nutrient ratios in aquatic systems are more closely matched than in terrestrial systems, and in predator–herbivore versus herbivore–producer interactions. These facts have been invoked to explain why low transfer efficiencies are generally associated with herbivore–producer interactions and occur in terrestrial systems, while higher transfer efficiencies are characteristic of predator–prey interactions and occur in aquatic systems.

Patterns of biomass accumulation

As NPP and/or transfer efficiency increases, bottom-up control predicts an increasing number of trophic levels as well as an increase in biomass at each trophic level. As producer biomass changes over time, the effect `trickles up' to produce correlated changes in each of the higher trophic levels (Fig. 1, left panel).

Top-Down Control

Top-down control means that predation by higher trophic levels affect the accumulation of biomass at lower trophic levels. Top-down control does not negate the importance of energy input into the basal trophic level; however, it suggests that biomass accumulation at any one trophic level depends on the intensity of predation from the trophic level above.

The `green world' hypothesis

The concept of top-down control first gained widespread attention as a result of the `green world' hypothesis developed by Hairston, Smith, and Slobodkin (hereafter `HSS') in 1960. In brief, HSS posited that the relative rarity of natural disasters and obvious abundance of plant life implied that the producer trophic level was generally limited by competition for light, nutrients, space, and other resources. HSS further reasoned that the `green world' around us is *prima facie* evidence that herbivores do not limit plant abundance; if they did, herbivores would be far more common and plants far less. Given that herbivores seem surrounded by more food than they can eat, it seems unlikely that resource competition limits them; HSS argued that predators are responsible for suppressing herbivore abundance below the level at which they can regulate plant biomass. Predators, in turn, are often territorial and wide ranging in their search for food; this implies that they are self-limited by competition for their herbivore prey. Finally, the fact that we are not surrounded by masses of decaying matter suggests that decomposers quickly and effectively exploit virtually all of their food resources; as a result, this trophic level is likely self-limited as well. While numerous researchers have subsequently identified potential flaws, limitations, and inconsistencies in the HSS hypothesis, its simplicity, clarity, and intuitive logic catalyzed research into the potentially far-reaching consequences of trophic interactions.

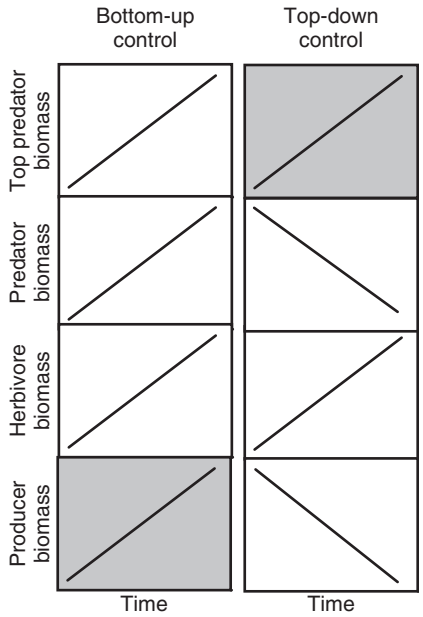


Fig. 1 (Left panel) Bottom-up control of a food chain. As producer biomass (gray box) increases over time, all other trophic levels show correlated increases in biomass. (Right panel) Top-down control of a food chain. As top predator biomass (gray box) increases over time, biomass in the trophic levels below either increase (herbivores) or decrease (predators and producers) in response.

Patterns of biomass accumulation

The hypothesis of top-down control predicts that trophic-level biomass is a function of the trophic interaction most influencing that level. The highest trophic level is always self-limited by competition, making the next-lowest trophic level limited by predation, which in turn allows the trophic level below it to again be limited by competition. In a three-level system, this means that predators and producers are limited by competition (thin top-down arrow) while herbivores are limited by predation (thick top-down arrow) (Fig. 2, a); in a four-level system, the top predators and herbivores are limited by competition while the predators and producers are limited by predation/herbivory (Fig. 2, b). Control exerted via the top trophic level also produces patterns of biomass accumulation distinct from those seen in bottom-up control (Fig. 1, right panel). In comparison to Fig. 1, an increase in top predator biomass leads to decreased predator biomass, thereby releasing herbivore populations which subsequently depress producer biomass.

Trophic cascades

The archetypal form of top-down control involves trophic cascades, where predators indirectly benefit producers by suppressing herbivores (Fig. 3). Such top-down control can be important in freshwater, marine, terrestrial, and belowground systems; in temperate lakes, it can produce visually spectacular differences in producer biomass. While trophic cascades are demonstrably important in many aquatic food webs, their importance in terrestrial systems has been the subject of vigorous debate. Current research seems to indicate that while predators suppress herbivores in both aquatic and terrestrial systems, indirect predator effects on producer biomass occur predominantly in aquatic systems. In terrestrial systems, predator addition often decreases herbivore damage to producers but has less of an impact on overall producer biomass.

Importance of Bottom-Up versus Top-Down Control

Biomass production at all trophic levels is ultimately dependent on the quantity and quality of resources comprising the basal trophic level. Experimental manipulations have generally found that biomass at all trophic levels increases with increased NPP. The ecosystem-level importance of bottom-up control is further underlined by the fact that global patterns of NPP correspond generally to predictions generated by models using only data on abiotic factors such as light, temperature, and water availability. There are also many systems, however, where top-down control clearly acts as a regulatory force; when unchecked by natural

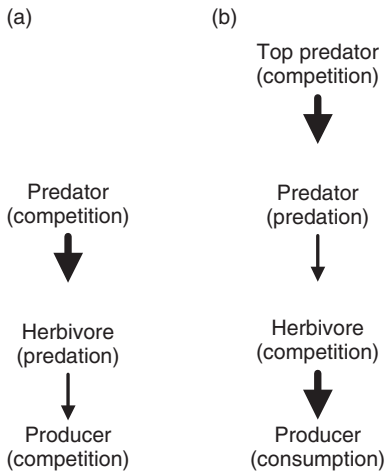


Fig. 2 Top-down control of a food chain. In a three-level food chain (a), predators are limited by competition for resources (thick arrow), herbivores are limited by predation and so cannot limit producers (thin arrow), which are thus limited by competition. In a four-level food chain (b) the pattern is reversed, with predators and producers limited by consumption and top predators and herbivores limited by resource competition.

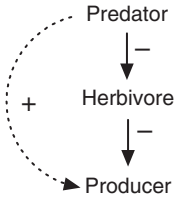


Fig. 3 A trophic cascade. Predators suppress herbivores (`−' arrow), which suppress producers (`−' arrow). By suppressing herbivore biomass, predators indirectly benefit plants (`+' dotted arrow).

enemies, herbivores are capable of population outbreaks that can devastate producer biomass. In agricultural systems, the biological control of crop pests is predicated on the ability of natural enemies to suppress herbivore abundance, reduce producer damage, and increase overall yield. Given that bottom-up control is essential to determining biomass production, the critical issue becomes understanding the conditions and systems in which top-down processes are also important.

Factors Affecting Control of Trophic Structure

A variety of factors affect the magnitude of top-down versus bottom-up control over trophic structure. While these factors can be a function of among- versus within-level trophic interactions, they can also emerge from the linkage between ecological communities and the surrounding environment.

Productivity

The bottom-up effects of increased productivity of the basal trophic level may be capable of influencing the strength of top-down control in a system and the patterns of biomass accumulation at subsequent trophic levels. The relationship between productivity and top-down control was first developed by Oksanen, Fretwell, and others as the `ecosystem exploitation hypothesis' (`EEH'). EEH suggests that as potential primary productivity (`PPP') increases, the equilibrium biomass at each trophic level in a food chain either increases linearly or shows no response (Fig. 4). At very low levels of PPP, there is insufficient producer biomass to support herbivores; as a result, producers are limited by resource competition and their abundance increases linearly as PPP increases (Fig. 4, part A). As PPP continues to increase, however, herbivores can enter the system and divert the increased production of producer biomass into herbivore flesh. Part B of Fig. 4 shows the result: higher PPP yields an increase in herbivore biomass while producer biomass remains unchanged. This continues until herbivore biomass is sufficiently abundant to support predators (Fig. 4, part C). The introduction of predators to the system diverts increased herbivore biomass into predator biomass, freeing producer biomass from herbivore control and allowing it to increase with PPP. When a fourth trophic level enters the system at high PPP, it diverts predator biomass and allows herbivore biomass to increase at the expense of further increases in producer biomass (Fig. 4, part D).

Aquatic versus Terrestrial Ecosystems

Ecologists have long speculated that differences between aquatic and terrestrial ecosystems might affect trophic structure. Early researchers suggested that factors like producer size (predominantly small and short generation times in aquatic systems, predominantly large and long generation times in terrestrial systems) might explain the inverted biomass pyramids found in aquatic systems. Limiting nutrients like nitrogen and phosphorus are relatively more abundant in aquatic food webs; other terrestrial–aquatic differences include the fact that aquatic herbivores consume a larger fraction of available producer biomass, their herbivore–producer transfer efficiencies are higher, and aquatic herbivores are more abundant than their terrestrial counterparts.

Differences in food chain length

One hypothesis for the relative abundance of herbivores and scarcity of producers in aquatic versus terrestrial systems involves inherent between-system differences in the number of `effective' trophic levels (i.e., levels that contribute substantially to top-down control of trophic structure). This argument suggests that terrestrial systems possess three trophic levels (Fig. 2, a) while aquatic systems possess four. Four-level aquatic systems occur via the addition of top predators, for example, piscivorous fish that eat planktivorous fish. In such an example, planktivorous fish are thus predation-limited and cannot control increases in herbivorous

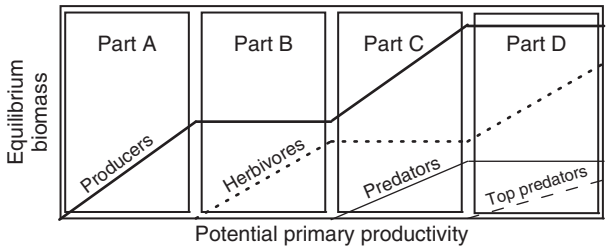


Fig. 4 The ecosystem exploitation hypothesis for trophic-level biomass accumulation as a function of potential primary productivity. Part A: As productivity increases, producer biomass increases. Part B: When producer abundance is sufficiently high to support a second trophic level, consumers enter the system and turn excess producer biomass into herbivore biomass. Part C: When herbivore abundance is sufficiently high to support a third trophic level, predators enter the system and turn excess herbivore biomass into predator biomass. This releases producers from herbivore control and allows producer biomass to increase. Part D: As in part C, but with a fourth trophic level (top predators) entering the system.

zooplankton that suppress phytoplankton biomass (Fig. 2, b). As a result, terrestrial systems have relatively few herbivores and appear `green,' while many (but not all) aquatic systems have abundant herbivores and relatively little in the way of producer biomass.

Differences in herbivore–producer linkage strength

While research supports the contention that aquatic producers experience greater grazing intensity than their terrestrial counterparts, there is less evidence that the apparent abundance of terrestrial producer biomass is due to predator suppression of herbivores. Studies manipulating top predator communities in aquatic versus terrestrial systems demonstrate that the resulting trophic cascades greatly affect producer biomass in aquatic systems. In contrast, the effect of predator addition on terrestrial producers is seen most strongly in reduced producer damage and less in increased producer biomass. Abundant research has shown, however, that manipulating the abundance of top and intermediate predators can strongly affect the biomass of both aquatic and terrestrial herbivores; this implies that system-specific variation in the top-down effect of herbivores on producers may be responsible for the observed differences between aquatic and terrestrial systems.

There are several suggestions for why aquatic versus terrestrial herbivores might have a greater impact on their resources. Phytoplankton, the primary producers in many aquatic systems, need to remain buoyant and absorb limiting nutrients across their cell walls; these requirements constrain them from reaching large sizes and may, in turn, preclude large investments in structural compounds while selecting for faster generation times. Terrestrial producers, in contrast, compete for light by investing in structural compounds that allow them to outgrow their neighbors. System-specific forces should thus select for small size (and rapid generation times) in aquatic producers and large size (and slower generation times) in terrestrial producers. These varying selective pressures may also mean that long-lived terrestrial producers are more apparent to herbivores and invest more heavily in defensive compounds that reduce the impact of herbivory. Ecological modeling has also shown that herbivore control over producer biomass is greatest when herbivores are larger than their resources; this is often the case in aquatic systems but rarely true in terrestrial food webs.

Trophic Position

The Menge–Sutherland hypothesis suggests how the trophic level of organisms within a food web may itself influence the factors that control biomass accumulation. In systems with multiple predator and prey species, herbivorous organisms are often preyed upon by many predators. In such systems, predation thus may play a more important role than resource competition in controlling the biomass accumulated by low-trophic-level organisms (Fig. 5). Although organisms at higher trophic levels are preyed upon by few (if any) predators, they compete fiercely with other members of their trophic level for relatively scarce prey resources. As a result, resource competition should be more influential than predation in determining high-trophic-level biomass. This hypothesis was developed in the context of marine inertial systems, and support for its general applicability has been mixed. A review of experimental literature found that predator effects were strongest on the lowest trophic levels in a food web, supporting the above argument; however, there is less evidence for the corollary that herbivore–herbivore competition is generally weak.

Heterogeneity Within Trophic Levels

Categorizing organisms within a community into discrete trophic levels can conceal a wide range of ecologically relevant differences between and among species. There are several ways in which the species-specific traits of organisms within a trophic level can affect biomass accumulation.

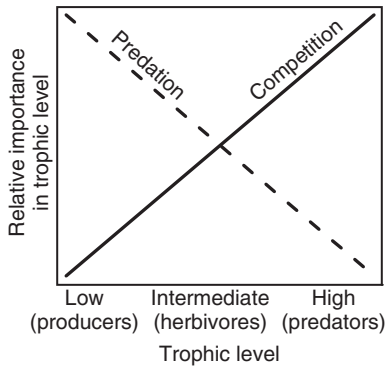


Fig. 5 The Menge–Sutherland hypothesis for within-community dynamics describing the relative importance of competition versus predation in structuring basal (producers) vs. higher trophic levels. Low trophic levels tend to be structured by predation, while higher trophic levels are increasingly structured by competition for resources.

Antiherbivore defenses

An early challenge to HSS came from the argument that the apparent abundance of `green' in terrestrial systems ignores the fact that terrestrial plants possess an array of chemical and physical defenses against herbivory. As such defenses become more effective, an increasing fraction of terrestrial plant biomass becomes effectively invulnerable to herbivory. In the midst of a `green world,' herbivores may thus in fact be forced to compete for access to a limited pool of edible resources. The array of plant defenses found in nature suggests that they are at least somewhat effective in suppressing herbivory; a counterargument points to the fact that even the best-defended plants have at least one herbivore species capable of devastating it in the absence of predators. Moreover, even if plant defenses reduce the overall impact of herbivory, digestibility-reducing compounds like tannins may force herbivores to develop more slowly and remain vulnerable to predators for a longer period of time, while herbivore-induced plant volatiles may attract predators to feeding herbivores. In both of these scenarios, plant defenses may actually serve to increase the efficacy of top-down control.

Heterogeneity in resource edibility

Several models address how interspecific variation in edibility within the basal trophic level might affect trophic structure. These models assume a tradeoff between defense (reduced edibility) and growth, with fast-growing species investing little in defense while slower-growing species are heavily defended and relatively invulnerable to consumption. In low-productivity environments with few consumers, species excelling in resource acquisition will predominate. As productivity and consumer abundance increases, however, the high-growth species will suffer disproportionately from consumption and the slow-growing but highly defended species will become increasingly abundant.

Edible–inedible resource model

The edible–inedible resource model (**Fig. 6**) posits EEH-type control over trophic biomass and predicts that low-productivity environments are initially inhabited only by rapidly growing edible producer species that competitively exclude inedible producers (**Fig. 6**, part A). Its predictions parallel those of EEH for the addition of a second trophic level (**Fig. 6**, part B). It diverges from EEH with the addition of a third trophic level that keeps herbivores from regulating producer biomass; in the edible–inedible resource model, the subsequent absence of herbivore control allows inedible producer species to invade the environment (**Fig. 6**, part C). Since inedible producers cannot be controlled by predation; further increases in PPP yields an increasingly large inedible fraction of total producer biomass.

Keystone predation model

A modification of the edible–inedible prey model, Liebold's keystone-predation model (**Fig. 7**) views resources as varying continuously rather than categorically in their degree of edibility; it predicts a series of species replacements of less- by more-defended resource species as consumer abundance increases. While it generates some of the same predictions as the edible–inedible resource model, it differs in that both consumer and resource biomass increase at a decreasing rate as PPP increases (**Fig. 7**, top panel). This is due to top-down control decreasing as better-and better-defended resource species come to dominate at high PPP (**Fig. 7**, bottom panel). While no resource species is completely invulnerable, consumers gain less and less biomass from preying upon marginally edible species and the system becomes increasingly bottom-up controlled.

Antipredator behavior

Most ecologists have traditionally seen the effect of predators on their prey in terms of the number of prey consumed by predators. A mounting array of evidence suggests that prey are far from helpless victims, however, and that they employ a wide array of defensive strategies. The costs of these strategies can include reduced energy income, lower mating success, or increased vulnerability to other predators. Predators can thus reduce prey density both through direct consumption as well as through the costs arising from antipredator strategies. The `nonlethal' consequences of altered and/or reduced prey foraging in the presence of

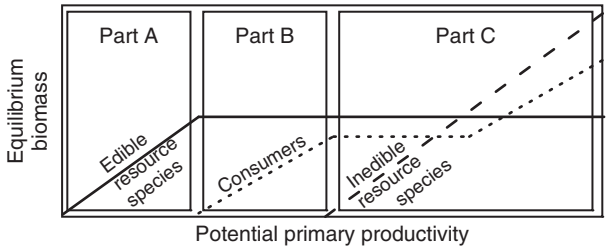


Fig. 6 The edible–inedible resource model showing the trophic-level effect of tradeoffs between species that are either fast growing and edible or slow growing but invulnerable to consumption. Part A: At low productivities, edible resource species outcompete inedible species and increase their biomass linearly with productivity. Part B: When edible resource abundance is sufficiently high to support a second trophic level, consumers enter the system and turn excess resource biomass into consumer biomass. Part C: When consumer biomass is sufficiently abundant to allow the entry of a third trophic level, predators enter (line not shown in figure) and turn excess consumer biomass into predator biomass. This allows inedible resource species to enter the system and increase linearly with productivity while edible species' biomass remains unchanged.

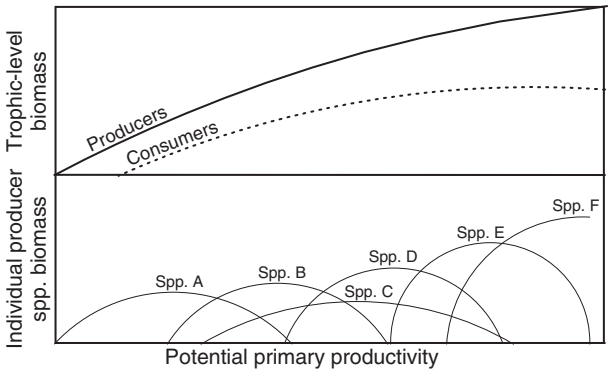


Fig. 7 The keystone predation model showing the trophic-level effect of species replacements by resource species that vary continuously in their growth rate and edibility. At low productivities, communities tend to be dominated by edible resource species with high growth rates. As productivity increases, communities come to be increasingly dominated by mostly inedible resource species with low growth rates. Bottom-up control thus increases as a function of productivity.

predators can profoundly affect the biomass of prey resources. For prey capable of antipredator behavior, predator-induced trophic cascades affecting the prey's resources may thus occur despite minimal prey mortality.

Generalist versus specialist predators

Species identity at the secondary consumer trophic level may also influence trophic structure. Effective top-down control in diverse ecological communities requires that predators consume a wide variety of prey species. As a consequence, generalist predators feeding on an array of species may be most effective at controlling trophic-level biomass accumulation. In contrast, specialist predators affect only a restricted subset of all prey. Any reduction in the biomass of prey targeted by specialist predators may simply release nontargeted species from competition and allow them to increase in abundance; in such a situation, even high densities of specialist predators should produce little change in overall trophic-level biomass. Differences in trophic structure due to predator identity should be most apparent when contrasting communities dominated by large generalist predators (strong top-down control) with communities characterized by small specialist predators (strong bottom-up control).

Omnivory and intraguild predation

Intraguild predators (organisms capable of eating their competitors) and omnivores (organisms that eat both autotrophs and primary producers) pose a major challenge to simple views of trophic structure; even their trophic classification is questionable. While omnivory and intraguild predation have traditionally been considered rare, there is an emerging consensus that both feeding modes occur in (and sometimes dominate) many food webs.

Abundant omnivory and intraguild predation in an ecological community can alter the strength of top-down control. Both feeding modes may act to reduce the strength of top-down control; this occurs because (1) the top-down effect of feeding is diluted across multiple trophic levels; (2) eating other predators may decrease the total predator impact on lower trophic levels; and (3) changes in abundance and feeding rates affect different trophic levels similarly. For example, increased abundance of an omnivorous crayfish should decrease the abundance of both snails (herbivores) and algae (producers). By feeding on the basal trophic level during periods when animal prey are scarce, however, omnivores may sustain high population densities capable of suppressing any future increases in prey biomass. Omnivory in such systems may thus actually serve to increase the strength of top-down control.

Species/trophic diversity

There are several arguments for how high `diversity' (a combined function of the number of trophic levels, species per level, and within-trophic-level foraging strategies) in food webs should alter the relative importance of competition versus predation in controlling trophic-level biomass. One argument suggests that low-diversity communities will tend to have fewer trophic levels, with the vast majority of species occupying the lower trophic levels; in such systems, the relative lack of predator pressure will mean that biomass accumulation will be determined primarily by resource competition (**Fig. 8**). As food web diversity increases, both the number and importance of higher trophic levels increase. This leads to an increased number of predator–prey interactions, and a corresponding rise in the relative importance of predation in structuring the community. As a result, more-diverse communities are structured primarily by predation, while competition plays a predominant role in less-diverse food webs. Contrary to this, an array of empirical work seems to show that less-diverse communities (`food chains') are more likely to show strong top-down control, while more-diverse communities (`food webs') tend to diffuse top-down control and be more affected by bottom-up factors.

Ecological models of even simple food webs incorporating linked food chains and multiple species per trophic level show that such changes may alter the bottom-up importance of increased productivity. Bottom-up effects on predator biomass in a simple food chain can be reduced by predator–predator competition, well-defended herbivores, or herbivores with shared predators and

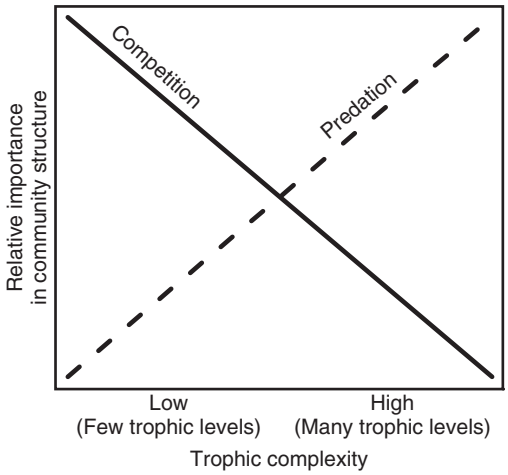


Fig. 8 The Menge–Sutherland hypothesis for between-community dynamics describing the relative importance of competition versus predation in structuring communities of low versus high complexity. Low-complexity communities tend to have few trophic levels and be structured primarily by competition, while high-complexity communities tend to have many trophic levels and are structured primarily by predation.

resources. As models grow more complex, outcomes ranging from strong top-down to strong bottom-up effects, and a range of intermediate conditions, are possible.

Temporal/Spatial Heterogeneity

Variation in both space and time can also affect the accumulation of biomass at different trophic levels. In the broadest sense, spatial and temporal variation in light, temperature, nutrients, and water availability sets the upper limit for ecosystem-level NPP. Systems characterized by wide seasonal variation often feature similarly wide swings in the strength of competitive interactions. Even when less-seasonal systems are included, temporal variation in competitive interactions appear to be the rule rather than the exception; as a result, food webs in a variety of systems may shift between top-down and bottom-up control over time.

Stressful environments

Trophic structure in harsh environments may be controlled by a different suite of factors than those operating in less-stressed systems. Models developed in the context of marine intertidal systems suggest that the effect of abiotic stress on trophic structure should occur because predators are, on average, more mobile than their prey. In relatively benign environments, predators should find it easy to search for and consume prey; higher trophic levels should be fully represented in such environments and be predominantly structured by predator–prey interactions. As abiotic stress increases, however, mobile predators can flee harsh environments while their sessile prey must remain. As a consequence, the importance of predator–prey interactions decreases sharply and competition, not predation, primarily determines trophic-level biomass accumulation.

Refuge habitats

Refuge habitats, areas where prey are free from the threat of predation, were at one time thought to play a major role in determining trophic structure. Since predators can only consume the fraction of prey that have lost in competition for refuge access, bottom-up control will dominate in refuge-rich areas. While such refuge-rich habitats are now generally considered to be the exception rather than the rule, they may explain systems where the changes in predator abundance have little overall impact on prey biomass.

See also: Behavioral Ecology: Optimal Foraging Theory; The Marginal Value Theorem in a Nutshell. Conservation Ecology: Trophic Index and Efficiency; Turnover Time; Trophic Classification for Lakes. Ecological Data Analysis and Modelling: Conceptual Diagrams and Flow Diagrams. General Ecology: Ecological Stoichiometry: Overview; Applied Ecology. Global Change Ecology: Nitrogen Cycle

Further Reading

Abrams, P., 1993. Effects of increased productivity on the abundance of trophic levels. American Naturalist 141, 351–371.
Carpenter, S., Kitchell, J. (Eds.), 1993. The Trophic Cascade in Lakes. Cambridge: Cambridge University Press.
Hairston Jr., N., Hairston, N., 1997. Does foodweb complexity eliminate trophic level dynamics? American Naturalist 149, 1001–1007.
Hairston, N., Smith, F., Slobodkin, L., 1960. Community structure, population control, and competition. American Naturalist 94, 421–425.
Hunter, M., Price, P., 1992. Playing chutes and ladders: Bottom-up and top-down forces in natural communities. Ecology 73, 724–732.

Jiang, L., Morin, P., 2005. Predator diet breadth influences the relative importance of bottom-up and top-down control of prey biomass and diversity. American Naturalist 165, 350–363.
Leibold, M., 1996. A graphical model of keystone predators in foodwebs: Trophic regulation of abundance, incidence, and diversity patterns in communities. American Naturalist 147, 784–812.
Leibold, M., Chase, J., Shurin, J., Downing, A., 1997. Species turnover and the regulation of trophic structure. Annual Review of Ecology and Systematics 28, 467–494.
Menge, B., Sutherland, J., 1976. Species diversity gradients: Synthesis of the roles of predation, competition, and temporal heterogeneity. American Naturalist 110, 351–369.
Morin, P., 1999. Community Ecology. Oxford: Blackwell Science.
Oksanen, L., Fretwell, S., Arruda, J., Niemela, P., 1981. Exploitation ecosystems in gradients of primary productivity. American Naturalist 118, 240–261.
Polis, G., 1999. Why are parts of the world green? Multiple factors control productivity and the distribution of biomass. Oikos 86, 3–15.
Polis, G., Strong, D., 1996. Foodweb complexity and community dynamics. American Naturalist 147, 813–846.
Schmitz, O., Krivan, V., Ovadia, O., 2004. Trophic cascades: The primacy of trait-mediated indirect interactions. Ecology Letters 7, 153–163.
Shurin, J., Gruner, D., Hillebrand, H., 2006. All wet or dried up? Real differences between aquatic and terrestrial foodwebs. Proceedings of the Royal Society of London, Series B: Biological Sciences 273, 1–9.

Water Availability☆

Gage H Dayton, Moss Landing Marine Laboratories, Moss Landing, CA, United States

Introduction

Freshwater is a vital resource for all living organisms. Most organisms have a water content of 50%–90% and there is a critical threshold that they must maintain in order to survive and reproduce. In addition, water also serves as critical habitat for numerous organisms in a wide variety of environments, ranging from extremely arid deserts to tropical rainforests. Regardless of the environment, water availability is a driving force that shapes both aquatic and terrestrial species assemblages. Structure and function of freshwater environments varies across a wide gradient of aquatic environments ranging from small ephemeral pools to expansive permanent lakes. This gradient is controlled by both abiotic factors (e.g., rainfall, aquifer level, soil type, springs, and exposure to sun and wind) and biotic factors (e.g., shading from trees, vegetation, consumption, and biomodification). The hydroperiod (the period of time and area is covered by water) of a water body significantly influences the distribution and abundance of various species, and ultimately plays a significant role in influencing community structure.

Freshwater environments support a vast array of biological diversity, ranging from microbes to trees, and are often referred to as among the most significant and productive ecosystems on Earth. The distribution and abundance of species that occur in a particular freshwater habitat is shaped by a complex system of interactions between abiotic and biotic factors. Most of our knowledge regarding mechanisms that structure freshwater communities comes from studies conducted in the eastern United States and tropical forests of Central and South America. High rainfall over large areas in these regions results in a heterogeneous environment of pools that persist for variable lengths of times, and allow organisms with differing developmental rates to successfully reproduce.

Permanent water bodies are maintained by springs, shallow aquifers, glacial runoff, and snowmelt. Water input from these sources is predictable on an annual basis and results in permanent or long-lived sites. Seasonal rainfall and snowmelt raises aquifers, which in turn influences surface water availability, recharges lakes, and causes streams and rivers to swell. In contrast to permanent water sources, temporary aquatic habitats rely largely upon input from seasonal rains and persist only during wet periods of the year. After a rain event, water is retained for varying durations depending upon subsequent rainfall, runoff, evaporation, biotic uptake, and infiltration. Temporary aquatic environments are usually physically isolated from other seasonal and permanent water bodies (with the exception of flood events) and persist in patchy distributions across the landscape. In temperate and tropical regions where humidity is relatively high, the ground is saturated, and dense vegetation results in shading, temporary aquatic environments tend to persist for many months. However, in arid regions where temperatures are hot, vegetation is sparse, and soils are often course (which results in rapid water infiltration), hydroperiods are characteristically very short. Ultimately, these abiotic factors that regulate water input and output are responsible for influencing water availability and, consequently, the persistence of various organisms across the landscape. Thus, when examining the influence of water availability on the ecology of living organisms, it is important to realize that abiotic factors serve as an initial filter that influence which species can utilize a particular water body. However, as will be explained in more detail later on, biological factors such as resource competition and predation play an important role in influencing which species can persist in a particular water body. These biological factors can be interpreted as secondary filters influencing species assemblages in particular aquatic habitats.

☆*Change History:* March 2018. Irene Martins made minor changes to the text and references.

Aquatic Organisms

The ecology of freshwater organisms can be generalized as a gradient between temporary and permanent pools that is largely influenced by competition and predation. Small highly ephemeral pools often have low species richness due to rapid drying that results in short duration of the site. The short hydroperiod of highly ephemeral pools restricts which organisms can successfully use a particular site. For example, pools that persist for <3 weeks cannot be successfully utilized by organisms that require permanent water (e.g., fish), nor can they be utilized by organisms that require the presence of water for >3 weeks for part of their life cycle (e.g., most amphibians and aquatic insects). In temporary water bodies there is a positive relationship between hydroperiod and species richness of both aquatic invertebrates and amphibians. However, as pools transition from temporary to permanent sites there is a decrease in species richness of many taxa (Fig. 1). This decrease in richness is largely attributed to the presence of fish and other predators that are dependent upon long-lived or permanent water. Thus, hydroperiod is a driving abiotic factor that regulates which species can successfully use a site, and ultimately influences community composition of aquatic environments.

Temporary Aquatic Environments

Aquatic organisms inhabiting temporary water bodies require life cycles that enable them to survive during dry periods. Of the many species adapted to live in temporary water bodies, each has its own threshold for the minimum time period in which a site must be inundated in order for members of that species to successfully complete the aquatic component of their life cycle. Although there are a wide variety of adaptations that enable organisms to persist in temporary ponds, the limiting factor is that, prior to the site drying up, individuals must reach a critical life stage that enables them to survive outside of an aquatic environment. Most species that inhabit temporary aquatic environments must either metamorphose into a terrestrial form (as seen in most amphibians and many insects) or lay eggs that are capable of withstanding dry phases (exemplified in fairy shrimp, mayflies, and cladocera). For example, in North America the larval stage of most amphibians is at least 2 months. Thus, the lower temporal threshold of pond hydroperiod for the majority of amphibians that occur in North America is 2 months. If pools dry up prior to this time period, larvae are unable to complete metamorphosis and subsequently there are no adult recruits. In the case of many aquatic invertebrates, such as fairy shrimp, the entire life cycle is completed in the temporary pool. Eggs remain dormant in the dirt until the following wet season, whereupon eggs hatch and the cycle repeats itself.

Species that inhabit short-lived pools are generally superior resource competitors to species that have relatively long larval periods. Competitive ability enables organisms living in short-lived pools to grow quickly and obtain resources necessary for metamorphosis and/or reproduction. However, there is often a fitness cost to being competitively dominant. Aquatic organisms that inhabit temporary pools are typically very active foragers that exhibit high activity rates; this high activity rate makes them susceptible to a wide variety of predators. These species are more vulnerable to predation due to the fact that activity is correlated with predator encounter rates through both enhanced detection by visual predators as well as increased probability of encountering a "sit and wait" predator through spatial movement. Thus, aquatic organisms associated with highly ephemeral ponds are often restricted to these types of environments due to their inability to coexist with predators that are abundant in longer-lived and permanent sites. The unique adaptations required to successfully persist in highly ephemeral pools results in a disproportionately

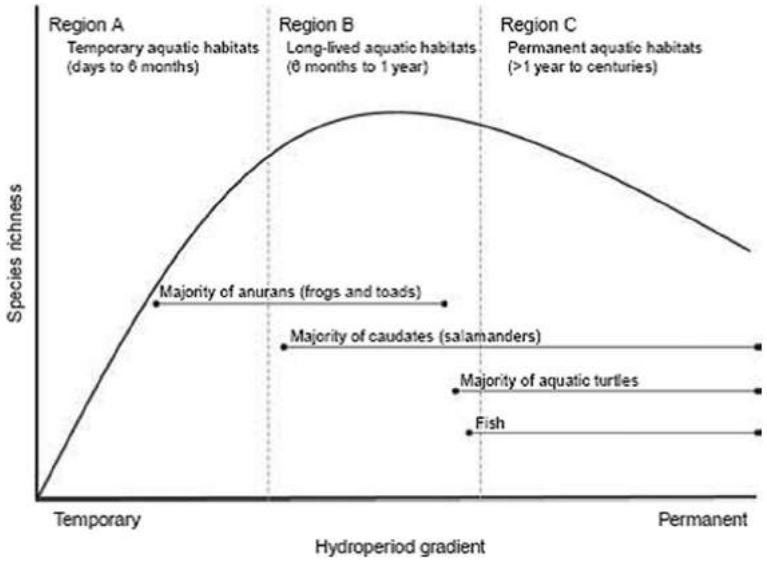


Fig. 1 Water permanency and species richness of several vertebrate groups that utilize aquatic habitats. Adapted from Wellborn, G. A., Skelly, D. K. and Werner, E. E. (1996) Mechanisms creating community structure across a freshwater habitat gradient. *Annual Review of Ecology and Systematics* **27**, 337–363.

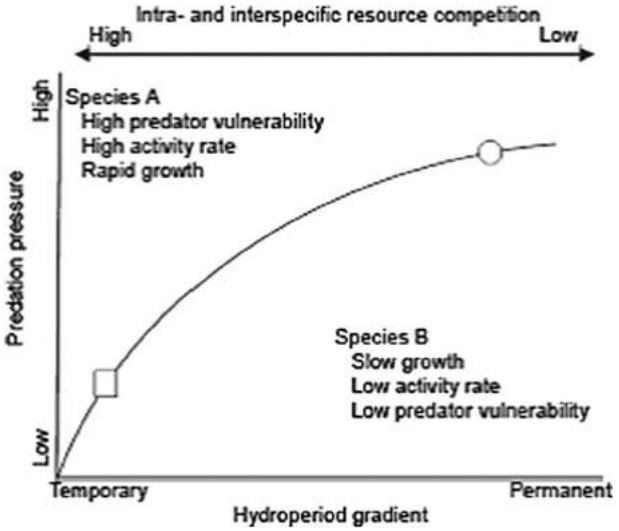


Fig. 2 Conceptual model that depicts the relationship of predation pressure, competitive environment, and hydroperiod gradient with predictions of prey characteristics along the gradient. Species A is likely to be susceptible to predation and show few defensive traits in the presence of predators because the cost of altering behavior and growth in delaying developmental time are too great in sites where desiccation is the primary cause of death. As a consequence of high susceptibility to predation, species A should be excluded from sites with long hydroperiods and high predator loads. Species B exhibits predator-induced defenses in the presence of predators. As a result, developmental times are longer, and species B is excluded from short-lived pools.

high number of endemic species associated with ephemeral water bodies, especially those that occur in dry environments where temporary pools are often very patchily distributed across both space and time.

As hydroperiod lengthens, there is a positive correlation with species richness. Many of the species that inhabit longer-lived pools are predators. As predation pressure increases in longer-lived pools, species composition shifts from an assemblage adapted to rapid growth and short life cycles, to an assemblage dominated by relatively slow-growing species adapted to coexisting with predators (Fig. 2). Organisms that inhabit aquatic environments where predators are common exhibit a wide array of defense strategies. Primary defense mechanisms include cryptic coloration, reduced activity, chemical compounds, and morphological adaptations. All of these defense strategies have a cost to the organism in that resources are either allocated toward producing defensive structures, or energy intake is decreased due to reduced feeding and lower activity rates. Consequently, the time required to reach critical developmental stages necessary for metamorphosis, or to produce gametes, is increased.

Permanent Aquatic Environments

Although temporary and permanent water bodies vary tremendously, competition, and predation are also important mechanisms affecting species assemblages in permanent aquatic habitats. In permanent aquatic habitats that lack predators, resource competition among organisms is believed to be important in structuring the composition of zooplankton communities. When few or no fish are present, zooplankton assemblages tend to consist of different species that are typically much larger in size than those that occur in lakes where fish are present. These relatively large zooplankton are believed to be superior resource competitors over smaller zooplankton; however, as discussed above for temporary pools, with competitive ability comes the cost of susceptibility to predation. In lakes where fish occur naturally, large zooplankton species are absent and assemblages are dominated by small species that are capable of withstanding predation pressure. Clear evidence for the impact of top predators on the structure of zooplankton, as well as amphibian, assemblages in permanent water bodies has been observed in alpine lakes of the Sierra Nevada Mountains in California. Many of the lakebeds in the Sierra Nevada Mountains were formed by glaciers carving out bedrock basins, which were later filled by annual snowmelt. As a result, these lakes are often very isolated and occur at high elevation and subsequently were never colonized by fish. However, beginning in the late 1800s, pioneers and sheepherders began stocking many alpine lakes with fish. The introduction of a top predator into these lakes eliminated many of the invertebrate and vertebrate species that naturally inhabited these sites. Thus, even though water availability remains constant among permanent sites (e.g., it is by definition always present), the presence of fish and other predators play an important role in structuring the species assemblages of permanent water bodies.

The Temporary/Permanent Aquatic Boundary

The potential fitness benefit for aquatic organisms that can utilize both temporary and permanent water bodies is very high. However, there are very few species that exhibit a generalist strategy that enables them to use both types of aquatic environments. This suggests that the selective pressures, and ultimately the costs, associated with being specialized for either fast growth (able to

Fig. 3 Couch's Spadefoot toad (*Scaphiopus couchii*). This individual recently emerged from estivating beneath the surface (notice mud on top of its head). Spadefoot toads remain underground for the majority of the year, emerging during heavy seasonal rains to breed and feed. Tadpoles will complete their life cycle in 7–10 days.

persist in temporary pools) or the ability to withstand predators (able to persist in permanent lakes and rivers) are high. However, there are many species that exhibit plasticity (ability to alter or change) in behavior, morphology, and or developmental rates. The ability to alter behavior or morphology, and thus developmental rates, by increasing feeding or portioning resources for defense mechanisms, enables some organisms to live in environments where competition and predation pressures vary from year to year. Plasticity is common in many organisms that experience fluctuating temporary aquatic environments that last for several months. Some of the best examples of phenotypic plasticity in aquatic organisms can be highlighted by examining behavior and morphological modifications that occur in tadpoles that inhabit long-lived temporary to permanent water bodies in the northeastern United States. In this region, permanent water bodies are abundant and temporary water bodies typically last for several months and exhibit a relatively linear increase in tadpole predators as hydroperiod increases. As predator density increases, many amphibian species reduce their activity rates, thus reducing the detection probability of predators. In addition, they will exhibit morphological changes in body shape that enable them to swim faster and evade predators. For example, when Gray Treefrog tadpoles inhabit waters with predators they reduce their activity rates and grow large tails that are tinted orange. The large colored tails are believed to help them evade predators by focusing attacks to regions of their tails that will rip easily and allow them to escape predation events. Because water is either permanent or present for a relatively long period of time, these plastic responses to shifting predatory regimes is an adaptation that likely increases the fitness of the species (e.g., time is not the limiting factor influencing tadpole survivorship, rather being eaten by a predator is the most likely cause of death).

Conversely, in very short-lived desert pools in the Chihuahuan Desert where Couch's Spadefoot (*Scaphiopus couchii*) breed (Fig. 3), desiccation is the most important factor influencing tadpole survivorship. Couch's Spadefoot tadpoles are very active foragers and can metamorphose in only 7–10 days after eggs are deposited (this is the shortest larval period for all North American amphibians). High densities of predators are very uncommon in temporary desert pools and the limiting factor for successful metamorphosis of Spadefoot tadpoles is time (rv75% of all breeding pools dry up prior to any metamorphosis occurring). Couch's Spadefoot tadpoles are very active foragers as they need to meet energy requirements for metamorphosis prior to ponds drying up. These high activity rates in turn make them susceptible to predators; thus, they are not capable of persisting in pools with high predator loads. The evidence that Couch's Spadefoot larvae are not adapted to coexisting with predators is further supported by studies that have exposed tadpoles to predators in controlled experimental settings and found that they do not exhibit plasticity in behavior or morphology. Together, these traits suggest that selective forces on rapid growth for Couch's Spadefoot tadpoles outweigh the costs of potential behavioral or morphological responses to predators.

Thus, while very few species can persist in both temporary and permanent water bodies, many organisms that inhabit intermediate sites tend to exhibit plasticity in behavior and/or morphology that enables them to adapt to varying competitive and predatory environments. However, in very short-lived aquatic environments, desiccation drives selection for rapid growth rather than antipredator responses. Finally, in permanent environments, predation fuels selection of behavioral and morphological traits that enable species to coexist with predators.

Water Availability and Terrestrial Organisms in Desert Environments

In addition to water availability being critical for all life, it is perhaps the single most important factor influencing the distribution of organisms that inhabit desert environments. Water availability for terrestrial organisms is important at both the surface and subsurface level. Surface water is an essential resource for terrestrial vertebrates and invertebrates (imagine a mountain lion lapping out of a pool or bees buzzing around your dogs water bowl). Subsurface water is critical for the persistence of most plants, benthic

invertebrates, and a wide diversity of fungi. Once water enters the soil (via precipitation, ground water, snowmelt, or runoff from surrounding water bodies), its availability to plants and other organisms depends upon the water potential of the soil. Water potential is the tendency of any system (e.g., soil, plant, and animal) to transfer water to its surroundings. Water always moves down the gradient from a system with high water potential to one with lower water potential.

The transfer of water from soil to roots is mediated by soil properties (e.g., pore space and texture). Fine texture soils, such as clay, are capable of holding more water than course sediments. However, because of the tight binding of water molecules in clay soils, it takes a much higher difference in water potential between soil and roots for plants to utilize the water. The capacity of plants to utilize water from the soil is determined by variation in water potential across the soil–root interface. If the water potential of roots is lower than the water potential of the soil, water moves from the soil to the plant and the plant grows. Plants in turn are the primary producers that are the basis of all trophic levels, produce oxygen, act as carbon sink, and provide essential habitat for numerous organisms.

Seasonal Vegetation as a Food Source

Desert habitats are defined as regions of the Earth's surface that receive less than 50 cm $year^{-1}$. Throughout desert environments much of the landscape is covered by open soil with sparse patches of vegetation. However, these areas are also dotted with springs where surface water is available throughout the year, as well as sites where water is captured during rain events, or runs at the surface, for several months at a time. The density of these water sources across the landscape varies depending upon the underlying geology of the area as well as the location of aquifers. Throughout much of the desert, vegetation is largely driven by seasonal rain events that bring patchy but predictable amounts of plant growth to the area. As water begins to dry up and become less available, vegetation begins to die off. The ebb and flow of vegetation in desert habitats has dramatic impacts on food availability for numerous species. Mammals are a conspicuous group of organisms that are significantly impacted by water availability in desert region. For example, as vegetation dries up, jackrabbit decline as individuals die of dehydration and malnutrition. However, populations are replenished by copious reproduction after seasonal rains when vegetation is once again abundant. Some species of kangaroo rats are highly adapted to living in desert areas and do not drink surface water, rather they get their water from seeds as well as feeding upon green vegetation during wet periods. It is believed that the water kangaroo rats get from eating green foliage plays a large role in supporting energetic costs of reproduction. Thus, in the case of the kangaroo rat, subsurface water availability provides both food and water requirements via plant growth.

As water availability decreases, due to factors such as reduced input and/or lowering of the aquifers through ground pumping, the impacts on vegetation are significant and have cascading effects on assemblages of species throughout the ecosystem. A classic example of this effect is the reduction in avian reproduction during drought years. Over a 2 year period in coastal California, avian reproduction output has been shown to drop from 2.37 fledglings per pair the first year to 0.07 fledglings per pair the second year. The low levels of nesting success may have been partially due to physiological constraints of not having enough water to utilize for egg production. However, the drop in reproductive success was likely due to the fact that arthropod abundance (food source for birds) was significantly impacted because plant growth was hindered as well. This illustrates an example of how water availability plays a fundamental role in shaping the community structure of organisms in an arid environment.

Influence of Spatial Heterogeneity of Water

The spatial distribution of water across the landscape is an important factor influencing the movement of organisms and, ultimately, the connectivity of populations. For example, many species of mammals cannot go without water for more than a few days; thus, their home ranges are limited by distances between reliable water sources. When water is abundant, home ranges of many organisms are significantly larger and there is more connectivity among populations as animals are able to move great distances.

Reduced water availability across the landscape can act as an isolating mechanism by separating populations. As populations become isolated, they are less likely to be recolonized or supplemented by immigration. In many arid regions, springs provide the only reliable year-round water sources. As spring habitats exhibit cycles of increase and decrease over the years, so do the populations of organisms that depend upon them. The temporal variation in the presence and absence of springs can impact species by spatially connecting or isolating disparate populations that are restricted to a particular range due to limited water availability. When springs go dry, due to natural or anthropogenic effects, the metapopulation structure of various species is altered. Thus, the spatial distribution of water bodies plays a critical role in the distribution and abundance of organisms across the landscape.

Conservation

Freshwater ecosystems are some of the most diverse environments in the world. Thus, the global loss of freshwater habitats poses a significant threat to plants, animals, and humans. In North America, 27% of freshwater fauna are considered threatened with extinction. In fact, United States Environmental Protection Agency reports that all watersheds analyzed in the continental United States have at least one species at risk. This fact is not hard to imagine when one considers that nearly half of all the wetlands in the United States are gone. In addition to playing a critical role in the survival of all living organisms, freshwater habitats serve as

biological filters, degrading contaminants through structural filtration, and chemical and photodegradation, as well as provide a significant increase in flood storage capacity, aquifer recharge, and surface water reuse. As water availability is reduced, community composition of organisms will shift to adapt to different conditions. The resulting shift in community structure will have significant impacts on species assemblages and will greatly alter the food web as well as the physical structure of freshwater habitats. In the coming century, climatic shifts and anthropogenic alterations of freshwater are likely to have profound effects on not only threatened and endangered species, but also ecosystem function. Preservation and restoration of freshwater ecosystems must be a top priority in order to maintain biodiversity as well as critical ecosystem processes.

Summary

The availability of water has numerous direct and indirect effects on the ecology of all living organisms. The presence and duration of water that is available for plant utilization has direct impacts on plant growth, which in turn regulates primary productivity and provides critical resources for numerous species. Freshwater habitats ranging from small temporary pools to permanent lakes support a tremendous amount of aquatic diversity. Community structure of freshwater habitats is largely determined by the duration a site remains inundated and frequency of desiccation. The pool of species that are able to exist in a particular site is determined by both abiotic and biotic factors. Abiotic factors control how long a site remains inundated, and biotic interactions play an important role in determining the species composition of each site.

Further Reading

Bolger DT, Patten MA, and Bostock DC (2005) Avian reproductive failure in response to an extreme climatic event. *Oecologia* 142: 398–406.
Brooks JL and Dodson SI (1965) Predation, body size, and composition of plankton. *Science* 150: 467–478.
Colburn EA (2004) *Vernal pools: Natural history and conservation*. Blacksburgm, VA: McDonald and Woodward Publishing Company.
Dayton GH, Saenz D, Baum KA, Langerhans RB, and DeWitt TJ (2005) Body shape, burst speed and escape behavior of larval anurans. *Oikos* 111: 582.
Hui-Mean F, Yusop Z, and Yusof F (2018) Drought analysis and water resource availability using standardized precipitation evapotranspiration index. *Atmospheric Research* 201: 102–115.
Kadlec RH and Knight RL (1996) *Treatment wetlands*. Boca Raton, FL: CRC Press.
Liu D, Guo S, Shao Q, Liu P, Xiong L, Wang L, Hong X, Xu Y, and Wang Z (2018) Assessing the effects of adaptation measures on optimal water resources allocation under varied water availability conditions. *Journal of Hydrology* 556: 759–774.
Mitsch WJ and Gosselink JG (2000) *Wetlands*. New York: Van Nostrand Reinhold.
Molles MC (2005) *Ecology: Concepts and applications*. Boston, MA: McGraw Hill.
Morin PJ (1983) Predation, competition, and the composition of larval anuran guilds. *Ecological Monographs* 53: 119–138.
Nagy KA (1994) Seasonal water, energy and food use by free-living, arid habitat mammals. *Australian Journal of Zoology* 42: 55–63.
Newman RA (1994) Effects of changing density and food level on metamorphosis of a desert amphibian, Scaphiopus couchii. *Ecology* 75: 1085–1096.
Relyea RA and Werner EE (2000) Morphological plasticity in four larval anurans distributed along an environmental gradient. *Copeia* 2000: 178–190.
Skelly DK (1997) Tadpole communities. *American Scientist* 85: 36–45.
Snodgrass JW, Komoroski MJ, Bryan AL, and Burger J (2000) Relationships among isolated wetland size, hydroperiod, and amphibian species richness: Implications for wetland regulations. *Conservation Biology* 14: 414–419.
Valenzuela D and Macdonald DW (2002) Home-range use by whitenosed coatis (Nasua narica): Limited water and a test of the resource dispersion hypothesis. *Journal of Zoology* 258: 247–256.
Wang W and Fu J (2018) Global assessment of predictability of water availability: A bivariate probabilistic Budyko analysis. *Journal of Hydrology* 557: 643–650.
Wellborn GA, Skelly DK, and Werner EE (1996) Mechanisms creating community structure across a freshwater habitat gradient. *Annual Review of Ecology and Systematics* 27: 337–363.
Wurbs RA and Hoffpauir RJ (2017) Environmental flow requirements in a water availability modeling system. *Sustainability of Water Quality and Ecology* 9–10: 9–21.

Anthropospheric and Anthropogenic Impact on the Biosphere

S Pegov, Russian Academy of Sciences, Moscow, Russia

Introduction

Industrial growth proceeded at such a fast pace that in the second half of the eighteenth century it became globally important and resulted in what was called the industrial, or second technological, revolution. Approximately 100 years later, the use of new sources of raw materials and energy brought to life high-efficiency technologies of mass production to produce machine tools and consumption goods. In the later part of the twentieth century, scientific and technical progress stimulated development of high technologies and the advent of space, petrochemical, electronic, pharmaceutical, and other industries. Further progress has brought enormous achievements in the field of information technologies. The rates of dissemination of new technological achievements and economic growth were amazing. Unparalleled high rates of technological development led to a multifold increase in industrial production and consumption of energy resources. The gross world product increased from about US$ 60 million up to US$ 39.3 billion (more than 650 times) between 1900 and the end of the twentieth century. If it took several millennia for agriculture to win the world, then the industrial revolution became a global phenomenon within 1.5–2 centuries.

There were unprecedented rates achieved of burning fossil fuels that had been created by ancient biospheres during a long geological history. For the period from 1950 to 1998, the consumption of various kinds of fossil fuels, expressed in the oil equivalent, increased by 2.1 times for coal, 7.8 times for oil, and 11.8 times for natural gas. While per capita energy consumption was 4000 kcal d^{-1} in the Stone Age, it rose to 12 000 kcal d^{-1} during the era of agricultural technologies, and reached 23 000–250 000 kcal d^{-1} at present. Technogenic interventions in the environment began to compete with many natural processes. Extraction of solid minerals and, hence, the massive impact on the lithosphere sharply increased. About 100 billion tons of raw material is excavated from the Earth's crust annually, or 15 t per inhabitant of our planet.

Studies of ice cores taken from depths of glaciers in Antarctica and Greenland show that such rates of change in biogenic concentrations in the atmosphere did not happen for more than 150 000 years during the overall modern Holocene period.

Studies of carbon isotopes, C^{13} and C^{14}, show that the growth in CO_2 concentrations in the atmosphere for the recent decades is connected with combustion of mineral fuels (Fig. 1). Thus, a huge amount of carbon – up to 180 Gt – had been emitted in the atmosphere as a result of various forms of human land use since its establishment as a planetary phenomenon before 1980, while industrial emissions from the period of industrial revolution to 1980 contributed only 160 Gt of carbon. Thus, a share of land use in CO_2 concentration changes in the atmosphere exceeds 50%.

However, if one compares anthropogenic contribution to the basic biogeochemical cycles, which constitute `biosphera machina' (see more about it below), they do not appear to be too great. At the same time, we feel that there is something odd in our human environment, which leads; us to be concerned about a potential ecological crisis. What is the impact of a dominant anthroposphere on the ecosphere? Is harmonious coexistence of the anthroposphere and the ecosphere possible?

Let us note that unlike such biosphere components as the atmosphere, biota, soils, hydrosphere, and stratosphere, each of which has had more or less clear spatial localization, the anthroposphere has lacked it and has always permeated the above media, even penetrating in the Earth's crust.

World Human Population, Energy Food Demand, and Energy Consumption

It is natural that the intensity of anthropogenic impact on the ecosphere depends (not usually in a linear way) on the size of human population, which grows as shown in Fig. 2.

Two thousand years ago, there were a quarter of a billion people living on the planet. This had doubled to about half a billion by the sixteenth to seventeenth centuries. The next doubling required two centuries (from the middle of the seventeenth century to 1800), the following doubling occurred over only 100 years, while the last one took only 39 years.

Homo sapiens belongs to both the biosphere and anthroposphere. If we consider humans as animals, then all human energy requirements are satisfied through food, and the annual energy food demand per individual is 4×10^9 J. Thus, in the year 2000, the annual energy food demand that determines the annual trophic flow to species *H. sapiens* in the world ecosystem must be 2.4×10^9 J.

The Earth receives 3.5×10^{24} J of solar energy annually, providing the work of the `green cover' with net primary production (NPP) equal to 5.5×10^{21} J yr^{-1} of new biomass. This energy flow also provides a steady state for 1.84×10^{18} g of living

 doi:10.1016/B978-0-444-63768-0.09800-0

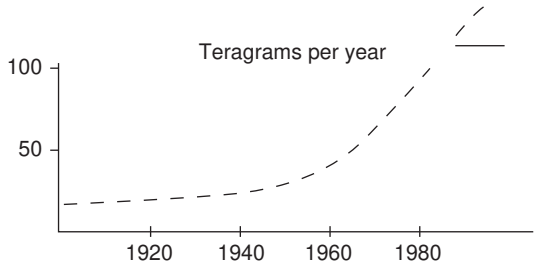


Fig. 1 Natural (solid line) and anthropogenic (dashed line) nitrogen fluxes in the twentieth century. From Vitousek, P.M., 1994. Beyond global warming: Ecology and global change. Ecology 75 (7), 1861–187.

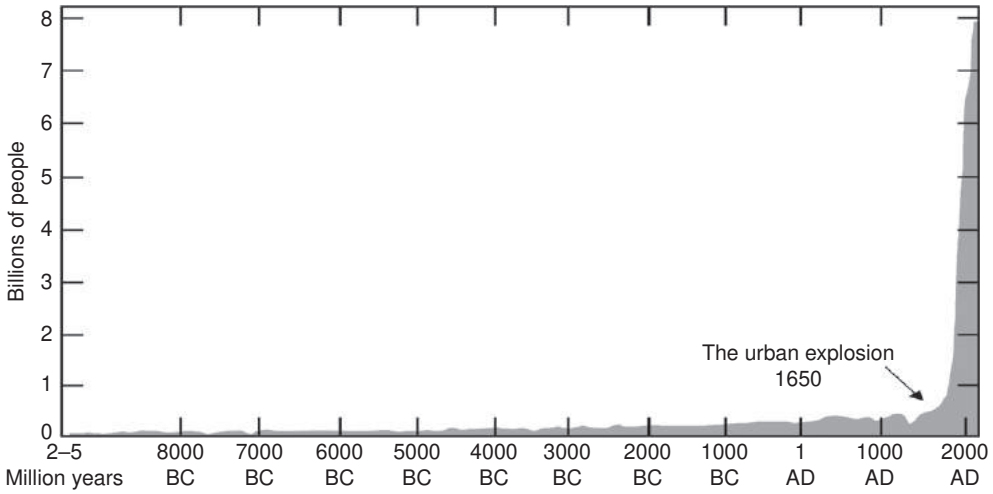


Fig. 2 Dynamics of the world population. From Heinke, G.W., 1997. The challenge of urban growth and sustainable development for Asian cities in the 21st century. AMBIO 8, 130–143.

biomass (or 3.5×10^{22} J), and animal biomass constitutes only 0.8% of it, that is, 1.46×10^{16} g. Animals consume only 3% of the NPP (7.35×10^{19} J yr^{-1}). *Homo sapiens* is one of the animal species with biomass 4.2×10^{14} g (in the year 2000), constituting 2.8% of the total biomass of animals. Therefore, humans can use only $2.8 \times 3 = 0.084\%$ of the NPP, that is, 2×10^{18} J. Thus, the food demand of mankind is more by almost 1 order of magnitude than the trophic flow, that is, the trophic chains including *H. sapiens* are very strained. It may bring in turn either global starvation or destruction of this chain, elimination of many species from the chain (or its elimination in the whole from the global ecosystem).

In 1650, human population was approximately 600 million, that is, an order of magnitude less than today (Fig. 2). From this, it follows that that the trophic flow was equal to food demand, and the corresponding trophic chain was not strained. In other words, humans were still one of many species, coexisting within the biosphere.

On the other hand, if we consider the fate of *H. sapiens* from the point of view of physical theory of fluctuations, the probability of fluctuation, which could cause the elimination of *H. sapiens*, is equal to

$$\mathrm{Pr} = \exp\left[-\frac{\text{energy demand for human population}}{\text{energy supply for all animals}}\right].$$

At the time of the Neolithic revolution, the human population consisted of around 4×10^6 individuals, and required an energy supply of 1.6×10^{16} J yr^{-1}, then $\mathrm{Pr} = \exp[-1.6 \times 10^{16}/7.35 \times 10^{19}] \approx 99.98\%$. If we estimate this probability for the year 2000, we get $\mathrm{Pr}' = \exp[-2.4 \times 10^{19}/7.35 \times 10^{19}] \approx 72.2\%$. Looking at these numbers one can say that *H. sapiens* as a biological species was very fortunate that it has not been eliminated before the anthroposphere arose. Namely, the industrial and accompanying agricultural revolution could mask the consequences of growing strain in the trophic chain.

One of the main characteristics of the anthroposhere is the use of fossil fuels (traces of the past biospheres), and (at present) such `nonbiosphere' energy as nuclear, with an accelerating rate (see Fig. 3).

At the present time, the anthroposphere spends about 3×10^{20} J yr^{-1} to provide for its functioning. This is mainly energy of fossil fuels and nuclear energy (fraction of the `pure' biosphere energy – hydropower station and firewood – in this balance is $\sim 5\%$), and it constitutes about 13% of the global NPP, 2.3×10^{21} J yr^{-1}. Nevertheless, this percentage is enough for the biosphere and anthroposphere to strongly compete for common resources, such as land area and freshwater. Contamination of the environment and reduction of biotic diversity are typical consequences of the competition.

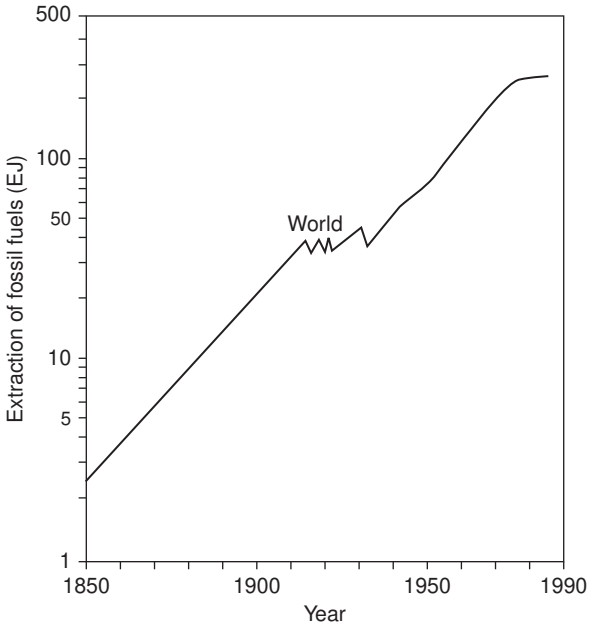


Fig. 3 Accelerating rate of use of fossil fuels and nuclear energy.

Since the biosphere (considered as an open thermodynamic system) is at a dynamic equilibrium, all entropy flows have to be balanced as well. Therefore, the entropy excess, which is created by the anthroposphere, has to be compensated by means of two processes: (1) degradation of the biosphere, and (2) changes in the work of the Earth's climate machine (in particular, through increases in the Earth's average temperature).

The energy of dissipation, corresponding to the full destruction of biota (equivalent to its complete combustion), is equal to 3.5×10^{22} J, while the energy dissipated by the anthroposphere is 3×10^{20} J. Even if the rate of the energy consumption in the anthroposphere does not increase, then this `anti-entropy storage' of biota can make up for the entropy, produced by the anthroposphere, in the next 120 years. If this `technogeneric' entropy could be compensated by soil destruction, then the agony would continue in the course of 300–400 years, since the storage of organic matter in soil is three- to fourfold larger than in biota.

Anthropogenic Impact on the Global Biogeochemical Cycles

It is known that all biogeochemical work of the biosphere is performed by the global biogeochemical cycles. The principal ones, which are, in particular, responsible for the contemporary global climate change, are the global carbon, nitrogen, and sulfur cycles.

Carbon. Gaseous carbon compounds of the global cycle include carbon oxides (CO_2, CO), methane (CH_4), and a great amount of different volatile hydrocarbons that are released as a result of vegetation metabolism and fuel combustion. The main problem here is to estimate flows of the main `greenhouse gases', such as carbon dioxide and methane, into the atmosphere, and their anthropogenic components.

The CO_2 flow into the atmosphere from anthropogenic sources results mainly (75%) from organic fuel combustion (coal, oil, gas) and also from other kinds of economic activities (cement production, flue gas burning), making 20 billion tons yr^{-1}. One should add about 7 billion tons of CO_2 due to annual destruction of forests and loss of vegetative cover. The overall CO_2 anthropogenic flow into the atmosphere reaches about 27 billion tons yr^{-1}, that is, less than 0.01% from the CO_2 total amount in the atmosphere. According to earlier data, the CO_2 anthropogenic emission into the air amounted to 21.3 billion tons yr^{-1} in 1990. Thus, estimating the proportion of anthropogenic and natural components in the CO_2 flux into the atmosphere, one should note that the natural component is approximately 25–30 times more than the human-made one.

Methane inflows to the atmosphere are subdivided into two groups:

- natural biogenic and abiogenic;
- anthropogenic that consists of two subgroups: sources relating to human activity as a biological species and technogenic sources.

An analysis of different data by Adushkin *et al.* in 1998 allows us to conclude that:

1. natural biogenic sources are responsible for an annual average flow of methane equal to about 540 million tons yr^{-1};
2. abiogenic natural sources from lithosphere and hydrosphere make up *c.* 1360 million tons of methane annually (therefore, a ratio between biogenic and abiogenic methane is 1:2.5 in natural sources);
3. anthropogenic sources, including methane resulting from human agricultural activity, losses of methane during extraction of fossil fuels, and its industrial emissions produce an average annual flow of methane equal to about 1100 million tons yr^{-1}.

Table 1 Global gas fluxes in the atmosphere from biosphere and anthroposphere

Source	*CO (bln. t yr^{-1})*	*CH_4 (10^6 t yr^{-1})*	*SO_2 (10^6 t yr^{-1})*	*NO_2 (10^6 t yr^{-1})*	*Total fluxes (bln. t yr^{-1})*
Natural	700	1900	200–300	310–1090	707.41–708.29
Anthropogenic	21.3–27	1100	130–210	30–110	22.92–29.12
Common	721.3–727	3000	330–510	340–1200	730.33–737.41

Therefore, the natural component of methane in the atmosphere estimated at 1900 million tons yr^{-1} is 1.7 times larger than its anthropogenic component.

Nitrogen. There are three kinds of nitrogen oxides – nitrous oxide (N_2O), nitrogen oxide (NO), nitrogen dioxide (NO_2) – and some ammonia. Nitrous oxide has the greatest concentration in the atmosphere (=270–280 ppbv).

Nitrogen oxides reach the atmosphere from different natural sources, such as decomposition of nitrogen-based compounds in the ground by anaerobic bacteria, forest and peat fires, hydrolysis, and sedimentation of nitrates. Nitrogen oxides give rise to aerosols of nitric acid, which is one of the basic components of acid deposits. Total emissions of nitrogen oxides from natural sources are estimated to be 310 million tons yr^{-1}, 540 million tons yr^{-1}, or 1090 million tons yr^{-1} depending on the source.

Sources of the anthropogenic flux of nitrogen oxides are industrial emissions of thermal power stations, chemical and iron and steel industry enterprises, waste dumps of coal and sulfur mines, motor transport, burning of biomass, etc. Total emissions of nitrogen oxides from anthropogenic sources are estimated to be from 30–55 million to 100–110 million tons yr^{-1}.

Therefore, a ratio of anthropogenic and natural components in a flux of nitrogen oxides is 1:10, that is, the anthropogenic flux is 10 times less than the natural one.

Sulfur. In nature, sulfurous gas, hydrogen sulfide, and other gaseous compounds containing sulfur are formed in large quantities as a result of processes of biological decomposition, decomposition of sulfur-containing ores, volcanic activity, and geothermal sources. Hydrogen sulfide getting in the atmosphere is quickly oxidized to make sulfurous gas; therefore, it can be considered one of the significant sources of SO_2.

A wide spectrum of gaseous sulfur compounds is released in the atmosphere after eruptions of volcanoes. Over a 25-year period, annual SO_2 emissions by subareal volcanoes changed from 10 to 30 million tons yr^{-1}. Volcanoes are responsible for approximately 7% of sulfur compounds getting to the atmosphere.

Thus, a total flux of gaseous sulfur compounds from natural sources (mainly gaseous sulfur dioxide) is estimated at 200–300 million tons yr^{-1}.

Anthropogenic sources of gaseous sulfur compounds are metallurgical enterprises, thermal power stations, cheminasescal and coke plants, oxidated landfills of collieries and sulfidic ores, transport, and explosive works. In addition, anthropogenic hydrogen sulfide is formed at factories manufacturing kraft pulp, mineral oil and natural gas treatment facilities, and enterprises making artificial silk and nylon. Global emissions of anthropogenic sulfur dioxide increased during 1950–90 from 20 to 160 million tons yr^{-1}.

The total emissions of anthropogenic sulfur oxides in the world are estimated at 130–200 million tons yr^{-1}. As a result, we observe that the anthropogenic flux of sulfur oxides is practically same, as its natural counterpart. Hence, an impact of anthropogenic sulfur oxide emissions on the environment, in particular, as regards atmospheric pollution, is comparable to the one from natural sources (**Table 1**).

Anthropogenic Impact on Chemical Composition of the Biosphere

The biosphere represents an immense equilibrium system of chemical reactions. Perturbation of the equilibrium at one site may provoke uncontrolled change in the whole system, in spite of the fact that there are different compensating mechanisms (Le Chatelier's principle). We can say that chemical activity of mankind is almost compared now with the chemical work of all living matter. For instance, about 10^{17} g of minerals are excavated annually from the Earth; this value already constitutes 5.5% in relation to 1.84×10^{18} g of all living biomass.

This is in regard to the so-called `gross' characteristics; if we look at `information' ones, in particular atomic composition of excavated matter, then one can see that its composition significantly differs from the compositions of living matter, soil, and oceanic waters. Note that all these minerals are dispersed finally over the Earth surface. The impact on the metal cycles is most significant (**Table 2**).

Our technocivilization is a civilization of iron. About 10% of iron used is destroyed as a result of corrosion, friction, etc. If the amount of lost iron increases by a factor of 2, then, in accordance with our table, soil concentrations of lead increase more than tenfold, and mercury concentrations by 100 times, with toxic contamination of these substances.

Global Land Use: Agriculture and Urbanization

One of the main spatial factors of anthropogenic impact on the biosphere is the rapid growth of agricultural lands, with accompanying change in their land use. Human activity to produce food leads to the reduction of areas of habitat for natural organisms and to a sharp increase in the area of marginal ecosystems. Improvement of agricultural technologies and wide application of fertilizers led to a fourfold rise in land productivity and sixfold rise of agricultural yield in the twentieth century.

Table 2 Relation of metals in soil, ocean, living matter, and world economy with respect to iron concentration

Element	*Soil*	*Ocean*	*Living matter*	*World economy*
Fe	1	1	1	1
Al	1.8	1	0.5	1.5×10^{-2}
Be	1.5×10^{-4}	6×10^{-5}	Traces	2×10^{-5}
Cr	5×10^{-3}	2×10^{-3}	1×10^{-2}	2×10^{-2}
Mn	2.1×10^{-2}	2×10^{-1}	1×10^{-1}	4×10^{-3}
Co	2.5×10^{-4}	5×10^{-2}	2×10^{-3}	3×10^{-4}
Ni	1×10^{-3}	2×10^{-1}	5×10^{-3}	4.5×10^{-4}
Cu	5×10^{-4}	3×10^{-1}	2×10^{-2}	1×10^{-2}
Zn	1×10^{-3}	1	5×10^{-2}	5×10^{-3}
Mo	5×10^{-5}	1	1×10^{-3}	3×10^{-5}
Ag	2.5×10^{-6}	3×10^{-2}	Traces	1.7×10^{-4}
Sn	2.5×10^{-4}	3×10^{-2}	5×10^{-3}	1.3×10^{-4}
Sb		5×10^{-2}	Traces	3×10^{-5}
W		10	Traces	2.5×10^{-5}
Hg	2.5×10^{-5}	3×10^{-3}	1×10^{-5}	1×10^{-5}
Au		4×10^{-4}	Traces	3×10^{-6}
Pb	2.5×10^{-4}	3×10^{-3}	5×10^{-3}	4.5×10^{-3}

Vinogradov, A.P., 1959. Chemical Evolution of the Earth. Moscow: USSR Academy Scientific Publisher.

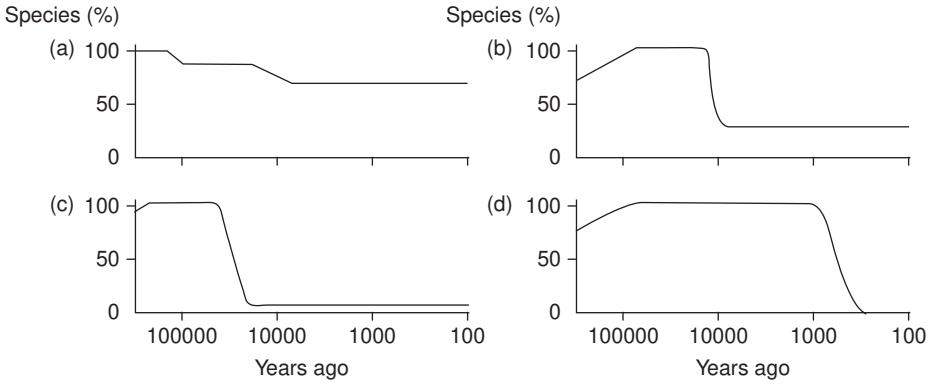


Fig. 4 Loss of large animal species in Africa (a), North America (b), Australia (c), Madagascar and New Zeeland (d) (The World Environment, 1992).

However, this was accomplished by reducing populations of organisms and biodiversity of natural ecosystems (Fig. 4). The biomass of agrocenoses never reaches the biomass of forests, while agrocenosis productivity is lower than that of natural ecosystems. Replacement of natural ecosystems by agrocenoses results in an 11.7% loss of the net primary product, while about 27% of NPP is lost in all human-degraded ecosystems.

About 23% of all usable lands in the world are subject to degradation, which leads to a reduction in its productivity. Agricultural technologies also lead to the destruction of a mid-term reservoir of biogenes, that is, soils. Significant amounts of soil are washed away. As a result of desertification, about 3% of NPP is lost, but soil organisms essentially suffer since they perish due to soil erosion and compression by agricultural implements, plowing, and application of fertilizers. For example, administration of nitrogen in the ground amounting to 3 g m^{-2} a year, with an unchanging amount of other fertilizers, would reduce the population of species by 20%–50% (Fig. 5).

Cities exert a spatially concentrated impact on the environment. While the world population has grown, since 1976, by 1.7% a year on average, population of cities increased by 4% annually. Accelerated urban growth leads to pollution of water, soil, and the air, making their inhabitants live in an unfavorable ecological and social environment. In addition, urbanization is accompanied by a sharp decrease in resistance of urban area territories to technogenic and technonatural hazards. This raises risks of urban dwellers and requires huge efforts of municipal authorities to maintain viability of urban infrastructure.

Industrial Revolution, Anthropocentrism, and the Biosphere Degradation

Industrial revolution unequivocally established an anthropocentric ideology in the human–nature relations. Humans placed themselves at the center of the biosphere, giving it a role of a huge pantry from which it is possible to extract resources beyond all

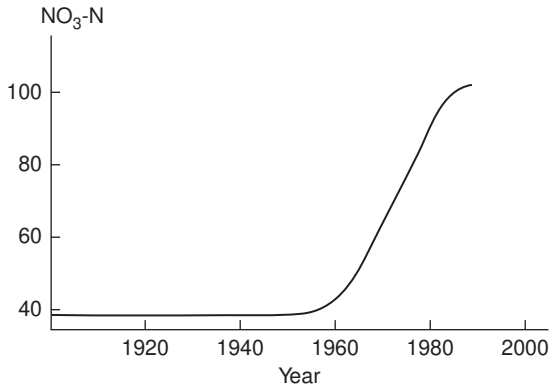


Fig. 5 Change in concentration of nitrogen compounds in estuary of the Mississippi River since the beginning of the twentieth century. From Vitousek, P.M., 1994. Beyond global warming: Ecology and global change. Ecology 75 (7), 1861–1876.

Table 3 Human-disturbed terrestrial ecosystems (not including glaciers and bare lands)

Land area	*Undisturbed area*	*Partly undisturbed areas*	*Totally disturbed area*
134 904 471 km^2	27%	36.7%	36.3%

bounds and, in return, store resulting waste. From the point of view of preservation of the global ecosystem, such relations are unpromising. Calculations show that the twenty-first century will see the exhaustion of many kinds of natural resources of our planet with perhaps unrealistic expectations that further technological advances and economic growth will open up new vistas for solving environmental problems.

Environmental degradation in the latter part of the twentieth century reached global scales. Notwithstanding that about US$ 1.2 billion was spent over the 20 years between the UN conferences in Stockholm (1972) and Rio de Janeiro (1992) on environmental protection, the state of the Earth's environment was worsening. Industrial development that should have strengthened economic advances went into contradiction with the environment since it failed to take into account real limits to biosphere sustainability. Two opposite trends prevail in the global economy: gross world income is growing while the global wealth (first of all, life-supporting resources) is shrinking.

Industrial revolution has led to further pressure of technically and technologically equipped humans on the environment and has created conditions for a new ecological crisis. The consequences of such processes are hard to predict. It is clear that the coming crisis will essentially differ from the previous crises.

Data on disturbed ecosystems is also given in **Table 3**.

Conclusion: Philosophy of the Biosphere

A concern over an imminent catastrophe is growing in the enlightened sectors of society. One of the first among the outstanding thinkers who have realized all the gravity of consequences of industrial revolution was Vernadsky, who developed a scientific concept about the biosphere as a synthesis of knowledge about humans, biology, and sciences about nature, closely connected historically. Dominant in this doctrine is belief in an indestructible power of scientific ideas as a planetary phenomenon capable to reconstruct the biosphere in a noosphere – the sphere of reason.

Many scientists and public and political leaders have understood this idea as a philosophical doctrine of the future development of the world. At the same time, the doctrine about a noosphere remains hardly worked out even at the conceptual level. At the world summit in Rio de Janeiro (1992), an attempt was made to suggest a global program of development of civilization. The document accepted at the conference was named as a concept of sustainable development.

The biosphere as a self-developing system for all its history has gone through a large number of local and global crises, every time reviving and continuing its development at a new evolutionary level. Humans as any biological species are temporary inhabitants on the Earth. Studies of biologists show that mechanisms of constant change of species incorporated in evolution of fauna provide existence in the biosphere of one species during about 3.5 million years on average. Therefore the modern human – Cro-Magnon man – that appeared 60 000–30 000 years ago as a biological species is at its initial stage of development. However, his activity for rather a short term placed him against the biosphere and he created conditions for an anthropogenic crisis.

Considering prospects of the postindustrial development of society, it is necessary to return to ecological understanding of sustainable development. Development can be considered sustainable if it remains within the limits of economic capacity of the biosphere, and maintains its functions as a self-organized and self-adjusted system.

See also: Evolutionary Ecology: Coevolution. Global Change Ecology: Biosphere: Vernadsky's Concept. Terrestrial and Landscape Ecology: Anthropogenic Landscapes

Further Reading

Barnola, J.M., Pimienta, P., Korotkevich, Y.S., 1991. CO_2 climate relationship as deduced Vostok ice core: A re-examination based on new measurements and re-evolution of the air dating. Tellus 43B (2), 83–90.
Coldy, M.E., 1990. Environmental management in development: The evolution of paradigm. World Bank Discussion Paper No. 80. Washington, DC: The World Bank.
Dobrecov, N.L., Kovalenko, V.I., 1995. Global environmental changes. Geology and Geophysics 36 (8), 7–29. (in Russian).
Golubev GN (2002) Global Ecological Perspective-3: Past, Present, Future. UNEP Moscow Interdialekt (in Russian).
Hannah, L., Lohse, D., Hutchinson, Ch., Carr, J.L., Lankerani, A., 1994. A preliminary inventory of human disturbance of world ecosystems. AMBIO 4–5, 246–250.
Heinke, G.W., 1997. The challenge of urban growth and sustainable development for Asian cities in the 21st century. AMBIO 8, 130–143.
Jorgensen, S.E., Svirezhev, Yu M., 2004. In: Towards a Thermodynamics Theory for Ecological Systems. Amsterdam: Elsevier, p. 370.
Laverov, N.P., *et al*., 1997. In: Global Environment and Climate Change. Moscow: Minnauki of Russia, RAN (in Russian), p. 430.
Pegov, S.A., Homiakov, P.M., 2005. In: Influence of the Global Climatic Change on the Economy and Human Health in Russia. Moscow: URSS (in Russian), p. 424.
Tolba, M.K., El-Kholy, O.A., El-Hinnawi, E., Holdgate, M.W., McMichael, D.F. (Eds.), 1992. The World Environment 1972–1992. London: Chapman and Hall, p. 884.
Vernadsky, V.I., 1998. In: The Biosphere. New York: Copernicus.
Vinogradov, A.P., 1959. Chemical Evolution of the Earth. Moscow: USSR Academy Scientific Publisher.
Vitousek, P.M., 1994. Beyond global warming: Ecology and global change. Ecology 75 (7), 1861–1876.
Vitousek, P.M., Erlich, P.R., Erlich, A.H.E., Matson, P.A., 1986. Human appropriation of the products of photosynthesis. Bioscience 36, 368–373.
Zavarzin, G.A., 1995. Circulation of methane in the ecosystems. Nature 6, 3–14. (in Russian).
Zimmerman, P.R., Greenbery, J.P., Wandiga, S.O., Crutzen, P.J., 1982. Termites: A potentially large source of atmospheric methane, carbon dioxide and molecular hydrogen. Science 218 (4572), 563–565.

Biogeocoenosis as an Elementary Unit of Biogeochemical Work in the Biosphere

J Puzachenko, Russian Academy of Sciences, Moscow, Russia

Introduction

Biogeocoenosis belongs to a class of ecological concepts such as phytocoenosis, landscape, units, and sites based on the ideas of spatial uniform units that are distinguished in a given area and separated by visible boundaries. The concept originated by realizing the necessity to study and display the interactions among soil-forming rocks, soil (edaphotope), atmosphere (climatope) with vegetation (phytocoenosis), animal population (biocoenosis), and microorganisms (microbocoenosis). The author of this concept is a Russian geobotanist and paleogeographer V. N. Sukachev.

Biogeocoenosis

The ideas of Sukachev as a geobotanist were close to those of Clements, although Sukachev never recognized phytocoenosis as an organism motivated by the fact that, unlike an organism, elements and parts of phytocoenosis and biogeocoenosis can exist out of the whole. However, he also did not accept the individualistic concept of Glizon-Ramenskii concerning the plant cover organization. According to the author's definition, on a specific area of the Earth's surface, biogeocoenosis is a combination of homogenous natural phenomena (atmosphere, rocks, vegetation, animal and microorganisms, and soil and water conditions). These components possess specific types of interactions and a definite type of interchange of matter and energy occurs between them and with other natural phenomena, thus representing an internally contradictory dialectical unity, being in constant movement and development. N.V. Timofeev-Resovskii determined biogeocoenosis as a biochorological unit, within which there exist no biocoenotical, geomorphological, hydrological, climatic, or pedological–geochemical boundaries. Biogeocoenosis is implied as an integral discrete elementary natural cell of the biosphere that realizes the function of matter and energy transformation. Although the boundaries of each biogeocoenosis may be distinguished according to any of its components, practically it is better to accomplish it using boundaries of the best-observed component, namely vegetation, that is, according to the boundaries of phytocoenosis. Different biogeocoenoses interact with one another in space forming the biogeocoenotic cover. Sukachev did not consider specially the spatial dimension of the biogeocoenotic cover, but as it follows from the context, it corresponds to a rather vast territory commensurable with a floristic district or area. Sukachev discussed in detail the correlation of his concept with Tansley's concept of an ecosystem, different variants of its definition, and the concept of landscape and its morphological units, mainly in the interpretation of the adherents of the Russian school. He paid attention rightly to the fact that an ecosystem is considered (according to Tansley) as an abstract physical system uniting organisms with their environment. It is worthwhile to recall that Tansley actively objected to Clements's holistic concept of organism and considered ecosystem as a set of relations within different spatial–temporal intervals and at different hierarchical levels rather than a reality. Later on, this methodological content of the ecosystem concept disappeared almost completely, and ecosystem has been considered as a natural unit representing a totality of biotic and abiotic elements and as a functional system. Nevertheless, the concept of ecosystem maintains its general meaning along with its traditional interpretation as a chorological unit. Sukachev insisted that the concept of biogeocoenosis as a strictly territorial unit was more definite than the uncertain concept of ecosystem. One can accept this to be true to some extent, but the history of development of science showed that precisely some uncertainty inherent to the concept of ecosystem ensured its viability and incorporation into the general scientific basis. In the light of general system concepts, biogeocoenosis may be considered as a kind of ecosystem which possesses relatively spatially homogenous or stable (random or specific quasi-regular variation) properties in terms of its components within the framework of their observed boundaries. At the same time, the reality and commonness of the distinguished boundaries are not proved specially, but accepted *a priori*, assuming that these boundaries are relatively gradual.

Comparing the concept of biogeocoenosis with the modern concepts of landscape, it is worthwhile to note that the latter are interpreted differently. The concept of biogeocoenosis is most likely to be close to that of units accepted in the Canadian and Australian schools. However, a unit in landscape science is a functional unit rather than an operational one. In American forest science, the notion of biogeocoenosis is comparable territorially with that of stand.

Sukachev, who fully accepted the concept of the biosphere proposed by Vernadsky, regarded biogeocoenosis as an elementary cell of the biosphere.

Researchers who accepted the concept of biogeocoenosis differentiated between the spatial structural elements of biogeocoenosis: vertical layers and horizontal occasionally or quasi-regularly alternating parcels (parts), which are usually distinguished by the shrub, grass, and moss layers commensurable with microassociations. The genesis of parcels was mainly related to the heterogeneity of the tree layer, and they may be associated with gaps. Different parcels are often related to different pedons of soil. Sometimes, parcels are determined by the initial pattern of the nanorelief and soil-forming rocks.

 doi:10.1016/B978-0-444-63768-0.00722-8

Biogeocoenotic Process

Sukachev considered studies of biogeocoenoses as an independent science – biogeocoenology, which studies the biogeocoenotic process. The idea of the biogeocoenotic process proper, being formed after the definition of the notion `biogeocoenosis', is the content of the concept and supplements the Vernadskii ideas for the local level of the biosphere organization. In addition, the elaboration of the concept of biogeocoenotic process was based on the ideas of materialistic dialectics (in this aspect, Sukachev was close to Tansley), kibernetics, and systems theory affecting greatly the development of science in the 1960s. The abundant experience of Sukachev himself as a paleobotanist, geobotanist, geographer, and naturalist was also of great importance.

The dialectic law of the unity and conflict of opposites is postulated as the basis of self-development of biogeocoenosis, with its existing discontinuities or disruptions, destruction of the old, and initiation of the new. Although a biogeocoenosis is an open system, all of its components together still form a certain integral dialectical unity characterized by internal contradictory interactions, which never produce a state of equilibrium within that unity (system). Climatope, edaphotope, phytocoenosis, zoocoenosis, and microbocoenosis are considered as components of biogeocoenoses.

The action of these interior forces leads to self-development, whereas the effect of the external ones leads to some variation and disturbance of the developmental process proper. It is practically useful to consider the mechanisms of nonequilibrium thermodynamics discussed by I. Prigogene, and the dynamics of nonlinear dissipative systems with positive and negative feedbacks capable of innovations discussed by G. Hacken.

The biogeocoenotic process is understood as a change in the matter and energy exchange due to the interaction of organisms with each other and with the environment, as well as between components of biogeocoenosis. The biogeocoenotic process includes not only interactions and exchange of matter and energy between biogeocoenose components, but also interactions and exchange of matter and energy between biogeocoenoses and their surroundings – the environment, in which they exist, and other biogeocoenoses (both adjacent and more remote ones). Since the process of interaction of a biogeocoenosis with its environment is partly expressed in terms of the incessant outflow of energy into space, it has, as it were, an entropic character. But, at the same time, new matter and energy are constantly entering the biogeocoenosis. A biogeocoenosis is considered as an elementary cell, and the biogeocoenotic process in each biogeocoenosis is typical due to specific relations between the biogeocoenosis components and the interaction with their environment. Under similar environmental conditions, biogeocoenoses with similar composition and structure also realize similar biogeocoenotic processes. Evidently, this model is the basis for the development of spatial hierarchical organization of biogeocoenosis, and the author of the concept suggests that the biogeocoenotic cover is a set of interacting biogeocoenoses over a rather vast territory.

The biogeocoenotic process unites four relatively independent processes:

1. The interactions of biogeocoenosis components and elements among themselves, which do not remain constant, but change in time and alter the course of the biogeocoenotic process. This process is a purely internal one and may be called `endocoaction'.
2. The introduction of microorganism germs, plants or new species of organisms by wind and water, and of some organisms from outside, which can change somewhat the biogeocoenotic process. This process was proposed to be called `inspermation'.
3. The introduction of mineral and partly organic matter with dust, surface, and intrasoil runoff. This process is called `inpulverization'.
4. The removal of mineral and organic matter by water and other organisms. This process is called `expulverization'.

The process of internal interactions never ceases; it slows down or accelerates to some extent. The slowing down is determined by a gradual increase in the resilience of biogeocoenosis, but the acceleration is determined by the disturbance of this stability via both settling of new species and changes in the structure of the interactions in the course of self-development. The second and third processes change at the level of the biogeocoenotic cover resulting from climatic and geodynamic fluctuations and asynchronous self-development of neighboring biogeocoenoses as well. The fourth process may be considered as an irreversible one to a considerable degree, and if it is not compensated for the third process, the changes in biogeocoenosis are determined by slow but permanent removal of mineral and organic substances from it. Finally, within the biogeocoenotic cover, the process of formation related to the origin of new phenotypes, genotypes, and morphofunctional forms of organisms is also realized.

A rather strict definition of the biogeocoenotic process as a change of states determined by different mechanisms allowed Sukachev to construct a harmonious classification of the dynamics of biogeocoenoses and biogeocoenotic cover on the following basis: equilibrium process with natural reversibility; nonequilibrium irreversible process; self-development (autogenous or endogenous processes), processes under the influence of external forces (exogenous); according to variation in time and space.

The classification of types of dynamics of the forest biogeocoenoses elaborated by Sukachev is given below:

A. Cyclic (periodic) dynamics of forest biogeocoenoses (reversible changes in forest biogeocoenoses).
   (1) Daily changes in biocoenoses.
   (2) Seasonal changes in biocoenoses.
   (3) Annual (weather) changes in biocoenoses.
   (4) Changes in biocoenoses due to the process of regeneration and growth of woody and other vegetation:
      (a) regular regeneration of woody plants;
      (b) irregular (wave) regeneration of tree stands;

(c) synusial dynamics, especially parcel dynamics (these variants of the dynamics were likely to be associated with a gap dynamics model; models of these types of relationships reproduce usually restricted quasi-cyclic fluctuations of productivity, biomass, and species composition).

B. Dynamics of the forest biogeocoenotic cover of the earth, or successions of forest biogeocoenoses.
   I. Autogenous (irreversible) successions of biogeocoenoses (developments of the forest phytogeosphere, of forest biogeocoenogenesis).
      (1) Syngenetic succession of biogeocoenoses.
      (2) Endogenous (endodynamic) successions of biogeocoenoses.
      (3) Phylocoenogenetic successions of biogeocoenoses:
         (a) phytophylocoenogenetic successions of biogeocoenoses;
         (b) zoophylocoenogenetic successions of biogeocoenoses.

(*Note*. Syngenetic processes are irreversible ones that proceed only due to alterations in the species structure without irreversible environmental changes (typical processes are the development of high bogs, progressive development of eluvial and illuvial horizons of soils). Phylogenetic successions imply processes determined by the origin of new forms. Probably, such processes are useful to be included into the dynamics determined by phylocoenogenesis of viruses and bacteria, including also the saprophytic microorganisms).

   II. Exogenous (reversible and irreversible) successions of biogeocoenoses.
      1. Hologenetic (irreversible) successions of biogeocoenoses:
         (1) climatogenic successions of biogeocoenoses;
         (2) geomorphogenic successions of biogeocoenoses;
         (3) selectocoenogenetic or areogenic successions of biogeocoenoses;
            (a) phytoareogenic successions of biogeocoenoses;
            (b) zooareogenic successions of biogeocoenoses.

(*Note*. Hologenetic processes are realizable at the regional level of the biogeocoenotic cover organization. It is worth noting that Sukachev did not extend the principle of actualism and reversibility to climatogenic, that is, paleoclimatic successions. Selectocoenogenetic successions may appear due to the invasions of alien species. Changes determined by invasions of agents of feral herd diseases of plants, animals, and saprophytic microorganisms are expedient to be included into this type of dynamics. A typical example is the mass and, most likely, irreversible death of American chestnut (*Castanea dentata* (Marsh.) Borkh.) in the Appalachians. If species change their properties in the process of settling, selectocoenogenetic successions are indistinguishable from phylocoenogenetic ones).

      2. Local (reversible and irreversible) catastrophic successions of biogeocoenoses.
         (a) anthropogenic successions of biogeocoenoses;
         (b) zoogenic successions of biogeocoenoses;
         (c) pyrogenic successions of biogeocoenoses;
         (d) windfall successions of biogeocoenoses;
         (e) successions of biogeocoenoses produced by mud streams, landslides, sudden inundations, and other causes.

Probably, this classification of the dynamics may be recognized as the most complete. For the modern ecology, it contains all the bases for particular and integrating models of dynamics and research programs (e.g., programs directed to the accumulation of data on the irreversibility of self-development processes). However, in order that the concept of biogeocoenosis might create the necessary bases for studies and simulation of biogeochemical cycles, it should contain some concrete system definition and refinement of ideas of the spatial–temporal hierarchy and elimination or weakening of contradictions between individualistic and organism concepts of spatial organization of the biosphere and its components.

Biogeocoenosis and the Biosphere

The system that specifies the biogeocoenosis concept is rigorously introduced in works by Vernadsky, who was not only a naturalist, but also a physicist and chemist; he possessed knowledge in thermodynamics and thermostatics. In complete accordance with concepts of thermostatics, he determined an object and its elements in the following way: ``I will call a set of organisms participating in geochemical processes living matter. Organisms composing this set will be elements of living matter. With all this going on, we will pay attention not to all the properties of the living matter, but only to those which are related to its mass (weight), chemical composition, and energy. In such a comprehension, living matter is a new scientific notion''. Later, Vernadsky directly associates individuals with molecules of gases and suggests to consider living matter as a statistical ensemble of elements. Thus determining the concept of the biosphere, he states that laws of equilibrium (equilibrium process) in general mathematical form as revealed by J. Gibbs (1884–87) (who reduced them to relationships between independent variables, such as temperature, pressure, physical state, and chemical composition, which characterize the chemical and physical processes and participate in system processes) could be applied to a living system of bodies. According to this statement, one can distinguish ``thermodynamic spheres as areas of equilibrium of thermodynamic variables that are determined by values of temperature and pressure; phase

spheres that are characterized by the physical state (solid, liquid, etc.) of bodies in their composition, chemical spheres different in the chemical composition. Only one sphere distinguished by E. Suess – the biosphere – remained aside. Undoubtedly, all the reactions of the biosphere follow the laws of equilibrium, but they include a new characteristic, new independent variable which was not taken into account by J. Gibbs and is very important in other equilibrium forms (in the context of thermodynamics). A special reaction is the phenomenon of photosynthesis, with radiant light energy as an independent variable. Therefore, ``living organisms, introducing the radiant light energy to physicochemical processes of the earth crust, drastically differ from other independent variables of the biosphere. Like these variables, living organisms change the course of equilibrium, but unlike them, they represent specific autonomous formations as specific secondary systems of dynamic equilibrium in the primary thermodynamic field of the biosphere. The autonomy of living organisms reflects the fact that the thermodynamic field, which inherently has quite other parameters than those observed in the biosphere. Therefore, organisms retain their own temperature (many organisms do so strongly) within the medium at another temperature and have their interior pressure. They are isolated in the biosphere, and its thermodynamic field is important for these organisms only due to the fact that it determines the area of existence of these autonomous systems, but not their interior field. From the chemical standpoint, their autonomy is expressed in the fact that chemical compounds produced in these systems cannot be synthesized beyond them under usual inanimate conditions of the biosphere. Being fallen into the conditions of this medium, they turned out to be unstable, are decomposed, transformed to other bodies, and in that way, they become disturbers of the equilibrium and represent a source of free energy in the biosphere''. Vernadsky discusses in detail all the properties of living matter known by that time, including basic mechanisms of its evolution. Generalizing his writings and using the modern terminology, one can define living matter as a stationary dissipative system of organism elements, which is far from thermodynamic equilibrium with free energy and exergy. The stationary state of this system is supported by the absorption of solar energy, which is responsible for the permanent conversion of the chemical element flux into a new organic form, realizing the cycle with a release of free energy to the environment, and transforming the latter as a result of useful work (exergy). The simplest example of this work is the intensification of the water cycle in the biosphere with appropriate contribution to climate control, that is, changes of equilibrium correspond to thermodynamic variables that change climate. So, when combining the concept of biogeocoenosis with the concept of `living matter', we obtain rather strict thermodynamic bases for the characterization of the biogeocoenotic process, as well as all the necessary fundamentals for consideration of their autochthonous (endogenous) dynamics and self-development of biogeocoenoses and biogeocoenotic cover as a nonequilibrium, stationary thermodynamic dissipative system. However, all this is insufficient to consider biogeocoenosis as an elementary cell of the biosphere, within which nonliving matter is converted to living one and, conversely, incomplete transformation of the former to mobile chemical compounds occurs with a release of free energy and changes in the environment (thermodynamic variables of the atmosphere, hydrosphere, and lithosphere).

Reductionism and Holism

According to the Gleason–Ramenskii continuum individualistic concept (reductionism), there are no necessary bases for the initiation of spatial cells as relatively discrete formations without any additional conditions. The existence of such cells is the basis of Clements's organism concept (holism). It is worthwhile to note that, strictly speaking, Clements may not be its original author. Even at the dawn of the development of geography, in 1811, Butte stated that none of the scientists had any doubts regarding the existence of earth organisms. Within any specific field, a combination of all the phenomena is not a simple set; they represent a holon. Butte assumed individual countries and districts (including humans), as `organisms', which, as any organism, may be considered both in terms of their physical and psychical aspects. He wrote that ``areas as a holon assimilate the human population'', and ``population assimilates these areas not less constantly''. At the same time, opponents of the hyperholon paid attention to the fact that it was difficult to find districts the boundaries of which could be determined as the basis of all the phenomena. The most complete criticism of this integral concept was given by A. L. Bucher in 1827. As a result, he concluded that there was no necessity to study boundaries, and regions might be distinguished in any arbitrary manner. He proved that geography should study relations between particular phenomena in any area of the earth's surface. Even now the same contradictions exist: on the one hand, the individualistic concept has been fully recognized; on the other, Gaia's superorganism concept is very popular. The criticism of this concept rests on traditional bases and factually repeats the discussion that has been continued for almost 200 years. If to leave aside these disputes, we can state that the two models of living matter – individualistic (reductionism) and organism (holism) – may be considered as those reflected in real natural phenomena. Developing these models up to possible logical limits, in both cases we obtain incompatible constructions. In the first variant, it is a construction similar to Dawkins's selfish gene; in the second one, it is a superorganism with its own purposeful development and superstability similar to Gaia's model. The individualistic model has been well substantiated theoretically and realized in microcosm and perfusion cultures. For the simplest linear variant, a theorem has been proved asserting that, in the homogenous environment, the number of stable coexisting populations is equal to the number of resources or, in general, to the number of any operating factors. The relations following this model were obtained by the methods of ordination for a wide diversity of plants and animals in direct terrain investigations. Particularly, such relations between different layers of a forest community and main tree species were shown for the Eurasian forest zone.

To prove the integrity of biogeocoenosis, ecosystem or plant community and their emergent properties should be understood more completely. Raised bogs may be referred to the formations of this type, the progressive growth of which is

supported by the positive feedback between the groundwater table at the territory adjacent to the bog and development of sphagnum mosses. The accumulation of dead parts of mosses raises the groundwater table, and this process promotes the further moss growth and peat accumulation. Raised bogs form their own dynamic spatial structure, and minimization of moisture evaporation in hot summer months may be accepted as its emergent feature. Such a raised bog in fact resembles a superorganism, which occupies slowly (up to 10 cm per year) the neighboring territories displacing forest communities. True, this superorganism exists primarily due to the almost complete cessation of the cycle of matter, representing an essential deceleration of the water cycle at the exergy lowest for the forest zone. Such organism features are difficult to find for many typical cases. If not to ignore the traditional experience to distinguish phytocoenoses as relatively homogenous spatial formations that indicate biogeocoenosis, their integrity may be accepted as an empirical fact. At the same time, it is admitted that the corresponding mechanisms are poorly known.

From the standpoint of postmodern science, there is no necessity to create a single eternal theory. The most topical concept is one that initiates research and provides foundations for verification of competitive hypotheses, as well as stimulates their diversification and does not eliminate their joint acceptability. From these positions, *a priori* denial of these two models is identical to a nonacceptance of liberal or social views in the organization of human society. It is evident that the individualistic concept is mainly close to the thermodynamic model of the world in its movement to equilibrium and higher entropy. The basis of the individualistic model is maximization of independency of each component and its resilience within the holon that is a rather satisfactory strategy for its survival. But at states far from equilibrium, positive correlation and effects of self-organization and relatively discrete spatial structures arise in the thermodynamic system in accordance to the theory of nonequilibrium thermodynamics. This is a good hypothesis, which is useful to be verified for the biosphere. If to lean upon the theory of dynamic systems, the entire biosphere and its patches may be considered with certainty as nonlinear oscillators of high dimension. From these positions, the efficiency of the fractal model for characterizing the diverse natural processes is well explained. Formally, a fractal set is continuous but undifferentiated, and displays a cascade of bifurcations in the spatial–temporal dynamics of nonlinear oscillators. In nature, it manifests itself in the possibility to distinguish between different-scaled and hierarchically subordinate relatively homogenous formations, boundaries of which may also be divided into such structures. The formal fractal model assumes a self-similar division into indefinitely small units. Real natural objects do not possess this specific feature – their fractality has a finite range of dimensions. Taking into account this property, the model is sufficient for the theoretical definition of a biogeocoenosis as a spatial–temporal cell commensurable with linear dimensions of dominant plant species in it, that is, including some minimal population stable at least in one generation. Direct measurements of the fractal landscape cover the structure using data of remotely sensed investigations and three-dimensional models of relief show that, almost everywhere, a fractal spectrum connecting the amplitude of spatial variation of the variables measured with the spatial wave number reveals quasi-harmonic fluctuations, but describes only some percentage of the spatial variation. However, the relative peaks of the spectrum, corresponding to definite linear sizes, allow correcting a choice of scale for different hierarchical levels. The nature of local spatial homogenous structures may be determined by the organization of relief, soil-forming rocks, soil, dynamics of vegetation, effects of animals, tree windfalls, fires, and so on. The spatial–temporal dynamics of each of these components are stipulated by both their own fluctuations and those originated due to their interactions. As a result, the spatial structure is fractal, and distinguishing the relatively even territories is possible and strictly realizable on the basis of classification of multispectral images, in particular. In the framework of the model of a nonlinear oscillator, the individualistic concept does not contradict the different-scale discontinuity, and although nonlinear oscillators produce effects of self-organization under definite conditions, these models do not contain mechanisms of structural stability. On the basis of these models, a holistic model is impossible to construct. On the other hand, using the multifractal model, the proportion between total energy, free energy, and entropy is deduced resulting in the natural generalization of two models of reality. Following this method, the concept of biogeocoenosis as an elementary cell of the biosphere may be of constructive importance in both the organization of terrain investigations for assessment of biogeocoenotic and biosphere processes and the elaboration of corresponding models. The fractal scheme of organization of the biogeocoenotic cover gives prerequisites for the recalculation of parameters obtained in large-scale studies to those corresponding to the high level of organization. In order to obtain the behavior similar to that of an organism, it is necessary to add contours of positive feedback providing relationships between components of the system and supporting system resilience under conditions far from the thermodynamic equilibrium in the environment to the model of nonlinear dynamics. The fact that such relations are realizable in organizing the components of the biosphere was shown from the example of the bog. A similar type of relation holds for a tropical forest that evaporates moisture intensely. The same is true for boreal spruce forest that evaporates more moisture than a deciduous forest and supports low temperatures favorable for spruce due to expenditures of heat for evaporation. The positive feedback is characteristic of mycorrhizae fungi and their hosts. There are many examples of positive feedbacks in a plant community (mutualism). However, the conditions under which they determine holistic features of biogeocoenosis and those of higher levels of its organizations are not evident and need special investigation. At the same time, their potentially significant role in the maintenance of homeostasis in an aggressive medium is evident, as is the nature of spasmodic and catastrophic transformations at small disturbances, primarily in the margin areas of the system tolerance.

Summary

A discussion of the general concepts of ecology and attempting to specify their physical sense are *a priori* ungrateful tasks. In ecology, as in any natural science, notions or definitions had a quite uncertain content at the time of their introduction and they determine an approach to studies rather than their object. Later, these notions and definitions were redetermined many times and differently by different researchers. The multidimensional subject of ecology stipulates such an uncertainty. The uncertainty causes periodically renewed discussions of the theoretical bases of science that are inevitable in formulating the concept of biogeocoenosis. The latter may be considered as a specific one in relation to a more general concept of an ecosystem. In the framework of the modern theory of thermodynamics and nonlinear dynamical systems, accepting the existence of self-similar quasi-discrete territorial units, the concept of biogeocoenosis is interpreted via living matter as a thermodynamic variable. On the other hand, the ideas of the dynamics of biogeocoenosis elaborated by Sukachev represent good bases to formulate verifiable hypotheses. They allow combining the concepts of reductionism and holism as interconnected (but not contradictory) models, the contribution of which to the spatial–temporal dynamics depends on geographic conditions, the current status, and the time of self-development.

See also: Global Change Ecology: Biosphere: Vernadsky's Concept

Further Reading

Abrosov, N.S., Kovrov, B.G., Cherepanov, O.A., 1982. Ecological Mechanisms of Co-Existence and Species Regulation. Novosibirsk: Nauka, 287pp.
Alcock, J., Hoffman, F.M., Schwartz, P.M., 2003. Positive feedback and system resilience from graphical and finite-difference models: The Amazon ecosystem – an example. Earth interactions 7.23pp. Paper No.5.
Hargrove, W.W., Hoffman, F.M., Schwartz, P.M., 2002. A fractal landscape realizer for generating synthetic maps. Conservation Ecology 6 (1), 2.
Hartshorne, R., 1939. The Nature of Geography. Lancaster, PA: Association of American Geographers.
http://www.consecol.org/vol6/iss1/art2 (accessed December 2007).
Ilya, P., 1997. The End of Certainty: Time, Chaos, and the New Laws of Nature. New York: Free Press.
Jorgensen, S.E., 2000. 25 Years of ecological modelling by ecological modelling. Ecological Modelling 126 (2–3), 95–99.
Lovelock, J.E., 1979. Gaia: A New Look at Life on Earth. Oxford: Oxford University Press, 252pp.
Prigogine, I., Stengers, I., 1990. Order Out of Chaos: Man's New Dialogue with Nature. London: Flamingo, (First publ. 1984).
Puzachenko, G., Yu, D'yakonov, K.N., Aleshenko, G.M., 2002. Diversity of landscape and methods of its measurement. Geography and biodiversity monitoring. Series of manuals. In Conservation of Biodiversity. Moscow: NUMTs, 143–302.
Puzachenko, Yu G., Skulkin, V.S., 1982. The Structure of Forest Vegetation. Moscow: Nauka, 320pp.
Sagoff, M., 2003. The plaza and the pendulum: Two concepts of ecological science. Biology and Philosophy 18, 529–552.
Schroeder, M., 1991. Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise. New York: W.H. Freeman and Company, 429pp.
Shugart, H.H., 1984. A Theory of Forest Dynamics. The Ecological Implications of Forest Succession Models. New York: Springer, 278pp.
Sukachev, V.N., Dylis, N.V. (Eds.), 1964. Fundamentals of Forest Biogeocoenology. Moscow: Nauka (Science), p. 574.
Tansley, A.G., 1935. The use and abuse of vegetational concepts and terms. Ecology 16, 284–307.
Turcotte, D.L., 1997. Fractals and Chaos in Geology and Geophysics, 2nd edn. Cambridge: Cambridge University Press, 367pp.
Vernadsky, W., 1929. La Biosphere. Paris: Librairie/Feliz Alcan.
Vernadsky, V.I., 1926. The Biosphere. New York: Copernicus, Leningrad: Nauchtekhizdat (in Russian). English version: Vernadsky VI (1998) *The Biosphere* (complete annotated edn.), 192pp.

Biosphere: Vernadsky's Concept☆

Yuri M Svirezhev and Anastasia Svirejeva-Hopkins, Potsdam Institute for Climate Impact Research, Potsdam, Germany

Glossary

Biosphere Earth's crust and lower levers of Atmosphere, where living matter is present (see text for detailed definition).

Biogeocenosis An interrelated complex of living and inert components associated with each other by material and energy exchange; one of the most complex systems in nature.

Vladimir Ivanovich Vernadsky (1863–1945) Russian mineralogist and geochemist, one of the founders of biogeochemistry.

Introduction

Vladimir Ivanovich Vernadsky (1863–1945) (Fig. 1. V.I. Vernadsky, Paris 1889), the great Russian scientist and thinker, the founder of modern concept of the biosphere has shown that during all geological epochs on Earth, the life was developing as interconnected group of organisms (as he calls it "living matter") that provided and provides the continuous flow of elements in biogenic turnover of matter and energy on the surface of our planet. This could easily be called scientific revolution at that time. Indeed, as Thomas Kuhl points out, scientific revolutions occur when someone creates a new perspective, a model used for understanding reality. Only after introduction of such an idea, great progress, that was previously impossible, could be made, opening new ways of thinking. In recent years it becomes even clearer that all the works of Vernadsky were devoted to the further development of scientific thought as a planetary phenomenon. The significance of his ideas could be compared to the teachings of great philosophers of Ancient Greece, Roman Empire and Eastern world, of the Renaissance; the developers of the basics of mathematics and physics, such as I. Newton; the creators of system thinking about the origin and functioning of life on Earth, C. Linnaeus, G. Buffon, J.-B. Lamarck, C. Darwin, A. Humbolt, G. Mendel, etc.; as well as famous Russian scientists, M. V. Lomonosov, D. I. Mendeleev, I. M. Sechenov, I. N. Mechnikov, V. V. Dokuchaev, and others. As A. E. Fersman, Vernadsky's devoted pupil and successor in the area of geochemistry wrote about his Teacher: "His general ideas will be studied and elaborated during centuries and one will discover new pages in his works which will serve as the source for new searches. Many scientists will learn his creative thought, which is acute, stubborn and articulated, always genial, but sometimes poorly understood. As for young generations, he always will be a Teacher in science and a striking example of a fruitfully lived life."

Fundamental Idea

In Vernadsky's book *Biosfera*, first published in 1926, Biosphere is simply the surface or the outer domain of the planet, separating it from its cosmic surroundings. The author calls it the "Face of the Earth."

But what is crucially new, that in accordance with dialectic principle, the process of cosmogonic evolution of the Earth is considered in the light of dynamics of the environment, which includes the system of many different forms of matter turnover; while its highest form, life itself, is determining other planetary processes, the latter being the very central idea in Vernadsky's teachings. Namely this concept served as a necessary and desired base for the development of modern ecology.

Living and Nonliving Matter, Their Interaction and Cosmogonic View

The first step toward changing the world's picture, as natural scientists see it now, was the introduction of the concept of living matter by Vernadsky, and the second step was considering it as a cosmoplanetary phenomenon. Vernadsky has defined the living matter as "the existing at present time unity of organisms with the mass, chemical composition and energy" connected with its environment by constant processes such as breathing, feeding, and procreation, but we shall further address it sometimes as organic matter. Verndsky often mentioned that life propagates and dissipated non-living matter in a gas-like manner, very fast and thorough. Vasily Dokuchaev, regarded as the "Father of Soil science" had played a big role in a formation of Vernadsky as a scientist. Studying mineralogy and crystallography during his student years, Vernadsky later has classified the dead organic matter

☆*Change History*: March 2018. Yuri M Svirezhev. Changed Fig. 1 to another photo. Some additional thoughts and details that are described in the newly published translation and collected works are interweaved in the text. Some minor text editing were done.

This is an update of Y.M. Svirezhev and A. Svirejva–Hopkins, Biosphere: Vernadsky's Concept, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 467–471.

 doi:10.1016/B978-0-12-409548-9.11154-6

Fig. 1 V.I. Vernadsky, Paris 1889.

of soils and sediments as "bio-generic matter" or bio-inert bodies (the natural structures, consisting of inert and living bodies). One can call it "the remnants of the past biospheres."

The processes of interaction between the living (organic) and nonliving (inorganic) matter, considered as the most important initial stage of cosmoplanetary evolution, could be observed since the very first stages of planet's existence. As an example, proving Vernadsky's generalizations, the recent determination of time of the beginning of formation of primary sedimentary rocks (the *Stratispehere*, as Vernadsky named it) is 3.7 billions years ago, while slightly younger age (3.4–3.5 billions years ago) is determined for the formation of first organic compounds, that is, "islands of living matter." The origin of life on our planet is directly connected to the origin of the biosphere; and evolution, as we perceive it, always takes place inside the biosphere, involving exclusively living matter. However, Vernadsky has expressed the fundamental importance of eternal interconnection between the living and nonliving matter, in the following very significant paragraph: "The Earth cover, Biosphere, while fully embracing the globe, has limits that are strictly determined by the existence of living matter in it—it is populated by it. Between its inorganic "lifeless" and living parts, inhabiting it, continuous exchange of matter and energy exists, expressed by atomic movement caused by living matter. With the time course, this exchange is expressed by constantly changing and tending to steady-state equilibrium. This equilibrium threads through the entire biosphere and this biogenic atomic flux to a large extent creates and maintains it. Hence, in this manner and during all geological epochs, biosphere is connected with the living matter that populates it. And namely by this biogenic flux of atoms and energy, the strong planetary cosmic significance of living matter is determined." This view through the "cosmic prism," so to speak, radically changes our understanding of dialectical interconnection of living and nonliving matter, that originally differ in their composition of elements. As one can add, nowadays there is certainly more data and hypotheses about the influence of solar and other cosmic radiation on the living organisms at different levels of their structural development.

As Vernadsky writes in his Essays on Geochemistry: "Radiations of immaterial substance (pure energy) embrace not only biosphere but all space, with wavelengths ranging from millionth part to a millimeter up to several km. Cosmic radiations received by our planet creating the biosphere extend over only 4,5 of the forty known octaves of frequencies. In the most familiar cosmic radiations those from the Sun we distinguish one octave of light, three octaves of heat (infrared) and half an octave of ultraviolet radiation, the last one being a small remnant that passed through the stratosphere. It seems improbable that the other octaves should be missing in cosmic space. We see their absence as an illusion caused by the absorption of the rays by the rarefied material of the upper layers of the atmosphere."

Indeed, quite a lot of data have been accumulated on the influence of weak electromagnetic fields on the information exchange between living cells. Cosmic rays could in certain way influence the information exchange that is conducted by means of weak electromagnetic fields between cells of a living organism, and therefore alter functioning of the multi-cell structures. "When the substance of the biosphere is penetrated by energy it becomes active, gathers and distributes the energy and turns it, in terrestrial organisms, into free energy capable of performing work. Therefore biosphere is not just a domain of sheer substance, but of energy, the site of transformation of the planet by external cosmic forces. The history of the biosphere is sharply distinguished from that of

the rest of the planet. The larger part of the biosphere is of non-terrestrial origin, since this substance from outside—cosmic dust and meteorites—is of the same structure as terrestrial substance. It is not limited to biosphere, it is the same for the whole crust of the Earth—for the kilometers of lithosphere of which the upper part is the biosphere. Lithosphere and biosphere gradually and inseparably merge with each other. The substance of the deeper parts of the planet is different, it hardly penetrates the biosphere in larger amounts during short periods of time, therefore can be ignored." Vernadsky believes that it is impossible to explain the definite composition of the Earth's crust and the whole planet by the different atomic masses of the elements comprising it. The difference in composition between Earth core and crust is due to the similarity between the composition of meteorites and of deeper layers of the planet. Relatively light elements prevail in Earth's crust but it is also rather heavy in iron, which is related to cosmic history and perhaps to the structure of chemical elements. Vernadsky also stresses the similarities in atomic structure between the surfaces of the Sun, the Earth, and the stars.

"By the way of radiation the elements exert a reciprocal effect on one another." Vernadsky begins to see them not entirely as purely terrestrial phenomena but related to the structure of atoms though their location in the cosmos, and to their evolution in the history of the cosmos. "The biosphere is the place for us for search not only for the reflections of accidental unique geological facts, but also for manifestations of the structure of the cosmos related to the structure and history of chemical atoms.

The mechanism of chemical action of living matter is unknown to us but it is obviously becoming clear that from the standpoint of energy phenomena in living matter, photosynthesis takes place not only in some chemical surrounding but also in a special thermodynamic field of the biosphere where they prove unstable and serve as sources of free energy (the thermodynamical field of living matter) is thermodynamically and chemically different from that of the biosphere".

Two Main Principles of Interactions Between Living and Nonliving Matter

It is necessary also to say a few words here about the two main principles of interactions between living and nonliving matter that Vernadsky has formulated, namely two biogeochemical principles that describe the nature of energy fluxes in the biosphere:

1. Geochemical biogenic energy of the biosphere tends to maximum.
2. During the evolution of species, only the organisms that increase this biogenic geochemical energy in a process of their life, will survive.

Vernadsky also writes about the irreversibility of life's processes and the increase in life's free energy, expressed in dissymmetry of composition of living matter.

In connection to the first biogeochemical principle, it becomes important to mention the similar work of Russian theoretical biologist, E. Bauer, who has formulated the fundamental principle of the permanent inequilibrium of living matter and the principle of maximum effect of external work. These principles, describing thermodynamics of evolution and organization of living matter, are called "the law of Bauer-Vernadsky."

Concept of the Biosphere, Definition of Term, and Method of Analysis

At the beginning of the nineteenth century J.-B. Lamarck had introduced the term "biosphere." He considered it as the "scope of life" and some sort of external cover for the Earth. In 1875 the same term was introduced in geology by E. Suss, who distinguished the biosphere as one of the Earth's covers. But V. Vernadsky was the first person who created the modern concept of the biosphere. This concept was first introduced in his two lectures in Paris, published in 1926. In a sequel it was further developed by Vernadsky himself, and by V. Kostitzin, V. Sukhachev, N. Timofeev—Resovsky, and other Russian scientists.

Biosphere includes all the hydrosphere, troposphere to the height of 30 km, and the upper part of the Earth's crust down to a depth of 2–3 km, for living bacteria still may be found at this depth in the underground waters and in the soil. It is an open thermodynamic system that exists with a permanent flow of solar energy (1.2×10^{22} kcal year^{-1}) since the very beginning of the Earth's history.

According to Vernadsky, the biosphere is an external Earth cover, the "scope of life" (as Lamarck named it). He also notes that this definition (as just the "scope of life") is not complete. Vernadsky's biosphere includes.

1. living matter;
2. "biogeneric matter," that is, organic and mineral substances, created by living matter (for instance, coal, peat, litter, humus, etc.); and
3. "bioinert matter," created by living organisms together with inorganic nature (water, atmosphere, sediment rocks).

Empirical Generalization Method

Two components constitute Vernadsky's concept of the biosphere. The first is the actual biosphere concept, which can be called a verbal model of the biosphere. The second component is the method of study of such a complex system as the biosphere, which he called "empirical generalization method" (EGM). Vernadsky opposes the reliance on mere hypotheses, repeatedly insisting that the

better suited method for a scientist is the Baconian system of accumulation of facts, as generalizations are becoming apparent from the data. That is essentially an inductive method and it indeed lies at the heart of the modern science. A perfect example of this method is Mendeleev Periodic Table of the elements. Certainly, the EGM is essentially wider than simply a method for the study of biosphere processes; it is the general scientific method. Let us remember Descartes' principle that "Science is a method." Speaking in modern terms, the EGM is a typical method of systems analysis.

How the IGM Works

The empirical generalization is based on real facts collected in an inductive way, but always keeping in mind to not leave the domain of these facts. At this first stage all possible scientifically established facts about studied phenomenon must be collected. The next stage is the aggregation of collected facts into some more general categories, called proper empirical generalizations. Basically, it gives us the possibility to pass from a huge number of accumulated facts to a considerably lesser number of statements that, in turn, allows us to truly speak about the possibility to describe the studied large (complex) system quantitatively. This stage does not allow for formulating of any kind of hypothesis, on which there is inevitably a mapping not only of scientific ideas, but also of nonscientific ones. Really an empirical generalization is a system of axioms, reflecting our level of empirical knowledge, which could be used as a basis for any formal theory developed in the future. Hence, having the system of empirical generalizations, we can follow either of two ways to construct models. Either we remain within the framework of this system, constructing so-called "phenomenological" models; or complementing some hypotheses relating to the existing empirical generalizations, we shall get some new models. In accordance with Vernadsky's opinion, in order to choose on the type of these models, one must produce hypotheses, based on coincidence of predicted and newly observed facts. If this coincidence takes place, then the hypothesis becomes an empirical generalization of a higher level. From this point of view, for example, the practical astronomy of Ancient World was a typical empirical generalization, and ancient astronomers were successfully using the phenomenological model created on its basis. The same underlying empirical generalization is the basis of two principally different cosmogonic hypotheses of Ptolemee and Copernicus. If and only if new facts had appeared, the Copernicus cosmogony would have become a new empirical generalization. Therefore, the same empirical generalization can be a basis of different models.

However, the reciprocal picture is possible, when an empirical generalization exists separately, without some kind of hypotheses and explanation from the viewpoint of contemporary science. For example, the radioactivity phenomenon could not have been explained by the state of the nineteenth century physics.

The System of Axioms

Which empirical generalizations constitute the base of Vernadsky's biosphere? In this case we will call this system of axioms "Vernadsky's biosphere;" however, these axioms will be presented in slightly more formal way than in Vernadsky's original work.

1. *During all geological periods on Earth, living organisms have never been created directly from inorganic matter*. This is the homogeneity axiom. Note that in mathematics, the operators that transform a zero into zero are called homogeneous, too. There is also an analog of this axiom in biology, called the Redi's law ("life comes only from life").
2. *The presently known facts cannot answer the question about the origin of life on Earth*. To get an answer, we must go beyond the framework of the EGM and use different speculations. There is only one way to resolve this contradiction, namely, to postulate the following: whatever the pre-biosphere history of the Earth was, the evolution of the biosphere during all geological periods must have produced the contemporary biosphere as a result. This is the ergodicity axiom. It postulates that to a large degree the process of the biosphere's evolution is deterministic and stable in respect to the initial periods of its history.
3. *There were no lifeless geological epochs*. This means that the contemporary living matter is genetically connected with living matter of all the previous epochs. It is natural to call this axiom the continuity axiom.

   The following empirical generalizations are, actually, a form of the conservation law. On the other hand, since they generalize some equilibrium properties of the biosphere, it is logical to call them: the axioms of stationary state.
4. The chemical composition of living matter was always, on average, the same as it is now.
5. *The amount of living matter, on average, was the same for all geological time*.

   These two above generalizations of Vernadsky cause a lot of objections at the present time. However, there are not enough new facts to formulate new empirical generalizations. Therefore, it is quite possible to consider the changes of the total amount of living matter, observed during different geological epochs, as fluctuations around some constant average level. (The same can be also said about the chemical composition of living matter in the terrestrial crust.)

   And, lastly, below are generalizations that determine the principles of functioning of the biosphere mechanisms.
6. *The energy which is being stored and emitted by living organisms is the solar energy*. With the help of organisms this energy is controlling the chemical processes on the Earth (in particular, the global biogeochemical cycles).
7. *Vegetation plays the main role in the assimilation and allocation of the solar energy*. If we agree with the axiom about the constancy of the total amount of living matter during the whole lifetime of the biosphere, then we have to assume that its evolution only followed the path of the structural complication of living matter, either by increasing the number of species (there are 3106 species on Earth), or by making the structure of biological communities more complex.

The Future of the Concept: Noosphere and Modern Perspective

Biogeocoenosis

While Vernadsky's concept can be considered as maximally aggregated (i.e., like a view on the biosphere from the outside), the concept of the biogeocoenosis (BGC) first suggested by V. Sukhachev in 1945 and later developed by N. Timofeev-Resovsky, relates to the elementary units of the biosphere, and is one that is basically atomistic in nature. In accordance with the definition by Timofeev-Resovsky's, the BGC is the part (area) of the biosphere, which has no essential ecological, geomorphological, hydrological, microclimatic, or any other boundary inside itself. Hence, the biosphere is divided into elementary systems, naturally separated from one another. According to Vernadsky, the biogeocenosis have appeared immediately upon the beginning of existence of the biosphere.

Noosphere

In his last years of life, Vernadsky's works were directed toward the future development of scientific thought as a planetary phenomenon. He perceived civilization as a form of a new geological force—scientific force performance. Vernadsky wrote:

"In the 20th century man for the first time in history of the Earth knew and embraced the whole Biosphere… That mineralogical rarity, native (pure) iron, is now being produced by billions of tons. Native aluminum, which never before existed on our planet, is now produced in any quantity. The same is true with regard to the countless number of artificial chemical combinations newly created on our planet. Chemically, the face of our planet, the biosphere, is being sharply changed by man… New species of animals and plants are being created by man."

The biosphere is a powerful geological force that has transformed this planet and its geochemistry in a most spectacular way. Toward the end of his life, Vernadsky planned to consider the Noosphere—the supremacy of scientific thought that always existed (term introduced in 1922 by a French philosopher and mathematician Edouard Le Roy) in more detail, but sadly could not finish developing this in his lifetime.

Vernadsky's ideas act on the contrary to doomsday scenarios since he views our civilization as a form of a new geological force—scientific thought, and therefore it cannot destroy itself. Indeed, we observe now the World's population growth stagnation, contrary to the predictions and calculations from the mid 20th century. This could be an example of the influence induced by intelligent thought and man's growing ability to forecast. Again, it is important to stress, that Noosphere is far from being Utopia, but rather the next stage of biosphere's development and these forecasting scenarios could change its future. It seems that we already live in Noosphere, even though most changes are subtle and happen slowly even with new technologies introduced.

Indeed, at present, despite its growing industrial power, our science and civilization will not yet be able to reconstruct the entire biosphere in the desirable way, if it comes near some critical conditions. It is practically impossible to predict the new quasi-stationary state, as no other possible states of the biosphere's equilibrium are known. The model of coevolution of "man" and the "biosphere" as the main principle of global coexistence was suggested by Yu. Svirezhev. And, indeed, Vernadsky has pointed out that humans and biosphere are not "enemies." The basic idea is to study the dynamics of the biosphere as an entity and its reactions to human impact. In the framework of this research the study of possible ways of development of human society as a natural component of the biosphere is of a special importance. The mechanism of coexistence of human society and the biosphere, which is actually the way of mutual coevolution and harmonic coupling of humans and their environment, is now becoming one of the most important scientific and social questions of the modern society. This problem, integrating natural dynamic processes operating within the biosphere, with human dynamics is of a multiscale and very complex.

Conclusions

In conclusion, we would like to stress again the great significance of Vermadsky's ideas for the future science and say that the concept of the biosphere and the next new state of it, Noosphere (from the Greek noó(s) (mind) and sphere), has shaped, and is continuing to do so, the global understanding of the origin and evolution of mankind, since as the Teacher himself points out: It "is a new geological phenomenon on our planet. In it for the first time man becomes a large-scale geological force… Wider and wider creative possibilities open before him. It may be that the generation of our grandchildren will approach their blossoming… Fairy-tale dreams appear possible in the future; man is striving to emerge beyond the boundaries of his planet into cosmic space. And he probably will do so."

See also: General Ecology: Biomass. Global Change Ecology: Noosphere; Anthropospheric and Anthropogenic Impact on the Biosphere; The Earth System and Climate Science: Understanding a Very Complex Entity

Further Reading

Bauer, E., 1935. In: Moskva-Leningrad, M.-L. (Ed.), Theoretical biology. Izd: Vsesoiuz nogo Instituta Eksperimentalnoi Mediciny (VIEM), p. 206.

Documentary popular, 2016—http://www.imdb.com/videoplayer/vi2148316441. Documentary popular science film about Vernadsky and the modern science: Vernadsky New Age (2016), Russian with English subtitles.

Kuhn, T.S., 1996. The structure of scientific revolutions. Chicago: University of Chicago Press, p. 222.
Svirezhev, Y.M., 1998. Globalistics: A new synthesis philosophy of global modelling. Ecological Modelling 108 (1–3), 53–65.
Timofeev-Resovsky, N.V., 1968. Biosphere and mankind. UNESCO Bulletin 1, 3–10.
Vernadsky, V.I., 1926. Biosphere. Leningrad: Gostekhizdat.
Vernadsky, V.I., 1945. The biosphere and the noosphere. American Scientist 33, 1–12.
Vernadsky V.I. (1991) Scientific Thought as Planetary Phenomenon, 271pp. (in Russian). Moscow: Nauka.
Vernadsky V.I. (1997) Biosfera. Langmuir D (trans) and McMenamin MAS (revised by). New York: Springer.
Vernadsky, V.I., 2007. In: Salisbury, F.B. (Ed.), Geochemistry and the biosphere: Essays by Vladimir I. Vernadsky First English Translation from the 1967 Russian Edition of Selected works. Santa Fe, New Mexico: Sciences Synergetic Press.

Climate Change 2: Long-Term Dynamics☆

Werner von Bloh, Potsdam Institute for Climate Impact Research, Potsdam, Germany

Glossary

Archean Geologic eon from 4000 to 2500 Myr ago.
Eucaryotes Organisms with cell nucleus and organelles.
Kerogen Reduced organic carbon in rocks.
Phanerozoic Current geologic eon from 542 Myr ago to present.
Procaryotes Unicellular organism lacking a nucleus.
Proterozoic Geologic eon before the onset of complex life from 2500 to 542 Myr ago.

Nomenclature

Symbol	Description (Unit)
C_{ker}	Carbon stored in kerogen pool (kg)
β	Biological enhancement factor of weathering (none)
$\delta^{13}C$	Isotopic signature ($^0/_{00}$)
F_{weath}	Normalized weathering flux (none)
Π	Biological productivity (kg/year)
P_{CO_2}	CO_2 partial pressure in atmosphere (bar)
P_{soil}	CO_2 partial pressure in soil (bar)
p_{O_2}	O_2 partial pressure (bar)
P_{min}	Minimum CO_2 partial pressure for photosynthesis (bar)
T_{max}	Upper temperature tolerance (°C)
T_{min}	Lower temperature tolerance (°C)

☆*Change History*: March 2018. Werner von Bloh. Sections "Continental Growth" and "Evolution of Climate" have been updated. Figs. 3 and 5 have been replaced by new figures. New Fig. 6 added. More up-to-date references have been added in the Further Reading section.

Introduction

The main component of the early Earth's atmosphere after the formation of the oceans was CO_2 as the second most abundant volatile in the accreting material. On planetary timescales the variation of the CO_2 content in the atmosphere is most important in investigating the relation between the evolution of the Sun as a main-sequence star and the stabilization of the Earth's surface temperature. During the history of the Earth the luminosity of the Sun has increased to the present level starting with a 30% lower value. There must be a mechanism that provides a feedback, generating a high concentration of atmospheric CO_2 in the past to prevent the Earth from freezing while solar luminosity was low. On the other hand a progressive lowering of CO_2 concentration is necessary for an increasing solar luminosity. Such a negative feedback mechanism is provided by the global carbon cycle among the surface reservoirs of carbon (atmosphere, ocean, crust) and the mantle reservoir. The overall chemical reactions for the weathering processes are

$$CO_2 + CaSiO_3 \leftrightarrow CaCO_3 + SiO_2$$
$$CO_2 + MgSiO_3 \leftrightarrow MgCO_3 + SiO_2$$

The main idea of this abiotic feedback is the interplay between weathering rate, surface temperature, and atmospheric CO_2 pressure. An increase of the luminosity leads to a higher mean global temperature causing an increase in weathering. Then more CO_2 is extracted from the atmosphere weakening the greenhouse effect. Overall the temperature is lowered and homeostasis, that is, self-stabilization of the global surface temperature is achieved. Plate tectonics is a necessary condition for closing the carbon cycle. Without spreading and subduction carbonates would be buried on the seafloor and not brought back to the atmosphere via regassing at mid-ocean ridges or andesitic/back-arc volcanism.

During the evolution of the Earth the global carbon cycle has been mediated by the biosphere. The occurrence of microfossils and stromatolites in rocks has shown that life had originated at least 3.5 Gyr ago. Isotopic signatures of $^{13}C/^{12}C$ suggest the presence of organic carbon in 3.8 Gyr rocks from Greenland. Photosynthetic fixation leads to buildup of organic material consuming CO_2 from the atmosphere. On the other hand, organic material is decomposed and CO_2 is remobilized:

$$CO_2 + H_2O \leftrightarrow CH_2O + O_2$$

Carbon isotopes imply that the enzyme Rubisco preferring the lighter isotope ^{12}C and therefore oxygenic photosynthesis controlled the global distribution of carbon in the atmosphere–ocean system for at least 3.5 Gyr and gave an imprint on the isotopic signature of the carbon reservoirs. Direct evidence of oxygenic photosynthesis is given by the steep rise of oxygen in the atmosphere 2.2 Gyr ago. Life is unable to influence Earth's carbon cycle in the absence of photosynthesis.

The biosphere can in principle be divided into three life forms that appeared on Earth in consecutive order (see Fig. 1). First Archaean and prokaryotic bacteria appeared, second eukaryotic life, and third complex multicellular life. In particular complex multicellular organisms have a strong influence on weathering by amplifying the weathering rates.

Global Carbon Cycle and the Biosphere

The global carbon cycle can be described by a box model of the surface reservoirs of carbon and the mantle reservoir. Fig. 2 denotes the most pertinent fluxes between the storage of carbon in the mantle, the combined reservoir consisting of ocean and atmosphere, the continental crust, the ocean crust and floor, the kerogen, and the three different biospheres. The efficiency of carbon transport between the reservoirs takes into account mantle de- and regassing, carbonate precipitation, carbonate accretion, hydrothermal reactions at mid-ocean ridges, decay of dead organic matter, and weathering processes. The Earth's crust contains the carbon storage of the continents, kerogen, and the ocean crust and seafloor. Biomass is accumulated from the atmosphere by photosynthesis. A fraction of the dead organic matter is buried in the kerogen pool. Additionally to these direct effects of the biosphere on the surface reservoirs of carbon there is an indirect effect due to the increase in weathering rates denoted by the dashed arrow in Fig. 2.

Continental Growth

The continental crust is especially diverse and heterogeneous, and its formation is less understood than that of the geologically relatively simple oceanic crust. In contrast to the seafloor which is primarily made of basalts it is granitic in composition. The onset of the granitic continents might have been triggered by the origin of life. Two different hypotheses have been suggested to explain the evolution of the continental crust. The first proposes that the present continental crust formed very early in the Earth's history and has been recycled through the mantle in steadily decreasing fashion such that new additions are balanced by losses, resulting in a steady-state system. The return of the continental material to the mantle and its replacement by new younger additions reduces its mean age, both of which keep the mass of the continents constant. The second hypothesis proposes crustal growth throughout geological time without recycling into the mantle.

Fig. 3 summarizes the different classes of continental growth scenarios. The continental area can be assumed to be fixed at its present value, can be grown linearly with a constant growth rate, or linearly with a delay. Geological investigations of the best-studied regions, North America and Europe, which formed a single land mass for most of the Proterozoic suggests that the

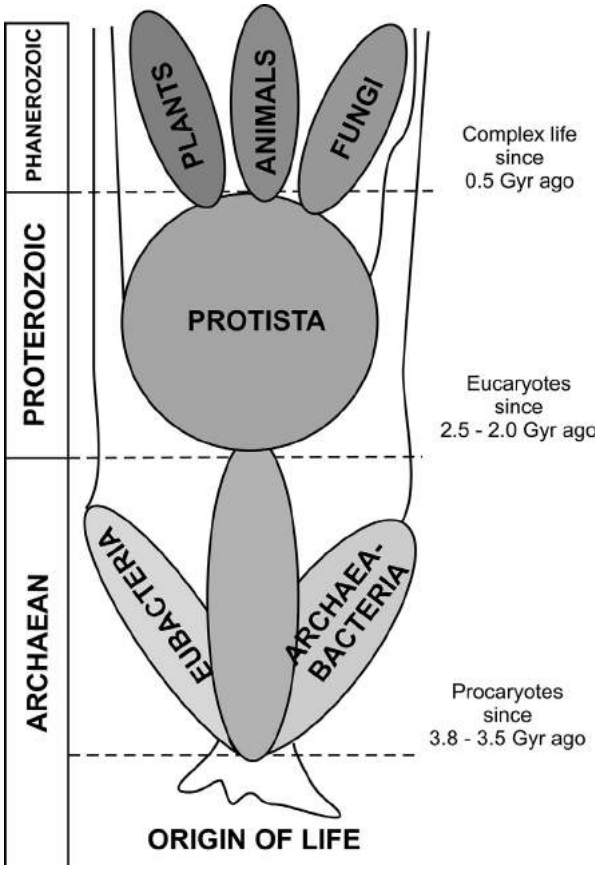


Fig. 1 Evolutionary path of life on Earth.

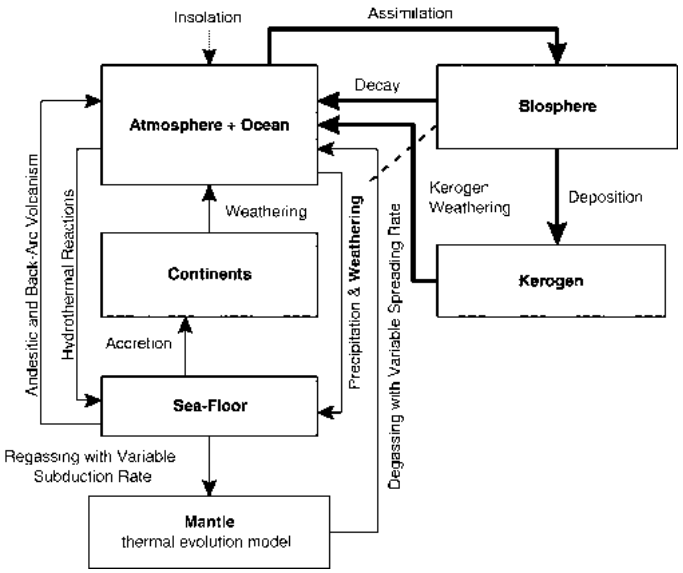


Fig. 2 Box diagram illustrating the basic mechanisms and interactions of the global carbon cycle. *Gray* boxes are the surface reservoirs of carbon. The fluxes from and to the different pools are indicated by *arrows*. *Bold arrows* denote fluxes affected by the biosphere.

continental crust grows episodically, and it is concluded that at least 60% of the crust has been replaced by the late Archaean. Furthermore, Thorium–Uranium–Niobium systematics of the depleted mantle can be used to derive the evolution of the continents. These data-based descriptions are clearly more realistic than the theoretical models. The continental area is directly related to the weathering process. A larger continental area leads to proportionally higher weathering rates.

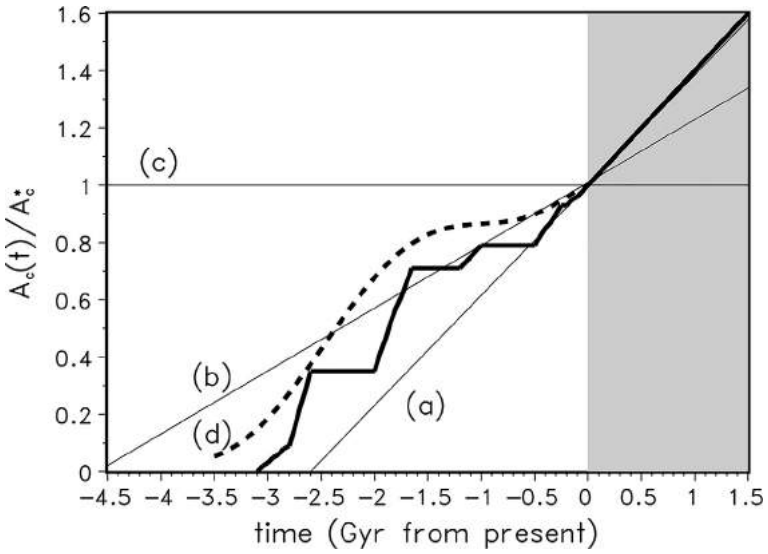


Fig. 3 Normalized continental area as a consequence of the following continental-growth scenarios: (a) delayed growth, (b) linear growth, (c) constant area, (d) growth derived from geological investigations. From Condie, K.C. (1990). Growth and accretion of continental crust: Inferences based on Laurentia. *Chemical Geology* **83**, 183–194 (*solid line*) and Collerson, K.D. and Kamber, B.S. (1999). Evolution of the continents and the atmosphere inferred from Th-U-Nb systematics of the depleted mantle *Science* **283**, 1519–1522 (*dashed line*).

Atmospheric Carbon Dioxide and Climate

The climate of the Earth is governed by the energy-balance equation between incoming and outgoing radiation and depends on the concentration of greenhouse gases and the solar luminosity. CO_2 and H_2O are the most abundant greenhouse gases in the atmosphere. The partial pressure of H_2O, p_{H_2O}, can be expressed as a function of temperature and relative humidity using the Clausius–Clapeyron equation. Therefore, the global mean temperature of the Earth depends primarily on the CO_2 concentration in the atmosphere and the solar luminosity. During the Earth's history the luminosity of the Sun has increased at a rate of about 10% per Gyr and will increase in the future (up to the next 5 Gyr) at approximately the same rate.

The efficiency of weathering processes and the biosphere productivity strongly depend on the partial pressure of atmospheric CO_2, p_{CO_2}. On geological timescales atmospheric carbon content is always in chemical equilibrium with the carbon content in the ocean and depends on the pH value of the ocean. A more acidic ocean lead to a an accumulation of CO_2 in atmosphere while an alkaline ocean results in lower atmospheric CO_2. In the case of a constant density profile of carbon dioxide in the atmosphere the distribution of carbon can be calculated from the condition of equal partial pressures of CO_2 at the interface between atmosphere and ocean.

Biotic Enhancement of Weathering

The rate of weathering is greatly amplified by a range of biological processes that respond to photosynthetic productivity. First, there is an increase of soil CO_2 partial pressure due to respiration of soil organisms and due to the respiration from the roots of vascular plants. Furthermore litter is decomposed by microorganisms, through providing organic matter for the formation of humic acids, and through mycorrhizal and fungal digestion. This can be expressed by a direct dependence of weathering on biological productivity by a factor β mediating the weathering rate, F_{weath}:

$$F_{\text{weath}} \propto \beta \cdot \left(\frac{a_{H^+,s}}{a^*_{H^+,s}}\right)^{0.5} \exp\left(\frac{T_s - 288\text{K}}{13\text{K}}\right),$$

where $a_{H^+,s}$ denotes the activity of fresh soil water $a^*_{H^+}$ is the corresponding present-day values. The activity $a_{H^+,s}$ itself is a function of the surface temperature and the CO_2 partial pressure in the soil, p_{soil}, where the equilibrium constants for the chemical activities of the carbon and sulfur systems have been taken into account. p_{soil} depends on the terrestrial biological productivity Π, the atmospheric CO_2 partial pressure, p_{CO_2}, and their corresponding (long-term mean pre-industrial) present values p^*_{soil}, Π^*, $p_{CO_2}{}^*$:

$$\frac{p_{\text{soil}}}{p^*_{\text{soil}}} = \frac{\Pi}{\Pi^*}\left(1 - \frac{p^*_{CO_2}}{p_{CO_2}}\right) + \frac{p_{CO_2}}{p^*_{\text{soil}}}$$

It is assumed that $p^*_{soil} = 10 p_{CO_2}{}^*$. The parameterization of weathering that considers only the variation of soil carbon dioxide levels gives a biotic amplification of weathering of 1.56 for the present state, which is a significant underestimate. Furthermore vascular land plants increasing the partial pressure of CO_2 in the soil appeared on Earth only 0.35 Gyr ago. The total weathering amplification due to land life is at least a factor of 10. This indicates that much of the observed biotic amplification of weathering is due to processes other than increased soil p_{CO_2}. This is considered by the prefactor β that reflects the biotic enhancement of weathering by the biosphere types, i:

$$\beta = \prod_{1=1}^{3} \left(\frac{1}{\beta_i} + \left(1 - \frac{1}{\beta_i} \right) \frac{\Pi_i}{\Pi_i^*} \right)$$

The factor β_i denotes the specific biotic amplification of weathering, Π_i the specific biological productivity, and Π_i^* the respective present-day value of biosphere type i. Biotic enhancement of weathering is only affected by complex multicellular life ($\beta_1 = \beta_2 = 1$, $\beta_3 > 1$). Complex multicellular life contributes about 10–100 times more to the biotic enhancement of weathering than primitive life.

Biological Productivity

The biological productivity Π is the amount of biomass that is produced by photosynthesis per unit time. In reality, Π is a function of various parameters as water supply, photosynthetically active radiation (PAR), nutrients (N, P, etc.), atmospheric CO_2 content and surface temperature:

$$\frac{\Pi_i}{\Pi_{\max,i}} = f_{T_s,i}(T_s) \cdot f_{CO_2,i}(p_{CO_2}) \cdot f(\mathrm{N}, \mathrm{P}, \mathrm{H_2O}, \mathrm{PAR}, \ldots)$$

where $\Pi_{max,\,i}$ is the maximum productivity of biosphere type i. For simplification biological productivity should depend only on the mean global surface temperature, T_s, and on the CO_2 partial pressure of the atmosphere, p_{CO_2}. Both variables are affected by the global carbon cycle. The qualitative dependence on CO_2 partial pressure and temperature is shown in Fig. 4. The function for the temperature dependence, $f_{T_{s,\,i}}$, can described by a parabola and the function for the p_{CO_2} dependence is an increasing function with a saturation level. A minimum CO_2 atmospheric partial pressure, $p_{min,I}$, allowing photosynthesis is necessary for all biosphere types. A biosphere based on C3 photosynthesis has a minimum value of 150×10^{-6} bar, while C4 photosynthesis results in a value of 10^{-5} bar. The interval $[T_{min,i} \ldots T_{max,i}]$ is the temperature tolerance window for the biosphere. If the global surface temperature is inside this window a global abundance of biosphere type i is possible. It must be emphasized that this window is related to the mean global surface temperature. Latitudinal differences in temperature decrease as global mean temperature increases and might vanish for $T > 30°C$. Table 1 contains estimated parameter ranges for the prokaryotic, eukaryotic, and complex multicellular biosphere, respectively.

Carbonate Precipitation

Weathering products are transported to the ocean and, depending on the solubility product, precipitated to the ocean floor. Because there exists a calcium carbonate compensation depth level in the present ocean, carbonates can precipitate only in the shallower regions such as around the mid-ocean ridges and the continental shelves. A total of 8% of the Earth's area is covered with ocean less shallow than 10^3 m. The change of equilibrium concentrations of Ca and Mg in water results in a change of solubility of carbonates in ocean water. Furthermore, oceanic photosynthesis provides an additional way to sequester carbon on the seafloor.

Hydrothermal Reactions

Due to hydrothermal reactions CO_2 dissolved in the oceans reacts with fresh mid-ocean basalts and precipitates in the form of carbonates to the ocean floor. Therefore it is an additional sink in the atmosphere-ocean reservoir. The hydrothermal flux is proportional to the production of fresh basalt at mid-ocean ridges, which in turn is proportional to the areal spreading rate. Assuming hydrothermal reactions to be dependent on the dissolved CO_2 content lead to a stronger sink for CO_2 in the atmosphere ocean system and can be an explanation for relatively low atmospheric CO_2 concentration and corresponding moderate temperatures in the Archean. The area around the spreading centers is likely to be one of the most habitable environments for a subsurface biosphere. It is porous and characterized by extensive hydrothermal circulation. Such hydrothermal systems provided a site for the rapid emergence of life through a sequence of abiotic synthesis.

Kerogen

Kerogen comprehends the dispersed, insoluble, organic carbon in rock including coal and mineral oil deposits. It is probably the least important reservoir from the point of view of carbon cycling because it is relatively inert. However, there are processes of kerogen weathering and kerogen formation. Kerogen is formed from ~0.1% of the dead biomass that is not returned to the atmosphere through litter decomposition. The present size of the kerogen reservoir of 10%–20% of the surface reservoirs is

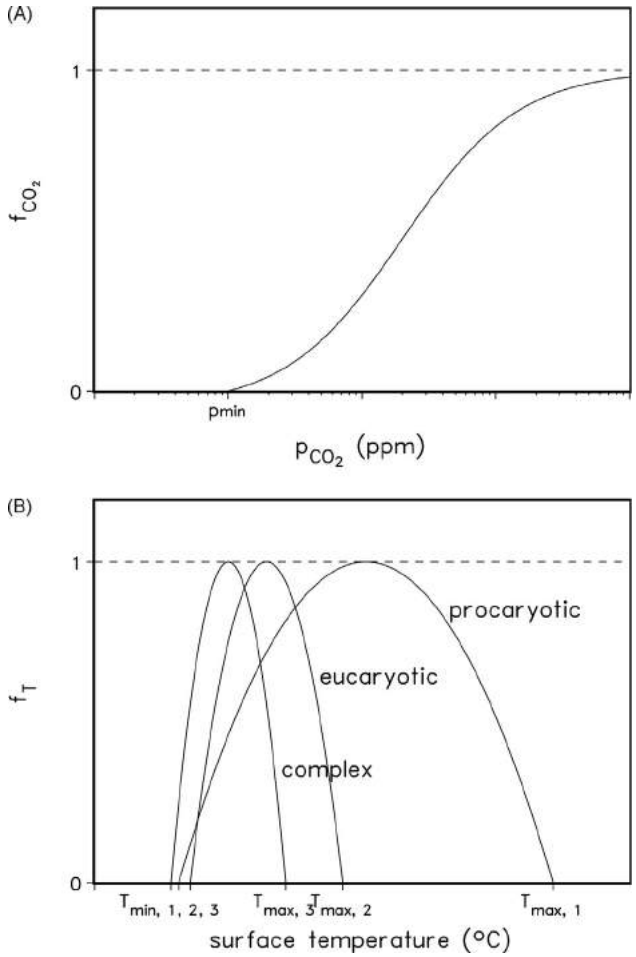


Fig. 4 (A) The dependence of biosphere productivity on CO_2 partial pressure in the atmosphere. (B) The dependence of biosphere productivity on global surface temperature for prokaryotes, eukaryotes, and complex multicellular life.

Table 1 Parameter estimates for the three different life forms (prokaryotes, eukaryotes, complex multicellular life)

Life form	*Prokaryotes*	*Eukaryotes*	*Complex multicellular*
T_{min} (°C)	2	5	0
T_{max} (°C)	100–130	45–60	30–45
P_{min} (10^{-6} bar)	10^a–150^b	10^a–150^b	10^a–150^b
β	1	1	4–20

[a]For C4 plants.
[b]For C3 plants.

obviously the net result of these processes. The main constraint for the reservoir size results from isotopic geochemistry. Since kerogen is isotopically light due to its biological origin it sequesters preferentially ^{12}C, while the continental carbon reservoir must get enriched in the heavier isotope ^{13}C. The isotopic signature is measured as a difference to a standard sample:

$$\delta^{13}C = \left[\frac{(^{13}C/^{12}C)\,\text{sample}}{(^{13}C/^{12}C)\,\text{standard}} - 1\right] \times 1000‰$$

The kerogen has a $\delta^{13}C$ value of $\sim -20‰$. The isotopic composition of the two carbon reservoirs kerogen and continental crust might have been constant over the last 3.5 Gyr. The ratio of kerogen carbon to continental carbon would also have been constant at a value of 1:4 taking into account the isotopic signature of the mantle carbon of $\delta^{13}C \sim -5‰$.

Atmospheric Oxygen

The evolution of the atmospheric partial pressure of oxygen, p_{O_2}, can be derived from the evolution of the kerogen pool C_{ker}, that is, the long-term deposition of reduced organic carbon. Between about 2.2 and 2.0 Gyr ago there was a global oxidation event in which atmospheric p_{O_2} rose from <0.0008 to >0.002 bar. Under the assumption that before 2.2 Gyr all oxygen had been chemically bound it is possible to make the following simple estimate:

$$p_{O_2}(t) = p^*_{O_2} \cdot \frac{C_{ker(t)} - C_{ker}(t = -2.2\,\mathrm{Gyr})}{C^*_{ker} - C_{ker}(t - 2.2\,\mathrm{Gyr})}$$

where $p_{O_2}{}^*$ is the present atmospheric O_2 level and C^*_{ker} is the size of the present kerogen pool.

Coevolution of the Biosphere–Geosphere System

The feedback between the biosphere and the surface reservoirs of carbon leads to several bifurcation points in Earth's history. In particular the evolution of the climate is affected by the change in CO_2 concentration in the atmosphere. Atmospheric CO_2 concentration is regulated by biologically mediated weathering processes and is driven by an increase in solar luminosity, continental growth, and lowering mantle temperatures. The decline of mantle temperature is causing a decrease in the spreading rate with lower outgassing at mid-ocean ridges.

Evolution of the Climate

Fig. 5A shows the results for the evolution of the mean global surface temperature (solid line). The figure has been derived from a coupled model of the global carbon cycle including the biosphere. The modeled surface temperature curve is in good agreement with the ^{18}O chert thermometer. According to these data, the ocean surface water has cooled from 70°C (±15°C) in the Archaean to the present value. There is, however, a recent debate of how strongly these isotope data reflect the ocean temperature. It might be that the Archean ocean temperatures were below 40°C. Such moderate Archean temperatures can be caused by more intensive hydrothermal reactions reducing the ocean/atmosphere CO_2 concentrations in combination with an acid ocean (see Fig. 6). The decrease of the global temperature from the Archean to the present time is caused by the growth of the continental area increasing the weathering processes and decreasing spreading rates lowering CO_2 outgassing at mid-ocean ridges. There was a drop in temperature 0.54 Gyr ago due to an increase in weathering rates caused by the first occurrence of complex life. After that event temperatures have roughly stabilized around the optimum growth temperatures for complex life. In the future the global surface temperature will rise because the increase in solar luminosity cannot be balanced by intensified weathering rates.

Evolution of the Biosphere

Fig. 5B shows the cumulative biomasses for the three life forms. From the Archaean to the future there always exists a prokaryotic biosphere. At 2 Gyr ago eukaryotic life first appears because the global surface temperature reaches the tolerance window for eukaryotes. This moment correlates with the onset of a temperature fall caused by an increasing continental area. The resulting enlargement in the weathering flux takes out CO_2 from the atmosphere. In contrast to the eukaryotes the first appearance of complex multicellular life starts with an explosive increase in biomass connected with a strong decrease in Cambrian global surface temperature at about 0.54 Gyr ago. The biological colonization of land surface by metaphyta and the consequent increase in silicate weathering rates caused a reduction in atmospheric CO_2 and planetary cooling. Protein sequence analysis has shown that a first appearance of land plants at this time was already possible. Metazoan fecal pellets supplied a new and important transport mechanism of organic carbon to the deep ocean. This provides an additional sink for CO_2 in the atmosphere–ocean system.

Cambrian Explosion

The Cambrian explosion is known as the Big Bang in biology. It began 542 million years ago and ended about 40 million years later. This period is characterized by the first appearance of abundant skeletonized metazoans, a sudden increase in biodiversity, and the emergence of most modern lines. In the Vendian (0.56–0.54 Gyr ago) first animals with soft bodies appeared announcing the Cambrian explosion. Before the Vendian period life was microscopic, vegetative, and mainly prokaryotic and eukaryotic.

There is still a lot of speculation about what caused the Cambrian explosion and why it happened when it did after 3 billion years of potential evolutionary time. The approaches that have been put forward to solve the puzzle of what triggered the explosion can be split into extrinsic (environmental) factors, intrinsic (biological) factors, or a mixture of both. Extrinsic factors are physical changes in the Precambrian environment. Among these changes are the breakup of the supercontinent Rodinia and the Neoproterozoic glaciations known as snowball Earth events. The snowball Earth events and the continental breakup are associated with genetic isolation but also with a reorganization of oceanic flow patterns causing upwelling, with increasing primary production, and with a consequently higher atmospheric oxygen level. Another cause is given by the rise of atmospheric oxygen as a trigger of the Cambrian explosion. This higher oxygen level can be caused by an intensified phosphorus flux into the ocean. The phosphorus is

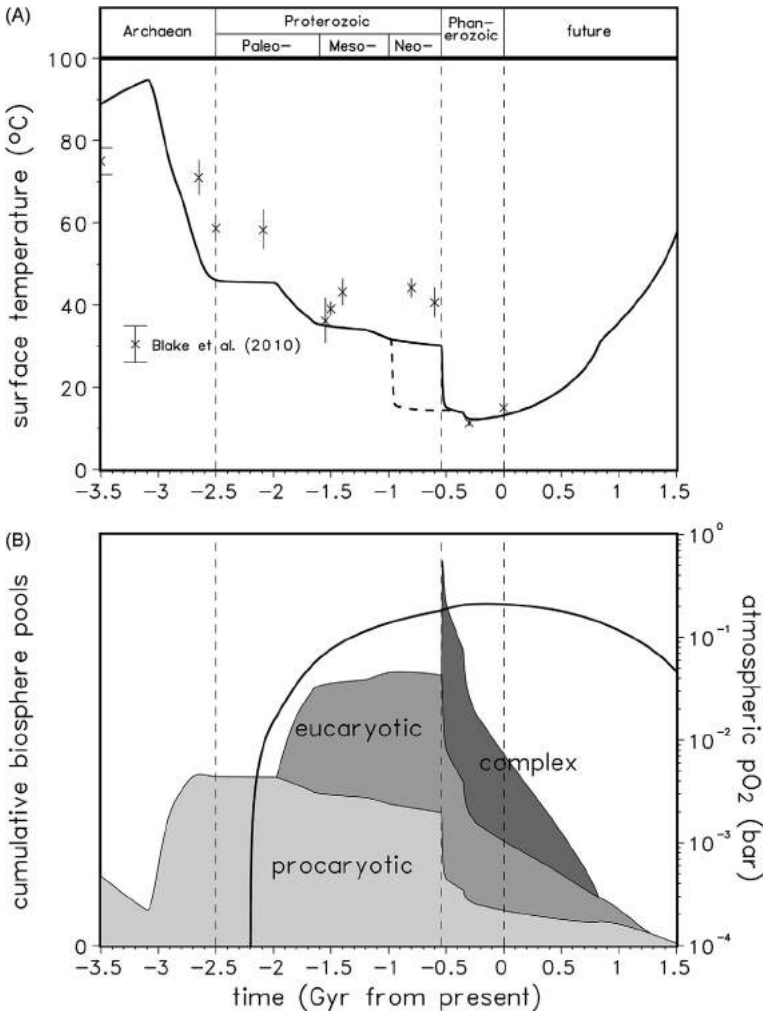


Fig. 5 (A) Evolution of global surface temperature (*solid line*). The *dashed line* denotes a second possible evolution path triggered by a temperature perturbation in the Neoproterozoic era. Vertical bars denote temperature estimates derived from cherts. (B) Evolution of the cumulative biosphere pools for prokaryotes, eukaryotes, and complex multicellular life. From Von Bloh, W., Bounama, C. and Franck, S. (2003). Cambrian explosion triggered by geosphere–biosphere feedbacks. *Geophysical Research Letters* **30**(*18*), 1963–1967.

released by weathering rates biotically enhanced by the first colonization of continents. Intrinsic causes involve some mechanisms within the Precambrian biosphere itself, which enabled evolution and diversification to start. An example is the finding in developmental genetics that the mutation of an ancestral metazoan could potentially initiate a large morphological change in the animal.

The dashed line in Fig. 5A shows a possible second evolutionary path. A cooling event can cause a premature rising of complex life. Up to 1.75 Gyr ago there is only a unique evolutionary path. After that time more than one stable state of the Earth system exists (bistability). It depends on the environmental conditions which state (with or without complex life forms) is realized.

Phanerozoic Time

After the Cambrian explosion, there was a continuous decrease of biomass in all pools. At 0.35 Gyr ago there was a slight drop in the global temperature connected with the rise of vascular plants. At this time weathering rates were increased due the elevated partial pressure of CO_2 in the soil by root respiration. The continuous decrease in biomass of primitive life forms (prokaryotes and eukaryotes) since the Cambrian explosion is related to the fact that Phanerozoic surface temperatures are below the optimum for these life forms. The decrease in biomass of complex life forms is due to the fact that there is a continuous decrease in Phanerozoic atmospheric carbon content. At present the biomass is almost equally distributed between the three pools and the mean global surface temperature of about 15°C is near the optimum value for complex multicellular life.

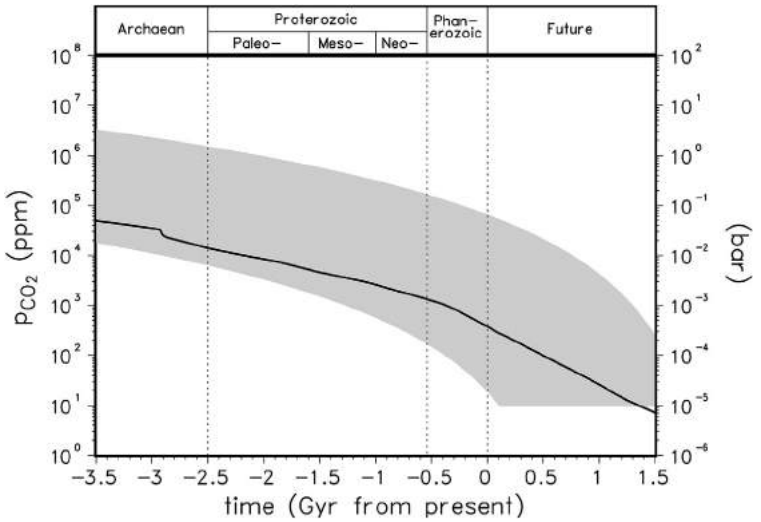


Fig. 6 Evolution of atmosphere CO_2 for strong hydrothermal reactions and acid ocean pH in the Archean. The *gray shaded* area denotes corresponding climate conditions with global surface temperatures $0°C < T < 100°C$.

Future Evolution

In the future we can observe a further continuous decrease of biomass with the strongest decrease in complex multicellular life. The life spans of complex multicellular life and of eukaryotes end at about 0.8 Gyr and 1.3 Gyr from present, respectively. In both cases the extinction will be caused by reaching the upper limit of the temperature tolerance window. In contrast to the first appearance of complex multicellular life via the Cambrian explosion, its extinction proceeds more or less continuously. In the future there will be no bistability, that is, the extinction of complex multicellular life will not proceed as an implosion. Comparing these results with the life span for an Earth without biotic enhancement of weathering ($\sim$0.5 Gyr) the life spans are extended.

Summary

The global temperature on Earth is regulated by the global carbon cycle. The main negative feedback is provided by the weathering processes mediated by the biosphere. In the past the Earth system was characterized by lowering temperatures caused by continental growth and declining outgassing at mid-ocean ridges. In the future, however, temperatures will rise due to the increase in solar luminosity.

The Cambrian explosion can be explained by extrinsic environmental causes, that is, a gradual cooling of the Earth. The Cambrian explosion was so rapid because of a positive feedback between the spread of biosphere, increased silicate weathering, and a consequent cooling of the climate. The environment itself has been actively changed by the biosphere maintaining the temperature conditions for its existence. Therefore, this explanation of the Cambrian explosion is in line with the Gaia theory of the Earth as a self-regulating system.

Prokaryotes, eukaryotes, and complex multicellular life forms will become extinct in reverse sequence of their appearance. Nonlinear interactions in the biosphere–geosphere system caused bistability during the Neo- and Mesoproterozoic era. There is no bistability in the future solutions for complex life. Therefore, complex organisms will not become extinct by an implosion (in comparison to the Cambrian explosion). Eukaryotes and complex life become extinct because of too high surface temperatures in the future. The time of extinction is mainly determined by the upper temperature tolerance limit of these life forms. The ultimate life span of the biosphere is defined by the extinction of prokaryotes in about 1.6 Gyr because of CO_2 starvation. Only in a small fraction (1.3–1.7 Gyr) of its habitability time (6.2 Gyr) can our home planet harbor advanced life forms.

Further Reading

Blake RE, Chang SJ, and Lepland A (2010) Phosphate oxygen isotopic evidence for a temperate and biologically active Archaean Ocean. *Nature* 464: 1029–1032.
Berner RA and Kothavala Z (2003) GEOCARB III: A revised model of atmospheric CO_2 over phanerozoic time. *American Journal of Science* 301: 182–204.
Caldeira K and Kasting JF (1992) The life span of the biosphere revisited. *Nature* 360: 721–723.
Collerson KD and Kamber BS (1999) Evolution of the continents and the atmosphere inferred from Th-U-Nb systematics of the depleted mantle. *Science* 283: 1519–1522.
Feulner G (2012) The faint young Sun problem. *Reviews of Geophysics* 50(2). https://doi.org/10.1029/2011RG000375.
Franck S, Kossacki KJ, von Bloh W, and Bounama C (2002) Long-term evolution of the global carbon cycle: Historic minimum of global surface temperature at present. *Tellus* 54B: 325–343.

Franck S, Bounama C, and von Bloh W (2006) Causes and timing of future biosphere extinctions. *Biogeosciences* 3: 85–92.
Knauth LP and Lowe DR (2003) High Archean climatic temperature inferred from oxygen isotope chemistry of cherts in the 3.5 Ga Swaziland supergroup, South Africa. *GSA Bulletin* 115(5): 566–580.
Condie KC (1990) Growth and accretion of continental crust: Inferences based on Laurentia. *Chemical Geology* 83: 183–194.
T.M. Lenton, A.J. Watson, Biotic enhancement of weathering, atmospheric oxygen and carbon dioxide in the Neoproterozoic, Geophysical Research Letters 31 (5) (2004); L05202, https://doi.org/10.1029/2003GLO18802.
Lenton TM and von Bloh W (2001) Biotic feedback extends the life span of the biosphere. *Geophysical Research Letters* 28(9): 1715–1718.
Lovelock JE and Watson A (1982) The regulation of carbon dioxide and climate. *Planetary and Space Science* 30: 795–802.
Nisbet EG, Cann JR, and Dover CL (1995) Origins of photosynthesis. *Nature* 373: 479–480.
Robert F and Chaussidon M (2006) A paleotemperature curve for the Precambrian oceans based on silicon isotopes in cherts. *Nature* 443: 969–972.
Schidlowski M (2001) Carbon isotopes as biogeochemical recorders of life over 3.8 Ga of earth history: Evolution of a concept. *Precambrian Research* 106: 117–134.
D. Schwartzman, T. Volk, Biotic enhancement of weathering and the habitability of earth, Nature 340 (1991) 457–460.
Smil V (2002) *The Earth's biosphere: Evolution, dynamics, and change*. Cambridge, MA: MIT Press.
Tajika E and Matsui T (1990) The evolution of the terrestrial environment. In: Newsom HE and Jones JH (eds.) *Origin of the earth*, pp. 347–370. Oxford: Oxford University Press.
Volk T (1987) Feedbacks between weathering and atmospheric CO_2 over the last 100 million years. *American Journal of Science* 287: 763–779.
von Bloh W, Bounama C, and Franck S (2003) Cambrian explosion triggered by geosphere–biosphere feedbacks. *Geophysical Research Letters* 30(18): 1963–1967.
von Bloh W, Bounama C, Eisenack K, Knopf B, and Walkenhorst O (2008) Estimating the biogenic enhancement factor of weathering using an inverse viability method. *Ecological Modelling* 216: 245–251.
Westbroek P (1991) *Life as a geological force: Dynamics of the earth*. New York: W. W. Norton & Company.

Deforestation

A Shvidenko, International Institute for Applied Systems Analysis, Laxenburg, Austria

Introduction

Conversion of forest to other land-use is an inherent feature in the history of human civilization. During the last 8000 years, the planet has lost ~40% of its original forest cover. By estimates, mature primary tropical forests once occupied 1600×10^6 ha; today their area is about 900×10^6 ha. Latin America and Asia have already lost 40% of their original forest cover, and Africa almost 50%. Most of this loss has occurred during the last two to three centuries. By the beginning of the third millennium, forests have completely disappeared in 25 countries, and forest of 57 countries covers less than 10% of their total land area. Historically, converting forest land to agricultural, infrastructural, industrial, and urban uses was ubiquitous process for development and progress. However, human interactions with forests have often resulted in conversion to unsustainable land uses that led to substantial environmental, social, and economic losses. During the last decades, deforestation and degradation of forests, mostly in the Tropics, have been continuing at an alarming rate. Increasing forest plantations do not change this trend – they account for less than 5% of the world's forest area. Between 1980 and 2005, the Tropics experienced much more intensive forest-cover loss than other regions, with the largest concentration of deforestation occurring in the Amazon Basin, Southeast Asia, and the Congo Basin.

Deforestation and other processes of impoverishment of the world's forests are one of the major drivers of undesirable transformation of the planet. It causes substantial losses of different nature (ecological, economic, social, etc.), dramatically accelerates global biogeochemical cycling, and negatively impacts the Earth system.

Major Definitions

Forest Cover

The global Forest Resource Assessment (FRA), which is provided regularly by the United Nations Food and Agricultural Organization (UN FAO), uses three major terms related to tree cover. Forest is defined as a land class with tree cover more than 10%, and the trees should be able to reach a minimum height of 5 m. Other wooded land (OWL) has either a tree canopy cover of 5–10%, or is presented by a combined cover of shrubs, bushes, and trees above 10%. Other land with tree cover (OLTC) should meet the above criteria for forest but is related to land classified as `other land' (e.g., groups of trees on agricultural land, parks, gardens, etc.). All three definitions require a minimum area of 0.5 ha. The last FRA (2005) estimated the world's forest area as 3952×10^6 ha of which the Tropics (~46%) and the boreal domain (~29%) are major forest biomes. About 30% of land surface is covered by forest. In addition, 1376×10^6 ha was classified as OWL. Global data on OLTC are incomplete – the area of this land-cover category was estimated at 76×10^6 ha in 61 countries out of 229. These estimates are mostly based on national forest statistics and individual country's reports. Another information source on the world's forests is observation of the Earth from satellites. Four different global remote sensing (RS) land-cover products, which have been reported during the two last decades, indicated the area of the world's forests on average as ~20% smaller than the FAO estimate. Reasons for this inconsistency stem from coarse resolution of the imagery used (1 km), not completely compatible definitions of forest, differences in classification, fragmentation of forests in many regions, long-period cloudiness in some (particularly tropical) regions, and lack of satisfactory ground truth data for proper validation of RS imagery. On the other side, national forest inventories are neither complete nor reliable for many developing countries.

Deforestation

Trends in dynamics of global forest cover are defined by two groups of processes. A major driver of decreasing the world forest area is deforestation. Two definitions of deforestation are widely used in their respective international frameworks. By the definition accepted by the UN Framework Convention for Climate Change (UNFCCC) this is an anthropogenic process: deforestation is defined as the direct human-induced conversion of forest to nonforest land. The FRA did not distinguish natural loss of forest from that caused by human impacts: deforestation was defined as the conversion of forest to another land-use or the long-term reduction of the tree canopy cover below the minimum 10% threshold. Both definitions refer to long-term or permanent change from forest to nonforest. Major alternative processes to deforestation are afforestation (establishment of forest on previously nonforest land), reforestation (natural or artificial development of forest on recently forest land), and natural expansion of forests into previously nonforest land. A superimposition of these four major processes results in net change of forest areas.

Here we apply the FAO definition which is used in most assessments and inventories: deforestation includes areas of forest converted to agriculture, natural nonforests such as shrubs and savannas due to anthropogenic impacts or natural disasters,

nonvegetative land (e.g., water reservoirs and urban territories), etc. The term does not include areas where the trees have been removed as a result of forest management activities (e.g., logging) or due to natural disturbances (fire, insects), and where forest is expected to regenerate in a natural way or by silviculture activities. Deforestation also includes areas where some permanently impacted drivers (e.g., disturbance, overutilization, pollution, or other changing environmental conditions) do not allow for maintaining the tree cover above the 10% threshold during a long period of time.

Some other processes like degradation and fragmentation also contribute to the impoverishment of the world's forests and impact global biogeochemical cycling. Three mostly used definitions of degradation (of FRA, International Timber Trade Organization (ITTO), and UN Convention to Combat Desertification (UNCCD)) are comparable with respect to the main clusters. Forest degradation means a process leading to a temporary or permanent decline in the density or structure of forest cover or its species composition, and thus leading to a lower capacity of forest to supply products and/or services, and finally to reduction or loss of the biological productivity of the land. Many reasons can contribute to forest degradation, including diverse human-induced disturbances; unsustainable, excessive forest exploitation; insufficient logging; short rotation periods; etc. In many regions of the world, particularly where forest ecosystems are impacted by accelerated regimes of natural and human-induced disturbances, deforestation and degradation are usually closely interconnected. Fragmentation of forests often leads to decreasing vitality of remaining pieces of forest cover, acceleration of disturbance regimes, and, under the lack of integrated land management, can be a component of structural degradation of forest cover, impoverishment of ecosystem structure, and decline of productivity. Only 22% of the world's forests are classified now as `intact' forests; 82 of 148 countries lying within the forest zone have lost all their intact forest landscapes.

Desertification

The process of (forest) desertification is a specific type of combining deforestation and degradation. The UNCCD defines desertification as land degradation in the arid, semiarid, and dry subhumid areas resulting from various factors, including climatic variations and human activities. Current drylands occupy 5.4×10^9 ha ($\sim$40% of the world land's area) of which $3.5–4.0 \times 10^9$ ha (57–65%) is either desertified or prone to desertification. Recently, a term of `green desertification' was introduced for the boreal biome, meaning a long-term (more than the life span of major forest-forming tree species) replacement of forests by grass-, shrub-, and wetlands, mostly due to disturbances (e.g., fire), the extent, frequency, and severity of which exceed the restoration capacity of forest ecosystems.

Major Drivers of Deforestation and Degradation

Tropical deforestation is driven by a sophisticated combination of direct and indirect drivers of different nature (social, ecological, economic, environmental, biophysical), which interact with each other, often synergistically; the specific combinations of drivers vary within a region of the globe, by countries, and across localities within countries.

Direct drivers are basically human activities at the local level and can be broadly categorized into those related to agricultural expansion, wood extraction, and infrastructure extension. Agriculture expansion is the most important direct driver of deforestation in practically all tropical regions and includes shifting cultivation, permanent agriculture, pasture creation, and resettlement programs, following converting the forest to other land uses. Wood extraction includes commercial logging, fuelwood harvesting, and charcoal production. A substantial negative effect is provided by illegal harvest: over 70 countries have problems with illegal logging that leads to dramatic ecological and economic losses. Commercial logging is an important direct driver in Asia and Latin America while fuelwood gathering is one of the most important drivers in Africa. Infrastructure extension includes construction of transport ways; development of new industrial enterprises; settlement expansion; and a variety of other activities (oil exploration and extraction, mining, construction of hydropower stations, pipeline and electric grids). The construction or paving of roads in forested areas is among the principal causes of deforestation. For example, in the Brazilian Amazon, above 80% of deforestation occurs in a 100 km band along major roads. During recent decades, wildfires have been recognized as a new actor of deforestation and degradation in the Tropics, as a rule following land-use change and fragmentation of forest cover. Exceptional fires took place in east South Asia and the Amazon in 1997–98, provoked by the severe droughts due to the El Niño event. In Indonesia alone, these fires enveloped 2.4 million ha of forest and peatland. Other drivers can be important in different regions of the globe, such as insect damage, drainage or other forms of alteration of wetlands, permafrost destruction in high latitudes, etc.

Indirect drivers of deforestation are caused by fundamental social processes which are usually revealed as a sophisticated interplay of factors of different nature. Economic factors (e.g., rapid market growth and incorporation into the global economy, commercialization, urbanization and industrialization, growth of demand for forest-related consumer goods, poverty, etc.) are crucial across many tropical regions. Institutional factors (taxation, subsidies, corruption, property rights, etc.) are frequently tied to economic drivers. Cultural and sociopolitical factors like lack of public support for forest protection and sustainable use, low educational level, and low perception of public responsibilities also play a substantial role. Population growth, density, and spatial distribution are usually not a primary driver of deforestation: these are always combined with other factors. Nevertheless, in a number of studies, population density has been shown to be highly correlated with the determination of certain land-use patterns often connected to deforestation. Impacts of some of the above factors are often difficult to separate.

Ecological Consequences of Deforestation

The primary ecological consequences of deforestation are decline in biodiversity, invasion of exotic species, destruction of hydrological cycle, increase in water runoff and decrease in water quality, and acceleration of soil erosion. Tropical forests contain between 70% and 90% of all of the world species, and as a result of deforestation the planet is losing between 50 and 130 animal and plant species each day. Deforestation dramatically impacts runoff and hydrological regime, which threatens, for example, about 2000 known species in the waters of the Amazon Basin – 10 times the number found in Europe. Clearing of tropical forests substantially impacts fertility of soil. About 80% of soil in the humid Tropics is acid and infertile. Once the soil temperature exceeds 25 °C, volatile nutrient ingredients like nitrogen can be lost. Due to intensive rainfalls and soil specifics in the Tropics, a single storm can remove up to 100–150 t topsoil per hectare after deforestation. Deforestation can also decrease the social, esthetic, and spiritual values of forested landscapes. The extent and magnitude of these impacts are influenced by the size, connectivity, shape, context, and heterogeneity of the forest patch remnants. A critical point of negative change in landscape functionality – when fragmentation increased rapidly – on average occurs when mature forest declined to 30–35% of the landscape area.

Forest-cover decline alters regional and potentially global climate system by affecting surface energy, water, and greenhouse gas (GHG) fluxes. Deforestation of temperate and boreal forests has a cooling effect on near-surface climate by increasing surface albedo because cultivated fields generally have a higher surface albedo than natural forests. In tropical regions, deforestation generally leads to the opposite response where the prevailing effect is a decrease of evapotranspiration due to lower surface roughness and a shallower rooting zone. The associated decrease in the latent heat flux suggests a warming trend. Change of evapotranspiration and sensible heat flux impacts the low-level atmosphere, regional, and, potentially, global-scale atmospheric circulation. For instance, higher rainfall and warmer temperature are already observed due to recent large-scale deforestation in the Amazon Basin. However, the length of the dry season increases due to deforestation-induced rainfall inhibition, which can be accelerated by rainfall reduction in future due to global warming. The changes in forest cover have consequences far beyond the Amazon Basin. Regional-scale deforestation in the Tropics has been observed in a number of modeling results to lead to remote temperature and precipitation changes. Simulations for the twenty-first century give regional anomalies (due to human-induced land-cover change) as ± 2 K in magnitude.

There are already recognized impacts of current climate change on tropical forests. Biomass and production of pristine tropical forests is increasing but it expected to be reversed. Climate change and fragmentation substantially increased vulnerability of tropical forests to fire. The species composition is changing even in remote areas. Strong negative relationship is recognized between changes of precipitation regime and net primary productivity (NPP) in the humid Tropics.

Global simulations show a clear decline of vegetation productivity with increasing values of its fraction that is appropriated to human use. This decline is the consequence of two effects: reduction of biomass and climatic differences associated with a reduced vegetation cover. An important biospheric feedback of decreasing biomass is changes in the strengths of dissipative processes in terrestrial ecosystems.

Territories prone to deforestation, degradation, and desertification contain a huge amount of organic carbon: world tropical forests contain 220–270 Pg C in vegetation and 220 Pg C in soil (down to a depth of 1 m); drylands and boreal forests contain, respectively, *c.* 240 Pg C and 470 Pg C in soil. Human-induced land-use–land-cover change (LULCC) destroys the equilibrium state of these carbon pools and eventually impacts stability of the Earth system due to large emissions of major GHGs to the atmosphere because the biomass stock per hectare in standing forest is much higher than in any replacement use, including tree crops and silviculture, and as a result of substantial decrease of soil carbon after the conversion.

Understanding the Extent of Deforestation and Degradation of Forests

Lengths of retrospective periods of documented LULCC, as well as reliability of data, vary by continents and countries. Relatively reliable `reconstructions' of global land-use dynamics have been done since the 1850s. According to these estimates, between 1850 and 1990, the area of cultivated lands, worldwide, has been estimated to have increased by more than a factor of 4, from 320×10^6 ha in 1850 to 1360×10^6 in 1990. The most rapid increase occurred in tropical regions after the 1940s – about half of the increase of $\sim 1000 \times 10^6$ ha occurred during this period. About 730×10^6 ha of agricultural lands was cleared from forests and woodlands that reduced the area of the world forests by 17% since 1850. The total net flux of carbon to the atmosphere from changes in land use was estimated to be 124 Pg C over the period 1850–1990. Changes in the forest area accounted for almost 90% of the net long-term carbon flux.

Currently, there are two major sources for monitoring LULCC including deforestation and degradation: RS and data of national forest inventories. RS was provided either at global or biome scale using imagery of coarse spatial resolution, or by statistical sampling of fine resolution. The basic tradeoff between these two groups of RS instruments is between spatial and temporal resolution. For example, Landsat's revisit time is 16 days – for satellites of coarse resolution, near-daily. Due to the high probability of cloud cover in many regions and the presence of smoke from vegetation fires, instruments of fine resolution cannot provide a satisfactory complete coverage. However, fine-resolution sensors cannot be avoided due to patchy structure of tropical deforestation with many small plots. Currently, a number of different satellites (optical bands) are used: SPOT (20 m resolution), IRS-2 (6–56 m), Landsat 5 and 7 (30 m), Terra (250–1000 m), ENVISAT (300 m), some others. Use of radars, which can penetrate the cloud cover, is one of the alternatives for estimating deforestation by satellites. Applications of radars from new satellites (e.g., SAR

Table 1 Annual average rates of tropical deforestation (10^6 ha yr^{-1})

Region	Average annual rates of deforestation			Net loss of forest area	
	1980s[a]	1990s[a]	2000–05[b]	1990–2000[c]	2000–05[c]
America	4.4–7.4	4.0–5.2	4.6	−4.5	−4.6
Asia	2.2–3.9	2.7–5.9	3.8	−0.8	+1.0
Africa	1.5–4.0	1.3–5.6	4.1	−4.4	−4.0
Total	8.1–15.3	8.0–16.7	12.5	−10.3	−7.6

[a]Range due to available publications and results of surveys.
[b]FAO (2005) data estimated as the total area for countries with net loss of forest area; data for Asia additionally include 0.4 × 10^6 ha for Oceania.
[c]FAO (2005) Global Resources Assessment 2005. Progress towards sustainable forest management. *FAO Forestry 147*, 350pp. Rome: Food and Agriculture Organization of the United Nations; the area of net change of forest area for USA and Canada are deducted from the total area for the American continent.

from ENVISAT, 75 m resolution, or ALOS, 50 m) show promising results. However, identification of small patches of deforested land, distinguishing degraded forests, and indicating regrowth can still not reliably be done from space for large areas. Thus, available estimates of global deforestation are not consistent and reliable enough yet.

Reported areas of deforestation in the Tropics vary substantially (**Table 1**). Several subsequent estimates of dynamics of the global forest cover were provided by FAO FRA, mostly using country surveys which are based on compilation and standardization of data of national statistics. Recently, FRA-2005 presented revised estimates for 1990–2000 and new results for 2000–05. As a total conclusion for 1990–05, the global deforestation rate was estimated at 13 × 10^6 ha per year, almost completely in the Tropics. Taking into account increased areas of forest plantations and natural expansion of forests, particularly in temperate and boreal zones, the global net change of forest area was estimated at −8.9 × 10^6 ha in 1990–2000 (equivalent to loss of 0.22% to remaining area annually) and −7.3 × 10^6 ha in 2000–05 (−0.18%). The largest net loss of forests in 2000–05 was estimated at 4.6 × 10^6 ha yr^{-1} for Central and Southern America followed by Africa, which lost 4.0 × 10^6 ha annually. During the last 5 years, the previously negative trend in Asia has reversed due to large-scale plantations established mostly in China and India. Other estimates of tropical deforestation for the last two decades of the twenty-first century vary from 8 × 10^6 to >16 × 10^6 ha yr^{-1}. RS estimates report that over 20 years (from the 1970s to the 1990s) the area of global forest decreased by 6%. On average, the RS estimates report lesser areas of tropical deforestation than FAO estimates. Likely, the above estimates of deforestation rate are slightly overestimated due to the fact that national inventories and RS data do not adequately record the regrowth. However, from another side, small deforested patches and selective logging, as a rule, are not included in the reported area. Probably, an aggregated conclusion on the current level of deforestation in the Tropics of *c.* 10 × 10^6 ha yr^{-1} can be considered as `the best' conservative estimate of this process.

Considering the regional aspect, Brazil reported 21% of the net global loss for 1990–2000 and 24% for 2000–05, but this country has probably the best national RS system of deforestation monitoring: since 1997, the Brazilian National Institute of Space Research (INPE) has been monitoring deforestation down to 6.25 ha. Estimated areas for the three years 2002–05 (August to August) were on average 2.37 × 10^6 ha yr^{-1} with reported error ±4%. Overall, during the last 25 years, the Brazilian Amazon lost an area of forest greater than the size of Germany. For ten countries of Southeast Asia, about 2.3 × 10^6 ha of forests was cleared every year between 1990 and 2000 and transferred to other forms of land use. Annual deforestation in Indonesia was estimated some 1.7 × 10^6 ha in 1987–97 with the increase to 2.1 × 10^6 ha in 2003.

Estimation of Carbon Emissions

Major results for assessing emissions due to land-use change were received using inventory-based approaches or models of different type. Inventory-based models consider all or some of the basic processes: (1) the immediate release of carbon to the atmosphere from organic matter burned at the time of clearing, (2) postdisturbance flux of carbon from decay of slash, (3) accumulation of carbon during regrowth, and (4) changes in soil carbon.

Table 2 contains data on carbon emissions caused by deforestation in the Tropics. The estimates differ substantially: the average annual carbon emissions for 1990–2005 are estimated in the range 0.8–2.2 Pg C yr^{-1} (15–35% of the annual global emissions from fossil fuels approximately during this period) with the overall average at about 1.5 Pg C yr^{-1}. This estimate corresponds well to the estimate of the third IPCC assessment of 1.6 ± 0.8 Pg C yr^{-1} for the period 1987–98 and to recent estimates for 2000–06. Simulations done with the model IMAGE 2.1 estimated C emissions from deforestation from 0.83 Pg C yr^{-1} in 1995, 1.04 in 2000, 1.58 in 2005, to 2.16 Pg C yr^{-1} in 2015. Several estimates of aggregated carbon fluxes from tropical land given by inverse modeling vary from 1.2 to1.5 Pg C yr^{-1}, if both fluxes to the atmosphere and hydrosphere are accounted for.

These estimates do not include carbon emissions from wildfire which could be very high, particularly during years of severe droughts. For instance, recent estimates put global carbon emissions from fires during 1997–98 El Niño event at 2.1 ± 0.8 Pg C, particularly in Indonesia.

The carbon stocks in forests may change without a change in forest area (e.g., selective harvest, forest fragmentation, non-stand-replacing disturbances, shifting cultivation, browsing, and grazing) and accumulation of biomass in growing and recovering

Table 2 Annual carbon emissions from tropical deforestation

Region	Carbon emissions due to deforestation (Pg C yr^{-1})		Total emissions in 1990–2005 (Pg C)
	1990s[a]	2000–05[b]	
America	0.55 (0.35–0.75)	0.55	8.3
Asia	0.72 (0.35–1.09)	0.64	10.4
Africa	0.24 (0.12–0.35)	0.29	3.8
Total	1.5 (0.8–2.2)	1.5	22.5

[a]Range due to available estimates.
[b]Emissions are calculated based on the average estimate for 1990–2000.

forests. During the last two decades, the area of primary natural forests decreased or modified through human intervention by 6×10^6 ha yr^{-1}. Due to FAO estimates, degraded and secondary forests in Africa, America, and Asia covered about 850×10^6 ha in 2002. While deforestation can be measured from space with relatively high accuracy, this is not the case for degradation and secondary regrowth; usually regrowth is spectrally indistinguishable from mature forests as early as after 15–20 years. Forest inventories, as a rule, do not contain any specific data on forest degradation. FAO (2000) estimated the area of disturbances that can be labeled as forest degradation at 24×10^6 ha yr^{-1} in the period 1990–2000; another recent estimate is at 10×10^6 ha yr^{-1}. Estimates of carbon emissions from the degradation of forests (expressed as a percentage of the emission from deforestation) vary greatly – from 5% for the world's humid Tropics to 25–42% for tropical Asia and above 100% for tropical Africa. Another study reports the global net emissions from land-use change in the Tropics including emissions from conversion of forest to other land use (71%) and loss of soil carbon after deforestation (20%), emissions from forest degradation (4.4%), emissions from the 1997–98 fires (8.3%), and sinks from regrowth (–3.7%).

Uncertainties of the above data are high. A number of reasons impact reliability of carbon emissions from deforestation and forest degradation: (1) accuracy of recognizing the areas of tropical deforestation and degradation; (2) weak knowledge of the amount of biomass and soil carbon on areas impacted by the land-use change; (3) fate of deforested land, that is, how much is reverting to secondary forests; (4) how much forests are burnt; and (5) how forest disturbance is affecting soil and forest floor carbon stores. In a number of studies, uncertainties on the amount of CO_2 released are estimated to be 25–50%. For the Brazilian Amazon, for example, a range of 150–280 Mt C yr^{-1} was reported.

The greenhouse impact of deforestation is greater than the difference in carbon stock between the forested and replacement landscapes due to releases of other GHGs, basically methane (CH_4) and nitrous oxide (N_2O) (ozone, carbon monoxide, and some other gases which are produced by deforestation are not direct GHGs; nevertheless, they impact concentrations of CO_2 and CH_4 in the atmosphere). The emissions of these gases do not occur directly with deforestation, but basically with the following land use such as rice cultivation, cattle breeding, application of fertilizers, etc. IPCC-2001 assesses the following contribution of the major GHGs to the enhanced greenhouse effect in 1750–2000: CO_2 – 60%, CH_4 – 20%, and N_2O – 6% (the other 14% are caused by halocarbons which are not produced by the biosphere). The contribution of deforestation to the global greenhouse effect is estimated in the range of 25–35%. Of this total, the contribution of CO_2 is about 15% (or about one-fourth of the global CO_2 emissions), CH_4 9–11% (40–50% of the global methane emissions), and N_2O 2% (from one-fifth to one-third of the global nitrous oxide emissions). Available regional estimates are of a similar magnitude. In the case of Brazilian Amazonia, for example, gases other than CO_2 increase the greenhouse effect by about 35%.

For decades, deforestation and degradation were considered as an almost exceptional phenomenon of the Tropics and arid lands. However, recent years have brought much evidence of possible damage to forests due to ongoing and expected global change in the boreal biome. Forest degradation and deforestation here mostly relate to the increase in frequency and severity of large-scale disturbances, change of hydrological regimes mostly related to permafrost destruction, industrial pressure on landscapes, pollution, and unsustainable logging. For instance, wild vegetation fires enveloped 23×10^6 ha (of which 17×10^6 ha on forest land) in Russia in 2003; during the first years of this century, outbreaks of dangerous insects in boreal forests exceeded 20×10^6 ha in the circumpolar boreal zone, of about the same area in American and Asian continents. The increase in the area of `green desertification' in the Russian taiga zone is estimated to be about 5×10^6 ha during the last two decades. The direct carbon emissions due to a fire in 2003 are estimated to be about 200 Tg C yr^{-1}. Very likely, the expected dramatic warming in high latitudes (up to 6–10 °C) will substantially accelerate processes of northern deforestation and degradation.

Conclusion: Managing Deforestation

While many governments try to provide a legislative basis and to realize measures to slow deforestation, the most recent and thorough deforestation studies offer no suggestion that deforestation is decreasing, either of its own accord or in consequence of policy interventions. On the contrary, increasing global integration of markets and growing demand for agricultural commodities and fuelwood in many regions of the developing world appear to be driving substantial increases in deforestation rates that will result in unsustainable forest management and further declining diverse forest services.

Some models and scenarios predict a substantial `baseline' deforestation, for example, for 2005–2015 ($\times$ 10^6 ha yr^{-1}): South America 3.9, Central America 1.2, Southeast Asia 2.6, Africa 5.2, and the total 12.9, with the average annual carbon efflux at the level of 1.2–2.0 Pg C yr^{-1} during the next two decades. The Special Report of IPCC on LULCC (2001) predicts the average annual accounted carbon stock change due to deforestation at –1.8 Pg C yr^{-1} of which –1.6 Pg C yr^{-1} is expected in the Tropics, and the global result of ARD activities between –1.2 and –1.6 Pg C yr^{-1}. The ongoing climatic change will accelerate negative consequences of the human-induced deforestation: the expected significant warming by the end of the century suggests dangerous implications for forests and human welfare. Studies report that the warming turns more and more tropical rainforest into steppe, and will transform up to 60% of this forest into dry land, dramatically impacting the region's richest biodiversity. Very likely, a similar process will be accelerated in the forest–steppe ecotone of the Northern Hemisphere, with substantial (up to 30%) increase in the area of the desertified steppe.

Tropical countries can reduce deforestation through adequate funding or programs designed to enforce environmental legislation; support for economic alternatives to extensive forest clearing, including carbon crediting; building institutional capacity in remote forest regions; and increase in areas of protective forests. Planted forests provide an opportunity to sequester carbon in vegetation and soils: afforestation and reforestation potentially could achieve annual carbon sequestration rates in live biomass in tropical regions 4–8 versus 0.4–1.2 t C ha^{-1} yr^{-1} in boreal regions, and 1.5–4.5 t C ha^{-1} yr^{-1} in temperate regions. An IPCC scenario (2000) predicts that the maximal amount of carbon that can be sequestered by global afforestation and reforestation activities is 60–87 Pg C on 344 $\times$ 10^6 ha during the first half of the twenty-first century with 70% in tropical, 25% in temperate, and 5% in boreal forests, provided the average annual carbon uptake is at 1.1–1.6 Pg C yr^{-1}. Of course, vast areas of forests converted to agriculture use, particularly to pastures, cannot be expected to recover forests of the original type on a timescale relevant to human planning: secondary forests differ in structure, composition, and productivity from their predecessors.

Reducing the rate of deforestation is another major way to decrease GHG emissions. However, neither the UNFCCC nor the Kyoto Protocol has introduced a satisfactory mechanism reducing GHG emissions from deforestation. Avoided deforestation was excluded from the Clean Development Mechanism, and the current international climate policy regime does not provide incentives for developing countries to reduce carbon emissions from tropical deforestation. This problem is under intensive international debates. One of the relevant ways how to curb emissions from deforestation is a so-called compensated reduction of tropical deforestation – the idea that tropical countries might reduce national deforestation under a historical baseline and be allowed internationally tradable carbon offsets having demonstrated reductions. Recent estimates assume that net deforestation would continue until the price of 1 t of sequestered carbon will be less than \$100 t^{-1} C. Such a price could give a possible decrease of carbon fluxes due to avoided deforestation at 300–650 Tg C yr^{-1}.

Tropical deforestation may be decisive in global efforts to stabilize GHG concentrations at levels that avoid dangerous interference in the Earth system. However, it will require substantial international and national efforts in many aspects, for many nations, at all times.

See also: Ecological Data Analysis and Modelling: Forest Models. Terrestrial and Landscape Ecology: Forestry Management

Further Reading

Achard, F., Eva, H.D., Mayaux, P., Stibig, H.-J., Belward, A., 2004. Improved estimates of net carbon emissions from land cover change in the Tropics for the 1990s. Global Biogeochemical Cycles 18, GB2008. doi:10.1029/2003GB002142.

DeFries, R.S., Houghton, R.A., Hansen, M.C., *et al.*, 2002. Carbon emission from tropical deforestation and regrowth based on satellite observations for the 1980s and 90s. Proceedings of the National Academy of Sciences of the United States of America 99, 14256–14261.

FAO, 2005. Global Resources Assessment 2005. Progress towards sustainable forest management. *FAO Forestry 147* Rome: Food and Agriculture Organization of the United Nations, 350pp.

Ferniside, P.M., 1997. Greenhouse gases from deforestation in Brazilian Amazonia: Net committed emissions. Climatic Change 35 (3), 321–360.

Ferniside, P.M., 2000. Global warming and tropical land-use change: Greenhouse gas emissions from biomass burning, decomposition and soils in forest conversion, shifting cultivation and secondary vegetation. Climatic Change 46, 115–158.

Geist, H.J., Lambin, E.F., 2002. Proximate causes and underlying driving forces of tropical deforestation. BioScience 52, 143–150.

Hirsch, A.I., Little, W.S., Houghton, R.A., Scott, N.A., White, J.D., 2004. The net carbon flux due to deforestation and forest re-growth in the Brazilian Amazon: Analysis using a process-based model. Global Change Biology 10, 908–924.

Houghton, R.A., 2003. Revised estimates of the annual flux of carbon to the atmosphere from changes of land use and land management 1850–2000. Tellus 53B, 378–390.

Houghton, R.A., Joos, F., Asner, G.P., 2004. The effect of land use and management on the global carbon cycle. In: Gutman, G., Janetos, A.C., Justice, C.O., *et al.* (Eds.), Remote Sensing and Digital Processing Series, Vol. 6: Land Change Science. Amsterdam: Kluwer Academic, pp. 237–256.

Lambin, E.F., Geist, H., Lepers, E., 2003. Dynamics of land use and cover change in tropical regions. Annual Review of Environment and Resources 28, 205–241.

Lepers, E., Lambin, E.F., Janetos, A.C., *et al.*, 2005. A synthesis of information on rapid land-cover change for the period 1981–2000. BioScience 55 (2), 115–124.

Moutinho, P., Schwartzman, S. (Eds.), 2005. Tropical Deforestation and Climate Change. Washington, DC: Amazon Institute for Environmental Research, p. 132.

Phillips, O.L., Malhy, J., Vinceti, B., *et al.*, 2002. Changes in growth of tropical forests: Evaluating potential biases. Ecological Applications 12, 576–587.

Santili, M., Moutinho, P., Schwartzman, S., *et al.*, 2005. Tropical deforestation and the Kyoto Protocol: An editorial essay. Climatic Change 71, 267–276.

Shvidenko, A., Barber, C.V., Persson, R., *et al.*, 2005. Forest and woodland systems. In: Hassan, R., Scholes, R., Ash, N. (Eds.), The Millennium Ecosystem Assessment Series, vol. 1: Ecosystems and Human Well-Being: Current State and Trends. Washington, DC: Island Press, pp. 585–621.

Watson, R.T., Nobble, I.R., Bolin, B., *et al.* (Eds.), 2000. Special Report of the Intergovernmental Panel on Climate Change: Land Use, Land-Use Change, and Forestry. Cambridge: Cambridge University Press.

The Earth System and Climate Science: Understanding a Very Complex Entity

Hans Joachim Schellnhuber and Maria A Martin, Potsdam Institute for Climate Impact Research, Potsdam, Germany

Glossary

2° guardrail Upper limit of global warming that is considered to most likely avoid catastrophic changes.

Anthroposphere Manifestation of human presence within the Earth System.

Ecosphere Biosphere-geosphere complex.

Paris agreement International political agreement from 2015, adopting the 2° guardrail.

Tipping element Element of the climate system that can undergo rapid and potentially irreversible, qualitative changes under global warming.

Ecosphere, Anthroposphere and Earth System Science

What is the "ecosphere"? Most of us will nowadays quickly type an unknown term into their search engine. In this case we would be confronted with an advertisement for a little world under glass, an aquatic system containing algea, bacteria, and even small shrimps (**Fig. 1**).

It is fascinating to observe a little world like this on its own, from a god-like perspective—but shatter the glass and you will destroy it. The real ecosphere on planet Earth is so much more robust, of course … or is it?

The history of the ecosphere began about 4 billion years ago, when the conditions on planet Earth finally became habitable for the first simple organisms. The interactions between them and their environment started with the development of basic forms of metabolic engines, life forms which drew on chemical compounds from their environment, and later recycled the waste products of other life forms as their food. In turn, the environment was changed by living things early on, the composition of the atmosphere being altered by the output of organisms, aiding with its oxidation. There has been a vivid partnership where life on Earth and the environment jointly developed, climbing up the *coevolution ladder*. This concept has been described over a decade ago by Lenton *et al.* (2004) (**Fig. 2**).

The authors extend it from its application to the early development of life to the major stages of human history—from the appearance of the hunter-gatherer *Homo sapiens*, via the *hydraulic societies* in the valleys of the Nile, Euphrates, Tigris and Indus, to our current state as a modern technical civilization in the trap of our own productivity; namely in a state of global industrialization. The degree to which we are currently interacting with the environment is unprecedented: The global metabolism revolving vast amounts of carbon, nitrogen, phosphorus and sulfur has been shifted markedly, land use alters the face of the Earth, and ocean acidification threatens marine life.

The study of the ecosphere, that is, the biosphere-geosphere complex, therefore needs to be complemented by consideration of its most recent offspring, the anthroposphere. This is indispensable in order to fully understand the design of the most recent bars of the coevolution ladder. The (inter)discipline aiming at achieving this completion is *Earth System Science*. Embracing this new scientific endeavor brings along a veritable shift in understanding, which has been called the *Second Copernican Revolution* (Schellnhuber, 1999). While the first Copernican Revolution was based on the desire to understand stellar phenomena and thereby place Earth in its proper astrophysical context, this one is directed at our planet itself with its seemingly endless complexity, weaving together the interacting physical, chemical, biological and not the least human components (**Fig. 3**).

It is revolutionary indeed because it again places us, humanity, in a new context: our presence and our activities on Earth have shifted, altered and even destroyed multiple elements of the system or interaction pathways. We are on the verge of triggering large-scale regime shifts, of which climate change is undoubtedly the most dangerous one. In other words: The anthroposphere is now dominating the ecosphere. Welcome to the *Anthropocene* (Crutzen, 2002)!

In order to direct this domination into a sustainable regime, we need to identify a safe operating space for humanity, respecting certain *Planetary Boundaries*. The principle was first introduced as a feature in *Nature* as part of the *Road to Copenhagen* program (Rockström *et al.*, 2009). It is based on the identification of crucial biophysical subsystems or processes in the Earth System like biochemical flows, land-use or climate change and their corresponding control variables, like flows of phosphorus to the ocean and biological nitrogen fixation, the percentage of forested land or atmospheric CO_2 concentration. A recent update on the state of the system (Steffen *et al.*, 2015) quantifies the control variables for seven of these planetary boundaries, several of which have been already transgressed. Climate change must be placed in the "zone of uncertainty," where thresholds might have been crossed, moving the climate system into a fundamentally different regime (**Fig. 4**).

Those thresholds play a vital role in a powerful paradigm of sustainability, epitomized by the concept of tipping elements in the climate system, which will be discussed in some detail in the following.

Fig. 1 A search-engine result for the term *Ecosphere*. From Flickr with CC-BY: https://www.flickr.com/photos/wicker-furniture/14808673001.

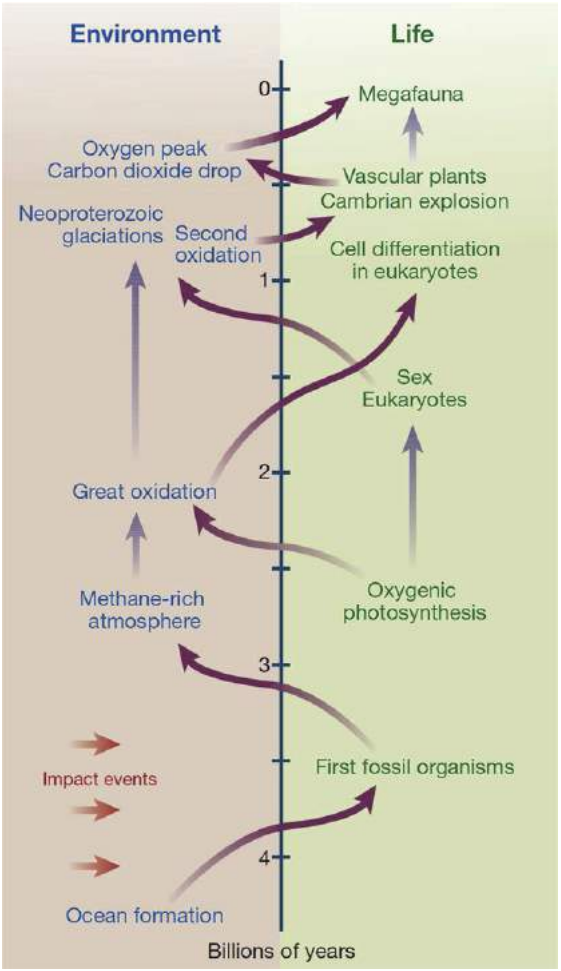


Fig. 2 The Coevolution ladder (Lenton *et al.*, 2004).

Tipping Elements

In the year 2003, a Dahlem Workshop was held in Berlin, Germany, bringing together pioneers at the new scientific frontier of Earth System science. The discussion covered long-term geosphere-biosphere interactions, the mode of operation of the Earth System during the Quaternary, its current state in the age of the Anthropocene and the potential transition towards sustainability. The freely available introductory chapter of the resulting book *Earth System Science for Sustainability* (Schellnhuber *et al.*, 2004) covers these topics and gives an excellent overview of the prospects for this new field of science (Clark *et al.*, 2005). It also

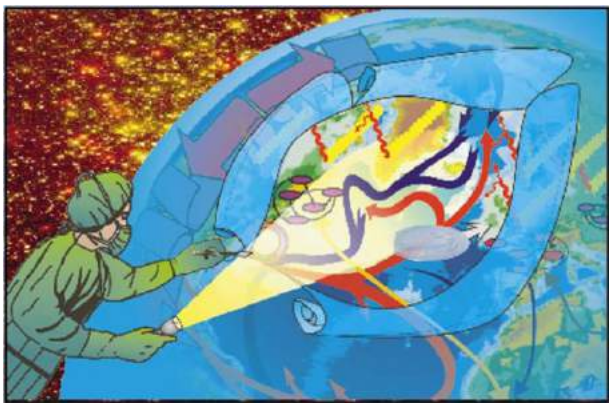


Fig. 3 Dissecting the Earth System (Schellnhuber, 1999).

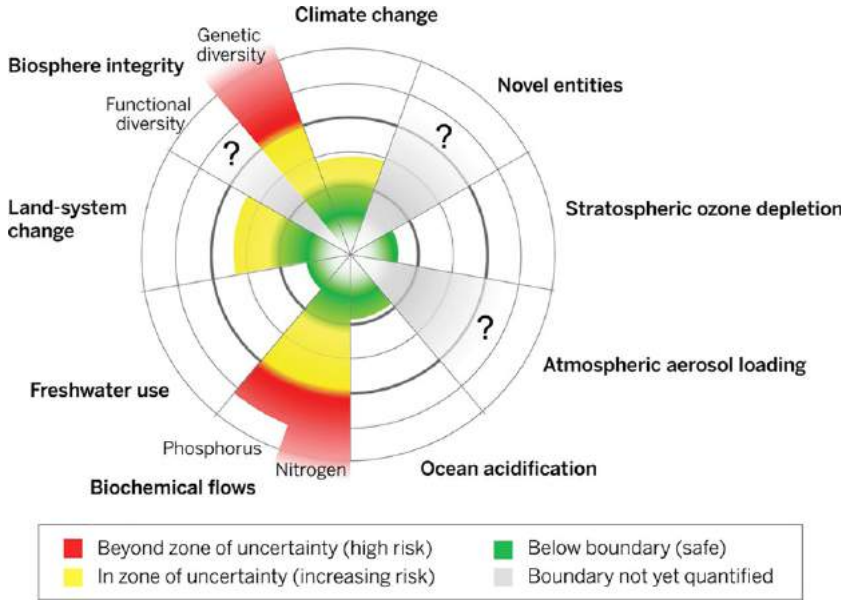


Fig. 4 Planetary Boundaries. From Steffen, W., Richardson, K., Rockström, J., Cornell, S., Fetzer, I., Bennett, E.M., Biggs, R., Carpenter, S.R., de Wit, C.A., Folke, C., Mace, G.M., Persson, L.M., Veerabhadran, R., Reyers, B., Sörlin, S., Cornell, E., Fetzer, I., Bennett, E.M., Biggs, R., Stephen, R., De Vries, W., De Wit, C.A., Folke, C., Gerten, D., Heinke, J., Mace, G.M., Linn, M., 2015. Planetary Boundaries: Guiding human development on a changing planet. Science 347 (6223), 1259855. https://doi.org/10.1126/science.1259855. Reprinted with permission from AAAS.

elaborates on the fact that Earth System science cannot rely on the valuable knowledge generated by natural sciences alone, but has to from an alliance with the social sciences in order to truly promote sustainability and not again step into the old pitfall of underestimating the human dimensions.

Although published around 12 years ago, these thoughts and concepts are still up to date. For instance, the early notion of tipping elements, at the time still called "switch and choke elements," is described and visualized in a manner largely resembling today's iconic graph (**Fig. 5**).

In 2008, a systematic revision of the concept, giving a formal definition along with an extensive literature review resulting in a short list of policy-relevant tipping elements, was published in PNAS (Lenton *et al.*, 2008) and one year later integrated into a special feature of the same journal (Schellnhuber, 2009).

In short, a tipping element denotes a component of the Earth System that is of at least subcontinental scale and enters a qualitative different state if a certain control parameter exceeds a critical value. This change of state might be irreversible, even if the control parameter (which could be the global mean temperature, for example) turns around to resume its original value. The pace at which the state of the tipping element changes after being triggered, however, differs greatly, depending on the nature of the system. Some examples may shed more light on the actual meaning of these general statements.

The Atlantic overturning circulation is responsible for the transport of vast amounts of warm surface water from equatorial regions towards northern latitudes near Greenland and the Labrador coast, where this water cools down and sinks. The resulting

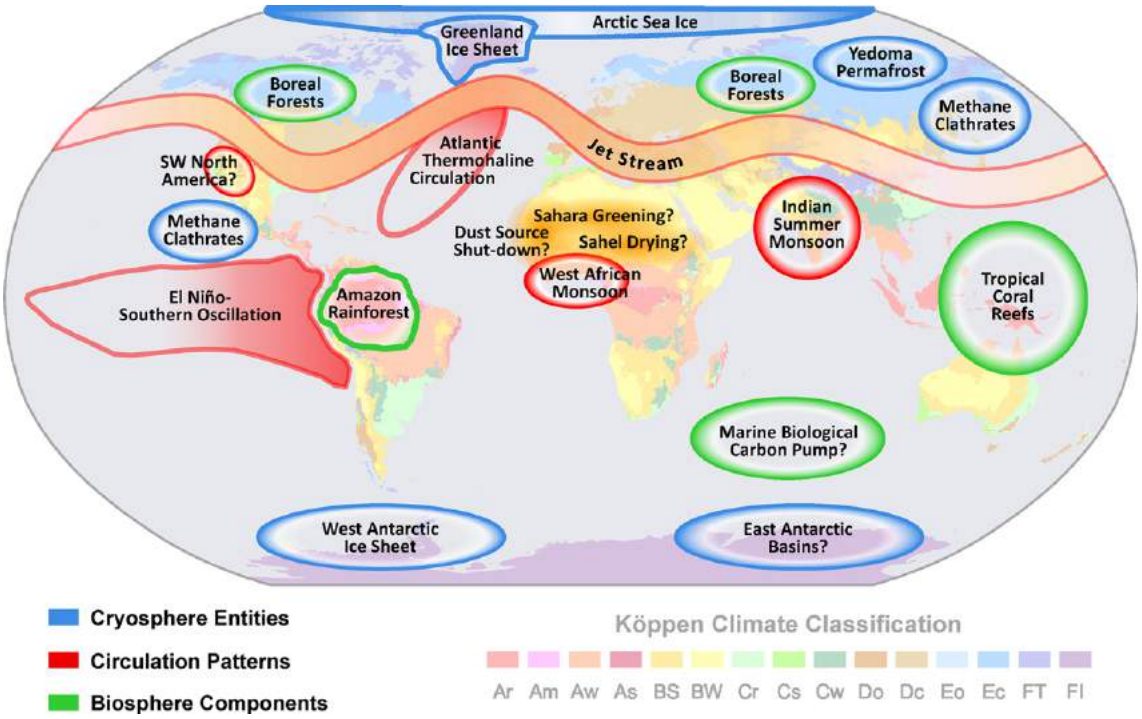


Fig. 5 Tipping Elements in the Earth System. Potsdam Institute for Climate Impact Research. (2017). Website: https://www.pik-potsdam.de/services/infodesk/tipping-elements/kippelemente.

cold, salty masses then flow southwards along the ocean floor. The cooling and sinking acts as a motor for the circulation system, which includes the Gulf Stream. Actually, the fuel for this motor is the very same northward stream of salty water from southern regions with high evaporation rates. This interplay constitutes a self-sustaining cycle, which cannot necessarily re-start once it is interrupted. Such an interruption could be brought about by a freshwater influx from melting polar ice, reducing the density of the water and thereby hindering downwelling—the motor begins to stutter or even shuts down. There is evidence that an exceptional weakening of the thermohaline circulation is currently underway (Rahmstorf *et al.*, 2015). The impacts of such a weakening range from disturbances of marine ecosystems, via cooling of the North Atlantic region to regional sea-level rise on the US-Atlantic coast.

There are more examples for circulation systems within the global environment that can be considered tipping elements. They include the tropical Pacific climate oscillation giving rise to the El Niño phenomenon or the planetary waves of the Jet Stream, which can give rise to extreme weather events in the mid-latitudes when blocking events occur and the waves lock into a certain position, thereby isolating cold polar and warm equatorial air for weeks. Monsoon systems are also prone to disturbance as they are based on a self-reinforcing feedback involving latent heat release during rainfall events.

In West Antarctica, there is enough ice stored to raise global sea level by around three meters. A special topographic situation, with kilometer-thick ice resting on bedrock deep below sea level at the center of the ice sheet, causes an unstable flow regime: Once too much ice is lost at the thinner borders of the ice sheet due to ocean warming or the reduction of buttressing by the floating ice shelves, the outflow keeps growing in line with the ice retreating inland (Joughin and Alley, 2011). The ice sheet may already have crossed a tipping point (see for example Mouginot *et al.*, 2014), leading to unstoppable sea-level rise in the centuries to come. In Greenland, on the other hand, the interplay between warming, melting and lowering of the surface leads to a self-reinforcing feedback (surface-elevation feedback), because the sinking ice surface is exposed to warmer masses of air, inducing further melt. A threshold might be crossed at a global warming level as "low" as 1.6°C (Robinson *et al.*, 2012).

But it is not only the inanimate parts of the Earth System that can undergo disruptive transitions. The unique terrestrial biodiversity of the Amazon rainforest, for example, depends on its own capacity to capture and store rainfall and to evaporate the water again, which will be eventually transported further inland by horizontal advection. If the abundant vegetation is destroyed or replaced with plantations in too many places, the "flying river" carrier of rain, conveying water from the oceans to the core of the continent, might be intercepted. Also, it has become painfully clear in recent years that coral reefs, the beautiful homes of thousands of types of marine life and the sources of nourishment for many more, are in acute danger because of warmer waters and less light received due to sea-level rise. Even if the 2°C guardrail agreed on in the Paris is respected, a breakdown of this marine sphere—a unique world of its own—has to be expected. Ocean acidification also plays a decisive role in the loss of corals, because the build-up of calcium carbonate structures requires waters of relatively mild acidity only. And ocean acidification plays the main part in one of the most exciting recent insights of Earth System science.

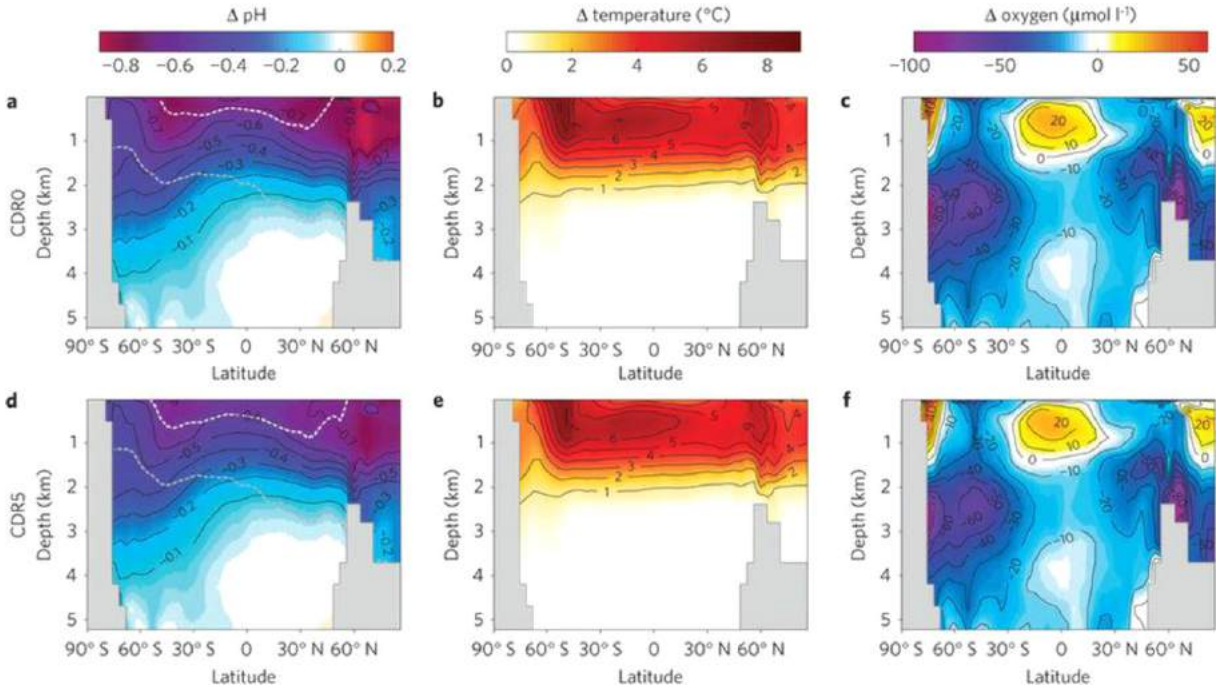


Fig. 6 The virtually nonexistent effect of massive carbon dioxide extraction from the atmosphere on the ocean (Mathesius *et al.*, 2015).

Topical Results of Earth System Science

It is well known that high atmospheric carbon dioxide levels due to anthropogenic emissions indirectly disturb the Earth System by enhancing the greenhouse effect, basically hindering long-wave radiation escaping into space. But there is also a very direct effect: CO_2 dissolves into the oceans, thereby altering the chemical composition of sea water and leading to acidification. This means that even if—by the improper means of climate manipulation (euphemistically called geoengineering) through solar radiation management—global mean temperature was kept stable, the acidification process would keep going and increase in severity. Humanity might feel safe to continue emitting CO_2, because the impacts on land would be stalled—marine life, however, would be subject to major disruptions. But what if we extracted CO_2 directly from the atmosphere? Would ocean acidification then disappear? This question has been answered with a big *NO* by a study involving scientists from the Carnegie Institution for Science in Stanford and the Potsdam Institute for Climate Impact Research (Mathesius *et al.*, 2015). If—in a dire scenario—humanity initially follows a business-as-usual scenario and in a late attempt to repair things starts reducing the net CO_2 content of the atmosphere around 2250, not even the utterly challenging scenario of extracting 18 billion tons of CO_2 each year would be a remedy. Even if extraction then continued for centuries, ocean chemistry would remain basically unimpressed by this effort (second row in **Fig. 6**, compared to the upper row without CO_2 extraction).

Another example of the dimension of the ongoing human interference with the Earth System concerns the natural cycle of ice ages, triggered by delicate changes in the way our planet orbits the sun (*Milankovic cycles*). The distribution of solar radiation impinging on Earth changes according to quasi-periods, for example due to axis precession (26,000 years period) or reflecting eccentricity changes of the elliptical trajectory of the planet (100,000 years period). The interplay of these periods with the nonsymmetric distribution of land on Earth is the underlying cause of drastic effects: ice ages and warm periods (*interglacials*) have now dominated the state of the Earth System and shaped the planet for 2.6 million years (*Quaternary glaciation*). But another player also has a key role here: atmospheric carbon dioxide concentration. Subtle differences in CO_2 levels can make all the difference!

On the basis of careful analysis of the conditions during past glaciations, a critical relationship between summer insolation at 65°N and CO_2 concentration has been derived (Ganopolski *et al.*, 2016). Based on this relationship it can be inferred that the next glaciation, which—without human interference—would be scheduled in about 50,000 years from now, will be canceled assuming even relatively moderate emission scenarios! (**Fig. 7**)

Understanding the Earth System, as has been discussed in this article, is both a tremendous scientific challenge and an absolute necessity. After decades of research and infinite struggles to make the significance of this research apparent to both the public and decision makers worldwide, an international agreement has been reached in 2015 in Paris to keep global warming below the threshold of 2°C, and possibly even below 1.5°C. The contours of this so-called Paris range have not been the result of political negotiations, but were ultimately derived from our best knowledge about the complex system constituting our environment. Recently, an analysis of an extensive body of literature on tipping points in the climate system has been condensed in one single graphic (**Fig. 8**). There are a number of tipping elements that might possibly be switched even within the Paris range, and several

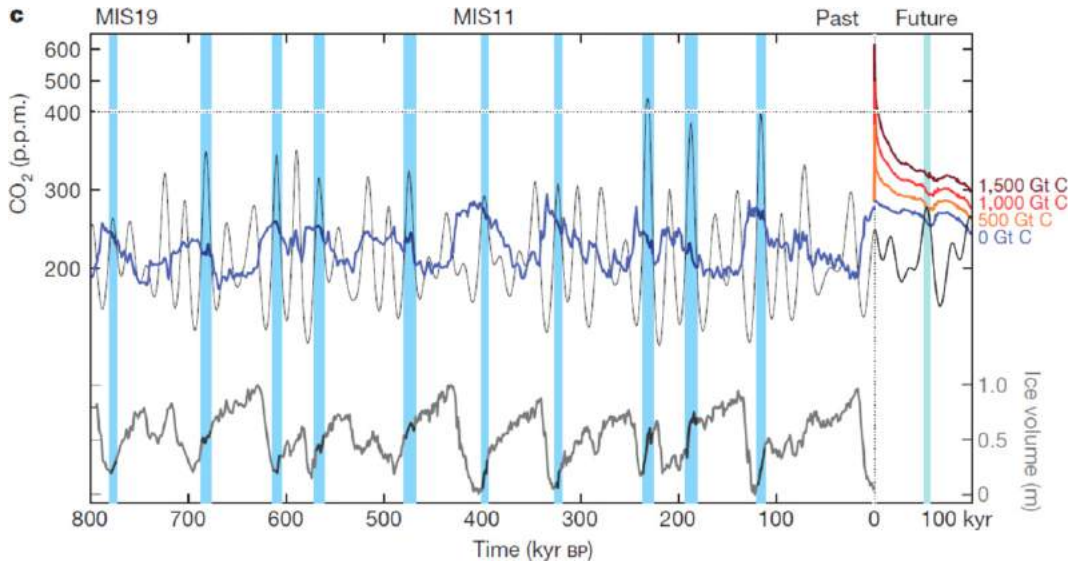


Fig. 7 Past glacial inceptions (*vertical blue bars*) as derived from the empirical relationship between maximum summer insolation at 65°N and CO_2 concentration. The next glacial inception—in 50,000 years from now (*vertical green bar*)—will be missed if either the red or dark red scenarios for anthropogenic carbon emission become reality (Ganopolski *et al.*, 2016).

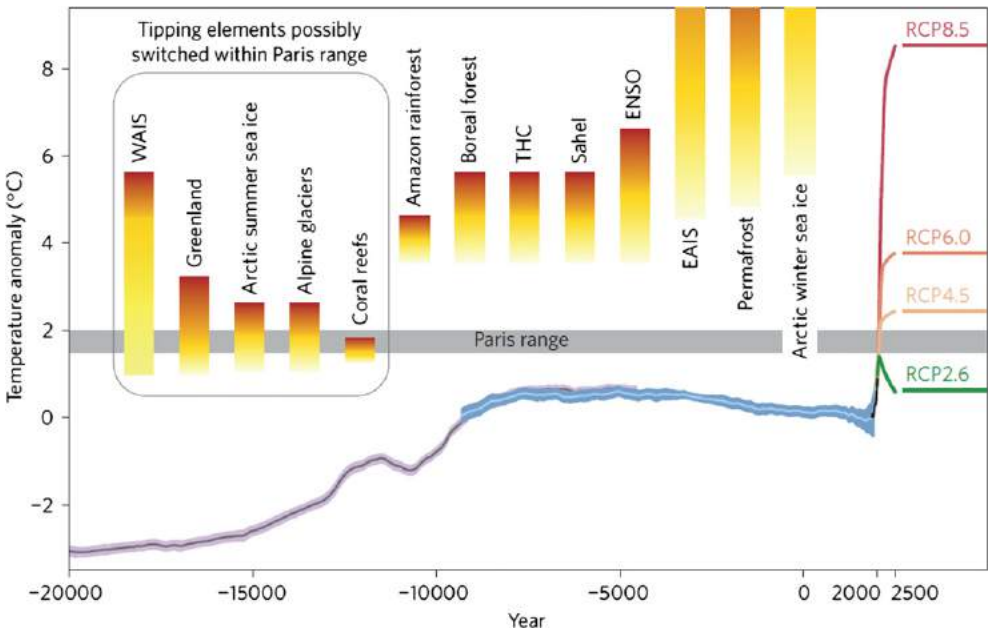


Fig. 8 Tipping elements in the context of 20,000 years of global mean temperature evolution and possible scenarios for the future (Schellnhuber *et al.*, 2016).

more looming not far above. This should make clear, without ambiguity, why the right climate target was agreed on in Paris (Schellnhuber *et al.*, 2016).

See also: Ecological Complexity: Hierarchy Theory in Ecology; Self-Organization; Complex Systems; Systems Ecology. Human Ecology and Sustainability: Socioecological Systems

References

Clark, W., Crutzen, P., Schellnhuber, H.J., 2005. Science for global sustainability: Toward a new paradigm. Working Paper 120 (120), 1–28.

Crutzen, P.J., 2002. Geology of mankind. Nature 415 (6867), 23. doi:10.1038/415023a.

Ganopolski, A., Winkelmann, R., Schellnhuber, H.J., 2016. Critical insolation–CO2 relation for diagnosing past and future glacial inception. Nature 529 (7585), 200–203. doi:10.1038/nature16494.

Joughin, I., Alley, R.B., 2011. Stability of the West Antarctic ice sheet in a warming world. Nature Geoscience 4 (8), 506–513. doi:10.1038/ngeo1194.

Lenton, T.M., Schellnhuber, H.J., Szathmáry, E., 2004. Climbing the co-evolution ladder. Nature 431 (7011), 913. doi:10.1038/431913a.

Lenton, T.M., Held, H., Kriegler, E., Hall, J.W., Lucht, W., Rahmstorf, S., Schellnhuber, H.J., 2008. Tipping elements in the Earth's climate system. Proceedings of the National Academy of Sciences 105 (6), 1786–1793. doi:10.1073/pnas.0705414105.

Mathesius, S., Hofmann, M., Caldeira, K., Schellnhuber, H.J., 2015. Long-term response of oceans to CO_2 removal from the atmosphere. Nature Climate Change 5 (August), 1–8. doi:10.1038/nclimate2729.

Mouginot, J., Rignot, E., Scheuchl, B., 2014. Sustained increase in ice discharge from the Amundsen Sea Embayment, West Antarctica, from 1973 to 2013. In: Geophysical Research Letters., pp. pp. 1576–1584. doi:10.1002/2013GL059069.1.

Rahmstorf, S., Box, J.E., Feulner, G., Mann, M.E., Robinson, A., Rutherford, S., Schaaernicht, E.J., 2015. Exceptional twentieth-century slowdown in Atlantic Ocean overturning circulation. Nature Climate Change. 1–6. (March).https://doi.org/10.1038/NCLIMATE2554

Robinson, A., Calov, R., Ganopolski, A., 2012. Multistability and critical thresholds of the Greenland ice sheet. Nature Climate Change 2 (4), 1–4. doi:10.1038/nclimate1449.

Rockström, J., Steffen, W., Noone, K., Persson, A., Chapin, F.S., Lambin, E.F., Lenton, T.M., Scheffer, M., Folke, C., Schellnhuber, H.J., Nykvist, B., de Wit, C.A., Hughes, T., van der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P.K., Costanza, R., Svedin, U., Falkenmark, M., Karlberg, L., Corell, R.W., Fabry, V.J., Hansen, J., Walker, B., Liverman, D., Richardson, K., Crutzen, P., Foley, J.A., 2009. A safe operating space for humanity. Nature 461 (7263), 472–475.

Schellnhuber, H.J., 1999. 'Earth system' analysis and the second Copernican revolution. Nature 402 (December),

Schellnhuber, H.J., 2009. Tipping elements in the earth system. Proceedings of the National Academy of Sciences 106 (49), 20561–20563. doi:10.1073/pnas.0911106106.

Schellnhuber, H.J., Crutzen, P.J., Clark, W.C., Claussen, M., Held, H., 2004. In: Schellnhuber, H.J., Crutzen, P.J., Clark, W.C., Claussen, M., Held, H. (Eds.), Earth system analysis for sustainability. Cambridge, MA, London, UK: MIT Press. Dahlem Wor.

Schellnhuber, H.J., Rahmstorf, S., Winkelmann, R., 2016. Why the right climate target was agreed in Paris. Nature Climate Change 6 (7), 649–653. doi:10.1038/nclimate3013.

Steffen, W., Richardson, K., Rockström, J., Cornell, S., Fetzer, I., Bennett, E.M., Biggs, R., Carpenter, S.R., de Wit, C.A., Folke, C., Mace, G.M., Persson, L.M., Veerabhadran, R., Reyers, B., Sörlin, S., Cornell, E., Fetzer, I., Bennett, E.M., Biggs, R., Stephen, R., De Vries, W., De Wit, C.A., Folke, C., Gerten, D., Heinke, J., Mace, G.M., Linn, M., 2015. Planetary Boundaries: Guiding human development on a changing planet. Science 347 (6223), 1259855. https://doi.org/10.1126/science.1259855

Emergence of Climate Change Ecology

Sergey Venevsky, Tsinghua University, Beijing, China

Introduction

The term "climate change ecology" is rather recent. Scopus search for this term results in only 42 articles with the term firstly mentioned in 2004 (Forchhammer and Post, 2004). Meanwhile, the leading scientific journal like Nature and Science are publishing Special Issues or sections with the title "Climate Change Ecology" (e.g. "Science" from August 2, 2013). Here, I discuss a subject of this newly appeared branch of ecology, its methods and current state.

What Is Climate Change Ecology

It can be defined that climate change ecology is a part of ecology dealing with climate change impacts upon terrestrial and marine ecosystems.

Climate Change

External and internal forces of climate

Climate at a certain location is a feature of Earth described by a sequence of long period (decadal, or century) averaged climate variables with temperature, precipitation, wind speed and direction, short wave radiation and relative humidity most important variables for the atmosphere (for the oceans the most important climate variable is sea surface temperature). Climatic variables can be considered at different time scale (larger than daily, i.e., 5-days, weekly, monthly, annually or for decades). Climate is different around the globe it may be polar, temperate, tropical (or has other subdivision) depending on temperature or may be arid or humid (depending on precipitation). Classification of Earth climates was suggested by Koppen (Wladimir and Alfred, 2005) and refined by Geiger in 1961 (see **Fig. 1**).

Climate of the Earth is controlled by external and internal forces. External driving forces for the Earth's climate are gravitation of nearest space bodies, which determine position of the planet to the Sun, and Solar irradiation. Climate internal driving forces are more numerous. They include gas and particle composition of the atmosphere, distribution of land and ocean masses, topography of land surface and bottom of the ocean. Especially important for providing climate suitable for life is gas content of the atmosphere. Greenhouse gases like water vapor, carbon dioxide, methane and nitrogen oxides are absorbing outgoing long wave radiation from the planet surface into outer space and, thus keep the average temperature of the Earth being 14°C. Humans by actively exploring Earth natural resources changing gas composition of the atmosphere. Exponentially increasing anthropogenic greenhouse gases emissions are responsible for additional warming of the atmosphere to almost 1°C since preindustrial period. The observed and projected increase of average global temperature induced by humans has differential regional manifestations, including not only seasonal warming, but cooling, as well as extreme climate events like prolonged droughts or floods, hurricanes etc. Projected climate change to the next century will substantially reshape Koppen-Geiger climate zones (see **Fig. 2**). Impact of climate change upon ecosystems and humans at different spatial scales at recent or at geological time steps is a subject of climate change ecology.

Climate change: Natural and man-made

Quantitatively climate change is determined as long-term (decadal to millions of years) change in statistical parameters of regional or global weather patterns. Mean and spread are major statistical parameters of climate change. For instance, annual average global temperature is a main statistical parameter used in description of global climate change. It was shown that annual average global temperature had significant variation (3–8°C) during history of the Earth. Evidence of climate change at decadal to geological time scales is confirmed by different observation sources. Temperature direct measurements from surface meteorological stations or radio balloons in troposphere allows direct record of climate change on decadal scale. Recent satellite observations allow conduction of observations of climate variables with fine spatial resolution using different proxies. Glaciers considered the best indicators of climate change in the past. Analysis of ice cores from glaciers revealed that since 3 million years glacial and interglacial cycles caused by orbital forcing were major regulators of climate change. However, data from ice cores drilled in central Greenland demonstrated that abrupt climate change during glaciation may persist without orbital forcing due to fast inner restructuring of climate system. Dendrochronology, which is focusing on width of tree rings as an indicator of dry and wet and warm and cold years is demonstrating evidence of climate change at century time scale. Palynology looks in distribution of fossil pollens reflecting pattern of past vegetation zones which were constrained by past climates.

 doi:10.1016/B978-0-12-409548-9.10583-4

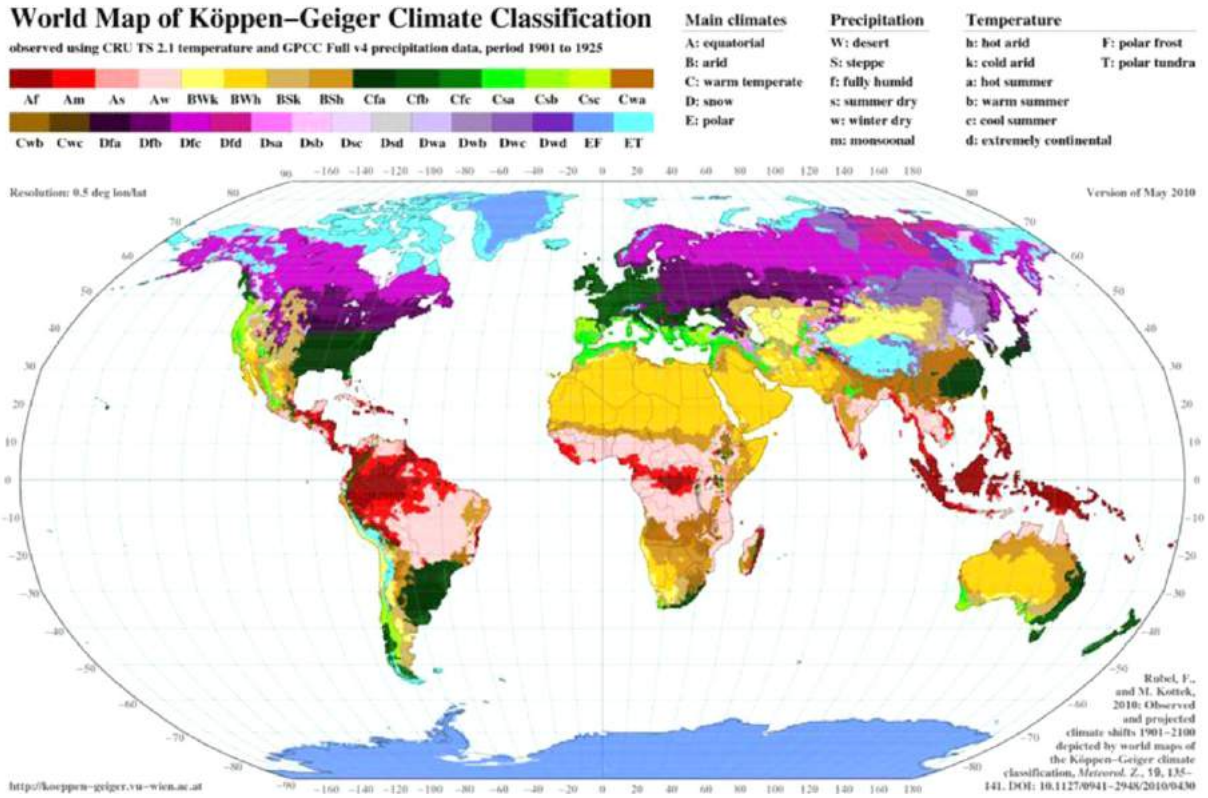


Fig. 1 Koppen-Geiger classification of climate zones 1901–1925 (Rubel and Kottek, 2010).

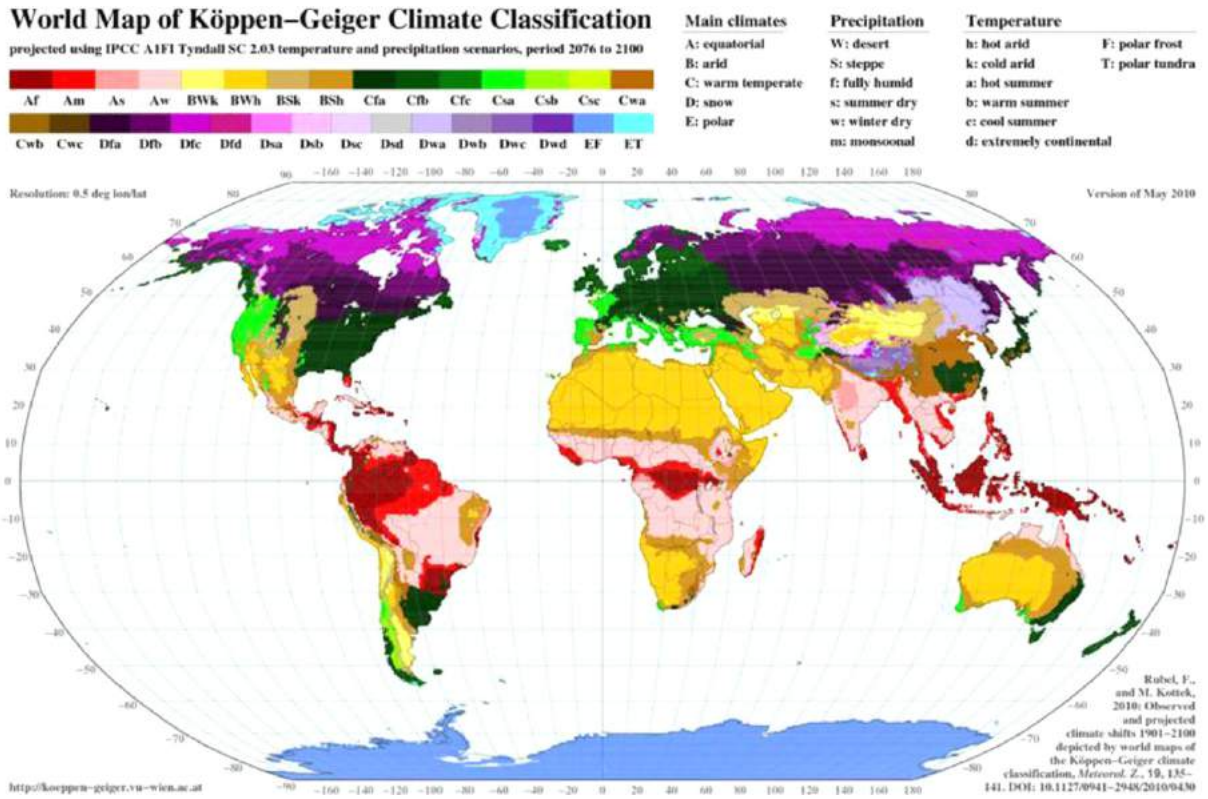


Fig. 2 Koppen-Geiger classification of climate zones 2076–2100 (Rubel and Kottek, 2010).

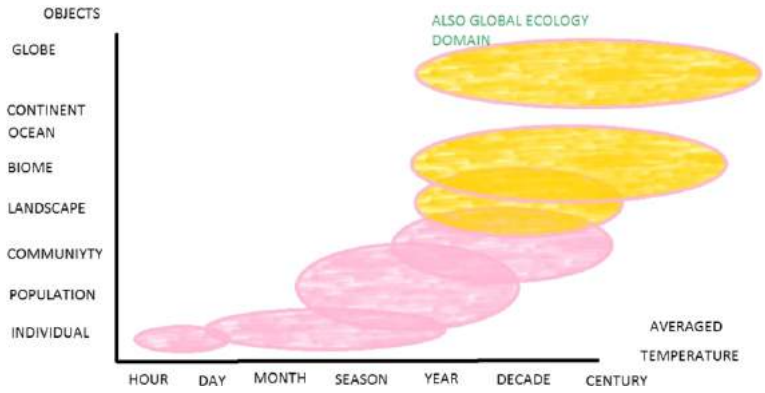


Fig. 3 Variety of objects of climate change ecology. Objects are shown in a space of one from averaged over long period climatic variables. Golden color depicts objects for which consideration of biophysical changes are considered in line with ecological changes. Intersection of objects assume consideration of intrainfluence of objects with different level of organization. Globe is an object not only of "climate change ecology," but as well an object of "global change ecology."

Climatic system consists of several elements: atmosphere, ocean, biosphere and cryosphere. Interaction of atmosphere and ocean due to transfer of heat, distributed differentially in space, creates inner climate variation. Biosphere is another important regulator of inner climate variability because of its role in global carbon and water cycle. Humans actively exploiting carbon stored in former biospheres as fossil fuel are changing gas content of the atmosphere and its radiative transfer abilities. Excessive amounts of greenhouse gases, first of all of carbon dioxide, may the reason of recent increase in global temperature named by general public as global warming, The so named "global warming," associated with change of global surface temperature, in regional realization can be seen as drying, wetting or cooling, as well as change in frequency of extreme weather events like hurricanes, because of complexity of interactions of elements of climate system and its highly nonlinear dynamics (Shepherd, 2014).

In course of time anthropogenic and natural forcing mechanisms may change each other even at decadal time scale. This can be seen for instance in XX century, where in the beginning of century increasing trend in global surface temperature is attributed to variation of solar irradiance, while relatively quick rise in global surface temperature in the last decades of the century is associated with rapid rise in anthropogenic greenhouse emissions (Stott *et al.*, 2000). The climate change caused by humans in the last few decades is a subject of great concern of society as consequences of climate change are already observed everywhere in biosphere and anthroposphere.

Ecological Response to Recent Climate Change

Impacts of recent climate change are observed almost everywhere, despite value of global surface temperature changed only to several percent to absolute value in last decades. We can divide conditionally these impacts to (a) biophysical changes which related to change in chemical fluxes and stores in ecosystems of different scales from the global to local both natural and human-dominated (urban and agricultural); (b) ecological changes, which include changes of abundance ranges and interactions between biological objects (organisms, populations, biomes). Accordingly climate change ecology studies impacts of climate change upon as living so nonliving objects of different temporal and spatial scales (see simplified scheme at **Fig. 3**). We figure further an overview of (a) most cited and (b) most recent findings of climate change ecology for different scales and objects.

Biophysical changes

Studying of global warming induced biophysical changes (changes of energy, information and matter fluxes and pools) at different spatial and temporal scales are most popular topics in climate change ecology. Research of alterations of biogeochemical cycles in terrestrial and marine ecosystems are central in studying of climate driven biophysical changes.

Climate change and global biogeochemical cycles

Climate change and water cycle

Global water cycle is significantly influenced by climate. There is a general agreement in projections that global water cycle accelerates with increase of global surface temperature. However, despite of amount of fresh water available for humankind is suggested to increase, regional distribution of water is supposed to has a great disparities in space (Oki and Kanae, 2006). This can increase vulnerability of water-stressed areas where according to several estimates leaves almost one third of human population of the Earth. Change of terrestrial water cycle is the key research focus in climate change ecology, as it determines future of humankind. Unfortunately, recent studies confirm that aridity over the land is being accelerated in course of climate change due to the land-atmospheric feedback, provides by soil moisture, and reaction of vegetation to CO_2 enhancement in the atmosphere (Berg *et al.*, 2016).

Climate change and carbon cycle

Alterations of global carbon cycle are closely related to climate change. Study of such alterations constitutes a core of climate change ecology equally to research in water cycle alterations. Anthropogenic carbon dioxide emissions are affecting climate due to greenhouse effect. Around half of these emissions are absorbed by oceans and marine and land ecosystems currently, while second half is gradually accumulated in the atmosphere. The absorbing ability of the ocean and the ecosystems are climate sensitive, so that increase of global surface temperature may decrease global absorption of carbon from the atmosphere and accelerate climate change even in the nearest future (year 2050 see Cox *et al.*, 2000)). Recent findings for Arctic underline close interconnection between marine and terrestrial part of carbon cycle in course of global warming. It appeared that sea-ice melt, permafrost thaw and release of fresh water in the Polar Ocean regulate carbon storage and release as from terrestrial so for marine ecosystems (Parmentier *et al.*, 2017).

Climate change and nitrogen cycle

Humankind accelerates global nitrogen cycle by industrial production of fertilizers to solve food problem. This acceleration, however, has significant direct effect on global carbon cycle, especially to absorbing abilities of the biosphere. Question of mitigation of climate change by artificial controlling of nitrogen cycle, however, is open due to counterbalancing negative effects of euthropication and global acidification of terrestrial and aquatic ecosystems (Gruber and Galloway, 2008). Recent manipulation mesocosm experiments over Tibetan alpine wetlands demonstrated that lowering of water table observed here due to climate change now days in combination with nitrogen deposition from human land use synergistically influence emissions of greenhouse gases from the area because of reactions of microbial communities in soil (Wang *et al.*, 2017).

Climate change and terrestrial ecosystems

Impact of climate change to ecosystems as terrestrial ones so marine ones has principal difference with the impact upon biogeochemical cycles. Indeed, climate driven alterations of biogeochemical cycles described by a set of rather similar (although rather complicated) chemical reactions in living and nonliving objects. Climate control of ecosystems is more subtle as living objects with their unique climate response are interacting within an ecosystem at individual, population and community levels. This interaction is a consequence of evolutionary history, sharpened by climate at different time scale and unequally distributed in geographical domains. Thus, a question of location and timing will be defining when studying climate change impact upon ecosystems. Besides, marine ecosystems, living in environment strikingly different by chemical and physical features from terrestrial environment have their own peculiar features of climate control over them.

Climate change and natural terrestrial ecosystems

Research of impact of climate change to terrestrial ecosystems is not new. Simultaneous observation records for weather and changes in phenologies, life cycles and ranges for birds, animals, herbs and trees are coming back to 1700 in Northern Europe. Especially, charismatic taxons like butterflies or birds have well documented responses of body size, reproductive functions and ranges to climate variation.

Now days, the number of published studies, observational or experimental or modeling, relating climate change to terrestrial ecosystems is increasing exponentially since 2003 (see Parmesan, 2006).

Mostly these studies orient to phenological responses to climate change. Plants as primary producers in terrestrial ecosystems are in major focus of phenological observations. So, it was observed using remote sensing that normalized difference vegetation index (NDVI), which describes summer photosynthetic activity, was increasing in the Northern Hemisphere during 1981–1991 in line with increase of the hemisphere average surface temperature (Myneni *et al.*, 1997). On-ground observations in the network of European Phenological Botanical Gardens revealed lengthening of growing season to 10.8 days in the period 1953–1993 (Menzel, 2000) as a response to global warming. However, recent research revealed that systematic identifying of climate windows, which include number and type of variable (if it is temperature, rain, air humidity, wind e. t. c.) and critical timing for physiological, behavioral, life-history, demographic, population and community traits is lacking (van de Pol *et al.*, 2016). More rigorous identification of climate windows at each spatial and temporal levels of living objects is necessary.

Climate change and agricultural ecosystems

Projected climate change can significantly influence crop production mainly due to availability of water. Combination with rapid growth of population in some areas makes climate change a major threat for food security. So, an analysis of climate risks to food security conducted on scenarios for 20 global circulation models till year 2030 confirmed that regions of South Africa and South Asia will most likely suffer from climate induced drop for important crop yields, which will lead to food insecurity for their fast growing populations (Lobell *et al.*, 2008). Using of cover crops (legumes, clover, mustard etc.) which are historically aimed for nitrogen fixation or reduction of soil erosion for mitigating of climate change is a new and blooming idea relating climate change and agricultural ecosystems (Kaye and Quemada, 2017).

Climate change and aquatic ecosystems

Climate change and marine ecosystems

Although there is a clear bias toward study climate change impact to terrestrial ecosystems, phenological response of marine species also got attention recently. For example, Arctic seabird *Uria lomvia*, has advanced its egg-laying date few days at its southern

boundary (Hudson Bay). The change of date is closely correlated to change in sea-ice cover (Gaston *et al.*, 2005). However, recent analysis of research of response to climate change of living marine resources is still rather poor and faces significant modeling and observational challenges (Tommasi *et al.*, 2017).

Climate change and freshwater ecosystems

Freshwater ecosystems demonstrate phenological response for climate change similar to terrestrial ecosystems. So, phytoplankton bloom, moved gradually by 19 days earlier in the period 1962–2002 in a lake in North-Western USA, while zooplankton development showed more erratic behavior (Winder and Schindler, 2004). Fresh study of rivers across a continent of Australia highlighted that climate change case a significant turnover and extinction of riverine fauna species additionally to a main threat coming from anthropogenic land use and water use (James *et al.*, 2017).

Ecological changes per se

Climate change and biogeography

The coherent signals of climate change on broad scale plant and animal communities are seen now days almost everywhere. One can expect poleward and upward shifts in warming regions and opposite direction development in cooling regions with restructuring of communities. Such studies at communities level are however rather rare. The theoretical models as a generalization of communities use biome concept or plant functional type concept related to it (Sitch *et al.*, 2003). However, observations confirming broad scale redistribution of biomes are rather rare. Here, observations of tree line change play major supporting tool for climate driven biogeography impacts. So, northward migration of tree line in warm decades of 1990s are seen in Scandinavia (Kullman, 2001) eastern Canada (Lescop-Sinclair and Payette, 1995) and in Europe (Meshinev *et al.*, 2000). It is found in recent study, however, that Alpine vegetation community has little control of multilevel hierarchical organization of species composition, but rather controlled by climate variables, which are different at different spatial scales (Malanson *et al.*, 2017). This means that this "biome" is either supporting "neutral" theory of biogeography (stochastic composition of species) or system in disequilibrium.

Climate change and spatial distribution of species

There are numerous confirmations to climate change related species ranges. Shifts in species abundance ranges are seen recently at polar latitudes, in Northern Hemisphere temperate species, in tropical species, mountain species and marine species (see review of Parmesan, 2006). It should be noted, however, that whole species range/climate change studies are rather rare. Such studies focus at the moment to plants, mammals and amphibian. The difficulty in research of climate change and spatial distribution of species are arising from patchiness of habitats with different size populations of a species. It was found recently from experimental manipulations and modeling that increases in climatic variance may as decrease, so increase a certain population growth making predictions rather uncertain (Lawson *et al.*, 2015).

Climate change and interaction between species

Species involved in trophic networks have different climate tolerances, variety of phenological responses to climate change. This may influence their interactions in prey–predator relationships. Climate change related asynchronies in timing of appearance/senescing of species of related trophic interactions may bring some species to the verge of extinctions or benefit them. Studies focusing on impact of climate change to interactions of species are, however, still rare and related mainly to plants–insects or plants–birds interactions. So, range contractions or expansions of common butterfly *Enphydra editha* were documented in relation with the time of blooming or senescing of nectar flowers (Thomas *et al.*, 1996). Flower life events in its turn were shown to be related to droughts or low snow pack years (Thomas *et al.*, 1996). Species–species tolerances determined by climate suitability may serve as a predictor tool for management and eradication of future invasive species as it was demonstrated on example of Australian Alps region recently (Harris *et al.*, 2017).

Methods of Climate Change Ecology

Climate change ecology by definition should focus on living and nonliving objects at different spatial scales, which function at different time steps. This requires combination of all available in ecology methods from traditional (as observations) to most recent (like large scale terrestrial ecosystem modeling).

Observations

Field observations

Analysis of climate change impact upon ecosystems requires systematic laborious monitoring. It is now widely accepted that so named research observatories networks are suitable for these purposes. Typically, such research networks are focusing on some rather large scientific questions (like status of carbon in an ecosystem, or status of biodiversity in an area). Most famous like Carbo-North, LTER and FLUXNET are building international networks based on predefined principles. Such principle can be, for

example, using watersheds of different level if we concentrating on soil moisture alterations due to climate change (Bogena *et al.*, 2006). Important global conclusions were found recently at FLUXNET observations: it was found that enhanced aerosol loading increase water use efficiency for all plant types due to diffuse radiation (Lu *et al.*, 2017).

Set experiments across climate gradients: FACE

Rising level of CO_2 in the atmosphere, which comes alone with climate change, may heavily influence plant physiology and accelerate or decelerate further global warming. Free-air CO_2 experiment (FACE) allows study of carbon dioxide enrichment on plant and ecosystem growth in natural conditions. The experiment is going already more than 20 years in different ecosystems types around the world. It was confirmed in the course of the experiment that above-ground production of plants is increasing in conditions of elevated CO_2, thus, making negative feedback to climate change. The effect, however, depends on functional type of ecosystem. Trees are reacting to carbon dioxide increase more than herbaceous vegetation (Ainsworth and Long, 2005). Now days FACE experiment also looks at the intercanopy structure and position of leaf for a control of possible response to elevated CO_2 (see example of Duke Forest FACE experiment (Paschalis *et al.*, 2017)).

Remote Sensing

Satellite remote sensing measuring entering and leaving flux of radiation from top of the atmosphere is one of the most powerful tools in climate change ecology. NASA at the moment uses 14 satellites of different types of orbits (sun-synchronous and geo-synchronous) with different type of sensors (passive sensors, which record naturally occurring electromagnetic radiation at top of the atmosphere and active sensors which emit electromagnetic radiation toward the Earth and measure its scattering/reflection (http:/earthobservatory.nasa.gov). Satellite remote sensing allowed to study large scale areas and make findings which cannot be done by modeling or field observations. An example is consequence of global sea-rise, which cannot be detected by the models at the moment (Yang *et al.*, 2013).

Geo-Spatial and Statistical Analysis

Statistical analysis of time series plays key role in quantitative climate change ecology. So, two thirds of all studies related to climate change impacts to marine ecosystems are studying statistical relationships between climate drivers and quantitative indicators of the ecosystems (Brown *et al.*, 2011). Visualization using geo-information systems is another important tool in climate change studies. Analysis of 30 visualizations for climate change impacts in Australia with stakeholders reveals high effectiveness of these tools for understanding of climate change.

Process-Based Modeling

Landscape scale modeling

Recent development of computer technics and systematic field observations made it possible to describe certain features of functioning of terrestrial and marine ecosystems at landscape scale as a set of related processes described by physical (e.g., mass conservation), chemical (e.g., kinetic) and biological equations (e.g., Lotka–Volterra). Outcome of response to climate change can be simulated using system of these equations (Allen *et al.*, 2010). Examples of such models are forest gap models which explicitly simulate process of tree growth for focused species, competition for light, water and space with favorite outcome for some trees and dying of other trees and opening the gap in the canopy. Processes of growth, competition and mortality are all climate specific. Thus, impact of climate change in form of composition of forest stand, above ground and underground biomass can be directly estimated by these simulation models. Recent experimental studies strongly support approach of forest gap models demonstrating that local interactions within stand and edaphic factors play almost no role in long-term vegetation dynamics (Murphy and McCarthy, 2017).

Continental scale and global modeling

The most important goal of climate change ecology is an understanding of functioning of Earth as a separate ecosystem, where interaction of biosphere, atmosphere, hydrosphere and oceans is considered. This requires using of modeling living objects of continental or global scale. It should be noted that continental/global scale objects of climate change ecology are potential vegetation types, while global change ecology considers actual vegetation cover and focuses on global change as synergy between climate change and land use/land cover change. An example of modeling object of climate change ecology at continental scale is a grid cell of dynamic global vegetation models (Sitch *et al.*, 2003), which is seen in these models as a separate ecosystem situated in terrestrial rectangle size of 1 to 1 km or 50 by 50 km. Grid cell is described as a set of percentage cover of certain functional types of vegetation (like boreal evergreen or grassland or tropical evergreen) with related carbon and water pools and fluxes. Listed grid cell ecosystem indicators are all climate sensitive. Thus, global or continental scale vegetation distribution with associated pools and fluxes can be simulated as an outcome of global or continental climate change scenarios. Recent studies of continental scale potential vegetation are concentrating on boosting of their performance (Khvostikov *et al.*, 2015).

Conclusion: Current State and Possible Future Trends in Climate Change Ecology

Climate change ecology appeared to become research field well established in recent decades. Future trends in climate change ecology will emerge as more close fusion of field and remote sensing observations as well as statistical and process-based models at fine spatial and temporal resolution. This will be secured by use of recently developed supercomputers and micro computational technologies. In the nearest future climate change ecology is going to focus on interecological disciplines like combining approaches of marine science and plant science, or climate envelop approach with invasive species theory and so on.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (31570475).

See also: Ecological Data Analysis and Modelling: Carbon Biogeochemical Cycle and Consequences of Climate Changes. Global Change Ecology: Climate Change 2: Long-Term Dynamics; Global Carbon Cycle 1: Short-Term Dynamics. Human Ecology and Sustainability: The Sustainable Development Goals

References

Ainsworth, E.A., Long, S.P., 2005. What have we learned from 15 years of free-air CO2 enrichment (FACE)? A meta-analytic review of the responses of photosynthesis, canopy properties and plant production to rising CO2. New Phytologist 165 (2), 351–372. doi:10.1111/j.1469-8137.2004.01224.x.

Allen, C.D., *et al.*, 2010. A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. Forest Ecology and Management 259 (4), 660–684. doi:10.1016/j.foreco.2009.09.001.

Berg, A., *et al.*, 2016. Land-atmosphere feedbacks amplify aridity increase over land under global warming. Nature Climate Change 6 (9), 869–874. doi:10.1038/nclimate3029.

Bogena, H., Schulz, K., Vereecken, H., 2006. Towards a network of observatories in terrestrial environmental research. Advances in Geosciences 9, 109–114.

Brown, C.J., *et al.*, 2011. Quantitative approaches in climate change ecology. Global Change Biology 17 (12), 3697–3713. doi:10.1111/j.1365-2486.2011.02531.x.

Cox, P.M., Betts, R.A., Jones, C.D., Spall, S.A., Totterdell, I.J., 2000. Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. Nature 408 (6809), 184–187. doi:10.1038/35041539.

Forchhammer, M.C., Post, E., 2004. Using large-scale climate indices in climate change ecology studies. Population Ecology 46 (1), 1–12.

Gaston, A.J., Gilchrist, H.G., Hipfner, J.M., 2005. Climate change, ice conditions and reproduction in an Arctic nesting marine bird: Brunnich's guillemot (Uria lomvia L.). Journal of Animal Ecology 74 (5), 832–841. doi:10.1111/j.1365-2656.2005.00982.x.

Gruber, N., Galloway, J.N., 2008. An Earth-system perspective of the global nitrogen cycle. Nature 451 (7176), 293–296. doi:10.1038/nature06592.

Harris, R.M.B., Kriticos, D.J., Remenyi, T., Bindoff, N., 2017. Unusual suspects in the usual places: A phylo-climatic framework to identify potential future invasive species. Biological Invasions 19 (2), 577–596. doi:10.1007/s10530-016-1334-8.

James, C.S., Reside, A.E., VanDerWal, J., Pearson, R.G., Burrows, D., Capon, S.J., Harwood, T.D., Hodgson, L., Waltham, N.J., 2017. Sink or swim? Potential for high faunal turnover in Australian rivers under climate change. Journal of Biogeography 44 (3), 489–501. doi:10.1111/jbi.12926.

Kaye, J.P., Quemada, M., 2017. Using cover crops to mitigate and adapt to climate change. A review. Agronomy for Sustainable Development 37 (1), doi:10.1007/s13593-016-0410-x.

Khvostikov, S., Venevsky, S., Bartalev, S., 2015. Regional adaptation of a dynamic global vegetation model using a remote sensing data derived land cover map of Russia. Environmental Research Letters 10 (12), doi:10.1088/1748-9326/10/12/125007.

Kullman, L., 2001. 20th century climate warming and tree-limit rise in the Southern Scandes of Sweden. Ambio 30 (2), 72–80.

Lawson, C.R., Vindenes, Y., Bailey, L., van de Pol, M., 2015. Environmental variation and population responses to global change. Ecology Letters 18 (7), 724–736. doi:10.1111/ele.12437.

Lescop-Sinclair, K., Payette, S., 1995. Recent advance of the arctic treeline along the eastern coast of Hudson Bay. Journal of Ecology 83 (6), 929–936.

Lobell, D.B., Burke, M.B., Tebaldi, C., Mastrandrea, M.D., Falcon, W.P., Naylor, R.L., 2008. Prioritizing climate change adaptation needs for food security in 2030. Science 319 (5863), 607–610. doi:10.1126/science.1152339.

Lu, X., Chen, M., Liu, Y., Miralles, D.G., Wang, F., 2017. Enhanced water use efficiency in global terrestrial ecosystems under increasing aerosol loadings. Agricultural and Forest Meteorology 237–238, 39–49. doi:10.1016/j.agrformet.2017.02.002.

Malanson, G.P., Zimmerman, D.L., Kinney, M., Fagre, D.B., 2017. Relations of alpine plant communities across environmental gradients: Multilevel versus multiscale analyses. Annals of the American Association of Geographers 107 (1), 41–53. doi:10.1080/24694452.2016.1218267.

Menzel, A., 2000. Trends in phenological phases in Europe between 1951 and 1996. International Journal of Biometeorology 44 (2), 76–81.

Meshinev, T., Apostolova, I., Koleva, E., 2000. Influence of warming on timberline rising: A case study on Pinus peuce Griseb. In Bulgaria. Phytocoenologia 30 (3–4), 431–438.

Murphy, S.J., McCarthy, B.C., 2017. Neighborhood interactions and local edaphic gradients have a weak influence on long-term vegetation dynamics in an old-growth forest community. Forest Ecology and Management 389, 314–322. doi:10.1016/j.foreco.2016.12.032.

Myneni, R.B., Keeling, C.D., Tucker, C.J., Asrar, G., Nemani, R.R., 1997. Increased plant growth in the northern high latitudes from 1981 to 1991. Nature 386 (6626), 698–702.

Oki, T., Kanae, S., 2006. Global hydrological cycles and world water resources. Science 313 (5790), 1068–1072. doi:10.1126/science.1128845.

Parmentier, F.J.W., Christensen, T.R., Rysgaard, S., Bendtsen, J., Glud, R.N., Else, B., van Huissteden, J., Sachs, T., Vonk, J.E., Sejr, M.K., 2017. A synthesis of the arctic terrestrial and marine carbon cycles under pressure from a dwindling cryosphere. Ambio 46, 53–69. doi:10.1007/s13280-016-0872-8.

Parmesan, C. (2006), Ecological and evolutionary responses to recent climate change, in Annual Review of Ecology, Evolution, and Systematics, edited, pp. 637–669, https://doi.org/10.1146/annurev.ecolsys.37.091305.110100.

Paschalis, A., Katul, G.G., Fatichi, S., Palmroth, S., Way, D., 2017. On the variability of the ecosystem response to elevated atmospheric CO2 across spatial and temporal scales at the Duke Forest FACE experiment. Agricultural and Forest Meteorology 232, 367–383. doi:10.1016/j.agrformet.2016.09.003.

van de Pol, M., Bailey, L.D., McLean, N., Rijsdijk, L., Lawson, C.R., Brouwer, L., 2016. Identifying the best climatic predictors in ecology and evolution. Methods in Ecology and Evolution 7 (10), 1246–1257. doi:10.1111/2041-210X.12590.

Rubel, F., Kottek, M., 2010. Observed and projected climate shifts 1901–2100 depicted by world maps of the Koppen-Geiger climate classification. Meteorologische Zeitschrift 19 (2), 135–141.

Shepherd, T.G., 2014. Atmospheric circulation as a source of uncertainty in climate change projections. Nature Geoscience 7 (10), 703–708. doi:10.1038/NGEO2253.

Sitch, S., *et al*., 2003. Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. Global Change Biology 9 (2), 161–185. doi:10.1046/j.1365-2486.2003.00569.x.

Stott, P.A., Tett, S.F.B., Jones, G.S., Allen, M.R., Mitchell, J.F.B., Jenkins, G.J., 2000. External control of 20th century temperature by natural and anthropogenic forcings. Science 290 (5499), 2133–2137. doi:10.1126/science.290.5499.2133.

Thomas, C.D., Singer, M.C., Boughton, D.A., 1996. Catastrophic extinction of population sources in a butterfly metapopulation. American Naturalist 148 (6), 957–975. doi:10.1086/285966.

Tommasi, D., *et al*., 2017. Managing living marine resources in a dynamic environment: The role of seasonal to decadal climate forecasts. Progress in Oceanography 152, 15–49. doi:10.1016/j.pocean.2016.12.011.

Wang, H., *et al*., 2017. Molecular mechanisms of water table lowering and nitrogen deposition in affecting greenhouse gas emissions from a Tibetan alpine wetland. Global Change Biology 23 (2), 815–829. doi:10.1111/gcb.13467.

Winder, M., Schindler, D.E., 2004. Climate change uncouples trophic interactions in an aquatic ecosystem. Ecology 85 (8), 2100–2106.

Wladimir, K., Alfred, W., 2005. The climates of the geological past. Stuttgart: Borntraeger.

Yang, J., Gong, P., Fu, R., Zhang, M., Chen, J., Liang, S., Xu, B., Shi, J., Dickinson, R., 2013. The role of satellite remote sensing in climate change studies. Nature Climate Change 3 (10), 875–883. doi:10.1038/nclimate1908.

Energy Balance☆

Axel Kleidon, Max Planck Institute for Biogeochemistry, Jena, Germany

Nomenclature

a_p	Planetary albedo
E	Evaporation rate (mm d^{-1})
H	Sensible heat flux (W m^{-2})
J	Ground heat flux (land) or horizontal heat transport (ocean) (W m^{-2})
L	Latent heat of vaporization ($\approx 2.5 \times 10^6$ J kg^{-1})
L_0	Solar luminosity (3.9×10^{26} W)
LE	Latent heat flux, composed of the latent heat of vaporization L and the evaporation rate (E) (W m^{-2})
R_0	Solar constant (1367 W m^{-2})
$R_{l,d}$	Downwelling flux of terrestrial radiation at the surface (greenhouse effect) (W m^{-2})
$R_{s,d}$	Downwelling flux of solar radiation at the surface (W m^{-2})
$R_{s,toa}$	Incoming flux of solar radiation at the top of the atmosphere (W m^{-2})
S	Entropy of a system (J K^{-1})
T_r	Radiative temperature (K)
T_s	Surface temperature (K)
λ	Wavelength of radiation (m)
σ	Stefan-Boltzmann constant (5.67×10^{-8} W m^{-2} K^{-4})
$\sigma_{process}$	Entropy production of a process (W m^{-2} K^{-1})

Glossary

Albedo The albedo describes the reflectivity of the atmosphere or surface. A high value of the albedo is associated with a highly reflecting object, such as snow or clouds.

Entropy production Entropy production refers to the general characteristic of Earth system processes to increasingly disperse energy at the scale of atoms, electrons, and molecules. It reflects that a system is far from thermodynamic equilibrium with its dynamics proceeding in the direction imposed by the second law of thermodynamics.

Feedback A feedback refers to a chain of functional relationships that modulate the response of a system to an externally caused change. A positive feedback refers to a chain of relationships which enhances the response of a system to the initial change, while a negative feedback stabilizes and dampens the response. Important feedbacks for the energy balance are the ice-albedo feedback and the water-vapor feedback.

Gaia hypothesis The Gaia hypothesis states that the state of the Earth system is regulated for and by the biosphere to maintain habitable conditions. It originated from the notion that organisms as well as the whole Earth system are complex thermodynamic systems that are maintained far from equilibrium.

Greenhouse effect The greenhouse effect refers to the absorption and subsequent re-emission of radiation within the atmosphere, which enhances the downward flux of radiation to the surface. The greenhouse effect is mostly caused by clouds, water vapor, and other greenhouse gases such as carbon dioxide, methane, and nitrous oxide.

Radiative temperature The radiative temperature is derived from a radiative flux using the Stefan-Boltzmann law. It thus refers to the equivalent temperature at which emission takes place to cause a certain flux of radiation.

Solar radiation The radiation emitted by the Sun is referred to as solar radiation. Its wavelength composition has a peak at visible light. It is also referred to as shortwave radiation.

Terrestrial radiation Terrestrial radiation, or longwave radiation, refers to radiation that is emitted by the Earth's surface or by the atmosphere. It is emitted at much lower temperatures of the Earth (when compared to the Sun) so that its wavelength composition is shifted to longer wavelengths with a peak in the infrared range.

Turbulent fluxes Turbulent fluxes refer to heat and mass fluxes associated with turbulent motion near the surface. These fluxes consist of the sensible heat flux and the latent heat flux which is associated with evaporation. These fluxes typically cool the surface and represent major components of the surface energy balance.

Introduction

All ecosystems are affected by and interact with their physical environment. At the global scale, the Earth's environment is characterized by the global energy balance, the balance of all heating and cooling terms that shape the climatological variations in

☆*Change History*: March 2018. Axel Kleidon updated all sections. Figures 1, 4, and 5 were updated with more recent data. The section on the global entropy balance was moved to a later section and was revised. The text was shortened to reduce overall length and also feedback diagram in Fig. 7(D) was updated. This is an update of A. Kleidon, Energy Balance, In Reference Module in Earth Systems and Environmental Sciences, Elsevier, 2013.

 doi:10.1016/B978-0-12-409548-9.00927-1

space and time, especially with respect to surface temperature, precipitation, and the availability of light. From an energy balance viewpoint, the interrelationships between ecosystems and their environment are threefold: (1) ecosystems utilize energy sources from their environment, and thereby are a part—though small—of the energy balance; (2) ecosystem processes are affected by environmental conditions that are directly or indirectly connected to the energy balance (e.g., precipitation affects the levels of water limitation of terrestrial productivity); and (3) the form and functioning of ecosystems affect energy balance terms. This article reviews the basics of the global energy balance, how it is reflected in the geographic distribution of mean climatological properties, and how it interacts with life through ecosystem functioning.

Global Energy Balance

At the planetary scale, the energy balance is driven by the absorption of sunlight and the emission of radiation to space. Planetary properties and the global energy balance give a first impression of the relevant processes that shape the environmental conditions at the surface and how habitable these are to life.

Planetary Energy Balance

The planetary energy balance is driven by the absorption of solar radiation of about 240 W m^{-2} in the global mean, which is then re-emitted into space at longer wavelengths as terrestrial radiation. The planetary energy balance is approximately at a steady state when the amount of absorbed radiation is balanced by the emission of radiation. In this case, the planetary energy balance is:

$$R_{s,\,toa}\left(1-a_p\right)=\sigma T_r^4$$

where the amount of absorbed solar radiation is expressed by the mean incident solar radiation at the Earth's orbit $R_{s,\,toa}$ = 340 W m^{-2} and the Earth's planetary albedo (or reflectivity) a_p = 0.29, which combined yield about 240 W m^{-2}. Emitted radiation is expressed by the Stefan–Boltzmann radiation law as σT_r^4, with σ = 5.67 10^{-8} W m^{-2} K^{-4} and T_r being the so-called radiative temperature. These numbers yield a value of T_r = 255 K for present-day Earth. The cooling of the Earth's interior adds <0.1 W m^{-2}, which is very small in comparison to the amount of absorbed solar radiation, and can therefore be neglected in the Earth's energy balance.

The observed global mean surface temperature of T_s = 288 K is notably higher by 33 K than the radiative temperature. This additional warming of the surface is due to the atmospheric greenhouse effect. It results from the absorption of long-wave radiation by greenhouse gases in the atmosphere that was emitted from the surface (Fig. 1). The absorbed radiation is re-emitted to space, but also back to the surface, thereby providing an additional radiative flux incident at the surface. The comparison of Earth's planetary characteristics to those of the planetary neighbors shows the importance of a well-balanced greenhouse effect in providing a habitable environment, with the hot surface temperatures of Venus explained by a very strong greenhouse effect while the greenhouse effect is absent on Mars (Table 1).

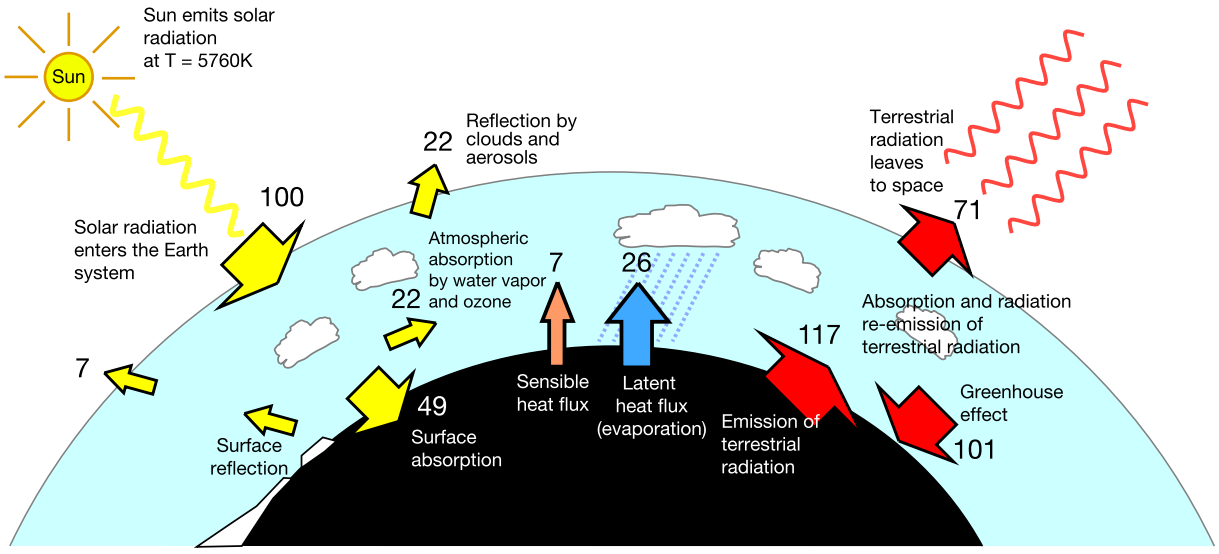


Fig. 1 Earth's global energy balance in terms of its dominant energy fluxes, their brief description, and their magnitudes, expressed as percentage of the average amount of incoming solar radiation of 340 W m^{-2}. The numbers are derived from the climatological mean of the years 2000–10 based on Stephens et al. (2012).

Table 1 The Earth in comparison to its planetary neighbors

	Earth	*Venus*	*Mars*	*Moon*
Orbital characteristics				
Distance to Sun	150×10^6 km	108×10^6 km	228×10^6 km	[b]
Obliquity	23.45°	<3°[a]	25.2°	6.7°
Eccentricity	0.017	0.007	0.094	0.055
Length of day	24 h	2802 h	24.7 h	708.7 h
Length of year	365.2 days	224.7 days	687 days	27.3 days
Atmospheric composition				
Surface pressure	101.3 kPa	9.2 MPa	640 Pa	3×10^{-10} Pa
Carbon dioxide (CO_2)	360 ppm	96.5%	95%	
Nitrogen (N_2)	78%	3.5%	2.7%	
Oxygen (O_2)	21%	0%	0.13%	
Climate parameters				
Planetary albedo	0.30	0.71	0.16	0.11
Absorbed solar radiation	239 W m^{-2}	190 W m^{-2}	124 W m^{-2}	304 W m^{-2}
Radiative temperature	255 K	233 K	210 K	275 K
Surface temperature	288 K	737 K	210 K	100–400 K
Greenhouse effect	+33 K	+504 K	+0 K	

[a]Venus rotates in the opposite sense than Earth.

[b]Distance Moon–Earth is 0.378×10^6 km.

Orbital characteristics, atmospheric composition, albedo, absorbed radiation, radiative temperature, surface temperature, and the strength of the atmospheric greenhouse effect are given for selected inner planets and the Earth's moon.

Surface Energy Balance

In addition to the absorption of solar radiation and the emission of terrestrial radiation, the surface energy balance needs to account for additional terms that heat and cool the surface:

$$R_{s,d}\,(1 - a_s) + R_{l,d} = R_{l,u} + H + LE + J$$

where the terms on the left hand side express the two radiative fluxes that heat the surface: the downwelling flux of solar radiation, $R_{s,d}$, modulated by the surface albedo a_s that describes its reflectivity, so that the product $R_{s,d}\,(1 - a_s)$ describes the absorption of solar radiation at the surface, and the downwelling flux of longwave radiation, $R_{l,d}$, that is emitted by the atmosphere (the greenhouse effect). The right hand side contains terms that generally cool the surface: The emission of radiation from the surface, $R_{l,u}$, which is directly linked to surface temperature by the Stefan-Boltzmann radiation law (as mentioned above), but also the turbulent fluxes of sensible and latent heat, H and LE, where the latent heat flux is written as the product of the latent heat of vaporization, $L \approx 2.5 \times 10^6$ J kg^{-1} K^{-1} and E is the evaporation rate. In addition, J is a term that either describes heat storage changes in the soil or the surface ocean and/or the horizontal transport of heat (relevant for oceans, but not for land).

Note that photosynthesis utilizes a fraction of the absorbed solar radiation and thus, in principle, also represents a term of the surface energy balance. However, photosynthesis utilizes <3% of the absorbed solar radiation, so that it represents a minor, often neglected term in the surface energy balance.

Global Energy Balance Components

In the global climatological mean, heat storage and transport terms in the surface energy balance average out to zero (i.e., $J \approx 0$). The estimates of each of the energy balance terms are shown in **Fig. 1**. Of the incoming 340 W m^{-2} of solar radiation at the top of the atmosphere, $R_{s,toa}$, 22% is reflected by clouds and aerosols in the atmosphere and another 7% by the surface. These two numbers add up to the planetary albedo of about $a_p = 0.29$. The remaining radiation is absorbed in the atmosphere (22% of $R_{s,toa}$, by ozone in the stratosphere and by water vapor) and at the surface (49% of $R_{s,toa}$). Additional surface heating is provided by the atmospheric greenhouse effect (*red arrow* in **Fig. 1**), which adds twice as much energy to the surface than solar radiation. These heating terms are balanced by cooling through emission of terrestrial radiation and the turbulent fluxes of sensible and latent heat. The atmosphere is heated by the absorption of solar radiation (22% of $R_{s,toa}$), absorption of terrestrial radiation emitted by the surface (117% of $R_{s,toa}$), turbulent fluxes (33% of $R_{s,toa}$), and cooled by the emission of terrestrial radiation to space (71% of $R_{s,toa}$) and to the surface (101% of $R_{s,toa}$, the greenhouse effect).

Radiation

Absorption, reflection, and emission of radiation are the dominant processes that shape the global energy balance and its regional variations. To understand these variations, the nature of radiation, the processes that reflect and absorb it, as well as the resulting latitudinal variation of radiative fluxes for the present-day climate are explained in the following.

Electromagnetic Radiation

Electromagnetic radiation is characterized by its wavelength λ, or alternatively by its frequency v. The two variables are related by $\lambda v = c$, with c being the speed of light ($c = 3 \times 10^8$ m s^{-1} in vacuum). Radiation with shorter wavelengths is referred to as more energetic. Relevant for climate are mainly the following wavelength ranges: (1) ultraviolet radiation, corresponding to wavelengths of <400 nm, is highly energetic and harmful for life. It plays a central role in the production and destruction of ozone in the stratosphere; (2) visible light, ranging from 400 nm (blue light) to 750 nm (red light), which is the energy source for photosynthesis; and (3) infrared radiation, referring to wavelengths longer than 750 nm that are associated with absorption and emission processes within the atmosphere that result in the atmospheric greenhouse effect.

Solar radiation is composed of a range of wavelengths centered around 550 nm (green light), while the Earth with its much lower emission temperature emits radiation around 11 μm (infrared). The peak in the wavelength composition of radiation is described by Wien's law in terms of the radiative temperature ($\lambda_{peak} = 0.2898 \times 10^{-3}$ m K/T_r). Since these peak wavelengths and the associated distributions are well separated, electromagnetic radiation in climatology is generally classified into two types: solar (or shortwave) radiation that is emitted by the Sun, and terrestrial (or longwave) radiation associated with emission of radiation within the Earth system.

Solar Radiation

Variations in the surface energy balance, and therefore temperature, relate mostly to variations in solar radiation. The main factors that cause variability in solar radiation are:

1. *The amount of emitted radiation by the Sun (solar luminosity L_0).* The typical value of the solar luminosity is $L_0 = 3.9 \times 10^{26}$ W, corresponding to a surface emission temperature of about 5760 K. The actual value of L_0 varies, for instance, on decadal timescales through the sunspot cycle (11 years, by <1.5 W m^{-2} at the Earth's orbit), and has increased over geologic time (4.5 billion years ago, L_0 was about 70% of the present-day value).
2. *The distance d_{Earth} of the Earth to the Sun.* The flux of solar energy remains constant through any surface around the Sun, so that the flux density decreases quadratically with distance. At the mean distance of the Earth's orbit of about $d_{Earth} = 150 \times 10^6$ km, an average amount of $R_0 = L_0/4\pi\, d_{Earth}^2 = 1367$ W m^{-2} illuminates the Earth. The value of R_0 is referred to as the solar constant. Considering that the Sun illuminates the Earth's cross section of size πr_{Earth}^2 (with the radius of the Earth denoted by r_{Earth}), but the surface area of the Earth is $4\pi r_{Earth}^2$, the mean solar radiation used above is obtained by $R_{s,toa} = R_0/4 = 340$ W m^{-2}. The mean distance of the Earth varies between 147×10^6 and 152×10^6 km throughout the year, due to Earth's slightly eccentric orbit (Fig. 2). The location of the Earth's orbit that is closest (farthest) to the Sun is called the perihelion (aphelion). The perihelion

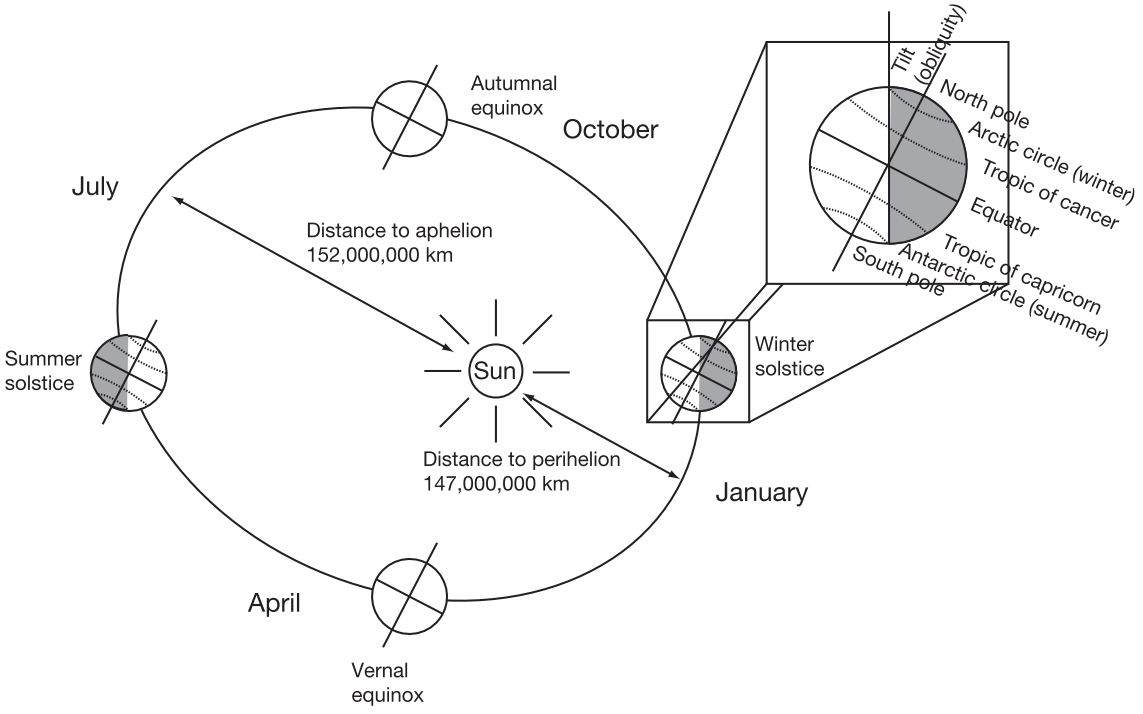


Fig. 2 The orbit of the Earth around the Sun and its relation to seasons. The orbit of the tilted Earth around the Sun results in the seasons, as indicated for the Northern Hemisphere (NH). In the NH winter, the Earth's axis of rotation is pointed away from the Sun, resulting in less incident solar radiation and the polar night at latitudes above the Arctic Circle. This situation is reversed in the summer.

currently occurs in early January. At this time, the Earth in total receives about 7% more sunlight than it does in July. These orbital parameters—perihelion, eccentricity, and tilt (or obliquity)—vary on longer timescales and relate to the timing of ice ages. The direct impacts of solar radiation and its variations are well understood, yet the indirect effects and feedbacks that amplify the Earth system response to these changes are not yet fully understood.

3. *The orientation of the surface toward the Sun*. The orientation of the surface to the incident solar radiation is characterized by two angles, the solar zenith angle and the declination angle. The incident solar radiation for a given location at the surface depends on latitude and time within the year. It is calculated from the zenith angle θ, which measures the position of the Sun to the vertical, and the declination angle δ, which characterizes the relation of the Earth's tilt to the direction of sunlight (**Fig. 3**). Integration yields a global mean solar radiation $R_{s,toa} = R_0/4$.

Reflection of Radiation

Reflection of radiation applies mostly to solar radiation and is mainly due to scattering. The size of the scattering particle plays an important role and affects the amount of radiation of a certain wavelength λ that is scattered, resulting in three types of scattering:

1. *Rayleigh scattering* applies to very small particles, such as electrons of air molecules. The intensity of scattering varies with λ^{-4}, therefore affecting primarily radiation of short wavelength. This form of scattering results in blue skies since blue light has a short wavelength and is therefore scattered much more strongly than red light.
2. *Mie scattering* involves particles of intermediate sizes, such as aerosols. The intensity of scattering varies with λ^{-1}, so that scattering is more evenly spread across wavelengths. This form of scattering results, for example, in hazy skies at a windy day at the beach due to sea spray, or over cities with air pollution due to aerosols from traffic.
3. *Geometric scattering* applies to large particles such as cloud droplets. The intensity of scattering does not vary with wavelength. This form of scattering makes clouds appear white.

The overall fraction of scattered radiation is described by the albedo (or reflectivity). Typical values of albedo of different surfaces are summarized in **Table 2**, with greater values of the albedo meaning greater reflectivity. The albedo also depends on other factors, such as the zenith angle, and wavelength. For instance, vegetated surfaces are generally much more reflective

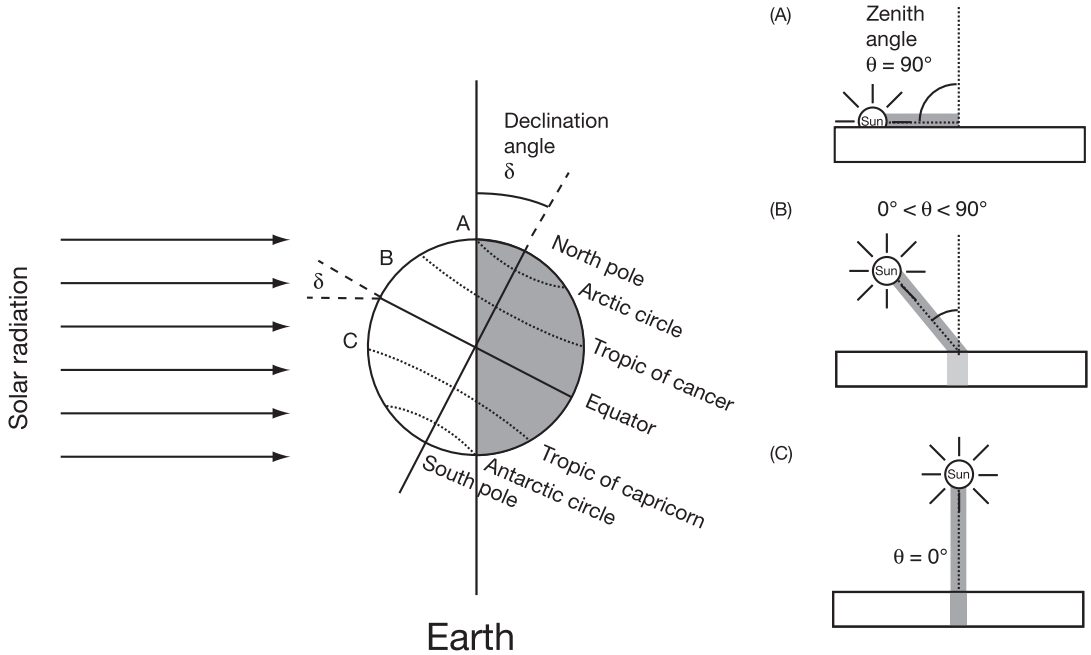


Fig. 3 Effects of the orientation of the Earth's surfaces toward the Sun on the amount of incident solar radiation at different locations labeled by A, B, and C. Left: The amount of solar radiation that reaches the surface at a given latitude depends on the declination angle δ. The declination angle measures the angle between the Earth's axis of rotation and the vertical plane of the orbit, or, alternatively, the angle between the direction of solar radiation and the Earth's equator. The declination angle defines Earth's major regions: the tropics (latitudes $-\delta$ to δ) and the polar regions ($90^\circ - \delta$ to 90°). Earth's declination angle is currently at 23.45°. Right: At a given location on Earth, the zenith angle θ measures the angle between the vertical and the Sun. It depends on hour, latitude, and time of year. In the situation shown on the left (Northern Hemisphere winter solstice), the zenith angle at location A at noon is 90°, that is, the Sun does not rise above the horizon and no solar radiation is incident at the surface. At location B, the zenith angle is in between 0° and 90°. At location C at the tropic of Capricorn, the zenith angle is 0° at noon, and the incoming solar radiation is vertical to the surface. In sloped terrain, a correction needs to be applied for the calculation of incident radiation to correct for the slope.

Table 2 Typical values and ranges for albedo for different surfaces and clouds

	Range (%)	*Typical (%)*
Atmosphere		
Cirrus clouds		21
Cumulus clouds		48
Stratus clouds		69
Ice, snow, and water		
Deep water (small zenith angle)	3–10	7
Deep water (large zenith angle)	10–100	
Sea ice	30–45	30
Snow (fresh)	70–95	80
Snow (old)	35–65	50
Snow (with forest)	11–35	25
Bare land		
Sand (wet)	20–30	25
Sand (dry)	30–45	35
Clay (wet)	10–20	15
Clay (dry)	20–40	30
Humus (moist)	5–15	10
Desert	20–45	30
Concrete	15–35	20
Asphalt	5–10	7
Vegetation		
Tundra	18–25	15
Grassland	16–26	19
Coniferous forest	5–15	12
Deciduous forest	10–20	17
Evergreen forest	12–25	13
Cropland		18

(30%–50%) in the near infrared (at wavelengths of 0.8–1.0 μm) but absorbent in the red part of the spectrum at about 0.6 μm, with a low reflectivity of around 5%. This difference in absorptive characteristics is used for the remote sensing of vegetation greenness.

Absorption of Radiation

Radiation is absorbed by different processes and at different intensities, depending on material characteristics and the wavelength of the radiation:

1. *Photoionization* refers to a process in which highly energetic radiation with wavelengths of <100 nm is absorbed by removing electrons from atoms, resulting in ionized atoms. This process can be found in the higher atmosphere at heights of 100 km and above in the so-called ionosphere.
2. *Photodissociation* is a process which also absorbs highly energetic radiation by breaking up molecular bonds. This process occurs in the atmosphere mainly for wavelengths shorter than visible light. An example is the absorption of ultraviolet radiation by molecular oxygen and ozone in the stratosphere.
3. *Electronic absorption* is associated with the absorption of visible light. Radiation is absorbed by raising electrons into excited states. While this form of absorption has little relevance in atmospheric absorption, it is essential for photosynthesis, where electronic absorption is used to separate hydrogen ions from the water molecule.
4. *Absorption by rotational and vibrational modes of molecules* takes place with radiation of low energy and long wavelengths (near infrared and longer) and causes molecules to rotate or vibrate. This form of absorption requires molecules with an uneven distribution of electrons, so that these molecules have a dipole moment. In the atmosphere, water vapor (H_2O) absorbs very well by its rotational and vibrational modes due to the architecture of the molecule, where the oxygen atom attracts the electrons more than the two hydrogen atoms. Other relevant gases that absorb by this mechanism are carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O).

Because the Earth's surface emits radiation mainly in the infrared, gases that absorb in these wavelengths are called greenhouse gases. Water vapor and clouds are by far the most important contributors to the strength of the present-day greenhouse effect. The special role of carbon dioxide as a greenhouse gas originates from two facts: (1) water vapor absorbs poorly at the peak of the Earth's surface emission at about 11 μm at which the CO_2 molecule has a dominant absorption peak nearby at 15 μm and therefore absorbs very well; and (2) the concentration of water vapor in the atmosphere is constrained by its saturation level, which in turn depends on the ambient air temperature. Hence, the concentration of water vapor reacts to other prevailing

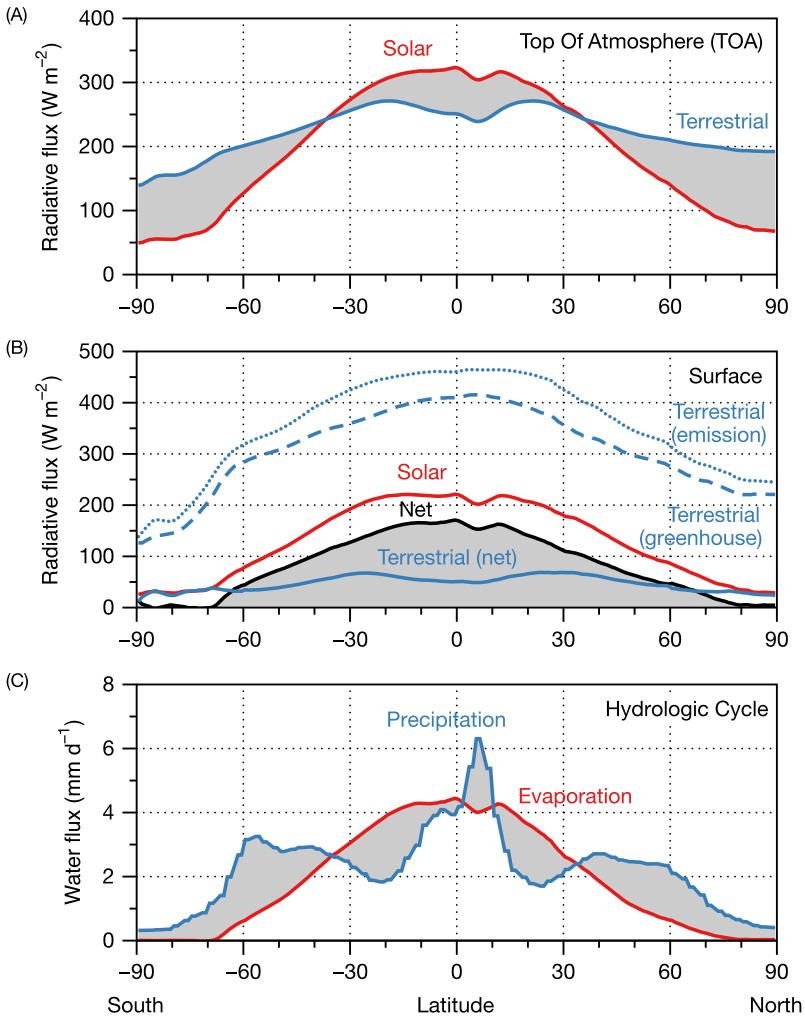


Fig. 4 Annual mean components of the energy and water balance along latitudinal zones at the top of the atmosphere and the surface for the time period 2001 to 2015. (A) The graph shows the fluxes of net solar radiation (incoming minus reflected, red line, "Solar") and terrestrial radiation (outgoing long-wave radiation, blue line, "Terrestrial") at the top of the atmosphere. The positive values of net radiation in the Tropics (i.e., more solar radiation is absorbed than terrestrial radiation emitted to space) indicates that heat is transported by the atmosphere and ocean systems toward the polar regions, where emission of terrestrial radiation exceeds absorption of solar radiation. This effect of heat transport is marked by the *gray shading*. (B) Radiative fluxes at the Earth's surface: absorbed solar radiation (*red line*, "Solar"), emission of terrestrial radiation (*blue dotted line*, "Terrestrial (emission)"), downward flux of terrestrial radiation (the greenhouse effect, *blue dashed line*, "Terrestrial (greenhouse)"), net flux of terrestrial radiation at the surface (*blue solid line*, "Terrestrial (net)") and net radiation (*black solid line*, "Net"). Net radiation is mostly balanced by the turbulent fluxes of sensible and latent heat and thus atmospheric motion (shaded *gray*), although ocean heat transport also contributes a minor part to this component. (C) The fluxes of precipitation (*blue line*) and evaporation (*red line*) that remove and add moisture to the atmosphere. Regions where evaporation exceeds precipitation are regions where the atmosphere gains moisture, which is transported by the atmospheric circulation to regions where precipitation exceeds evaporation. This effect of atmospheric transport is indicated by the *gray shading*. The plots were created using the CERES radiation datasets for the radiative fluxes, the GCPC precipitation data set for precipitation, and evaporation estimated from net radiation and temperature.

conditions and by itself does not act as an independent driver for change. For instance, cold air is unable to hold large amounts of water vapor, and consequently water vapor plays a less important role, for example, in cold regions of the atmosphere, and in winter seasons in polar regions.

Emission of Radiation

Emission of radiation is described by the Stefan-Boltzmann radiation law which relates the temperature to the rate of emission. Emission also depends on the emissivity of the emitting mass. While surfaces generally emit with emissivities near one, emission within the atmosphere typically takes place at lower emissivities, depending on the concentration of greenhouse gases and clouds, which vary with height, location and time.

Energy Balance and Climate

The mean variation in solar radiation shapes the climatological variations of the surface energy balance, atmospheric dynamics and the hydrologic cycle. These variations are described here in terms of the latitudinal variations of the annual means and is shown in Fig. 4.

Radiative Forcing

Solar radiation varies mostly due to the orientation of surfaces across latitudes, with slight variations in reflectivity due to cloudiness and ice cover, particularly at latitudes 70°S poleward. The flux of terrestrial radiation to space shows a less pronounced latitudinal variation. The imbalance at the top of the atmosphere between the net fluxes of solar and terrestrial radiation reflect the overall heat transport within the climate system. The surplus of heat from greater absorption of solar radiation than emission of terrestrial radiation in the tropical regions is transported toward the extratropics where it maintains a greater flux of terrestrial radiation compared to the absorption of solar radiation. This heat transport is accomplished by the large-scale circulation of the atmosphere and ocean.

Surface Fluxes

The components of the surface energy balance are also strongly shaped by the zonal variation of absorbed solar radiation. Fig. 4b shows the zonal variation of the radiative fluxes at the surface and the net radiation, the sum of all radiative fluxes at the surface. Net radiation is balanced by the heat transported by turbulent fluxes and by oceanic heat transport (which is comparatively small when averaged over a latitudinal band). Note how the downward flux of terrestrial radiation (the greenhouse effect) is considerably larger than the absorption of solar radiation, yet as emission of radiation is also a large flux, the net cooling of the surface by terrestrial radiation is comparatively small across latitudes. Much more relevant for the cooling of the surface are the turbulent fluxes of sensible and latent heat. As solar radiation is absorbed mostly at the surface and the atmosphere aloft cools by emitting terrestrial radiation to space, a difference in density is generated that drives vertical convection which transports sensible and latent heat from the surface to the atmosphere. The sensible heat flux directly relates to dry convection, that is, the upward transport of heat by motion generated by the heating of near-surface air from the surface, which gains a lower density and thus becomes buoyant. The latent heat flux links directly to evaporation, which acts as a moisture source for the atmosphere. It thus plays a central part for the hydrologic cycle (see also below). When the moisture condenses within the atmosphere, it creates another heating source that creates lower density and drives moist convective motion.

Large-Scale Motion

The zonal variation in solar radiation causes heating differences that drive the large-scale atmospheric circulation. Heating differences cause differences in temperature and density that result in large-scale pressure differences within the atmosphere that then generate planetary-scale motion. This motion redistributes heat, water, and mass at the hemispheric scale, and overall aims to deplete the differences in heating caused by the zonal variation in solar radiation. The ocean is maintained in motion due to the wind stress at the ocean surface, and by differences in density caused by temperature variations and the removal or addition of freshwater by evaporation and precipitation which affects the density of seawater by affecting its salinity. The effects of large-scale motion are then noticeable in the imbalance in radiative fluxes at the top of the atmosphere (as marked by the gray shading in Fig. 4a). Most of this imbalance results from the heat transported by the atmosphere, with a comparatively smaller contribution by the oceanic circulation.

Hydrologic Cycling

The latent heat flux within the surface energy balance directly relates to evaporation, which moistens the atmosphere. The drying of the atmosphere is accomplished by precipitation, which releases the latent heat into the atmosphere, causing a strong heating source and moist convection. The zonal variation in evaporation follows strongly the zonal variation of solar radiation, but precipitation occurs more concentrated in the tropics and mid-latitude regions (Fig. 4c). The difference between evaporation and precipitation indicates that moisture is transported by the atmospheric circulation (indicated by the *gray shading* in Fig. 4c). In the tropics, precipitation exceeds evaporation because the atmosphere imports moisture from the subtropics and condenses it in a relatively narrow band of the tropics within the innertropical convergence zone. The large-scale atmospheric circulation also transports moisture from the subtropics into the mid-latitudes, resulting in precipitation exceeding evaporation in those latitudes as well. Hence, hydrologic cycling directly follows from the latent heat term of the energy balance, yet with a strong imprint of the atmospheric circulation.

Feedbacks

Changes in climate affect the energy balance, and these effects are often modulated by a chain of processes resulting in feedbacks. Examples for important feedbacks that affect the energy balance are given in Fig. 5. Feedbacks characterize the response of the energy balance to a perturbation in the external forcing. They formalize the nonlinearities and interactions among the processes

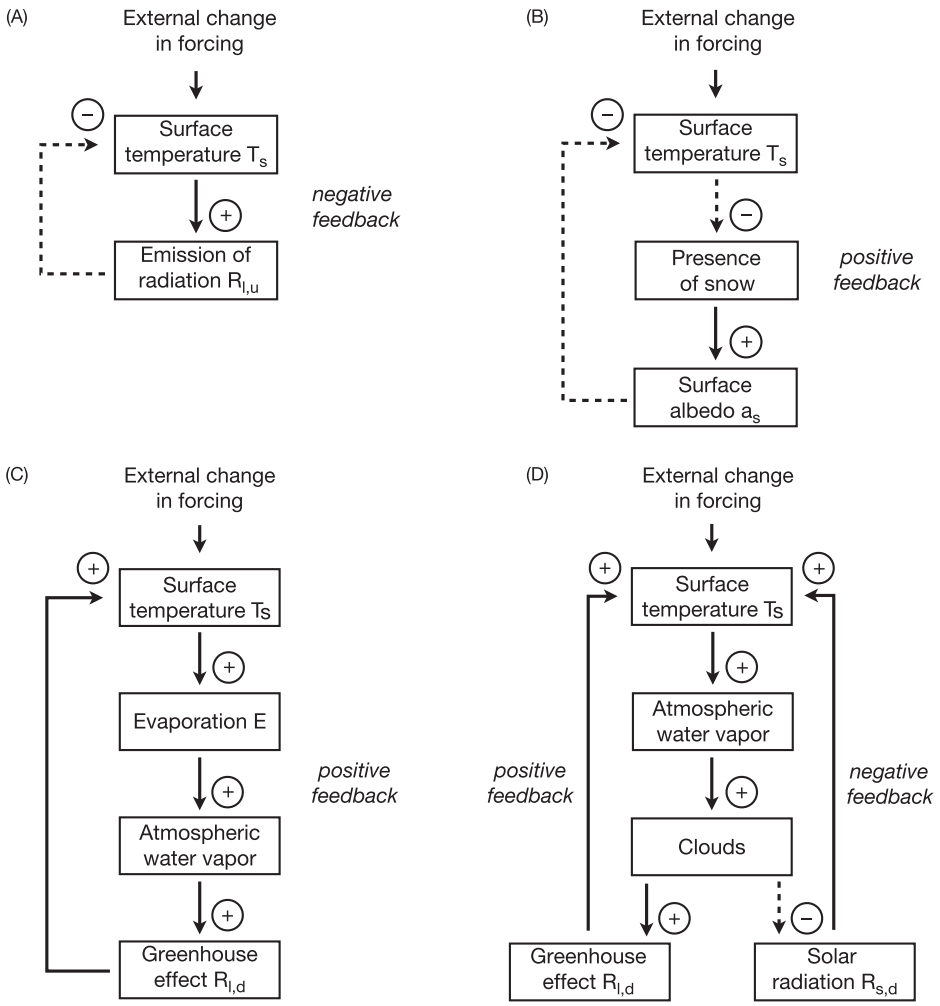


Fig. 5 Dominant feedback processes that shape the response of the global energy balance and surface temperature to external change. The diagrams show the variables involved in four important feedback loops. The (+/−) signs at the arrows indicate positive/negative influences. (A) The thermal radiation feedback. An external change in forcing that would increase surface temperature would also increase the emission of terrestrial radiation (a "+" influence). An increased emission would result in a lower surface temperature (a "−" influence). The enhanced emission of terrestrial radiation therefore counteracts the initial change, resulting in a negative feedback loop. The same line of reasoning also applies for an external change that would reduce surface temperature. (B) The snow (or ice) albedo feedback. An external change that warms the surface reduces the presence of snow, lowers the surface albedo, thereby amplifying the warming (a positive feedback). (C) The water vapor feedback. An external change that warms the surface heats the lower atmosphere. Since warmer air can hold more moisture, this enhances surface evaporation and the amount of water vapor in the atmosphere. More water vapor results in a stronger atmospheric greenhouse effect, thereby amplifying the initial change (a positive feedback). (D) Two types of cloud feedbacks. Continuing from the water vapor feedback, more water vapor in the atmosphere can result in more clouds. Depending on the balance of increased cloud cover on increased greenhouse forcing (left loop) or shortwave reflection (right loop), cloud feedbacks can form both positive and negative feedback loops on surface temperature.

within the climate system. Feedbacks are classified into positive and negative feedbacks. Positive feedbacks enhance the response of a chosen variable (mostly temperature) to the external forcing, while negative feedbacks stabilize the system, making it less responsive. In other words, a positive feedback makes a positive (negative) external change more positive (negative).

Thermodynamics and the Energy Balance

The general direction into which processes take place within the energy balance is described by thermodynamics, particularly the second law. This law states that the entropy, a measure for how well energy is dispersed at the microscopic scale of photons (carriers of radiative energy), electrons and molecules, can only increase in an isolated system. This implies that energy tends to become increasingly dispersed, a direction that is irreversible, meaning that it cannot be undone. An example for this dispersal is that heat fluxes are directed from hot to cold, thereby aiming to deplete temperature differences. The associated increase in entropy

is a general characteristic of any process and thus sets a basis to compare processes of different kinds in terms of how they follow the second law.

This direction applies to the processes involved in the energy balance and the Earth system in general. In addition, entropy sets an important constraint on how much physical work can be done, which, for example, is needed to accelerate the atmosphere and sustain the large-scale circulation against friction or to maintain hydrologic cycling.

Entropy Budget of the Earth System

By exchanging radiation with space, the Earth system is not an isolated system. The second law of thermodynamics then needs to be evaluated in the context of the entropy budget. The entropy budget relates the change of entropy within the system, dS/dt, to the entropy production of processes, $\sigma_{process}$, within the system, and to the net entropy exchange across the system boundary ($J_{s,in} - J_{s,out}$):

$$dS/dt = \sigma_{process} + J_{s,in} - J_{s,\,out}$$

The entropy S relates to the extent of thermodynamic disequilibrium within the system, with thermodynamic equilibrium achieved when S reaches its maximum value. A steady state of the entropy budget is defined by $dS/dt = 0$, and this state can be maintained in thermodynamic disequilibrium as long as the entropy production within the system is balanced by the net entropy exchange. Note that this is a different, less common definition of the climatological steady state as it relates to the disequilibrium within the Earth system.

The entropy budget of the Earth can be estimated from this steady-state condition by the difference of entropy fluxes across the Earth–space boundary. The entropy flux associated with a heat flux is described by the heat flux divided by its temperature. Planetary entropy production results from various irreversible processes and is maintained by entropy exchange associated with the radiative fluxes of solar and terrestrial radiation. Solar radiation has a very low entropy due to the high temperature at which emission takes place at the Sun's surface. In contrast, terrestrial radiation has much higher entropy due to the comparatively low emission temperature of the Earth. Among the processes that produce entropy within the Earth system, radiative processes are by far the dominant contributors. They produce entropy by scattering the direct, focused beam of solar radiation to diffuse radiation and by the absorption and re-emission of radiation at different temperatures. Entropy production by planetary motion results from the heat redistribution and mixing that depletes temperature gradients that are generated by solar radiation. Dry convection is driven by the sensible heat flux which depletes the temperature difference between the surface and the atmosphere, thereby producing entropy. Large-scale motion of the atmosphere produces entropy by transporting heat from the warm tropics toward the cold poles. Hydrologic cycling includes several processes that produce entropy, including the evaporation of water vapor into unsaturated air, the generation of motion and subsequent frictional dissipation by latent heating in moist convection, and the dissipation of kinetic energy of falling raindrops. Biotic activity uses a fraction of low entropy solar radiation, converts it into chemical energy in form of organic carbon compounds, and produces entropy by the respiration of these compounds into heat. There are various other processes that produce entropy, such as seasonal freeze–thaw associated with sea ice and snow cover, seasonal storage and release of heat, wetting and drying of soils, geochemical reactions and metabolisms.

Thermodynamic Limits

The entropy balance sets an important constraint on how dissipative the Earth system can be, and, more specifically, how much physical work can be generated. This work is needed to maintain motion, hydrologic cycling and other processes in states of thermodynamic disequilibrium.

The limit to work is set by the Carnot limit, which describes how much work can maximally be generated (at a rate G) out of a heat flux J between two heat reservoirs at temperatures T_w and T_c with $T_w > T_c$:

$$G = J\,(T_w - T_c)/T_w$$

where the temperature ratio term is referred to as the Carnot efficiency. This limit is a direct consequence of the first and second law of thermodynamics.

This limit constrains the large-scale atmospheric circulation. The work needed to drive the atmospheric circulation is derived from the heat flux from the tropics to the poles. Yet, the more heat is transported toward the poles, the lower the temperature difference between the tropics and the poles. This effect is shown in the three plots in **Fig. 6**, which illustrate three cases of heat transport and their effect on top-of-atmosphere radiative fluxes. By reducing temperature differences, heat transport thus reduces the Carnot efficiency, and thereby sets a maximum power limit at an intermediate value of heat transport and temperature differences. This limit yields about 2 W m^{-2} of power to accelerate the large-scale circulation, which is much less than the total influx of solar radiation of 340 W m^{-2}. Comparison to observations and modeling studies suggest that the atmosphere operates near this limit.

The work done balances frictional dissipation in steady state, which produces entropy. The maximum power limit therefore closely relates to the hypothesis of Maximum Entropy Production (MEP), which states that complex systems with sufficient degrees of freedom maintain states at which entropy production is at a maximum. However, this hypothesis is not widely accepted within climatology.

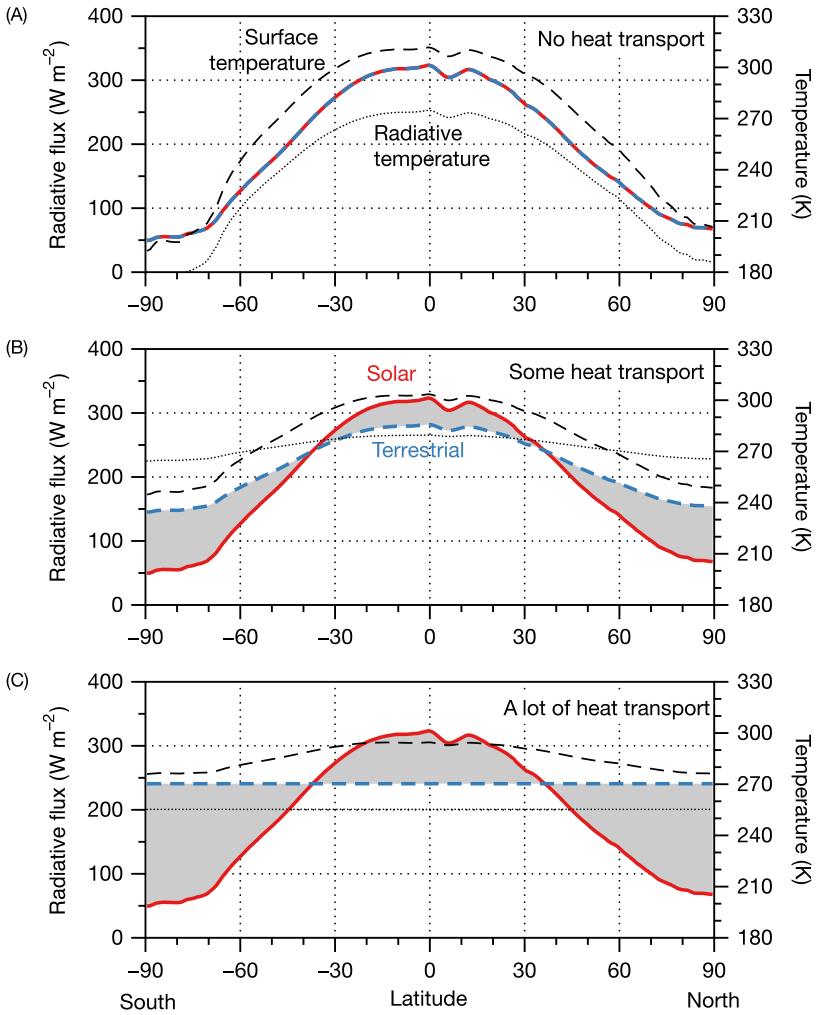


Fig. 6 The effect of poleward heat transport on radiative fluxes at the top of the atmosphere (net solar flux $R_{s,toa}$, *red lines*; net terrestrial flux, $R_{l,toa}$, *blue dashed lines*) and temperatures (*black lines*). The three panels show the cases of differing heat transport (A: no heat transport, B: some heat transport, and C: a lot of heat transport) which result in a different latitudinal distribution of terrestrial radiation to space and affect temperatures. Estimates for surface temperature, T_s, (*black dashed line*) and radiative temperature, T_r (*black dotted line*) are also shown. The reduction of the temperature difference by heat transport sets a thermodynamic limit on how much work the global atmosphere can maximally perform to maintain motion. Of the three cases shown here, the intermediate case (B) has greatest power and is closest to observations (cf. Fig. 4).

Energy Balance and Ecosystems

Ecosystems affect the energy balance directly by utilizing solar radiation by photosynthesis, and indirectly by altering the surface energy balance components, for instance by changing the surface albedo and evapotranspiration rates, and by affecting biogeochemical fluxes and atmospheric composition, which alters radiative transfer, for example, by scattering by aerosols or by changing the magnitude of the greenhouse effect.

Direct Biotic Effects

Photosynthesis utilizes $<3\%$ of the absorbed solar radiation. Since most of the carbohydrates are respired within relatively short time at the same location, most of the energy is released as heat by respiration. Hence, the energy fluxes associated with photosynthesis and respiration are generally neglected in considerations of the surface energy balance.

Indirect Biotic Effects and Feedbacks

Ecosystems interact with their physical and geochemical environment. The effect of ecosystems on the energy balance are categorized into two types of effects:

1. *Biogeophysical effects* modify components of the surface energy balance and the physical functioning of the climate system. These effects are strongest for terrestrial vegetation, which affects the energy balance over land in various ways: (a) the surface albedo of vegetated surfaces is generally darker than bare surfaces (**Table 2**), thereby enhancing absorption of solar radiation; (b) vegetation root systems enhance the ability to recycle soil moisture through transpiration, thereby affecting the latent heat flux; and (c) vegetated surfaces modulate the partitioning of sensible and latent heat through stomatal functioning. These effects have considerable effects on the surface energy and water balance on land and the overlying atmosphere, particularly on evaporation, (**Fig. 7**) and can result in two biogeophysical feedback processes that enhance productivity (**Fig. 8**).
2. *Biogeochemical effects* modify geochemical cycling of elements and the atmospheric composition (**Table 1**). Some of these have important consequences for the global energy balance and atmospheric dynamics, such as (a) carbon cycling affects the

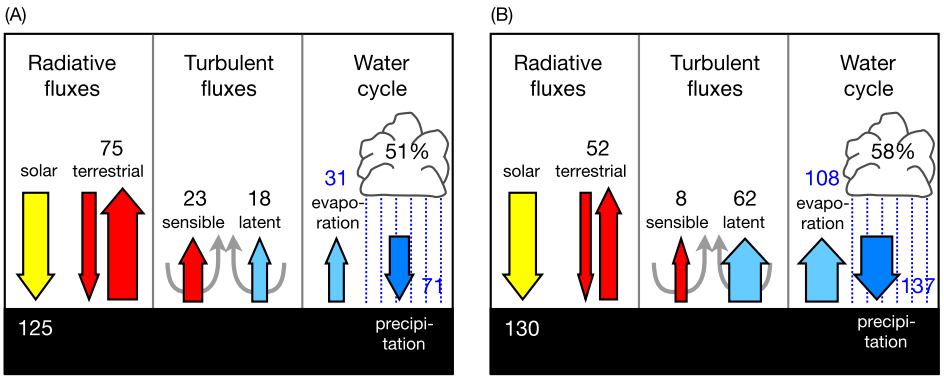


Fig. 7 Estimates of vegetation effects on the energy- and water balances on land in the climatological mean. The diagrams show results from climate model simulations of (A) a "desert world" void of terrestrial vegetation and (B) a "green planet" where all land regions acts like a rainforest. Radiative and turbulent fluxes are given in W m^{-2} and water fluxes in 1000 km^3 per year. These climatological differences result mostly from the effect of vegetation on surface albedo and the depth of the rooting zone.

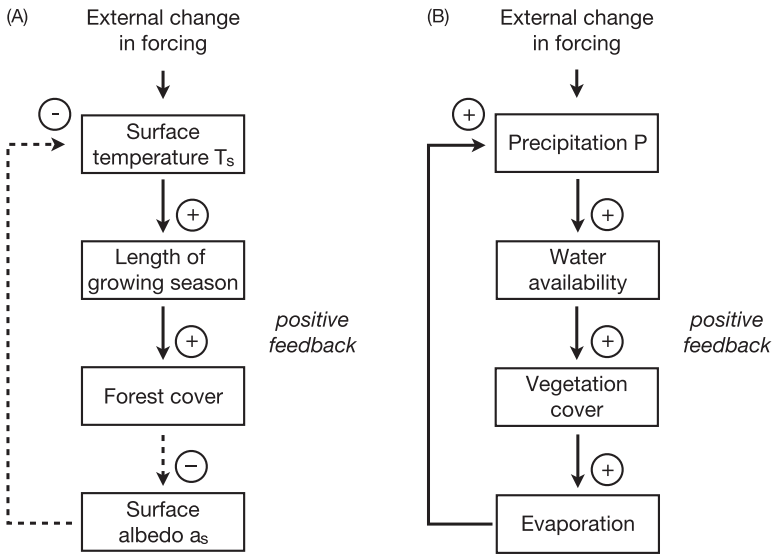


Fig. 8 Vegetation feedbacks on the surface energy balance. The diagrams show the two major feedback loops by which vegetation affects the surface energy balance. (A) The snow masking feedback. An external change in forcing that would increase surface temperature in regions where temperature limits terrestrial productivity (such as the Arctic) increases the length of the growing season. A longer growing season would result in higher productivity, which extends the boreal forest cover in temperature-limited regions. Enhanced boreal forest cover masks the presence of snow at the surface, thereby lowering the surface albedo. This results in enhanced absorption of solar radiation and a warmer temperature, which amplifies the initial change, resulting in a positive feedback loop. (B) The water cycling feedback. An external change that results in enhanced precipitation in regions where water limits productivity (such as the semiarid regions) increases the productivity. This can sustain a greater vegetation cover and thereby evaporation, which can enhance precipitation further. The initial change is hence amplified, resulting in a positive feedback.

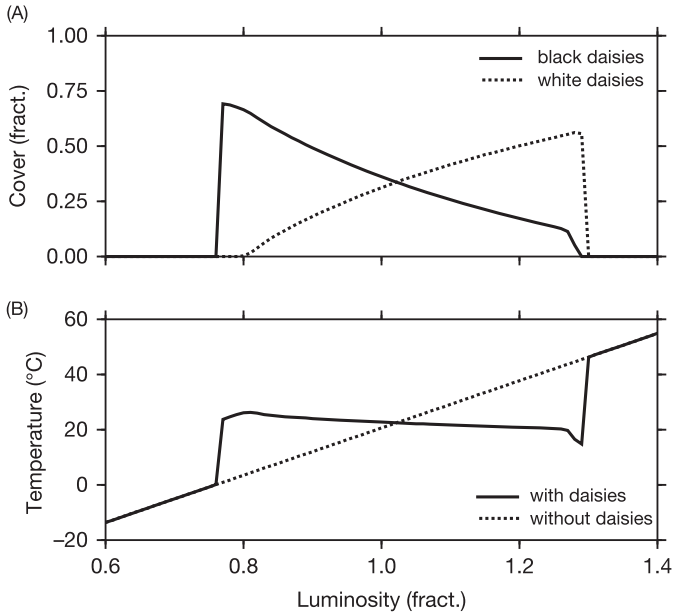


Fig. 9 Emergence of temperature regulation in the conceptual "Daisyworld" model. The "Daisyworld" model represents a virtual world in which the planetary albedo is regulated by the population dynamics of *black* and *white* daisies. (A) The fractional cover of *black* and *white* daisies for different values of solar luminosity, expressed as the fraction of its present-day value. (B) The different proportions of daisies result in an overall planetary albedo that results in constant temperature conditions over a wide range of solar luminosity values.

strength of the greenhouse effect caused by carbon dioxide, (b) oxygen concentrations affect the ozone layer and stratospheric absorption of solar radiation, and (c) the production of some compounds, such as dimethyl sulfide by marine algae or volatile organic compounds (VOC) by terrestrial vegetation, act as cloud condensation nuclei, thereby modifying cloud cover.

Gaia Hypothesis

In the extreme form, strong negative biotic feedbacks on temperature can regulate the global energy balance into a state of homeostasis (i.e., no temperature sensitivity to changes in external forcing). This has been suggested by the Gaia hypothesis of James Lovelock, stating that the atmosphere is regulated for and by the biosphere. This notion was originally motivated by the observation that the Earth's atmospheric composition is far from thermodynamic equilibrium, and that this state of dis-equilibrium is maintained by the biota.

The conceptual "Daisyworld" model was developed to demonstrate the possibility of global homeostasis. This model describes a world where the planetary albedo is determined by the fractions of *black* and *white* daisies, and of bare ground. Using equations of population dynamics and a temperature-dependent growth, "Daisyworld" demonstrates that homeostasis is a possible outcome of population dynamics interacting with the global energy balance (**Fig. 9**).

However, the notions of Gaia and Daisyworld remain controversial. Surface temperatures in Earth's history have been far from constant, and the representation of biospheric dynamics in Daisyworld is highly simplistic. Yet the challenge to find and apply general organizing principles, such as the maximum power limit of the previous section, that can explain the interactions of the biosphere with the global energy balance remains an active research topic.

See also: Global Change Ecology: Energy Flows in the Biosphere

Reference

Stephens, G.L., Li, J., Wild, M., Clayson, C.A., Loeb, N., Kato, S., L'Ecuyer, T., Stackhouse, P.W., Lebsock, M., Andrews, T., 2012. An update on Earth's energy balance in light of the latest global observations. Nature Geoscience 5, 691–696.

Further Reading

Bonan, G.B., 2016. Ecological climatology, 3rd edn Cambridge, UK: Cambridge University Press.
Hartmann, D.L., 2015. Global physical climatology. Amsterdam, The Netherlands: Elsevier.

Kleidon, A., 2012. How does the Earth system generate and maintain thermodynamic disequilibrium and what does it imply for the future of the planet? Philosophical Transactions of the Royal Society A 370, 1012–1040.
Kleidon, A., Fraedrich, K., Heimann, M., 2000. A green planet versus a desert world: Estimating the maximum effect of vegetation on land surface climate. Climatic Change 44, 471–493.
Lovelock, J.E., 1965. A physical basis for life detection experiments. Nature 207, 568–570.
Lovelock, J.E., Margulis, L., 1974. Atmospheric homeostasis by and for the biosphere: The Gaia hypothesis. Tellus 26, 2–9.
Oke, T.R., 1987. Boundary layer climates, 2nd edn New York: Halsted.
Ozawa, H., Ohmura, A., Lorenz, R.D., Pujol, T., 2003. The second law of thermodynamics and the global climate system: A review of the maximum entropy production principle. Reviews of Geophysics 41, 1018.
Peixoto, J.P., Oort, A.H., 1992. Physics of climate. New York: American Institute of Physics.
Schneider, S.H., Miller, J.R., Crist, E., Boston, P.J., 2004. Scientists Debate Gaia. The Next Century. Cambridge, USA: MIT Press.
Watson, A.J., Lovelock, J.E., 1983. Biological homeostasis of the global environment: The parable of Daisyworld. Tellus 35B, 284–289.

Relevant Websites

http://nssdc.gsfc.nasa.gov—Lunar and Planetary Science at the National Space Science Data Center (NSSDC).
https://ceres.larc.nasa.gov/index.php—Global energy balance data and additional information at the Clouds and the Earth's Radiant Energy Systems (CERES) website of NASA.
https://www.esrl.noaa.gov/psd/data/gridded/data.gpcp.html—Global precipitation data and additional information on the website of the Earth System Research Laboratory of NOAA, USA.

Energy Flows in the Biosphere

YM Svirezhev, Potsdam Institute for Climate Impact Research, Potsdam, Germany

Introduction

Life on Earth is a product of so-called 'photon's mill', which has started to function when our solar system was in the form of matter's 'clots' embedded into the ocean of 'cold' photons with temperature $T=2.7$ K. Evolution and self-organization of planets (including life on our planet) is a result of this mill's functioning, which is happening due to the fact that Sun's surface irradiates the 'hot photons' at $T=5800$ K, and these photons reach the cold planet's surfaces. Then they cool down to the temperature of the surface and irradiate back into the space. For the Earth, this temperature is equal to 253 K; it is the temperature, which could be measured by an observer at the top of atmosphere. Formally, the photon's mill is a typical 'heat machine' that is functioning by Carnot cycle, but its working body is the photon gas.

Incoming and Outgoing Radiations and the Planetary Energy Balance

A parallel flow of solar radiation at the Earth's mean distance from the Sun is equal to 1368 W m^{-2}; this is an irradiation of blackbody with $T=5800$ K (see Fig. 1). Since the Sun's flow is parallel its effective section is πr_{Erth}^2, where r_{Erth} is the Earth's radius. An area of the Earth's surface is $4\pi r_{\text{Erth}}^2$, that is, four times larger than the section's area; therefore, only one-fourth of the total flow, 342 W m^{-2}, is coming to the area unit of the upper boundary of the 'Earth + atmosphere system' (EAS). It is obvious that specific flow is varying from point to point on the globe and in time within 1-year interval, so that these and other local values connected with energy flows are averaged over the globe and the year.

Incoming radiation is described by the energy spectrum $E^{\text{in}}(\lambda, x, y, t)$, where λ is a wavelength, x and y are geographic coordinates, and t is a time. By interacting with a surface, it is transformed onto the energy spectrum of outgoing radiation, $E^{\text{out}}(\lambda, x, y, t)$:

$$E^{\text{in}}(\lambda, x, y, t) \overset{F(\lambda,x,y,t)}{\Rightarrow} E^{\text{out}}(\lambda, x, y, t) \qquad [1]$$

where $F(\lambda,x,y,t)$ is the transition operator. The spectrum of outgoing radiation is close to the blackbody spectrum with $T=253$ K.

Really, we have information only about these two spectra measured sufficiently frequently, sufficiently densely, in sufficiently big number of spectral bands (today up to 120), in the course of sufficiently long time. Outside of the EAS, satellites are carrying out these measurements today. At the level of the Earth's surface (the ground), it is performed also by satellites, and when they are corrected by data of the ground measurements. Note that outgoing radiation contains a lot of information about a surface, interacting with incoming radiation. Spectra of incoming and outgoing radiations are shown in Fig. 1.

The simplest form of $F(\lambda, x, y, t)$ is a shift operator $R(\lambda, x, y, t) = E^{\text{in}}(\lambda, x, y, t) - E^{\text{out}}(\lambda, x, y, t)$; a convolution $\overline{R}(x, y, t) = \int_\Omega R(\lambda, x, y, t)\mathrm{d}\lambda$ over all wavelengths is named a local 'radiative (radiation) balance'. Convolutions $\overline{E}^{\text{in}}(x, y, t) = \int_\Omega E^{\text{in}}(\lambda, x, y, t)\mathrm{d}\lambda$ and $\overline{E}^{\text{out}}(x, y, t) = \int_\Omega E^{\text{out}}(\lambda, x, y, t)\mathrm{d}\lambda$ are the total energy of incoming and outgoing radiation at the given point and in the given moment. If we average $\overline{R}(x, y, t)$ over the EAS surface, S_{EAS}, and 1-year interval, t_1, we get the 'annual planetary radiative balance'

$$\hat{R} = \frac{1}{S_{\text{EAS}} \cdot t_1} \int_G \int_T R(x, y, t)\mathrm{d}x\, \mathrm{d}y\, \mathrm{d}t \qquad [2]$$

is equal to zero. This is a typical 'empirical generalization'.

The wavelengths λ is usually measured in the band Ω: (0.2, 50 μm), which contains almost 100% of the total energy of incoming and outgoing radiations. Its most part (~99%) is a shortwave radiation (SWR) with wavelengths λ lying within the spectral band S: (0.2, 5.0 μm), where 53.5% constitutes a radiation with $\lambda \in (0.4, 0.7\ \mu\text{m})$, so-called photosynthetically active radiation (PAR). This spectral band is called 'visible'. The radiation with $\lambda \in L(5, 50\ \mu\text{m})$, the long-wave radiation, LWR, constitutes only 0.45% of the total radiation. About 0.5% constitutes an ultraviolet radiation, $\lambda < 0.2$ μm, that is, fortunately, almost completely detained by ozone layer. Note that other divisions of the total spectral band are often used, for instance, S: (0.3, 3.0 μm), etc.

Later on we shall operate with values of energy integrated over these two spectral bands and averaged on the total EAS' surface and 1-year interval: $\hat{E}_S^{\text{in}}, \hat{E}_L^{\text{in}}$ and $\hat{E}_S^{\text{out}}, \hat{E}_L^{\text{out}}$. In accordance with observed data we have with sufficient accuracy that for the EAS: $\left(\hat{E}_S^{\text{in}}\right)_{\text{EAS}} = 340$ W m^{-2}, $\left(\hat{E}_L^{\text{in}}\right)_{\text{EAS}} = 0$, and $\left(\hat{E}_S^{\text{out}}\right)_{\text{EAS}} = 102$ W m^{-2}, $\left(\hat{E}_L^{\text{out}}\right)_{\text{EAS}} = 238$ W m^{-2}, so that $\left(\hat{E}^{\text{in}}\right)_{\text{EAS}} = \left(\hat{E}^{\text{out}}\right)_{\text{EAS}}$. Earth on the whole gets 238 W m$^{-2} \times 5 \times 10^{14}$ m$^2 = 1.2 \times 10^{17}$ W of the SWR, and irradiates as much LWR again.

 doi:10.1016/B978-0-444-63768-0.00604-1

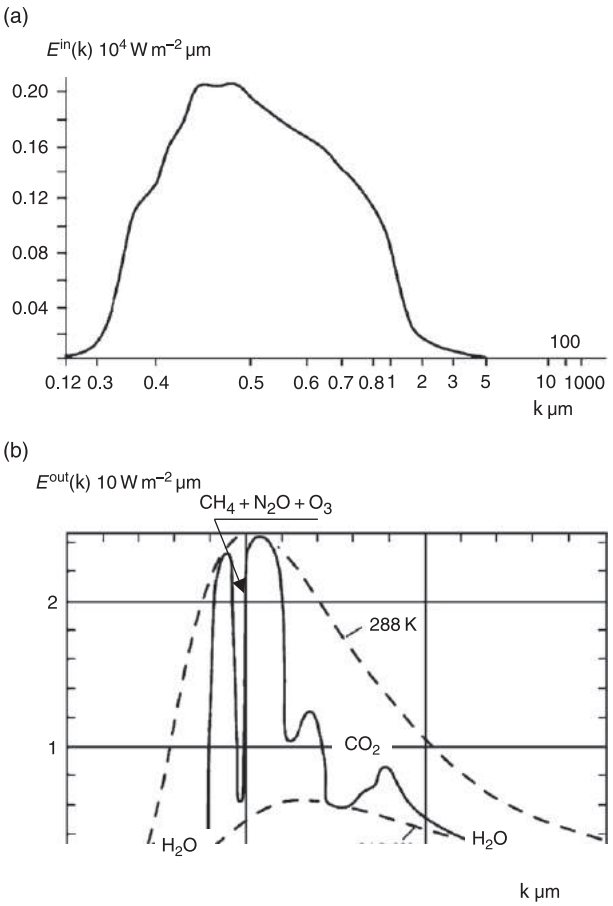


Fig. 1 Spectra of incoming and outgoing solar radiation. (a) Standard spectrum of solar radiation at the top of the atmosphere. (b) Typical spectrum of the Earth's thermal radiation.

The relation between these values can also be presented as

$$\hat{R}_{EAS} = \left(\hat{E}_S^{in}\right)_{EAS}(1 - \alpha_{EAS}) - \left(\hat{E}_L^{out}\right)_{EAS} \quad [3]$$

where $\alpha_{EAS} = \left(\hat{E}_S^{out}\right)_{EAS} \Big/ \left(\hat{E}_S^{in}\right)_{EAS} = 0.3$ is so-called planetary 'albedo' (from Latin 'whiteness'), that is, a coefficient of reflectance of the EAS with respect to incoming radiation.

Transformation of Solar Energy Inside the EAS

What is the fate of 340 W m^{-2} of SWR coming into the top of the atmosphere? Clouds reflect about 68 W m^{-2} of the total incoming radiation. Molecules of atmospheric gases and aerosols still scatter 16 W m^{-2} by forming so-called diffuse radiation, half of which finally goes out to the space, and other half reaches the ground. The 26 W m^{-2} of SWR is reflected by the ground. So, the $68 + 8 + 26 = 102$ J m^{-2} of energy in the form of SWR leaves our planet every second. The rest, 238 W m^{-2}, is consumed and used by the EAS to maintain the work of the Earth's climate machine: evaporation and 'precipitation', 'oceanic currents', atmospheric circulation, etc.

About 78 W m^{-2} is absorbed by the atmosphere and then used in different phase transitions: clouds formation, precipitation, formation of dew, etc. Another part, 160 W m^{-2}, is absorbed by the ground. The measurements show that the spectrum of radiation coming into the Earth's surface is a mixture of SWR, $\left(\hat{E}_S^{in}\right)_G = 186$ W m^{-2}, and LWR (a counter-radiation from the warmed atmosphere), $\left(\hat{E}_L^{in}\right)_G = 102$ W m^{-2}. The ground irradiates into the atmosphere $\left(\hat{E}_S^{out}\right)_G = 26$ W m^{-2} of SWR that corresponds to the mean albedo $\alpha_G = \left(\hat{E}_S^{out}\right)_G \Big/ \left(\hat{E}_S^{in}\right)_G \approx 0.14$ and $\left(\hat{E}_L^{out}\right)_G = 160$ W m^{-2} of LWR. The radiation absorbed by the ground, 160 W m^{-2}, is transformed to heat. The depth of heat penetration depends on the properties of underlying surface. For the land, the depth depends on the properties of soil (for instance, soil moisture) and equals a few meters. As to the ocean, oceanic waves intermix effectively the surface layer, so that the depth is about 100–200 m. If we take into account that the mean depth of

the ocean is 3800 m, then the depth of heat penetration is negligibly small in comparison with the mean depth. Therefore, the warmed body is a very thin film, which is not able to warm the 'lithosphere' (their masses are not commensurable), but it is warming up the atmosphere. Seasonal and diurnal oscillations of the temperature at the levels lying below the film stopped. The film is considered as a low boundary (basement) for the EAS, and it is namely identified with the Earth's surface. A heat flow across the low boundary is negligibly small. For instance, the thermal flow from the Earth's upper mantle is maximally 0.2 W m^{-2}. So, one can say that the EAS consists of the atmosphere and the upper layers of the 'hydrosphere and lithosphere'.

Greenhouse Effect

If Earth would not have the atmosphere, then all absorbed 160 W m^{-2} has to be irradiated into space in the form of LWR, since the temperature of the Earth's surface does not change. However, different atmospheric gases, transparent for SWR, may be weakly transparent for the LWR. For instance, the water vapor strongly absorbs LWR (in the cloudless atmosphere) in the spectral band (5–7.5 μm) and the carbon dioxide in the band (13–17 μm); absorption in the band (9–12 μm) is relatively small (see Fig. 1). As a result, a part of the ground infrared radiation is detained, the atmosphere is warmed, and becomes in turn a source of LWR. Appearing here as a 'counter-radiation', $E_{\text{count}} = \left(\hat{E}_L^{\text{in}}\right)_G = 102$ W m^{-2}, it compensates a significant part of the ground LWR. A difference between the ground LWR and the counter-radiation is called an 'effective radiation' of the Earth's surface, $E_{\text{eff}} = \left(\hat{E}_L^{\text{out}}\right)_G - \left(\hat{E}_L^{\text{in}}\right)_G = 58$ W m^{-2}. Its spectrum is close to the blackbody one with $T = 288$ K, and namely this amount of LWR is irradiated by the Earth's surface into the space, the rest 180 W m^{-2} is irradiated by the atmosphere. This is an essence of the 'greenhouse effect'.

Albedo

Theoretically, the albedo's value may change from 0 of a blackbody until 1 of a 'white body' completely reflecting the solar radiation. It is natural that an albedo depends on spectrum of the radiation, since different surfaces reflect differently in different spectral bands.

Albedos of typical underlying surfaces in visible light range from 0.04 for charcoal, one of the darkest substances, up to 0.90–0.95 for fresh snow. Albedo of salt and sand deserts are 0.45–0.5, while the albedo of coniferous forest is 0.1. Note that the maximal albedo of a surface, covered by vegetation (meadow), is 0.25. Albedo of wet soils is usually less than the albedo of dry ones, for instance, the albedo of chernozem (~ 0.15) is reduced to 0.05 under moistening conditions.

The classic examples of albedo's effect are the snow–, vegetation–, and moisture–temperature feedbacks. If a snow-covered area warms and the snow melts, the albedo decreases down to 0.4–0.5, more sunlight is absorbed, and the temperature tends to increase. If a desert is covered by vegetation, then the albedo decreases, and the temperature has to increase. While the increase in plant biomass tends to slow down in the carbon dioxide concentration in the atmosphere that, in turn, tends to decrease the temperature, so that the balance of these feedbacks may become very complex. Similar considerations are valid also for 'soil moisture– temperature' feedback.

Albedo of water bodies differs from albedo of land surfaces, since the reflection of SWR from water depends on the angle of incidence. At small angles, most part of radiation is reflected from the surface, not penetrating deeply into water body. As a result the albedo increases up by a few tenths, while the albedo at the great angles, that is, when the Sun elevation is high, is equal to a few hundredths. For instance, if the angle of incidence $\beta < 10^\circ$, then $\alpha > 0.22$; if $\beta > 45^\circ$, then $\alpha < 0.05$. Albedo of scattered radiation does not really depend on the angle of incidence, and it is almost constant, about 0.10.

The significant part of incoming radiation is reflected by clouds; their albedo, depending on the thickness of cloudiness, is equal on average to 0.4–0.5, so that the mean albedo of the Earth is about 0.29. This is far higher than for the ocean primarily.

The Earth's surface albedo is regularly estimated via 'Earth observation satellite sensors' such as NASA's MODIS instruments onboard the Terra and Aqua satellites.

Equations of Radiative Balance

Due to the greenhouse effect, the Earth's surface gets the 102 W m^{-2} of radiative heat additionally. An amount of 80 W m^{-2} of this heat is used in the process of evaporation and transpiration of water by plants, evapotranspiration, and 20 W m^{-2} are transported into the atmosphere by a turbulent (sensible) heat flow, E_{turb}, caused by a difference in the temperatures of the ground and the atmosphere. The first term is named a 'latent' flow; it is equal to $L \cdot Q$, where $L = 2453$ J g^{-1} is the specific enthalpy of evaporation (heat content) and Q is the flux of water, evaporated from the surface of water-bodies, soils, and plants, and also water, condensed on these surfaces. To close the balance, we add the value of $E_{\text{mech}} = 2$ W m^{-2} that is a dissipated mechanical energy (friction). The corresponding equation of radiative balance for the Earth's surface is

$$\hat{R}_G = \left(\hat{E}_S^{\text{in}}\right)_G (1 - \alpha_G) - E_{\text{eff}} \qquad [4]$$

which is positive, $\hat{R}_G = 102$ W m^{-2}.

Since the radiative balance of the EAS is equal to zero, the radiative balance of the atmosphere

$$\hat{R}_a = \hat{R}_{EAS} - \hat{R}_G = \left(\hat{E}_S^{in}\right)_{EAS}(1-\alpha_{EAS}) - \left(\hat{E}_S^{in}\right)_G(1-\alpha_G) - \left[\left(\hat{E}_L^{out}\right)_{EAS} - E_{eff}\right] \quad [5]$$

has to be negative. The negativeness is compensated by the latent and turbulent heat flows.

The main carriers of heat between the ground and the atmosphere are precipitation and water vapor. Then the radiative balance for the EAS can be represented as

$$\hat{R}_{EAS} = F_s + L(Q-P) \quad [6]$$

where the term F_s is the sum of the heat inflows and outflows across the vertical walls of the EAS column with unit basement, the term $L(Q-P)$ is a difference between the flows of latent heat $L \cdot Q$ and heat brought by precipitation, $L \cdot P$, where P is the sum of all precipitations. Since for the globe and 1-year interval $Q=P$ and $F_s=0$, then the equation of energy (heat) balance for the EAS has a simple form:

$$\hat{R}_G = 0 \quad [7]$$

The balance equations for the atmosphere and the Earth's surface are

$$\begin{aligned} \hat{R}_a &= -L \cdot P - E_{turb} - E_{mech}(\text{atmosphere}) \\ \hat{R}_G &= L \cdot Q + E_{turb} + E_{mech}(\text{Earth's surface}) \end{aligned} \quad [8]$$

A generalized scheme of energy flows and their transformation is shown in Fig. 2.

At least, we can estimate the internal energy of the atmosphere, which is equal to 8.6×10^{23} J (1.7×10^9 J m^{-2}), the storage of latent heat 3×10^{22} J (6×10^7 J m^{-2}), and the storage of mechanical energy 2.5×10^{15} (5×10^5 J m^{-2}). About 40% of the total atmospheric internal energy constitutes a potential energy (0.7×10^9 J m^{-2}), but only 4 W m^{-2} is necessary to maintain turbulent flows.

Parallel to the vertical redistribution of solar energy, there are powerful energy flows redistributing it over the Earth's surface. All of them form the complex system of atmospheric circulation and oceanic currents that provides to transport heat from the low latitudes to the high latitudes by 'softening' the Earth's climate.

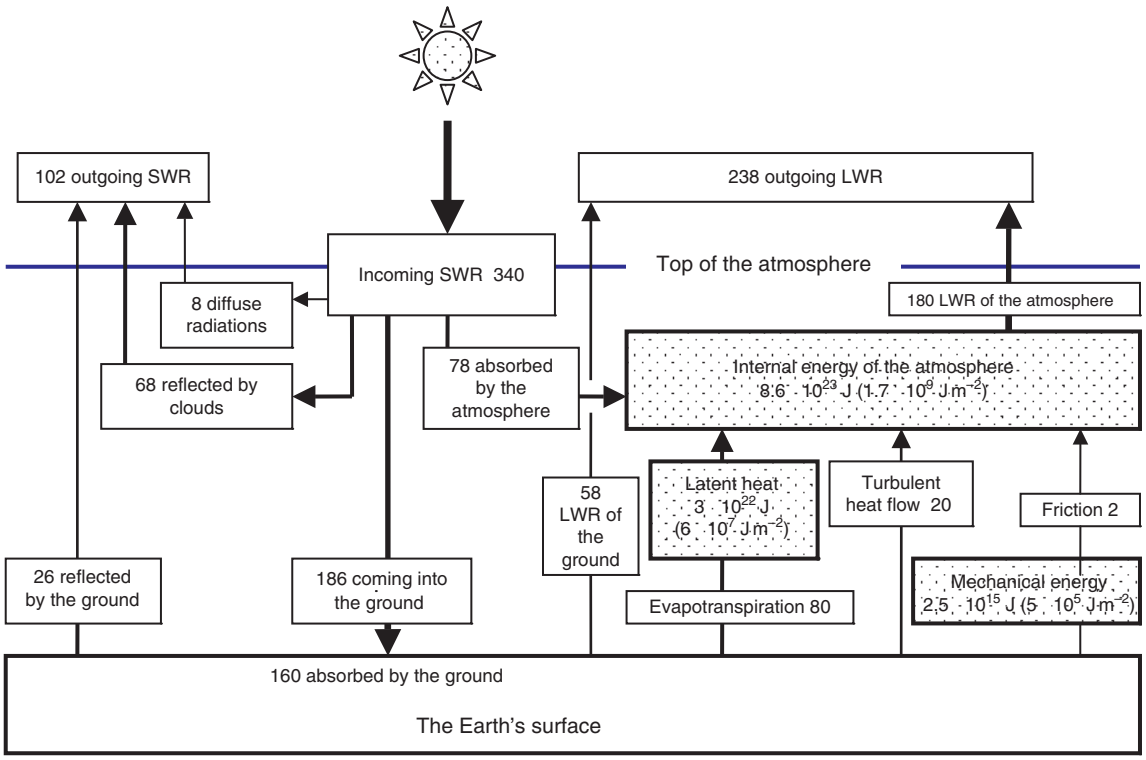


Fig. 2 Energy flows in the system 'the Earth's surface + atmosphere'.

Energetics of Photosynthesis and Vegetation

We described above the main processes of the transformation of solar energy, which are the principal components of the global energy balance, forming in essence the thermostat for the biosphere. However, there are other processes, which do not really influence the energy balance. Their heat flows are very small (mentioned above as the dissipation of mechanical energy, dew condensation, etc.). Some of them, nevertheless, play the principal role in the biosphere, for instance, photosynthesis.

Green plants (autotrophs) convert solar energy into the chemical energy of new living biomass in the process of photosynthesis. The process uses energy of visible light, which is absorbed by the chlorophyll molecules of plants to convert carbon dioxide and water into carbohydrates and oxygen. Note that the presence of oxygen in the Earth's atmosphere is a result of photosynthesis. Proteins, fats, nucleic acids, and other compounds are also synthesized during the process, as long as elements such as nitrogen, sulfur, and phosphorus are available.

Then the stored chemical energy flows into herbivores, carnivores (predators), parasites, decomposers, and all other forms of life. Photosynthesis produces the living biomass of vegetation, constituting more than 95% of the global biomass and being a main agent in the 'global biogeochemical cycle of carbon'.

Photosynthesis

The basic equation of photosynthesis is

$$CO_2 + H_2O + h\nu \rightarrow (CH_2O) + O_2 + 470\ \text{kJ mol}^{-1}$$

where $h\nu$ is a photon energy and (CH_2O) is a fragment of carbohydrate molecule, releasing 470 kJ mol^{-1} of energy (that equals an increase in enthalpy, ΔH). Since the change of free energy is equal to $\Delta G = 504$ kJ mol^{-1}, and $G\Delta = \Delta H - T\Delta S$, the change of entropy, ΔS, is equal to $(470 - 504)/273 = 116$ J K^{-1} mol^{-1} ($T = 293$ K), that is, photosynthesis is an antientropic process.

Efficiency of photosynthesis is defined in different ways. Its theoretically maximal value is the ratio of ΔG to the total energy of eight photons ($E_{ph} = 1470$ kJ mol^{-1}), which are necessary to get one molecule of O_2, $\eta_{max} = 504/1470 = 34\%$. On the other hand, since the 'useful' work, which can be performed by photosynthesis, is 'exergy', $Ex = -T\Delta S = 34$ kJ mol^{-1}, then efficiency is defined as $\eta_{ex} = Ex/E_{ph} = 34/1470 = 2.3\%$.

Efficiency of Vegetation

If the working process creating a new biomass is photosynthesis, then the working machine is plant. Therefore, it is natural to say about efficiency of plant (efficiency of vegetation) or 'green leaf' than about efficiency of photosynthesis. The rate of photosynthesis depends on the amount of light reaching the leaves, the temperature of surrounding air, and the availability of water and other nutrients such as nitrogen and phosphorus. One of these factors ('limiting factors') already limits the rate, so that the real efficiency of vegetation is lower than its theoretical value, 34%, and what is more, this efficiency should not exceed the 'exergic' efficiency, 2.3%.

From the thermodynamic point of view, a 'green leaf' is a heat machine with photosynthesis as the working process, and molecules of chlorophyll, adenosine triphosphate (ATP) etc., transferring energy of photons into leaves as the working body. Efficiency of the heat machine is $\eta_{leaf} = (T_{leaf} - T_{air})/T_{leaf}$, where T_{leaf} and T_{air} are the mean daily temperatures of leaves and surrounding air; since the reaction of photosynthesis is exogenous, the leaf is warmed, $T_{leaf} > T_{air}$. Under summer conditions in temperate forest, $T \sim 5\ ^\circ$C and $T_{air} \sim 20\ ^\circ$C on average; therefore, $\eta_{leaf} = 5/298 = 1.7\%$.

It is known that about 98–99% of solar energy, reaching the Earth's surface, is reflected from leaves and other surfaces and absorbed by other molecules, which convert it to heat. Thus, vegetation is available to catch about 1–2% of incident solar energy, that is, these numbers constitute its efficiency.

The rate at which plants convert PAR (or inorganic chemical energy) to the chemical energy of organic matter is named gross primary production (productivity) (GPP). This value (as well as biomass) is often reported in grams or metric tons of either dry weight or carbon (the latter is about one-half of the first). Since enthalpy of 1 of carbon is equal to ~ 42 kJ g^{-1}, then production and biomass can be also reported in joules.

Fifteen to sixty percent of the energy assimilated by plants immediately is spent in cellular respiration, when carbohydrates, proteins, and fats are broken down, or oxidized, to provide energy (in the form of ATP) for the cell's metabolic needs. The residual (40–85%) is stored in biomass as net primary production (NPP). The highest annual NPP, 2000 gC m^{-2} yr^{-1}, occurs in swamps, marshes, and tropical rainforests; the lowest, 20 gC m^{-2} yr^{-1}, occurs in deserts. The mean NPP for terrestrial ecosystem is about 400 gC m^{-2} yr^{-1}. Among aquatic ecosystems, the highest NPP, 2000 gC m^{-2} yr^{-1}, occurs in estuaries; the mean NPP in the ocean is 75 gC m^{-2} yr^{-1}, so that the ocean is a desert (see **Table 1**).

Efficiency of solar energy utilization by vegetation can be defined as the ratio of enthalpy, contained in the NPP, to the solar radiation, reaching to the Earth's surface and integrated over the vegetation period. The corresponding values for continents and for land overall are shown in **Table 1**.

Table 1 The NPP and the efficiency of utilization

Continents	*NPP* ($gC\ m^{-2}\ yr^{-1}$)	*Efficiency (%)*
Europe	365	0.54
Asia	421	0.38
North America	353	0.40
South America	899	0.49
Africa	443	0.25
Australia with Oceania	370	0.19
Land on average	408	0.37

One square meter of the terrestrial vegetation on average utilizes in the course of 1 year about 17 million joules of solar energy, but this gigantic number constitutes only 0.37% of the total solar energy that comes into the Earth's surface.

The total annual production of terrestrial vegetation is about 60 gigatons (Gt, 1 Gt $= 10^9$ t) of carbon, while for the ocean this value is estimated as 25 Gt with NPP $= 75$ gC m^{-2} yr^{-1}. Thus, the mean global NPP is 186 gC m^{-2} yr^{-1}, and the mean efficiency of global vegetation is about 0.1%. However, it is necessary to take into account that much energy is consumed in the process of forming and maintaining of the thermostat for vegetation. It is very similar to the situation with greenhouse, where the most part of energy is used for its heating.

Energy Transfers, Trophic Chains, and Trophic Networks

If we look at a global pattern of pathways on which the solar energy stored in biomass is flowing within the gigantic (and unique) ecosystem (often associated with the biosphere), we see the network entangling the Globe. It is named a 'trophic network or a food web', and as a rule subdivided on local networks. The trophic network is described by an oriented graph with vertices corresponding to species that constitute the ecosystem, and links indicating trophic interaction between them (their directions show the energy flowing, for instance, prey → predator). In the network structure the 'trophic levels' are naturally distinguished, that is, groups of species having no direct trophic interactions; however, species of one level usually either compete for life resource or cooperate in its utilization. It is natural that some part of energy is spent (and later on dissipated as heat) in such kind of interactions – this is a payment by means of energy for stability of the network structure. Another significant part of consumed energy (from 30% to 70%) is spent for maintaining life in the process of 'metabolism' (respiration).

In any trophic network a structure, in which every two adjacent species form a prey–predator pair and which is described by a linear graph, can be distinguished. It is called a 'trophic (food) chain'; their interlacing and branching form a trophic network. Since the energy dissipation along the chains is very high they are usually short (their length measured in the number of links is about 4–6).

The basic trophic species of chain usually are 'producers' (plants, autotrophic organisms that accumulate the solar energy and 'nutrients' – carbon, nitrogen, phosphorus, etc.); the next species are 'primary consumers' (herbivorous, heterotrophic organisms), and 'secondary consumers' (carnivores, predators, preying on herbivorous). Really, the chain may be longer. It is not necessary that the chain be originated by an autotroph: it may be any species considered as a resource for consequent ones. For instance, if a resource is 'detritus' (faeces, dead organic matter) then a special 'detritus chain' can be considered. At last, trophic chains could be 'open' and 'closed'; as a rule they are open in relation to the energy flowing through an ecosystem and carbon that is accumulated in the process of photosynthesis and spent in the respiration, and closed in relation to nutrients turning in the ecosystem.

In order to start a 'biogeochemical machine' we have to 'close' the chain by species named 'decomposers' (protozoa, bacteria, fungi, scavengers, and carrion eaters), which in the course of their vital activity split complex organic compounds into simpler mineral substrates (nutrients) for autotrophs. A principal scheme of such kind of 'biogeochemical machine' is shown in **Fig. 3**.

Really, the closure is not complete: about 1% of dead organic matter is deposited in the deep ground ('kerogen'), and has accumulated over long periods of geologic time (oil and coal repositories).

A small amount of the energy passes from one trophic level to another; for instance, only from 5% to 25% of plant biomass is consumed by herbivores, the rest falls out and becomes a resource for decomposers (detritophages). Efficiency of this passing is called 'ecological efficiency'; on average, it equals 10%. The rate at which these consumers use the chemical energy of their food for growth and reproduction is called 'assimilation efficiency'. For instance, assimilation efficiency of herbivores lies in the interval from 15% to 80%, while the interval for carnivores is from 60% to 90%.

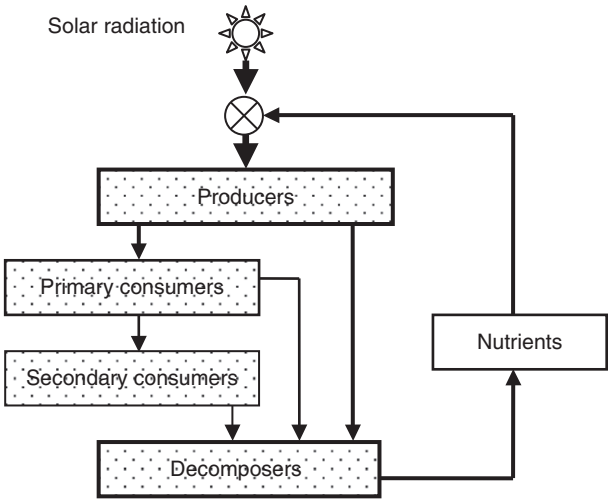


Fig. 3 Flows of mass and energy in an elementary biogeochemical machine.

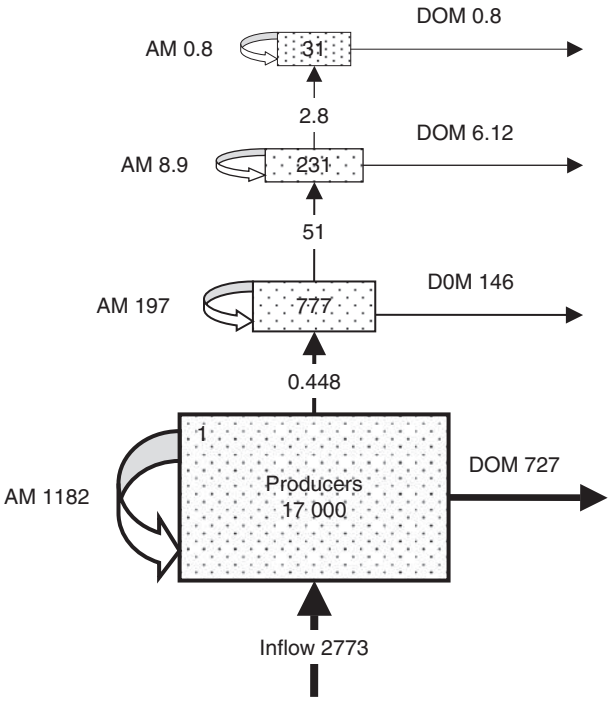


Fig. 4 Trophic chain of the ecosystem of Silver Springs. Energy flows and biomasses are measured in mW m^{-2} and kJ m^{-2}, respectively. DOM, dead organic matter; AM, assimilation and metabolism.

It is easy to estimate that such a predator as *Homo sapiens* (the third trophic level) gets only 1% of solar energy stored by plants. Unfortunately, the situation has not improved; if he would be a vegetarian, by winning in the ecological efficiency, he would lose in the assimilation efficiency.

The result of such kind of consequent energetic transitions is a pyramid of energy, with most energy concentrated by autotrophs at the bottom of trophic chain and less energy at each higher trophic level. As an example the trophic chain and the pyramid of biomass of the concrete ecosystem of warm Silver Springs in Florida are presented in **Fig. 4**. Note that it is a classic object that has been studied by H. T. Odum. The ecosystem has four trophic levels: (1) producers (phytoplankton), (2) herbivores (zooplankton), (3) carnivores (fish), (4) higher predators (predacious fish), and one special level, decomposers, with biomass equal to 105 kJ m^{-2}. Since the system is through-flowing, that is, described in the terms of energy flows, therefore the chain may be considered open, without decomposers.

Conclusion

As described above biosphere machines from the anthropocentric point of view are badly made, with very low efficiency. They dissipate the solar energy by heating the environment more than perform some useful work. Nevertheless, they are significantly reliable. Since it is necessary to pay for their reliability and stability, they pay by high dissipation of energy that in turn decreases their efficiency.

See also: Ecological Complexity: Thermodynamics in Ecology. Global Change Ecology: Energy Balance; Biosphere: Vernadsky?s Concept. Human Ecology and Sustainability: Emergy and Sustainability

Further Reading

Budyko, M.I., 2001. Evolution of the Biosphere (Atmospheric and Oceanographic Sciences Library). Berlin: Springer, 444p.
Jørgensen, S.E., Svirezhev, Yu M., 2004. Towards a Thermodynamic Theory for Ecological Systems. Amsterdam: Elsevier, 366p.
Morowitz, H.J., 1978. Foundations of Bioenergetics. New York: Academic Press.
Smile, V., 2002. The Earth's Biosphere: Evolution, Dynamics, and Change. Cambridge, MA: MIT Press.

Entropy and Entropy Flows in the Biosphere

YM Svirezhev, Potsdam Institute for Climate Impact Research, Potsdam, Germany

Introduction

From the thermodynamic point of view, Earth is a closed system, which gets 1.2×10^{17} J of energy from the Sun every second in the form of a short-wave radiation, which corresponds to the density of energy flow in 238 W m^{-2}. The same amount of energy is irradiated into the space in the form of a long-wave (infrared) radiation. We assume that the Earth's total mass and its mean temperature (more precisely, the temperature of the Earth–Atmosphere System (EAS)) are not changing in the course of rather long time ($\sim 10^3$ years). The latter means that the planetary radiative balance is constant. These are plausible hypotheses, which can be considered as 'empirical generalizations'.

Carriers of energy are 'hot' photons with temperature $T_S = 5800$ K of the Sun's surface, and the energy is carried away by 'cooled' photons at $T_E = 253$ K. This is the so-called 'photon mill'; evolution and self-organization of planets (including life on Earth) is a result of its work.

Formally, the photon mill is a typical 'heat machine' functioning as a Carnot cycle, but its working body is the photon gas, whose 'molecules' have no mass, so that in this case it becomes slightly incorrect to talk about a heat machine (although Gibbs has indicated it). Later on, Prigogine stated that such a classic thermodynamic concept as the heat machine is also applicable to the photon gas. Note that this 'roughness' is not necessarily present, if the concept of 'exergy' is used.

Let $d_i\sigma/dt$ be the internal production of entropy by the EAS, and $d_e\sigma/dt$ be the exchange flow of entropy between the Sun and the EAS, then the change in the total entropy of the EAS is

$$\frac{d\sigma}{dt} = \frac{d_e\sigma}{dt} + \frac{d_i\sigma}{dt} \qquad [1]$$

The value of $d_e\sigma/dt$ can be estimated as the algebraic sum of the entropy flow from Sun to Earth, $q_{SE} = (4/3)\ (238\ \mathrm{W\ m^{-2}})\ (1/T_S)$, and the entropy flow from the EAS to space, $q_{ES} = -(4/3)\ (238\ \mathrm{W\ m^{-2}})\ (1/T_E)$:

$$\frac{d_e\sigma}{dt} = \frac{4}{3}\left(238\mathrm{W}\,m^{-2}\right)\left(\frac{1}{5800} - \frac{1}{253}\right) \approx -1.2\mathrm{W\,K^{-1}m^{-2}} \qquad [2]$$

where factor 4/3 is the so-called Planck's form factor. The annual entropic balance for the globe overall is equal to -2×10^{22} J K^{-1} yr^{-1}.

We assume here implicitly that the irradiation of the EAS is the blackbody irradiation with $T_A = T_E = 253$ K. Indeed, the irradiation is a sum of the blackbody irradiations with the temperatures from 215 to 288 K, so that this estimation is a zero approximation.

In accordance with Prigogine's theorem, at the dynamic equilibrium the system's entropy must be constant, that is, $d\sigma/dt = 0$; whence

$$-\frac{d_e\sigma}{dt} = \frac{d_i\sigma}{dt} \qquad [3]$$

The value of $d_e\sigma/dt$ is known; if we could estimate the value of $d_i\sigma/dt$, and if equality [3] holds, or, in other words, if the internal production of entropy is balanced by its export into the environment, we would prove one important statement: the EAS is at the dynamic equilibrium with its environment, the space. Note that equality [3] has to hold overall for the EAS, but each of its subsystems may be in nonequilibrium, so that the main statement of Prigogine's theorem (equality [3]) in relation to each subsystem does not necessarily hold.

Entropy Flows in the EAS

Let us look at the simplified scheme of the energy flows in the EAS shown in **Fig. 1**.

Using **Fig. 1** we can write equations of entropic balance for the EAS (atmosphere + ground) in a more detailed way:

$$\begin{aligned}\frac{d\sigma_G}{dt} &= \frac{4}{3}\frac{q(SG)}{T_S} - \frac{q_1(GA)}{T_G} - \frac{q_2(GA)}{T_G} - \frac{4}{3}\frac{q(GS)}{T_G} + \frac{d_i\sigma_G}{dt} \\ &\approx -0.586\,\mathrm{W\,K^{-1}\,m^{-2}} + \frac{d_i\sigma_G}{dt}\end{aligned} \qquad [4]$$

 doi:10.1016/B978-0-444-63768-0.00888-X

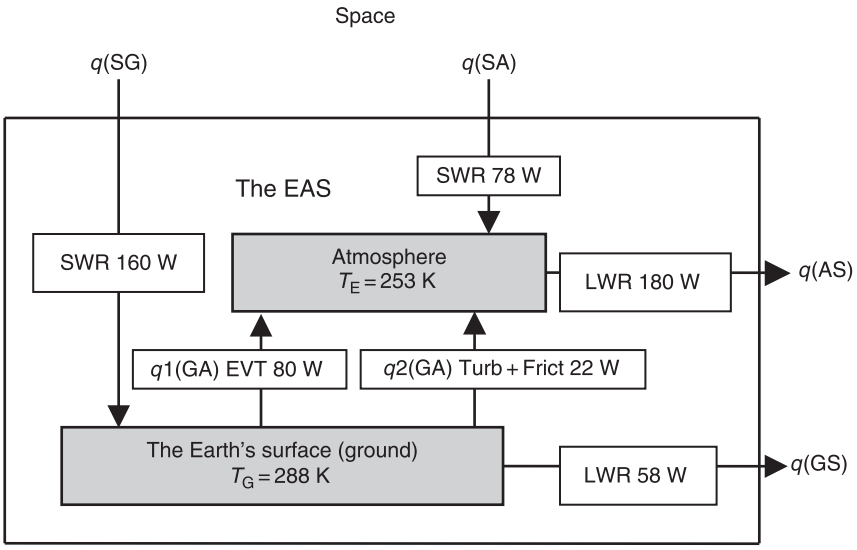


Fig. 1 Energy flows in the EAS. SWR is the flow of short-wave radiation with $T_S = 5800$ K, LWR is the flow of long-wave radiation with the temperatures of the Earth's surface (ground), $T_G = 288$ K, or the atmosphere $T_E = 253$ K; EVT is the flow of latent heat (evapotranspiration), and Turb + Frict is the flow of turbulent heat (sensible flow) + the flow of heat discharged in mechanical movements (friction).

$$\begin{aligned}\frac{d\sigma_A}{dt} &= \frac{4}{3}\frac{q(SA)}{T_S} + \frac{q_1(GA)}{T_G} + \frac{q_2(GA)}{T_G} - \frac{4}{3}\frac{q(AS)}{T_E} + \frac{d_i\sigma_A}{dt} \\ &\approx -0.576\,W\,K^{-1}\,m^{-2} + \frac{d_i\sigma_A}{dt}\end{aligned} \quad [5]$$

It is interesting that the exchange entropic flows, $d_e\sigma_G/dt \approx -0.586\,W\,K^{-1}\,m^{-2}$ for the ground and $d_e\sigma_A/dt \approx -0.576\,W\,K^{-1}\,m^{-2}$ for the atmosphere, are almost equal, and their sum is equal to $d_e\sigma/dt \approx -1.16\,W\,K^{-1}\,m^{-2}$, that almost coincides with the value given by eqn [2]. In accordance with Prigogine's theorem,

$$-\frac{d_e\sigma}{dt} = \frac{d_i\sigma}{dt} = \frac{d_i\sigma_A}{dt} + \frac{d_i\sigma_B}{dt} \approx 1.16\,W\,K^{-1}\,m^{-2} \quad [6]$$

Entropy Storage of the Biota

The EAS is divided into four subsystems: atmosphere (A), hydrosphere (H), pedosphere (P), and biota (B). The atmosphere is a mixture of different gases: mainly nitrogen and oxygen; in lesser concentrations, carbon dioxide, water vapour, argon, etc., which determine the thermal regime of our planet. The hydrosphere's mass is a mass of all water (including salt dilutions and excluding polar ice and glaciers). The pedosphere is soils. All these are exchanging energy and matter with each other, and in turn the EAS is exchanging, however, only energy with space. In particular, the matter exchange is realized by means of the global biogeochemical cycles.

The atmosphere, hydrosphere, and pedosphere have stored gigantic amounts of entropy. For instance, the entropy storage of atmosphere is $3.5 \times 10^{22}\,J\,K^{-1}$ that in general is close to the global entropy balance; the storages of other subsystems are significantly larger. The exchange entropy flows that bound them with the biota are relatively weak with respect to their entropy storages, and do not really change their state, but they are able to change the state of biota. Thus, the latter is important for us.

Since the atmospheric CO_2 is one of the 'life-forming' gases, it is interesting to estimate its entropy, which is equal to $1 \times 10^{19}\,J\,K^{-1}$.

The biota is defined as all of the Earth's living matter. Apparently, this is one of the reasons why the term 'biosphere' is often used (especially in Anglo-Saxon literature) in the sense of 'biota'. The present bulk of living organisms are confined to land, and their mass (on dry basis) amounts to 1.88×10^{18} g. For instance, the oceanic biomass is about 0.5% of that in land. Since the terrestrial vegetation constitutes the most part of biota, mainly contributing to its dynamics, the biota is identified with the terrestrial vegetation.

So, biota is the terrestrial phytomass, put into a thermostat with the mean annual temperature of the Earth's surface, $T_B = 15\,^{\circ}C$. The total phytomass is known; hence, if only the specific entropy of living matter is also known, there is no problem in calculating the total entropy of biota. However, here we deal with a strongly nonequilibrium system, and it is unknown how to define the entropy in this case. What can be done here is to calculate the entropy of dead organic matter (DOM; in dry weight), $s(DOM) = h(DOM)/T_B$, where $h(DOM) = (16.4 - 18.4)\,kJ\,g^{-1}$ is its specific enthalpy. Therefore, $s(DOM) = 60.4\,J\,K^{-1}\,g^{-1}$, and the total entropy of 'dead biota' is $S(DOM) \approx 1.1 \times 10^{20}\,J\,K^{-1}$, which is less by two orders of magnitude than the atmosphere entropy.

Change of Entropy in the Terrestrial Biota

Let $d_iS_B/dt = \dot{S}_B^i$ be the annual internal production of entropy by the global biota, and $d_eS_{jB}/dt = \dot{S}_{jB}$, j=S, P, H, A be the flows of entropy from jth subsystem into biota, then the rate of its entropy change is

$$dS_B/dt = (\dot{S}_{SB} + \dot{S}_{PB} + \dot{S}_{HB} + \dot{S}_{AB}) + \dot{S}_B^i \quad [7]$$

Here we make a very important assumption: although we do not know what the specific entropy of living matter is, we can still speak about the change of entropy. For instance, the value of $(-T_B \cdot \Delta S)$ can be interpreted as an ability of a living system to perform the work, which in turn can be measured. This ability is named 'exergy'.

For the Earth's surface, using the data given above, $d_e\sigma_G/dt \approx -0.586\ W\ K^{-1}\ m^{-2}$. Since the area of globe is $5.1 \times 10^{14}\ m^2$, $d_eS_G/dt \approx (-0.586\ W\ K^{-1}\ m^{-2})(5.1 \times 10^{14})\ (3.15 \times 10^7) = -9.41 \times 10^{21}\ J\ K^{-1}\ yr^{-1}$.

Exchange between Space and Biota

We assume that plants use only the solar short-wave radiation, absorbed by the Earth's surface, $160\ W\ m^{-2}$. Since plants absorb only 53.5% of this energy (photosynthetically active radiation), $85.6\ W\ m^{-2}$, and vegetation covers 72.5% of land area, $A_{veg} \approx 1.1 \times 10^{14}\ m^2$, the biota gets annually

$$\begin{aligned} \dot{S}_{SB} &= \frac{4}{3}(85.6\ W\ K^{-1}\ m^{-2})\frac{A_{veg}}{T_S}(3.15 \times 10^7) \\ &= 0.53 \times 10^{20}\ J\ K^{-1}\ yr^{-1} \end{aligned} \quad [8]$$

Here T_S=5800 K.

Exchange between Pedosphere and Biota

This flow is mainly determined by the flow of DOM from terrestrial biota into pedosphere q_{DOM}. Assume that annual flow of DOM is equal to the net primary production (NPP), 140 Gt of dry matter, then $q_{DOM} = -1.4 \times 10^{17}$ g d.w. of DOM per year. Since a living biomass contains about 65% (on average) of water, then it is natural to assume that standing dead vegetation contains the same percentage of water, and the flow of dry DOM, -1.4×10^{17} g, has to be accompanied by the water flow, -2.6×10^{17} g H_2O. By taking into account that specific entropy of H_2O is $3.89\ J\ K^{-1}\ g^{-1}$, the total entropy flow $\dot{S}_{PB} = -(60.4 \times 1.4 + 3.89 \times 2.6) \times 10^{17} = -0.947 \times 10^{19}\ J\ K^{-1}\ yr^{-1}$.

There is also a reversible flow of minerals (the nutrients: nitrogen, phosphorus, potassium, etc.), which are used in the process of creation of new biomass. All these substances come into the biota in the form of water solutions, entropy of which is the sum of the water entropy and exactly these elements. Note that their contribution constitutes less than 1% of the contribution of water.

Exchange between Hydrosphere and Biota

This flow is defined as $\dot{S}_{HB} = s(H_2O) \cdot q_{HB}(H_2O)$, where $q_{HB}(H_2O)$ is the annual flow of water consumed by biota, and $s(H_2O) = 3.89\ J\ K^{-1}\ g^{-1}$ is the specific entropy of liquid water at T=288 K. We assume that $q_{HB}(H_2O)$ is equal to the annual transpiration of global vegetation, 4.8×10^{19} g H_2O. Then $\dot{S}_{HB} = 3.89 \times 4.8 \times 10^{19} = 1.87 \times 10^{20}\ J\ K^{-1}\ yr^{-1}$.

Exchange between Atmosphere and Biota

This flow, $\dot{S}_{AB}$, is a sum of the following flows:

1. entropy flow caused by diffusion of CO_2 through stomata into leaves, $\dot{S}_{AB}(CO_2)$;
2. entropy flow caused by diffusion of O_2 through stomata into the atmosphere, $\dot{S}_{BA}(O_2)$; and
3. entropy flow caused by the transpiration of water, $\dot{S}_{BA}(H_2O)$.

The first and second flows are defined as $\dot{S}_{AB}(CO)_2) = s(CO_2)_{AB} \cdot q_{AB}(CO_2)$ and $\dot{S}_{BA}(O_2) = -s(O_2) \cdot q_{BA}(O_2)$, where $q_{AB}(CO_2) = NPP[gC] \cdot (44/12) = 6.6 \times 10^{16} gC(44/12) = 2.42 \times 10^{17} gCO_2\ yr^{-1}$ and $q_{BA}(O_2) = -(6.6 \times 10^{16} gC)(32/12) = -1.76 \times 10^{17} gO_2 yr^{-1}$ are the rates of consumption and release of carbon dioxide and oxygen by plants in the process of photosynthesis. The specific entropies are: $s(CO_2) = 4.86\ J\ K^{-1}\ g^{-1}$ and $s(O_2) = 6.41\ J\ K^{-1}\ g^{-1}$. Then $\dot{S}_{AB}(CO_2) = 1.18 \times 10^{18}\ J\ K^{-1}\ yr^{-1}$ and $\dot{S}_{BA}(O_2) = -1.13 \times 10^{18}\ JK^{-1}\ yr^{-1}$. The summation of these flows gives $\dot{S}_{AB}(CO_2) + \dot{S}_{BA}(O_2) = (1.18 - 1.13) \times 10^{18} = 5 \times 10^{16}\ J\ K^{-1}\ yr^{-1}$, that is, the exchange flows of entropy related to CO_2 and O_2 are almost balanced by each other, so that their sum is reduced by two orders of magnitude.

The entropy flow $\dot{S}_{BA}(H_2O) = s(WV) \cdot q_{BA}(H_2O)$, where $q_{BA}(H_2O) = -q_{HB}(H_2O) = -4.8 \times 10^{19} gH_2O\ yr^{-1}$ is the annual transpiration through stomata (we assume that all consumed water is transpired), and s(WV) is the specific entropy of water vapor at T_B=288 K and 1 atm. We see that maximal entropy flows are associated with water in liquid and vapor forms, that is, with the global water cycle. Their total balance $\dot{S}(H_2O) = \dot{S}_{HB} + \dot{S}_{BA}(H_2O) = q_{BA}(H_2O) \cdot (h_{ev}/T_B)$, where h_{ev}=2462 J per g H_2O is the

specific enthalpy of evaporation, and is equal to a jump of entropy caused by the phase transition 'liquid water→water vapor', $\dot{S}(H_2O) = -4.1 \times 10^{20}\,J\,K^{-1}\,yr^{-1}$.

The internal production of entropy, $\dot{S}_B^i$, can be represented as a sum of two terms: $\dot{S}_B^i = \dot{S}_B^{DOM} + \dot{S}_B^{Work}$. The first is mainly connected with chemical reactions forming structural molecules of biomass (cellulose, proteins, carbohydrates, lipids, etc.). Organic compounds containing phosphorus take an active part in such type of reactions. All these processes are associated mainly with the carbon, nitrogen, and phosphorus biochemical cycles, the entropy flows of which are less than in the water cycle approximately by two (and less) orders of magnitude. Certainly, knowing the chemical composition of living matter, we can calculate its chemical entropy as a sum of corresponding specific entropies weighted proportionally to their percentages. However, since the dead matter has the same composition, then the specific chemical entropies of living and dead matter do not differ from each other. Hence, we can assume that the processes of forming of the new biomass and falling off the DOM with respect to their chemical composition are mutually reversible, that is, $\dot{S}_B^{DOM} + \dot{S}_{PB} = 0$.

The second term, $\dot{S}_B^{Work}$, is the entropy produced by the biota during its working cycle (see details below).

By summing all these flows, we get

$$dS_B/dt \approx [(0.53 - 4.11) \times 10^{20} = -3.58 \times 10^{20}\,J\,K^{-1}\,yr^{-1}] + \dot{S}_B^{Work} \quad [9]$$

that is, from the thermodynamic point of view, biota is a strongly nonequilibrium system. The structure of the exchange entropic flows for the biota is shown in Fig. 2.

If we compare $d_eS_B/dt \approx -3.57 \times 10^{20}\,J\,K^{-1}\,yr^{-1}$ and $d_eS_G/dt \approx -9.41 \times 10^{21}\,J\,K^{-1}yr^{-1}$, then it is easy to see that these values differ by 26 times. Even if we take into account that in the case of biota we deal with the land area (covered by vegetation), which is less by almost fivefolds than the globe area, then we have almost five-multiple excess. Nevertheless, if we now compare the biota and the Earth's surface with respect to the energy obtained (the first obtains less than 1% in comparison with the second), then we can conclude that the biota is one of the main actors on the entropic scene.

Biota Performs the Work

We have an argument carrying the concept of self-organization of the biosphere: the very existence of the living biota is in necessary disequilibrium with the nonliving part of the biosphere. The main consequence of this disequilibrium is that the biota is able to perform some useful work, and a measure of this work is 'exergy'. By performing work, the biota produces additional entropy; namely, this entropy 'closes' its balance. This in turn allows us to say that the biota is in dynamic equilibrium at least in the course of last millenaries. What is this work? This is mainly the chemical work of the biogeochemical cycles, the work forcing to move the matter flows, that is, forcing the 'wheels' of the 'biosphere machine' to be turned, and then to move evolution. In particular, all the work to produce and maintain the gigantic overproduction of offspring (namely this is one of the main ideas of Darwin, while the concept of natural selection dates back to antiquity) is the work of biota.

So, this work (more correctly the ability to perform the work, i.e., 'exergy') is equal to

$$Ex = -T_B(\text{change of entropy}) = -T_B(d_eS_B/dt) \approx 1 \times 10^{23}\,J\,yr^{-1} \quad [10]$$

Exergy is also defined as $Ex = \beta\, h_{DOM} \cdot NPP$, where $h_{DOM} = 17.4\,kJ\,g^{-1}$ is the specific enthalpy of DOM, $NPP = 1.4 \times 10^{17}$ g d.w. is the annual NPP, and the factor β is some specific genetic characteristic of a living organism defined by a number of nonrepetitive genes in its genome. For plants, this value lies in the range 29–87. In our case $\beta \approx 42$, which is close to the mean of this interval, $\hat{\beta} = 58$.

It is not a secret that all these estimations for the total biomass of vegetation, its annual production, the volume of water transpired by plants, etc., are rather conditional and strongly varying. Nevertheless, the corresponding values of β in most cases get into the interval from 29 to 87.

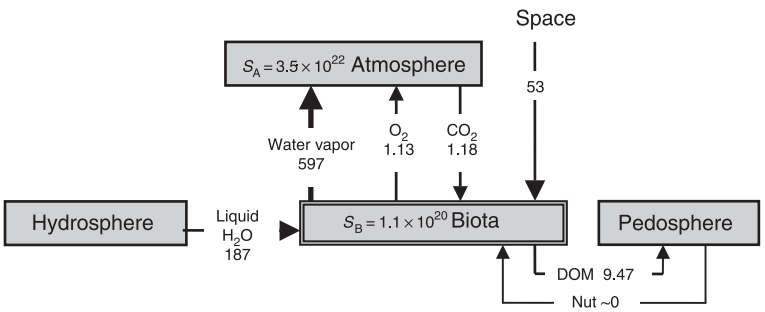


Fig. 2 Exchange entropic flows for the biota (all the flows are shown in $10^{18}\,J\,K^{-1}\,yr^{-1}$). DOM is the entropy flow determined by falling off the dead organic matter, Nut is the entropy flow of nutrients (nitrogen, phosphorus, potassium, etc.).

Humans and Biota

The annual production of artificial energy consumed by the anthroposphere constitutes about 3×10^{20} J yr^{-1}, which is $\sim$12% of the global terrestrial NPP. If all the energy is transformed into heat, then the annual production of entropy is $\dot{S}^{\mathrm{Art}} \approx 1 \times 10^{18}$ J K^{-1} yr^{-1}, which is comparable with some entropic flows in the biota (e.g., with flows caused by uptake of CO_2 and release of O_2).

Now the biosphere and anthroposphere are in the state of strong competition for common resources, such as land area and fresh water. Contamination of the environment and reduction of the biota diversity are the consequences of the competition.

Since the biosphere (considered as an open thermodynamic system) is in dynamic equilibrium, then all entropy flows have to be balanced too. Therefore, the entropy excess, which is created by the anthroposphere, has to be compensated by means of two processes: (1) reduction of the biota and degradation of the biosphere, and (2) change in the work of the Earth's climate machine (in particular, an increase in the Earth's mean temperature). Note that in any case it is desirable to include the entropic flow $\dot{S}^{\mathrm{Art}}$ into the total balance of entropy for both, the atmosphere and biota, but we shall assume that the anthropogenic impact concentrates only on the biota. Then $\dot{S}_{\mathrm{B}}^{\mathrm{i}} = \dot{S}_{\mathrm{B}}^{\mathrm{DOM}} + \dot{S}_{\mathrm{B}}^{\mathrm{Work}} + \dot{S}^{\mathrm{Art}}$. By assuming that equalities $\dot{S}_{\mathrm{AB}}(\mathrm{CO_2}) + \dot{S}_{\mathrm{BA}}(\mathrm{O_2}) \approx 0$ and $\dot{S}_{\mathrm{B}}^{\mathrm{DOM}} + \dot{S}_{\mathrm{PB}} = 0$ hold in this case also, we get the following simplified equation:

$$\mathrm{d}S_{\mathrm{B}}/\mathrm{d}t \approx \dot{S}(\mathrm{H_2O}) + \dot{S}_{\mathrm{B}}^{\mathrm{Work}} + \dot{S}^{\mathrm{Art}} \qquad [11]$$

The total flow of transpiration can be represented as $|q_{\mathrm{BA}}(\mathrm{H_2O})| = bB$, where B is the total mass of biota (in d.w. of DOM) and $b = |q_{\mathrm{BA}}(\mathrm{H_2O})|/B$ is the specific intensity of transpiration (in g H_2O g^{-1} d.w.), which is constant. We implicitly assume here that the power of transpiration 'pump' is proportional to the biomass of plant.

On the other hand, since this value of water is necessary to transpire in order to create P units of a new biomass (P=NPP in d.w.), then $|q_{\mathrm{BA}}(\mathrm{H_2O})| = pP$, where p is the amount of transpired water, which is necessary for creating 1 g of biomass. Therefore, the coefficient $P/B = b/p$. It is known that the P/B coefficient is a biome-specific value; apparently, we can let it be a constant. Since $|q_{\mathrm{BA}}(\mathrm{H_2O})| = 4.8 \times 10^{19}$ g H_2O yr^{-1}, $B = 1.86 \times 10^{18}$ g d.w. and $P = 1.4 \times 10^{17}$ g d.w. yr^{-1}, then b=25.8 g H_2O g^{-1} d.w. per year, p=343 g H_2O g^{-1} d.w., and P/B=0.075 yr^{-1}. So, $S\cdot(\mathrm{H_2O}) = -0.221 \times 10^3 B$ J K^{-1} yr^{-1} or $S\cdot(\mathrm{H_2O}) = -2.93 \times 10^3 P$ J K^{-1} yr^{-1}.

Let us consider the entropic flow $\dot{S}_{\mathrm{SB}}$, which is proportional to area A_{veg} covered by vegetation. Since vegetation covers the globe by a relatively thin layer, then the equality $\dot{S}_{\mathrm{SB}} = aB$ or $\dot{S}_{\mathrm{SB}} = [a/(P/B)]P$ are rather plausible hypotheses. The value of a is easily found from eqn [8]: a=28.5 J K^{-1} per g d.w. per year.

The entropic flow $S_{\mathrm{B}}^{\mathrm{Work}} = [\beta s(\mathrm{DOM})]P = [\beta s(\mathrm{DOM})(P/B)]B$, so that $S_{\mathrm{B}}^{\mathrm{Work}} = 60.4\beta P = 4.53\beta B$.

Finally, eqn [11] is rewritten as

$$\begin{aligned} \mathrm{d}S_{\mathrm{B}}/\mathrm{d}t &\approx (-192 + 4.53\beta)B + \dot{S}^{\mathrm{Art}} \\ &\approx (-25.5 + 0.604\beta) \times 10^2 P + \dot{S}^{\mathrm{Art}} \end{aligned} \qquad [12]$$

This equation allows us to estimate different critical bounds of the impact of humankind on the biosphere. The impact may be manifested through: (a) increase in energy, $E = T_{\mathrm{B}}\dot{S}^{\mathrm{Art}}$ leads to decrease in the total biomass, B; (b) increase in energy inhibits the NPP, that is, $B = B(E)$, $\partial B/\partial E \leq 0$ and $P = P(E)$, $\partial P/\partial E \leq 0$. The simplest form of these functions may be linear, $B = B_{\mathrm{nat}}\left(1 - E/E_{\mathrm{crit}}^{\mathrm{B}}\right)$ and $P = P_{\mathrm{nat}}\left(1 - E/E_{\mathrm{crit}}^{\mathrm{P}}\right)$, where $B_{\mathrm{nat}} = 1.86 \times 10^{18}$ g and $P_{\mathrm{nat}} = 1.4 \times 10^{17}$ g are natural (without anthropogenic impact) values of biomass and NPP, $E_{\mathrm{crit}}^{\mathrm{B}}$ and $E_{\mathrm{crit}}^{\mathrm{P}}$ are critical values of energy with respect to the biomass and NPP (they vanish at these values).

The biota is living if $\mathrm{d}S_{\mathrm{B}}/\mathrm{d}t < 0$; therefore, the upper bound for human energy production, E^* is

$$E^* = \frac{E_{\mathrm{crit}}^{\mathrm{B}}\xi^{\mathrm{B}}}{E_{\mathrm{crit}}^{\mathrm{B}} + \xi^{\mathrm{B}}} = \frac{E_{\mathrm{crit}}^{\mathrm{P}}\xi^{\mathrm{B}}}{E_{\mathrm{crit}}^{\mathrm{B}} + \xi^{\mathrm{B}}} \qquad [13]$$

where $\xi^{\mathrm{B}} = T_{\mathrm{B}}(192 - 4.53\beta)B_{\mathrm{nat}}$, $\xi^{\mathrm{P}} = T_{\mathrm{B}}(25.5 - 0.604\beta) \times 10^2\ P_{\mathrm{nat}}$. In the previous section, we gave some meaningful interpretation to the parameter β, but here β is regarded as a free parameter.

Let us consider two simple examples.

Example 1: *The lower bound of β is $\beta_* = 29$, which is equivalent to full disappearance of the biosphere of vascular plants. From eqn [13] we get, for $\beta = 29$: $\xi^B \approx 3.2 \times 10^{22}$ J yr^{-1}. In order to estimate $E_{\mathrm{crit}}^{\mathrm{B}}$ we assume that this value is equal to the full enthalpy of biota, that is, $E_{\mathrm{crit}}^{\mathrm{B}} \approx 3.2 \times 10^{22}$ J yr^{-1}, then $E^* \approx (1/2)E_{\mathrm{crit}}^{\mathrm{B}} = 1.62 \times 10^{22}$ J yr^{-1}. Today humans are consuming about 3.24×10^{20} J annually. If humans would be doubling their energy consumption by every decade, then they would reach and exceed this bound during the next 70 years.*

Example 2: *The work performed annually by the biota is $W_{\mathrm{B}} = T_{\mathrm{B}}\dot{S}_{\mathrm{B}}^{\mathrm{Work}} = T_{\mathrm{B}}h_{\mathrm{DOM}}\beta P$. Since $P = P_{\mathrm{nat}}\left(1 - E/E_{\mathrm{crit}}^{\mathrm{P}}\right)$ where $P_{nat} = 1.4 \times 10^{17}$ g, then $W_{\mathrm{B}} = W_{\mathrm{B}}^{\mathrm{nat}}\left(1 - E/E_{\mathrm{crit}}^{\mathrm{P}}\right)$, where $W_{\mathrm{B}}^{\mathrm{nat}} = h_{\mathrm{DOM}}\beta P_{\mathrm{nat}} \approx 1 \times 10^{23}$ J yr^{-1} is the work of the 'natural' biota. Therefore, the relative work corresponding to the bound E^* is $W_{\mathrm{B}}^*/W_{\mathrm{B}}^{\mathrm{nat}} = E_{\mathrm{crit}}^{\mathrm{P}}/(E_{\mathrm{crit}}^{\mathrm{P}} + \xi^{\mathrm{P}})$. One of the possible estimations of $W_{\mathrm{B}}^*/W_{\mathrm{B}}^{\mathrm{nat}} \approx 0.95$, that is, only 5% of the potential work of the biosphere can be used to maintain its structure (in particular, animals) and its evolution; the rest is spent to turn the 'wheels' of the global biogeochemical cycles, so that $\xi^{\mathrm{P}}(\beta) = 0.0526E_{\mathrm{crit}}^{\mathrm{P}}$. By substituting this value into eqn [13], we get $E^* = 0.05E_{\mathrm{crit}}^{\mathrm{P}}$. We assume that $E_{\mathrm{crit}}^{\mathrm{P}}$ is equal to the total enthalpy of the NPP $\sim 2.44 \times 10^{21}$ J yr^{-1}, then $\mathrm{E}^* = 1.22 \times 10^{20}$ J yr^{-1}. By comparing this value with the current energy uptake, 3×10^{20} J yr^{-1}, we see that we already have serious problems today.*

Entropy Balance in Elementary Ecosystems

From the thermodynamic point of view, any ecosystem is an open system. An ecosystem being in a 'climax' state corresponds to a dynamic equilibrium, in which the internal production of entropy is balanced by the entropic outflow to the environment.

An 'elementary ecosystem' is the area unit of land, covered by some type of vegetation, which is properly the main part of any terrestrial ecosystem, and upper layer of soil with litter, in which DOM is decomposed. We neglect horizontal exchange flows of matter, energy, and entropy between this and other ecosystems.

The equation of energy balance for this area is $R = h_{evp} q_W + Q_{turb} + h_{DOM}\mathrm{GPP}$. Here $h_{evp} = 2462\ \mathrm{J\ g^{-1}}$ H_2O is the specific enthalpy of evaporation, q_W is the flow of evapotranspiration, Q is the turbulent heat flow, transporting heat from the surface into the atmosphere, $h_{DOM} = 17.4\ \mathrm{kJ\ g^{-1}}$ is the specific enthalpy of DOM, and GPP is the gross primary production (in g d.w.). Oxidation of biomass (respiration and decomposition of DOM) gives an additional source of heat, therefore the left side of the balance equation has to be $R + (Q_{met} + Q_{dec})$, where Q_{met} is a metabolic heat and Q_{dec} is heat releasing in the process of decomposition.

Let us group items of the radiative balance into two classes (in square brackets), which differ by values of their elements: $[R - h_{ev} q_W - Q_{turb}] + [Q_{met} + Q_{dec} - \mathrm{GPP}] = 0$, where the difference may constitute a few orders. For instance, the energy acting in the process of evapotranspiration is higher by two orders of magnitude than the energy of photosynthesis. Then we can equate each of the brackets to zero (it is the so-called 'asymptotic splitting': $[R - h_{ev} q_W - Q_{turb}] = 0$ and $[Q_{met} + Q_{dec} - \mathrm{GPP}] = 0$.

We assume that the fulfilment of the first equality provides the existence of some 'thermostat', which should be called the 'environment'. Then the fulfilment of the second equality is determined by a consistency of the processes of production, on the one hand, and metabolism of plants and decomposition of DOM in litter and soil, on the other.

In accordance with a standard definition, the internal production of entropy is equal to $d_iS/dt \approx Q_{ox}/T$, where T is the system temperature, and Q_{ox} is the heat generated by the system. The total heat production is a result of two processes: metabolism or respiration (Q_{met}) and decomposition of DOM (Q_{dec}). Since these processes can be considered as a burning of corresponding amount of organic matter, then the values of Q_{met} and Q_{dec} can be also expressed in enthalpy's units. Thus, $d_iS/dt \approx (Q_{met} + Q_{dec})/T$. The mean annual temperature at the surface of given site is the system temperature.

Since the equality $[Q_{met} + Q_{dec} - \mathrm{GPP}] = 0$ must hold, then $d_iS/dt \approx \mathrm{GPP}/T$. At the dynamic equilibrium the internal entropy production must be compensated by the entropy export from the system, so that

$$\frac{d_iS}{dt} = \frac{d_eS}{dt} = \frac{\mathrm{GPP}}{T} \qquad [14]$$

where $|d_eS/dt|$ is so-called 'entropy pump', which 'sucks' the redundant entropy (that is existing in the system for a long time), out of the ecosystem. We assume the local climatic, hydrological, soil, and other environmental conditions are adjusted in such a way that only one natural ecosystem corresponding specifically to these conditions can exist at this site and be in dynamic equilibrium. This is a concept of 'entropy pump'.

Any natural ecosystem is in dynamic equilibrium if and only if the internal entropy production within the system is balanced by an entropic outflow from the system to its environment (the 'entropy pump' is working). Suppose that additional inflows of artificial energy (energy load, W_{ae}) and chemical substances (chemical load, W_{ch}) start entering into the system. This is a typical impact of industry (or, in a broader sense, technological civilization) and industrialized agriculture on the environment. The internal production of entropy by the 'disturbed' ecosystem is given by

$$\frac{d_iS}{dt} = \frac{1}{T}[W + \mathrm{GPP}(W)] \qquad [15]$$

where $W = W_{ae} + W_{ch}$ is the total anthropogenic impact. Since a certain part of the entropy is released by the 'entropy pump' with power $|d_eS/dt| = \mathrm{GPP}_0/T$, where GPP_0 is the gross primary production of undisturbed 'wild' ecosystem located at a given point, then the total entropy balance is given by

$$\frac{dS}{dt} = \sigma = \frac{1}{T}[W + \mathrm{GPP}(W) - \mathrm{GPP}_0] \qquad [16]$$

Under the anthropogenic pressure, the system moves toward a new state, gaining the ability to perform some work, then it returns to the initial state, performing the work and producing the entropy. This is a typical two-time working cycle of a thermodynamic machine called an 'elementary ecosystem'.

If this system tends to some stable dynamic equilibrium with respect to W ($W_{eq} = W^*$) and, in addition, satisfies to Prigogine's theorem, then $\mathrm{GPP}(W^*) + W^* = \mathrm{GPP}_0'$ and $\mathrm{GPP}(W) + (W) \to \min_W$ at $W^* \neq 0$. Here GPP_0' is a new value of the power of 'entropy pump', corresponding to a new equilibrium, which is established in the process of succession from natural to 'anthropogenic' ecosystem. Unfortunately, the proper time of this transition is rather long, and often the transition is not successfully finished (e.g., the 'old field' succession recovers a structure of pre-anthropogenic natural ecosystem, and does it very fast).

As a rule, the decrease in entropy, obtained at the first stage of the cycle, does not compensate its increase at the second stage. The further destiny of this 'superfluous' entropy could be different: (1) it is accumulated by the system, the system (in particular, its environment) degrades, and after a while, dies; (2) entropy may be exported from the system, the initial state is reestablished, and the system is again ready for the next cycle. The latter strategy may be realized by means of an import of additional low-entropy energy that could be used for the system restoration: soil reclamation, pollution control, or generally speaking, ecological

technologies, etc. In other words, this refers to the so-called 'ecological management'. Using such entropy calculation, we can estimate the necessary investments (in energy units).

Unfortunately, there is a 'third alternative' to restore the initial state: to divide the system on two parts – a proper biological community and its abiotic environment, pumping over the superfluous entropy from one to another. In other words, we try to resolve the problem at the expense of environmental degradation. Note that the value of entropy excess σ could be used as a measure of the latter, or, as the entropy fee which has to be paid by society (actually suffering from the degradation of environment) for modern industrial technologies.

From the thermodynamic point of view, the environmental degradation leading to a decrease in the GPP is a typical system's reaction tending the internal entropy production to decrease (Prigogine's theorem and Le Chatelier's principle), while it may be considered as a disaster from the anthropocentric position. Thus, in order to avoid the anthropogenic disaster, we have to compensate the positive increment of entropy at each working cycle of this machine.

All these concepts are visibly illustrated in the case of agroecosystems.

Agricultural (Agro-) Ecosystems

What concerns agroecosystems, which are typical representatives in the class of anthropogenic ecosystems exploited by *Homo sapiens*, it is obvious that by increasing the input of artificial energy we increase their (agricultural) production. Note that the increase does not have an upper boundary and can continue infinitely. However, this is not the case, and there are certain limits, determined by the second law of thermodynamics. In other words, we pay the cost for increasing of agricultural productivity, which is a degradation of the physical environment, in particular, soil degradation. As an example, we shall analyze, as a case study, the maize production in Hungary of 1980s.

To start with, we apply the previous results to the case of agroecosystems. By taking into account that only some fraction of the GPP, $(1-k)(1-r)$GPP, participates in the local production of entropy, another fraction, $\gamma = k(1-r)$GPP, is exported from the system as a crop yield. Here r is the respiration coefficient and k is the fraction of biomass corresponding to the crop yield γ. Note also that the latter and the flow of artificial energy is usually bounded by some linear relation, $\gamma = \eta W$, where η is the so-called Pimentel's coefficient. Then instead of eqn [16] we write

$$\begin{aligned} \frac{dS}{dt} &= \sigma = \frac{1}{T}\left[\gamma\left(\frac{1}{\eta}+\frac{1}{s}-1\right) - \mathrm{GPP}_0\right] \\ &= \frac{1}{T}\left[W\left(1-\eta+\frac{\eta}{s}\right) - \mathrm{GPP}_0\right],\ s = k(1-r) \end{aligned} \qquad [17]$$

The agroecosystem will exist for an infinitely long time without degradation if the annual overproduction of entropy will be equal to zero ($\sigma = 0$). This is a typical situation of the local sustainability.

Therefore, eqn [17] under the condition $\sigma = 0$ gives us the value of 'limit energy load':

$$W_{\mathrm{sust}} = W_{\mathrm{sust}} = \frac{\mathrm{GPP}_0}{1-\eta+\eta/s} \qquad [18a]$$

which provides sustainability of the agroecosystem, if $W \le W_{\mathrm{sust}}$. Using another form of eqn [18] we get

$$\gamma_{\mathrm{sust}} = \frac{\mathrm{GPP}_0}{1/s+1/\eta-1} \qquad [18b]$$

This is an evaluation of some sustainable yield, that is, the maximal crop production, which could be obtained without a degradation of agroecosystem, in other words, in a sustainable manner.

In our case $W = 27$ GJ ha^{-1}, $\gamma = 4.9$ ton d.m. per hectare $= 73.5$ GJ ha^{-1}, $\eta = 2.7$, $r = 0.4$, $k = 0.5$, $s = 0.3$. It is natural to take the Hungarian steppe as a reference natural ecosystem with $\mathrm{GPP}_0 = 118$ GJ ha^{-1}. By substituting these values into [17] we get $\sigma T = 81$ GJ ha^{-1}; therefore, to compensate for the environmental degradation we must increase the energy input by three times, when two thirds of it is used only for soil reclamation, pollution control, etc., with no increase in the crop production.

Using eqns [18a] and [18b] we get $W_{\mathrm{sust}} = 16$ GJ ha^{-1} and $\gamma_{\mathrm{sust}} = 2.9$ ton d. m. per hectare. It is interesting that the first value is very close to different estimations of the 'limit energy load', 14–15 GJ ha^{-1}, derived from economical considerations or empirically. It is the maximal value of the total anthropogenic impact (including tillage, fertilization, irrigation, pest control, harvesting, grain transportation and drying, etc.) on 1 ha of agricultural land; and if the anthropogenic impact exceeds this limit, an agroecosystem is destroyed (soil acidification and erosion, chemical contamination, etc.).

As to the second value, let us now keep in mind that the contemporary maize yield in the USA is equal to 3 ton; and also after 'black storms of 1930s', the modern agricultural technologies allow us to avoid the strong soil erosion.

Entropy (more correctly the dissipative function, $\sigma_{\mathrm{er}}T$) corresponding to the destruction of one ton of soil in the Hungarian case is $\sigma_{\mathrm{er}}T = 2.54$ Gt ha^{-1}; then the annual loss of soil per one hectare is $\sigma T/\sigma_{\mathrm{er}}T \approx 32$. Therefore, the high maize production would cost us 32 ton of soil loss annually. It is obvious that the value of 32 ton per hectare is an extreme value: the actual losses are less, approximately 13–15 ton. This means that also other degradation processes take place, such as environmental pollution, soil acidification (the latter is very significant for Hungary), etc.

Myth of Sustainable Development

Thanks to the Brundtland Commission book *Our Common Future. From One Earth to One World*, the concept of sustainability has become rather 'fashionable' today. Unfortunately, the sustainable development runs counter the second law of thermodynamics. What kind of arguments could be used to prove this thesis?

Our technological civilization: (a) uses nonbiospheric, nonrenewable sources of energy (fossil fuels and nuclear energy); (b) applies technological processes, which increase concentrations of chemical elements in comparison with their concentrations in the biosphere (metallurgy, chemical industry, etc.); (c) disperses chemical elements decreasing their concentrations in comparison with their biotic concentrations. All these processes produce redundant entropy, which is not sucked out by the biosphere's entropy pump, which is tuned in natural conditions. Thus, degradation of the environment is the only way to compensate for the entropy overproduction. Of course, we can avoid the degradation by applying ecological technologies, but they are rather expensive. Therefore, another way is often used by TNC. What is this way?

Since the overproduction is spatially heterogeneous, the redundant entropy naturally overflows from one site with high entropy to others with lower entropy, or it is artificially transported. If in the first case the process manifests as spreading of different pollutants by natural agents (wind, rivers, etc.), then in the second case this is either a purposeful export of industrial waste and polluting technologies to other regions, or import of low-entropy energy (e.g., fossil fuels) from other regions. Finally, we formulate the following thesis: sustainable development is possible only locally, in selective areas of the planet, and only at the expense of creating 'entropy dumps' elsewhere.

Note that in order to 'save' the sustainability concept, in the sustainability literature one talks about the so-called 'strong' sustainability, which is impossible due to the second law, and then 'weak' sustainability where losses are replaced by other gains. For instance, our technological civilization is generally using nonrenewable energy resources and materials that inevitably will lead to a loss of sustainability, but if we develop new technologies based on renewable sources of energy and materials, we are still doing well with respect to weak sustainability. However, in this case we shall deal with some slow movement of the biosphere from its contemporary equilibrium to some new unknown one. Certainly, the equilibrium might either be more suitable and comfortable for *Homo sapiens*, or might not be – that we do not know. There is one more rock in this slow movement: the small changes are accumulated without some visible effect, but sooner or later it could result in a disaster. This behavior is typical for nonlinear system such as the biosphere.

Conclusion

The author would like to complete the article by quoting the British physicist Robert Emden:

> When I have been a student, I have read with pleasure F. Wald's small book under the title "The Queen of the World and her Shadow". Energy and entropy were kept in mind. Now, when I understand these concepts deeper, I think that their positions should be interchanged. In the giant factory of natural processes, the entropy law is a director who controls and manages all the business, while the energy conservation law is only an accountant who is keeping a balance between debit and credit. (Emden, 1938).

See also: Ecological Complexity: Thermodynamics in Ecology. Global Change Ecology: Energy Flows in the Biosphere. Human Ecology and Sustainability: Emergy and Sustainability

Further Reading

Aoki, I., 1995. Entropy production in living systems: From organisms to ecosystems. Thermochimica Acta 250, 359–370.
Ebeling, W., Engel, A., Feistel, R., 1990. In: Physik der Evolutionsprozesse. Berlin: Akademie Verlag, p. 374.
Jørgensen, S.E., Svirezhev, Yu M., 2004. In: Towards a Thermodynamic Theory for Ecological Systems. Amsterdam: Elsevier, p. 366.
Kleidon, A., Lorenz, R.D. (Eds.), 2005. Non-Equilibrium Thermodynamics and the Production of Entropy. Life, Earth, and Beyond, Series: Understanding Complex Systems, XIX. New York: Springer, p. 260.
Morowitz, H.J., 1970. In: Entropy for Biologists: An Introduction to Thermodynamics. New York: Academic Press, p. 195pp.
Morowitz, H.J., 1978. Foundations of Bioenergetics. New York: Academic Press.
Svirezhev, Yu M., 2005. Application of thermodynamic indices to agro-ecosystems. In: Jørgensen, S.-E., Costanza, R., Xu, F.-L. (Eds.), Handbook of Ecological Indicators for Assessment of Ecosystem Health. New York: CRS, Lewis Publishers, pp. 249–277.

Environmental and Biospheric Impacts of Nuclear War

P Carl, Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany
Y Svirezhev†, Potsdam Institute for Climate Impact Research, Potsdam, Germany
G Stenchikov, Rutgers University, New Brunswick, NJ, USA

Introduction

Apprehensions about uncontrolled thermonuclear fusion arose early in the Manhattan Project: Could explosion of a hydrogen bomb trigger a global physical catastrophe in starting chain reactions that seize the light elements and thus wipe out all life on Earth? It was not without grave anxiety that Emil Konopinski, Cloyd Marvin, Jr., and Gregory Breit ruled out this possibility. Since the end of World War II, public knowledge about `the unthinkable', the consequences of nuclear war, was largely shaped by the horrible direct health effects of the 1945 atomic bombing of Hiroshima and Nagasaki, by the devastations and disruptions of life and infrastructure due to heat, blast and electromagnetic waves, prompt ionizing radiation, and radioactive fallout.

Public awareness of worldwide fallout risks was triggered by the fatal outcome of the atmospheric test Bravo, the largest weapon exploded by the United States (Bikini atoll, 28 February 1954; 15 Mt TNT equivalent). Temporary geophysical effects of atmospheric tests, such as planetary pressure waves, magnetic field distortions, or ionospheric disruptions causing blackout in radio communication, were frequently recorded. A signature in worldwide weather was not found, however. The largest weapon ever tested, the `Tsar of Bombs', a more than 50 Mt `clean' bomb (with a nonfissionable mantle), was exploded on 30 October 1961, above the northern Soviet test site at Novaja Zemlya. Its pressure wave circled the Earth several times. The most obvious long-term, large-scale direct geophysical effect of nuclear explosions appears to have been caused by the Starfish Prime test on 9 July 1962 – a 1.4 Mt detonation some 400 km above the Johnston Island area in the tropical central North Pacific. An artificial (mini-van Allen radiation) belt of charged particles was trapped by the Earth's magnetic field and traced for a couple of years. High-altitude nuclear explosions may `blind' reconnaissance satellites, impair electronics over vast areas, and even inflict on a missile attack, by their electromagnetic pulse (EMP).

Radioactive tracers from the 539 atmospheric test explosions until 1980, with an aggregate yield of about 440 Mt, will remain identifiable worldwide for millennia. Hot spots at test sites and the unresolved issue of low-dose radiation effects notwithstanding, though, their health effects are far from endangering the species of man. In a nuclear war, the number of warheads and their total explosive yield might exceed these figures by an order of magnitude, and the period of 35 years would reduce to a couple of days, if not hours. Such a 10^5-fold `compaction' of interacting effects rules out extrapolation from the test series, as do consequences of a decisive distinction in targeting: The deadly logic of `mutual assured destruction' (MAD) bears attacks on large population centers. Cities would also not escape `countervalue' and `counterforce' strikes against the economic and military potential, notably the command, control, communication, and intelligence (C^3I) structures. Not only does this turn `warfare' into `exchange', it also gives birth to a new quality of risks – the long-term, worldwide indirect aftereffects that add to and interfere with the disastrous direct effects of nuclear explosions.

Studies on Indirect Effects of Nuclear War

Multiple upper-atmosphere nuclear bursts might be scheduled for a ballistic missile defense (BMD) system's `terminal phase defense' under the threat of a `decapitation' strike. This would endanger the Earth's `ozone screen' in generating high amounts of nitrogen oxides (NO_x). Rising fireballs of tropospheric bursts would have the same effect. The biologically active part of the solar ultraviolet radiation (UV-B), which causes structural change in amino acids and is normally absorbed by stratospheric ozone molecules, would then reach the surface. According to the US National Academy of Sciences (NAS; 1975), a 10 000 Mt exchange could cause a hemispheric ozone loss by 30–70% that would extend globally, with potentially grave impacts on terrestrial and aquatic ecosystems, and recovery over years. The amount of stratospheric dust, injected by megaton-yield near-surface explosions, may resemble the aerosol load due to the Krakatau eruption in 1883. NAS thus took a surface cooling of only a few tenths of a degree centigrade for plausible, but expressed another concern: ``It is not known whether climatic variables have stable equilibrium values and a tendency to relax after an impulse disturbance such as that generated by a nuclear exchange.'' This led the authors to conclude that irreversible climatic shifts cannot be ruled out.

Criticism by the Federation of American Scientists (FAS) for too `optimistic' NAS conclusions about potential impacts on remote, noncombatant countries became justified 7 years later. Invited to contribute on ozone impacts to a first international, comprehensive nuclear war risk study commissioned by the Royal Swedish Academy of Sciences and first published in its environmental journal *AMBIO* (1982), Paul Crutzen and John Birks were surprised to find unknown, severe effects due to the

†We are mournful about the loss of our friend and colleague Yuri Mikhailovich Svirezhev who passed away during the time of working on this paper. We dedicate our own contribution to his memory. (GS & PC)

 doi:10.1016/B978-0-444-63768-0.00611-9

smoke from `postnuclear' fires. A key mechanism like this was missing since the early speculations on changes of weather and climate due to nuclear war.

To grip consequences of the changing military policy away from `assured destruction', the US Office of Technology Assessment (OTA; 1979) had just analyzed a range of scenarios in an influential war risk study, from attacks on cities and oil refineries to one-sided counterforce and countervalue strikes. High-altitude bursts were mentioned as critical but not addressed, though doubts were cast on massive stratospheric ozone depletion. The chemical system was better known then, and high-yield weapons had given way to missiles with multiple warheads of lower yield each. Focusing on direct effects of nuclear attacks, civil defense, economic breakdown, recovery and societal impacts, OTA suggests that extreme uncertainties, and certainty about disastrous `minimum' consequences, both ``play a role in the deterrent effect of nuclear weapons''.

Other than OTA, the *AMBIO* study used a global reference exchange that comprises ground and tropospheric bursts with a total yield of about 5750 Mt. Crutzen and Birks confirmed both the NAS estimates of ozone depletion and the OTA doubts, that is, their advanced ozone model did not qualitatively alter the results, but the evolving strategic arsenals apparently did. In addition, they identified large-scale forest fires and intense urban and industrial conflagrations as sources of a long-lasting photochemical smog over large areas of the Northern Hemisphere. The sunlight needed for its formation, however, could be blocked by high amounts of smoke, notably when oil and gas fields or refineries were targeted. Smoke absorbs the short-wave solar flux and heats up, but interferes much less with the outgoing long-wave thermal radiation. The surface cools therefore, which would also suppress convection and reduce precipitation. Cold and darkness after a nuclear war would be severe for terrestrial ecosystems, but especially grave for oceanic food chains given the quick consumption of phytoplankton at the very base of the trophic web.

The anticipated climatic disruption triggered inquiry by Richard Turco, Owen Toon, Thomas Ackerman, James Pollack, and Carl Sagan (TTAPS), who were able to marshal the data for urban mass fires and fire storms not available in time to Crutzen and Birks, and to quantify the effect they coined `nuclear winter' for a broad range of scenarios (100 Mt `countercity', 5000 Mt `baseline', 10 000 Mt full-scale exchange, for example): a massive thermal inversion of the planetary atmosphere and its climatic consequences. A line of research which turned out to be essential addresses mass extinctions in Earth history, including hypotheses of extraterrestrial impacts and those that blame geological periods of enhanced volcanism or worldwide forest fires, maybe even all in combination. An authoritative circle of biologists and ecologists, led by a group including Paul Ehrlich, John Harte, Mark Harwell, Peter Raven, and George Woodwell, concluded that the extinction of man after a large nuclear war could no longer be ruled out. A public conference ``The World after Nuclear War'' (Washington, 31 October–1 November 1983) attracted unprecedented attention by communicating these findings, by the participation of scientists of both superpowers, and by a technique used on this occasion: a satellite TV bridge (Moscow–Washington) between both academies of sciences. Two research groups from either side of the globe exchanged their first `nuclear winter' results, based on global climate models, via this public `Moscow link:' Vladimir Aleksandrov and Georgiy Stenchikov, and Curt Covey, Steven Schneider and Starley Thompson.

Rethinking the Unthinkable

An unparalleled, open activity toward a worldwide process of research and education, which goes beyond the traditional understanding of scientific responsibility in terms of specialist's denial, the public action of the scientific elite, and the work in closed circles, was launched after delivery of the *AMBIO* study by the International Council of Scientific Unions (ICSU). Steered by the project ``Environmental Consequences of Nuclear War'' (ENUWAR) of the Scientific Committee on Problems of the Environment (SCOPE), resulted it by the end of 1985 in the two-volume report SCOPE-28. This study cannot offer an overall view on the complex entity `ecosphere' after nuclear war. None of the major physical control parameters (light, temperature, water) are expected to be disturbed that hard or weak at planetary scale so as to justify `simple' conclusions. A substantial impact on the stratospheric ozone budget, however, maintained by the US National Research Council (NRC; 1985), holds the more, since smoke-induced heating would change all chemical reaction rates.

Long-term problems due to radioactive exposure are borne in the selective vulnerability of forest communities. Nuclear war might change the biogeography of vast areas, notably where coniferous temperate forests dominate and the `radiation shock' combines with large-scale fire, climatic change, UV-B effects, etc. SCOPE-28 notes specific sensitivities: ``Temperature effects would be dominant for terrestrial ecosystems in the Northern Hemisphere and in the Tropics and subtropics; light reductions would be most important for oceanic ecosystems; precipitation effects would be more important to grasslands and to many Southern Hemisphere ecosystems.'' Such a `distributed vulnerability' structure and its more subtle patterns, the other side of the `biogeography' medal, reflects a range of stabilizing feedbacks against gradual or abrupt transitions in atmospheric or oceanic conditions and related compositional, thermal, meteorological, and hydrological regimes. The `acute' phase of nuclear winter would bear structural changes in the physical environment of a dimension that might transgress those stability limits, maybe at hemispheric scale. The crucial question is the one posed by the NAS 10 years before: Will new feedbacks take over to stabilize the system in a regime different from present day, or will it return? Latest, the transition to `chronic' response would be influenced by climate–biosphere feedbacks. The result of acute-phase environmental devastations, like the patchiness of surviving communities, may thus attain a structural role in shaping a postwar environment.

The vulnerability of the `noosphere' – taken as the complex entity of man's society and managed environment – against the direct effects of nuclear war is much higher than that of the (natural) ecosphere. Agricultural systems may only exist due to human maintenance. Worldwide disruption of functioning agriculture and food supply after nuclear war would expose the majority of

survivors to the risk of starvation. The OTA study suggests that this would even hold without severe environmental aftereffects. Climatic impacts that hit agriculture at vulnerable spots (length of the growing season, hydrological change, etc.) may bear just that sort of feedback, however, which keeps surviving humans stuck to marginal subsistence for any period relevant to societal restoration. The impact on the Southern Hemisphere of nuclear war in the north is a key issue in view of a postwar noosphere. Beyond the `import' of climatic effects due to interhemispheric smoke transport, with their ecological and agricultural consequences, a major disturbance would be caused by interruption of the lifeline of international trade, even if agricultural productivity could be maintained at a level of sufficiency. The risk consists in a large societal setback to which a modern society may not adapt without existential disturbance.

A convincing approach to environmental impacts addresses productivity limits and `convolves' this knowledge with a realistic range of stresses derived from climate model output. The resulting `response surfaces' to stress factors as expected after nuclear war (changes in temperature, light level, precipitation) are not of a simple shape for grassland ecosystems, notably when secondary productivity of herbivores is considered. Regulatory feedbacks act together, that is, stresses are not generally additive or even mutually enhancing (`synergistic'). The `nonsevere factor space' of functioning ecosystem response may have rather sharp boundaries, however, beyond which the `message' becomes simple: ecosystem productivity reduces almost abruptly to a level that would not support a human population. An important conclusion is on uncertainty again: a group of survivors who found an ecological niche in a postwar environment may be ``plunged into destruction by seemingly minor drift away from those conditions.'' Ecosystems become unpredictable when driven to marginal existence, in the vicinity of critical transition, by a changing climate.

In two authoritative assessments, the World Meteorological Organization (WMO) confirmed the risk of a severe smoke-induced climatic impact. Among the pertaining uncertainties, a potential modification of the hydrological cycle was emphasized. This concerns the `Hadley circulation' above all, a double-cell of upward and poleward circulation (in the zonal mean) flanking the meteorological equator, which is driven by the strongest heating there. As the season advances, the southern `Hadley cell' shifts northward and blows up to form the `monsoon cell' in boreal summer, with upward legs as far north as the Tibetan and Mexican plateaus. A smoke veil above would attenuate this structure, that is, weaken or disrupt the monsoons, but depending on season, injection heights, and location of the smoke source, the Hadley circulation may also become enhanced. The WMO assessments posed into doubt that climate models may reach the required realism and reliability soon, but confirmed both a potential monsoon disruption in boreal summer and smoke lofting as well as transport into the Southern Hemisphere.

A minimum demand, the realistic simulation of seasonally varying rainfall, is not easily met by global climate models. Even more challenging is the agriculturally important intraseasonal activity, notably the active–break cycles of the major monsoon branches and their dynamic interplay. Key knowledge about monsoon dynamics, and thus about the atmospheric hydrological cycle and its interactions, did just settle when the scientific consensus formed about major climatic effects of nuclear war. This bears potential for surprise, and the consensus may deserve further development just where it directly concerns half the world population. Abrupt onset and retreat of the boreal summer monsoon are known for decades, its oscillatory (interhemispheric) nature since the mid-1970s. These are features typical of a dynamic system that passes a critical transition. Long-term consequences for man and the biosphere of the climatic response to nuclear war may thus be borne in the potential for structural recovery of the present-day `monsoon climate' on Earth. This includes monsoon interactions with the El Niño–Southern Oscillation (ENSO) system. Both dynamic subsystems mediate climatic and environmental impacts on the Southern Hemisphere, beyond the more direct effects of interhemispheric smoke transport.

SCOPE successor studies to ENUWAR (1982–88) took another turn: the projects RADPATH (1988–93) and RADTEST (1993–99) addressed the pathways of radionuclides across the environment, exemplified by field studies into the consequences of the 1986 Chernobyl nuclear reactor accident and at selected nuclear test sites worldwide. These projects mounted a substantial database, improved the knowledge of processes, and identified gaps in understanding the biogeochemical dispersion of radionuclides. All three projects were led by Sir Frederick Warner. Their documented results, SCOPE report nos. 28, 50, and 59, are reference sources for the state of scientific knowledge toward the turn of the twentieth century about the gravest risk and challenge of its second half – the `doomsday' of man in an all-out nuclear conflict.

Behind and Beyond the Scenarios

The idea of a `doomsday machine' as an *ultima ratio* of deterrence is due to Herman Kahn. As a `terminal' retaliation should deterrence fail, such a hypothetical device was thought to automatically kill the majority of mankind, if not the species of man or all life on Earth. MAD was a sort of `homicide pact' indeed, settled by the Antiballistic Missile Treaty of 1972 (which allowed one BMD system at either side). Negotiations could give MAD a frame as long as it was accepted as a matter of fact and a relatively stable island was sought within the sea of inherent risks. In a severe crisis, however, a strategic exchange could have been initiated just by technical failure, misinterpretation, false information, or madness. Aimed to balance Soviet conventional forces, the US nuclear guarantee for Western Europe established the principal context of the doctrine of extended deterrence. The ability to control escalation, a prerequisite of this posture, was its dilemma as well. The myth, the adversaries in a nuclear war could climb a fictitious `escalation ladder' up and down at will, is not backed by any realistic view on the dynamics of escalation, be it only due to the vulnerability of the very means of control, the C^3I systems, which are primary targets in the earliest phase of war. Moreover, tactical nuclear weapons are an escalation-prone arming *per se*. Their massive deployment along the European front made an early,

uncontrolled use in any armed conflict nearly certain. Postures other than MAD were also delusionary due to the `third power problem': the nuclear forces of Britain and France, maintained in part in recognition of the US dilemma with extended deterrence, were `MAD forces' by intention, with a substantial destruction potential. `Escalation control' and `limited nuclear war' were sold by a `nuclear utilization theory' (NUT) as alternatives to MAD. Soviet strategic forces in `launch on warning' alert, however, and a doctrine of earliest possible, massive infliction of (not just response to) any nuclear attack would have left no space for bargaining after crossing the threshold to war. NUT did not replace MAD, but increased the risk of strategic instability.

When Herman Kahn died on 3 July 1983, a revision of his classic *Thinking about the Unthinkable* had been caught up with `nuclear winter'. In a comment, the editors admit strategic consequences, excepting the `war fighting' postures. A similar view was held in a brief report delivered by the US Secretary of Defense, Caspar Weinberger. It focuses on the early TTAPS study and uncertainties discussed there, cites with the same bias the reasoning of the NRC study, `massacres' Soviet contributions as `propaganda', and praises escalation control as one of the means to avoid nuclear winter. That resistance of the military bureaucracy to new knowledge drives the `overkill' arsenals beyond any justification shows also the example of the Pacific-Sierra Research Corporation, where smoke emissions and fire effects have been studied with a primary view on target planning for nuclear war: the `blast model' of casualty estimation survived any `fiery' challenge. For a 10 000 Mt war, the World Health Organization (WHO) estimated a short-term toll of 2.2–2.5 billion casualties, with a ratio of deaths to injured from 1.1 to 1.6. Lacking appropriate medical care, many of the injured would be doomed to die. Immediate casualty estimates of the Greater London Area War Risk Study (GLAWARS; 1986), the most comprehensive public assessment of the impact of nuclear war on a region, range from 1 to 6.2 millions (97%) of the London population. At a symposium at the NAS Institute of Medicine (IOM; September 1985) such estimates were challenged by a new model that takes `postnuclear' fires into account. Immediate fatalities had been underestimated by a factor of 2 or more when `only' prompt radiation, heat and blast waves, as well as local radioactive fallout were considered (blast model). The lower-edge figures for London increase substantially when using the `conflagration model'.

Difficulties in `translating' climatic into health effects are partly due to missing local information, neither provided by climate models nor easily derived: fog or haze, storminess, chemical and radioactive load of precipitation, etc. For the longer term, GLAWARS' gravest concern is food supply for survivors. Genuine medical aspects include enteric diseases and those spread by insects or due to poor sanitation and nutrition, all favored in victims who became `immunocompromised'. The key point here, also identified at the IOM symposium, is just the combined action of stresses in the nuclear aftermath to impair the immune system. Factors causing immune suppression include radioactive and UV-B radiation, malnutrition, burns and trauma, as well as psychosocial stress. Clinical evidence indicates that these factors all converge in their action on a single element of the immune system, the T-lymphocyte, of which also the `helper-to-suppressor ratio' is crucial. The Acquired Immune Deficiency Syndrome (AIDS) is characterized by deficiencies of the T-lymphocyte variety similar to those expected due to the combined stresses after nuclear war. The list of factors is certainly not exhaustive. The `clinical record' of today's monsoon and ENSO variability, from lasting hot-dry to torrential flooding, should bear medical implications of structural impacts on the tropic–subtropical climate. Coming to grips with these dynamic systems challenges climate modeling today, as did a smoky atmosphere in the 1980s.

`Nuclear Winter' Modeling – A Sketch

To follow solar and thermal radiations through a smoke- and dust-laden atmosphere, TTAPS had used a height-resolved (one-dimensional; 1-D) radiative–convective model (RCM) with annual mean insolation. An RCM describes these processes in greater detail than general circulation models (GCMs) do but misses their horizontal motions. Further 1-D and 2-D studies, at the Lawrence Livermore National Laboratory (LLNL), the US National Aeronautics and Space Administration (NASA), and the University of Maryland, helped clarifying basic effects and feedbacks including smoke uplift and changes in the snow-ice albedo, both of which may protract climatic effects. Like Covey, Schneider, and Thompson of the US National Center for Atmospheric Research (NCAR), who used a GCM of Australian origin with 7 atmospheric layers, Aleksandrov and Stenchikov of the Moscow Computing Center of the USSR Academy of Sciences (CCAS) confirmed TTAPS' major results. They used a coarse-resolution two-layer tropospheric GCM that had been adapted for different purposes in cooperation with Lawrence Gates of the Oregon State University (OSU), and equipped it with a simple ocean model. In addition to severe surface air temperature drops in continental interiors (mitigated near oceanic coasts) and large-scale thermal inversions of the atmosphere, clear signs of interhemispheric smoke transport due to a structural response of the Hadley circulation were noted by both groups.

These early 3-D `nuclear winter' studies were admittedly quick shots: their immobile smoke stayed uninfluenced by atmospheric motions, did not interact with the hydrological cycle to become washed out, and could not buoyantly rise by solar heating. Though state-of-the-art in the early 1980s, artificial model climates had also to be left behind for more realistic assessments. Michael MacCracken and John Walton of the LLNL and the CCAS team introduced more realistic feedbacks into their two-layer GCMs, whereas the NCAR group focused on the model `physics' first to keep firm footing. A visit in Moscow, coincidentally just before the 1983 Washington conference, triggered a study series by Stenchikov and Carl that addressed a `minimum' disturbance (without minimizing the problem), traced conditions for Southern Hemisphere impacts, and explored the transient response for hints to answer the 1975 NAS question on `postnuclear' climate relaxation. This induced a closer view on the complexity of the acute phase of perturbation and its implications. Just during startup of this common work in Berlin, on 31 March 1985, Vladimir Aleksandrov vanished without a trace in Madrid. The shock and irritation about his disappearance and fate (which made him even

an `unperson' for a couple of months) drove the first of those studies into an unexpected tension field. Nevertheless, it helped to overcome the Soviet `hard scenario' attitude and to tear down a barrier to public information at the German east side of the Iron Curtain.

The most detailed results were due to Thomas Malone and co-workers of the Los Alamos National Laboratory (LANL), who extended the NCAR GCM in the vertical to address the smoke transport more precisely. They confirmed the expected lofting into the lower stratosphere and thus a much prolonged residence and forcing. Though the NCAR authors fixed the important issue of `quick freeze' beneath smoke clouds, notably in the subtropics and Tropics at startup of interhemispheric transport, a controversy arose from their inquiries suggesting change of the popular metaphor into `nuclear fall'. Until 1987, persistent efforts to deblur longer-term effects due to the oceanic response have only been undertaken at the CCAS. In a more realistic atmosphere–ocean GCM study, virtually the last `nuclear winter' publication for 15 years, Steve Ghan confirms Alan Robock's (1-D) finding that the acute-phase ocean and sea-ice response may bear climatic impacts for years.

Regional Conflicts and Their Global Effects

The `doomsday scenario', executed by retreating Iraqi troops in February 1991 in setting the Kuwaiti oil fields alight, was meant as a modern version of Kahn's ultimate deterrent – an idea that failed. Two climate modeling responses, from the United Kingdom Meteorological Office and the Max Planck Institute for Meteorology, denied an attenuating impact on the Indian monsoon (a concern that had been expressed before). Successors of their GCMs did correctly represent the major Asian rainbelts as part of a planetary system, and their seasonal migration, but not the seasonal mean distribution, to say nothing about intraseasonal activity. Just those 30–60-day active–break monsoon cycles, including realistic motions of the major Asian systems, were now found in the Berlin version (`CCAS-B') of the CCAS GCM in an own Kuwait oil well fire study. This GCM version is a completely regenerated, flexible tool of dynamic systems analysis. Its boreal summer monsoons turned out indeed to be part of an interhemispheric, oscillatory seasonal climate regime between critical transitions in June and September. The Kuwait oil fire smoke caused a regional lower-troposphere heating anomaly, and thus an `exciting' disturbance that fanned the GCM's dynamics in a way not dissimilar to the observed 1991 season. Such a type of monsoon climate may thus be the `playing ground' for martial adventures seizing the source regions of the atmospheric water cycle. Its structural robustness is unknown.

The theme is again put on the agenda by recent studies into the climatic effects of a potential regional nuclear conflict of 1.5 Mt `size' in Southeast Asia, using the full atmosphere–ocean GCM with high vertical resolution of NASA's Goddard Institute of Space Studies (GISS), which has been successfully applied to study the climatic impact of volcanic eruptions. The GISS model shows extremely long smoke residence times, up to a decade, due to efficient lofting into the upper stratosphere, all year round in these latitudes. The surface air temperature drop is much less than in the `nuclear winter' case, of course, but still considerable if compared with the climate record: a global cooling from 1.25 to 0.5 K over a decade, with minima of several kelvin (degrees centigrade) over large areas of North America and Eurasia. A 10% weakening of the global precipitation is concentrated in the Tropics, but substantial (seasonal mean) reductions of the Asian subtropical summer monsoons are also found, with potentially serious human impacts.

Final Remarks

The scientific consensus as settled on the pages of SCOPE-28 was a snapshot taken from a dynamic research process. Shortly after the second edition, the Cold War ended, and modeling the climatic response to massive smoke injections was terminated just when it had reached more firm grounds. The theme was picked up not before another recent revision using the GISS model. Questions like that of the 1975 NAS study on climate relaxation remain unanswered as yet. Summarizing the status of the smoke source term discussion in their last common paper, though, TTAPS had shown that figures which were finally used in the climate model studies of the 1980s remained in the vicinity of earlier assumptions – a consequence of mutually balancing changes in detail. It has been learnt, for example, that smoke consists of fractal aggregates which have little in common with the earlier picture of largely spherical objects. This reduces the rate at which their short-wave absorptivity decreases and prolongs the direct radiative forcing of climatic effects. The debate about nuclear `winter' or `fall' occupied the community but did not fundamentally change the perspective as well. A detailed study of atmospheric coastal flow fields did not even confirm a mitigating oceanic impact on the surface air temperature drop over land.

We do not mirror and discuss the points here that have been made with due justification concerning the political response to nuclear winter. Science itself is the addressee of a disturbing question: Was there a potential to substantially influence public and strategic thinking by timely, deliberate inquiry? A `doomsday potential' was inherent to nuclear deterrence since the 1960s, at the latest, and it may be questioned that the `policy war' between MAD and NUT was predicated to end at the terms of the war-fighting strategists. Game theory was abused to justify NUT, risky nuclear weapons tests were conducted, and the Cuban missile crisis made humankind totter at the brink of its ultimate catastrophe. Though lately a result of the arms race, MAD was a vulnerable and immoral posture. Remarkable activities of the 1960s notwithstanding, though, did nuclear deterrence and nuclear war become great themes for the general scientific community only during the 1980s. A largely unmonitored evolution toward `wars of the

twenty-first century' is likewise a risky habit. It has `tradition' in military politics to occupy gray zones of knowledge, and in scientific `surveillance' to lag behind the arsenals and strategies of war.

Further Reading

Ball, D., 1981. Can nuclear war be controlled? Adelphi Paper No. 169 London: International Institute for Strategic Studies, p. 51.

Bergström, S., Bochkov, N.P., Leaf, A., *et al.*, 1987. Effects of nuclear war on health and health services. 2nd edn. Geneva: World Health Organization, p. 179. *Report A40/11.*

Carl, P., Worbs, K.D., Tschentscher, I., 1995. On a dynamic systems approach to atmospheric model intercomparison.Report of the World Climate Research Programme (WCRP–92), WMO/TD-No. 732 Geneva: World Climate Research Programme, pp. 445–450.

Carrier, G.F., Moran, W.J., Birks, J.W., *et al.*, 1985. The Effects on the Atmosphere of a Major Nuclear Exchange. Washington, DC: National Research Council, p. 193.

Ehrlich, A., Gunn, S.W., Horner, J.S., *et al.*, 1986. London under Attack. Oxford: Basil Blackwell, p. 397.

Ehrlich, P.R., Sagan, C., Kennedy, D., Roberts, W.O. (Eds.), 1984. The Cold and the Dark: The World after Nuclear War. New York: Norton & Co, p. 229.

Gadgil, S., Sajani, S., 1998. Monsoon precipitation in AMIP runs.Report of the World Climate Research Programme (WCRP-100), WMO/TD-No. 837 Geneva: World Climate Research Programme, p. 86.

Golitsyn, G.S., MacCracken, M.C., 1987. Atmospheric and climatic consequences of a major nuclear war: Results of recent research.Report of the World Climate Research Programme (WCP-142), WMO/TD-No. 201 Geneva: World Climate Research Programme.

Harwell, M.A., Hutchinson, T.C., Cropper Jr., W.P., Harwell, C.C., Grover, H.D., 1985. SCOPE 28 – Environmental Consequences of Nuclear War, Vol. 2: Ecological and Agricultural Effects. Chichester, UK: Wiley, p. 523. (2nd edn. with a 31pp. updating preface, 1989).

Johns, L.S., Sharfman, P., Medalia, J., *et al.*, 2005. The Effects of Nuclear War. Washington, DC: Congress of the U.S., Office of Technology Assessment, p. 151.

Kahn, H., 1984. Thinking about the Unthinkable in the 1980s. New York: Simon & Schuster, p. 250.

McNaughton, S.J., Ruess, R.W., Coughenour, M.B., 1986. Ecological consequences of nuclear war. Nature 321, 483–487.

Nier, A.O.C., Friend, J.P., Hempelmann, L.H., *et al.*, 1975. Long-Term Worldwide Effects of Multiple Nuclear-Weapons Detonations. Washington, DC: National Academy of Sciences, p. 213.

Nuclear War: The Aftermath. In: Peterson, J., Hinrichsen, D. (Eds.), AMBIO. Oxford: Pergamon, p. 196. 11(2/3).

Pittock, A.B., Ackerman, T.P., Crutzen, P.J., *et al.*, 1986. SCOPE 28 – Environmental Consequences of Nuclear War, Vol. 1: Physical and Atmospheric Effects. Chichester, UK: Wiley, p. 359. (2nd edn. with a 36pp. updating preface, 1989).

Robock, A., Oman, L., Stenchikov, G.L., 2007. Nuclear winter revisited with a modern climate model and current nuclear arsenals: Still catastrophic consequences. Journal of Geophysical Research 112, D13107. doi:10.1029/2006JD008235.

Sagan, C., Turco, R., 1990. A Path Where no Man Thought: Nuclear Winter and the End of the Arms Race. New York: Random House, p. 499.

Solomon, F., Marston, R.Q. (Eds.), 1986. The Medical Implications of Nuclear War. Washington, DC: Institute of Medicine, National Academy of Sciences, p. 619.

Stenchikov, G.L., 1985. Climatic consequences of nuclear war. In: Velikhov Ye, P. (Ed.), The Night After Climatic and Biological Consequences of a Nuclear War. Moscow: Mir Publishers, pp. 53–82. (Russian edn.: Nauka, Moscow 1986).

Svirezhev, Ju M., Carl, P., *et al.*, 1990. Götterdämmerung. Globale Folgen eines atomaren Konflikts. (substantially revised and extended German edn. of Svirezhev YM, Alexandrov GA, Arkhipov PI, *et al.* (1985) *Ecological and Demographic Consequences of Nuclear War*, 267pp. Moscow: USSR Academy of Sciences, Computer Center) Berlin: Akademie-Verlag, p. 261.

Thompson, S.L., Aleksandrov, V.V., Stenchikov, G.L., *et al.*, 1984. Global climatic consequences of nuclear war: Simulations with three dimensional models. AMBIO 13, 236–243.

Toon, O.B., Robock, A., Turco, R.P., *et al.*, 2007. Consequences of regional-scale nuclear conflicts. Science 315, 1224–1225.

Turco, R.P., Toon, O.B., Ackerman, T.P., Pollack, J.B., Sagan, C., 1990. Climate and smoke: An appraisal of nuclear winter. Science 247, 166–176.

Gaia Hypothesis

PJ Boston, New Mexico Institute of Mining and Technology, Socorro, NM, USA

Introduction

The Gaia hypothesis, named after the ancient Greek goddess of Earth, posits that Earth and its biological systems behave as a huge single entity. This entity has closely controlled self-regulatory negative feedback loops that keep the conditions on the planet within boundaries that are favorable to life. Introduced in the early 1970s, the idea was conceived by chemist and inventor James E. Lovelock and biologist Lynn Margulis. This new way of looking at global ecology and evolution differs from the classical picture of ecology as a biological response to a menu of physical conditions. The idea of co-evolution of biology and the physical environment where each influences the other was suggested as early as the mid-1700s, but never as strongly as Gaia, which claims the power of biology to control the nonliving environment. More recently, the terms Gaian science or Gaian theory have become more common than the original Gaia hypothesis because of modifications in response to criticisms and expansion of our scientific understanding.

Gaia – Original Versions

In the late 1960s, James Lovelock was working for NASA on life detection methods for Mars. With his chemistry training, this experience caused him to think deeply about what makes Earth different from Mars or her other neighbors in the solar system and the role that life might be playing in those differences. The imprint that life leaves on the chemistry of our own atmosphere stood out as a significant fingerprint of Earth's ecosystems. These musings led to the formulation of the first incarnation of the Gaia hypothesis. The early notion advanced by Lovelock is summarized in his 1972 paper, ``Life regulates the climate and the chemical composition of the atmosphere at an optimum for itself.'' Novelist William Golding, who lived near Lovelock, suggested naming the idea after the Greek goddess. This was lovely and poetic, but probably contributed to early perceptions that the concept was cultic or New Age, not scientific.

After significant initial criticism, Lovelock and Margulis realized the flaws in the initial version that laid them open to criticism. Biologist Ford Doolittle was particularly helpful in pointing out that the hypothesis as stated required foresight and planning on the part of collections of organisms toward a common goal. This appeared to be a teleological (purposeful or designed) notion that is not in keeping with the scientific view of causality.

Later, the revised formulation appeared in a number of written and oral presentations that can be paraphrased as: ``The whole system of life and its material environment is self-regulating at a state comfortable for the organisms.'' This was eventually restated by Lovelock in 1988 in his book *The Ages of Gaia* as ``Living organisms and their material environment are tightly coupled. The coupled system is a superorganism, and as it evolves there emerges a new property, the ability to self-regulate climate and chemistry.'' Lynn Margulis, the innovator of the endosymbiotic theory of eukaryotic cell origins, emphasizes the role of symbiosis in biology. Her statements about Gaia usually include the phrase superorganismic system. In her view, evolution is the result of cooperation, not competition, and this is in keeping with the Gaian interpretation of global ecology.

The initial conception involved the idea of homeostasis, that is, regulation around a narrow range of physical variables and resistance to perturbation via cybernetic feedback loops. However, Margulis particularly argued that Gaian systems are rather homeorhetic, meaning that the Earth's atmosphere, hydrosphere, and lithosphere are regulated around set points that can change in time as the whole system evolves essentially through a life cycle. The basic logic of negative and positive feedback loops is illustrated in **Fig. 1**.

Possible Evidence of Gaian Mechanisms

Biogeochemistry and Gaia

The chemical disequilibrium of the Earth's atmosphere is the feature that first captured Lovelock's attention. He noticed that on Venus and Mars (planets apparently with no life at least on the surface) the atmospheres are primarily CO_2. On Earth, the dominant constituents are reactive species of nitrogen, oxygen, and minor constituents (methane, ammonia, nitrous oxide, etc.). In the absence of other factors, over time, one would predict that Earth would resemble her neighbor planets but she does not. According to Gaia, life is the factor that maintains this disequilibrium over time. **Table 1** shows various parameters that have been suggested as possibly Gaia-controlled. The chemical compounds involved and their various reactions and fluxes control the large-scale biogeochemical cycles that enable Earth to constantly recycle materials and make them available for succeeding generations of life.

 doi:10.1016/B978-0-444-63768-0.00735-6

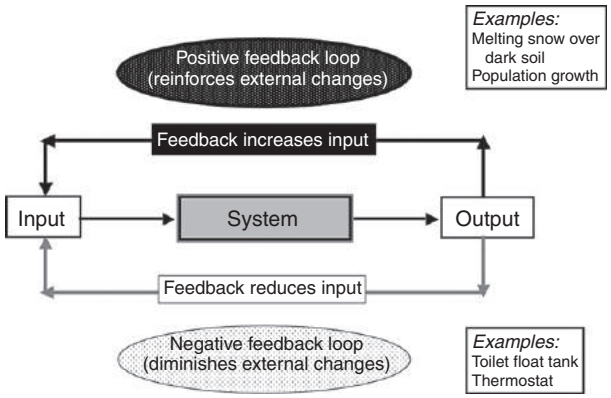


Fig. 1 Logic of negative and positive feedback loops.

Table 1 Proposed Gaian-controlled parameters

- Temperature, gas balance, greenhouse feedback
- Plant-albedo feedback (e.g., Daisyworld-like mechanisms)
- Evapotranspiration, latent heat, climate feedback
- Photosynthetic manipulation of air composition
- Dimethyl sulfide (DMS), marine cloud, algae association
- Microbial respiration rates and the carbon cycle
- Methanogenesis and greenhouse warming
- Carbon dioxide levels and carbonate cycle
- Carbonate-shelled organisms as long-term carbon sink
- Continental weathering rates via lichen, other microorganisms, etc.
- Oxygen levels, biomass burning feedback
- Ocean salinity levels

Bioweathering and Gaia

In 1989, Tyler Volk and Dave Schwartzman showed convincingly that the rock weathering rate increased by three orders of magnitude in the presence of life compared to the lifeless case. Gaian proponents viewed this as a major piece of supporting evidence in their contention that CO_2 effects on climate, known to be very powerful, could be significantly affected, even controlled, by the biologically enhanced rate of weathering. Lovelock has said of their work ``This is much more than is needed to enable a powerful physiological regulation of climate and carbon dioxide. We think it could account for the 300-fold decline in carbon dioxide since life began on Earth."

A Controversial Idea from the Beginning

Early Criticisms

Since its inception, the Gaia hypothesis has been controversial. For a few years, it was simply ignored. Further papers and presentations caught attention and the notion was widely castigated. It was criticized as being merely a restatement of ideas that already had a long history, as early as the work of James Hutton (1727–97), the founder of modern geoscience, who suggested that the study of Earth should be considered geophysiology. Further, Earth viewed as a single entity is in conflict with the fundamental ecological ideas of organisms engaged in Darwinian competition and narrowly defined survival and reproductive success. It has also been pointed out that maybe we do not need to invoke Gaia because proposed geochemical mechanisms can adequately explain many aspects of the Earth system without biological processes. Besides genuine weaknesses in the arguments, initial negative reactions to Gaia may in part be blamed on the lack of common language between Earth sciences and biology in the early 1970s.

Kirchner's Formulations

The best critical analysis of Gaian ideas was done by Jim Kirchner, at a Chapman Conference (American Geophysical Union) in San Diego in 1988 that was devoted to the scientific consideration of Gaia. He separated the jumble of ideas into four clear levels in order of increasing strength of claims from weak to strong (Table 2). These ranged from (1) `co-evolutionary Gaia' that merely

Table 2 The many types of Gaia according to Kirchner[a]

Hypothesis type	*Properties*	*Likely consensus*
Influential	Biology exerts significant influence over some aspects of theplanetary system	Testable and supported by evidence
Co-evolutionary	Darwinian process in which biota affects nonliving systems,in turn they affect biota	Testable and under active debate
Homeostatic	System is stabilized by negative feedback loops involvingbiota and physical/chemical systems	Testable and under active debate
Teleological	Conditions maintained by the biosphere for its own benefit	Testable, refuted by the Daisyworld demonstration
Optimizing	The biosphere directly manipulates its environment to provide optimum conditions for itself	Skeptically received, possibly not testable, not self-consistent

[a]Strength of statement in order from highest (influential) to lowest (optimizing).

claimed life and Earth had evolved together over time affecting each other; (2) `homeostatic Gaia' involving self-regulation around set points; (3) `geophysical Gaia', which overlapped significantly with the physical Earth sciences; and (4) the most extreme claim of `optimizing Gaia' that life was molding the planet's behavior into a state most favorable toward all life. The latter notion came in for the most criticism as being scientifically untestable and the most radical Gaian idea.

Further General Criticisms

Our planet has a long and dynamic history. How narrowly can Gaia be said to have constrained conditions? As we learn more about Earth's history, it is clear that huge changes have occurred in the climate, position of land masses, ocean currents, and other global-scale properties. For example, several times in the planet's history, we believe that it has been largely covered with ice. During the Mesozoic period, it appears that the planet was much warmer than it has been since. The atmosphere has evolved from anaerobic to a high level of free oxygen and many other major chemical changes have occurred. Against the dramatic backdrop of these changes, it is hard to claim that Gaia has held conditions constant and the window of variability seems very large even to qualify as homeorhesis.

Gaia as an organism has foundered on another point. Organisms reproduce. How can an entity the size of a whole planet reproduce? Gaia has not yet done so, but it has been suggested by some that space colonization may be the biosphere's first attempt to reproduce itself on other planetary bodies. The notion of Earth as superorganism may be specious and not central to the idea of global homeorhesis; thus, this may be a fairly trivial semantic criticism.

Because the notion has evoked visions of Gaia as the `mother goddess', as a benign entity and protector of life, it is appealing to people outside the scientific community. This has also been another point of attack for its critics, who view it as overly romanticized, more philosophical in nature, and scientifically untestable, thus, not of value in the strict scientific sense. If Gaia is the mother goddess, then her first-born were probably bacteria-like organisms who would be poisoned by our current oxygen-containing atmosphere. Her enormous family includes countless species whose individual needs and welfare conflict with each other. This has resulted in the natural extinctions of the bulk of all species that have ever arisen. The interpretation of such a system as benign or nurturing is stretching it too far.

Specific Criticisms

There are numerous specific criticisms that have been leveled against examples that Lovelock, Margulis, and other proponents have put forth as evidence for Gaia. Only one is given here for brevity. For example, Lovelock invoked Gaia to explain life's survival during the rise of oxygen in the early Earth atmosphere. Photosynthetically produced high levels of oxygen (the so-called `oxygen crisis') were lethal for the largely anaerobic lifeforms of the day. Ultimately, the oxygen atmosphere probably enabled a net increase in biodiversity and total biomass over time, but it spelled doom for many of the organisms then alive. If Gaia favors life, how can changes that favor some life but destroy large numbers of other organisms be reconciled? How can a mechanism be proposed that would amount to altruistic suicide on the part of many organisms on behalf of unrelated organisms? Kirchner summed up the dilemma in his 1989 paper, ``If the most destabilizing biotic event in Earth's history can be construed as evidence for Gaia, and the relative stability since then can also be cited as evidence for Gaia, one wonders what conceivable events could not be interpreted as supporting the Gaia hypothesis. If there are none, Gaia cannot be tested against the geologic record.... If Gaia stabilizes and Gaia destabilizes... is there any possible behavior which is not Gaian?"

If Gaia cannot be disproved in any case, then it does not meet the criterion for falsifiability developed by philosopher of science, Karl Popper, in the 1930s. An idea must, in principle, be able to be proved false for it to be considered testable. Popperian falsifiability has itself been attacked, notably by Alan Sokal and Jean Bricmont in their 1998 book, *Fashionable Nonsense*. Some argue that Popperian falsifiability is already biased toward only the methodology of reductionism, and that it may be inherently unable to fully define the essence of extremely complex and closely coupled systems, especially those that change over time in some sort of ontological process. Nevertheless, it is a useful indicator of whether a concept can be considered scientifically

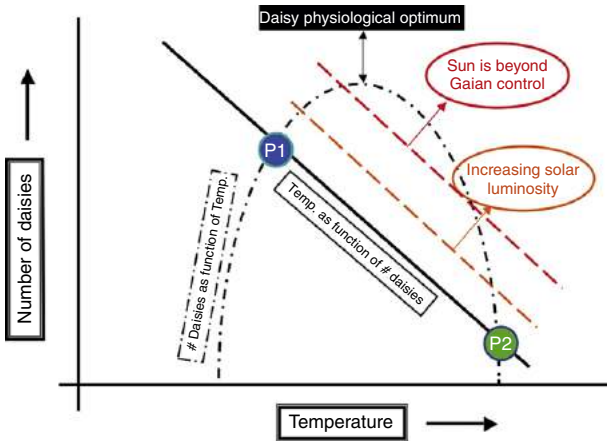


Fig. 2 The Daisyworld computer model.

testable at least in the narrow sense. The major result of Kirchner's criticism has resulted in attempts at crisper formulations of the ideas and has pushed Gaia to the weaker versions.

The highly reductionist geneticist Richard Dawkins believes that individual genes are in control of evolution and has entirely dismissed Gaia on that basis. Genes are grouped together into replicator packets. These are the functional units that Dawkins contends were both the first form of life and still remain the functional unit of selection. He dubs cells and organisms survival machines, and claims that they serve only to help the replicators propagate. Certainly any superorganismic concept violates Dawkins' notion of a single exclusive level of selection. Such an intensely reductionistic view does not take into account higher-order properties that may emerge from complex system interactions, and some scholars working on the mathematics of complex systems have in turn dismissed Dawkins' views as overly simplified.

Countering the reductionist view, J. Z. Young pointed out in the book *Doubt and Certainty in Science* that ``Biology, like physics, has ceased to be materialist. Its basic unit is a non-material entity, namely an organization.'' Here the emphasis is on pattern, because matter is frequently replaced and thus, transient, in biological processes. An organism is not a particular chunk of matter, but a persistent pattern through which material flows. With such a definition, the notion of the Earth system as a superorganism becomes less strained and does not require a slavish point-by-point comparison to the properties of individual organisms.

Daisyworld

Lovelock and his collaborators' greatest effort to counter the teleological criticism and show that undirected negative feedback can result in homeostatic regulation came in the form of a computer model, Daisyworld (Fig. 2). In the first and simplest version of the model, the imaginary planet Daisyworld has only one species of plant, white daisies. It has soil that is darker than the daisies. The star of this planet grows more and more luminous as we believe our sun to have done during the early history of the Earth. The relative abundance of daisies versus soil controls the temperature environment of the planet. The daisies have a physiological temperature window within which they are viable and reproductive. Using very simple rules, the differential albedos of soil and daisies combine to enable Daisyworld to remain habitable even as its sun is growing brighter. Of course, later versions of the model have added more biological variables and more complex physics of the environment, but the essence of the demonstration remains the same.

Where Is Gaian Science Headed?

Because of the global scale of Gaian processes, field observations of potential evidence supporting or refuting Gaia is difficult to obtain. Efforts continue sporadically to advance on this front, usually as a by-product of investigators' more mainstream activities. Possibly the most promising arena for testing Gaian ideas currently available lies in modeling and the understanding of complex systems. An early paper by Tregonning and Roberts in 1979 looked at how simple models of complex systems could develop homeostasis, and was seminal in early thinking about Gaia. Recently, the study of complex adaptive systems (CAS) has begun to advance our general understanding of the behavior of massively coupled complex systems. As this science progresses, insights applicable to testing Gaian predictions may well emerge.

See also: Global Change Ecology: The Earth System and Climate Science: Understanding a Very Complex Entity; Biosphere: Vernadsky?s Concept

Further Reading

Charlson, R., Lovelock, J., Andreas, M., Warren, S., 1987. Oceanic phytoplankton, atmospheric sulfur, cloud albedo, and climate. Nature 326, 655–661.
Kirchner, J.W., 1989. The Gaia hypothesis: Can it be tested? Reviews of Geophysics 27 (2), 223–235.
Kirchner, J.W., 2003. The Gaia hypothesis: Conjectures and refutations. Climatic Change 58 (1–2), 21–45.
Lovelock, J., 1972. Gaia as seen through the atmosphere. Atmospheric Environment 6, 579–580.
Lovelock, J.E., 1979. Gaia: A New Look at Life on Earth. Oxford: Oxford University Press.
Lovelock, J., 1983. Daisy world: A cybernetic proof of the Gaia hypothesis. Coevolution Quarterly 38, 66–72.
Lovelock, J.E., 1995. The Ages of Gaia. A Biography of Our Living Earth, 2nd edn. Oxford: Oxford University Press.
Lovelock, J.E., Margulis, L., 1974. Atmospheric homeostasis by and for the biosphere: The Gaia hypothesis. Tellus 26, 2–9.
Schneider, S.H., Boston, P.J. (Eds.), 1991. Scientists on Gaia. Cambridge, MA: MIT Press.
Schneider, S.H., Londer, R., 1984. Coevolution of Climate and Life. Berkeley, CA: Sierra Club Books.
Schneider, S.H., Miller, J.E., Crist, E., Boston, P.J. (Eds.), 2004. Scientists Debate Gaia: The Next Century. Cambridge, MA: MIT Press.
Schwartzman, D.W., Volk, T., 1989. Biotic enhancement of weathering and the habitability of Earth. Nature 340, 457–460.
Tregonning, K., Roberts, A., 1979. Complex systems which evolve towards homeostasis. Nature 281, 563–564.
Volk, T., 1998. Gaia's Body: Toward a Physiology of Earth. New York: Springer.
Watson, A.J., Lovelock, J.E., 1983. Biological homeostasis of the global environment: The parable of Daisyworld. Tellus 35B, 284–289.

Global Carbon Cycle 1: Short-Term Dynamics☆

GA Alexandrov, Russian Academy of Sciences, Moscow, Russia

Introduction

Short-term dynamics of the global carbon cycle is closely related to the concept of climate system: the totality of the atmosphere, hydrosphere, biosphere, and their interactions. Human activities have been substantially increasing the concentrations of carbon dioxide and other greenhouse gases in the atmosphere and thus inducing potentially adverse changes in the climate system. This tendency has become of public concern that led to the United Nations Framework Convention on Climate Change (UNFCCC). This convention suggests protection of carbon pools, enhancement of carbon sinks, and reduction of emissions from carbon sources.

Carbon Pools

Carbon pool (or reservoir, or storage) is a system that has the capacity to accumulate or release carbon. The absolute quantity of carbon held within at a specified time is called carbon stock. Transfer of carbon from one carbon pool to another is called carbon flux. Transfer from the atmosphere to any other carbon pool is said to be carbon sequestration. The addition of carbon to a pool is referred to as uptake.

Carbon Sink

Carbon sink is a process or mechanism that removes carbon dioxide from the atmosphere. A given carbon pool can be a sink, during a given time interval, if carbon inflow exceeds carbon outflow.

Carbon Source

Carbon source is a process or mechanism that releases carbon dioxide to the atmosphere. A given carbon pool can be a source, during a given time interval, if carbon outflow exceeds carbon inflow.

☆*Change History:* December 2015. GA Alexandrov updated the sections, further readings and Figure to this entire article.

Carbon Budget

The estimates of carbon stocks and carbon fluxes form the carbon budget, which is normally used as a kind of diagnostic tool in the studies of the short-term dynamics of the global carbon cycle.

Carbon Budget Components

The components of the global carbon budget may be subdivided into fossil and dynamic categories (Figure 1).

Fossil Components

The fossil components are naturally inert. The stock of fossil organic carbon and mineral carbonates (estimated at 65.5×10^6 PgC) is relatively constant and would not dramatically change within a century. The lithospheric part of the carbon cycle is very slow; all the fluxes are less than 1 PgC year^{-1}. For example, volcanic emissions are estimated at 0.15–0.25 PgC year^{-1}. Therefore, the turnover time of the storage amounts to millions or hundred millions of years.

Dynamic Components

The dynamic components are not inert. The carbon stocks in the atmosphere, ocean, soil, and vegetation remain constant as long as they are at dynamic equilibrium.

Atmosphere

The turnover time of atmospheric carbon is very short. It is less than 5 years. Therefore, the balance of atmospheric carbon strongly depends on the gas exchange between the atmosphere and ocean as well as on the gas exchange between the atmosphere and terrestrial ecosystems.

Atmospheric CO_2 content

The atmospheric content of carbon dioxide gradually increased from 680 PgC in 1960–69 to 760 PgC in 1990–99 and reached 830 PgC in 2011. The rate of increase was 3.2 ± 0.2 PgC year^{-1} during 1980–89, 3.1 ± 0.2 PgC year^{-1} during 1990–99, 4.0 ± 0.2 PgC year^{-1} during 2000–09, and 4.3 ± 0.2 PgC year^{-1} during 2002–11. Interannual variations were significantly wider. The lowest rate of 1.9 PgC year^{-1} was observed in 1992, and the highest rate of 6.0 PgC year^{-1} was observed in 1998. Since fossil fuel emission does not show short-term variability of this magnitude, the interannual variations are normally attributed to the climate-induced variations in the land–atmosphere flux, or the ocean–atmosphere flux, or both.

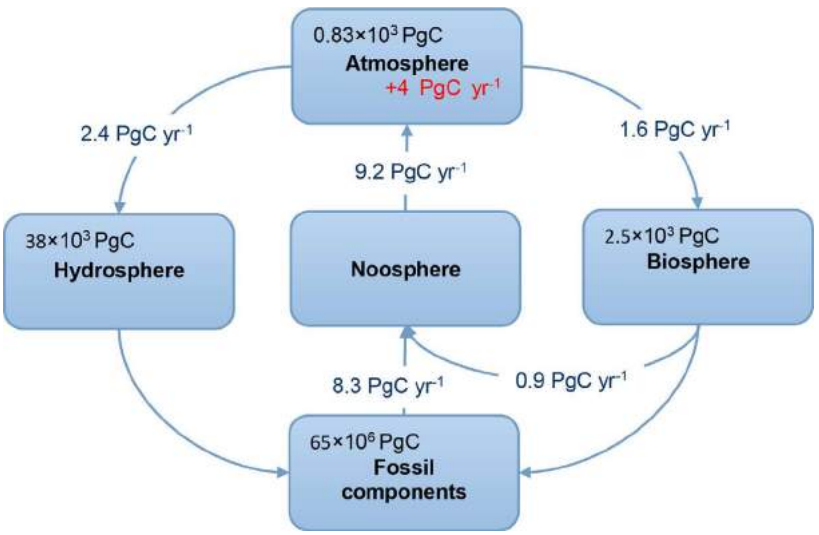


Figure 1 Dynamic components of the global carbon cycle.

Air/sea exchange

The gross carbon exchange between the atmosphere and ocean is estimated at 80 PgC year^{-1}. Since atmospheric CO_2 concentration is increasing, there is net uptake of carbon by the ocean, driven by the atmosphere–ocean difference in partial pressure of CO_2. The magnitude of the uptake slightly increases on decadal scale: 2.0 ± 0.7 PgC year^{-1} for 1980–89, 2.2 ± 0.7 PgC year^{-1} for 1990–99, 2.3 ± 0.7 PgC year^{-1} for 2000–09, 2.4 ± 0.7 PgC year^{-1} for 2002–11. The weak upward trend in the the magnitude of net oceanic uptake is attributed to large scale climate re-organizations.

Air/land exchange

The gross carbon exchange between the atmosphere and land is estimated at 60 PgC year^{-1}. Despite the net release of CO_2 associated with land use, the balance between emissions from fossil fuel combustion, net ocean uptake, and accumulation in the atmosphere is positive suggesting net uptake of carbon by the land. The magnitude of the net terrestrial uptake dramatically increased in 1990s, and remained quite stable in this century: 0.1 ± 0.8 PgC year^{-1} for 1980–89, 1.1 ± 0.9 PgC year^{-1} for 1990–99, 1.5 ± 0.9 PgC year^{-1} for 2000–09, 1.6 ± 1.0 PgC year^{-1} for 2002–11.

Ocean

The total amount of inorganic carbon in the sea is about 38,000 PgC. This includes dissolved carbon dioxide, HCO_3^-, and CO_3^{2-}. The dissolved carbon dioxide forms only one per mille of the total. Most of carbon is stored in the form of HCO_3^-. The share of CO_3^{2-} is less than 15% (about 13%).

The ocean storage is divided into surface water and deep sea. The surface water is in turn divided into cold surface water and warm surface water, both extending to 75 m depth – that is, the depth of seasonal thermocline. Warm surface water is a part of ocean between 40°N and 40°S where a well-defined permanent thermocline exists. Cold surface water is the rest of the ocean, which exchanges with deeper layers of the ocean by convection.

The surface water contains about 10% more carbon than the atmosphere that constitutes about 2% of the total carbon content in the ocean.

The concentration of dissolved carbon dioxide is proportional to its partial pressure above ocean surface (P_{CO_2}). The coefficient of proportionality is called solubility coefficient. The solubility of CO_2 varies with temperature suggesting a transfer of CO_2 in the atmosphere from warm regions to polar regions where CO_2 solubility is higher. This atmospheric transfer is balanced by the backward net oceanic transfer, the so-called 'conveyer belt'.

The dissolved carbon dioxide forms about 0.5% of the total dissolved inorganic carbon (DIC). The concentration of DIC also depends on P_{CO_2}, but in a more complicated way:

$$\frac{\Delta \mathrm{DIC}}{\mathrm{DIC}} = \frac{1}{\xi} \frac{\Delta P_{CO_2}}{P_{CO_2}}$$

where ξ is the buffer factor.

The surface water contains a significant amount of organic carbon: about 50 PgC of dissolved organic carbon (DOC), 30 PgC of particulate organic carbon (POC), and 3 PgC of plankton. The rate of photosynthesis varies from 0.06 gC m^{-2} day^{-1} in the desert regions which are characterized by downwelling and lack of nutrients to 0.6 gC m^{-2} day^{-1} in areas of intense upwelling. The total primary production amounts to 50 PgC year^{-1}. About 25% of the primary production reaches the deep water that contains about 700 PgC of DOC.

Land

Terrestrial ecosystems store about 2500 PgC: 500 PgC in vegetation and 2000 PgC in soil. The amount of carbon stored in vegetation varies significantly depending on vegetation type. Forests generally store ten times more carbon than grasslands. However, forest soils do not necessarily contain more carbon than grassland soils, for carbon stock in soil depends on the factors that control the rate of organic matter decomposition.

The global net production of organic matter by plants, net primary production (NPP), is about 60 PgC year^{-1}. This input to the pool of living organic matter is compensated by litter fall. The residence time of the pool is 1 year in case of annual grasslands, but it may be 100 years in case of pristine forests.

The litter fall in its turn is the input to the pool of nonliving organic matter, which is compensated by heterotrophic respiration (i.e., by CO_2 released with the respiration of soil biota decomposing organic debris). The net accumulation of carbon by ecosystem, including both soil and vegetation, is called net ecosystem production (NEP).

NEP of an undisturbed ecosystem should be close to 0. However, most of ecosystems are disturbed in some way (harvesting, fire, etc.). Therefore, global NEP is estimated at 10 PgC year^{-1}. This value characterizes nonrespiratory losses such as release of carbon due to forest fires, or relocation of carbon to wood products and other components of urban metabolism.

The net accumulation of carbon that includes both respiratory and nonrespiratory losses is called net biome production (NBP). NBP for a relatively short period of time may differ from 0, reflecting continued effect of the losses occurred in the past. Thus, global NBP (i.e., net terrestrial uptake) for the decade 2002–11 has been estimated to be positive (1.6 ± 0.7 PgC year^{-1}).

Human Intervention

Fossil Fuel Combustion and Cement Production

The main anthropogenic source of CO_2 emissions is the combustion of carbon-based fuels. Since 1860, industrialization has progressed with an increase in the use of fuels – especially fossil fuels – and a corresponding increase in CO_2 emissions. The growth has been steady and exponential, interrupted only by the two World Wars and the Great Crash of 1929. In 2103 the emissions reached a maximum of 9.9 PgC year^{-1} (0.2 PgC year^{-1} of this was from cement production). The average value of emissions increased from 5.5 ± 0.4 PgC year^{-1} in 1980s to 8.3 ± 0.7 PgC year^{-1} in 2002–11.

Land-Use Change

About 10–30% of the current total anthropogenic emissions of CO_2 are estimated to be caused by land-use conversion. In the historical perspective, the share of land use is larger. From 1750 to 2011, 375 ± 30 PgC has been emitted as a result of fossil fuel burning, and 180 ± 55 PgC as a result of land-use change.

The current net land-use flux comprises the balance of positive terms due to deforestation and negative terms due to regrowth on abandoned agricultural land. During 1980s the net land-use flux of 1.4 ± 0.8 PgC year^{-1} was almost entirely due to deforestation of tropical regions. Temperate forests show approximate balance between carbon uptake in regrowing forests and carbon lost in oxidation of wood products. Since the rates of deforestation are declining, the net land-use flux was slightly smaller in the beginning of this century: 1.1 ± 0.8 PgC year^{-1} in 2000–09 and 0.9 ± 0.8 PgC year^{-1} in 2002–11.

Biosphere Response

Growth Rate of Atmospheric CO_2

From 1750 to 2011, atmospheric concentration of CO_2 increased by 40%, from 280 to 390 ppm, that corresponds to 240 ± 10 PgC increase in the atmospheric content of CO_2 and to 43% of the total anthropogenic emission over this time. More than half (57%) of the anthropogenic emission has been taken up in the ocean and the terrestrial ecosystems.

Net Oceanic Uptake

The cumulative ocean uptake during the period from 1750 to 2011 is estimated to be 155 ± 30 PgC – that is, about 30% of anthropogenic emission has been taken up in the ocean.

The fraction of anthropogenic CO_2 that is taken up in the ocean declines with increasing CO_2 concentration in the atmosphere. Increasing atmospheric CO_2 concentration maintains the atmosphere–ocean difference in partial pressure of CO_2 that causes net uptake of carbon by the ocean. However, increasing DIC reduces the buffer capacity of the carbonate system, and thus weakens the capacity of the oceanic uptake.

The capacity of the ocean uptake is also sensitive to the rate of increase of atmospheric CO_2. The uptake is limited with the rate of mixing between deep water and surface water, and hence the lower the growth rate of atmospheric CO_2, the higher the rate of CO_2 sequestration.

Residual Terrestrial Uptake

Balancing the carbon budget for the period from 1750 to 2011 yields a global net terrestrial source of 30 ± 45 PgC. Hence, a residual terrestrial sink of 160 ± 90 PgC is required to reconcile the difference between the relatively small net terrestrial source and the relatively large terrestrial source resulted from land-use change (180 ± 80 PgC). This residual terrestrial uptake is sometime referred to as the 'missing carbon sink', although it can be attributed to well-known biophysical mechanisms.

Carbon dioxide fertilization

Stimulation of photosynthesis at higher CO_2 is one of the well-known mechanisms to which the 'missing carbon sink' is normally attributed. Carbon dioxide fertilization effect on plant productivity is not linear:

$$\frac{\Delta \mathrm{NPP}}{\mathrm{NPP}} = \gamma \ln\left(1 + \frac{\Delta P_{\mathrm{CO_2}}}{P_{\mathrm{CO_2}}}\right)$$

It is weakening at high atmospheric concentration of CO_2. If $\gamma = 0.35$, then NPP increases by 24%, when CO_2 increases two times. Thus, the growth of atmospheric concentration of CO_2 enhanced plant productivity by 10% or little more in comparison to 1750.

Carbon dioxide fertilization produces only an excess amount of organic matter. The mechanism of carbon sequestration associated with CO_2 growth is more complicated. Carbon sequestration stems from imbalance between production and decomposition of organic matter – that is, from NEP. NEP approaches to naught when ecosystem approaches to equilibrium. Continuous

growth of CO_2 maintains ecosystem disequilibrium. Therefore, the rate of carbon sequestration is determined by the rate of atmospheric CO_2 growth and the rate of carbon turnover in terrestrial ecosystems.

Nitrogen deposition

Production of organic matter is generally limited with nitrogen and some other nutrients. Therefore, NPP is expected to increase with a rapid growth in reactive nitrogen deposition over the last 150 years. Reactive nitrogen is released into the atmosphere in the form of nitrogen oxides (NO_x) during fossil fuel and biomass combustion. The annual deposition even in rural areas may amount to 40 kg ha^{-1}. Another source (mainly of ammonia) is animal husbandry and fertilizer use. The nitrogen deposition mainly affects ecosystems close to cities and industrial centers as well as in the vicinity of intensive agricultural enterprises.

Land-use management

Agricultural and forest management practices significantly affect carbon stocks in managed ecosystems. These practices dramatically changed since 1750; they began to be oriented to the sustainable use of natural resources. A forest that is managed in sustainable manner operates as a machine that removes carbon from atmosphere and exports it as forest products. Similarly, alteration of tillage practices leads to protection and increase of soil carbon content.

Concluding Remarks

This article presents an overview of basic biogeochemical concepts related to short-term dynamics of global carbon cycle. It is intended to make short-term dynamics of global carbon cycle understandable to ecologists who are involved in carbon management and related ecological applications and targeted to a reader who has basic background in ecology or is familiar with the basic ideas of sustainable development, biosphere equilibrium, and environmental protection. The 'Further reading' list provides the information on a broad context in which the basic concepts overviewed in this article (or their modifications) normally appear.

Further Reading

Bolin B, Degens ET, Kempe S, and Ketner P (eds.) (1979) *The global carbon cycle, SCOPE 13*. Chichester: Wiley.
Bolin B, Döös BR, Warrick RA, and Jäger J (1986) *The greenhouse effect, climatic change, and ecosystems, SCOPE 29*. Chichester: Wiley.
Ciais P, Sabine C, Bala G, Bopp L, Brovkin V, Canadell J, Chhabra A, DeFries R, Galloway J, Heimann M, Jones C, Le Quéré C, Myneni RB, Piao S, and Thornton P (2013) Carbon and other biogeochemical cycles. In: Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, and Midgley PM (eds.) *Climate change 2013: The physical science basis. contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge/New York, NY: Cambridge University Press.
Field C and Raupach MR (eds.) (2004) *The global carbon cycle: Integrating humans, climate and the natural world, SCOPE 62*. Washington, DC: Island Press.
Global Carbon Project (2003) *Science framework and implementation*. Earth System Science Partnership Report No. 1 Canberra: Earth System Science Partnership (IGBP, IHDP, WCRP, DIVERSITAS).
Houghton JT, Ding Y, and Griggs DJ, et al. (eds.) (2001) *Climate change 2001: The scientific basis*. New York, NY: Cambridge University Press.
Jørgensen SE and Svirezhev YM (2004) *Towards a thermodynamic theory for ecological systems*. Amsterdam: Elsevier.
Melilo JM, Field CB, and Moldan B (eds.) (2003) *Interactions of the major biogeochemical cycles: Global change and human impacts, SCOPE 61*. Washington, DC: Island Press.
Schellnhuber HJ, Crutzen PJ, and Clark WC, et al. (eds.) (2004) *Earth system analysis for sustainability*. Cambridge: MIT Press.
Watson RT, Noble IR, and Bolin B, et al. (eds.) (2000) *Land use, land-use change, and forestry*. New York, NY: Cambridge University Press.

Global Negative Emission Land Use Scenarios and Their Ecological Implications

Yoshiki Yamagata, National Institute for Environmental Studies, Tsukuba, Japan

Glossary

Negative emission In order to limit global temperature rises to below 2°C, "negative emissions" technologies are expected to be applied to remove carbon from the atmosphere. This can be done easily at small scale by restoring forests, planting trees to absorb carbon dioxide from the atmosphere. However, most of the current IPCC scenario suggest that large scale deployment of negative emission is necessary to achieve the low carbon target by growing bioenergy crop and using bioenergy, capturing the carbon emitted and pumping it into underground geological reservoirs as bioenergy, carbon, capture and storage (BECCS).

Land use scenario Land use has been changing due to the change of socio-economic development as well as natural environment. IPCC and other scientific assessments are using land use scenarios for the future to assess the impact of climate change and evaluate the potential of mitigation and adaptation policies. Most land use changes are induced by mainly by economic growth, migration or agricultural and industrial productivity changes. Scenario analysis comparing business as usual and possible narrative cases allow assessing the influence of policies.

Ecosystem services There are many human benefits that are contributed as services from the natural ecosystems. Such ecosystem services are also supplied from semi-natural agroecosystems, forest ecosystems, grassland ecosystems and aquatic ecosystems. These benefits are in combination recognized Nature's Contribution to People (NCP) in the recent IPBES assessment report. In general, ecosystem services are grouped into four categories: provisioning (e.g., production of food and water); regulating (e.g., control of climate and disease); supporting (e.g., nutrient cycles and crop pollination); and cultural (e.g., such as spiritual and recreational benefits). Usually these values are not capitalized and often neglected. There have been a lot of economic studies to support decision-makers by valuing ecosystem services in economic term.

Paris agreement At the Paris climate conference (COP21) in December 2015, 195 countries adopted this global climate deal. The agreement sets out to avoid dangerous climate change by limiting global warming to long-term goal of keeping the increase in global average temperature to well below 2°C above pre-industrial levels. It also aim to limit the increase to 1.5°C, since this would significantly reduce risks and the impacts of climate change. There is the need for global emissions to peak as soon as possible to undertake rapid reductions thereafter. Before the Paris conference, countries submitted comprehensive national climate action plans (INDCs).

SDGs In September 2015, the UN committed to the 2030 agenda for sustainable development (2030 agenda) containing 17 Goals and 169 targets called sustainable development goals (SDGs). SDGs include: poverty, hunger, wellbeing, education, equity, water and sanitation, energy, economic growth, employment, infrastructure and innovation, sustainable cities, sustainable consumption and production, climate change, oceans, terrestrial ecosystems, societies, global partnership. Most of goals need to be achieved by addressing the need for sustainable development that works for all people. Encompassing universal, transformative, inclusive and integrated goals and targets is one of the most comprehensive global agenda adopted.

Introduction

In December 2015, the Conference of the Parties (COP) to the United Nations Framework Convention on Climate Change (UNFCCC) adopted the "Paris Agreement" that stipulates "holding the increase in the global average temperature to well below 2°C above preindustrial levels" (Article 2). Considering the risk of crossing the dangerous tipping point of abrupt and irreversible change above a certain temperature (Schellnhuber *et al.*, 2016; Schleussner *et al.*, 2016), this "2°C target," or even "1.5°C target" if possible, is internationally agreed as a very important global common goal to achieve.

In fact, from a scenario point of view, "2°C target" corresponds to the IPCC representative concentration pathway (RCP) 2.6 scenario (van Vuuren *et al.*, 2011). According to the underlying scenarios corresponding to the RCP2.6 scenario, it is supposed that the global total carbon emission (human emissions minus natural absorption) will go negative near the end of this century (Fuss *et al.*, 2014).

On the other hand, regarding more comprehensive global sustainability, heads of states also agreed on yet another very important treaty in the same year. Namely, on September 25, 2015, the United Nations adopted the "sustainable development goals (SDGs)" which demand all UN countries to make all efforts to end poverty, protect the planet, and ensure prosperity for all as part of a new sustainable development agenda. Although the 17 SDG compliances are voluntary, each goal has specific targets (measured by corresponding indicators) to be achieved by 2030. Actually, "Climate action" is the 13th goal of the 17 SDGs. So, the

 doi:10.1016/B978-0-12-409548-9.10893-0

climate change mitigation policy (until 2030) of the Paris Agreement is a part of SDGs. Under the SDGs, it is recommended that governments achieve all goals at the same time. However, in fact, there are trade-offs and synergies between the goals. So, we need to understand the interactions between the goals to come up with a better scenario to achieve the SDGs. This is the background why we need to study the impacts of climate change mitigation for other types of sustainability such as water, food and ecosystems.

In fact, considering the cumulative emission of anthropogenic greenhouse gases (GHGs) until today and projected emissions accompanying human activities in the future (Creutzig *et al.*, 2016), it is quite challenging to achieve the "2°C target" by only reducing GHGs. In this regard, BioEnergy with Carbon Capture and Storage (BECCS) is highlighted as a promising "Negative Emission" technology which allows us both to earn more carbon-neutral energy and reduce atmospheric CO_2 concentrations at the same time (Smith *et al.*, 2016). The technology would produce electricity by bioenergy combustion, then capture CO_2 emissions and store it into deep ground.

However, producing a massive amount of bioenergy crop requires the vast use of agricultural cropland. Smith *et al.* (2016) estimated that the mean land requirement for BECCS would be 380–700 106 ha in 2100. The land requirement would be partly suppressed by enhancing agricultural productivity by irrigation. Hejazi *et al.* (2015) projected the total water use in the United States under a stringent GHG emission reduction policy. They found that the policy could increase water stress more than the climate change itself mainly due to bioenergy irrigation. Bonsch *et al.* (2016) estimated how much land and water was required to produce 300 EJ $year^{-1}$ of bioenergy. They found that 486 106 ha of cropland and 3000 km^3 $year^{-1}$ of irrigation water withdrawal were needed. In case no irrigation was applied, 41% of cropland was additionally required (total 689 106 ha).

Although the abovementioned two studies have quantified the trade-offs between mitigation and water scarcity, and land and water, respectively, further investigation is needed to explore whether irrigation water is stably and sustainably available. From the perspective of its impact on land use and ecosystem services, deployment of BECCS may have ramifications such as loss of biodiversity, deterioration of water quality, and additional emissions of nitrous oxide to the atmosphere (Melillo *et al.*, 2009; Smith *et al.*, 2016). Expansion of plantation of bioenergy crops (e.g., oil-palm) would exert influences on atmospheric quality by emitting volatile organic compounds (Misztal *et al.*, 2011). However, our knowledge on the direct and indirect impacts of BECCS on ecological systems is far from sufficient to conduct a reliable evaluation and to plan feasible management.

To fill the gap of the knowledge required from the urgency to the need to implement climate change mitigation activities and their sufficient assessments regarding their impact to other sustainability indicators, many researchers used multiple models to simulate the trade-offs between water, food, and ecosystems under three different land-use scenarios to produce bioenergy crop in agricultural land with/without irrigation and in converted forest lands. Using these models, the impacts from the BECCS deployment scenarios are assessed with the total amount of up to 3.3 GtC $year^{-1}$ (annual negative emission potential required for RCP2.6; see Smith *et al.* 2016) of bioenergy by growing bio-crops with substantive use of global agricultural and forest lands. Especially, the effects and sustainability of irrigation for global massive production of bioenergy is investigated by using the hydrological models (Hanasaki *et al.*, 2008a,b, 2010). The impacts of massive use of converted forest land are assessed using terrestrial ecosystem models (Ito and Inatomi, 2012).

BECCS Assessment Methods

There are some studies that have assessed the impacts of BECCS deployment scenarios on land systems including land use, water resources, and ecosystem services. **Fig. 1** shows the general explanation of the models used in a study (Yamagata *et al.*, 2018) and illustrates the parameters and variables exchanged between the water resources, ecosystems, and land-use models. There are some dependences between the models. For example, irrigation for bioenergy crop will decrease renewable water resources, while land conversion for bioenergy cropland will decrease the forest area etc.

Land use change impacts are tested using land-use scenarios (**Fig. 2**) to achieve the annual emission reduction of 3.3 GtC $year^{-1}$ (required for IPCC-RCP 2.6). Cropland scenarios can be used such as harmonized global land use (Chini *et al.*, 2014) for RCP2.6. It projects that global total cropland area reaches 2.12 billion ha in 2100. Since this kind of scenario does not specify the land used for food and bioenergy production, researchers need to assume how much land is required to achieve 3.3 GtC of BECCS. In the case of (Yamagata *et al.*, 2018), it indicates that cropland for food production would be 1.62 billion ha, or approximately 10% larger than the present areal extent. However, as it could be necessary to save cropland for food from the view point of food security in the future, because difficult food supply and demand condition are expected due to climatic change impacts on food production, population growth etc. So, Yamagata *et al.* (2018) considered two additional land use scenarios other than the use of rainfed cropland for the bio-crop production (S2), namely intensive irrigation for bioenergy crop (S1) and additional use of forest land (S3) as is explained in **Fig. 2**.

In case of S2, assuming BECCS for the RCP2.6, very large areas (500 million ha, or up to 25% of the global farm lands) need to be used for bioenergy crop production (rainfed) all over the world. Comparing the case of S1, the demand for farm lands is relaxed by assuming that bio-energy crop is irrigated to increase productivity to reduce the required land to half. In the case of the S3 scenario, large natural lands (500 million ha, or up to 10% of the total current forest land area) are assumed to be converted into bioenergy crop lands.

Another approach is to consider two subscenarios: (S3-1) no reserved area, allowing conversion of high biodiversity tropical forests, and (S3-2) biodiversity hotspots are reserved on the basis of the map by World Wide Fund for Nature. By assessing such

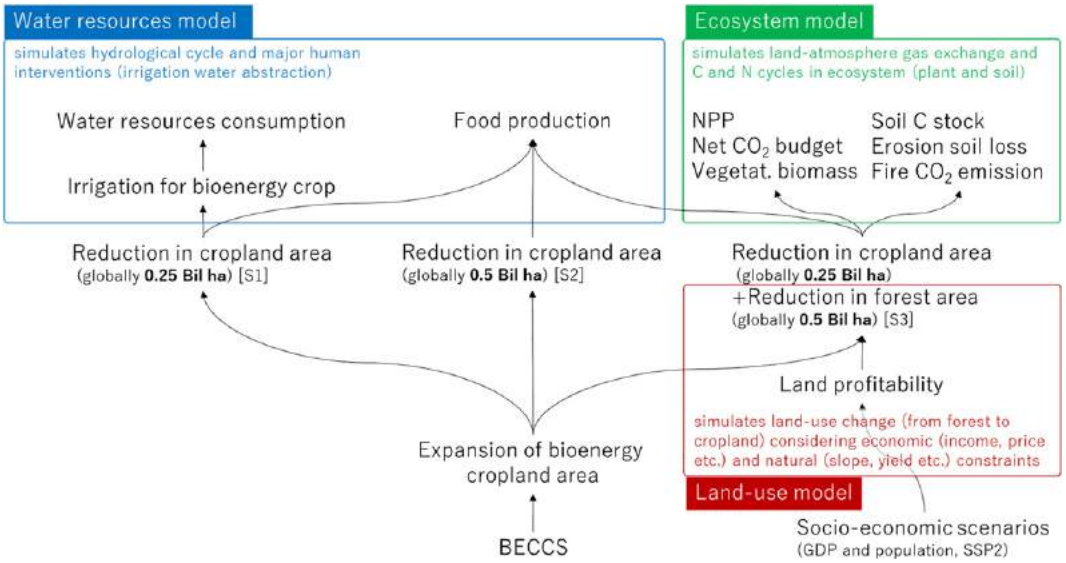


Fig. 1 Explanation of the models used to assess the impacts of different land- use scenarios and interactions between model parameters and variables. They are used to project situations in 2100 (Yamagata et al., 2018).

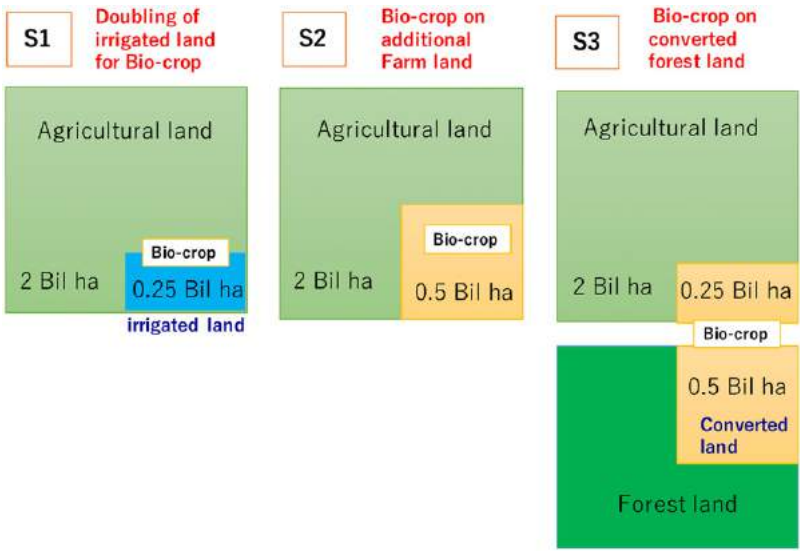


Fig. 2 Land-use scenarios for bioenergy crop production for RCP2.6 in 2100 (Yamagata *et al.*, 2018).

variety of land use scenarios, we can clarify the potential impacts and risks of the land use scenarios more clearly in a spatially explicit manner. For creating these variety of scenarios, land-use change models can be used by considering the land productivity for growing crops with allocation of the necessary areas for conversion.

Using these S1, S2, S3 land-use scenarios, the models such as described in **Fig. 1** can used to analyze the impacts for sustainability of land-use scenarios under RCP 2.6 from the perspectives of ecosystem and water resource in the future. In the land-use modeling, the suitability of bioenergy crop cultivation can be estimated in a similar manner to other food crops (i.e., wheat and maize) for simplicity. The base cropland area used in future land-use scenario such as S1 (with irrigation) and S2 (without irrigation) in 2000 can be derived from harmonized global land use (Chini *et al.*, 2014) which is consistent with RCP 2.6 (van Vuuren *et al.*, 2011) and cropland expansion for BECCS was assumed to occur in the 21st century at a constant rate. The scenario type such as S3 in which forests are consumed for bio-crop production can be derived from a land-use model considering socio-economic factors.

The spatial distribution of forest conversion into cropland for bio-energy can be developed based on the land suitability for agriculture. In the case of Yamagata *et al.* (2018), this was estimated in each 30 arc *sec* grid cell using explanatory variables such as wage, slope angle of land, bio-crop price and bio-crop yield defined by the socio economic scenario using GTOPO30 (available from USGS). The spatial distribution of bio-crop yield can be given by downscaling the results of a model whose resolution was

half-degree grid cell. It can be assumed that bioenergy crops such as Miscanthus and Switchgrass would be used and their yields were estimated by a model (Kato *et al.*, 2013). The socio-economic scenario such as wage and crop price can be provided by Integrated Assessment Model such as Asia-Pacific Integrated Model (Fujimori *et al.*, 2014). However, for simplicity, it was also assumed that the spatial distribution of cropland for food and pasture land would not change from 2000.

Simulations can be conducted using the climate scenario developed for the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP; Hempel *et al.*, 2013). In case of Yamagata *et al.* (2018), scenarios is conducted using the MIROC-ESM-CHEM Earth System Model under the RCP2.6 scenario (van Vuuren *et al.*, 2011). The MIROC-ESM-CHEM tends to give a high warming trend, making it easy to assess possible climate impacts. Nevertheless, the magnitude of warming in the 21st century under RCP2.6 scenario (about 2.3 K land-average) was comparable with other climate models.

Water Resource Models

The H08 model is a physically-based global hydrological model (Hanasaki *et al.*, 2008a,b, 2010). It simulates basic hydrological components globally at $0.5^\circ \times 0.5^\circ$spatial resolution by solving the energy and water balance at land surface. H08 explicitly expresses major human interventions in the natural hydrological cycle, namely, water abstraction for irrigation, industrial, and domestic use, and reservoir operation of major dams. All the natural and human processes interact at a daily interval.

H08 incorporates submodels to estimate the potential crop yield, the cropping calendar, and irrigation water requirement for annual food crops, but not for perennial bioenergy crops. These submodels were enhanced to deal with giant miscanthus (Miscanthus giganteus) and switchgrass (*Panicum virgatum*) which are both so-called second-generation bioenergy crops as follows. First, the crop-specific parameters for miscanthus and switchgrass were added to the crop yield submodel which were taken from the SWAT model version 2012 (Arnold *et al.*, 2012). Next the cropping calendar submodel was enhanced to deal with perennial plants. The growing period was estimated by searching the longest continuous days above the base air temperature (10°C for miscanthus and 12°C for switchgrass) in a year. If it exceeded 300 days, the continuous 300 days which produced the maximum mean yield was selected. Irrigation water was applied to keep soil moisture above 75% of the field capacity during the cropping period. Note that these submodels are independent from that of Kato *et al.* (2013) which was used to develop the land-use scenarios: the submodels are tightly incorporated into the H08 model and it was unable to replace them with Kato *et al.* (2013). In H08, soil moisture (or water and energy balance) is calculated for each land use. The standard H08 subdivides a grid cell into four different land uses, namely double-crop irrigated food cropland, single-crop irrigated food cropland, rainfed food cropland, and nonagricultural land. We newly added two land uses for irrigated and rainfed bio energy cropland.

Using the enhanced H08, three simulations were conducted. The first simulation was the base simulation. The simulation period is 1996–2005. It assumed no production of massive second-generation bioenergy crop in this period. The H08 simulated the natural hydrological cycle and human water use using the present spatial distribution of irrigated and rainfed cropland. The second simulation assumed that 250 million ha of cropland was changed into irrigated bioenergy cropland by 2100 (Scenario S1). The simulation period is 2006–2100. Irrigation water was primarily taken from rivers in the same grid cell. We estimated how much additional water was required when the rivers were depleted. The third simulation assumed that 500 million ha of cropland was changed into rainfed bioenergy cropland by 2100 (Scenario S2). The boundary conditions given to H08 were identical to those of Hanasaki *et al.* (2008a,b). Irrigated cropland area, and crop type were obtained from Siebert *et al.* (2005) and Monfreda *et al.* (2008), respectively and fixed throughout the simulation period. In this study, the effects of CO_2 fertilizer were not considered. It was expected that crop yield would grow by time due to technological advancement, that effect was not considered either.

Ecosystem Models

To evaluate the impacts to the terrestrial ecosystem services, various ecosystem models can be used. In Yamagata *et al.* (2018), Vegetation Integrated SImulator for Trace gases (VISIT) model was employed. The model is a process-based model of the terrestrial biogeochemical cycle and ecosystem dynamics (Ito and Inatomi, 2012), focusing on atmosphere–ecosystem interactions under changing climate conditions. The model simulates water, carbon, and nitrogen cycles in terrestrial ecosystems using a simple box-flow framework, enabling us to apply this model to point to global scales. Particularly, this model simulates atmosphere–ecosystem exchange of trace gases such as greenhouse gases (CO_2, CH_4, and N_2O), biomass-burning emissions (e.g., CO and black carbon), and biogenic volatile organic compounds.

In Yamagata *et al.* (2018), terrestrial ecosystem services was evaluated for: (1) net primary production (NPP) related to fundamental and provisional services, (2) net ecosystem CO_2 exchange related to regulation services, (3) vegetation biomass related to fundamental, provisional, and cultural (by scenery) services, (4) soil carbon stock related to fundamental services, (5) soil loss due to erosion, and (6) biomass burning related to degradation of ecosystem services.

In general, carbon dynamics in terrestrial ecosystems can be estimated on the basis of leaf-level gas exchange and canopy radiation transfer (Ito and Oikawa, 2002). Mass balance of carbon stocks in vegetation and soil pools can be estimated by accounting the carbon input and output for each of the eight carbon pools: three for C3-type vegetation, three for C4-type

vegetation, and two soil organic carbon. Soil loss by water erosion was estimated using the revised universal soil loss equation, which accounts for slope, soil stability, precipitation, vegetation cover, and human management factors (Ito, 2007). Biomass burning can be simulated using an empirical scheme to estimate burnt area and combustion intensity, which are functions of fuel load and soil wetness.

Land-Use Scenarios

Fig. 3 displays projected bio-crops distributions in 2100 under S1 and S3 (a), and S2 (b) in Yamagata *et al.* (2018). The figure shows that S2 increases land for bio-crops globally. The large increase is especially expected in South Africa, North America, and Europe. **Fig. 4** shows the distribution under S3 with assumption of additional use of natural forest land without or with biodiversity reserved lands. In the case of S3-1 which does not take into account forest protection regulations such as REDD, the rain forest of Brazil is predicted to change into bioenergy crop land first, followed by the rain forest of Congo. On the other hand, if the biodiversity rich tropical forests are protected (S3-2), bioenergy crop land is predicted to increase more in semi-arid land in Australia and southern Africa and boreal forest land in Canada and Russia.

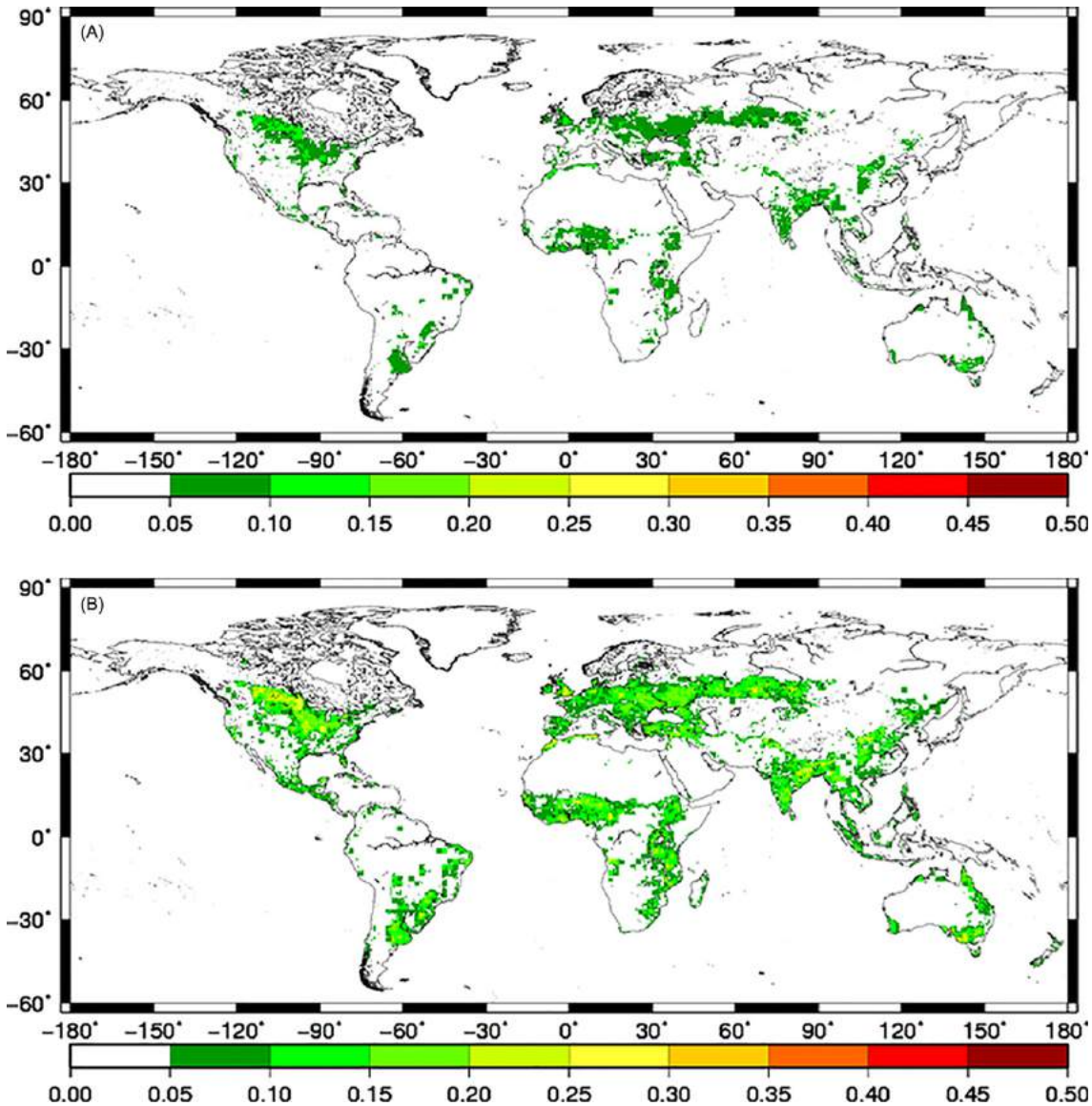


Fig. 3 Areal fraction of bio-crop farmland in 2100. (A) S1 and S3 (excluding farmland transferred from forest) and (B) S2 (Yamagata *et al.*, 2018).

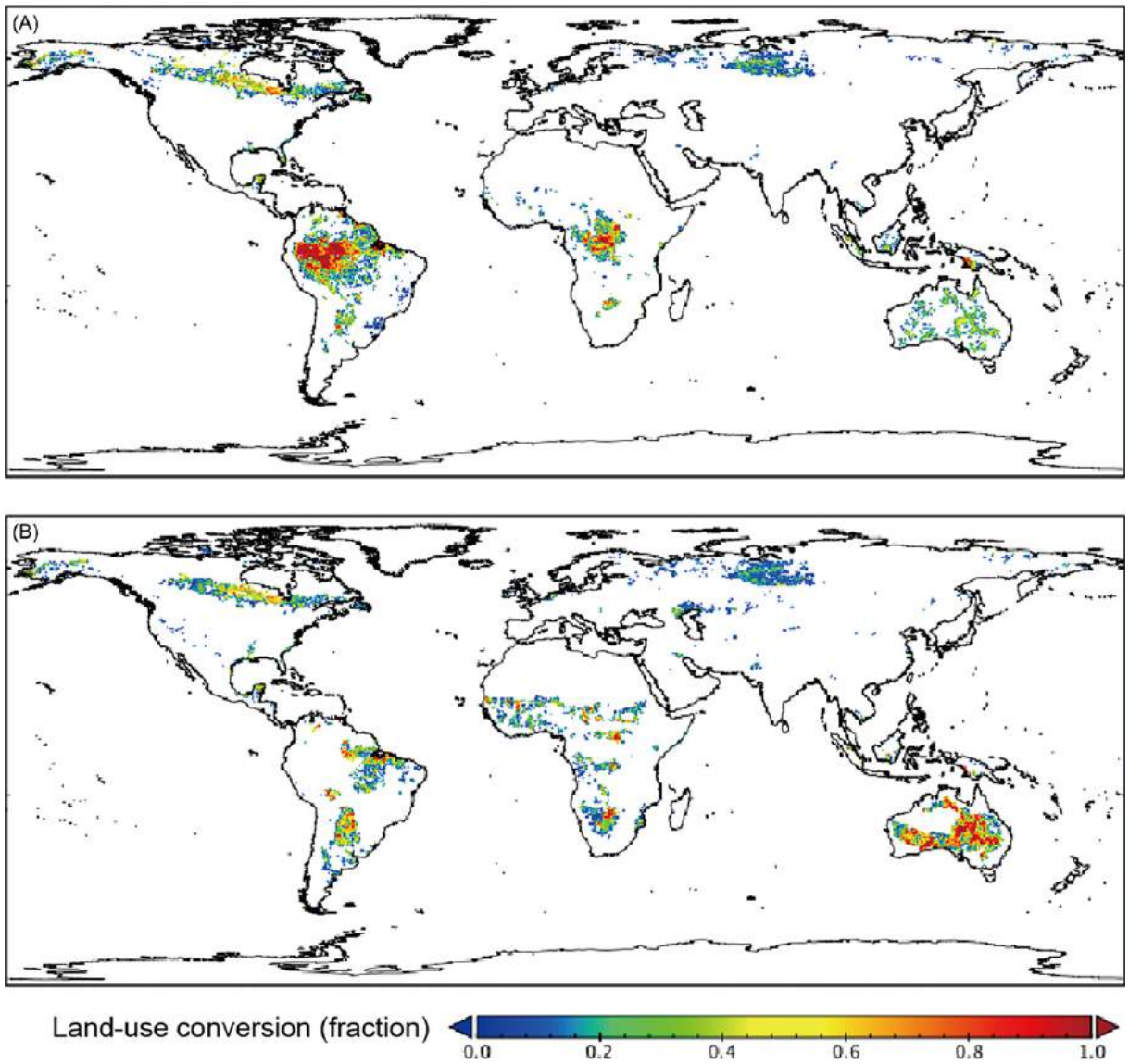


Fig. 4 Areal fraction of bio-crop farmland in 2099 transferred from forest (S3) [land fraction]. (A) No reserved land (S3–1) and (B) with reserved lands for biodiversity hot spots (S3–2) (Yamagata *et al.*, 2018).

Assessment of Land Use Scenarios With Water Resource Models

Irrigation water use for bioenergy can be calculated as shown in Yamagata *et al.* (2018), for Scenario S1, the volume of consumptive irrigation water use to produce bioenergy was estimated at 1910 km^3 $year^{-1}$ in 2090s (the mean of 2091–2100). This volume is as much as 135% of the volume for food production (1420 km^3 $year^{-1}$) of the base simulation which is fairly comparable with earlier independent estimates (e.g., 1231 km^3 $year^{-1}$ in Döll *et al.*, 2012). Irrigation water use for bioenergy is expected to concentrate in some parts of South America, central Sahel, eastern India, and northern and southern Australia (**Fig. 5**).

The spatial distribution in the figure reflects three aspects: the distribution of irrigated bioenergy cropland, soil moisture deficit, and length of growing period. In the paper (Yamagata *et al.*, 2018), Scenario S1 assumed that 250 million ha of cropland was converted into that for bioenergy proportional to the total cropland, hence irrigation is concentrated in the world's major breadbasket areas. Second, since irrigation was applied to maintain soil moisture above 75% of the field capacity, it tends to be concentrated in semi-arid regions. Third, since this study assumed that irrigation was applied throughout the cropping period, warm regions with long cropping period tend to require a large volume of irrigation. Of the total irrigation water requirement, rivers supplied only 1580 km^3 $year^{-1}$ of water. The remaining 1910 km^3 $year^{-1}$ should be supplied from other sources. This is mainly explained by the opposite temporal phase between river discharge and irrigation water: irrigation water requirement is

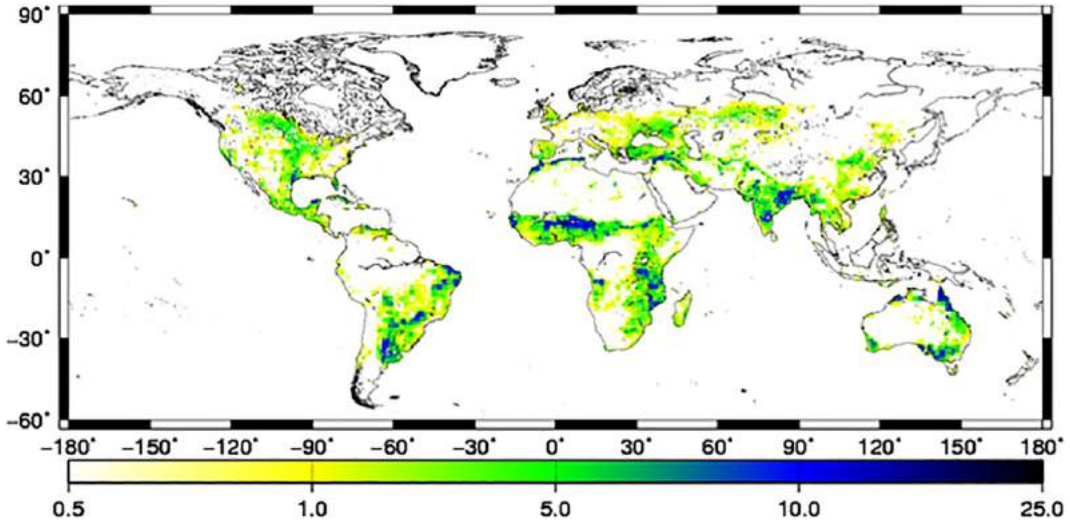


Fig. 5 Irrigation water requirement for bioenergy crops under Scenario S1 in the 2090s ($m^3\ s^{-1}$) (Yamagata *et al.*, 2018).

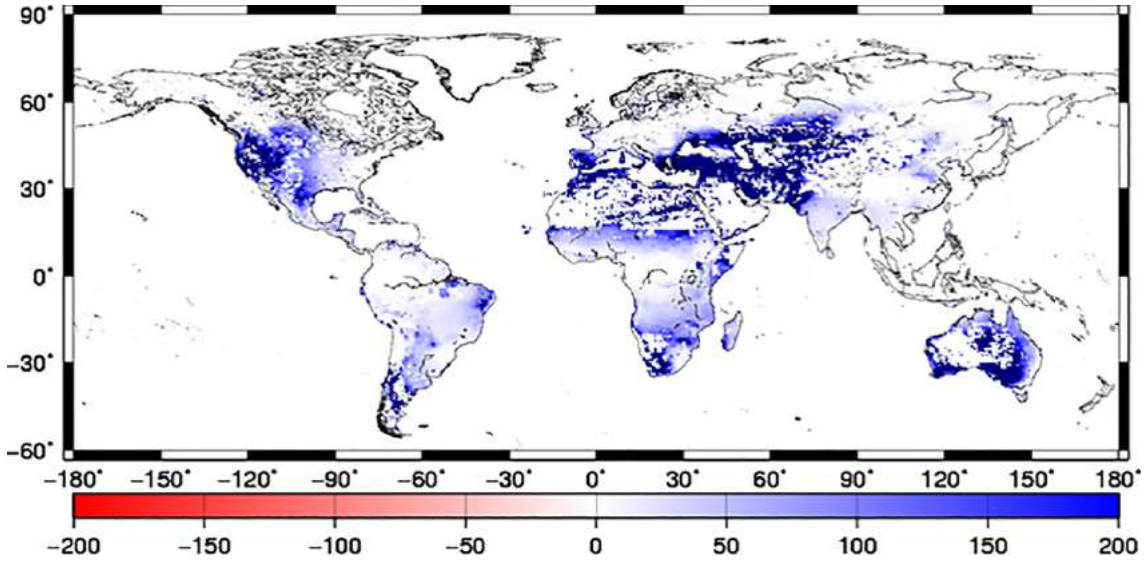


Fig. 6 The effect of irrigation on yield of bioenergy crop or the percentage change of the yield between irrigated and rainfed bioenergy crop in the 2090s [%] (Yamagata *et al.*, 2018).

intensive when the soil gets drier, and the condition also restricts the runoff of the surrounding regions (for further discussion, see Hanasaki *et al.*, 2017).

The production of food in the rainfed cropland is primarily influenced by the cropland area, but it is also affected by climatic change. The production of bioenergy crop was estimated at 8800 106 t for Scenario S1 and 12,300 106 t for Scenario S2. The estimated total bioenergy production goes along with the 3.3 GtC $year^{-1}$ of BECCS or the primary assumption of this study. The relationship between bio-energy crop production and BECCS is expressed as follows (Eq. 5 of Kato and Yamagata, 2014).

The global average yield of bioenergy crop was estimated 35.2 t/ha (with irrigation) and 24.6 t/ha (without irrigation) respectively, indicating that irrigation increased the yield by approximately 50% (i.e., required land use became 2/3). The effect of irrigation, or the fractional change in crop yield due to application of irrigation is shown in **Fig. 6**. The effect is prominent in arid and semiarid regions in western North America, the Mediterranean, southern Africa, Central Asia, western South Asia, and eastern Australia. The abundant irrigation boosted the crop yield in these regions, because limitation in precipitation is the key restricting factor of the crop growth. However, available water from river is quite limited in these regions, which pushed up the fraction of non-river-originated water resources of Scenario S1. Moreover, vast irrigated food cropland area is concentrated in these regions as of today. Further enhancement of irrigation for growing bioenergy crop would likely conflict with food production. This could be a major problem because climate models are projecting a large decrease of food crop yield due to the climatic change even during this century.

Assessment of Land Use Scenarios With Ecosystem Models

Under the S3-1 (no reserved land) and S3-2 (with reserved land) scenarios, a vast area of natural ecosystems is converted to bioenergy crop cultivation. In our simulation, global NPP increased from 56.5 GtC $year^{-1}$ in the 1990s to 65.6 GtC $year^{-1}$ in the 2090s, mainly because of the effects of CO_2 fertilization. Because crops have high productivity comparable with those in forests, the prescribed land-use conversion did not largely affect global total NPP.

In the Yamagata *et al.* (2018), as shown in **Fig. 7**A, global NPP increased almost linearly until around 2060 when it peaked; such trend is in parallel with atmospheric CO_2 concentrations. In each case, terrestrial ecosystems acted as a small net sink of CO_2 including emissions from land-use change and biomass burning, with a considerable range of interannual variability due to climate condition (**Fig. 7**B). Note that this net CO_2 sink is only by ecosystem carbon stock, and sequestration by CCS should be evaluated separately. Carbon stocks in vegetation biomass (**Fig. 7**C) and soil organic carbon (**Fig. 7**D) showed, in our simulation, clear difference among scenarios with a small range of decadal variability. With no land-use change after 2000, vegetation biomass increased from 483 GtC in the 1990s to 544 GtC in the 2090s.

Under the S3-1 scenario, it decreased to 460 GtC due to deforestation in tropical forests. In contrast, under the S3-2 scenario, vegetation biomass increased slightly (495 GtC in the 2090s, a bit lower than 509 GtC of the RCP2.6-based case), indicating the effectiveness of reservation for biodiversity hotspots. Soil carbon stock decreased in the land-use cases to some extent in the early 21st century and then increased gradually due to accumulation in temperate and boreal ecosystems. Here, difference between the results of S3-1 and S3-2 was not so large, because soil carbon stock in tropical rainforests is low and comparable with rangelands. The average rate of net carbon sequestration under S3-2 scenario (about 0.3 GtC $year^{-1}$) is comparable with the present level of terrestrial uptake or climate regulation services including the effect of land-use change (Le Quéré *et al.*, 2016).

Interestingly, these simulations for land-use cases (RCP2.6, S3-1, and S3-2) shows that soil loss by water erosion would be increasing during the 21st century, in contrast with the constant to weak decreasing trends during the 21st century of the fixed land-use case. Because soils provide fundamental support for many ecosystem services, such a loss of soil carbon could result in ecosystem degradation that could have adverse influences on the human society. On the other hand, biomass burning would increase until around 2040 probably due to the increase of fuel supplied from vegetation biomass production.

The terrestrial ecosystem functions are expected to change over the land surface. For example, vegetation biomass is expected to increase in the middle to high latitudes, because plants in these regions would enjoy favorable effects from higher atmospheric CO_2 concentrations and global warming even under the RCP2.6-based climate. In contrast, under the S3-1 scenario, vegetation biomass in lower latitudes such as forests in Amazon Basin, Central Africa, and Southeast Asia is estimated to decrease largely during the 21st century (**Fig. 8**A–C), as a result of land-use conversion for bioenergy crop production. Because these tropical forests support important biodiversity and associated ecosystem services, such an intense biomass decrease should bring about adverse influences on local communities.

Under the S3-2 scenario, biodiversity hotspots such as central Amazon are preserved. However, surrounding natural ecosystems were still seriously affected by land-use for bioenergy production (**Fig. 8**D–E). In Australia, expansion of bioenergy cultivation led to a slight increase of vegetation biomass, but it could not compensate for the massive loss in other forests. In our simulation, the loss of vegetation cover resulted also in deterioration of soil loss due to water erosion. As shown in **Fig. 8**F, the present soil loss by water erosion occurs mainly in mountain areas, croplands, and monsoon Asia with high precipitation. Soils of tropical forests in the Amazon Basin and Central Africa are protected by dense vegetation cover, leading to lower soil loss due to water erosion than high precipitation. The intense land-cover change for bioenergy crop production in the tropics (S3-1) and subtropics (S3-2) could cause serious deterioration of soil loss (**Fig. 8**G–J) accompanied with degradation of vegetation productivity, hydrologic regulation, and biodiversity. Such soil loss also occurred in boreal regions but with lower intensity.

Total BECCS Impacts on Land Systems

By using state-of-the-art global land models and systematic land-use scenarios, a set of comprehensive simulations on the massive production of bioenergy can be conducted. If approximately one eighth of the cropland in 2090s was transferred into cropland for irrigated bioenergy, or in case of S1, 8800 106 t of bioenergy crop together with 9% of increase in total food would be produced instead of 135% of additional water consumption. In case of S2, 12,300 106 t would be produced instead of 5% of reduction in food production (Yamagata *et al.*, 2018).

From the view point of water resources, S1 is challenging because it more than doubles the present water use. Water scarcity is observed in many parts of the world, and further increase in river water would exacerbate the problem. The additional water would be abstracted from other than river water, which implies the need for intensive water resources development (dams, aqueducts, groundwater development).

From the view point of food production, S2 is also challenging because it decreases food production. The global average crop yield of food for S1 and S2 is approximately 7% smaller than the present level (notice that the CO_2 fertilization effect and yield growth in the future are not taken into account in this study), indicating that a considerable improvement in efficiency is needed for food distribution to feed the world.

From the view point of ecosystem services, S3 could be problematic with regard to ecosystem sustainability, because the extensive conversion of natural forests exerts undesirable impacts to ecosystem integrity as it was the case where large-scale

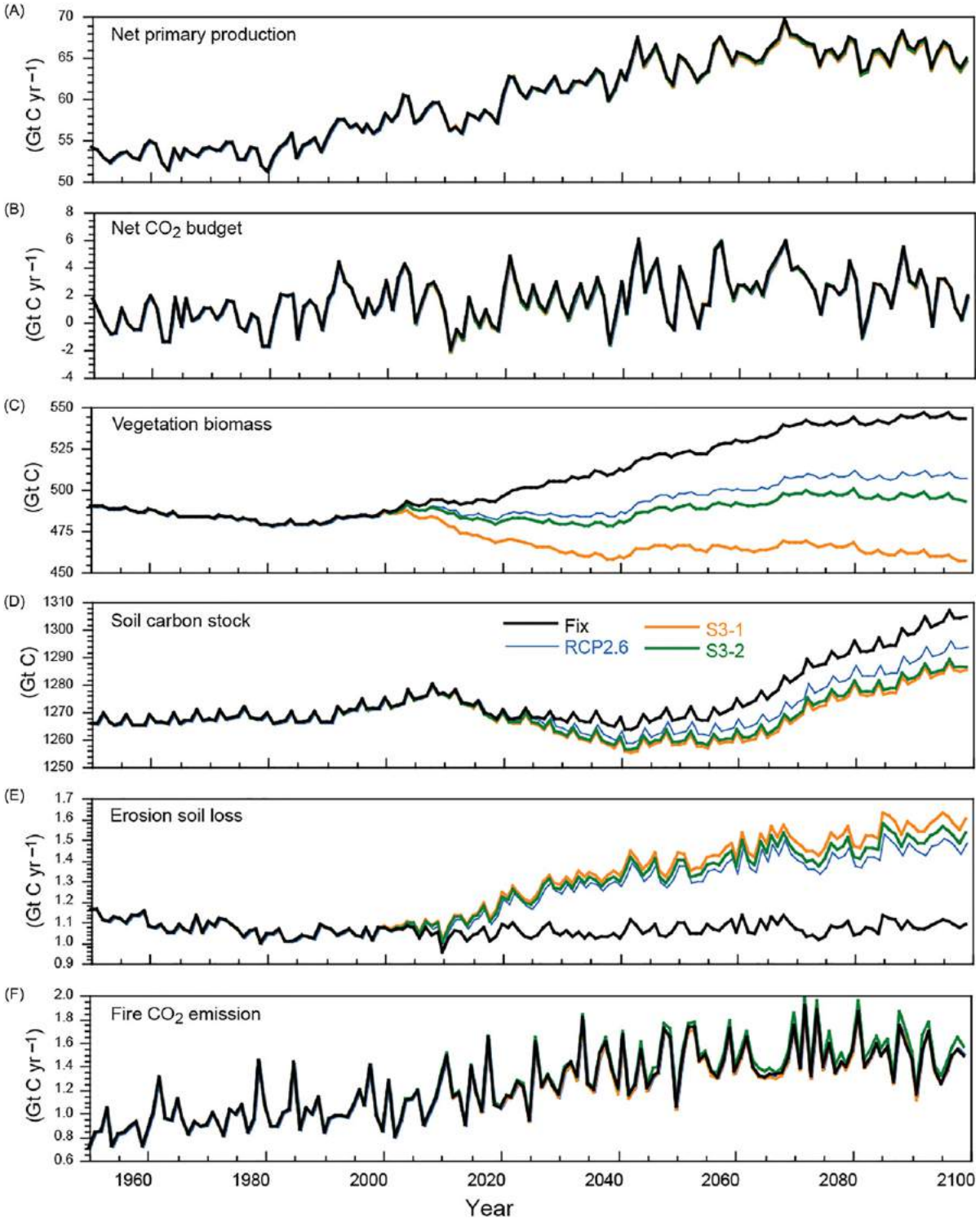


Fig. 7 Time-series of simulated terrestrial properties related to ecosystem services using the S3 scenarios. (A) Net primary production, (B) Net ecosystem CO_2 exchange, (C) Vegetation biomass, (D) Soil carbon stock, (E) Soil loss from erosion, and (F) CO_2 emissions from biomass burning. Thick *black line* shows the result for fixed land-use case, thin blue for RCP2.6-based land-use case, thick orange for S3-1 (no reserved land) case, and thick green for S3-2 (with reserved land) case (Yamagata *et al.*, 2018).

deforestations occurred in the tropical regions due to the rapid expansion of palm oil plantations. Moreover, loss of tropical forests in Amazon and Congo Basin would bring about serious decline of biodiversity in these regions. Soil loss could be caused by exacerbated water erosion due to land-use conversion, if no strong deforestation regulation (such as REDD +) is implemented.

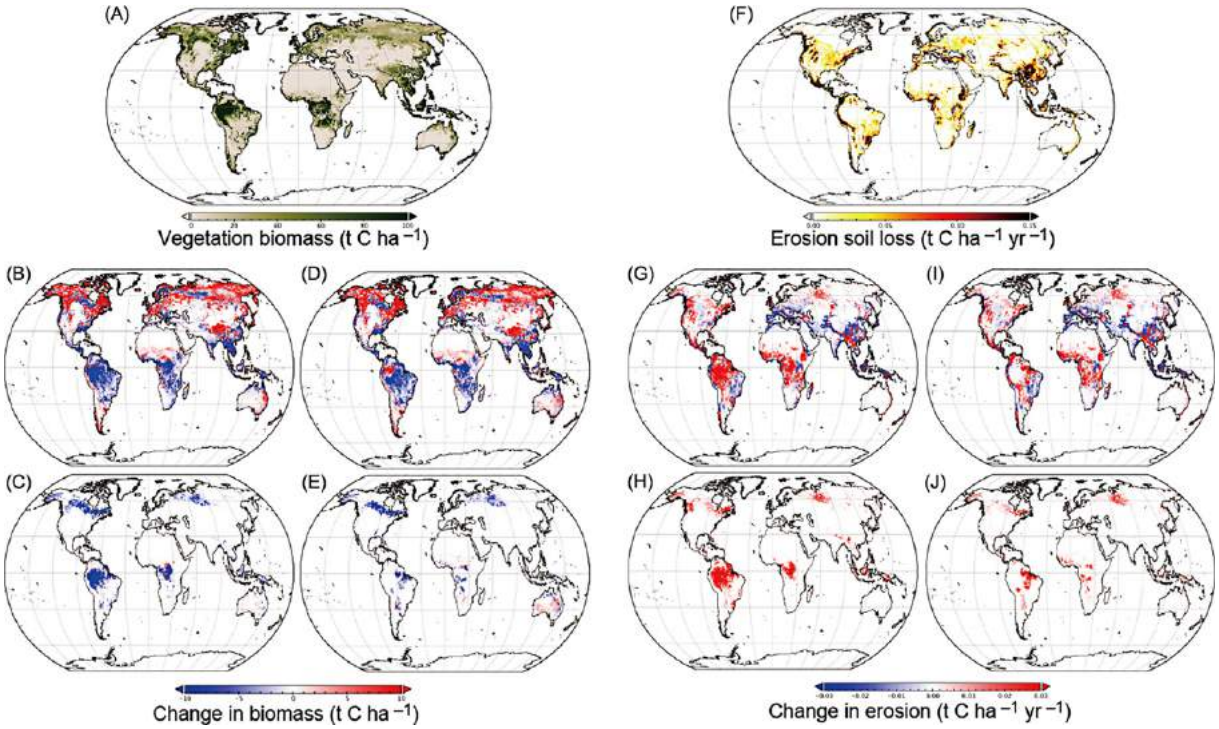


Fig. 8 Maps of simulated terrestrial properties related to ecosystem services and their change in the 21st century based on the S3 scenarios. (A) Vegetation biomass and (F) soil erosion loss in the 2090s, respectively. Difference between the 1990s and 2090s for (B) vegetation biomass under S3-1 (no reserve land), (D) vegetation biomass under S3-2 (with reserve land), (G) soil erosion under S3-1, (I) oil erosion under S3-2, respectively. (C, E, H, J) Difference between the S3-based results in (B, D, G, I) and RCP2.6-based ones, showing the impacts of BECCS deployment (Yamagata *et al.*, 2018).

Large scale bioenergy crop deployment could also cause land degradations such as collapse of ecosystem structure and declined productivity.

It might be over-simplistic to assume that a vast extent of food cropland could be successfully transferred into bioenergy cropland in proportion to the total cropland area. It would likely be more realistic to assume an expansion of bioenergy cropland utilizing the scenarios developed by more sophisticated land use models.

In general, complex trade-offs could occur among land (and ecosystem conservation), water, and food when a massive amount of bioenergy is produced. Internationally coordinated developments of integrated models that can deal with these interactions are urgently needed. For example, trade-offs between biodiversity conservation and land use become complicated when considering the existence of prioritized biodiversity hot-spots (e.g., Myers *et al.*, 2000; Newbold *et al.*, 2016). Our knowledge on ecosystem functions and services are inadequate, such that forests can exert both positive and negative feedback effects on climate change depending on climatic conditions (Betts, 2011). We need to deepen our understanding on specific land processes and their interactions.

Several current land use scenario-based assessments show that, especially in terms of sustainability of ecosystem services, unrestricted expansion of bioenergy crop cultivation at the expense of natural forests will not be feasible, because it can cause serious extensive decline in carbon stock and related ecosystem services, although several regions receive some benefits. In this regard, we should pay more attention to the cobenefits and synergies between the biodiversity conservation and climatic change mitigation activities for optimizing various sustainability benefits.

Spatially Explicit Socio-Economic Scenarios

Another remaining issue for this kind of land use modeling is a more detailed consideration of socio-economic scenarios such as the shared socioeconomic pathways (SSPs; see Riahi *et al.*, 2017). SSPs, which are scenarios describing alternative future developments, consist of the sustainability (SSP1), middle of the road (SSP2), fragmentation (SSP3), inequality (SSP4), and fossil-fueled development (SSP5) scenarios. While we constrained land use assuming SSP2, different scenarios can lead to different conclusions. Besides, while SSPs and other socio-economic scenarios are typically country-level scenarios, regional/local-level socio-economic development is actually influential on water-food-ecosystems and land use in the regions. Use of spatially fine dataset on SSP1-5 would be an interesting next topic.

There are already several studies that have downscaled country-level socioeconomic scenarios into spatially fine scenarios (e.g., Bengtsson *et al.*, 2006; Grübler *et al.*, 2007; van Vuuren *et al.*, 2007; Gaffin *et al.*, 2004; Hachadoorian *et al.*, 2011; McKee *et al.*, 2015; Nam and Reilly, 2013; Yamagata *et al.*, 2015, 2018; Jones and O'Neill, 2016; Murakami and Yamagata, 2016). Especially, as it could influence the future land use drastically, we need to study spatially explicit economic growth (GDP) impacts on food preferences and food security and the trade-offs between water, food, and ecosystems. To support this kind of studies, we have also developed gridded GDP scenarios for SSP 1–3 (Murakami and Yamagata, 2016) in addition to SSP2. Our newly developed socio-economic dataset will be used for the ISIMIP as one of the standardized input datasets. The datasets are downloadable from GCP website (http://www.cger.nies.go.jp/gcp/population-and-gdp.html).

See also: Global Change Ecology: Climate Change 2: Long-Term Dynamics; Global Carbon Cycle 1: Short-Term Dynamics

References

Arnold, J.G., Kiniry, J.R., Srinivasan, R., Williams, J.R., Haney, E.B., Neitsch, S.L., 2012. SWAT input/output documentation version 2012. Texas Water Resources Institute 654, 1–646.

Bengtsson, M., Shen, Y., Oki, T., 2006. A SRES-based gridded global population dataset for 1990–2100. Population and Environment 28, 113–131.

Betts, R.A., 2011. Afforestation cools more or less. Nature Geoscience 4, 504–505.

Bonsch, M., Humpenöder, F., Popp, A., Bodirsky, B., Dietrich, J.P., Rolinski, S., Biewald, A., Lotze-Campen, H., Weindl, I., Gerten, D., Stevanovic, M., 2016. Trade-offs between land and water requirements for large-scale bioenergy production. GCB Bioenergy 8, 11–24. doi:10.1111/gcbb.12226.

Chini, L.P., Hurtt, G.C., Frolking, S., 2014. Harmonized Global Land Use for Years 1500–2100, V1, Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA. doi:10.3334/ORNLDAAC/1248.

Creutzig, D., Agoston, P., Minx, J.X., Canadell, J.G., Andrew, R.M., Quéré, C., Peters, G.P., Sharifi, A., Yamagata, Y., Dhakal, S., 2016. Urban infrastructure choices structure climate solutions. Nature Climate Change 6, 1054–1056.

Döll, P., Hoffmann-Dobrev, H., Portmann, F.T., Siebert, S., Eicker, A., Rodell, M., Strassberg, G., Scanlon, B.R., 2012. Impact of water withdrawals from groundwater and surface water on continental water storage variations. Journal of Geodynamics 59/60, 143–156. doi:10.1016/j.jog.2011.05.001.

Fujimori, S., Masui, T., Matsuoka, Y., 2014. Development of a global computable general equilibrium model coupled with detailed energy end-use technology. Applied Energy 128, 296–306.

Fuss, S., Canadell, J.G., Peters, G.P., Tavoni, M., Andrew, R.M., Ciais, P., Jackson, R.B., Jones, C.D., Kraxner, F., Nakicenovic, N., Le Quéré, C., Raupach, M.R., Sharifi, A., Smith, P., Yamagata, Y., 2014. Betting on negative emission. Nature Climate Change 4, 850–853.

Gaffin, S.R., Rosenzweig, C., Xing, X., Yetman, G., 2004. Downscaling and geo-spatial gridding of socio-economic projections from the IPCC special report one missions scenarios (SRES). Global Environmental Change 14, 105–123.

Grübler, A., O'Neill, B., Riahi, K., Chirkov, V., Goujon, A., Kolp, P., Prommer, I., Scherbov, S., Slentoe, E., 2007. Regional, national, and spatially explicit scenarios of demographic and economic change based on SRES. Technological Forecasting and Social Change 74, 980–1029.

Hachadoorian, L., Gaffin, S., Engelman, R., 2011. Projecting a gridded population of the world using ratio methods of trend extrapolation. In: Cincotta, R., Gorenflo, L. (Eds.), Human population. New York: Springer, pp. 13–25.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., Tanaka, K., 2008a. An integrated model for the assessment of global water resources—Part 1: Model description and input meteorological forcing. Hydrology and Earth System Sciences 12, 1007–1025. doi:10.5194/hess-12-1007-2008.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., Tanaka, K., 2008b. An integrated model for the assessment of global water resources—Part 2: Applications and assessments. Hydrology and Earth System Sciences 12, 1027–1037. doi:10.5194/hess-12-1027-2008.

Hanasaki, N., Inuzuka, T., Kanae, S., Oki, T., 2010. An estimation of global virtual water flow and sources of water withdrawal for major crops and livestock products using a global hydrological model. Journal of Hydrology 384, 232–244. doi:10.1016/j.jhydrol.2009.09.028.

Hanasaki, N., Yoshikawa, S., Pokhrel, Y., Kanae, S., 2017. A global hydrological simulation to specify the sources of water used by humans. Hydrology and Earth System Sciences Discussions 2017, 1–53. doi:10.5194/hess-2017-280.

Hejazi, M.I., Voisin, N., Liu, L., Bramer, L.M., Fortin, D.C., Hathaway, J.E., Huang, M., Kyle, P., Leung, L.R., Li, H.-Y., Liu, Y., Patel, P.L., Pulsipher, T.C., Rice, J.S., Tesfa, T. K., Vernon, C.R., Zhou, Y., 2015. 21st century United States emissions mitigation could increase water stress more than the climate change it is mitigating. Proceedings of the National Academy of Sciences of the United States of America 112, 10635–10640. doi:10.1073/pnas.1421675112.

Hempel, S., Frieler, K., Warszawski, L., Schewe, J., Piontek, F., 2013. A trend-preserving bias correction—The ISI-MIP approach. Earth System Dynamics 4, 219–236. doi:10.5194/esd-4-219-2013.

Ito, A., 2007. Simulated impacts of climate and land-cover change on soil erosion and implication for the carbon cycle, 1901 to 2100. Geophysical Research Letters 34. L09403. https://doi.org/10.1029/2007GL029342

Ito, A., Inatomi, M., 2012. Water-use efficiency of the terrestrial biosphere: A model analysis on interactions between the global carbon and water cycles. Journal of Hydrometeorology 13, 681–694.

Ito, A., Oikawa, T., 2002. A simulation model of the carbon cycle in land ecosystems (Sim-CYCLE): A description based on dry-matter production theory and plot-scale validation. Ecological Modelling 151, 147–179.

Jones, B., O'Neill, B.C., 2016. Spatially explicit global population scenarios consistent with the shared socioeconomic pathways. Environmental Research Letters 11.084003

Kato, E., Yamagata, Y., 2014. BECCS capability of dedicated bioenergy crops under a future land-use scenario targeting net negative carbon emissions. Earth's Future 2, 421–439. doi:10.1002/2014EF000249.

Kato, E., Kinoshita, T., Ito, A., Kawamiya, M., Yamagata, Y., 2013. Evaluation of spatially explicit emission scenario of land-use change and biomass burning using a process based biogeochemical model. Journal of Land Use Science 8, 104–122.

Le Quéré, C., Andrew, R.M., Canadell, J.G., Sitch, S., Korsbakken, J.I., Peters, G.P., Manning, A.C., Boden, T.A., Tans, P.P., Houghton, R.A., Keeling, R.F., Alin, S., Andrews, O.D., Anthoni, P., Barbero, L., Bopp, L., Chevallier, F., Chini, L.P., Ciais, P., Currie, K., Delire, C., Doney, S.C., Friedlingstein, P., Gkritzalis, T., Harris, I., Hauck, J., Haverd, V., Hoppema, M., Klein Goldewijk, K., Jain, A.K., Kato, E., Körtzinger, A., Landschützer, P., Lefèvre, N., Lenton, A., Lienert, S., Lombrardozzi, D., Melton, J.R., Metzl, N., Millero, F., Monteiro, P.M.S., Munro, D.R., Nabel, J.E.M.S., Nakaoka, S., O'Brien, K., Olsen, A., Omar, A.M., Ono, T., Pierrot, D., Poulter, B., Rödenbeck, C., Salisbury, J., Schuster, U., Schwinger, J., Séférian, R., Skjelvan, I., Stocker, B.D., Sutton, A.J., Takahashi, T., Tian, H., Tilbrook, B., van der Laan-Luijkx, I.T., van der Werf, G.R., Viovy, N., Walker, A.P., Wiltshire, A.J., Zaehle, S., 2016. Global carbon budget 2016. Earth System Science Data 8, 605–649. doi:10.5194/essd-8-605-2016.

McKee, J.J., Rose, A.N., Bright, E.A., Huynh, T., Bhaduri, B.L., 2015. Locally adaptive, spatially explicit projection of US population for 2030 and 2050. PNAS 112, 1344–1349.

Melillo, J.M., Reilly, J.M., Kicklighter, D.W., Gurgel, A.C., Cronin, T.W., Paltsev, S., Felzer, B.S., Wang, X., Sokolov, A.P., Schlosser, C.A., 2009. Indirect emissions from biofuels: How important? Science 326, 1397–1399. doi:10.1126/science.1180251.

Misztal, P.K., Nemitz, E., Langford, B., Di Marco, C.F., Phillips, G.J., Hewitt, C.N., MacKenzie, A.R., Owen, S.M., Fowler, D., Heal, M.R., Cape, J.N., 2011. Direct ecosystem fluxes of volatile organic compounds from oil palms in South-East Asia. Atmospheric Chemistry and Physics 11, 8995–9017. doi:10.5194/acp-11-8995-2011.

Monfreda, C., Ramankutty, N., Foley, J.A., 2008. Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. Global Biogeochem Cycles 22.GB1022. https://doi.org/10.1029/2007GB002947

Murakami, D., Yamagata, Y., 2016. Estimation of gridded population and GDP scenarios with spatially explicit statistical downscaling. https://arxiv.org/pdf/1610.09041.pdf

Myers, N., Mittermeier, R.A., Mittermeier, C.G., da Fanseca, G.A.B., Kent, J., 2000. Biodiversity hotspots for conservation priorities. Nature 403, 853–858.

Nam, K.-M., Reilly, J.M., 2013. City size distribution as a function of socioeconomic conditions: An eclectic approach to downscaling global population. Urban Studies 50, 208–225.

Newbold, T., Hudson, L.N., Arnell, A.P., Contu, S., De Palma, A., Ferrier, S., Hill, S.L.L., Hoskins, A.J., Lysenko, I., Phillips, H.R.P., Burton, V.J., Chng, C.W.T., Emerson, S., Gao, D., Pask-Hale, G., Hutton, J., Jung, M., Sanchez-Ortiz, K., Simmonds, B.I., Whitmee, S., Zhang, H., Scharlemann, J.P.W., Purvis, A., 2016. Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. Science 353, 288–291. doi:10.1126/science.aaf2201.

Riahi, K., Van Vuuren, D.P., Kriegler, E., Edmonds, J., O'neill, B.C., Fujimori, S., Bauer, N., Calvin, K., Dellink, R., Fricko, O., Lutz, W., Popp A Cuaresma, J.C., Samir, K.C., Leimbach, M., Jiang, L., Kram, T., Rao, S., Emmerling, J., Ebi, K., Hasegawa, T., Havlik, P., Humpenoder, F., Da Silva, L.A., Smith, S., Stehfest, E., Bosetti, V., Eom, J., Gernaat, D., Masui, T., Rogeli, J., Strefler, J., Drouet, L., Krey, V., Luderer, G., Harmsen, M., Takahashi, K., Baumstark, L., Doelman, J.C., Kainuma, M., Kilmont, Z., Marangoni, G., Lotze-Campen, H., Obersteiner, M., Tabeau, A., Tavoni, M., 2017. The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: An overview. Global Environmental Change 42, 153–168. doi:10.1016/j.gloenvcha.2016.05.009.

Schellnhuber, H.J., Rahmstorf, S., Winkelmann, R., 2016. Why the right climate target was agreed in Paris. Nature Climate Change 6, 649–653. doi:10.1038/nclimate3013.

Schleussner, C.F., Lissner, T.K., Fischer, E.M., Wohland, J., Perrette, M., Golly, A., Rogelj, J., Childers, K., Schewe, J., Frieler, K., Mengel, M., Hare, W., Schaeffer, M., 2016. Differential climate impacts for policy-relevant limits to global warming: The case of 1.5°C and 2°C. Earth System Dynamics 7, 327–351. doi:10.5194/esd-7-327-2016.

Siebert, S., Döll, P., Hoogeveen, J., Faures, J.M., Frenken, K., Feick, S., 2005. Development and validation of the global map of irrigation areas. Hydrology and Earth System Sciences 9, 535–547. doi:10.5194/hess-9-535-2005.

Smith, P., Davis, S.J., Creutzig, F., Fuss, S., Minx, J., Gabrielle, B., Kato, E., Jackson, R.B., Cowie, A., Kriegler, E., van Vuuren, D.P., Rogelj, J., Ciais, P., Milne, J., Canadell, J.G., McCollum, D., Peters, G., Andrew, R., Krey, V., Shrestha, G., Friedlingstein, P., Gasser, T., Grübler, A., Heidug, W.K., Jonas, M., Jones, C.D., Kraxner, F., Littleton, E., Lowe, J., Moreira, J.R., Nakicenovic, N., Obersteiner, M., Patwardhan, A., Rogner, M., Rubin, E., Sharifi, A., Torvanger, A., Yamagata, Y., Edmonds, J., Yongsung, C., 2016. Biophysical and economic limits to negative CO_2 emissions. Nature Climate Change 6, 42–50. doi:10.1038/NCLIMATE2870.

van Vuuren, D.P., Lucas, P.L., Hilderink, H., 2007. Downscaling drivers of global environmental change. Global Environmental Change 17, 114–130.

van Vuuren, D.P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G.C., Kram, T., Krey, V., Lamarque, J.-F., Masui, T., Meinshausen, M., Nakicenovic, N., Smith, S.J., Rose, S.K., 2011. The representation concentration pathways: An overview. Climatic Change 109, 5–31. doi:10.1007/s10584-011-0148-z.

Yamagata, Y., Murakami, D., Seya, H., 2015. A comparison of grid-level residential electricity demand scenarios in Japan for 2050. Applied Energy 158, 255–262.

Yamagata, Y., Hanasaki, N., Ito, A., *et al.*, 2018. Estimating water–food–ecosystem trade-offs for the global negative emission scenario (IPCC-RCP2.6). Sustainability Science 13, 1–13. doi:10.1007/s11625-017-0522-5.

Further Reading

Dellink, R., Chateau, J., Lanzi, E., Magné, B., 2017. Long-term economic growth projections in the shared socioeconomic pathways. Global Environmental Change 47, 200–214.

Ramankutty, N., Evan, A.T., Monfreda, C., Foley, J.A., 2008. Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000. Global Biogeochemical Cycles 22. doi:10.1029/2007GB002952.

Information and Information Flows in the Biosphere

PJ Georgievich, Russian Academy of Sciences, Moscow, Russia

Information is a concept intuitively clear to everybody and quite correctly associated with knowledge or – which is similar in meaning – with eliminating uncertainties. Knowledge is naturally considered to be useful as it increases efficiency of person's activities, ensures his better adaptability to changing environments, and therefore enhances his vital capacity and sustainability. In fact, however, that is not always so and excessive knowledge may be dangerous. Whether the knowledge appears to be constructive or destructive for an entity (subject) depends mostly on the subject's own state or own structure. When turning our own attention to ourselves, we could confidently assert that information (knowledge) obtained is capable of making changes in organization (structure, order, regulations) of our thoughts, organization of technological processes, engineering structures, social communities, etc., the latter gaining in efficiency and stability in the process.

C. Shannon worked out the law of the rate of information transfer from transmitter to receiver over a channel of any physical nature with noise. The law has been derived from a random process model and has direct analogy with models of thermostatics and therefore with models of diversity.

The capacity of a channel of band W perturbed by white thermal noise power N when the average transmitter power is limited to P is given by

$$C = w\log\left(1 + \frac{P}{N}\right) \qquad [1]$$

where $N = wN_0$ (N_0 is the power of noise on unit of a band of frequencies).

It is easily seen that this law is nothing else than a logarithmical form of allometric relationship which widely occurs both in living organisms and in inorganic nature. By conditions, the law of information transfer gives rise to a fractal set. In linguistics, the frequency band is related with the alphabet length, and the signal power with the length of a word. For biological systems, the frequency band is associated with the specialization level (the narrower is the band, the more specialized is the system). The signal power (dispersion) depends on the environment strength (energy) and diversity. There is a linear dependence between noise and frequency band – the narrower is the band, the less are errors, though communication channel capacity decreases accordingly. Assuming that there is some probability of errors becoming lethal on accumulation, the individual stability of the receiver (in terms of error-free operation) would increase with narrowing of the frequency band (increasing specialization). Taking improvement of stability to be a target function of evolution, we come to a conclusion that specialization is a natural way to this target. A specialized system, however, has a lower channel capacity and therefore lesser resistance to fluctuations in environments. If we assume, for example, that a population of organisms should have a certain minimum of diversity, it is easy to see that the most specialized and least fertile organisms are likely to inhabit environments of tropical rainforests, while the least specialized organisms of maximum fertility would be found under conditions of cold climate of taiga and tundra, and partly in deserts. This dependence is a matter of common knowledge. It follows from the law of communication channel capacity that there exists a limit of information amount that can be transmitted per unit of frequency band equal to 1.443 natural units. All the limitations bring us to the conclusion that no supersystem can exist that could receive information within an arbitrary large frequency band; a number of receiving systems (mutually complementary by alphabet) are necessary for effective transformation of the information. Hence it immediately follows that a diversity is necessary in the receiving system, and the more powerful is the transmitter, the greater is the number of various receivers needed for complete transformation of information. Under certain simple assumptions, it may be inferred from eqn [1] that the diversity is given approximately by

$$\text{Number of species } S = aN^b$$

where $b < 1$ and N is sample size and $N = f$(area, habitat capacity). This is identical to the relationships derived from thermostatics. A connection between the quantitative information model and thermostatic model is determined by their common mathematical basis: information is defined as the inverse of the entropy.

If a noisy channel is fed by a source, there are two statistical processes at work: the source and the noise. Thus there are a number of entropies that can be calculated. First there is the entropy $H(x)$ of the source or of the input to the channel (these will be equal if the transmitter is nonsingular). The entropy of the output of the channel, that is, the received signal, will be denoted by $H(y)$. In the noiseless case $H(y) = H(x)$. The joint entropy of input and output will be $H(xy)$. Finally, there are two conditional entropies, $H_x(y)$ and $H_y(x)$ – the entropy of the output when the input is known and conversely. Among these quantities, we have the relations $H(x,y) = H(x) + H_x(y) = H(y) + H_y(x)$ All of these entropies can be measured on a per-second or a per-symbol basis. The rate of transmission I can be written in two other forms due to the identities noted above. We have $I = H(x) - H_y(x) = H(y) - H_x(y) = H(x) + H(y) - H(x,y)$.

Entropy differs from diversity in that a quantity of information within a closed system transmitter increases, and not decreases, with time, as uncertainty of the transmitter decreases with time and its behavior may be more reliably predicted by the receiver. If, however, the transmitter is an open system, its uncertainties do not depend generally on the duration of transmission, and its

 doi:10.1016/B978-0-444-63768-0.00742-3

behavior keeps up at a constant level of unpredictability. As the model of the communication channel capacity is homologous to the thermostatic model, information may be considered a phenomenon of universal occurrence. It is the transmission of information from environment to any object within a certain frequency band that controls the existing order or structure of the object. Even in case of ceasing external action or transfer of information from outside, the structure appears steady for a long time in the existing environments. In that case, there is a good reason to speak of stored information.

The rate of information flow received by the system within a certain time interval may be estimated in terms of difference in diversity at the moments of time under comparison. Information may also be measured by Kulback entropy in comparison with a diversity under conditions of equilibrium or steady (stationary) state. There is practically no study aimed at measurement of the quantitative information flows. There was a rather keen interest in information theory as applied to natural sciences, and to biology in particular, in the 1950s and the 1960s. One of the sections of information theory (i.e., theory of coding) made a considerable contribution to solving problems of genetic code and molecular synthesis. Limited possibilities for measurements and inadequate equipment hindered fruitful application of information theory for ecological research. Though a connection between information theory and thermodynamics was evident as early as the 1950s, a real integration of the two branches became possible only on a basis of developed theory of nonequilibrium thermodynamics and synergetics. All the above accounts for an exponential growth of published papers dealing with the considered problem during the last 10–15 years. The studies are mainly focused on explaining the evolution of both living matter and human society.

Evidently, the law of quantitative transfer of information does not cover all the aspects of what we instinctively associate with knowledge. A signal received may be meaningless in the receiver's perception and would not change its state, and, vice versa, a signal of negligible strength may induce drastic changes. Accordingly, information includes both quantitative and semantic components. In the simplest case, the latter may be dealt with in terms of decoding of signals coming from transmitter to receiver. It implies that there is an outside observer who establishes rules of decoding, and records signal characteristics at the input and consequent changes in the receiver state. Formally, it is a problem of statistic analysis aimed at a search for invariants with respect to signal receiver toward the transmitter. This important and by no means trivial problem of biosemiotics is related to partial interaction analysis and is potentially capable of simulation of all possible partial relations. However, its solution does not necessarily give an insight into the problem at the macroscopic level.

It should be stressed that, as follows unambiguously from our experience, an interaction between two systems may produce some new systems, and structure and properties of the latter may appear completely unpredictable, even if we have a complete knowledge of the initially interacting systems (emergence). Generally, it is impossible even to define a set of possible outcomes, that is, expected uncertainty. Therefore, the appearance (emergence) of a new, earlier unknown structure may be defined as origination of new information. The only condition for it is some energy input to the system. On the other hand, any former locally steady structure may disappear, together with related information. It seems conceivable, therefore, that conservation laws do not apply to information at the macroscopic level. Being a measure of order, information arises from chaos and returns to it; the evolution based on memory (selection of locally stable structures) proceeds by progressive retrieval of order and accumulation of information. Open macrosystem receives energy in various forms from conventionally separated environment and generates flow of information and its continuous increase.

Actually, it is this phenomenon that brings us to change our understanding of entropy as a measure of disorder; it makes us revise the classic thermodynamic model (that admits only mechanical forms of energy conversion), thus eliminating discrepancy between the observed evolution and the second principle of thermodynamics. There are numerous researches dealing with this problem. More than 60 monographs have been recently published by Springer-Verlag publishers. S. D. Khaitun, in particular, gives a meticulous review of existing opinions on thermodynamic irreversibility and concludes on advisability of coming back to wording of the second principle as stated by W. Thomson (Lord Kelvin); according to the latter, mechanical energy dissipates (depreciates) in the course of irreversible processes – its amount decreases when passing into other kinds of energy. It is a mechanical approach, where all the processes in the system are described by movement of constituent particles and its state is exhaustingly characterized in terms of coordinates and impulses so that the energy appears to be their function. Mechanical energy differs from nonmechanical in that its movement may be completely described by a set of coordinates and impulses; in other words, the energy is described by the Hamiltonian function. Nonmechanical energy is related to entropy information, because a part of the energy is spent for new structure synthesis and maintenance and for synthesis of new information. When considered together with the law of the information transfer rate over a communication channel, the results bring us to a conclusion that power (energy) of any external action is spent partly for synthesis of some elements of known type and partly for creation of new structures, with unknown characteristics; those enlarge the band (where the external actions are reproduced) and reduce the noise level in every individual case of the information reception.

S. E. Jorgensen and Yu. M. Svirezhev introduce information into a biological system through Kulback entropy, the latter being a measure of the system deviation from the stationary state. In their model, the system evolution is governed by consumed energy and inner order generated by the system itself and controlling the exergy (useful work). The evolution is aimed at increase in exergy, that is, at a synthesis of structures far from equilibrium or stationary state. Demonstrating a fact of information synthesis, A. M. Khasen supplemented the nonequilibrium thermodynamics model developed by I. Prigogine. He considered entropy information as a function of complex variables, which permitted to recognize it in two constituents, namely basic information and semantic information. The expanded model generates new structures and increases entropy information within the self-developing and self-organizing system. It would be natural to suggest that a constant analogous to Boltzmann constant (length of a word or width of frequency band) appears as a function of self-development and creates a hierarchy (of the word–phrase–paragraph type).

The hierarchy arises from a limited transmission capacity at a currently accepted level of energy transformation. Increase of the transmissivity is due to self-organization of the synthesized systems into systems of the next (higher) level, with narrower frequency band. Accordingly, the number of hierarchic levels increases with total signal strength, while diversity decreases at every higher level.

The chosen descriptive (qualitative) models of information synthesis and transformation, in common with other analogous models, predict an exponential growth of information in biosphere and therefore `cancel' a danger of the `heat death' imminent according to the second principle of thermodynamics. Within the frame of those models, the biosphere is considered a system of a practically unlimited growth of information complexity. That does not mean that individual elements cannot fail; but every lost element would be replaced by two to four new ones, so that the rate of diversity synthesis grows progressively. It should be noted, however, that there is no universally accepted model of information processes in the biosphere; at present, we can only speak about a search for an adequate theory.

Summary

At present, there is no general information theory. In the theory of quantitative information, it is described as elimination of uncertainty. This definition implies existence of a finite set of possible states or relationships and their prior probabilities. In a more comprehensive sense, information is understood as the appearance (emergence) of order or structure with unknown characteristics from the chaos. In that case, there is no closed set of states or their prior probabilities. The information may be measured post factum, for example, in terms of distance between the emerged structure and its stationary analog, by some other means. There exists a distinct trend toward inclusion of information into thermostatic model as a missing variable which controls evolution and its irreversibility. The very fact of living matter evolution (including evolution of human beings) demonstrates that in the course of time the set of its stages gains in power and new locally stabilized systems appear; they become more and more complicated and require increasing flow of energy for their maintenance. A growth of the consumed energy flow is compensated by enhanced total transmission capacity. A great problem in synthesis of new structures consists of balance between the memory controlling admissible variants of new structures (targeted evolution) and environmental influence either through selection or by way of direct or indirect perception of its properties by the evolving object. Under actual conditions, the information flows are measurable, though an experience in such measurements is rather scarce.

Further Reading

Ashby, W.R., 1956. An Introduction to Cybernetics. London: Chapman and Hall.
Jorgensen, S.E., Svirezhev, Yu M., 2004. Towards a Thermodynamic Theory for Ecological Systems. Amsterdam: Elsevier Science, 366pp.
Khaitun, S.D., 1996. *Mechanika I Neobratimost* (Russ.) (Mechanics and Irreversibility). Moscow: Janus, 445pp.
Khazen, A.M., 2000. *Razum Prirodi I Razum Cheloveka* (Russ.) (Nature's Intelligence and Intelligence of Man). Moscow: Mosobluprpoligrafizdat (ISBN 5-7953-0044-6), 608pp.
Shannon, C.E., 1948. The mathematical theory of communication. Bell Systems Technology Journal 27, 379–423. 623–656.
Shannon, C.E., 1949. Communication in the presence of noise. Proceedings of the Institute of Radio Engineers 37, 10–21.

Iron Cycle

KA Hunter and R Strzepek, University of Otago, Dunedin, New Zealand

Introduction

This article presents a scientific overview of the biogeochemical cycling of iron in the ocean, focusing in particular on what is currently known about the importance of this element as a micronutrient for the growth of oceanic phytoplankton. The first section focuses on the basic biochemistry of iron in phytoplankton metabolism, followed by consideration of the biogeochemistry of this element and how this affects its chemical speciation and bioavailability. The remaining sections deal with large-scale experiments involving iron enrichment in the ocean and the mechanisms that phytoplankton have developed to acquire iron for metabolic processes.

Why Is Iron Important to Phytoplankton?

It has long been known that iron is an essential element for the metabolism of many organisms, including humans. Iron is one of the most abundant elements in the Earth's crust, and the Fe(II)–Fe(III) redox couple provides for facile electron-transfer reactions:

$$Fe^{3+} + e^- \rightarrow Fe^{2+}$$

$$E^0 = 0.77\ \text{V}$$

As a result, iron-containing enzyme proteins are among the most common electron-transfer catalysts (**Table 1**).

Of particular importance to photosynthesizing algae are the photosystem proteins which are involved in the splitting of water to form O_2 and which contain a number of Fe redox centers. Prokaryotic phytoplankton (microbes lacking a cell nucleus), which evolved early on in the evolution of the ocean during the Archean period at least 2.5 billion years ago, used iron as the basis of these redox catalysts because of the high abundance of iron in the early ocean. In the absence of free O_2, iron forms the quite soluble form Fe(II) and probably had concentrations on the order of 1 mM during Archean times.

Ironically, photosynthesis began to win out over other biological strategies, and free O_2 became increasingly available during Proterozoic times. This change in the oxygen status of the ocean had serious consequences for the photosynthesizing algae, for it led to the oxidation of Fe(II) to the much less soluble form Fe(III). During the latter stages of the Archean, this gave rise to immense deposits of Fe(III) oxides on the seafloor known as the `banded iron formations' (BIFs). The banded nature of BIFs resulted from periodic cycles of `boom and bust' as the algae coped with the episodic delivery of Fe(II) through upwelling from the anaerobic deep ocean and its subsequent oxidation by the O_2 generated by the algae. This process resembles a giant titration of the Earth's Fe (II), both in the ocean and on the land, by the algal waste product O_2.

Abundance and Sources of Fe in the Ocean

Eventually, the modern ocean evolved about 1.2 billion years ago in which O_2 of photosynthetic origin permeated almost all of the ocean's depth, not to mention the atmosphere, laying the foundation for the rich and complex biological systems we know today. Under these conditions, iron has become a very rare trace element in the oceans, having concentrations in most regions of the order of 1 nM or less, except in coastal and estuarine regions under the influence of terrestrial runoff.

In surface waters, where Fe is needed for phytoplankton growth, the lowest Fe concentrations are found in those oceanic regions most remote from land. More specifically, away from the direct runoff of Fe in rivers, the main external source of Fe entering oceanic surface waters in the modern Earth system is soil-derived dust transported over great distances from the arid areas of the Earth's surface (**Table 2**). Of particular importance are the Sahara and Sahel desert regions which deliver Fe to the equatorial and North Atlantic Ocean, and the Asian deserts which are a major source for the western North Pacific Ocean. The Southern Ocean contains no major dust sources other than the desert regions of Australia and Patagonia well to the north. Not surprisingly, this region turns out to be particularly depleted in iron.

Iron Limitation and Iron-Enrichment Experiments

Less than two decades ago it became clear that the low abundance of iron in certain remote areas of the surface ocean represented an important limitation to phytoplankton growth. Over most of the temperate and tropical latitudes of the world's oceans, the main factor controlling phytoplankton growth rates is thought to be the availability of the nutrients nitrate and phosphate. The

Table 1 Some iron-containing enzyme proteins and their functions

Cytochromes	*Photosynthetic and respiratory e^- transfer*
Cytochrome oxidase	$O_2 + 4H^+ + 4e^- \rightarrow 2H_2O$
Fe-superoxide dismutase	$O_2^- + 2H^+ \rightarrow H_2O + O_2$
Catalase	$2H_2O_2 \rightarrow 2H_2O + O_2$
Peroxidase	$R(OH)_2 + H_2O_2 \rightarrow RO_2 + 2H_2O$
Ferredoxin	e^- to $NADP^+$, NO_3^-, SO_2, N_2, thioredoxin
Succinate dehydrogenase	$FAD + \text{succinate} \rightarrow FADH_2 + \text{fumarate}$
Nitrate reductase	$NO_3^- + 2H^+ + 2e^- \rightarrow NO_2^- + H_2O$
Nitrite reductase	$NO_2^- + 8H^+ + 6e^- \rightarrow NH_4^+ + 2H_2O$
Nitrogenase	$N_2 + 8H^+ + 6e^- \rightarrow 2NH_4^+$

Table 2 Annual flux of dust delivered to different ocean basins

Region	*Dust flux (Tg yr^{-1})*
N. Pacific	480
S. Pacific	39
N. Atlantic	220
S. Atlantic	24
N. Indian	100
S. Indian	44
Global	910

concentrations of both of these nutrients are extremely low in such waters, having been efficiently consumed by plankton. Indeed, as originally postulated by Arthur Redfield, the molar ratio of nitrate to phosphate in the global ocean is remarkably constant at about 15:1, almost exactly the same as the requirements of phytoplankton for these elements. This constant ratio is probably maintained by a balance between phosphate availability and the more biochemical alternative of nitrogen fixation that is available to nitrogen-fixing plankton.

However, in certain areas of the ocean, these nutrients remain at high residual concentrations, suggesting that another factor has come into play as a limitation on growth. These regions, which are characterized by high nutrient but low chlorophyll (HNLC), became strikingly obvious once both detailed surface maps of nutrients became available (starting with the pioneering GEOSCS Program in the 1960s and later through the programs JGOFS and WOCE) (e.g., **Fig. 1**) and also the ability to map surface water chlorophyll concentrations using satellites such as the Coastal Zone Color Scanner (**Fig. 2**). A comparison of these figures indicates that relatively low chlorophyll concentrations are found in regions of high nutrients, especially the Southern Ocean HNLC region.

The late John Martin, of Moss Landing Marine Laboratory in California, first made the suggestion that a lack of iron inhibited phytoplankton growth in these HNLC waters. He did this using incubation experiments in which seawater samples were inoculated with small additions (1–2 nM) of iron. Several days after inoculation, considerable increases in chlorophyll and plankton growth rate were observed compared to controls. A key to the success of these experiments was the ability to collect and handle seawater samples under scrupulously clean conditions that minimized the influence of dust contamination introduced by the experimenter. From these results, Martin speculated that iron was the growth-limiting factor in HNLC waters. He also went on to claim that periods of enhanced growth during glacial times might have been a result of enhanced dust input during more arid glacial climates. Periodic inputs of such dust are recorded in the polar ice core record, and seem to correlate well with periods of low atmospheric CO_2, consistent with enhanced plankton growth.

In spite of these convincing arguments, there were many skeptics. A major criticism centered on the artificiality of the small bottle incubation experiments. Grazing is also an important controlling factor on phytoplankton populations, and small bottles would not contain a sufficient population of the larger grazers. This criticism was settled by several mesoscale iron-enrichment experiments initiated in the mid-1990s. In these, a large area typically $8 \times 8\ km^2$ was fertilized with several tonnes of iron (as $FeSO_4$) along with an inert tracer SF_6 to mark the patch of iron-fertilized water. The first two experiments, IronEx I and II, took place in the equatorial Pacific Ocean, which is mildly HNLC. However, in 2002 a group of NZ and British scientists conducted the Southern Ocean Iron Enrichment Experiment (SOIREE) in the HNLC waters of the Southern Ocean south of Tasmania.

In these experiments a dramatic increase in chlorophyll as a result of a phytoplankton bloom was observed several days after the initial infusion of iron. This was accompanied by a decrease in the CO_2 equilibrium partial pressure in the water, indicating biological uptake of CO_2 by plankton. More detailed examination showed that the main beneficiaries of the added iron, and thus the main source of the new chlorophyll, were large pennate diatoms such as *Fragilieria kergulensis*. These are not the dominant organisms under normal, low-Fe conditions. All other things being equal, the best strategy for surviving under limited iron

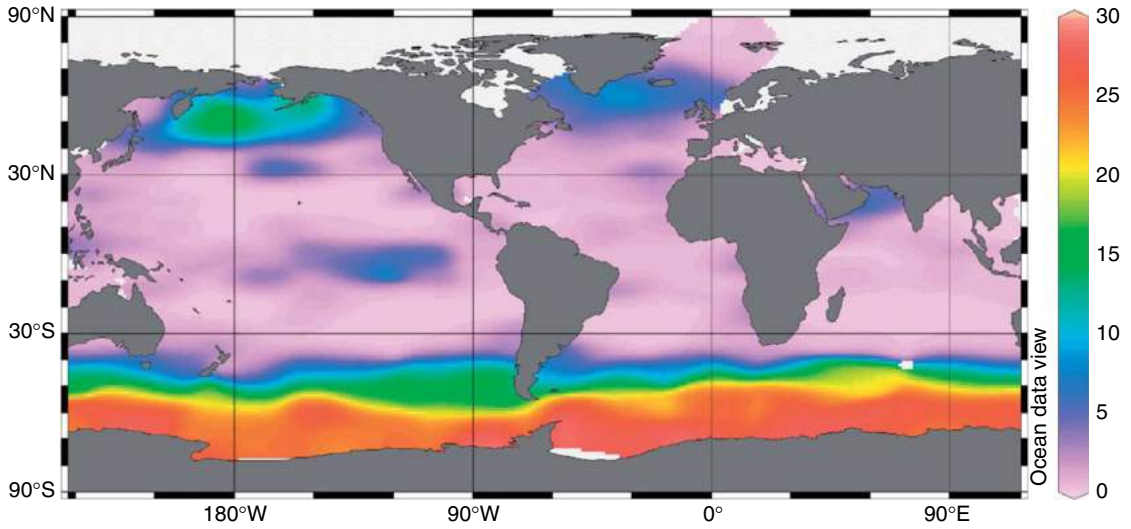


Fig. 1 Map showing the annual mean concentration of nitrate in ocean surface waters. Drawn by the authors using data collected during the World Ocean Circulation Experiment (WOCE).

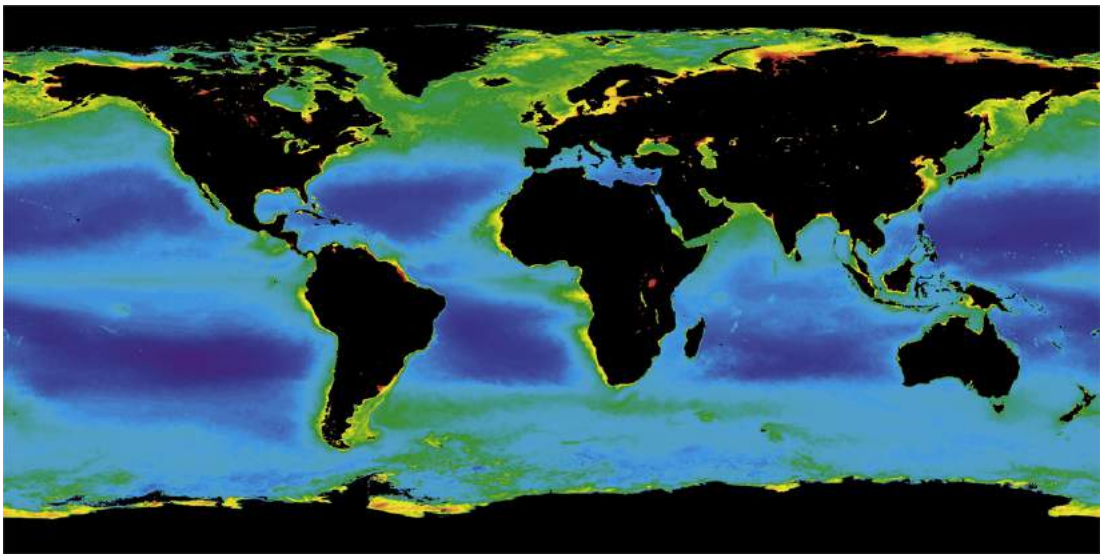


Fig. 2 Satellite map showing the annual mean chlorophyll concentration of ocean surface waters (blue indicates low values; red indicates high values). Provided by the SeaWiFS Project, NASA/Goddard Space Flight Center, and ORBIMAGE.

conditions is to have as small a cell as possible. The SOIREE iron-induced bloom was particularly intense, and evidence was still visible in chlorophyll satellite images up to 55 days after the initial iron infusion.

The Geritol Fix

These iron-enrichment experiments engendered considerable interest both inside and outside the scientific community. They raised the possibility of bioengineering of the oceanic ecosystem as a palliative measure against rising levels of fossil fuel CO_2 by the controlled addition of iron to HNLC areas of the ocean. This became known as the `Geritol fix', named after a popular iron-containing tonic for `tired blood' that was popular many decades ago.

Initially, this idea looked promising. Calculations based on the biological response in these enrichment experiments indicated that a single Fe atom could theoretically initiate the uptake of many thousands of CO_2 molecules. This means that to sequester billions of tonnes of fossil fuel carbon (the current global input from fossil fuels) might only require a few million tonnes per year of iron. This is a very small fraction of the total amount of iron smelted each year. Iron is, after all, an extremely abundant element in the Earth's crust. It has even been reported that John Martin quipped ``Give me half a tanker of iron and I'll give you an ice age.''

However, things are not that simple. It is not sufficient for iron enrichment to stimulate new CO_2 uptake through additional phytoplankton growth. It is also necessary for the biological carbon sequestered in this way to survive respiration long enough to sink out of the mixed layer, thus removing the sequestered carbon from the ocean–atmosphere system. Once `pumped' into deep water in this manner, the sequestered carbon will not return to equilibrate with the atmosphere for 1000–2000 years, the turnover time of the deep water circulation system.

Thus, experiments were conducted to measure the flux of biological carbon sinking into deep water as a result of iron enrichment, and here things began to look less promising. In the initial IronEx experiment, some sinking of carbon into deep water was observed, but this may have merely been a result of subduction of the water itself. In the remaining experiments, especially during SOIREE, no increase in the flux of biological carbon to deep water was observed. This was in spite of the fact that the majority of organisms that bloomed were relatively large diatoms.

In reality, the sinking of biological carbon into deep water is a much more complex process because an entire food web is involved. Much of the carbon flux is mediated by grazing zooplankton which produce large, rapidly settling fecal pellets. During artificial experiments like SOIREE, the principal grazers of the large diatoms are probably not very abundant before iron infusion stimulates the rapid growth in numbers of their prey, and the predominant grazers may have been too small to take advantage of a bloom of very large phytoplankton. However, if iron infusions were carried out on a semicontinuous basis, who knows what permanent changes to the food web might be induced? Although this offers tantalizing benefits for mitigating climate change, it does seem to be a very dangerous experiment.

Speciation and the Bioavailability Conundrum

Not surprisingly, the discovery of the importance of iron in regulating plankton productivity in HNLC areas of the ocean stimulated a renaissance of interest in the marine chemistry of this element. Very quickly, new knowledge began to emerge that made out understanding of this complex situation even more difficult. As already mentioned, iron is very difficult to measure accurately at the very low concentrations observed in seawater, and even now there is no universal agreement on its distribution in ocean waters. This is in spite of some carefully designed intercalibration experiments that have attempted to sort out the best experimental methods for sample collection, handling, and analysis. Nonetheless, some features are now clear.

As mentioned, in the modern ocean iron is present mostly as Fe(III); this oxidation state is very insoluble in seawater at its normal pH of about 8 because of the very insoluble hydroxide $Fe(OH)_3$. Careful laboratory measurements using purely inorganic salt solutions suggest that at this pH the solubility of $Fe(OH)_3$ is about 0.2 nM. Yet the so-called `dissolved' Fe concentrations, measured using filtered samples of seawater, are invariably up to 3–4 times higher, even in remote regions. One reason for this discrepancy is that Fe(III) readily forms colloidal particles of $Fe(OH)_3$ which are small enough to pass through most filters, thus masquerading as `dissolved' Fe. However, very small ultrafilters can be used to eliminate a lot of the colloidal fraction, but even then the concentrations of the apparently soluble fraction still exceeds the theoretical solubility limit of 0.2 nM.

We now know that this is a result of the interaction of Fe(III) with natural organic matter (NOM) dissolved in seawater which form coordination complexes with NOM ligands. A number of very sensitive techniques are now available to probe the nature of these NOM complexes, and while there is some variation in the reported results, some general trends are clear. Seawater appears to universally contain an excess of NOM ligands that bind Fe(III), some of which are extremely strong in a thermodynamic sense (large equilibrium constant for formation). In surface waters, there is mounting evidence that the main NOM ligands are of direct biological origin, similar to the `siderophore' compounds known to be produced by certain terrestrial microorganisms as a mechanism to sequester iron in, for example, soil waters. Iron-binding NOM persists throughout the oceanic water column, and it has been estimated that as a result of their presence, the total oceanic inventory of Fe(III) is raised by a factor of at least 4 over the solubility limit. Clearly this is very important, especially for phytoplankton growing in HNLC areas such as the Southern Ocean, where the main Fe supply may well be the upwelling of deep waters rich (relatively speaking) in NOM-bound iron.

Increasing the solubility of dissolved Fe is advantageous only if the Fe bound to NOM can then be rapidly taken up and released as inorganic Fe inside the cell. This may not be a problem for marine prokaryotes. Heterotrophic bacteria and cyanobacteria isolated from marine habitats also produce siderophores when Fe-limited, some of which have been isolated and chemically characterized. Moreover, marine bacteria transport Fe bound to siderophores regardless of whether or not they produce their own. Little is known of the mechanism by which marine bacteria obtain siderophore-bound Fe, but there is evidence that its fundamental features resemble those of terrestrial bacteria, which possess outer-membrane receptors that transport a wide range of intact Fe(III)–siderophore complexes through the cell wall.

However, the binding of Fe by NOM ligands generates a puzzling conundrum for marine eukaryotic phytoplankton. For these organisms, the principal effect of the formation of a coordination complex by a metal ion with NOM ligands is considered to be a `reduction' in bioavailability. In this paradigm, in order for a metal ion to become available for cellular uptake, it must first dissociate from the NOM complex and become converted into a kinetically available inorganic form such as the free ion Fe^{3+} or its complexes formed with simple ligands such as OH^- or Cl^-. Only these forms are considered kinetically accessible to ion-uptake mechanisms on the cell wall.

This is why the chelator ethylenediamine tetraacetic acid (EDTA) is added to many culture media. Without it, the metal ions present as impurities in the salts used to prepare the media would be far too toxic for any phytoplankton to grow. Similarly, chelators like EDTA are used to strip metal ions like Pb^{2+} when people suffer from lead poisoning.

The conundrum is that Fe, a biologically essential element in drastically short supply in HNLC areas, appears to be bound up by NOM that ought to make it unavailable to much of the phytoplankton community. Worse still, the NOM appears to be of biological origin. So what is really going on with Fe(III) and the NOM complexes it forms? It does not make sense that phytoplankton, already struggling with a lack of iron supply, should synthesize iron-binding compounds like siderophores unless the formation of Fe(III) complexes by these materials actually assists them in acquiring iron. That implies that they have some specific mechanisms on the cell surface for unlocking Fe bound by NOM. In support of this, some very elegant culture experiments using radiolabeled Fe conducted on board ship made it clear that oceanic plankton from HNLC areas were able to take up iron much faster than it could possibly dissociate from NOM complexes to form readily available inorganic forms of Fe(III).

However, at the time of writing, we have no clear idea how this works. One possibility is that photochemistry may play a role. Fe(III)-containing complexes can be photochemically reduced to Fe(II) in seawater, in which form the Fe is much more biologically available. However, although recent work has shown that Fe(II) is generated during daylight hours in seawater, the amount of Fe(II) produced does not seem to be enough to support much plankton growth.

Biologically mediated reduction of Fe may be an alternative means to increase the biological availability of Fe bound to NOM. Experiments conducted on marine diatoms have shown that Fe(III)–NOM complexes can be accessed through use of a cell membrane Fe(III) reductase, similar to systems found in some vascular plants and other eukaryotes. Under Fe deficiency the activity of the reductase is enhanced, enabling these diatoms to acquire Fe bound to a number of natural and synthetic Fe chelators and to grow rapidly. In this type of non-ligand-specific system, reduction of organically bound Fe(III) results in dissociation of the complex, allowing uptake as inorganic Fe(II) or as Fe(III) after reoxidation.

An interesting twist in the reductive uptake process of Fe NOM complexes is the possible involvement of copper. There is evidence from a marine diatom that Fe acquisition involves two consecutive redox transformations of Fe. First Fe(III) is enzymatically reduced to Fe(II) by cell membrane reductases, then Fe is taken up by a protein complex containing a multicopper oxidase, which oxidizes Fe(II) back to Fe(III) during the membrane transport step. Even though the oxidation of Fe(II) occurs spontaneously and rapidly in oxygenated seawater, a multicopper oxidase may be important in order to acquire Fe before it diffuses away from the cell. This Fe transport pathway is highly analogous to that identified in common yeast, and some fungi and green algae. Genes homologous to those that encode for the proteins of this pathway have been identified in the recently sequenced genome of the diatom *Thalassiosira pseudonana*.

See also: Global Change Ecology: Biosphere: Vernadsky's Concept

Further Reading

Boyd, P., Watson, A.J., Law, C.S., *et al*., 2000. A mesoscale phytoplankton bloom in the polar Southern Ocean stimulated by iron fertilization. Nature 407, 695–702.
Hunter, K.A., Turner, D. (Eds.), 2001. The Biogeochemistry of Iron in the Ocean. New York: Wiley.
Jickells, T.D., An, Z.S., Anderson, K.K., *et al*., 2005. Global iron connections between desert dust, ocean biogeochemistry and climate. Science 308, 67–71.
Saito, M., Sigman, D., Morel, F.M.M., 2003. The bioinorganic chemistry of the ancient ocean: The co-evolution of cyanobacterial metal requirements and biogeochemical cycles at the Archean/Proterozoic boundary? Inorganica Chimica Acta 356, 308–318.

Material and Metal Ecology

MA Reuter, Ausmelt Ltd, Melbourne, VIC, Australia
A van Schaik, MARAS (Material Recycling and Sustainability), Den Haag, The Netherlands

Introduction

`Metals and materials´ are used in a wide range of products and applications ranging from consumer products (cars, electronics, white and brown goods, etc.) to constructions (buildings, roads) and agriculture (fertilizers), etc. The social and ecological value of the materials in these applications is not only determined by the `in-use value´ of these applications such as functionality, durability, safety, reduced energy consumption, esthetics, etc., but also by the possibility of these materials to return from their original application into the `resource cycle´ after their functional lives/use at the lowest environmental impact. The design of the product determines the selection of materials to be applied in the products as well as the complexity of the material combinations and interactions within this product (e.g., welded, glued, alloyed, layered). These actions directly affect the recyclability of the materials, that is, whether the material cycle can be closed and whether one can speak of an industrial ecological system.

Fig. 1 indicates that the social/environmental value of materials and metals can only be properly determined if both the resource cycle and the `technology/design cycle´ are fundamentally understood and described, but more important that tools are available to link these three inseparable disciplines. The interconnectivity between the resource cycle (i.e., the primary and secondary material cycles), the technology/design cycle (i.e., product design, recycling technology, materials processing, etc.) and the nature (social/environmental) cycle is depicted by Fig. 1.

This section shows how complexly linked (nano-/micro-) metals (with their associated materials, plastics, etc.) that is, the resource cycle is related to products, production, manufacturing technology, and the product designer, that is, the technology and design. Ultimately these cycles intersect with nature, that is, the nature cycle, but this is disregarded from the discussion here; however, the link to geology, ores, and metal-containing minerals is considered as is the environmental impact of waste and residues from anthropogenic activity. This together composes the complex `web of metals and materials´ from an anthropogenic point of view. This web depicts the flow of metals/materials into consumer products; subsequently as consumer products into most regions of the world and finally either recycled back into consumer products and/or into nature and/or humans and/or disposed. Central to this is the metallurgical processing technology, which constitutes the `ecological organism´ in this industrial ecological system; the `organism´ that closes the anthropogenic material and metal cycles.

The web of metals and materials or `industrial ecological metal and material system´ shows how metals and associated materials flow through the resources industry as well as the consumer product and waste system. It describes this flow on the basis of the first principles of recycling technology and metallurgical processing technology theory which are closely linked to product design. This technology- and engineering-based approach provides insight into the complex web of metals and materials providing information on how to monitor, control, and improve the system (and on this basis its economics), at the same time linking this information back to product design. The best manner in which to map the complex interactions between metals is by the application of dynamic modeling, which is more advanced than material flow analysis (MFA) and provides valuable insights into the dynamic interaction and movement of metals and materials linked to consumer products. This is crucial to ensure that valuable minor elements find their way back into products, visualizing and controlling the distribution of these elements onto the surface of the Earth due to the action of consumer society and original equipment manufacturers (OEMs). Therefore, the following will be discussed providing a holistic and fundamental approach to material and metal ecology.

- Resource efficiency and future availability of materials/minor elements is of environmental, economic, and societal concern. Therefore, the recovery of material and metals within the highly connected web of metals and materials in the resource cycles of both base metals and especially environmentally relevant and valuable minor elements for automobiles, consumer electronics, miniaturization applications, and nanotechnology is of crucial importance from a sustainability perspective. Crucial here is also the effect of the closely connected materials in consumer products that cannot be separated due to their close association in end-of-life products.
- Time dependencies as well as process dynamics often have a crucial impact on the web of metals and materials and hence on the impact these metals and materials have on the environment. This is important since it takes time for consumer products and their associated metals and materials to flow through the system. Also the complex connection between the linked primary ore and secondary recycled materials chains, and rapidly changing product compositions have an important effect on determining where metals and materials report to. Social aspects such as the concentration of labor in certain parts of the world associated with metal production and recycling could be dynamically illustrated by such a dynamic visualization.
- Energy and climate change effects are directly connected to recycling as many of the environmental impacts are dominated by the energy needed and CO_2 produced to extract materials associated with products or by the prevention of this by proper recycling activities. Recycling is therefore of extreme importance to lower the consumption of energy during their production, hence directly having an impact on greenhouse gas emissions. It is shown how the web of metals and materials can be arranged

 doi:10.1016/B978-0-444-63768-0.00739-3

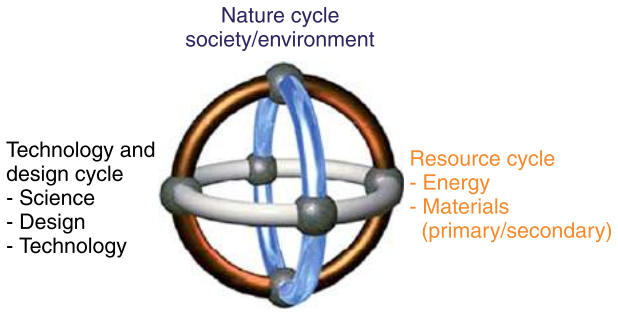


Fig. 1 Philosophy: toward sustainability and material and metal ecology by linking the indicated cycles. Reproduced from van Schaik, A., Reuter, M.A., 2004. The time-varying factors influencing the recycling rate of products. Resources, Conservation and Recycling 40 (4), 301–328, with permission from Elsevier.

to maximize energy recovery, support light-weight design, minimize toxic emissions by basing the system models on the first principles of process engineering.

- The role of product design is demonstrated by discussing the design wheel, which shows how computer-aided design (CAD) is linked to recycling, recyclate quality, metal and energy recovery, and waste creation/prevention. Often materials are connected to each other, alloyed, welded, glued, etc., which makes it impossible to consider the ecology of metals and materials only on a linear and single-element basis.

This rigorous approach provides a basis for quantifying legislation on a more technological and fundamental basis (i.e., physics, chemistry, and thermodynamics) hence providing a first-principles basis for recycling targets and a solid legal basis which OEMs can safely operate on and manufacture products (e.g., future energy-efficient recyclable light-weight cars). This dynamic and technological detailed first-principles approach supports the simplistic life cycle assessment (LCA) approach and provides the consumer with transparent information on all issues surrounding the ecological safe production and use of the product until its end-of-life phase and subsequently its recycling back into metals, materials, and energy.

The Metal Wheel – Material and Metal Ecology

Consumer products are a complex mixture of closely associated metals, plastics, chemicals, inorganic compounds, and materials. These complex connections are often difficult to separate due to the limitations of the applicable separation physics as well as incompatible thermodynamics, which sometimes render the complete recycling chain uneconomical, subject to product type. The result could be that these end-of-life products are then shipped to low-cost countries where these are hand-processed (often more efficiently than present technology permits) and/or dumped or even reused but then eventually finding their plight on an unsafe dump in an uncontrolled economic environment. The result is obviously that hazardous materials could report to the ground-water with all subsequent consequences to health by for example uncontrolled burning of these goods.

Therefore, any modeling and assessment approaches should provide valuable information to the legislator to provide a fundamental basis for global environmental legislation based on achievable technology and economics incorporating the dynamics of market flexibility and consumer behavior. Mapping of materials will inform the original equipment manufacturers (OEMs) on a technology basis where materials and elements in their products are reporting to, ensuring that a solid legal and environmental basis is maintained for marketing these products. On the other hand, the consumer can be transparently informed of all benefits and risks of using the products of the OEM, as well as providing the legislator the means to monitor the (likely) movement of end-of-life products across the globe to ensure that nature and humans do not ultimately come to harm. Thus, a first-principles modeling and simulation approach ensures that the recycling loop can be mapped and subsequently `closed´ for metals and related materials in relation to design. This approach that quantifies the material and metal flows is the only manner to ensure that sustainable resource usage will be attained.

Therefore, judicious management and understanding of the plight of valuable (and also possibly toxic) minor elements is a matter of extreme importance to OEMs, legislators, recyclers, ecologists, environmentalists as well as sociologists, general population, nongovernmental organizations (NGOs), to name but a few. Therefore, for all the issues that will be discussed below to discuss the `ecology of materials and metals´, a fundamental understanding of the size and nature of resource cycles over time is needed in order to address environmental impacts and to formulate policy, design, technological and system organizational strategies for more sustainable global resource cycles.

Fig. 2 shows that each carrier commodity metal is associated in nature (geology) by a unique blend of valuable minor elements (with or without own processing infrastructure) as well as harmful elements of no economic value that are lost due to unfavorable thermodynamic and other conditions within the processing chain. The carrier element can in some cases be only the secondary material being recovered since the minor elements are of much higher economic value. Therefore, affecting the production of these carrier elements could in the end adversely affect the production of the valuable minor elements. Green processing would imply minimizing the losses of elements to the green outside band of Fig. 2.

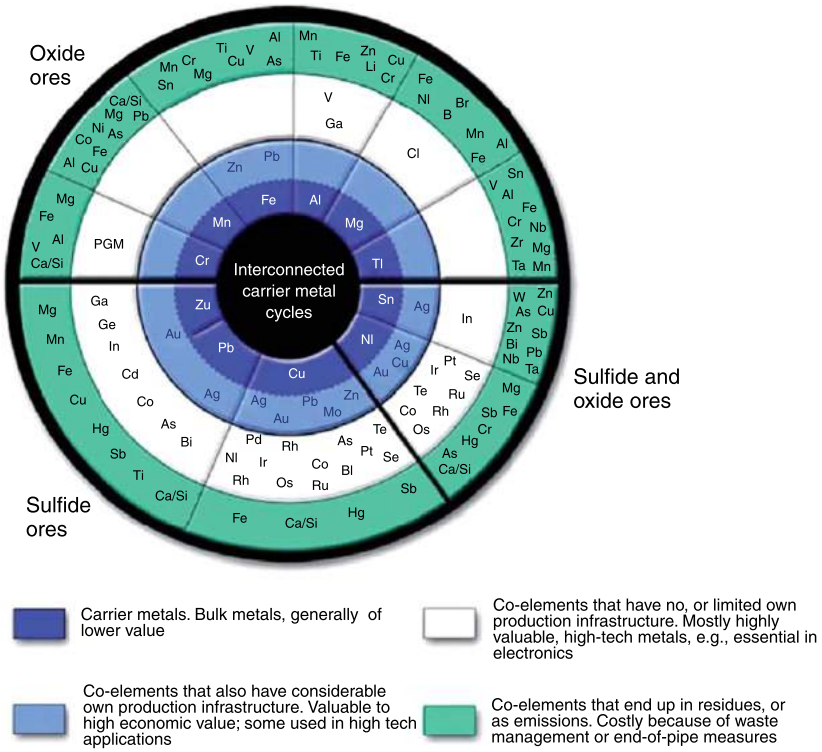


Fig. 2 The `metal wheel´ showing the complex interactions between different metals as well as the economically and thermodynamically recoverability of (co-)elements. Reproduced from Reuter, M.A., Heiskanen, K., Boin, U.M.J., *et al.*, 2005. The Metrics of Material and Metal Ecology. 760 pp Amsterdam: Elsevier Science, ISBN - 13: 978-0-444-51137-9, with permission from Elsevier.

The intricate and unique blend of elements within each ore has led to metallurgical processing being honed to effectively recover and contain most elements economically. This complex link of materials, processing of ores, metals, and end-of-life products has led the creation of a complex web of metals and materials, in which each element has a unique position. Hence affecting the production of one element has an effect over other elements within the web.

An interesting example is indium that is used in flat TV panels. The question is how much is required, where, when, and how this changes dynamically due to the metal (especially zinc) market? For example, a country such as Japan requires around 160 t per annum but can only obtain a fraction of this via primary ore sources. Therefore, recycling has to supply the rest, but this is only achieved partially, that is, in total primary and secondary recovery only reaches around $\gtrsim 100$ t.

Fig. 3 shows a model that simulates this complex link between various elements and predicts their global flow over time through the complete metal and material chain, also for example the passage of the metal Indium (used in flat panel displays) through the complete material and metal system in time. Environmental indicators have been linked to the output in order to quantify the environmental performance of the complete global anthropogenic system.

Product Design and Fundamental Recycling Optimization Models

The design of a product is linked to recycling as depicted in Fig. 4. The design does affect how materials are liberated during shredding, how efficient materials can be separated, and what the composition and quality of the recyclates will be. This determines if these recyclates can be recycled or not, therefore what the losses will be from the recycling chain. The control of the recycling chain determines what the qualities of the streams are and whether or not the recycling rate is high. This in essence determines whether materials and metals can be recycled, hence the pivot of `industrial ecology of the materials and metals´ within the car. This section will discuss the various factors influencing the `material and metal ecology´ of a consumer product, for example, the car.

Product Design, the `Metal Wheel´, and Recycling Technology

Product designers select the produced metals and materials from (primary) resources as depicted by Fig. 2 and apply them in products and applications. The product designers determine which interconnected materials are to be separated and recovered from primary ores for application in the car.

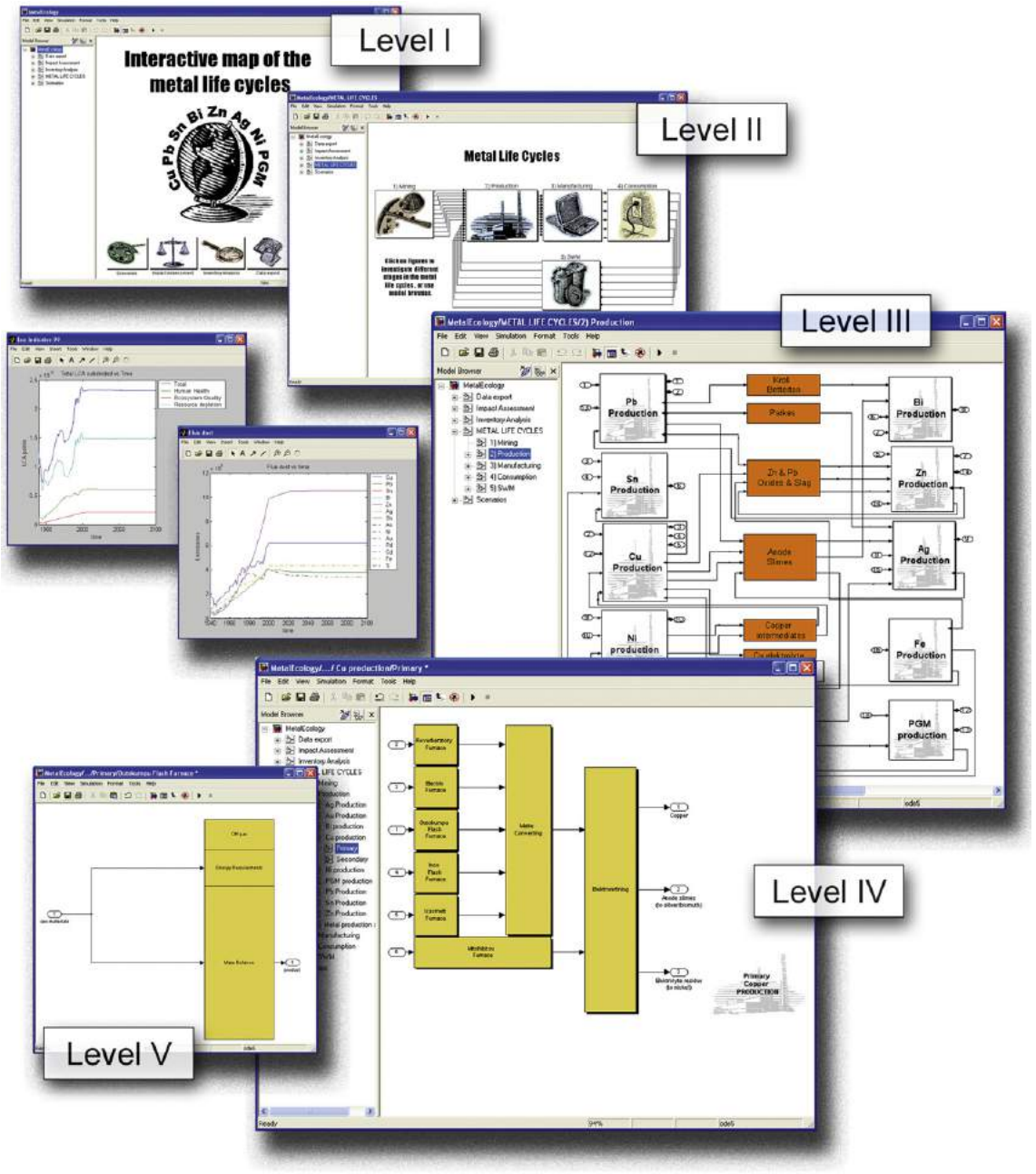


Fig. 3 The various levels of the dynamic Simulink model that dynamically links the metal flow of various metals as shown in the level III slide, producing a dynamic LCA environmental score for the complete system (two small gray windows left middle). Reproduced from Reuter, M.A., Heiskanen, K., Boin, U.M.J., *et al.*, 2005. The Metrics of Material and Metal Ecology. 760 pp Amsterdam: Elsevier Science, ISBN - 13: 978-0-444-51137-9, with permission from Elsevier.

During the design of the product a range of materials/elements are once again mixed and complexly connected (gluing, welding, alloying, etc.). Modern products contain a combination of metals that are not necessarily linked in the natural resource systems as shown in **Fig. 2**. As a consequence, these materials are not always compatible with the current processes in the metals production network, which was developed and optimized for the processing of primary natural resources and associated minor elements.

In general an increased complexity of recycling pyrometallurgy has arisen through the development and design of modern consumer products (such as passenger vehicles and consumer electronics). The consequence is the formation of complex residue streams or undesired harmful emissions that cannot be handled in the current system (thus the processing and recycling of those products at their end-of-life). This can be prevented by linking product design with optimized recycling technology, therefore minimizing the loss of valuable material and preventing the decrease of both quality of recyclates and recycling rates of these products.

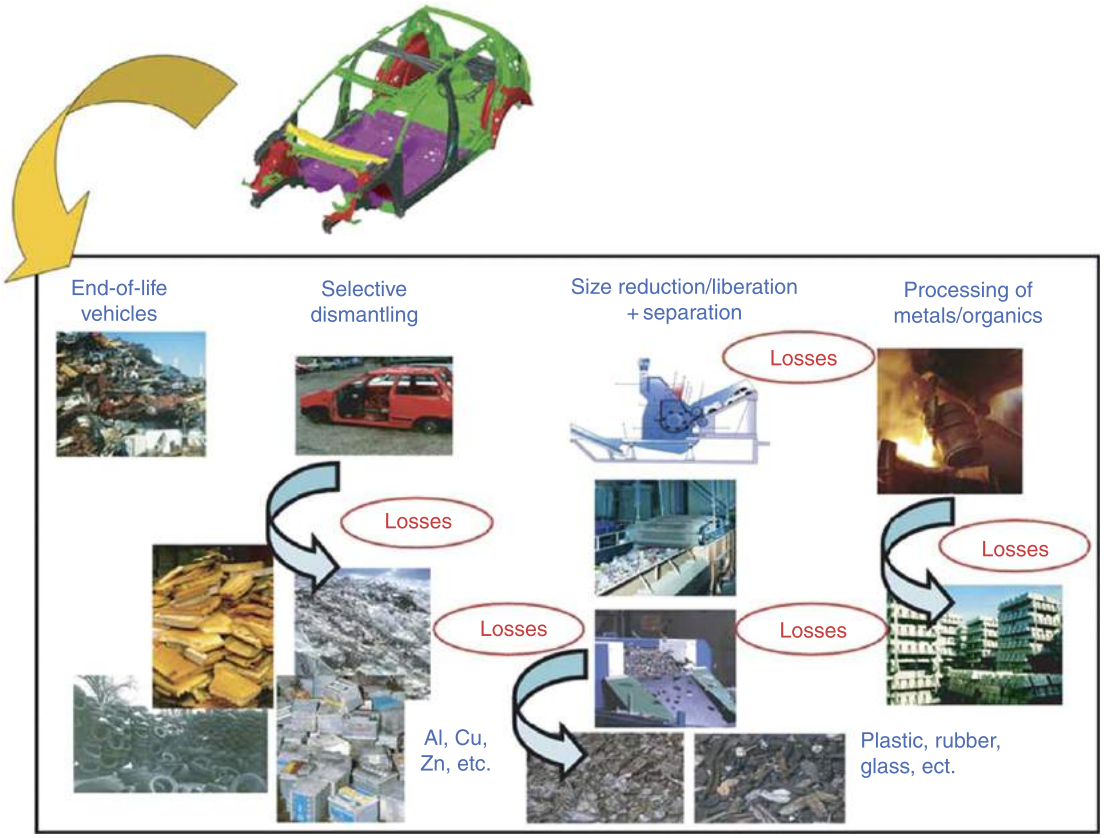


Fig. 4 The link of a car design to the recycling chain, which includes selective dismantling, size reduction, and subsequent physical separation, metallurgical, and thermal treatment.

Recycling Optimization Models

Commercial recycling systems never create pure material streams (see **Fig. 5** for materials from end-of-life-vehicles (ELVs) recycling collected from various shredder plants), never achieves 100% material recovery (recycling) during physical separation (dictated by separation physics), neither achieves 100% material recovery (recycling) during high-temperature metal production (dictated by thermodynamics), and nor achieves 100% energy recovery (dictated by thermodynamics).

The recyclability of a product is not only determined by the intrinsic property of the different materials used, but by the quality of the recycling streams (see **Fig. 5**), which is determined by the mineral classes (combination of materials due to design, shredding, and separation), particle-size distribution, and degree of liberation (multimaterial particles) (see **Fig. 6**). All these affect the physical separation efficiency, metallurgical and energy recovery, which all in turn determine the quality and economic value of the recycling (intermediate) products in the recycling system, which can be applied as secondary resources.

Since the quality of recyclates and the recycling rate of a product is largely affected by the design of the product (see **Fig. 7**), it is required that tools are available that link CAD software and recycling models in order to predict recyclability of the car during the design phase. In addition, this predicts and determines the social and environmental value of the materials applied in the product. Fundamental knowledge of recycling processes, such as shredding, mechanical separation processes and metallurgy, and material characteristics of recycling (intermediate) products (material type, liberation, etc.) have to be combined with that of the design of the product (material combinations and connections). In order to optimize the resource cycle and maximize the recycling rate of (future) products all the parameters determining the recycling/recovery rate for each of the materials present in the multimaterial designs and applications of the present and future have to be fully understood. This should all embrace the dynamics and statistically distributed nature of the resource cycle system.

The prediction of the recyclability and recoverability of products already in the design stage requires the exploration of the limits of recycling on a fundamental basis as has been discussed by Reuter *et al.* Recycling models have been developed by Reuter and van Schaik. These recycling models take into consideration (1) material quality (physical and chemical) and calorific values of the (intermediate) recycling streams being a function of material/mineral classes, particle-size classes, liberation classes (degree of liberation); (2) the value of intermediate streams; (3) separation physics and thermodynamics; (4) losses and emissions; (5) harmonization of plant/flowsheet architecture with changing product design; and (6) distributed and dynamic properties of present and future product designs (see **Fig. 8** for a detailed flowsheet of the recycling optimization model).

Fig. 5 Impure quality materials created during physical separation of shredded ELVs (clockwise top left: steel, wires, Mg/Al/Zn/Cu/SS, steel/Cu, Mg/Al/Zn/Cu/stainless steel (SS), and plastics). Reproduced by permission of Elsevier.

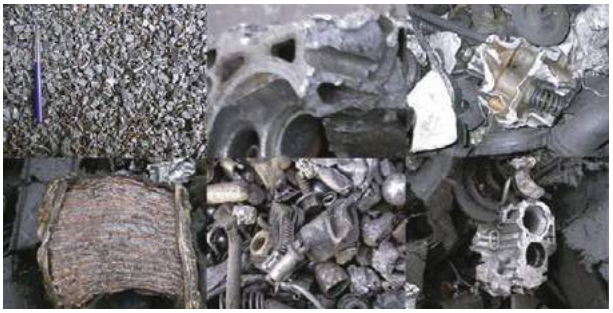


Fig. 6 The `mineral´ aluminum in its different appearances (liberation and particle-size classes) as a high-quality liberated fraction (top left) to unliberated radiator (bottom left) and various unliberated mixed fractions that cannot necessarily be recycled directly.

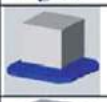
Connections types	*Before shredding*	*After shredding*
Bolting/riveting		
Gluing		
Insertion		
Coating/painting		

Fig. 7 Possible connection types in car design with distinctive liberation behavior. From van Schaik, A., Reuter, M.A., Richard, A., 2005. A comparison of the modelling and liberation in minerals processing and shredding of passenger vehicles. In: Schlesinger, M. E. (Ed.), EDP Congress 2005. Warrendale: TMS (The Minerals, Metals & Materials Society), pp. 1039–1052.

Fig. 9 shows how after shredding, shredded particles have different degrees of liberation, therefore creating streams of different composition (quality and economic value). This is partially caused by imperfect separation in the physical separation stage of these particles and also by the design choices as shown in Fig. 7, affecting particle composition after shredding. The quality of recyclate streams ultimately determines in which processing steps depicted in the detail flow sheet in Fig. 8 these materials can be processed and hence how much material of sufficient quality and economic value can be recycled.

Fig. 10 depicts how product design selects materials from the primary metal and material cycles and combines them into a complex multimaterial design, in which the various materials (metals and nonmetals) are complexly integrated. The combination and connection of the materials in the product design is linked (on the basis of the discussed recycling models) to the quality of recyclates as a function of the degree of liberation of the various particles after shredding. The colors in the `design wheel´ of Fig. 10 reflect the (in)compatibility of material combinations in the recyclates (either due to imperfect liberation or separation) based on the material combination matrix given in Fig. 11, in which the (in)compatibility of material combinations is based on the thermodynamics and kinetics of metallurgical processing (see also Fig. 2).

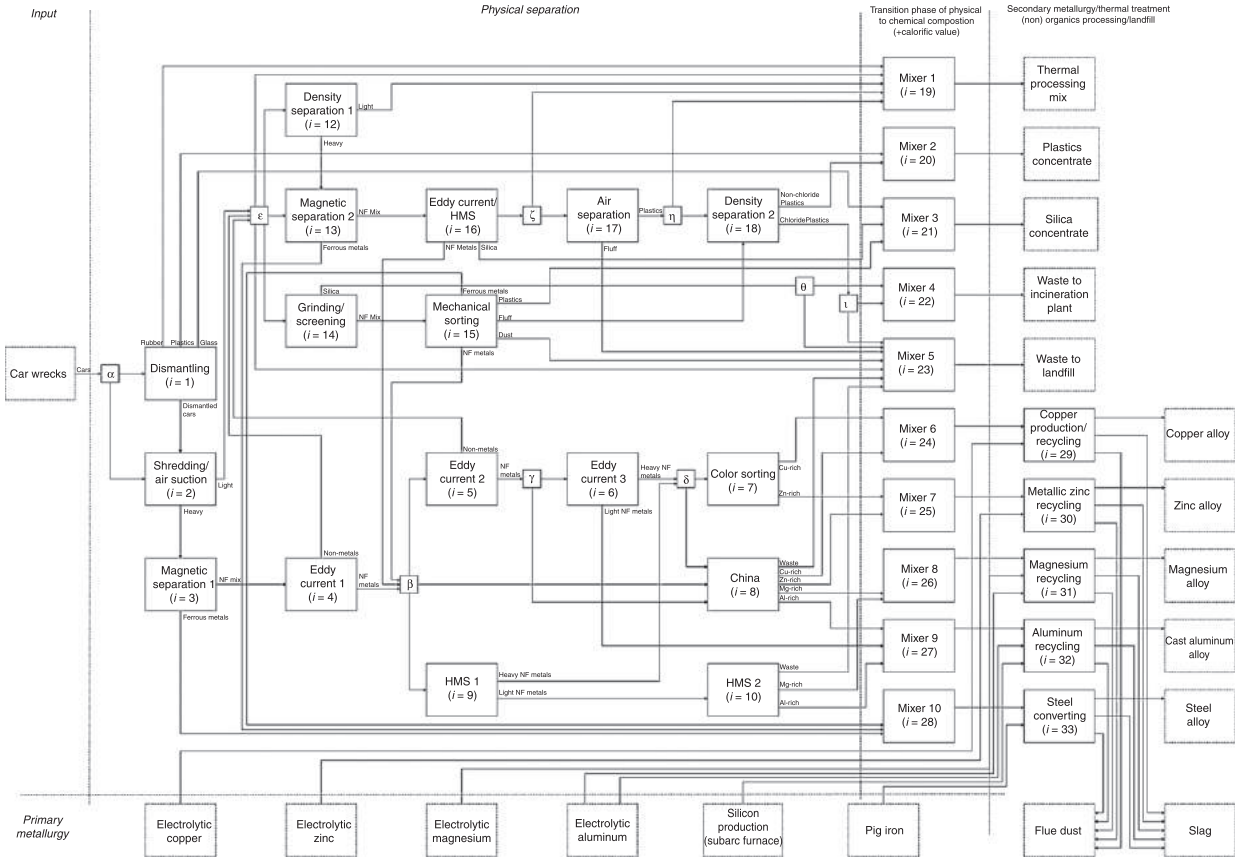


Fig. 8 Flowsheet of detailed recycling system optimization model (programmed in AMPL). Reproduced from Reuter, M.A., van Schaik, A., Ignatenko, O., de Hann, G.J., 2006. Fundamental limits for the recycling of end-of-life vehicles. Minerals Engineering 19, 433–449, with permission from Elsevier.

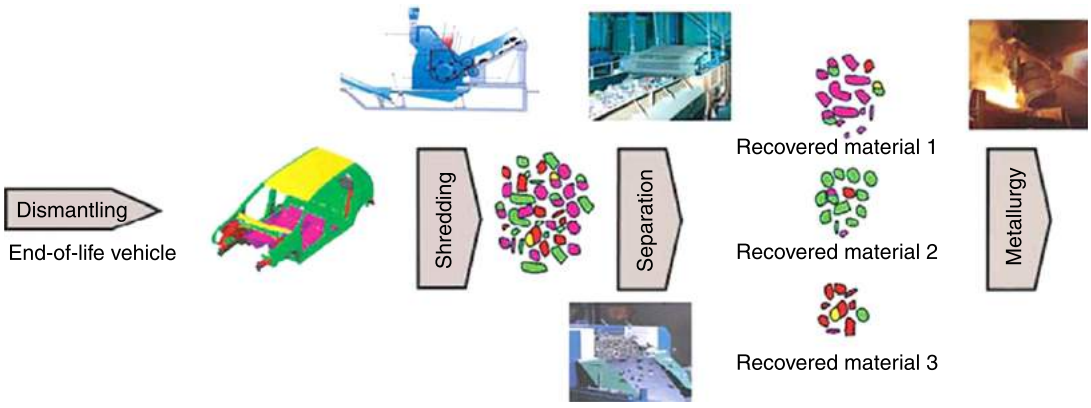


Fig. 9 Shredding of a car and the creation of liberated and unliberated materials.

Fig. 10 reflects the knowledge and modeling detail captured by the developed recycling optimization tools and provides feedback to the designer on desired and undesired material combinations in the design. The wheel acts as a preliminary design for recycling (DfR) tool, reflecting the complexity and detail of the developed recycling models to ensure a proper reflection of the reality of recycling system behavior and the quality and value of produced recyclates. The wheel enables DfR based on the limits and possibilities of recycling technology and recyclates quality as a function of design and separation efficiency. In summary therefore this wheel shows what can be achieved as a function of design in combination with `best available technology´ (BAT) and hence the limits of physics and chemistry as taking place in the technology.

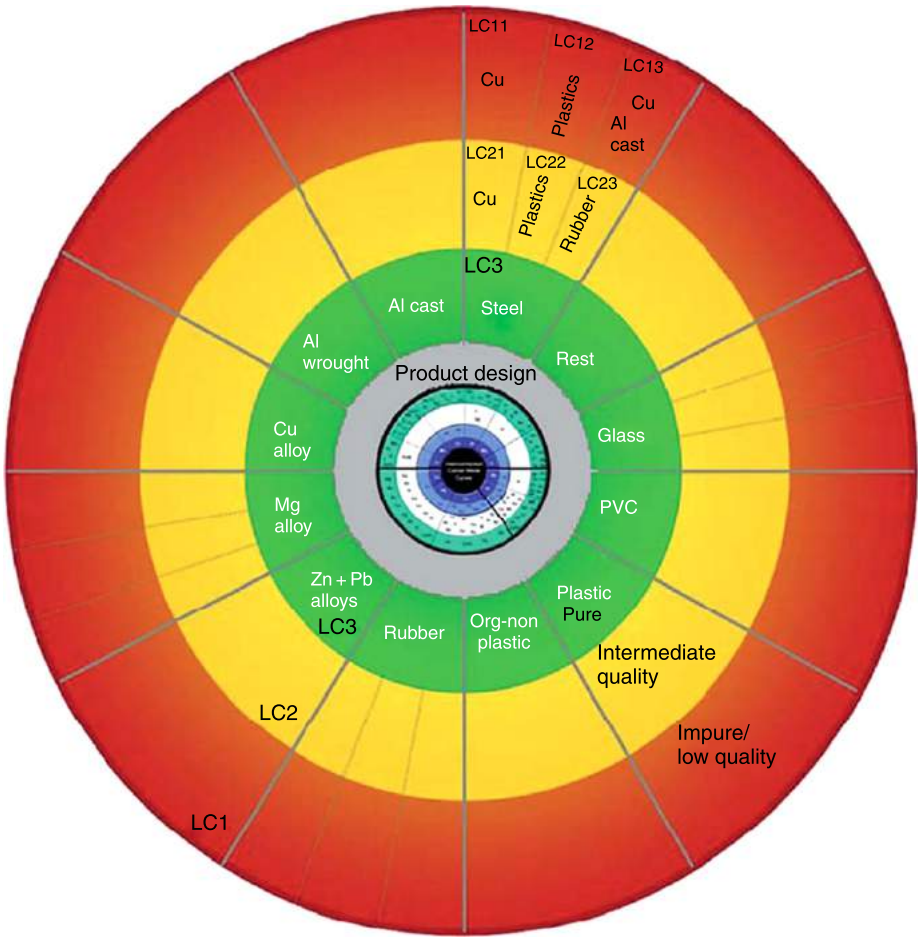


Fig. 10 The `design wheel´ illustrating the underlying liberation classes that are created as a function of product design as predicted by the recycling model depicted in Fig. 8. Reproduced from van Schaik, A., Reuter, M.A., 2007. The use of fuzzy rule models to link product design to recycling rate calculations. Minerals Engineering 20, 875 -890, with permission from Elsevier.

Linking Design and Recycling

Figs. 4, 8, and 9, respectively, depict a simple scheme for car recycling and a complex optimization model for recycling of end-of-life products such as a car. Fig. 12 depicts the (un)liberated particles after shredding (Figs. 6 and 9) which determines its recyclability due its quality (and hence its economic value). These determine whether or not the material chain can be closed.

Table 1 explains why, if certain fractions are liberated, unliberated, whether they can be fully recycled or not. For example, copper connected to steel will dissolve in steel. Since copper is more noble (less reactive to oxygen) than steel it cannot readily be removed from the steel. This affects the steel quality negatively (e.g., its mechanical properties) and therefore it is given a red color in Table 1. Note, that this is dependent on the amount of the one material connected to the other (the concentration of the contaminant). In many cases, although red material combinations exist, shredding liberates the materials, which are subsequently separated during physical processing and hence they are recyclable. Fig. 11 and Table 1 are only true if there are reasonable amounts of materials connected in the recyclates (exceeding the contamination limits), hence producing alloys outside their normal definitions. The type of models as discussed above can predict the recyclate quality and therefore link design to recycling and restrictions as indicated in Table 1.

The optimization model of which the flowsheet basis is depicted by Fig. 8 is far too complex to link to CAD directly, therefore fuzzy logic rule-based models that mirror the results of this complex model have been developed from the numerical results of the recycling models, and linked to CAD software and the design of the product (© MARAS). Not only can these fuzzy logic models be linked to CAD software, but they can also be integrated into LCA tools, in order to ensure that environmental models are provided with fundamental information on the end-of-life behavior of products which include (1) physics and thermodynamics of separation processes, (2) the quality and value of recyclates as a function of physical design choices (material combinations and connections), and (3) physical separation and metallurgical and thermal processing technology on a statistical basis.

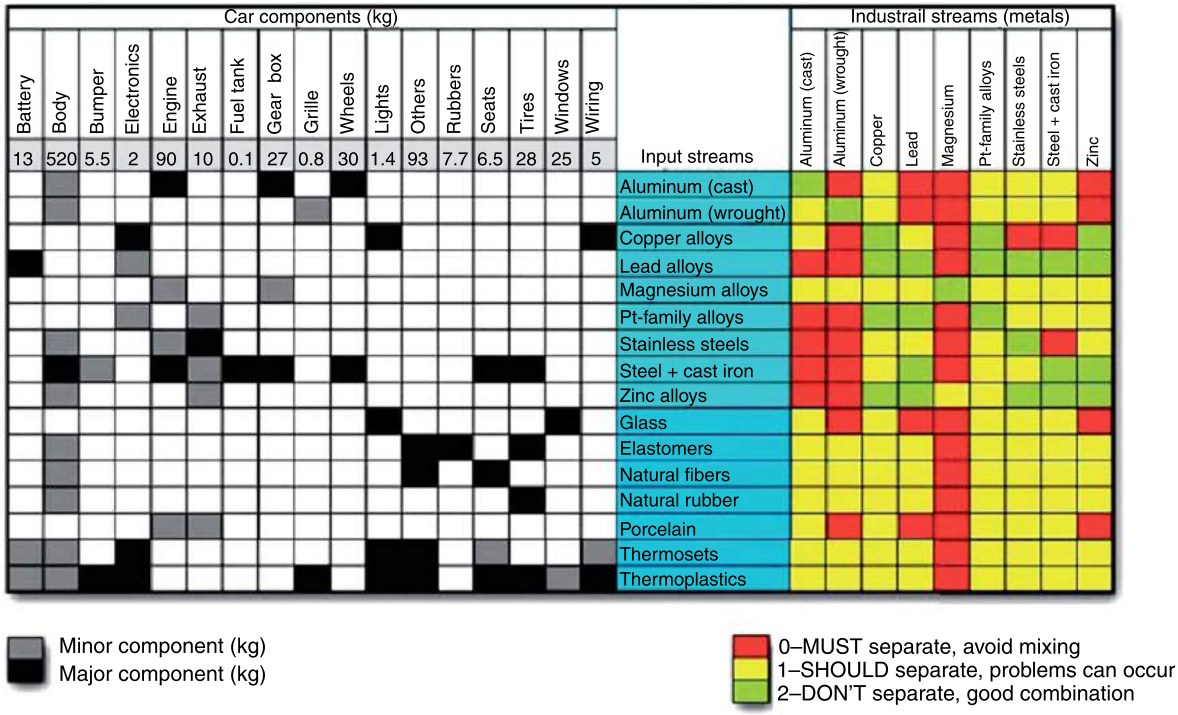


Fig. 11 Material combination matrix: permitted connections and nonpermitted connections and combinations in particles after shredding and separation. Reproduced from Reuter, M.A., Heiskanen, K., Boin, U.M.J., *et al.*, 2005. The Metrics of Material and Metal Ecology. 760pp Amsterdam: Elsevier Science, ISBN - 13: 978-0-444-51137-9, with permission from Elsevier.

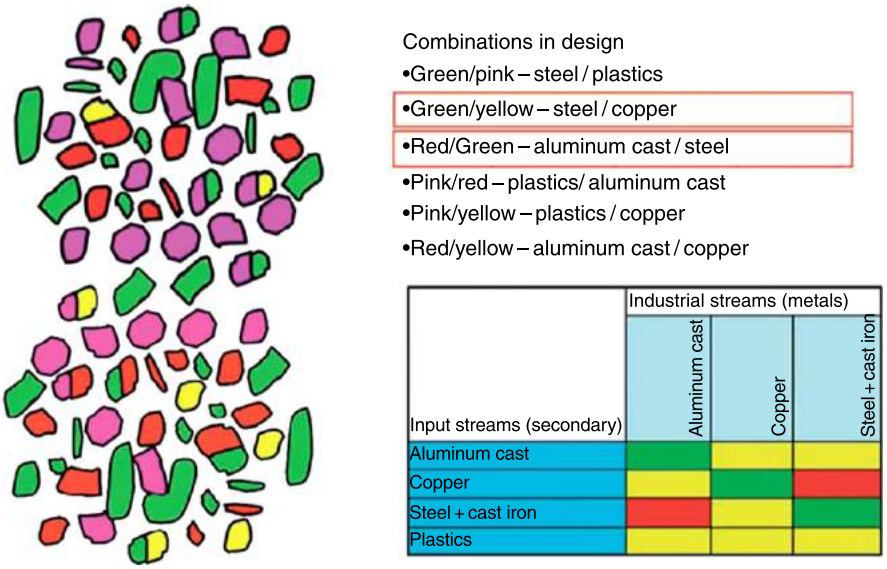


Fig. 12 A selection of liberated and unliberated particles from the car model given in Fig. 4. A section from the compatibility matrix (Table 1 and Fig. 11 – please note that colours of particles are those of the body-in-white (BIW) in Figs. 4 and 9).

Recycling 1153 ELVs – From Theory to Practice

The developed theory as described in the previous section provides a fundamental basis for proper collection of data, supported by a good mass balance based on data reconciliation, and the corresponding statistics and how this should be performed when carrying out experiments or auditing a plant. This theory is essential to characterize and control the material and element flows in recycling plants and through the complete recycling system, which is extremely important for good metal/material accounting, the

Table 1 The reasons for certain materials being compatible or not (also see Figs. 11 and 12) explained on a thermodynamic basis

Input streams (secondary) = recyclates	*Industrial streams (metals)* *Aluminum cast*	*Copper*	*Steel + cast iron*
Aluminum cast	Similar material	During copper processing Al is lost to slag	Loss of Al; Al less noble
Copper	Cu is more noble than Al cast; a certain % of Cu is allowed being one of the alloying elements for Al cast	Similar material	Cu is more noble than steel
Steel + cast iron	Steel+cast iron more noble than Al cast	Creates excessive slags, loss of steel to slag	Similar material
Plastics	Limited due to reaction of Al with C and subsequent loss of Al (Al_4C_3)	Affects energy balance of processing; fillers affect slag properties; possible dioxine creation	Affects energy balance of processing; fillers affect slag properties; possible dioxine creation

calculations of recycling rate on a sound statistical basis, as well as quality control of recycling streams. In fact this is the basis of any meaningful discussion on `material and metal ecology´.

Experimental and industrial data on the composition of the car, the separation efficiency of the various processes, liberation and particle-size reduction in the shredder, the quality (or grade) of the recycling (intermediate) material streams is typical information that becomes available through a good understanding of the theory of recycling as discussed in the previous section. Furthermore, the collection of industrial data on recycling based on best available technology is essential to predict and calculate the recyclability of passenger vehicles, using the developed models. This is of extreme importance for a realistic definition of the type approval and end-of-life legislation of vehicles or any other consumer product. This type of data hence underpins the viability of material and metal ecology.

The theory is applied to provide a procedural basis from which the recycling rate can be calculated from an industrial experiment, in which 1153 ELVs were recycled. This experiment was executed at a large-scale industrial recycling plant and clearly illustrates how statistically sound recycling rates can and therefore should be calculated from data collected from recycling experiments based on the developed theory and classical sampling theory and statistics.

Practical Procedures for Performing Large-Scale Industrial Recycling Experiments

Mass balances of plants based on measured data mostly do not close due to inevitable weighing and sampling errors, as is also the case for the shredding and `postshredder technology (PST)´ trial as discussed here. Data reconciliation has been applied to close total and element/compound mass balances over the plant and its unit operations. A large body of data renders the mass balance more accurate and makes it possible to calculate the recovery and grade for each of the different materials over the various process steps. These data are used to calibrate the models in the optimization and dynamic models mentioned above. Classical sampling theory has been applied to calculate statistically correct sample sizes for analyses of the various material flows throughout the plant (see Fig. 4). The mass flows and composition of the streams were measured and analyzed over all unit operations in the plant that is, on the input, intermediate, and output streams, in order to increase the amount of data available for data reconciliation, which increases the accuracy of the mass balance and its statistics.

Calculation of Recycling/Recovery Rate

Based on the mass balance and its statistics, the recycling rate of ELVs based on the discussed test could be calculated for best available technology as shown in Fig. 13. For the first time a test was therefore concluded in which the recycling rate was calculated within a statistical framework, crucial to proving the validity of the recycling rate calculation. Ultimately the recycling rate is determined by the possibility of the market to absorb the produced output streams (either for direct application or in metallurgical or thermal processes) and is therefore determined by the quality of the recycling (intermediate) products as well as by the geographic location of the plant (due to local environmental legislation).

Statistics

Only data reported within a statistical and theoretical framework can have a legal basis and can find their way into design software for cars in order to perform `DfR´ and hence real `material and metal ecology´ on a large industrial scale. Moreover the statistics around the calculation of the recycling rate based on plant data indicates that the (calculations for the) recycling/recovery rates and requirements for type-approval of cars as imposed by legislation in Europe should also be based on a statistical basis and are meaningless if represented by a single value as is required at the moment. Any methodology to assess end-of-life systems has to take into account the statistics of design and end-of-life technology.

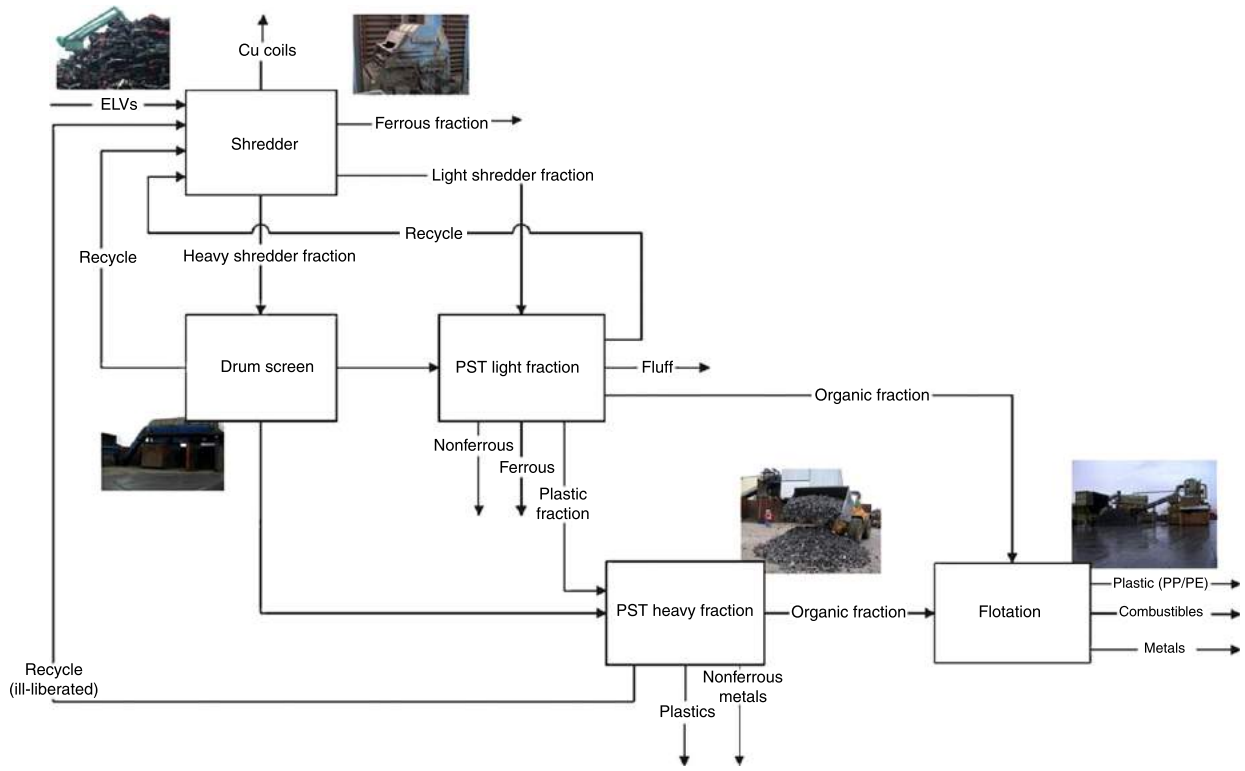


Fig. 13 Simplified flowsheet of the shredding and postshredding technology (PST) plant. Reproduced from Reuter, M.A., Heiskanen, K., Boin, U. M.J., *et al.*, 2005. The Metrics of Material and Metal Ecology. 760pp Amsterdam: Elsevier Science, ISBN – 13: 978-0-444-51137-9, with permission from Elsevier.

Material and Metal Ecology

The fundamental basis of `material and metal ecology´ is discussed in this section. The discussed aspects provide fundamentally based answers and approaches to questions of crucial importance to the industry and environmentalists such as the magnitude of the legally required recycling rate of presently designed products, emissions of processes and end-of-life products to nature, etc. This approach is being adopted by the automobile industry to ensure among others that recycling rates have the fundamental basis that challenges but also provides the legal basis for recycling legislation. It can also provide the basis for risk analyses for new car designs from a recycling point of view.

Any evaluation of the economic and/or environmental consequences and calculation of environmental scores of product and material applications can only be conducted if the interconnectivity of material/metal cycles is recognized and recycling rate calculations are based on fundamental recycling models. The model structure should be linkable to CAD product design activities, material choices, joints, etc. This is not possible with a LCA approach on its own, since it is not a simulation or predictive tool, it only represents the present and does not give technological advice about the future, how technology should be controlled and adapted, how the physical design of products have to be changed to ensure that economic recyclates can be created.

In summary the discussed approach provides a basis to `material and metal ecology´ in the following areas, which should ideally be integrated:

- Creation of a fundamental basis to define and realize `material and metal ecology´. All the models are predictive as a function of physics, thermodynamics, and chemistry as well as time, hence they are dynamic.
- Development of fundamental recycling models for interconnected metals and materials applied in various products/applications, for example, the models are applied by the authors for the recycling of passenger vehicles, waste electric and electronic equipment (WEEE), as well as for other waste/material systems.
- Development of fuzzy logic rule-based models to link fundamental recycling models on a simplified basis to design tools and material choices, while still maintaining the detail knowledge (process, material, quality, etc.) as captured by the complex models, hence providing simple risk models for designers.
- Fundamental recycling models (calibrated with data based on a statistical basis) can be linked to environmental LCA tools/software in order to provide a fundamental technological and statistical basis for the calculations of recycling/recovery rates, prediction of quality of recyclates, process operation, recycling system architecture, etc.

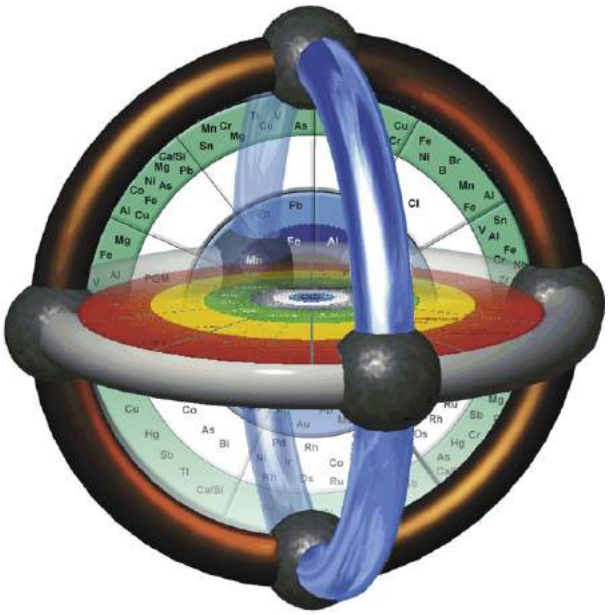


Fig. 14 The interlinked disciplines – applying fundamental models to link design to recycling and interconnected metal cycles providing fundamental knowledge and data to environmental models (LCA, MFA, etc.).

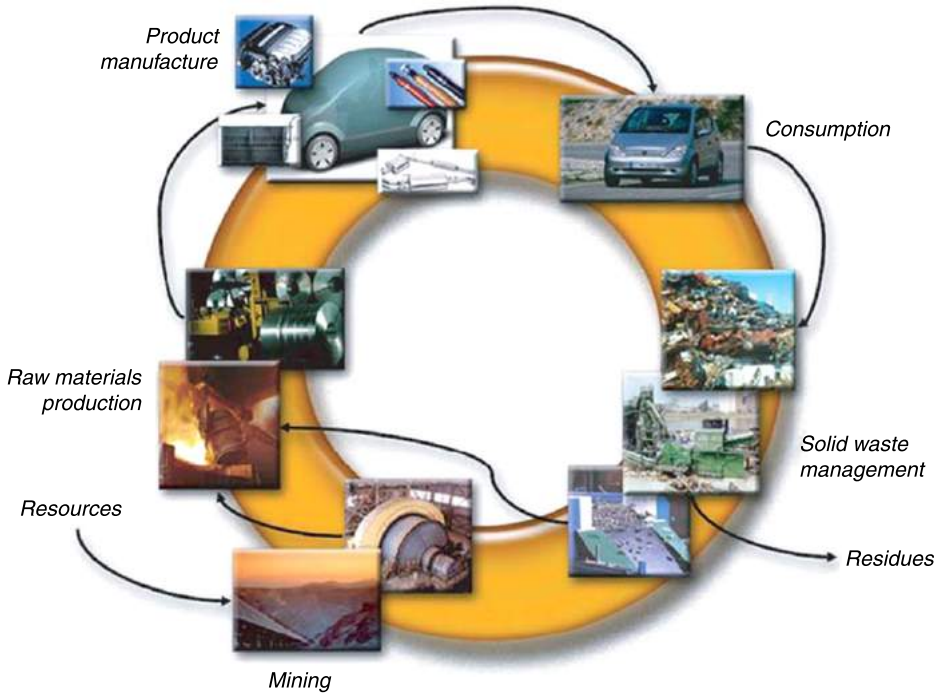


Fig. 15 The anthropogenic material and metal cycle. Reproduced from Reuter, M.A., Heiskanen, K., Boin, U.M.J., *et al.*, 2005. The Metrics of Material and Metal Ecology. 760 pp Amsterdam: Elsevier Science, ISBN - 13: 978-0-444-51137-9, with permission from Elsevier.

- A first-principles and technological basis for `DfR´ guidelines is provided which include (1) the influence of material combinations and connections, (2) liberation, (3) particle size, and (4) physical/chemical/thermodynamical process efficiency on the quality of recyclates and the maximum achievable recycling rates.
- Support eco-design by providing fundamental knowledge on recycling systems and DfR.

Fig. 14 summarizes the application of material and metal ecology, in which fundamental models link material choices in product design to recycling and interconnected material cycles. This provides fundamental knowledge and data for environmental

models (LCA, MFA, etc.), therefore linking the various disciplines related to the ecological value of materials in our society. It is still required to link these material and metal ecology models to the environment so that the effect of a product design on the environment can be directly determined. This would then constitute the final objective of true material and metal ecology in the present industrialized society. **Fig. 3** depicts a dynamic model that provides a basis for mapping dynamically the flow of elements, while the model depicted by **Fig. 8** depicts a recycling system model. The true innovation of this work is depicted by **Fig. 14** shows how **Fig. 2** (metal wheel) is linked to product design (design wheel). This innovation is a key issue in controlling the anthropogenic material and metal cycle (resource cycle), ensuring that the positive interaction with the nature cycle as shown in **Fig. 1** is optimized.

In order to realize a `sustainable´ material and metal ecology, the system depicted by **Fig. 15** should be in balance with the material and metal cycles in nature. In summary, visualizing and describing mathematically the link between our industrialized societies with nature, as discussed above, is key to providing a measure for `sustainability´ in our present consumer society. This measure provides a tool to shape a more harmonious future!

Further Reading

Pitard, F.F., 1993. Pierre Gy's Sampling Theory and Sampling Practice, 2nd edn. Boca Raton, FL: CRC Press, 488pp.
Reuter, M.A., Heiskanen, K., Boin, U.M.J., *et al*., 2005. The Metrics of Material and Metal Ecology. 760pp Amsterdam: Elsevier Science, ISBN – 13: 978-0-444-51137-9.
Reuter, M.A., van Schaik, A., Ignatenko, O., de Hann, G.J., 2006. Fundamental limits for the recycling of end-of-life vehicles. Minerals Engineering 19, 433–449.
van Schaik, A., Reuter, M.A., 2007. The use of fuzzy rule models to link product design to recycling rate calculations. Minerals Engineering 20, 875–890.
van Schaik, A., Reuter, M.A., 2004. The time-varying factors influencing the recycling rate of products. Resources, Conservation and Recycling 40 (4), 301–328.
van Schaik, A., Reuter, M.A., 2004. The effect of design on recycling rates for cars – Theory and practice. In Proceedings of the International Automobile Recycling Congress. Geneva, Switzerland, March 10–13, 2004, 21pp.
van Schaik, A., Reuter, M.A., Heiskanen, K., 2004. The influence of particle size reduction and liberation on the recycling rate of end-of-life vehicles. Minerals Engineering 17 (2), 331–347.
van Schaik, A., Reuter, M.A., Richard, A., 2005. A comparison of the modelling and liberation in minerals processing and shredding of passenger vehicles. In: Schlesinger, M. E. (Ed.), EDP Congress 2005. Warrendale: TMS (The Minerals, Metals & Materials Society), pp. 1039–1052.
Verhoef, E.V., Reuter, M.A., Dijkema, G.P.J., 2004. Process knowledge, system dynamics and metal ecology. Journal of Industrial Ecology 8 (1-2), 23–43.
Ververka, V., Madron, F., 1997. Material and Energy Balancing in the Process Industries. Amsterdam, The Netherlands: Elsevier.

Microbial Cycles

GA Zavarzin, Russian Academy of Sciences, Moscow, Russia

Introduction

Microorganisms, particularly bacteria (=prokaryotes), represent the first sustainable system in the biosphere into which all other living beings are superimposed and included. Sustainability of the system depends on the catalytic role of bacteria in the cycles of biogenic elements and their mediating role in the transformation of other elements. Development of cooperative microbial community led to the biogeochemical succession, the most prominent result being oxygenation of the atmosphere around 2.4 billion years ago with interconnected changes for chemical compounds. The role of bacteria in the biosphere depends on their functional diversity and formation of cooperative trophic systems, scaling up from the local ecosystems to the biosphere as a whole. The limits of life are delineated by the topic adaptability of bacteria, while all other living beings remain within these frames. The number of bacteria exceeds 10^{28} in the active layers with rapid turnover and might be 10^{30} in total. It makes them by far the most important group in the conceptual structure of the sustainable biosphere. The trophic structure of the microbial system makes the framework of the biosphere. The interconnection of biogeochemical cycles makes the functional role of microorganisms in the biosphere most fundamental.

Biogeochemical cycles represent the main system by which the energy of the Sun is transformed into energy of the chemical compounds by living beings and products of their activity. The cyclic arrangement is the main principle of sustainability in the Earth system. It means that the compound involved in the process after sequential transformations is regenerated as its end product. Cycles are regarded as the cycles of the elements. Stepwise reactions of the cycles are catalyzed by specific groups of microorganisms. The system of higher organisms is superimposed into the initial cooperative system constructed by bacteria.

C_{org}-Cycle

The driving force of the system of interlinked cycles is the cycle of organic carbon (C_{org}). The cycle involves two steps: production and destruction. During production CO_2 is assimilated in the biomass; during destruction dead biomass is decomposed into CO_2. Composition of biomass includes in addition to C_{org}, as the main components N_{org} and P_{org} in approximate molar ratio 106:16:1. This ratio calculated for marine phytoplankton is quoted as 'Redfield ratio'. H and O are included in the biomass in water in the ratio 2:1, making the reductive level of C_{org} close to $[CH_2O]$. Strong deviations from Redfield ratio are known for terrestrial biomass with organic supportive structures as in trees with C_{org}:N_{org} about 500; minor deviations are caused by storage products. There are other elements included in the biomass such as S_{org}, and a number of essential 'mineral' elements beginning with K, Fe, Mg, Ca, and microelements. Composition of the living biomass might be considered as invariable with minor deviations (Fig. 1).

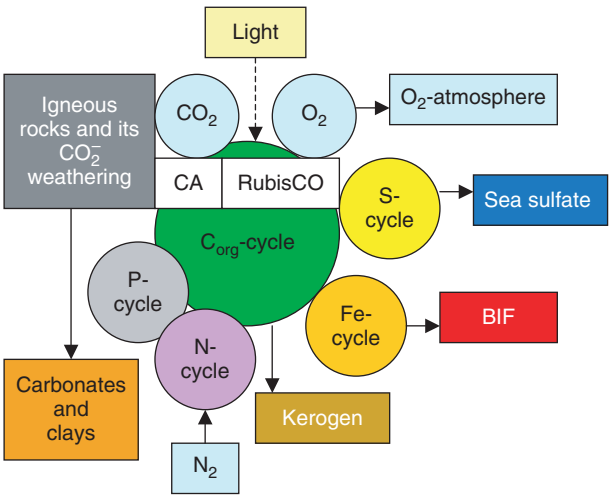


Fig. 1 Interlink between the cycles of the main biogenic elements. Cycle of C_{org} makes the main driving force of machinery coupled to the cycles of biomass constituents N_{org} and P_{org} and catabolic cycles of oxygen, sulfur, and iron. Cycles of these elements are coupled to reservoirs of inorganic matter in the geosphere. For each time period approximate material balance should be sustained. Misbalance leads to the biogeochemical succession on the large time scale. Modified from Zavarzin GA (2004) *Lekcii po Prirodovedcheskoi Mikrobiologii* (*Lectures in Environmental Microbiology*). Moscow: Nauka.

 doi:10.1016/B978-0-444-63768-0.00745-9

Transformation of CO_2 into C_{org} is performed by autotrophic organisms by the metabolic pathways where the Calvin cycle is quantitatively dominating; other autotrophic reactions seem not important quantitatively on the global scale. The key enzyme is ribulose-1,5-bisphosphate carboxylase/oxygenase (RubisCO) which carboxylates phosphopentose regenerated in the cyclic metabolic pathway (Calvin cycle). Due to the discrimination of ^{13}C during autotrophic assimilation isotopically lighter carbon with $\delta^{13}C_{org} \sim -25‰$ is produced, which is considered as the isotopic signature of a biotic source. Careful interpretation of isotope fractionation data is strongly recommended since they depend both on biotic pathways and inorganic diffusion. Assimilation depends on the source of energy, light of the Sun for photosynthesis, or oxido-reductive reaction of inorganic compounds for chemosynthesis (chemolithotrophy is the later synonym) in subterranean systems. In photosynthesis the overall reaction $CO_2 + H_2O \rightarrow [CH_2O] + O_2$ takes place. It makes a coupled cycle with an equimolar ratio of CO_2/O_2. The quantity of O_2 liberated is equivalent to the total C_{org} extracted from the system into the biomass and the reduced products of its decomposition. The link between reservoirs of inorganic and organic carbon is performed by enzyme carboanhydrase (CA) in the reaction $CO_2 + H_2O + CA \leftrightarrow H_2O - CA - CO_2 \leftrightarrow HCO_3^- + H^+ + CA$. In cyanobacteria, CA and RubisCO are integrated into the structural unit carboxysome. In eukaryotes, intracellular localization of enzymes is different. CA is responsible for CO_2 evolution during respiration. The production is measured either by O_2 production in water systems or by ^{14}C-bicarbonate assimilation. The cycle of C_{org} is linked to the reservoir of C_{inorg} with a strong influence of the calcium cycle.

Photosynthesis is the dominating process in production. Primary production is proportional to the illuminated surface or more precisely to the density of chlorophyll, with an approximate ratio of annual assimilation 145 kg C_{org} per kg of chlorophyll in terrestrial boreal ecosystems. Formation of C_{org}-pool occurs through several steps. The first one is gross primary production (GPP), counterbalanced by photorespiration in which approximately half of carbon is lost. Netto primary production (NPP) is calculated on the annual basis of the growing season for a plant. Evidently for algae with a short life cycle, the concept is different. C_{org} balance in the ecosystem is different and includes losses by the respiration of decomposers; it is referred to as Netto ecosystem production (NEP). Optimal conditions for photosynthesis and destruction are different: destruction has higher optimal temperature than photosynthesis, and different dependence on water, being suppressed by the excess of water, causing anaerobiosis. This causes zonal variance for biomes. The accumulation of C_{org} on the decadal scale is designated as Netto biome production (NBP) for which accumulation of nondecomposed C_{org} as humic substances and peat is the main parameter. For marine ecosystems, 'dissolved C_{org}' substitutes soil humus. The recalcitrant C_{org} of humic substances has a residence time of about millennia. It is converted into carbon of sedimentary rocks known as 'kerogen', which makes the main reservoir of reduced carbon on the planetary scale with a residence time of more than millions of years, depending on geological recycle. The reservoir of kerogen is sufficient to balance oxygen in marine sulfates and iron oxides deposits. Only 5% of the total oxygen produced remains in the transitionary reservoir of the atmosphere.

From a brief description of C_{org}-cycle, it is evident that the residence time in reservoirs is to be included into consideration. Seasonal variations in CO_2 fluxes are illustrated by the annual oscillations of atmospheric CO_2 in the continental Northern Hemisphere with an amplitude of about 20 ppm (parts per million) in Hawai and increasing in higher latitudes. In the oceanic Southern Hemisphere, oscillations are smoothened by the carbonate/bicarbonate system of the ocean.

Destructive pathways begin by decomposition of the dead biomass. Transition from living to dead biomass is accompanied by autolysis, which liberates part of the organic matter. For cyanobacteria and algae, lysis induced by viruses or phagi is quite important. Density of population is important, and below 10^5 cells per milliliter, phagolysis is ineffective. Lysis produces two components: dissolved organic compounds (DOCs) and particulated organic compounds (POCs), which consist mainly of structural components of the cell. Osmotrophic microorganisms can immediately use DOC; the threshold depends on the dilution with 1–10 mg l^{-1} still utilizable depending on the inflow. POC is to be hydrolyzed by hydrolytic exoenzymes before osmotrophic organisms can utilize it – bacteria in the sea or fungi in terrestrial ecosystems. Destructive pathways are formed by organotrophic organisms, which traditionally are named heterotrophs. This term is imprecise since it refers to the assimilative pathway leading at the end to the secondary production. There are three main metabolic pathways for C_{org}: proteolytic, saccharolytic, and lypolytic, according to the composition of biomass. The Winogradsky rule (1896) says that each natural compound has its specific microbial decomposer. The number of species of prokaryotes exceeds 5000 of cultivated and 2×10^4 clones of noncultivated. That gives sufficient functional diversity to perform biogeochemical essential reactions. As a result, specific trophic groups of organisms characterized by the utilizable substrate (e.g., cellulolytic, or lypolytic, or lignin-decomposing fungi), appear. The set of these organisms makes the functional biodiversity in the trophic system, which should make a complete community for each habitat. In the terrestrial environment, mycelial fungi are most important. Wood consists of 20%–30% lignin, which is decomposed by fungi and that gives the lower limit of their involvement in terrestrial C_{org}-cycle as 1/3–1/4 of CO_2 producers. Adding cellulose decomposition would at least double their contribution.

In the presence of O_2, aerobic organisms regenerate approximately one-third of C_{org} in secondary production with CO_2 as the product of respiration. Consumption of O_2 in the dark is by the usual estimation of respiration by the so-called biochemical oxygen demand (BOD) test.

In an anoxic environment, a cascade of reactions begins with fermentation, which is also the main pathway for many hydrolytic decomposers. As products, a mixture of organic acids and H_2 appears, and this is the reason why this stage is designated as acidogenic or hydrogen producing. Organic acids as nonfermentable compounds can be utilized only with an external oxidant, such as nitrate or ferric iron, sulfate, or CO_2. A cascade of anaerobic reactions makes a community function as an entity with an integrated trophic network.

Without an external oxidant, anaerobic decomposition is completed by methanogenesis, a process which dominates in terrestrial mires and lake mud. In the sea, it takes place in deep layers of sediments, when sulfate is exhausted from interstitial water. Methanogens make a group of Euryarchaeota usually named as *Methano-*…. There are three pathways for methanogenesis:

either hydrogenotrophic with $H_2 + CO_2$, or acetoclastic, or methylotrophic for C-1 compounds. Acetoclastic pathway dominates in C_{org}-abundant environment, for instance, in methane tanks. Methylotrophic methanogens develop noncompetitive pathways in saline environment, while they do not compete with sulfate reducers. Hydrogenotrophic methanogens can use endogenous H_2 formed by reaction of water with superheated rocks and belong to hyperthermophiles, for example, *Methanobacterium fervidus*, which develops at temperatures over 100 °C. More important is the role of hydrogenotrophic methanogens in community, where they act as H_2-sink. They establish H_2-concentration below 10^5 ppm and this allows us to oxidize acetate and other nonfermentable substrates in cooperative action with H_2-producing syntrophic organisms. Biogenic methane is identified by its isotopically light composition. Most of methane is ^{13}C-depleted.

Methane either remains in the sediments or escapes into the oxic zone where it is oxidized with O_2 by a specific group of methanotrophs. Under geologically favorable conditions, methane is stored in sedimentary rocks. Another possibility is the formation of crystallohydrates, which at appropriate hydrostatic pressure and low temperature make an ice-like cover for deep methane. At present, about 500 Mt yr^{-1} of methane comes into the atmosphere, where it is oxidized photochemically. Many times more than this quantity is oxidized by methanotrophs, which form an oxidative filter on the path of CH_4 to the atmosphere. The genera of methanotrophs are designated as *Methylo-*…. Oxidation of CH_4 includes its enzymatic transformation in C-1 compounds in the cell by a special metabolic pathway and thus methanotrophs represent a specialized group of the Proteobacteria. In the ocean, methane is oxidized by anaerobic consortia of methanogens with sulfate-reducing or denitrifying bacteria. The microbial cycle of CH_4 is most important for the biosphere.

Involvement of oxidized N, Fe, and S compounds as oxidants conjugates C_{org}-cycle with cycles of other elements. Transition from oxic to anoxic zone favors retainment of nondecomposed organic matter and leads to formation of oil and gas deposits in aquatic environments and coal in terrestrial ecosystems.

Particulated components including bodies of bacteria are consumed by phagotrophic Protists or/and by zootrophic multicellular animals. The trophic chains of animals are arranged into a trophic pyramid with a number of levels, including herbivorous and carnivorous. The size of the prey determines the nutritional pyramid. Animals that use filtration for nutrition are important in aquatic environment, keeping the density of microorganisms on the threshold level of about 10^5 cells ml^{-1}. The total amount of bacteria in the active zone of the ocean and soil is at least of the order 10^{28}, with the biomass of each cell about 10^{-12} g.

Microbial Nitrogen Cycle

C_{org}-cycle is coupled with N_{org}-cycle. Nitrogen cycle begins by nitrogen fixation. Nitrogenase enzyme is present in prokaryotes exclusively and distributed among different groups. N_2-fixation is an energetically expensive process. It occurs only on severe limitation on the bound nitrogen. It is facilitated in a reductive environment. The main groups of nitrogen-fixing bacteria are cyanobacteria and anaerobic bacteria. Aerobic nitrogen-fixing bacteria are either plant symbionts or organisms within the community supplied in excess by the nitrogen-free organic matter. C_{org}:N_{org} ratio over 20 stimulates N_2-fixation. Fixed N_2 is included into the biomass. Then the regenerative cycle begins. In the sea, its zone is just below the photic zone. Nitrogen is liberated in the proteolytic pathway with ammonification as summation of the process. NH_4^+ is either re-assimilated, or is metabolized by the two-step conversion into nitrate by chemosynthetic nitrifiers in the presence of O_2. Nitrate is assimilated by phytoplankton or plants. Part of it escapes from the productive zone. In the anoxic zone, in the presence of available organic matter, various denitrifiers use nitrate and nitrite as oxidants and reduce it to N_2 closing the cycle. As a variant, nitrate reduction to ammonia occurs but this is a less important pathway. Denitrification is the closing step in the nitrogen cycle. Nitrate makes a reservoir of bound nitrogen in the deep cold ocean of $\sim$20Eg. It is noticeable how important denitrification is in the marine environment for decomposition of organic compounds including hydrocarbons. Limitation of the availability of bound nitrogen is the major problem for plant productivity. In the sea, seasonal exhaust of nitrates definitely determines algal development as it was demonstrated for the Northern Sea. For the Pacific, a peculiar change in plankton composition indicates an N- and P-cycle interrelation: in nitrate-limited conditions cyanobacteria dominate when there is enough phosphate, whereas in phosphate-limited conditions algae are the main group in phytoplankton.

Phosphorus

Phosphorus comes into the ecosystem due to the weathering of rocks. The productivity of the lakes is proportional to the phosphate availability. For instance, soda lakes are often eutrophic in spite of extreme environment and limitation by other elements. Phosphorus is mobilized from its minerals by many acid-producing microorganisms, which make dissolution zones on the plates with an enamel of phosphate containing minerals. Phosphorus of the sea has terrestrial origin. Assimilated phosphate is included into the nucleic acids and phospholipids of the biomass. Regenerative cycle of phosphorus includes liberation of phosphate by the action of phosphatases. Phosphorus escapes from the cycle by binding into insoluble compounds of phosphorites on reaction with Ca and F. It should be noted that deposits of micritic phosphorites were formed by cyano-bacterial mat; phosphatisized microfossils of cyanobacteria are clearly visible in the scanning electron microscope. Cyanobacteria store phosphate as intracellular polyphosphate, which is the transitional source for rapid phosphatization. The iron pump demonstrates liberation of phosphate in anoxic environment: ferric iron binds phosphate in an insoluble compound but reduction to ferrous state liberates phosphates. On a large scale, phosphorus is the limiting element for primary production, which depends on its availability.

Sulfur Cycle

Sulfur cycle is the most important cycle conjugated to C_{org}. Assimilation of sulfate into S_{org} is quantitatively of minor importance in spite of the fact that it is the main source of dimethylsulfide – a volatile compound contributing to the source of S in the atmosphere. Its photochemical oxidation leads to the formation of aerosol in the stratosphere and is most important for the climate. In the destructive pathway coupled to the C_{org}-cycle, sulfate is reduced to H_2S by sulfate-reducing bacteria (SRBs), which by now are taxonomically numerous but functionally uniform. There are the following trophic groups: H_2-utilizers, SRBs with incomplete oxidation of organic acids (*Desulfovibrio*-type) producing acetate, and complete oxidizers (*Desulfobacter*-type), which use various unfermentable products of fermentations. Hydrogenotrophic SRBs are H_2-scavengers, which allow them to serve as intermediary oxidants to H_2-producing syntrophic bacteria and thus oxidize a variety of organic compounds. Most interesting is their interaction with methanogens in anaerobic methane oxidation in marine sediments. Methane is oxidized by reversed methanogenesis with the formation of isotopically light carbonates and evolution of H_2S by SRBs. H_2S, if not bound by iron into pyrite, escapes to the surface of the mud where it is oxidized by pelophilic sulfur bacteria, which can either use intracellular S_0 for oxidation into sulfate, if O_2 or NO_3^- is available, or use it as an oxidant in sulfur reduction. Magnificent benthic mats of trichomic sulfur bacteria are found on the shelf close to Chile and West Africa. Here, the so-called thiobios is formed by sulfur bacteria of *Thioploca*-type. Large filamentous bacteria cross the ox–red boundary and receive H_2S from the anaerobic layer and the current near the surface of the mud brings oxidant as nitrate or O_2, which is used for chemosynthesis. H_2S escaping in the water mass in the bodies of water with limited circulation makes a chemocline with the reductive zone below the oxic zone; Black Sea is a conventional example. The same occurs in stratified lakes. It is supposed that Mid-Proterozoic stratified ocean had the same structure. If H_2S zone comes to the photic zone, anoxic sulfur phototrophs develop. There are a variety of anoxygenic phototrophic bacteria, which belong to phylogenetically distant phyla. Purple layers of phototrophs make a remarkable landscape when they come up to the surface on the beach or in the soda lakes. In oxygenated photic zone, H_2S is oxidized into sulfate by various thionic bacteria. It is noteworthy that the appearance of sulfates in the palaeocean correlates with the oxygenation of the atmosphere around 2.4 Ga, and before 1 Ga its composition became close to the present one. It might be speculated that sulfates of the sea are biogenic in their origin. What was the initial source of mineral S? If the source was massive sulfides, then for their mobilization oxidative step was needed by aerobic acidothiobacteria used now in biohydrometallurgy in the general reaction, $FeS + O_2 \rightarrow Fe^{3+} + SO_4{}^{2-}$. The reaction strongly depends on the availability of O_2. Oxidation of sulfides leads to the formation of extreme acid conditions. It is most spectacular on volcanic thermal fields with sulfur exhalations, so-called solfataras. When A. Humboldt visited Vesuvius before its eruption, he noted that hot vapors were neutral in spite of possible SO_2 production in the heat, while cold walls of the crater were strongly acidic for Lakmus paper. Now it is known that oxidation of sulfur occurs mainly by acidophilic thionic bacteria and only in outlets of fumaroles, extremely thermophilic archaea are active. In the deeper parts of thermal fields, S_0 is used as an oxidant by anaerobic archaea with H_2S production. Short cycle $S_0 \leftrightarrow H_2S$ works also in microbial mats where white sulfur is deposited from H_2S by microaerobic sulfur bacteria and reduced when oxidant is not available. A large variety of microorganisms are involved in the cycle. Sulfur cycle closes destruction of organic matter in anoxic zone with sulfate regeneration either by anaerobic phototrophs or by aerobic sulfur oxidizers. Its function strongly depends on the transport processes across chemocline. The outcome from the cycle depends on availability of iron, which forms insoluble sulfides first as hydrotroillit and then pyrite.

Iron Cycle

The production of H_2S is environmentally incompatible with dissolved iron because of the formation of sulfides. Thus in terrestrial wetlands, where sulfate is limiting, iron cycle develops. Bacterial Fe-cycle takes place now in swamps. It includes oxidation of Fe^{2+}-bicarbonate under O_2-limited conditions with formation of $Fe(OH)_3$ ferrihydrite. Energy of oxidation might be used for chemosynthesis by *Gallionella* with precipitation of Fe(III) on the slimy stalks. Precipitation of Fe(III) on slimy structures is well known for the number of so-called 'iron bacteria', among which *Leptothrix ochracea* is best known for large deposits of ochre in slow-flowing water. Historically, that was the first example of geological activity of microorganisms described by Ehrenberg. Ochre-forming deposits were used as a 'swamp-ore' in the beginning of the Iron Age. However, two processes should be distinguished: chemosynthetic oxidation of iron and precipitation of iron hydroxides on mucous polysaccharides. Both processes are geologically significant. Ferrihydrite is readily reduced by iron-reducing bacteria, which use H_2, acetate, and a number of other C_{org}-compounds as electron donors. There are two possible end products: siderite $FeCO_3$ is formed in excess of organic matter and magnetite Fe_3O_4 under more restricted conditions. Iron-reducing bacteria substitute nitrate reducers in moderately reductive habitats. There are also thermophilic iron reducers. For formation of ferrihydrite in anoxic environment, there are two possible pathways, both phototrophic: one possibility is oxidation of Fe^{2+} by cyanobacteria but it is unclear if it is direct or indirect and caused by O_2 production; and the other is definitely direct and is performed by nonsulfur purple bacteria such as *Rhodomicrobium*. Oxidation of Fe^{2+} by anoxygenic phototrophic nonsulfur bacteria was described only recently. Product of oxidation in the light is ferrihydrite. This process closes the iron cycle in anoxic environment. Large deposits of layered silicified iron oxides composed of alternating layers of hematite and magnetite known as banded-iron formations (BIFs) were formed during the Early Proterozoic 1.8 billion years ago. Their origin remains unclarified. Fe is of hydrothermal origin. Total amount of iron oxides contains about 40% of O_2 evolved corresponding to C_{org} of kerogen. Iron migrated in the ancient ocean most probably as bicarbonate. Period of BIF is clearly incompatible with sulfate reduction.

Oxidation of sulfides, first of all pyrite, involves both cycles of iron and sulfur. Oxidation involves two functions. At low pH Fe^{2+} is stable in the air. Oxidation of sulfide produces sulfuric acid with a drop to pH <2. Some pyrite-oxidizing chemosynthetic bacteria such as *Acidothiobacillus ferrooxidans* use energy of both sulfur and iron oxidation. However in nature, these two functions are often divided between iron-oxidizing *Leptospirillum* and sulfur-oxidizing *Acidothiobacteria* working in concert. There are also other examples of these bacteria, especially thermophilic, which are most important in bacterial hydrometallurgy because they are able to dissolve various sulfides, and copper, gold, and other metals that also come into solution.

In addition to the cycles of major elements used by chemosynthetic bacteria, it is worth mentioning cycles of arsenic, manganese, and selenium, where both oxidative and reductive pathways operate. The general rule is, chemosynthetic microbes use oxido-reductive reaction and develop in the thermodynamic field of stability of the product of this reaction. Energy generated in the reaction must be sufficient to support the formation of ATP.

Biologically Mediated Reactions

The so-called biologically mediated reactions are also very significant in the involvement of microorganisms in the biogeochemical cycles. In these reactions, microbes form an environment in which certain forms of minerals are stable according to the fields of thermodynamic stability. It is most important for rare elements, whose amount is insufficient to ensure the existence of certain specific groups. These trace elements act as indicators of the environment. Uranium is one of the examples. Another example is the deposition of metals in the zone of sulfate reduction. Such microbially mediated pathways form many sedimentary deposits. The so-called biogeochemical barriers represent the sites of drastic changes in the environment, which cause precipitation of minerals. Biogeochemical barriers are sustained by countercurrents of solutes and by the activity of microorganisms. There are different kinds of barriers: oxido-reductive, alkaline, sulfidic, etc. Chemocline in the lakes is an example. Change in the state of environment often causes transition from migrating chemical species to insoluble one.

The cycle of calcium belongs to biologically mediated reactions. It begins by CO_2 weathering of rocks. The essential point here is the concentration of carbon in the biomass and concentrated release of CO_2 during decomposition. Formation of active products of decomposition as acids or chelators is also important. Released Ca^{2+} comes into waterways as $Ca(HCO_3)_2$ and migrates to the ocean. Here it might be used by calcareous eukaryotes for the formation of skeleton and release of CO_2. The main part of $CaCO_3$ at present is biogenic by origin. Another possibility is the release of CO_2 in warm shallow water when its solubility decreases and the reaction, $Ca(HCO_3)_2 \rightarrow \downarrow CaCO_3 + \uparrow CO_2 + H_2O$ develops. The surface of precipitated $CaCO_3$ is covered by a microbial biofilm. It gives to the precipitate a laminated texture due to the slime produced. If the microbial biofilm is formed by cyanobacteria, then the additional sink of CO_2 assimilated in C_{org} might result in a drop of pH and precipitation of carbonate. The layered structures of precipitated carbonates are known as stromatolites. They are recorded for Proterozoic as the most important deposits. Their abundance indicates that they represent significant deposits of CO_2. Stromatolites correlate with deposits of dolomites, but sometimes pieces of black chert are included and in these cherts, silicified microfossils of cyanobacteria are observed. Preservation is excellent and one can identify taxa, with the aid of books on systematics, of extant cyanobacteria to approximately 2.4 billion years ago. The scale of stromatolite formation delineating ancient warm shallow water environment is of the order of millions of square kilometers. Height of such deposits is up to hundreds of meters. Mass development of stromatolites ended with the end of bacterial exclusive domination in the biosphere. Calcium cycle contributes to the neutrophilic conditions on the Earth.

Trophic Organization of Microbial Communities

The organization of the prokaryotic community is most clearly demonstrated by the cyano-bacterial mat. Sign (-) denotes here two components: cyanos as prime producers and bacteria as decomposers in regenerative cycle. In the mat, distinct layers are found: the upper illuminated level is occupied by cyanobacteria; below is the white layer of sulfur bacteria, followed by the purple layer of anoxygenic phototrophs, and then the black layer of sulfide-producing bacteria. Still below are the layers of dead bacteria. The whole system has dimensions of 2–4 mm. It is called in German 'Streiffarbsandwatt'. The architecture of cyano-bacterial mats is similar in hypersaline lagoons, soda lakes, thermal springs, etc. The structure of mat is maintained by exopolysaccharides produced by cyanobacteria, which are edificators (from 'edifice') for the community. The main factor is illumination and self-shadowing by the upper layers of cyanobacteria, which move to the optimal illumination. Minor differences in the composition of mats are caused, for instance, by the absence of purple bacteria in thermal habitats or strong development of planktonic forms in soda lakes. The 'Winogradsky column' illustrates stratified planktonic microbial community: cylinder with water from the site supplemented by mud with organic debris and gypsum at the bottom. Blooming microbes in the column produce alternating black, purple, and green layers. The column might sustain for years.

Trophic links in the microbial community are organized into the trophic network of a cascade of degradative reactions. The rule is that each step should be sufficient to support the species performing the transformation. Degradation begins with hydrolysis of biopolymers, the most resistant being structural components of the cell walls. Aerobic and anaerobic dissipotrophic bacteria in the cascade of reactions utilize low-molecular-weight compounds, dissipating from the sites of hydrolysis. The final result should be complete decomposition of organic matter, so-called 'mineralization'. In fact, decomposition is not complete and recalcitrant

substances are formed in minor part, giving rise to humic substances, and dispersed organic matter. It should be noted that the physical environment strongly contributes to the trapping of undecomposed organic matter, preventing microbial activity.

This is a brief description of biogeochemical cycles catalyzed by bacteria. The main conclusion is that bacteria act as a cooperative community with the interlinked metabolic pathways of the main elements; only the cooperative community is autonomous due to the links between productive and regenerative cycles. Each step of catalysis is performed by a functional or trophic group of specialized bacteria. Links provide the trophic network. Cyclic pathways make such a community autonomous, depending mainly on the energy for photosynthesis. Cooperative community is an operational unit for the ecosystem at the landscape level. However, biogeochemical cycles are not entirely closed: there is formation of products, which escape recycle. The changes in the community composition known as succession are caused by the accumulation of products as well as exhaust of substrates. In the microbial community, it is the development from fast-growing copiotrophs to a climacteric community with well-balanced interactions. In fact, the microbial community exists all the time in a transitional state.

When we consider a larger temporal scale, the most important concept of biogeochemical succession arises. It may be illustrated by the composition of the atmosphere, which is formed by microbial activity, since the main components of the atmosphere are metabolized by bacteria: CO_2, O_2, CH_4, CO, H_2, N_2, NO_x, NH_3, and sulfur species. Due to the accumulation of the waste products – oxygen is the most evident example – biosphere becomes uncomfortable to the community, here the initial anoxic microbial community. Less evident but more important is accumulation of C_{org} in sedimentary rocks, which rolls cycles with the passage of time. As a result, the atmosphere moves from the neutral to the oxygenated state. It changed conditions on the Earth's surface. Biosphere overturned: anaerobic pockets remain under the shield of aerobic O_2 consumers. Much the same occurred with the ocean (or hydrosphere), which is in equilibrium with the atmosphere. Three main steps seem to be identified: the first before approximately 2.4 billion years with the domination of iron cycle; beginning of the biosphere with O_2 in the atmosphere and pronounced S-cycle; and present-day biosphere with O_2-atmosphere where biogenic O_2 substituted part of CO_2. Pathways in the community were reoriented to the new environment. Consider that any sustainable system should be able to support its own existence by effective feedbacks otherwise it is not sustainable. S. Winogradsky in 1896 suggested the qualitative concept of cycle of life as a 'huge organism' (or the goal-oriented system) with microorganisms acting as the main catalysts. Later V. Vernadsky in *The Biosphere* (1926) introduced the quantitative approach, considering biogeochemical cycles as the main mechanisms. The Geospheric–Biospheric Program and Global Change concept represent the contemporary approach to the problem. The expediency of the links in the biosphere leads to its interpretation as 'Gaia'.

However, biosphere was always within the geographic envelope of the Earth. This means that there was always a mosaic of landscapes arranged in climatic zones. Landscapes give the possibility of lateral interaction and formation of geochemical barriers. The mosaic of landscapes furnishes refugia, places for survival of particular communities. Landscapes on geological timescale are dependent on the tectonic. Weathering–sedimentation pathway leads to equilibrium if no metamorphism and geological cycle occurs.

Since bacteria catalyzed main cycles and established the primary biogeochemical system, they form the dynamic environment, into which Protists, Metaphyta, and Metazoa were incorporated in the course of evolution. The system was developed by the substitution of prime producers by algae, kelps, and plants. The terrestrial system changed with the appearance, about 300 Ma ago, of vascular plants, which changed the atmospheric hydrological cycle by the involvement of deeper layers of ground water and producing a new illuminated surface within the leaves for derivates of cyanobacteria converted into chloroplasts. Plant cover significantly changed the terrestrial environment. However, microbes remain as the main catalysts in the system of biogeochemical cycles.

Further Reading

Brock, T.D., Madigan, M.T., 1991. Biology of Microorganisms, 6th edn. Englewood Cliffs: Prentice-Hall, 874pp.
Lengeler, J.W., Drews, G., Schlegel, H.G. (Eds.), 1999. Biology of the Prokaryotes. Stuttgart: Thieme, p. 955.
Schlesinger, W.D., 1997. Biogeochemistry: An Analysis of Global Change, second ed. San Diego: Academic Press, 588pp.
Zavarzin, G.A., 2004. Lekcii po Prirodovedcheskoi Mikrobiologii (Lectures in Environmental Microbiology). Moscow: Nauka.

Nitrogen Cycle☆

TP Burt, Durham University, Durham, UK

Introduction

The nitrogen cycle is arguably the second most important cycle, after the carbon cycle, to living organisms. Nitrogen is essential to plant growth, and therefore is a significant contributor to the human food chain, but its presence in the environment is strongly influenced by anthropogenic activities.

The global nitrogen cycle is outlined first; then long-term trends are examined at national and global scales for both terrestrial and aquatic ecosystems; next the transport of nitrogen at local and long-distance scales is described; finally there is consideration of how public policy for environmental protection can seek to mitigate against pollution effects.

Nitrogen was discovered in 1772 by Daniel Rutherford, who called the gas 'noxious air'. During the late eighteenth century other chemists, such as Scheele, Cavendish, Priestley, and Lavoisier were also studying 'dephlogisticated' air, the term then used for air without oxygen. By the late nineteenth century its vital role as a plant nutrient was understood and by the early twentieth century, the Haber–Bosch process was able to 'fix' nitrogen from the atmosphere on an industrial scale. Nitrogen fixation influences the amount of food present within an ecosystem. Prior to the industrial process of N production, crop growth was sustained by recycling crop residues and manures on the same land where food was grown. Any 'new' N was created by growing rice or legumes, or by mining guano and nitrate deposits. However, as the human population increased, so has the demand for food and with that the dependence on inorganic fertilizers to sustain agriculture. This trend has affected the nitrogen cycle at global, national, and local scales.

The Nitrogen Cycle

Nitrogen comprises approximately 79% of the Earth's atmosphere in the form of biologically unavailable dinitrogen (N_2) gas. This reservoir is estimated to be in the order of 3.8×10^9 kg N, approximately 90% of the global reservoir. Crustal reservoirs comprise the remaining 10% (Figure 1). By comparison, the amount of N stored in the biomass (terrestrial and oceanic) and soil is small, but this, of course, is the vital component as far as living organisms are concerned.

The global nitrogen cycle (Figure 2) is driven by biological and physical processes, which depend on a variety of environmental factors such as solar energy, precipitation, temperature, soil texture, soil moisture, the presence of other nutrients, and atmospheric CO_2 concentrations.

These factors control N fluxes into and out of soils and vegetation, thereby influencing the mass of N in these compartments, and therefore its availability. Figure 3 illustrates the global distribution of nitrogen in soil and vegetation. Tropical forest soils show the least amount of storage because of high decomposition and uptake rates but have higher N storage. In general terms, human activity has tended to accelerate nitrogen cycling, increasing flux rates from one store to another.

In order for nitrogen to be used for plant growth, it must be available in inorganic formal ammonia (NH_3), ammonium (NH_4), nitrite, (NO_2), or nitrate (NO_3). In the terrestrial nitrogen cycle (Figure 4), soil nitrogen cycling processes dominate, with surface application (fertilizer and manure) providing most of the nitrogen inputs. Microbes break down organic matter to produce much of the available nitrogen in soils. Mineralization/immobilization, nitrification, nitrate leaching, denitrification, and plant uptake can then occur. Nitrate is completely soluble in water and since it is not adsorbed to clay particles, it is vulnerable to being leached out of the soil by percolating rainfall or irrigation water. Generally, the movement of nitrogen can occur in one of three directions: (1) upward – crop uptake and gaseous loss, (2) downward – as leaching to groundwater, and (3) lateral – via surface and subsurface flow to surface waters.

The nitrogen cycle is strongly influenced by anthropogenic activities. During the twentieth century, land-use changes, such as intensive agriculture, over-fertilization, deforestation, biomass burning, combustion of fossil fuels, industrial activities, and energy production, have significantly disturbed 'natural' N biogeochemical cycling. In natural ecosystems plant growth rates are low and annual uptake of N is relatively small. Cultivated crops are much more demanding with nutrient uptake ranging from about 100 kg N ha $year^{-1}$ for wheat and up to 450 kg N ha $year^{-1}$ for sugar cane. Improved grasslands for livestock rearing typically require 250 kg N ha $year^{-1}$. The mineralization capacity of soils is almost always insufficient to maintain optimum

☆*Change History:* March 2013. TP Burt updated Figure 6.

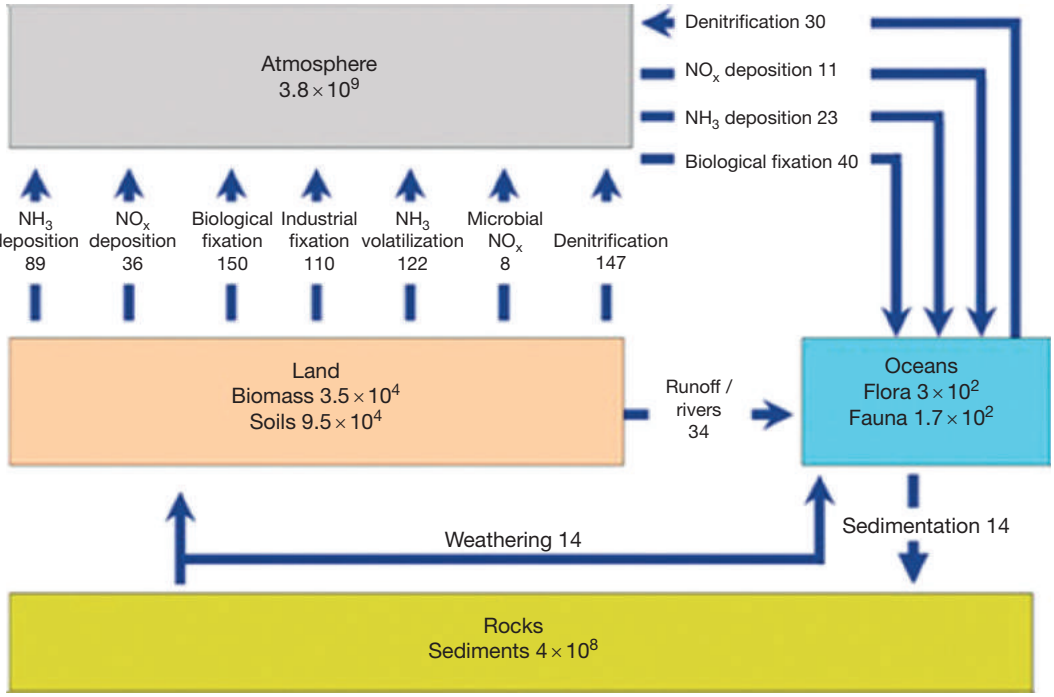


Figure 1 Global nitrogen reservoirs (units kg N year^{-1}) and fluxes (units $\times 10^9$ kg N year^{-1}).

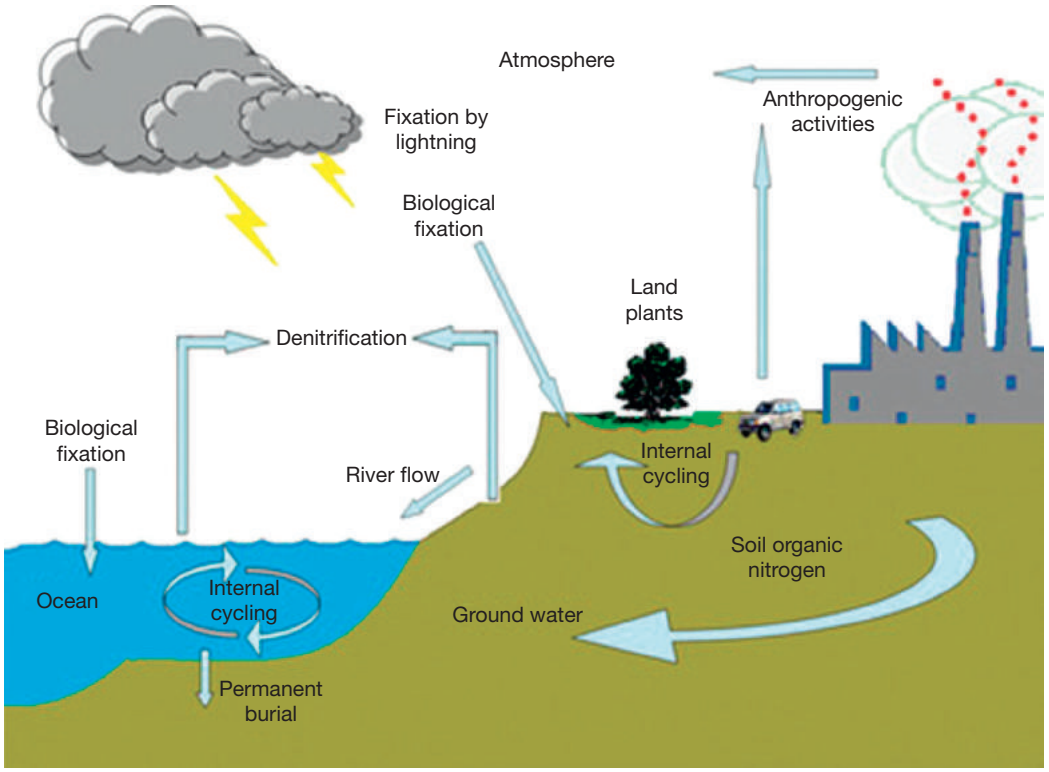


Figure 2 Global nitrogen cycle.

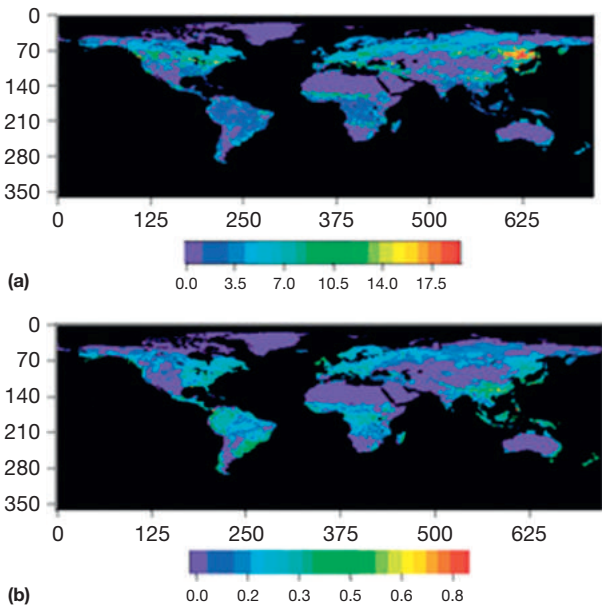


Figure 3 Global distribution of nitrogen storage (kg m^{-2}) in soil (a) and vegetation (b). Reproduced from Bin-Le Lin, Sakoda A, Shibasaki R, Gato N, and Suzuki M (2000). Modelling a global biochemical nitrogen cycle model in terrestrial ecosystems. *Ecological Modelling* **135**(1), 89–110, with permission from Elsevier.

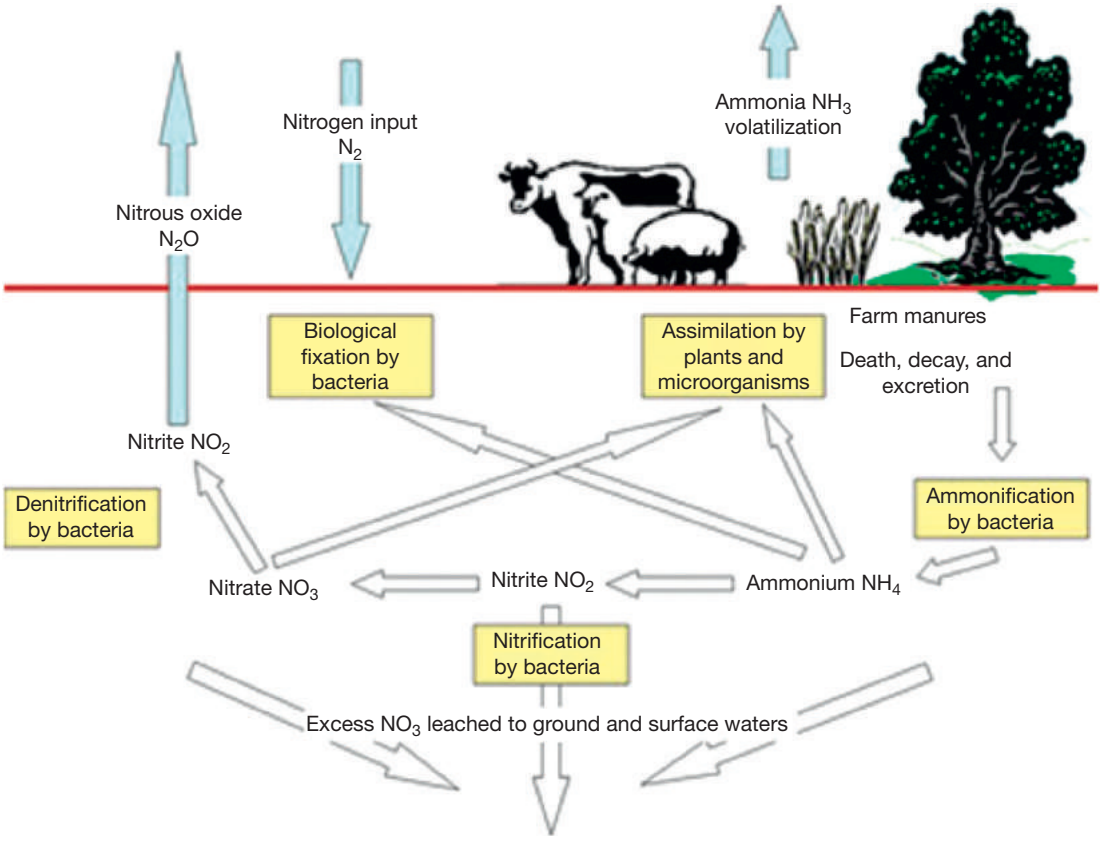


Figure 4 The terrestrial nitrogen cycle.

growth; therefore, chemical fertilizers and manures are required to supply N for intensive agriculture. This has resulted in changes to the long-term trends within the N cycle at global, regional, and local scales.

Long-Term Global and Regional Trends in the Nitrogen Cycle

Globally, nitrogen is found in the terrestrial ecosystem as dead organic matter (89.5%), with live biomass accounting for 4% and inorganic nitrogen 6.5% of this source. Natural sources of nitrogen have seen a small decline since 1890 (Table 1). Losses of biomass due to large-scale burning and forest clearances during the late twentieth century have contributed to the decline of this reservoir.

Natural reservoirs now cannot provide nitrogen in the quantity required for global food production. In 1890, total anthropogenic N production was approximately 15×10^9 kg N year^{-1}, but by 1990 this had risen by approximately an order of magnitude to 140×10^9 kg N year^{-1}.

In the terrestrial ecosystem, globally, nitrogen production is generally driven by the use of fertilizers for intensive agriculture, with cultivation and combustion contributing approximately 38% to all anthropogenic sources. However, this is not evenly distributed across the world regions. Asia produces almost half the world's nitrogen fertilizers, followed by Europe then North America (Table 2). Africa, Latin America, and Oceania combined contribute less than 12% of global nitrogen production.

Significant changes to the nitrogen cycle have been apparent since the 1960s. This is closely linked to expanding human populations and an increasing demand for food and energy. Creation of anthropogenic nitrogen in Asia increased from $\sim 14.4 \times 10^9$ kg N year^{-1} in 1961 to $\sim 68 \times 10^9$ kg N year^{-1} by 2000, and is set to increase to 105×10^9 kg N year^{-1} by 2030. North America doubled its N production between 1961 and 1997, with most of the increase occurring during the 1960s and 1970s. Although the largest increase was in use of inorganic N fertilizer, emissions of NO_x from fossil fuel combustion also increased substantially. By 1997, even though N fixation had increased, fertilizer use and NO_x emissions had increased more rapidly and two-thirds of reactive N inputs were denitrified or stored in soils and biota, while one-third was exported, the largest export being in riverine flux to coastal oceans, followed by export in food and feeds, and atmospheric advection to the oceans. The consumption of meat protein is a major driver behind N use in agriculture in North America. Without changes in diet or agricultural practices, fertilizer use will increase over the next 30 years, and fluxes to coastal oceans may increase by another 30%.

Similar trends are mirrored in the European N budget (Table 3). By 1990, N inputs are approximately 34.5×10^9 kg N year^{-1}, the major process of N fixation being fertilizer production at $\sim 14.0 \times 10^9$ kg N year^{-1}, with industry and combustion accounting for a further $\sim 3.3 \times 10^9$ kg N year^{-1}. Imported N from products such as animals, animal feeds, food, fertilizers, forestry products, exceeds the amount of N exported outside Europe. Furthermore, exports in riverine flux to oceans accounts for $\sim 4.0 \times 10^9$ kg N year^{-1}.

Nitrogen Export by Rivers

Water is a carrier of N from pollution source to river outlet. The fraction that ultimately reaches the outlet depends on amount of runoff and distribution between different runoff components. Time delay between inputs at the soil surface and inputs to surface water additionally depends on groundwater residence times. The natural water quality of a river will be determined primarily by the catchment soil type and underlying geology to which water, falling on the catchment as rain, is exposed as it drains to the river. Climate provides an important context for nitrogen cycling by controlling the propensity for carbon and nitrogen to be stored within the catchment; thus in the UK, upland soils tend to conserve organic matter as peat, whereas organic matter tends to decompose much more readily in lowland soils. Deviations from this baseline water quality are generally caused by the influence of human activities through point and diffuse pollution sources. Up to 40% of the total nitrogen flux reaches the aquatic system through direct surface runoff or subsurface flow. Nitrogen delivery to surface waters is further controlled by (1) soil structure and type, (2) rainfall, (3) the amount of nitrate supplied by fertilizers, and (4) plant cover and root activity.

In pristine river systems, the average level of nitrate is about 0.1 mg l^{-1} as nitrogen (mg l^{-1} NO_3-N). However, in Western Europe, high atmospheric nitrogen deposition results in nitrogen levels of relatively unpolluted rivers to range from 0.1 to 0.5 mg l^{-1} NO_3-N. In recent years, nitrate concentrations in European rivers have remained relatively stable after rising rapidly towards the end of the twentieth century; progress still needs to be made in reducing the concentration of nitrate in Europe's rivers. High rates of nitrogen input to rivers and coastal waters are not confined to Europe. In USA as late as 1998, more than one-third of

Table 1 Nitrogen production (10^9 kg N year^{-1})

Nitrogen production (10^9 kg N year^{-1})	*1890*	*1990*
Anthropogenic sources	15.0	140.6
Terrestrial ecosystem	100.0	89.0
Marine ecosystems	140.0	140.0
Fixation by lightning	5.0	5.0
Total	~260	~374

Table 2 Global anthropogenic nitrogen production 1990 ($\times 10^9$ kg N year^{-1})

World region	*Fertilizer production*	*Cultivation*	*Combustion*	*Total*
Africa	2.5	1.8	0.8	5.1
Asia	40.1	13.7	6.4	60.2
Europe and former Soviet Union	21.6	3.9	6.6	32.1
Latin America	3.2	5.0	1.4	9.6
North America	18.3	6.0	7.4	31.7
Oceania	0.4	1.1	0.4	1.9
Total	~86	~31	~23	140.6

Table 3 European N Budget 1990

N input	*$\times 10^9$ kg N year^{-1}*	*N output*	*$\times 10^9$ kg N year^{-1}*
N-fertilizer production	14.0	Denitrification	13.8
Combustion and industry	3.3	Emissions of NH_3 and NO_x	7.8
Biological N fixation	2.2	Sewage and industry	2.6
Deposition	7.3	Riverine flux to oceans	4.0
Imported products	7.6	Exported products	6.3
Total	34.5	Total	34.5

Table 4 Nitrogen inputs to rivers and coastal waters

River	*N inputs to rivers (kg year^{-1})*	*N exports to coastal waters (kg year^{-1})*
Mississippi	7 489	597
Amazon	3 034	692
Nile	3 601	268
Zaire	3 427	632
Zambezi	3 175	330
Rhine	13 941	2 795
Po	9 060	1 840
Ganges	9 366	1 269
Chang Jiang	11 823	2 237
Juang He	5 159	214

all river miles, lakes (excluding the Great Lakes), and estuaries did not support the uses for which they were designated under the Clean Water Act (1987). For example, Table 4 illustrates the extent of N inputs to rivers and coasts in areas of America, Africa, and Asia. These trends are cause for concern as seasonal hypoxia develops during the summer months, resulting in a depletion of sea bed vegetation and changes in fish stocks.

It is now widely acknowledged that agriculture is the main source of N pollution in surface waters and groundwater in rural areas of Western Europe and USA. The UK House of Lords' report Nitrate in Water (1989) commented on the conflicts that can arise when the use of land for farming comes into conflict with the use of land for water supply. Concern for this initially focused on the alleged links between high nitrate concentrations in drinking water and two health problems in humans: the 'blue-baby' syndrome (*methaemoglobinaemia*) and gastric cancer. Now, there are also major concerns about environmental degradation. Nutrient enrichment in water bodies encourages the growth of aquatic plants (see Figure 5).

Reed beds and other marginal plants may be attractive on a small scale, but when these and, particularly, underwater plant growth are excessive, this can cause a narrowing of waterways, and become a nuisance to recreational users of rivers and lakes. Furthermore, eutrophication (a group of effects caused by nutrient enrichment of water bodies) can adversely affect the aquatic ecosystem, especially in marine systems where nitrogen may be the limiting nutrient. An algal bloom may cut out light to the subsurface, and when it dies, decomposition uses the oxygen supply needed by other species. Some algae are toxic to fish, while others, for example, cyanobacterial species, are toxic to mammals including domestic pets. Studies in Asia have demonstrated the link between increasing use of fertilizers and increasing incidence of algal blooms. For example in some Chinese provinces, fertilizer application is greater than 400 kg N ha^{-1}. This is usually applied as a single application and with crop utilization efficiency as little as 30–40%, a high proportion is lost to rivers, lakes, and coastal waters. The environmental impact at the regional level is the incidence of red tides (algal blooms). During the 1960s less than 10 red tides per year were recorded, but in the late 1990s over 300 per year were being recorded.

Figure 5 Choked watercourse, River Skerne, UK. Source: P. Widdison.

Land-Use Controls to Reduce N Enrichment to Surface Waters

The popular misconception that the nitrate problem is caused by farmers applying too much nitrate fertilizer is too simplistic. Nevertheless, there is now little doubt that the high concentrations of nitrate in freshwaters noted in recent years have mainly resulted from runoff from agricultural land and that the progressive intensification of agricultural practices, with increasing reliance on the use of nitrogenous fertilizer, has contributed significantly to this problem. Since 1945, agriculture in the industrialized world has become much more intensive. Fields are ploughed more frequently; more land is devoted to arable crops, most of which demand large amounts of fertilizer; grassland too receives large applications of fertilizer to ensure a high-quality silage for winter feed; stocking densities in general are higher leading to increased inputs of manure on grassland and problems of disposal of stored slurry; cattle often have direct access to water courses resulting in soil and bank erosion and direct contamination from animal waste; many low-lying fields are now under-drained, encouraging more productive use of the land and speeding the transport of leached nitrate to surface water courses. Lowland rivers close to urban areas may receive larger quantities of nitrogen from sewage effluent, but budgetting studies confirm that agriculture is still the main source of nitrate in river water, except in the most urbanized river basins. Nevertheless, sewage now provides a greater proportion of total nitrogen flux than in former decades following policy initiatives to limit nitrogen losses from agriculture.

In mainland Britain mapped nitrate concentrations demonstrate a marked northwest to southeast gradient, reflecting relief, climatic conditions, and agricultural activity. Upland areas in the north and west are usually characterized by nitrate concentrations below 1 mg NO_3-N l^{-1}. This reflects the high rainfall and low temperatures of such areas: upland soils tend to conserve organic matter and mineralization rates are low. In contrast, a decreasing ratio of runoff to rainfall and an increasing intensity of agricultural land use toward the south and east of Britain results in higher mean concentrations of nitrate in river water. Many of the lowland rivers are characterized by concentrations above 5 mg NO_3-N l^{-1}; in East Anglia and parts of the Thames basin, mean nitrate concentrations in rivers are close to the European Union limit of 11.3 mg NO_3-N l^{-1}, a level exceeded in some spring waters especially in the Jurassic limestone of the Cotswold's and Lincolnshire Wolds.

The changing pattern of British lowland agriculture since 1945 is reflected in long-term records of nitrate for surface and groundwaters (Figure 6). Such graphs confirm the accelerated nitrogen cycling in recent decades and increasing fluxes from the terrestrial to aquatic compartments of the N cycle.

For both large and small rivers, there has been a relatively steady upward trend in nitrate concentrations, often of the order of 0.1–0.2 mg NO_3-N l^{-1} a^{-1}. Analyses for relatively short time series of just a few years have shown that the upward trend may be interrupted, either because of climatic variability (drier years are associated with lower nitrate concentrations) or because of land-use change. Nevertheless, analysis of long time series shows that the main effect is a steady increase in nitrate levels over time which is independent of climate. If trends continue, the mean nitrate concentration of many rivers in Europe will soon be above the EU limit; in many cases this level is already exceeded during the winter when nitrate concentrations reach their maximum. However, modelling studies now suggest that trends are beginning to level off and will gradually fall over the next few decades, although the downward trend is at a disappointingly slow rate. In catchments where groundwater is the dominant discharge source, this

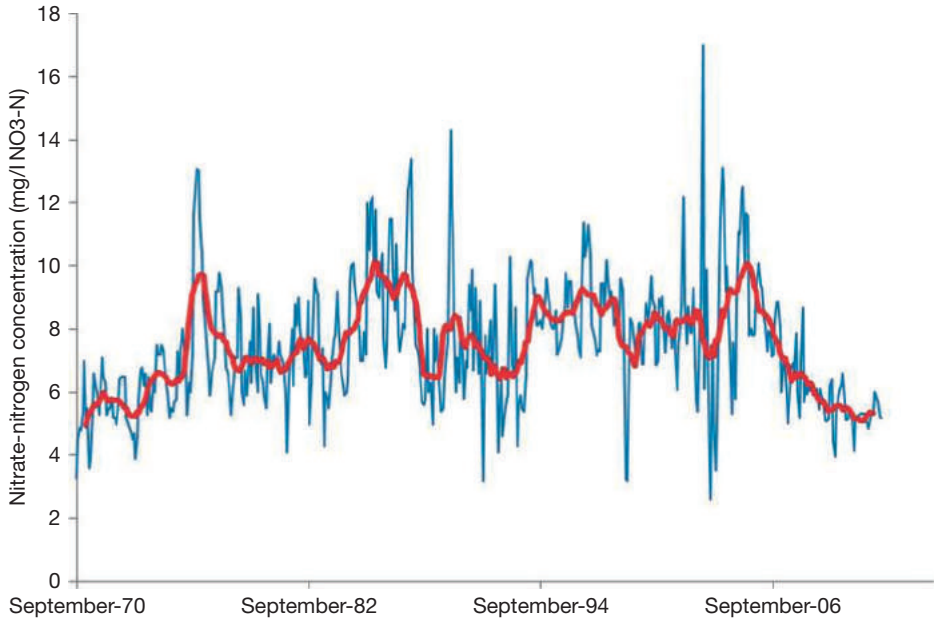


Figure 6 Long-term nitrate record 1970–2012: Slapton Wood catchment (UK).

long-term trend may be prolonged since it may take years for nitrate to percolate down to the saturated zone. In such basins, nitrate pollution will remain a problem for decades to come. In recent years, a number of options have been considered as a means of halting the upward trend.

Trends in water management in Europe include moves toward catchment-level management, improved intersectoral coordination and cooperation, and frameworks facilitating stakeholder participation. This approach is developed by the European Union in its Water Framework Directive (2000/60/EC), which sets targets for good ecological status for all types of surface water bodies and good quantitative status for groundwater. More localized schemes, like the UK Nitrate Vulnerable Zones, involve greater restrictions on farming practice, such as restricting the amount and timing of organic and inorganic fertilizer application. The EU Common Agricultural Policy has changed the way payments are made to farmers. Single-farm payments encourage farming in a more environmentally friendly way. Financial payments are available to farmers for loss of income or for changing farming practice such as improving slurry storage and fencing off watercourses to restrict livestock access. Much interest currently focuses on the use of riparian land as nitrate buffer zones.

The terrestrial–aquatic ecotone occupies the boundary zone between the hillslope and the river channel, usually coinciding with the floodplain. Given their position, near-stream ecotones can potentially function as natural sinks for sediment and nutrients emanating from farmland. Observed denitrification rates in floodplain sediments may be sufficient to remove all nitrate from groundwater flowing under a riparian woodland, with a floodplain width of 30 m. Saturated, anoxic soils, rich in carbon, are exposed to nitrate-rich groundwater. Rates of denitrification are high within this zone since the nutrients required by denitrifying bacteria are abundant. Wetlands and wet meadows (defined as areas where the water table is at or above land surface for long enough each year to promote the formation of hydric soils and support the growth of aquatic vegetation) also have potential as nitrogen sinks. High production rates by wetland vegetation result in an abundance of carbon providing an organic substrate for bacterial processes. Wetland plants transport oxygen into anaerobic sediments which can enhance denitrification leading to losses of nitrogen as N_2O or N_2 from wetland sediments.

The type of vegetation found on the floodplain controlling the efficiency of nitrate absorption is the subject of much debate. Several studies have argued the presence of trees is crucial, yet others state the role of surface vegetation is secondary to presence of saturated conditions together with a carbon-rich sediment. Denitrifying bacteria operate best at the junction anaerobic/aerobic zones where both carbon and nitrate are abundant. It is clear that nitrate losses may be reduced by creating a nutrient-retention zone between the farmland and the river. Given that many floodplains around the world are part of an intensive agricultural system, creating permanently vegetated buffer strips between field and water courses is an idea that should be actively promoted. However, buffer strips will only be successful nutrient sinks if they are managed in an appropriate way. Underlying artificial drainage should be broken or blocked up to prevent a direct route to the watercourse for solutes and grassland strips need maintenance to prevent them becoming choked with sediment and losing their sediment retention potential.

Solving the problem of nutrient enrichment of surface waters cannot be seen in the short-term. Long-term land-use change is needed. Taking farm land immediately adjacent to water courses out of production is one option that could go some way to allow modern agriculture and water supply to coexist in the same basin. Such proposals inevitably raise questions about who pays for them – farmers, water supply companies, or the taxpayers.

Further Reading

Addiscott TM (1996) Fertilizers and nitrate leaching. In: Hester RE and Harrison RM (eds.) *Agricultural chemicals and the environment. Issues in environmental science and technology*. Cambridge, UK: Royal Society of Chemists.

Betton C, Webb BW, and Walling DE (1991) *Recent trends in NO_3-N concentration and load in British rivers*. Wallingford: IH Press IAHS publication 203, pp. 169–180.

Burt TP and Johnes PJ (1997) Managing water quality in agricultural catchments. *Transactions of the Institute of British Geographers* 22(1): 61–68.

Burt TP, Heathwaite AL, and Trudgill ST (eds.) (1993) *Nitrate: Processes, patterns and management*. Oxford: Wiley.

Butcher SS, Charlson RJ, Orians GH, and Wolfe GV (eds.) (1992) *Global biogeochemical cycles*. London: Academic Press p. 379.

De Wit M, Behrendt H, Bendoricchio G, Bleuten W, and van Gaans P (2002) The contribution of agriculture to nutrient pollution in three European rivers, with reference to the European nitrates directive. *European Water Management Online*.

Eckerberg K and Forsberg B (1996) Policy strategies to reduce nutrient leaching from agriculture and forestry and their local implementation: A case study of Laholm Bay, Sweden. *Journal of Environmental Planning and Management* 39(2): 223–242.

Haycock NE, Burt TP, Goulding KWT, and Pinay G (eds.) (1997) *Buffer zones: Their processes and potential in water protection*. Harpenden: Quest Environmental.

Hem JD (1970) *Study and interpretation of the chemical characteristics of natural water*. Washington: United States Government Printing Office 363 pp.

Howden NJK, Burt TP, Mathias SA, Worrall F, and Whelan MJ (2011a) Modelling long-term diffuse nitrate pollution at the catchment-scale: Data, parameter and epistemic uncertainty. *Journal of Hydrology* 403(3–4): 337–351.

Howden NJK, Burt TP, Worrall F, Mathias S, and Whelan MJ (2011b) Nitrate pollution in intensively farmed regions: What are the prospects for sustaining high-quality groundwater? *Water Resources Research* 47: W00L02. http://dx.doi.org/10.1029/2011WR010843.

Kessler E (ed.) (2002) Special report. *Ambio* 31(2).

Lin B-L, Sakoda A, Shibasaki R, Goto N, and Suzuki M (2000) Modelling a global biogeochemical nitrogen cycle model in terrestrial ecosystems. *Ecological Modelling* 135(1): 89–110.

Norse D (2003) Fertilisers and world food demand. Implications for environmental stress. In: *IFA-FAO Agriculture Conference Rome*http://www.fertilizer.org/ifa/publicat/PDF/2003_rome_norse.pdf.

Ribaudo M (2001) Non-point source pollution control policy in the USA. In: Shortle JS and Abler DG (eds.) *Environmental Policies for Agricultural Pollution Control*. Oxford: CAB International.

Sprent JI (1987) *The ecology of the nitrogen cycles*. Cambridge: Cambridge University Press 151 pp.

White RE (1987) *Introduction to the principles and practice of soil science*. New York: Blackwell 244 pp.

Noosphere

C Jäger, Potsdam Institute for Climate Impact Research, Potsdam, Germany

The Noosphere Concept

The noosphere concept is best developed before the background of the related concept of ecosphere. The ecosphere is usually understood to be the space inhabited by living beings. It comprises the living organisms (biosphere), the lower atmosphere, the hydrosphere (oceans, lakes, glaciers, etc.), and the highest layer of the lithosphere (topsoil as well as various kinds of rocky ground). The word biosphere was invented by the Austrian geologist Eduard Süß, who used it more or less in passing, in an influential textbook on the formation of the Alps. In 1911, Süß met the Russian-Ukrainian mineralogist and geochemist Vladimir Vernadsky, who gave the word its current meaning. This meaning includes the fact that the biosphere is connected in space and time, that all living beings are related to each other by evolution, and that not only the biological, but also the chemical and physical, processes in the biosphere are shaped to a considerable extent by the functioning of living beings. A major example is the oxygen content of the atmosphere resulting from photosynthesis.

In the 1920s, Vernadsky was staying in Paris where he met the philosopher and mathematician Edouard LeRoy, whose lectures on biogeochemistry he attended. Through LeRoy, Vernadsky got exposed to a concept that Teilhard de Chardin, who also attended LeRoy's lectures, was developing in those days: the concept of noosphere. (The term noosphere, is derived from the Greek root *nous* meaning mind.)

Teilhard, a French geologist and Catholic priest, saw the emergence of the human species out of biological evolution as the beginning of a far-reaching transformation of the world we live in. The human mind would gradually learn to shape the world to a larger and larger extent, transforming the biosphere into the noosphere. Vernadsky related the concept to the historical dimension he had experienced in World War II. In his mind, this war showed that humankind was beginning to act on a global scale, but was not yet able to do so in a responsible way. The development of nuclear physics – that Vernadsky had been following already before World War I – presented the same challenge in an even more dramatic form. The transition from the biosphere to the noosphere, then, was to be the process in which humankind would learn to consciously and responsibly shape the ecosphere. This idea has been taken up in various forms by current authors interested in global environmental change.

Related Concepts

According to Venadsky, ``The Noosphere is the last of many stages in the evolution of the biosphere in geological history'' (Vernadsky, 1945, p. 10). The word ``evolution'' here does not refer to the interplay of variation and selection that Darwin saw at work in the evolution of biological species. Rather, it hints at a process in which new realities emerge in the course of time without any need for inheritance of traits between biological generations. This line of thinking is related to the idea of `emergent evolution' proposed by the psychologist Lloyd Morgan and further developed by LeRoy. Today, the emergence of new realities in the course of time is often described as a process of self-organization in complex systems. Evolutionary history then becomes an overarching narrative telling the story of the world as a whole. It tells how physical matter rearranged itself up to the point where portions of it became the first living organisms, how these then evolved into species of increasing organic complexity, how the complexity of some organisms enabled them to develop the mental faculties that characterize humankind, and how humankind is now beginning to understand its own global environmental impacts.

The concept of the noosphere is also related to the concept of Gaia proposed by Lovelock and Margulis. The Gaia concept pictures the Earth as a complex, self-regulating system, a kind of organism that maintains conditions favorable to life despite a variety of disturbances. The emergence of the noosphere then means that some living beings – humans – became aware of this larger organism they are part of, of their capability to modify it by technological means, and of their responsibility to develop these means in ways that do not disrupt Gaia.

Closely related is a new concept of Earth system. Traditionally, Earth scientists considered as the Earth system those physical and chemical processes taking place on planet Earth that shaped oceans and continents, forming rocks, causing earthquakes, etc. Living beings were seen as playing a rather peripheral role (although for obvious reasons fossil fuels always were a big topic for the Earth sciences), and the influence of human beings on the Earth system was considered negligible. The debate about global environmental change and sustainability has changed this situation. As a result, a broader concept of Earth system has been proposed by Schellnhuber and others. In this perspective, the Earth system is seen as a complex system including physical, chemical, biological, as well as social and mental processes. Some sort of emergent evolution is seen as leading from a purely physicochemical system first to a biogeochemical system and then to one including human beings and their interactions. The first transition can be described as the emergence of the ecosphere, the latter as the emergence of the noosphere.

Finally, the role of humankind in shaping the face of the Earth has been used to propose a new geological epoch, the Anthropocene, supposed to start more or less with increased control over natural resources due to application of fossil energies during the industrial revolution in the nineteenth century. So far, geological epochs were defined to be periods of millions of years, and the last such epoch, the Holocene, has been defined to start just about 10 000 years ago. The concept of the Anthropocene marks a clear break with the previous practice of structuring a geological timeline. However, others have suggested that humankind significantly altered the climate system already some 8000 years ago by clearing forests. On a timescale of millions of years, this would make the beginning of the Holocene and the Anthropocene indistinguishable. On a conceptual level, of course, there still is a major difference between defining the current geological epoch in terms of an ice age that came to an end independent from any human action or in terms of the emergence of humankind as a new geological force. It is the latter approach that clearly relates to the concept of the noosphere.

Mechanisms and Institutions

As Vernadsky realized, the concept of the noosphere implies a causal chain from human thoughts to large-scale physical effects. This poses two challenges for research. First, there is the question of how the movements of human hands, legs, and bodies can be amplified so as to have effects that are observable at a planetary scale. And second, there is the question of how human thoughts can cause movements of hands, legs, and bodies.

As for the first question, fire has been a key amplification mechanism of human action since prehistorical times. Clearly, the burning of fossil fuels with the resulting emission of greenhouse gases is a related mechanism today. Vernadsky was particularly impressed by an amplification mechanism that was developed during his lifetime. The human capability to think had led to an understanding of subatomic processes that enabled human beings to build atomic bombs as well as to generate electricity from nuclear power plants. Vernadsky had studied radioactive materials already before World War I; during World War II he played a key role in triggering the nuclear weapons program under Stalin, and he forcefully supported the Soviet nuclear energy program. It is noteworthy that Lovelock, champion of the Gaia concept, strongly advocates nuclear power as the way to meet the challenge of anthropogenic climate change.

Nuclear physics is a prime example of human thoughts whose material impacts – while clearly being huge – depend mainly on political decisions. However, it is clear that the market institution is one of the most effective mechanisms to enlarge the range of human actions. The market economy has enabled human beings to develop global patterns of division of labor, of cooperation and competition. So far, research drawing directly or indirectly on the noosphere concept has not paid much attention to the economic links in the causal chain from thoughts to material impacts. This clearly is a major research challenge for the future. It includes the task of distinguishing those impacts of the market economy that change our global environment without impairing it from those that jeopardize properties of our environment that we value and need. Will the noosphere concept be helpful in new discoveries about how markets work and how key instances of market failure can be addressed?

Body and Soul

Vernadsky was fully aware of the fact that the second question – how human thoughts can cause changes in the material environment – was a key research challenge posed by the noosphere concept. Nowadays, brain research holds promise of important elements to address that question. However, when imagining that these elements will be sufficient to answer the question, a simple fact is ignored: what can be found in the human skull are neurons, synapses, electrochemical reactions, but no thoughts. One may expect that some day we will be able to establish a one-to-one relation between certain brain processes and certain thoughts, a bit as playing music from notes is based on a correspondence between certain marks on paper and certain sounds. But this does not mean that marks on paper and sounds are the same things. The noosphere concept challenges environmental research to reflect on one of the weak points of contemporary scientific culture: the difficulty in developing coherent arguments about the relations between movements of the human body and what was once called the human soul.

Research in logic has helped to clarify the role of domains of discourse for the development of arguments. For logical inference to be possible, participants in a debate must share the ability to refer to individuals – stones, dreams, numbers, rainbows, people, whatever – in some reasonably well-defined domain. This ability has a price, however: the domain itself must be presupposed; attempts to refer to it within the logical discourse it supports lead to paradoxes and eventually contradictions. Discourse A can refer to the domain of discourse B, but not to its own domain. The domains of discourse used in biogeochemistry, however, are quite different from the ones needed to talk about human thoughts. Perhaps a new domain of discourse needs to be established before the intuition conveyed by the noosphere concept can be used in reliable professional research. Using a word like `noosphere' as if one had a great unified domain of discourse at hand, however, can be not only inspiring, but also seriously confusing.

The world as a whole is not a possible subject of logical inferences. This led Wittgenstein to suggest that accepting silence, mysticism if one wishes, was the appropriate stance toward the world in its entirety. Later, however, he realized that this silence was interwoven with a different kind of speech. In a letter to his friend Drury, a psychiatrist who at one stage wondered whether it would not have been better to become an academic, he wrote: ``Look at your patients more closely as human beings in trouble and enjoy more the opportunity you have to say `good night' to so many people'' (Rhees, 1984, p.109f). We may call this way of using

words – as in honestly wishing `good night' to somebody in trouble – poetic. Developing a domain of discourse is a poetic craft, a way of world-making, perhaps. Of course, the argumentative and the poetic use of words are not mutually exclusive; but sometimes the former is more appropriate, sometimes the latter. And this can lead one to wonder whether the noosphere concept does not fit a poetic use of language more than an argumentative one.

See also: Global Change Ecology: Biosphere: Vernadsky's Concept

Further Reading

Crutzen, P.J., 2002. The Anthropocene: Geology of mankind. Nature 415, 23.
Jaeger C (2003) A note on domains of discourse. Logical know-how for integrated environmental modelling. *PIK-Report No 86*. Potsdam: Potsdam Institute of Climate Impact Research.
LeRoy, E., 1928. Les Origines Humaines et l'Åvolution de l'intelligence. Paris: Bolvin.
Lloyd Morgan, C., 1923. Emergent Evolution. London: William & Norgate.
Lovelock, J.E., Margulis, L., 1974. Atmospheric homeostasis by and for the biosphere: The gaia hypothesis. Tellus 26, 2–10.
Rhees, R., 1984. Ludwig Wittgenstein, Personal Recollections. Oxford: University Press.
Ruddiman, W.F., 2003. The anthropogenic greenhouse era began thousands of years ago. Climatic Change 61, 261–293.
Samson, P.R., Pitt, D. (Eds.), 1999. The Biosphere and Noosphere Reader: Global Environment, Society and Change. London: Routledge.
Schellnhuber, H.J., Wenzel, V., 1999. Earth System Analysis. Integrating Science for Sustainability. Berlin: Springer.
Schneider, S.H., Miller, J.R., Crist, E., Boston, P.J. (Eds.), 2004. Scientists Debate Gaia: The Next Century. Cambridge, MA: MIT Press.
Süß, E., 1875. Die Entstehung der Alpen (The Origin of the Alps). Vienna: W. Braunmuller.
Teilhard De Chardin, P., 2004. The Future of Man. Garden City, NY: Doubleday, (first published during 1920–1952).
Vernadsky, V.I., 1945. The biosphere and the noosphere. Scientific American 33 (1), 1–12.
Vernadsky, L., 1997. The Biosphere. New York: Springer, (first published in 1926).
Wittgenstein, 2001. Tractatus Logico-Philosophicus. London: Routledge, (first published in 1921).
Wittgenstein, L., 2001. Philosophical Investigations. London: Routledge, (first published in 1953).

Oxygen Cycle☆

DJ Wuebbles, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Introduction

Without atmospheric oxygen, life on Earth would be extremely different. The increase in atmospheric oxygen to present levels, and the corresponding increase in atmospheric levels of ozone, with their ability to absorb biologically harmful high-energy radiation from the Sun, allowed life to evolve and to emerge from the oceans to the land, undergoing an amazing evolution to its present diversity.

Oxygen is one of the most abundant elements on our planet. It is the most abundant element by mass in Earth's crust and is found in most rocks. Oxygen also accounts for 89% of the mass of the oceans. After molecular nitrogen, oxygen is the second most abundant element in Earth's atmosphere, with molecular oxygen accounting for 20.95% of the atmospheric content. There is nearly uniform mixing of molecular oxygen in the atmosphere until above the mesosphere, roughly 80 km above Earth's surface. Therefore, because of the nearly exponential decrease in pressure with altitude, the bulk of molecular oxygen is found in the first few kilometers above Earth's surface. However, although the turbulent mixing in the lower atmosphere keeps the molecular oxygen in a nearly constant mixing ratio (relative to the total air density), measurements indicate that there are seasonal latitudinal variations of as much as 15 ppm (parts per million molecules of air). These seasonal variations are most pronounced at high latitudes in the Northern Hemisphere, where the seasonal cyclic variations in photosynthesis and respiration are most strongly felt.

While there are many gases and particles in the atmosphere containing oxygen, the other gas that needs to be discussed as part of the oxygen cycle because of its great importance to both the atmosphere and life on Earth is ozone. While atmospheric concentrations of ozone are much smaller (ppm levels) than that for molecular oxygen, ozone is important for several reasons: (1) it absorbs biologically harmful levels of ultraviolet (UV) radiation, keeping this radiation from reaching the Earth's surface; (2) it is a 'greenhouse' gas that influences the Earth's climate; and (3) the direct contact with ozone pollution in the lower atmosphere can be harmful to plants, animals, and humans.

The following sections discuss the oxygen cycle in more detail, focusing on the processes affecting molecular oxygen and ozone in the atmosphere, the historical changes in the amounts of these important gases, and their projections for the future.

The Oxygen Budget

Atmospheric molecular oxygen is generated and consumed by a wide range of processes in the Earth system. Biological, chemical, and physical processes, both on and beneath Earth's surface, all contribute to the budget for oxygen in Earth's atmosphere. **Figure 1** shows a simple representation of the budget for oxygen. This figure also shows how the oxygen budget is closely linked to the carbon cycle. While it is difficult to measure the annual fluxes shown in **Figure 1**, oxygen isotopes (^{17}O, ^{18}O) do provide constraints leading to the approximate values shown.

The major mechanism by which molecular oxygen is produced on our planet is through photosynthesis. As seen in **Figures 2** and **3**, the net reaction of photosynthesis is to convert carbon dioxide (CO_2) and water to molecular oxygen. This large annual flux is counteracted by the equally large annual flux from the effects of respiration in removing oxygen from the atmosphere. Thus, as suggested by **Figures 1** and **2**, the current atmospheric equilibrium is maintained by this cycle between photosynthesis and respiration. As indicated in **Figure 3**, there are several slower removal processes in addition to this more rapid cycling.

The mean residence time of molecular oxygen in the atmosphere is roughly 4000 years. With the approximate balance between photolysis and respiration, this residence time is largely controlled by the long-term burial of reduced carbon in ocean sediments. This burial of organic matter results is initiated via the photosynthesis from organisms in the upper levels of the ocean, whose carbon is then deposited to deeper levels of the ocean. The actual rate of burial depends on the area of the ocean floor subjected to anoxic conditions, which in turn varies inversely with the concentration of atmospheric oxygen. The balance between the burial of

☆*Change History:* February 2013. DJ Wuebbles updated the following sections: The Recent Decline in Atmospheric Oxygen and Human Effects on Ozone, Figures 5 and 6, and the Further Reading.

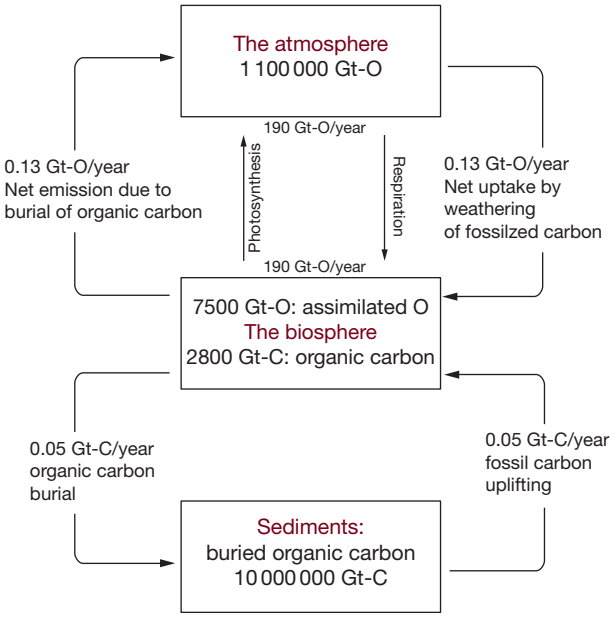


Figure 1 A simple model of the oxygen cycle, and its relationship to the carbon cycle, that considers the three major reservoirs: the atmosphere, the biosphere, and the sediments. Redrawn from graph presented by http://atoc.colorado.edu/~fasullo/pjw_class/images/oxygencycle.gif.

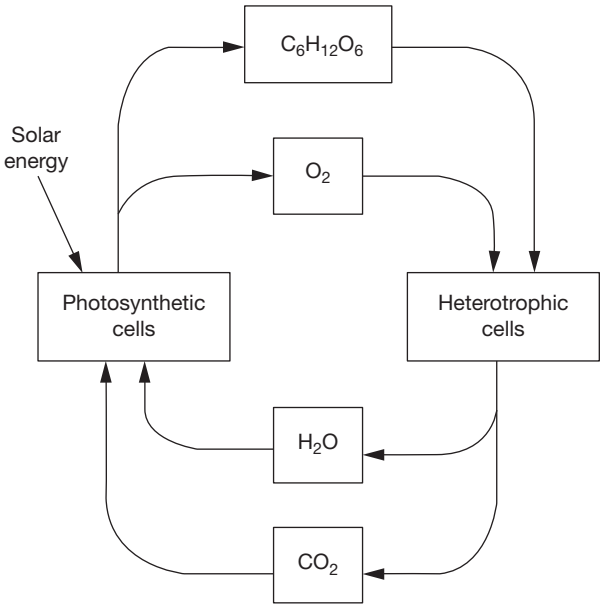


Figure 2 A representation of the basic exchange of atmospheric oxygen between biospheric systems.

organic matter and its oxidation thus plays a significant role in the maintenance of the atmospheric molecular oxygen at 20.95% of the atmospheric density.

A large amount of oxygen has also been consumed by weathering of reduced crustal materials, such as those containing iron and sulfur, through the geologic history. However, the residence time of atmospheric oxygen at the current rate of exposure would be roughly 2 My because of removal mechanisms.

The History of Atmospheric Oxygen

Primitive life began in the absence of free oxygen. However, in the first 400 My of Earth, bacteria-like organisms developed that could take advantage of the light energy from the Sun to initiate photosynthesis, although early production of oxygen likely oxidized crustal materials instead of building up in the atmosphere. Geochemical evidence suggests that there was little oxygen in

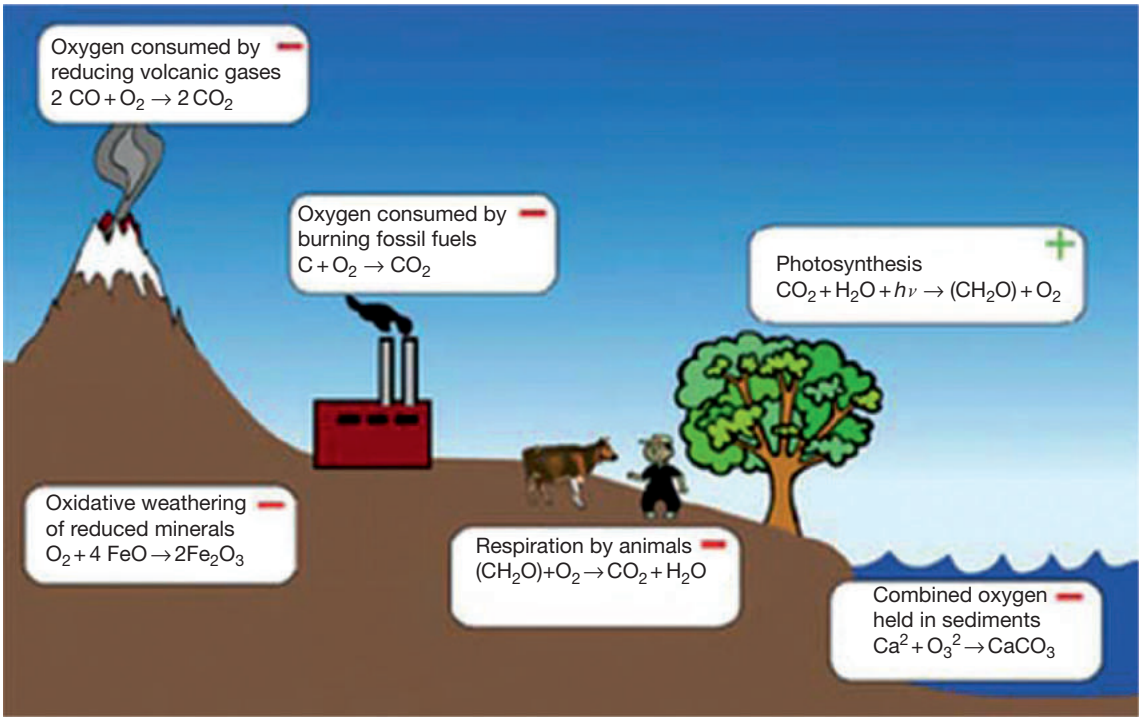


Figure 3 A simple representation of the key chemical reactions controlling the fluxes of oxygen to and from the atmosphere.

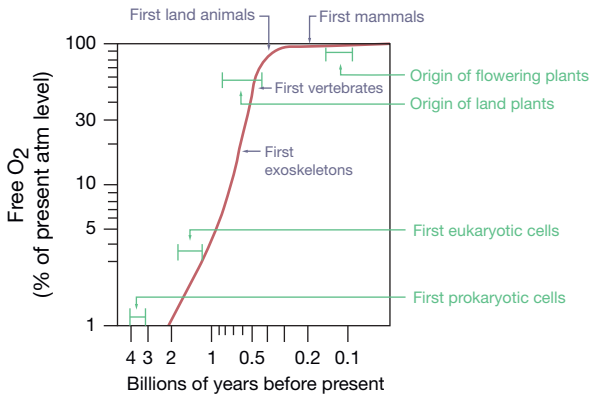


Figure 4 Estimated growth of free oxygen in Earth's atmosphere. Based on graph from http://www.ldeo.columbia.edu/edu/dees/U4735/projections/free_o.html.

the atmosphere during the Achaean (~2.5–4 billion years ago), but near the beginning of the Proterozoic, about 2.5 billion years ago, cyanobacteria – photosynthetic prokaryotes also known as blue-green algae – started the process that eventually led to a massive increase in the concentration of atmospheric oxygen. The definitive worldwide change that signaled the appearance of a significant increase in free oxygen was the appearance of red beds, stratified layers of sedimentary rocks, characterized by abundant red oxides of iron (ferric oxides), which occurred about 2 billion years ago. When the oxygen began to accumulate, the resulting oxidation of other atmospheric constituents totally changed the nature of the atmosphere and the oxygen 'poisoning' was devastating to many of the existing life forms then found on Earth and its oceans. Some bacteria, however, were able to endure the oxygen atmosphere. A symbiosis between bacteria and the former free-living mitochondria enabled eukaryotes to eventually evolve in response to the crisis. Oxygen-based metabolism came into being. The environment changed and life began to evolve.

Figure 4 provides an estimate of the growth of free oxygen in Earth's atmosphere, along with the corresponding evolution of various types of life forms found on our planet. With the increasing proliferation of life, the amount of photosynthesis increased and atmospheric oxygen also continued to increase. The oxygen buildup eventually resulted in the appearance of the first higher order cells with nuclei, called eukaryotes or eukaryotic cells (some analyses suggest this occurred earlier than shown in the figure). These cells depend on atmospheric oxygen for the formation of complex energy-producing compounds. Such cells provide the makeup of all nonbacterial (or nonbacteria-like) forms of life on Earth.

Atmospheric oxygen grew to its present levels of roughly 21% of the atmospheric content about 400 Ma. While it was likely slowed in reaching this state by consumption in terrestrial weathering processes, the amount of atmospheric oxygen, as suggested by the paleontologic record, appears to have varied little since then. The variations that have occurred (while we have little evidence of actual variations, concentrations may have varied from as little as 15% to as much as 30% during the last 500 My) are likely related to the periods when there were strong variations in the deposition of organic matter.

The Recent Decline in Atmospheric Oxygen

Since 1989, when Ralph Keeling of the Scripps Institution of Oceanography began extremely high precision measurements, there has been a steady decline of molecular oxygen in our atmosphere. The decrease is small, just 0.04% through 2009 or an annual rate of about 20 molecules per million molecules of atmospheric oxygen. Measurements of the bubbles of air trapped in ice cores from Greenland and Antarctica suggest that this oxygen decline started in the late eighteenth century, at the beginning of the industrial revolution, when fossil-fuel burning increased dramatically.

This decrease in atmospheric oxygen is not unexpected, for the combustion of fossil fuels, while in the process of producing carbon dioxide, destroys O_2. For every 100 atoms of fossil-fuel carbon burned, it has been estimated that roughly 140 molecules of O_2 are consumed. Since every molecule of additional carbon dioxide locks up two oxygen atoms, the free oxygen decline is greater than the carbon dioxide increase. However, the rate of decline in molecular oxygen is only about two-thirds of that expected from fossil-fuel combustion. While it has not been fully verified, the difference may be explained by the increase in biomass known to occur as a result of the increase in CO_2. Plants grow a bit faster than before, leading to a greater storage of carbon in tree wood and soil humus. For each atom of extra carbon stored in this way, roughly one molecule of extra oxygen accumulates in the atmosphere.

Although the oxygen decrease is unlikely to get to be large enough an effect to be of major concern, it is expected to continue in the future as fossil burning continues and concentrations of atmospheric carbon dioxide continue to increase.

The Role of Ozone

As oxygen increased in Earth's atmosphere, so did the levels of ozone, O_3. The formation of the ozone layer in the upper atmosphere is generally believed to have played an important role in the development of life on Earth, particular in the development of life on land. The accumulation of oxygen molecules in the atmosphere allowed for the production of ozone. Gradually, the increasing levels of ozone led to the formation of the stratosphere, a region of the upper atmosphere where temperature increases with altitude largely as a result of the absorption of solar radiation by ozone. The resulting screening of lethal levels of solar UV radiation by the ozone layer is thought to have been important in allowing life to migrate from the oceans onto the land.

Ozone, O_3, is composed of three oxygen atoms and is a gas at atmospheric pressures and temperatures. Approximately 90% of the atmospheric ozone is in the stratosphere. Most of the remaining ozone is in the troposphere, the lower region of the atmosphere extending from Earth's surface up to roughly 10 km at midlatitudes and 16 km in the tropics. At midlatitudes, the peak concentrations of ozone occur at altitudes between 20 and 30 km. At high latitudes, the peak occurs at lower altitudes, largely as the result of atmospheric transport processes and the lower height of the tropopause (the transition region between the troposphere and the stratosphere).

Ozone in the stratosphere is often called 'good' ozone because it protects life on Earth from harmful levels of UV radiation from the Sun. Therefore, a decrease in the amount of ozone would allow an increase in the amount of the UV radiation from the Sun to reach Earth's surface. Corresponding to an increase in UV are projected significant impacts on ecosystems and human health, including increase in incidences of skin cancer, eye cataracts, damage to genetic DNA, and suppression of the efficiency of the immune system. It is the concern about the increased biologically harmful levels of UV from the decreasing levels of ozone that has largely been the driver for policy actions to protect the ozone layer.

On the other hand, ozone near Earth's surface is called 'bad' ozone because of its direct effects on plants, ecosystems, and humans. Ozone pollution is a concern during the summer months because strong sunlight and hot weather result in harmful ozone concentrations in the air we breathe. Many urban and suburban areas throughout the world have high levels of 'bad' ozone during summer months, and the effects of winds can carry these high ozone levels to rural areas.

Breathing ozone can trigger a variety of health problems including chest pain, coughing, throat irritation, and congestion. It can worsen bronchitis, emphysema, and asthma. 'Bad' ozone can also reduce lung function and inflame the linings of the lungs. Repeated exposure may permanently scar lung tissue. Ground-level ozone also damages vegetation and ecosystems. It leads to reduced agricultural crops and commercial forest yields, reduced growth and survivability of tree seedlings, and increased susceptibility to diseases, pests, and other stresses such as severe weather.

Ozone can also radiatively affect Earth's climate. Ozone absorbs solar radiation but it is also a so-called greenhouse gas that can absorb infrared radiation from Earth that otherwise would be emitted to space. It is the balance between the solar and infrared radiative processes that determines the net effect of ozone on climate. Decreases in ozone in the stratosphere above about 30 km (roughly 18 miles above Earth's surface) tend to increase the surface temperature as a result of the increased absorption of solar radiation, effectively increasing the solar energy that warms Earth's surface. Below about 30 km, decreases in ozone tend to cool the

surface temperature, as the infrared greenhouse effect dominates in this region. Scientific analyses have shown that the decrease in stratospheric ozone over recent decades has had a cooling effect, counteracting a fraction of the warming effect over this time from increasing concentrations of carbon dioxide and other greenhouse gases.

The Production and Destruction of Ozone

Without human intervention, the stratospheric ozone layer would be produced and destroyed through natural processes. The amounts of ozone in the stratosphere vary naturally throughout the year as a result of production and destruction processes, and as a result of winds and other transport processes that move the ozone molecules around the planet. In addition, changes in ozone occur associated with changes in the solar radiation reaching the Earth during the 11-year solar cycle and with various events such as large explosive volcanic eruptions.

Production of ozone in the stratosphere results primarily from photodissociation of oxygen, O_2, molecules. The breaking of the molecular bond by high-energy solar photons at wavelengths less than 242 nm results in oxygen atoms that generally react rapidly with an oxygen molecule to form ozone. The sequence of reactions to form ozone is represented as

$$O_2 + h\nu(\lambda < 242\,\text{nm}) \rightarrow O + O \quad [1]$$

$$O + O_2 + M \rightarrow O_3 + M \quad [2]$$

where $h\nu$ represents a photon, λ is wavelength, and M is a third atmospheric gas, normally N_2 or O_2, the primary components of air.

Since the atmosphere is uniformly filled with oxygen molecules, most of the ozone is generated where there is a balance between the decreasing atmospheric density with altitude and where there is available UV solar radiation in the wavelength region less than 242 nm. This occurs primarily in the upper stratosphere.

The high-energy solar radiation needed to produce ozone is largely absorbed in the stratosphere resulting in the production of the ozone layer. Too little UV radiation reaches the troposphere for it to be a major cause of ozone production in this region. In contrast, the lesser amount of ozone formed in the troposphere is largely through a series of chemical reactions, generally referred to as smog reactions. The primary source of ozone in the troposphere (and in the smog in urban areas) is the conversion of nitric oxide, NO, to nitrogen dioxide, NO_2, which then photolyzes at visible wavelengths to release an oxygen atom that produces ozone through reaction (see eqn [2]). The transport of stratospheric ozone to the troposphere is also important to the budget of tropospheric ozone.

If there was no natural ozone destruction, most of the oxygen in the stratosphere, and perhaps throughout the atmosphere, would eventually be converted to ozone. Such concentrations of ozone would be intolerable to many forms of life on Earth, both because of the direct toxicity of ozone and because of the resultant elimination of any UV radiation reaching Earth's surface.

Ozone photodissociates at UV and visible wavelengths to produce O and O_2. However, because the oxygen atom will generally react to reform ozone, this mechanism produces no net change in the amount of odd oxygen. The actual destruction of ozone and odd oxygen in the stratosphere occurs mainly through catalytic reactions with other gases. In the stratosphere, important catalysts for ozone destruction include nitric oxide (NO), hydroxyl (OH), chlorine (Cl), and bromine (Br). Such gases can be directly substituted for X in the following catalytic mechanism,

$$X + O_3 \rightarrow XO + O_2 \quad [3]$$

$$XO + O \rightarrow X + O_2 \quad [4]$$

which results in the net reaction of $O + O_3$ being converted to two oxygen molecules. As gas X is recycled through these reactions, it continues to destroy ozone until some other reaction converts X to a less reactive form, such as HNO_3 or HCl. In addition to the mechanism described by reactions [3] and [4], there are a number of other catalytic mechanisms also affecting ozone. In this way, a single chlorine atom can lead to the net destruction of thousands of ozone molecules. Because of such cyclic reaction mechanisms, relatively small concentrations of reactive chlorine in the stratosphere can have a significant impact on the amount and distribution of stratospheric ozone.

Human Effects on Ozone

Human activities have had a devastating effect on the concentration and distribution of stratospheric ozone over the last few decades. Measurements of ozone by satellite and ground-based measurements over the last several decades indicate that stratospheric ozone levels have decreased. Atmospheric ozone has decreased globally by more than 10% since 1970 (see **Figure 5**). Atmospheric measurements have also shown that the depleted levels of ozone have indeed increased the amount of UV at the surface. The significant global decrease in stratospheric ozone since the 1970s is well correlated with the increasing amounts of chlorine and bromine in the stratosphere. The sources of this chlorine and bromine are chlorofluorocarbons (CFCs) and other halocarbons produced industrially for a variety of uses such as refrigerants in refrigerators, air conditioners, and large chillers; as propellants for aerosol cans; as blowing agents for making plastic foams; and as solvents for dry-cleaning and for degreasing of

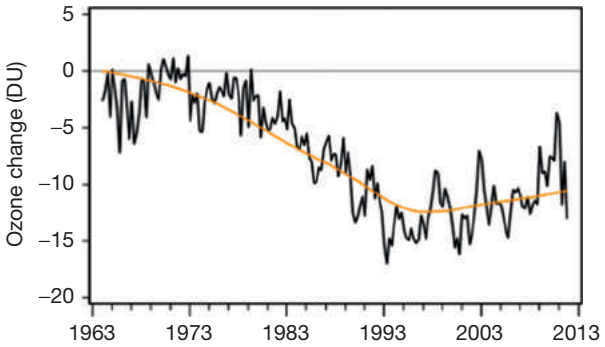


Figure 5 Deviations in total ozone with time relative to 1964 from various ground-based and satellite measurements (TOMS and SBUV). The data are area weighted over 90° S–90° N with seasonal, quasi-biennial oscillations and 11-year solar cycle variations removed. Graph provided by Vitali Fioletov as an update to earlier analyses presented in Fioletov, V.E. (2008). Ozone climatology, trends, and substances that control ozone. *Atmosphere–Ocean* **46**(1), 39–67, http://dx.doi.org/10.3137/ao.460103.

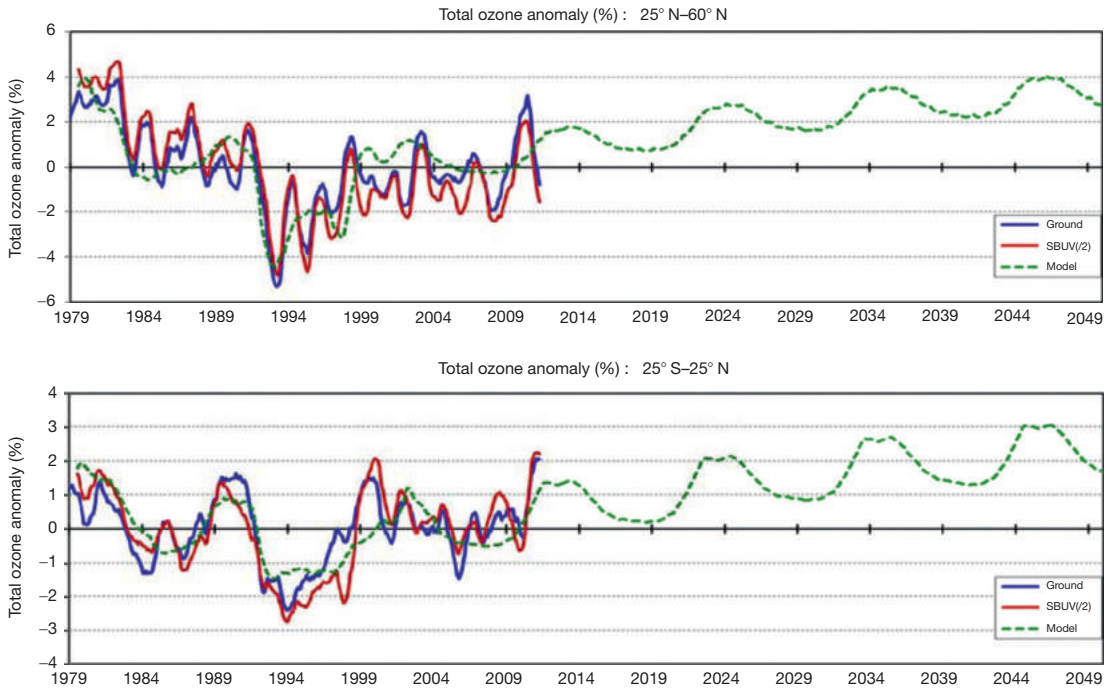


Figure 6 Annual average total ozone for region 25–60 N (top) and 25° S–25° N (bottom) as percent departure from the monthly average climatology from 1979 to 2011. Comparison of trend based on measurements from the solar backscatter ultraviolet (SBUV) satellite instrument relative to ground-based observations and with the derived trend from the University of Illinois zonally averaged chemistry-transport model of the global atmosphere. The effects of the natural cycle due to the quasi-biennial oscillation (QBO) have been removed from the satellite data. While the overall decline in ozone results from increases in concentrations of stratospheric chlorine and bromine from human-related emissions of various halocarbons, the cyclic variations in the observations and model results are due to the effects of the 11-year solar sunspot cycle, while the extra deep minimum in ozone in the early 1990s is due to the effects of emissions from the Mount Pinatubo volcanic eruption in 1991. The model results also show the projected recovery of ozone if the Montreal Protocol reduces halocarbon emissions and atmospheric concentrations as expected. Note that the zonally averaged model does not represent QBO effects on variations in ozone.

materials. Atmospheric measurements have clearly corroborated theoretical studies showing that the chlorine and bromine released from the destruction of these halocarbons in the stratosphere are reacting to destroy ozone. Figure 6 shows an excellent comparison between the observed trend in ozone based on measurements from the total ozone mapping spectrometer (TOMS) and the solar backscatter ultraviolet (SBUV) satellite instrument along with ground-based data and results from the University of Illinois zonally averaged model that is based on the changes in atmospheric gases and particles. In addition to the human-related emissions, natural forcings on ozone from the effects of the 1991 Mount Pinatubo volcanic eruption (due to sulfur emissions) and from the effects of solar flux variations during the 11-year sunspot cycle also contribute to the observed trends. Figures 5 and 6 also show that since the late 1990s, ozone has begun to increase slowly as the effects of the Montreal Protocol to protect the ozone layer have reduced the amounts of chlorine and bromine in the stratosphere. However, as discussed later, full recovery from these effects will take many more decades.

Beginning in the late 1970s, a special phenomenon began to occur in the springtime over Antarctica, referred to as the Antarctic ozone 'hole.' A large decrease in total ozone, now over 60% relative to prehole levels, has been observed in the springtime (September–November) over Antarctica. Dr. Joseph Farman and colleagues first documented this rapid springtime decrease in Antarctic ozone over their British Antarctic Survey station at Halley Bay, Antarctica. These analyses attracted the attention of the scientific community, who soon found that decreases in the total ozone column were greater than 50% compared with the historical values observed by both ground-based and satellite techniques. At the time, there was no expectation that such a phenomenon would be discovered. As a result of the Farman paper, a number of hypotheses arose attempting to explain the large ozone destruction in the springtime over Antarctica. It was initially proposed that the chlorine catalytic cycle might explain the observed ozone decrease, but this did not match the expected ozone decrease possible from the reactive chlorine available at the high latitudes. A special measurement campaign in 1987, as well as later measurements, proved that chlorine and bromine chemistry was indeed responsible for the ozone 'hole,' but because of heterogeneous reactions occurring on polar stratospheric clouds in the lower stratosphere.

The air over the Antarctic becomes extremely cold during the winter as a result of the lack of sunlight over the polar region and because of the greatly reduced mixing of the lower stratospheric air over this region with the air outside this region. During the winter, a circumpolar vortex, also called the polar winter vortex, forms, which isolates the air in the polar region from that outside the region as a result of a stratospheric jet of wind circulating between approximately 50° and 65° S. The extremely cold temperatures inside the vortex lead to the formation of clouds in the lower stratosphere (from roughly 12 to 22 km), called 'polar stratospheric clouds.' Heterogeneous reactions occur on these particles that convert less reactive forms of chlorine to ones that are much more reactive with ozone. When daylight starts to occur over Antarctica in the early spring, this chlorine is available to react with and destroy ozone. Bromine compounds and nitrogen oxides can also react heterogeneously on the particles of these clouds. The ozone destruction continues until the polar vortex breaks up, usually in November.

In the late 1980s, it was generally thought that the Arctic lower stratosphere did not get cold enough to lead to decreases in ozone during the winter and springtime such as those found in the Antarctic. The polar vortex is not generally as strong in the Northern Hemisphere, and, although polar stratospheric clouds would form, it is not likely that they would last long enough for extensive decreases in ozone. However, since 1990, ozone decreases of as much as 30% have been found in the Arctic in those years when lower stratospheric temperatures in the Arctic vortex have been sufficiently low to lead to ozone destruction processes similar to those found in the Antarctic ozone 'hole.' As with Antarctica, large increases in reactive chlorine concentrations have been measured in the regions where the large ozone destruction is occurring.

The recognition of the harmful effect of chlorine and bromine on ozone triggered international policies to restrict the production and use of CFCs and halons and protect stratospheric ozone. These included the 1987 Montreal Protocol on substances that deplete the ozone layer, the subsequent 1990 London Amendment, the 1992 Copenhagen Amendment, and the 1997 Montreal Amendment. In the Montreal Protocol and its Amendments, there is a distinction between the control measures in developed and developing countries. These agreements initially called for reduction of CFC consumption in developed countries. A November 1992 meeting of the United Nations Environment Program held in Copenhagen resulted in substantial modifications to the protocol because of the large observed decrease in ozone, and called for the phase-out of CFCs, carbon tetrachloride (CCl_4), and methyl chloroform (CH_3CCl_3) by 1996 in developed countries. As part of this, the United States has, through the Clean Air Act, eliminated production and import of these chemicals. Production of these compounds is to be totally phased out in developing countries by 2006, while production of halons in developed countries was stopped in 1994. Human-related production and emissions of methyl bromide were not to increase after 1994 in developed countries, with total elimination by 2005. Hydrochlorofluorocarbons, many of which have been used as replacements for the CFCs, still contain chlorine that can destroy ozone, and are to be phased out in the developed countries by 2030.

Worldwide compliance with the international agreements to protect ozone is resulting in significant reductions in the emissions of the CFCs, halons, and other halocarbons having the largest effects on ozone; as a result, levels of stratospheric ozone should slowly begin to recover over the coming decades as the reactive chlorine and bromine in the stratosphere declines. This recovery will be gradual, primarily because of the long time it takes for CFCs and halons to be removed from the atmosphere. Atmospheric model results, such as those shown in **Figure 6**, suggest that the effects of halocarbons on ozone should return to 1980 ozone levels by 2040–50. At the same time, changes in Earth's climate and in the amounts of atmospheric methane and nitrous oxide are likely to further affect the amount and distribution of stratospheric ozone.

While ozone slowly recovers, scientists and policymakers will need to work together to ensure that new problems do not develop as a result of the introduction of new chemicals into the marketplace. They will also need to interact with industry and governments to ensure that the potential effects of increasing concentrations of other gases changing as a result of human activities, such as methane and nitrous oxide, do not produce their own significant impacts on ozone. Further understanding of evaluating the potential effects of various natural events, such as volcanic eruptions and solar events, as well as other possible human activities, including nuclear explosions is also needed.

Further Reading

Fioletov VE, Bodeker GE, Miller AJ, McPeters RD, and Stolarski R (2002) Global and zonal total ozone variations estimated from ground-based and satellite measurements 1964–2000. *Journal of Geophysical Research* 107: 4647. http://dx.doi.org/10.1029/2001JD001350.

Fioletov VE (2008) Ozone climatology, trends, and substances that control ozone. *Atmosphere-Ocean* 46(1): 39–67. http://dx.doi.org/10.3137/ao.460103.
Heimann M (2001) The cycle of atmospheric molecular oxygen and its isotopes. In: Schultze ED, Heimann M, and Harrison S, et al. (eds.) *Global biogeochemical cycles in the climate system*, pp. 235–244. San Diego: Academic Press.
Holland HD and Turekian KK (eds.) (2005) *Biogeochemistry, Treatise on Geochemistry* 8. Oxford: Elsevier – Pergamon.
Jacobson MZ (2002) *Atmospheric pollution: History, science, and regulation*. New York: Cambridge University Press.
Keeling RF (1995) The atmospheric oxygen cycle: The oxygen isotopes of atmospheric CO and O and the O/N ratio (U.S. National Report to IUGG, 1991–1994). *Reviews of Geophysics* 33(supplement). http://dx.doi.org/10.1029/95RG00438.
Margulis L and Schwartz KV (1988) *Five kingdoms*, 2nd ed. New York: W. H. Freeman.
Schlesinger WH (1997) *Biogeochemistry: An analysis of global change*. San Diego: Academic Press.
Turco RP (2002) *Earth under siege: From air pollution to global change*, 2nd ed. New York: Oxford University Press.
Turekian KK (1996) *Global environmental change: Past, present, and future*. Upper Saddle River, NJ: Prentice Hall.
Volk T (1998) *Gaia's body: Toward a physiology of earth*. New York: Copernicus and Springer.
World Meteorological Organization, WMO. (2011). Scientific Assessment of Ozone Depletion: 2010, Technical Report Number 52, *World Meteorological Organization, Global Ozone Research and Monitoring Project*, Geneva, Switzerland.

Paleoclimatology

Marcus J Thomson, University of California, Los Angeles, California, United States and International Institute of Applied Systems Analysis (IIASA), Laxenburg, Austria

Nomenclature

BP Before present (1950 CE, by convention)
ka 10^3 years
Ma 10^6 years
Ga 10^9 years

Glossary

Holocene The present geological era, 11.5 ka BP–present.
Isotopes Atoms with the same number of protons and different numbers of neutrons.
Isotopologues Molecules of the same chemical structure comprised of isotopes and combinations of isotopes.
LGM Last glacial maximum (MIS2).
LIG Last interglacial period (MIS3).
MIS Marine isotope stage, counting major interglacial-glacial periods back from the present era, MIS1; where odd numbers stand for interglacials and even numbers stand for glacial periods.
Pleistocene The previous geological era, 2.55 Ma—11.5 ka BP.
Quaternary The present geological period, comprised of the Pleistocene and Holocene eras.

Paleoclimatology

Paleoclimatology is the study of Earth's climate from the formation of the planet to the present. It is distinguished from historical and "modern" climatology by its use of proxies, rather than historical documents and instrumental measurements, to reconstruct past climatic conditions. Proxies are naturally occurring archives of environmental change which are interpreted to infer climate histories. For those interested in past ecosystems, paleoclimatology is important to inform the contexts in which plant and animal communities evolved and lived; for those interested in present and future ecosystems in the context of climate change, paleoclimatology is the only means to establish a baseline against which to assess degrees of change.

Climate Change and Its Causes

Climate is a loose term for the state of the weather in an area over a period of time, usually months or years. Climate change is a result of dynamics of solar insolation, the geometry of the Earth's orbit, the geographical disposition of the continents and oceans, and atmospheric greenhouse gas (GHG) concentrations. Important second-order causes of climate change include dynamics of energy circulation in the ocean and atmosphere, volcanism, ice sheet scale and areal coverage, topography, and vegetative land cover. The Earth's mean annual diurnal variations in temperature, one measure of the planet's climate, for example, is moderated by the general circulation of the atmosphere-ocean system and greenhouse effect of the atmospheric gases.

The degree to which the climate may be said to change depends on its normative state, for which are several scientific conventions. For instance, in the chapter on paleoclimate in its fifth assessment report, the International Panel on Climate Change pays special attention to the past 2000 years (2 ka BP). On a planet whose climate record may be reconstructed or inferred from nearly 4.6 billion years (Ga) BP, choosing the most recent four ten-thousandths of a percent of that period may seem odd. It may even have engendered controversy among observers who are not aware of the vastly different scales over which proxies of paleoclimate change operate. Focusing on the past 2 ka permits paleoclimatologists to draw on data from multiple proxy sources which have good chronological and geographical resolution; and ignore parts of the planet's history which have no modern or probable near-term future analog.

Solar intensity and Milankovitch cycles

The flux of solar energy from the top of the atmosphere to the earth's surface (insolation) is the overwhelming determinant of the planet's environment. The sun's output varies over time as a consequence of complex patterns of solar cycles and, over its long lifetime, the availability of nuclear fuel. Solar cycles occur with fairly regular 11 year periods, although these short cycles are superposed on far longer and less predictable oscillations in output. They manifest as variations in the sun's radiative output of about 0.08%, coronal mass ejections, and visually as changes in sunspots and the auroras seen on Earth. Because of historical sunspot observations in particular, we may be confident that the sun has behaved similarly over the past several centuries. Using cosmogenic radiocarbon (^{14}C) and beryllium (^{10}Be) archived in tree growth rings and ice cores, patterns in solar intensity have been reconstructed for millennia. The sensitivity of the Earth's climate system to these small variations is not well understood.

 doi:10.1016/B978-0-12-409548-9.10894-2

Over timescales on the order of a thousand years and more, variations in the sun's radiative output are less important to the Earth's climate than the solar flux (energy per unit area per unit time) over the northern hemisphere, which depends primarily on orbital geometry. The lower the sun rises in the sky, the lower the flux at the surface. The seasonal position of the sun in the sky is determined by orbital mechanics. Over long timescales these are called Milankovitch cycles. Milankovitch cycles are the summation of several distinct oscillations due to the force of gravity: first is the axial-tilt (obliquity) of the planet's axis of rotation with respect to the plane of its orbit (the ecliptic); second is the precession of the Earth as it spins on its axis, due to net torques on its nonuniform distribution of mass; third is the (apsidal) precession of the Earth's slightly elliptical orbit due to the influence of mass asymmetries in the solar system; and fourth, in addition to those Milankovitch computed, are included variations in the inclination of the ecliptic. The combined effect of these oscillations are modes with beat periods of roughly 100, 41, and 23 ka. The remarkable (although imperfect) correspondence between Milankovitch cycles and glacial-interglacial intervals of over the quaternary period (2.44 Ma BP—present) is the strongest evidence for the dominant Milankovitch Theory of Quaternary glaciation.

Geographic disposition of the continents

The Earth's surface is subdivided into dynamic tectonic plates. Over millions of years, continental plates drifted apart, producing rifts and oceans, and ran together, lifting mountain ranges and forming supercontinents. The present conformation of the continents, with the land area in the northern twice that of the southern hemisphere, was a major factor contributing to Quaternary glaciation, as it enabled the accumulation of ice sheets on continental scale. During cold periods, multiyear snow accumulated as it does presently on Baffin Island, Greenland, and Antarctica. Whereas snow accumulation on Antarctica is restricted by the Southern Ocean, the vast tracts of the North American and Eurasian landmasses straddling the Arctic Circle allowed ice sheets to grow almost without constraint.

Further, the geographic disposition of land and sea controls the general circulation of the oceans and atmosphere. The thermohaline circulation (THC) of the oceans is driven by thermal and salinity gradients. A sudden disruption to the THC in the North Atlantic may have caused the Younger Dryas (12.9–11.7 ka BP), a sudden and short-lived rebound glacial-like conditions at the beginning of the present interglacial period. Measuring climate change in response to internal forcing mechanisms, like emergent disruptions in the THC, are challenging, requiring sophisticated and computationally expensive numerical models. Observing past changes in the THC by proxy is possible, to a limited degree, by comparing isotopic ratios of from disparate sediment cores.

Atmospheric greenhouse gas concentrations

The atmosphere helps to thermally regulate the surface of the planet. Because some of its constituent gases are transparent to incident solar radiation and tend to block lower frequency (thermal) energy that is re-radiated from the surface, it functions like a greenhouse: when there is a net surplus of energy into the Earth-system, it warms up. These constituents are termed greenhouse gases (GHGs). The present composition of dry air in the atmosphere is about 78.08% nitrogen (N_2), 20.95% oxygen (O_2), 0.93% argon (Ar), and 0.04% carbon dioxide (CO_2) by volume, with trace gases. The concentration of water vapor in the atmosphere varies, but is approximately 1%. Water vapor is a critical greenhouse gas (GHG), particularly for its role in cloud formation and due to its high heat capacity.

The atmosphere is structured by gradients in temperature and its constituent gases. The bottom two layers are the stratosphere and, below it, the troposphere. The boundary between troposphere and stratosphere is defined by the tropopause, usually marked by a temperature inversion. This is the altitude at which thermal radiation from the earth is balanced by radiation from the sun. Just above the tropopause is the layer of ozone (O_3) gas which is a critical barrier against high energy cosmic rays which would significantly disrupt the biosphere. There are still higher-elevation structures that shield against particulate radiation harmful to organic life.

The role of CO_2 as a GHG is well understood. The present concentration of CO_2 in the atmosphere is about 410 ppm. This concentration depends on the Carbon Cycle, a biogeochemical exchange between the geosphere, pedosphere, hydrosphere, biosphere, atmosphere, cryosphere, and increasingly the anthropogenic built environment. Archives of CO_2 concentration in the atmosphere (hereafter $[CO_2]_{atm}$) demonstrate that this is an increase of at least 33% from 1800 CE due to the industrial-scale conversion of fossil fuels to energy and waste gases. Over the last 400 ka, $[CO_2]_{atm}$ had a concentration of 180–210 ppm during glacial periods, when ice sheets covered substantial parts of the northern hemisphere, and 280–300 ppm during warm interglacial periods. Finding periods in the atmosphere's history wherein $[CO_2]_{atm}$ was equal to or greater than 400 ppm, and comparing the consequences on climate records, is a major preoccupation of many paleoclimatologists today. For instance, about 125 ka BP, during the Last Interglacial (LIG; 129–118 ka BP), was a period sometimes called the Eemian: this was the warmest period of the past 200 ka, with a climate similar to that of today, and we have no evidence that $[CO_2]_{atm}$ was greater than about 300 ppm. However, there is a growing consensus that regional differences in warming, particularly over the Arctic, were substantially different from the modern (and probable future) case as compared to the Eemian.

Over geological history, which spans about 23,000 times 200 ka, concentrations of constituent gases in the atmosphere have varied so radically that the planet is considered to have experienced at least three entirely different atmospheres. The present composition of the atmosphere was established roughly 3.5 Ga BP, although the relative concentrations of its constituent gases have varied since, primarily as a consequence of biochemical feedbacks within the Earth system.

Paleoclimate Reconstruction Methods

Landforms were long ago recognized as probable archives of climate change. Scientists compared processes of mountain glaciology to features of the landscape of Europe and North America, and saw relicts of enormous ice sheets which predated historical records. Yet in spite of the evidence of environmental change by proxy, there were no reliable methods to determine the age of prehistoric phenomena. The development of radiometric methods after the middle 20th century was revolutionary, as now different proxies could be independently dated and ordered in time to produce a synoptic record of paleoenvironmental change.

Proxies

Proxies are naturally occurring phenomena whose intrinsic alteration, such as growth or accumulation, is sensitive to environmental conditions and whose impressed remains are stable archives through time. Paleoclimatologists interpret patterns in proxy records to reconstruct the environmental conditions under which they were produced. As various proxies respond differently to the same climatic mechanism, paleoclimatologists seek multiple proxies to develop complete synoptic records of climate change. Multiproxy records show that the Earth's mean annual temperature, for example, has followed a linear cooling trend over the past 800 thousand years (800 ka), but also displayed quasi-periodic oscillations from cold glacial to warm interglacial conditions over thousands of years, as well as the seasonal and diurnal ups and downs of everyday experience.

Stable isotopes

When plants respire and process CO_2, for example, plants that follow the C_3 photosynthetic pathway take up less of the heavier isotope than plants which follow the C_4 pathway. This fractionation is written in terms of a standard measure, $\delta^{13}C$, and means that C_3 plants are relatively depleted in ^{13}C ($\delta^{13}C$ of between -33 and $-24‰$) as compared to C_4 plants ($\delta^{13}C$ of between -16 and $-10‰$). In lake sediment cores, $\delta^{13}C$ and $\delta^{15}N$ are used to discriminate between endogenous and exogenous sources of organic carbon.

All water (H_2O) in nature contains the most common isotope of oxygen ($^{16}_{8}O$) and a trace fraction of $^{18}_{8}O$, a stable isotope of oxygen with two additional neutrons. During glaciations, the accumulation of continental scale ice sheets sequestered vast amounts of precipitated water on land. Evaporation thermodynamically favors $H_2^{16}O$, due to its lower mass, so meteoric water tends to be depleted in $H_2^{18}O$; that is, during periods of glaciation, the seawater from which meteoric water was ultimately sourced was enriched in $H_2^{18}O$. This is measured in terms of $\delta^{18}O$, an isotopic fraction of heavy oxygen normalized by a standard value.

Strontium is abundant in nature, readily substitutes for Ca^{2+} in crystals of $CaCO_3$ wherein resists diagenetic alteration. Ratios of strontium isotopes ($^{87}Sr/^{86}Sr$) can be uses as geological fingerprints to determine the provenience of sediment. This has been applied to reconstruct the Quaternary paleohydroclimate of the Nile River, whose main tributaries, the Blue Nile and the White Nile are driven by different climatic mechanisms (Talbot *et al.*, 2000). The relative strength of these mechanisms is important to questions of Holocene climate change over Africa and the Indian Ocean.

One of the most exciting methodological developments in paleoclimatology is the application of 'clumped' isotopes to temperature reconstruction (Eiler, 2011). For example, carbonate (CO_3^{2-}) crystals appear in nature as molecular combinations (isotopologues) of stable isotopes of carbon (^{13}C, ^{12}C) and oxygen (^{16}O, ^{17}O, ^{18}O). As a crystal lattice forms, isotopologues will equilibrate according to their slight mass differences, with higher mass isotopologues thermodynamically favored at lower temperatures and vice versa. By comparing isotopologue abundances in CO_2 from carbonate crystals (written as Δ_{47}) to that which would be expected from their natural abundances alone (i.e., no "clumping"), researchers estimate isotopologues equilibrium constants and thus create a paleothermometer. The great strength of this method is that it depends exclusively on thermodynamics; the weakness, for the moment, is that the small temperature dependence of "clumping" means that carbonate samples must be uncontaminated and large.

Proxies for the quaternary period

Ice cores

Ice records changes in polar and high elevation glaciers from entrained dust, pollen, and charcoal, or from ancient gases trapped in interstitial pockets. Ice sheets form from snow that continues to accumulate over many years. Sublimation and melting causes snow to recrystallize as "firn," a granular manifestation of snow-ice that forms the uppermost laminae of an ice core. Firn sequesters air bubbles that sample the atmosphere at the time of its formation. With depth, the considerable weight of multiyear firn reorganizes water-ice crystals to produce solid ice below about 10 m; and ice sheets on Greenland and Antarctica are kilometers thick.

Speleothems

Common salts of carbonate (CO_3^{2-}) dissolve readily in water. A consequence is that sedimentary carbonate formations of limestone, dolomite, gypsum and like underlying water-saturated soils are characterized by subterranean cave systems; and within the caves, carbonate salts precipitate onto the various surfaces of the void spaces, producing cave deposits, termed speleothems. The most common minerals comprising speleothems of interest to paleoclimatologists are calcareous carbonates ($CaCO_3$), such as calcite and aragonite.

The dissolution of atmospheric carbon dioxide into water forms weakly acidic carbonic acid (H_2CO_3). In the archetypical case, this percolates through the subsoil and weathered limestone substrate, dissolving and mobilizing calcium and bicarbonate, via the

reaction, $CaCO_3 + H_2CO_3 \rightarrow Ca^{2+} + 2HCO_3^-$. When this solution percolates into the cave system, it loses dissolved carbon dioxide to the low partial pressure of CO_2 in the void space, thereby precipitating calcium carbonate via the reaction, $Ca^{2+} + 2HCO^{3-} \rightarrow CaCO_3 + H_2O + CO_2$. A proportion of the molecular carbon and oxygen archived in speleothem carbonates was sourced from the atmosphere and the hydrosphere, respectively. Thus speleothem carbonates may be analyzed for isotopic signatures (e.g., $\delta^{13}C$, $\delta^{18}O$) of the sources of these environmental inputs. Although usually analyzed for patterns of change in the hydrosphere, in some circumstances be dominated by changes in temperature.

Marine and lake sediment cores

A core is a column of sediment that preserves the stratigraphy of its parent material. From top to bottom, they archive sediments from youngest to oldest. The length of a core, compared to the range of time in archives, is determined by the sedimentation rate, the speed at which substrate accumulates. In deep ocean cores, this may be just millimeters per thousands of years, while on a continental shelf, where it is augmented by alluvial matter discharged from rivers, the sedimentation rate may be much faster. Higher sedimentation rates typically produce higher resolution records. Terrestrial wetlands and lakes may accumulate sediment very quickly, particularly if they are susceptible to inwash from flooding, high biotic productivity, and accumulation of peat.

Foraminifera (see below) records from cores of the ocean floor were the first truly global-scale paleoclimatic proxies; and therefore, many other proxies are compared to the sequence of $\delta^{18}O$ lows and highs, numbered down-core, or from youngest to oldest, starting with the present, Holocene Epoch ($\sim$11.5 ka BP to present). Each oddly numbered marine isotope stage (MIS) represents a warm interglacial (or interstadial), such as the Holocene (MIS1); each evenly numbered MIS represents a cold glacial (stadial), such as the LGM (MIS2). Numbering periods according to MIS avoids overtly phenomenological descriptions of climate change, whose effects vary by location. We now understand that large ice sheets persisted over parts of the Northern Hemisphere throughout MIS3, and that the last period which may be practicably termed 'interglacial' was MIS5 (130–80 ka BP). Careful investigation of the MIS sequence and comparisons to other multiproxy records revealed substages, termed "horizons," within the stages themselves. For instance, MIS5 began with the warmest period of the Pleistocene, the Eemian (130–115 ka BP), which is alphanumerically labeled MIS5e to distinguish it from markedly different late-Stage 5 conditions.

Lake and other terrestrial wetland environments readily sequester material from their surroundings, incorporating it into their sediment. Over long periods of time, lakes can significantly change in depth-profile (bathymetry) and volume, based on their water-balance, or net water inflow. Over the short term, on the order of days to decades, water-balance is affected by changes in precipitation, streamflow, and evaporation. Over centuries and more, water-balance can be affected by geomorphic changes to the lake's catchment, chemical changes to the soil profile and drainage, and sediment buildup. When a lake sediments in and dries out, it is called a "playa." The famous Bonneville Salt Flats playa, for example, was formed when the Pleistocene Lake Bonneville dried out ($\sim$14.5 ka BP) and left behind a thick evaporite of salts. The Fayum, in northern Egypt, is an ancient lake basin in what has been, for the past several thousand years, a hyperarid desert. Its modern lake, Birket Qarun, is a fraction the size of its premodern predecessor, Lake Moeris, mostly due to flood control of the Nile River. However, even ancient Lake Moeris was a fraction the scale of the Pleistocene-era Fayum paleolake, whose shorelines are visible in the desert tens of meters above the Qarun shoreline.

Most lakes more than a few meters in depth naturally stratify by temperature due to small differences in density. At the top of the water column is the epilimnion, a stratum of warm, well-mixed water. At the base of the water column is the hypolimnion, a stratum of cold, hypoxic or anoxic water. Between these strata is the metalimnion, sometimes loosely referred to as the thermocline, which is the sudden transition with depth from warm to cold water. Holomictic lakes mix at least once a year; a subset of these, dimictic lakes, mix in both the spring and the fall. In a stratified lake there is no replenishment of dissolved oxygen to the base of the water column; so it is used up quickly by microorganisms, and hypoxic or anoxic conditions develop. As the lake absorbs thermal energy over the course of the summer, the epilimnion will expand until the water column is isothermic due to mixing. The water-sediment interface will oxygenate and the lake is said to have "turned over." In dimictic lakes, turnover will occur again in the spring: when water nears its freezing point, water molecules tend to reorder into complexes which reduce its mass density, making it more buoyant. After the spring melt (ice-out), warming surface water passes through its maximum density at 4°C (at 1 atm of pressure), at which point it sinks, displacing the deeper water below. Over many seasons of regular mixing, "varves," which are laminae in the sediment, develop. Paleoclimatologists seek to core varved lake sediments because varves may often be counted as tree rings are, improving the quality of an age model. Varved sediments from Lake Van in eastern Turkey, for instance, have been used to reconstruct millennia of paleoenvironmental change in the Near East. Lake sediments have proved some of the most geographically distributed and informative sources of paleoclimatic data for the northern hemisphere over the late quaternary period.

Pollen

Palynology is the study of microscopic pollen and spores, collectively termed palynomorphs. Pollen grains and spores have a hard outer coating (exine) made largely of sporopollenin, an inert polymer which makes them highly resilient to diagenesis. Pollen assemblages are used to reconstruct the nearby vegetative environment contemporaneous with their position in a sediment core. Fossil pollen samples are typically prepared by chemically reducing and sieving unwanted sediment substrate, chemically darkening and dyeing grain exines to highlight diagnostic features, and mounting on optical microscope slides. Pollen grains are counted relative to a tracer spore prepared within the sample. Identification to better than genus level is not usually possible; and some grain morphotypes are difficult to identify at better than family level, such as nondomesticated grasses. Fossil pollen-based

paleobotanical reconstructions are approximations limited by identification resolution and bias towards copious, anemophilous (wind-dependent pollinating) plants.

Tree rings

Woody plants grow their trunks by means of a live tissue just beneath the bark called "cambium." For most trees, the cambium is bifacial: it sets down parallel undifferentiated cell pairs, the outer layer of which is the "phloem," and the inner layer of which becomes the "xylem." The phloem transports water and nutrients critical for the tree's growth; the xylem hardens into wood, a material of dense cellulosic fibers. Newer wood layers grow over old layers, forming annuli. Trees that grow seasonally, either because of warm-cold or wet-dry transitions in the climate, suffer periods of growth-stress or dormancy. In temperate climates, where growth stress is a function of the seasonal cycle, common tree rings are produced because wood laid down in the spring and summer ("early wood") is structurally different from that which is laid down in the late fall ("late wood"). A single tree therefore captures snapshots of its local environmental stress history; a copse of trees captures a redundant record of the paleoenvironment for the area. By matching patterns of growth rings between living trees and deadfall, it is also possible to build environmental histories that reach well beyond the lifetime of a tree.

Trees consume CO_2 from the atmosphere via the stomata in their leaves and water (usually via their roots) during photosynthesis. Their tissue is therefore composed of carbon and oxygen atoms in part from the atmosphere and protons and oxygen atoms donated from water. Meteoric and groundwater tend to have different isotopic signatures. Therefore, trees pick up environmentally determined isotopic signals in terms of $\delta^{13}C$, $\delta^{18}O$ and deuterium (δD, used to infer water-balance) from these inputs alone. Tree rings have been used to reconstruct records of prehistoric hurricanes over the southeastern United States, rainfall over isolated parts of Amazonia, and the timing of important modes of internal variability in the Pacific Ocean, such as the El Niño Southern Oscillation (ENSO).

Other biogenic proxies

Many organisms archive records of environmental change over time in the hard tissues (sclera) they grow. Common applications are the cross-sectional analysis of mollusk shells, and piscine otoliths for changes in stable isotope ratios. For instance, corals are important foundational members of many marine ecosystems. Corals are complexes built from the calcareous secretions of coral polyps. Assemblages of coral species are specific to local environmental conditions. Complexes of coral skeletons may be analyzed for changes in marine conditions and even variations in mean sea level.

Midges (Insecta: Diptera: Chironomidae) are flying insects that lay their eggs in water. In fact, the majority of a midge's existence is spent as an aquatic instar; its adult stage is short, merely to translocate and reproduce. Midges are good indicator species, as they reproduce only within a narrow niche defined by environmental factors like water temperature and various solute concentrations. When midge instars undergo ecdysis, their exoskeletons (cuticles) sink and collect in sediment. The cuticles of late-stage instars are especially robust, and may remain preserved in sediment for millennia. Diagnostic features of fossil cuticles are used to identify midge taxa recovered from sediment cores. Assemblages have been used to reconstruct ecological changes to wetland systems, including local air temperature.

Diatoms (Bacillariophyta) are a major class of phytoplankton which have become especially influential in terrestrial wetland paleolimnology, particularly in determining water balance and solute concentration. Diatoms are unicellular autotrophic organisms whose cell walls (frustules) are composed of silica. Frustules of expired diatoms readily collect in sediment and are highly resilient. Fossilized diatoms are found after about 185 Ma BP; and the earliest diatoms may, in fact, date to the permian mass extinction (250 Ma BP).

Otoliths are calcium carbonate masses that grow within the inner ears of vertebrates. Fish otoliths found in sediment cores are analyzed for isotopic differences, such as $\delta^{18}O$ variations down-core to determine the history of evaporative enrichment in a lake. Such data may be inform an estimate of a lake basin's water balance in time series or drought history. In a classic study of paleoclimate change in the Near East, $\delta^{13}C$ in the shells of land snails showed a change in their diet, from C_3 to C_4 pathway plants, indicating a shift in the dominant precipitation regime from relatively wet mid-Holocene to dry modern conditions (Goodfriend, 1988).

Neotoma

Pack rats (*Neotoma* spp.) are a genus of small rodents found throughout northern Mexico, the southern and western United States, and western Canada. The nests, or middens, that they build are highly resilient and may be occupied, and reoccupied, by generations of neotoma over thousands of years. In constructing their middens, neotoma hoard nearby materials and cement them with secretions of their urine. Neotoma urine readily desiccates in their dry habitats becoming amberat, the substance that cements their middens. Additionally, neotoma will seal off midden chambers occupied by previous generations, producing nearly hermetic time capsules of highly localized assemblages of plant macrofossils, shells, bones, insect exoskeletons, and pollen. Neotoma middens containing remarkably well preserved paleoecological material have been radiocarbon dated to the Pleistocene.

Geomorphology

The Earth's surface is constantly under the influence of biotic and abiotic agents of change, whose activity is largely driven by solar radiation and the force of gravity over time. While the lithosphere is a good archive of hundreds of millions of years of this influence, is it commonly encountered in quaternary-age reconstructions. Throughout northern Europe, northeastern North

America, and in mountain valleys all over the world, there is ample evidence of glaciation in the form of moraines, enormous nonvolcanic boulders scattered haphazardly on the landscape (glacial "erratics"), striations in bedrock due to abrasion, and deep reservoirs of rock flour in ancient basins. Isostatic rebound, which is the slow adjustment of continental plates to having the great weight of ice sheets removed, is raising Scotland and making new dry land in northern Canada and the Baltic in spite of sea level rise due to global warming.

Physical histories in basins with no outflow, such as endorheic lakes, can be useful in reconstructing local water balance over many thousands of years. Streams modify the landscape by downcutting into the sediment (degradation) and by building up alluvial deposits (aggradation). The difference between a stream's tendency to downcut or deposit may be due to discharge, the amount of material it carries in suspension, or whether it freezes. Stream terraces form from repeated and alternating episodes of channel degradation and aggradation. Terrace gravels are generally dated radiometrically to unravel the channelization sequence. These have clear dependencies on prevailing climate, as well as dramatic weather related events like floods; and it is a useful technique for Holocene-age xeric environments. Analysis of paleochannels and paleolake shorelines has been particularly successful in the development of evidence for an expansive mid-Holocene subpluvial ("wet" period), sometimes called the "Green Sahara," some 9–5 ka BP across northern Africa; and in reconstructing hydrologic water balance in the Pleistocene Southwest United States. Both results have significant implications for large scale ocean circulation patterns which drive prevailing climates in those regions.

Loess is eolian, or wind-deposited, dust that often accretes in periods of large scale drought. Eolian dust forms when, driven by wind, loose grains skip along the ground with sufficient speed to cause microscale fragments to eject from the grains, a process known as saltation bombardment. In spite of their poor aerodynamics, the fragments become suspended in turbulent air flows and can rise kilometers into the troposphere. Airborne dust can cause moisture to condense and precipitate, and block sunlight. Loess deposits are often interpreted as evidence of drought; but also more subtly as evidence of a breakup of the biological soil crusts in dryland ecosystems. Patterns of cross-bedding in loess and sand dune deposits even carries information on the strength and direction of the prevailing wind.

Proxies for prequaternary ages

Foraminifera and coccoliths

Foraminifera (or commonly just "forams") are protists that inhabit nearly every marine environment, including estuaries. Forams are best known among paleoclimatologists for their utility in reconstructing past sea-surface temperature (SST) variations, because they incorporate the environmental isotopes of their local environments when they grow their shells, termed "tests." Tests from expired forams accumulate in marine sediment. Morphological variations in tests enable researchers to discriminate between different taxa in a fossil foram assemblage from a core. Thus, researchers can infer the former presence of microenvironments in which similar assemblages are found presently, or select the tests of specific taxa for isotopic analyses. An example of the former case is in the reconstruction of mean sea level (MSL), as foram assemblages of from the intertidal and subtidal zones from coastal sediments are very specific; and an example of the latter is SST reconstructions from pelagic forams, or those that live in the open ocean. In both cases, fossil assemblages must be converted to MSL or SST estimates by means of transfer functions, which are usually regressions of instrumental data on modern foram assemblages.

Another important fossil resource found in marine sediments are "coccoliths," calcium carbonate scales produced within an ancient group of phytoplanktons, coccolithophores. As phytoplankton, they are obliged to live at the top of the water column, where there is sufficient light for photosynthesis, in the euphotic zone. The euphotic is the brightest part of the epipelagic zone, or depth over which any light penetrates at all. As such, it can be radiatively warmed and remains relatively well mixed; that is to say, in good contact with the atmosphere and the isotopic ratios of its constituent gases.

Age modeling

Age models are built in a variety of ways. It is often impossible or impractical to determine the age for every element in a series of proxies, such as a sediment core. In such a case, material from a subset of the core sediment is sampled for radiometric dating. The result will be an ordered sequence of radiometric ages which, following calibration to some standard, are interpolated to produce a sediment age-depth model. Similarly, age models are generated for speleothems using uranium-series or uranium'thorium dating, for dunes and sediment deposits using luminescence techniques, and for glacial erratics and moraines using cosmogenic nuclides.

Prior to the advent of radiometric methods after the middle 20th century, ages of phenomena were estimated by their relative positions within geological sequences or by comparison with documentary historical evidence, such as Egyptian papyri and ships' logs. Geological phenomena, such as the uplift and degradation of mountains, downcutting of river channels, and sedimentation and lithification of stone, were clearly older by orders of magnitude. Comparative anatomists could build up a sequence of similar life forms from their fossils in different geological facies, and thus deduce which were likely more or less recent in time. These were methods of relative age determination. Methods of direct age determination require independent clocks, such as radioactive decay, and detectors of sufficient sensitivity that only became practicable after the middle 20th century.

Dendrochronology and sclerochronology

Tree rings were long known to represent the conditions under which trees grew, but the first systematic use of tree rings for age determination (dendrochronology) was not developed until the early 20th century by astronomer A. E. Douglass, who realized that growth-stressed trees were good independent clocks. Trees grow by laying down new cellulosic matrix on their cambium layer,

just beneath their bark. In extratropical environments, there are seasonally expressed differences in environmental stress, and trees cycle through periods of dormancy and growth. Similarly, the use of time-dependent information archived in biotic sclera is termed sclerochronology.

Radiometric dating

Radiometric methods rely on radioactive beta-decay, a fundamental process in nature, during which a neutron spontaneously transforms into a proton and an electron, and an energetic but weakly interacting neutrino. Individual neutron decays are unpredictable, but the average rate of decay of a large population of neutrons follows a simple exponential decay, the rate of which depends on the mass of the particle into which a neutron is bound. The radioactive decay rate of any isotope may therefore be represented by a single number, its characteristic "half-life," $t_{\frac{1}{2}}$, the average time it would take for a population of radioactive particles to decline by one-half, or its inverse, the decay constant, λ.

Radiocarbon (^{14}C), with a half-life of about 5730 years, is ideally suited to radiometric age determination because it is abundant in nature, chemically indistinguishable from common ^{12}C. Carbon-14 is produced in the upper atmosphere by collisions between nitrogen (^{14}N) and cosmic rays, mixes to sea level in about 6 months, and is promptly insinuated into the carbon cycle of the biosphere. The ratio of $^{14}C/^{12}C$ remains in equilibrium with the environment in an organism until death, whereupon no new ^{14}C is contributed and the $^{14}C/^{12}C$ ratio is determined by the rate of decay of ^{14}C (into ^{14}N). Careful measurement of this $^{14}C/^{12}C$ ratio enables age determination with a precision of as little as a few percent over the half-life of ^{14}C and with a practicable maximum age of about seven half-lives using a state-of-the-art, accelerated mass spectrometer (AMS) measurement. The $^{14}C/^{12}C$ ratio is given in terms of ^{14}C-years, which are nonuniform units of time, whose zero-point is defined as 1950 CE. Converting ^{14}C-years to a meaningful age determination requires correcting for nonequilibrium ancient carbon (reservoir effect), differences in physiological and taphonomic processing of carbon (fractionalization), and postdepositional changes (diagenesis) to the organic crystal; and most important, calibrating for the variable rate of production of atmospheric ^{14}C. In spite of its long history of use, radiocarbon research remains an active area of research, with considerable resources dedicated to improving calibration curves.

Many other radionuclides are important for age determination. For example, an isotope of chlorine (^{36}Cl) is produced by cosmogenic and terrigenous nuclear reactions in the atmosphere in immediately beneath the Earth's surface, respectively, is used for dating samples of middle to late-Pleistocene age. Uranium-series dating is used extensively for corals and speleothems. The uranium-series method measures the proportion of ^{234}U to its radioactive daughter product, ^{230}Th; and because the half-life of ^{234}U is much greater than that of ^{230}Th, over time, the rate of production of ^{230}Th will equal its rate of decay. At this point, ^{234}U and ^{230}Th are in secular equilibrium. Since uranium is water-soluble and thorium is not, thorium atoms will accumulate in precipitated carbonate (modulated by their decay rate) while the concentration of uranium remains constant. The method can be used to determine carbonate ages to some 500 ka BP.

Secondary effects of exposure to radiation on crystals includes two classes of very successful methods: fission-track and luminescence dating. Fissile atoms eject energetic charged particles which damage or disrupt crystal lattices. In fission-track dating, the cumulative damage as a consequence to exposure to ^{238}U provides an estimate of the duration of the crystal's exposure to the source. In luminescence dating, the method takes advantage of naturally occurring anomalies, termed traps, in the lattice that tend to accumulate small amounts of charge over time, partly due to exposure to low levels of background radiation (see Rhodes, 2011). Exciting the crystals with light or by heating causes the traps to discharge, which is monitored in the lab. By estimating the time required to accumulate the excess of electrons in crystal traps, based on a gross estimate of ambient radioactivity, the dose rate, researchers make a clock which can determine the duration a crystal has been buried (not exposed to light or excessive heat).

Tephrochronology

Tephrochronology uses layers of volcanic ash (tephra) to confine age limits of sedimentary facies. Where a tephra is often traceable to a specific volcanic eruption of known age, a sediment may be confidently dated. Bracketing is the term used to describe the method of bookending facies of interest between two or more tephras, thus finding upper and lower age limits.

Synopsis

The study of paleoclimates is, by the nature of the problems involved, methodologically flexible and interdisciplinary. Methods are unified by the use of proxies, a category of phenomena whose only shared characteristic is their response to environmental change. The range of timescales requires synthesizing archives whose comparison is not straightforward, or whose individual responses to the same environmental stimulus changes in time and by context.

Increasingly sophisticated computer models, including fully coupled Earth system models, are tasked with reproducing patterns of climate change at 2° of resolution or better for the last millennium. The architecture of these models, the community of experts they engender, and the software tools that they produce are dramatically improving the general understanding of climate and the complex dynamics of the Earth system. Modelers can incorporate hundreds of influences, over dozens of scenarios, to estimate climatic variables, and thereby test hypotheses of the underlying mechanisms of climate change, patterns of teleconnections (covariances between different elements of the climate system), and the effects of combined changes on the Earth system through time. Although computers have been applied to problems of climate since the middle 20th century, the growth rate of affordable

computing power and data science analytics is another revolution in the offing. Yet collecting, analyzing, and producing new and better paleoclimate proxy data remains essential.

Every discipline cherishes its own conventions. Paleoclimatology is therefore complicated by procedural questions such as which is or is not the 'best' practice or proxy. Normalizing the practice of science, in general, remains a challenge in many disciplines, and therefore a constant tension in a diverse one like paleoclimatology. Nevertheless, paleoclimatologists have demonstrated sincere commitment to publicly databasing their datasets, including raw measurements. This is a critical piece of the evolution towards "big data" analytics, which will almost certainly become an essential element of paleoclimatology, on par with radiometric dating, in the near future.

References

Eiler, J.M., 2011. Paleoclimate reconstruction using carbonate clumped isotope thermometry. Quaternary Science Reviews 30 (25), 3575–3588.
Goodfriend, G.A., 1988. Mid-Holocene rainfall in the Negev Desert from ^{13}C of land snail shell organic matter. Nature 333 (6175), 757–760.
Rhodes, E.J., 2011. Optically stimulated luminescence dating of sediments over the past 200,000 years. Annual Review of Earth and Planetary Sciences 39, 461–488.
Talbot, M.R., Williams, M.A.J., Adamson, D.A., 2000. Strontium isotope evidence for late Pleistocene reestablishment of an integrated Nile drainage network. Geology 28 (4), 343–346.

Further Reading

Juggins, S., Birks, H.J.B., 2012. Quantitative environmental reconstructions from biological data. In: Tracking environmental change using lake sediments. Netherlands: Springer, pp. 431–494.
Masson-Delmotte, V., Schulz, M., Abe-Ouchi, A., Beer, J., Ganopolski, A., González Rouco, J.F., Jansen, E., Lambeck, K., Luterbacher, J., Naish, T., Osborn, T., Otto-Bliesner, B., Quinn, T., Ramesh, R., Rojas, M., Shao, X., Timmermann, A., 2013. Information from paleoclimate archives. In: Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgle, P.M. (Eds.), Climate change 2013: The physical science basis, Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change. Cambridge, United Kingdom and New York: Cambridge University Press.
Shellito, C., 2016.November 15. InTeGrate Paleoclimatology https://serc.carleton.edu/integrate/workshops/methods2012/courses/shellito.html Accessed 30 September 2017.

Relevant Websites

http://www.carbonateresearch.com/home—Carbonate Research ICL. Retrieved 27 September 2017.
https://www.neotomadb.org/—Neotoma Paleoclimatology database. Retrieved 20 September 2017.
https://www.ncdc.noaa.gov/data-access/paleoclimatology-data—NOAA Paleoclimatology database. Retrieved 20 September 2017.

Pedosphere☆

Victor O Targulian, Russian Academy of Sciences, Moscow, Russia
Richard W Arnold, USDA Natural Resources Conservation Service, Washington, DC, United States
Bradley A Miller, Iowa State University, Ames, IA, United States
Eric C Brevik, Dickinson State University, Dickinson, ND, United States

Glossary

Anthroposphere The part of the environment made or modified by humans.
Atmosphere The gases surrounding Earth.
Biosphere The living organisms occupying Earth.
Geomorphic Relating to the Earth's surface features.
Horizons As relates to soils, a layer of uniform properties that forms parallel to Earth's surface.
Hydrosphere All waters found at Earth's surface and the interactions between them.
Lithosphere The solid outer part of the Earth, composed of minerals and rocks.
Pedogenic Processes related to the soil or soil formation.

Concepts

The pedosphere is the soil mantle of the Earth. This concept evolved from the basic scientific concept of soils as specific bodies in nature that developed in time and space in situ at the land surface due to processes resulting from interactions of soil-forming factors. These factors are parent material, climate, organisms, topography, and time. In this way, the pedosphere represents the intersection of the lithosphere, atmosphere, hydrosphere, and biosphere (Fig. 1).

This basic concept of soils as accepted by modern scientists was described by V.V. Dokuchaev in the 19th century and has generally been accepted worldwide. Humans as components of the biosphere have increasingly become a significant factor interacting with the other spheres; consequently, the anthroposphere (realm of human society) is now considered to be a major influence. Being at the intersection of all the other spheres, fully understanding the pedosphere requires a systems approach that considers the large number of processes that interact in unique combinations around the world. All of these interactions make soil complex, open, bio-abiotic, nonlinear, multifunctional, multiphased, and spatially diverse.

Soils cover much of the Earth's land surface and the bottom of stable waterbodies as part of a continuum or mantle. This continuum, called the pedosphere (from Greek *pedon* meaning ground), serves as the Earth's biogeomembrane, which is somewhat analogous to the biomembranes of living organisms. As a biogeomembrane, the pedosphere facilitates and regulates the exchange of substances and fluxes of energy among the land biota, atmosphere, hydrosphere, and lithosphere. Additions, removals, translocations, and transformations occur in the pedosphere depending on the interplay of local environmental conditions and the inherent properties within the soil bodies (Fig. 2). Paleosols represent earlier periods of soil formation, and it is expected that in the Mesozoic and Paleozoic eras some extinct types of soils and pedogenic processes could be found. Emphasis has more commonly been given to the major climatic and geomorphic effects on the pedosphere that existed during the Pleistocene and Holocene epochs. Currently, soil properties and functions as influenced by extensive exploitation of soil by humans during the last two centuries are receiving more attention.

☆*Change History*: March 2018. E Brevik and B Miller updated all sections of the entry, the further reading, and Figs. 2 and 5. Fig. 1 was added.

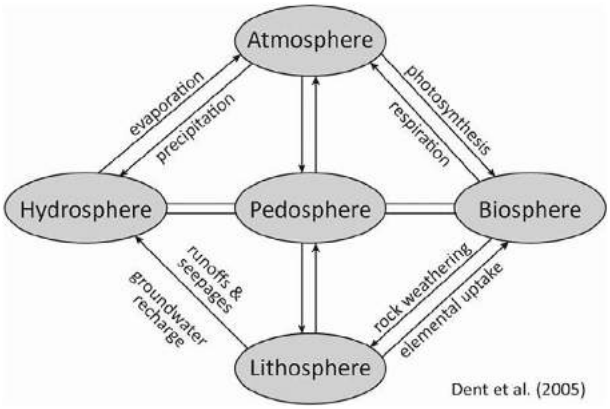


Fig. 1 The pedosphere as the intersection of the lithosphere, atmosphere, hydrosphere, and biosphere. Modified from Schaetzl, R. and Thompson, M.L. (2015). Soils: Genesis and Geomorphology, 2nd edn. New York: Cambridge University Press.

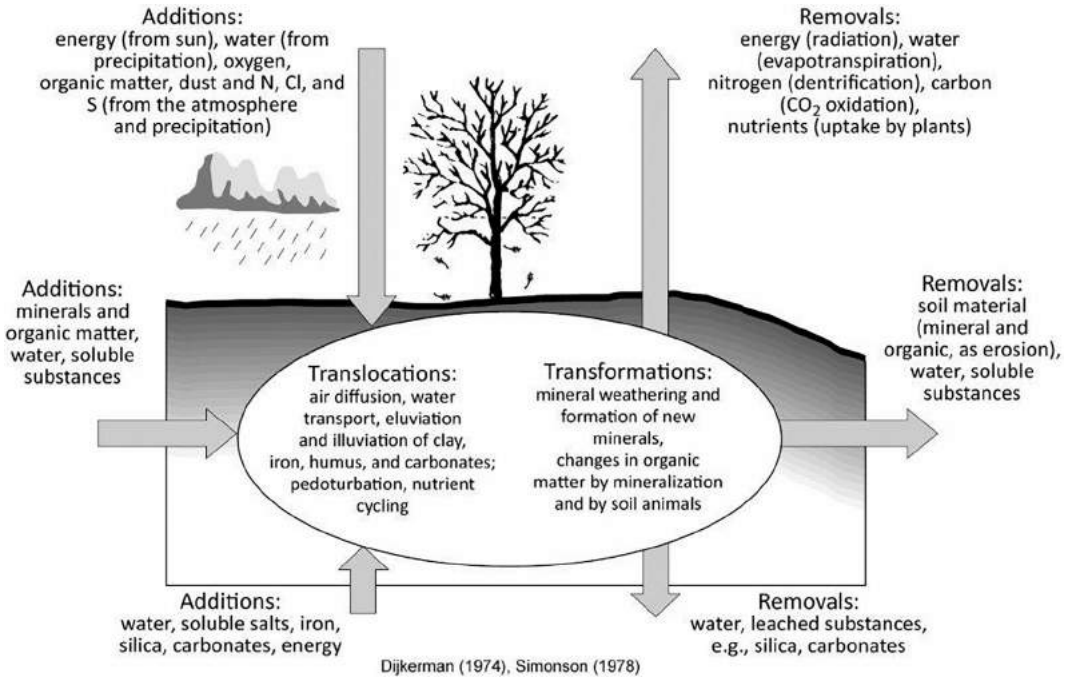


Fig. 2 Generalized processes active in developing soil features and horizons. Modified from Schaetzl, R. and Thompson, M.L. (2015). Soils: Genesis and Geomorphology, 2nd edn. New York: Cambridge University Press.

Processes

Most processes of pedosphere functioning operate in an open system, and although some appear to be cyclic and reversible, for example, biogeochemical cycling of C and N, many of them are unidirectional and irreversible, such as weathering of silicates in a soil and leaching of substances out of a soil. Due to the open and irreversible nature of some processes, there are many residual products, especially solid-phase materials, both organic and mineral, that are produced and retained in the parent materials. The annual formation of such components is very small and hardly detectable; however, when the soil forming processes occur for a long time (10^2–10^6 years), the gradual long-term accumulation of pedogenic solid compounds alters parent materials into soil horizons within soil profiles. Such processes of solid-phase, macrofeature formation during long-term functioning of a soil system can be perceived as a synergetic self-organization of the system—pedogenesis. Pedogenic features making up the solid-phase structure and composition of portions of the pedosphere are more pronounced where the upper unconsolidated layer of the lithosphere has neither been renewed by erosion or sedimentation nor mixed with deeper layers. Where landscapes have been stable and have had long-term functioning of soil-forming processes, gradual accumulation of pedogenic products occurs and well-differentiated soils form. The general development of the pedosphere is conceptually a sequence. There is an accumulation of earthy

materials that over time are altered by processes of interaction with the atmosphere, hydrosphere, and biosphere. A general rule of pedogenesis is: interacting factors open system processes formation of pedogenic properties and features (Fig. 3). Eventually the pedosphere forms into a three-dimensional structure that covers most of the lithosphere's interface with the other spheres.

Structure

The pedosphere has its own specific structure. Vertical variability is the result of processes altering parent materials in situ into pedogenic features and properties that make up horizons and soil profiles; see Fig. 4. These processes are usually called soil-forming, or pedogenic, processes. Many variations are possible due to the wide range of environmental conditions and scope of the factors interacting to form and develop soils. Translocation and transformation processes form a natural sequence of soil horizons that are indicative of the soil's genesis.

Soils, as multiphase bodies in the pedosphere, have several kinds of depth distributions at any moment. There are profiles of temperature, moisture, gases, soil solution and nutrients, macro- and microbiota, and solid-phase properties. The first three or four are mainly functional, that is, they are very labile and change quickly (10^{-1}–10^{1} years). The solid-phase profile is more stable, changing slowly ($10^{1\text{-}2}$–$10^{5\text{-}6}$ years), and is characterized by interrelated horizons with variable texture, structure, and mineralogical and chemical composition (Fig. 5). Many kinds of diagnostic features and horizons are recognized, and their combinations give rise to many unique soils throughout the pedosphere.

Classification systems such as the World Reference Base for Soil Resources and Soil Taxonomy are based on combinations of defined pedogenic properties, mainly solid-phase ones. The organization of these classification systems facilitates representations of the pedosphere at the global scale which facilitates recognition, identification, and classification, as noted by the color patterns in Fig. 6. The lateral combinations of individual soil bodies comprise the continuous soil cover of land, the pedosphere. Spatial patterns or structures of soil cover exist at all scales. In accounting for the dominant pattern at different scales, a general hierarchy relating to the scale at which soil forming factors operate has been observed. Bioclimatic factors tend to relate well to patterns shown on soil maps at the global scale. At more local scales, bioclimatic factors are less variable, but variability relating to differences in parent material are more easily represented. When more detail can be included in the soil map, then the effect of topography on soil variability becomes more evident. However, there are differences of opinion about what and how to define the combinations of pedogenic properties for organizing classification. Because of this, the percentage of soils versus non-soils are not well known partly due to different mapping conventions used by national organizations. Unlike a biological classification system, a soil classification system can be open to integrate both natural patterns as well as perceptions of importance.

Due to major climate changes throughout the Pleistocene and Holocene landscape evolution altered many areas of the pedosphere mainly through erosional and depositional cycles. More recently, human activities have modified most of the land, so that few truly natural soils exist. Those that remain are typically in areas that are largely in-accessible or unused by humans (e.g., tundra and boreal taiga zones, high mountains, tropical rainforests, and extreme deserts). New kinds of anthropic features and soil horizons are being identified, described, and recognized as significant features of the pedosphere. The World Reference Base for

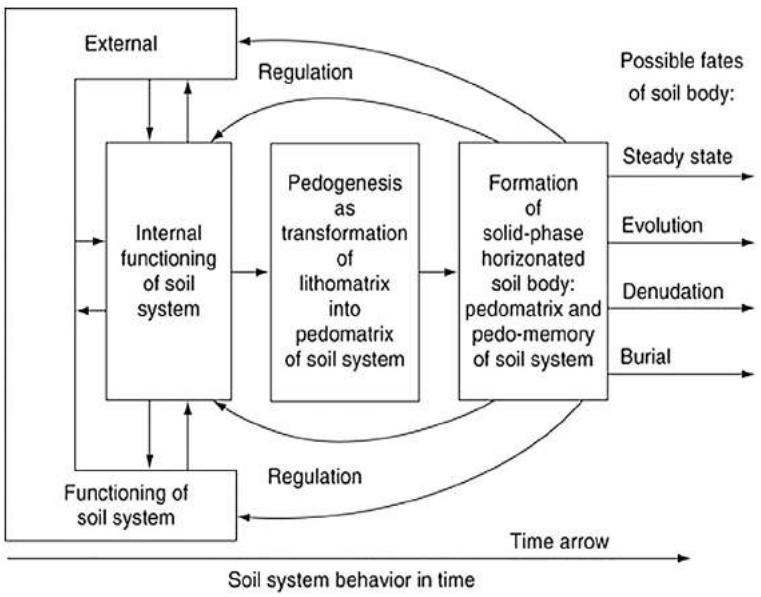


Fig. 3 Functioning of a soil system and possible future condition of a soil body.

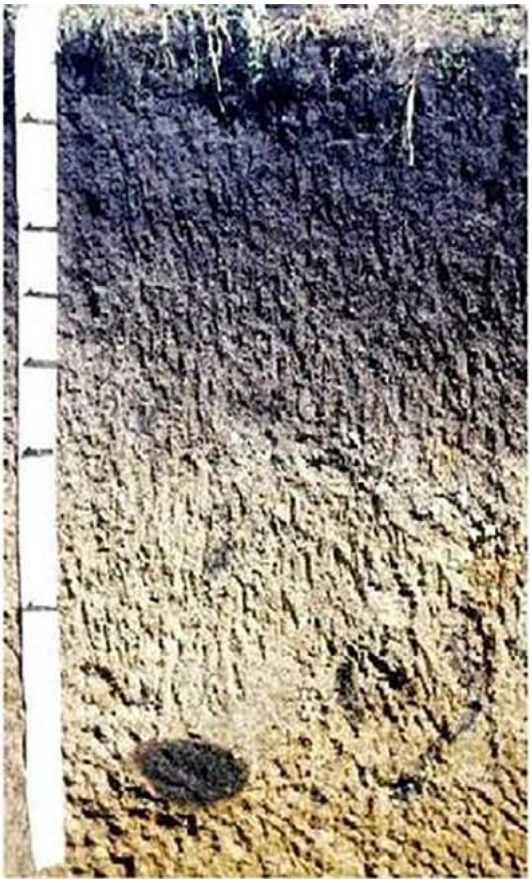


Fig. 4 Vertical variability revealed as horizons in a drained and cultivated Mollisol soil derived from calcareous till in Iowa, USA. Photo credit: R. W. Arnold.

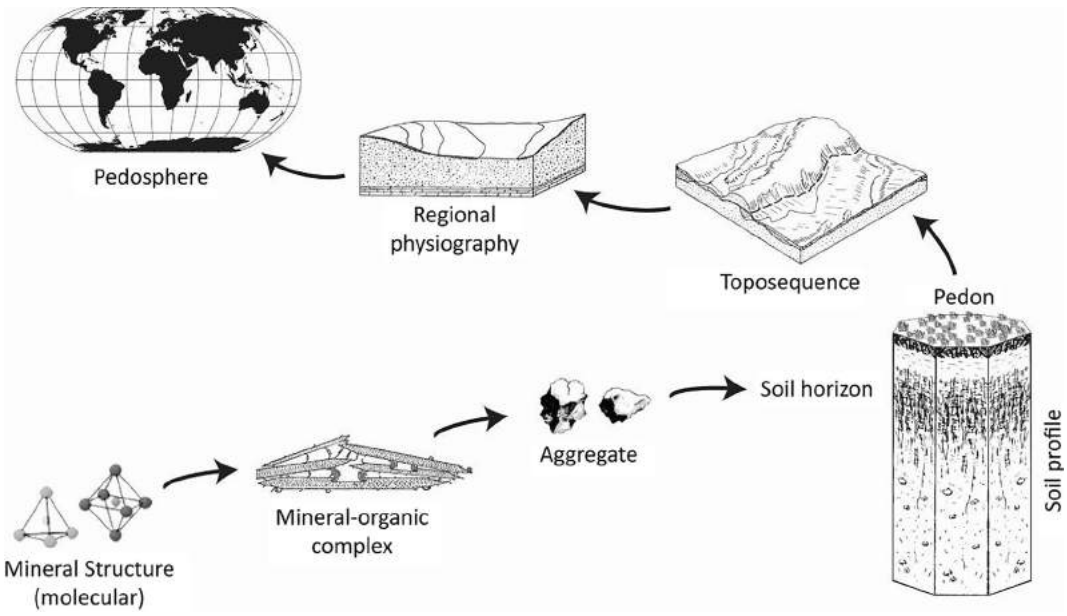


Fig. 5 Schematic of hierarchal scales involving soil solid phase components that combine to form horizons, profiles, local and regional landscapes, and the global pedosphere. Modified from Schaetzl, R. and Thompson, M.L. (2015). Soils: Genesis and Geomorphology, 2nd edn. New York: Cambridge University Press.

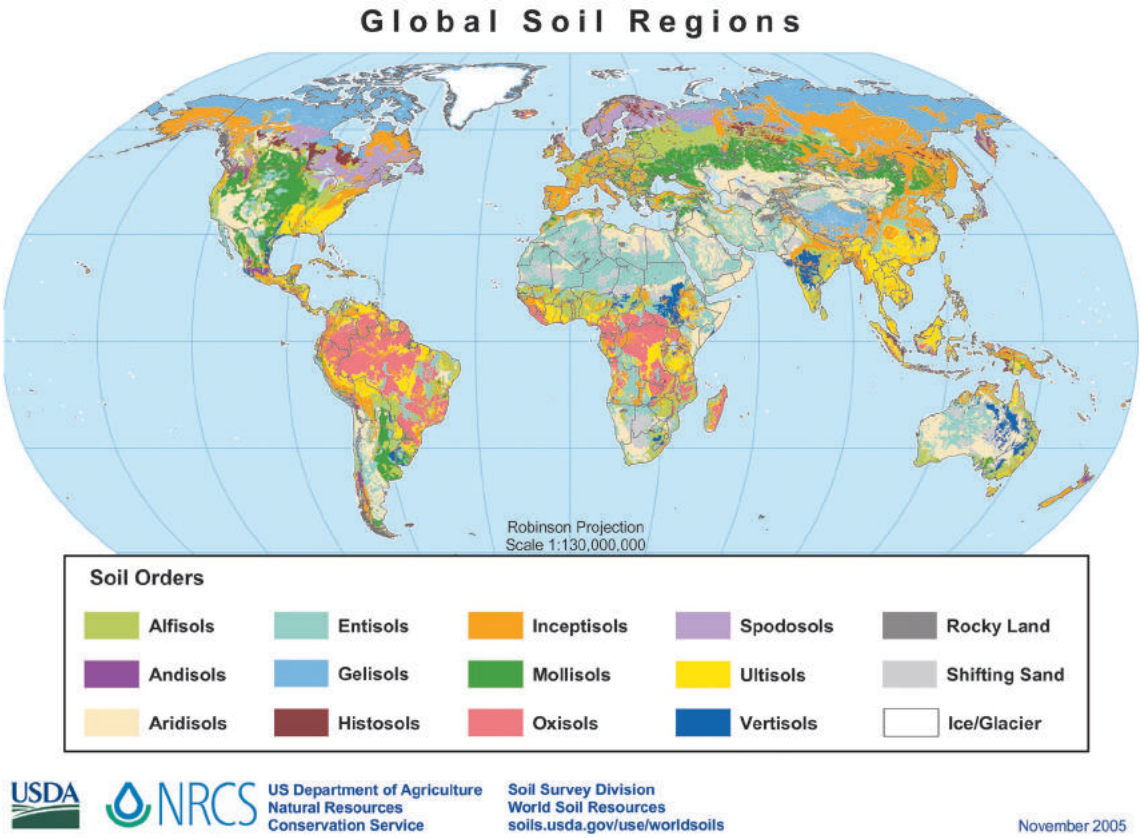


Fig. 6 Map of world soil resources using the orders of Soil Taxonomy. Areas marked as rocky land, shifting sand, and ice/glacier are essentially free of soil, but all other terrestrial areas have soil cover. Map produced by USDA-NRCS.

Soil Resources and Soil Taxonomy now reflect such changes. Refinements of the concept of the pedosphere will be and are being made as improved techniques for their examination and measurement become available.

Pedo-Memory

Most soils are organized, structured, natural entities whose pedogenic properties have recorded the main features of environments and ecosystems that existed during their formation and subsequent changes. Soil, therefore, is commonly a product and an imprint of long-term interactions and functioning in nature. During the past two or three centuries, much of the pedosphere has also recorded many anthropotechnogenic impacts and those portions now have memories of complex biosphere–lithosphere–anthroposphere interactions. Soils have different capacities for recording past and present environments depending on the time required for processes to come into quasi-equilibrium with environmental conditions (characteristic times, CTs). General CTs are: for gaseous phase, CT $\sim 10^{-1}$–10^{1} years; liquid phase, CT $\sim 10^{-1}$–10^{2} years; micro- and macrobiota, CT $\sim 10^{-1}$–10^{3} years; and solid phase, CT $\sim 10^{1}$–10^{6} years. These orders of magnitude are only indicative of the wide ranges involved. Although solid-phase features reflect environmental changes more slowly than the other phases, they retain the changes much longer and are the major recorders of prior environmental conditions.

Pedogenic solid-phase properties also have different characteristic memory retention times; the more rapidly formed properties may record changes for years, decades, and even centuries. Litter leaching and decomposition, soil structure formation and degradation, salinization and desalinization, and reduction and oxidation are examples. The slower-formed properties may record changes for millennia to millions of years, for example, deep and strong weathering, transformation and translocation of clays, and alteration and accumulation of iron-rich compounds. The age of soil memory depends on the duration and interactions of soil-forming and weathering processes that occurred at a specific place.

Soils of the existing pedosphere generally consist of complex combinations of inherited properties of pre-Pleistocene and Pleistocene weathering, landscape evolution, and pedogenesis, as well as more recent Holocene and Anthropocene impacts. Some features of pedogenic properties are partially erased by erosion and other degrading processes such as excessive leaching or acidity, and later processes of landscape and soil evolution overprint properties and memories (a palimpsest phenomena). Usually local knowledge of geomorphology, sedimentation, hydrology, and past climates provide a foundation on which to base pedogenic

interpretations. The complicated records of the pedosphere are slowly being read by pedologists to provide more information about past environmental conditions. Understanding soil components as carriers of pedo-memory and the rates of change of solid-phase properties remains a challenge to understanding and predicting future changes of the pedosphere.

Functions

The pedosphere is an extremely active terrestrial and subaqueous layer surrounding the Earth whose functions are closely linked with other spheres. The biospheric function is the major production function as it provides soil fertility and a suitable habitat for most species of organisms, thereby supporting land biodiversity. By this function, biomass transformations occur, nutrients are supplied and cycled, and the myriad microorganisms in soil enable sustainable biological productivity, diversity, and activity. Their metabolism is the primary basis for regulation and production functions in soils. Most biogenic substance fluxes are known as biogeochemical turnovers. The Millennium Assessment indicates that more land was converted to cropland since 1945 than in the 18th and 19th centuries combined, and that agricultural land uses now cover a quarter of the terrestrial surface.

Because the pedosphere is the zone of interaction between the biosphere and the atmo-hydro-lithospheres, it is commonly thought of as a reactor and regulator that functions to mediate and control fluxes of energy and substances. For example, temperatures are modified by the pedosphere and make most life, as we know it, a possibility.

The atmospheric function includes energy and moisture exchanges, respiration, and transfer of gases, including oxygen and greenhouse gases, and the force that transports and deposits dust derived from soils. Because of porosity, permeability, and absorption, soils have a hydrospheric function to partition water in, through, and out of the pedosphere. The geochemistry of the Earth's waters are mainly determined by the influences of the pedosphere. Where resistance thresholds are exceeded, water erodes surface particles from soils and deposits sediments downstream. Soil erosion degrades soil quality, often jeopardizing sustainable uses of soils, and can cause considerable downstream damages and economic loss. The lithosphere function of soils is that of providing a suitable medium for the biosphere components that support most terrestrial life. The large variability of the geosphere indicates that no single strategy for use and preservation is likely possible. The pedosphere has an important utilization or carrier function manifested as building sites for communities and transportation networks. Soils supply materials for many types of construction, and also are critical areas for waste disposal.

Last but not least is the cultural and historical function of the pedosphere. Society's interactions with soil were initially for agricultural purposes and the lore is rich with stories and myths of the power of unseen forces to help sustain soil fertility. Soils also serve as a repository of archeological artifacts, stratigraphic markers, and memory of ancient settlement environments. In general, human attitudes that define "self" in a context and in relation to nature result in religious beliefs as ways of bringing order into the seeming chaos of nature. The biogeochemical cycling of life, from dust to dust, is such a concept. Sanctity and stewardship of resources have their roots in the pedosphere.

Some Limiting Conditions

The Atlas of the World Reference Base for Soil Resources illustrates the striking variability of soils in the pedosphere, reminding us that there is a lot of uncertainty in the details of spatial patterns and explanations of soil evolution. Because soil conditions such as fertility, drainage, and topography can be artificially modified and changed by external activities, it is often assumed that the pedosphere is a renewable resource. However, experience has demonstrated that maintaining soil functions desired by society is not ecologically sustainable; rather, they must be reinforced with external energy and substances. The characteristic times of formation and/or resilience of many ecologically and agriculturally important soil features are much longer than human lives and even longer than some civilizations. The interactions of environmental conditions in natural ecosystems produce modifications much more slowly in soil than needed by modern society to provide expected products and services. During the next 50 years, demand for food crops is projected to grow by 70%–85% under the Millennium Assessment scenarios, and demand for water by 30%–85%.

The pedosphere with its functional and structural features has its own space and time limitations. Thickness and area are spatial limitations, whereas temporal functions and soil processes vary so widely that incongruencies and inconsistencies often make successful management or control very difficult. Soil thickness is not the thickness of the rooting zone, rather it is the unspecified thickness of an upper layer of the lithosphere involved in regular bio–litho–atmo–hydrosphere interactions. All of the interactions and resulting processes are relevant to defining the functional thickness of soils. This pedosphere thickness strongly controls and regulates the interactions—it is a real biogeomembrane of the Earth. The shallowness of fertile topsoil limits agricultural use and is susceptible to contamination by pollutants, in addition to degradation and destruction due to human-induced erosion.

Assuming the ice-free land area is about 131 Mkm^2, it has been estimated that about 93Mkm^2 is biologically productive land, of which forests are about 33%, pastures 32%, and cropland 11%. Only about a third of the land surface has pedosphere components that can reasonably be expected to provide sufficient food to support our current human civilization. Major limitations for agriculture include drought, nutrient deficiency, pollution, shallow soil depth, excess water, and permafrost. Other use limitations involve expansion of urban areas and transportation networks, small isolated tracts of suitable land, traditional parceling of land ownership, high costs of preparing land for cultivation, and loss of productive land to non-sustainable practices that cause degradation.

Why are temporal functions a limitation? As mentioned, natural changes of the pedosphere occur at rates too slow to satisfy the desires of modern society. Rates and characteristic times of soil functions, formation, and evolution processes cover at least nine orders of magnitude (from 10^{-3} to 10^{6} years). During the Anthropocene, humans have exploited the pedosphere's "treasure trove" that accumulated over millennia and hundreds of thousands of years of natural soil formation and evolution, creating a modern-day dilemma.

Further Reading

Arnold RW, Szabolcs I, and Targulian VO (eds.) (1990) *Global soil change. Report of an IIASA-ISSS-UNEP task force on the role of soil in global Change.CP-90-2*. Laxenburg, Austria: IIASA. http://www.iiasa.ac.at/Admin/PUB/Documents/CP-90-002.pdf. Accessed 10 June 2017.

Brevik EC, Calzolari C, Miller BA, Pereira P, Kabala C, Baumgarten A, and Jordán A (2016) Soil mapping, classification, and modeling: History and future directions. *Geoderma* 264: 256–274.

Buol SW, Southard RJ, Graham RC, and McDaniel PA (2011) *Soil genesis and classification*, 6th edn. Ames, IA: Wiley-Blackwell.

Certina G and Scalenghe R (eds.) (2006) *Soils: Basic concepts and future challenges*. Cambridge: Cambridge University Press.

Hempel J, Micheli E, Owens P, and McBratney A (2013) Universal soil classification system report from the International Union of Soil Sciences working group. *Soil Horizons* 54(2): 1–4. https://doi.org/10.2136/sh12–12-0035.

Howard JL and Shuster WD (2015) Experimental order 1 soil survey of vacant urban land, Detroit, Michigan, USA. *Catena* 126: 220–230.

Indorante SJ and Jansen IJ (1984) Perceiving and defining soils on disturbed land. *Soil Science Society of America Journal* 48: 1334–1337.

IUSS Working Group WRB (2014) *World Reference Base for soil resources 2014: International soil classification system for naming soils and creating legends for soil maps*. Rome: FAO.

Mellody M (2010) *Can Earth's and Society's systems meet the needs of 10 billion people? Summary of a workshop National Research Council*. Washington, DC: The National Academic Press.

Schaetzl R and Thompson ML (2015) *Soils: Genesis and geomorphology*, 2nd edn. New York: Cambridge University Press.

Soil Survey Staff (1999) Soil taxonomy: A basic system of soil classification for making and interpreting soil surveys. In: *Natural Resources Conservation Service. U.S. Department of Agriculture Handbook 436*, 2nd edn. Washington, D.C: U.S. Government Print Office.

Soil Survey Staff (2014) *Keys to soil taxonomy. USDA-Natural Resources Conservation Service*, 12th edn. Washington, D.C: U.S. Government Print Office.

Targulian VO and Krasilnikov PV (2007) Soil system and pedogenic processes: Self-organization, time scales and environmental significance. *Catena* 71(3): 373–381.

Ugolini FC and Spaltenstein H (1992) The pedosphere. In: Charlson R, Orions G, Butcher S, and Wolf G (eds.) *Global biogeochemical cycles*, pp. 85–153. San Diego, CA: Academic Press.

Viscarra Rossel RA, Adumchuk VI, Sudduth KA, McKenzie NJ, and Loberg C (2011) Proximal soil sensing: And effective approach for soil measurements in space and time. *Advances in Agronomy* 113: 2–72.

Phenomenon of Life: General Aspects

SV Chernyshenko, Dnipropetrovsk National University, Dnipropetrovsk, Ukraine

Introduction

The life phenomenon is one of the basic problems of understanding the universe. It is extremely important for both natural sciences (physics, chemistry, biology, etc.) and humanities (philosophy, psychology, etc.). The process of perception is a loop leading through inorganic nature, life, and consciousness back to reflection of the foundations of nature; so its understanding cannot be complete without answering the question: ``What is life?''

The life phenomenon problem can include two important aspects concerning life, as a general concept, a logical scheme, on the one hand; and as the real object, special natural realization, on the other. This article is devoted to the first approach. The universal definition of life (including, e.g., its potential electronic forms) cannot be complete at the moment in absence of the practical experience of dealing with extraterrestrial or artificial life. However, it can be obtained by way of extrapolation of stored knowledge and is useful for the study of real life forms as a theoretical background, helping interpretation of real observations and giving general perspectives of life science development.

General Principles of Life

Life is a form of matter organization. It is an extremely complex phenomenon, which is still poorly comprehended by both science and common sense. The main features of life as a general phenomenon are the following:

- It is a dynamic process. It is impossible to stop it (even mentally) for investigation. Stopped life is death.
- It is superposition of many different scales. One cannot understand life without understanding of the different level processes: from the microlevel (down to quantum processes) to the macrolevel (up to planetary and space processes). The levels are in permanent interaction. During the evolution of life, both corpusclarization and globalization took place; they have been consistent in both directions.
- It is a hierarchical system of numerous elements. Biological systems can be described by laws of the systems theory and cybernetics. They have abilities for homeostasis, adaptation, use of information, self-organization, and evolution.

Dynamic Nature of Life

Life is not a structure, it is a process. Life units are similar to waves, they permanently renew their composition. The normal state of a biological system is a state of `dynamic equilibrium', when inflow and outflow of matter compensate each other. Metabolism is one of indispensable conditions of a living organism. There is entry of the matter as a source of energy and constructional material, its use (assimilation), and excretion of decay product.

Balance of synthesis and destruction is one of the explanations of the cyclic nature of life. A more general explanation is that, the necessity for a stationary dynamic process to be cyclic, it must coil up in bounded space. At the level of a cell or organism, the cycle is shown as metabolism; at the level of ecosystems it is biogeochemical cycling. The concept of the cyclic character of natural processes is a part of many philosophical and religious doctrines. A good illustration of this fact is the well-known Buddhist Wheel of Life (see **Fig. 1**). The most interesting details are the central circle, where one can see a naive image of closed nutrient cycling and the figure of the demon, personifying time, which gobbles all that is existing.

The next step of development of cyclic movement is iterative dynamics. Recursion, unlimited repetition of itself, can be considered as an important form of nonlinearity. It begets fractals in structure and iterations in dynamics. The main form of iterations in biological systems is replication or reproduction. It is extremely important at least in two aspects. First, it is a way to transmit information from micro- to macrolevel. Second, it is a prerequisite for evolution on the basis of Darwinian natural selection.

Multilayer Character of Life

One of the peculiarities of biology is the fact that it embraces many levels of matter organization, from molecules to biosphere. It results in a large complexity of life, and sometimes complexity and diversity are considered as important characteristics of biological systems. But complexity as such is not a solution; uncontrolled growth of complexity either leads to the reduction of stability, or does not influence it. The stability of real biological systems is a result of very specific interactions between its elements; complex systems must be very well organized. In accordance with the pronouncement of W. Weaver (1948), the subject of biology is `organized complexity', contrary to classical physics (`organized simplicity') and statistical physics (`chaotic

Fig. 1 The Wheel of Life.

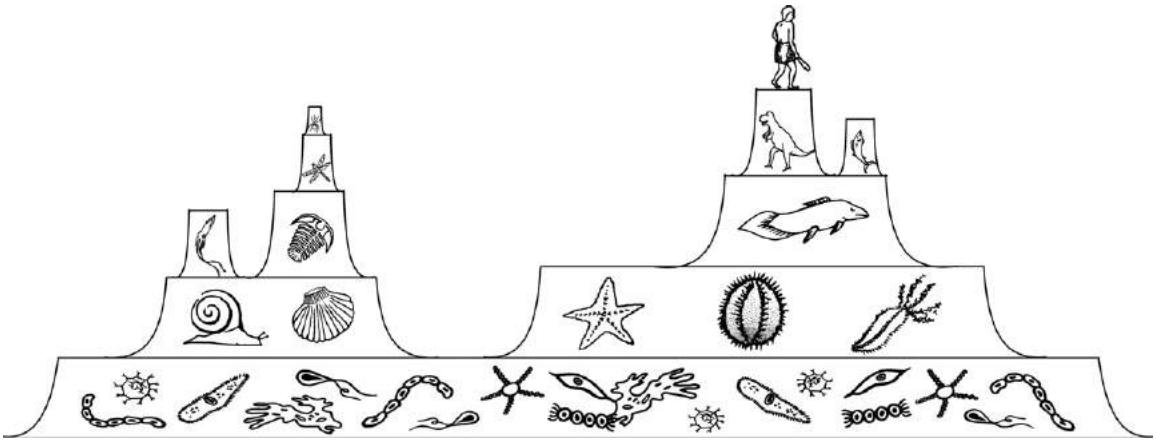


Fig. 2 Pyramid of differentiation of living matter.

complexity'). Dynamic laws should be appropriate for ensuring self-organization. Subjects of biological processes (cells, specimens, etc.) behave not chaotically, but coordinately.

Life development is, particularly, a process of matter differentiation. Step-by-step life makes the world more complex, changes `the space of abilities', creates potential wells in this space (new niches for itself), and fills the wells with new species. In Fig. 2, the niches are shown as a set of more and more narrow trapeziums, one originating from the other.

While `inventing' new levels of matter organization, life keeps previous ones. Usually new forms of matter differentiation cannot exist without older forms, which are parts of their usual environment from their origin. Each step to deeper differentiation needs huge amount of less differentiate matter. Essential progress in producing new abilities usually accelerates the development of living matter, but this acceleration concerns a decreasingly small part of the matter.

Although sometimes a new form can essentially transform or even annihilate a previous one, the latter usually continues to exist as a basis and environment for the former. Life forms itself as a multilayer object. Each new layer emerges by using energy of the predecessors and establishing new forms of connections between the previous layers' elements.

Such an elementary type of formation does not give an optimal result. New objects' functions can duplicate functionality of lower layers or even be at variance with it. The design of the objects would be more rational in the case of starting from the very beginning, without context of previous stages. But nature prefers to build on the old basements from available bricks, which were not initially planned for forming new buildings. Such a choice has some advantages. Losing optimality, nature saves time and gets reliability. Systems with duplicated (and coordinated) functions are more stable; keeping of low-level reactions is useful in case of temporary degradation of environment and so on.

Interaction of different layers is not trivial; their structure and functions are in the permanent process of mutual coordination. The formation of the life multilayer structure was not a unidirectional movement from the lowest level to the highest one. In the course of history of the Earth, after the first chemical layer, the planetary layer of biosphere was formed. All the other levels (cellular, organism, etc.) were wedged between these extreme layers in the course of the process of `discretization'.

Life, Death, and Immortality

The borders between living and inanimate objects are intuitively clear, but not very strict. Such micro-objects as viruses or plasmids are evidently a part of life, but at the same time they are chemical molecules (or a static group of several molecules) only. These obligatory parasites cannot exist as independent organisms, but they should be considered as living because they are part of biological macrosystems and even play an important role in their evolution. It is a good illustration of the fact that living matter sometimes cannot be divided for separate organisms.

Viruses form crystals, which can be disassembled and assembled again. Their individuality is interrupted; life and death lose their usual meaning. Death is not underside of life, and the syllogism ``If there is no life on Mars, it means that there is no death there'' is only a joke. The idea of death corresponds to high form of life only; it is inapplicable to unicellular organisms, which reproduce by division.

For life as a global phenomenon, death does not exist (at least, we know nothing about its imminence for biosphere). Death of individuals is a peculiarity of the life dynamics; it is explained by inexpediency to continue life of organisms, which have functioned their reproductive period.

Immortality, naturally, is impossible; there is nothing eternal in this world. However, individual life of multicellular organisms can be prolonged. Physiological limits of a lifetime are connected with a restriction for the number of divisions of somatic cells, which is connected, in its turn, with genome spoiling. There have now appeared the first ideas of how to struggle against this spoiling; present-day people have a chance for essential prolongation of their lives.

Extraterrestrial and Artificial Life

Our understanding of life is limited by its earth forms. Unfortunately, we do not know about extraterrestrial life, although Epicure spoke about it more than 2000 years ago, and J. Bruno was fagoted in 1600 because of his propagation of ideas about its possibility.

Now it is clear that the solar system planets are not really appropriate places for life, at least in its known forms. Jupiter's atmosphere is, probably, similar to that of the ancient Earth, and life can take its first steps there. Venus is too hot because of the greenhouse effect, and life is possible at some height in the atmosphere only. Mars is too cold, but in rocks found in the Antarctica and, probably, originating from Mars, scientists found microstructures resembling structures of leftovers from bacteria on the Earth. This indication of the existence of ancient life on Mars is very controversial, and the fact is only one `collateral evidence' of extraterrestrial life.

Searching for sentient life in space was started in 1960 by the project OZMA, which was followed by the Cyclops program in 1971 and many others later. The search for artificial radio radiation and other indirect indications of life is still unsuccessful, provoking pessimistic opinion that mankind is alone in the universe.

We are also quite far from the origination of artificial life produced by man. In principle, it is possible to design self-assembling robots, but they cannot be reliable and self-sufficient. Modern electronic devices have some properties of living beings, but they are a part of the global noosphere system (combined biological and technical elements) and cannot exist for a long time without the environment of human civilization. Even if the perspective of electronic life exists, it is a very remote one.

Life as a System

Modern view on biological objects as complex systems is proposed by the prominent Austrian biologist L. von Bertalanfy (1901–84), who established a new scientific discipline – `systems theory'. According to his definition, ``system can be determined as a complex of interacting elements.'' Bertalanfy proposed to consider the role of the systems theory regarding living matter as similar to the role of physics regarding abiotic world.

The systems analysis plays the role of methodological background of biology. Fundamental laws of life (such as the law of natural selection) can be interpreted as universal laws for complex dynamical systems. And, vice versa, systems laws are organic for biology and allow for solving many of its theoretical and practical problems.

The effectiveness of the systems approach in biology is closely connected with high level of emergency, which is typical for biological systems. A biological object, as well as all stable complex systems, cannot be understood as a set of separate elements only. Each new layer of hierarchy is a new special object with its own properties, which are based on properties of its elements, but are not their direct consequence. There are two aspects which can partially explain the phenomenon of emergency:

- A higher level is a very special result of self-organization processes in the lower one. It is a summary of huge current and past processes at the lower level. At the same time, in biology, the higher level plays the role of regulatory mechanism and can radically influence low-level processes. Thus, both levels determine each other.
- Nature prefers `economy' in principles of system organization. Systems can have similar structure, irrespective of elements' nature, and vice versa. It gives a possibility to study systems, abstracting internal elements' organization.

Nonlinearity of Biological Systems

Biological systems can have both linear and nonlinear properties; during their evolution they used all possible types of dynamics to increase their effectiveness and stability. Nevertheless, most biological processes are nonlinear. One can mention the following nonlinear effects: system's state jumping (bifurcation or `transformation of quantity to quality'); system's transition between deterministic and chaotic behaviors; hysteretic effect, that is, the system `remembers' its history; self-organization (purposeful decrease of the system entropy). Examples of evidently nonlinear biological processes are autocatalysis, reproduction, evolution of species, etc.

The analysis of critical regimes and singularities of the parametric space can be used for revealing `acupuncture points', where small local perturbations provoke great large-scale metamorphoses of the system. A spectrum of quasistationary solutions is realized as a set of possible forms of morphogenesis. The discarded forms are still within system's reach but remain dormant, unknown to observers in the course of evolution.

Nonlinear dynamics of living beings is often intuitively incomprehensible; admiration of nonlinear algorithms of life produces paradoxical ideas about intelligence of cytoplasm or bacteria.

Structure and Hierarchy

One of the important characteristics of a system is its structure – a set of links between elements and their space distribution. There are two main structural forms of matter: centralized (hierarchical) and distributed (skeleton). Physical fields have a distributed organization, whereas atoms are centralized. A cell has the center (nucleus); a colony of cells is homogenous; and the organism is centralized again. Centralized organization is rational in the case of high-level differentiation of elements; the distributed one is more typical for systems with homogenous elements.

Often the structure of biological systems is tree-like; they are hierarchical systems. It is the result of two important processes: differentiation of living matter and bunching (oligomerization) of its elements. An important feature of hierarchical systems is the fact that each level is characterized by new emerging properties not presented at lower levels. A scheme of the general hierarchy of biological systems is represented in **Table 1**.

The structure of real systems can be changed, but its dynamics is very slow in comparison with other processes, called functionality of the system. A set of characteristic times can also form a hierarchy. For each process, other ones can be considered as part of the environment: slower ones because of their relative stability, and faster ones because of rapid running to equilibrium.

Table 1 Hierarchy of biological systems

Science	*System*	*Elements*	*Interactions*	*Elements' state*
Biochemistry	Chemicalreaction	Organic macromolecules	Chemical	Form, position, energy, etc.
Cellular biology and genetics	Cell	Organelles and genome	Endoplasmic and nuclear	Kind, size, position
Morphology and anatomy	Organ	Types of cells and tissues	Intercellular, chemical, and electrical	Kind, vitality, phase of development, etc.
Physiology	Specimen	Organs	Interorgans, by hormones and neural impulses	Kind, vitality, state of health, etc.
Population ecology	Population	Specimens	Cooperative and competitive	Age, sex, physiological state, etc.
Global ecology, biogeocoenology	Ecosystems, biogeocoenose	Populations	Trophic, competitive, and cooperative	Size, age, sexual, genetic structure
Biosphere ecology	Biosphere	Regional ecosystems	Through climate, atmosphere, etc.	Productivity, sustainability, disturbance

Cybernetic Principles in Biosystems

Adaptation and self-organization are impossible without information and controlling processes, that is, without realization of cybernetic principles. Cybernetic mechanisms can be found at all the levels of life, from biochemical processes to biosphere.

Self-regulation is a process of changing functionality of the system directed at its conservation. It is development of property of inanimate systems expressed by the Le Chatelier principle (1884) – external influence on the system's state is compensated by internal processes, influenced in the opposite direction. The law's version for open systems can be formulated as the following: an increase of the system's input leads to corresponding increase of its output. This reaction is passive and does not need energy.

For biological systems, it is very typical to use active methods to keep the system's steady state (maintain `homeostasis'). One of the ways is to follow the cybernetic principle of `negative feedback'; the output of some part of the system must influence its input – if the output is too large, the input is decreased, and vice versa.

Nature of Life: Mathematical, Physical, and Chemical Approaches

The multilevel and multimedium nature of life necessitates considering it in different aspects, in the framework of different sciences. Really, many definitions of life have no structure as ``Life is …,'' but only as ``Life can be considered as…." Probably, one can expect gradual synthesis of various approaches to the life problem, but, for the time being, the integrated picture has not been formed.

Mathematical View on Life

Mathematics is a tool for abstraction, a way to the core of scientific knowledge. Mathematics is not interested in details; for it, life is a kind of complex systems with special relations between elements. For the description of different properties of life, there are various mathematical models. Probably, it is impossible to design a universal model of life; each model has its own field of application and level of approximation. Attractiveness of mathematical models does not consist in their complexity, but in their lucidity and explanatory power. According to Einstein, ``Models should be as simple as possible, but not more so."

One of the first biological models was the Malthus model of exponential growth (1800). It was developed for the field of population dynamics by the Lotka–Volterra models (1925–31). Models of life were proposed by J. von Neuman, R. Tom, H. Meinhard, and others; mostly they were differential models. Their use was very productive; in particular, they are a basis for the nonlinear analysis.

Another interesting mathematical tool for life description is the theory of cellular automata. This kind of discrete model has a property to be chaotic at the microlevel and ordered at the global level. In principle, one can imagine the world as a cellular automaton with elements – physical particles. The well-known Game of Life of Conway (1970), which really reflects some features of real life, is also a cellular automaton.

Usually mathematical methods are numerical, but it seems to be a very perspective way to use topological and algebraic approaches also. The above-mentioned topological theory of ecological niches can be considered as an example of this way.

Physical Principles of Life

According to J. S. Mill (1806–73), laws of life cannot be something other than laws of behavior of molecules, interacting as parts of a living organism. But, because of emergence of biological systems, it is not easy to reduce biological laws to physical ones. Such a way, called `reductionism', does not always give practical results, but it is important as the theoretical basis for searching borders of the possible for living objects. It is not easy to predict fundamental consequences from fundamental laws; each forecast of possible effects is a discovery.

Biology is a continuation of physics and chemistry and chemistry is a continuation of physics. One can understand biology as `new physics' and pose a problem to find its form, which corresponds to physical traditions. Particularly, physics of life is possible only under very special values of the world constants; traditional physics `does not know' what life is, and cannot explain these values. It is necessary to use the `anthropic principle': our existence as intellectual beings, studying the world, presupposes its features ensuring origin of man.

In biology, as well as in the other sciences, the problem of energetic balance is very essential. It gives a general estimation of the process of life functioning. As open systems, living objects need permanent energy income; they use it step by step and finally transform it to thermal energy of the environment. The main source of energy for life as a whole is the radiation of the Sun (and, insufficiently, energy of the Earth's interior: chemical, thermal, and, probably, radioactive). Plants (`phototrophs') use the solar energy for chemical synthesis of organic substances (the process of photosynthesis), supporting their own existence and providing chemical energy for all other forms of life: `heterotrophs' (herbivores and carnivores) and `saprotrophs'. Physically one can say that the solar energy in the course of photosynthesis raises energetic levels of electrons in some atoms of living matter; then the electrons gradually and purposefully descend, executing chemical and mechanical work.

Life directs energy flux to itself and uses it. According to the I. Prigogine theorem, an open system, in the case of linearity of the energy flux through it, produces minimum likely entropy. Life is an inconvertible process, going in a linear area of forces–flow rates; it endeavors to keep this linearity. But irreducible small nonlinearity produces stochastic noise, finally destroying each living organism.

Contrary to general physical tendency, postulated in the second thermodynamic law, life as a global process is characterized by gradual decrease of entropy. (Separate organisms also decrease its entropy during most periods of their life, but after their death the entropy `gains revenge'.) The paradox was already pointed out by the father of statistical physics L. Boltzmann (1844–1906), and later was deeply analyzed by A. J. Lotka (1880–1949). In 1944, the Nobel Prize winner physicist E. Schrödinger (1887–1961) published his famous book *What Is Life*? devoted to this problem.

The general explanation why entropy can be decreased in living systems is evident; these systems are open; they use external energy to decrease their own entropy and, at the same time, increase entropy of the environment. In general, both first and second thermodynamic laws hold true. But the ways of converting the energy income to entropy reduction (or maintaining order) is not so clear. According to E. Schrödinger, organisms `drink orderliness' from a suitable environment. He explains about flux of `negative entropy' (negentropy) to organism, which compensates natural increasing entropy. He does not explain the process in detail, but stresses that life's tools for this aim are `aperiodic solids' – the chromosome molecules. Schrödinger's book had an essential influence on molecular biology; particularly, it stimulated J. D. Watson and F. Crick to discover the DNA structure (1953) and explore in that way, the physical explanation of life.

It is not very clear yet what Schrödinger's negentropy is – free or stored energy, information, organization, or something else? Probably, a perspective conception is the idea about necessity for life of two coupled processes. The first (energetic) one accepts energy from environment and provides it to the second (information) process, which is responsible for the living system's development. A disproportion of entropy takes place; the second process presupposes decrease of entropy; the first one, correspondingly, increases it. Such processes are observed in inanimate nature; for example, explosion of an ultranew star transforms it into a primitive clot of neutrons, but, at the same time, heavy elements of the periodic system (prerequisites of life) are synthesized and spread in the universe. High-ordered entropy disproportion in living organisms presupposes very exact coordination of biological processes; in accordance with Schrödinger's opinion, information DNA molecules play the role of the coordination center.

Life is not contradictory to the second thermodynamic law, but uses it in a special way. Excluding from reproduction all the descendants of a couple except two of them, death of prey killed by predator, extinction of species in the course of evolution – all these events on the one hand increase entropy, but on the other hand they lead to general progress, to ordering matter in some local areas (from which, because of reproduction, the new forms spread as widely as possible).

As for inanimate nature, many scientists see in unidirectionality of the entropy change the basis of the time phenomenon; the tendency of entropy reduction in living systems can give a key to understanding of the general laws of living matter evolution. A. J. Lotka in the article `Contribution to the energetics of evolution biology', published in 1922, proposed to consider energetic power of organisms as the main criterion maximized in the course of evolution. Later, he called this maximum power principle, the `fourth thermodynamic law'. The approach is still under discussion; it was supported and developed by such prominent scientists as V. I. Vernadsky and H. T. Odum.

The law is based on the consideration of species' evolution, when in conditions of ``the struggle for existence, the advantage must go to those organisms whose energy-capturing devices are most efficient in directing available energy into channels favorable to the preservation of the species'' (A. J. Lotka). A capability of better assimilation of solar energy or energy collected by other organisms is a prior evolutional advantage.

It is quite right at the level of ecosystems, when stochastic fluctuations and individual peculiarities at the level of species are integrated and averaged out. More and more effective populations are involved into biological cycling, increasing its intensity. As a result, the ecosystem power (consumed energy per unit time) permanently grows. It is mainly the result of competition from plants (producers), which are forced to maximize production for keeping their place in the ecosystem. Another extremely important factor is the activity of animals (consumers). They withdraw producers' biomass and additionally intensify cycling. Probably, the global role of consumers in biosphere consists exactly in the spinning up of ecological cycles.

At the level of concrete species, classical power is not the only parameter determining its evolutionary perspectives. One should take into account, for example, the efficiency of the species in limitation of entropy growth. As a result, it is more reasonable to speak not about all the available energy, but about `exergy'. The latter shows an ability of the organism to make the work relative to the surrounding; it is the `co-property' of a system and a reservoir.

Another important aspect, influencing vitality of the species, is the integrated character of energetic abilities of living organisms. H. T. Odum proposed a concept of emergy (embodied energy) as ``a measure of energy used in the past'' and stored in the system's structure. The concept is being developed by S. E. Jørgensen and others. The maximum `empower principle' is proposed by H. T. Odum as ``a unifying concept that explains why there are material cycles, autocatalytic feedbacks, succession stages, spatial concentrations in centers, and pulsing over time.'' Generalization of the approach is possible by way of taking into consideration `population strategies' of species. For example, one can base on the r/K concept or its modification the r/C model (in the context of which population preferences in division of its energetic recourses between the processes of growth and competition are considered).

Information Basis of Life

The notion about the information nature of life is generally accepted. At the same time, even the term `information' is interpreted in biological literature in various ways. Starting from the classical works of the founders of the information theory, C. Shannon (1948) and J. von Neumann (1951), different directions of the generalization of the term were proposed.

Concerning the information character of internal biological processes, a reasonable approach is based on I. I. Shmalhausen's views (1968) about the resonance nature of biological information. Most of the relations between elements in biological systems are based on special resonance organization of living objects. Very often, an energetically weak influence of one element on another one produces its powerful reaction. This interaction cannot be interpreted as pure energetic; we call it `information interaction'. In this context, information cannot be transmitted; it is a relation between two elements. There is connection between exergy as the object's energy with relation to its surrounding and the possibility of the object to realize information actions.

For realization of an information action, the initial influence (signal) must exceed some critical value: `excitability threshold' or `reobase'. For example, the maximal tension of cell electric discharge is 0.1 V; its reobase is 10% of this value. Sporadic resonance effects take place in inanimate nature too, but for living matter it is its basis; there are special mechanisms of energy charging for the creation of prerequisites for resonance (information) action. From the level of cells, there are extremely complex structures of signal relations, separated from processes of energy and matter transmission. In multicellular organisms and ecosystems, special information subsystems were originated. They influence each other by means of special media (substances and fields), bearing precisely signal character. Special systems of coding and decoding become more and more sophisticated. Information is the main instrument in all homeostatic processes; it is the main way to organize negative feedbacks in living systems.

Physically, the resonance is realization of potential energy; mathematically the same is bifurcation, nonlinear effect of steady-state (attractor) change. One can interpret information processes in a living system as a sequence of `internal bifurcation'. As life is an information process, it exists near separatrixes, divided areas of steady states' attraction in the space of system parameters. It is a mathematical illustration of life's fragility: directed small changes of system critical parameters can easily upset its dynamic equilibrium.

Quite often, the concept of information is used in biological literature for designation of a measure of living matter ordering. In this case, information is the opposition to entropy, something like Schrödinger's negentropy. In the course of life development, its information content gradually grows. The philosophic question about the origin of information in the universe is still open. There are two main opinions: either the Big Bang in the result of a symmetry breach created all existing information and it is gradually sowing up, or its quantity equaled zero at the beginning and it is in the process of permanent growth. In the second case, the information creation is bound with Darwinian natural selection. According to G. Kastler ``the information creation is storage of random selection.'' It happened also in the inanimate nature, but became much more intensive in the living one. The history of the universe is characterized by exponential growth of information.

General Chemical Principles of Life

Living matter is a direct sequel of the chemical level of organization. It is not invariant to its chemical composition; the material determines key properties of life.

Chemical reactions in living systems are cyclic and autocatalytic. According to the hypothesis of S. Kauffman (1993), big chemical systems of interacting polymers, which reach a critical level of complexity, necessarily become autocatalytic and self-replicating. Their elaboration can be ordered by natural selection. As examples of autocatalytic reactions, one can consider Calvin's cycle (`propagation of sugar phosphates') or replication of ATP, also connected with photosynthesis. The cyclic chemical reactions discovered by A. Szent-Györgyi and represented by the famous Belousov–Zhabotinsky reaction can play the role of `soft clocks' in living organisms.

Life is a positive connection between information molecules and proteins. The most important chemical cycle is the following: DNA produces enzymes, which, in their turn, ensure its replication.

Organic macromolecules are not thermodynamic objects; they have no aggregate state and are naturally far from a steady state. It is a prebiological stage of matter development. According to Schrödinger's simile, a living organism as well as a pendulum clock are not thermodynamic objects; because of solidity of the clock and stability of the hereditary substance of the organism, room temperature for them is practically equivalent to zero.

Polymerization as well as enzymes decrease entropy, and thereby decrease molecules' freedom to move and select their co-reactants. A principal difference of life chemistry from ordinary one is the matrix synthesis. For living matter, contrary to inanimate one, a strict order of extremely long chain of reactions is possible.

Life Is a Way of Matter Self-Organization

It is common knowledge that life is the manifestation and result of a general tendency of matter to self-organization. The difference between opinions consists in understanding of the predetermination of self-organization steps. Really, this difference is not so essential: on the one hand, random mutations in pure Darwinism lead to realization of more or less predetermined process of adaptation to existing conditions, whereas on the other hand, a `vital force' is needed in some mechanisms (why not stochastic?) to conduct their programs. The self-organization algorithms must have some physical basis (self-organization can be considered as a physical principle according to M. Eigen); it is not an opposition to Darwinian selection.

Self-organization is a process of forming order from disorder. It is possible in the following conditions: (1) there is a big quantity of simple components; (2) the components can constitute mutual relations; (3) there is a source of energy, supporting the

relations formed; (4) external conditions are suitable for stability of the new system; and (5) the systems can play the role of elements in the process of forming systems of a next level. Evolution of living matter is a combination of rising tendency, connected with growth of complexity, and descending one, connected with tendency to stability.

There are no special biological fundamental forces or elementary particles. Life is the continuation of the inanimate world; it uses all opportunities given by the laws of nature, particularly nonlinear scenarios such as phase transformation, hysteresis effect, the formation of dissipative structures, etc. All the possibilities should be groped for, probably by accident.

Biological Self-Organization as a Process of Niche Creation and Filling

Life can be considered as existing in its own space of possibilities (a `space of life'). Mathematically, this space can be called as a `phase space'. Potential wells in this nonlinear space are occupied by different forms of life; they are the so-called ecological `niches'. Usually the niche of species is defined as a set of environmental factors corresponding to the needs of the species. A linear world (without sources of energy) would have no niches. The inflow of energy creates or allows to create wells in the life space, where life can consolidate. The situation is similar to the physical picture of the world; elementary particles, probably, exist in potential wells and the world space is furrowed by some fundamental processes (e.g., the energy conservation law can be approximate, and leaking energy furrows space).

Evolution is a process of optimization. Biological systems evolve to a steady state or, in other words, optimize some criterion (Lyapunov's function), looking for a potential well (a niche) in a phase space. One of the optimization criteria is entropy production, which must be minimal in a state of dynamic equilibrium in accordance with the Glansdorff–Prigogine theorem.

In a niche, the system possess properties of homeostasis; after small distortions, it will return to the initial state. The evident discrete character of the steady states' set produces separateness of niches. Heterogeneity of the life space explains existence of strictly separated species and other taxonomic units. Living organisms are substantially casts of their niches. There are many examples of convergence, when similarity of niches leads to similarity of organization of living beings.

Tendency of nonlinear systems to keep in equilibrium is well known as a general systems property of `equifinality', postulated by L. Bertalanfy. It is a spontaneous reaction of the system; it is an important factor of evolution, but it does not direct the life progress. Optimality does not mean progressive character; the level of amoeba fitness is not lower than that of man. Evolution uses the gradient method, which allows finding a local extremum only; for improving living conditions, species must jump over unacceptable ones. According to I. Prigogine, self-organization is a process of step-by-step loss of stability. The change of a steady state can be a result of either essential external influence upsetting the system from a previous equilibrium and forcing it to transit to a new one, or a gradual change of system parameters leading to a change of the phase space topology, to disappearance of the current steady state. Both of these ways were combined during life history; the model corresponds to the saltation conception about evolution as a sequence of catastrophes. External influences (geological, cosmic, anthropogenic) cause disappearance and reorganization of niches. New niches stimulate evolution of their potential hosts, which fill them because of reproductive abilities producing `pressure of life' (see Fig. 3). Species, which have lost their niches, should progress or become extinct.

In accordance with another topological model, catastrophes (and accompanying biological innovations) create new dimensions of the life space. In the niche (a point of local minimum), a new dimension appears; correspondingly, the host of the niche loses its stability and gets a possibility to find a better position in a new direction. At the result of such step-by-step changes, the development trajectory of the species is formed as a chain of orthogonal line segments. The path of progress can be compared with a `bobsleigh track'.

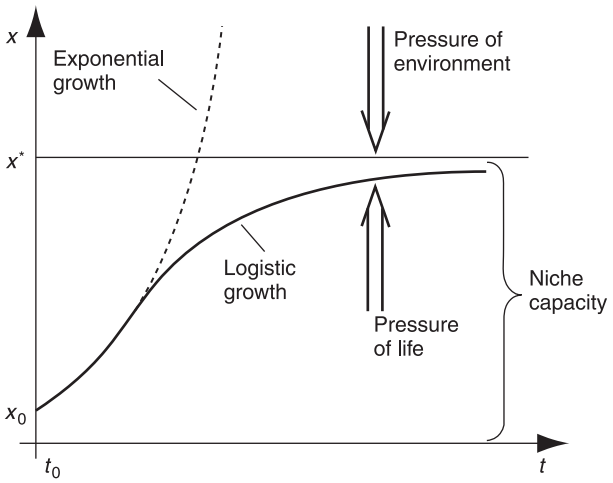


Fig. 3 Niche capacity as a result of dynamic equilibrium between pressure of life and pressure of environment.

Can evolution proceed under the constant conditions of environment? Darwinism does not exclude this possibility; the development of life organization can go on in the course of successive small improvements (decreasing entropy), but this movement is very slow. Usually it does not presuppose essential change of phase space topology; although it can sometimes stimulate radical changes: for example, the development of photosynthesis brought to the oxygen revolution and total reorganization of the system of niches. As a rule, a host of a stable niche slowly evolves and can meet competition only from the side of kindred species or, quite rarely, introduced ones. The Black Queen hypothesis about the necessity of permanent improvement of all species (``to stop means to die'') is not true for stable niches. This fact is illustrated by the existence of a great number of primitive species formed millions and billions of years ago. For continuation of the race, it is necessary to change topology of the phase space, to disturb the system.

Relatively fast evolution of life has become possible because of general instability of niches. The history of life is a sequence of actions of forming and filling niches. As a result of divergence, niches bifurcate; the process of `niche proliferation' takes place. Quite fast processes of niches interaction are ecological successions, when niches regularly change and supplement each other. Often infill of a niche creates a number of new ones, and, importantly, the complexity of the derived niches is usually higher than that of the initial one. It creates prerequisites for progressive evolution of life, an increase of its complexity. Newer forms of life produce newer local worlds, up to virtual worlds in the human mind, which are entirely real as both processes in brains and a plan for the real world change. Man with his imaginary worlds creates principally new powerful niches, particularly, for the development of artificial `electronic beings'.

Principally, the evolution process concerns not separate organisms, but the biosphere as a whole. The optimality of organisms is not an absolute value; it has a sense in the context of environment including other organisms. In other worlds, self-organization of life cannot be understood at the level of organisms; it is necessary to consider the general life space, separated species in corresponding niches, and their interference.

General Evolutionary Rules

1. Evolution is irreversible (the principle of L. Dollo, 1893). Why does the return to former conditions not lead to recurrence of old forms? There are at least two explanations. First, energetic innovations found during the previous stages of the evolution will not be lost if they are effective. Second, change of a living form is reflected in surrounding ones; the form can regress only together with its community. It is not so probable, although sometimes an ecosystem can degrade – not because of degradation of developed forms, but as a result of their elimination and return to the forefront of primitive forms.
2. Evolution is a movement from simple to complex, but the simple is not destroyed – it is included in the new structure of life. According to A. Szent-Gyorgyi, life never revises what has been made, but it builds above the existing. The cell is similar to a lot of archeological diggings, where one can see a number of strata – the older, the deeper. The structure is not optimal, but is reliable; the damage of upper layers is not fatal – ancient mechanisms can smooth the blow.
3. Complexity growth stimulates self-organization process. As is shown by J. von Neumann's model of self-organizing automata, the ability to self-organize depends on complexity of the object. There is a critical level, beginning from which automata can reproduce more complex forms than themselves. During its development, life also passed through critical levels, which forced its tendency to progressive evolution.
4. Evolution has often a binary character. Evolving systems can consist of two closely connected parts such as DNA–proteins, cell–nucleus, male–female, etc. During active stages of evolution one part outstrips the other; the first one becomes an object for experiments, and the second one guarantees preservation of attained level.
5. Life uses for its self-organization, natural frequencies (modes) of component systems. Some modes can intercept incoming energy from other ones and intensify themselves. This effect has resonance nature and is close to information phenomenon. Octave principles can take place – structures are formed by series on multiple frequencies. It is one of the ways of forming hierarchy of living systems.
6. Spatial self-organization of life can be considered as a process of dissipative structures' formation – one of the nonlinear mechanisms of originating order from disorder.
7. Life evolution is directed from profound symmetry to absolute asymmetry. The main stages are ball, radial, axial, and bilateral symmetries, and triaxial asymmetry. The asymmetry is a reaction on anisotropy of the living space, for example, gravity anisotropy.
8. Evolutionary process is not so much an invention of new forms as a search for effective combinations of existing ones. Various living beings are built from the same standard bricks, which were formed during early stages of evolution.
9. Evolution evolves itself. In the history of life, there were several `evolutionary formations', characterized by special evolutionary factors and features of self-organization forms.

Basics of Darwinism

Although people have used selection of domestic animals for a long time, and first guesses about the development of life were stated in the antiquity (e.g., according to Aristotle it is driven by a special living force – entelechy), the theory of `transformism' (about changeability of living forms), which opposed the theory of creationism (about constancy of organisms, created by God), was formed only in the eighteenth century. The first fundamental theory of biological evolution was proposed by J. B. Lamarck

(1744–1829) in 1809. This theory, progressive for its time, did not include the idea of natural selection and assumed inheritance of acquired characters. In 1858, Charles Robert Darwin (1809–82) proposed a new evolutionary theory based on the mechanism of natural selection.

Darwinism includes two key ideas: undirected variation (`mutations') of discrete hereditary codes (`genes') passed from parents to children, and elimination of less-adapted individuals in the course of `struggle for existence'. The codes, better reflecting the external conditions, gradually become dominant; average characteristics of individuals are changed. Selection transfers information about environment in the hereditary code. Useful negative fluctuations of entropy are spread over the species.

The Darwinian idea of natural selection can be considered as a very general explanation of matter self-organization – the tendency of entropy decrease in particular objects. This thought corresponds to the opinions of the founders of mathematical genetics, such as R. E. Fisher (1890–1962): ``Natural selection is a mechanism for generation improbability,'' and A. Lotka: ``The principle of natural selection reveals itself as capable of yielding information which the first and second laws of thermodynamics are not competent to furnish.''

Inheritance, Variability, and Natural Selection

For Darwinian self-organization, life must be structured for discrete generations and lifetime of each generation must be limited. To be effective, transfer of hereditary information from a previous generation to the next one cannot be absolutely free. Each species is divided into more or less isolated populations, where panmixia takes place. New specimens' characters are examined at the level of populations; only in the case of success, they spread for the whole species. Sometimes there are additional levels of such hierarchy: subspecies, races, subpopulations, etc. Structural units of species, uniting closely related individuals, are partly reproductively isolated. It is still not clear whether it is a natural result of hereditary remoteness or it is a manifestation of special isolation mechanisms forming an optimal structure of the hereditary field.

An important fact is discreteness of the hereditary code; genes are indivisible. This consideration eliminates Jenkin's nightmare: a useful character cannot resolve in descendants of reiterative coenobium.

Inherited characters cannot be absolutely independent; some of them are more or less correlated. On the one hand, it is a destabilizing factor; characters are selected in the context of other ones. On the other hand, it leads to tendentiousness of genes' variability, joint manifestation of correlated characters in accordance with the homologous series law of N. I. Vavilov (1920).

Darwinian variability is a principally random phenomenon. It is impossible to predict dynamics of external conditions; evolutionary perspective living forms must have multidirectional hereditary deviations. Initially directed evolution (as Berg's nomogenesis) is not sufficiently flexible to be effective.

Genes' variability (mutations) must be within reasonable limits. If it is too small, the progress will be too slow and can stop far from a local extremum. If the variability is too big, it will lead to system's chaotic behavior. These parametric effects are well illustrated by mathematical models based on the well-known `genetic algorithm'.

Darwinian natural selection examines the character of different specimens: stability, amativeness, reproductive potential, etc. It is not always a struggle for existence; often it is a struggle for leaving sufficient number of descendants. The main criterion is birth rate. If it is less than one, the species is doomed.

There are three levels of natural selection; only such forms of life can exist, which: (1) are stable and can physiologically give a breed; (2) allow origin of intellectual man (the anthropic principle); and (3) survive in the course of competition with other species (Darwinian selection). Evolution eliminates evidently defective individuals; other ones are not really exterminated, but rather `squeezed' from the ecosystem because of low birth rate.

Natural selection can be classified into three forms: stabilizing (supporting existing adaptations in stable environment); motivating (producing new adaptations); and disruptive (leading to separation of the population in condition of heterogenic environment).

Selection leads to harmony of the organism and its environment. Similar conditions produce similar organism's forms of adaptation to them. This effect is called `convergence'. Adaptation can follow a limited number of ways; in particular cases, it is not evident whether this form is a result of selection, or it is a direct consequence of physical laws.

Evolution can be divided into micro- and macroevolution. Microevolution consists in accumulation of hereditary changes in the population. At this stage, the most essential effects are the change of statistical distribution of different genes within the population and a search for their optimal combinations. The main factors of population evolution are maintenance of genetic heterogeneity, population size fluctuations, reproductive isolation, and natural selection.

Macroevolution is the evolution at the level of ecosystems; its result is the origin of new species. Its nature is not absolutely clear yet. It can be explained by either multistep microevolution or a random essential change, appearance of `perspective freaks'.

New Tendencies in Darwinism

The modern evolutionary theory is an elaboration of the classical Darwinian theory. Apart from such extremely important fields as `genetics', Darwinism accepted a number of new ideas such as genetic drift and recombination, cooperative evolution, and global biospheric context of evolution.

Mutations do not so often revolutionize populations; mostly, they support its genetic heterogeneity. In accordance with modern views, genetic mutations are not obviously new forms of genes. Usually they are recombinations of existing hereditary material, at the molecular, cellular, organism, and ecosystem levels. Such recombinations can produce a fast evolutionary leap forward. There is a common genetic pool of life; similarity of genes does not necessarily mean cognation.

According to Kimura's theory of `genetic drift', genes can exist and even breed in a latent form and then rapidly declare themselves. The effect does not contradict Darwinism; it only proposes a broadened understanding of mutations and their formation.

Variability as a result of genetic change is quite typical at the lower level of life. In cells, there is dissipated genetic material, which does not influence its characters, in silent parts in DNA (introns), in cellular parasites, viruses and plasmids (DNA molecules in the cellular protoplasm, which can be transmitted not only to descendants, but also to neighboring cells).

Although multicellular organisms are protected from genetic material damage by special mechanisms, cases of genetic material transfer can take place for them as well: by viruses, as a result of distant hybridization, etc. Besides, genes, reflecting environment, reflect genes of neighboring organisms. Thus, hereditary units (genes) of all living beings form a closely connected system, a general `gene pool' of biosphere.

One of the factors of horizontal genetic transfer, great importance of which has been understood lately, is the formation of symbiotic beings. A well-known example of symbiotic organisms is lichen, consisting of two species, fungi and alga, which, in principle, can exist separately. Because of long-standing coevolution, these two species have adjusted their biochemistry and synchronized reproduction. There is a small green sea worm *Convoluta roscoffensis*, feeding on symbiotic algae; algae germs transfer to the next worm generation through gametes. Most animals need symbiotic microorganisms for effective digestion, luminescent organs of animals are a result of symbiosis with bacteria, etc. One of the most important revolutions in the history of life was the origin of eukaryotic cells as a result of step-by-step symbiotic integration of several prokaryotic ones that finally shared their genes.

A subject of competitive selection can be a symbiotic system, unifying initial forms, which before had struggled for existence separately. The symbiotic evolutionary theory complements Darwinism with a deeper appreciation of the fundamental cooperative processes, which accompanied the origin and evolution of life.

Each organism has its own place in the biosphere. It cannot evolve separately; it should coordinate its change with connected species. Particularly, evolution of separate organisms should not damage biogeochemical cycles. In the course of evolution, the cycles, as well as the biosphere, reproduce themselves as comprehensive wholes.

Although the natural selection operates at the level of individuals, increase in information takes place only at the levels of species and ecosystems. Evolution of the biosphere is a grand process of information collection. The main source of the information storage is the biospheric gene pool.

Summary

Although life is an extremely complex phenomenon, including processes of different levels, nature, and duration, modern science has elaborated general approaches to its understanding. First of all, it is the systems approach which allows describing multilevel structure of life and modeling some general principles of its dynamics. The physical view on life is interesting due to understanding of biological laws as continuation of physical ones and the energy approach to living systems, particularly, investigation of role of entropy, exergy, emergy, and so on in biological evolution. Another interesting subject is the information nature of life, which is, finally, the main manifestation of universal information processes.

Dynamics of life is a permanent process of self-organization. A few very general principles of this process can be formulated: equifinality of biological processes; development of life as a process of niches proliferation; and Darwinian natural selection. There are a lot of questions science has no answers for, but general scope of the life phenomenon problem becomes more or less clear now.

See also: Global Change Ecology: Biosphere: Vernadsky's Concept

Further Reading

Bonner, J.T., 1988. The Evolution of Complexity by Means of Natural Selection. Chicago: University of Chicago Press.
Brooks, D.R., Wiley, E.O., 1988. Evolution as Entropy: Toward a Unified Theory of Biology. Chicago: The University of Chicago Press.
Camazine, S., Deneubourg, J.L., Franks, N.R. (Eds.), 2001. Self-Organization in Biological Systems. Princeton, NJ: Princeton University Press.
Eigen, M., Schuster, P., 1979. The Hypercycle. A Principle of Natural Self-Organization. Berlin: Springer.
Jørgensen, S.E., Brown, M.T., Odum, H.T., 2004. Energy hierarchy and transformity in the universe. Ecological Modelling 178, 17–28.
Kauffman, S.A., 1993. The Origins of Order: Self-Organization and Selection in Evolution. Oxford: Oxford University Press.
Lotka, A.J., 1925. Elements of Physical Biology. Baltimore, MD: Williams and Wilkins.
Margulis, L., Sagan, D., 2000. What Is Life?: The Eternal Enigma. Princeton, NJ: University of California Press.
Odum, H.T., 1994. Ecological and General Systems: An Introduction to Systems Ecology. Niwot, CO: Colorado University Press.
Pahl-Wostl, C., 1995. The Dynamic Nature of Ecosystems: Chaos and Order Intertwined. New York: Wiley.
Pimm, S.L., 1991. The Balance of Nature? Chicago: University of Chicago Press.

Rossi, E., 1992. What is life: From quantum flux to the self. Psychological Perspectives 26, 6–22.
Rosen, R., 1967. Optimality Principles in Biology. New York: Plenum.
Rowe, G., 1994. Theoretical Models in Biology. New York: Springer.
Schrödinger, E., 1944. What Is Life? Cambridge: Cambridge University Press.
Seifert, J., 1997. *What Is Life? The Originality, Irreducibility*, and Value of Life. Amsterdam: Rodopi.
Sheldrake, A.R., 1981. A New Science of Life: The Hypothesis of Formative Causation. London: Blond and Briggs.
Svirezhev, Yu.M., 2000. Thermodynamics and ecology. Ecological Modelling 132, 11–22.
Ulanowich, R., 1986. Growth and Development: Ecosystems Phenomenology. New York: Springer.
von Bertalanfy, L., 1952. Problems of Life. New York: Wiley.

Phosphorus Cycle☆

Y Liu and J Chen, Tsinghua University, Beijing, China

Introduction

Phosphorus (P) is important because it is a nonsubstitutable element in sustaining all life and food production on our planet. It is one of the three macronutrients needed by all crops (together with N and K). Human activity has quadrupled the mobilization of phosphorus, and although the availability of this nonrenewable resource does not seem to pose a problem at the moment, there are other aspects of our phosphorus metabolism that do require our attention, namely, the wastes (water and soil sinks) and how we affect the biogeochemical cycle of phosphorus on Earth. Throughout the metabolism of phosphorus in our economy, there are large amounts of wastes and emissions as will be shown later. P sources coming from industry, farmland, animal feed, and household consumption are all main contributors to overnutrient water bodies, causing eutrophication. For the soil sink, P is accumulated in both agricultural and natural soils due to fertilizer application exceeding crop assimilation and due to dumped industrial, agricultural, and animal wastes, which slowly leach into the soil. Thus, a huge amount of P is immobilized in soils, which results in deterioration of farmlands and the inefficient use of P resource.

The Human-Intensified Phosphorus Cycles

Inorganic Cycle

Phosphorus circulates through the environment in three natural cycles. The first of these is the inorganic cycle, which refers to phosphorus in the crust of the Earth. In geologic time, phosphorus has moved slowly through the inorganic cycle, starting with the rocks, which slowly weather to form soil, from which the phosphorus is gradually leached from the land into rivers and onward to the sea, where it eventually forms insoluble calcium phosphate and buries in aquatic, primarily oceanic, sediments (Follmi, 1996). There it remains until it is converted to new, so-called sedimentary rocks as a result of geologic pressure. On a timescale of hundreds of millions of years, these sediments are uplifted to form new dryland and the rocks are subject to weathering, completing the global cycle (Schlesinger, 1991). In addition, some phosphorus can be transferred back from the ocean to the land by fish-eating birds whose droppings have built up sizable deposits of phosphate as guano on Pacific coastal regions and islands and by ocean currents that convey phosphorus from the seawater to these regions. A simplified schematic of the global phosphorus cycle is presented in **Figure 1**.

The global cycle of phosphorus is unique among the cycles of the major biogeochemical elements in having no significant gaseous compounds. The biospheric phosphorus flows have no atmospheric link from ocean to land. A little phosphorus does get into the atmosphere as dust or sea spray, accounting for 4.3 million metric tons of phosphorus per year (MMT P $year^{-1}$) and

☆*Change History:* November 2013. Y Liu and J Chen updated all parts of the text and references.

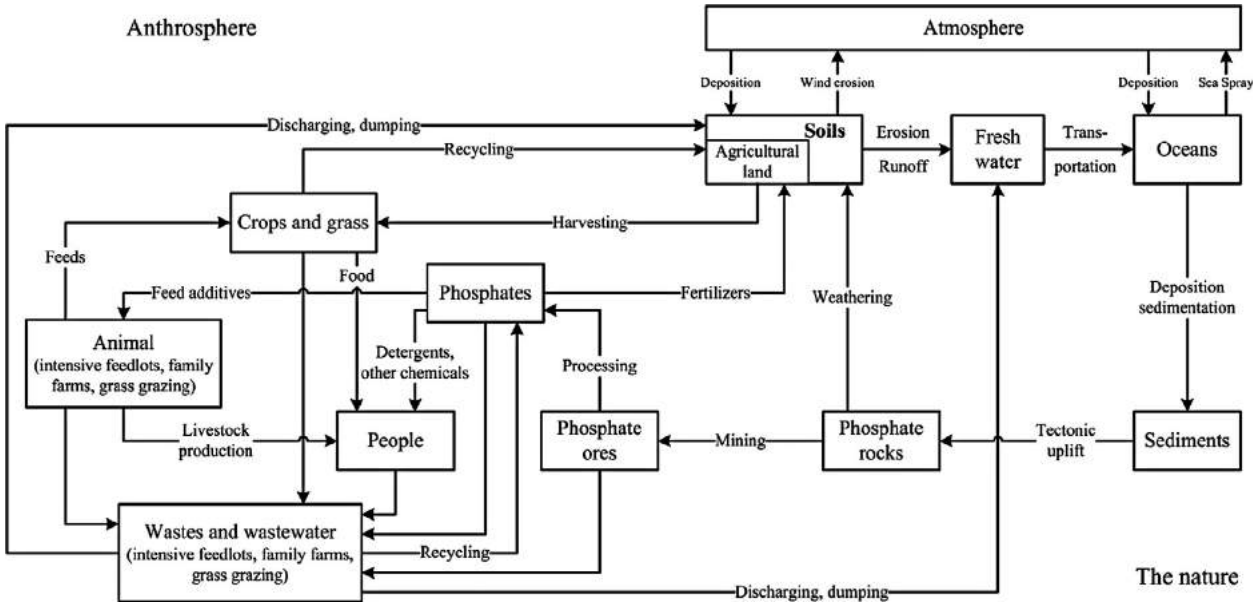


Figure 1 The human-intensified global phosphorus cycles. Reproduced from Liu et al. (2005) *Phosphorus Flows in China: Physical Profiles and Environmental Regulation*. PhD Thesis, Environmental Policy Group, Wageningen University, Wageningen, The Netherlands; Smil, V. (2000). Phosphorus in the environment: natural flows and human interferences. *Annual Review of Energy and the Environment* **25**, 53–88.

0.3 MMT P year^{-1}(Richey, 1983), respectively, but the amounts are several orders less important than other transfers in the global phosphorus cycle. The amount 4.6 MMT P year^{-1} of atmospheric phosphorus deposition, being balanced by the phosphorus carried by the wind and the sea spray, cannot offset the endless drain of this element from the land due to erosion and river transportation. Fortunately, increased anthropogenic mobilization of the element has no direct atmospheric consequences.

Nearly all the phosphorus on land is originally derived from the weathering of calcium phosphate minerals, especially apatite [$Ca_5(PO_4)_3OH$]. Around 13 MMT P year^{-1} of this is released to form soils each year (Emsley, 2000). However, this amount cannot offset the annual losses of phosphorus from the land. Taking into account all four forms of phosphorus (dissolved and particulate and organic and inorganic), the total amount of annual phosphorus losses from the lithosphere into freshwaters is estimated at 18.7–31.4 MMT P year^{-1} (Compton et al., 2000).

The uncertainty in the estimate is mainly due to a lack of knowledge on the biogeochemical processes of the particulate inorganic phosphorus, which constitutes the major component in the total loss. Not all of the eroding phosphorus can eventually reach the ocean. About 3.0 MMT P year^{-1} is carried away by wind into atmosphere, and at least 25% of that is redeposited on adjacent cropland and grassland or on more distant alluvia (Smil, 2000). Consequently, the amount of phosphorus transported by freshwaters into the ocean is probably in the range 12–21 MMT P year^{-1}. This result agrees with the most likely value of 17–22 MMT P year^{-1} given by some previous estimates (Emsley, 1980; Richey, 1983). The river-borne transport of phosphorus constitutes the main flux of the global phosphorus cycle. The loss, as a result of erosion, pollution, and fertilizer runoff, must be considerably higher than it was in prehuman times. It can be argued that the human-intensified phosphorus flux caused by wind and water erosion is at least two or even three times its prehistoric level (Compton et al., 2000; Smil, 2000).

Organic Cycles

Imposed on the inorganic cycle are two organic cycles that move phosphorus through living organisms as part of the food chain. These are a land-based phosphorus cycle that transfers it from soil to plants, to animals, and back to soil again and a water-based organic cycle that circulates it among the creatures living in rivers, lakes, and seas. The land-based cycle takes a year on average and the water-based cycle organic cycle only weeks. It is the amount of phosphorus in these two cycles that governs the biomass of living forms that land and sea can sustain.

The amount of phosphorus in the world's soils is roughly (90–200) $\times 10^3$ MMT P according to various estimates (Emsley, 1980; Filippelli, 2002). While the total phosphorus content of soils is large, only a small fraction is available to biota in most soils. This constitutes an available phosphorus pool containing 1805–3000 MMT P, most likely 2000–2600 MMT P (Emsley, 1980; Richey, 1983). A larger amount, in the range (27–840) $\times 10^6$ MMT P, can be found in the oceans. The surface seawater contains (80–3097) $\times 10^3$ MMT P and the rest is accumulated in deep seawater (Paytan and McLaughlin, 2007).

It is estimated that the oceanic residence time of dissolved P is between 20 and 100 kiloyears, although P is extensively cycled within the ocean on much shorter timescale. The ocean water loses phosphorus continually in a steady drizzle of detritus to the bottom, where it builds up in the sediments as insoluble calcium phosphate. Despite the geologic remobilization, there is a net annual loss of millions of tons of phosphate a year from the marine biosphere. Thus, the ocean sediments are by far the largest stock

in the biogeochemical cycles of phosphorus. Estimates of total P burial in open ocean marine sediments range from 2.8×10^3 to 10.5×10^3 MMT P per year, and the total marine sediments is estimated to be 840×10^3 MMT P(Paytan and McLaughlin, 2007). The majority of this burial flux is reactive P, with most of the nonreactive P having been deposited in the continental shelves.

Global Phosphate Consumption

The phosphate rock is initially converted to phosphoric acid (P_2O_5) by reaction with sulfuric acid. The phosphoric acid is further processed to produce fertilizers, food-grade and feed-grade additives, and detergents. Other marginal applications include metal surface treatment, corrosion inhibition, flame retardants, water treatment, and ceramic production. Despite such widespread use, the latter applications represented only ~3% of the total consumption of various phosphates in the 1990s (CEEP, 1997).

The global consumption of all phosphate fertilizers surpassed 1 MMT P year^{-1} during the late 1930s. After reaching 14 MMT P year^{-1} in 1980, the world consumption of phosphate fertilizers has been relatively stable. It was 14.8 MMT P (34 MMT P_2O_5) in statistical year 2002–03 and increased to 19.8 MMT P (45.4 MMT P_2O_5) in statistical year 2009–10, roughly accounting for 81.1% of the global extraction of phosphate rock (FAO, 2013). The top three economies, including China, the United States, and India, accounted for one-half of the world consumption. The area of the world current cropland is about 1.4 billion ha, implying that the global fertilizer application intensity averages 14 kg P ha^{-1}. The application rate varies significantly among continents, ranging from about 3 kg P ha^{-1} in Africa to over 28 kg P ha^{-1} in Asia. Annual P application in Western European countries was 24 kg P ha^{-1} in 1965, peaked at 34 kg P ha^{-1} in the 1980s, and then gradually decreased to 17 kg P ha^{-1} in 2007 (Sattari et al., 2012).

Crop Harvests

The use of phosphates to nourish agricultural soils aims to replenish the removal of phosphorus from soil by harvests and erosion losses. Adopting the average phosphorus contents in crops and the harvest index, the global crop production harvested 13.9 MMT P from soils in 2010 as shown in **Table 1**. A study of Chinese phosphorus flows suggested that the national harvest in 2004 removed 3.2 MMT P from croplands, based on a set of 'domestic' data for the phosphorus contents and the harvest index (Liu, 2005). These two estimates agree that (1) cereals accounted for a major part of the harvested phosphorus, that is, 70% at the global level and/or 68% in China, and (2) about two-thirds of the annually harvested phosphorus is contained in grains and the rest is contained in straw and other agricultural waste.

Since natural weathering and atmospheric deposition, as discussed elsewhere, cannot compensate the amount of phosphorus uptake from soils, application of phosphates, in both inorganic and organic forms, becomes essential to sustain today's harvests. There are several means of organic phosphorus reuse. The most direct means is to recycle crop residues *in situ*. Assuming roughly half of the annual output of crop residues (mostly cereal straw) is not removed from fields, the amount of the direct reuse of crop residues is about 2.4 MMT P year^{-1}.

Livestock and Animal Wastes

Animal wastes have been applied as organic manure in traditional farming and remain a relative large source of recyclable phosphorus in modern agriculture. According to the latest estimate from China, beef cattle, dairy cattle, swine, and poultry

Table 1 Allocation of phosphorus in world harvest in 2005

	Harvested crops			*Crop residues*		
	Fresh weight (MMT)	*Dry matter (MMT)*	*P in grains (MMT P)*	*Residues (MMT)*	*P in straws (MMT P)*	*Total P uptake (MMT P)*
Cereals	2474	2175	6.5	3256	3.2	9.7
Sugar crops	1917	595	0.6	462	0.9	1.5
Roots and tubers	750	150	0.1	230	0.2	0.3
Vegetables	1049	105	0.1	175	0.1	0.2
Fruit	738	111	0.1	184	0.1	0.3
Pulses	69	66	0.3	69	0.1	0.5
Oil crops	169	124	0.1	112	0.1	0.2
Other	56	45	0.1	112	0.1	0.2
Forages		500	1.0	0	0.0	1.0
Total	7196	3850	9.0	4601	4.9	13.9

Source: FAO (2013). *FAOSTAT*. Statistics Division, Food and Agriculture Organization of the United Nations; Smil, V. (1999). Crop residues: agriculture's largest harvest. *BioScience* **49**(4), 299–308; Liu et al. (2005).

produced 4.9 MMT P contained in animal manures in 2004 (Chen et al., 2008). The livestock population in China accounts for about 30% of the world total in 2004 and the proportion has remained fairly constant (FAO, 2013). On this basis, the global production of animal wastes would be 16.0 MMT P $year^{-1}$. However, the real figure may be somewhat larger, because the animals in China are not so well fed compared with developed countries. For this reason, the estimate of global production of 16–20 MMT P $year^{-1}$ in animal wastes, applying an average concentration of 0.8–1% of phosphorus for both confined and unconfined animal wastes, is probably more accurate.

Only the phosphorus in confined animal wastes is considered to be recyclable for croplands, while unconfined animal wastes mainly return to pastures. Assuming that animal biomass remains relatively constant, the amount of phosphorus in animal wastes is equal to the consumption of phosphorus contained in all kinds of feeds. In 2003, livestock consumed 36% of the harvested cereals (excluding the amount of cereals processed for beer), 21% of the harvested starchy roots, and 20% of the harvested pulses (FAO, 2013). Consequently, the annual livestock consumption of phosphorus in the harvests is accounted as about 2.4 MMT P $year^{-1}$.

Some part of crop residues is used as animal fodder. However, the reuse ratio of crop residues as fodder considerably varies globally. For instance, it was reported that the percentage of crop residues – mostly the straws of rice, wheat, and corn (maize) – used as fodder ranged from 3.6% in Shanghai to 42.8% in Gansu Province in 2000 in China, depending on crop and livestock species, farming and feeding traditions, and local economic profiles, and averaged 22.6% across the nation. Since over 70% of world livestock are raised in developing countries where commercial feeds are less used, the global recycling rate of crop straws as fodder is probably about 25%, leading to an absolute quantity of 1.2 MMT P $year^{-1}$.

Another major source of animal daily phosphorus intake is via feed additives. Around 6% of the global yield of phosphoric acid has been processed as animal feed-grade additives since 2000 (Brasnett, 2002). This constitutes an annual phosphorus flux of 1.0 MMT P $year^{-1}$ input to livestock husbandry.

Adding all the previously mentioned three sources, the global livestock consumption of phosphorus amounts to *c.* 4.6 MMT P $year^{-1}$. Taking into account the recycling of various industrial by-products and kitchen organic wastes (which is prevalent in rural family-based farms in developing countries), this figure could be as much as 20% higher, resulting in a total of 5.6 MMT P $year^{-1}$. Of course, the phosphorus flux to livestock of 5.6 MMT P $year^{-1}$ is mainly consumed by animals in confined facilities, while the world's cultivated and natural pastures provide a major source of phosphorus for unconfined animals. If one-half of the organic phosphorus in confined animal wastes is subject to recycling, animal manure is responsible for about 2.8 MMT P $year^{-1}$ returns to global croplands.

Food Consumption and Human Wastes

The third source of organic phosphorus available, in principle, for cropland is human waste. Assuming the world human body mass averaging 45 kg per capita (reflecting a higher proportion of children in the total population of low-income countries) and phosphorus content in human body averaging 470 g P per capita implies a global anthropomass contains approximately 3.0 MMT P. The typical daily consumption is about 1500 mg P per capita for adults (CEEP, 1997). This is well above the dietary reference intake (DRI), the amount human individual should take each day, as recommended by the Food and Nutrition Board, Institute of Medicine, US National Academy of Science. The US-recommended intakes are 700 mg per capita for adults over 18 years of age, 1250 mg per capita for young adults between 9 and 18 years of age, and 500 mg per capita for children.

A similar estimate for China is derived from a previous study: the individual daily intake of phosphorus was 1400 mg P per capita for urban residents and 1470 mg P per capita for rural residents in 2000. This exceeds the DRI of 1000 mg per capita recommended by the Chinese Nutrient Society. In addition, livestock products provided 30% of daily phosphorus intake for Chinese urban residents and 14% of that for rural population.

According to total protein intake and percentage of protein from vegetable sources, the mass phosphorus produced from human wastes annually on a capita basis in combined urine and feces ranges from 0.18 to 0.73 kg in 2009. Human excreta and urine contained about 3.4 and 1.7 MMT P $year^{-1}$, respectively. Urban and rural population generated 2.7 and 2.4 MMT P $year^{-1}$, respectively (Mihelcic et al., 2011).

Application of human excreta as organic fertilizer is common both in Asia and in Europe, but less prevalent elsewhere in the world. The nutrient linkage between farmers and croplands has been relatively stable, but the human wastes in urban areas are less recycled than in rural areas. For instance, less than 30% of human wastes in urban areas were recycled for agricultural uses in the late 1990s in China. This percentage dramatically decreased from 90% in 1980. In contrast, about 94% of human wastes in rural areas were returned to croplands in the 1990s. In European countries, the recycling rate of urban sewage averaged about 50% over the 1990s. Globally, it could be appropriate to assume that about 20% of urban human wastes and about 70% of rural human wastes are recycled at present. Therefore, recycled human wastes amount to 2.2 MMT P $year^{-1}$.

Adding the quantities of the phosphorus recycled as crop residues, animal manures, and human wastes, the total organic fertilizers applied to croplands amount to 6.4 MMT P $year^{-1}$. This is equivalent to 24% of the applied amount of inorganic fertilizers. Thus, the global input of phosphorus to croplands is probably 26.2 MMT P $year^{-1}$ in total or 1.9 times the amount of the phosphorus removed from the soil by harvesting. This leads to a net accumulation of 12.3 MMT P $year^{-1}$ or 7.9 kg P ha^{-1} in global soils, disregarding erosion and runoff losses.

Phosphates in Soil

The distribution, dynamics, and availability of phosphorus in soil are controlled by a combination of biological, chemical, and physical processes. These processes deserve special attention, as a considerable proportion of the applied phosphate is transformed into insoluble calcium, iron, or aluminum phosphates. On average, only a small proportion, perhaps 15–20% of the total amount of phosphorus in the plant, comes directly from the fertilizer applied to the crop. Readily available phosphorus in the soil solutions, however, provides most of the plant-available phosphorus. For most of the twentieth century, farmers in Western countries were advised to add more than double the amount of phosphate required by a crop, because these immobilized calcium, iron, and aluminum phosphates had been assumed to be permanently unavailable to plants.

Concentration of phosphate ions in the solution and the P-buffer capacity are the two main factors controlling the availability of soil phosphorus to plant roots. But the equilibrium concentration of phosphate present in soil solution is commonly very low, below 5 μmol (Condron and Tiessen, 2005). At any given time, soil water contains only about 1% of the phosphorus required to sustain normal plant growth for a season (Emsley, 2000). Thus, phosphates removed by plant and microbial uptake must be continually replenished from the inorganic, organic, and microbial phosphorus pools in the soil. These continuous processes dominate contemporary agricultural production to remove about 8.9 kg P ha^{-1} from cropland each year on a world average (based on our own estimate) and commonly 30 kg P ha^{-1} from the US and European fertile agricultural soils.

Each phosphate mineral has a characteristic solubility under defined conditions (Valsami-Jones, 2004). The solubility of many compounds is a function of acidity (pH; **Figure 2**). An increase in pH can release sediment-bound phosphorus by increasing the charge of iron and aluminum hydrous oxides and therefore increasing the competition between hydroxide and phosphate anions for sorption sites. Also, organic acids can inhibit the crystallization of aluminum and iron hydrous oxides, reducing the rate of phosphorus occlusion (Schlesinger, 1991). The production and release of oxalic acid by fungi explain their importance in maintaining and supplying phosphorus to plants.

The relative sizes of the sources and stocks of phosphate in soil change as a function of soil development (**Figure 3**). The buildup of organic phosphates in the soil is the most dramatic change. As time goes on, this becomes the chief reservoir of reserve phosphate in the soil. In most soils, organic phosphates range from 30% to 65% of total phosphorus, and it may account for as high as 90%, especially in tropical soils (Condron and Tiessen, 2005). The reasons for this are their insolubility and chemical stability. It has been noted that acid soils tend to accumulate more total organic phosphorus than do alkaline soils. This is almost certainly because organic phosphates react with iron and aluminum under acid conditions and become insoluble. Being the salts or metal complexes of phosphate esters, they release their phosphate by hydrolysis, but only very slowly. Phosphate esters can have half-lives of hydrolysis of hundreds of years. This process can be greatly speeded up by the action of phosphatase enzymes in the soil whose function is to facilitate reaction by catalyzing it. At later stages of soil development, phosphorus is progressively transformed into less-soluble iron- and aluminum-associated forms, and organic phosphorus contents of the soil decline. At this stage, almost all available phosphorus is found in a biogeochemical cycle in the upper soil profile, while phosphorus in lower depths is primarily involved in geochemical reactions with secondary sediments.

When the supply of dissolved phosphates to growing biomass is abundant, a net immobilization of inorganic phosphorus into organic forms will occur. Vice versa, inadequate inorganic phosphorus supply will stimulate the production of phosphatases and

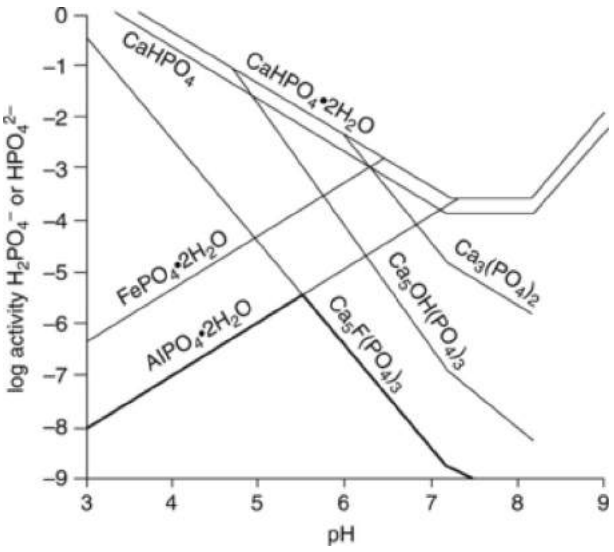


Figure 2 The solubility of phosphorus in the soil solution as a function of pH. Adapted from Schlesinger, W. H. (1991). *Biogeochemistry*: *an analysis of global change*. San Diego, CA: Academic Press.

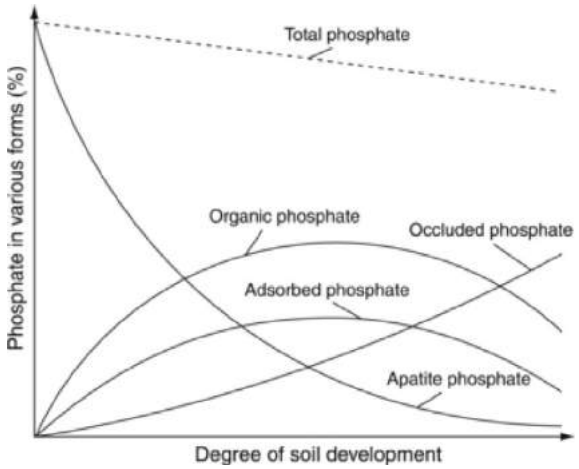


Figure 3 Phosphates in the soil vary with soil development. Adapted from Emsley, J. (1980). The phosphorus cycle. In: Hutzinger, O. (ed.) *The handbook of environmental chemistry*: *the natural environment and the biogeochemical cycles*, pp. 147–167. Heidelberg, NY: Springer.

the mineralization of labile organic forms of phosphorus for microbial uptake. A continuous drain on the soil phosphorus pools by cultivation and crop removal will rapidly deplete both labile inorganic and organic phosphorus in soils.

Allowing soil reserves of readily available phosphorus to fall below a critical value, determined by field experiments, can result in a loss of yield. The turnover of available phosphates by plants in soil solution is determined by rates of releases of phosphorus from insoluble forms to soluble phosphates. All kinds of soil particles can contribute to this process, and in some cases, it is not only the chemical balance that maintains the supply but also the action of microbes and enzymes that releases phosphate from organic debris in the soil. It is believed that the biogeochemical control of phosphorus availability by symbiotic fungi is a precursor to the successful establishment of plants on land.

However, our existing knowledge – briefly discussed earlier – cannot yet provide a comprehensive understanding of the complex movements and transformations of phosphorus, especially of its organic forms (Turner et al., 2005). This hampers the efficient application of phosphate fertilizers and the efficient control of phosphorus losses.

Phosphorus Losses

Phosphorus is lost from croplands via erosion or runoff. Quantifying phosphorus losses in eroding agricultural soils is particularly uncertain, as erosion rates vary widely even within a single field. It is also because few nations have comprehensive, periodic inventories of their soil erosion.

The phosphorus loss from croplands can be roughly estimated based on the amount of topsoil erosion and average phosphorus content. A crop takes up the majority of the nutrients it requires from topsoil. The topsoil is often identified as the 'plow layer,' that is, the 20–30 cm depth of soil, which is turned over before seedbed preparation. The volume of topsoil in the plow layer is around 2500 m^3 $ha^{-1}$and weighs approximately 2000 t (Johnston and Steen, 2000). A ton of fertile topsoil can contain 0.6–3.0 kg of phosphorus, based mainly on US and European agricultural practices.

At a global scale, 20–30 MMT P $year^{-1}$ is lost via erosion, and this would be equivalent to an annual loss of 15–20 kg P ha^{-1}, assuming that erosion from land other than arable land is negligible. Estimates of the amount of soil lost via erosion range from 5×10^3 to 40×10^3 kg ha^{-1} $year^{-1}$ for an average European arable soil to 10×10^3 kg ha^{-1} $year^{-1}$ at most for the major part of Europe. As the phosphorus concentration in an average European soil range between 0.05% and 0.2%, erosion of 10×10^3 kg ha^{-1} $year^{-1}$ of soil represents an annual loss of 5–10 kg P ha^{-1}. Annual soil erosion from agricultural systems of the United States, China, and India was estimated to be 3.0, 5.5, and 6.6 Gt $year^{-1}$, respectively. Some of the most serious soil erosion takes place in the agricultural systems of Southeast Asia, Africa, and South America. Hence, the real global soil erosion loss could be even higher, perhaps as much as 75 Gt $year^{-1}$.

The erosion intensity from croplands varies a lot among countries, ranging from 0.5 to 400 t ha^{-1} $year^{-1}$. Worldwide, soil erosion is highest in Asia, Africa, and South America, averaging 30–40 t ha^{-1} $year^{-1}$ of soil loss (Pimentel, 2006). It was suggested that the global average erosion rate is at least 20 t ha^{-1} $year^{-1}$. The lowest erosion rates occur in the United States and Europe where they average about 10 t ha^{-1} each year. It is evident that soil erosion in the United States has been reduced by soil conservation policies; a national survey showed that the total soil erosion between 1982 and 1992 decreased by 32%. The annual sheet and rill erosion rate in the United States fell from an average of 10 t ha^{-1} in 1982 to 7.7 t ha^{-1} in 1992, and the wind erosion rate fell from an average of 8.1–5.9 t ha^{-1} $year^{-1}$ over the same period. Assuming global erosion rate averaging 25 t ha^{-1}, the soil loss from crop land was estimated to be 38.5 Gt $year^{-1}$ from cropland (cf. **Table 1**). Furthermore, most of the loss is permanent and may not be replenished by weathering. For instance, the excessive soil loss, a rate that would impair long-term crop productivity, is estimated at about 25.4 Gt $year^{-1}$ in around 1980.

Table 2 Global soil erosion and phosphorus losses from agricultural land (2003)

	Permanent pasture		
	Cropland	*Overgrazed*	*Ordinary*
Total area (million ha)	1541	1720	1720
Soil erosion			
Erosion rate (t ha^{-1} $year^{-1}$)	25	15	5
Erosion quantity (Gt $year^{-1}$)	38.5	25.8	8.6
Phosphorus loss			
P content in topsoil (kg P ha^{-1})	0.5	0.5	0.5
P loss (MMT P $year^{-1}$)	19.3	12.9	4.3

Erosion from pastures is commonly less intensive than that from plowed fields. However, soil losses have been greatly increased by overgrazing, which now affects more than half (i.e., at least 1720×10^6 ha) of the world's permanent pastures with a high erosion rate of 15 t ha^{-1} each year. This leads to 25.8 Gt $year^{-1}$ of the soil loss from overgrazed pastures. Together with the amount of soil loss from cultivated grassland, the world's permanent pastures lose their topsoil at an annual rate of 34.4 Gt $year^{-1}$. Adding the losses from cropland and pastures, the world soil erosion from agricultural areas amounts to 72.9 Gt $year^{-1}$ in total, or 15 t ha^{-1} $year^{-1}$ on average. This is similar to previous estimates as discussed earlier. Allowing for the poor condition of topsoil in developing countries, it might be appropriate to assume that the global phosphorus content in topsoil averages about 0.5 kg P t^{-1}, or 1.0 t P ha^{-1}. This gives the world phosphorus losses at 19.3 and 17.2 MMT P $year^{-1}$ from cropland and pastures, respectively, as shown in Table 2.

The surface runoff loss of applied inorganic phosphate fertilizer varies significantly with a number of agronomic factors. Typical runoff rates of phosphorus in European countries range from 0.2% to 6.7%, an average of 3.5% (Hart et al., 2004). Worldwide, the maximum rate can reach 10% under certain soil characteristics and climatic condition. Roughly, the world total phosphate fertilizer application can lead to a loss of 0.5 MMT P $year^{-1}$ in surface runoff.

Phosphate Balance in Cropland

The national phosphorus balance varies significantly from one country to another, due to differences in the use of mineral fertilizers and manure and differences in animal husbandry practices. Broadly speaking, in developing countries, soils tend to be deficient in phosphorus, while in developed countries, the phosphorus content of the soils is adequate or even excessive. Taking into account applications of mineral fertilizers and manure, the balance for some West European countries is positive, particularly in the Netherlands where it exceeds 39 kg P ha^{-1} each year. For the majority of Western European countries, the phosphorus balance ranges from 8.7 to 17.5 kg P ha^{-1} annually. China, one of the largest agricultural systems in transition, also achieved a positive balance around 1980 at the national level, in parallel with increasing application of synthetic fertilizers. In 2004, the national balances of phosphorus in Chinese soils for arable land and grassland were estimated at an average of 59 kg P ha^{-1} (surplus) and −5 kg P ha^{-1} (deficit), respectively (Chen et al., 2008).

The world phosphorus budget for cropland is summarized in Table 3. To balance the phosphorus budget for the world's cropland, two natural inflows of phosphorus to croplands should be taken into account. Based on the ratio of cropland area to world total land areas, weathering and atmospheric deposition contribute 2.0 MMT P $year^{-1}$ as inputs to world croplands. Although the magnitudes of recycling of animal wastes and soil erosion need further verification, the budget provides a comprehensive overview on the global phosphorus flows associated with the farming sector, which is the most intensive and complicated subsystem of the anthropogenic phosphorus cycle.

Although there is a global phosphorus deficit for cropland, in regions with developed agriculture, residual soil phosphorus is very high, and soil phosphorus status has been improved over the past decades by applying fertilizer and manure. Cumulative inputs of phosphorus fertilizer and manure in Western Europe, Asia, North America, and Oceania for the period 1965–2007 amounted to 1115, 700, 500, and 560 kg P ha^{-1}, respectively. Since the 1980s, phosphorus application rates have declined in many European countries, but uptake continued to increase. This finding is possibly due to the continued supply of plant-available phosphorus from the residual soil phosphorus pool.

Ecological Impacts of Phosphorus Use

The phosphorus-related ecological issues fall into a broad range. Some of them are caused by inappropriate use of the material, some are not. Eutrophication, being regarded as the most immediate environmental consequences of extensive phosphorus usage in contemporary societies, has received wide attention. However, it is not the whole story, and other issues deserve to be taken into consideration.

Table 3 Phosphorus budget for the world's cropland in recent year (2004)

Flows	*Annual fluxes (MMT P)*
Inputs	29.2
Weathering	1.6
Atmospheric deposition	0.4
Synthetic fertilizers	19.8
Organic recycling	7.4
Crop residues	2.4
Animal wastes	2.8
Human wastes	2.2
Removals	13.9
Crops	9.0
Crop residues	4.9
Losses	19.8
Erosion	19.3
Runoff	0.5
Balance	−4.5
Input shares	
Fertilizer application	68%
Organic recycling	25%
Uptake efficiency	48%

Mineral Conservation

Estimate of world phosphorus reserves varies from 15 to 1200×10^3 MMT. But the most credible data come from the US Geological Survey, which gave estimates of 19.1×10^3 for reserve and 71×10^3 for reserve base in 2010. Regardless, it is widely accepted that phosphorus is nonrenewable within the timescale of human society, and it is becoming increasingly scarce and expensive. Peak phosphorus was estimated to occur by 2035, after which demand would outstrip supply (Cordell et al., 2009; Filippelli, 2008, 2011). Already in 1972, the Institute of Ecology reported that phosphorus could be depleted before the end of the twenty-first century. Present estimates concluded that world-known phosphorus would be depleted in the next 50–150 years; otherwise, additional resource could be found. Moreover, it has been projected that the utilization trend is unlikely to decline in next 20 years. It will instead probably increase at the rate of 0.7–1.3% annually (van Vuuren et al., 2010). This strongly suggests that phosphate rock, as a finite nonrenewable resource, may be exhausted in a much shorter time. It has been shown that the global average phosphorus content in raw ores dropped to 29.5% in 1996 from 32.7% in 1980 and that global reserves can sustain the current mining intensity for only another 80 years. Some phosphorus-rich deposits around the world can be exploited much sooner. China's phosphorus reserves, for instance, constitute 26% of the world's total reserve base, second only to Morocco and the Western Sahara. With a highly intensive extraction activity as well as losses incurred during mining, the basic reserve of the nation's phosphorus resources, that is, 4054 million tons with average P_2O_5 content of 17–22%, could be exhausted in 50–70 years. Certainly, the larger reserve base and probably more reserves to be discovered in the future guarantee a longer life span of the extraction. Even so, the deposits of phosphorus in the lithosphere will inevitably be depleted before new igneous or sedimentary rocks can be formed via the biogeochemical process at the timescale of millions of years.

Soil Erosion

One of the most important results derived from the phosphorus budget (cf. **Table 3**) is that the world cropland has lost phosphorus at a surprising rate of 10.5 MMT P $year^{-1}$. This massive loss from croplands is mainly caused by wind and water erosion of topsoil. Soil erosion has been recognized as one of the most serious environmental crisis the world is suffering from. It is estimated that 10 million ha of cropland is abandoned each year worldwide due to lack of productivity caused by the soil erosion (Pimentel, 2006). Nearly 60% of present soil erosion is induced by human activity, increasing by 17% since the early 1900s.

In contrast to the erosion loss, a huge amount of phosphorus has been mobilized in cultivated soils. Contemporary scientific knowledge cannot fully explain the complex transportation of phosphorus between plant roots, soil waters, and soil particulates. More complete understanding of these processes might suggest a possibility of controllable remobilization of soil phosphorus that would benefit the environment via reducing both the input of fertilizers and the loss of phosphorus from soils.

Animal Wastes

Livestock husbandry, in particular large intensive feedlots, has become a major problem both for recycling of organic phosphorus and for emission of phosphorus pollutants. Worldwide, the structure of animal agriculture has changed as livestock are concentrated in fewer but larger operations. In the United States, in spite of losing nearly a fourth of the livestock operations between 1982 and 1997,

the total number of animal units (an equivalent that converts various kinds of animals into cattle based on individual nutrient excretion) has remained fairly constant at *c.* 91–95 million. In China, a large number of intensive feedlots appeared in suburbs and rural areas during the last decade. According to a national investigation, the output of hogs, meat chicken (broilers), and egg chicken (layers) produced by intensive feedlots and farms accounted for 23%, 48%, and 44% of the national total in 1999, respectively.

As livestock operations have become fewer, larger, and more spatially concentrated in specific areas, animal wastes have also become more concentrated in those regions. This leads to a considerable phosphorus surplus in manure, as the amount of manure nutrients relative to the assimilative capacity of land available on farms for application has grown, especially in specific high-production areas. Consequently, off-farm manure export requirements are increasing.

But because of its bulk, uneven distribution, and prohibitive cost of transport beyond a limited radius, a large proportion of manure phosphorus is now subject to disposal instead of recycling. If construction of necessary infrastructures for appropriate disposal of manure lags behind, animal wastes become a major source of phosphorus loads in surface waters. Uncontrolled phosphorus emission from intensive feedlots and farms in China has escalated in parallel with the gradual growth in total animal feeding operations and the rapid shift in breeding structure. The emission of China's livestock was estimated at 39% of its national phosphorus load to aquatic environments in 2004. Thus, livestock husbandry is the most significant source of phosphorus flux to surface waters in China, similar to the situation in European countries in the early 1990s.

Sewage Treatment

In the 1960s, many developed countries began to alleviate the pollution in surface water by constructing municipal sewage infrastructures and implementing phosphorus discharge restrictions on production sectors. The giant infrastructure of centralized wastewater treatment has drastically reshaped the phosphorus cycle within modern cities. Despite high economic costs, its environmental benefits with regard to removal of phosphorus from wastewater are far less than satisfactory worldwide. However, some progress has been achieved in European countries and the United States (Litke, 1999). As the centralized control strategy just removes 'pollutants' into sewage sludge rather than promotes a recovery and recycling of resources, including phosphorus, it does not really solve the long-term ecological problem. The costly and rigid infrastructures have significantly reduced agricultural reuse of urban human excreta and contributed to a disconnect of nutrient cycles between urban areas and croplands. Unfortunately, no available technologies for stable recovery and recycling of phosphorus are likely to be successfully commercialized in the near future. Hence, most of the phosphorus in urban human wastes is not subject to efficient recycling and is permanently lost from the land.

Proposals for recovery of phosphorus via decentralized source-separated strategies have received increasing attention since the mid-1990s. This decentralized and downsized sanitation concept, focused on ecologically sustainable and economically feasible closed-loop systems rather than on expensive end-of-pipe technologies, advances a new philosophy. It departs from the one-way flow of excreta from terrestrial to aquatic environments, as introduced by the conventional flush-and-discharge sewage system. The new alternative separates nutrients and domestic used water at source and handles both components individually based on material flows approaches. Thus, it avoids the disadvantages of conventional wastewater solutions and enables and facilitates nutrient recycling. Although the reinvention and transition of urban wastewater systems poses a major challenge, it does provide a promising prospect for future phosphorus recovery and recycling in an ecological and economic efficient way. (Detailed studies are essential as a first step, inter alia, of technological, organizational, economic, and social aspects. In addition, the involvement of multistakeholders, such as residents, building owners, farmers, politicians, officials, and other interested parties, from the start seems essential. All these problems cannot be solved overnight, as it requires nothing less than a paradigmatic change of a large sociotechnical system.)

Detergent Use

The use of sodium tripolyphosphate ($Na_5P_3O_{10}$, STPP), the most widely used detergent additive, has been identified as a significant contributor to eutrophication. STPP was first introduced in the United States in 1946 (Emsley, 2000). After reaching a peak in the 1960s, global production has finally fallen down to one-half of the peak level, *c.* 1.0 MMT P $year^{-1}$, mainly due to bans on phosphorus-containing detergents in developed countries. The total quantity of STPP production was estimated at 0.865 MMT P in 2004. In the late 1990s, phosphorus-free detergents accounted for 45%, 97%, and 100%, respectively, in the United States, Japan, and the European countries (Litke, 1999).

There has been a controversy on the environmental impacts of STPP since the mid-1980s. Today, it is acknowledged that limiting or banning household consumption of phosphorus-containing detergents would not lead to a significant or a perceivable improvement of eutrophication. It would have little impact on water and human health compared to other substitute chemicals (sodium carbonates, sodium silicates or zeolites A, and sodium nitrilotriacetate), from both an ecological and an economic perspective. In parallel with these discoveries, some Nordic countries ecolabeled STPP as an environmentally friendly component of detergents in 1997 and have repromoted the production and consumption of STPP since then.

Eutrophication

Eutrophication is an unwanted explosion of living aquatic-based organisms in lakes and estuaries that results in oxygen depletion that can destroy an aquatic ecosystem. It has been regarded as the most important environmental problem caused by phosphorus

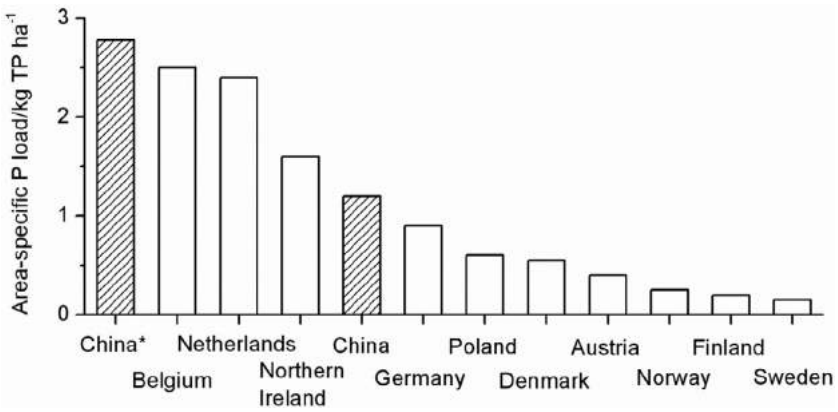


Figure 4 Comparison of phosphorus loads to aquatic environments by country (unit: kg P ha^{-1}). Reproduced from EEA (2005). *Source apportionment of nitrogen and phosphorus inputs into the aquatic environment*, No.7/2005. Copenhagen: European Environment Agency; Chen, M., Chen, J., Sun, F. (2008). Agricultural phosphorus flow and its environmental impacts in China. *Science of the Total Environment* **405**(1), 140–152.

losses. Significant eutrophication took place in the 1950s in the Great Lakes of North America and has been prevalent in many lakes and estuaries around the world. Phosphorus is often the limiting factor responsible for eutrophication, since nitrogen fluxes to water bodies are relative large.

Phosphorus losses from industries, croplands, animal farms, and households constitute the main sources. Figure 4 illustrates a cross-country comparison of the phosphorus loads in European countries and in China to their domestic aquatic environments.

The results show that the phosphorus loads range from 0.2 kg P ha^{-1} in Sweden to 2.5 kg P ha^{-1} in Belgium at the national level. China lies between Germany and Northern Ireland in turns of the load per unit land area. At the basin level, the comparison of phosphorus loads shows a similar figure. These results suggest that the Chinese economy is in general processing phosphorus 'wastes' less efficiently than developed countries. However, regarding phosphorus control strategy, it is nearly impossible to determine a common benchmark (of a desired phosphorus load) to prevent eutrophication in water bodies. This is because the complex interrelations between the amount of aquatic biomass and the phosphorus load are affected by a number of hydrological, meteorological, and biochemical factors that remain unclear under current knowledge. Further improvement of our current understanding of phosphorus movement and transformation within aquatic and territorial ecology, across socioeconomic and ecological boundaries, is desired.

Regulating the Societal Phosphorus Flows

Phosphorus (P), intensively extracted from the natural sink in the lithosphere and processed through various production–consumption cycles, ultimately deposits in soil or reaches the water body by different pathways. The anthropogenic P flows are characterized by complicated physical interconnections in high intensities among a number of production and consumption sectors. Hence, it is vitally important to connect the P flows with environmental regulations introduced to intervene in social practices and human behaviors.

Instead of continuously trying to limit the growth of P (like the bans of detergent P), there is a great need to reconstruct the physical structure of P flows, in particular by redirecting the crucial P flows with highly negative environmental impacts. The ecological restructuring of the current one-through mode of societal P metabolism is thus desired, leading to a structural shift in the societal production and consumption of P flows. The ecologically rational switch can contribute to a substantial decline of P outflows by minimizing P input and maximizing P recycling. Since ecologizing the P flows only succeeds when measures are institutionalized into the economy and society as a whole, this process will most likely be a gradual one rather than a radical revolution.

References

Brasnett R (2002) Feed phosphates: Their role in animal feeding and prospects for demand growth. In: *Paper Presented at 2003 Fertilizer Outlook Conference, 14–14 November, Arlington, VA, USA*.

CEEP (Center European d'Etudes des Polyphosphates) *Phosphate*. (1997) Brussels: CEEP.

Chen M, Chen J, and Sun F (2008) Agricultural phosphorus flow and its environmental impacts in China. *Science of the Total Environment* 405(1): 140–152.

Compton JS, Mallinson DJ, Glenn CR, et al. (2000) Variations in the global phosphorus cycle. In: Glenn CR (ed.) *Marine Authigenesis: From Global to Microbial*, pp. 21–33. Tulsa, OK: SEPM.

Condron LM and Tiessen H (2005) Interactions of organic phosphorus in terrestrial ecosystems. In: Turner BL, Frossard E, and Baldwin DS (eds.) *Organic Phosphorus in the Environment* Wallingford, UK: CAB International.

Cordell D, Drangert JO, and White S (2009) The story of phosphorus: Global food security and food for thought. *Global Environmental Change* 19(2): 292–305.
EEA (2005) *Source Apportionment of Nitrogen and Phosphorus Inputs into the Aquatic Environment*. Copenhagen: European Environment Agency, No.7/2005.
Emsley J (1980) The phosphorus cycle. In: Hutzinger O (ed.) *The Handbook of Environmental Chemistry: The Natural Environment and the Biogeochemical Cycles*, pp. 147–167. Heidelberg, NY: Springer.
Emsley J (2000) *The Shocking History of Phosphorus*. London: Macmillan.
FAO (2013) *FAOSTAT*. Statistics Division, Food and Agriculture Organization of the United Nations.
Filippelli GM (2002) The global phosphorus cycle. *Reviews in Mineralogy and Geochemistry* 48: 391–425.
Filippelli GM (2008) The global phosphorus cycle: Past, present, and future. *Elements* 4(2): 89–95.
Filippelli GM (2011) Phosphate rock formation and marine phosphorus geochemistry: The deep time perspective. *Chemosphere* 84(6): 759–766.
Follmi KB (1996) The phosphorus cycle, phosphogenesis and marine phosphate-rich deposits. *Earth-Science Reviews* 40(1–2): 55–124.
Hart MR, Quin BF, and Nguyen ML (2004) Phosphorus runoff from agricultural land and direct fertilizer effects: A review. *Journal of Environ Quality* 33(6): 1954–1972.
Johnston AE and Steen J (2000) *Understanding Phosphorus and its Use in Agriculture*. Brussels, Belgium: EEMA.
Litke DW (1999). Review of phosphorus control measures in the United States and their effects on water quality. *Water–Resources Investigations Report 99–4007*. Denver, CO: US Geological Survey.
Liu Y (2005). *Phosphorus Flows in China*: *Physical Profiles and Environmental Regulation*. PhD Thesis, Environmental Policy Group, Wageningen University, Wageningen, The Netherlands.
Mihelcic JR, Fry LM, and Shaw R (2011) Global potential of phosphorus recovery from human urine and feces. *Chemosphere* 84(6): 832–839.
Paytan A and McLaughlin K (2007) The oceanic phosphorus cycle. *Chemical Reviews* 107(2): 563–576.
Pimentel D (2006) Soil erosion: A food and environmental threat. *Environment, Development and Sustainability* 8: 119–137.
Richey JE (1983) The phosphorus cycle. In: Bolin B and Cook RB (eds.) *The Major Biogeochemical Cycles and their Interactions*, pp. 51–56. New York: Wiley.
Sattari SZ, Bouwman AF, Giller KE, and van Ittersum MK (2012) Residual soil phosphorus as the missing piece in the global phosphorus crisis puzzle. *Proceedings of the National Academy of Sciences* 109(16): 6348–6353.
Schlesinger WH (1991) *Biogeochemistry: An Analysis of Global Change*. San Diego, CA: Academic Press.
Smil V (1999) Crop residues: agriculture's largest harvest. *BioScience* 49(4): 299–308.
Smil V (2000) Phosphorus in the environment: Natural flows and human interferences. *Annual Review of Energy and the Environment* 25: 53–88.
Turner BL, Frossard E, and Baldwin DS (2005) *Organic phosphorus in the environment*. Wallingford: CABI Publishing.
Valsami-Jones E *Phosphorus in Environmental Technology: Principles and Applications*. (2004) *Integrated Environmental Technology Series*. Cornwall: IWA Publishing.
Van Vuuren DP, Bouwman AF, and Beusen AHW (2010) Phosphorus demand for the 1970–2100 period: A scenario analysis of resource depletion. *Global Environmental Change* 20(3): 428–439.

Sulfur Cycle

PA Loka Bharathi, National Institute of Oceanography, Panaji, India

Sulfur Cycle

Most elemental cycles are operative in both oxidative and reductive mode, each fueling the other, either in a dynamic instantaneous manner in space or sequentially over time. Sulfur and its species are important geochemical agents. While the element sulfur is the fourteenth most abundant element on Earth, sulfate ion is the second most abundant ion next only to chloride in seawater and carbonate in freshwater. Elemental sulfur is produced hydrothermally and also by oxidation of sulfide by weathering. The element is also formed as an intermediate of sulfide oxidation or sulfate reduction. Sulfides exist in a variety of forms, most of which are solids. However, dissolved sulfide can occur as bisulfide (HS^-) at neutral pH, sulfide ions (S^{2-}) at alkaline pH, and H_2S at acidic pH, which is volatile and has a rotten egg smell.

Sulfur transformations govern the compositions of the oceans, and the redox balance on the Earth's surface. It is complex due to a variety of oxidation states. Besides, some transformations occur at significant rates both bacteriologically as well as chemically.

The sulfur cycle involves eight electron oxidation/reduction reactions between the most reduced H_2S (-2) to the most oxidized SO_4^{2-} ($+6$). It acts as either electron donor or acceptor in many bacterially mediated reactions. The oxidation states of key sulfur compounds are given in the following table:

Organic	S (R–SH)	−2
Sulfide	(H_2S)	−2
Elemental S	(S^0)	0
Thiosulfate	($S_2O_3^{2-}$)	+2 (av./S)
Tetrathionate	($S_4O_6^{2-}$)	+2 (av./S)
Sulfur dioxide	(SO_2)	+4
Sulfite	(SO_3^{2-})	+4
Sulfur trioxide	(SO_3)	+6
Sulfate	(SO_4^{2-})	+8

Bacteria can mediate these oxidations. While bacterial oxidation of sulfur at the expense of oxygen or nitrate generally leads to chemosynthetic carbon fixation, the reductive S cycle is respiratory.

Sulfur Oxidation

Although a variety of oxidation states exist, only three forms are important, namely the sulfhydryl and elemental form besides the sulfate radical.

Reduced sulfur compounds can be used either by colorless sulfur bacteria or the colored photosynthetic bacteria. These bacteria notably belong to β purple bacteria group, namely the colored *Chromatium* sp. or the colorless *Thiobacillus* sp. The others include *Thiosphaera*, *Thiomicrospira*, *Thermothrix*, *Beggiatoa*, and the archaean *Sulfolobulus*. The final product sulfate and different amounts of energy are available depending on the oxidation state of sulfur used as the electron donor:

$$H_2S + 2O_2 \rightarrow SO_4^{2-} + 2H^+ \quad -798.2\ \text{kJ}$$

$$HS^- + 1/2O_2 + H^+ \rightarrow S^0 + H_2O \quad -209.4\ \text{kJ}$$

$$S^0 + H_2O + 11/2O_2 \rightarrow SO_4^{2-} + 2H^+ \quad -587.1\ \text{kJ}$$

$$S_2O_3^{2-} + H_2O + 2O_2 \rightarrow SO_4^{2-} + 2H^+ \quad -822.6\ \text{kJ}$$

Those forms that can oxidize sulfur under acid conditions are also able to oxidize iron. Yet others grow at neutral pH.

$$S_2O_3^{2-} \rightarrow S_4O_6^{2-} + 2e^-$$

The oxidation of sulfur involves the reaction of sulfhydryl groups of the cell, like glutathione, with the formation of sulfide–sulfhydryl complex. The enzyme sulfide oxidase oxidizes the sulfide to sulfite.

Chemosynthetic Sulfur Oxidation

The chemotrophic pathways are involved in the oxidation of reduced sulfur compounds, H_2S, S, and $S_2O_3^{2-}$. The oxidation of $S_2O_3^{2-}$ to tetrathionates $S_4O_6^{2-}$, trithionate $S_3O_6^{2-}$, pentathionate $S_5O_6^{2-}$, and elemental S^0 depends on environmental factors like

 doi:10.1016/B978-0-444-63768-0.00761-7

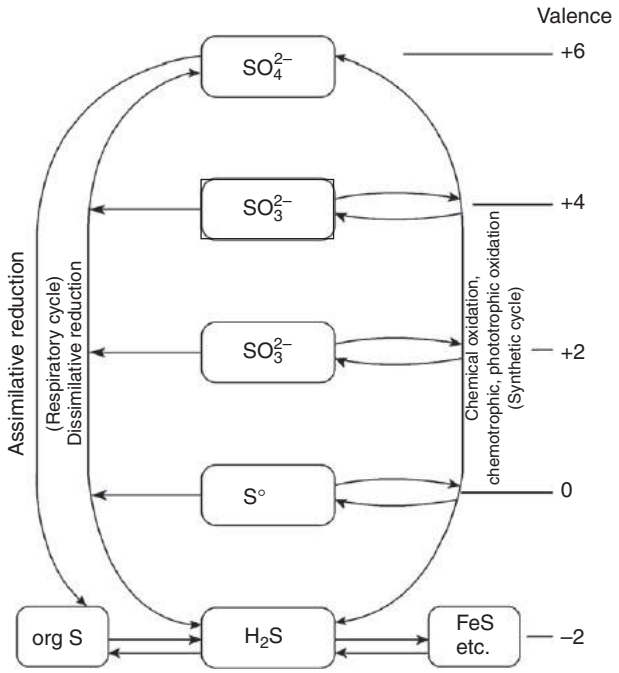


Fig. 1 The microbial sulfur cycle. Modified from Fenchel T and Blackburn TH (1979) *Bacteria and Mineral Cycling*. London: Academic Press.

oxygen and pH. The trithionate and pentathionate are formed from tetrathionate with sulfite and thiosulfite, respectively:

$$S_4O_6^{2-} + SO_3^{2-} \leftrightarrow S_3O_6^{2-} + S_2O_3^{2-}$$

$$S_4O_6^{2-} + S_2O_3^{2-} \leftrightarrow S_5O_6^{2-} + SO_3^{2-}$$

The chemotrophic bacteria have to compete with the spontaneous oxidation of sulfide. However, this chemical oxidation rates are speeded up by bacterial intervention – for example, bacterial oxidation of sulfide ores by thiobacilli. Thiobacilli-like bacteria play an important role in thiosulfate oxidation and sulfide oxidation, sometimes at the expense of nitrate as in *Thiobacillus denitrificans*:

$$5S_2O_3^{2-} + 8NO_3^- + H_2O \leftrightarrow 10SO_4^{2-} + 4N_2 + 2H^+$$

Sulfide oxidation coupled to nitrate reduction could be an important process in some coastal ecosystems where sediment-produced sulfide encounters nitrate-rich overlying waters that have been depleted in oxygen.

In other ecosystems like deep-sea hydrothermal vents, sulfur/sulfide oxidation is one of the main processes that bacteria utilize for chemosynthetic production of organic matter. Gigantic tubeworms in and around the vent fields harbor symbiotic sulfide-oxidizing bacteria. Special hemoglobins that bind H_2S as well as O_2 transport both substrates to the trophosome where they are released to the bacterial symbiont, thus preventing sulfide poisoning of the host.

Sulfide concentration in some vents measured as the sum of H_2S, HS^-, and S_2 can be related to prevailing temperature, but where biological uptake is rapid, the relationship can become nonlinear. The proportion of these species is strictly determined by pH of the surrounding. At the near-neutral pH from 7 to 7.9 of low-temperature fluids, HS^- is the prevailing species. Sulfide exposed to oxygen could be inorganically oxidized with a half-life of *c.* 380 h at 2 °C, and pH 7.8 and 110μm O_2 and 10μm H_2S. Biological oxidation of sulfide by macroorganisms and microorganisms at Galapagos vents could be 4–5 orders of magnitude greater than spontaneous sulfide oxidation in the laboratory. Generally in nature, the transition zones where the anaerobic zone meet the aerobic, the sulfide-oxidizing bacteria can form a sumptuous source of food to protozoans, microzooplanktons, and other higher forms of life. The oxidation of reduced sulfur compounds generates organic compounds (CH_2O) from inorganic substrates and is akin to primary production. In the marine microbial sulfur cycle, there is no net gain of organic material (Fig. 1). This is because organic material must be oxidized to generate the sulfide that is required for chemosynthetic production of organic carbon. However, at geothermal vents sulfide is released from the geochemical interaction of seawater and hot rock deep within the Earth crust. Under these conditions there is a net gain of organic material through the oxidation of sulfide and production of new biomass. The actual biochemical transformations are complex with light or chemical energy used to generate reducing power like NADH that is coupled to CO_2 fixation generally through Calvin–Benson cycle. Though chemosynthesis was described more than a century ago by Winogradsky, it was only with the discovery of hydrothermal vents that its significant quantitative role got established.

Geochemical Implications of Sulfur Oxidation

The sulfur-oxidizing microbes act as competitors and sinks for inorganic sulfur along with other reduced compounds as producers of organic biomass as food for zooplankton and a variety of benthic organisms.

Just as dissolved sulfides support chemosynthetic production, particulate sulfides too can support autotrophic growth. Metal sulfides, including pyrite, FeS_2, pyrrhotite, Fe_5S_6-$Fe_{16}S_{17}$, chalcopyrite, $CuFeS_2$, and sphalerite, ZnS, form widespread active and relict hydrothermal sites. Though sulfide oxidation of metal sulfides generally takes place at low pH, nonacidophiles from vent's sulfides are capable of autotrophic growth. Thus massive sulfide deposits on the seafloor may serve as potential source of electrons for autotrophic growth even when the vents are extinct. Such systems are also known to support high biomass of invertebrates.

Chemosynthetic systems in extreme environments like hydrothermal vents are likened to extreme environment systems on other planets or other extraterrestrial bodies. The search is on for chemoautotrophic forms on Mars. Such systems are also suspected to occur on Europa, a moon of Jupiter. They could represent the type of nonequilibrium systems which are thought to have been important in the origin of life on our planet.

Applications

These microbes evoke deep scientific interest because of their metabolic diversity and their adaptability to extreme environments. These characteristics can be gainfully harnessed by biotechnological firms for their enzymes or other metabolic products.

Sulfur-oxidizing bacteria could be judiciously used to contain `crude oil souring' due to excess sulfide production in oil wells.

Sulfur Bacteria in Symbiosis

The association of thioautotrophic (autotrophic at the expense of reduced sulfur compounds) symbiotic bacteria could be very specific. Most of them are related to subdivision of Gamma proteobacteria. Mariculture of symbiont-bearing invertebrate bivalves has also been suggested as a means of treating industrial sulfur waste.

Photosynthetic Sulfur Oxidation

Anoxygenic photosynthesis is also responsible for converting reduced sulfur to sulfate, thus forming habitats called `sulfureta'. The purple sulfur bacteria like *Chromatium* can oxidize sulfide internally through elemental sulfur to sulfate. The green sulfur bacteria do so externally. Consequently, the former are less tolerant to sulfide (-0.8 to 4 M) as compared to the latter (4–8 mM). The purple nonsulfur bacteria are least tolerant (0.4–2 mM).

When an anoxygenic bacteria grows on CO_2 as sole source of carbon, besides the formation of ATP, reducing-power NADPH must also be made available so that CO_2 can be reduced to form cell material. The source of reducing power is not water as in photosynthetic plants but reduced sulfur compounds like H_2S, S°, or thiosulfate. Hydrogen or organic compounds, such as lactate, succinate, butyrate, and malate, can also donate electrons for the activity:

$$6CO_2 + 12H_2S \rightarrow C_6H_{12}O_6 + 6H_2O + 12S^\circ$$

Habitats in which oxygenic photosynthesis is important is relatively limited in distribution and is therefore restricted to shallow coastal sediments or coral lagoon or sediments on beaches. These bacteria are at the end of anaerobic food chain that operates within microbial level. Heterotrophs feed sulfate-reducing bacteria (SRBs) and SRBs in turn feed photosynthetic bacteria, which can act as food sources for protozoans or microzooplankton.

Mixotrophy

Sulfur oxidation is carried out not only by chemolithotrophs but also by other groups like (1) mixotrophs (capable of autotrophic and heterotrophic growth); (2) chemolithotrophic heterotrophs; (3) heterotrophs which do not gain energy but derive benefits; (4) heterotrophs which gain nothing from the oxidation. Most pseudomonads are capable of growing mixotrophically on organic compound and reduced inorganic sulfur. Both marine and freshwater pseudomonads are capable of growing on thiosulfate and oxidizing it to tetrathionate.

The metabolic capabilities of a microbe sometimes cannot be too specific. It is argued that many microbes could be facultative, autotrophic at times, and heterotrophic at other, assimilating simple organic substrates that are available. Thus microbes like *Pseudomonas* sp. and *Alcaligenes* sp. have also been implicated in the heterotrophic sulfide oxidation. They could also behave like *Thiobacillus denitrificans*-like organisms (TDLOs) oxidizing reduced sulfide at the expense of nitrate and fixing carbon dioxide in the process. Such metabolic flexibility increases their competitive edge.

Sulfate Reduction

Assimilatory Sulfate Reduction

Sulfate-reducing activity (SRA) that takes place for the incorporation of sulfide radical for biosynthetic cycle is referred to as assimilatory sulfate reduction (ASR). Sulfate is reduced to sulfide in the assimilatory cycle which combines with serine to form cysteine. This in turn can be converted to methionine. These two amino acids are the main constituents of sulfur-containing molecules in the cells. Sulfur content can vary from 0.3% in eel grass to 3.3% in marine algae. As the N:S ratio in land plants is only 30:1, the reductive assimilation of sulfate is less important than nitrate. Assimilatory reduction is common among organisms and does not lead to the production of sulfide.

The eight-electron reduction of sulfate to sulfide proceeds in different stages. As the ion is stable it needs to be activated with ATP. The enzyme ATP sulfurylase catalyzes the attachment of sulfate ion to phosphate of ATP to form adenosine phosphosulfate (APS). Another P is added to APS to form phosphoadenosine phosphosulfate (PAPS) before it gets reduced to form sulfite.

Dissimilatory Reduction

The SRA that takes place in anaerobic respiration is termed as dissimilatory sulfate reduction (DSR). Here sulfate is used as terminal electron acceptor leading to the production of sulfide. Here too the sulfate ion is activated by ATP to form APS. However, in this case, the sulfate moiety of APS is reduced directly to form sulfite with the release of AMP. Thus the first product of ASR and DSR is sulfite.

Dissimilatory SRBs act as agents of synergy in sulfur cycle and bring about syntrophic associations. The end products from organic substrate oxidation and sulfate reduction lead to the formation of sulfide and carbon dioxide. SRA can account for nearly 80% of organic carbon mineralization in marine environment, especially in coastal regimes where nearly 5×10^{12} kg yr^{-1} of sulfate gets reduced.

SRA follows zero-order kinetics in marine sediments with respect to sulfate up to a concentration of *c*. 2 mM . The rate of SRA depends on both quantity and type of organic matter and the sulfate ion available which is generally not limiting in the marine environment. There is in general 2:1 molar relationship between the labile carbon utilized and the sulfate reduced. Sulfate-reducing rates (SRRs) can span several orders of magnitude: 50–500 nm cm^{-3} day^{-1} in coastal zones; 2785 nm g^{-1} day^{-1} in salt pans; 4000 nm ml^{-1} day^{-1} in salt marshes; and up to 14 000 nm ml^{-1} day^{-1} in microbial mats.

The other environmental parameter that affects SRA is temperature. Though the rates of activity can vary by factors ranging from <5 in winter to >0.30 in summer in the temperate region, in the tropics it is not very marked.

SRA predominates in marine sediments but the accumulation of the end products of sulfate respiration is pH dependent. Both metal sulfide formation and rapid biological oxidation are responsible for controlling the amount of sulfide that eventually escapes any system. Sediments harboring vegetation tend to emit less sulfide due to rapid oxidation by oxygen emitted from roots.

Though SRA is largely anaerobic there have been observations of its occurrence in the surficial oxic layers of microbial mats. Sometimes the rate measured in these oxic layers can equal or exceed the SRA of the deeper anoxic layers.

Abundance, Physiological Groups, and Taxonomic Diversity

Though SRBs are high in activity, they are low in abundance contributing to a maximum of 5–6% of total counts of bacteria. While culturable forms retrieved as colony-forming units (CFUs) range from 10^2 to 10^4 l^{-1} in agar shake tubes, most probable numbers (MPNs) methods yield 10^6–10^8 ml^{-1} and fluorescent *in situ* hybridization method (FISH) up to 10^7 ml^{-1}.

SRBs form two main groups based on their ability to utilize carbon sources completely or incompletely. In incomplete organic carbon oxidation, SRBs utilize a variety of organic substrates and oxidize it to acetate:

$$2(\text{lactate}) + SO_4^{2-} \rightarrow 2(\text{acetate}) + CO_2$$

In complete organic carbon oxidation, the organic substrate is totally oxidized to carbon dioxide, water, and sulfide:

$$\text{Acetate} + SO_4^{2-} \rightarrow 2CO_2 + S^{2-} + H_2O$$

Though the above two groups of SRBs are physiologically distinct they can coexist with the latter using the metabolic end products of the former.

SRBs also fall in the following major groups, namely Gram-negative mesophilic, Gram-positive spore-forming thermophilic bacteria belonging to δ subgroup of Proteobacteria and Archaea. The former includes two main families Desulfovibrionaceae and Desulfobacteriaceae. Desulfovibrionaceae includes the genera Desulfovibrio and Desulfomicrobium. Desulfobacteriaceae includes at least 20 genera, most of which are complete oxidizers of organic acids. The Gram-positive group comprises of *Desulfotomaculum*.

Desulfobacteriaceae are metabolically more versatile and are highly adapted to environments that undergo drastic redox changes as in salt marsh sediments or intertidal regions. Rhizosphere habitats are replete with *Desulfobulbus* species which are capable of sulfur disproportionation.

Sulfur-reducing activity

Bacteria like *Desulfuromonas acetoxidans* are capable of reducing sulfur at the expense of acetate. Some SRBs and iron-reducing bacteria are also capable of reducing sulfur. Many of these bacteria are able to generate ATP during sulfur reduction. These groups can also use organic disulfide molecules like cysteine or glutathione. Though sulfur and sulfate reducers can coexist, the latter can produce more sulfide. Most of these bacteria belong to Archaea. Methanogenic thermophilic Archaea reduce sulfur to sulfide while methane generation gets retarded. The process of sulfur reduction is an ancient process, as is suggestive from their presence in the deep branches of the phylogenetic tree. Though some sulfur reducers phylogenetically belong to δ subclass of Proteobacteria, they show affinity to other unrelated classes as well. Metabolic flexibility assures ecological competitiveness.

Sulfur disproportionation

Inorganic fermentation or disproportionation of sulfur and sulfur compounds, thiosulfate, and sulfite has been frequently encountered in SRB:

$$S_2O_3^{2-} + H_2O \rightarrow SO_4^{2-} + HS^- + H^+$$

$$4S^\circ + 4H_2O \rightarrow SO_4^{2-} + 3HS^- + 5H^+$$

$$4SO_3^{2-} + H^+ \rightarrow 3SO_4^{2-} + HS^-$$

Disproportionation seems to be very important and therefore widespread. Thiosulfate is an important intermediate as it can act as an electron acceptor or donor and thus mediate both oxidative and reductive cycle. Thus the thiosulfate shunt provides for complete anaerobic sulfur cycling. Though this is not energetically very viable, the bacteria are able to grow in the presence of metal oxides which can scavenge sulfide.

Use of Heavy Isotopes in Ecology

Sulfur exists primarily as two stable forms of isotopes: ^{32}S and to a certain extent as ^{34}S. Heavier isotopes are discriminated against; that is, most biochemical reactions prefer the lighter isotope and this preference is useful in elucidating microbial interactions. Thus sulfide production by bacterial reduction is much lighter than sulfide of strictly chemical or geothermal origin. Also the biological oxidation of sulfide to sulfur or sulfate either aerobically or anaerobically shows a preference for the lighter isotope. However, this fraction is not as great as that occurring through sulfate reduction as respiratory rates are faster and higher than synthetic ones.

Geochemical Implications of SRBs and SRA

Though the abundance of SRBs is generally low, their high respiratory activity mediates many other activities. Sulfate reduction sets other geochemical reactions in pace. The sulfide formed is responsible for the precipitation of metal sulfides which are available for autotrophic sulfide-oxidizing bacteria. It has been argued that about 90% of the sulfide produced by SRA is recycled back to sulfate to complete the cycle. The rest gets buried to form FeS_2. However, the energy gain from dissimilatory SRA is relatively low: at an energy yield $\Delta G'$ of -128 kJ with lactate and -48 kJ with acetate. Nevertheless, the sulfide produced by these bacteria can act as an energy source for other autotrophic bacteria.

The SRBs are also capable of using a variety of inorganic sulfur compounds as electron acceptors. These include dithionite, tetrathionite, thiosulfate, sulfite, bisulfite, metabisulfite, sulfur, sulfur dioxide, and dimethyl sulfoxide.

Sulfate reducers can use other electron acceptors like nitrate; group 4 oxyanions like molybdate and selenate; and even metals like uranium, chromium, technetium, gold, iron, and manganese(IV).

SRA Implications on Climate

SRBs can not only mediate synergistic reactions locally but also impact the climate on a wider scale. They are known to participate not only in the degradation of dimethylsulfoniopropionate (DMSP) but also in the flux of degradation product, dimethyl sulfide (DMS).

The SRBs are involved in the demethylation of DMSP to yield methylmercaptopropionate (MMPA), carbonate, and sulfide or oxidation of DMS to yield bicarbonate and sulfide.

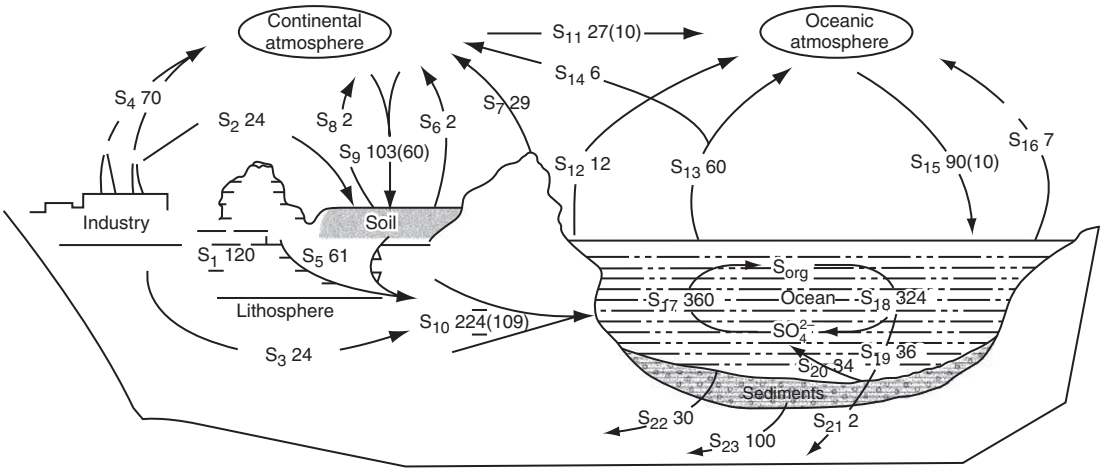
DMSP demethylation

$$3/4SO_4^{2-} + DMSP \rightarrow MMPA + HCO_3^- + 3/4HS^- + 5/4H^+$$

DMS oxidation

$$3/4SO_4^{2-} + DMS \rightarrow HCO_3^- + 5/2HS^- + 3/2H^+$$

Intertidal sediments harbor algal osmolyte DMSP. DMSP could release DMS by the intervention of SRB. This could have countereffect on global warming. This reaction also decreases the effect of the potent greenhouse gas methane.



Sulfur flux with anthropogenic contributions in parenthesis

P1	Mining from lithosphere	P2	Fertilizers from soil
P3	Industrial sewage	P4	Anthropogenic sulfur to atmosphere
P5	Erosion	P6	Biogenic sulfur
P7	Volcanic	P8	Dust emission
P9	To land from atmospheric precipitation	P10	From river runoff
P11	Anthropogenic and natural flux from continent to oceans	P12	Biogenic H_2S from shallow coastal sediments
P13	Marine sulfur from sea spray	P14	Marine sulfur to continents
P15	Ocean atmosphere to ocean	P16	Reduced sulfure mission from ocean
P17	Biomass from marine plants	P18	Mineralized sulfur from dead marine organisms
P19	Organic sulfur to sea bottom	P20	Sulfate oxidized from organic sulfur returns to sea
P21	Organic sulfur buried in marine sediment	P22	Sulfate buried in marine sediments
P23	Reduced sulfur buried in marine sediments		

Fig. 2 Fluxes of the global biogeochemical sulfur cycle. Modified from Ivanov MV (1981) Global biogeochemical sulfur cycle. In: G E Likens (ed.) *Some Perspectives of the Major Biogeochemical Cycles*, ch. 4. Chichester, UK: Wiley.

DMS emissions form an important bulk of sulfur that enters the atmosphere and affects the climate. It accounts for 90% of the biogenic sulfur emissions from the marine ecosystem. DMS produced from DMSP breakdown which reaches the atmosphere serves to decrease warming by radiative backscatter from aerosols. It also reflects radiation from increased cloud cover. Both methanogens and SRBs compete for DMS but methanogens outcompete SRBs when DMS concentrations are high.

Similarly, the breakdown of the osmoregulant glycine betaine in marine sediments releases acetate and trimethylamine. The former is a preferred substrate for SRBs and the latter for methanogens. Other sulfur compounds like carbonyl sulfide (OCS) and carbon disulfide (CS_2) species could be formed photochemically or biologically for bacterial consumption.

Thus, the gaseous products of sulfur cycle interlink land, water, and atmosphere. These include hydrogen sulfide, DMS, methane thiol, carbonyl sulfide, and carbon disulfide. These volatile sulfur compounds get photochemically oxidized to produce acid rain or aerosol sulfate particles that decrease the incoming solar radiation and lead to cloud condensation nuclei. These processes influence the global radiative balance and consequently the climate.

Applications

The activity of SRB could be deleterious to all underground constructions because of their involvement in corrosion. The sulfide they produce is responsible for anodic corrosion, and their propensity to scavenge hydrogen generated in underwater metal structures could cause cathodic corrosion. However, some of these activities could be used in metal recovery from wastewater treatment as metal sulfides. The synergy existing between SRB and other microbes could be effectively used in bioremediation and ecosystem management. This trait could be exploited to contain mercury and other heavy metal contamination in water bodies.

Fluxes of the Global Biogeochemical Sulfur Cycle

The sulfur fluxes, both natural and anthropogenic, have been derived from various studies. A summary diagram of the global sulfur cycle with quantitative estimates of the sulfur fluxes is given in Fig. 2. The numbers near the arrows designate the total sulfur flux in Tg S yr^{-1} for all compounds. The contributions from anthropogenic activities are indicated by numbers in parentheses. About 120 Tg S are extracted annually by man from the lithosphere in fossil fuels and sulfur-containing raw materials for the chemical industry of which about 58% (70 TgS) gets emitted to the atmosphere. About half of the remaining 50 Tg S directly enters rivers through sewage and residual waters, and another part from fertilizers to agricultural land. Simultaneously, volcanic gases contribute markedly to the atmospheric sulfur cycle over continents amounting to 29 Tg yr^{-1}. The major transfer of sulfur from continents to the ocean by river runoff amounts to 224 Tg of which anthropogenic contribution is about 109 Tg. The total flux of various sulfur forms, that is, organic, sulfate, and pyrite from oceanic water to sediments and further to the lithosphere, amounts to 130 Tg yr^{-1}. Thus the estimates suggest that the anthropogenic sulfur fluxes to the atmosphere and hydrosphere have reached a level comparable with that of natural fluxes. The natural sulfur flux from the lithosphere, its main reservoir, is compensated by the reverse flux of sulfur compounds to the lithospheric sediments of the ocean. Further, there is also indication that by the end of this century the anthropogenic sulfur fluxes could notably increase all over the world.

All of the main reactions of the sulfur cycle involving living organisms are closely related to the carbon cycle. The amount of carbon involved in the fluxes of the sulfur cycle through biogenic processes varies depending on the type of organisms undertaking the metabolism of the sulfur compounds. In the processes of bacterial chemosynthesis, which are characterized by low amounts of energy utilized for the CO_2 assimilation, only relatively small amounts of carbon are transformed into organic matter. In anaerobic bacterial photoassimilation of CO_2 where sulfur compounds are used as electron donors, the amounts of oxidized sulfur and assimilated carbon are comparable. In anaerobic sulfate reduction, 24 g of organic carbon is mineralized for each 32 g of reduced sulfate sulfur. Thus, in ecosystems with an advanced development of photoauotrophic bacteria and SRBs, both groups of microorganisms transform significant amounts of carbon compounds and, consequently, these organisms should be considered not only as participants in the sulfur cycle but also as active biogeochemical agents of the carbon cycle.

Summary

Microbes, especially bacteria, play an important role in oxidative and reductive cycle of sulfur. The oxidative part of the cycle is mediated by photosynthetic bacteria in the presence of light energy and chemosynthetic forms in the absence of light energy. At the end of the anaerobic food chain in bacteria they serve to purify the system of sulfide and other metabolic end products. In the process sulfur is returned to the system as sulfate. In transition zones from anaerobic to aerobic, photosynthetic bacteria can form a food source to protozoans and microzooplankton. Chemosynthetic sulfur-oxidizing bacteria are the dominant bacterial forms that support thriving ecosystems in hydrothermal vents. Scientists are seeking evidences from such extreme environment for similar life on other planetary bodies.

The reductive cycle on the other hand is mostly driven by the sulfate/sulfur-reducing bacteria which use sulfate as the electron acceptor in anaerobic respiration to produce sulfide. Their close association with other microbes can have profound geochemical influence. Their metabolic activity dictates the availability of trace metals to other forms of life. While sulfide gets precipitated, phosphate gets released into the systems. Nitrogen fixation by these anaerobes also adds to the nitrogen economy of the environment they inhabit. In sediments of continental shelves that hold the reserve of gas hydrates, these microbes can modulate the concentration of methane in such ecosystems. Most importantly, the interaction with DMSP, an osmolyte from phytoplankton, can have wide-ranging climatic implications.

The main reactions of the sulfur cycle involving living organisms are closely related to the carbon cycle. The amount of carbon involved in the fluxes of the sulfur cycle through biogenic processes varies depending on the type of organisms undertaking the metabolism of the sulfur compounds. The estimates suggest that the anthropogenic sulfur fluxes to the atmosphere and hydrosphere have reached a level comparable with that of natural fluxes. The natural sulfur flux from the lithosphere, its main reservoir, is compensated by the reverse flux of sulfur compounds to the lithospheric sediments of the ocean. Further, there is also indication that by the end of this century the anthropogenic sulfur fluxes could notably increase all over the world.

This is NIO Contribution No. 4296.

Furhter Reading

Fenchel, T., Blackburn, T.H., 1979. Bacteria and Mineral Cycling. London: Academic Press.

Hines, M., 1996. Emission of sulfur gases from wetlands. In: Adams, D.D., Crill, P.M., Seitzinger, S.P. (Eds.), Mitteilungen der IVL, Vol. 25: Cycling of Reduced Gases in the Hydrosphere. Stuttgart: Science Publishers, pp. 153–161.

Ivanov, M.V., 1981. Global biogeochemical sulfur cycle. In: Likens, G.E. (Ed.), Some Perspectives of the Major Biogeochemical Cycles. Chichester, UK: Wiley. ch. 4.

Jørgensen, B.B., 1988. Ecology of the sulfur cycle: Oxidative pathways in sediments. In: Cole, J.A., Ferguson, S.J. (Eds.), The Nitrogen and Sulfur Cyles. Cambridge: Cambridge University Press, pp. 31–63.

Loka Bharathi, P.A., 2004. Synergy in sulfur cycle: The biogeochemical significance of sulfate reducing bacteria in syntrophic associations. In: Ramaiah, N.N. (Ed.), Marine Microbiology Facets and Opportunities. Panaji, India: National Institute of Oceanography, pp. 39–51.
Madigan, M.T., Martinko, J.M., Brock, P.J., 1997. Brock Biology of Microorganisms, 8th edn. Upper Saddle River, NJ: Prentice-Hall.
Van Dover, C.L., 2000. The Ecology of Deep-Sea Hydrothermal Vents. Princeton, NJ: Princeton University Press, pp. 115–226.

Sustainable Cropping Systems

Shabtai Bittman and Derek Hunt, Agriculture and Agri-Food Canada, Agassiz, BC, Canada
Cynthia Grant, Agriculture and Agri-Food Canada, Brandon, MB, Canada
William Deen, University of Guelph, Guelph, ON, Canada

Introduction

As the world population has grown so has the portion of surface land area of the globe that is used for crop systems to produce food for humanity (11% of 13.6 billion ha, according to FAO). Also, the productivity of food production per unit of land area has increased so that land required for food production per capita has declined from 0.45 to 0.25 ha since 1965, thanks to a compliment of diverse technologies in the fields of genetics, engineering, chemistry, ecology and others. Genetics has provided more productive and hardy crop plants, engineering has provided more efficient and more rapid tools and machines, chemistry has provided more effective fertilizers and pesticides, and ecology has helped famers better understand how to manage their fields and landscapes sustainably (e.g., effective crop rotations and pest management). Most of the world's food is produced on the richest soils with the most favorable climate (temperatures and precipitation) although in some regions supplementary irrigation is required during dry spells. Everywhere, crops have displaced natural vegetation; in moist areas crops have displaced forests and in dry areas grasslands and shrubs. Original plant communities, like tall grass prairie, have almost completely disappeared due to crop production in some countries.

Crops typically need at least 100 days of favorable weather, free of frost, and for a typical wheat crop about 1 mm of rain per 5–22 kg of grain per ha (or 0.7–3 kg of protein per ha). Besides providing most of the carbohydrates, proteins and fats in our diets, crops also give us essential nutrients like minerals and vitamins, health promoters like flavonoids, and amenities like flavors, medications and even recreation. And a substantial proportion of all crops produced are fed to livestock that provide a more varied plate but also help to balance dietary fats, proteins and supplement certain minerals (e.g., iron and selenium) obtained directly from plants.

The study of crops and their relationship to other plants (weeds and invasive plants), macro- and micro-fauna, microbes, and the soil substrate is called crop ecology. This is a field of study gaining worldwide interest, as can be seen by increasing university positions and publications, because crop ecology is expected to help sustain crop production and ecosystem services like biodiversity and soil carbon sequestration, while also helping to protect surface water from phosphorous eutrophication, ground water from nitrate contamination, air from fine particles and trace gases pollution, and climate from emissions of greenhouse gases. This article will describe three contrasting cropping ecosystems that are typical of major crop growing regions around the globe. The article will also discuss the long term sustainability of these three diverse crop ecosystems.

Semi-Arid Small Grain Cropping Systems

Semi-arid small grain cropping systems refer to rain-fed agriculture in areas where precipitation ranges between 250 and 500 mm per year (Fig. 1). These dryland cropping systems are prevalent in the Great Plains region of North America, as well as in areas of Australia, Africa, and Asia. By definition water availability is restricted in semi-arid cropping systems and is generally the most limiting environmental production factor, particularly in warmer regions where evaporative losses are high. Cropping choices and productivity are also restricted by the length of growing season which is constrained by temperature and/or water seasonality. Crop productivity is also affected by nutrient limitations, particularly nitrogen deficiency, and by crop losses from weed competition, insects and diseases.

Management practices in semi-arid cropping systems are designed to optimize crop yield potential and long-term sustainability. Crop yield is optimized by using incoming moisture as efficiently as possible, utilizing as much of the growing season as possibly, supplying the nutrients needed to support the crop yield potential and removing the weeds, diseases and insects that compete with the crop for resources such as light, water, and nutrients. Environmental sustainability requires that the soil, air, water and biological resources in the ecosystem are not degraded. Soil erosion and degradation from salinization, organic matter loss and nutrient depletion are of major concern in semi-arid regions. Nutrient movement from field areas to the air and water may affect environmental quality. Introduction of foreign pests and diseases and development of genetic resistance by pests to commonly used control practices present ongoing challenges to effective pest management and economic sustainability of production.

Efficient semi-arid cropping systems rely on the capture, conservation, and utilization of the limited available precipitation. In the past, use of a fallow year between one or two crop years was common to capture water and lower production risk. Tillage during the fallow phase was used to control weeds and release nutrients through organic matter decomposition. The moisture and nutrients conserved during the fallow period increased the productivity of the following crop. However, availability of improved equipment, herbicides, crop genetics and fertilizers has allowed a reduction in tillage. Reduced tillage and increased crop residue cover lead to greater moisture retention, lower water run-off, reduced evaporative losses and higher available moisture for crop growth, allowing for intensification of crop production and the reduction or elimination of fallow.

 doi:10.1016/B978-0-12-409548-9.10580-9

Fig. 1 Small-grain (wheat in foreground) based cropping systems in the semi-arid Northern Great Plains.

A key principle in sustainability of semi-arid cropping systems is intensification of crop production to balance crop water extraction with available moisture. Optimizing crop production per unit available moisture will normally optimize economics of production. Environmentally, it is also important that incoming moisture is used by the crop, rather than being lost from the soil–crop system through infiltration below the rooting depth, run-off or evaporation. Water moving through the soil by infiltration can lead to leaching of nutrients, while run-off can cause soil erosion and nutrient movement to surface waters. Evaporative loss of water from soil rather than transpiration through plants enhances the movement of salts towards the soil surface through capillary action, increasing the risk of salinization. Intensified crop rotations effectively use the available water for crop production, reducing the risk of leaching, salinization and surface run-off. In addition, intensified rotations increase the amount of crop biomass produced and returned to the soil, potentially increasing soil organic matter content and contributing to a healthy soil biological community. Diversification of semi-arid cropping systems will also contribute to economic and environmental sustainability. Including crop species with different rooting and water use patterns can improve water use efficiency. For example, a deep-rooted crop can be used to capture water that may have moved below the rooting zone of a more shallow-rooted crop. Keeping an actively growing crop in the field for as long as possible will help to increase organic matter input and reduce the risk of soil erosion, run-off, and leaching. Inclusion of perennial crops in a rotation where possible, is therefore of particular value.

As with moisture, crop nutrient supply and demand must be balanced to ensure crop yield and water uptake and avoid environmental problems. Intensification of crop production increases nutrient removal. For long-term sustainability, nutrients removed from the soil in the harvested crop must be replaced to maintain crop yields and avoid soil nutrient depletion and degradation. However, excess or poorly managed fertilizer inputs, whether from organic or inorganic sources can harm ecosystem and human health by contributing to nitrate accumulation in groundwater, eutrophication of surface water, soil and water acidification, greenhouse gas production, formation of ground-level ozone and particulate matter, and loss of biodiversity. Phosphorus movement to surface waters can lead to eutrophication while potentially toxic trace elements such as cadmium present in phosphorus fertilizers can accumulate in the soil over time. Carbon dioxide emissions from the large amount of fossil fuel used in fertilizer production and transport can also contribute to climate change. Rate, source, timing and placement of nutrient application must be carefully selected to match supply to the crop uptake in order to improve nutrient use efficiency and reduce potential environmental impacts. Crop nutrient use efficiency can be increased by including crops with different nutrient requirements, for example following a nitrogen-fixing legume, such as pulse crops or alfalfa (*Medicago sativa* L.), by a crop with a high nitrogen demand to utilize nitrogen left behind in the soil or mineralized from high nitrogen crop residues.

Efficient crop production in semi-arid systems requires that pests are controlled to optimize crop vigor and reduce competition for water, light, and nutrients. Integrated pest management using a variety of tools is commonly practiced. Effective tillage management and seed-bed preparation allows for establishment of a vigorous crop that can compete with pests. A diversified crop rotation that includes cereals, legumes and oilseed crops can reduce the build-up of insects and diseases specific to certain plant species. Integrated disease management uses a range of practices to reduce disease pressure including residue management, balanced crop nutrition, disease-resistant cultivars, crop rotation, field sanitation, and chemical application if necessary. Insect management also utilizes crop rotation, cultivar selection, and chemical application where insect populations warrant. Weed control is a major challenge in semi-arid agriculture and the majority of semi-arid cropping systems rely on judicious use of herbicides to control weeds, which would otherwise drastically reduce crop yield and quality. Use of herbicide-resistant crop cultivars that allow the application of relatively non-selective herbicides has been widely adopted, particularly in North America, closely tied to the expansion of reduced tillage systems. Overuse of a limited range of herbicides has led to selection of some weeds that are no longer controlled by certain herbicide families. Diversified cropping systems allow for use of herbicides with varying modes of action, slowing the development of herbicide resistant weeds.

Organic systems make up a small but significant portion of semi-arid small grain production systems. Organic systems attempt to improve environmental sustainability by minimizing external inputs that may be disruptive to the agroecosystem. Organic farmers normally manage crop nutrient supply through the use of diversified crop rotations, including legume crops for nitrogen fixation, production of green manures, and addition of manures or composts. Weed populations are managed by non-chemical means such as crop rotation and production of perennial crops, use of competitive cultivars, high seeding rates, field sanitation, and tillage. Just as with conventional production systems, crop yields in organic systems can vary widely, but are normally lower than in conventional systems. Profitability of organic systems can be increased because of reduced input costs and availability of price premiums associated with organic products.

Environmental benefits attributed to organic systems include enhanced biodiversity, higher energy use efficiency, and reduced risk of nutrient loss to air and water. Organic systems that include addition of livestock manure or compost as a nutrient source may increase organic carbon levels in the soil thus improving soil quality. Manure and compost inputs in organic systems must be managed carefully, just as chemical fertilizers must be, in order to limit the risk of off-site nutrient movement. However, systems that do not include replacement of exported nutrients can deplete the soil, resulting in lower crop productivity, lower long-term inputs of organic carbon, and reduced sustainability. While nitrogen can be provided internally to the system through production of nitrogen-fixing legume crops, exported nutrients such as phosphorus cannot be generated internally and their depletion may be problematic for long-term sustainability. Another concern related to organic production systems is the reliance on tillage for pest control. Excessive tillage can increase the risk of soil erosion and lead to depletion of soil organic matter. Development of management practices that maintain residue cover and restrict the amount of tillage are helping to improve sustainability of organic systems in semi-arid regions.

Mesic Corn (*Zea mays* L.) and Soybean (*Glycine* max (L.) Merr.) Cropping Systems

A large part of former deciduous forest in North Central United States and Ontario is now in cropping systems planted to predominantly corn which has been dubbed the Northern Corn Belt (Fig. 2). The corn, which is often grown from hybrids, is used for livestock feed, ethanol production, and a variety of food uses. The national average corn grain yield (per land area) in this region has steadily increased since the 1940s and is now nearly five times greater than 70 years earlier. Yield increases resulted from the adoption of hybrid corn, improved genetics, availability of nitrogen fertilizer and chemical pesticides, improved agronomy, and mechanization. Soybean expansion into the Northern Corn Belt began in the 1920s. Soybeans are produced for their oil and protein, which compliment carbohydrate-rich corn in livestock feed rations. A smaller percentage of soybean is processed for human consumption or other non-food (industrial) products. Like corn, soybean yield per land area has steadily increased due to improved genetics, availability of chemical pesticides, improved agronomy, and mechanization.

As corn and soybean production expanded from about 1860 to 1970, the Northern Corn Belt was transformed from a mixed crop-and-livestock farming system that included rotations with perennial forages to a highly specialized and simplified cash-grain farming system dominated by corn/soybean crop rotation (Fig. 3). Whereas in the past, farmers relied upon crop rotation (and manures) to provide weed and pest control, fertility, and labor distribution, various technologies now available provide farmers an alternative to crop rotation. For example, farmers do not require crop rotation to manage weeds and pests as they can achieve this with herbicides, insecticides, fungicides and improved crop varieties with resistance to pests.

Fig. 2 Coarse grain cropping systems (corn in mid-frame) in the mesic (deciduous forest) Northern Corn Belt region.

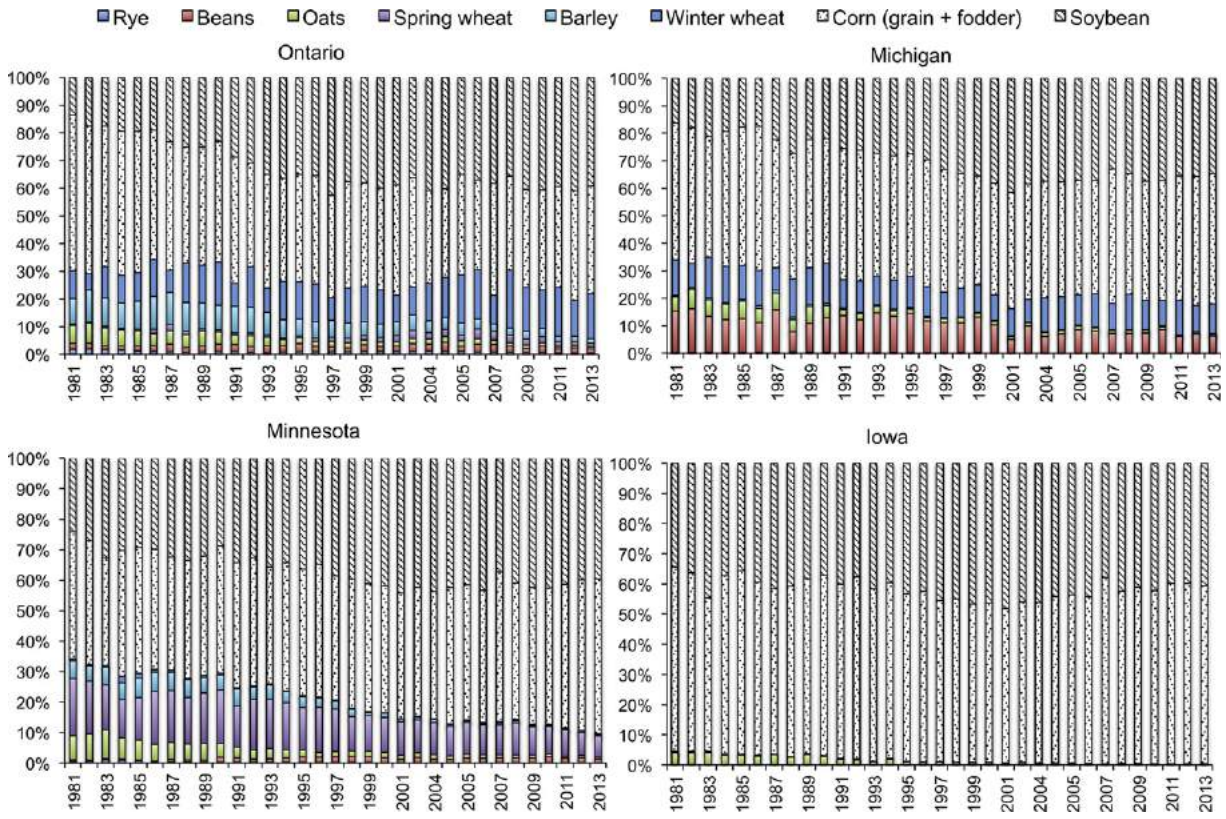


Fig. 3 Harvested areas (as % of total harvested area) of 8 major field crops in four states/provinces of the Northern Corn Belt from 1981–2013. Areas of canola and hay were not included for clarity. Sources OMAFRA, 2014, USDA-NASS, 2014; Reproduced from Gaudin, A. C. M., Janovicek, K., Deen, B. and Hooker, D.C. (2015). Wheat improves nitrogen use efficiency of maize and soybean-based cropping systems. *Agriculture, Ecosystems and Environment* **210**, 1–10. https://doi.org/10.1016/j.agee.2015.04.034.

The decline in bio-diversity of the Northern Corn Belt has been manifested not only through the dominance of the simple 2-crop corn/soybean rotation system but also through (1) an increase in field size which facilitates use of larger equipment, and (2) a decline of diversity at the sub-field scale. As field size has increased there has been a corresponding decline in fence rows, grass waterways, and buffer zones to environmentally sensitive areas such as streams. Also, over the past three decades, homogeneity of within-field management has increased; whereas formerly many farmers made agronomic decisions at the subfield scale, the increase in size of farms and equipment, and improved herbicides has encouraged farmers to apply agronomic management practices at a much larger scale. It should be noted that this particular trend is currently being challenged by the introduction of precision agriculture techniques (Berry *et al.*, 2003). These techniques involve collecting data at the subfield scale in order to inform decision tools that interact with agricultural equipment to vary management spatially in real time.

There is growing concern that there is a loss of biodiversity in the Northern Corn Belt agroecosystem, particularly due to less diverse crop rotations. It is well established that highly diverse cropping systems that more closely mimic the diversity of natural ecosystems are more resilient and environmentally sustainable due to their biological diversity. The decline in rotation diversity has been shown to be associated with reduction in soil quality parameters such as organic matter and aggregate stability (or tilth), increased soil erosion, increased greenhouse gas emissions, and ultimately decreased yield potential and yield stability. These agronomic and environmental consequences have also negatively influenced crop response to nitrogen and soil nitrogen processes leading to nitrogen losses from fields. In fact, it was shown removal of either a legume, such as alfalfa, red clover (*Trifolium pratense* L.) or soybean, or a non-legume such as winter wheat (*Triticum aestivum* L.), from a corn based rotation increased corn requirement for nitrogen fertilizer. This demonstrates that simple rotation systems have a greater reliance on inputs such as nitrogen fertilizer to achieve maximum yields. Also, crop yields respond more poorly to no-till (direct planting without any soil disturbance) under simple than under diverse crop rotation. This is a significant concern since no-till systems are associated with benefits such as reduced soil erosion and reduced energy consumption. Hence, under a simple crop rotation, farmers are more likely to forego the benefits of no-till which may lead to reduced soil health and poor nutrient use efficiency, and increased contamination of water sources by sediment and nutrients. Finally, since low biodiversity systems have reduced water use efficiency, simple crop rotations are associated with reduced resilience to extreme weather events, particularly drought.

The trend towards crop rotation simplification in the Northern Corn Belt indicates that farmers and society have been willing to accept the negative consequences of these practices. There is, however, growing evidence that costs associated with maintaining simple rotations are increasing and that there may be increasing incentives for farmers and society to reintroduce diversity into the system. (1) The direct

costs of yield loss and yield variability are expected to increase under simple rotations due to warmer temperatures, more variable precipitation patterns, and greater frequency of drought anticipated under a changing climate. This effect will be exacerbated by increased yield potential of new corn and soybean varieties, which will place further strains on an increasingly limiting water supply. (2) The indirect benefits of increasing soil organic matter associated with complex rotations are becoming increasingly valuable to farmers in a changing climate. Increasing soil organic matter represents sequestered carbon which can meaningfully mitigate increasing greenhouse gas levels. Soil organic matter is also correlated with improved soil structure rendering soil more fertile and less vulnerable to soil erosion resulting from expected extreme weather events. (3) Reductions in soil erosion will help reduce water quality concerns resulting from deposition of nitrogen, phosphorus, and soil sediment from farmland into waterways. Complex rotation also mitigates greenhouse gases by reducing nitrogen fertilizer requirements and improving nitrogen fertilizer use efficiency. (4) Weed management may re-emerge as an incentive for farmers to diversify crop rotation. Simple rotations have been facilitated by herbicides and the emergence of crops that have resistance to herbicides that are effective across a range of weed species, soil types, and timings. However, simple rotations are also associated with larger weed seedbanks, increased weed pressure and increased rapidity of herbicide resistance. Many herbicides that have been commonly used in corn and soybean are now less useful due to widespread weed resistance in the Northern Corn Belt. By necessity, farmers are again looking to complex crop rotations to enable more integrated weed management strategies with less reliance of herbicides for controlling weeds.

The fact that many farmers in the Northern Corn Belt have already begun reintroducing diversity is encouraging for sustainability of cropping systems in the Northern Corn Belt. The following are some modern examples of farmer efforts to reintroduce diversity. (1) As mentioned, new precision agriculture techniques are being employed to enable agronomic decisions at a smaller scale. (2) To overcome the poor results from no-till systems, particularly in a simple corn/soybean rotation, farmers are increasingly using strip tillage or other innovative reduced tillage systems. These systems provide some of the soil protection under extreme weather events that no-till provides, but minimize the yield lag associated with no-till. (3) Farmers are increasingly using cover crops to protect the soil over winter, to capture excess nutrients and to add diversity to their cropping system. In a corn/soybean rotation, cover crops are typically inter-seeded into juvenile corn plants or planted following soybean harvest. Biomass production of cover crops in a corn soybean rotation is limited by the fact that both corn and soybean are full season crops (i.e., high yielding long-season crops); however, even the small amounts of organic matter added to the soil helps to reduce soil erosion risks. (4) Finally, farmers are increasingly adding crops to their rotations. In Ontario, for example, winter wheat is again gaining in popularity as farmers are increasingly recognizing the extensive benefits of adding a third crop to their corn soybean rotation. Since winter wheat is planted after soybean in late summer, it also functions as a cover crop protecting the soil over winter. Moreover, since winter wheat is harvested in July–August, it provides an ample 2-month window to establish a cover crop with sufficient biomass to contribute significantly to soil organic matter. While these changes are incremental, over time they will provide substantial improvements in the sustainability of production in the Northern Corn Belt. This ability of farmers in the Northern Corn Belt to respond to shifting realities underpins the resiliency and sustainability of both human and biological systems.

Intensive Forage Cropping Systems in a Humid Region

Cattle and other ruminants comprise a major source of meat, milk, fiber, crop nutrients (manure), fuel (manure and tallow), and draft power. Ruminants have a nutritional requirement for forage crops that are high in cellulosic fiber that cannot be digested by monogastric animals (including humans). The public observes dryland forage landscapes with beef cattle in western "cowboy" films and moist forage systems with dairy cows in the bucolic images of the English and Swiss country-sides (**Fig. 4**). The former are expansive water-limited grasslands that are given few inputs and often comprise largely of native flora (i.e., native grasslands). The latter are planted by farmers, typically replacing forests, and often receiving more nutrients and pest control measures than the dry grasslands but less than arable crops as described above. For society, production of animal products is often seen as inefficient since the yield of protein per land area is much lower than arable crops. However, forage crops are usually perennial and are often much better suited than arable crops to difficult soils and landscapes that may be steep, stony, flooded, or with extremely fine or coarse soil texture that have, respectively, poor drainage and poor water holding capacity. Production of cattle and other ruminant

Fig. 4 Forage and ruminant pastoral cropping systems in a moist region.

species can make use of substandard soils or inedible crops so forage-livestock systems are contribute value to society. Also, when properly managed, perennial crops may be benign and even beneficial to the environment, relative to arable crops, by preventing erosion, sequestering C and improving fertility.

Feeding cattle for production of milk has, over recent decades, become a very exacting practice with very high milk production resulting when well managed cows are provided a high-quality, balanced ration. Most often on dairy farms there are at least two crops with complimentary nutritional attributes that are grown and fed whole (stems, leaves and grain) to the cows. There is typically an annual crop with high grain content (over 40%) that is highly digestible and energy-rich like corn or barley (*Hordeum vulgare* L.). There is also a fiber and protein rich forage, typically a perennial grass like timothy (*Phleum pretense* L.) or perennial ryegrass (*Lolium perenne* L.) or a legume like alfalfa or red clover, or a mix of one to many grasses and legumes. Both crop types need to be managed precisely, especially in terms of timing of harvest, to ensure that large amounts of very high quality feeds are produced and not spoiled during preservation.

Corn is a remarkable crop with tremendous capacity for producing large amounts of highly digestible feed. However, corn has a tropical origin and grows only in warm weather. In temperate regions, this means that for the cooler part of the year the field may be bare and therefore, as stated above, subject to soil loss by erosion, and nutrient loss into water bodies by runoff and leaching. Furthermore, nutrient uptake ends in August, 1–2 months before harvest, so nutrients released in late summer from applied manure, from crop residues, and from the soil itself, may accumulate in the soils and eventually leach even when the crop is still standing in the field. However, corn is often coupled with a winter cover crop which provides for some fall nutrient capture and soil protection over winter. Due to its relatively shallow root system and the distance between rows, corn is also subject to nutrient loss during its early growth period. An innovative cropping technique is to plant the cover crop into the young corn crop so that it captures the sunlight that falls wastefully between corn rows until the canopy closes, and resumes growth rapidly after the corn is harvested (**Fig. 5**). This technique provides more fall growth, more nutrient capture, and better overwinter soil cover than on conventional fall cover crops. And more high quality feed is produced in spring before the corn is replanted. For the somewhat staggered growth periods of the winter and summer crop, including the time of overlap, this progressive cropping system has been dubbed "relay cropping". Relay cropping is widely used by dairy farmers in northwest Washington State and nearby areas.

In contrast to corn, widely grown temperate perennial forages such as alfalfa and tall fescue (*Festuca arundinacea* L.) are better adapted to temperate conditions and hence have a long growing period, which may start even before the snow is gone and continue after initial frosts. To harden before winter, temperate perennial crops have a tendency towards becoming dormant in fall, triggered by the short days, then break dormancy when the days start to get longer, even in sub-zero temperatures. Crops that have broken dormancy especially in spring are subject to injury from a late freeze. Depending on the region and the amount of land available, cattle producers may grow grass, alfalfa or a mixture to complement their corn (or in cooler regions a small grain like barley).

Nutrient Management Using Cow Manure

The challenge for all sustainable livestock-crop farms is to use manure as the primary or, ideally, the only nutrient source for crops. Most dairy farms in industrialized countries handle manure as a liquid or slurry. Slurry has a very high water content (90%–95%) and low nutrient concentrations (0.3% nitrogen and 0.05% phosphorous) and is thus heavy and laborious to haul to fields and spreading can cause soil compaction and rutting in fields. The ratio of nitrogen to phosphorous in dairy slurry is about 5:1 but half the nitrogen is bound in organic molecules, such as proteins, that resist mineralization so that the ratio of available nitrogen to phosphorous is about 3:1. This contrasts with plant requirement which is about 10:1. Furthermore, most of the available (mineral) nitrogen in manure in the ammonia form (NH_3 or NH_4^+ depending on pH) which is subject to loss by volatilization especially if the slurry is broadcast on the soil or crop surface, which is the cheapest and most convenient method for farmers. Volatilization of

Fig. 5 Cover crop planted into a juvenile corn crop as it looks in the following spring in coastal British Columbia.

ammonia from land applied manure is one of the main nitrogen loss pathways from cropping systems to the environment; in the environment reactive forms of nitrogen (i.e., all chemical forms except dinitrogen, N_2) can acidify soil and water, create greenhouse gas emissions, reduce biodiversity and form fine particulates with significant human toxicity.

Many farmers abate ammonia loss by applying slurry to grass with a low emission method such as surface banding or by injecting into shallow (5-cm deep) trenches; both methods reduce emissions by maximizing contact of the manure, hence ammonia, with the soil. This approach also improves recovery of manure N by crops which improves yield and protein content of the herbage. However, even with reduced emissions, application of slurry for optimum crop production inevitably results in excessive loading of phosphorus which will accumulate in soils and be at risk for loss to waterways by runoff and leaching.

To help address the problems of low efficiency of both nitrogen and phosphorous in manure, a strategy has been developed which involves separating the dairy slurry into two products, one with low solids and high proportion of mineral nitrogen and the other with high solids and high ratio of phosphorus. This is done by first allowing the phosphorous-rich particles to passively settle to the bottom of the storage tank. The next step is to decant the thin nitrogen-rich supernatant and apply it with a low emission applicator to grass, to meet its high nitrogen requirement. Due to its low viscosity, this fraction rapidly infiltrates into the soil, thereby limiting ammonia volatilization and the saved nitrogen increases crop and protein production. And, importantly, there is little buildup of phosphorous in the soil as uptake by the grass is nearly equal to the amount applied. The final step is to use the phosphorous-rich sludge for corn by placing the sludge in furrows which were precisely 10 cm or less from the corn seed furrows. In practice, the corn is precision seeded near the sludge a few days after the sludge is injected. The precision injected sludge obviates the need for any fertilizer phosphorous and also reduces volatilization and the need for nitrogen.

Is it Sustainable?

There are no biological systems, natural or anthropogenic, that are hermitically sealed and even some natural systems are structured to benefit from the leakage of other systems. **Fig. 6** shows a field of alfalfa in Wisconsin where rising smoke illustrates worm channels and other macro-pores that are potential pathways for rapid nutrient loss especially in un-tilled crops.

For cropping system managers there is always concern that measures to reduce environmental impact may have unexpected side-effects or unintended consequences. As an example, **Fig. 7** shows that a sharp spike in emissions of a potent greenhouse gas, nitrous oxide (N_2O), followed the precision injection of slurry as described above. The emission spike is due to the moist and anaerobic conditions in the manure furrow which favors microbial denitrification of soil nitrate which often contributes to nitrous oxide emissions. This spike is obviously an unintended consequence. However the emission spike was controlled by adding a commercially available product which inhibits oxidation of ammonia to nitrate (NO_3) thereby curtailing N_2O formation. While the inhibitor was very effective in controlling emissions, unfortunately it did not result in more corn yield, so for now there is little economic incentive for farmers to purchase this product. However, where the technical tools exist there is always the option that policy makers will intervene to help make cropping systems more sustainable.

Is Yield Sustainable?

There are many factors that affect crop productivity so long term yield trends need to be carefully considered. **Fig. 8** shows yields from 23 years of corn hybrid testing on a dairy farm in the Lower Fraser Valley of British Columbia. The corn was grown under actual farm

Fig. 6 Rising smoke in this alfalfa field illustrates the presence of macropores that may be channels for nutrient loss from un-tilled fields. Photo courtesy of Michael Russelle.

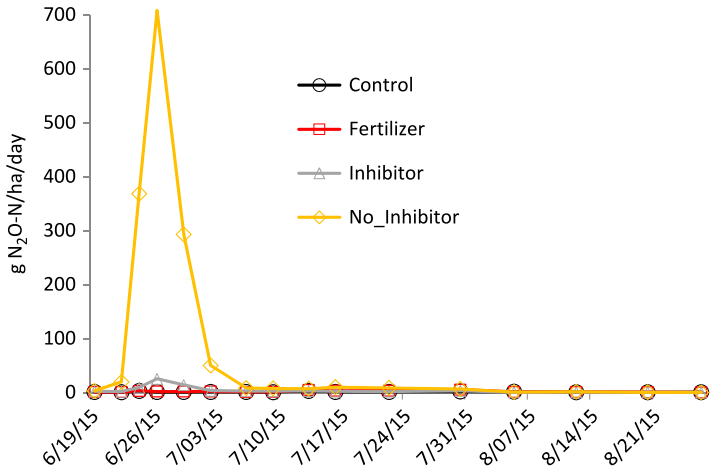


Fig. 7 Emissions of nitrous oxide from the soil surface following application of dairy slurry by two methods, broadcasting and injection. The spike in emissions after injection was controlled by adding a chemical which hinders the conversion of ammonium to nitrate (called nitrification).

conditions including rotation with grass, and received typical inputs of manure, fertilizer, and weed control measures. The unmistakable gradual rise in yield (mean of the top ten hybrids tested each year) is a reflection of a combination of many natural and manmade factors which cannot be teased out: crop genetics, farm equipment, soil quality, nutrient inputs, weed control practices, and climate. Also, often forgotten is the important aspect of farmer's crop husbandry skills. While we cannot know if yield will continue to rise, there is certainly no evidence of a threat to corn production on dairy farms in this region. The rise in crop yield and crop quality has supported consistently increasing milk production per land area over the past few decades in this region.

Conclusion: Cropping Systems and the Environment

Cropping ecosystems are intended to produce food efficiently and economically, and by their nature are more simple and homogeneous than most natural systems. However, this is mitigated by a number of factors. Diversity is added by tree-belts along field edges which provide habitat for wildlife including pollinators and help to sequester carbon along the field edges. There are now more often strips of natural or natural-like vegetation along riparian (streambank) areas that help prevent nutrient loss and protect fish habitat in streams. In some topographies, such as the prairie potholes, there are frequently ponds or "sloughs" which are important breeding grounds for migratory waterfowl. There is growing appreciation and adoption of crop rotations and even fairly complex multi species rotations with winter crops which have both agronomic benefits like break in disease cycles and cycling of nutrients, and ecosystem benefits such as soil biodiversity and health. Diverse crops may also support more varied wildlife. A study in coastal British Columbia showed that overwintering shorebirds (Dunlin, *Calidris alpine* L.) which are typically seen on the mudflats outside the dikes move to farm fields at night especially during high tide and stormy weather, and for the juvenile birds, the fields provide almost half their feed. The shorebirds clearly preferred bare fields due to supply of invertebrates and facility for detecting predators. In the same landscape it was determined that waterfowl preferred fields with grass or winter cereals while raptors preferred perennial shrubs and grass set-asides where they would find an abundance of rodents and roosting areas. Thus a diversity of crops, not within fields, supported a diversity of wildlife in this landscape. The open landscape, compared to the native cottonwood (*Populus trichocarpa* L.) forests, was also favorable for observing wildlife, a valuable ecosystem service to add to that of food security.

Epilogue: Cropping Systems in the Context of Society

While it is true that almost all our nutrition is derived from crop production, the corollary that all crop production is consumed for nutrients is not true at all. Currently topical is the use of scarce land resources to produce bioenergy crops ranging from corn grain for ethanol, to rapeseed (canola) for biodiesel, to whole corn for biogas, to *Miscanthus x giganteus* or switchgrass (*Panicum virgatum* L.) for direct combustion. No less controversial is the amount of cropland and resources (especially nitrogen and water) and the magnitude of environmental impact associated with crops used to feed farm animals, from the grain used to feed mainly chickens and pigs (and cattle) to non-edible forages used mainly for cattle and other ruminants. There is little doubt that these lands can be returned to more natural ecosystems. Since meat and energy consumption are largely discretionary, it must follow that large amounts of crops are not used to mitigate hunger or malnutrition. The situation is even more complex when we take into account the many crops we produce, often with high inputs, for amenities like wine, sugar and chocolate, but even vegetables like onions and coriander which we use mostly to flavor our food, and colored fruits to meet uncertain requirements for anti-oxidants. Even production of

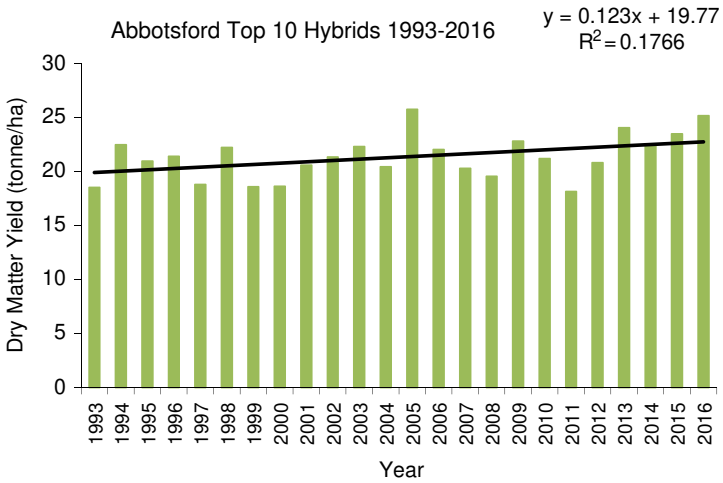


Fig. 8 Trend in yield of corn on a dairy farm between 1993 and 2016. Yield trend is due to crop genetics, management and climate.

vegetables in glasshouses is often more for their value as amenities than nutrients, and concentrations of glasshouses (and other protected cropping systems agriculture) are changing the land-cover in some regions to the possible detriment of natural species like migratory birds. The debate is over allocation of land, depleting inputs like phosphorous and natural gas, and environmental impact from pollution of air, fresh and marine waters, and even the soil itself (especially by salinization).

But, at the center of the discussion is water. Water is the ultimate paradox of a limitless renewable resource that is clearly depleting due to overuse, contamination and most of all, due to irregular climate (affecting both recharge and consumption). Water more than any other aspect of cropping systems is subject to society competition because of its many uses. It can be vital for food production, although the foods may not be vital themselves, but it is also needed for many of our other economic activities and comforts. In many parts of the world there is a direct competition for water and hence also growing pressure to clean it up. It is the worry about shortages that has led some countries to adopt new water recycling technologies in, for example, urban waste water treatment plants.

Mankind has probably always suffered during periods of drought, which have been tied to famines and migrations, and still are a force in our modern world. But one may ask why, during the Millennium Drought in Australia and the persistent multi-year drought in California, there was barely a ripple in the food supply, the almond wars in Fresno notwithstanding. This is an important question for society. To some extent this is thanks to globalization and the ability to move food around cheaply via marine waters. It is also due to very clever agronomic practices that have dampened the effect of water shortages. Water use is greatly improved with nitrogen inputs, and yields are enhanced if the crop can be planted rapidly to take advantage of more favorable spring water balances. Even crop rotations favor better exploitation of water from different soil depths. Plant breeders have improved yields with better plant architecture and pest resistance to take better advantage of agronomic strategies and inputs. Building organic matter helps soils hold water against powerful physical forces like evaporation and drainage. It is noteworthy that less successful have been the efforts to directly improve the physiology of drought tolerance and water use efficiency in plants. There have only recently been drought adapted corn hybrids licensed and they are only modestly better than standard cultivars. So, perhaps the message here is that we are not likely to have great breakthroughs in plant water use, either directly or indirectly. The future of our food security lies in a climate of changing weather patterns and changing human expectations. Hunger has become internationalized and solutions will likely be increasingly dependent on public policy which includes placing crop production systems within the panoply of human endeavors.

See also: Ecological Processes: Biological Nitrogen Fixation; Grazing. Ecosystems: Agriculture Systems. Terrestrial and Landscape Ecology: Integrated Farming Systems; Organic Farming

Reference

Berry, J.K., Delgado, J.A., Khosla, R., Pierce, F.J., Precision conservation for environmental sustainability. Journal of Soil and Water Conservation 2003; 58(6): 332–339. http://www.jswconline.org/content/58/6/332.

Further Reading

FAO, (2003), FAO World Agriculture: Towards 2015/2030—An FAO perspective www.fao.org/docrep/005/y4252e/y4252e06.htm.

Gaudin, A.C.M., Janovicek, K., Deen, B., Hooker, D.C., 2015. Wheat improves nitrogen use efficiency of maize and soybean-based cropping systems. Agriculture, Ecosystems and Environment 210, 1–10. doi:10.1016/j.agee.2015.04.034.

Urbanization as a Biospheric Process: Carbon, Nitrogen, and Energy Fluxes☆

Anastasia Svirejeva-Hopkins, Potsdam Institute for Climate Impact Research, Potsdam, Germany

Glossary

Anthropogenic Heat Flux The heat flux resulting from human activities such as traffic, heating and cooling of buildings, industrial processes and the metabolic heat release by people.

Exergy Exergy is the potential of a system to cause a change as it achieves equilibrium with its environment, i.e. is the energy that is available to be used.

Urban density The degree of people's concentration $D = P/S$, where P is population size and S is the area it occupies.

Urbanization dynamics The change of urban population and or area.

Urbanization (*Encyclopaedia Britannica)* This is the process by which large numbers of people become permanently concentrated in relatively small areas, forming cities and their suburbs.

Introduction

While the global population growth process has come to the stagnation, urban population continues growing and in some regions exponentially. The article will consider the ongoing process of urbanization and the interaction of urban areas with other compartments of the Vernadsky's biosphere. Hence, we must study the material and energy fluxes that take place. Cities are looked at as some entities that change the characteristics of land cover and the direction of matter and energy flows, not only on the territories they occupy but on much larger areas. As one can see from the above definition of urbanization, there are two aspects of it: the population and area, interconnected by the notion or urban population density. Population, P, and territory, S, of these formations are changing in time, so that urbanization is a dynamic process controlled by these main variables. The degree of people concentration is determined by the "population density," $D = P/S$, where P is population size and S is the area it occupies.

Different national statistics operate with different definitions and meanings of the terms "urban," "urban area," or "urbanized territory." For instance, the United Nations defines all places with $>20{,}000$ inhabitants living close together as urban, while the US Census Bureau uses "urban area" as a densely populated area (built-up area) with $D > 1000$ inhabitants per square mile and $P > 50{,}000$. Thus, the minimal urban area is equal to 50 square miles. At the same time, settlements with more than 2500 inhabitants are considered to be urban areas in the United States; while 2000 inhabitants living in contiguous housing form an urban area in France, and in the Netherlands it is municipalities with 2000 or more inhabitants. The highest urban density of 45,700 km^{-2} was registered in Dhaka (Bangladesh) (Demographia, 2009, 2017). Settlement densities can be orders of magnitude higher than agricultural rates, although residential densities in some urban areas are only marginally higher than the farmland densities in the most intensively cultivated agricultural areas (compare 2500 people per square kilometer in Los Angeles suburbs with 2000 peasants per square kilometer of arable land in Sichuan, China). However, maximum residential densities, *c*. 90,000 people per square kilometer (downtown of Hong Kong), translate into an anthropomass of 36 MJ m^2. This is roughly 200 times higher than the density of large herbivorous ungulates in Africa's richest ecosystem.

The definition of urban area in some other regions differs significantly from the above-mentioned one, with the concept of urban area somewhat based on the ancient structure of a city. All this shows a significant uncertainty in the term "urban(ized) area" and which is necessary to take into account in any quantitative estimations.

Past, Present, and Future of Urbanization

Two thousand years ago, there were about a quarter of a billion people living on our planet. The global population doubled to about half a billion by the 16th to 17th centuries. The next doubling required two centuries (from the middle of the 17 century to 1850, when the size of two European cities, Paris and London, exceeded 1 million inhabitants); the following doubling occurred just over the next 100 years, while the last one took only 39 years. The year 1650 is named as the start of "the urban explosion." Generally speaking, beginning from this date, the enormous population growth started.

☆*Change History:* March 2018. Anastasia Svirejeva-Hopkins slightly changed the title, hence added nitrogen flux description as well as exergy and index to make up for the "biospheric content." Introduction has been altered and Figure 1 has been updated with the latest data (2014). New sections Nitrogen Balance in Urbanized Territories, Cities: Sources or Sinks of Nitrogen?, Food-Print, Nitrogen Budget for the Paris Metropolitan Area, and Cities From the Thermodynamic Point of View: Sustainability of Cities and Exergy Maps are added. New Figures 5 and 6 and references are added.

This is an update of A. Svirejeva–Hopkins, Urbanization as a Global Ecological Process, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 3672–3678.

Nevertheless, it is a common practice to perceive the beginning of urbanization to coincide with the start of the agricultural revolution (7000–5000 BCE). It was at this time that nomadic hunters settled down and began to grow their food. A food surplus was created, and the division of labor made it possible to evolve gradually into the complex, interrelated social structures we know now as cities. The first cities were located along the Tigris and Euphrates Rivers (4000–3000 BCE) in contemporary Iraq, with urbanization then occurring also in Egypt, North Africa, India, China, Japan, and Europe—the Americas being the regions of most recent urbanization. Environmental factors were the major driving forces in the development of earlier cities. Fertile soils and easy access to water bodies, as well as adequate water supply, were essential. The first environmental disaster was triggered by the deforestation of the Middle East that led to soil degradation in the area, followed by a collapse of irrigations systems, and, as a consequence, to famine. Ancient cities were extremely dependent on the surrounding ecosystems, in particular, on agricultural lands. In Europe, since the 11th century, there has been a historical continuing flow of people from the countryside to the cities, although the "Black Death" in the 14th century impaired the process of urbanization. Europe recovered from the effect of this pandemic only by the middle of the 17th century, when the "urban explosion" occurred. Urbanization had also been occurring worldwide for at least two centuries. During the 18th century we have seen modern urbanization due to technological development, while earlier the process was driven by the migration of people from rural areas, since they were not needed in farming anymore.

However, in the last decades of the last century, we observed the unprecedented global population growth and the accompanying process of urbanization (Fig. 1).

Generally speaking, this enormous population growth was accompanied by other significant changes:

1. the rise of each person's ability to affect the natural environment through energy sources' exploitation;
2. the rise of the unevenness of the spatial distribution of people through development of "cities";
3. migration and travels' increase, while contacts between cultures also rise.

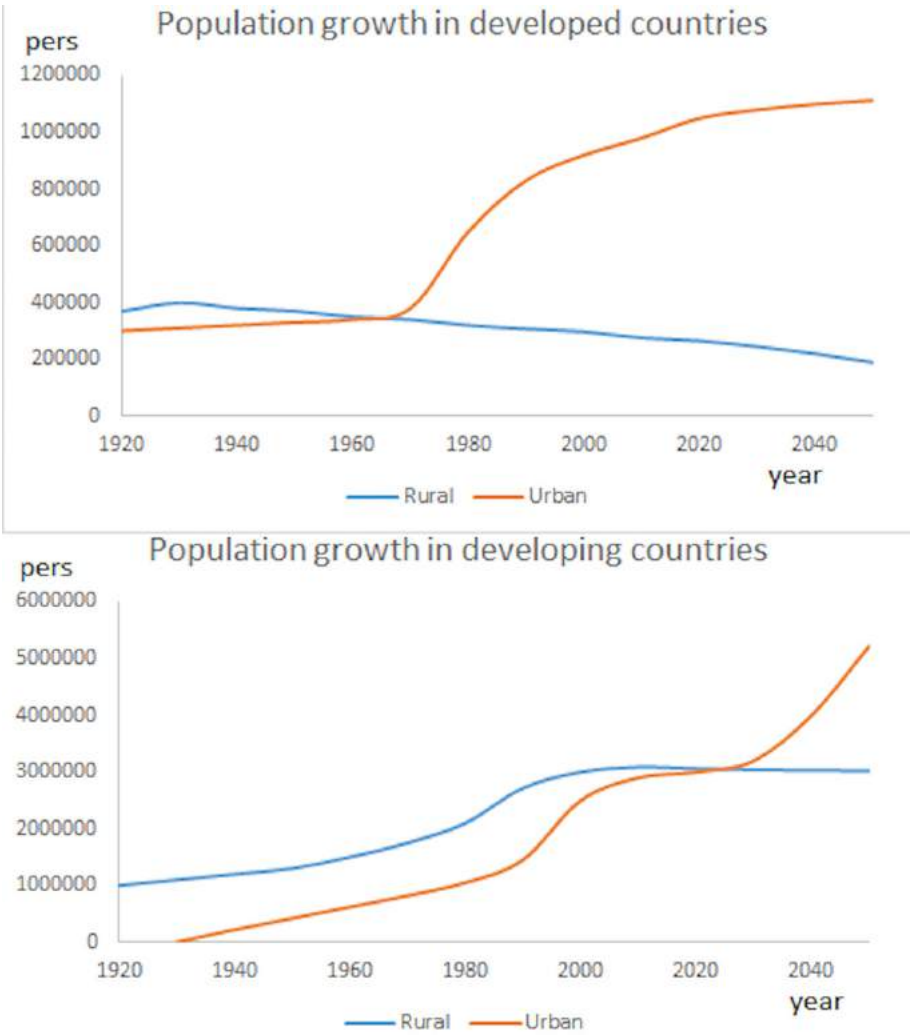


Fig. 1 Comparison of urban and rural population growth in developed and in developing countries (urban is defined as settlements of 20,000 people and above). Source: UN, World Urbanization Prospects, 1990, 1991 and 2014 revision.

Although only 12% of the world's population lived in urban centers in 1940, this percentage had risen to 33% by 1980. After World War II, a 2% urbanization rate was observed in the developed world, while it was almost 4% in the developing countries.

However, urban growth rates doubled that of the total population. And while the total population growth rate in the developed world has been decreasing, the urban population's proportion has increased from 55% to 70% of the total population. The major reason is the decline in rural population, as well as the arrival of new immigrants to the cities of some countries.

Since the middle of the last century, the following trend is observed in the percentages of urban population: 1950, 29%; 1960, 34%; 1970, 37%; 1980, 40%; 1990, 43.5%, and in 1995, 45%. However, the definition of what is urban varies greatly, which is why the estimations differ as well. For example, another source estimates 20% in 1950.

Urbanization growth rate has significantly left total population growth behind. From the year 1800 to 1990, the absolute number of city dwellers increased from 18 million to 2.3 billion, a 128-fold increase, while the total population has increased by only 6 times (from 0.9 to 5.3 billion). Furthermore, more than 1.4 billion city dwellers live in the less-developed world. If we look at the rates of urban area growth and compare them with urban population growth, we find that the first grows faster than the population, which in turn grows at a faster rate than the total country population, and this is a common phenomenon.

In 1990–95, the world's urban population grew by 2%–4% per year, while rural population only grew by 0.7% per year. Urban population increased by 2.1% during the 1995–2000 period, while rural population only by 0.7%. Today, 75% of the world's population lives in the less-developed countries, and 58% in Asia. In 1999, 19 urban settlements had 10 million or more inhabitants, and 47% of all people lived in cities. The number of "mega-cities," that is, giant urban agglomerations with a densely settled urban core of the original city, is increasing. Around this core, the satellite cities have grown, either planned or unplanned, linked to the central core by transport, communication, economic interdependence, and political-administrative structures. This tendency is confirmed by the following statistics: from 1950 until 1975, many cities with population of 5 million people have doubled in total urban population, while at the same time, cities with $<100{,}000$ people declined in their relative importance. In 1992, there were 23 M-cities with populations >8 million: 6 in the developed world (Tokyo, New York, Los Angeles, Osaka, Paris, and Moscow), and 17 in the developing world (from which 11 were in Asia). For most Asian cities, the shortage of water will be the most critical issue and is the limiting factor for the further growth of Beijing, Manila, Bangkok, Jakarta, and other cities. Also, while during the 19th century water and air pollution were associated with only a few larger cities, they are now becoming a global problem due to the rapid industrialization and the simultaneous concentration of people in cities.

While the past and current demographic situations are estimated more or less accurately, future dynamics are forecast with a very high level of uncertainty. The UN vividly illustrates that if current exponential and hyperbolic growth continues in each major region and at the current rates, then the population will increase by >130-fold in 160 years, from 5.3 billion in 1990 to 694 billion in 2150. However, eventually, the problem will be how to feed these people, since food and water limitations will certainly arise. The UN also shows that future global population size is very sensitive to the future level of average fertility.

Projections of global population dynamics are also uncertain, because external factors such as climate may change unexpectedly. Furthermore, even if external factors change as expected, the relationship between those factors and demographic rates may change.

We have the following hypothetical picture for the next half of the century. The global population will grow by 2–4 billion people, mostly in poor, but not rich, countries. It will also increase less rapidly than before and will become more urban than now. Hence, "from here on it is an urban world." Most of all, the additional people will be living in cities in poor countries, which can become an epidemiological danger. Population of the more developed countries will decline slightly, but increase substantially in less-developed countries.

In this century, global urban population will increase 1.8 times by 2020 (relative to 1990), while the total population will grow by only 1.4 times, and almost all population growth will be associated with cities in the developing countries. By the year 2030, the world urban population will reach 4.9 billion (1 billion in developed countries and 3.9 billion in developing countries). The global rural population will remain constant at 3.2 or 3.3 billion, although in the developed countries the rural population will decline. The trend in the developing countries is that rural population will slowly rise through the next couple of decades, reaching 3.1 billion, and then will slowly decline.

Present and Future Dynamics of Urban Areas

In 1985, 43% of the world population lived in cities while urban settlements covered just over 1% of the Earth's surface. In 1990, 50% of the global population of *Homo sapiens* inhabited $<3\%$ of the Earth's ice-free land area. However in the near future we may expect a further growth of the urban territories' area. World dynamics of this process is shown in **Fig. 2**. All these prognoses are based on two models: the first is a regression model, connecting urban population and urban area, and giving a minimal estimation, and the second, uses the spatial distribution of population density and gives a maximal estimation.

Note that the dynamics of urban areas are significantly different in the major world regions (**Fig. 3**). For instance, the fast growth in African region differs from almost constant dynamics in the highly industrialized countries.

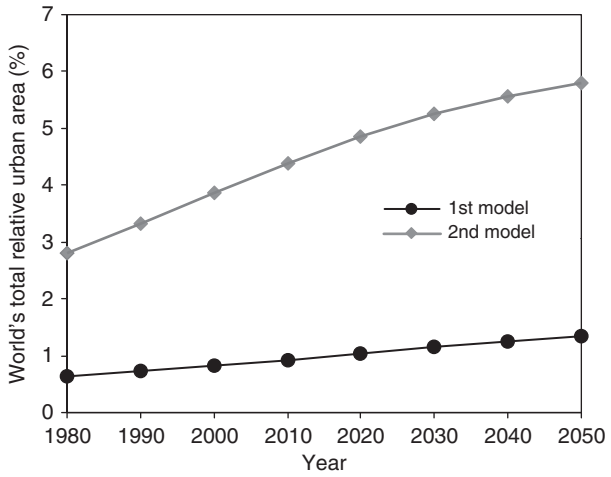


Fig. 2 Dynamics of the relative world urban area (in percentage of the total world area) between 1980 and 2050.

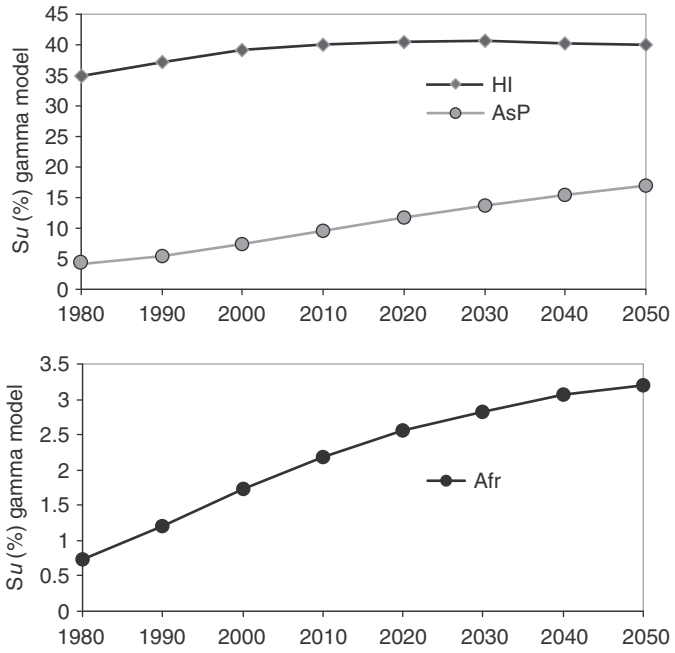


Fig. 3 Dynamics of the relative urban areas (% of the total regional area) for the three world regions: HI, highly industrialized European; AsP, Asia and Pacific; and Afr, African.

The City as a Specific Heterotrophic Ecosystem

From an ecological point of view, any city is a heterotrophic system maintained by external inflows of energy, food, water, and other substances. Thermodynamically, any city (and generally, any urbanized territory) is an open system that is far from thermodynamic equilibrium. All matter and energy needed for a city's functioning are collected from external territories that are significantly larger than the area of the city itself and very often are located quite far away.

The heterotrophic ecosystem "city" differs very much from a natural heterotrophic ecosystem. In fact, a city has a more intensive metabolism per area unit, requiring a significant inflow of artificial energy. Its consumption per urban area unit may be 3–4 orders of magnitude higher than the same for rural area. For instance, the annual subsidies in fuel, fertilizers, labor, etc., required to maintain a lawn in the Madison metropolitan area (Wisconsin, USA) is equal to 22 GJ ha^{-1}, which is approximately equal to the artificial energy input for a maize field. During the process of its own metabolism, a city consumes large amounts of various materials: food, water, wood, metals, etc., all that we call "gray energy." Products of city's metabolism have larger volumes of, and more toxic, substances than the same of natural ecosystems.

If we compare cities and natural forests in Wisconsin, USA, we can see that the number of species in a city forest (75 tree and 74 bush species) is more than in natural one (10 tree and 20 bush species). The annual net production and the amount of living biomass (in

carbon units) in city's forests are equal to 500 t C $km^{-2}year^{-1}$ and 7000 t C km^{-2}; while in a natural forest the corresponding values are 400 t C km^{-2} $year^{-1}$ and 13,000 t C km^{-2}. The greater values of species diversity and production are provided in city's forests at the expense of additional inflow of "gray energy." For instance, the annual import of fertilizers is about 140 t km^{-2}.

While natural forest is almost a closed system, a city forest is a typical through-flow system with "gray energy" input and output in the form of dead organic matter: about half of the annual accretion is exported from the city to waste treatment plants. This is one possible explanation why the amount of biomass in the city forest is lesser than in the natural ecosystem. The carbon storage in urban forests with their relatively low tree cover (25.1 t C ha^{-1} in average for United States) is less than in natural forest stands (53.5 t C ha^{-1}). The gross sequestration rate, that is, the fraction of the gross annual production accumulated in wood, in urban forests is equal to 0.8 t C ha^{-1} $year^{-1}$, which is also less than in natural ones (for instance, 1.0 t C ha $year^{-1}$ for a 25-year-old natural regeneration spruce–fir forest with 0.1 kg C m^{-2} cover), although the difference is insignificant. However, on a per-unit tree cover basis, carbon storage by urban tree and gross sequestration may be greater than in natural forests, 92.5 t C ha^{-1} and 3.0 t C ha^{-1} $year^{-1}$, due to a larger proportion of large trees and the more open structure (that leads to the weakness of competition) in urban forests.

Environmental Effect of Urbanization and Ecological Footprint

Retrospectively, most humans have lived in small settlements dispersed within larger biomes (Fig. 4). We see there is a certain correlation between the type of biomes and the degree of urbanization, some biomes being more preferable for cities.

The growth of cities in these "patches," absorbing and transforming nearby natural ecosystems and agricultural lands, modifying energy and matter flows, typical for these ecosystems, negatively influences the local and regional biodiversity, increasing fragmentation of large rural areas and natural zones. This process (and contamination of the atmosphere) leads to the changing of the nature of land surfaces and near-surface atmospheric layer, and therefore its reflection and absorption of solar radiation and aerodynamic properties. This in turn leads to raising urban temperatures and the changing of the local climate, creating so-called "urban heat islands," which is warmer by 1–2°C than surrounding territories. The "urban heat island" effect occurs mainly at night, when the buildings, etc., release heat absorbed during day.

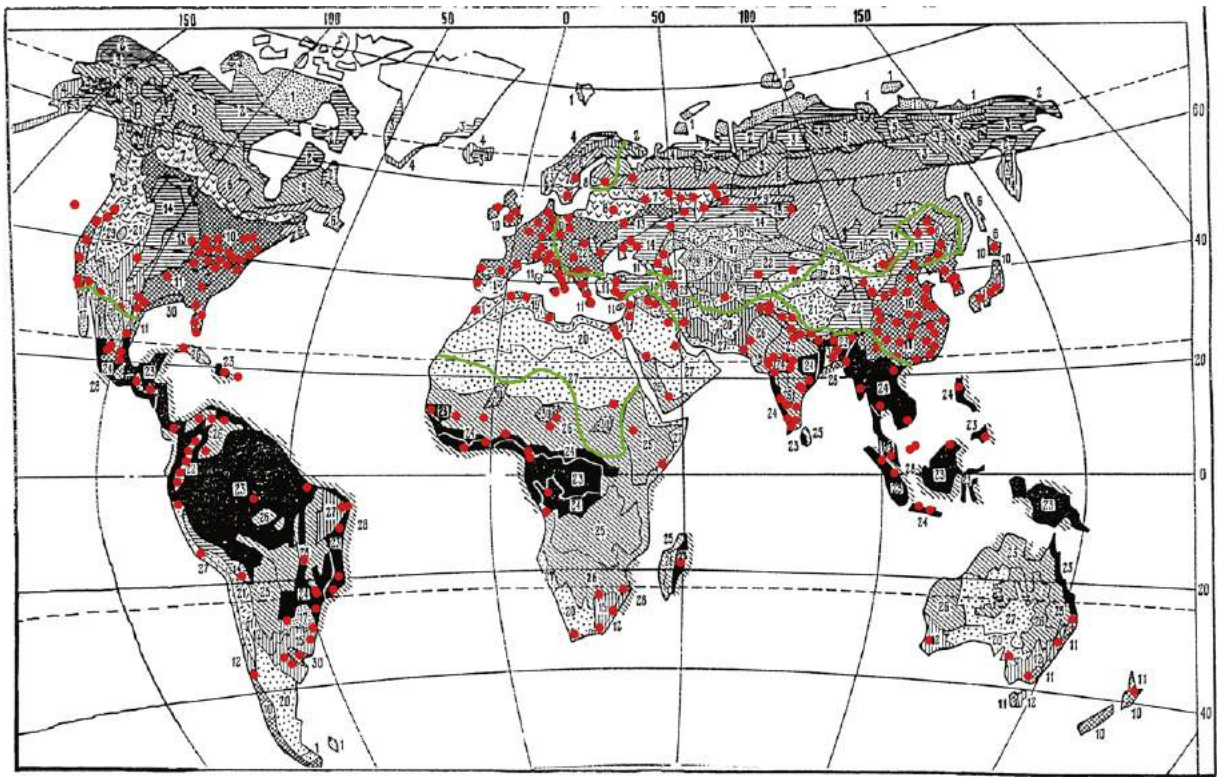


Fig. 4 Different types of global vegetation (biomes). Red dots represent the concentration of cities; the green line is the border of the regions. 1, Polar desert, polar tundra; 2, tundra; 3, mountainous tundra; 4, forest tundra; 5, north taiga; 6, middle taiga; 7, south taiga; 8, temperate mixed forest; 9, aspen–birch lower taiga; 10, deciduous forest; 11, subtropical deciduous and coniferous forest; 12, xerophyte woods and shrubs; 13, forest steppe; 14, temperate dry steppe (including mountainous); 15, savanna; 16, dry steppe; 17, sub-boreal desert; 18, sub-boreal saline "desert"; 19, subtropical semidesert; 20, subtropical desert; 21, mountainous desert; 22, alpine and subalpine meadows; 23, evergreen tropical rainforest; 24, deciduous tropical forest; 25, tropical xerophyte woodland; 26, tropical savanna; 27, tropical desert; 28, mangrove forest; 29, saline land; 30, subtropical and tropical woodland and Tugay shrubs.

In addition, metropolitan agglomerations influence the local and global environment through their consumption of nonnative resources and their concentrated production of waste and consumables.

The footprint is the quantitative conversion of the material and energy flows required to support human population in cities into the land area required to produce these flows. Although cities occupy a relatively small area on the planet, they are the dominant human ecosystem and the ecological space taken up by humans as a species is much higher. Every city depends (for its existence and growth) on a globally diffuse productive hinterland up to 200 times the size of a city itself. To illustrate this, let us examine the following case studies.

For instance, one person of the US urban population consumes daily (1) food produced by 0.75 ha of agriculture land, (2) paper and wood by 0.4 ha of forest, (3) water by about 7.5 m^3. So, a 1 million-populated city with the population density of 4000 persons per kilometer (city area is 250 km^2 correspondingly) needs a significantly larger area for its support: 7500 km^2 of agriculture land and 4000 km^2 of forest. A rather large river watershed (presuming abundant precipitation) can provide inflow of 7.5 million m^3 of water daily. If we take into account that the area of such watershed is about 15,000 km^2 then the total footprint area will be about 26,500 km^2, that is, 106 times the city area.

The city of Vancouver (Canada) had in the year 1991 a population of 47,2000 living in an area of 11,400 ha. If we assume that the per capita land consumption rate is 4.3 ha, then the people in this city would require 2.03 Mha of land. Hence, the inhabitants would require a land area 180 times larger than their habitat. Furthermore, adding a marine footprint of 0.7 ha per person, the total area needed to support the city becomes 2.36 Mha, or 200 times larger than the geographic area of the city. For London, the equivalent footprint is 120 times the area of the city itself. The New York metropolitan area annually consumes the equivalent of 800,000 ha of wheat, or approximately the total amount of wheat grown yearly in the state of Nebraska.

So, if we presume that the world urban area constitutes 1% of the total land area, and the footprint is 100 times this area (note that these are minimal estimations), even in this case the urban footprint exceeds all the world land area.

Carbon Balance in Urbanized Territories

Therefore, we can say that although the total area of urbanized territories is relatively small (1%–2% in 1990s), they play an ever-increasing role in global change in general and in the global carbon cycle, the main biogeochemical cycle of the biosphere, in particular.

Urban areas emit (in accordance with different estimations) 78%–97% of the total anthropogenic carbon emission. Up to 60% of this emission comes from the transportation and building sectors, while the rest are from industry. Of course, all of these emissions are "spread" and mixed in the entire atmosphere over 3–4 months period, but they are generated in particular by urban point sources.

Cities transform the natural territories they occupy, partially obliterating vegetation and soil, partially modifying them. Similarly, urbanization changes the structure and function of the local carbon flows within these territories. Note that the process often involves considerably larger territories than the exact city areas.

Cities consume a lot of organic carbon in the form of food and other agricultural products, as well as wood, etc., produced, as a rule, far from the urban territories, transforming them into other forms of carbon (feces, exhaled CO_2, residues of food processing, dead organics of "green zones," etc.) in the process of urban and purely physiological human metabolism. In other words, cities destroy the spatial entity of the processes of production and decomposition of living matter that is typical for natural ecosystems. Note that this entity provides the closure of any local carbon cycle.

Urban territories have more carbon stored per unit area than natural ecosystems. Organic carbon is stored in soils and vegetation of urban territories, but also in people and pets; nonorganic carbon includes carbon transported into the cities and stored in buildings, etc., but most of this carbon is transformed into waste. Processes similar to those in peatlands accompany carbon fluxes from mineralization, incineration (rapid oxidation of carbon), and landfilling of solid waste. For instance, the global input of carbon into solid waste (sludge and industrial waste) is estimated to be 0.16 Gt year $^{-1}$ (1 Gt 1/4 $\times$ 10^9 t).

Long-term organic carbon in urban territories is accumulated in

1. *Biomass in humans and animals*: For the world population of 6 billion, the total amount of carbon is equal to 45 million tons of carbon that constitutes about 10% of the total biomass of land animals. A biological metabolism of 6 billion people is accompanied by exhalation of 0.34 Gt C $year^{-1}$ and secretion of 0.18 Gt $year^{-1}$ with feces and other discharges that, respectively, give a total 0.52 Gt C $year^{-1}$. The value is entirely comparable with components of the global carbon cycle. For instance, soil erosion releases 0.98 Gt C $year^{-1}$.
2. *Biomass in trees and other plants*: The mean global value of living plant biomass in cities is 3500 t C km^{-2}, and the mean net primary production is 500 t C km^{-2} $year^{-1}$. By taking the global urban area in 1980 as 2 $\times$ 10^6 km^2 and assuming that 50% of it is covered by city vegetation, we find that urban territories contain 3.5 Gt C in living vegetation biomass, while a global figure for net plant assimilation of carbon in urban territories is approximately 0.5 Gt C $year^{-1}$.
3. *Carbon in construction material, furniture, books*: Extensive amounts of carbon are accumulated for long time period in building constructions, furniture, books, and other articles made of organic materials. For instance, *c.* 3 Gt C is fixed in houses in the whole of Europe, North America, Japan, and Australia, and about 0.4 Gt C in other regions.
4. *Carbon in solid waste*: Most products of forestry and agriculture are turned eventually into waste. Solid waste is either deposited in sanitary landfills or incinerated. Carbon stored in landfills experiences slow decomposition rates and is gradually released due to microbiological activity. The annual world solid-waste production is equal to 170–180 million tons of carbon.

Approximately 60–70 million tons is released into the atmosphere by burning, while about 110–120 million tons is deposited in landfills followed by slow release into the atmosphere.

Landfills are often regarded as long-term accumulators of carbon and in this respect can be compared with natural peatland ecosystems (even after 30 years one-third of the organic carbon remains nonmineralized). This carbon is bound in long-lived humus and has not been mineralized for a very long time.

Nitrogen Balance in Urbanized Territories

Let us have a look at another important element in the biosphere, nitrogen, and study how "urban compartment" changes its fluxes. It is important to mention that we consider the so-called reactive Nitrogen (Nr) and that in its case, and as compared to carbon, there is downscaling to the level of landscapes. Cities import and concentrate Nr in the form of food and fuel, and then disperse it as air and water pollution to other ecosystems covering much larger areas. Certainly, they also dramatically change the global nitrogen cycle, but it is difficult to quantify these changes at that level. In detail, firstly, urban lands fix substantial amounts of atmospheric N_2 to Nr as NO_x through the high temperature combustion of fossil fuels. Secondly, they drive the industrial fixation of Nr to fertilizers, importing the Nr produced in food for burgeoning urban populations, subsequently dispersing it in air and water to other ecosystems over much larger areas than the cities themselves.

In other words, cities act as concentrators, transformers and dispersers of nitrogen, therefore acting as new entities of the Earth System, or compartments of the biosphere.

Cities: Sources or Sinks of Nitrogen?

What are the important issues concerning N management in a certain city, or cities of a certain region? Cities can be characterized either as a source of Nr (i.e., emitting large amounts as liquid or solid household waste, automobile exhaust, air pollution from power plants) or a sink of Nr (either through importing more food, fossil fuels, etc., and/or having fewer emissions to the air and water). Let's have a look at Paris Metropolitan Area for a more in-depth quantification of terrestrial and atmospheric fluxes, and input and output to the system-city. Besides, by constructing mass balances at scales from the household to the city, human choice can be linked directly to biogeochemical cycling.

Paris Metropolitan Area changed from being a sink in the 18th and 19th centuries to a source of Nr today. Major changes in the city functioning occurred before 1950, but especially recent decades have been characterized by an unprecedented amplification of those changes.

1. The major part of Nr output is attributed to the combustion of fossil fuels for transport and energy, which converts both atmospheric N_2 and fossil Nr to reactive NO_x. The second largest Nr contribution comes from incineration of solid waste, and third highest emissions come from sewage water treatment plants.
2. Urban wastewater discharge into rivers largely contributes to N contamination of the aquatic environment, although sophisticated and expensive tertiary treatment techniques are now available to drastically reduce Nr emissions.
3. Denitrification in urban landscapes is controlled by the presence of water bodies and green areas. These areas have lower biomass and decomposition rates than natural ecosystems.
4. To achieve sustainability of urban systems in relation to the N cycle, road transport of goods and passengers has to be reduced, household waste generation minimized, and wastewater treatment improved.
5. More attention should be given to future sewage processing systems that process Nr (and other nutrients) for reuse as a fertilizer rather than losing the Nr resource by denitrifying it back to N_2.
6. Such measures could eventually turn urban areas from sources of Nr to N-neutral or even N-sink areas. Regional adaptation measures should be specifically tailored to the individual urban ecosystems of Europe.

Food-Print

Apart from ecological footprint, there is so-called "food-print" indicator that takes into account the area required for producing agricultural goods, being expressed in terms of the effective surface of the surrounding territory needed to support the food requirements of a city. It was found, for example, that despite a 50-fold increase in the population of Paris since the 15th century, the food-print of the city barely increased twofold. In contrast, the further doubling of the population in the 20th century was paradoxically accompanied by a fivefold decrease in the food-print, because of the intensification of agricultural production. However, "food-print" is only a partial indicator of the impact of a city. For example, it does not include a comprehensive account of all resources required, such as energy, fertilizers and waste disposal.

Nitrogen Budget for the Paris Metropolitan Area

The urban nitrogen budget (balance) can be considered as a subset of national and regional N budgets. It incorporates imports of Nr-containing products into the city, their conversion within the city boundaries and exports outside of the urban

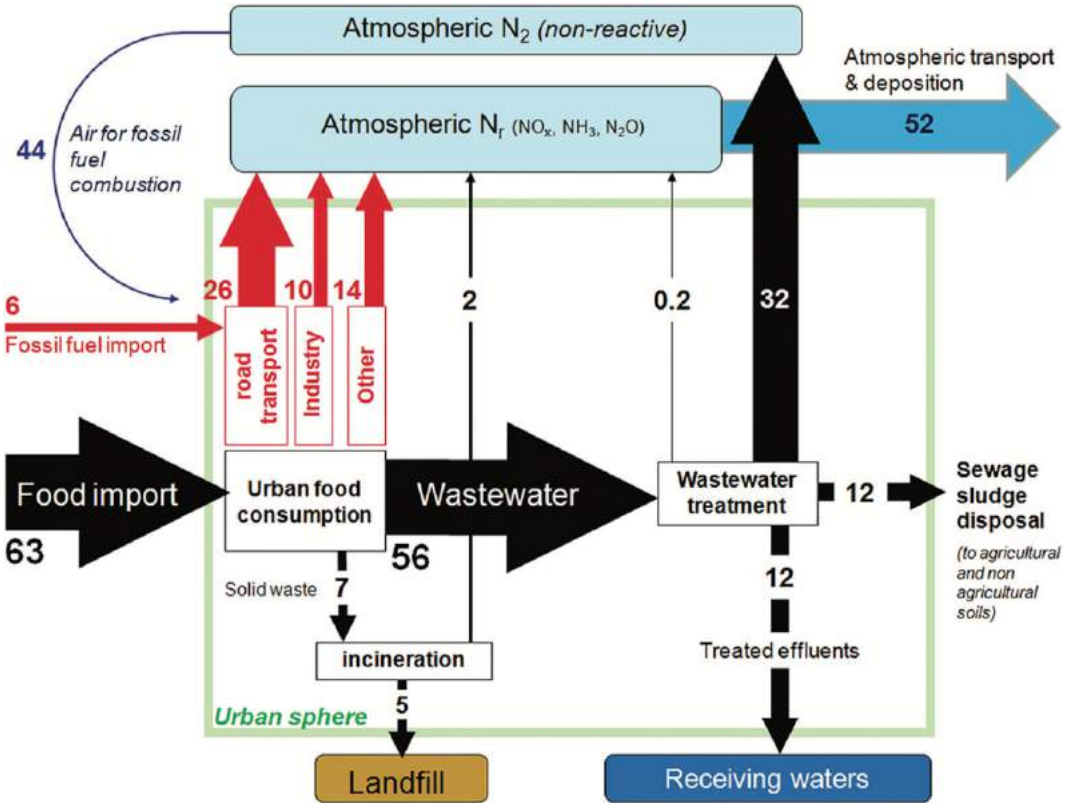


Fig. 5 Nitrogen flows quantified for the Paris Metropolitan Area for the year 2006 (PAM, numbers in Gg N $year^{-1}$). The quantified fluxes displayed reflect major N flows through the PAM originating from food import and fossil fuel use, as well as N_2 out-flux from wastewater treatment. From Svirejeva-Hopkins, A., Reis, S., Magid, J., Nardoto, G. B., Barles, S., Bouwman, A. F., et al. (2012). Nitrogen flows and fate in urban landscapes. In: M. A. Sutton, et al. (eds.), *The European nitrogen assessment: Sources, effects and policy perspectives*. Cambridge: Cambridge University Press.

sphere. The urban sphere incorporates the three dimensional space surrounding the urban habitat and spans all environmental media, water, air, and soil.

We made the first step by creating a detailed Nr mass balance for Paris and its urbanized surroundings in order to estimate the magnitude of major fluxes across the urban landscape and to see how N cycling varies among urban system components. It will help to determine which budget terms are most open to management in order to reduce N pollution to recipient systems. The budget is shown in Fig. 5.

The Paris Metropolitan Area is a source of Nr, emitting in total the amount of 50 Gg $year^{-1}$ to the atmosphere, the major part being attributed to the emissions from transport and energy. Although much smaller, emissions of Nr to air from the incineration of solid waste are also substantial, contributing 2 Gg $year^{-1}$. The amount emitted to the aquatic environment, at about 12 Gg N $year^{-1}$, greatly depends on the type of wastewater treatment adopted. Disposal of solid wastes and incineration residues in landfills or of sewage sludge on agricultural and nonagricultural soils (potentially leaking to the ground water over time), together amount to 17 Gg N $year^{-1}$.

Regarding the transformations between N_2 and Nr, the largest of these occurs outside this budget, in the production of fertilizer Nr to provide food. Overall, the food Nr import of 63 Gg $year^{-1}$ is of a similar order of magnitude to the inadvertent fixation of N_2 to Nr through combustion processes.

In the case of Nr from combustion processes, all of this is exported from Paris as Nr. Thus four times as much Nr is released to the environment of Paris from combustion processes than from the Nr originating in food imported to Paris. If this highlights the problem of Nr emissions from combustion sources for this city, it should not be forgotten that the Nr, denitrified in wastewater treatment, represents the loss of a valuable resource. Without commenting here on the economic viability of recycling wastewater Nr, it may simply be noted that, at an indicative value of €1 per kg fertilizer Nr, the denitrification of wastewater Nr in Paris represents an annual resource loss of €32 million per year.

Urban Areas From the Thermodynamic Point of View: Sustainability of Cities and Exergy Maps

Urban areas "modify" the surrounding landscape by changing the fluxes of matter and energy. The urban area is viewed as an open thermodynamic system, maintaining its structure through the conversion of energy. During its evolutionary process including

growth (increase in size) and structural development (increase in the system's organization), the system-city fluctuates from one state to another and its exergy grows and becomes the highest when the system is farthest from an equilibrium state that has local maximum of entropy.

We estimate numerically this state (and the degree of nonequilibrium) by recalculating standard satellite data, deriving and combining global maps of entropy, information gain and exergy, as well as proposing a new approach for the study of urban sustainability questions. Human decisions are also affected by various ecological factors, so that urban structural organization then determines the parameters of the thermodynamic system and its many specialized sub systems and forms of energy transfer. Spatially, cities constitute a mosaic of open, built-up and vegetated spaces, which creates high internal gradients of entropy, exergy and heat flux. Urban areas are also known as the biggest producers of the anthropogenic heat flux (AHF, or thermal pollution). The certain Index (anthropogenic heat flux/exergy) is chosen to serve as the measure of the degree of disturbance in urban areas at certain geographical locations, which, recalculated per capita, reflects the level of urban infrastructure in cities of different regions.

Fig. 6 shows the global map of annual average exergy in the year 2002 overlaid by the 23 largest urban areas. The highest values of exergy are attributed to the tropical forest zones with the highest biomass density.

AHF/Ex shows the degree of urban system disturbance, calculated "per million inhabitants", since that better reflects the level of a city's infrastructure/energy efficiency. The lower the index, the less disturbed is the system.

We then overlay this map with the map of Anthropogenic Heat Flux (*AHF*) for 2005, or thermal pollution (http://www. cgd. ucar.edu/tss/ahf/). Preliminary analysis shows that New York is quite energy efficient, especially when the *AHF/Ex* index value is compared to Seoul, considering that their populations are the same. Osaka-Kobe urban area is also similar to Seoul in terms of index value, while Mumbai's value is the lowest. We notice that there is no relationship between population size, energy and thermodynamic characteristics of cities and attribute the thermodynamic characteristics to the structural characteristics of urban areas. Exergy of cities mainly depends on their geographical location. Tropical cities have larger values of exergy in comparison with temperate and subtropical cities, as they are situated in more unstable and environmentally sensitive zones. The largest heat flux is produced by cities of developed countries, particularly those that are concentrated in the temperate zone (with the exception of New York), while the smallest values naturally exist in the least developed cities (e.g., Kinshasa, Lima, Lagos). Similarly, when we recalculate the index of disturbance per million inhabitants, we see that the least disturbance is found in undeveloped countries and the greatest levels of disturbance are in cities of developed countries (Paris, London, Osaka), most likely due to the long history of human impacts. Only one city from of the developed countries, New York, is in the least disturbed state, while BRIC countries (Brazil, Russia, India, and China) have medium values. However, this method cannot answer the question about the actual ways of degradation, that is, it only shows the possibility of degradation and estimates the value of the entropy over-production (energy of disorder) that can eventually destroy the system—but it cannot show the pathways along which this destruction will really occur.

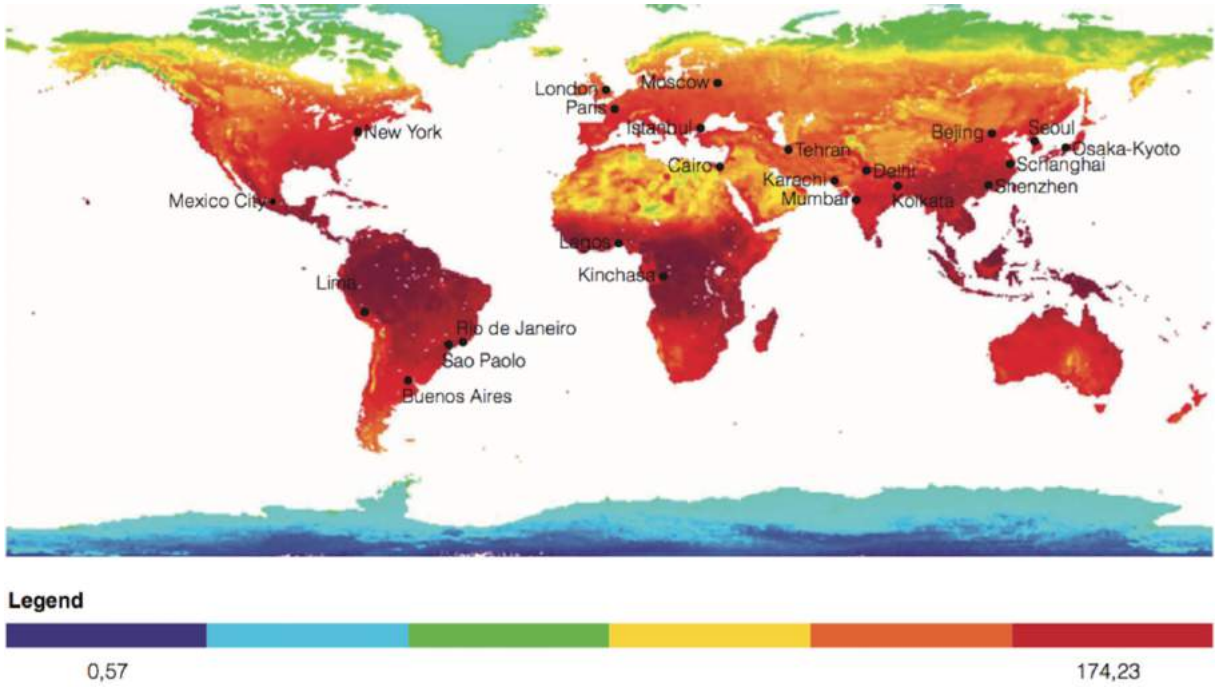


Fig. 6 Annual average exergy W/km^2 and 23 largest UA locations (Demographia, 2009, 2017). Svirejeva-Hopkins, A., et al. (2011). City systems, their growth and sustainability: An urban biogeochemistry approach. UGEC Viewpoints No. 5, April 2011, www.ugec.org.

Conclusions

Urbanized territories dominate the surrounding environment in a number of ways—the growth of cities, and absorbing and transforming nearby natural ecosystems and agricultural lands. This process leads to the changing of the nature of land surfaces. Therefore, we can say that although the total area of urbanized territories is relatively small (1%–3%), they play an ever-increasing role.

We can summarize this influence as the following:

- Cities transform the natural territories they occupy, partially obliterating vegetation and soil, partially modifying them. By the same token, urbanization changes the structure and function of the local carbon flows within these territories. Note that the process often involves considerably larger territories than the exact city areas.
- Cities consume a lot of organic carbon in the form of food and other agricultural products, as well as wood, etc., produced, as a rule, far from the urban territories, transforming it into other forms of carbon (feces, exalted CO_2, residues of food processing, dead organics of "green zones," etc.) in the process of urban and purely human metabolism. In other words, cities destroy the spatial entity of the processes of production and decomposition of living matter that is typical for natural ecosystems. Note that this entity provides the closure of any local carbon cycle.

See also: General Ecology: Biomass; Demography. Human Ecology and Sustainability: Urban Systems; Ecological Footprint; The Anthropocene; Carbon Footprint; Nitrogen Footprints; Human Population Growth; Urban Metabolism

References

Demographia, 2009. World Urban Areas. In *Database*. www.demographia.com
Demographia, 2017. World Urban Areas. In *Database*. www.demographia.com

Further Reading

Billen, G., Barles, S., Garnier, J., *et al*., 2009. e food-print of Paris: Long-term reconstruction of the nitrogen flows imported into the city from its rural hinterland. Regional Environmental Change 9, 1436–1498.
Brundtland's World Commission on Environment and Development, 1987. Our common future. Oxford: Oxford University Press.
Encyclopædia Britannica Inc., 2005. Encyclopædia britannica. CD-ROM/Ultimate Reference Suite, DVD.
Hauser, J.A., 1992. Population, ecology and the new economics: Guidelines for a steady-state economy. Futures 24 (4), 364–387.
Heinke, G.W., 1997. The challenge of urban growth and sustainable development for Asian cities in the 21st century. Environmental Monitoring and Assessment 44, 155–171.
Miller, G.T., 1988. Living in the environment, 6th edn Belmont, CA: Wadsworth.
Small, C., and Cohen, J.E., (1999) Continental physiography, climate and the global distribution of human population. In: Svirezher Yu (ed.) Proceedings of the International Symposium on Digital Earth, pp. 965–971.Beijing: Chinese Academy of Science.
Svirejeva-Hopkins, A., Schellnhuber, H.-J., 2006. Modelling carbondynamics from urban land conversion: Fundamental model of city in relation to a local carbon cycle. Carbon Balance and Management 1, 8.
Svirejeva-Hopkins, A., Schellnhuber, H.-J., Pomaz, V.L., 2004. Urbanised territories as a specific component of the global carbon cycle. Ecological Modelling 173, 295–312.
Svirejeva-Hopkins, A., et al., (2011). City systems, their growth and sustainability: An urban biogeochemistry approach. UGEC Viewpoints No. 5, April 2011, www.ugec.org.
Svirejeva-Hopkins, A., Reis, S., Magid, J., Nardoto, G.B., Barles, S., Bouwman, A.F., *et al*., 2012. Nitrogen flows and fate in urban landscapes. In: Sutton, M.A., *et al*. (Eds.), The European nitrogen assessment: Sources, effects and policy perspectives. Cambridge: Cambridge University Press.
United Nations (UN) 1999. Prospects for urbanization—1999. Revision. New York: United Nations, ST/ESA/SER.A/166, Sales No. E.97.XIII.3.
United Nations (UN) 2000. The state of the world cities 2001, 121 pp., Nairobi: United Nations Centre for Human Settlements (UNCHS).
World Urban Areas (2009, 2017) database www.demographia.com.

Relevant Websites

https://unhabitat.org/books/global-urban-indicators-database/—United Nations.
http://www.demographia.com—Demographia.
www.ugec.org—Urbanization and Global Environmental Change.

Water Cycle☆

Zbigniew W Kundzewicz, Polish Academy of Sciences, Poznan, Poland

Introduction

Water resources of our planet take part in an infinitely recurrent water cycle. It is the largest movement of matter in the Earth's system. The Earth is indeed a blue planet, since the oceans cover nearly 71% of its surface, that is, over 361 million km^2, while the continents and islands—the solid surface of the Earth—make up only 29% of the total Earth area. The hydrosphere includes all water on the planet and is interconnected with all the other "spheres" in the Earth system, that is, the geosphere (lithosphere and atmosphere), biosphere, and human-related anthroposphere.

Water is the most widespread substance in the natural environment of our planet. It is available everywhere on Earth, albeit its abundance largely differs in space and time. Water molecules take one of three states—liquid, solid, and gaseous (vapor), with liquid state being most commonly occurring in the Earth's conditions. Water undergoes phase changes: from liquid to gaseous phase—by evaporation (evapotranspiration); from gaseous to liquid phase—by condensation; from liquid to solid state—by freezing; and from solid to liquid state—by thawing. Direct phase change between the solid and the gaseous phase is also possible, in the process of sublimation.

Water is the basic element of the life-support system of the planet. It is a constituent in plant and animal tissues. Water transfer plays an essential physiological role in human organisms—for example, physical workers in warm climate lose water and drink water in the same time. Due to the thermoregulation mechanism of the human body, they are sweating and, simultaneously, feeling thirst.

Water cannot be substituted by any other substance. By its capacity to dissolve substances, water plays an essential role in the chemistry of life. Water is an excellent solvent, able to dissolve many chemical compounds, for example, mineral salts. It also plays a substantial role in biogeochemical cycles of carbon, phosphorus, nitrogen, as solvent and carrier. Water interacts with both the atmosphere and the lithosphere, acquiring solutes from each. Water cycle contains natural purification mechanisms. Evaporation purifies (distills) salty oceanic water. Evaporate is freshwater, but salt remains in the oceans as water evaporates. Moreover, water self-purification takes place in rivers and wetlands.

Movement of Water Between Stores

There is water in the hydrosphere (oceans and seas, polar ice, lakes, rivers and streams, wetlands and marshes, snow pack and glaciers; containing liquid and solid water) and in the lithosphere (solid Earth), under the Earth's surface (in the rocks and soil, including permafrost, and deeper in the ground, down to the Earth crust—in liquid, solid, and gaseous phases), in the atmosphere and in the biosphere (living organisms, flora, and fauna).

The total global water resources constitute approximately 1.385 billion km^3 (0.17% of Earth's volume) but only 2.5% of global water resources are fresh. Water is the most abundant substance at the Earth's surface, with 96.5% of its volume (1.338 billion km^3) contained in salty oceans. Oceans, the largest water store, play an essential role in the water cycle as the main source of water for the atmosphere. Other water stores on Earth contain much smaller volumes. Glaciers and permanent snow cover contain 24.3 million km^3 of water, that is over 50 times less than the ocean volume (c. 1.72% of global water resources). However, this solid water store (whose prevailing part is ice and permanent snow cover in the Antarctic, the Arctic, and mountainous regions) contains freshwater, making up most (69.6%) of the total freshwater resources. The third largest global water store is groundwater (23.4 million km^3), but more than half of it is not fresh. Even if frozen hydrosphere (cryosphere) is the largest reservoir of freshwater, groundwater is the largest available source of freshwater. All the lakes on Earth contain 176,400 km^3 of water, with freshwater constituting more than half of the total volume. Approximately 16,500 km^3 of water is stored in the soil (0.05% of total freshwater), while all the rivers of the world carry, on average, in any time instant about 2120 km^3 of water, being only 0.006% of total freshwater. The atmosphere itself stores approximately 13,000 km^3 (0.001% of total water, 0.04% of total freshwater) and

☆*Change History:* March 2018. Z W Kundzewicz updated all the sections in this chapter, as well as Fig. 1.

wetlands about 11,500 km^3 of water. Biological water has a global volume of 1120 km^3 that is 0.0001% of total water and 0.003% of freshwater. Total freshwater resources are estimated to be in excess of 35 million km^3.

The water is on a perpetual move, converting from liquid to solid or gaseous phase, or back. It partakes in processes of exchange of mass and energy between the various spheres of the Earth system. The main, in volumetric terms, water transfer takes place between the hydrosphere and the atmosphere in processes of evaporation and precipitation. The evaporation process purifies (distills) salty oceanic water into freshwater. Annual precipitation total largely depends on the latitude. Globally averaged latitudinal precipitation is highest near the equator and relatively high at the latitude around 60 degree (where upward lift of air masses is dominating). It is lower at the latitude around 30 degree and near the poles (where downward movement of air masses dominates). Also the altitude above the sea level is an important control of the amount of precipitation. Among further factors of importance are distance from source of water, exposition to prevailing wind, and large-scale landscape structure. Water moves not only in the processes of evaporation, precipitation, and infiltration, or flow in rivers and streams, plants and animals, but also in oceans, seas, and lakes, in snow pack, and in even seemingly immobile glaciers.

The most essential, and universal, law guiding the water cycle is the rule of balance (expressed by the continuity equation, also called equation of conservation of mass). It reads, for any fixed control volume:

$$\text{Inflow} - \text{Outflow} = \text{Change of storage}$$

Considering only the most essential hydrological processes, that is, the total precipitation on the basin, evaporation from the basin, runoff (river flow in a cross section terminating the basin), and change of storage in the basin (manifesting itself via surface waters—rivers, lakes, ponds, wetlands; soil moisture, groundwater, and intercepted water), one can formulate the continuity equation for a river basin as:

$$\text{Precipitation} - \text{Evaporation} - \text{Runoff} = \text{Change of storage}$$

The total volume of water in the hydrosphere has been nearly constant over a longer timescale. Hydrosphere is a closed system and water takes part in recycling rather than loss and replenishment processes.

The major water fluxes are evaporation and precipitation. Every year, solar energy lifts about 500,000 km^3 of water, 86% of which (i.e., 430,000 km^3) evaporates from the oceanic surface and 14% (i.e., 70,000 km^3) from land. About 90% of the volume of water evaporating from oceans precipitates back onto oceans, while 10% is transported to areas over land, where it precipitates. About two-thirds of the latter evaporate again and one-third runs off to the ocean. By virtue of the continuity equation for stationary conditions, the global volume of precipitation is equal to that of evaporation, that is, 500,000 km^3 of water falls as atmospheric precipitation (on the ocean 390,000 km^3 and on land 110,000 km^3). The resulting imbalance—difference between precipitation on and evaporation from land surface (110000–70,000 = 40,000 km^3 $year^{-1}$)—represents the water vapor movement from oceans to terrestrial atmosphere over continents and islands, being equal to the total runoff of Earth's rivers and direct groundwater runoff to the ocean. Fig. 1 illustrates the principle of the global water cycle, as explained above. Solid arrows represent movement of water in liquid and solid phases, while broken-line arrows represent movement of water vapor.

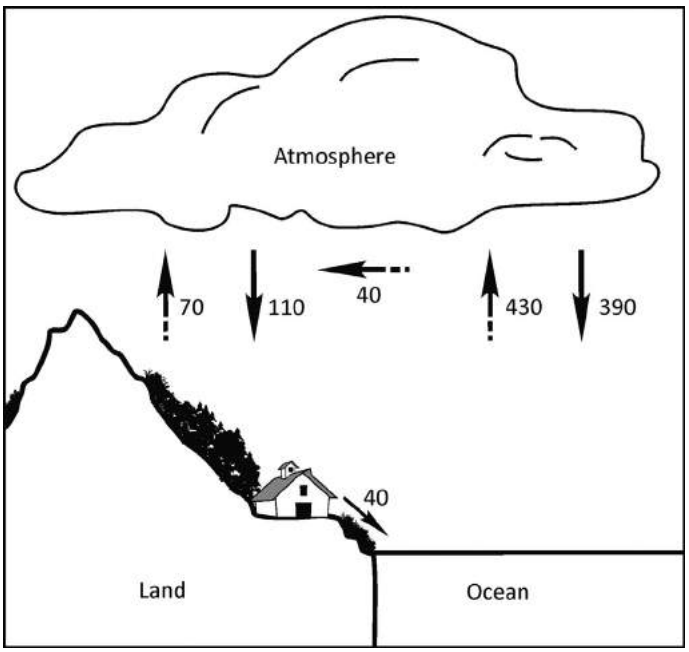


Fig. 1 Principle of the global water cycle. Numbers refer to annual fluxes in thousands of cubic kilometers. The processes of vertical movement of water are precipitation and evaporation, and processes of nearly lateral movement are runoff and advection of atmospheric moisture from ocean areas to the land areas.

Since the large volumes in hundreds of thousands of cubic kilometers are not easy to interpret, the fluxes in the global water cycle can be expressed in units of length (thickness of water layer). On average, a layer of 140 cm of water evaporates from the oceans, and 127 cm of water precipitates onto the oceans. The difference of 13 cm is very important, as it drives the continental phase of the water cycle. Since more water evaporates from the ocean than precipitates on it, there is a surplus of moisture, which moves over land and precipitates there. In result, precipitation on the land (80 cm) is much higher than evaporation from land (48 cm).

The mean sojourn time of a water particle in different stores varies from hours to millennia. Slow turnover is typical in ocean bodies, large lakes, and deep groundwater, where mean residence time of a water particle depends on the depth, and in ice sheet and glaciers, due to their frozen, immobile nature in cold (low-energy) climates. A water particle spends, on average, about 10,000 years in underground ice in the permafrost area or the eternal snows and polar ice, 2500 years in the ocean, 1600 years in mountain glaciers, and 1400 years in groundwater. In lakes, wetlands, and the soil, the mean residence times of a water particle read 15–17, 5, and 1 year, respectively. Much faster is the turnover of stored water in rivers (16 days), atmospheric water (8–10 days), and biological water (a few hours). Since water in the atmosphere is completely replaced once every 8–10 days, one can state that the atmosphere recycles its contents about 40 times per year.

For millennia, people have not properly interpreted the hydrological cycle, even if they understood water as an indispensable condition of life and carried out advanced water management. In the ancient times, water was treated as one of four elements (beyond fire, air, and Earth). It was easy to comprehend that the gravity force dominates in the atmospheric precipitation, overland (surface) flow, river runoff, and infiltration. It was followed in the water supply technology (aqueducts). However, it was not clear how the water got up to the source areas located in higher altitudes, against the force of gravity. It was difficult to understand how the loop of the water cycle was closed—what was the force lifting water to the atmosphere, so that it could precipitate onto the ground. It was also not clear why the sea level does not grow despite the perpetual inflow of gigantic rivers. In short, many thinkers falsely interpreted the closing of the water cycle, expecting an underground connection. A concept of the water cycle, similar to the present interpretation, was known in ancient Rome over 2000 years ago, but possibly earlier in China and Greece.

Hydrological Processes

Water cycle can be interpreted as a set of hydrological processes that transfer water between different stores (reservoirs).

The water cycle is powered by solar energy, mainly through direct vaporization. Water evaporates from the Earth's surface, vapor is lifted into the atmosphere to form clouds, and transported in the atmosphere. The heating is spatially uneven, as most of the solar energy warms tropical seas and drives evaporation there. Water vapor, which has risen to the atmosphere, is carried by winds away from the tropics, where it condenses, releasing latent heat and precipitates over oceans and land. Vertical processes of evaporation and precipitation are the (two) major fluxes in the water cycle.

Precipitation water falling down on land is the main source of the formation of land waters: rivers, lakes, groundwater, and glaciers. Precipitated water may be intercepted by vegetation, infiltrate into the ground, be stored in ponds, lakes, and depressions at the Earth's surface, or run off. A portion of atmospheric precipitation evaporates, a part infiltrates and contributes to groundwater, and the rest gets as river flow to the ocean, then evaporates, so that the process repeats again and again. A portion of global river

Table 1 Major hydrological processes

Process	*Description*
Evaporation	Transformation of liquid water from the Earth's surface into the vapor state
Transpiration	Transformation of liquid water into the vapor state and transfer of water vapor to the atmosphere via plant metabolism
Evapotranspiration	Joint category embracing evaporation and transpiration. Potential evapotranspiration is the upper limit to evaporation and transpiration, assuming unlimited supply of water and full opening of the stomata (daylight hour)
Condensation	Transformation of water from vapor into (denser) liquid form in the air, producing clouds or fog
Advection	Transport of water by horizontal movement of mass of air
Precipitation	Transfer of water from the atmosphere to the Earth's surface in liquid (rainfall, fog drip) or solid (snow, graupel, hail, and sleet) state
Runoff	Transfer of water across the land. This category includes surface runoff (overland flow), subsurface flow, groundwater runoff, and river flow
Sublimation	Change of water state directly from solid (snow or ice) to gaseous
Infiltration	Flow of water from the ground surface (also bottom of water body) into the ground, under the combined forces of gravity, viscosity, and capillarity. Infiltrating water contributes to soil moisture (within the vadose or aeration zone), or percolates deeper to become groundwater (in the aquifer, that is, saturated zone)
Snowmelt	Transfer of water from snow cover (solid state) to liquid state, by melting process
Interception	Storage of precipitated water (in liquid or solid state) by plant foliage. Intercepted water may evaporate back to the atmosphere or fall onto the ground. The amount of intercepted water depends on the duration of the storm, wind speed, temperature, and the density of foliage. A dense forest can intercept nearly all water from a low intensity rainfall
Subsurface flow	Flow of water under the ground surface, in the vadose (aeration) zone or saturation zone (aquifer)
Capillary rise/ exfiltration	Movement of infiltrated water back toward the Earth surface, driven by an upward capillary potential gradient (caused by evaporation)

discharge, in the drainless areas of endorheic runoff, does not reach the ocean. Water moves also in the biosphere, in oceans, seas, and lakes, in snow pack, and even in seemingly immobile glaciers.

A list of major hydrological processes partaking in the water cycle, and their short characteristics, are compiled in Table 1.

One can conceptually divide precipitated water into "green" water and "blue" water. The former is a part of precipitation that evaporates and sustains plant growth, while the latter is liquid water in surface and subsurface water bodies, which can be withdrawn for human use (e.g., for irrigation). Blue water turns to green water in the ecosystems (including agriculture).

There is a considerable movement of water within the ocean bodies. Mixing between two stratified layers (upper—warmer; deeper—colder), separated by the thermocline, is very slow. Motions (currents) of ocean's waters result from the density differences, dependent on temperature and salinity. The thermohaline circulation is intense—the Gulf Stream is a conveyor belt of heat responsible for the relatively mild climate of Europe.

Extremes

At times, volumes of water in stores and fluxes of water between stores, at various spatial scales, take extremely high or extremely low values. For over a century, highest precipitation values have been recorded in meteorological stations worldwide, for various time intervals. The list of global records range in different time intervals include 38 mm of precipitation in 1 min, through 1825 mm in 1 day (24 h), to 9.3 m in 1 month, 22.45 m in 6 months, 26.46 m in 1 year, and 40.76 m in 2 years.

Too much rainfall can cause excess runoff and inundation (flooding) of terrain, while too little rainfall leads to drought, decrease in water level, or even drying out of surface-water bodies, drop of groundwater level and soil moisture, often accompanied by failure of crops, and adverse effects, for example, related to water supply, navigation, hydropower, and wild fire.

Amounts of precipitation may considerably vary from year to year. The same areas may experience a drought in 1 year and a flood in next year (or sometimes—even drought and flood in the same season). For instance, the River Elbe flooded Dresden (Germany) in summer 2002 with a record high level and featured extremely low flow in summer 2003. In 1988, an intense drought disrupted agriculture in the Midwestern United States (too little water), while in 1993 the same area was subjected to severe inundations, greatly reducing the annual harvest (too much water).

Human Impacts

The Earth's freshwater resources remain constant, but man is capable of altering the water cycle and the water resource itself, in both quantity and quality context.

Humans have always interacted with the water cycle, drinking freshwater and using it for various purposes. They have influenced hydrological processes, in order to accelerate the water movement (e.g., improving conveyance in open channels), or to slow it down (by damming a river and catching water in a reservoir rather than letting it flow promptly to the sea). Man has benefited from the kinetic or dynamic energy of water and the value of water itself.

The water resources have always been distributed unevenly in space and time and man has tried to reduce this unevenness and smoothen the spatial–temporal variability. Regulating flow in space can be achieved via water transfer, while regulating flow in time to suit human needs can be achieved by storage reservoirs (capturing water when abundant and using it when it is scarce).

Water transfer is an old idea. Man-made water conduits (aqueducts) date back to the ancient world. Already 6000 years ago, river water was diverted for irrigation agriculture in Mesopotamia. In Mesopotamia and Egypt, *qanats*—underground lateral canals—were used, through which water was conveyed without losses by evaporation.

The rationale for a large-scale water transfer is backed by the following observations:

- The aggregate renewable freshwater resources on the Earth (total river discharge) are sufficient to meet the human water demands for many decades ahead;
- Freshwater resources on the Earth are spatially distributed in a very uneven way; and;
- Man's economic activities additionally exacerbate natural unevenness in spatial distribution of water resources.

It is therefore tempting to transfer water from the regions where it is abundant to water-scarce regions. However, effects of large-scale water transfers on the natural environment have to be thoroughly analyzed, because adverse side effects may prove to be serious.

Today, water storage reservoirs serve multiple purposes—water supply for agriculture, households, industry, hydropower, navigation, and recreation. Irrigated agriculture is indispensable to feed the increasing population of the globe. Hydropower developments have been enhanced by increasing fossil fuel cost and the advent of climate mitigation policy via renewable energy sources. However, beneficial impacts of reservoirs are not for free. Among adverse effects of dams and reservoirs are disturbances to ecosystems, barrier to fish, inundation of fertile land, and relocation of people. Moreover, enhanced evaporation from a large water surface reduces available water resources of the region. Hence, reservoirs should be taken into account, while estimating water consumption. Dams have been built for millennia, but most large dams have been constructed since the mid-20th century. At present, the total design volume of world reservoirs exceeds 6000 km^3, and their total water surface area reaches 500,000 km^2.

Beyond water storage and water transfer schemes, the water cycle is exposed to many other human impacts: changes in land use and land cover, deforestation or afforestation, modification of soil layers, urbanization, and agricultural activities. Water withdrawals and uptake from rivers, lakes and wells for irrigation, as well as municipal and industrial water usages modify water cycle significantly in both quantitative and qualitative aspects. Impacts of population growth, economic activities, and consumptive lifestyle on hydrological cycle and water withdrawals result in rising water stress. Water withdrawals increase directly with the growth of population and water usage per capita, and indirectly through the increase in food production.

Particularly important are easily accessible freshwater resources, such as surface waters and shallow groundwater, part of which is accessible to plant roots. At present, about 600–700 km^3 of annual water withdrawal stem from groundwater. A large part of this groundwater is used for irrigation and municipal needs. For areas with almost no river runoff (e.g., Arabian Peninsula, Libya), groundwater and desalinated seawater are the main water sources, but much of the groundwater is nonrenewable. The recharge took place in past climates, and after withdrawal, the groundwater resources will not be replenished.

The ongoing globalization has increased the transport and trade of "virtual water" worldwide. The virtual water content of goods is equal to the amount of water required if the transferred goods are produced in the importing and consuming area. Hence, in destinations of virtual water trade, water encapsulated in products (e.g., food) is imported from source regions.

Access to freshwater is now being regarded as a universal human right and extending access to safe potable water is one of the Millennium Development Goals—to halve, between 1990 and 2015, the proportion of people without sustainable access to safe drinking water. However, there are still more than 880 million people that have no access to safe freshwater. Human population of 7.5 billion people drink more than 7 km^3 of water per year.

Until a century ago, the number of people on Earth was not high, and human impact on water resources was generally insignificant and local rather than global. Thanks to the renewal process of the water cycle and its self-purification properties, on average, the quantity and quality of fresh waters had not changed much (except for climate-driven natural variability). The process of evaporation and surface water systems (rivers, lakes, and, in particular, wetlands) remove a large portion of pollutants from the water, in liquid or gaseous state. There had been an illusion that water resources are infinite, inexhaustible, and perfectly renewable, free goods. The situation has dramatically changed over the last century, when water withdrawals strongly increased due to the dynamic population growth and socioeconomic development driving the increase of human living standards. There has been a dramatic expansion of irrigated agricultural areas, growth of industrial water use (including the power sector), and intensive construction of storage reservoirs worldwide.

Indeed, water is not a free goods any more. A future-oriented water resources management should emphasize shaping demands rather than supply extension. It is a must to improve the efficiency of water use, trying to "do more with less" ("more crop per drop"). Financial instruments, such as the water pricing not only granting full cost recovery but also accounting the cost of the resource, in the sense of foregone opportunities, can generally improve the efficiency of water use.

Global water consumption has increased more than sixfold since the beginning of the 20th century, being twice stronger than the population growth. Facing the increasing pressures, the business-as-usual approach to water development and management cannot be globally sustainable.

The problems of water shortage are likely to be aggravated in the 21st century, which was baptized "the age of water scarcity." Population growth, economic development, and increasingly consumptive lifestyle impact on the hydrological cycle, boosting water withdrawals and increasing the hazard of water stress and water scarcity.

Irrigated agriculture consumes, globally, 70% of the world water withdrawals. More and more water is needed to produce food for the ever-increasing population of the globe. Since projections for the future foresee further growth of population, the consequences to food and fiber production are clear and the global demand for water will grow further. Faster growth is expected in less developed countries: in the whole of Africa and much of Asia.

Man has changed the quality of the world's water, creating acute problems in densely populated regions of the Earth, where no efficient wastewater purification takes place. It is estimated that only 5% of the world's wastewater is treated. The structure of human-caused water pollution problems has changed in time, with fecal coliform bacteria and organic pollution being the oldest.

Today, important water pollution problems are caused by bacteriological and organic contamination, salinization of freshwater (groundwater, rivers, lakes), driven by irrigation, groundwater overexploitation or saltwater intrusion, water withdrawals, pollution by nutrients (nitrogen, phosphorus), whose abundance leads to eutrophication and toxic algae blooms, remains of agricultural chemistry products, metals, and radioactive material, organic micropollutants, and acidification. Remains of agricultural chemistry products, artificial fertilizers, pesticides, and herbicides, are particularly difficult to eliminate, due to the distributed nature of the source. Some synthetic chemicals, for example, organochlorines (organohalides) have a long half-life time: 8 years in the case of DDT.

Even when perennial surface water source is available in a given location, water consumption in untreated state may present a risk to human health because of contamination by pathogens or waste. The number of people dying each year of water-related diseases is of the order of millions. Particularly burning water supply problems occur in informal human settlements, for example, slums around mega-cities in developing countries, where the poor have no access to public, safe, tap water. They have to buy lower-quality water from vendors and pay much more than the price charged to more wealthy citizens who have access to the public supply of safe water.

The need for protection of the aquatic ecosystems is being increasingly recognized. Despite the rising human demand for water, it is necessary to allocate a share of water to maintain the functioning of freshwater-dependent ecosystems, thus meeting conditions of environmental flow and environmental water requirements. This would allow (if flows are regulated) to maintain the water

regime within a river or a wetland, that suits aquatic and riparian ecosystems. However, earmarking water for environmental requirements is very difficult in some areas. At times, river flow may not reach the sea due to excessive human water withdrawal.

In order to improve the quality of water in the countries of the European Union (EU), the Water Framework Directive entered into force in December 2000, setting out a framework for actions in the field of water policy in the European Union. The key objective of the Directive, which imposes legal obligations on the authorities in EU member states, is to achieve a "good water status" for all waters of the European Union.

Climate Change Impact on Water Resources

Climatic and freshwater systems are interconnected in a complex way, so that any change in one of these systems induces a change in the other. Climate and water on the planet Earth are closely linked. Climate change exerts considerable impact on the water cycle and all the hydrological processes partaking in it.

Water takes part in a large-scale exchange of mass and heat between the atmosphere, the ocean, and the land surface, thus influencing the climate, and also being influenced by the climate. Water plays a pivotal role in the redistribution of heat in the Earth's atmosphere, and in the Earth's thermal system. Due to its high specific and latent heat, water moderates the Earth's climate, acting as air-conditioner in the Earth system. Large hidden energy is released in the atmosphere when water vapor condenses and latent heat (water vapor) transport is a major component of the Earth's heat balance. Some 23% of the solar radiation that reaches the Earth is used for evaporating water. Most of the Earth's waters is contained in the oceans and the very high heat capacity of this large volume of water buffers the Earth surface from strong temperature changes such as those occurring on the waterless Moon. Ocean acts as the principal heat storage component in the Earth system, a regulating flywheel in the Earth's heat engine.

The water cycle affects the energy budget of the Earth. Clouds alter Earth's radiation balance. Atmospheric water vapor (along with carbon dioxide and methane) is a powerful greenhouse gas, playing a significant role in the greenhouse effect. This effect, which can be described as absorbing the long-wavelength infrared radiation emitted by the Earth's surface, is responsible for maintaining the mean surface temperature about 33°C higher than would be the case in the absence of the atmosphere.

In the history of Earth's climate, there were time periods when much of the hydrosphere on the surface of the planet was in the solid form of glacial ice. Possibly, during the Cryogenian period, the range of sea ice extended nearly to the equator. There have been several ice ages in the history of the Earth, and the most recent retreat of glaciation is dated at some 10,000 years ago. Range and extent of ice sheets, glacier, and permanent snow areas remain a sensitive indicator of changes in the Earth's climate. After expansion during the Little Ice Age, they have been shrinking recently in response to the ongoing global warming.

Earth's climate has always been changing, reflecting regular shifts in Earth's orbit and solar activity and radiation, and irregular volcanic eruptions. However, a large part of the ongoing climate change is due to human activity. Man has been carrying out a planetary-scale experiment, disturbing the natural composition of the atmosphere by increasing the contents of greenhouse gases, by unprecedented level of burning of fossil carbon (coal) and hydrocarbons (oil and natural gas), and large-scale deforestation (reduction of carbon sink). In consequence, the greenhouse effect becomes more intense, leading to global warming.

The globally averaged Earth (land and ocean) surface temperature data as calculated by a linear trend, show a warming of 0.85°C [0.65–1.06°C], over the period 1880–2012. Further temperature increase is projected, depending on the socioeconomic (hence, greenhouse gas emission) scenarios. If effective global climate change mitigation will not take place (development pathway, RCP8.5), the global mean temperature may grow in 2081–100 to 2.6–4.8°C relative to 1986–2005. The mean warming over land will be larger than over the ocean and the Arctic region will warm more rapidly than the global mean.

Apart from the warming, there are several further manifestations of climate change and its impacts on freshwater resources, many of which have already been observed, and further (and more pronounced) impacts have been projected. Observational evidence indicates an ongoing intensification of the water cycle, with increasing rates of evaporation and precipitation. There is more water vapor in the warmer atmosphere, and this creates potential for enhanced intense precipitation. There is a poleward shift of the belt of higher precipitation. Increase in midsummer dryness in continental interiors has been observed and further increase is projected.

Changes in streamflow volume, both increases and decreases, have been recorded in many regions, but often these trends cannot be definitively attributed to changes in climate, due to existence of several other factors. The effect of climate change on streamflow, lake levels, and groundwater recharge, which varies regionally, largely follows changes in precipitation. A robust finding is that warming leads to changes in the timing of river flows where much of the winter precipitation currently falls as snow. The effect is greatest at lower elevations (where snowfall is more marginal). Winter flows increase and summer flows decrease.

Carbon dioxide enrichment improves efficiency of plant water use, reducing stomatal conductance and leaf-scale evaporation, but this is partly offset by increased plant growth. Warmer temperatures generate increased glacier melt; hence, widespread glacier retreat has been already observed, and many rivers draining glaciated regions have increasing flows, due to increase of melt. Many small glaciers disappear.

Water quality is likely generally to be degraded by higher water temperature, but this may be offset regionally by the dilution effect of increased flows. Warming-enhanced sea-level rise can lead to saltwater intrusion into fresh groundwater bodies. Thus, freshwater availability in coastal areas is likely to decrease in the warmer climate.

Effects of future climate change on average annual river runoff across the world indicate some generally consistent patterns of change—increases in high latitudes and the wet tropics, and decreases in mid-latitudes and some parts of the dry tropics. High reductions in the mass of Northern Hemisphere glaciers are expected in the warming climate. As these glaciers retreat, rivers, which

are sustained by glacier melt during the summer season, feature flow increase, but the contribution of glacier melt will gradually fall over the next few decades.

Further Reading

Bates BC, Kundzewicz ZW, Wu S, and Palutikof JP (eds.) (2008) *Climate change and water. Technical Paper of the Intergovernmental Panel on Climate Change*, p. 210. Geneva: IPCC Secretariat.

Eagleson PS (1970) *Dynamic hydrology*. xvi, 462, New York: McGraw-Hill.

IPCC (2014). Climate change 2014: Synthesis report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, Pachauri, R.K. and Meyer, L.A. (eds.)]. Geneva: IPCC, p. 151.

Singh V (ed.) (2016) *Handbook of applied hydrology*. New York: McGraw-Hill.

WWAP (United Nations World Water Assessment Programme) (2015) *The United Nations world water development report 2015: Water for a sustainable world*. Paris: UNESCO.

Xenobiotic (Pesticides, PCB, Dioxins) Cycles

VN Bashkin, VNIIGAZ/Gazprom, Moscow, Russia

Introduction

Persistent organic pollutants, POPs, are a wide class of chemical species with different physicochemical properties and toxicology. Here we will consider the following priority list of POPs: 1,1,1-trichloro-2,2-bis (4-chlorophenyl) ethane, DDT; hexachlorocyclohexanes, HCHs; hexachlorobenzene, HCBs; polychlorinated dibenzo-*p*-dioxins and dibenzofurans, PCDD/Fs; polychlorinated biphenyls, PCBs; polycyclic aromatic hydrocarbons, PAHs. Environmental pollution by POPs is one of the global problems that is drawing attention at national and international levels. The transboundary aspects of POP transport and pollution of various environmental media require study of the relevant effects on human health and the environment, including quantification of those effects. In accordance to Protocol on POPs to the UN ECE Convention on Long-Range Trans-Boundary Air Pollution that entered into force in October 2003, the parties to the protocol shall encourage research, development, monitoring, and cooperation related, in particular, to an effects-based approach which integrates appropriate information on measured or modeled environmental levels, pathways, and risk to human health and the environment.

Evaluation of POP Deposition

Calculated fields of depositions and concentrations give the opportunity to assess the changes in atmospheric contamination and deposition of POPs and to select `hot spots´ of contamination. As an example, the spatial distribution of PCDD/Fs depositions to the EMEP region, calculated for the beginning and the end of the considered period, is given in **Fig. 1**. `Hot spots´ are particular cells of the EMEP grid characterized by the highest values of PCDD/Fs deposition fluxes in both years (marked by arrows). As seen from the data presented, deposition fluxes over the European countries decreased substantially. PCDD/F deposition at one `hot spot´ near Prague (the Czech Republic) decreased more than 2 times. Such calculated fields of depositions for other considered POPs are also available on the Internet (http://www.msceast.org).

Spatial Pattern of PCDD/Fs Contents in Various Environmental Compartments

For PCDD/Fs the spatial distribution of concentrations in air in comparison with that for soil concentrations in 2001 is shown in **Fig. 2**. Note that significant differences in the spatial distribution of air and soil concentrations in most European countries are observed. This fact can be explained by the long-term accumulation of PCDD/Fs in soil and relatively low degradation rates in this medium in combination with changes in the emissions during a long time period.

To take into account the effect of accumulation of POPs in different environmental compartments (soil, seawater, and vegetation) the modeling of their long-range transport was performed for a more prolonged period of time (1970–2001). In general, the trends of PCDD/F content in air and seawater followed the emission variation. The trend of PCDD/F accumulation in soil was strongly different from that of emissions. Emissions began to reduce in 1980, whereas the decrease in soil contamination started in 1990. The rate of the soil content decrease is much lower. This causes substantial PCDD/F re-emission flux from soil, which slows down the tendency for a decrease in PCDD/F content in the atmosphere.

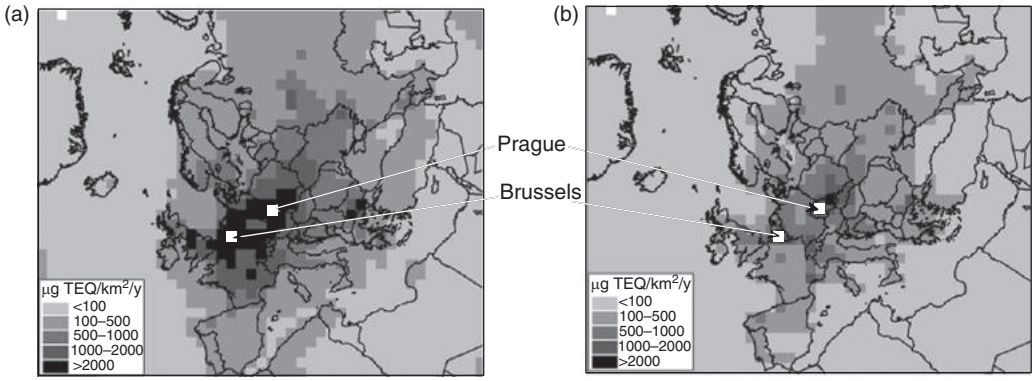


Fig. 1 Spatial distribution of PCDD/Fs depositions, (a) 1990 and (b) 2001 (the arrows show `hot spots').

 doi:10.1016/B978-0-444-63768-0.00767-8

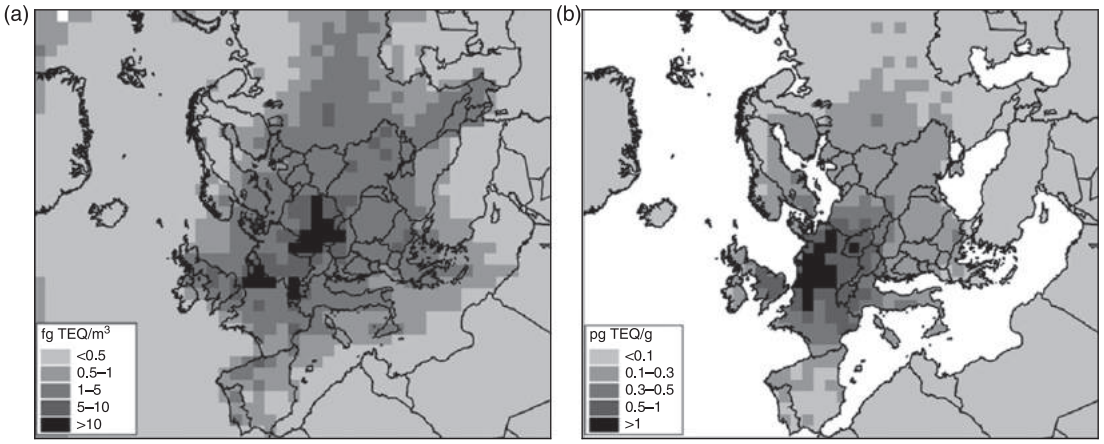


Fig. 2 Spatial distribution of PCDD/F concentrations in the (a) air and (b) soils of Europe in 2001.

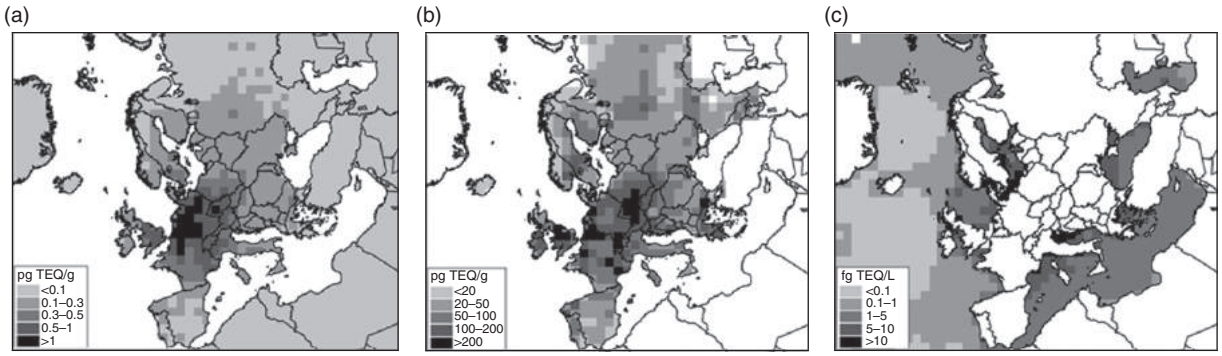


Fig. 3 Spatial distribution of PCDD/F concentrations in (a) soil, (b) vegetation, and (c) seawater in 2001.

Such pollutants as PAHs and PCBs also tend to be accumulated in the terrestrial environment but HCB and γ-HCH in the marine ones. Thus, this information gives us an idea of the POP exposure pathways to human beings.

To identify the areas and regions which were the most polluted by the considered POPs, the preliminary model results on the spatial distribution of their concentrations in different environmental media of the EMEP region were obtained. As an example, the spatial distributions of PCDD/F concentrations in soil, vegetation, and seawater with a spatial resolution of 50×50 are presented in **Fig. 3**.

POP Transport in the Northern Hemisphere

To evaluate the long-range transport ability of the considered POPs, the amount of each of these pollutants emitted in Europe and transported outside the EMEP region (outflow) was estimated. For the considered POPs, the percentage ratio of outflow to annual emissions ranged from 20% to 80%. For pollutants with the highest long-range transport potential, such as PCBs, HCHs, and HCB, calculations on the hemispheric scale were made. To evaluate the importance of intercontinental transport for these pollutants, calculations of their transport from different groups of sources such as European, American, and so on were carried out. To make these calculations tentative, hemispheric emission data for these pollutants were used.

On the basis of calculations made, contributions of different groups of emission sources located in the Northern Hemisphere, for instance, to HCB depositions over Europe and the Arctic were evaluated (**Figs. 4a** and **4b**). The contributions of remote source groups in the contamination of these regions are essential. Contributions of Russian emission sources to the European and Arctic contamination amount to about 19% and 31%, respectively. The relevant sum values of Canada and USA are 7% for the European domain and 17% for the Arctic.

At present evaluation of POP depositions to various types of the underlying surface are under investigations. The spatial distribution of PCB-153 depositions to areas covered with forests, soil, and seawater in 2000 is demonstrated in **Fig. 5**. Depositions of this pollutant to forests, soil, and seawater were estimated using different parametrizations of dry deposition velocities for different types of underlying surfaces. This resulted in considerable differences in depositions to the considered areas. As seen from the maps, the highest levels of PCB-153 depositions were characteristic of forested areas.

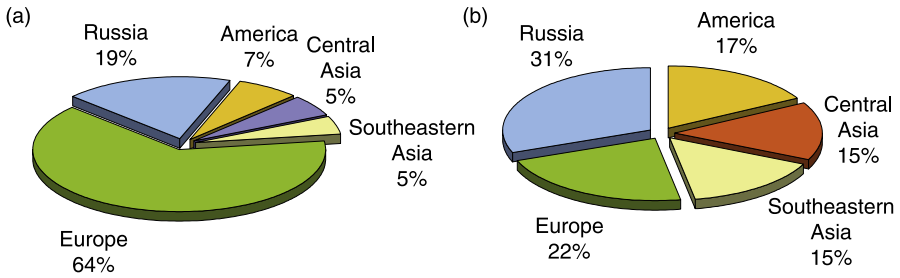


Fig. 4 Contributions of emission sources located in the Northern Hemisphere to depositions of HCB over Europe (a) and the Arctic (b), 2000.

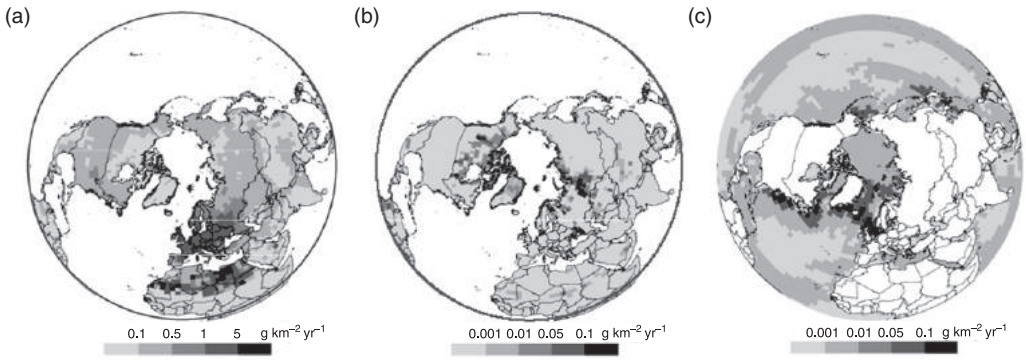


Fig. 5 Spatial distribution of PCB-153 depositions on (a) forests, (b) soil, and (c) sea in 2000.

Exposure Pathways of Dioxins and Dioxin-Like PCBs to Human

General Description of Dioxins

In this section, we consider the exposure pathways of POPs to human beings on the example of PCDD/Fs, often called just `dioxins'. These species consist of two groups of tricyclic aromatic compounds with similar chemical and physical properties. The number of chlorine atoms in each molecule can vary from one to eight. The number of chlorine atoms and their positions are of utmost importance for the toxicological potency of each congener. PCDD/Fs have never been produced intentionally, except for pure substances used as references in analytical and toxicological research, and have never served any useful purpose, unlike many other POPs such as PCBs and DDT. PCDD/Fs are formed as unwanted by-products in many industrial and combustion processes. They have also been shown to be formed in the environment by forest fires and volcanoes, and also via enzymatically catalyzed processes.

Primary sources of environmental contamination with PCDD/Fs in the past were the production and use of organic chemicals containing chlorine. PCDFs were formed as inadvertent by-products in the production and use of PCBs and, in combination with PCDDs, in such high-temperature processes as waste incineration, the metal industry, home heating, and other energy production processes.

PCDFs are also found in residual waste from the production of vinyl chloride and the chlor-alkali process for chlorine production. Factors favorable for the formation of PCDD/Fs are high temperatures, alkaline media, the presence of ultraviolet light, and the presence of radicals in the reaction mixture/chemical process.

Previous production of pentachlorophenol, as well as the bleaching process in pulp and paper mills, has been shown to be a major source. Changes in industrial processes have resulted in a reduction of PCDD/Fs concentration in products. Whereas in the past the chemical industry and, to a lesser extent, the pulp and paper industry were considered to be the main sources of PCDD/Fs (and also the cause of many of today's contaminated sites in several industrialized countries), today's dioxin input is mainly due to thermal processes. There is still a considerable focus on waste incineration but, owing to requirements for dioxin reduction in stack gases set by several national authorities, the importance of this category has declined during the last years. Examples can be seen especially in the European emission inventories. An overview of combustion sources known to generate and emit PCDD/Fs is presented in **Table 1**.

PCDD/PCDFs are found not only in stack gases but also in solid residues from any combustion process such as bottom ash, slag, and fly ash. With advanced technology and better burnout of the ashes and slag (characterized by a low content of organic carbon), PCDD/F concentrations have declined.

Secondary sources of PCDD/Fs, their reservoirs, are those matrices where they are already present, either in the environment or as products. Product reservoirs include PCP-treated wood, PCB-containing transformers and sewage sludge, compost and liquid

Table 1 Sources of emission of PCDD/Fs

Stationary sources	
Waste incineration	Municipal solid waste, clinical waste, hazardous waste, sewage sludge
Steel industry	Steel mills, sintering plants, hot-strip mills
Recycling plants	Nonferrous metals (melting, foundry: Al, Cu, Pb, Zn, Sn)
Energy production	Fossil fuel power plants, wood combustion, landfill gas
Diffuse sources	
Traffic	Cars
Home heating	Coal, oil, gas, wood
Accidents	PCB fires, fires in building, forest fires, volcanic eruptions

From Fielder H (1999) Sources of PCDD/PCDF and impact on the environment. *Chemosphere* 32: 55–64.

manure, which can be used as fertilizers in agriculture and gardens. Reservoirs in the environment are, for example, landfills and waste dumps, contaminated soils (mainly from former chemical production or handling sites), and contaminated sediments (especially in harbors and rivers with industries discharging directly to the waterways).

Although these reservoirs may be highly contaminated with PCDD/Fs, the chemical and physical properties of these compounds imply that dioxins and furans will stay adsorbed to organic carbon in soils or other particles. On the other hand, mobilization can occur in the presence of lipophilic solvents (leaching into deeper layers of soils and/or groundwater) or in cases of erosion or runoff from topsoil (translocation into the neighborhood). Experience has shown that transport of PCDD/Fs due to soil erosion and runoff does not play a major role in environmental contamination and human exposure.

PCBs have been used commercially since 1929 as dielectric and heat exchange fluids and in a variety of other applications. The presence of PCBs in human and wildlife tissues was first recognized in 1966. Investigations in many parts of the world have since revealed widespread distribution of PCBs in the environment, including remote areas with no PCB production or use. There is evidence that the major source of PCB exposure in the general environment is the redistribution of PCBs previously introduced into the environment. It is believed that large bodies of water, such as the Baltic Sea and the Canadian Great Lakes, may release significant amounts of PCB residues from previous uses into the atmosphere. The fact that PCB levels seem to decline in a similar way at different latitudes indicates that primary sources may still play an important role. The amount of dioxin-like PCBs might vary in the environment but the sources, transport, and distribution, as well as persistence, show similarities with the general properties of PCBs.

Potential for Long-Range Transboundary Air Pollution

PCDD/Fs are very persistent compounds; as their Kow and Koc are very high, they will intensively adsorb on to particles in air, soil, and sediment and accumulate in fat-containing tissues. The strong adsorption of PCDD/Fs and related compounds to soil and sediment particles means that their mobility in these environmental compartments is negligible. Their mobility may be increased by the simultaneous presence of organic solvents such as mineral oil. The air compartment is probably the most significant compartment for the environmental distribution and fate of these compounds.

Some of the PCDD/Fs emitted into air will be bound to particles while the rest will be in the gaseous phase, which can be subject to long-range transport (up to thousands of kilometers). In the gaseous phase, removal processes include chemical and photochemical degradation. In the particulate phase, these processes are of minor importance and the transport range of the particulate phase will primarily depend on the particle size. PCDD/Fs are extremely resistant to chemical oxidation and hydrolysis, and hence these processes are not expected to be significant in the aquatic environment. Photodegradation and microbial transformation are probably the most important degradation routes in surface water and sediment.

The number of chlorine atoms in each molecule can vary from one to eight. Among the possible 210 compounds, 17 congeners have chlorine atoms at least in the positions 2, 3, 7, and 8 of the parent molecule and these are the most toxic, bioaccumulative, and persistent ones compared to congeners lacking this configuration. All the 2,3,7,8-substituted PCDDs and PCDFs plus coplanar PCBs (with no chlorine substitution at the *ortho*-positions) show the same type of biological and toxic response.

PCDD/Fs are characterized by their lipophilicity, semivolatility, and resistance to degradation. The photodegradation of particle-bound PCCD/Fs in air was found to be negligible. These characteristics predispose these substances to long environmental persistence and to long-range transport. They are also known for their ability to bioconcentrate and biomagnify under typical environmental conditions, thereby potentially achieving toxicologically relevant concentrations. The tetra–octa PCCD/PCDFs have lower vapor pressures than PCBs and are therefore not expected to undergo long-range transport to the same extent; nevertheless, there is evidence for deposition in Arctic soils and sediments.

Persistence in water, soil, and sediment

Owing to their chemical, physical, and biological stability, PCDD/Fs are able to remain in the environment for a long time. As a consequence, dioxins from so-called `primary sources' (formed in industrial or combustion processes) are transferred to other matrices and enter the environment. Such secondary sources are sewage sludge, compost, landfills, and other contaminated areas. PCBs and PCDD/Fs are lipophilic (lipophilicity increases with increasing chlorination) and have very low water solubility. Because of their persistent nature and lipophilicity, once PCDD/Fs enter the environment and living organisms, they will remain for a very long time, like many other halogenated aromatic compounds. As log Kow (typically 6–8) or log Koc are very high for all these compounds, they will intensively adsorb on to particles in air, soil, and sediment. The strong adsorption of PCDD/Fs and related compounds to soil and sediment particles causes their mobility in these environmental compartments to be negligible.

Their mobility may be increased by the simultaneous presence of organic solvents such as mineral oil. The half-life of TCDD in soil has been reported as 10–12 years, whereas photochemical degradation seems to be considerably faster but with a large variation that might be explained by experimental differences (solvents used, etc.). Highly chlorinated PCDD/Fs seem to be more resistant to degradation than those with just a few chlorine atoms.

Bioaccumulation

The physicochemical properties of PCBs and their metabolites enable these compounds to be absorbed readily by organisms. The high lipid solubility and the low water solubility lead to the retention of PCCD/Fs, PCBs, and their metabolites in fatty tissues. Protein binding may also contribute to their tissue retention. The rates of accumulation into organisms vary with the species, the duration and concentration of exposure, and the environmental conditions. The high retention of PCDD/Fs and PCBs, including their metabolites, implies that toxic effects can occur in organisms spatially and temporally remote from the original release.

Gastrointestinal absorption of tetrachlorodibenzo-*p*-dioxin (TCDD) in rodents has been reported to be in the range of 50–85% of the dose given. The half-life in rodents ranges from 12 to 31 days except for guinea pigs, which show slower elimination ranging from 22 to 94 days. The half-life in larger animals is much longer, being around 1 year in rhesus monkeys and 7–10 years in humans.

Monitoring

PCCD/Fs have been found to be present in Arctic air samples, for example, during the winter of 2000/2001 in weekly filter samples (particulate phase) collected at Alert in Canada. PCDD/PCDFs have been monitored since 1969 in fish and fish-eating birds from the Baltic. The levels of PCDD/Fs in guillemot eggs, expressed as TEQ, decreased from 3.3 ng/g lipids to around 1 ng/g between 1969 and 1990. Since 1990, this reduction seems to have leveled off and today it is uncertain whether there is a decrease or not. Fish (herring) show a similar picture.

Thus both physical characteristics and environmental findings support the long-range transport of PCCD/Fs and PCBs. There are differences, however, both between and within the groups regarding ability to undergo LRTAP.

Pathways of LRTAP-Derived Human Exposure

For decades, many countries and intergovernmental organizations have taken measures to prevent the formation and release of PCDD/Fs, and have also banned or severely restricted the production, use, handling, transport, and disposal of PCBs. As a consequence, release of these substances into the environment has decreased in many developed countries. Nevertheless, analysis of food and breast milk show that they are still present, although in levels lower than those measured in the 1960s and 1970s. At present, the major source of PCB exposure in the general environment appears to be the redistribution of previously introduced PCBs.

Significant sources and magnitude of human exposure

PCDD/Fs are today found in almost all compartments of the global ecosystem in at least trace amounts. They are ubiquitous in soil, sediments, and air. Excluding occupational or accidental exposures, most human background exposure to dioxins and PCBs occurs through the diet, with food of animal origin being the major source, as they are persistent in the environment and accumulate in animal fat.

Importantly, past and present human exposure to PCDD/Fs and PCBs results primarily from their transfer along the pathway: atmospheric emissions → air → deposition → terrestrial/aquatic food chains → human diet. Information from food surveys in industrialized countries indicates a daily intake of PCDD/Fs on the order of 50–200 pg I-TEQ/person per day for a 60 kg adult, or 1–3 pg I-TEQ/kg bw per day. If dioxin-like PCBs are also included, the daily total TEQ intake can be higher by a factor of 2–3. Recent studies from countries that started to implement measures to reduce dioxin emissions in the late 1980s clearly show decreasing PCDD/F and PCB levels in food and, consequently, a lower dietary intake of these compounds by almost a factor of 2 within the past 7 years.

Biota from the Baltic have, however, not shown any clear trend for dioxins or PCBs since 1990. Occupational exposures to both PCDDs and PCDFs at higher levels have occurred since the 1940s as a result of the production and use of chlorophenols and chlorophenoxy herbicides and to PCDFs in metal production and recycling. Even higher exposures to PCDDs have occurred sporadically in relation to accidents in these industries. High exposures to PCDFs have occurred in relation to accidents such as the Yusho (Japan) and Yucheng (Taiwan) incidents, involving contamination of rice oil and accidents involving electrical equipment containing PCBs.

Exposure levels in adults

PCDD/Fs accumulate in human adipose tissue, and the level reflects the history of intake by the individual. Several factors have been shown to affect adipose tissue concentrations/body burdens, notably age, the number of children and period of breast-feeding, and dietary habits. Breast milk represents the most useful matrix for evaluating time trends of dioxins and many other POPs. Several factors affect the PCDD/PCDF content of human breast milk, most notably the mother's age, the duration of breastfeeding, and the fat content of the milk. Studies should therefore ideally be performed on samples from a large number of mothers, taking these variables into account.

The WHO Regional Office for Europe carried out a series of exposure studies aimed at detecting PCBs, PCDDs, and PCDFs in human milk. The first round took place in 1987–88 and the second in 1992. In 2001–02, a third round was organized in collaboration with the WHO Global Environmental Monitoring System/Food Contamination Monitoring and Assessment Programme (GEMS Food) and the International Programme on Chemical Safety (IPCS). Results are currently available from 21 countries. **Fig. 6** presents the temporal trends of levels of PCDDs and PCDFs expressed in WHO–TEQ for those countries participating in all three rounds or in the last two rounds of the WHO study. A clear decline can be seen, with the largest decline for countries originally having the highest level of dioxin-like compounds in human milk.

The general population is mainly exposed to PCBs through common food items. Fatty food of animal origin, such as meat, certain fish, and dairy products, is the major source of human exposure. Owing to considerable differences in the kinetic behavior of individual PCB congeners, human exposure to PCB from food items differs markedly in composition compared to the composition of commercial PCB mixtures.

PCB levels in fish have been decreasing in many areas since the 1970s, but the decrease has leveled off during the last couple of years. Today, the daily PCB intake is estimated to be around 10 ng/kg bw for an adult.

Exposure levels in children (including prenatal exposure)

Once in the body, PCBs and PCDD/Fs accumulate in fatty tissues and are slowly released. Lactation or significant weight loss increases the release of the substances into the blood. PCBs can cross the placenta from mother to fetus, and are also excreted into the breast milk. PCB and PCDD/F concentrations in human milk are usually higher than in cow's milk or other infant foods. As a

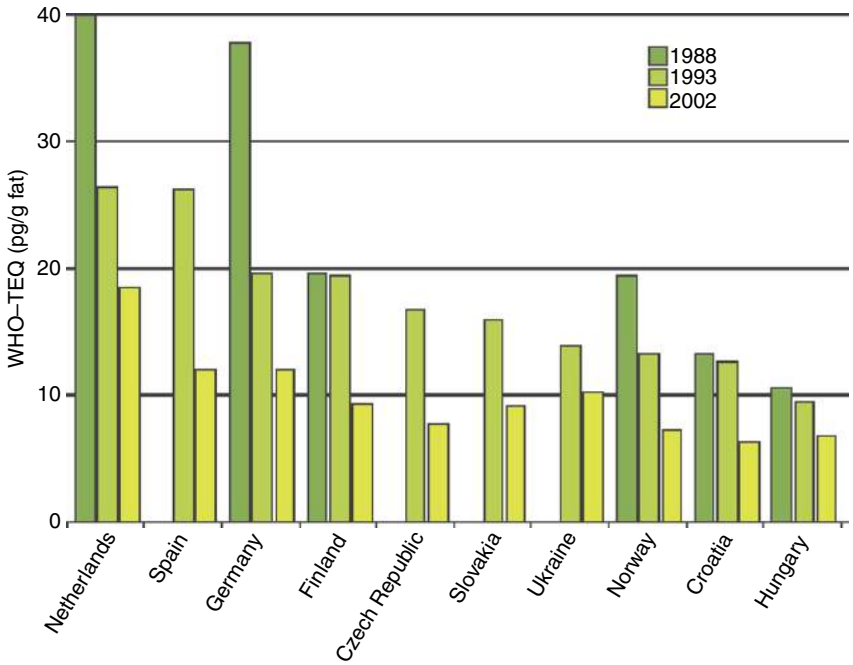


Fig. 6 Temporal trends in the levels of dioxins and furans in human milk in various countries participating in consecutive rounds of the WHO exposure study. From van Leeuwen FXR and Malisch R (2002) Results of the third round of the WHO-coordinated exposure study on the levels of PCBs, PCDDs and PCDFs in human milk. *Organohalogen Compounds* 56: 311–316.

result, breastfed infants undergo higher dietary exposure than those who are not breastfed. This concerns particularly breastfed infants of women exposed to high levels of PCBs, including Inuit and women whose diet is mainly based on fish from highly contaminated rivers and lakes, such as the Great Lakes and the Baltic Sea. Time-trend information suggests that PCDD/F and PCB concentrations in human milk have decreased significantly since the 1970s in countries that have taken measures against these substances. However, the decrease has leveled off during the last couple of years. Therefore, current fetal and neonatal exposures continue to raise serious concerns regarding potential health effects on developing infants.

Compared to adults, the daily intake of PCDD/Fs and PCBs by breastfed babies is 1–2 orders of magnitude higher. A recent field study showed higher mean levels of PCDD/Fs and PCBs in human milk in industrialized areas (10–35 pg I-TEQ/g milk fat) and lower levels in developing countries (<10 pg I-TEQ/g milk fat). Very few studies have been performed on Arctic populations with respect to the exposure of children to these substances. It is likely, however, that the differences in exposure between children and adults demonstrated in many industrialized regions also exist in Arctic regions.

Potential for high-exposure situations

It has been shown that these substances, and especially PCBs, can occur in elevated concentration in Arctic fauna. As the diet of many Arctic populations relies to a vast extent on marine mammals that represent high trophic levels, human exposure has been shown to be considerably high compared to industrialized areas.

Health effects in humans

There are many studies on the carcinogenicity of 2,3,7,8-TCDD in accidentally exposed workers. Epidemiological studies on people exposed in connection with the accident in Seveso have generated valuable information. Excess risks were observed for ovarian and thyroid cancers and for some neoplasia of the haematopoietic tissue; these results were, however, based on small numbers. Epidemiological studies on the cohorts most highly exposed to 2,3,7,8-TCDD produced the strongest evidence of increased risks for all cancers combined, along with less strong evidence of increased risks for cancers of particular sites. The relative risk for all cancers combined in the most highly exposed and longer-latency subcohorts is 1.4.

Studies of noncancer effects in children have indicated neurodevelopmental delays and neurobehavioral effects, including neonatal hypotonia. In children in Seveso who were highly exposed to TCDD, small, transient increases in hepatic enzymes, total lymphocyte counts and subsets, complement activity, and nonpermanent chloracne were observed. Also, an alteration of the sex ratio (excess female to male) was observed in children born to parents highly exposed to TCDD.

Critical outcomes and existing reference values

During the last two decades, a number of different risk assessments of dioxins and related compounds have been performed. Since the mid-1990s, coplanar PCBs have often been included in the assessments. In 1997, WHO established an expert group on dioxins and related compounds. It proposed, based on the toxic equivalency Factor (TEF) scheme shown in **Table 2**, a TDI for dioxins and

Table 2 WHO TEF values for human risk assessment

Congener	*TEF value*	*Congener*	*TEF value*
Dibenzo-p-*dioxins*		*Non*-ortho-*PCB*	
2,3,7,8-TCDD	1	PCB 77	0.0001
1,2,3,7,8-PnCDD	1	PCB 81	0.0001
1,2,3,4,7,8-HxCDD	0.1	PCB 126	0.1
1,2,3,6,7,8-HxCDD	0.1	PCB 169	0.01
1,2,3,7,8,9-HxCDD	0.1		
1,2,3,4,6,7,8-HpCDD	0.01		
OCDD	0.0001		
Dibenzofurans		*Mono*-ortho-*PCB*	
2,3,7,8-TCDF	0.1	PCB 105	0.0001
1,2,3,7,8-PnCDF	0.05	PCB 114	0.0005
2,3,4,7,8-PnCDF	0.5	PCB 118	0.0001
1,2,3,4,7,8-HxCDF	0.1	PCB 123	0.0001
1,2,3,6,7,8-HxCDF	0.1	PCB 156	0.0005
1,2,3,7,8,9-HxCDF	0.1	PCB 157	0.0005
2,3,4,6,7,8-HxCDF	0.1	PCB 167	0.00001
1,2,3,4,6,7,8-HpCDF	0.01	PCB 189	0.0001
1,2,3,4,7,8,9-HpCDF	0.01		
OCDF	0.0001		

related compounds. The proposal was based on kinetic calculations of doses to body burden and vice versa. The body burden approach resulted in a reduced need for a safety factor for extrapolation between species. The WHO expert group calculated that a reliable LOAEL probably could be found in the range of 14–37 pg/kg bw per day. By applying a safety factor of 10 to this range, it proposed a TDI of 1–4 pg/kg bw. The group emphasized that the TDI represents a tolerable daily intake for lifetime exposure, and that occasional short-term excursions above the TDI would have no health consequences provided that the averaged intake over long periods was not exceeded. In addition, it recognized that certain subtle effects may be occurring in some sections of the general populations of industrialized countries at current intake levels (2–6 TEQ /kg bw per day), but found it tolerable on a provisional basis since these reported subtle effects were not considered overtly adverse and there were questions as to the contribution of non-dioxin-like compounds to the observed effects. The group therefore stressed that the upper range of the TDI of 4 pg TEQ /kg bw should be considered a maximum tolerable intake on a provisional basis, and that the ultimate goal was to reduce human intake levels to below 1 pg TEQ /kg bw per day. In 2001, the European Commission and the Scientific Committee for Food proposed a temporary TWI of 14 pg/kg bw for 2,3,7,8-PCDD/Fs and dioxin-like PCBs.

Summary

It has been demonstrated that dioxins and many PCBs resist degradation, bioaccumulate, are transported through air, water, and migratory species across international boundaries, and are finally deposited far from the place of release where they can accumulate in terrestrial and aquatic ecosystems. The clearest evidence for this long-range transport derives from the levels of PCDD/Fs and PCBs measured in the Arctic. Owing to long-range transboundary transport, these substances are nowadays ubiquitous contaminants of the ecosystem and are also present in the food chain. Therefore, most of the human population is exposed to PCDD/Fs and PCBs. Moreover, since dioxins and PCBs pass from mother to fetus through the placenta, and from mother to newborn through breastfeeding, infants are at risk of harmful effects in the most critical period of their development. There are just a few reports of dioxins in humans from Arctic regions, but there are plenty of animal samples analyzed for dioxins and PCBs that give information on human exposure through food. As many people living in the Arctic still practice hunting and fishing for an important part of their diet, their exposure to dioxins, PCBs, and other contaminants could be elevated compared to people living in industrialized parts of the world.

Further Reading

Alcock, R., Bashkin, V., Bisson, M., *et al*., 2003. Health Risk of Persistent Organic Pollutants from Long-Range Transboundary Air Pollution. Bonn: WHO, 252p.

Bashkin, V.N., 2003. Environmental Chemistry: Asian Lessons. Singapore: Kluwer Academic, 490p.

Bertazzi, P.A., Bernucci, I., Brambilla, G., Consonni, D., Pesatori, A.C., 1998. The Seveso studies on early and long-term effects of dioxin exposure: A review. Environmental Health Perspectives 106 (supplement 2), 625–633.

Brzuzy, L.R., Hites, R.A., 1996. Global mass balance for polychlorinated dibenzo-*p*-dioxins and dibenzofurans. Environmental Science and Technology 30, 1797–1804.

Dutchak S, Shatalov V, Mantseva E, *et al.* (2004) *Persistent Organic Pollutants in the Environment.* MSC-E and CCC. EMEP Status Report 3/2004, Jun. 2004.

Fiedler, H., 1999. Sources of PCDD/PCDF and impact on the environment. Chemosphere 32, 55–64.

Galiulin, R.V., Bashkin, V.N., Galiulina, R.A., 2005. Ecological risk assessment of riverine contamination in the Caspian Sea basin: A conceptual model for persistent organochlorine compounds. Water, Air, and Soil Pollution 163 (1–4), 33–51.

Gray, L.E., Ostby, J.S., Kelce, W.R., 1997. A dose-response analysis of the reproductive effects of a single gestational dose of 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) in male Long Evans hooded rat offspring. Toxicology and Applied Pharmacology 146, 11–20.

Gray, L.E., Wolf, C., Ostby, J.S., 1997. *In utero* exposure to low doses of 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) alters reproductive development in female Long Evans hooded rat offspring. Toxicology and Applied Pharmacology 146, 237–244.

Mackay, D., Shiu, W.Y., Ma, K.C., 1992. Illustrated Handbook of Physical–Chemical Properties and Environmental Fate for Organic Chemicals. Boca Raton FL: Lewis.

Mantseva E, Dutchak S, Rozovskaya O, and Shatalov V (2004) *EMEP Contribution to the Preparatory Work for the Review of the CLRTAP Protocol on Persistent Organic Pollutants.* EMEP MSC-E Information Note 5/2004.

Oehme, M., Schlabach, M., Hummert, K., Luckas, B., Nordoy, E.S., 1994. Levels of polychlorinated-*p*-dioxins, dibenzofurans, biphenyls and pesticides in harp seals from the Greenland Sea. Organohalogen Compounds 20, 517–522.

van den Berg, M., Bimbaum, L., Bosveld, A.T., *et al*., 1998. Toxic equivalency factors (TEFs) for PCB, PCDDs, PCDFs for humans and wildlife. Environmental Health Perspectives 106, 775–792.

van Leeuwen, F.X.R., Malisch, R., 2002. Results of the third round of the WHO-coordinated exposure study on the levels of PCBs, PCDDs and PCDFs in human milk. Organohalogen Compounds 56, 311–316.

Wagrowski, D.M., Hites, R.A., 2000. Insights into the global distribution of polychlorinated dibenzo-*p*-dioxins and dibenzofurans. Environmental Science and Technology 34, 2952–2958.

Relevant Website

http://www.msceast.org– Meteorological Synthesizing Centre – East.

Adaptive Management and Integrative Assessments

L Gunderson, Emory University, Atlanta, GA, USA

Adaptive management is an approach for planning and managing natural resource systems. It is based on a perception of ecosystems as complex, dynamic systems that have a large degree of unpredictability. Hence, it was proposed as a learning-based approach that confronts inherent uncertainties and complexities of resource systems. As originally described, adaptive management is not a trial and error approach to management. That is, it utilizes scientific approaches to design actions that generate learning and understanding. Nor is it an approach that updates management action as new information becomes available. While it has these elements, it is an approach that is structured to learn while doing, and to attempt to make that learning efficient in order to winnow key resource uncertainties.

Background

A sequence of works has outlined, tested, and expanded the theory and practice of adaptive environmental assessment and management since the late 1960s. The first exercise in adaptive management was the Gulf Island Recreational Land Simulation study in 1968, where the participants attempted to explore ways to bridge gaps among scientific disciplines, technical experts, and policy designers. C. S. Holling and his colleagues introduced the concepts of adaptive management and results of various early attempts at implementation in the 1978 compendium *Adaptive Environmental Assessment and Management*. In 1982, the Canadian government commissioned a review and evaluation of the approach. Carl Walters presented theory and methods for dealing with the uncertainties of managing resources in the classic book *Adaptive Management of Renewable Resources*, which was published in 1986. In 1993, Kai Lee related the experience in the Columbia River basin of using adaptive management concepts to guide decision making in a social and political arena in his book, *Compass and Gyroscope*. Continuing the series in 1995 is the volume *Barriers and Bridges to the Renewal of Ecosystems and Institutions*, which presents a series of case studies on resource management histories and comments by social scientists in order to test ideas about the coevolution of ecosystems and management institutions. Recent ideas of adaptive management appeared in 2002 in *Panarchy*; *Understanding Transformations in Systems of Humans and Nature* which focuses on developing theoretical frameworks for why systems are so surprising based upon cross-scale dynamics (in space and time), and *Navigating Social–Ecological Systems*: *Building Resilience for Complexity and Change* which focused on approaches to management in social–ecological systems.

Adaptive management was proposed to fill three perceived gaps in extant management approaches. The first is to bridge diverging assumptions (mental models or paradigms) of resource dynamics; the second is to integrate differing perspectives among scientific disciplines; and third is to fill the breach between knowledge and action. Adaptive management was originally treated as adaptive environmental assessment and management (AEAM) to describe separate but linked processes of integrated assessment and active management. The main process during the integrative assessment aims to articulate assumptions of resource dynamics and integrate disciplinary perspectives and assess what is known and not known about resource issues. The most common approach at integration is to construct a computer model in a series of workshops that attempts to summarize and synthesize existing information and understanding in order to propose a set of policy options. Those options then generate a set of plausible management actions that help test the uncertainties of managing with incomplete knowledge and information, while confronting complexities of resource systems.

Confronting Complexity

Managed resource systems are inherently complex. One source of complexity is the number of dimensions in resource systems, including ecological, economic, social, political, organizational, among others. Indeed, even each of these dimensions are complex themselves, which leads to difficulties in tractability. The high dimensionality of these issues as well as the lack of common methods and institutions to manage these multiple dimensions has led some authors to apply the phrase `wicked problems' to natural resource issues. Practitioners of adaptive management acknowledge these multiple dimensions and attempt to separate resource issues into scientific (primarily ecological) and social (political) components. Although not always separable, these two categories provide a useful way of discussing the application of adaptive management to resource issues.

One source of complexity arises from differences in paradigms, theories, methods, and practices among academic disciplines that underlie resource issues. Even though many would argue the pragmatic and applied aspects of their work, all applied disciplines have a corollary academic field. For example, conservationists' ideas are rooted in understanding of ecology or biology; engineers in mathematics and physics; planners in geography or architecture. Some ecologists may be very good at understanding and evaluating ecological impacts and yet they often make myopic or inadequate suggestions for action because they are unaware

 doi:10.1016/B978-0-444-63768-0.00635-1

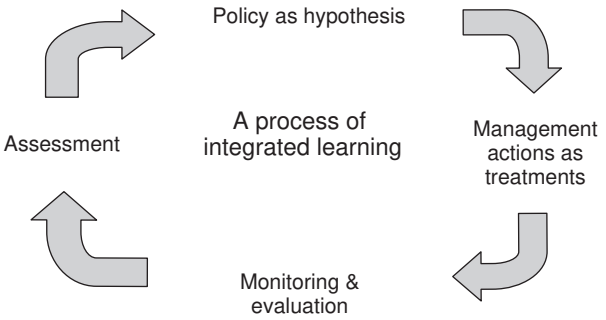


Fig. 1 Flowchart of adaptive environmental assessment and management. The assessment phase integrates existing knowledge to develop policies that are structured for learning about system uncertainties. The policies are tested by management actions and evaluated through monitoring and research.

or dismissive of the realities of human behavior, organizational structures, and institutional arrangements. Another such example is when engineers or managers assume that the uncertainty of nature can be replaced by human attempts to control and stabilize ecological systems. It is not that these approaches are wrong, but that they are partial and require more disciplinary integration. Overcoming disciplinary differences and bridging the gap between theory and praxis has been one of the central purposes of adaptive environmental assessment and management.

Complexity in resource systems leads to difficulties in prediction. All phases of environmental management, assessment, and planning activities involve some estimation about `what will happen'. However, these activities are problematic because of difficulties relating to our abilities to forecast. Certainly, many things are known, especially the broad and the general. However, there are many reasons why our predictive abilities of ecological systems are limited. These include sheer complexities of ecological systems, the wide range of scales of ecological structures and processes, competing ideas about how ecological systems are structured and function, and lack of data/information to test ideas about ecosystems across scales. These same arguments could be applied to the social dimensions of assessment and planning, but include problems of shifting and multiple values that society places on different aspects of ecosystems.

The wide range of spatial and temporal scales over which ecological systems are structured and operate provides another source of complexity. Ecosystems operate over spatial scales from millimeters to thousands of kilometers, and temporal scales of milliseconds to millennia. The cross-scale interaction of processes and structures in ecosystems can lead to surprising behavior. The concept of ecological resilience suggests that ecosystem shifts among alternate states are linked to cross-scale interactions. Examples of such ecological shifts include the transition from grass to shrub dominance in semiarid rangelands, population outbreaks of forest pests, shifts from clear water to turbid water in shallow lakes and similar structural shifts in inland sea systems, as well as shifts between coral and algae dominance in coral reefs. In all of these cases, the transition is mediated by the interaction between slower and faster components in ecosystems. Nonlinearities, alternative ecological states, and the inherent unpredictability of ecosystems suggested by resilience theory are all ecological concepts that underpin adaptive management.

The original works of adaptive management introduced the idea that it is equally important to address uncertainties of resource issues and management actions as it is to state what is known. A number of other approaches have been developed to address and winnow uncertainties. These include processes such as qualitative assessments by knowledgeable persons, judgment of experts, the development of narratives or scenarios, strategic gaming, statistical analyses of empirical or historical data, and causal modeling. Expert opinions are often used in legal situations or where causal understanding or data are lacking. The development of narratives or scenarios has been used successfully to highlight emerging issues or to roughly define feasible alternative futures that contain great uncertainty (such as Rachel Carson's 1962 book *Silent Spring*). Scenario-based planning has been used in business, to address issues of climate change, to explore sustainable futures, or in resource issues fraught with a large number of unknowns (such as the recent Millennium Ecosystem Assessment). While a growing number of theories, methods, and approaches have been developed to confront the ecological and technical components of resource systems, they are but part of the overall complexity of resource issues. As mentioned above, the social dimensions of these problems are as vexing and problematic, as discussed in the next section.

Other sources of complexity in managed resource systems arise from conflicting worldviews and interests of stakeholders. Resource issues most often engage governmental agencies, resources users, and other stakeholder groups, all with competing views and expectations of how ecosystems are valued and should be managed. In the United States, for example, federal agencies such as the Fish and Wildlife Service have responsibilities under the Endangered Species Act, while the Environmental Protection Agency and Army Corps of Engineers each have specific mandates that are not overlapping and can come into conflict with other governmental goals and claims. Nongovernmental organizations (NGOs) and even sovereign entities (First Nations) have taken larger roles in the lobbying, planning, design, and management of resources. Yet difficulties persist and protract the planning and design process due to the number and types of agencies and stakeholders that are engaged. In addition to the complexities due to the number and types of competing interests, another obstacle to integration is in the ways in which these groups interact and resolve differences so that actions can be taken.

The set of laws and bureaucratic structures developed to implement laws appear to generate partial solutions to the myriad problems faced in integrating ecology into all elements of resource management. Moreover, the legal and bureaucratic frameworks are set up in a way that defaults first to administrative processes and second to legal institutions to resolve conflicts or fill in missing pieces. In these frameworks, integration of ecological concerns into other social objectives is rarely achieved. Rather, these partial solutions shortchange ecological concerns. Often this is due to a lack of understanding about the natural systems, or unrealistic expectations about how these systems behave. Sometimes, these partial solutions generate greater ecological problems.

In order to manage these different forms of complexity, adaptive assessment and management was developed to simplify the complex and manage different forms of uncertainty. That is, processes were developed to make the issues tractable, and embrace the recognized uncertainties of managing complex systems. Two separable but related processes were developed and refined over the past three decades (**Fig. 1**). The first process is an integrative assessment that is usually followed by a management or implementation phase. The assessment process identifies uncertainties, points of intervention, and alternative explanations around resource issues. The management phase designs actions that can sort among competing hypotheses while achieving social objectives. Each of these phases is described in the following sections.

Integrative Assessments

Developers of adaptive environmental assessment suggest that an assessment must deal with three questions: (1) how to decompose usually complex resource issues so that they are tractable; (2) how to bring dispersed expertise and information to the problem; and (3) how to effectively communicate results to decision makers and the public. The questions are addressed by a number of techniques all of which have been designed and refined to process information developed in a series of workshops. In the workshops, constructing a computer model is the primary focus of information processing and compilation. The building of the model forces the group to bound the assessment, by defining what components go into a computer model. The model addresses the second question also, by consolidating disparate views of the issue. A number of techniques, including visualization and storytelling, are important and effective in transferring and communicating the results of the integrated understanding.

The objective in building computer models during adaptive assessments is to synthesize and reflect on current modes of understanding, rather than to predict ecosystem behavior. This is because uncertainties arise that are not amenable to uncovering by existing scientific approaches and techniques. Developers of this approach outline three steps in the assessment process: (1) determine resource issues and generate alternative hypotheses of system behavior; (2) develop quantitative approaches to evaluate how uncertainties relate to management options and actions; and (3) use a combination of approaches involving gaming and formal optimization to winnow these options.

The integrative assessment of resource issues begins with identifying management objectives and constraints. Often an explicit set of management objectives such as sustainable or maximum sustainable yield are stated. Implicit objectives arise from people with different backgrounds, training, and without a clear model. Constraints are identified in a larger social or political setting, and hence determine whether policy development is possible or not. These policy and management objectives and constraints are incorporated into a simple computer model.

The construction of computer models is used to integrate disciplinary understanding of ecological dynamics with management objectives. The construction of the model itself is fraught with uncertainty. At least three levels of uncertainty can be dealt with by this technique. Background noise or variation can be dealt with by including feedbacks on variables in a model. Statistical or parametric uncertainty about forms and values of relationships can be assessed by evaluating alternative sets of equations or by estimating ranges or variations of parameters used in the equation sets. Finally, structural uncertainty in models, or what variables to consider in the model, is a matter of judgment, because situations in nature, such as surprises, cannot easily be dealt with even in the most sophisticated models. The body of literature on adaptive management describes how models should be developed with respect to the tradeoffs between model complexity and utility. These models help clarify uncertainties, but usually cannot resolve them by decomposition or research. Some uncertainty can be highlighted by the use of alternative statistical models that assign probabilities or odds on the possible outcomes.

The models are viewed as hypotheses, and as such cannot be validated; they can only be invalidated in the Popperian view of science. The models are caricatures of reality, only including what is essential. Therefore, what is important is model credibility, not validity. It is only after resisting attempts at invalidation that a model becomes credible. One way of attempting invalidation is to compare the model output with historical data (verified data, not interpreted). Another way is that correlation between the model and historical data does not imply causation. Other means of invalidation include trial and error approaches, that is, compare the model predictions with what happens in the real world. There may be natural trials where model output can be compared to natural experiments. One can also compare the behavior of alternative models. Once the models (or sets of models) have resisted invalidation, they can be used to evaluate alternative actions. Proponents of adaptive management suggest that the adaptive assessment phase utilizes models to highlight uncertainties and foster communication but they cannot and should not be used to prescribe specific management actions. Most management decisions are gambles because of the inherent uncertainty of outcomes of a management action. In essence, the development of credible or plausible models helps determine what is not feasible, rather than specify a particular set of actions. Indeed, the most important outcome from the assessment phase is to identify a set of alternative policies that can be discussed in the social–political arena.

Once a set of alternative policies has been developed, it is important to communicate these results to policy makers. A number of techniques have been developed, including audiovisual presentations (slide shows, story boards, and computer animations) that attempt to simplify the complexities of both the analysis and results. Communication of alternative potential policies to social and political entities of the resource system should start early in the assessment process. Decision makers and the public should be actively involved from the beginning, rather than passively informed at the end. There are three arguments made for doing this: (1) that it is a fairer way of making social decisions; (2) that participation of stakeholders in a decision process increases the likelihood that they will accept the outcome; and (3) that people bring important knowledge to assessment process that will improve its quality.

Adaptive assessments have been applied to hundreds of resource issues around the world over the past three decades. Some of the assessments resulted in no future activity. Some have led to the development of new monitoring programs, especially in the cases where insufficient data were available for sorting among competing explanations, whereas other assessments led to dramatic transformations in understanding and new management efforts. Unless the assessment can generate an agreed-upon set of actions for testing the uncertainties of the system, it is unlikely that the system will move into the adaptive or active management phase.

Adaptive Management

Once a policy is chosen, then two implementation activities are launched: management actions and monitoring. The implementation phase is the experimental phase, in which hypotheses are challenged, that is, the policy is subject to invalidation.

The design of management actions covers a wide range of options, in terms of the scientific rigor. In an experiment, all sources of variation are identified, and used in the design. In the adaptive sense, some of the variations can be accounted for, but unforeseen dynamics are likely to appear. Two types of adaptive management have been identified, active and passive. Actively adaptive actions probe the sources of variation; hence, actions cover a wide range of possible treatments in order to uncover a wide range of system response. The passively adaptive approach lets natural processes provide the sources of variation.

Monitoring should be the means by which policies are tested and evaluated, rather than collecting data for the sake of gathering information. Monitoring is the way in which policies are put at risk of invalidation. It should involve key variables in the system and choice of these key variables is critical. Variables to be monitored are critical ecosystem processes identified in the modeling exercises. The information developed from monitoring should be useful for future learning. The usual limitations on fiscal and human resources should not decide what variables to monitor, but be used as way of forcing the important variables into monitoring.

Summary

Prior experience indicates that adaptive environmental assessment and management has been successful at identifying key points of intervention and developing lurches of understanding. That is, they have led to new management actions and improved understanding of resource dynamics. These include situations of adaptive waterfowl harvest in North America, harvest of fisheries in the reefs of Australia, and sediment transport in the Grand Canyon. Many other cases, however, point to a failure of adaptive management due to institutional rigidity, bureaucratic inertia, lack of ecological resilience, and lack of social capital in the form of trust and cooperation. Adaptive assessment and management is one approach that can help bridge the gap between science and resource management, but its implementation requires a small set of agreed-upon hypotheses, sufficient resources to conduct and evaluate experiments, and a social willingness to accept failure at appropriate scales.

Adaptive management proposes that ecological management be viewed as a series of experiments rather than the application of a solution. Adaptive management is an ongoing process that combines assessment with management actions that are structured in a way to learn about the complexities of system dynamics as well as to achieve social objectives. Assessing a system requires synthesizing available data to generate a set of competing alternative hypotheses about particular sets of resource problems and social objectives. Management actions are designed to sort among alternative hypotheses, prior to implementation. These actions are evaluated by monitoring key system indicators. While these activities are described linearly, adaptive management usually is designed to be an iterative process that develops a social dialog about a system. Modifying and creating ecological management institutions is difficult. Consequently, one of the main challenges of adaptive management is to develop new ways to enable social learning and experimentation.

Further Reading

Carson, R., 1962. Silent Spring. Boston: Houghton Mifflin Company.
Gunderson, L., Holling, C.S., 2002. Panarchy: Understanding Transformations in Human and Natural Systems. Washington, DC: Island Press.
Gunderson, L., Holling, C.S., Light, S.S., 1995. Barriers and Bridges to the Renewal of Ecosystems and Institutions. New York: Columbia University Press.
Holling, C.S., 1978. Adaptive Environmental Assessment and Management. Caldwell, NJ: Blackburn Press.
Lee, K., 1993. Compass and Gyroscope: Integrating Science and Politics for the Environment. Washington, DC: Island Press.
Ludwig, D., Hilborn, R., Walters, C., 1993. Uncertainty, resource exploitation, and conservation: Lessons from history. Science 260, 17–36.

Margoluis, R., Salafsky, N., 1998. Measures of Success: Designing, Managing, and Monitoring Conservation and Development Projects. Washington, DC: Island Press.
Millennium Assessment, 2003. Ecosystems and Human Well-Being a Framework for Assessment. Washington, DC: Island Press.
Nyberg, B., 1999. An Introductory Guide to Adaptive Management for Project Leaders and Participants. Victoria, BC: BC Forest Service.
Pahl-Wostl, C., 2002. Towards sustainability in the water sector – The importance of human actors and processes of social learning. Aquatic Sciences 64, 394–411.
Pahl-Wostl, C., Jaeger, C.C., Rayner, S., *et al.*, 1998. Regional integrated assessment and the problem of indeterminacy. In: Cebon, P., Dahinden, U., Davies, H.C., Imboden, D.M., Jaeger, C.C. (Eds.), Views from the Alps: Regional Perspectives on Climate Change. Cambridge: MIT Press, pp. 435–497.
Salafsky, N., Margoluis, R., 1999. Greater than the Sum of Their Parts: Designing Conservation and Development Programs to Maximize Results and Learning. Washington, DC: Biodiversity Support Program.
Schön, D.A., 1983. The Reflective Practitioner: How Professionals Think in Action. New York: Basic Books.
Walters, C.J., 1986. Adaptive Management of Renewable Resources. New York: McGraw-Hill.

The Anthropocene

Clive Hamilton, Charles Sturt University, Canberra, Australia

Origin of the Term

At a workshop in Mexico in 2000, the atmospheric chemist Paul Crutzen became frustrated at the repeated references to the Holocene. "Stop using the word Holocene", he interrupted. "We're not in the Holocene any more. We're in the . . . the . . . the . . . (searching for the right word) . . . the Anthropocene" (Steffen, 2013: 486). Those assembled immediately concurred that the Earth had entered into a new geological epoch to succeed the Holocene. The rapid increase in the concentration of carbon dioxide in the atmosphere due to the burning of fossil carbon and its cascading effects throughout the Earth System is the principal reason for agreeing that the planet had shifted out of the previous epoch.

The new idea quickly caught on and spread, initially though two short articles by Crutzen (one coauthored) (Crutzen and Stoermer, 2000; Crutzen, 2002). The work of Earth System scientists around the world soon became organized around this powerful, new concept.

At one level, the definition of the Anthropocene is straightforward. It refers to the very recent rupture in the functioning of the Earth System as a whole arising from the impact of human activity (Hamilton, 2016a: 251). This definition is conceptually rich and requires explanation. In particular, the concept of "the Earth System" is very new and often misunderstood. The kinds of human activity capable of disrupting the Earth System have to be separated from those that merely alter the landscape or interfere in an ecosystem. And (as we will see) the claim that the rupture is "very recent" has been a matter of intense scientific debate (although it is now largely settled).

Before beginning to explore these ideas it's important to remind ourselves that the Anthropocene is put forward as a new epoch to be added to the Geological Time Scale (GTS). The Scale divides the Earth's 4.5 billion-year history into ages, epochs, periods, eras and eons in ascending order of significance. The Earth entered the Holocene epoch around 11,700 years ago when the planet began to warm rapidly from an ice-age freeze. It is a division of the Quaternary period (of the Cenozoic era) that began some 2.58 million years ago. Around 10,000 years ago the Earth's mean temperature stabilized at a level close to the one that prevailed until the start of the industrial revolution some two centuries ago.

The International Commission on Stratigraphy (ICS) is the body formally charged with responsibility for deciding on the divisions of the Geological Time Scale (GTS). Stratigraphy is the branch of geology specializing in the study of rock layering and what it can tell us. In 2009 it created the Anthropocene Working Group (AWG), chaired by geologist Jan Zalasiewicz, to prepare a report on the case for adding the Anthropocene as a new epoch to the GTS to succeed the Holocene. The AWG now faces an intellectual challenge because the new division has been proposed not by geologists digging down into the rock layers but by Earth System scientists observing changes in the Earth System "above ground." The coming together of stratigraphy and Earth System science is provoking creative tension leading to new scientific insights. At the time of publication, a formal decision about the Anthropocene had not been made, but there is no doubt that work on this rich and potent idea will continue to grow.

Earth System Science

The idea of the Anthropocene cannot be properly understood except in the context of the emergence of Earth System science. Thomas Kuhn's language of "paradigm shift" and "scientific revolution" (Kuhn, 1962) has often been applied too freely, yet it seems to be justified in this case. (This section and the next draw heavily from Hamilton (2016b) and Hamilton and Grinevald (2015). I owe a large debt to Jacques Grinevald for helping develop much of the analysis in this section, including the intellectual history of the concept.) If a paradigm is a distinct set of assumptions and patterns of thought then there can be no doubt that Earth System science represents a strikingly new way of thinking about the Earth as an object.

The emergence of Earth System science can be traced in the following terms (Hamilton and Grinevald, 2015). The foundations were laid in the 1950s, during the Cold War, when oceanographic and atmospheric sciences were transformed and globalized with the emergence of early computers (Edwards, 2010: 129ff). Numerical weather prediction for large regions became routine only in the 1960s, but it is worth noting that the notion of a *global* climate (one of the components of the Earth System) became widely accepted by scientists only after the Second World War. Except for a few speculative commentaries, "the climate" had previously been considered a local or regional phenomenon. A popular handbook still being published in 1961 argued that "the notion of a global climate made little sense" because the weather is too changeable between the poles and the tropics (Edwards, 2010: 67–9).

Systems ecology was developed in the 1960s, notably within the Radiation Ecology Section of Oak Ridge National Laboratory, itself an outgrowth of the Manhattan Project that developed the atom bomb (Coleman, 2010). The Laboratory became an important center for understanding the CO_2-energy-climate problem, leading to the creation of the Carbon Dioxide Information Analysis Center in 1982. Biophysical modeling of the biosphere (building on legacy of the great Russian geologist and geochemist Vladimir Vernadsky) was developed in the 1980s by Russian scientists in collaboration with Western colleagues at the International Institute for Applied Systems Analysis and later at the Potsdam Institute for Climate (Jørgensen, 2010).

Within these increasingly global ways of thinking, scientists began to understand the Earth as a total functioning entity composed of a number of interacting "spheres"—the atmosphere, the hydrosphere, the cryosphere, the biosphere and the lithosphere. In particular, climatologists built a deeper conception of the climate as a global phenomenon, and not one confined to the atmosphere, with expanding understanding of the ocean-atmosphere circulation.

This emerging notion of the Earth as a total functioning entity would be married to the new conceptual approach known as system dynamics. The computer-based methodology of system dynamics was developed by Jay W. Forrester at Massachusetts Institute of Technology in the 1950s. Initially it was applied to social systems. An invitation from the Club of Rome to apply it to the functioning of Earth as a complex "world ecosystem" led to the influential and controversial book *The Limits to Growth*, published in 1972 (Meadows et al., 1972). System dynamics brought to Earth science essential concepts for thinking about the Earth as an entity that behaves as a *complex system* rather than as a set of cause-and-effect processes.

The complexity of a complex system goes beyond the idea that relationships between variables are complicated. It means the system is more than the sum of its parts because it has certain properties that are inherently difficult to represent in mathematical models. They include: nonlinearity, so that a tiny change somewhere can bring about a very large change somewhere else; feedback loops, which dampen or amplify a change leading either to system stability or run-away change; self-organization, so that the system spontaneously creates order out of chaos; and, most puzzlingly, emergent properties, that is, properties that belong to the system but cannot be found in any individual element of it, so that something new can emerge with no apparent cause.

It was into this flux of emerging ideas that in 1974 James Lovelock and Lynn Margulis introduced the Gaia hypothesis, the idea that life and the Earth's nonliving environment form "a self-regulating system that maintains the Earth's climate and the composition of the atmosphere in a habitable state" (Lenton 2016: 4–5). The Gaia concept was a forerunner of the Earth System; both reject the idea that the Earth is a fixed, static ball of rock on which the other components develop and move. The scientific establishment (with rare exceptions) rejected the Gaia hypothesis, often vehemently, not least because it challenged the basis of the theory of evolution, according to which all causation runs one way, from the natural environment to the development of living entities (Lovelock, 1988: xiv–xv). The hypothesis has undergone substantial modification since then, but its foundational idea of the Earth as a system whose components evolve together is close to that of the conception of the Earth being developed by Earth System science.

Meanwhile, scientists had been developing instruments to measure changes in global processes. Sometimes in science, new instruments give rise to new thinking and occasionally to great conceptual breakthroughs (Shapin and Schaffer, 2011). Three measurement projects should be noted: the launch from the 1970s of artificial satellites to monitor, inter alia, global change in air pollution, land cover and solar radiation fluxes; the recording of variations in atmospheric CO_2 at the Mauna Loa Observatory in Hawaii by Charles David Keeling from the late 1950s (Keeling, 1970); and, the revelations of ice core drilling in Antarctica which began in the early 1980s (Jouzel et al., 2013). Insights from these measurement projects nudged scientists toward a firmer grasp of the globality, complexity and systemic nature of the Earth.

In addition, worldwide programs on measuring biogeochemical cycles had begun in the 1970s. The first reports on the global carbon (and other biogeochemical) cycles, the greenhouse effect and climatic change (Bolin et al., 1986) played a vital role leading into the first report of the Intergovernmental Panel on Climate Change in 1990. As climate scientists have explored the global climate system more deeply they have come to realize that climate change is not only (or even predominantly) a phenomenon of the atmosphere but is tightly connected with changes in the hydrosphere, the cryosphere and the biosphere, and even the lithosphere in the form of earthquakes, volcanism and new kinds of sedimentation (McGuire, 2012).

In the 1980s, scientists became increasingly concerned about the impact of human activity on the planet as a whole, beyond the damage being done to local or regional environments. Evidence of the dangers to life from the expanding hole in the stratosphere's ozone layer and from global warming due to burning fossil fuels led them to build global monitoring networks and to explore the interactions of the various components of planetary processes through elaborate models that attempted to capture the connections between the major processes governing the Earth. As Zalasiewicz et al. (2017: 208) would observe: "it was long the case that most of the geological community thought of the human impact on the Earth's geology as trivial and fleeting by comparison with large-scale geological processes acting over millions of years. That general opinion began to change in the second half of the 20th century . . .". The emerging paradigm of Earth System science led in 1986 to the launch of the International Geosphere-Biosphere Programme (IGBP), which would become the institutional heart of global ecology and Earth System thinking (Grinevald, 1990; Steffen and Tyson, 2001).

The Earth System and the Anthropocene

In their textbook, Charles Langmuir and Wally Broecker provide a concise definition of the Earth System.

> The various parts of the Earth system – rock, water, atmosphere – are all involved in interrelated cycles where matter is continually in motion and is used and reused in the various planetary processes. Without interlocked cycles and recycling, Earth could not function as a system. . . . In the last fifty years or so we have come to recognize the movements in all Earth's layers, including the plates at the surface, the mantle and the core as well as the atmosphere and ocean (Langmuir and Broecker, 2012: 20, 22).

Earth System thinking, which emerged fully in the 1990s, is the integrative metascience of the whole planet as a unified, complex, evolving system beyond the sum of its parts. It is a transdisciplinary and holistic approach integrating earth sciences and life sciences, as well as the "industrial metabolism" of humankind, all within a systems way of thinking, with special focus on the nonlinear dynamics of a system. (With thanks to Jacques Grinevald for helping formulate this definition.) By contrast, ecological thinking, which emerged fully in the 1960s and 1970s, is the biological science of the relationship between communities of organisms and their local environments. While ecological thinking may or may not draw on systems concepts (much practical ecology deploys cause-and-effect thinking), Earth System science could not exist without systems thinking. The gulf between the two remains even when local environments are aggregated up to the "global environment." The global environment thought this way is not the Earth System, understood as the Earth taken as a whole in a constant state of movement driven by interconnected cycles and forces, from the planet's core to the atmosphere and out to the Moon, and powered by the flow of energy from the Sun. It is a single, dynamic, integrated system, rather than a collection of ecosystems.

Now that we are clear about what the Earth System is (and when the idea emerged), we can gain a proper understand of the Anthropocene. In what might be regarded as the canonical statement on the Anthropocene, the new epoch is defined by the fact that the "human imprint on the global environment has now become so large and active that it rivals some of the great forces of Nature in its impact on the functioning of the Earth system" (Steffen et al., 2011: 843). Some authors have claimed to have found "precursors" of the concept in the work of natural scientists going back as far as the late 19th century. It is claimed that, among others, Antonio Stoppani (in 1873), G.P. Marsh (in 1874), Vladimir Vernadsky (in 1929), Pierre Teilhard de Chardin (in 1925) and Edmund Le Roy (in 1927) developed the idea of humankind as a significant geological or morphological force and thereby discovered the essential idea of the Anthropocene.

However, Hamilton and Grinevald (2015) argue that none of these earlier scientists could have grasped the Anthropocene as a disruption to the functioning of the Earth System because the concept of the Earth System did not emerge until the 1980s and 1990s. Instead, they were referring to spreading human impact across the Earth's surface, a phenomenon of the landscape or ecosystems or in some cases the local or regional climate. And instead of being a rupture, the change they identified was understood as a continuous one. None had a conception of disturbance to global biogeochemical cycles; as we saw, even the climate was understood as a regional phenomenon rather than a globally integrated system until the 1950s.

The effect of finding precursors is to gradualize the new epoch so that it is no longer a break in the functioning of the Earth System due primarily to the burning of fossil fuels. Finding these alleged precursor concepts actually "undermines the radical novelty of the concept and the actuality of the proposed new geological epoch" and thereby "misconstrues the suddenness, severity, duration and irreversibility of the Anthropocene . . ." (Hamilton and Grinevald 2015: 3, 8–9).

Not long after Crutzen conceived the concept of the Anthropocene, others began to appropriate the concept of the Anthropocene into their own ways of understanding, at times distorting its meaning. The pioneers of the Anthropocene—mainly Paul Crutzen, Will Steffen, John McNeill, Colin Waters and Jan Zalasiewicz—have repeatedly reminded us that the concept holds water only if it can be shown that humans have had a *detectable* impact on the *functioning of the Earth System*. It's worth belaboring this point: the fundamental test of the Anthropocene is whether human activity disturbs the functioning of the Earth System as a whole, does so discernibly, and is outside the range of natural variability (Steffen et al., 2007).

Much of the debate has centered on the question of when the Anthropocene can be said to have begun. Initially, Crutzen and his coauthors identified its beginning at the end of the 18th century with the onset of large-scale coal burning to power the Industrial Revolution in England. More recently, the Anthropocene Working Group has argued that the new epoch is better dated from the years after the Second World War when the impact of humans on the Earth System as a whole was unmistakable (Zalasiewicz et al., 2015a). The decades after the War have been dubbed "the Great Acceleration," years in which there was "a sharp step change in the nature, magnitude, and rate of human pressures on the Earth System, driving impacts that push the system beyond the Holocene basin of attraction" (Steffen et al., 2016: 336). Within that basin of attraction negative feedbacks keep the system in a state that is recognizably Holocene.

In 2003, palaeoclimatologist William Ruddiman published a paper arguing that the Holocene-Anthropocene shift occurred not at the end of the 18th century with the Industrial Revolution but 5000–8000 years ago with the onset of forest clearing and farming, which led to enhanced levels of CO_2 and CH_4 in the atmosphere (Ruddiman, 2003). However, Ruddiman's interpretation of the data turned out to be unpersuasive. Crutzen and Steffen (2003) pointed out that human impact on the Earth System 5000 to 8000 years ago is not discernible in the data, and certainly was not large enough to upset permanently the stability of the Holocene Earth. The data do show a shift occurring in the late 18th century, the beginning of the Industrial Revolution. And the charts also show an incontrovertible leap after World War Two.

> The mid-twentieth century was a pivotal point of change in the relationship between humans and their life support systems. The period of the Anthropocene since 1950 stands out as the one in which human activities rapidly changed from merely influencing the global environment in some ways to dominating it in many ways (Crutzen and Steffen, 2003).

A number of other analyses have rejected the evidence for Ruddiman's early Anthropocene hypothesis, leading the Intergovernmental Panel on Climate Change to conclude that it is not clear that the small and very slow changes in CO_2 and CH_4 from around 8000 years ago were due to human activity, let alone were sufficient to change the course of the Earth System (Ciais et al. 2013, 483–5 & Fig. 6.6).

A number of analysts have interpreted the Anthropocene as no more than the continuation of human impacts on the landscape or ecosystems. If true, then it is not a disruption to the functioning of the Earth System. This has been taken to its furthest point by Ellis who claims that humans

> have been reshaping the terrestrial biosphere, and perhaps even the global climate, for millennia. The entire past 11,000 years of the Holocene might simply be renamed the Anthropocene (Ellis, 2013: 32).

In this view the Anthropocene is just another name for the Holocene. How is such a conclusion possible? The words to notice in the quoted passage are "the terrestrial biosphere," human changes to which are enough, in Ellis's view, to define a new geological epoch. This misinterpretation of the Anthropocene arises from a misunderstanding of Earth System science. Ellis believes earlier work on biomes and anthromes—where anthromes are global ecological patterns influenced by human activity—can be simply scaled up to get to the Earth System, a reading that misses the significance of the Earth System.

Elsewhere, Ellis and Ruddiman have asked: "Does it really make sense to define the start of a human-dominated era millennia after most forests in arable regions had been cut for agriculture . . .?" (Ruddiman et al., 2015). The answer is "yes," if those human activities did not change the functioning of the Earth System. The pioneers of the Anthropocene concept have always written of the new epoch *in contrast to* the Holocene as a geological epoch, never in terms of landscapes or ecosystems modified in the Holocene. None of the leading exponents of Earth System science believes that changes in the terrestrial biosphere alone can bring about a new epoch, and even less so if we are thinking of vegetation and landscape ecology (Lenton and Williams, 2013: 382).

The writing of the new geological epoch into established ways of thinking is an instance of what Kuhn called "drastically restricted vision" (Kuhn, 1962). A similarly restricted vision has come from a contribution from archeology. Again, the issue is the starting date of the new epoch. In a paper titled "The onset of the Anthropocene", the abstract begins:

> A number of different starting dates for the Anthropocene epoch have been proposed, reflecting different disciplinary perspectives and criteria regarding when human societies first began to play a significant role in shaping the earth's ecosystems (Smith and Zeder, 2013).

One need not read past this sentence to know that the authors have misconstrued the new epoch, and that their conclusions about the onset of the new epoch must be mistaken. It's the very last letter, the "s" in ecosystems, that gives it away. The Anthropocene does not begin when humans first play "a significant role in shaping the earth's ecosystems"; it begins when humans first play a significant role in shaping the *Earth*, that is, the Earth that evolves as a totality, as a unified, complex system comprised of the tightly linked atmosphere, hydrosphere, cryosphere, biosphere and lithosphere. It is not about changes to ecosystems except insofar as ecosystems are affected by changes in the functioning of the Earth System.

It is also possible to misread the nature and significance of the Anthropocene by viewing it through the lens of geography. Reprising the "pre-Columbian Anthropocene hypothesis" of Dull et al. (2010), Lewis and Maslin (2015) locate the start of the new epoch in 1610. They put forward a complex narrative covering the colonization of South America, introduced diseases, depopulation, forest regrowth, trans-continental trade, species exchange and pollen counts, all of which are said to be associated with a small dip of 10 ppm in the atmospheric concentration of CO_2 in 1610. However, the analysis failed to show numerically that the dip in CO_2 changed the functioning of the Earth System or was caused by human activity (Hamilton 2015), and a number of Earth System scientists pointed out that in the pre-industrial Holocene there were many comparable dips in atmospheric CO_2 concentration and that a change of 10 ppm is well within the range of natural variability in the Holocene (Zalasiewicz et al., 2015a, b).

A common feature of these misreadings of the Anthropocene through lenses other than that of Earth System science is that, by treating the new epoch as a continuation of landscape or ecosystem change going back centuries or millennia, they divorce it from modern industrialization and the burning of fossil fuels. In this way they miss the central fact that the Anthropocene represents a rupture in Earth history (and so deprive it of its dangerous quality).

An interesting attempt to bridge the gap between Earth System science and other disciplines has been put forward in the form of the concept of the Palaeoanthropocene," "the period from the beginning of human effects on the environment to the beginning of the Anthropocene" (Foley et al., 2013). The beginning of the Palaeoanthropocene would then be diffuse (including all of the Holocene and much of the Pleistocene), associated with local rather than global events, and not be linked to geological boundaries or changes in the functioning of the Earth System. All of the erroneous concepts of the Anthropocene discussed above would fall into this pre-Anthropocene zone, with the Anthropocene reserved for describing the era of disruption in Earth System processes.

Reconciling Earth System Science With Stratigraphy

One of the prime objectives of ongoing research on the Anthropocene is to identify a suitable stratigraphic signature in rock or soil sediments or in ice layers that can be used to meet the criterion for the declaration of a new geological epoch (Waters et al., 2016). There are plenty of deposits that might suit, including "technofossils" like concrete and plastics, black carbon and spherical carbonaceous particles showing up in sediments worldwide from around 1950, and geochemical signatures like pesticide residues and polyaromatic hydrocarbons. Along with elevated CO_2 and CH_4 concentrations in the atmosphere, these "novel signatures" signal the impact of humankind on the functioning of the Earth System and provide indicators of the persistence of that impact.

There have, however, been some important criticisms of the proposal to add the Anthropocene to the GTS. A full review is not possible here, but they have been detailed and replied to by Zalasiewicz et al. (2017). One consequence of these critiques is that the AWG has been more explicit about its determination to make the case for a new epoch using strictly stratigraphic criteria and to eschew "above ground" indicators of Earth System change.

One of the more telling criticisms of the Anthropocene concept is that, unlike all previous divisions in the GTS, the proposed new epoch emerged not out of stratigraphy (digging down through rock layers) but by observation of human impact on the Earth System (Finney and Edwards, 2016). It therefore represents a kind of "geology of the future", which is highly uncertain. Zalasiewicz et al. (2017) acknowledge this break with established methods, and concede that formal acceptance of the proposed epoch must hinge on stratigraphic evidence. But they note that the Anthropocene concept arises from the entirely unexpected overlap of geological time with (human) historical time. This suggests two important observations.

First, the Anthropocene proposal poses the new question of whether and how the established boundaries in the GTS—which are often defined by a major change in the Earth's biota captured in the fossil record—demarcate shifts in the functioning of the Earth System. The relationship between them is described by Zalasiewicz et al. (2015a) as follows:

> An effective geochronological and chronostratigraphical boundary often reflects a substantial change in the Earth system, so that the physical and chemical nature of the deposits, and their fossil contents, are recognizably different above and below the boundary.

The association between the boundaries and shifts is prompting new thinking about what happened in the Earth System at some of those major boundaries.

Second, the mixing of geological time and historical time gave rise to a seminal observation that has had a profound impact among social scientists and humanities scholars thinking about the Anthropocene. The observation was made by the historian Dipesh Chakrabarty who noted that the arrival of the Anthropocene represents the "the collapse of the age-old humanist distinction between natural history and human history," two kinds of history that had always existed in entirely separate realms (Chakrabarty, 2009: 201).

Social Science and the Anthropocene

Kinds of Critique

No more than a quick overview is possible here, but the idea of the Anthropocene has attracted intense interest from social scientists and humanities scholars. It has attracted very little interest from philosophers, although Danowski and Viveiros de Castro (2017) is a notable exception. For social scientists, the new epoch's arrival is telling us something novel and profound about humankind's relationship with the Earth. In addition, the idea itself is lobbed like a hand grenade into a deep epistemological divide among social scientists, that is, a disagreement over the nature and scope of kinds of knowledge. While some have fruitfully analyzed the scientific and para-scientific discourses around the concept (Bonneuil and Fressoz, 2016), among the interpretations and developments of the new concept two broad epistemological approaches are evident.

One, which I will call humanist, takes as given the facts presented by Earth System science and asks about the social and philosophical meaning of the changed physical relationship between humans and the Earth, and how to respond to it. In this view, the advent of the Anthropocene shows up the disastrous consequences of the contradiction between human separation and domination of nature and our unavoidable dependence on it. In the light of our deeper understanding of our ecological entanglement in ecological processes, the modern philosophical distinction between nature and society remains valid, although the collision of natural history and human history in the Anthropocene, arising from the fact that humans have become a force of nature (Chakrabarty 2009), complicates and weakens the distinction. However, Hornborg argues that the "physical mixing" of nature and society does not invalidate the analytical distinction between them (2015: 58).

For the other approach, which I will call posthumanist (also known as new materialist), the distinction between the human and the natural, known as Cartesian dualism or the subject-object split, has never been valid, and the arrival of the Anthropocene is seen as a vindication (Latour 2017: 3). (Posthumanists writing in this domain are often influenced by the work of Bruno Latour (1993, 2014).) Unlike the humanists, the posthumanists challenge the epistemological priority given to the Earth System science underlying the Anthropocene, arguing that since the Anthropocene is a human-dominated epoch "the facts" are as much in the domain of the social sciences as the physical sciences (although to make this claim they must draw first on evidence generated by Earth System science.)

The difference between the two approaches reflects the split in philosophy and the social sciences that opened up in the 1970s with the emergence of what is sometimes named postmodernism, that is (in the present context), a rejection of the notion of objective truth and the assertion that all claims to knowledge are socially constructed and therefore the products of social and historical conditions. In this view, the scientists' Anthropocene—both the idea of it and even the epoch itself—is subject to interpretation by understanding the social conditions that gave rise to it.

Humanist Analysis

Scholars in both camps have taken exception to Paul Crutzen's term for the new epoch. To name the new epoch "the Anthropocene" seems to attribute the shift in the Earth System to the actions of humankind as such, *anthropos*, whereas in fact it was brought on by the actions of a rich, white minority. Malm and Hornborg (2014) write that if we talk of "the geology of mankind in general" then the new epoch "must have its roots in the properties of that being." Yet it is *divisions* between humans rather than any homogeneity that we must look to for the origins of "sociogenic" climate change. This argument has led to a proliferation of alternative names, such as the Capitalocene, the Technocene and the Econocene.

Scholars working in the broad Marxist tradition have integrated the new science of the Anthropocene into their critique of capitalism. Foster et al. (2010) write that the Anthropocene highlights the disastrous "ecological rift" between humanity and nature that is embedded in modern capitalist society. In that society, humanity is alienated both from itself and from nature, and this alienation is so entrenched that humans are on the verge of destroying the conditions of life. In a way that is structurally inescapable, the system places its survival first, and solutions to the ecological crisis must be made to work around it. They identify traditional social science as part of the problem.

Angus (2016) analyzes the contradictions of "fossil capitalism," in which fossil fuels in the hands of powerful corporations have permitted enormous improvements of human health and wealth yet which now threatens to destroy it all. The Anthropocene is properly understood as a biophysical phenomenon, but it is also a socio-ecological one representing a shift in the relationship between human society and the rest of the natural world. Only by understanding the social origins of the Anthropocene in capitalism can we find an answer to the looming crisis, and that is the replacement of capitalism with societies that are radically egalitarian and fully democratic, operating on the best ecological principles.

Humanists accept that our embeddedness in networks of biophysical relationships constrains and complicates what humans can do; but, in contrast with the posthumanists, they do not accept that this embeddedness redefines and deflates human *agency*. Traditionally in sociology, "agency" refers to the capacity of free beings to act independently, and is posed in opposition to the power of social structures (like social classes) and institutions (like governments) to guide and constrain individual behavior. While posthumanists (as we will see) take agency away from humans and discount political action to effect social change, humanists reject what they see as an essentially conservative political philosophy that denies political agency and therefore the capacity to bring about the radical social changes needed to respond to the climate crisis and the Anthropocene.

Posthumanist Critique

Many social scientists have interpreted the arrival of the Anthropocene as proving the danger of anthropocentrism—the predisposition of modernity to which they are most strongly opposed. Anthropocentrism is the assignment to humankind a moral and practical right to exploit nature in its own interests. *Anthropos* is "central" in the sense that the rest of nature is understood through human experience and values. For moderns, this standpoint gives humans the right to dominate and exploit other creatures and natural systems for their own benefit. Although the critique of anthropocentrism had been made by eco-philosophers for some decades, it has been taken up by posthumanist philosophers. For them, the idea of the Anthropocene can be critiqued by extending the catalogue of "essentialist" assumptions embedded in European colonialism, that is, Eurocentrism, androcentrism and racism. Moore puts it bluntly: "Just as we have been learning to move beyond the dualisms of race, gender, sexuality, and Eurocentrism over the past four decades, it is now time to deal with the source of them all: the Nature/Society binary" (Moore 2015: 4).

Humanists, by contrast, do not identify a generalized belief in anthropocentrism as the source of our predicament. They locate the problem in a social structure that treats the natural environment as an endless, exploitable resource in the pursuit of profits and growth.

For posthumanists, anthropocentrism is possible because of the modern Cartesian split between subject (the human agent) and object (the world of things it acts on), a division now shown as impossible by the evidence of humankind's deep entanglements with the natural world (Bennett, 2009). So, for example, primates are more like us than we realize; natural systems and resources have shaped our cultures deeply; our gut fauna influences how our brains work. Morton (2015) writes that "simply allowing nonhumans to be what they are, namely entangled with us in all kinds of strange ways, neither absolutely reducible to human access nor completely divorced from it". Haraway (2015) tries to overcome dualism by arguing that both "nature" and "human" are cultural categories.

Drawing on ecological science, the posthumanists extend the postmodernist argument that ideas have no objective truth, because they are products of social conditions and structures, to argue that our *physical selves* have no independence because we are deeply embedded in the natural world, just as other creatures are. They use this fact to criticize all beliefs in anthropocentrism. (Curiously, they reject the central role of humans precisely at the time, in the Anthropocene, of humankind's greatest power.)

Determined to prick the bubble of human hubris and deny any view that humans have the right to dominate nature, posthumanists draw on the entanglement of humans in nature to diminish or "redistribute" agency through the natural world. Other creatures and even natural processes and objects are attributed agency. For critics of posthumanism, it makes no sense to attribute agency to a mollusk except by changing the meaning of agency. Hornborg (2017) helpfully draws a distinction between nonliving objects that have *consequences*, living entities other than humans that have *purposes* (due to their sentience and communication), and humans who have *intentions* (because they can reflect on their purposes). When agency is stripped of all element of choice it becomes mere influence.

And, argues Hamilton (2017), posthumanists make an epistemological mistake in regarding human domination of nature as a form of oppression equivalent to racism or androcentrism. The other targets of postmodern critique were forms of discrimination characteristic of relations between people (among human subjects), while anthropocentrism applied to the relationship between humans and the natural world (blithely crossing the boundaries between subjects and objects or nonhuman subjects). The arrival of the Anthropocene is a resounding affirmation of the extraordinary power of humans, yet at this very moment posthumanist theorists are attempting to cut humans down to size and deflate their power and significance to the planet.

In the Anthropocene, it is necessary to recognize *both* the ecological truth that humans are embedded in nature in a thousand ways, but also that we have unprecedented power in nature. We therefore need to draw a distinction between, on the one hand, anthropocentrism as a description of the uniqueness of humans as a species and our actual power on Earth, and, on the other, the attitude of arrogance and mastery over nature that has led to the new epoch, and which manifests in ecomodernists' calls for geoengineering schemes that aim to regulate the planet's functioning. A "new anthropocentrism" has been proposed (Hamilton 2017) as a way of acknowledging humankind's exceptional, real power on Earth and an attitude of caution and modesty in the face of the Earth System's essentially uncontrollable and unpredictable power.

Acknowledgments

I am grateful to Alf Hornborg for helpful comments on a draft.

References

Angus I (2016) *Facing the Anthropocene: Fossil capitalism and the crisis of the Earth System*. New York: Monthly Review Press.
Bennett J (2009) *Vibrant matter: A political ecology of things*. Durham, NC: Duke University Press.
Bolin B, et al. (1986) The greenhouse effect, climatic change and ecosystems. *SCOPE*, vol. 29. Chichester: John Wiley & Sons.
Bonneuil C and Fressoz J-B (2016) *The shock of the Anthropocene*. London: Verso, London.
Chakrabarty D (2009) The climate of history: Four theses. *Critical Inquiry* 35: 197–222. Winter.
Ciais P, et al. (2013) Carbon and other biogeochemical cycles. In: Stocker TF, Qin D, Plattner G-K, and Tignor M, et al. (eds.) *Climate change 2013: The physical science basis*. Cambridge: Cambridge University Press.
Coleman DC (2010) *Big ecology: The emergence of ecosystem science*. Berkeley: University of California Press.
Crutzen P (2002) Geology of mankind. *Nature* 413: 23.
Crutzen P and Steffen W (2003) How long have we been in the Anthropocene era? An editorial comment. *Climatic Change* 61: 253.
Crutzen P and Stoermer E (2000) The "Anthropocene". reprinted in In: Robin L, Sörlin S, and Warde P (eds.) *The Future of Nature*, pp. 483–485. New Haven: Yale University Press.
Danowski D and Viveiros de Castro E (2017) *The ends of the world*. Cambridge: Polity Press.
Dull R, et al. (2010) The Columbian encounter and the little ice age: Abrupt land use change, fire, and greenhouse forcing. *Annals of the Association of American Geographers* 100(4): 755–771.
Edwards P (2010) A vast machine. In: *Computer models, climate data, and the politics of global warming*. Cambridge, MA: MIT Press.
Ellis E (2013) Using the planet. *Global Change* 81: 32–35.
Finney SC and Edwards LE (2016) The "Anthropocene" epoch: Scientific decision or political statement? *GSA Today* 26(3): 4–10.
Foley SF, et al. (2013) The Palaeoanthropocene – The beginnings of anthropogenic environmental change. *Anthropocene* 3: 83–88.
Foster JB, Clark B, and York R (2010) *The ecological rift: Capitalism's war on the Earth*. New York: Monthly Review Press.
Grinevald J (1990) L'effet de serre de la Biosphère: de la révolution thermo-industrielle à l'écologie globale. *Stratégies énergétiques, Biosphère et Société* 1: 9–34. available online.
Hamilton C (2015) Getting the Anthropocene so wrong. *The Anthropocene Review* 2(2): 102–107.
Hamilton C (2016a) Define the Anthropocene in terms of the whole Earth. *Nature* 536: 251.
Hamilton C (2016b) The Anthropocene as rupture. *The Anthropocene Review* 3(2): 93–106.
Hamilton C (2017) *Defiant Earth: The fate of humans in the Anthropocene*. Cambridge: Polity Press.
Hamilton C and Grinevald J (2015) Was the Anthropocene anticipated? *The Anthropocene Review* 2(1): 59–72.
Haraway D (2015) Anthropocene, Capitalocene, Plantationocene, Chthulucene: Making kin. *Environmental Humanities* 6: 159–165.
Hornborg A (2015) The political ecology of the Technocene: Uncovering ecologically unequal exchange in the world system. In: Hamilton C, Bonneuil C, and Gemenne F (eds.) *The Anthropocene and the global environmental crisis*, pp. 57–69. London: Routledge.
Hornborg A (2017) Artifacts have consequences, not agency: Toward a critical theory of global environmental history. *European Journal of Social Theory* 20(1).
Jørgensen S (ed.) (2010) *Global ecology. A derivative of encyclopedia of ecology*. Amsterdam: Elsevier.
Jouzel J, Lorius C, and Raynaud D (2013) *The White Planet. The evolution and future of our frozen world*. trans. from the French [2008] by Teresa Lavender Fagan Princeton: Princeton University Press.
Keeling CD (1970) Is carbon dioxide from fossil fuel changing Man's environment? *Proceedings of the American Philosophical Society* 114(1): 10–17.
Kuhn T (1962) *The structure of scientific revolutions*. Chicago: University of Chicago Press.
Langmuir C and Broecker W (2012) *How to build a habitable planet*, Revised edition. Princeton: Princeton University Press.
Latour B (1993) *We have never been modern*. Cambridge Mass: Harvard University Press.
Latour B (2014) *Politics of nature: How to bring the sciences into democracy*. Cambridge Mass: Harvard University Press.

Latour B (2017) *Facing Gaia: Eight lectures on the new climatic regime*. Cambridge: Polity Press.
Lenton T (2016) *Earth system science: A very short introduction*. Oxford: Oxford University Press.
Lenton T and Williams H (2013) On the origin of planetary scale tipping points. *Trends in Ecology & Evolution* 28(7).
Lewis S and Maslin M (2015) Defining the Anthropocene. *Nature* 519: 171–180.
Lovelock J (1988) *The Ages of Gaia: A biography of our living earth*. Oxford: Oxford University Press.
Malm A and Hornborg A (2014) The geology of mankind? A critique of the Anthropocene narrative. *The Anthropocene Review* 1(1): 62–69.
McGuire B (2012) *Waking the Giant: How a changing climate triggers earthquakes, tsunamis, and volcanoes*. Oxford: Oxford University Press.
Meadows D, et al. (1972) The limits to growth. In: *A report for the Club of Rome's project on the predicament of mankind*. New York: Universe Books.
Moore J (2015) *Capitalism in the web of life*. London: Verso.
Morton, Tim (2015), Anna Lowenhaupt Tsing's Mushroom at the End of the World, Somatosphere (website), December 8, 2015.
Ruddiman W (2003) The anthropogenic greenhouse era began thousands of years ago. *Climatic Change* 61(3): 261–293.
Ruddiman W, et al. (2015) Defining the epoch we live in. *Science* 348(6230): 38–39.
Shapin S and Schaffer S (2011) *Leviathan and the air-pump: Hobbes, Boyle, and the experimental life*. Princeton: Princeton University Press.
Smith B and Zeder M (2013) The onset of the Anthropocene. *Anthropocene* 4: 8–13.
Steffen W (2013) Commentary on 'The "Anthropocene"'. In: Robin L, Sörlin S, and Warde P (eds.) *The future of nature: Documents of global change*. New Haven: Yale University Press.
Steffen W and Tyson P (eds.) (2001) *The Global Environmental Change Programmes, IGBP Science, Vol. 4. Global Change and the Earth System: A planet under pressure,* p. 33, IGBP Secretariat: Stockholm.
Steffen W, Crutzen P, and McNeill J (2007) The Anthropocene: Are humans now overwhelming the great forces of nature? *Ambio* 36(8): 614–621.
Steffen W, Grinevald J, Crutzen P, and McNeil J (2011) The Anthropocene: Conceptual and historical perspectives. *Philosophical Transactions of the Royal Society A* 369: 842–867.
Steffen W, et al. (2016) Stratigraphic and Earth System approaches to defining the Anthropocene. *Earth's Future* 4: 324–345.
Waters C, et al. (2016) The Anthropocene is functionally and stratigraphically distinct from the Holocene. *Science* 351: 6269.
Zalasiewicz J, et al. (2015a) When did the Anthropocene begin? A mid-twentieth century boundary level is stratigraphically optimal. *Quaternary International* 383: 196–203.
Zalasiewicz J, et al. (2015b) Colonization of the Americas, 'little ice age' climate, and bomb-produced carbon: Their role in defining the Anthropocene. *The Anthropocene Review* 2(2): 117–127.
Zalasiewicz J, et al. (2017) Making the case for a formal Anthropocene epoch: An analysis of ongoing critiques. *Newsletters on Stratigraphy* 50(2): 205–226.

Further Reading

Zalasiewicz J, et al. (2016) Scale and diversity of the physical technosphere: A geological perspective. *The Anthropocene Review* 4(1): 9–22.

Biophilia

SR Kellert, Yale University, New Haven, CT, USA

Biophilia is defined here as the inherent human inclination to affiliate with natural systems and processes, most particularly life and life-like (e.g., ecosystems) features of the nonhuman environment. The notion of biophilia advances the idea that people's physical and mental well-being remains reliant on contact with natural systems and processes. This dependency reflects the fact that humans evolved in adaptive response to the fitness and survival requirements of a largely natural and not artificial or human-constructed world. In other words, the evolutionary context for the development of the human mind and body largely was a sensory challenging and diverse natural environment with critical elements being light, sound, odor, wind, vegetation, landscape, water, animals, and more that provided much of the basis for human learning and maturation.

The emergence during the past 5000 years of small- and large-scale agriculture, technology, industry, and the city represents but a small fraction of human evolutionary history. It is unreasonable to assume, thus, that human physical and mental well-being may have escaped the dictates of a largely natural environmental evolutionary context. Yet, many people in modern society assume that human distinctiveness, progress, and civilization reflect our species' capacity to separate from, if not transcend, our biological roots. This entry views this assumption as an illusion, and instead advances the idea that contact with natural systems and processes remains an anvil on which human fitness, health, and productivity continue to be forged, even in an increasingly fabricated urban world.

Biophilia is manifest in a range of genetically encoded tendencies to attach meaning to and derive benefit from, in effect value, the natural world. Nine biological inclinations to value nature have been identified, each reflecting a range of adaptive benefits instrumental in human physical, emotional, and intellectual fitness and well-being. This view of ``the human mind…in evolutionary perspective,'' presumes when these values are adaptively expressed they enhance human performance and development. In other words, these values are regarded as not vestigial in the sense of having evolved in an environmental context no longer relevant to modern life. Collectively, the nine values continue to reflect the richness of the human reliance on natural systems for fitness and security, a web of relational dependency so pronounced an ethic of concern for nature can emerge from a broad understanding of human self-interest from continuing contact with natural systems and processes. **Table 1** provides brief definitions of each of the nine values of biophilia and associated adaptive benefits.

Biophilia is, however, a `weak' biological tendency greatly dependent on learning, experience, and sociocultural support to become functionally manifest. The biophilic values are not `hard-wired' instincts, but rather genetically programmed tendencies that rely on sufficient stimulation and reinforcement to functionally occur, although as `genetically prepared learning rules' they can be easily triggered and learned relatively quickly. Like so much of what makes humans unique, biophilia is subject to people's choice and free will as they respond to their weak biological urges. We may be born with the inclination to affiliate with natural systems and processes, but this tendency remains nascent and atrophied in absence of stimulation and reinforcement. The `genius' of humanity is its extraordinary capacity for learning, creativity, and cultural construction in response to weak inherent tendencies. This ability to affirm or deny much of our biology has fostered considerable human invention and the ability to change and progress. Yet, this adaptability and innovative capacity is ultimately subject to the dictates of biology. Not all individual and cultural creations are functional over time, and some eventually prove damaging and destructive. Human creativity and free will is, thus, a two-edged sword, carrying both the potential for positive innovation and negative self-destruction.

As noted, biophilia is highly dependent on learning, experience, and social support to become functionally manifest. Biophilia can, therefore, be classified as a `biocultural' phenomena, where each biophilic value is the product of both genetics and cultural construction. Considerable diversity consequently occurs in the content and intensity of each value in response to the influences of learning, history, and geography. Yet, this variability is bound over time by the requirements of biological fitness. Each value hypothetically occurs along a continuum where considerable `normal and functional' variation is encountered, but where

Table 1 Typology of values of biophilia

	Definition	*Adaptive benefits*
Esthetic	Physical appeal and beauty of nature	Inspiration, harmony, security
Dominionistic	Mastery and physical control	Physical prowess, self-confidence, mastery skills
Humanistic	Emotional attachment to aspects of nature	Bonding, cooperation, companionship
Moralistic	Spiritual reverence and ethical concern	Order, meaning, kinship
Naturalistic	Direct experience and exploration	Curiosity, discovery
Negativistic	Fear and aversion	Security, protection, awe
Scientific	Systematic and empirical study	Knowledge, understanding, critical thinking skills
Symbolic	Nature in language and expressive thought	Communication, mental development
Utilitarian	Practical and material exploitation	Physical sustenance and security

dysfunction also takes place at the extremes – atrophied development at one end and excessive expression at the other, both resulting in emotional and intellectual deficits and maladaptive behavior.

The biophilia hypothesis that contact with natural systems and processes fosters human health, productivity, and well-being will be briefly reviewed, drawing largely on evidence from recent conducted scientific studies. Following this review, the article will conclude with a limited discussion of how modern, especially urban, society has increasingly diminished biophilic contact with the natural environment, and how this situation can be improved through a different design approach to the human built environment. The brief review of data bearing on the benefits of contact with nature will focus on studies of outdoor experience, healthcare, work, and community relationships.

Evidence Relating Contact with Nature to Human Health and Well-Being

Exposure to parks and gardens has long been associated with restorative and healing affects. Presumed benefits include rest, relaxation, physical restoration, even spiritual renewal. This assumption is reflected in such ancient practices as establishing `sacred' groves, mountains, water bodies, and other prominent natural features. These natural areas were frequently viewed as places where benign deities and guardian spirits resided. In more secular societies, parks and gardens, in both wilderness and more densely populated areas, have also been rationalized by their presumed physical, psychological, and even moral benefits.

The association of protected areas and open spaces with human physical and mental well-being has been more systematically studied in recent years, although the data remain sparse. Several investigations have documented various benefits of human contact with the nonhuman environment in such settings including stress relief, peace of mind, enhanced coping capacity, physical health, illness recovery, and improved critical thinking and problem-solving abilities. Even passive viewing of nature has been correlated with stress reduction, `emotional balance', and improved cognitive functioning. Most of the data have been derived from social surveys, but some investigations have employed more reliable and valid physiological indicators including measures of blood pressure and muscle tension. Overall, these studies have consistently demonstrated largely positive affects on people exposed to park-like settings, gardens, and landscapes. In addition, the natural areas studied have often been characterized by savanna-like meadows, large and mature canopy trees, forest edges, brightly colored flowers and shrubs, the presence of water, and other natural features that have largely proven instrumental in human evolution and survival.

An illustrative study examined the physical, emotional, and intellectual effects of park-like settings in urban areas. Randomly chosen college students were subjected to a series of intellectually demanding and mentally fatiguing tasks and assigned to three groups: one who took a 40 min walk in an urban park, another who walked for an equivalent period in an attractive urban area dominated by buildings and human activity, and a third who remained indoors in a comfortable chair listening to music and reading. After 40 min, each group completed a demanding proofreading assignment. Significantly higher scores on measures of emotional restoration (`levels of positive and negative affect') and cognitive functioning (attentiveness and concentration) occurred among students who walked in the park.

A number of studies have also examined the effects on participants of diverse outdoor recreational activities. Both anecdotal evidence and research results found immersion in relatively undisturbed natural settings often exerted significant and sometimes life-changing impacts, especially on late adolescents in the company of peers. Among the documented effects included: significant improvements in physical fitness, stamina, endurance, coordination, self-esteem, self-confidence, independence, initiative, personal responsibility, coping abilities, seeing tasks to completion, critical thinking, problem solving, curiosity, inventiveness, school and work performance, cooperation, teamwork, conflict avoidance, respect for others, peace of mind, and spiritual well-being.

Another body of research has focused on the effects of domesticated nature on people, particularly gardens and companion animals (pets). The presence of flowers and appealing vegetation has often been correlated in these studies with healing and calming affects, especially among the physically and mentally impaired. Plants and flowers were also found to be nearly universal in modern hospitals, their presence often among the most widely stated healthcare preferences of patients, and generally assumed to exert stress- and symptom-relieving benefits.

Several studies reported simply viewing nature reduced stress, relieved tension, and enhanced recovery among patients suffering from clinically diagnosed disorders. One investigation of patients recovering from gall bladder surgery randomly assigned patients to two types of hospital rooms – one with window views of trees and vegetation, the other of a brick wall. The investigators reported patients with the view of vegetation had significantly faster recovery rates, more positive treatment responses, and required less pain medication than the patients who only looked at the brick wall. Another study of patients recovering from heart surgery involved their being randomly assigned to three types of recovery rooms – one with pictures of water and trees, the second with pictures of abstract art, and the third with only blank walls. Significantly less anxiety and fewer demands for strong pain medication occurred among patients with the pictures of water and trees. Still, a third study of persons facing dental surgery examined three groups of patients – one, exposed to `serene' pictures of nature, a second, who viewed pictures of highly active outdoor scenes (ocean surfing), and a third who saw no pictures at all. Significantly lower blood pressure levels were observed among patients exposed to the serene nature scenes. Finally, another study focused on patients facing dental surgery who were randomly assigned to three treatments: one group exposed to a live fish tank, a second who observed pleasant pictures of nature, and a third who viewed only a blank wall. Nearly all patients showed signs of distress, but significantly lower levels of physical discomfort and `treatment-aversive behaviors' occurred among patients exposed to the fish tank.

A different body of research focused on the effects of companion animals. These studies generally revealed that the presence of companion animals enhanced calm and peace of mind, fostered physical and mental restoration, and relieved loneliness. Various studies of the physically and mentally disturbed reported major improvements in symptoms, more rapid healing, and faster recovery times among persons who had contact with companion animals. One illustrative study examined the effect of companion animals on persons recovering from heart attacks. Patients were matched demographically and symptomatically and randomly assigned to two groups: one who received conventional treatment and the other who were carefully exposed to companion animals. Researchers reported the presence of companion animals resulted in a one-third decrease in mortality rates and corresponding increases in survival and recovery. Another mental health study focused on boys suffering from attention-deficit hyperactive disorder who were randomly assigned to two nature-related activities – an outdoor challenge program that included canoeing and rock climbing, and an animal interaction activity focusing on care for companion animals, with the two groups switched to the other activity midway through the study. Therapeutic gains were reported for both activities, although significantly more positive and lasting benefits were observed among the animal care activity. Caring for companion animals resulted in significant symptomatic improvements, greater learning, improved school performance, speech gains, better attentiveness, and control over impulsive behavior. Moreover, differences were evident 6 months following the program, prompting the researchers to conclude a caring relationship for companion animals can relieve stress, improve social interaction, increase empathy, and contribute to task performance.

Benefits from contact with nature have been further studied in the modern workplace. Most environmental contact was limited to exposure to plants, outside views, natural lighting, natural ventilation, and pictures of nature and decorative art. Yet, the data have been consistent and generally supporting the contention that even restricted contact with nature in the modern work place can enhance health, productivity, and well-being. Most of the initial research focused on work settings with problems resulting from air-tight construction, artificial lighting, and widespread use of toxic chemicals in paints, coverings, furnishings, and other products and materials. The presence of these conditions often correlated with higher levels of respiratory and skin disorders, increased fatigue and absenteeism, diverse physical and psychological ailments, poor morale, and lower productivity. These conditions also prompted the identification of `building-related illness' and `sick building syndrome'.

More recent studies have begun to focus on the positive affects of exposure to nature in the work place, particularly from improvements in natural lighting, natural ventilation, the use of natural materials, and increased direct and representational contact with nature. These studies have revealed these positive environmental conditions can result in significant improvements in employee comfort, satisfaction, morale, health, well being, and productivity. For example, studies of European office and factory workers found viewing nature or the presence of plants could reduce job-related stress, allergies, and improve emotional well-being. Studies in the US also found window views and plants reduced job-related frustration and improved physical and mental well-being. Productivity studies of workers who had contact with plants and views of nature demonstrated fewer errors, more efficient work performance, lower blood pressure, and better attentiveness. Several studies reported workers with improved natural lighting and natural ventilation had significantly better cognitive performance. One particularly ambitious study examined office and manufacturing workers at a furniture company immediately before, immediately following, and 9 months after workers moved from facilities with minimal environmental amenities to new facilities with extensive natural lighting, natural ventilation, natural materials, improved energy efficiency, restored natural areas, greater open space, and walking trails. These changes were correlated with a 20% gain in productivity 9 months after the move to new facilities, as well as significant improvements in job satisfaction, physical health, relaxation, peace of mind, and contentment at work.

A limited number of community-based studies have also been conducted. One ambitious large-scale investigation focused on 18 rural, suburban, and urban neighborhoods within a single overall watershed. This study found a strong correlation between environmental quality, environmental values, and people's physical and mental well-being in the communities studied. Neighborhoods characterized by better environmental quality tended to express more positive values of nature and had a higher quality of life, whereas lower environmental quality communities generally demonstrated less environmental interest and typically had a lower quality of life. Moreover, the relationship of environmental quality, environmental values, and quality of life occurred in both urban and nonurban communities, as well as independent of income and education level.

Regarding the latter finding, it is sometimes assumed that the positive effects of contact with nature is mainly relevant to socioeconomically secure or privileged persons, and that conversely nature is of limited significance for individuals and groups characterized by poverty and social oppression who must cope with more basic needs. This assumption is not only unsupported by the previously cited finding, but also by important studies of Chicago public housing projects home to mainly poor African-Americans. Residents of architecturally identical buildings in these projects were compared, the only difference being that some buildings were surrounded by limited grass and a few trees, while other buildings had no vegetation and were surrounded instead by asphalt and concrete. Residents were randomly assigned to the buildings and had no control over landscaping. Yet, the study found residents of the vegetated buildings had significantly higher levels of physical and emotional well-being, better coping and conflict management skills, and superior cognitive functioning. Moreover, residents of the buildings surrounded by trees and grass revealed better social ties, superior interpersonal relationships with neighbors and strangers, lower violence and crime rates, greater safety and security, and a stronger sense of community than residents of the buildings surrounded by concrete and asphalt.

The final briefly reviewed area of research focuses on the role of contact with nature in childhood development. As noted, the biophilic notion that people have an inherent need to affiliate with nature implies that even for a human species capable of

lifelong learning the most important period for the development of a biological tendency will be childhood. Limited research supports this contention, or as the psychiatrist Harold Searles concluded: ``The non-human environment, far from being of little or no account to human personality development, constitutes one of the most basically important ingredients of human psychological existence.'' Studies of adults reflecting on childhood have found that the outdoors was almost universally cited by them as their most important environment during childhood, although the outdoor areas were often modest and located in backyards and nearby open spaces. Several features of the natural environment were identified in these studies as especially powerful maturational influences on children including: the stimulation of diverse senses, variability of conditions, dynamic and changing circumstances, and the presence of animate and life-like features. These and other characteristics of the natural environment appeared to foster coping and problem solving, creativity and challenge, change and adaptation, and other critical elements of learning and development. Additional research has suggested the most important period for children's contact with nature is middle childhood, roughly between the ages of 6 and 10.

In the most general sense, the inherent tendency to affiliate with nature or biophilia is viewed as the primary causative agent for the association revealed in the various cited studies of contact with nature and human health and well-being. In other words, the assumption is that various genetically encoded tendencies to affiliate with nature developed during the course of human evolution – for example, responses to sunlight, landscape, habitat, weather, color, plant and animal associations, etc. These tendencies fostered varying states of comfort, satisfaction, relaxation, alertness, curiosity, imagination, and more that proved instrumental in human fitness and survival. The actual causal mechanisms or linkages involved in this relationship require far more explanation and, at this point, our understanding must be considered largely speculative.

Decline in Contact with Nature in Modern Urban Society

Despite the evidence presented of the benefits of contact with nature on human health, productivity, and maturation, various modern trends suggest a significant decline in the `direct' experience of the natural environment, especially among urban youth. It is important to note, however, people experience nature in direct, indirect, and vicarious or symbolic ways. Direct experience involves relatively unstructured contact with largely self-sustaining natural features and processes. Indirect experience involves highly structured and organized contact with natural features requiring extensive human input and management (e.g., zoological parks, gardens). Vicarious or symbolic experience involves no actual contact with real or living nature, but rather with the image or representational expression of nature through, for example, the media of books, film, and computers. The apparent decline in human contact with natural systems and processes is principally a decline in direct and spontaneous contact. By contrast, the indirect and vicarious experience of nature appears to have increased in modern times.

As intimated earlier, functional and adaptive behavior is highly reliant, however, on direct environmental experience, especially during childhood. A decline in direct contact with nature, especially among modern youth, has led some to refer to the related phenomena of `nature-deficit disorder' and an `extinction of experience' to describe this condition. Factors associated with a significant decline in adult and children's direct experiences of nature include major biodiversity loss, degradation of natural systems, declining open space, and chemical pollution and contamination. Most of these trends have been linked to rapid urban growth and sprawl, at least as urban development has occurred until now. It is sobering to recognize that nearly three-quarters of the industrially developed world now resides in an urban area, and the majority of the world's population was recently identified as living in or near a city for the first time in human history.

Unfortunately, the prevailing paradigm of design and development of the modern urban environment has relied on massive consumption of energy and natural resources, enormous generation of wastes and pollutants, extensive degradation of natural habitats and loss of biological diversity, and increasing separation if not alienation of people from contact with natural systems and processes. This urban development paradigm is viewed as a design flaw rather than an inevitable and intrinsic failure of modern life. Fundamentally altering this design paradigm will necessitate a radically different development strategy, one sometimes referred to as sustainable but which has been called here `restorative environmental design'.

Sustainable design to date has largely focused on minimizing and avoiding adverse impacts of the human built environment on natural systems and human health. Important elements of this `low environmental impact' approach have included: energy and resource efficiency, waste minimization, avoidance of toxic products and materials, and protecting and restoring natural systems. This approach is vital and necessary but not sufficient for mending the current `nature deficit'. In addition, we also require design strategies that foster beneficial contact between people and nature in places of ecological and cultural familiarity and significance. This latter approach can be called `positive environmental impact' or, for reasons apparent by now, `biophilic design'. Some specific elements of biophilic design in the built environment include: environmental features (e.g., natural materials, natural ventilation); natural shapes and forms (e.g., botanical and animal motifs); natural patterns and processes (sensory variability, aging, and change); light and space (natural light, spaciousness); place-based relationships (historic and ecologic connection to locality); and evolved humans relations to nature (e.g., prospect and refuge, organized complexity). Restorative environmental design is the integrated and complementary combination of both low environmental and biophilic design approaches, a necessary basis for a true and lasting sustainability. Hopefully, this new design paradigm will render more compatible if not harmonious the relationship between the natural and human built environments in even our modern cities, resulting in the protection and restoration of necessary ecosystem services, as well as enhancing people's biophilic needs for positive contact with nature.

Further Reading

Barkow, J., Cosmides, J.L., Tooby, J. (Eds.), 1993. The Adapted Mind: Evolutionary Psychology and the Evolution of Culture. New York: Oxford University Press.
Kahn, P.H., 1999. The Human Relationship with Nature: Development and Culture. Cambridge, MA: MIT Press.
Kahn, P.H., Kellert, S.R. (Eds.), 2002. Children and Nature: Psychological, Sociocultural, and Evolutionary Investigations. Cambridge, MA: MIT Press.
Kellert, S., 1997. Kinship to Mastery: Biophilia in Human Evolution and Development. Washington, DC: Island Press.
Kellert, S., 2005. Building for Life: Designing and Understanding the Human–Nature Connection. Washington, DC: Island Press.
Kellert, S.R., Wilson, E.O. (Eds.), 1993. The Biophilia Hypothesis. Washington, DC: Island Press.
Pyle, R.M., 1993. The Thunder Tree: Lesson from an Urban Wildland. Boston: Houghton Mifflin.
Wilson, E.O., 1984. Biophilia: The Human Bond with Other Species. Cambridge, MA: Harvard University Press.

Carbon Footprint

Dario Caro, Aarhus University, Roskilde, Denmark

Introduction

Climate change represents one of the most relevant challenges for human being. It evolves into a full range of issues such as economy, technology, society and ecology. Limiting the effects of climate change is necessary to achieve sustainable development and equity. Countries' future contributions to the accumulation of greenhouse gas (GHG) emissions in the atmosphere will play a key role in keeping warming below 2°C relative to preindustrial level (UNFCCC, 2015). In a globalized era, countries have different impact on climate change as well as different vulnerability to the effects of the resulting climate change. They also have different capacities to address mitigation and adaptation. Monitoring GHG emissions at different level of analysis (national, city, organization, product) constitutes an important reference point upon which environmental policies and strategies able to drive the GHG emission mitigation can be based. Indeed, emission drivers need to be considered at different scales for making possible to set goals for GHG emissions mitigation, develop a management plan and implement new policies aimed to reduce the impact of climate change.

In the last decades there has been a growing interest in estimating and revealing GHG emission drivers via analyses based on carbon footprint at different scales. The carbon footprint originates from the concept of ecological footprint which is a measure of the impact on the environment expressed as the amount of land required to sustain natural resources. However, the carbon footprint of a functional unit, when it is not associated with ecological footprint, is the climate impact under a specified metric that considers all relevant emission sources, sinks and storage in both consumption and production within the specified spatial and temporal system boundary (Peters, 2010). Although a specific definition of carbon footprint has been not stated, according to Wiedmann and Minx (2008) carbon footprint is the measure of the total amount of carbon dioxide emissions directly and indirectly caused by an activity or accumulated over the life stages of a product. Specifically, the carbon footprint is the overall amount of GHG emissions associated with a country, city or product, along its supply chain including end-of-life recovery and disposal. The carbon footprint is an environmental indicator and as such it needs to be used in appropriate contexts thus providing right information. When properly used, it is essential for making decisions and performance evaluations and for allowing policy-makers to have a solid basis upon which climate policies can be established and implemented.

Accounting Systems: Metrics and Standard

Carbon footprint analysis has provided important findings that have helped to shape climate policy at the global and national level. The carbon footprint approach gives consumers the responsibility for their environmental impacts in terms of GHG emissions, regardless of where those impacts occur. Two different sources of emissions are considered on the basis of the system analyzed: (a) direct emissions; occurring inside the boundary of the system, (b) indirect emissions; occurring outside the boundary of the system for satisfying the demand. In general, three scopes regarding three categories of emissions are considered. Scope 1 or direct emissions are defined as emissions from sources that are owned or controlled by the system. Scope 2 or energy indirect GHG emissions are defined as emissions from the consumption of purchased electricity, steam, or other sources of energy (e.g., chilled water) generated upstream from the system. Scope 3 or other indirect emissions are defined as emissions derived from upstream and downstream system activities occurring outside the boundary of the system analyzed. Indeed, the ideal accounting framework of the carbon footprint should reveal the emission source (type and location), the destination and the belonging scope.

In general, the accounting systems used to determine the carbon footprint should meet the requirements of the definition. As such, a carbon footprint can be developed for various different functional units at different scales and using different methods. Concerning the country scale, the way of taking into account GHG emissions is still largely discussed. In particular the IPCC-based accounting and the consumption-based accounting provide two different perspectives on how assign the responsibility of GHG emissions embodied in trade. Although presently, the IPCC inventories represent the official reference point to estimate GHG emission at the national level for countries within the UNFCCC, consumption-based accountings by using environmentally extended input–output (EEIO) analyses are largely developed by researchers worldwide. The IPCC inventories as well as EEIO analyses may be also performed at city scale. However, for several different reasons, their utilization at the city level may be insufficient in providing a complete overview of the system. In recent years, academic researchers proposed a combination of different methods to overcome their respective drawbacks. Such hybrid models aim to merge the strength of different methods becoming more complete with respect to the primary models. For example the combination of the EEIO with LCA represents an innovative hybrid model which is mainly used for estimating the carbon footprint at the city level. Carbon footprint also represents an important tool for estimating direct and indirect GHG emissions generated within the range defined by the organizations (industries or companies) themselves. Although, organizations, companies or industries substantially contribute to the economic growth, they also consume a lot of energy and produce a large amount of waste which also release GHG emissions in the atmosphere. Developing a carbon footprint analysis at organizational scale is therefore an important step for identifying the sources of GHG emissions occurring within the organizational boundaries and provide solutions for reducing them. It may be also aimed specifically at the product level and nowadays, a great consensus on the carbon footprint at the product level has been build up. The carbon footprint is applied by many companies for specific products in all sector worldwide and it represents a solid advanced tool for knowledge-based decision making in the context of sustainable consumption and production. It stimulates companies to achieve the best environmental performances in their production by suggesting actions aimed to reduce GHG emissions of their products and thus promoting the development of the sustainable innovation and technology.

The carbon footprint has not a standardized accounting system to be developed. Although the development of the carbon footprint follows basilar concepts to be performed, the method strictly depends on the different level of analysis. Following, the accounting systems for estimating carbon footprint at different scales are presented and discussed.

Country Scale

In general, GHG emissions accounting at the national level should reflect how national policies contribute to the reduction of the climate change. It means that actions that mitigate GHG emissions should be incentivized whereas action that increase GHG emissions should be discouraged. In this context, carbon footprint plays a key role, being an effective reference point for climate policies. Multiregional input–output (MRIO) models are one of the current instrument for estimating the carbon footprint at the national level. Multiregional input–output models have been developed to study the economic and production networks, and associated environmental impacts, that lie upstream of a given purchase. The use of MRIO models in environmental accounting is constantly increasing. Numerous improvements have been achieved in recent years for the assessment of the carbon footprint at the national level. In a MRIO model, national input–output tables, representing financial transactions between economic sectors within a country, and trade flow tables, showing the value of exports and imports by country and economic sector, are linked together in one coherent accounting framework (Lutter et al., 2016). This core of a combined, multinational interindustry transaction matrix is furthermore linked to primary inputs on one hand and final demand on the other hand. Multiregional input–output models allow analyses at a multiregional level, thus avoiding the double counting of emissions. It should be noted that carbon footprint at the national level represents an alternative GHG accounting with respect to the traditional approach based on the IPCC guidelines. Although the IPCC guidelines are widely used and provide punctual measures of GHG emissions at the national level, their geographical approach implies that the countries where the finished products are actually consumed take no responsibility for the environmental impacts generated by the producer countries, thus neglecting the international trade effect. A MRIO analysis complements the geographical approach by including all driving forces for GHG emissions associated with consumption and focusing on the localization of emissions, especially those embodied in international trade (Jakob and Marschinski, 2013). In short, the carbon footprint is capable to capture the environmental impact of a national economy in terms of GHG emissions. In a globalized era, estimating the trade effect in terms of GHG emissions is relevant as a large amount of goods are internationally traded. In this context, policy relevant issues connected with the trade such as carbon leakage, allocation of emissions embodied in trade and border-tax adjustments are a basic part of carbon footprint analysis. Although there are several advantages to develop a carbon footprint analysis at the national level, some questions arise concerning the feasibility of the actual implementation of the MRIO models. While the MRIO models have been widely approved within the scientific community their adoption when dealing with national emissions accounting has so far been limited by a number of implementation issues. In particular MRIO models are labor-intensive and not always available as a long, continuous time series. Moreover, currently MRIO databases provide results, without accompanying estimates of reliability and uncertainty. In this context, the recently developed multiregional, environmentally-extended database called “Exiobase” proves to be cutting-edge in the reduction of data uncertainty (Tukker and Dietzenbacher, 2013). Anyway, all the GHG accounting systems have their limitations especially when they deal with macro-scale analyses. A complementation of different methods, when properly developed may also be an important alternative approach. For instance, complementing carbon footprint with traditional IPCC-based accounting may reveal both the effects of policies aimed to

encourage the development of low carbon technologies within the national boundaries and the effects of policies addressed to the final consumption of goods, thus also including the trade effect in the computation.

City Scale

The carbon footprint evaluating the environmental impact of cities or subnational regions represents an important assessment to provide a valuable tool for local policy decision makers. An important aspect when a carbon footprint is applied to cities is where the city emissions occur which means the spatial boundary considered. In this context, the accounting system should be properly selected also considering the limited amount of data available at the city level. Once the accounting system has been selected it is necessary to apply a model to assess the carbon footprint of the city. Although there exist standardized methods used to calculate and report GHG emissions at different level of analysis (national, regional, city), hybrid models which combine the strength of different methods are presently an active area of research and are increasingly being used for estimating the carbon footprint at the city level.

IPCC guidelines

The Intergovernmental Panel on Climate Change (IPCC) has defined a complete method to standardize the computation of GHG emissions at national level. Currently, the monitoring of GHG emissions within a territorial system (country, region, etc.) uses IPCC guidelines to realize annual inventories assessing the amount of six main gases (CO_2; CH_4; N_2O; HFCs; PFCs; SF_6), outlined in the Kyoto Protocol (IPCC, 2006). Although the IPCC method has been performed for countries, it has been also adapted at the subnational level such as regions and cities. The estimation of GHG emissions by using the IPCC method cannot be considered a carbon footprint measure because it only reports direct emissions from sectors and subsectors within city boundaries regardless of where output of the production is consumed. It means that the IPCC method only takes in account the Scope 1 and some parts of the Scope 2, thus neglecting parts of the Scope 2 and the entire Scope 3. However some studies showed that a GHG inventory developed by using the IPCC method is able to suggest suitable plans to mitigate GHG emissions and provide important information to public authorities interested in reducing GHG emissions at the subnational level (Bastianoni et al., 2014).

Input–output analysis

The input–output analysis is a top down model able to take in account transactions between activities measured in monetary units and extend them at the environmental level in terms of GHG emissions (environmental extended input–output analysis, EEIO). It has long been recognized as a useful and consistent technique to estimate carbon footprint at different level of analysis. So, monetary flows are converted into GHG emissions thus considering the associated emissions embodied in trade. Indeed, some input–output based studies have revealed that substantial GHG emissions can be embodied in goods and services that are traded and, therefore, not typically included in the IPCC based accounting. On the basis of the data availability the input–output analysis can be used at each level of scale. In particular input–output matrixes are necessary to fulfill the analysis and they are not always available at the city level. However the use of input–output analysis in GHG accounting is increasing and numerous improvements have been achieved in recent years, such as specific advanced databases for the assessment of the carbon footprint at the city level (Dong et al., 2016).

Life cycle assessment

Life cycle assessment (LCA) is a very popular analysis for reporting potential environmental loads and resources consumed in each step of a product or service supply chain. Although its use is mainly associated with the estimation of the environmental impact of products, some studies have also used a LCA for cities. Since in LCA data are collected for all processes that have been identified as relevant to include within the chosen system boundary, its application at the city level imply a time-consuming process requiring a large amount of input data. Therefore, estimating the carbon footprint at the city level by LCA is not recommended due to its complexity. However, LCA has a great level of accuracy and is able to provide responses to actions as well as important information for policy makers interested in improving the environmental performances and understanding the potential of GHG mitigation art the city level.

Hybrid models

The use of LCA datasets in EEIO modeling is a practice to improve the EEIO analysis, as the detail, accuracy, representativeness and technological specificity of the life cycle processes are notably higher in the LCA than the environmental satellite accounts associated with IO tables. EEIO models are capable of tracing back all the intermediate steps of the supply chain and identifying the sources of impact. Service and capital goods expenses data are provided by EEIO whereas the process-based data are added by LCA. The complementation of EEIO with LCA can provide benefits, as LCA, unlike EEIO, is a standardized and continuously improved tool for environmental accounting. Although some still-existing methodological challenges (Caro et al., 2015) EEIO/LCA is one of the most promising models for estimating the carbon footprint at the city/region level. Additional hybrid models have been presented and performed complementing LCA and EEIO. For instance, a combination of material flow accounting with EEIO (Ramaswami et al., 2012) or emergy and LCA (Pincetl et al., 2012). Such models aim to overcome specific drawbacks in taking into account environmental impacts that mass and energy flow analyses neglect. Alternative dynamic methods for estimating the carbon footprint at city scale have been recently such as City Carbon Map and City Carbon Network, respectively introduced by Wiedmann

et al. (2016) and Chen et al. (2016). City Carbon Map was applied to Melbourne. It is a consistent method based on EEIO to report the carbon footprint, showing the origin (physical source) and the destination (allocation of supply chain to final products) of the emissions in a carbon map. Basically, Carbon Map is a two dimensional decomposition of the carbon footprint of a city's final demand. The City Carbon Network, starting from the Carbon Map approach, focuses on the intercity embodied emission flows across countries, giving more information about trade partners and thus providing important insights for local policy makers. Although an improvements in data and methodology is still required these two dynamic carbon footprint's accountings are capable to assess the total impact of the commitments that city made at the Paris Climate Conference thus being two promising methods for the next years.

Organizational Scale

Over the last years the application of carbon footprint at the organizational level has constantly increased with so many companies interested in improving their environmental performances. In the next two paragraphs the two assessment standards for developing a carbon footprint analysis at organizational scale are presented.

GHG protocol

The GHG Protocol Corporate Accounting and Reporting Standard, which has been the first standard for estimating carbon footprint at organizational scale, helps organizations to identify, calculate, and report GHG emissions, providing specific guidelines for performing a GHG emission inventory at organizational scale (WRI, 2011). The GHG protocol website provides freely available electronic GHG calculators able to estimate carbon footprint associated with specific sources or sectors of the organization. In connection with the United Nations Framework Convention on Climate Change, the GHG Protocol covers the accounting of seven GHGs such as carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), hydrofluorocarbons (HFCs), perfluorocarbons (PCFs), sulfur hexafluoride (SF_6) and nitrogen trifluoride (NF_3). The GHG Protocol not only focuses on results, but also includes a deepened analysis to identify the most effective reduction opportunities.

The key steps in calculating an organizational carbon footprint are (i) defining organizational boundaries; (ii) defining operational boundaries; (iii) estimating carbon footprint and (iv) reporting the final achievement. The definition of the organizational boundaries depends on the characteristic of the organization and several aspects should be considered such as the organizational structure, geographic location and purpose of information. The choice of the operational boundaries is also relevant because it determines which emission sources will be quantified such as Scope 1 and 2 (it should be included) and Scope 3 (not always included). The carbon footprint is generally estimated multiplying activity data by standard emissions factors. In some cases organizations may need to track emissions over time in order to identify their temporal advancements and a specific base year is selected. Finally, organizations develop reports to inform internal and external partners thus promoting actions addressed to mitigate GHG emissions at the organizational level.

ISO 14064

In 2006, ISO released the ISO 14064 standard, which is an international standard for the determination of boundaries, quantification, mitigation and removal, used to guide the companies to measure and control the GHG emissions (ISO 14064, 2006). ISO 14064 has been prepared in three parts. Part 1 provides the principles and requirements for designing, developing, managing and reporting organization level GHG inventories. It includes requirements for determining boundaries, quantifying emissions and removals, and identifying specific company actions or activities aimed at improving GHG management. Part 2 provides the requirements for determining scenarios and provides the basis for GHG analysis to be validated and verified. Part 3 provides principles, requirements and guidance for validation and verification of the analysis. Three key steps are required for developing carbon footprint at the organizational level: setting GHG inventory, quantifying GHG emissions, and finally verifying and reporting the final achievement. ISO 14064 provides the steps to develop an inventory that is not only able to be easily verified but can be compared to the inventories of other organizations.

Most cases, the methodology used for estimating GHG emissions is based on the general equation: emission = activity data × emission factor. Activity data is recommended to be collected at site level and data collection should be described in detail when reporting the final outcome. Emission factors relative to each activity data are generally obtained from literature such as IPCC guidelines or more specific national handbooks. It should be noted that if the analysis does not include the Scope 3, carbon footprint is not properly estimated because the total amount of GHG emissions indirectly caused by the organization is not captured. However, specific assumptions may be explicated to make the analysis valuable and representative as well.

Implementing the ISO 14064 allows organizations to promote credibility in GHG quantification, facilitate the trade of GHG credits, support the development of strategies and plans for reducing GHG emissions and finally disseminate their ability in achieving environmental targets. The application of standard measurement methodologies such as the GHG protocol and ISO 14064 have increased in the last decades and by now they are important reference points for estimating carbon footprint at the organizational level. The need to generate a common language and platform to address GHG reduction is more and more important and the public and the public and private body's efforts may play a key role in mitigating GHG emissions worldwide.

Product Scale

Life cycle assessment aims at estimating the magnitude and significance of the potential environmental impacts of a product or a service throughout its entire life cycle. The impact categories of a LCA represent environmental issues of concern to which LCA results may be assigned. One of the key impact categories considered in an LCA is climate change for the estimation of the GHG emissions released along the life cycle of products that is from the extraction of raw materials, the manufacture of goods, and their use by final consumers including recycling, energy recovery and ultimate disposal. The carbon footprint at the product level is a LCA focused on the climate change category. As a consequence, when the carbon footprint of a product is estimated, other important environmental impacts are not considered. This is mainly due to the relevance that climate change has with respect to other environmental impacts, especially in terms of imminence and magnitude of the issue. However, it should be highlighted that achieving sustainable consumption also requires the evaluation of all environmental impacts.

Data sources for estimating the carbon footprint of a products are therefore obtained from LCA databases with so many sources of data, software tools and handbooks on LCA available. The key steps in calculating the carbon footprint of a product are: (i) identifying all materials, activities and processes that contribute to the life cycle of the product; (ii) defining the system boundary; (iii) estimating the carbon footprint; (iv) reporting results obtained by specific report or declaration.

A typical problem with LCA has mainly concerned the comparability of different studies. For instance, assumptions are an important component of the LCA process and they may have effect on final results. In this context, a standardization of the process has been required and the LCA community has widely worked on the fulfillment of specific standard for calculating the carbon footprint of products. Currently the international standards ISO 14067 provide robust and practice-proven requirements for performing transparent and accepted carbon footprint calculations. It specifies principles, requirements and guidelines for the quantification and communication of the carbon footprint of a product which are based on International Standards on LCA (ISO 14040 and ISO 14044) for quantification and on environmental labels and declarations (ISO 14020, ISO 14024 and ISO 14025) for communication.

Conclusions

The decision by national and international governments to take measures to limit GHG emission calls for suitable tools for objectively monitoring and quantifying emission, as well as checking mitigation/reduction programs. As a consequence of potential exposure to carbon pricing, also companies are increasingly interested in understanding their carbon footprint. The carbon footprint represents a crucial tool for developing an environmental plan on the GHG emission assessment and their reduction. It acts as bridge between the scientific community and decision makers, suggesting the best environmental actions to policy makers and guiding future management choices. It plays a key role when used as tool for labeling products in order to incentivize consumers toward a more sustainable consumption and so improve their environmental lifestyle. It can also be a tool for generating a positive competition between companies in terms of environmental impact. In this context, the carbon footprint is capable to capture the trade effect, thus including companies outside the consuming country.

The recent standardization of carbon footprint at the city and product level is an important evidence of how companies and local institutions place the carbon footprint as tool in their decision making (Afionis et al., 2017). Instead, at the international level the utilization of carbon footprint and its potential ability in addressing issues such as carbon leakage and emissions embodied in trade have still been not given sufficient consideration. In facts, while carbon footprint has become an important indicator for supply chain improvements at micro level, its application as tool for evaluating strategies and taking decisions at macro scale has still not been sufficiently approved from the international community. However, the adoption of carbon footprint at the national level might have deep consequences on the global trading system as well as on the international environmental policies aimed at mitigating GHG emissions at country scale. Indeed, with respect to the traditional IPCC-based accounting, the carbon footprint would assign GHG responsibility to consumer countries in international trade, thus inducing all the subjects involved in commercial dynamics to improve their environmental performances. Anyway, there is a large potential for carbon footprint to be used in policy at a variety of scales and so far, the application of carbon footprint also at country scale has been so much useful for identifying important issues to a wide audience. While its utilization is expected to increase over the next years, further advancements in the development of the carbon footprint analysis are still needed. For instance, most of the studies have neglected emissions from land-use change and the proposed methodologies for including such substantial emissions source still suffer from different shortcomings. The lack of a standardized uncertainty analysis associated with the carbon footprint at different scales is also required. While in so many carbon footprint related studies the uncertainty analysis is not considered, it may add valuable information thus strengthening the case studies analyzed. Future standardization actions for an acceptable level of uncertainty are therefore needed at each level of analysis.

References

Afionis S, Sakai M, Scott K, Barrett J, and Gouldson A (2017) Consumption-based accounting: Does it have a future? *WIREs Climate Change* 8: e438. https://doi.org/10.1002/wcc.438.

Bastianoni S, Marchi M, Caro D, Casprini P, and Pulselli FM (2014) The connection between 2006 IPCC GHG inventory methodology and ISO 14064-1 certification standard. A reference point for environmental policies at sub-national scale. *Environmental Science and Policy* 44: 97–107.

Caro D, Rugani B, Pulselli FM, and Benetto E (2015) Implications of a consumer-based perspective for the estimation of GHG emissions: The illustrative case of Luxemburg. *Science of the Total Environment* 508: 67–75.

Chen G, Wiedmann T, Wang Y, and Hadjikakou M (2016) Transnational city carbon footprint networks—Exploring carbon links between Australian and Chinese cities. *Applied Energy* 184: 1082–1092.

Dong H, et al. (2016) A review on eco-city evaluation methods and highlights for integration. *Ecological Indicators* 60: 1184–1191.

IPCC-Intergovernmental Panel on Climate Change (2006) In: Eggleston HS, Buendia L, Miwa K, Ngara T, and Tanabe K (eds.) *2006 IPCC Guideline for National Greenhouse Gas Inventories*. Japan: IGES.

ISO14064 (2006) *2006-Geneva, Switzerland, International Organization for Standardization*. .

Jakob M and Marschinski R (2013) Interpreting trade-related CO_2 emission transfers. *Nature Climate Change* 3: 19–23.

Lutter S, Giljum S, and Bruckner M (2016) A review and comparative assessment of existing approaches to calculate material footprints. *Ecological Economics* 127: 1–10.

Peters G (2010) Carbon footprints and embodied carbon at multiple scales. *Current Opinion in Environmental Sustainability* 2: 245–250.

Pincetl S, Bunje P, and Holmes T (2012) An expanded urban metabolism method: Toward a systems approach for assessing urban energy processes and causes. *Landscape Urban Planning* 107: 193–202.

Ramaswami A, Chavez A, and Chertow M (2012) Carbon footprinting of cities and implications for analysis of urban material and energy flows. *Journal of Industrial Ecology* 16: 783–785.

Tukker A and Dietzenbacher E (2013) Global multiregional input-output frameworks: An introduction and outlook. *Economic Systems Research* 25: 1–19.

UNFCCC (2015) In: *Decision 1/CP.21: Adoption of the Paris Agreement, Paris Climate Change Conference; 2015 Nov 30–Dec 11; Paris, France*.

Wiedmann TO and Minx J (2008) A definition of "carbon footprint". In: Pertsova CC (ed.) *Ecological economics research trends*. Hauppauge, NY: Nova Science.

Wiedmann TO, Chen G, and Barrett J (2016) The concept of City carbon maps: A case study of Melbourne, Australia. *Journal of Industrial Ecolology* 20: 676–691.

WRI (2011) *The greenhouse gas protocol: A corporate accounting and reporting standard*. World Business Council for Sustainable Development: Geneva, Switzerland.

Ecological Economics 1☆

Robert Costanza, Australian National University, Canberra, ACT, Australia

Basic Worldview and Goals

Ecological economics starts with the observation that the human economy is a subsystem of society and the larger ecological life support system. It recognizes that humans are a part of this larger ecological system and not apart from it. Humans have shaped and modified their supporting ecosystems since the time of their appearance as a species, sometimes sustainably, sometimes not. In the past, this human presence (the economic subsystem) was relatively small in scale compared to the size of the rest of the supporting ecosystem. In the last century, due largely to the utilization of fossil fuels, the human subsystem has expanded so dramatically that it is now a major component of the overall system. Unlike the situation in the majority of human history, we now live in a relatively "full" world and have entered a new geologic epoch called the "anthropocene." This changes everything. In a full world context, the goal of the economic subsystem can no longer be simply expansion and growth with little regard to the rest of the system. We must now consider the whole system and the goal must shift from economic growth to sustainable development. Growth implies increasing in quantity or size, while development implies improvement in quality without necessarily increasing in size. In a full world context, the goal must shift from creating "more" to creating "better"—to create a sustainable and desirable future.

This shift in primary goals and vision for the future has profound implications for analysis and policy, across the full range of academic disciplines and human activities. For example, if one's goals include ecological sustainability then one cannot rely on the principle of "consumer sovereignty" on which most conventional economic solutions are based, but must allow for coevolving preferences, technology, and ecosystems. One of the basic organizing principles of ecological economics is thus a focus on this complex interrelationship between ecological sustainability (including system carrying capacity and resilience), social sustainability (including distribution of wealth and rights, social capital, and coevolving preferences), and economic sustainability (including allocative efficiency in the presence of highly incomplete and imperfect markets). The complexity of the many interacting systems that make up the biosphere means that this involves a very high level of uncertainty. Indeed, uncertainty is a fundamental characteristic of all complex systems involving irreversible processes, and ecological economics is particularly concerned with problems of uncertainty. More particularly, it is concerned with the problem of assuring sustainability under uncertainty. Instead of locking ourselves into development paths that may ultimately lead to ecological collapse, ecological economics seeks to maintain the resilience of the highly interconnected socioecological system by conserving and investing in natural and social capital assets in a balanced way with investments in human and built capital.

History

Ecology and economics share the same Greek root, *oikos*, meaning "house." Ecology is, literally, the "study of the house," while economics is the "management of the house," where the house is taken to be the world or any part of it. Thus ecological economics implies studying and managing the world in an integrated way, taking full advantage of our accumulated knowledge and understanding of both the natural and the social parts of the system.

Ecological economics has historical roots as long and deep as any field in economics or the natural sciences, going back to at least the 17th century. Nevertheless, its immediate roots lie in work done in the 1960s and the 1970s. Kenneth Boulding's classic *The Economics of the Coming Spaceship Earth* set the stage for ecological economics with its description of the transition from the "frontier economics" of the past, where growth in human welfare implied growth in material consumption, to the "spaceship economics" of the future, where growth in welfare can no longer be fueled by growth in material consumption. This fundamental difference in vision and worldview was elaborated further by Herman Daly, who in 1968 recast economics as a life science, akin to biology and especially ecology, rather than a physical science like chemistry or physics. The importance of this shift in "preanalytic vision" cannot be overemphasized. It implies a fundamental change in the perception of the problems of resource allocation and how such problems should be addressed. More particularly, it implies that the focus of analysis should be shifted from marketed resources in the economic system to the biophysical basis of interdependent ecological and economic systems and their coevolution over time.

Ecological economics is not, however, a single new paradigm based on shared assumptions and theory. It is instead a "metaparadigm." Rather than espousing and defending a single discipline or paradigm, it seeks to allow a broad, pluralistic range of viewpoints and models to be represented, compared, and ultimately synthesized into a richer understanding of the inherently complex systems it deals with. It represents a commitment among economists, ecologists, and other academics and practitioners to

☆*Change History:* February 2018. R. Costanza made changes to the text and references.

This is an update of R. Costanza, Ecological Economics 1, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 999–1006.

learn from each other, to explore new patterns of thinking together, and to facilitate the derivation and implementation of effective economic and environmental policies. Ecological economics is deliberately and consciously pluralistic in its conceptual underpinnings. Within this pluralistic metaparadigm, traditional disciplinary perspectives are perfectly valid "as part of the mix." Ecological economics therefore includes some aspects of neoclassical environmental economics, traditional ecology, and ecological impact studies, and several other disciplinary perspectives as components, but it also encourages completely new, more integrated ways to think about the linkages between ecological and economic systems.

Ecological economics has also developed a solid institutional base. After numerous experiments with joint meetings between economists and ecologists, the International Society for Ecological Economics (ISEE) was formed in 1988 and currently has over 3000 members worldwide. The journal of the society, *Ecological Economics*, published its first issue in February 1989 and is currently publishing 12 issues per year, with an impact factor taking it to the top one-fifth of all economics and all environmental journals. Major international conferences have been held since 1990, with attendance reaching as high as 1500. Several ecological economic institutes have been formed around the world, a significant number of books have appeared with the term ecological economics in their titles, and a fair number of university courses, certificate programs, and graduate degree programs have also developed.

Links With Natural Sciences

Ecological economics' explicit links with the natural sciences result in a more scientific approach, which is inherently more pluralistic and empirically grounded. It places humans and human behavior in a broader historical, evolutionary, and ecological context. Humans are seen as a part of the natural world, not abstractions in isolation from nature and each other. It is problem-based, not tool-based, and its methods include any that are applicable to the problems at hand. These include everything from participatory processes to envisioning alternative futures to complex systems simulation modeling. It recognizes the importance of envisioning and the limits of the positive-normative dichotomy. It goes well beyond interdisciplinary dialog. It aspires to be a truly transdisciplinary science.

The broad spectrum of relationships between ecosystems and economic systems is the locus of many of our most pressing current problems (i.e., sustainability, climate disruption, species extinction, income and wealth distribution) but it is not covered adequately by any existing discipline. Environmental and resource economics, as they are usually practiced, are subdisciplines of neoclassical economics focused on the efficient allocation of scarce environmental resources but generally ignoring ecosystem dynamics and scale issues, and paying only scant attention to wealth distribution issues. Ecology, as it has historically been practiced, sometimes dealt with human impacts on ecosystems, but the more common tendency was to stick to "natural" systems and exclude humans.

Ecological economics also focuses on a broader set of questions and goals than the traditional disciplines. Here, again, the differences are not so much the newness of the questions or goals, but rather the attempt to integrate them. They can be stated as both questions and goals, since they represent both complex questions requiring further research to fully understand and goals that most people would agree are worthy of the title:

1. Assessing and insuring that the scale of human activities within the biosphere are ecologically sustainable; how do we stay within the biophysical planetary boundaries?
2. Distributing resources and property rights fairly, both within the current generation of humans and between this and future generations, and also between humans and other species; and
3. Efficiently allocating resources as constrained and defined by (1) and (2) above, including both marketed and non-marketed resources, especially social and natural capital and ecosystem services.

These questions/goals are interdependent and yet need to be addressed hierarchically. The problem of ecological sustainability needs to be solved at the level of preferences or technology, not at the level of optimal prices. Only if the preferences and production possibility sets informing economic behavior are ecologically sustainable can the corresponding set of optimal and intertemporally efficient prices be ecologically sustainable. Thus the principle of "consumer sovereignty" on which most conventional economic solutions are based is only acceptable to the extent that consumer interests do not threaten the overall system, and through this the welfare of future generations. This implies that if one's goals include ecological and social sustainability then one cannot rely on consumer sovereignty, and must allow for coevolving preferences, technology, and ecosystems.

Material and Energy Flows

One focus of the work on joint ecological economic systems has been material and energy flows. A dominant theme in this body of work has been the grounding of conventional economic models in the biophysical realities of the economic process. This emphasis shifts the focus from exchange to the production of wealth itself. Cleveland traces the early roots of this work dating back to the physiocrats. The energy and environmental events of the 1960s and the 1970s pushed work in this area to new levels. Energy and material flow analysis in recent times is rooted in the work of a number of economists, ecologists, and physicists. Economists such as Boulding and Geogescu-Roegen demonstrated the environmental and economic implications of the mass and energy

balance principle. Ecologists such as Lotka and H. T. Odum pointed out the importance of energy in the structure and evolutionary dynamics of ecological and economic systems. And physicists such as Prigogine worked out the far-from-equilibrium thermodynamics of living systems.

The principle of the conservation of mass and energy has formed the basis for a number of important contributions. The assumption was first made explicit in the context of a general equilibrium model by Ayres and Kneese and subsequently by Mäler, but it also is a feature of the series of linear models developed after 1966. All reflect the assumption that a closed physical system must satisfy the conservation of mass condition, and hence that economic growth necessarily increases both the extraction of environmental resources and the volume of waste deposited in the environment.

Perrings developed a variant of the Neumann–Leontief–Sraffa general equilibrium model in the context of a jointly determined economy–environment system subject to a conservation of mass constraint. The model demonstrates that the conservation of mass contradicts the free disposal, free gifts, and noninnovation assumptions of such models. An expanding economy causes continuous disequilibrating change in the environment. Since market prices in an interdependent economy–environment system often do not accurately reflect environmental change, such transformations of the environment often will go unanticipated.

Ayres described some of the important implications of the laws of thermodynamics for the production process, including the limits they place on the substitution of human capital for natural capital and the ability of technical change to offset the depletion or degradation of natural capital. Although they may be substitutes in individual processes in the short run, natural capital and human-made capital ultimately are complements, because both manufactured and human capital require materials and energy for their own production and maintenance. The interpretation of traditional production functions such as the Cobb–Douglas or constant elasticity of substitution (CES) must be modified to avoid the erroneous conclusion that "self-generating technological change" can maintain a constant output with ever-decreasing amounts of energy and materials as long as ever-increasing amounts of human capital are available.

Furthermore, there are irreducible thermodynamic minimum amounts of energy and materials required to produce a unit of output that technical change cannot alter. In sectors that are largely concerned with processing and/or fabricating materials, technical change is subject to diminishing returns as it approaches these thermodynamic minimums. Ruth uses equilibrium and nonequilibrium thermodynamics to describe the materials–energy–information relationship in the biosphere and in economic systems. In addition to illuminating the boundaries for material and energy conversions in economic systems, thermodynamic assessments of material and energy flows, particularly in the case of effluents, can provide information about depletion and degradation that are not reflected in market price.

There is also the effect of the time rate of thermodynamic processes on their efficiency, and, more importantly, their power or rate of doing useful work. H. T. Odum and Pinkerton pointed out that to achieve the thermodynamic minimum energy requirements for a process implied running the process infinitely slowly. This means at a rate of production of useful work (power) of zero. Both ecological and economic systems must do useful work in order to compete and survive, and H. T. Odum and Pinkerton showed that for maximum power production an efficiency significantly worse than the thermodynamic minimum was required.

These biophysical foundations have been incorporated into models of natural resource supply and of the relationship between energy use and economic performance. Cleveland and Kaufmann developed econometric models that explicitly represent and integrate the geologic, economic, and political forces that determine the supply of oil in the United States. Those models are superior in explaining the historical record than those from any single discipline. Larsson et al. also use energy and material flows to demonstrate the dependence of a renewable resource such as commercial shrimp farming on the services generated by marine and agricultural ecosystems.

One important advance generated by this work is the economic importance of energy quality, namely, that a kilocalorie of primary electricity can produce more output than a kilocalorie of oil, a kilocalorie of oil can produce more output than a kilocalorie of coal, and so on. H. T. Odum describes how energy use in ecological and economic hierarchies tends to increase the quality of energy, and that significant amounts of energy are dissipated to produce higher-quality forms that perform critical control and feedback functions which enhance the survival of the system. Cleveland et al. and Kaufmann show that much of the decline in the energy/real GDP ratio in industrial nations is due to the shift from coal to petroleum and primary electricity. Their results show that autonomous energy-saving technical change has had little, if any, effect on the energy/real GDP ratio. Stern finds that accounting for fuel quality produces an unambiguous causal connection between energy use and economic growth in the United States, confirming the unique, critical role that energy plays in the production of material wealth.

The analysis of energy flows has also been used to illuminate the structure of ecosystems. Hannon applied input–output analysis (originally developed to study interdependence in economies) to the analysis of energy flow in ecosystems. This approach quantifies the direct plus indirect energy that connects an ecosystem component to the remainder of the ecosystem. Hannon demonstrates this methodology using energy flow data from the classic study of the Silver Springs (Florida) food web. These approaches hold the possibility of treating ecological and economic systems in the same conceptual framework, one of the primary goals of ecological economics.

Accounting for Natural Capital, Ecological Limits, and Sustainable Scale

Most current economic policies are largely based on the underlying assumption of continuing and unlimited material economic growth. Although this assumption is slowly beginning to change as the full implications of a commitment to sustainability sink in,

it is still deeply embedded in economic thinking as evidenced by the frequent (but mistaken) equation of "sustainable development" with "sustainable growth." The growth assumption allows problems of intergenerational, intragenerational, and interspecies equity and sustainability to be ignored (or at least postponed), since they are seen to be most easily solved by additional material growth. Indeed, most conventional economists define "health" in an economy as a stable and high "rate of growth." Energy and resource depletion, pollution, and other planetary boundaries and limits to growth, according to this view, will be eliminated as they arise by clever development and deployment of new technology. The assumption is that we can completely "decouple" the economy from environmental impacts. This line of thinking often is called "technological optimism."

An opposing line of thought (often called "technological skepticism") assumes that technology will not be able to circumvent fundamental energy, resource, or pollution constraints and that eventually material economic growth will need to stop. It has usually been ecologists or other life scientists that take this point of view (notable exceptions among economists are Boulding and Daly), largely because they study natural systems that invariably do stop growing when they reach fundamental resource constraints. A healthy ecosystem is one that maintains a relatively stable level. Unlimited growth is cancerous, not healthy, under this view.

Technological optimists argue that human systems are fundamentally different from other natural systems because of human intelligence and that history has shown that resource constraints can be circumvented by new ideas. They claim that Malthus' dire predictions about population pressures have not come to pass and the "energy crisis" of the late 1970s is behind us. Technological skeptics, on the other hand, argue that many natural systems also have "intelligence" in that they can evolve new behaviors and organisms (including humans themselves). Humans are therefore a part of nature, not apart from it. Just because we have circumvented local and artificial resource constraints in the past does not mean we can circumvent the fundamental ones that we will eventually face. Malthus' predictions have not come to pass yet for the entire world, the skeptics would argue, but many parts of the world are in a Malthusian trap now, and other parts may well fall into it. This is particularly important because many industrial nations have increased their numbers and standard of living by importing carrying capacity and exporting ecological degradation to other regions.

The debate has gone on for several decades now. It began with Barnett and Morse's *Scarcity and Growth* in 1963, but really got into high gear only with the publication of *The Limits to Growth* by Meadows et al. in 1972 and the Arab oil embargo in 1973. Several thousand studies over the last 45 years have considered aspects of our energy and resource future, and different points of view have waxed and waned. But the bottom line is that there is still considerable uncertainty about the impacts of energy and resource constraints. We have already hit real fossil fuel limits, not so much because of supply constraints but because of their impacts on climate. Will fusion energy or solar energy or conservation or some as yet unthought of energy source step in to save the day and keep economies growing? The technological optimists say "yes" and the technological skeptics say "maybe." Certainly we can "decarbonize" the economy by shifting to solar and wind energy to deal with climate impacts. The price of these sources have dropped dramatically in recent years. They could certainly sustain a steady state economy with vastly improved quality of life and wellbeing, but whether they could sustain a materially growing economy on a finite planet is another question.

The more specific issues of concern all revolve around the question of limits: the ability of technology to circumvent them, and the long-run costs of the technological "cures." Do we adapt to limits with technologies that have potentially large but uncertain future environmental costs or do we limit population and per capita consumption to levels sustainable with technologies which are known to be more environmentally benign? Must we always increase supply or can we also reduce demand? Is there an optimal mix of the two?

Issues of sustainability are ultimately issues about limits. If material economic growth is sustainable indefinitely by technology then all environmental problems can (in theory at least) be fixed technologically. Issues of fairness, equity, and distribution (between subgroups and generations of our species and between our species and others) are also issues of limits. We do not have to worry so much about how an expanding pie is divided, but a constant or shrinking pie presents real problems. Finally, dealing with uncertainty about limits is the fundamental issue. If we are unsure about future limits the prudent course is to assume they exist. One does not run blindly through a dark landscape that may contain crevasses. One assumes they are there and goes gingerly and with eyes wide open, at least until one can see a little better.

Vitousek et al., in an oft-cited paper, calculated the percent of the Earth's Net Primary Production (NPP) which is being appropriated by humans. This was the first attempt to estimate the "scale" or relative size of human economic activity compared to the ecological life-support system. They estimated that 25% of total NPP (including the oceans) and 40% of terrestrial NPP was currently being appropriated by humans. It left open the question of how much of NPP could be appropriated by humans without damaging the life-support functions of the biosphere, but it is clear that 100% is not sustainable and even the 40% of terrestrial NPP currently used may not be sustainable.

A related idea is that ecosystems represent a form of capital—defined as a stock yielding a flow of services—and that this stock of "natural capital" needs to be maintained intact independently in order to assure ecological sustainability. The question of whether natural capital needs to be maintained independently ("strong sustainability") or whether only the total of all capital stocks need to be maintained ("weak sustainability") has been the subject of some debate. It hinges on the degree to which human-made capital can substitute for natural capital, and, indeed, on how one defines capital generally. In general, conventional economists have argued that there is almost perfect substitutability between natural and human-made capital, while ecological economists generally argue on both theoretical and empirical grounds that the possibilities for substitution are severely limited. They therefore generally favor the strong sustainability position.

Another critical set of issues revolve around the way we define economic income, economic welfare, and total human welfare or wellbeing. Daly and Cobb clearly distinguish these concepts, and point out that conventional GDP is a poor measure even of

economic income. Yet GDP continues to be used in most policy discussions as the definitive measure of economic health and performance, and will continue to be until viable alternatives are available. According to Hicks, economic income is defined as the quantity we can consume without damaging our future consumption possibilities. This definition of income automatically embodies the idea of sustainability. GDP is a poor measure of income on a number of grounds, including the fact that it fails to account for the depletion of natural capital and thus is not "sustainable" income in the Hickian sense. GDP is an even poorer measure of economic welfare, since many components of welfare are not directly related to income and consumption. The Index of Sustainable Economic Welfare (ISEW) devised by Daly and Cobb (and its renamed, but almost identical successor, the Genuine Progress Indicator (GPI)) is one approach to estimating economic welfare (as distinct from income). GPI starts with Personal Consumption Expenditures (a major component of GDP) but adjusts them using 24 different components, including income distribution, environmental costs, and negative activities like crime and pollution, among others. GPI also adds positive components left out of GDP, including the benefits of volunteering and household work. By separating activities that diminish welfare from those that enhance it, GPI better approximates sustainable economic welfare. However GPI is not the perfect indicator of societal wellbeing and needs to be viewed alongside biophysical and other indicators.

Past national GPI studies have indicated that in many countries, beyond a certain point, GDP growth no longer correlates with increased economic welfare. An important function of GPI is to send up a red flag at that point. Since it is made up of many benefit and cost components, it also allows for the identification of which factors increase or decrease economic welfare. Other indicators are better guides of specific aspects.

For example, Life Satisfaction is a better measure of overall self-reported wellbeing. By observing the change in individual benefit and cost components, GPI reveals which factors cause economic welfare to rise or fall even if it does not always indicate what the driving forces are behind this. It can account for the underlying patterns of resource consumption, for example, but may not pick up the self-reinforcing evolution of markets or political power that drives change.

Valuation of Ecological Services

All decisions concerning the allocation of environmental resources imply the valuation of those resources. Ecological economics does not eschew valuation. It is recognized that the decisions we make, as a society, about ecosystems imply a valuation of those systems. We can choose to make these valuations explicit or not; we can undertake them using the best available ecological science and understanding or not; we can do them with an explicit acknowledgment of the huge uncertainties involved or not; but as long as we are forced to make choices about the use of resources we are valuing those resources. These values will reflect differences in the underlying worldview and culture of which we are a part, just as they will reflect differences in preferences, technology, assets, and income. An ecological economics approach to valuation implies an assessment of the spatial and temporal dynamics of ecosystem services, and their role in satisfying both individual and social preferences and their broader contribution to wellbeing that may not be perceived by most people. It also implies explicit treatment of the uncertainties associated with tracking these dynamics.

Ecological economics is different from environmental economics in terms of the latitude of approaches to the ecosystem valuation problem it allows. They include more conventional Willingness To Pay (WTP)-based approaches, but they also include other more novel methods, including deliberative and participatory approaches and approaches based on explicitly modeling the linkages between ecosystems and economic systems.

Preference formation is influenced by limited information in estimating WTP-based values for biodiversity preservation. Empirical studies show that a significant portion of individuals exhibit "lexicographic" preferences, that is, they refuse to make tradeoffs that require the substitution of biodiversity for other goods. This places significant constraints on the use of stated preferences, as used in contingent valuation studies, for valuation of ecosystem services and decision-making. It places more emphasis on the need to develop more direct methods to assess the value of these resources as a supplement to conventional WTP-based methods.

The valuation of ecological resources requires a deeper understanding of the ways in which economic activity depends on biogeophysical processes than is usually recognized. However, the issue of ecosystem valuation is far from solved. In fact, it is probably only in the early stages of development. Conventional WTP-based approaches have severe limitations. Key directions for the future include integrated ecological economic modeling, as elaborated in the next section.

Integrated Ecological Economic Modeling and Assessment

The emphasis on both (1) issues of scale and limits to the carrying and assimilative capacity of ecological systems and (2) underlying dynamics of those systems imply the need for a new approach to the modeling of joint systems. It is not surprising that this is an active area of research in ecological economics. Indeed, combining (sometimes implicit) models of ecological processes and economic decision models in a new that makes the feedbacks between the two sets of processes transparent is where we most expect new advances to be made as a result of the ongoing dialog between economists and ecologists.

Ecology and economics have long diverged both methodologically and conceptually. One reason for the difficulty in bridging the modeling gap is that economics, as a discipline, has developed almost no tools or concepts to handle spatial differentiation beyond the notions of transport cost and international trade. The spatial analysis of human activity has been seen as the domain of geographers, and has had remarkably little impact on the way economists have analyzed the allocation

of resources. This makes collaboration between economists and disciplines based more directly on spatial analysis difficult. Since the development of spatially explicit integrated models is one of the areas in which ecological economics is expected to develop most rapidly, it would seem that geographers are likely to become an increasingly important part of the research agenda in ecological economics.

A second characteristic of ecological–economic models concerns the way in which the valuation of ecological functions and processes is reflected in the model structure. The point was made in the previous section that valuation by stated preference methods (estimation of willingness to pay or accept using contingent valuation or contingent ranking) may capture the strength of people's perceptions and their level of income and endowments (their ability to pay), but it generally fails to capture the impact of a change in ecosystem functions and processes on the output of economically valued goods and services. Unless the role of non-marketed ecological functions and processes in the production of economically valued goods and services is explicitly modeled, it is hard to see how they can be properly accounted for in economic decision making.

A third characteristic concerns the role of integrated modeling in strategic decision making. One of the challenges to ecological economics has been to devise methods to address strategic "what if" questions in a way that reflects the dynamics of the jointly determined system. The general problem confronting anyone attempting to model long-run dynamics explicitly is that ecological–economic systems are complex nonlinear systems. The dynamics of economic systems are not independent of the dynamics of the ecological systems which constitute their environment, and that as economies grow relative to their environment, the dynamics of the jointly determined system can become increasingly complex. Indeed, the development of ecological economics can be thought of as part of a widespread reappraisal of such systems.

In ecology, this reappraisal has influenced recent research on scale, complexity, stability, and resilience, and is beginning to influence the theoretical treatment of the coevolution of species and systems. The results that are most important to the development of ecological economics concern the link between the spatial and temporal structure of coevolutionary hierarchical systems. Landscapes are conceptualized as hierarchies, each level of which involves a specific temporal and spatial scale. The dynamics of each level of the structure are predictable so long as the biotic potential of the level is consistent with bounds imposed by the remaining levels in the hierarchy. Change in either the structure of environmental constraints or the biotic potential of the level may induce threshold effects that lead to complete alteration in the state of the system.

In economics there is now considerable interest in the dynamics of complex nonlinear systems. Economists have paid less attention to spatial scale and its significance at or near system thresholds, but there is now a growing body of literature with roots in geography which seeks to inject a spatial dimension into nonlinear economic models. There is also an economic analog to the biologist's interest in evolution and the significance of co-dependence between gene landscapes. The steady accumulation of evidence that economic development is not a stationary process, that human understanding, preferences, and technology all change with development, and that such change is generally nonlinear and discontinuous, has prompted economists to seek to endogenize technological change. Although the adaptation of this work by environmental economists has been rather disappointing, the treatment of technology and consumption preferences as endogenous to the economic process is a fundamental change that brings economics much closer to ecology.

The challenge to ecological economics in the future is to develop models that capture these features well enough to incorporate at least the major risks in economic decisions that increase the level of stress on ecological systems.

Summary and Conclusions

This is a sample of the range of transdisciplinary thinking that can be put under the heading of ecological economics. While it is difficult to categorize ecological economics in the same way one would a normal academic discipline, some general characteristics can be enumerated.

- The core problem is the "sustainability" of interactions between economic and ecological systems.
- An explicit attempt is made at "pluralistic dialog" and integration across disciplines, rather than territorial disciplinary differentiation.
- An emphasis is placed on "integration" of the three hierarchical goals of sustainable scale, fair distribution, and efficient allocation.
- There is a deep concern with the "biophysical underpinnings" of the functioning of jointly determined ecological and economic systems.
- There is a deep concern with the relationship between the "scale" of economic activity and the nature of change in ecological systems.
- Since valuation based on stated willingness to pay reflects limitations in the valuer's knowledge of ecosystems functions, there is an emphasis on the "development of valuation techniques" that build on an understanding of the role of ecosystem functions in economic production and wellbeing.
- There is a broad focus on systems and "systems dynamics, scale, and hierarchy" and on "integrated modeling" of ecological economic systems.

These characteristics make ecological economics applicable to some of the major problems facing humanity today, which occur at the interfaces of human and natural systems, and especially to the problem of improving humanity's wellbeing and assuring its

survival within the biosphere into the indefinite future. It is not so much the individual core scientific questions that set ecological economics apart—since these questions are covered independently in other disciplines as well—but rather the treatment of these questions in an integrated, transdisciplinary way, which is essential to their understanding and effective use in policy.

See also: Human Ecology and Sustainability: Resilience; Ecosystem Services Evaluation; Ecological Systems Thinking

Further Reading

Barbier, E.B., Burgess, J.C., Folke, C., 1994. Paradise lost? The Ecological Economics of Biodiversity. London: Earthscan, p. 267.
Boulding, K.E., 1966. The economics of the coming spaceship earth. In: Jarrett, H. (Ed.), Environmental quality in a growing economy. Baltimore, MD: Resources for the Future/Johns Hopkins University Press, pp. 3–14.
Boumans, R., Costanza, R., Farley, J., *et al.*, 2002. Modeling the dynamics of the integrated earth system and the value of global ecosystem services using the GUMBO model. Ecological Economics 41, 529–560.
Costanza, R., 1991. Ecological economics: The science and management of sustainability. New York: Columbia University Press.
R. Costanza. Visions of alternative (unpredictable) futures and their use in policy analysis Conservation Ecology 4 1 2000 5, http://www.consecol.org/vol4/iss1/art5/ (accessed October 2007).
Costanza, R., 2001. Visions, values, valuation and the need for an ecological economics. Bioscience 51 (2001), 459–468.
Costanza, R., Wainger, L., Folke, C., Mäler, K.-G., 1993. Modeling complex ecological economic systems: Toward an evolutionary, dynamic understanding of people and nature. Bioscience 43, 545–555.
Costanza, R., Voinov, A., Boumans, R., *et al.*, 2002. Integrated ecological economic modeling of the Patuxent River watershed, Maryland. Ecological Monographs 72, 203–231.
Costanza, R., Cumberland, J.C., Daly, H.E., Goodland, R., Norgaard, R., Kubiszewski, I., Franco, C., 2014. An introduction to ecological economics, second edn. Boca Raton: Taylor and Francis.
Costanza, R., de Groot, R., Braat, L., Kubiszewski, I., Fioramonti, L., Sutton, P., Farber, S., Grasso, M., 2017. Twenty years of ecosystem services: How far have we come and how far do we still need to go? Ecosystem Services 28 (2017), 1–16.
Daly, H.E., 1968. On economics as a life science. Journal of Political Economy 76, 392–406.
J. Farley, R. Costanza. Envisioning shared goals for humanity: A detailed shared vision of a sustainable and desirable USA in 2100 Ecological Economics 43 2002 245–259.
Jansson, A.M., Hammer, M., Folke, C., Costanza, R., 1994. Investing in natural capital: The ecological economics approach to sustainability. Washington, DC: Island Press, p. 504.
Krishnan, R., Harris, J.M., Goodwin, N., 1995. A survey of ecological economics. Washington, DC: Island Press.
Martinez-Alier, J., 1987. Ecological economics: Energy, environment, and society. Oxford: Blackwell, p. 287.
Norton, B., Costanza, R., Bishop, R., 1998. The evolution of preferences: Why "sovereign" preferences may not lead to sustainable policies and what to do about it. Ecological Economics 24, 193–211.
Ward, J., Sutton, P., Werner, A., Costanza, R., Mohr, S., Simmons, C., 2016. Is decoupling GDP growth from environmental impact possible? PLoS One 11 (10), e0164733

Ecological Economics 2☆

Robert Costanza, Australian National University, Canberra, ACT, Australia

Introduction

Stories about the economy typically focus on gross domestic product (GDP), jobs, stock prices, interest rates, retail sales, consumer confidence, housing starts, taxes, and assorted other indicators. We hear things like "GDP grew at a 3% rate in the fourth quarter, indicating a recovering, healthy economy, but with room for further improvement." Or, "The Fed raised short-term interest rates again to head off inflation."

But do these reports, and the indicators they cite, really tell us how the economy is doing? What is the economy anyway? And what is this economy for?

Conventional reports on these questions are rather narrow. The "economy" we usually hear about refers only to the market economy—the value of those goods and services that are exchanged for money. Its purpose is usually taken to be to maximize the value of these goods and services—with the assumption that the more the activity, the better off we are. Thus, the more the GDP (which measures aggregate activity in the market economy), the better. Likewise the more contributors to GDP (such as retail sales and salaries paid to employees), the better. Predictors of more GDP in the future (such as housing starts and consumer confidence) are also important pieces of information from this perspective. Declining or even stable GDP is seen as a disaster. Growth in GDP is assumed to be a government's primary policy goal and also something that is sustainable indefinitely.

But is this what the economy is all about? Or more accurately, is this "all" that the economy is about? Or, is this what the economy "should be" about? The answer to all of these is an emphatic "no." Here's why.

Let's start with purpose. The purpose of the economy "should be" to provide for the sustainable well-being of people. That goal encompasses material well-being certainly—but also anything else that affects well-being and its sustainability. This seems obvious and noncontroversial. The problem comes in determining what things actually affect well-being and in what ways.

There is substantial new research on this "science of happiness" that shows the limits of conventional economic income and consumption in contributing to well-being. Psychologist Tim Kasser in his 2003 book *The High Price of Materialism* points out, for instance, that people who focus on material consumption as a path to happiness are actually less happy and even suffer higher rates of both physical and mental illnesses than those who do not. Material consumption beyond real need is a form of psychological "junk food" that only satisfies for the moment and ultimately leads to depression, Kasser says.

Economist Richard Easterlin, a noted researcher on the determinants of happiness, has shown that well-being tends to correlate well with health, level of education, and marital status, and with income only up to a fairly low threshold (**Fig. 1**). He concludes in a recent paper in the *Proceedings of the National Academy of Sciences* that,

> People make decisions assuming that more income, comfort, and positional goods will make them happier, failing to recognize that hedonic adaptation and social comparison will come into play, raise their aspirations to about the same extent as their actual gains, and leave them feeling no happier than before. As a result, most individuals spend a disproportionate amount of their lives working to make money, and sacrifice family life and health, domains in which aspirations remain fairly constant as actual circumstances change, and where the attainment of one's goals has a more lasting impact on happiness. Hence, a reallocation of time in favor of family life and health would, on average, increase individual happiness.

British economist Richard Layard's 2005 book *Happiness: Lessons from a New Science* echos many of these ideas and concludes that current economic policies are not improving happiness and that "happiness should become the goal of policy, and the progress of national happiness should be measured and analyzed as closely as the growth of GNP." Several countries are now interested in alternative measures of progress. For example, the country of Bhutan has recently announced that it will make "gross national happiness" its explicit policy goal.

Economist Robert Frank, in his 2000 book *Luxury Fever*, also concludes that the nation would be better off—overall national well-being would be higher, that is—if we actually consumed less and spent more time with family and friends, working for our communities, maintaining our physical and mental health, and enjoying nature.

On this last point, there is substantial and growing evidence that natural systems contribute heavily to human well-being (**Fig. 2**). In a paper published in 1997 in the journal *Nature*, the author with his co-workers estimated that the annual, nonmarket value of the Earth's ecosystem services is $33 trillion globally, substantially larger than global GDP. The just released UN Millennium Ecosystem Assessment is a global update and compendium of ecosystem services and their contributions to human well-being.

So, if we want to assess the "real" economy—all the things which contribute to real, sustainable, human welfare and quality of life—as opposed to only the "market" economy, we have to measure the nonmarketed contributions to human well-being from

☆*Change History:* February 2018. R. Costanza made minor changes to the text, figures and references.

This is an update of R. Costanza, Ecological Economics 2, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 1006–1011.

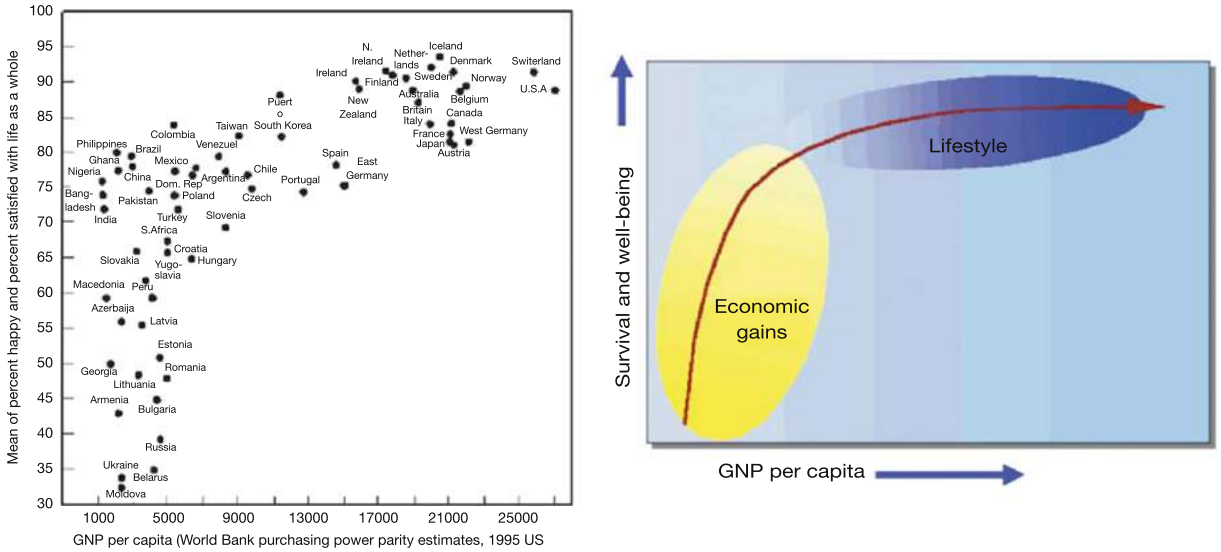


Fig. 1 Relationship between GNP per capita and life satisfaction. Source: World Development Report.

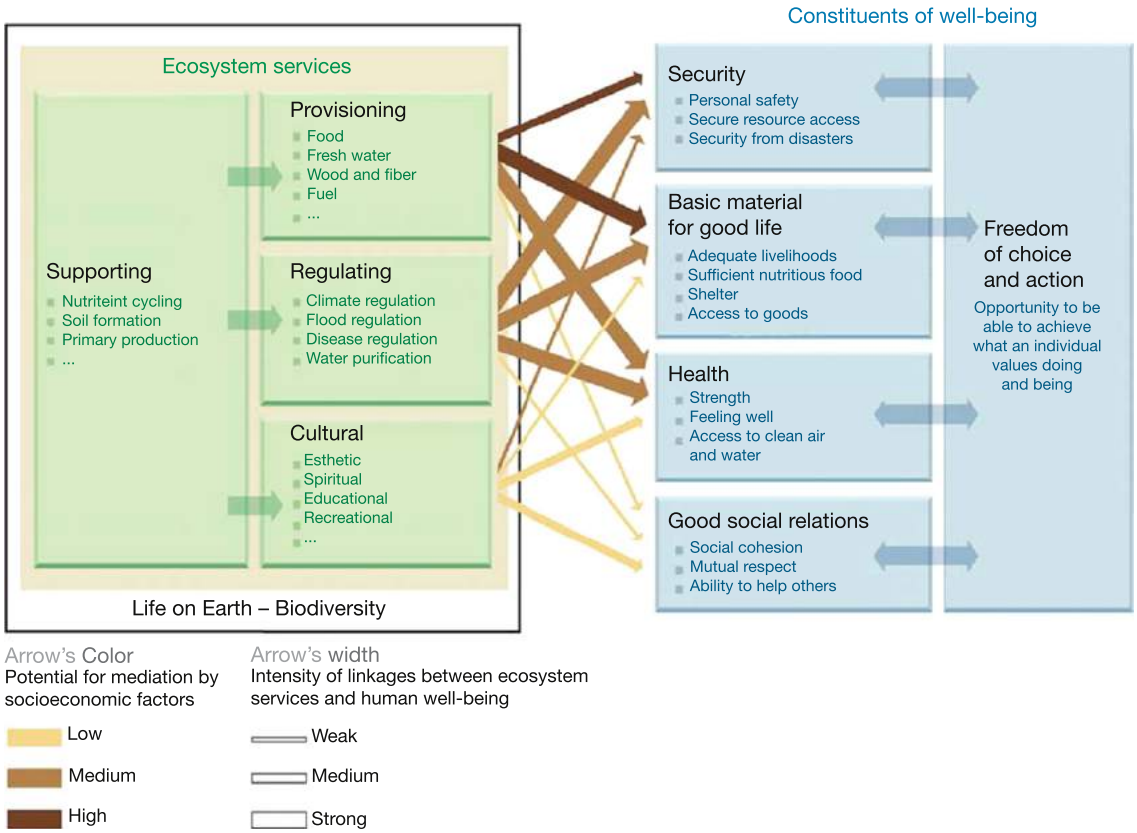


Fig. 2 Consequences of ecosystem change for human well-being. Source: Millennium Ecosystem Assessment.

nature, from family, friends, and other social relationships at many scales, and from health and education. One convenient way to summarize these contributions is to group them into four basic types of capital that are necessary to support the real, human-welfare-producing economy: built capital, human capital, social capital, and natural capital (**Fig. 3**).

The market economy covers mainly built capital (factories, offices, and other built infrastructure and their products) and part of human capital (spending on labor), with some limited spillover into the other two types. Human capital includes the health, knowledge, and all the other attributes of individual humans that allow them to function in a complex society. Social capital includes all the formal and informal networks among people: family, friends, and neighbors, as well as social institutions at all

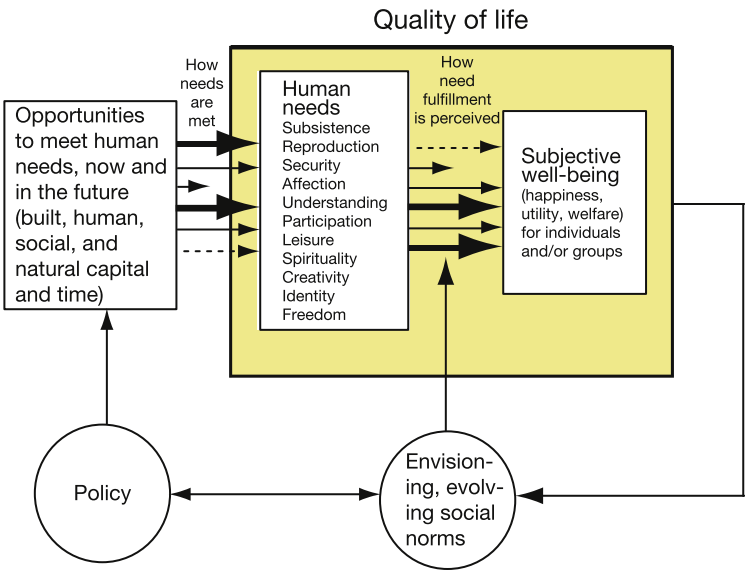


Fig. 3 Quality of life as the interaction of human needs and the subjective perception of their fulfillment, as mediated by the opportunities available to meet the needs. From Costanza, R. J., Fisher, S., Ali, C. et al. (2006) Quality of life: An approach integrating opportunities, human needs, and subjective well-being. *Ecological Economics* **61**, 267–276.

levels, like churches, social clubs, local, state, and national governments, NGOs, international organizations, and the institutions of the market itself. Natural capital includes the world's ecosystems and all the services they provide that support human well-being. Ecosystem services occur at many scales, from climate regulation at the global scale, to flood protection, soil formation, nutrient cycling, recreation, and esthetic services at the local and regional scales.

So, how have the world's real economies been doing recently, compared to their market economies? The short answer is, not so good. How do we know? One way is through surveys of people's life satisfaction, which in the United States has been decreasing slightly since about 1975. A second approach is an aggregate measure of the real economy that has been developed as an alternative to GDP called the genuine progress indicator, or GPI.

Let's first take a quick look at the problems with GDP as a measure of true human well-being. GDP is not only limited—measuring only marketed economic activity or gross income—it also counts all of this activity as positive. It does not separate desirable, well-being-enhancing activity from undesirable well-being-reducing activity. For example, an oil spill increases GDP because someone has to clean it up, but it obviously detracts from society's well-being. From the perspective of GDP, more crime, more sickness, more war, more pollution, more fires, storms, and pestilence are all potentially good things, because they can increase marketed activity in the economy.

GDP also leaves out many things that do enhance well-being but are outside the market. For example, the unpaid work of parents caring for their own children at home does not show up, but if these same parents decide to work outside the home to pay for child care, GDP suddenly increases. The nonmarketed work of natural capital in providing clean air and water, food, natural resources, and other ecosystem services does not adequately show up in GDP, either, but if those services are damaged and we have to pay to fix or replace them, then GDP suddenly increases. Finally, GDP takes no account of the distribution of income among individuals. But it is well-known that an additional $1 worth of income produces more well-being if one is poor rather than rich. It is also clear that a highly skewed income distribution has negative effects on a society's social capital.

The GPI addresses these problems by separating the positive from the negative components of marketed economic activity, adding in estimates of the value of nonmarketed goods and services provided by natural, human, and social capital, and adjusting for income-distribution effects (**Fig. 4** lists the components of the GPI). While it is by no means a perfect representation of the real well-being of a nation, GPI is a much better approximation than GDP. As Amartya Sen and others have noted, it is much better to be approximately right in these measures than precisely wrong.

Comparing GDP and GPI for several countries shows that in many "developed" countries the benefits of growth in the market economy are now being outweighed by the uncounted costs of that growth. For example, **Fig. 5** shows that globally, while GDP has steadily increased since 1950, with the occasional dip or recession, GPI peaked in about 1978 and has been gradually decreasing ever since. From the perspective of the real economy, as opposed to just the market economy, the world has been in recession since 1978. As already mentioned, this picture is also consistent with survey-based research on people's stated life-satisfaction. We are now in a period of what Herman Daly has called "un-economic growth," where further growth in marketed economic activity (GDP) is actually reducing well-being on balance rather than enhancing it. In terms of the four capitals, while built capital has grown, human, social, and natural capital have declined or remained constant and more than canceled out the gains in built capital.

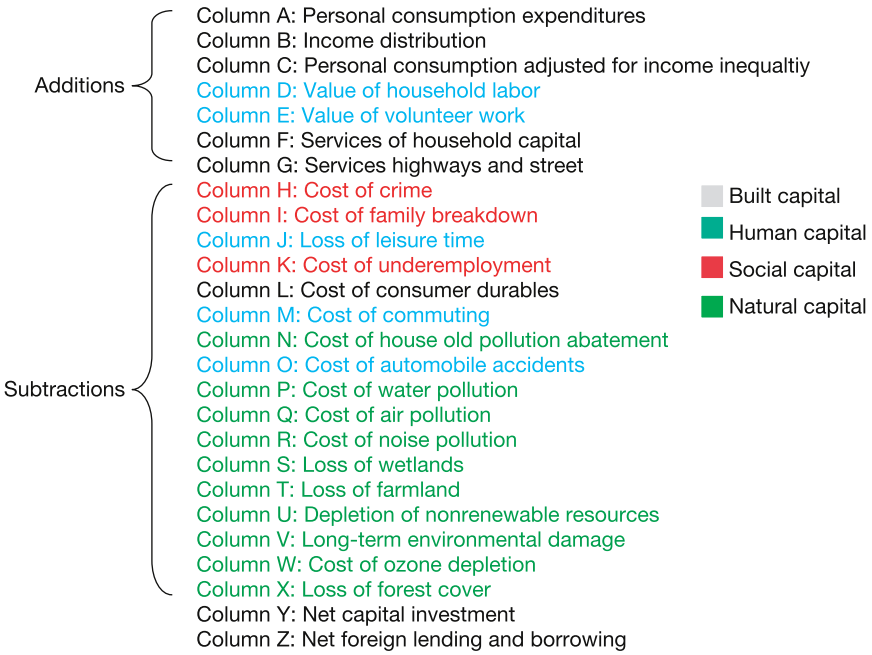


Fig. 4 The genuine progress indicator (GPI) by column.

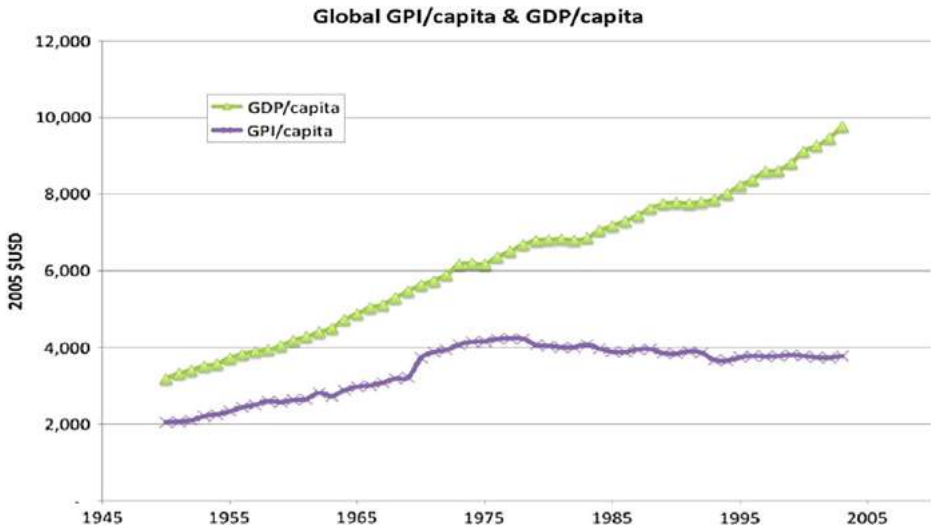


Fig. 5 Global GDP/capita vs. GPI/capita, 1950–2014. Source: Kubiszewski, I., Costanza, R., Franco, C., Lawn, P., Talberth, J., Jackson, T. and Aylmer, C. (2013). Beyond GDP: Measuring and achieving global genuine progress. *Ecological Economics*. **93**, 57–68.

From the perspective of the real economy, things are not improving.

Is the news all bad? No. We estimated the GPI of the State of Vermont and of Burlington, the state's largest city, and found that Vermont's and Burlington's GPI per capita had increased over the entire 1950–2000 period and is now more than double the national average. This was due to Vermont's attention to protecting and enhancing natural, human, and social capital in balance with gains in built capital—accomplished through the application of strong, local democratic principles, and processes still actively at work in Vermont. Recent GPI studies of all US states by M-V Fox and J. Erickson show that the range of variation is quite large.

The lesson is that there is significant variation within and across states and countries in trends in well-being and quality of life, and plenty of good examples we can learn from to improve overall well-being at multiple scales.

How can we apply these lessons to get out of the real recession in human well-being at the national scale that many countries are now in? Several policies have been suggested that would help to turn things around:

- Shifting our primary national policy goal from increasing marketed economic activity (GDP) to maximizing national well-being (GPI or something similar). This would allow us to see the interconnections between built, human, social, and natural capital and build well-being in a balanced and sustainable way.
- Reforming tax systems to send the right incentives by taxing negatives (pollution, depletion of natural capital, over-consumption) rather than positives (labor, savings, investment).
- Reforming international trade to promote well-being over mere GDP growth. This implies protecting natural capital, labor rights, and democratic self-determination first and then allowing trade, rather than promoting the current trade rules that ignore all nonmarket contributions to well-being.
- Implementing strong democracy, as Tom Prugh, Robert Costanza, and Herman Daly have argued in the book *The Local Politics of Global Sustainability*. Strong democracy implies true participation of all in governance and is an essential prerequisite to building a sustainable and desirable future.
- Increasing the size of the "common sector" of the economy (as opposed to the private and public sectors) but creating common property asset trusts to "propertize" natural and social capital assets, as described in Peter Barnes' book *Capitalism 3.0*.

Ultimately, getting out of the recession in well-being we are currently in will require us to look beyond the limited definition of the "economy" we read about in the newspapers, and recognize what the real economy is and what it is for. We must not allow deceptive accounting practices to paint an inaccurate and ultimately destructive picture of how "well" we are doing. Alternatives are available, but they need significant further discussion and research.

With nothing less than our current and future well-being at stake, we can certainly afford to devote greater effort to learning how to adequately understand and measure it. If we want things that really matter to our well-being to count, we must learn how to recognize and count them, use that information to inform policy in a real democracy, and create adaptive institutions that can effectively implement the policy.

See also: Human Ecology and Sustainability: Resilience; Ecosystem Services Evaluation; Ecological Systems Thinking

Further Reading

Barnes, P., 2006. Capitalism 3.0: A Guide to Reclaiming the Commons. San Francisco, CA: Berrett-Koehler.
Costanza, R., d'Arge, R., de Groot, R., *et al.*, 1997. The value of the world's ecosystem services and natural capital. Nature 387, 253–260.
Costanza, R.J., Erickson, K., Fligger, A., *et al.*, 2004. Estimates of the genuine progress Indicator (GPI) for Vermont, Chittenden County, an Burlington, from 1950 to 2000. Ecological Economics 51, 139–155.
Costanza, R.J., Fisher, S., Ali, C., *et al.*, 2006. Quality of life: An approach integrating opportunities, human needs, and subjective well-being. Ecological Economics 61, 267–276.
Daly, H.E., Farley, J., 2003. Ecological economics: Principles and applications. Washington, DC: Island Press.
Easterlin, R.A., 2003. Explaining happiness. Proceedings of the National Academy of Sciences 100 (19), 11176–11183.
Fox, M.-J.V., Erickson, J.D., 2018. Genuine economic progress in the United States: A fifty state study and comparative assessment. Ecological Economics 147 (2018), 29–35.
Frank, R., 1999. Luxury fever: Why money fails to satisfy in an era of excess. Mankato, MN: The Free Press.
Kahneman, D., Diener, E., Schwarz, N., 1999. Well-Being: The Foundations of Hedonic Psychology. New York: Russell Sage Foundation.
Kasser, T., 2003. The high price of materialism. Cambridge, MA: MIT Press.
Kubiszewski, I., Costanza, R., Franco, C., Lawn, P., Talberth, J., Jackson, T., Aylmer, C., 2013. Beyond GDP: Measuring and achieving global genuine progress. Ecological Economics 93, 57–68.
Layard, R., 2005. Happiness: Lessons from a new science. New York: Penguin.
Prugh, T., Costanza, R., Daly, H., 2000. The local politics of global sustainability. Washington, DC: Island Press, 173 pp.
World Health Organisation, 2006. Millennium ecosystem assessment: Synthesis. Washington, DC: Island Press.

Ecological Footprint☆

Mathis Wackernagel, David Lin, Laurel Hanscom, Alessandro Galli, and Katsunori Iha, Global Footprint Network, Oakland, CA, United States

Introduction

Ecosystems provide many critical services to human society, including both the direct provision of goods, such as food and fiber products, as well as less visible services, such as water filtration and climate stabilization. The availability of these ecosystem services depends closely on the functioning of biological capital. Broadly defined, biological capital contains all of the ecosystems and various components of the biosphere that directly or indirectly provide goods and services. Careful management of this capital is central to maintaining not only the health of the natural environment but also human well-being into the future.

Managing biological capital, however, requires tools that are able to track its availability and use. The Ecological Footprint is an accounting tool that calculates human demand on the biosphere, and compares this to the planet's ability to meet these demands. It does this by answering the specific research question, "how much of the Earth's regenerative capacity is occupied by human activities?"

The accounting builds on a simple thermodynamic principle. Since life competes for biologically productive space, every demand that competes for such areas can be added up. Therefore, demand on regenerative capacity can be approximated by adding up these mutually exclusive biologically productive areas occupied for providing the demanded services (the "Ecological Footprint"). This demanded space then can be compared to the space available (the "biocapacity"). Since not every unit of space has equal productivity, both Ecological Footprint and biocapacity are expressed in a productivity adjusted area unit, the global hectare. This is a hectare of biologically productive surface area of the planet with world-average productivity.

Footprint analysis helps governments, businesses, and individuals track economies' dependence on biological capital over time. Similar to financial balance sheets, the resulting ecological accounts can be used as a quantitative input into decision making at all levels.

Metabolism: An Ecological Model of Society

The Ecological Footprint applies principles of ecology and thermodynamics to human society to create a framework for mapping society's material and energy metabolism. As in any ecosystem, primary producers fix energy from sunlight through photosynthesis. This energy is then available for consumers, who use this primary production for powering their activities. All material ingested by consumers eventually returns to the biosphere as waste products. The biological waste products are broken down and recycled back into the raw materials for primary production.

This thermodynamic reality of every input eventually turning into waste, applies to any organism that exists in a natural environment, including human society. The human economy takes high-quality matter and energy as inputs from its environment and returns these in degraded form as material waste and heat. Societies consume resources in order to maintain themselves. Between resource intake and waste discharge, matter accumulates in these systems, leading to increased body mass in the case of animals, or an accumulation of material stocks in societies.

From a human perspective, the ability of the biosphere to absorb wastes and regenerate resources is known as the regenerative capacity of the planet. Although the Earth's regenerative capacity is robust, it can be eroded in a number of ways. The Natural Step framework puts them into three main categories (Robèrt *et al.*, 2002):

1. Harvesting renewably generated resources, such as trees or fish, faster than they can be replenished can overwhelm natural cycles. Direct physical interference, such as the paving over of green surfaces or farming practices that cause soil erosion, can also damage the underlying capital (i.e., loss of soil nutrients or ground water depletion) that creates these resources, further reducing the Earth's total regenerative capacity.
2. Substances produced by society that are persistent and do not readily break down, and which ecosystems have not developed abilities to assimilate, can compromise regenerative capacity as they accumulate in the biosphere. Examples include synthetic chemicals such as DDT and PCBs that persist and bio-accumulate through tropic levels. Many of these man-made substances have no natural analogues, and the biosphere as a whole has not evolved efficient means to break down and reassimilate these products on human time scales.
3. Substances normally buried deep within the Earth's crust can be extracted, refined, and introduced into the biosphere at quantities that ecosystems or the atmosphere are not able to assimilate. Examples include heavy metals, radioactive elements,

☆*Change History:* March 2018. M Wackernagel, D Lin, L Hanscom, A Galli, and K Iha updated all figures and tables.

This is an update of M. Wackernagel, E. Lazarus, D. Lin, K. Iha and J.A. Kitzes, Ecological Footprint, In Reference Module in Earth Systems and Environmental Sciences, Elsevier, 2015.

 doi:10.1016/B978-0-12-409548-9.09567-1

minerals, and mined carbon. Because there are few natural cycles that can return these substances to the crust within human time spans, these substances systematically accumulate in the biosphere and atmosphere.

The fundamental insight emerging from looking at the human economy from the perspective of its metabolism is that, in the long-term, the biosphere must be able to turn wastes back into resources faster than human society turns the resources into waste. If the extraction of resources becomes too large, or if nature's regenerative capacity is compromised, biological capital will be systematically degraded and wastes will accumulate.

While in the past, most "environmental" problems arose from poor management of some local aspects of society's metabolism, such as careless waste disposal, smokestacks, or overuse of a river basin, now the very size of society's global metabolism that has become the overarching concern. While local overuse of the biosphere has a long history (e.g., overfishing, deforestation, soil erosion, etc.), the global human economy has now become so large, relative to the regenerative capacity of planet Earth, that it is now, for the first time in human history, confronting global limits.

Fundamental Assumptions

In order to provide a quantitative answer to the research question of how much regenerative capacity is required to maintain a given resource flow through human society, Ecological Footprint analysis uses a methodology grounded on six basic assumptions:

1. *The annual amounts of resources consumed and wastes generated by countries are tracked by national and international organizations.* Most countries have extensive annual statistics documenting their resource use, particularly in the areas of energy, forest products and agricultural products. United Nations agencies, like the Food and Agriculture Organization (FAO), compile many of these national statistics in a consistent format.
2. *The quantity of biological resources appropriated for human use is directly related to the amount of bioproductive land area necessary for regeneration, accommodation of infrastructure, or the assimilation of waste.* Bioproductive processes are associated with surfaces that capture sunlight for photosynthesis. Even three-dimensional processes that represent layers of such surfaces, as in aquatic ecosystems or rainforests, can be mapped on the two-dimensional area on which they occur.
3. *By weighting each area in proportion to its usable biomass productivity (i.e., its potential annual production of usable biomass), the different areas can be expressed in terms of a standardized average productive hectare*. For this reason, Ecological Footprint accounting uses as its measurement the standardized global hectare with world-average productivity. (In other words, each of the global hectares represents an equal share of the planet's total regenerative capacity).
4. *The overall demand in global hectares can be aggregated by adding all mutually exclusive resource-providing and waste-assimilating areas required to support the demand*.
5. *Aggregate human demand (Ecological Footprint) and nature's supply (biocapacity) can be directly compared to each other*. By using the standardized unit of a global hectare, demand and supply can be compared, as can different components of demand and supply.
6. *Area demanded can exceed area supplied.* A Footprint greater than available biocapacity at any given scale indicates that demand exceeds the regenerative capacity of existing biological capital. This condition, known as "overshoot," is physically possible only in the short-term. For some limited time, resources can be harvested faster than they regenerate (e.g., deforestation) and wastes can accumulate (e.g., carbon dioxide in the atmosphere). In the long-term, however, such overshoot inevitably implies ecological degradation. This will reduce the amount that the overused ecosystems can deliver ever year. Overshoot may also lead to more abrupt change: demand can cross ecological thresholds that trigger nonlinear, large-scale, abrupt environmental change or even collapse (Rockstrom *et al.*, 2009).

Note the similarity and complementarity to the approach named "Planetary Boundaries" (Rockstrom *et al.*, 2009; Steffen *et al.*, 2015). Ecological Footprint accounting summarizes human demand against the planet's level of regeneration. In contrast, Planetary Boundaries investigates nine dimensions of human uses of the biosphere. By mapping overall demand on biocapacity only, Ecological Footprint accounting reveals overall limitations of the biosphere and implies trade-offs between the Planetary Boundaries dimensions.

Planetary Boundaries (Rockstrom *et al.*, 2009; Steffen *et al.*, 2015) identify key physical quantitative conditions that are needed to maintain the integrity of the biosphere. Nine areas have been identified, in which transgressions could lead to shifts in the biosphere's integrity, potentially irreversibly moving the biosphere out of the stable conditions which characterized the Holocene.

Therefore, one could interpret the Planetary Boundaries as the inverse of the ingredients for healthy, productive ecosystems that can maintain their integrity. In other words, staying within the "safe operating zone" for each of the nine ingredients enables the biosphere's productivity, which Ecological Footprint accounting calls the biosphere's biocapacity. These boundaries translate into the following specific ingredients, similar to what economists would call "the production factors" enabling biocapacity:

- Stable climate
- Healthy biodiversity
- Sufficient (yet not excessive amounts of) nutrients
- Protective ozone layer

- Absence of pollutants
- Clean and sufficient fresh water
- Absence of acidification in both water and soils

The Ecological Footprint does not quantify, in contrast to the planetary boundaries' trigger points. Rather it tracks human demand against biological regeneration. If demand exceeds regeneration, the resulting natural capital loss or degradation will eventually trigger irreversible change since one cannot indefinitely take more from ecosystems than they can renew. Ecological Footprint accounts summarize the overall outcome of transgressing planetary boundaries and quantify this transgression in one "biocapacity" number representing the regenerative capacity of the biosphere.

There is strong complementarity between planetary boundaries and Ecological Footprint. Where the planetary boundaries span a broad field of academic inquiry, Footprint assessments can be categorized as a simple biophysical accounting approach based on fundamental thermodynamic laws and biological principles.

The Planetary Boundaries approach mostly applies to the global level, while Ecological Footprint can be tracked at any scale. Planetary Boundaries are richer in describing the impact of the human economy on the biosphere; Ecological Footprint accounting's strength is built on comparing human demand against ecosystem regeneration, but is therefore in return restricted to only one core dimension—biocapacity. This uni-dimensionality makes it easy to apply and interpret Footprint results. Also, Ecological Footprint limits are based on tracking demand against "sustainable yields" or regeneration. Planetary Boundaries tracks human metabolism against trigger points or levels of irreversible change, many of which cannot be predicted, or even defined with accuracy.

Ecological Footprint and Biocapacity Accounting

Ecological Footprint analysis examines the size of society's metabolism with a specific research question: how much of the regenerative capacity of the biosphere is being occupied by human activities (Rees and Wackernagel, 1994; Wackernagel and Rees, 1996; Wackernagel *et al.*, 2014)? To answer this question, Footprint analysis measuring how much biologically productive land and water area an individual, a city, a country, a region, or humanity uses to produce the resources it consumes, accommodate its infrastructure, and to absorb the waste it generates, using prevailing technology and resource management schemes.

This demand can be compared with supply, or biocapacity, which is the total available biologically productive surface of the planet. The common unit used for this analysis, as the term "Footprint" suggests, is area, more specifically productivity adjusted hectares, called global hectares, representing hectares of land or sea with world-average biological productivity.

The most comprehensive Ecological Footprint accounts are the national calculations by Global Footprint Network, called "National Footprint Accounts." The 2017 edition of these accounts tracks Ecological Footprint and biocapacity of all countries accounts from 1961 to 2013. 2013 is the last year for which a complete set of input data is available (United Nations Food and Agriculture Organization, United Nations Comtrade, International Energy Agency, and others). National Footprint Accounts are updated every year—with a 4-year time lag in data. With the use of extrapolation, more recent proxy data (like GDP), or partial data (e.g., greenhouse gases which are reported with less time lag), it is possible to "nowcast" biocapacity and Footprint results to the present.

Calculating Demand: Footprint

At present, human demand on ecosystems, or a population's Footprint, is translated into demands for six major demand types—crops, grazing products, sea food, forest products, built-up land, and carbon Footprint (**Table 1**, Borucke *et al.*, 2013). Carbon Footprint and forest products both compete for productive forest areas. Hence only five area types accommodate the six demand types. The first four demand categories represent food, fiber, and timber products for human consumption. These products may be consumed directly or processed further before final consumption. Regardless of the forms they eventually take within society, however, all products produced from these four categories can be translated, through the use of yields (annual tons per hectare)

Table 1 Major land types in Ecological Footprint and biocapacity accounting

Ecological Footprint (demand types)	*Biocapacity (areas)*
Crops	Cropland
Grazing products	Grazing area
Sea food	Fishing grounds
Carbon Footprint	*Forest*
Forest products	
Built-up land	Built-up land

The forest biocapacity serves two competing uses: absorbing CO_2 for the carbon footprint and providing forest products such as timber and firewood.

and conversion factors (tons of processed product per tons of raw material), back into the amount of world-average bioproductive area required to produce the products. Note that this land and water area can be located anywhere on the planet.

The fifth land type, built-up land, represents the area required for the physical infrastructure associated with human society, such as cities and roads. The sixth demand, carbon Footprint, represents the amount of biologically productive space (forests) required to absorb one of the human economy's currently most significant waste products: carbon dioxide. This Footprint is calculated as the amount of forested area required to sequester a given amount of carbon dioxide, effectively removing it from the atmosphere, after accounting for absorption by the oceans (Mancini *et al.*, 2016). Of course, this sequestration also needs to be accompanied by biological long-term carbon storage (including the preservation of primarily old growth forests and peat bogs).

In earlier calculations, the Footprint of nuclear electricity was included as on par with coal power for lack of better information. But since 2009, only the CO_2 component of nuclear energy is included in Ecological Footprint accounts to make the assessment fully consistent with the research question. Yet, we also recognize and emphasize that the research question driving Ecological Footprint accounts is not the most relevant and critical one for assessing the viability of nuclear power. Other aspects such as costs, operational risks, the military nexus, and waste storage concerns are more relevant issues to be considered. Yet, the use of nuclear power can lead to the loss of significant amounts of biocapacity, as in the case of the Fukushima nuclear accidents with large, productive areas becoming unavailable to human use for decades if not centuries (WWF-Japan and Global Footprint Network, 2012).

The approach of translating fossil fuel use into bioproductive area needed to sequester the emitted carbon dioxide does not suggest that afforestation or other types of biological sequestration are the main solution to reducing atmospheric carbon dioxide concentrations. This measure does show, however, how much larger the biosphere would need to be to stabilize carbon dioxide concentrations in the atmosphere without further human intervention. It also emphasizes that a reduction in emissions is unavoidable if carbon concentration in the atmosphere is to be contained to an acceptable level. The 2017 Edition of the National Footprint Accounts, for example, calculates that in 2013 the release of 1 ton of carbon dioxide per year has a Footprint of approximately 0.34 global hectares. Other human-supported methods exist for sequestering carbon dioxide, and the use of these technologies will be reflected in a decrease in the energy Footprint as they are brought online. Similarly, the introduction of renewable energy technologies with lower carbon-intensities will lead to a reduced carbon Footprint in the future. In addition, land-use changes such as deforestation can add emissions and reduce biocapacity, while land-use changes from erosive to sustainable agriculture can strengthen or even increase biocapacity over the long-term.

Some have questioned the utility of including the carbon Footprint in the Ecological Footprint (e.g., van den Bergh and Verbruggen, 1999 or Blomqvist *et al.*, 2013). However, given the research question, which focuses on the competing demands for biocapacity, CO_2 emission from fossil fuel use is clearly a competing demand for biocapacity and needs to be included. Without the sequestration, natural capital is not being maintained, and ecological debt accumulates in terms of increasing CO_2 concentrations in the atmosphere. This is parallel to overharvesting timber in a forest, where overuse leads to a depletion of natural capital (or a build-up of ecological debt). Trade-offs in biocapacity demands are meaningfully depicted as emissions can be compensated by either allocating more biocapacity to carbon sequestration, or directly reducing emissions.

Therefore, Ecological Footprint accounting complements and strengthens the rationale for climate action in four ways:

1. *Easy, intuitive and transparent*. Ecological Footprint accounts confirm reduction requirements consistent with the Paris Agreement's 2°C or 1.5°C goal, complementing the more complex, dynamic, and less testable climate models. With basic, widely understood and easily testable scientific principles, the accounts can be understood and audited by anyone with basic science education. (Climate models are needed to estimate remaining carbon budget, estimating the links between carbon concentration, radiative forcing and temperature increase; Ecological Footprint accounting complement them by revealing the steady-state availability of renewable capacity through a simple resource balancing approach).

2. *Ecosystem productivity as a core resource*. The focus on ecosystem productivity (or biocapacity, as expressed in Footprint accounting) becomes a primary concern once we acknowledge commit to a societal transition from fossil fuel to renewable energy sources. Economies will therefore rely primarily on what ecosystems can provide for their physical inputs (plus that energy which is not in competition for bioproductive land such as photovoltaics in deserts or windmills off-shore). Without cheap energy, it might be difficult to maintain high agricultural yields, and the demand for fuel wood may go up, both potentially increasing demand on land. If humanity should continue to use fossil fuels beyond the carbon budget, the ensuing climatic changes would most likely reduce the planet's biocapacity, making it even more difficult to run the economy in the future. One main reason why climate change needs to be avoided is to not endanger biocapacity, on which the human economy inevitably depends.
3. *Connecting land and atmosphere*. The National Footprint accounts are consistent with the Paris Agreement's embracing of net-emissions. The focus on net-emissions recognizes the fundamental link between the atmosphere and the biosphere. It is not only about carbon emissions, but also about how much of the carbon can be sequestered, by biological, technical, or other means. This opens up broader opportunities for mitigation. Biocapacity is a safe and natural way to get rid of excess carbon. But biocapacity is also limited in how much excess carbon it can sequester (much less than current emissions from fossil fuel burning). In contrast, there are many geoengineering ideas but no meaningful governance mechanism to keep them under control.

4. *Emphasis on self-interest*. By putting the climate challenge into the context of biocapacity, the resource security perspective (and its link to an economy's self-interest to be resource secure) becomes more obvious, possibly helping to overcome the common misperception that climate change is an inevitable "tragedy of the commons." Rather, addressing resource security and climate change in all investment choices enables the development of successful, resilient economies.

Calculating Supply: Biocapacity

Human demand, or Footprint, can be compared to the total availability of biologically productive land and sea, or biocapacity. Biocapacity is currently measured in five major land types (**Table 1**), analogous to the six demand types of Footprint (where both the carbon Footprint and the forest product Footprint compete for forest land).

Globally, Footprint analysis identifies, for 2013, approximately 12.2 billion hectares of biologically productive land and sea that can provide economically useful concentrations of renewable resources. These 12.2 billion hectares cover just under one quarter of the planet's surface and include 1.6 billion hectares of cropland, 3.4 billion hectares of grazing land, 4 billion hectares of forest, 3 billion hectares of marine and inland fisheries, and 0.2 billion hectares of built-up land. Built-up (or urbanized) land is included because it represents in most cases highly productive areas that are either paved over or fenced in, excluding other uses. The "compromised biocapacity," that is, how much biocapacity was given up for this urbanized area, of paved over areas is accounted for.

These areas concentrate the bulk of the biosphere's regenerative capacity. Global Footprint Network has not yet been able to estimate or identify reliable existing estimates of how much of the total annual biomass generation or net primary production (NPP) is concentrated on these 12.2 billion hectares. But we assume that those productive areas contain the overwhelming majority of the planet's harvestable NPP. While the remaining areas of the planet are also biologically active and support life, such as the deep oceans or deserts, their renewable resources are not concentrated enough to be a significant addition to the overall biocapacity. Also, low concentration of biomass in those areas makes harvesting their products highly resource intensive.

Many materials and pollutants place demands on the biosphere primarily by reducing the ability of ecosystems to provide goods and services, which leads to a loss in biocapacity. Toxins, heavy metals, and other persistent pollutants fall into this category. The amount of bioproductive area required to mine mercury, for example, is insignificant compared to the extent of the ecosystems that this metal affects. Similarly, the area required to absorb this product is an undefined quantity, as ecosystems do not have a well-defined or understood ability to assimilate this metal naturally. As a result, the environmental impacts of mercury, as well as other toxins, do not appear, nor are they specifically tracked by the Ecological Footprint. The physical area occupied by mines is indeed captured as infrastructure or built-up area, while the energy used in production is captured within the carbon Footprint. The primary effect of such materials is represented by the widespread loss of biocapacity that would result if they were released widely into the environment.

The Common Unit: Global Hectares

Given the widely varying scope of human demands, and the wide variety of ecosystems available on the planet, any aggregate analysis or indicator requires a common metric for comparison. Ecological Footprint accounts compare different types of Footprint to each other and to available biocapacity using a global hectare, defined as a hectare with world-average productivity of the 12.2 billion biologically productive hectares on planet Earth.

In the context of global hectares, biological productivity refers to the productivity (or more precisely the maximum sustainable yields) of an area's products and services that the human economy demands. It is an agro-ecological measure of regeneration, which can directly be compared to harvests. In that way it is related to net primary production (NPP) measures for biomass used in ecological sciences. The agro-ecological approach increases the ability to compare harvest with regeneration, a particular weakness of the otherwise highly valuable NPP assessments (Rojstaczer *et al.*, 2001).

The agro-ecological approach inherent in Footprint accounting also helps to compare across various land-uses. For instance, one hectare of highly productive land (e.g., cropland) is worth more global hectares than one hectare of less productive land (e.g., pasture). Global hectares are normalized so that the number of actual hectares of biologically productive land and sea on the planet is equal to the total worldwide budget of global hectares in any given year (therefore, a global hectare is also an average hectare).

Ideally, hectares should be compared against each other in terms of their inherent potential NPP—their unaltered innate ability to regenerate, independent of the human-imposed land-use. But in absence of such data sets, the National Footprint Accounts use relative agricultural productivity as a first approximation of relative biocapacity.

Ecological conservation areas that are not harvested are counted at their full biocapacity in Global Footprint Network's National Footprint Accounts.

Note that comparing amount of biocapacity available to amount of biocapacity is merely one dimension of describing human dependence on ecological services. Using less than can be regenerated is a necessary, but not sufficient condition for sustainability. In addition to getting quantity right, there are a number of additional quality considerations. For instance, in order to meaningfully preserve biodiversity, intact old growth forests are needed, with no fragmentation or other major human intervention. In other words, a global hectare of plantation forest and a global hectare of old growth forest, while representing the same amount of timber productivity or carbon sequestration capacity, do not represent the same biodiversity value. Keeping the human Footprint

within the biocapacity of ecosystems is just one of the many conditions of biodiversity conservation. However, getting quantity right is a precondition for being able to achieve quality at scale. If the quantity of human demand exceeds what biocapacity can regenerate, the quality (and preservation) of biocapacity management of some of the biocapacity will inevitably come at increased pressure on the remainder of the biocapacity.

Methodology

Data Sources

The most robust Ecological Footprint accounts exist at the national scale. Subnational Footprint calculations are possible and increasingly common. They typically build on the national assessments, and the resolution of the assessments depends on the availability and accuracy of data sources at the subnational level. A few examples are available on Global Footprint Network's website at www.footprintnetwork.org/our-work/ecological-footprint/cities. Many other organizations have performed such calculations over the last decade.

The international community of Footprint practitioners developed standards for Footprint applications—the initial standards were launched in 2006 and updated in 2009 (www.footprintstandards.org).

The National Footprint Accounts are currently maintained by Global Footprint Network, a nonprofit organization headquartered in Oakland, California, and its over 65 partner organizations throughout the world. Results are published annually and made available on Global Footprint Network's data platform data.footprintnetwork.org. Other organizations' publications also publish the results, as for example WWF International's bi-annual *Living Planet Report* (The latest one is WWF, Global Footprint Network Zoological Society of London, 2016).

The national analysis relies heavily on data published by the Food and Agriculture Organization of the United Nations (FAO), the United Nations Statistics Division, the Intergovernmental Panel on Climate Change (IPCC) and the International Energy Agency (IEA). Other data sources, including meta-analyses, scientific publications, and thematic collections, are used to fill in the gaps between these international sources. The use of official, typically conservative sources points toward the high probability that biocapacity estimates may be exaggerated, while Footprints may be underreported.

Yield Factors and Equivalence Factors

In the national assessments, two conversion factors, yield factors and equivalence factors, are used in the calculation of Footprint and biocapacity, particularly for translating results into the common unit of global hectares.

Yield factors are calculated as the ratio of national, country-specific yields for a given product to the average world yield for that same product. This ratio describes the extent to which a biologically productive area in a given country is more (or less) productive than the global average of the same land type. Differences in national and global yields can be due to a wide variety of factors, including variation in climate, soil conditions, available technology, and management regimes. Yield factors are specific to individual land types, countries, and years.

Equivalence factors relate the average productivity of a given land type to the world-average productivity of all biologically productive land types. Cropland, for example, has a higher productivity than world-average land. Grazing land is, on average, less productive than the average of all land types. Equivalence factors are currently calculated using Global Agro- Ecological Zones (GAEZ) data, which provides a spatial model of potential agricultural yields. The equivalence factor for a land type depends on its level of potential agricultural productivity relative to other land types.

Calculating Footprint and Biocapacity

The general formula for calculating the Ecological Footprint associated with the consumption of a quantity of product is given as:

$$EF = (M/NY) \times YF \times EQF \quad (1)$$

where EF is the Ecological Footprint of a given product flow (in global hectares), M is the mass of the product flow (in tons per year), NY is the national yield of the country in which that product was produced (in annual tons per hectare), YF is a yield factor calculated as the ratio of national yields to world yields for a given product, and EQF is an equivalence factor reflecting the relative productivity of a given land type compared to world-average productivity.

This formula can be applied directly to all products harvested directly from the first four major productive land types: cropland, grazing land, fishing ground, and forest land. The Footprint of secondary products (e.g., flour) that are created from primary products (e.g., wheat) is calculated by converting them back into primary-product equivalents.

The Footprint of built-up land is calculated by using the physical extent of the area occupied in built area (in hectares) instead of the product of mass (M) and national yield (NY). Yield and equivalence factors for cropland are applied, reflecting the assumption that most built land occupies former cropland, unless more accurate data is available. The carbon Footprint is calculated using the total mass of carbon dioxide emissions released from a given activity and the world-average sequestration rate of forested land (after subtracting the portion absorbed by the oceans). This rate is used in place of the ratio of national yield (NY) and yield factor (YF).

To calculate the Footprint of a nation, the Footprint of all products consumed within that country is calculated using the formula above. All the product Footprints are then summed. The global Ecological Footprint is calculated as the sum of all national Footprints.

The biocapacity associated with a given productive land or sea area is calculated in a similar fashion:

$$BC = A \times YF \times EQF \tag{2}$$

where BC is the biocapacity of a given area (in global hectares), A is physical extent of the area under analysis (in country-specific hectares), YF is a yield factor for that country and land type, and EQF is an equivalence factor for that land type. Note that this formula is identical to that for Ecological Footprint, except here area substitutes for the product of mass and national yield. This formula can be applied equally to all five major productive land types on the planet: crop land, grazing land, fishing grounds, forests, and built-up land. The regenerative capacity of forests is used to meet the demands (i.e., the Ecological Footprint) for both forest products and carbon sequestration.

Key Global and Regional Results

At the largest scales, global Ecological Footprint analysis shows that the total human Footprint, or demand on ecosystems, exceeds the planet's available biocapacity, or its ability to supply resources and waste sinks. Estimates for 2017 (extrapolated from 2013 data) suggest that the Footprint of humanity exceeds biocapacity by 71%. Global overshoot has existed since the early 1970s (Fig. 1). The most significant growth in Ecological Footprint over this time period has been a result of an increase in the productive land area required to meet human demands for carbon sequestration. This carbon Footprint made up 60% of humanity's Ecological Footprint in 2013. The growth in available biocapacity over time largely reflects increases in the productivity of cropland. These increases were at least partially enabled, however, by increases in application of fertilizers and pesticides and use (and overuse) of freshwater, all of which contributed to the rapidly growing Ecological Footprint over this same time period. Also, it indicates that this increase may not be maintained forever: Overuse reduces availability and also reduces future productivity (as in the example of collapsed fisheries, depleted groundwater, overharvested forests, or climate change). When and how this plays out is still an open, significant research question.

These consistently growing global trends, however, mask significant regional variation (Fig. 2). In 2013, the per person Ecological Footprint of North America and in the EU ws 1.7-fold and 2.1-fold the biological capacity available within those regions. If everyone in the world lived with a level of ecological demand equal to the EU or North American average, humanity would require the regeneration of three or five planet Earths, respectively. Asia-Pacific has an average level of ecological demand that would require 1.4 planet Earths if all of humanity lived like them. A free public data package with all the results in a convenient EXCEL sheet can be downloaded from data.footprintnetwork.org.

Levels of demand also vary significantly between high, middle, and low-income countries. These income categories are defined by the World Bank. In 2013, high-income countries had an average Ecological Footprint of 6.4 global hectares per

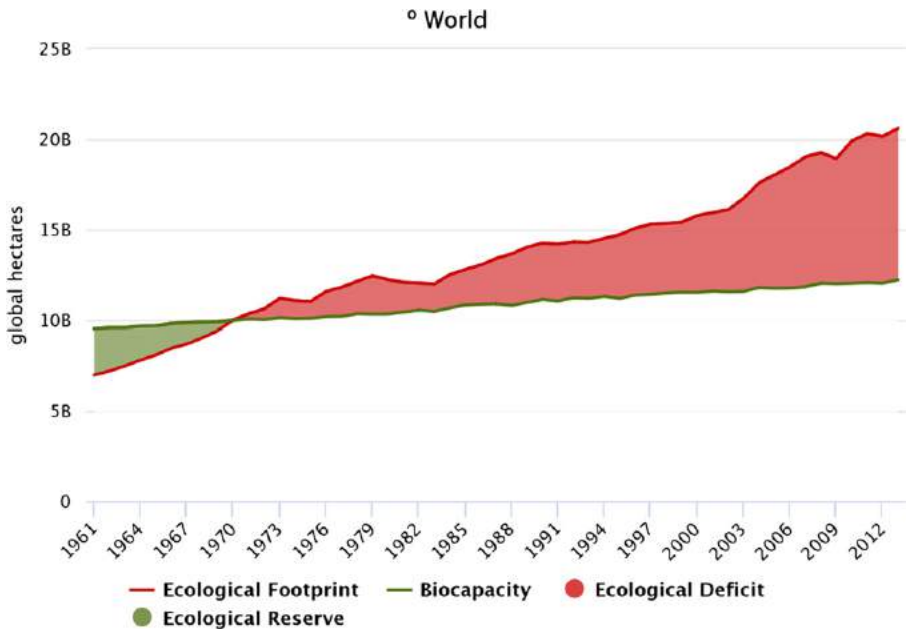


Fig. 1 Humanity's Ecological Footprint and the planet's biocapacity, 1961–2013. Both are measured in global hectares, with hectares representing the area with world-average productivity of the year 2013.

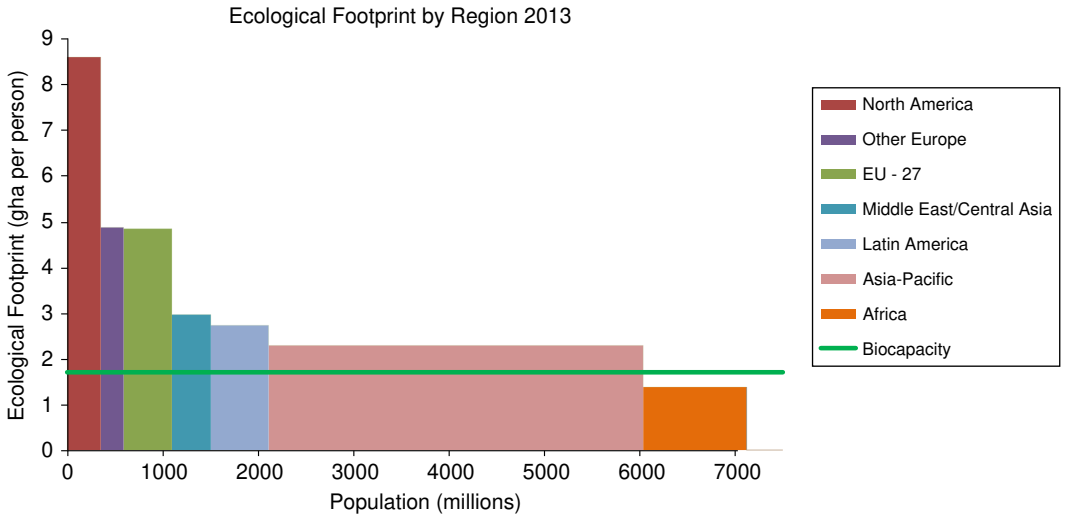


Fig. 2 Regional Ecological Footprint and population, 2013.

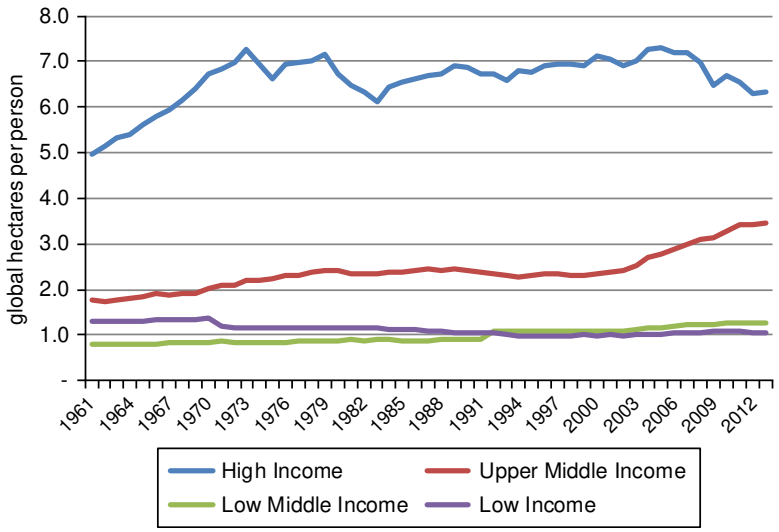


Fig. 3 Ecological Footprint and biocapacity by national income grouping, 1961–2013. Income groupings are defined by the World Bank. The most rapid per person increase took place in upper middle-income countries, mainly driven by the rise of Chinese resource demand (Global Footprint Network, 2017).

person, compared to 1.1 global hectares per person for low-income nations. Over the past 56 years, consumption of ecological resources, per capita, has increased by nearly 30% in high-income countries but fallen by 19% in low-income countries (**Fig. 3**).

More detailed national level Footprint results are published in annual editions by Global Footprint Network. Results for most countries with high enough quality scores are directly available from the open data platform at data.footprintnetwork.org. Further data sets and reports are available on Global Footprint Network's website (www.footprintnetwork.org).

Evaluating the Sustainable Development Performance of Countries

The Ecological Footprint can be applied to evaluate countries' overall performance toward sustainable development. Sustainable development is, as the two words reflect, a commitment to thriving lives for all (development) within the planetary budget (sustainable). Because of these two dimensions, the relationship between the two can be plotted in a graph with two axes (see **Fig. 4**).

The vertical axis (labeled SUSTAINABLE) stands for living within the means of planet Earth. This resource security condition requires an average Ecological Footprint per person of <1.7 global average hectares (the supply of biologically productive

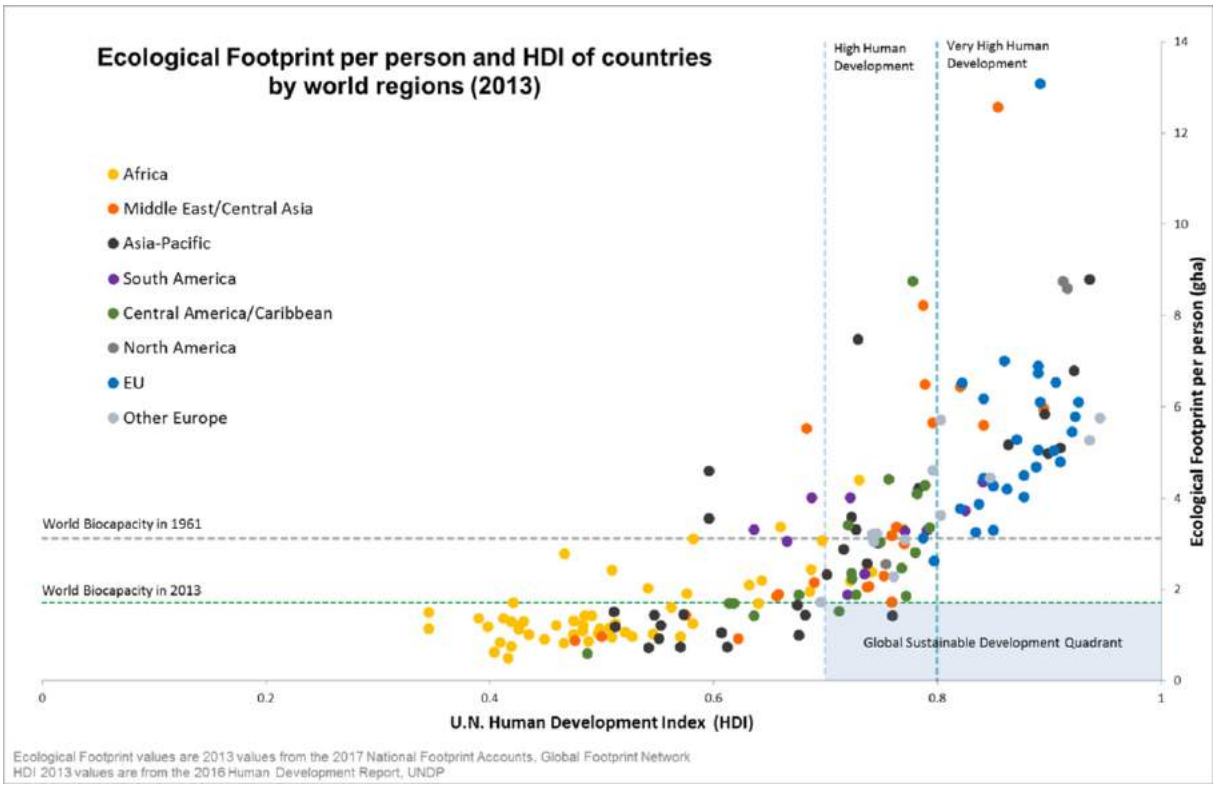


Fig. 4 National Sustainable Development performance can be plotted by comparing countries' per person Ecological Footprint to their HDI achievements (2013 data).

planetary surface area that exists per person in 2013). Note that this amount does not mean that 1.7 is the ultimate goal: the threshold for the Footprint would need to be even lower in order to also support wild species. For instance, if we followed E.O. Wilson's suggestion of leaving half the biocapacity wild there would be 0.85 global hectares per person available at current population levels (Wilson, 2016). In contrast, today's Footprint of humanity is 2.7 global hectares per person.

The horizontal axis represents DEVELOPMENT. Since "development" is not a specific outcome, but rather an issue area, it cannot be measured directly. But indices are available that approximate this concept. The most widely cited is the one conceived and promoted by the United Nations: UNDP's Human Development Index reflects human progress and development outcome through three key components: life expectancy; literacy & education, and per person income of the nation's residents (or more precisely the logarithm of income). On a scale of zero to one, 0.7 is UNDP considered to be the threshold for a high level of development. 0.8 is the threshold for very high.

The sustainability threshold (Ecological Footprint $<$ one Earth) and the development threshold (HDI > 0.7) define two minimum criteria for global sustainable development. The graph with 2013 data shows that few nations have a development model that creates high human development with a resource demand that is globally replicable.

To truly move toward sustainable development, a necessary condition would be to make the world average move toward the identified box of high development within the resource budget of planet earth (Raworth, 2017). Most efforts today do not yet live up to this challenge since on average, humanity's Ecological Footprint is still going up, and biocapacity per person is shrinking (Wackernagel *et al.*, 2017).

The Footprint of Consumption Activities

While the methods and national analysis presented above provide information on the Ecological Footprint and biocapacity of different land types, they do not indicate which types of consumption are responsible for placing these demands. The provision of food products, for example, requires significant quantities of cropland, grazing land, and fishing grounds, and emit significant amount of carbon dioxide. Dividing the Ecological Footprint into its specific consumption components can be valuable for policy applications and communication programs. One way of systematically allocating overall demand to economic activities is ee-MRIO, or environmentally extended Multi Regional Input Output analysis. Global Footprint Network uses a model building on Purdue University's GTAP database. These broken-down results are commonly used in scenario analyses, trade flow analyses, and

by individuals who wish to find ways to reduce their own personal Footprints. Also, while the Footprint does not measure biodiversity loss directly, it tracks global pressures on biodiversity and can be used to complement other measures of ecosystem-specific impacts on biodiversity (Galli *et al.*, 2014).

Various techniques, including input–output analysis and process-based allocation, can be applied to apportion the Ecological Footprint into consumption categories. All of these types of analyses generate a consumption land-use matrix, a table that allocates the total Footprint in each of the major land types across a series of consumption categories. An example for the province of Ontario, Canada, is included below.

By consumption category, personal transportation makes the largest contribution to the average Ontario resident's Ecological Footprint (**Table 2**). The finer level of detail in tables such as this can be used to suggest scenarios for policy making as well as possibilities for individual action.

The Future of Ecological Footprint Applications

Beyond the thousands of existing Footprint applications, much more is still possible. The open data platform, launched in April 2017, has made the results far more easily available, leading to a rapid increase of visits and downloads from the data platform data.footprintnetwork.org. Hopefully, this increased availability will also lead to a new wave of academic publications and fresh ways to analyze the data.

This increased transparency and availability also enables more scrutiny and engagement with the method. Already in the past, various authors have published criticisms about the Footprint and its applications. On the methodological side some of them have been addressed through methodological improvements (Kitzes *et al.*, 2009).

Other critics are uncomfortable with the interpretation of the implications of Ecological Footprint results, even though these may not be specific to the Footprint, but rather a consequence of the planet's finite surface area (Stiglitz *et al.*, 2009, and response by Global Footprint Network (2009). Yet other critics raise issues that are not pertinent to the research question driving Footprint accounting, criticism often inspired by the critics' interpretation of what the name "Footprint" should or could mean according to them. Some of those latter ones are discussed in Lin et al. (2015), Goldfinger et al. (2014), Wackernagel (2014), and Rees and Wackernagel (2014).

To make criticism useful, the first step is to establish whether the criticized method and the criticism both refer to the same research question. Some criticism reads aspects into the name of the methodology without defining or debating the actual research question (examples of such criticism include van den Bergh and Grazi, 2014a, 2014b, 2015, or Blomqvist *et al.*, 2013).

Criticism that actually addresses the pursued research question can legitimately question the method on three grounds:

1. *The method and its results may not be relevant to the identified policy concerns.* Obviously, this would make the method and its results useless. Andrea Collins and Andrew Flynn produced possibly the most comprehensive study to evaluate the utility of the Footprint, particularly in the United Kingdom. They concluded that while the approach shows strength in communicating ecological limits, differences in resource distribution, and interconnectedness of resource issues across scales and geographies, they found that the tool's ability to influence policy outcomes has so far been more limited (Collins and Flynn, 2015). Whether this is a consequence of the challenging sustainability topic per se, or a weakness in the Footprint's relevance still needs to be examined further.
2. *The method and its results, while theoretically relevant may in practice be less accurate than other methods answering the same research question.* van den Bergh and Grazi (2014a) do not question Footprint Accounts on the basis that more accurate methods are available. Rather, they note, "Perhaps it is of consolation to Footprint devotees to know that other efforts to arrive at an aggregate environmental indicator have failed as well." Clearly on specific aspects such as carbon, or particular substances such as nitrogen, other tools can be more accurate. But at the aggregate level of overarching human demand, Global Footprint Network has not been able to identify a method that answers the question with more rigor and robustness.
3. *Even if the method is considered relevant, and critics recognize that there is no better answer to the question available, they may still raise the argument that society would be better off without the results this method generates.* For instance, they may argue that the results are so inaccurate that not having the results is producing less confusion and misdirection than having them. Giampietro and Saltelli (2014) insinuate this even in their title of their 2014 paper "Footprints to Nowhere" in which they set out to make the case that society would be better off without the results this method generates since, according to them answers are so poor that they mislead more than they inform. However, as Global Footprint Network would argue that Footprint results are in line with many other studies addressing humanity's dependence on the Earth's resources and services (Millennium Ecosystems Assessment, 2005; Steffen *et al.*, 2015; IPCC, 2014).

Further, national government agencies of over 12 countries have tested Ecological Footprint accounts. They compared Global Footprint Network results with results they calculated based on their own data sets. For instance, the French statistical office of their environment ministry, on their own, reproduced the French Footprint time series within 1%–3% deviation from the Global Footprint Network numbers. Reports from national and international reviews are collected at www.footprintnetwork.org/reviews. Also international agencies such as UNDP or international organizations such as WWF or WBCSD have used Ecological Footprint results in their publications.

Table 2 Process-based consumption land-use matrix for Ontario (2010)

	(gha person^{-1})	*Crop Footprint*	*Grazing Footprint*	*Forest Products Footprint*	*Fish Footprint*	*Built-up Footprint*	*Carbon Footprint*	*Ecological Footprint*
Household	*Food*	0.56	0.19	0.03	0.07	0.01	0.12	0.98
	Solid food	0.45	0.14	0.02	0.06	0.01	0.09	0.77
	Nonalcoholic beverages	0.04	0.02	0.00	0.00	0.00	0.01	0.08
	Alcoholic beverages	0.07	0.03	0.00	0.01	0.00	0.02	0.13
	Housing	0.01	0.00	0.07	0.00	0.01	0.38	0.46
	Actual rentals for housing	0.00	0.00	0.00	0.00	0.00	0.01	0.01
	Imputed rentals for housing	0.00	0.00	0.01	0.00	0.01	0.01	0.03
	Maintenance and repair of the dwelling	0.00	0.00	0.03	0.00	0.00	0.02	0.06
	Water supply and miscellaneous dwelling services	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Electricity, gas other fuels	0.00	0.00	0.01	0.00	0.00	0.33	0.35
	Service for household maintenance	0.00	0.00	0.00	0.00	0.00	0.00	0.01
	Personal transportation	0.11	0.07	0.09	0.01	0.04	1.32	1.64
	Purchase of vehicles	0.01	0.00	0.02	0.00	0.01	0.12	0.17
	Operation of personal transport equipment	0.09	0.06	0.06	0.01	0.02	0.85	1.10
	Transport services	0.00	0.00	0.01	0.00	0.01	0.35	0.37
	Goods	0.17	0.10	0.18	0.01	0.02	0.23	0.71
	Clothing	0.04	0.03	0.00	0.00	0.01	0.04	0.12
	Footwear	0.01	0.01	0.00	0.00	0.00	0.01	0.02
	Furniture, furnishings, carpets etc.	0.01	0.01	0.01	0.00	0.00	0.02	0.05
	Household textiles	0.01	0.01	0.00	0.00	0.00	0.01	0.02
	Household appliances	0.00	0.00	0.00	0.00	0.00	0.02	0.03
	Glassware, tableware & household utensils	0.00	0.00	0.00	0.00	0.00	0.01	0.01
	Tools and equipment for house & garden	0.00	0.00	0.01	0.00	0.00	0.01	0.01
	Medical products, appliances & equipment	0.00	0.00	0.00	0.00	0.00	0.01	0.01
	Telephone & telefax equipment	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Audio-visual, photo & info. Processing equipment	0.00	0.00	0.01	0.00	0.00	0.03	0.05
	Other major durables for recreation & culture	0.00	0.00	0.00	0.00	0.00	0.01	0.01
	Other recreational equipment etc.	0.02	0.01	0.12	0.00	0.00	0.03	0.19
	Newspapers, books & stationery	0.00	0.00	0.01	0.00	0.00	0.01	0.03
	Goods for household maintenance	0.00	0.00	0.00	0.00	0.00	0.00	0.01
	Tobacco	0.08	0.03	0.00	0.01	0.00	0.02	0.14
	Services	0.04	0.02	0.08	0.00	0.02	0.28	0.45
	Out-patient services	0.00	0.00	0.00	0.00	0.00	0.01	0.01
	Hospital services	0.00	0.00	0.00	0.00	0.00	0.01	0.02
	Postal services	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Telephone & telefax services	0.00	0.00	0.01	0.00	0.00	0.03	0.04
	Recreational & cultural services	0.00	0.00	0.01	0.00	0.00	0.02	0.04
	Package holidays	–	–	–	–	–	–	–
	Education	0.00	0.00	0.01	0.00	0.00	0.04	0.06
	Catering services	0.01	0.00	0.01	0.00	0.00	0.03	0.05
	Accommodation services	0.00	0.00	0.00	0.00	0.00	0.00	0.01
	Personal care	0.00	0.00	0.01	0.00	0.00	0.03	0.04
	Personal effects nec	0.01	0.01	0.01	0.00	0.00	0.02	0.05
	Social protection	0.00	0.00	0.01	0.00	0.00	0.03	0.04
	Insurance	0.00	0.00	0.01	0.00	0.00	0.03	0.05
	Financial services nec	0.00	0.00	0.00	0.00	0.00	0.02	0.03
	Other services nec	0.00	0.00	0.00	0.00	0.00	0.01	0.02
	Subtotal short-term household consumption	0.88	0.39	0.44	0.09	0.10	2.33	4.23
Government paid short-term household consumption		0.04	0.02	0.13	0.00	0.03	0.35	0.56
Gross fixed capital formation		0.10	0.04	0.53	0.01	0.05	0.69	1.42
Total		1.02	0.45	1.10	0.10	0.18	3.36	6.21

The results are grouped by short-term direct demands in Food, Housing, Personal Mobility, Goods, and Services. Demands paid for by government, and investments in capital assets (such as houses, roads and infrastructure) are shown in separate categories. (Global Footprint Network (2015). The Footprint and Biocapacity of Ontario, Canada: Comparing Results for 2005 and 2010. California: Global Footprint Network. Produced for the Ontario Ministry of Natural Resources and Forestry. http://sobr.ca/_biosite/wp-content/uploads/Zokai-et-al.-2015_Ontario-Ecological-Footprint-and-Biocpacity-Technical-Report-2015-03-23-final.pdf.)

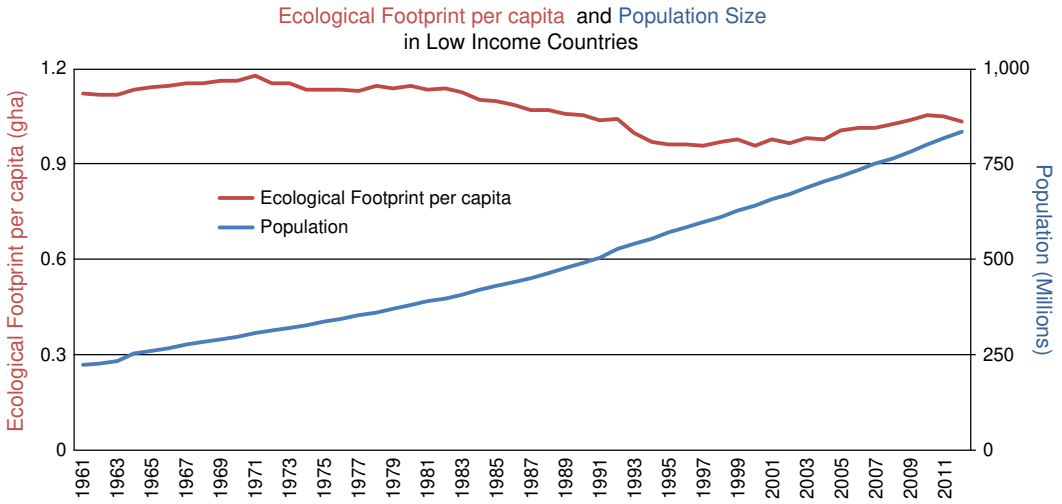


Fig. 5 Population size of current Low Income Countries over time, and their per person Ecological Footprint, 1961–2012. This graph shows in blue the population trend and in red the per person Ecological Footprint averaged across all people in Low Income Countries, from 1961 to 2012. The population nearly quadrupled in those countries, with their Ecological Footprint per person slightly declining. This decline is from a human health and development perspective particularly troubling given that the Ecological Footprint in these countries is already at a very low level. Also note that in these countries the Footprint per unit of cereal and animal protein is typically higher than world average because they need specialty cereals for marginal land areas and because animals are kept as savings leading to a high ratio between feed and fodder input to animal product output. Source: Global Footprint Network (2017). National Footprint Accounts, www.footprintnetwork.org.

By offering accounts that summarize Footprint and biocapacity, the main factors which determine a biocapacity deficit or a biocapacity reserve become clear. Similar to Holdren and Ehrlich's IPAT equation (Ehrlich and Holdren, 1971), there are five factors that shape the biocapacity reserve or deficit. The per person Footprint is determined by two factors: how much is consumed, as well as the efficiency by which the consumed items are being produced. The per person biocapacity is determined by three factors: amount of biologically productive space; the productivity per hectare of this space; and the number of people sharing this space. Dissecting the accounts therefore allows analysts to separate out the various contributions to the overall Footprint, and to develop scenarios about how the ratio between Footprint and biocapacity can be altered. For instance, by comparing changes in per person demand with changes in number of people, one can see which factor is moving more rapidly (**Fig. 5**).

Academic articles on the Footprint have multiplied. Particularly notable is the strong interest among Chinese scholars which could be driven by the Chinese government's focus on Eco Civilization, as noted in their 13th five-year plan. Google Scholar search finds, as of May 2017, 55,000 documents containing the phrase "Ecological Footprint."

An overview of the uses from a UK perspective has been provided by Collins and Flynn (2015), published in their above mentioned book. Earth Overshoot Day (www.overshootday.org), which marks every year the day by which humanity has used the annual renewable budget of planet Earth is reaching increasing number of audiences. For instance, Earth Overshoot Day 2016 (marked on Aug 8th of that year) generated at least 1.9 billion media impressions through 1800 media events in 84 countries. On September 25, 2016, Switzerland voted on whether it wanted to achieve the goal of living within the means of one planet by 2050. 36% of the voters agreed, recognizing that the current Footprint, which would take more than three planet Earths if everybody lived like the Swiss, does ultimately not serve Switzerland (www.achtung-schweiz.org/en).

Economic planning still has not embraced biocapacity as the ultimate biophysical resource on which economic activity depends. Providing robust, relevant data is only one piece for changing that mindset. But we hope it is a productive contribution to a necessary change in the way we hold and act upon humanity's dependence on Earth's regenerative capacity.

More information about Footprint results, methodologies, and applications can be found on Global Footprint Network's website at www.footprintnetwork.org.

See also: Human Ecology and Sustainability: Limits to Growth; Tragedy of the Ecological Commons; System Sustainability; Carbon Footprint; Human Population Growth; Ecosystem Services Evaluation

References

Blomqvist, L., Brook, B.W., Ellis, E.C., Kareiva, P.M., Nordhaus, T., *et al.*, 2013. Does the shoe fit? Real versus imagined ecological footprints. PLoS Biology 11 (11), e1001700. doi:10.1371/journal.pbio.1001700.

Borucke, M., Moore, D., Cranston, G., Gracey, K., Iha, K., *et al.*, 2013. Accounting for demand and supply of the Biosphere's regenerative capacity: The National Footprint Accounts' underlying methodology and framework. Ecological Indicators 24, 518–533.

Collins, A., Flynn, A., 2015. The ecological footprint: New developments in policy and practice. UK: Edward Elgar Publishers, p. 232.
Ehrlich, P.R., Holdren, J.P., 1971. Impact of population growth. Science. American Association for the Advancement of Science 171 (3977), 1212–1217. doi:10.1126/science.171.3977.1212.
Galli, A., Wackernagel, M., Iha, K., Lazarus, E., 2014. Ecological Footprint: Implications for biodiversity. Biological Conservation 173, 121–132.
Giampietro, M., Saltelli, A., 2014. Footprints to nowhere. Ecological Indicators 46, 610–621.
Global Footprint Network (2009). Response to the "Commission on the Measurement of Economic Performance and Social Progress" (or "Stiglitz Commission") Report, Oakland, CA. http://www.footprintnetwork.org/images/uploads/Global%20Footprint%20Network%20Stiglitz%20response.pdf.
Global Footprint Network, 2017. www.footprintnetwork.org
Goldfinger, S., Wackernagel, M., Galli, A., Lazarus, E., Lin, D., 2014. Footprint facts and fallacies: A response to Giampietro and Saltelli, (2014). "Footprints to nowhere". Ecological Indicators 46, 622–632.
IPCC, 2014. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Core Writing Team, Pachauri, R.K., Meyer, L.A. (Eds.), Geneva, Switzerland: IPCC, p. 151. https://www.ipcc.ch/pdf/assessment-report/ar5/syr/SYR_AR5_FINAL_full_wcover.pdf (accessed April 27, 2018).
Kitzes, J., Galli, A., Bagliani, M., Barrett, J., Dige, G., Ede, S., Erb, K., Giljum, S., Haberl, H., Hails, C., Jolia-Ferrier, L., Jungwirth, S., Lenzen, M., Lewis, K., Loh, J., Marchettini, N., Messinger, H., Milne, K., Moles, R., Monfreda, C., Moran, D., Nakano, K., Pyhälä, A., Rees, W., Simmons, C., Wackernagel, M., Wada, Y., Walsh, C., Wiedmann, T., 2009. A research agenda for improving national Ecological Footprint accounts. Ecological Economics 68 (7), 1991–2007.
Lin, D., Wackernagel, M., Galli, A., Kelly, R., 2015. Ecological Footprint: Informative and evolving—A response to van den Bergh and Grazi (2014). Ecological Indicators 58, 464–468 http://www.sciencedirect.com/science/article/pii/S1470160X15002186.
Mancini, M.S., Galli, A., Niccolucci, V., Lin, D., Bastianoni, S., Wackernagel, M., Marchettini, N., 2016. The Ecological footprint: Revisiting the carbon component. Ecological Indicators 61, 390–403.
Millennium Ecosystems Assessment, 2005. http://www.millenniumassessment.org (accessed April 2018).
Rees, W.E., Wackernagel, M., 1994. Ecological footprints and appropriated carrying capacity: Measuring the natural capital requirements of the human economy. Chapter 20 In: Jansson,Folke,Hammer,Costanza, Investing in natural capital. Washington DC: Island Press.
Robèrt, K.-H., Schmidt-Bleek, B., Aloisi de Larderel, J., Basile, G., Jansen, J.L., Kuehr, R., Price Thomas, P., Suzuki, M., Hawken, P., Wackernagel, M., 2002. Strategic sustainable development—Selection, design and synergies of applied tools. Journal of Cleaner Production 10 (3), 197–214.
Rockstrom, J., Steffen, W., Noone, K., Persson, A., Chapin III, F.S., Lambin, E., Lenton, T.M., Scheffer, M., Folke, C., Schellnhuber, H., Nykvist, B., De Wit, C.A., Hughes, T., van der Leeuw, S., Rodhe, H., Sorlin, S., Snyder, P.K., Costanza, R., Svedin, U., Falkenmark, M., Karlberg, L., Corell, R.W., Fabry, V.J., Hansen, J., Walker, B., Liverman, D., Richardson, K., Crutzen, P., Foley, J., 2009. Planetary boundaries: Exploring the safe operating space for humanity. Ecology and Society 14 (2), 32. (online). http://www.ecologyandsociety.org/vol14/iss2/art32/.
Rojstaczer, S., Sterling, S.M., Moore, M.N.J., 2001. Human appropriation of photosynthesis products. Science 294 (5551), 2549–2552.
Raworth, K., 2017. Doughnut Economics: Seven Ways to Think Like a 21st-Century Economist. Vermont: Chelsea Green Publishing.
Steffen, W., Richardson, K., Rockström, J., Cornell, S.E., Fetzer, I., Bennett, E.M., Elena, M., Biggs, R., Carpenter, S.R., de Vries, W., de Witt, C.A., Folke, C., Gerten, D., Heincke, J., Mace, G., Persson, G.L.M., Ramanathan, V., Reyers, B., Sörlin, S., 2015. Planetary boundaries: Guiding human development on a changing planet. Science 347 (6223), doi:10.1126/science.1259855.
Stiglitz, J., A. Sen, and J-P. Fitoussi (2009), Report by the Commission on the Measurement of Economic Performance and Social Progress, Available at: www.stiglitz-sen-fitoussi.fr.
van den Bergh, J.C.J.M., Grazi, F., 2014a. Ecological footprint policy? Land use as an environmental indicator. Journal of Industrial Ecology 18 (1), 10–19.
van den Bergh, J.C.J.M., Grazi, F., 2014b. Response to Wackernagel. Journal of Industrial Ecology 18 (1), 23–25.
van den Bergh, J.C.J.M., Grazi, F., 2015. Reply to the first systematic response by the global footprint network to criticism: A real debate finally? Ecological Indicators 58, 458–463.
van den Bergh, J.C.J.M., Verbruggen, H., 1999. Spatial sustainability, trade and indicators: An evaluation of the "Ecological Footprint". Ecological Economics 29 (1), 61–72.
Wackernagel, M., 2014. Comment on "ecological footprint policy? Land use as an environmental Indicator". Journal of Industrial Ecology 18 (1), 20–23. doi:10.1111/jiec.12094. http://onlinelibrary.wiley.com/doi/10.1111/jiec.12094/abstract.
Wackernagel, M., Rees, W.E., 1996. Our ecological footprint: Reducing human impact on the Earth. Gabriola Island: New Society Publishers.
Wackernagel, M., Cranston, G., Morales, J.C., Galli, A., 2014. Chapter 24: Ecological footprint accounts: From research question to application. In: Neumayer, E., Agarwala, M. (Eds.), Handbook of sustainable development, Second revised edition Cheltenham, UK: Edward Elgar Publishing.
Wackernagel, M., Hanscom, L., Lin, D., 2017. Making sustainable development goals (SDGs) consistent with sustainability. Frontiers in Energy Research 5, 18. doi:10.3389/fenrg.2017.00018.
Wilson, E.O., 2016. Half-earth: Our planet's fight for life. Liveright Publishers. p. 272.
World-Wide Fund for Nature International (WWF), Zoological Society of London, Global Footprint Network, 2016. Living Planet Report 2016. Gland, Switzerland: WWF, www.panda.org/livingplanet
WWF-Japan and Global Footprint Network, 2012. Japan ecological footprint report. Japan: Tokyo, p. 2012. http://www.footprintnetwork.org/images/article_uploads/Japan_Ecological_Footprint_2012_Eng.pdf

Further Reading

Barrett, J. (n.d.). Component ecological footprint: Developing sustainable scenarios. Impact Assessment and Project Appraisal 19:2, 107–118.
Chambers, N., Simmons, C., Wackernagel, M., 2000. Sharing nature's interest: Ecological footprints as an indicator for sustainability. London: EarthScan.
Costanza, R., Ayres, R., Deutsch, L., Jansson, A., Troell, M., Rönnbäck, P., Folke, C., Kautsky, N., Herendeen, R., Moffat, I., Opschoor, H., Rapport, D., Rees, W., Simmons, C., Lewis, K., Barrett, J., Templet, P., Van Kooten, C., Bulte, E., Wackernagel, M., Silverstein, J., 2000. Forum: The dynamics of the ecological footprint concept. Ecological Economics 32 (3), 341–394.
Global Footprint Network and the University of Sydney, 2005. The ecological footprint of Victoria: Assessing Victoria's demand on nature. Report prepared for EPA Victoria. Available at: .http://www.epa.vic.gov.au/Eco-footprint/.
Monfreda, C., Wackernagel, M., Deumling, D., 2004. Establishing national natural capital accounts based on detailed ecological footprint and biological capacity accounts. Land Use Policy 21, 231–246.
Rees, W.E., Wackernagel, M., 2013. The shoe fits, but the footprint is larger than Earth. A response to the Breakthrough Institute's PLoS paper. PLoS Biology 11 (11), e1001701. doi:10.1371/journal.pbio.1001701. http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001701.
Wackernagel, M., Schulz, N.B., Deumling, D., Linares, A.C., Jenkins, M., Kapos, V., Monfreda, C., Loh, J., Myers, N., Norgaard, R., Randers, J., 2002. Tracking the ecological overshoot of the human economy. Proceedings of the National Academy of Sciences, USA 99 (14), 9266–9271.
Wackernagel, M., Kitzes, J., Moran, D., Goldfinger, S., Thomas, M., 2005. The ecological footprint of cities and regions: Comparing resource availability with resource demand. Environment and Urbanization 18 (1), 103–112.
Wiedmann, T., Minx, J., Barrett, J., Wackernagel, M., 2006. Allocating ecological footprints to final consumption categories with input-output analysis. Ecological Economics 56 (1), 28–48.

Environmental Protection and Ecology☆

Clive Hamilton and Andrew Macintosh, Australian National University, Canberra, ACT, Australia
Nicoletta Patrizi and Simone Bastianoni, University of Siena, Siena, Italy

What Is Environmental Protection

Environmental protection can be defined as the prevention of unwanted changes to ecosystems and their constituent parts. This includes

- the protection of ecosystems and their constituent parts from changes associated with human activities; and
- the prevention of unwanted natural changes to ecosystems and their constituent parts.

One issue associated with this definition is whether "ecosystems and their constituent parts" include humans and communities, or whether environmental protection is only concerned with the protection of natural capital. From an ecological perspective, humans are regarded as an integral part of the ecosystem. Separating humanity from the natural environment can therefore be seen as artificial. While this is true, the phrase environmental protection is not used to refer to measures that are designed to regulate or mediate direct interaction between people. For example, laws prohibiting assault are not regarded as environmental protection measures. Environmental protection is concerned with the relationship between people and the natural environment rather than the relationships between people and communities.

Another issue is whether environmental protection relates to preservation, conservation, or both. Preservation refers to the protection of an ecosystem or natural environment from change, while conservation is generally associated with the sustainable use of natural resources. The objective of conservation is to ensure the maintenance of a stock of renewable resources that is being exploited for human purposes rather than the protection of the natural environment from any anthropogenic modifications. The exploitation of natural resources for human purposes is not environmental protection as it is not associated with the prevention of unwanted changes. The change associated with exploitation is deliberate and wanted, at least by those doing the exploitation. However, measures that are put in place to prevent overexploitation of natural resources do constitute environmental protection. They are designed to prevent exploitation beyond a point that is deemed desirable or sustainable. For example, catch quotas in fisheries and air pollution limits are environmental protection measures because, while they accept some environmental degradation, they aim to limit it.

The distinction between preservation and conservation has dissipated in recent years with growing recognition of the dynamic nature of natural systems, humanity's place in the biosphere, and the need for active human involvement to maintain the integrity of certain ecosystems. Consequently, environmental protection is now generally used to refer to measures that have traditionally been associated with preservation (e.g., reserves, including national parks), as well as conservation and natural resource management initiatives.

A critical aspect of environmental protection is that it is driven by the values that humans attribute to different aspects of the environment. These values need not be instrumental, but the motivating factor for environmental protection is always the prevention of changes to the environment that humans do not want. This is why measures associated with the prevention of unwanted natural changes to ecosystems—like the prevention of coastal erosion or systematic burning in reserves to reduce the risk of wildfires—can be included as environmental protection. Such measures do not aim to protect ecosystems from human activities but rather from natural forces that are deemed to threaten human interests.

☆*Change History:* March 2018. Simone Bastianoni and Nicoletta Patrizi updated the article. Recent initiatives on the field of Environmental Protection and Ecology have been added in the Conclusion section as well as their references.

Environmental remediation is distinct from environmental protection as its primary objective is to restore an ecosystem or natural environment to a previous state; that is, like exploitation, it is associated with deliberately induced change, as opposed to the prevention of change.

Domestic Policy Instruments

There are a number of ways to classify domestic policy instruments that are used to protect the environment. Some people divide them into voluntary and mandatory instruments, while others place them into regulation-based and incentive-based, or command-and-control and economic instruments. The method preferred here is to divide them into four broad categories: regulatory, economic, voluntary, and education and information.

Regulatory Instruments

Regulatory instruments impose legally enforceable restrictions on economic agents to realize environmental protection objectives. They are sometimes referred to as "command-and-control" mechanisms because they prohibit or mandate certain actions (i.e., the command), while using various forms of punishment to motivate compliance (i.e., the control mechanism). Environmental regulations can take many different forms, including

- prohibitions on specified activities (e.g., discharging pollutants into a water body or the atmosphere, or taking a threatened species);
- requirements to obtain a governmental approval (or permit) before undertaking a specified activity (e.g., pollution permits, operating licenses, and development approvals);
- requirements to follow certain procedures when carrying out specified activities (e.g., to use certain equipment, abide by operating standards, or to monitor pollution emissions); and
- requirements to undertake specified actions that are deemed to be environmentally beneficial (e.g., weed control in agricultural areas).

The use of regulatory instruments can be justified on the basis of deterrence or similar choice theories of criminology. According to this approach, people are assumed to make choices by rationally weighing the costs and benefits associated with alternative courses of action and selecting the action that is most likely to maximize their utility. Regulations can affect this process by altering the costs and benefits associated with environment-related activities. For example, by outlawing the emission of pollutants into a river and incarcerating or imposing fines on people who violate the law, the government is able to increase the costs of emitting pollutants. In doing so, it makes alternative nonpolluting options more attractive. In countries where the rule of law prevails, provided the penalties and probability of enforcement are high enough, and punishment is swift, the desired pattern of behavior should emerge.

The difficulty with this theory is that humans are not always rational utility maximizers, meaning that environment regulations may not always result in the desired outcomes. Regulatory approaches can also fail to achieve their objects because of poor design (e.g., ambiguous regulations and unworkable administrative arrangements), strategic avoidance by polluters, and an absence of monitoring and enforcement due to a lack of resources or political will. There are even cases where regulatory instruments have aggravated the environmental problems they were designed to solve.

Regulatory instruments are criticized not only for being ineffective, but also for their inefficiency. This is because they can impose inflexible restrictions on producers, thereby limiting the choices that are available as to how producers meet the desired environment target. Regulatory mechanisms can also have large administration and compliance costs. Government agencies are required to constantly monitor compliance and undertake costly litigation when breaches are discovered, while producers incur legal and other costs while attempting to abide by the environmental regulations. In addition, regulatory mechanisms do not provide incentives to encourage the reallocation of resources toward producers with the lowest marginal costs of environmental protection. As a result, it is difficult for regulatory mechanisms to satisfy the equimarginal principle, which requires that the marginal cost of environmental protection be equalized across polluters to achieve the desired environment target at the lowest possible cost.

The other main criticism of environment regulation is that it can be inequitable. Disputes about whether regulatory mechanisms are unfair are usually framed in terms of the alteration of preexisting property rights. Some people may believe they have a right to pollute or use the environment in a particular way. When environment regulations are introduced, these property rights may be taken away, prompting calls for compensation. In the absence of compensation, the people who are affected by the regulations may feel they are being forced to shoulder a disproportionate amount of the financial burden associated with the provision of public good environmental outcomes. This problem is often confronted with laws that prohibit the removal of native vegetation or the taking of native species, or that restrict the development of real property for commercial purposes.

In several countries disputes about the impact of environment regulations on property rights have been a central part of environmental policy debates. For example, in the United States, the Fifth Amendment to the Constitution provides that private property shall not be "taken for public use without just compensation." This has led to numerous Supreme Court cases and an extensive literature on so-called "regulatory takings." The Australian Constitution contains a similar provision, which provides the

Federal Government with the power to make laws with respect to the "acquisition of property on just terms." In both jurisdictions, courts have found that these constitutional provisions limit the extent to which the government can abolish property rights for environmental purposes.

While issues surrounding the abrogation of property rights have been influential in framing equity issues associated with environmental regulations, it is often forgotten that most regulations alter property rights. The elevated status of property right issues in the context of environmental debates is arguably due to the reverence accorded to real property and preferential treatment given to certain primary industries in western countries.

Despite the criticisms of regulatory measures, they remain the most widely used instrument for environmental protection. There is an ongoing debate about why states tend to prefer regulatory instruments, with supporters of economic instruments often attributing it to ignorance of alternatives, simplicity, and pressure by special interest groups. Yet, regulatory instruments have a number of attributes. The most important of these is the certainty they can provide when faced with imperfect information about environmental risks and the irreversibility of certain forms of environmental harm (e.g., species extinction). In these circumstances, relying on economic, voluntary, or information instruments can lead to suboptimal outcomes by permitting excessive exploitation of the environment. Regulatory instruments are well suited to these situations as they allow distinct boundaries to be placed on the use of environmental resources. This has led to the advocacy of decision-making tools like the precautionary principle and collective choice processes like the safe minimum standards approach.

Another benefit of regulatory instruments is they can be more cost-effective than other measures in dealing with certain types of environmental issues. As discussed, regulatory instruments are often criticized for having large administrative and compliance costs. Yet, in some cases, alternative economic and voluntary approaches may have higher administrative and compliance costs because of the nature of the environmental problem and the complexity in the required program. For example, addressing land clearing with voluntary measures can require agreements to be negotiated with large numbers of landholders, leading to high transaction costs and a plethora of different standards. In contrast, regulations can provide a uniform standard that does not require case-by-case negotiations and is easier to monitor and enforce. Regulatory instruments can also overcome free-rider problems by forcing recalcitrant polluters to abide by the necessary standards.

Economic Instruments (Market-Based Measures)

Economic instruments can be defined as mechanisms that force economic agents to internalize all or part of the social costs associated with environmentally harmful activities and that rely on market forces to promote efficiency. In doing so, they seek to impose additional costs on producers that harm the environment and reward those that improve environment outcomes, while utilizing market forces to improve the allocation of resources. (Some analysts include subsidies among economic instruments but as they are voluntary and economic agents are not forced to internalize the social costs they are more appropriately classified as voluntary instruments.)

This approach to environmental protection is usually associated with environmental economics, a school of economic thought that is a subdiscipline of neoclassical economics. According to environmental economists, environmental problems arise because of the existence of externalities—impacts involuntarily incurred by a person or persons without compensation or payment as a result of the actions of another. Because of the existence of externalities, markets are unable to guarantee the efficient allocation of resources. For example, if producers emit pollution into the atmosphere without paying for it, the price that consumers pay for the producers' outputs will not reflect the full social cost of the transaction. As a result, there will be excessive output and consumption of the relevant good or service. If producers are forced to internalize the social costs associated with the air pollution, there would be a more efficient tradeoff between air pollution and output, leading to higher net social welfare.

The more recent trend in environmental economics has been to characterize environmental problems as being a product of the incomplete allocation of property rights. According to this approach, if a property right over the relevant environmental resource were appropriately defined and allocated to individuals, and there was perfect information and no transaction costs, the operation of market forces would lead to efficient outcomes. For example, if the atmosphere were owned by someone and producers had to pay to emit pollution, then negotiation between the owner and producers would ensure the most efficient allocation of atmospheric resources. On the basis of these theories, economic instruments either

- require polluters to pay for all or part of the costs associated with pollution (e.g., pollution fees, individual liability, and removal of subsidies that promote overuse);
- place a restriction on the amount of pollution that can be emitted or resource that can be used and then allow pollution or resource entitlements to be traded among economic agents (called "marketable permit" or "cap-and-trade" schemes, e.g., tradable emission, water, catch and development rights schemes); or
- seek to create well-defined, secure, and transferable property rights over environmental resources and allocate these to relevant individuals or groups ("pure property rights" approaches, e.g., land titles and fishing area rights).

Marketable permit schemes and pure property rights approaches are similar in that both rely on the creation and exchange of property rights to promote environmental and economic outcomes. But pure property rights approaches place no external restrictions on the use of the relevant resource and rely on market incentives to achieve the desired environmental outcome,

while marketable permit schemes rely on a cap or limit on the use of the relevant resource to achieve the desired environmental outcome.

One of the major benefits associated with economic instruments is that by utilizing market forces they can encourage a more efficient allocation of resources. For example, when tradable emission quotas are used, the operation of market forces should ensure that the necessary emission reductions are achieved at least cost (i.e., the equimarginal principle should be satisfied). Further, economic instruments provide an incentive for producers to reduce pollution, which encourages innovation. Advocates of economic instruments also claim they are more flexible than regulatory instruments, although this is not always the case.

Although economic instruments can be more efficient than alternative policy mechanisms, they can suffer from a number of weaknesses. In relation to pollution fees, individual liability and pure property rights approaches, there can be a considerable amount of uncertainty associated with environmental outcomes. For example, producers may choose to absorb the increase in costs associated with a pollution fee, or demand may be unresponsive to price rises, meaning the level of pollution may not decline by the desired amount. Consequently, where policy-makers are faced with uncertainty regarding environmental risks and questions regarding irreversibility, alternative approaches can be preferable.

Like regulatory approaches, marketable permit schemes (or cap-and-trade approaches) can place an upper limit on the permissible amount of pollution or resource extraction. Hence, they can be useful in dealing with uncertainty and threshold effects. The advantage that marketable permit schemes offer is that having set a specified limit on pollution or resource extraction, they allow market forces to determine the allocation of pollution or extraction rights among producers. One of the most successful marketable pollution permit schemes has been the United States Environmental Protection Agency's Sulfur Dioxide Program, which is part of the broader Acid Rain Program. The cost of reducing emissions was substantially lower than predicted because producers had an incentive to find cheaper ways to do so.

Problems arise with marketable permit schemes when there is a lack of equivalence between the environment or pollution units that producers are expected to trade (i.e., the resource is not homogeneous). For example, tradable development permit schemes that place a limit on the amount of development in an area but allow developers to exchange development rights can lead to the rights moving toward the developments with the highest economic returns. However, they will not necessarily achieve biodiversity objectives as each parcel of land may contain different biodiversity values. Similar problems can arise with emission schemes that allow emission permits to be generated through the enhancement of sinks (i.e., there can be uncertainty about whether the enhancement of sinks will offset the additional emissions).

Transaction costs can also pose problems for economic instruments. Devising schemes that can be administered in a cost-effective manner can sometimes be difficult. Further, if there are excessive costs associated with the negotiation and exchange of marketable permits, the efficiency benefits may not materialize.

As with all environmental policy mechanisms, politics can impede the effective use of economic instruments. However, economic instruments can be especially vulnerable to political influences if it is necessary to constantly adjust the price signals provided through the scheme. For example, if a carbon tax is used to address climate change, it will be necessary to adjust the tax over time to account for unexpected events and new information. Special interest groups may impede this process, thereby undermining the efficacy of the tax.

There has been a tendency in the past for regulatory instruments and economic instruments to be presented as substitutes. In practice, these two types of instruments are generally used as complements and economic instruments always require a regulatory framework. Indeed, there is a growing recognition of the need for policy packages or policy mixes that use a range of instruments to achieve environmental protection objectives.

Voluntary Approaches

Voluntary approaches can be defined as any mechanism or program that aims to protect the environment where relevant economic agents are able to decide whether or not to participate; that is, involvement in the program is voluntary and no direct penalties are imposed on nonparticipants, although incentives may be used to encourage participation.

There are three broad types of voluntary approach.

- Unilateral initiatives where polluters act without direct government involvement to protect the environment. The defining features of unilateral initiatives are that they are initiated, designed, and operated by polluters. Hence, government involvement is generally limited, which raises questions about whether unilateral initiatives are a policy mechanism or a type of market behavior. Yet governments can encourage unilateral initiatives by suggesting them to polluters or threatening mandatory measures. There are three main types of unilateral approaches: voluntary adjustment of internal processes (e.g., under an environmental management plan); industry self-regulation (e.g., codes of conduct); and environmental certification schemes (e.g., organic producer associations).
- Bilateral agreements between the regulator and a polluter or group of polluters. These initiatives involve negotiation between the parties about how environmental protection will be achieved. Both parties have obligations under the agreement with polluters generally expected to meet certain targets and abide by conditions to protect the environment and the regulator generally expected to provide some sort of incentive. The incentives provided by regulators can include subsidies (e.g., monetary payments and technical assistance), public recognition, and undertakings not to enforce regulations or to introduce new regulations. The

agreements need not be legally binding, but there must be negotiation and some sort of understanding about the obligations of the parties.

- Voluntary public (or government) programs where the regulator determines who is eligible to participate, the obligations of participants, and the incentives used to encourage compliance. The key to these types of programs is that they are initiated and designed by the regulator, and relevant producers are invited and encouraged to participate. Again, the types of inducements include grants, technical assistance, and public praise.

The main advantages of voluntary approaches are that they are flexible (which provides polluters with the freedom to find cost-effective solutions) and noninterventionist. In addition, where disputes arise about the fairness of polluter-pays instruments like regulations and pollution fees, voluntary instruments can help overcome political resistance by enabling governments to pay polluters for the loss of property rights (i.e., they can resort to a beneficiary-pays approach). Due to these characteristics, voluntary approaches are often supported by polluters, which can assist in reducing the political costs for regulators. Further, it is sometimes claimed that voluntary approaches have lower administrative costs than other instruments and that in certain circumstances they can be more effective than mandatory approaches.

Although voluntary approaches can offer some benefits, because of the public good characteristics of many environmental goods and services (i.e., they are nonrival and nonexcludable) they are unlikely to result in an optimal level of environmental protection. In particular, there is a risk that some producers will seek to free-ride on the environmental protection measures undertaken by others. Like some economic instruments, voluntary mechanisms also lack certainty and are ill-suited to dealing with uncertainty and irreversibility.

Voluntary approaches can also be expensive to operate and administer. The incentives necessary to encourage participation can impose a significant burden on taxpayers. These problems can be exacerbated by gaming on behalf of polluters where they seek to take advantage of information asymmetries to extract excessive economic rent. In addition, the transaction costs associated with voluntary approaches can be high if there is a need to negotiate agreements with a significant number of producers.

Given the weaknesses associated with voluntary mechanisms, they are often seen as being used when political resistance blocks the introduction of more effective instruments or as a mechanism that supports or complements other programs.

Research has shown that if voluntary approaches are to be effective, the existence of a strong and credible threat of regulation is essential. The existence of an appropriate threat of regulation increases the incentive for polluters to participate and bolsters the bargaining position of regulators. It can also reduce the financial incentives needed to ensure participation, which can improve the cost-effectiveness of the program.

Information and Education Instruments

Information and education instruments aim to promote environmental protection by improving people's awareness and understanding of environment issues and building their capacity to respond to environmental threats. They include such things as environmental and sustainability reporting by governments and corporations, and advertising and education campaigns.

Many environmental problems arise at least partly because of imperfect information about environmental risks and a lack of awareness about how to respond. These types of instruments can help overcome these issues and can be effective where the threats to the environment are known and producers have an economic incentive to improve environmental outcomes. Information instruments can also be an important tool for encouraging greater support for environmental protection in the community, which can reduce the political costs associated with various environment policy instruments.

The main flaw associated with information instruments is that they will rarely get at the root causes of environmental degradation. As a result, on their own, they are generally an ineffective means of achieving environmental objectives. However, information instruments are commonly viewed as an essential part of environmental policy packages. Without information instruments it is very difficult for policy-makers to select appropriate policy instruments and it is unlikely that environmental protection measures will attract the political and community support that is necessary to ensure their success.

The advantages and disadvantages of the four approaches are summarized in the Table 1.

International Dimensions of Environmental Protection

Under international law, states have the sovereign right to exploit, manage, and conserve the natural resources and natural systems within their jurisdiction, including resources located in their territorial sea and exclusive economic zone, and sinks such as the atmosphere. States also have a broad right to engage in fishing on the high seas. However, the expansion of the world economy has placed increasing pressure on natural systems that overlap or transcend political boundaries. This has gradually led to the development of a large number of international agreements, systems, and processes to address transnational environmental issues.

The international governance system that has emerged over the past 60 years to facilitate environmental protection can be divided into three main parts: international environmental law, international environmental bureaucracy, and international environmental financial mechanisms.

Table 1 Domestic environmental policy instruments—pros and cons

Policy instrument	*Definition*	*Advantages*	*Disadvantages*
Regulatory	Instruments that impose legally enforceable restrictions on economic agents to realize environmental protection objectives	Certainty about environmental outcomes Ability to limit free-riding Clarity of standards—easy to comply with, monitor, and enforce No need to negotiate individual standards, which lowers administration costs	Inefficiencies—regulations impede the operation of market forces and can lead to the misallocation of resources Can stifle innovation Potentially inequitable because some may shoulder a disproportionate burden
Economic	Instruments that force economic agents to internalize all or part of the social costs associated with environmentally harmful activities and that rely on market forces to promote efficiency	By utilizing market forces they are able to achieve the desired outcomes in an efficient manner Promote innovation Provide flexibility as they often enable individuals to determine the best method of achieving the desired outcomes Ability to limit free-riding	Can lack certainty about environment outcomes Can be complex and have high administration costs Potentially inequitable
Voluntary	Any mechanism or program that aims to protect the environment where relevant economic agents are able to decide whether or not to participate	Noninterventionist, meaning they are likely to have high levels of acceptance by producers Low political costs for regulators High levels of flexibility	Lack certainty about environment outcomes Can have high administration costs Risk of free-riding Risk that producers will engage in gaming to extract excessive economic rents from regulators
Information and education	Instruments that aim to promote environmental protection by improving people's awareness and understanding of environment issues and building their capacity to respond to environmental threats	Noninterventionist Low political cost for regulators Flexibility Relatively low administration costs	Lack certainty about environment outcomes Risk of free-riding Inability to address main reasons for market failure

International Environmental Law

At the heart of the international environmental governance system lies the body of legal principles and agreements that collectively constitute international environmental law. Although international environmental agreements have existed for centuries, the number, scope, and complexity of these agreements has increased considerably since the 1940s. By the mid-2000s, there were more than 500 multilateral environment agreements (MEAs) in existence. Around 270 of these MEAs were broad international agreements, while the remainder had a regional focus and a relatively limited number of signatories. Not surprisingly, these agreements cover a wide range of topics, including climate change, biodiversity, transport and disposal of hazardous materials, and fisheries management.

One of the most basic principles of international environmental law is that, while states have sovereignty over the resources in their jurisdiction, no state has the right to use or permit the use of its territory in such a manner as to cause injury to the another's territory, person, or property. This principle, which is generally traced to the Trail Smelter Dispute that commenced in the 1920s between Canada and the United States, concerns environmental obligations between particular states (i.e., reciprocal obligations).

Growing awareness of the interrelated nature of natural systems and the scale of environmental problems in the later part of the twentieth century resulted in growing support for the notion that states should also have obligations to protect global commons and the interests of humanity, including future generations. This has led to the formation of a significant number of international agreements that promote the protection of a broader range of interests in the environment. The guiding principle regarding the rights of states to exploit natural resources is now viewed as incorporating a fundamental duty to protect global commons. For example, the Rio Declaration on Environment and Development (1992) provides that states have a "responsibility to ensure that activities within their jurisdiction or control do not cause damage to the environment of other States or of areas beyond the limits of national jurisdiction." There are also a number of international agreements, such as the United Nations Framework Convention on Climate Change (UNFCCC), that seek to protect certain aspects of the environment for "the benefit of present and future generations of humankind."

The notion of sustainable development has been a common element of MEAs since the late 1980s and 1990s. This has led to greater concern about intragenerational equity and encouraged the inclusion of the so-called principle of "common but differentiated responsibility" in MEAs. This principle has two parts. The first is that states have a shared responsibility for the protection of the environment, or relevant parts of it. The second part is that the extent to which each individual state is responsible for environmental protection should be determined with regard to the state's capacity to respond and historical contribution to the

relevant problem. One of the clearest articulations of the principle is found in the Rio Declaration on Environment and Development (1992), where it provides that

> States shall cooperate in a spirit of global partnership to conserve, protect and restore the health and integrity of the Earth's ecosystem. In view of the different contributions to global environmental degradation, States have common but differentiated responsibilities. The developed countries acknowledge the responsibility that they bear in the international pursuit to sustainable development in view of the pressures their societies place on the global environment and of the technologies and financial resources they command.

The principle of common but differentiated responsibility also features prominently in the UNFCCC (1992) and the Montreal Protocol on Substances that Deplete the Ozone Layer (Montreal Protocol) (1987). For example, Article 3(1) of the UNFCCC states that

> The Parties should protect the climate system for the benefit of present and future generations of humankind, on the basis of equity and in accordance with their common but differentiated responsibilities and respective capabilities.

International Environmental Bureaucracy

The second element of the international governance system is the collection of international bodies whose functions include the oversight of environmental issues. At the heart of these is the United Nations, which has established a collection of agencies who are responsible for developing policy and promoting better environmental stewardship. These include the United Nations Environment Programme (UNEP), United Nations Development Programme (UNDP), Commission on Sustainable Development (CSD), and the United Nations Educational, Scientific, and Cultural Organisation (UNESCO). The agencies associated with the United Nations are complemented by other international and regional bodies like the World Bank, Organisation for Economic Cooperation and Development (OECD), European Union, and the Organisation of American States (OAS) that play a role in the development and implementation of environment policy.

International Environmental Financial Mechanisms

The final element of the international governance system is the financial mechanisms that have developed to support the work of international environment agencies and to assist in the achievement of international environmental objectives. These range from the general (e.g., the financial arrangements governing the World Bank and United Nation) through to the more specific, which concentrate solely on environment issues. The Global Environment Fund (GEF), established by the OECD in the early 1990s, is one example. Consistent with the principle of common but differentiated responsibility, many MEAs now include provisions requiring developed countries to transfer technology and financial resources to developing countries to assist them to meet their environmental obligations. One of the earliest examples was the Multilateral Fund established under the Montreal Protocol to assist developing countries phase out the use of ozone-depleting substances. In some cases, including under the Montreal Protocol, the obligation of developing countries to comply with the terms of the agreement has been made contingent on the extent to which developed countries provide the specified financial and technical assistance. For example, Article 4(7) of the UNFCCC states that

> The extent to which developing country Parties will effectively implement their commitments under the Convention will depend on the effective implementation by developed country Parties of their commitments under the Convention related to financial resources and transfer of technology . . .

Despite considerable progress being made in some areas, the international environmental governance system has generally failed to bring about substantial and sustained changes in the stewardship of natural resources and environmental systems. In many cases, this is due to the difficulties associated with designing agreements and systems that can accommodate the divergent interests of the states that are involved in transnational environmental issues. Negotiations can be tediously slow and the need to reach consensus can lead to lowest common denominator outcomes. Similarly, due to the reluctance of states to place constraints on their sovereign rights over natural resources, it has been difficult to establish appropriate mechanisms for monitoring compliance and enforcing the terms of the agreements. The international governance system has also been hampered by a lack of resources for key institutions and programs.

Conclusion

Environmental protection has always been practiced by humans in one form or another. However, as anthropogenic pressures on the environment have escalated over the past century, the need for systematic environmental protection has increased. This has led to considerable experimentation with the domestic and international measures that are used to achieve environmental protection objectives. Some of these have been successful, but the overall picture is one of failure.

Due to the failings of the past and greater awareness of the complexity of environmental problems, there is a growing acceptance that environmental protection is best achieved through the use of a multipronged approach. This requires the use of a combination of regulatory, economic, voluntary, and information instruments, where the policy mix is determined on the basis of the available evidence regarding cost-effectiveness.

The international challenge lies in the development of effective and equitable approaches to global environmental problems that are supported by a well-resourced bureaucracy and appropriate financial mechanisms.

However, in the last few years, a new initiative has come out bringing together recent scientific advancements in our knowledge of the Earth System functioning, environmental and ecological law as well as economic instruments to propose a new global governance system aimed at favoring environmental protection. In 2009 a group of scientists developed the "Planetary Boundaries" framework to measure and monitor the state and functioning of the Earth System (Rockström et al. 2009; Steffen et al. 2015). The framework, has been developed to define a planetary safe operating space within which humanity can survive and thrive, and to highlight risks that destabilization of the system creates for human well-being. The identification of such safe operating space for humanity has then helped to recognize the need of an evolution of the current international legal system. In this light the concept of "The Common Home of Humankind" has been proposed as a social construct, based on legal solutions to represent the global natural reality (Magalhães et al., 2016). The legal recognition of a specific, functioning state of the Earth System (i.e., a Holocene-like state, as defined by the planetary boundaries framework) as a common natural intangible heritage of mankind would allow to include all positive and negative externalities in the governance and maintenance of the Earth System thus possibly constituting a new legal environmental protection framework.

That legal systems are not ready to face these challenges is shown also by the fact that a group of law experts have launched the Ecological Law & Governance Association (ELGA) to reframe law according to ecological limits. In their *Oslo Manifesto* it is reported that "Ecological Law requires human activities and aspirations to be determined by the need to protect the integrity of ecological systems. Ecological integrity becomes a precondition for human aspirations and a fundamental principle of law. In other words, ecological law reverses the principle of human dominance over nature, which the current iteration of environmental law tends to reinforce, to a principle of human responsibility for nature. This reversed logic is arguably the key challenge of the Anthropocene".

Further Reading

Birnie P and Boyle A (2002) *International law and the environment*, 2nd ed. Oxford, UK: Oxford University Press.

Gunningham N and Sinclair D (1999) Regulatory pluralism: Designing policy mixes for environmental protection. *Law and Policy* 21: 49–76.

Karamanos P (2001) Voluntary environmental agreements: Evolution and definition of a new environmental policy approach. *Journal of Environmental Planning and Management* 44: 67–84.

Keohane N, Revesz R, and Stavins R (1998) The choice of regulatory instruments in environmental policy. *Harvard Environmental Law Review* 22: 313–367.

Kolstad C (2004) *Environmental economics*. Oxford, UK: Oxford University Press.

Magalhães P, Steffen W, Bosselmann K, Aragão A, and Soromenho-Marques V (eds.) (2016) *SOS treaty—The safe operating space treaty a new approach to managing our use of the earth system*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Organisation for Economic Cooperation and Development (2003) *Voluntary approaches for environmental policy: Effectiveness, efficiency and usage in policy mixes*. OECD Paris: France2003.

Panayotou T (1994) Economic Instruments for Environmental Management and Sustainable Development, Environmental Economics Series, Paper No. 16. Nairobi: United Nations Environment Programme.

Rockström J, Steffen W, Noone K, Persson Å, Chapin FS III, Lambin EF, Lenton TM, Scheffer M, Folke C, Schellnhuber HJ, Nykvist B, de Wit CA, Hughes T, van der Leeuw S, Rodhe H, Sörlin S, Snyder PK, Costanza R, Svedin U, Falkenmark M, Karlberg L, Corell RW, Fabry VJ, Hansen J, Walker B, Liverman D, Richardson K, Crutzen P, and Foley JA (2009) A safe operating space for humanity. *Nature* 461: 472–475.

Steffen W, Richardson K, Rockström J, Cornell SE, Fetzer I, Bennett EM, Biggs R, Carpenter SR, de Vries W, de Wit CA, Folke C, Gerten D, Heinke J, Mace GM, Persson LM, Ramanathan V, Reyers B, and Sörlin S (2015) Planetary boundaries: Guiding human development on a changing planet. *Science* 347: 1–15. https://doi.org/10.1126/science.1259855.

Sunstein CR (1990) Paradoxes of the regulatory state. *The University of Chicago Law Review* 57: 407–441.

United Nations Development Programme, United Nations Environment Programme, World Bank and World Resources Institute, 2003. World Resources 2002–2004. World Resources Institute: Washington, DC.

United Nations Environment Programme (2004) *The use of economic instruments in environmental policy: Opportunities and challenges*. Geneva, Switzerland: United Nations.

Relevant Website

https://www.elga.world.

Ecological Systems Thinking☆

David W Orr, Oberlin College, Oberlin, OH, United States
Valentina Niccolucci and Simone Bastianoni, University of Siena, Siena, Italy

Introduction

The greatest discovery of the past century had nothing to do with nuclear physics, or computer science, or genetic engineering. Rather it was the discovery of the essential connectedness of life and environment. The primary discipline of interrelatedness is ecology beginning with the work of Ernst Haeckel in the nineteenth century. The discovery of evolution extended the awareness of our connections to life in time and more extensively to the story of life on Earth. Fields such as ecology, general systems theory, systems dynamics, operations research, and chaos theory added details and theoretical depth, but with each advance in the precision and extent of knowledge the larger story remained the same. Living systems are linked in food webs and ecological processes into larger systems whether called the noosphere, biosphere, ecosphere, or Gaia. The boundaries between life forms and between what we take to be living and nonliving things shift and sometimes morph into other forms and processes. In Earth systems, small changes can have large effects somewhere else and at some later time. Natural systems and the world made by humans are intertwined in more ways than we can possibly imagine. The result is less like a machine than it is like a web stretching across all life forms and back through time. The effects of human actions millennia ago still ripple forward, intersect with other changes sometimes amplifying, sometimes diminishing in intensity. Some human-wrought changes, such as deforestation and saline soils throughout much of the Middle East, are permanent as we measure time.

Nothing in the preceding paragraph is particularly new or controversial. But the idea of interrelatedness has yet to take hold of us in a deep way. We still live in thrall to a world created by Descartes, Bacon, Galileo, and their heirs who taught us to dissect, divide, parse, and analyze by reduction but not how to put things back together or see the world as systems and patterns. The results were intellectual power without perspective so that, in time, overspecialization became a kind of a cultural disease. There are many reasons why things do not change long after their deficiencies are apparent: the inertia of habit, economic inconvenience, the preservation of reputation, and intellectual laziness. But the most important barrier to change remains simply that science and the technology it spawned works and is a powerful presence in our daily lives. Automobiles, airplanes, the cornucopia evident in every supermarket, miracle cures, and the wonders of computers and communications are a constant reminder of the powers of a particular kind of science and a promise of things to come. That much of our technology also "bites back" and incurs costs that we do not see is mostly lost on us. Many live in what has been called a "consensus trance," believing that things will go well for us, which is to say that progress will continue indefinitely. Beneath such ideas is the faith that nature does not "set traps for unwary species," as biologist Robert Sinsheimer once put it or that progress itself is not a self-made trap.

There have always been skeptics, however. Toward the end of his life, H. G. Wells could see no grounds for hope. More recently, Joseph Tainter, Martin Rees, and Jared Diamond have expressed doubts about our longevity based in no small part on their views of scientific progress. Rees, for example, believes that our odds of making it to the year 2100 are no better than fifty-fifty. Diamond has cataloged the reasons why past societies have collapsed and they bear more than a passing resemblance to our present behavior. James Lovelock, coauthor of the Gaia hypothesis, believes that we are approaching a climate-tipping point somewhere between 400 and 500 ppm CO_2 in the atmosphere after which "nothing the nations of the world do will alter the outcome and the Earth will more irreversibly to a new hot state." In various ways, each of these attributes our vulnerability to the failure to see systems, patterns, and to exercise foresight. As a result, we stumble toward a time of severe climate destabilization, biotic impoverishment, and ecological surprises.

The failure of ecological knowledge to penetrate very deeply into the larger society and its decision-making systems ought to be a matter of grave concern. The early work of ecologists Howard and Eugene Odum on the productivity of salt marshes, for example, may have slowed but certainly did not stop the juggernaut of development that has severely damaged coastal ecosystems virtually everywhere. Similarly, we know a great deal about the services of natural systems and the impossibility of duplicating these by human means. Yet the drawdown of natural capital and the destruction of ecosystems are still trumped by narrow short-term

☆*Change History*: March 2018. S Bastianoni, editor in chief of this section of the Encyclopedia of Ecology, has updated the "Ecological Systems Thinking" chapter with the coauthor V. Niccolucci. Section "Applied System Thinking" has been integrate to better emphasize the importance of Systems thinking within sustainability science and the methods or indicators used to appraise it. Section "Environmental Education" has been updated with a paragraph dedicated to the central role played by Education within the SDGs framework proposed by UN in 2015. Edited entries were also added in Further Reading.

concerns of profit and economic expansion. Sometimes the costs of ecological folly become starkly apparent as they did following hurricane Katrina in the fall of 2005 in which the damage done by a class III hurricane (at landfall) was amplified by the removal of mangroves and coastal forests that would otherwise have absorbed much of its energy and dampened the destructive effects. That, too, was known in many circles but did not have much effect on the policies that prevailed along the Gulf Coast, where oil extraction, commerce, and gambling ruled the day.

Public attitudes toward science are often undermined by poor education, inadequate public funding, and, sometimes, religious dogma. In the United States, evolution, once thought to be an established part of science, is hotly contested as just another "theory" by advocates of "intelligent design." The scientific evidence about human-driven climate change is indisputable, but ignored or underestimated even when alternatives are economically advantageous. The results are evident in the considerable data describing ecological deterioration virtually everywhere and the failure to seize better alternatives as well. Law based on ecological knowledge and the hope that we might calibrate our public business with the way the world works as a physical system is under constant assault. Evidence about the health and ecological effects of toxins is downplayed. Public access to information about the release of toxics is restricted. The result is a significant gap between what is known about how the world works as a physical system and the public policy in every country. The cumulative result is that we are much more vulnerable to ecological ruin and extreme events than we might otherwise be.

What can be done with ecological knowledge? One answer is that ecology as a science ought to do what it has been doing, which is to say document the deterioration of ecosystems in ever finer detail. Ecology, the argument goes, is a science and its practitioners ought to maintain their credibility as scientists and not assume the role of advocates and risk losing their credibility even when they recognize folly disguised as public policy. If that is the future of the discipline it will, I think, flourish for a time while the human prospect withers.

There is, however, another perspective on the uses of ecology. Paul Sears in 1964 and later Paul Shepard and Daniel McKinley in 1969 once called the discipline "the subversive science." They proposed ecology as an integrative discipline, "a kind of vision across boundaries" and a "resistance movement"—an alternative to being "man fanatic." Ecology in their view "offers an essential factor . . . to all our engineering and social planning." In their perspective, the world needs to know what ecologists know and needs to take that knowledge seriously enough to transform the ways by which we provision ourselves with food, energy, materials, shelter, and livelihood. Ecology as a subversive science would be integrated with building, industry, agriculture, landscape management, economics, and governance. In short, the idea of interrelatedness would move from the pages of obscure scientific journals out to the main street, and into board rooms, editorial offices, courtrooms, legislatures, and classrooms. It would progress from being just one more interesting but obsolete idea to become the design principles for a better world—the default setting for everyday behavior.

Applied Systems Thinking

In this regard, the news is guardedly optimistic. The art and science of high-performance building is growing. The result is a new generation of buildings that require a fraction of the energy of conventional buildings, use materials screened for environmental effects, minimize water consumption, and are landscaped to promote biological diversity, moderate microclimates, and grow foods. The best of these are highly efficient, powered substantially by sunlight and feature daylight, water recycling, and interior green spaces. They are a finer calibration between our five senses and the built environment and tend to promote higher user satisfaction and productivity. The costs of building green, as it turns out, are not necessarily higher than conventional buildings while having lower operating costs. The goal is to design buildings as whole systems, not as disjointed components. The green building movement is now a worldwide movement and is transforming the practice of architecture, landscape architecture, and engineering. It could, in time, transform the design of communities and cities as well.

Business, too, is beginning to go green. The best example of a well-run environmentally sensitive business is that of Interface, Inc., a global manufacturer of carpet tiles and raised flooring. In the mid-1990s, company founder and CEO, Ray Anderson, decided to transform the company to eliminate waste and carbon emissions. Interface launched a pioneering effort to develop carpet products that were returned to the company as a "product of service" not otherwise discarded in a landfill. Interface now leases carpet to its customers and takes it back to be remade into new products, thereby eliminating much of the petrochemical sources at one end and waste at the other. In the past decade, the company has eliminated 56% of its carbon emissions and is on track to becoming carbon neutral. The model for the company is consciously that of ecology all the way down to carpet products that mimic a forest floor. Interface is not alone. Other companies like Wal-Mart and DuPont are beginning to transform themselves as well. Some day, perhaps, all business will be powered by sunlight with materials cycles that mimic the circular flow of nutrients in ecosystems.

In agriculture, Wes Jackson, cofounder of the Land Institute, is pioneering the development of natural systems agriculture. The goal is to model agriculture on ecological systems such as forests and prairies. If successful, the end product will be agricultural polycultures of high yield perennials, long thought to be a biological impossibility. The early results, however, have confirmed Jackson's hypothesis that the two can be stitched together, thereby eliminating a great deal of fossil energy and soil erosion.

Materials science is a fourth area in which ecology is being taken seriously. Nature, as chemist Terry Collins has noted, uses only a relatively few ingredients while industrial chemistry uses virtually the entire periodic table, creating ecological havoc. The field of biomimicry has grown in response by studying how nature works in fine detail. Natural systems are a carnival of color, for instance, but nature does not use paints. To answer such questions, Janine Benyus, author of *Biomimicry*, is developing a database of the ways

nature works to filter, reduce, recycle, color, purify, form, and join—all done without the use of toxics and fossil fuels and all of it biodegradable. The result could be a transformation of materials and industry that dramatically reduce pollution and energy use.

Systems thinking is also fundamental to provide useful support in addressing the complex challenges that sustainability science needs to solve. With the introduction of this concept, many interesting holistic methods or indicators have been proposed to appraise the human dependence from Nature and its resources, and to support policy making decisions. Among others, there are some methods that are widely used for the simplicity and the immediacy of their message, that is, ecological footprint accounting, emergy evaluation, material flow accounting, and (partially) life cycle assessment (LCA), to make some examples.

All these methods have different perspectives, rationales, pro and cons, but a common *modus operandi*, that is, all inputs needed for the functioning of the analyzed system are converted into a common unit. The unit is the element of diversity among the methods: it is expressed in term of solar emergy joules needed, directly and indirectly, to produce and input or product (in emergy evaluation), of biologically productive land needed to make available life supporting resources or to absorb wastes (ecological footprint), etc. In this way, flows of different kinds can be accounted for, at the same time, in order to make a picture of the system according to systems thinking.

In these examples and elsewhere, the science of applied ecology has begun to seriously influence decisions and behavior and the evolution of architecture, engineering, materials science, agronomy, urban planning, and economics. The driving force is partly economic (to reduce the costs of unnecessary energy, materials and water use) and partly a matter of conviction (that it is wrong to leave a legacy of ruin behind us). While promising, such measures are necessary but insufficient. Ecological thinking, in one way or another, must become a more central part of global society and this is the task of education.

Environmental Education

The idea of specifically environmental education entered the public discourse in the late 1960s. Among the recommendations of the Stockholm Conference in 1972 was to "establish an international program in environmental education." UNESCO and UNEP subsequently undertook to prepare curricular materials, establish priorities, develop pilot projects, and organize meetings. The result was a UN-sponsored Conference at Tbilisi, Georgia, in 1978 that produced a consensus statement including the words:

Environmental education . . . should constitute a comprehensive lifelong education . . . it should prepare the individual for life through an understanding of the major problems of the contemporary world, and the provision of skills and attributes needed to play a productive role toward improving life and protecting the environment with due regard given to ethical values. By adopting a holistic approach, rooted in a broad interdisciplinary base, it recreates an overall perspective which acknowledges the fact that natural environment and manmade environment are profoundly interdependent . . ..

The Tbilisi Conference produced 41 recommendations spanning the needs for environmental education between developed and less-developed countries. In the subsequent decades, initiatives, including those spawned by Agenda 21 **(see article 36)** and discussions about the Earth Charter, have advanced the discussion of environmental education into a major part of the dialog about the role of education relative to the human prospect. There is no serious discussion about the transition to sustainability launched by the Brundtland Report in 1987 that does not include changing the goals and methods of education. From Tbilisi (US Department of Health, Education, and Welfare, 1978) and Talloires (UNESCO, 1977), and subsequent international gatherings, a strong consensus about the importance of environment in higher education is clearly apparent.

Despite considerable progress, both conceptually and practically, there are serious differences about the goals and methods of environmental education that reflect and, in some ways, amplify larger disagreements about education. At the lowest level, there is a general consensus that the young ought to know something about how nature works as a physical system—the rudiments of biology and planetary science. There is considerably less agreement about how this should be incorporated into the standard curriculum or at what level. Most elementary schools include curricular components such as "project learning tree" or "Wet and Wild" that introduce children to what was once called natural history along with some field experience and practical outdoor skills. But the later inclusion of values or discussion about the causes of environmental ills has often been controversial, especially when it has led to questions about conventional economic or political wisdom.

In important respects, all education is environmental education, that is, by what is included or excluded students are taught that they are part of or apart from ecological systems. The standard, discipline-centric curriculum may have contributed to a mindset that helped to create environmental problems by separating subjects into boxes and conceptually by separating people from nature. As a result, graduates are often ignorant of ecological relationships or why they are worthy of consideration. Not surprisingly, the first response to proposals for environmental education attempted to accommodate environmental issues and ecology into formal education as a kind of add-on. More radical critics proposed that formal education ought to be reformed along ecological lines, raising another and no less contentious issues. From either perspective, environmental mismanagement and the larger discussion of sustainability raise questions about the meaning of human mastery over nature, or more accurately as C. S. Lewis once put it: what does it mean for some men to control other men through the mastery of some parts of nature? What is the core knowledge of the environment that ought to be standard in an educational curriculum? At the heart of such questions are important differences about what it means to be human, what part of that definition ought to remain inviolable, and about the manipulation of natural systems through technological means such as genetic engineering. Is the problem, in other words, one in education or one of education?

What can be said with certainty is that public schooling and higher education have been underachievers in the task of inculcating essential knowledge about the environment. Public opinion surveys show high levels of support for environmental quality but little

ecological knowledge. In the words of one typical survey, people have acquired a "substantial familiarity with environmental issues, but [have] a long way to go in developing a working environmental/energy knowledge." Much of what people know about the environment is derived from television in bits and pieces and not through direct experience with nature or through cultural transmission.

One particularly encouraging aspect is the development of environmental education in institutions of higher education. Stemming from innovations in the 1980s, a vibrant campus ecology movement has emerged in Europe, Australasia, and the United States, along with a wide discussion of sustainability of educational institutions. Beginning with the studies of college food, energy use, and pollution, the movement has grown in subsequent decades to a worldwide scale. Hundreds of colleges and universities globally have organized efforts to systematically reduce energy use, water consumption, and material flows. Campus sustainability and climate stability have come to the center of institutional planning, purchasing, and construction. Beginning in the late 1990s with the advent of means to promote and measure environmental performance of buildings, the construction of academic facilities is undergoing a rapid revolution. Green or high-performance building standards are increasingly regarded as necessary to reduce energy and maintenance costs as well as laboratories for research and education. Many of the problems of sustainability—ecological design, applications of solar energy, water purification, food production, ecological restoration, and landscape management—can be studied in buildings and adjacent landscapes at a scale that is both significant yet manageable. Given recent developments on many campuses, it is not inconceivable that educational institutions at all levels will one day become models of ecological design mirroring the larger solutions necessary to the transition to sustainability.

A decisive step forward in the awareness of the importance of education in sustainability has been done in 2015, when the state members of the United Nation approved the so-called sustainable development goals (SDGs), a set of 17 goals and 169 target, to support the country transition toward sustainable development, by 2030. The SDGs framework recognize a central relevance to Education and include it in most of the 17 SDGs. In particular, the SDG4 is specifically aimed to "Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all" and the target 4.7 entitled "Education for sustainable development and global citizenship" is dedicated to a sustainable and equitable education "*by 2030, ensure that all learners acquire the knowledge and skills needed to promote sustainable development, including, among others, through education for sustainable development and sustainable lifestyles, human rights, gender equality, promotion of a culture of peace and non -violence, global citizenship and appreciation of cultural diversity and of culture's contribution to sustainable development.*"

Education is specifically mentioned also in other targets: target 12.8 (*by 2030, ensure that people everywhere have the relevant information and awareness for sustainable development and lifestyles in harmony with nature*) and target 13.3 (*Improve education, awareness—raising and human and institutional capacity on climate change mitigation, adaptation, impact reduction and early warning*) are the most relevant among the others.

Summary

In the decades since the Stockholm Conference in 1972, environmental education has emerged as a significant component of education virtually everywhere in the world. It has, for the most part, flourished at all levels of education. There are magazines and journals such as *Sustainability in Higher Education*, professional associations, and regular conferences. It is not difficult to imagine all of this as the start of something like an ecological enlightenment emerging in the decades or centuries ahead. But no such thing is certain. If education is to be midwife to a deeper, broader, and sustainable transformation, it will have to surmount serious challenges.

References

UNESCO (1977) *Intergovernmental Conferenceon Environmental Education, Tbilisi (USSR)74–26 October 1977*. Final Report.

US Department of Health, Education, and Welfare Toward an Action Plan (1978) *A Report on the Tbilisi Conference on Environmental Education*. Washington, DC: US Government Printing Office.

Further Reading

Barlett P and Chase G (2004) *Sustainability on campus*. Cambridge, MA: MIT Press.

Benyus J (1998) *Biomimicry*. New York: William Morrow.

Bowers C (1993) *Education, Cultural Myths, and the Ecological Crisis*. Albany, NY: SUNY Press.

Bowers C (1995) *Educating for an ecologically sustainable culture*. Albany, NY: SUNY Press.

Corcoran P and Wals A (2004) *Higher education and the challenge of sustainability*. Dordrecht, The Netherlands: Kluwer Academic.

Coyle K (2005) *Environmental Literacy in America*. Washington, DC: The National Environmental Education & Training Foundation.

Creighton S (1998) *Greening the ivory tower*. Cambridge, MA: MIT Press.

de Chardin T (1965) *The phenomenon of man*. Harper Torchbooks: New York.

Fischetti M (2001) Drowning New Orleans Scientific American. In: *76–85 (October)*.

Kuhn T (1963) *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Lovelock J (2006) *The revenge of Gaia*. London: Penguin Books.

Lovelock J (1979) *Gaia: A new look at life on Earth*. New York: Oxford University Press.
Lovins A (2005) *Winning the oil endgame*. Snowmass, CO: Rocky Mountain Institute.
Meadows D (2008) *Thinking in systems—A primer*. UK: Earthscan.
Oakeshott M (1989) *The voice of liberal learning*. New Haven, CT: Yale University Press.
Orr D (1992) *Ecological Literacy*. Albany, NY: Suny Press.
Orr D (1994) *Earth in Mind*. Washington, DC: Island Press.
Orr D (2006) *Design on the edge*. Cambridge, MA: MIT Press.
O'Sullivan E (2005) *Millennium Ecosystem Assessment Report vols. 1–5*. Washington, DC: Island Press.
Rees M (2003) *Our final hour*. New York: Basic Books.
Sears P (1964) Ecology—A subversive subject. *Bioscience* 14(7): 11–13.
Shepard P and McKinley D (1969) *The subversive science*. Houghton Mifflin: Boston.
Sinsheimer R (1978) The presumptions of science. *Daedalus* 107: 23–36.
Sobel D (1996) *Beyond Ecophobia*. Great Barrington, MA: The Orion Society.
Steffen W, Sanderson A, Jäger J, et al. (2004) *Global change and the earth system*. Berlin: Springer.
Tenner E (1996) *Why things bite back: Technology and the revenge of unintended consequences*. Knopf: New York.
Union of Concerned Scientists World Scientists (1992) *Warning to Humankind*. Boston: Union of Concerned Scientists.
United Nation (UN) (2015) *Transforming our world: The 2030 agenda for sustainable development. General assembly. Seventieth session, A/RES/70/1*.
Vernadsky V (1998) *The biosphere*. New York: Springer.
Washburn J (2005) *University INC: The corporate corruption of higher education*. Basic Books: New York.
Wright R (2005) *A short history of progress*. New York: Carroll & Graf.
Wright T (2004) Evolution of sustainability declarations in higher education. In: Corcoran PB and Wals AEJ (eds.) *Higher education and the challenge of sustainability*, pp. 7–19. Dordrecht, The Netherlands: Kluwer Academic.

Ecosystem Services Evaluation

Luca Coscieme and Jane C Stout, Trinity College Dublin, Dublin 2, Ireland

Introduction

Ecosystem services are the outputs of nature which have benefits to people. These include provisioning services such as food and fresh water; regulating services such as climate control; and cultural services such as recreation and spiritual and educational values (Costanza et al., 1997; MA, 2005; TEEB, 2010). Ecosystem functions and biodiversity are part of the natural capital stock that yields ecosystem service flows into the future (Costanza and Daly, 1992; Maseyk et al., 2016). Nature cannot provide any benefits without the presence of people (human capital), their communities (social capital), and their built environments (built capital) (Costanza et al., 2014; Pulselli et al., 2015) (Fig. 1).

However, the contribution of natural capital to human well-being is often expressed purely in terms of human, social and built capital, whilst the contribution of many ecosystem services are largely neglected, poorly understood, and scarcely monitored. This biases cost-benefit analyses and trade-offs between positive and negative outcomes of policies and actions. Increasing well-being by expanding the role of the economy is a priori perceived as more valuable than increasing well-being by preserving or improving environmental quality and the flow of ecosystem services.

The fact that no, or little, value is associated with the provision of ecosystem services has resulted in conversion of ecosystems, habitat fragmentation, landscape alterations, and anthropization of the natural environments. It also means that we lack signals of increasing scarcity of ecosystem services, with consequent continuous depletion and overuse of natural resources. In more general terms, this gap in the way we evaluate factors that contribute to well-being fosters a worldview by which the conservation of biodiversity and functioning ecosystems constrains human development.

As a consequence, the loss of vital ecosystem services may produce short-term positive effects for the economy. For example, it has been suggested that a sharp decline in pollinator populations would actually produce immediate benefits for the economy because the pollination service will have to be replaced by, for instance, commercially produced managed pollinators or by hand-pollinating the crops (Allsopp et al., 2008). This will mean employment, economic growth and tax revenues. However, pollinator decline will be disastrous for the agricultural economy in the long-term (Gallai et al., 2009) and for the production of food crops essential for our health and well-being (Chaplin-Kramer et al., 2014; Smith et al., 2015; Garratt et al., 2014). Pollinators generally provide their services for free and on a massive scale, and their disappearance will influence the biosphere in unpredictable ways, not just in terms of crop production. This makes pollination an irreplaceable ecosystem service whose loss will have long-term negative outcomes that outweigh by far any positive short-term effect on economic growth.

An evaluation system that risks the production of such distorted signals is an evaluation system that is failing us. The fact that values are assigned to man-made contributions to well-being, while Nature's contributions are underestimated or ignored, explains why financial economists are the primary advisors of governments. Climate change and resource scarcity mean that it is necessary for natural/environmental scientists and environmental economists to have the same power in the decision-making process.

In this vein, an evaluation system for ecosystem services is needed for (1) informing policymakers on the scale of human activities in relation with the biosphere; (2) fairly distributing resources within this generation, among future generations, and between humans and the rest of Nature; (3) allocating in the most efficient way resources for maximizing sustainable human development and socio-ecological resilience (Liu et al., 2010).

Different Kinds of Values

A value can be defined as the degree by which an object contributes to reaching an objective. In the case of ecosystem services, values should reflect the contribution of the particular service to human well-being. A distinction can be made between "use" and "non-use" values. Use values derive from the direct or indirect use of ecosystem services, also including the option value, that is, the satisfaction from possible future uses. Non-use values reflect the intrinsic value of ecosystems and biodiversity. This refers, for example, to the satisfaction derived from knowing that a pristine environment, or a particular species, exists, and that other people and next generations will have the opportunity to enjoy it (existence and bequest value).

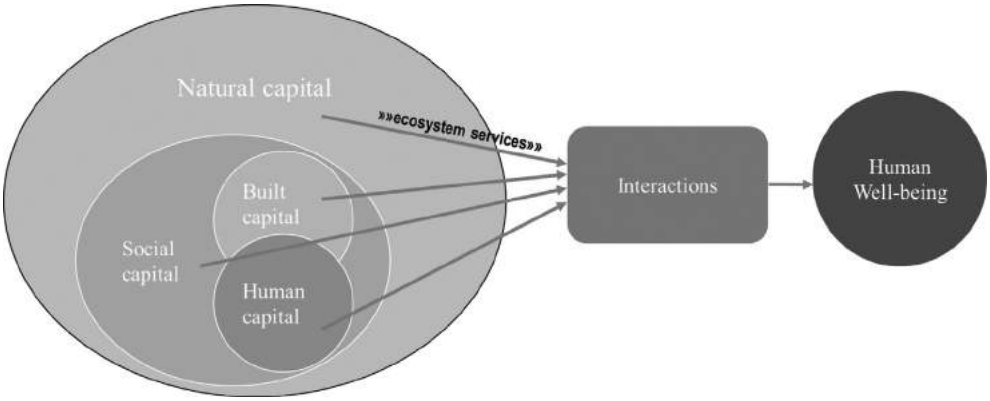


Fig. 1 Interactions between natural, social, built and human capital are required for human well-being. A relational order exists between the different forms of capital, with human and built capital (including the economy) embodied into social capital, and social capital embodied into natural capital. Adapted from Costanza, R. et al. (2014). Changes in the global value of ecosystem services. *Global Environmental Change* **26**, 152–158 and Turner, K.G., Anderson, S., Gonzales-Chang, M., Costanza, R., Courville, S., Dalgaard, T., Dominati, E., Kubiszewski, I., Ogilvy, S., Porfirio, L., Ratna, N., Sandhu, H., Sutton, P.C., Svenning, J-C., Turner, G.M., Varennes, Y-D., Voinov, A., Wratten, S. (2016). A review of methods, data, and models to assess changes in the value of ecosystem services from land degradation and restoration. *Ecological Modelling*, 319, 190–207.

Ecosystem Services Evaluation Methods

Broadly speaking, ecosystem services evaluation methods can be grouped into two main categories: (1) monetary valuation techniques, that require ecosystem services to be assessed in monetary terms; (2) biophysical accountings, that rely on more robust techniques and are less controversial, but at the same time are unsuitable for highlighting the importance of ecosystem services relative to, and in combination with, other contributors to human well-being.

Monetary valuation techniques of ecosystem services derive from environmental economics and aim at expressing in monetary terms ecosystem service values. These techniques can be divided into three categories: (1) direct market evaluations; (2) indirect market evaluations; (3) preference-based evaluations. Direct market evaluations are suitable to assess the monetary value of ecosystem services that already possess an explicit market. This is the case of a wide set of provisioning services (e.g., food, materials) that are sold and bought through conventional markets, but it is also the case of some cultural services that imply monetary transactions, such as the recreational service provided by protected areas and national parks that can be accessed by paying a fee. For these services, market prices are assumed to provide good estimations of the value people associate to the good or service provided by the ecosystem. However, market users might be faced with incorrect information through prices not embodying, for example, positive or negative externalities generated to include the ecosystem service in the market.

Indirect market evaluations aim at assessing the monetary value of ecosystem services that do not possess an explicit market. These evaluations rely on proxy measures of the contribution of the ecosystem service to human well-being, expressed in monetary terms. For example, coastal wetlands reduce the damaging effects of storms on coastal communities by absorbing the storm energy on a larger extent than solid land or open water (i.e., through increasing resistance on water motion, reducing direct wind effect on the water surface, absorbing wave energy and maintaining shallow water depths). This service can be directly relatable to the costs avoided by the presence of the coastal ecosystem that reduce the total damage to buildings and infrastructures during a storm. The monetary value associated with these avoided costs is intended to be a sufficiently accurate estimate of the contribution of the ecosystem service for human well-being. Coastal wetlands in the US were estimated to provide US $23.2 billions per year in storm protection services (Costanza et al., 2008). Similar assumptions are at the basis of indirect market evaluations of ecosystem services that rely on estimating the costs of artificially replacing ecosystem services. Preference-based methods rely on values expressed by stated preferences collected through surveying a sample of individuals/stakeholders, or by nonstated preferences revealed through observing people behaviour and choices (see De Groot et al., 2002; Christie et al., 2012 for an overview).

Ecosystem service values expressed in monetary terms can generate controversies when confused with prices (intentionally or unintentionally). The main critique of this evaluation approach is that by assigning a monetary value to services provided for free by Nature, there is a risk of commodification of ecosystems and biodiversity. Other critiques are based on assigning finite values to natural goods and services that are essential and irreplaceable, and thus of infinite value. Despite these critiques, monetary evaluation of ecosystem services is proving to be useful to accounting for the role of Nature in cost-benefit analysis, allowing the inclusion of the costs of environmental degradation and natural resource depletion in macroeconomic indicators. These methods are also effective in communicating the importance of environmental protection to policymakers and the business sector.

Biophysical accounting methods are evaluation techinques that do not express Nature's contribution to human well-being in monetary terms. For example, the value of provisioning services can be expressed in physical units, without relying on market prices. The value of a lake for fish production can be expressed in tons of fish per year, instead of using the market value of the fish. The carbon absorbed by a tropical forest can be expressed in terms of CO_2 equivalent, or the recreational value of a protected area in

number of visitors per year (Remme et al., 2014). Biophysical accounting is less controversial but produces results that are less impactful for policymakers, difficult to be aggregated in a total ecosystem value and to be compared with economic costs/benefits of alternative uses of the ecosystem.

Different evaluation methods are more suitable for evaluation of particular categories of ecosystem services. However, the choice of a specific method is largely dependent on the objectives of the evaluation and the specific audience and socio-economic context. This explains why the same ecosystem service type, provided by the same ecosystem, can be valued very differently case by case. For example, the value of coral reefs for tourism has been evaluated in a range going from tens to millions of US $ per hectare per year (*TEEB Summary: Responding to the value of nature*, 2009), depending on the geographical area, the kinds of tourist activities, and the environmental status of the reef, among other elements (Table 1).

Main Ecosystem Service Evaluation Initiatives

Ecosystem service evaluation can be performed at different spatial and temporal scales and refer to different categories of biomes, ecosystems, and socio-ecological systems. At the global scale, the earliest estimation of the overall ecosystem service values was performed by Robert Costanza and coauthors, and published in the journal *Nature* in Costanza et al. (1997). This analysis relied on a broad literature review, and some original calculations, to estimate the monetary value of the ecosystem services generated by our biosphere in a year. It was calculated that ecosystems at the global scale annually produce goods and services to a total value of around US $46 trillion, which is almost twice the total annual value produced by the global economy. This estimate has been updated in 2011 to US $125 trillion.

Ecosystem services evaluation gained broader attention after the publication of the Millennium Ecosystem Assessment (MA, 2005). Initiated by the United Nations in 2001, this global assessment provided a scientific evaluation of the conditions and trends of ecosystem services. The Millennium Ecosystem Assessment inspired several other international initiatives such as The

Table 1 List of the most common methods used to evaluate the different categories of ecosystem services

Approach	*Method*	*Ecosystem services*	*Rationale*
Monetary valuation	Market price	P, C	Actual market prices are used as a proxy for the good or service value
	Shadow price	All	Values are estimated as prices in hypothetical markets
	Production function	P	The indirect role of the ecosystem in the production of a marketable good is modeled and assessed
	Factor income	P, C	Some ecosystem services create or enhance incomes. The income (or income surplus) is considered as a proxy for the service value
	Hedonic pricing	All	The value of a good or service can be influenced by the quality of the environmental context where people benefit from it
	Travel cost	All	The money and time people spend to go enjoy an ecosystem service may serve as a proxy of its value
	Avoided cost	All	Ecosystem services allow society to avoid costs that would have been incurred in the absence of those services
	Replacement cost	P, R	The costs of artificial replacement of an ecosystem service are considered as a proxy of the service value
	Contingent valuation	All	People are asked about the value they attach to ecosystem services (through direct questions; hypothetical scenarios; trade-off choices between alternative services)
	Group valuation	All	A group of people is asked to discuss about the value they attach to ecosystem services, and provide a common value
	Benefit transfer	All	Ecosystem service values previously calculated are adapted to different contexts
Biophysical accounting	Changes in stocks/ quantities	P	Increasing/decreasing quantities of an ecosystem good indicate increasing/decreasing values
	Ecological integrity	All	The higher the ecological integrity, the higher the value of that ecosystem for that service
	Biodiversity indexes	All	The higher the biodiversity, the higher the service value of that ecosystem
	Ecological Footprint	P	The area and productivity of ecosystems within a territory (i.e., the Biocapacity) are used as proxies for the ecosystem service value of that territory
	Changes in efficiency	P, R	The more efficient an ecosystem is in producing goods and services, the higher its value
	Emergy	All	The more equivalent solar energy embodied in the ecosystem components, the higher the ecosystem service values
	Eco-exergy	All	The higher the level of genetic information and biomass in the ecosystem, the higher the ecosystem service values

P, Provisioning services; C, Cultural services; R, Regulating services.

Economics of Ecosystems and Biodiversity (TEEB, www.teebweb.org) and the ongoing Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES, www.ipbes.net). These initiatives are mainly based on state-of-the-art reviews and aim at highlighting knowledge gaps and promoting the message that a correct management and conservation of ecosystems is highly valuable (also in economic terms) for human well-being. The final reports produced are often addressed to policymakers and the business sector and the results are disseminated through impactful key messages shaped to reach the general public.

In some cases, ecosystem services evaluation have been implemented into policy strategies and business partnerships. The European Union Biodiversity Strategy for 2020 includes targets and actions specifically centred around the concept of ecosystem services and their evaluation. For example, Target 2 "Maintain and restore ecosystems and their services" dictates actions on mapping and assessing ecosystem service values in the member states' national territories, and ensure no net loss of ecosystem services. The Wealth Accounting and Valuation of Ecosystem Services (WAVES) project of the World Bank (https://www.wavespartnership.org/) aims to include the value of ecosystem services in development planning and national economic accounts. In 2008 and 2012 the World Resource Institute published The Corporate Ecosystem Services Review, that presents guidelines for including ecosystem services into business risks and opportunities management, recognizing the business sector's dependence and impact on ecosystems.

At country level, several nations produced National Ecosystem Services Assessments, such as the UK NEA. In Europe, Portugal, Spain, Norway, the Netherlands, Finland and Germany represent the most complete examples. Due to the very different approaches and methods used in these assessments, comparisons among them are not easy. A standarsization process of data collection, methods and indicators for ecosystem services evaluation is underway through intiatives such as the European Environmental Agency's Common International Classification of Ecosystem Services (CICES), or the EU Mapping and Assessment of Ecosystems and their Services (MAES).

Despite the diversity of methods used in ecosystem services evaluation, which ensures a plurality of perspectives and an interdisciplinary approach, standardization of evaluation and accounting methods is important for including ecosystem service values into alternative "beyond GDP" macroeconomic indicators. Through this operation, policymakers will be able to rely on development indicators that do not ignore the role of the environment for human well-being.

Thematic initiatives have been promoted to assess the value of specific ecosystem types, specific services, or with a focus on specific environmental elements such as soil, for example, or on particular species or, also, on the effects on ecosystem services of particular activities, such as deforestation or land degradation. The latter is the case of The Economics of Land Degradation (ELD) Initiative, a global study on the economic costs and benefits of land management practices, land degradation, and restoration.

Several initiatives, and research in general, have been recently focused on a particularly important ecosystem service, pollination, which is important for food production, but increasingly threatened by human activities, and thus a good example of the risks associated with ecosystem service loss. Due to the variety of methods used for assigning a value to pollination and pollinators, and due to the broad relevance of the status and trend of this service, we consider it a useful example of different approaches to ecosystem services evaluation.

Examples of Ecosystem Services Evaluation: Pollination Services

The majority (87.5%) of the world's flowering plants are pollinated by animals (Ollerton et al., 2011), including 87 of the leading 115 global food crops (Klein et al., 2007). Because of this, market values of food crops are often used to illustrate the economic value of pollination services. However, the total economic value of pollination services incorporates both market and non-market benefits (Hanley et al., 2015).

Market values of pollination service, in terms of the contribution to agricultural and horticultural production, have been quantified in a variety of ways. An early approach was to equate the total value of insect pollinated crops, or the cost of renting hives from beekeepers, with pollination services (Hanley et al., 2015). However, this approach is simplistic, as insect-pollinated crops are often able to produce some yield in the absence of pollinators (and thus the total crop value is an over-estimate of pollination value), hive-rental is usually not the only input to crop production, and in many places farmers do not pay to rent hives from local beekeepers (thus under-estimating pollination value) (Allsopp et al., 2008).

More recently, pollinator "dependence ratios" (DR) of crops have been used. A DR is calculated as the difference in yield in the presence and absence of pollinators (Klein et al., 2007) and thus provides a measure of crop loss in the absence of pollinators. It is usually measured in the field by comparing yield of plants protected from pollinator visitation (either grown in cages or covered with netting to exclude visitors) with plants open to natural levels of pollinator visitation. This ratio can then be multiplied by the area of crop planted and farm gate price of those crops in any given year, to give a figure representing the amount of farmer income attributable to pollination service. This method has been used to estimate the value of pollination services to crops for human consumption globally (e.g., Gallai et al., 2009), and regionally (e.g., Carreck and Williams, 1998; Losey and Vaughan, 2006). This enables identification of crops vulnerable to loss of pollination service and of areas where conservation of pollinators is critical, and is easily calculated using national statistics (Hanley et al., 2015). However, estimates vary temporally (year to year) and spatially (country to country), depending on crop types and market forces (Leonhardt et al., 2013).

In addition, there are several other problems with this approach. One is that data on DR are often taken from crops grown in different environmental conditions or of different varieties or cultivars and are not collected in a standardized way (Hanley et al., 2015). For example, yield may be expressed in terms of various functions of fruit or seed production from individual flowers or

whole plants. Another problem is that other outputs (such as crop quality) and inputs (such as agrochemical inputs, irrigation or managed pollination) are often not taken into account (Garratt et al., 2014), although some studies have incorporated crop quality (in terms of shelf-life) into estimates of pollinator value (Klatt et al., 2014). Thirdly, mean values do not incorporate the marginal variation in demands for pollinator dependent crops and how these change over space and time, that is, crop values are not "fixed" and vary according to consumer demand (Ricketts and Lonsdorf, 2013). This means that the assumption that loss of pollination will decrease crop value does not hold (Melathopoulos et al., 2015). In addition, these methods do not account for farmers switching to less pollinator dependent crops or varieties due to falling pollinator numbers.

Another approach to estimating the value of pollination services in agriculture is to estimate how much it would cost to pollinate crops using alternative methods. Replacement cost estimates have been based on pollen dusting and various methods of hand pollination by humans. An assessment of these methods, conducted on the Western Cape deciduous fruit industry of South Africa, valued pollination services higher than current market prices for commercial pollination (Allsopp et al., 2008). A recent study incorporated several methods to assess the value of pollination for a watermelon in the United States; subtracting the cost of inputs, valuing only the pollen deposition that is required for fruit production, and not excess pollen deposition, and attributing value to different pollinating taxa (Winfree et al., 2011). Obviously, detailed data are required for this approach. The most comprehensive approach so far involves modeling pollination efficiency, effectiveness, abundance and visitation rates of different taxa to derive a community production function (Hanley et al., 2015), but again, often the data required to make this complete assessment are not available.

It is predicted that in the future the demand for, and the area planted with, insect-pollinated crops (both field and covered crops) will increase (Aizen et al., 2009; Breeze et al., 2014). With changing climate, policy and consumer demands, insect pollinated crops could become increasingly important and so accurately valuing these services may become important in influencing agri-environmental decision-making to promote pollinator conservation. Furthermore, insect pollinated nonfood crops (e.g., fibers, fuel, animal fodder), as well as medicinal, cultural and/or symbolic plants, which may have market values, have not been subject to the same sort of analyses (but see Ollerton et al., 2016).

In addition to the short-falls in approaches used to value pollination services thus-far, few studies have attempted to consider the non-market values of pollination. To understand the real value of pollination services, non-market values need to be incorporated into assessments. Although a range of methods for capturing non-market values exist (e.g., contingent valuation such as willingness to pay and choice experiments, as outlined above), this has not yet been attempted for pollination services. In addition, studies so far have not addressed the value of services to maintenance of plant community composition, and how changes in plant community composition may affect the delivery of other ecosystem services, including supporting and regulating services, as well as cultural services.

Thus the valuation of pollination services is a good example of the overall approach to ecosystem services evaluation and illustrates that a joint effort among different disciplines is required. This enables a very broad variety of methods, some of which are already widely implemented, while some others need further improvements and applications. The need for a common framework, and for the standardization of the evaluation methods, is highlighted and partially fulfilled by initiatives such as the recent IPBES "Assessment report on pollinators, pollination and food production: summary for policymakers." Similar documents have been, and will be, produced regarding other types of ecosystem services.

Conclusions

Ecosystem service evaluation aims at considering the contribution of ecosystems to human well-being into cost-benefit analyses, macro-economic indicators and, finally, policies and decision making. A vast amount of literature testifies the efficacy of different evaluation methods, while a series of governmental and non-governmental intitatives testify the increasing interest in relying on ecosystem service evaluations to improve environmental management at different scales.

The many different evaluation methods developed so far represent the added value that a transdisciplinary approach can bring to theoretical and applied research. However, ecosystem services evaluation methods need to be further improved, in particular for including non-market values.

The criticisms, especially to monetary evaluation techniques, often address real issues such as the risks of commodification of the environment and of misusing the calculated values.

Finally, the rationale behind monetary evaluation of ecosystem services is to express the contribution of ecosystems to our well-being in terms comparable with those used to express the economy's contributions to well-being. It is in effect a communication strategy to reach policymakers.

The fact that ecosystem services and services from the economy are expressed through a common language does not mean that they automatically assume the same characteristics. Once a common language is defined, it is up to the speakers to choose what to say within the rules of that language. One of the most impactful uses of monetary ecosystem service values will be their inclusion into national accountings and well-being indicators, in order to produce a measure of development that considers the role of the environment (and society), instead of merely considering economic growth.

On the other hand, biophysical accounting methods do not produce results that are easily translated into policies, and this limit needs to be overcome.

References

Aizen MA, Garibaldi LA, Cunningham SA, and Klein AM (2009) How much does agriculture depend on pollinators? Lessons from long-term trends in crop production. *Annals of Botany* 103: 1579–1588.

Allsopp MH, de Lange WJ, and Veldtman R (2008) Valuing insect pollination services with cost of replacement. *PLoS One* 3(9): e3128.

Breeze TD, Vaissière BE, Bommarco R, Petanidou T, Seraphides N, Kozák L, Scheper J, Biesmeijer JC, Kleijn D, Gyldenkærne S, Moretti M, Holzschuh A, Steffan-Dewenter I, Stout JC, Pärtel M, Zobel M, and Potts SG (2014) Agricultural policies exacerbate honeybee pollination service supply-demand mismatches across Europe. *PLoS One* 9: e82996.

Carreck N and Williams I (1998) The economic value of bees in the UK. *Bee World* 79: 115–123.

Chaplin-Kramer R, Dombeck E, Gerber J, Knuth KA, Mueller ND, et al. (2014) Global malnutrition overlaps with pollinator-dependent micronutrient production. *Proceedings of the Royal Society B: Biological Sciences* 281(1794).

Christie M, Fazey I, Cooper T, Hyde T, and Kenter JO (2012) An evaluation of monetary and non-monetary techniques for assessing the importance of biodiversity and ecosystem services to people in countries with developing economies. *Ecological Economics* 83: 67–78.

Costanza R and Daly HE (1992) Natural capital and sustainable development. *Conservation Biology* 6: 37–46.

Costanza R, et al. (1997) The value of the world's ecosystem services and natural capital. *Nature* 387: 253–260.

Costanza R, Pérez-Maqueo O, Martinez ML, Sutton P, Anderson SJ, and Mulder K (2008) The value of coastal wetlands for hurricane protection. *Ambio* 37(4): 241–248.

Costanza R, et al. (2014) Changes in the global value of ecosystem services. *Global Environmental Change* 26: 152–158.

De Groot RS, Wilson MA, and Boumans RMJ (2002) A typology for the classification, description and valuation of ecosystem functions, goods and services. *Ecological Economics* 41: 393–408.

Gallai N, Salles J-M, Settele J, and Vaissière BE (2009) Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. *Ecological Economics* 68: 810–821.

Garratt MPD, Breeze TD, Jenner N, Polce C, Biesmeijer JC, and Potts SG (2014) Avoiding a bad apple: Insect pollination enhances fruit quality and economic value. *Agriculture, Ecosystems & Environment* 184: 34–40. https://doi.org/10.1016/j.agee.2013.10.032.

Hanley N, Breeze TD, Ellis C, and Goulson D (2015) Measuring the economic value of pollination services: Principles, evidence and knowledge gaps. *Ecosystem Services* 14: 124–132.

Klatt BK, Holzschuh A, Westphal C, Clough Y, Smit I, Pawelzik E, and Tscharntke T (2014) Bee pollination improves crop quality, shelf life and commercial value. *Proceedings of the Royal Society B: Biological Sciences*. 281.c.

Klein AM, Vaissière BE, Cane JH, Steffan-Dewenter I, Cunningham SA, Kremen C, and Tscharntke T (2007) Importance of pollinators in changing landscapes for world crops. *Proceedings of the Royal Society Series B - Biological Sciences* 274: 303–313.

Leonhardt SD, Gallai N, Garibaldi LA, Kuhlmann M, and Klein A-M (2013) Economic gain, stability of pollination and bee diversity decrease from southern to northern Europe. *Basic and Applied Ecology* 14: 461–471.

Liu S, Costanza R, Farber S, and Troy A (2010) Valuing ecosystem services. Theory, practice, and the need for a transdisciplinary synthesis. *Annals of the New York Academy of Sciences* 1185: 54–78.

Losey JE and Vaughan M (2006) The economic value of ecological services provided by insects. *Bioscience* 56: 311–323.

MA (2005) *Millennium Ecosystem Assessment*. Washington, DC: Island Press.

Maseyk FJF, Mackay AD, Possingham HP, Dominati EJ, and Buckley YM (2016) Managing natural capital stocks for the provision of ecosystem services. *Conservation Letters*. https://doi.org/10.1111/conl.12242.

Melathopoulos AP, Cutler GC, and Tyedmers P (2015) Where is the value in valuing pollination ecosystem services to agriculture? *Ecological Economics* 109: 59–70.

Ollerton J, Winfree R, and Tarrant S (2011) How many flowering plants are pollinated by animals? *Oikos* 120: 321–326.

Ollerton J, Rouquette J, and Breeze T (2016) Insect pollinators boost the market price of culturally important crops: Holly, mistletoe and the spirit of Christmas. *Journal of Pollination Ecology* 19(13): 93–97.

Pulselli FM, Coscieme L, Neri L, Regoli A, Sutton PC, Lemmi A, and Bastianoni S (2015) The world economy in a cube: A more rational structural representation of sustainability. *Global Environmental Change* 35: 41–51.

Remme RP, Schroter M, and Hein L (2014) Developing spatial biophysical accounting for multiple ecosystem services. *Ecosystem Services* 10: 6–18.

Ricketts TH and Lonsdorf E (2013) Mapping the margin: Comparing marginal values of tropical forest remnants for pollination services. *Ecological Applications* 23: 1113–1123.

Smith MR, Singh GM, Mozaffarian D, and Myers SS (2015) Effects of decreases of animal pollinators on human nutrition and global health: A modelling analysis. *The Lancet* 386(10007): 1964–1972.

TEEB Foundations (2010) *The economics of ecosystems and biodiversity: Ecological and economic foundations*. London and Washington: Earthscan.

TEEB Summary: Responding to the value of nature. (2009). London and Washington: Earthscan.

Winfree R, Gross BJ, and Kremen C (2011) Valuing pollination services to agriculture. *Ecological Economics* 71: 80–88.

Further Reading

Turner KG, Anderson S, Gonzales-Chang M, Costanza R, Courville S, Dalgaard T, Dominati E, Kubiszewski I, Ogilvy S, Porfirio L, Ratna N, Sandhu H, Sutton PC, Svenning J-C, Turner GM, Varennes Y-D, Voinov A, and Wratten S (2016) A review of methods, data, and models to assess changes in the value of ecosystem services from land degradation and restoration. *Ecological Modelling* 319: 190–207.

Relevant Websites

http://www.teebweb.org—"TEEB; The Economics of Ecosystems and Biodiversity".
http://www.millenniumassessment.org—"Millennium Ecosystem Assessment".
http://www.ipbes.net—"Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services".
http://theconversation.com/if-dollars-rule-the-world-why-dont-the-bees-get-a-bailout-38384—"The Conversation UK".
http://www.ecosystemvaluation.org/—"University of Maryland; Ecosystem Services Evaluation".

Emergy and Sustainability☆

Roberto Ridolfi, Federico M Pulselli, Fabiana Morandi, Mariana Oliveira, and Simone Bastianoni, University of Siena, Siena, Italy; State University of Campinas, São Paulo, Brazil

Introduction

Natural systems belong to an energy network in which transformation processes connect smaller scales and larger scales. In every transformation, energy remains constant while exergy (sometimes referred to as "available energy") at one level is consumed leaving a smaller amount at the next level.

It is recognized that different types of energy exist. Joules of energy of different kinds are not equivalent in their ability to do work. The energy that flows through all real processes is partially degraded. The second law of thermodynamics states that all real processes, including processing of energy and storage of materials, imply a dispersion of part of the energy in the form of heat. Nature is therefore organized in flows of energy of different qualities.

Odum recognized the implications of energy quality and introduced the concept of emergy (initially called "embodied energy") to quantify the available energy (exergy) of a given type directly and indirectly used to make a product. The type of energy chosen as reference was solar energy, since it is basically the source of all flows in the biosphere. The solar emergy (from now on simply emergy) of a product is, therefore, the solar energy (exergy) needed, directly and indirectly, to make that product.

A flow can be evaluated not only on the basis of the amount of exergy carried but also on the basis of the amount of solar energy directly and indirectly used to produce the flow. In this way, it is possible to differentiate flows of higher quality, that is, that need great quantities of energy or material for their production, from flows of low quality that need little energy or material.

Energy Language, Energy Modeling, and Hierarchical Web

To describe the flows of energy and matter in a system, a modeling language, called "energy system diagram," has been developed. Systems are made up of forces and energy pathways: the former are causal actions, the latter represent how and where these forces are directed. The symbols used in energy diagrams are summarized in Fig. 1.

The emergy concept was developed with hierarchical webs in mind (see Fig. 2). In these webs, quantity of energy associated with each transformation decreases at each step in the process. Every transformation is accompanied by production of "dispersed" heat.

In the energy diagram the productive units on the left produce goods and services for those on the right which return materials and control to the left. The different units reinforce each other so that the whole system self-organizes and maximizes power. Energy is transformed from left to right and in each transformation the output has less energy but of higher quality and controls other units of the system. To create quality on the right, a great amount of low-quality energy is necessary. The diagrams show an energy hierarchy with step-wise convergence from left to right. In a food web (see Fig. 2), the position in the chain represents different capacities for control and quality. For example, directly and indirectly, it takes about 1000 J of sunlight to make a joule of spatially dispersed organic matter, about 40,000 to make a joule of coal, and usually even more to make a joule of electrical energy. A joule of sunlight, a joule of coal, and a joule of human work have different qualities and different energy convergences to make them. According to this concept, Odum sustained that a more meaningful way to express quality is not to consider the energy content of a

☆*Change History*: March 2018. R. Ridolfi, F. M. Pulselli, F. Morandi and S. Bastianoni updated Emergy Algebra and further readings; added a new section "Emergy And Set Theory"; added new Fig. 4. The authors updated and added a new section "Emergy in an input-state-output framework."

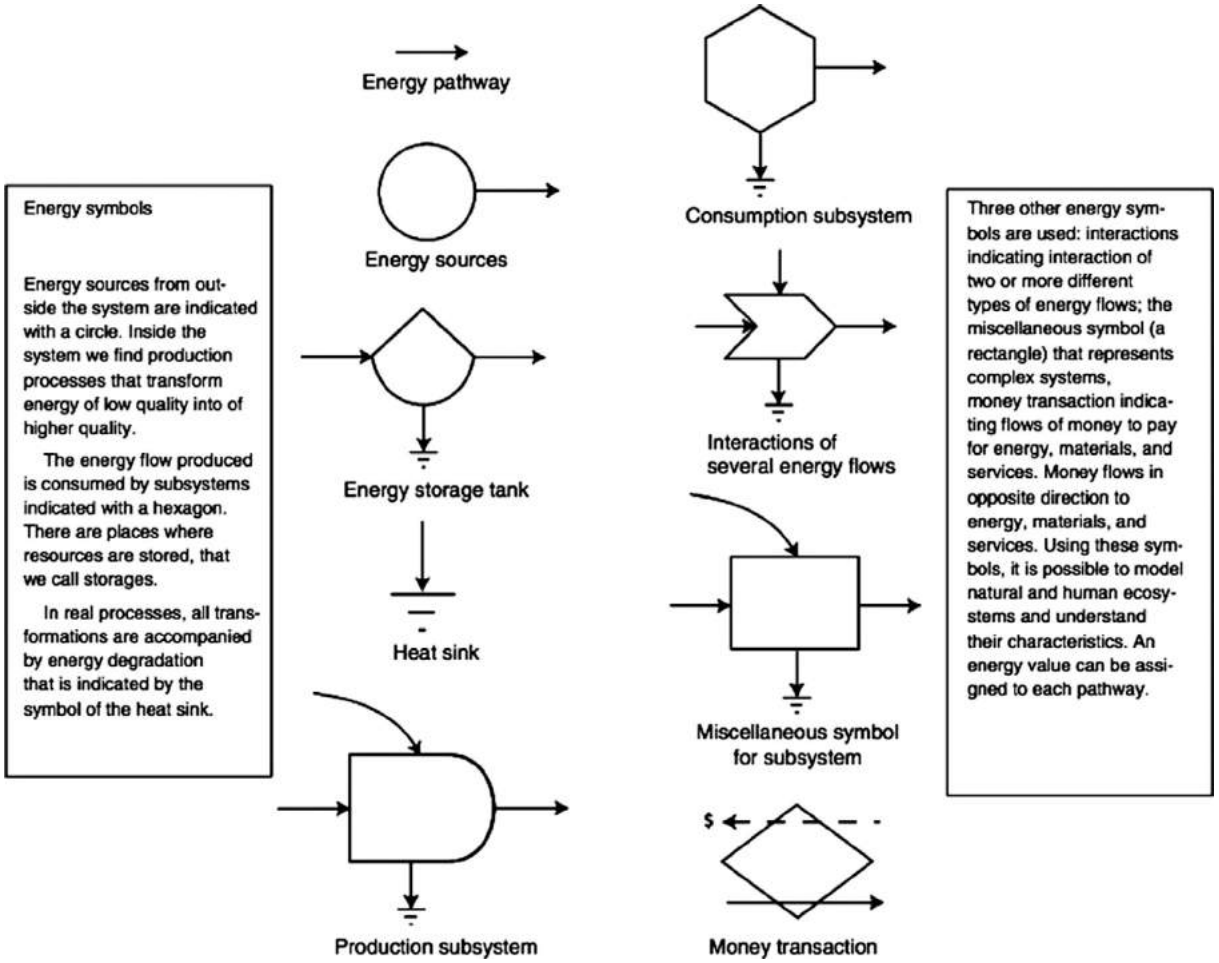


Fig. 1 Energy system symbols.

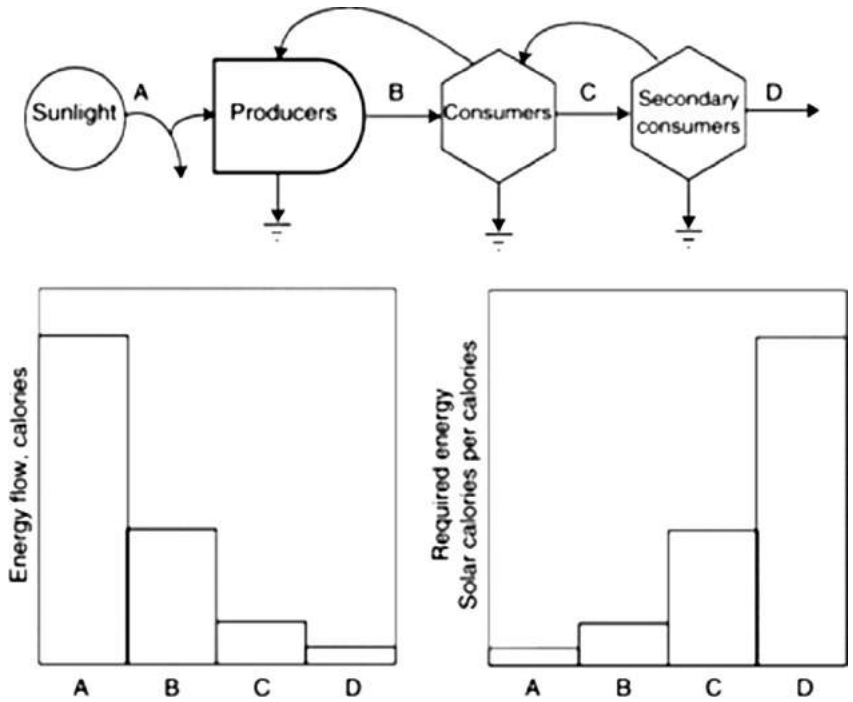


Fig. 2 Hierarchical webs and energy flows.

system but the energy embodied (emergy) in it, that is, how much energy was used to make or sustain the system starting from the lowest level of the web. For this reason, emergy is sometimes referred to as "energy memory."

Emergy

From its definition emergy is an extensive quantity; its unit is the solar emergy joule (sej) and its physical dimensions are those of energy (ML^2T^{-2}) of solar (or equivalent) type, where M = mass, L = length, T = time). It is not a state function, since it depends on the pathway that the process follows. In fact, the emergy of a product is related to the way it is produced. A joule of electricity generated from oil has a different emergy to one generated by wind power.

Emergy is a donor-referenced concept rather than a receiver-referenced one. It measures the source energies converging at system boundaries into processes or products. The total energy flowing through a system per unit time is its empower, with units sej/[time] and physical dimensions (ML^2T^{-3}) of solar type.

The basis of emergy evaluation is the conversion of all process inputs, including energy of different types and energy inherent in materials and services, into emergy by means of a conversion factor called transformity.

Transformity

Transformity is defined as the energy of one type directly and indirectly required to generate 1 J of another. Solar transformity is currently used in emergy evaluation. Unlike emergy, transformity is an intensive quantity, representing the inverse of classical energy efficiency and is measured in sej J^{-1}. It is dimensionless and system specific. To evaluate increased organization of concentrated matter, a mass-specific emergy (sej g^{-1}) is sometimes used instead of transformity. The emergy of a certain type of material is obtained by multiplying its mass with the emergy-to-mass ratio. As with energy-based transformities, matter evaluations are also system and process specific. In general we can call Unit Emergy Value (UEV) the ratio of emergy to any unit output.

Emergy and Transformity

As stated above energy decreases along transformation chains, therefore along a chain or a web, transformities increase and emergy, following "memorization" laws, may remain constant or grow down-chain. Fig. 2 is useful for evaluating emergy and transformity at different steps in an energy flow sequence. If the producers are forage plants and the consumers (C) are cows, the sun transfers energy to the plants from the boundary of the system, the plants use it by photosynthesis and transpiration, taking up soil nutrients and fertilizers. The transformity of the plants is clearly greater than 1 (i.e., the solar transformity of the sun) because the energy in the forage is obviously much less than that of the incoming solar energy, but the plant increases the organization of the sun–forage–cow system. Similarly, the forage transfers its solar-derived energy to the cow. The cow's transformity is much greater than 1 because of the cow's low energy with respect to the solar energy that has been used for it. Since the forage contains energy derived from the sun, so does the cow, and energy embodiment increases as energy dissipates along food chains.

As a general rule, the transformities of similar products are compared to obtain information about production efficiency. If, for instance, the transformity of forage from one field is 4×10^4 sej J^{-1} and that from another field is 1×10^5 sej J^{-1}, the forage from the first field can be said to be more efficient (less emergy per unit of product).

Transformities of different classes of product (e.g., forage and cow) can also be compared. In this case, transformities indicate the relative "position" in the global hierarchy of processes. All energy transformations can be represented in sequence and the position of each energy flow in the sequence is indicated by the transformity.

Energy system diagrams show energy hierarchy with small fast turnover units on the left and aggregations in space and time on the right (Fig. 3).

Emergy is the same from left to right (Fig. 3B and C), whereas available energy decreases after each transformation and the transformity increases. The higher the transformity, the greater the demand for energy to make that flow or product. So in order to overcome natural selection, systems reinforce their networks with feedbacks. Units receiving feedback from units further down the chain are reinforced by a small energy flow of high quality, that is, more concentrated and therefore more capable of doing work.

Emergy Algebra

As stated above, emergy analysis obeys a logic of memorization and therefore needs its own algebra, that is summarized in four main rules, which according to Brown are:

1. All source emergy to a process is assigned to the process's output (Em_k):

$$Em_k = \sum_i Tr_i E_i \tag{1}$$

where E_i is the energy of the ith component and

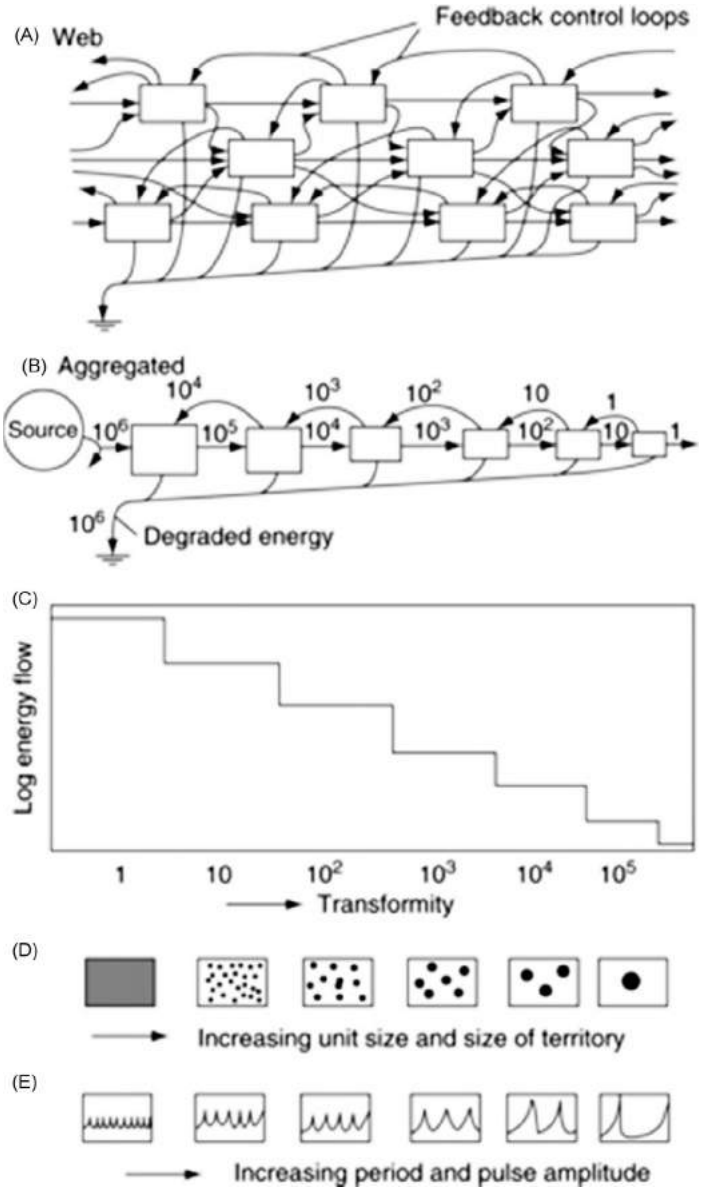


Fig. 3 Energy hierarchy in space and time. Reproduced with permission from Ridolfi, R., Pulselli, F. M., Morandi, F. and Bastianoni, S. (2013). Emergy. *Reference Module in Earth Systems and Environmental Sciences.* https://doi.org/10.1016/B978-0-12-409548-9.00590-X.

$$Tr_i = \frac{Em_i}{E_i} \tag{2}$$

2. By-products of a process have the total emergy assigned to each pathway.
3. When a pathway splits, the emergy assigned to each branch is based on its percentage of the total energy flow on the pathway (the two products are called splits).
4. Emergy cannot be counted twice in a system: (a) emergy in feedbacks is not counted twice; (b) when emergy of by-products is summed, it cannot be greater than the source emergy from which it was derived.

It follows that the transformities of two (or more) splits are identical, while their emergy contents are generally different (unless energy is distributed equally among the splits); on the contrary, the transformities of co-products are generally different (unless the same amount of energy went into the various co-products).

The first emergy rule indicates that it is necessary to know the emergy of the inputs in order to calculate the emergy of the output. Apart from the sun, that has a transformity of 1 by definition, it is necessary to calculate the transformity or emergy of natural resources. Emergy accounting starts with an evaluation of the Earth's emergy processes. Once the emergy of the main natural flows is calculated, it is possible to calculate the emergy of many resources, from minerals to fossil fuels. Over the years, depending on the

model used and on the combination of the main natural flows, other baselines have been developed but now an agreement has been reached: the reference baseline (that represents the reference for all the emergy calculations) has been updated to 12×10^{24}sej year^{-1}.

Emergy and Set Theory

A recent development has suggested that Emergy can also be defined as the set of the solar exergy that is directly and indirectly required to obtain a service or a product. In particular, emergy can be seen as the set of all the (equivalent) photons falling on certain portions of the biosphere in certain time intervals, that have been used directly and indirectly to make a product or service. In this way, because the emergy of a product (or service) is obtained through the convergence of several inputs, it can be seen as the union of the sets of solar exergy required to obtain each input, that is:

$$Em = \bigcup_i \{\text{direct and indirect solar exergies}\}_i \tag{3}$$

Eq. (1) is then corrected with:

$$Em = \bigcup_i Tr_i E_i \tag{4}$$

This formula is more general and correct than Eq. (1) because it is able to fully consider all the rules of emergy algebra. In fact, the union, with respect to the sum operation, does not require any restriction: if the emergy flows are independent, the union is replaced by the sum, otherwise only the independent parts (without intersections) should be added. In this way, all rules of emergy algebra can be encompassed in the property of union of sets because splits are represented by a partition of the out coming flow while by-products (that are associated to the operation of copy) and feedback are never counted twice when they interact in the process (Fig. 4).

Emergy, Sustainability, and Its Indicators

Unlike classical energy and economic analyses that only consider items that can be quantified in energy or money terms, thus omitting most free inputs from the environment, emergy analysis is a thermodynamic methodology which considers both the

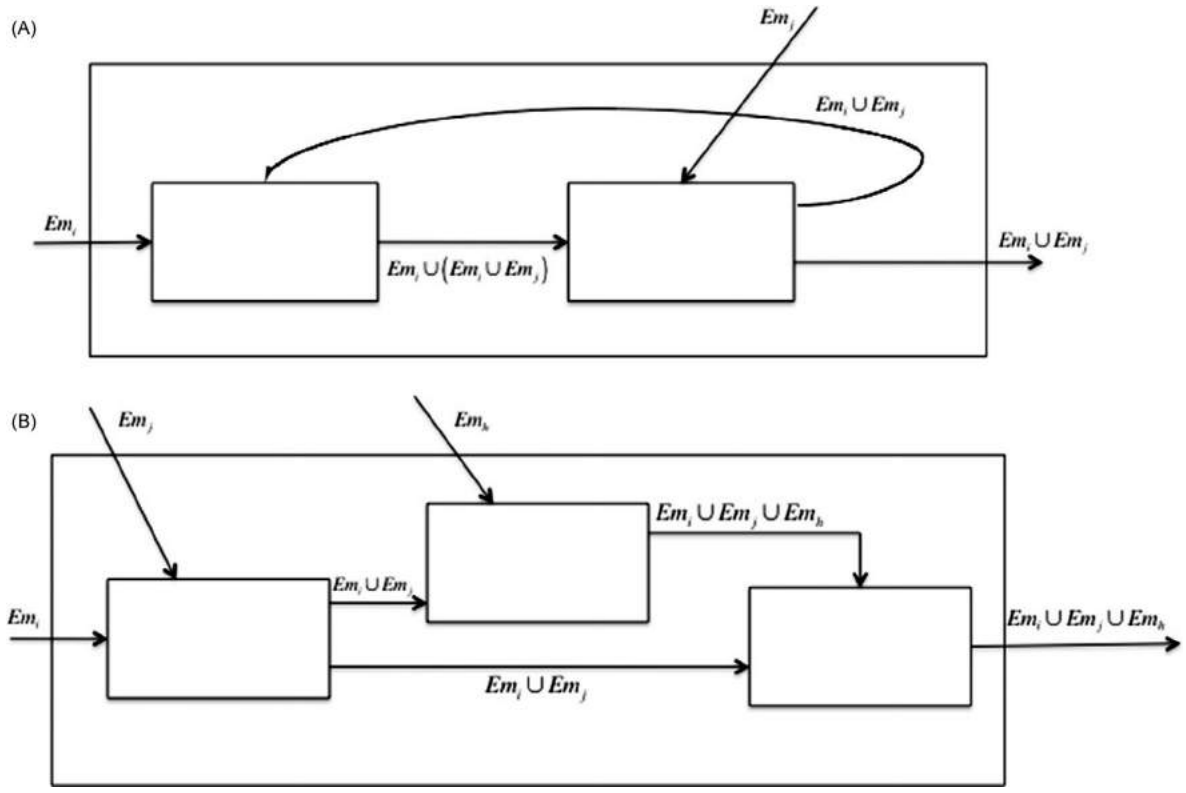


Fig. 4 Examples of (A) a process with a by-product that feeds back and (B) of a process where by-products are reunited.

economic and environmental aspects of a system by converting all inputs, flows, and outputs to the common denominator of solar energy, the basic energy behind all the processes of the biosphere.

This is a primary factor because, although the market only considers monetary value, the economy is also based on quantities from the environment, which must be considered and assigned a value, if resources are to be exploited sustainably in the long period.

Emergy analysis is useful to check applications of Herman Daly's first rule of sustainable development, the so-called sustainable yield principle, that states that resources should be exploited at a rate compatible with their replacement by nature. It can be used to define guidelines for consumption of resources compatible with their formation times.

Emergy can be regarded as the solar energy necessary to sustain a system; the greater the total emergy flow necessary for obtaining a product, the greater the consumption of solar energy necessary for its reformation once it has been used, and thus the greater the past and the present environmental cost to maintain it.

The intensive use of the services and products of an ecosystem can degrade its structures and functions, decreasing the capacity of the ecosystem to self-organize efficiently. In order to facilitate the measurement of a system's sustainability, some emergy indicators were introduced.

To explain them a simple model of a system is shown in Fig. 5. Emergy flows to the system are divided into the main categories as given in the following:

- local renewable resources (R);
- local nonrenewable resources (N);
- feedback (F): purchased resources and services from outside system; and
- the yield (Y): the output of the system ("virtual" in the case of territory).

These flows can be combined to obtain a set of indicators; here are most common ones.

1. Emergy yield ratio (EYR) is the ratio of the output of a system (Y) to the feedbacks from outside (F). Therefore, it is the ratio of total emergy to the nonrenewable, economic inputs. Considering that the total emergy is the sum of all local and external emergy inputs, the higher the ratio, the higher is the relative contribution of the local sources of emergy to the system. This index therefore shows the ability of a system to use the available local resources. Generally, EYR values less than 5 are typical of secondary energy sources and primary materials (e.g., cement and still), whereas values greater than 5 are shown by primary energy sources. In the case of processes that give products with EYR values less than 2, the processes are not considered an energy source but rather a consumer or a transformation process:

$$\mathrm{EYR} = \mathrm{Y}/\mathrm{F}$$

2. The environmental loading ratio (ELR) is the ratio of renewable to nonrenewable emergy use inside a system. It is a measure of the renewability of a production system or a system state. The higher the ratio, the lower the system's sustainability. A high ratio

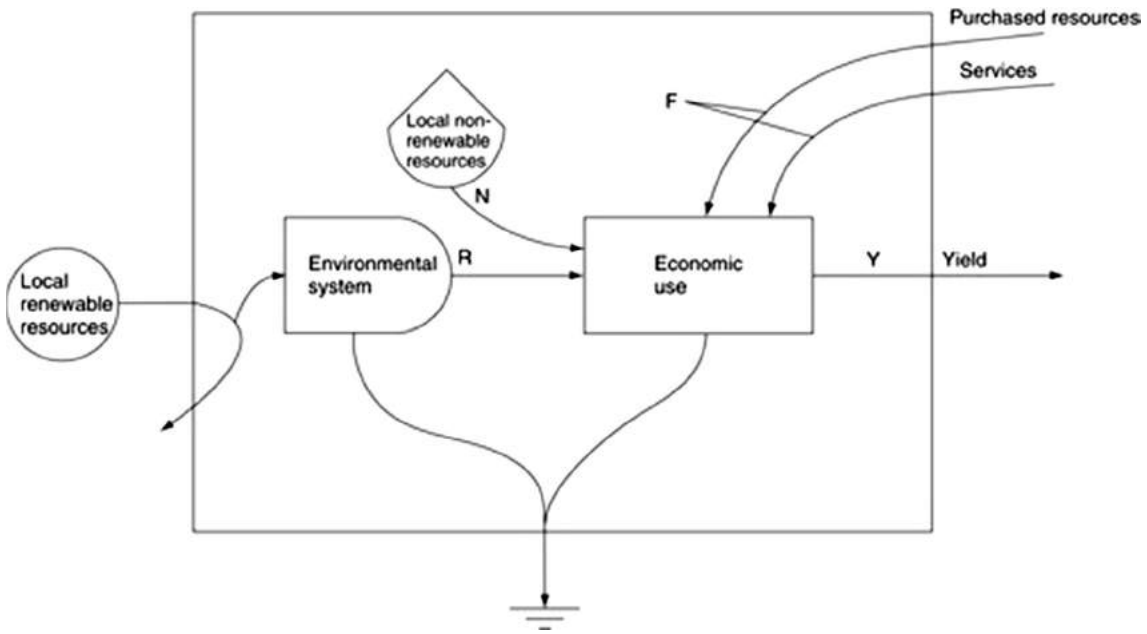


Fig. 5 Energy system diagram of a generic system.

suggests a high technological level of emergy use and/or a high level of environmental stress (either local or global). It sounds an alarm-bell of nonequilibrium which could become irreversible for a state or a production system:

$$\mathrm{ELR} = (\mathrm{F} + \mathrm{N})/\mathrm{R}$$

3. Emergy money ratio is the ratio of emergy use in a country (or a region) to its gross national product. It measures how much emergy is associated with the economic wealth of a state. The emergy money ratio provides a link between emergy evaluation and economics. The use of emergy with traditional economic evaluation methods can provide additional information to guide human activities, as in planning land use or designing production processes. This tool has a twofold function: (1) it provides information about systems, for example, comparison of national economies with different emergy flows and GDP, or evaluation of the trend of an economic system in time; (2) it can be used as a conversion factor for further emergy evaluations when monetary parameters must be converted into emergy and vice versa:

$$\text{Emergy to money ratio } (sej/\$ \text{ or } €) = \mathrm{Y}/GDP$$

where Y is the total emergy used in a territorial system in a year ($\mathrm{Y} = \mathrm{R} + \mathrm{N} + \mathrm{F}$) and the GDP is referred to the same year.

4. The emergy investment ratio (EIR) is the ratio of feedback to local resources (renewable and otherwise). It measures how much a system depends on the outside rather than on local resources, and how much a system or process uses invested emergy in comparison with alternatives:

$$\mathrm{EIR} = \mathrm{F}/(\mathrm{R} + \mathrm{N})$$

5. Emergy per person is the ratio of emergy used in a country divided to population. It measures how much emergy is available per person. This indicator suggests a measure of the standard of living in a country in terms of current availability of resources and goods, though it does not give any indication about the future availability of resources. Two different meanings are associated with emergy per person: in the case of a low ELR (low use of nonrenewable inputs), a high emergy per person ratio indicates resource availability, whereas when ELR is high (high use of nonrenewable inputs), emergy per person represents consumption and the system needs to invest in order to decrease emergy use and enhance renewable uses:

$$\text{Emergy flow per person } (sej/(\text{pers.year})) = \mathrm{Y}/\text{persons}$$

where Y is the total emergy used in a territorial system in a certain amount of time, typically 1 year ($\mathrm{Y} = \mathrm{R} + \mathrm{N} + \mathrm{F}$).

6. Empower density is the ratio of emergy flow into a system to area of the system. It measures the concentration of emergy in space and time. This indicator is a measure of spatial concentration of emergy flow within a process or system. It can be used to highlight areas under environmental pressure with respect to more natural areas. A high empower density can be found in processes in which emergy use is large with respect to the available area. This suggests that land is a limiting factor for the future economic growth. The emergy flow density can be used to plan environmental policy. Areas with high emergy densities have a concentration of emergy use and call for different policies than natural or agricultural areas. For example, cities and industrial districts should invest in increasing efficiency of energy use, transport and so on, whereas agriculture areas need more attention to soil management and fertilizer use:

$$\text{Empower density } \left(sej/\left(m^2 \text{ year}\right)\right) = \mathrm{Y}/\text{Area}$$

Emergy Applications

Ecosystems

Emergy evaluation is a method that works perfectly at the interface between human and natural systems. There have been many studies on this interaction; here we chose one that assessed the environmental impact of a fish farm. Aquaculture has many interactions with the surrounding environment using resources and producing changes in the ecological system.

The fish farm, in this application, is located in the Gulf of La Spezia (Ligurian Sea, NW Mediterranean Sea). The installation called Spezzina fish farm is a marine inshore fish farm located at Punta Pezzino; this area of Gulf of La Spezia is sheltered by the offshore breakwater and is thus characterized by a low circulation regime (Fig. 6). Reared fish are mostly gilthead sea bream (*Sparus aurata*) and sea bass (*Dicentrarchus labrax*) to a lesser extent. In the study only *S. aurata* production is considered in the emergy calculation.

The farm comprised of a total of 60 fish cages (for a total of 44,000 m^3) placed on floating wharves oriented from the coast toward the open sea (Fig. 6). The whole installation covers an area of 24,000 m^2 with a mean depth of about 12 m. Average production is reported to be 350 tons of fish per year.

Food supply takes place twice a day with an automatic system and the fish diet varies both in composition and quantity in relation to biological factors (i.e., age, biomass, and starving) and in relation to climatic and physical factors (i.e., water temperature, and dissolved oxygen).

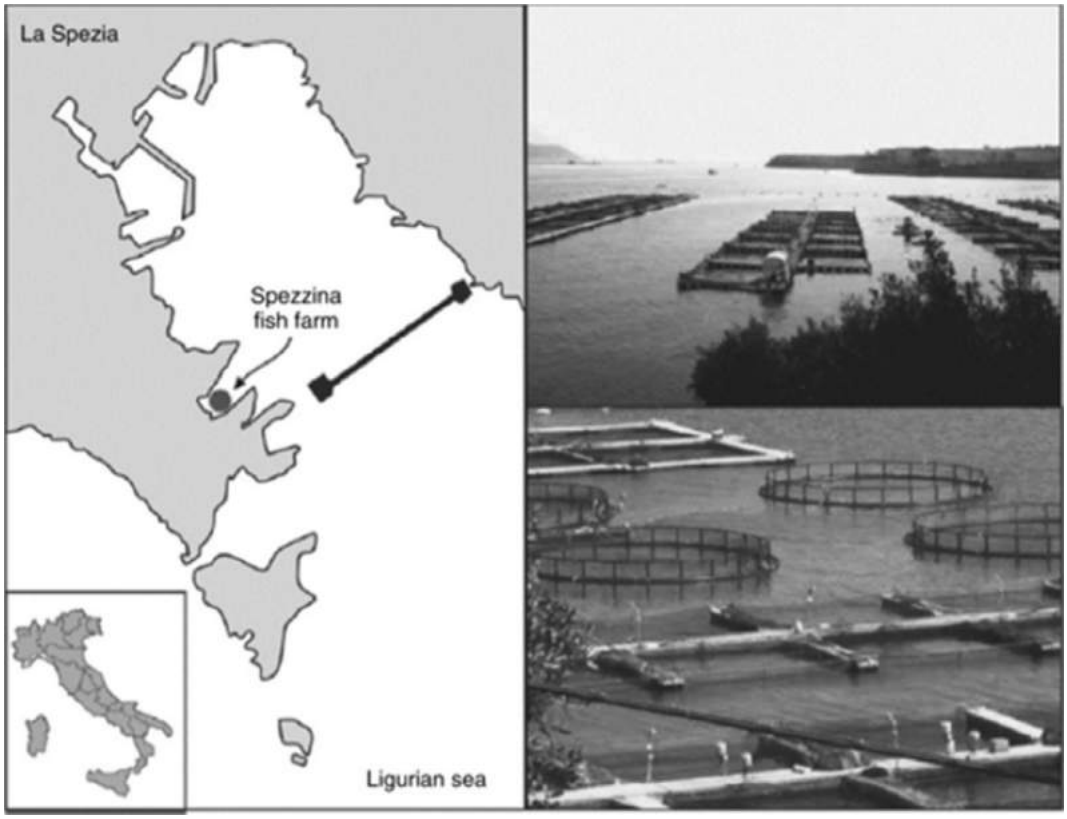


Fig. 6 Spezzina fish farm geographical position and cages disposition. Reproduced by permission of Elsevier.

Moreover, the farm is equipped with an automatic system for oxygen spreading installed under the cages; it is utilized when oxygen concentration falls below a threshold level considered dangerous for the health of the reared fish.

Fingerlings are introduced in to the cages in two periods: in spring (from March to June) and in autumn (from September to November). The total number of fish introduced is roughly 1,300,000 each year; fingerlings experience a 20% mortality rate. Fish reared in Spezzina fish farm spend 16–18 months (depending on weather, size, and period of introduction) to reach the size needed for the yield (350 g) and the farming is managed to obtain at least 7.5 tons of fish each week to meet the market demand.

The energy system diagram of the system, under study, is reported in Fig. 7.

Table 1, constructed from the diagram, shows the actual flows of materials, labor, and energy and all flows are evaluated.

For the analysis the different inputs have been identified and grouped in two categories: renewable and nonrenewable resources. The total emergy flow, obtained by summing all the independent contributions, made it possible to calculate the transformity of fish, given the quantities produced (see Table 1) and assumed that each kilogram of fish corresponds to 800 kcal (3.35×10^6 J kg^{-1}).

Three other systems producing fish are selected for comparison: two intensive fish-farming systems producing, respectively, salmons (*Salmo salar*) in the Umpqua river estuary and tilapia (*Tilapia mariae*) in the Nayarit, Mexico; the third system (Valle Figheri, in the Venice lagoon) produces *S. aurata* but in a seminatural extensive manner.

Assuming that the fish are all of the same quality if transformities are considered, *T. mariae* and *S. aurata* rearing showed the highest efficiencies, probably due to more favorable weather conditions of the areas that allow growth of the fish in an optimal temperature range. Furthermore, *T. mariae* is well known as an easily reared fish because of its very good adaptation to various conditions. Nevertheless, differences in transformities may also be due to the different species or to a different type of rearing system.

Considering the same type of systems, we can see that the transformities display increasing values from the lowest quality fish (*T. mariae*) to the highest one (*S. salar*) showing a good accordance with the maximum empower principle.

On the other hand, Spezzina fish farm and Valle Figheri produce the same kind of fish but in drastically different systems: one is very intensive in a narrow gulf in the Ligurian Sea; the other is an extensive system that tries to use all the natural resources available in this artificial ecosystem. The result is a much lower transformity for the system under study, meaning that Spezzina fish farm is more efficient in producing fishes, even though the production is much more dependent on external and nonrenewable inputs.

ELR for the Spezzina fish farm showed that fish rearing needs for the larger part, nonrenewable resources for its maintenance: nonrenewable external inputs are five times higher than renewable ones. If this result is compared with other similar production systems (Table 2), the dependence on nonrenewable resources of the *S. aurata* with respect to *S. salar* net pen culture in Oregon is higher, but much lower if compared with *T. mariae* rearing in brackish water in Mexico.

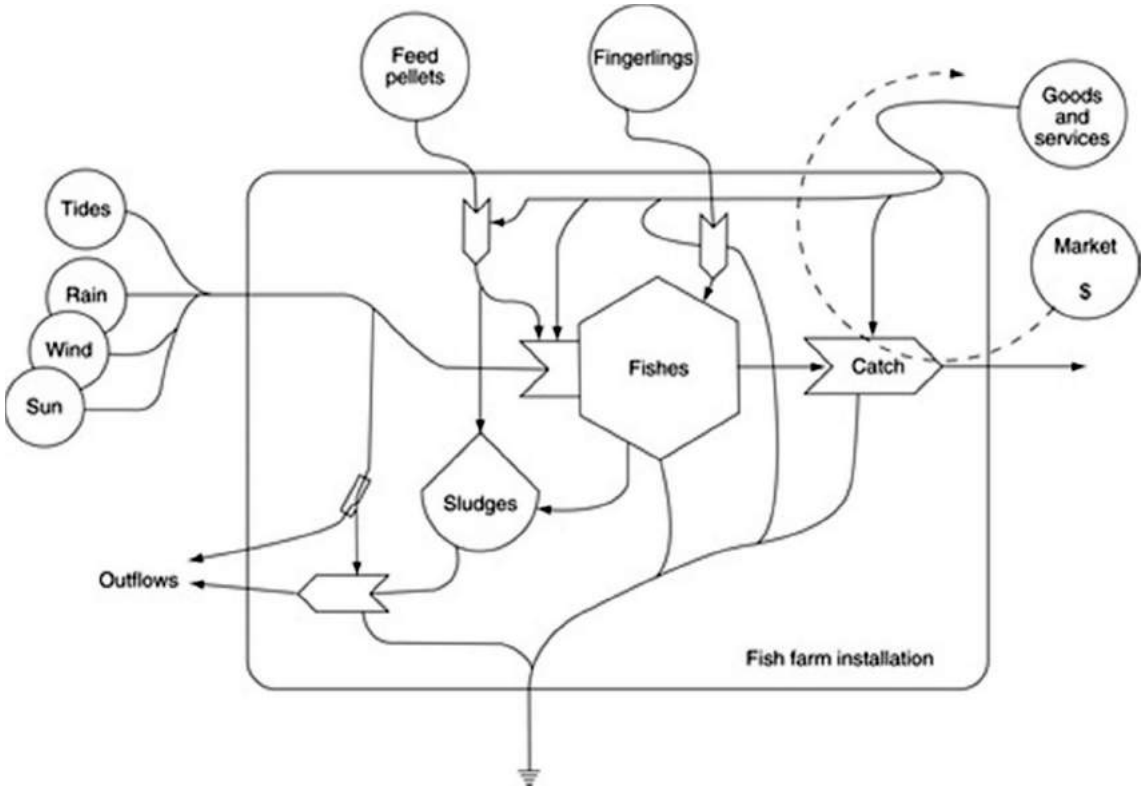


Fig. 7 Energy diagram of Spezzina fish farm. Reproduced by permission of Elsevier.

Table 1 Emergy synthesis of Spezzina fish farm

Inputs	*Unit of measure*	*Flux (unit year^{-1})*	*Emergy/unit (sej/unit)*	*Emergy flow (sej year^{-1})*
Renewable resources				
Solar radiation	J	8.61E + 15	1	8.61E + 15
Wind	J	2.44E + 12	2.45E + 03	5.97E + 15
Rain	J	1.63E + 13	1.54E + 04	2.51E + 17
Tides	J	9.96E + 10	1.68E + 04	1.67E + 15
Nonrenewable resources				
Fingerlings	€	3.25E + 05	2.22E + 12	7.22E + 17
Feed	J	1.81E + 11	1.00E + 06	1.81E + 17
Goods and services				4.34E + 17
Human labor	J	1.05E + 10	1.24E + 06	1.30E + 16
Fuel	J	4.00E + 11	5.30E + 04	2.12E + 16
Capital costs	€	1.00E + 05	2.22E + 12	2.22E + 17
Maintenance costs	€	8.00E + 04	2.22E + 12	1.78E + 17
Renewable resources (sum 3–4)				2.67 + E17
Nonrenewable resources (sum 5–7)				1.34E + 18
Total emergy flow (sum 3–7)				1.60E + 18
Fish (*S. aurata*)	J	1.22E + 12	1.32E + 06	
	kg	3.64E + 05	4.40E + 12	

Reproduced by permission of Elsevier.

These results could be related to differences in the role of natural systems for the *S. salar* culture that take advantage of the wide tidal range that characterizes the Oregon coasts; while for *T. mariae* culture, value is strictly affected by the external inputs and is higher than what is detected in other systems. Nevertheless, all these results are very relevant if compared with the seminatural system in the lagoon of Venice, where renewable inputs are three times larger than nonrenewable ones. As a direct consequence of what is observed for the ELR, values of EYR in Spezzina fish farm are quite low and close to the unit. This is an indication of the prevalent dependence of the analyzed system on the economy and on human control. This kind of result is typical of systems unable

Table 2 Comparison between emergy indices from Spezzina fish farm and similar production system

	S. aurata	S. salar	T. mariae	S. aurata
Transformity	1.32E + 06	9.70E + 06	5.61E + 05	2.45E + 07
ELR	5.00	4.24	46.52	0.34
EYR	1.20	1.23	1.02	3.95

to exploit natural resources and is based on the import of inputs characterized by high emergy. This is true also for the other intensive systems: especially *T. mariae* rearing, which, in fact, is much more dependent on expensive, nonrenewable resources with respect to what is observed in Spezzina fish farm. On the contrary, the extensive system shows a very high contribution from the environmental inputs, although at the cost of a lower productivity.

According to emergy results, a significant improvement toward the environmental sustainability of the examined process could be derived from the modification of the productive cycle that should, for example, include the in situ fingerlings production that have the highest emergy input in the evaluation. Moreover, it is advisable that in Mediterranean conditions the fish farm installation was disposed in the widest possible area so that natural resource contributions could be more significant.

Regional Analysis

Regional systems are complex systems that for the purpose of this type of evaluation can be considered to be in a steady state. Their structure has many levels of organizations, and their openings to the outside and irreversible dynamics are typical features of complex systems. They require resources and energy to enhance internal order and to self-organize, while increasing the entropy of their surroundings. The increase in entropy exceeds the internal entropy decrease of the system, so the second principle of thermodynamics is observed. Hence, the region behaves like a dissipative structure.

For a rigorous and successful analysis, much historical and scientific information about the region must be collected. The first step of an emergy evaluation is energy modeling of the region with its inflows, outflows, and local stocks. Many emergy evaluations of states and nations have been done and can be found in the scientific literature (see, e.g., emergy evaluation of Italy, environmental accounting using emergy: West Virgina, etc.). In the analysis of nations, flows from outside the system (F) and flows exploited inside the system (R and N) are identified. Imported flows are considered nonrenewable as the transport component is mostly based on fossil fuel. Once inputs are expressed in emergy terms, it is possible to calculate emergy indicators. These can be useful to describe the characteristics of a region and to formulate future policy. Fig. 8 shows a simplified energy diagram for West Virginia.

In the diagram, purchased resources are indicated F (fuels), G (Goods), and PI (services): R and N indicate the renewable and nonrenewable resources, respectively, while B and PE represent the exported goods and services, respectively. *Dashed lines* represent the counter-current to emergy in exchanges. The main characteristics and storage of the system are indicated. In particular, the storage of coal is very important for the state.

Once the diagram has been made, the next step is to quantify the inputs; it is necessary to transform resource and energy inputs into emergy units. The next step is to sum the resources in categories as stated above (renewable, nonrenewable, etc.). Then emergy indicators can be constructed. For a detailed analysis of West Virginia, see the "Further reading" section; here, we only present the main characteristics to illustrate an application of emergy. The analysis shows that the largest renewable emergy input is growth of timber (44×10^{20} sej year^{-1}), while, among nonrenewable inputs, coal is the greatest emergy flow for production and consumption, followed by electricity, petroleum, and natural gas. Among imports the largest are imported goods, services embodied in goods, and petroleum fuels. Natural gas is not considered as it only passes through the state of West Virginia and then it is not used here but somewhere else. Coal and electricity generated from coal amount to 63% of the total emergy exported.

The emergy results (Table 3) indicate that West Virginia has a low investment ratio (2.39 vs. 7 of United States) and a high ELR (20). Thus, it is sustained by a flow of nonrenewable resources that is 20 times the size of renewable flows and needs little investment to exploit it. The system could sustain its future development by using its stocks of nonrenewable resources (e.g., coal), but it has to account for past consumptions of renewable resources (e.g., timber) and consider the negative effects of using nonrenewable resources, such as greenhouse gas emissions. At the current rate of use, West Virginia coal resources will be finished within 300 years.

By emergy evaluation, it is possible to assess whatever the emergy paid for a resource balances the real value of what the system is exporting. In economics, it is customary to relate the cost of a good to its cost of extraction; the real value of natural resources is often underestimated. This occurs in West Virginia, which exports twice as much emergy as it receives, causing a deficit of 1.46×10^{23} sej year^{-1} (about 66% of annual emergy use in the state). This is mainly due to export of coal that is about 1.50×10^{23} sej year^{-1}. As stated by Campbell et al.,

> West Virginia received 3.56 billion dollars in net transfer payments ... from the federal government. This money makes up about 14% of the existing emergy deficit when converted to emergy using the West Virginia emergy to dollar ratio. The question of the equity of exchange between West Virginia and the nation could be further resolved using emergy methods to systematically consider all the benefits and costs accruing to both the state and the nation as a result of their relationship.

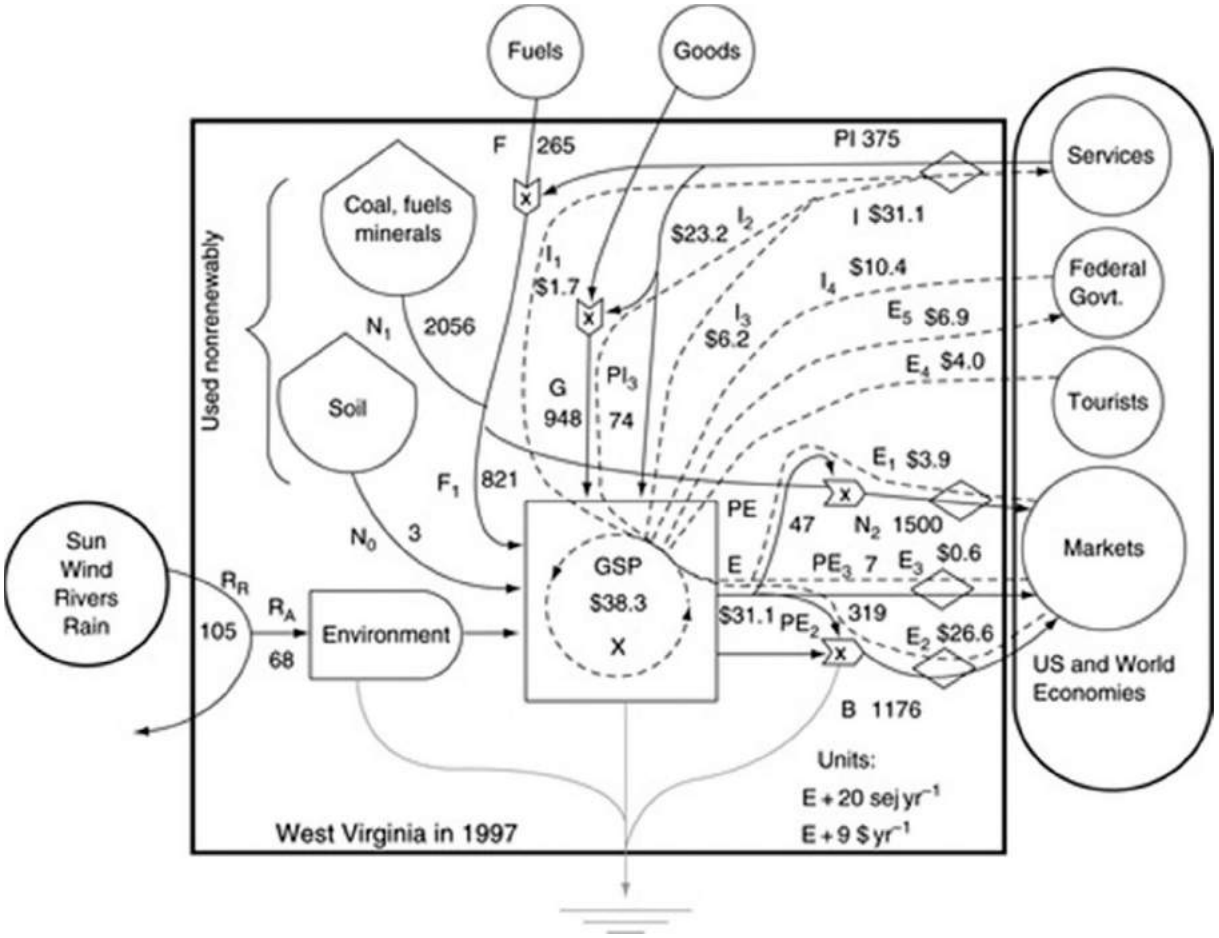


Fig. 8 Energy diagram of West Virginia. Reproduced from Campbell, D.E. and Brandt-Williams, S.L. (2005). Environmental accounting using emergy: Evaluation of the State of West Virginia, EPA/600/R-05/006, with permission from USEPA.

The emergy results also show a high emergy use per capita, in contrast to many social indicators. Emergy use is concentrated in the industrial sector while the rest of the population (58%) was in rural areas and does not benefit from the empower flows.

As in the case of West Virginia, it is appropriate to exclude flow that hides other characteristics of the system from emergy evaluation. In this evaluation export of coal and electricity were excluded, revealing that the development of West Virginia has similar characteristics to that of other American states, such as Maine and North Carolina.

Emergy in an Input-State-Output Framework

Ecosystems and human systems are open thermodynamic systems that self-organize, hopefully toward higher complexity and organization. An input-state-output framework has been developed to describe their functioning with a common feature: in both ecosystems and human systems emergy can represent the input upon which the structures of the systems are created, maintained and are able to produce outputs. For what concerns ecosystems, the "state" can be measured by systems indicators like ecoexergy (*link to entry*) and their outputs by means of the ecosystem services evaluation (*link to entry*). Starting from data available on different types of ecosystems and identifying a threshold at the average/median points of the three axes (emergy flow, ecoexergy and ecosystem services) determining a 3D space (a cube), eight categories, or classes of ecosystems, have been defined (high and low emergy flow, high and low ecoexergy, high and low ecosystem services). Results show that, in general, there is a good correlation between ecosystem services and ecoexergy, highlighting the fact that half of the subcubes obtained are practically empty: in case of high emergy those with high value of ecoexergy and low value of services or vice versa; in case of low emergy it seems impossible to have high ecosystem services regardless the degree of organization of ecosystems (ecoexergy).

This approach has been conceived also to depict the sustainability of economic systems starting from the environmental, social and economic viewpoints. In this representation, though, the three dimensions are taken as strictly separated and irreducible to each other: the environment provides the goods and services that are used by societies for their well-being; an organized society has a useful economic production that contributes to the well-being as well. We can therefore adopt a pyramidal representation of sustainability, that highlights the succession of stages rather than a partial overlapping among the three components.

Table 3 West Virginia emergy indices.

Name of index	*Expression*	*Quantity*	*Units*
Renewable emergy received	R_R	1.05E + 22	sej $year^{-1}$
Renewable emergy used	R_A	6.60E + 21	sej $year^{-1}$
In state nonrenewable	$N_0 + N_1$	2.06E + 23	sej $year^{-1}$
Imported emergy	F + G + PI	1.59E + 23	sej $year^{-1}$
Total emergy inflows	R_R + F + G + PI	1.70E + 23	sej $year^{-1}$
Total emergy used	$U = R_A + N_0 + F_1 + G + PI$	2.21E + 23	sej $year^{-1}$
Total exported emergy	$B + PE + N_2$	3.05E + 23	sej $year^{-1}$
Emergy used from home sources	$(N_0 + F_2 + R_A)/U$	0.282	
Imports–exports	$(F + G + PI) - (B + PE + N_2)$	−1.46E + 23	sej $year^{-1}$
Ratio of exports to imports	$(B + PE + N_2)/(F + G + PI)$	1.92	
Fraction of use, locally renewable	R_A/U	0.030	
Fraction of use, purchased import	(F + G + PI)/U	0.72	
Fraction of use, imported service	PI/U	0.17	
Fraction of use that is free	$(R_A + N_0)/U$	0.03	
Ratio of purchased to free	$(F_1 + G + PI)/(R_R + N_0)$	19.9	
Environmental loading ratio	$(F_1 + N_0 + G + PI)/R_R$	20.4	
Investment ratio	$(F + G + PI)/(R_R + N_0 + F_2)$	2.39	
Use per unit area	U/Area	3.55E + 12	sej m^{-2}
Use per person	U/Population	1.22E + 17	sej/ind.
Renewable carrying capacity at present	$(R_R/U)^*$(Population)	88.625	People
Standard of living			
Developed carrying capacity at same	8(R/U)*(Population)	709.003	People
Living standard			
WV State Econ. Product	GSP	3.83E + 10	$/year
Ratio of WV emergy use to GSP	U/GSP	5.78E + 12	sej/$
Ratio of US emergy use to GNP	U/GNP	1.20E + 12	sej/$
Ratio of electricity/emergy use	EI/U	0.072	
Ratio of Elec. Prod./emergy use	Elp/U	0.254	
Emergy of fuel use per person	Fuel use/population	3.41E + 16	sej/ind.
Population		1.81E + 06	People
Area		6.24E + 10	m^2

Reproduced from Campbell, D.E. and Brandt-Williams, S.L. (2005). Environmental accounting using emergy: Evaluation of the state of West Virginia, EPA/600/R-05/006, with permission from USEPA.

Several choices can be made for the three-axis representation. Emergy could be substituted for example by Ecological footprint (*link to entry*) on the environmental (input) axis. On the social axis indicators of Social Capital, indicators of equity within a society (e.g., Gini index or Palma Ratio), of labor or of education can be adopted; indicators like GDP, the Index of Sustainable Economic Welfare (ISEW) or Genuine Progress Indicator (GPI) on the economic one.

A first analysis was conducted by Pulselli et al. (2015) using the emergy flow per capita; the Gini index of income distribution and the Gross Domestic Product per capita (GDP) on the environmental, social and economic axes respectively. The Gini coefficient is a measure of inequality ranging from 0 (in the situation of perfect equality, where every unit has the same income) to 1 (in the situation of greatest inequality, where only one unit receives the whole income). Even if, strictly speaking, the Gini Index is an economic indicator it has been shown that inequality is strongly correlated with health and social problems (Wilkinson and Pickett, 2009).

GDP is the sum of the market value of the overall set of goods and services produced by an economic system in a given period of time (generally 1 year). It can thus be intended as an indicator of the economic output.

Pulselli et al. (2015) examined 99 national economies, grouping them into 8 subcubes separated by the median value of the three axes. Data are referred to the year 2008, since it was the last one available for emergy. Most of the economies (85) fall in four of the eight possible subcubes, while the other 14 points fall in less populated subcubes, but in zones that are very close to median values of at least one of the indicators.

From the results a strong relationship between resource use per capita and GDP per capita emerges, pointing out how economic growth drives, and depends on, an increasing requirement of energy and matter to be transformed by the economic system, while the level of inequality on societies is quite independent of the other two metrics.

The area of the cube where there should be nations with low level of resource use and high GDP is practically empty indicating that dematerialization is a goal hard to be reached.

The temporal dimension is a key factors of this representation: further than the ranking of the nations in 1 year, also the evolution of a single point in the diagram is meaningful in order to follow the behavior of a nation in time (see Fig. 5). This can be useful to evaluate the effects of national policies in economic but also in social and environmental terms.

References

Pulselli FM, Coscieme L, Neri L, Regoli A, Sutton PC, Lemmi A, and Bastianoni S (2015) The world economy in a cube: A more rational structural representation of sustainability. *Global Environmental Change* 35: 14–51.
Wilkinson R and Pickett K (2009) *The spirit level*. London: Penguin Books Ltd.

Further Reading

Bastianoni S, Morandi F, Flaminio T, Pulselli RM, and Tiezzi EBP (2011) Emergy and emergy algebra explained by means of ingenuous set theory. *Ecological Modelling* 222: 2903–2907.
Brown MT and Herendeen RA (1996) Embodied energy analysis and emergy analysis: A comparative view. *Ecological Economics* 19: 219–235.
Brown MT, Campbell DE, De Vilbiss C, and Ulgiati S (2016) The geobiosphere emergy baseline: A synthesis. *Ecological Modelling* 339(2016): 92–95.
Campbell DE and Brandt-Williams SL (2005) Environmental accounting using emergy: Evaluation of the state of West Virginia, EPA/600/R-05/006.
Coscieme L, Pulselli FM, Jørgensen SE, and Bastianoni S (2013) Thermodynamics-based categorization of ecosystems in a socio-ecological context. *Ecological Modelling* 258: 1–8.
Kazanci C, Schramski JR, and Bastianoni S (2012) Individual based emergy analysis: A Lagrangian model of energy memory. *Ecological Complexity* 11: 103–108.
Odum HT (1988) Self-organization, transformity and information. *Science* 242: 1132–1139.
Odum HT (1996) *Environmental accounting: emergy and decision making*. New York: Wiley.
Odum HT and Odum E (1981) *Energy basis for man and nature*. New York: McGraw-Hill.
Sciubba E and Ulgiati S (2005) Emergy and exergy analyses: Complementary methods or irreducible ideological options? *Energy* 30: 1953–1988.
Ulgiati S, Odum HT, and Bastianoni S (1994) Emergy use, environmental loading and sustainability, An emergy analysis of Italy. *Ecological Modelling* 73: 215–268.
Vassallo P, Bastianoni S, Beiso I, Ridolfi R, and Fabiano M (2006) Emergy analysis for the environmental sustainability of an inshore fish farming system. *Ecological Indicators* 7: 290–298.

Emergy Ecosystems and Network Analysis☆

Mark T Brown and Mathew Cohen, University of Florida, Gainesville, FL, United States

Definitions

Energy is sometimes referred to as the ability to do work. Energy is a property of all things which can be turned into heat, and is measured in heat units (BTUs, calories, or joules).

Emergy is the availability of energy (exergy) of one kind that is used up in transformations directly and indirectly to make a product or service. The unit of emergy is the emjoule (see below), a unit referring to the available energy of one kind consumed in transformations. For example, sunlight, fuel, electricity, and human service can be put on a common basis by expressing them all in the emjoules of solar energy that is required to produce them. In this case the value is a unit of solar emergy expressed in solar emjoules (abbreviated sej). Although other units have been used, such as coal emjoules or electrical emjoules, in most cases all emergy data are given in solar emjoules.

Emjoule is the unit of measure of emergy, the term is short for emergy joule. An emjoule is an expression of the units of energy previously used to generate a product; for instance the solar emergy of wood is expressed as joules of solar energy that were required to produce the wood. Solar emjoules is abbreviated "sej."

Empower is a flow of emergy (i.e., emergy per time). Emergy flows are usually expressed in units of solar empower (solar emjoules per time).

Exergy is the maximum useful work that can be extracted from a system as it reversibly comes into equilibrium with its environment. In other words exergy is the portion of the total energy of a system that is available for conversion to useful work.

Unit emergy values (UEVs) are based on the emergy required to produce something. UEVs are calculated by dividing the sum of all emergy required for producing an output by the units of product output. There are two types of UEVs appropriate for this article as follows:

Transformity is defined as the emergy per unit of available energy (exergy). For example, if 4000 solar emjoules are required to generate a joule of wood, then the solar transformity of that wood is 4000 solar emjoules per joule (abbreviated sej J^{-1}). Solar energy is the largest but most dispersed energy input to the earth. The solar transformity of the sunlight absorbed by the earth is 1.0 by definition.

Specific emergy is the unit emergy value of matter defined as the emergy per mass, usually expressed as solar emergy per gram (sej g^{-1}). Solids may be evaluated best with data on emergy per unit mass for its concentration. Because energy is required to concentrate materials, the unit emergy value of any substance increases with concentration. Elements and compounds not abundant in nature therefore have higher emergy/mass ratios when found in concentrated form since more work was required to concentrate them, both spatially and chemically.

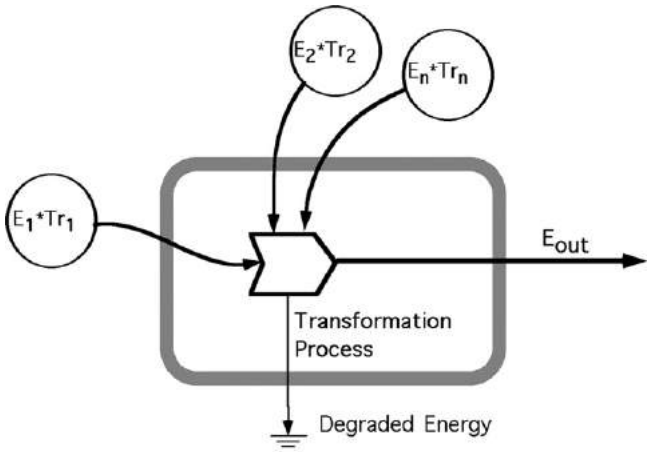
Emergy, Energy, and Quality

Emergy, defined as the available energy of one form (usually solar energy) required, directly and indirectly, through all processes and transformations to make a product or flow, provides a numeric framework for comparing different forms of energy and materials and providing a measure of quality in directly comparable units (solar emjoules, or sej). Emergy is often referred to as energy memory, reflecting that this system synthesis approach is effectively a form of accounting that traces energy flow and dissipation back through all necessary transformations to scale all flows relative to a common energy benchmark (solar emergy). Emergy accounting allows comparison of energy flows of different forms; Odum (1996) argues that different forms of energy have different qualities that arise from the energy required to make them. In self-organizing adaptive systems, he argues, forms of energy that require larger investment per unit available energy must provide commensurate higher quality cybernetic work in the form of feedback control. Transformity (T_R) is an index of quality and quantifies the emergy invested per unit available energy produced (i.e., emergy per exergy, sej J^{-1}) for each flow in a system of interest. In like manner, specific emergy (S_P) is an index of quality of material by quantifying the emergy investment per unit mass produced (sej g^{-1}).

The use and transformation of material and energy is system dependent; where the appropriateness of an energy in a particular system is dictated by its form and is related to its concentration. The processes of the biosphere are infinitely varied and are more than just thermodynamic heat engines. As a result, the use of heat measures of energy that can only recognize one aspect of energy, its ability to raise the temperature of things, cannot adequately quantify the work potential of energies used in more complex processes that span multiple levels of system processes, ranging from the smallest to the largest scales of the biosphere.

☆*Change History*: March 2018. Brown updated all sections Tables 1–3, Figs. 6.

This is an update of M.T. Brown and M.J. Cohen, Emergy and Network Analysis, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 1229–1239.



$$Em_{out} = \Sigma\, E_n{*}Tr_n$$

$$Tr_{out} = Em_{out} / E_{out}$$

Where;

$E_{1...n}$ = Available energy inputs

E_{out} = Available energy of output

Em = Emergy

Tr = Transformity

Fig. 1 All transformation processes utilize available energy inputs of varying qualities (E_1, E_2, ..., E_n) matching lower quality energy inputs (E_1) to higher quality inputs (E_2, ..., E_n). Energies are converted to emergies by multiplying by their appropriate Emergy Intensity (or transformities). In a static calculation the emergy of the output is equal to the sum of the emergy inputs over the time require to make the output.

Transformity and Specific Emergy

Transformity and specific emergy are unit emergy values calculated as the total amount of emergy required to make a product or service divided by the available energy of the product (resulting in a transformity) or divided by the mass of the product (resulting in a specific emergy). Figs. 1 and 2 illustrate a static method of calculating a transformity first in equation form (Fig. 1) and then with example numbers (Fig. 2). The transformity of the product is the *emergy* of the product divided by the *energy* of the product (units are sej J^{-1}). If the output flow is in mass then the specific emergy of the product is the emergy of the output divided by the mass (units are sej g^{-1}).

Emergy and Hierarchy

A hierarchy is a form of organization resembling a pyramid where each level is subordinate to the one above it. Depending on how one views a hierarchy, it can be an organization whose components are arranged in levels from a top level (small in number, but large in influence) down to a bottom level (many in number, but small n influence). In general, in ecology we consider hierarchical organization to be a group of processes arranged in order of rank or class in which the nature of function at each higher level becomes more broadly embracing than at the lower level. Thus we often speak of food-chains as hierarchical in organization.

Most if not all systems form hierarchical energy transformation series, where the scale of space and time increases along the series of energy transformations. Many small scale processes contribute to fewer and fewer of larger scale processes (Fig. 3). Energy is converged from lower to higher order processes, and with each transformation step, much energy loses its availability (a consequence of the 2nd Law of Thermodynamics), while only a small amount is passed along to the next step. In addition, some energy is fed back reinforcing power flows up the hierarchy. Note in Fig. 3 the reinforcing feedbacks by which some of each transformed power flow feeds backward so that its special properties can have amplifier actions.

Unit Emergy Values and Quality

Quality is a system property, which means that an "absolute" scale of quality cannot be made, nor can the usefulness of a measure of quality be assessed without first defining the structure and boundaries of the system. Self-organizing systems (be they the biosphere or an ecosystem) are organized with hierarchical levels (Fig. 3) and each level is composed of many parallel processes. This leads to two possible definitions of quality: (a) Parallel quality, and (b) Cross quality. The first, "parallel quality," is related to

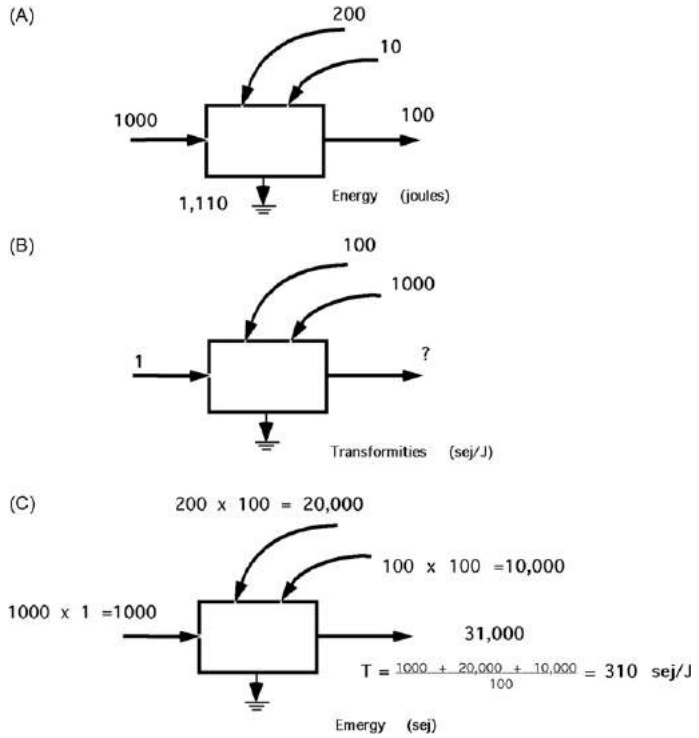


Fig. 2 The calculation of a transformity for a transformation process where (A) available energies of different qualities are used to produce a higher quality energy of a different form. (B) The different qualities are expressed by their different solar transformities. (C) The emergy on each pathway is determined by multiplying the available energy in (A) by its transformity in (B). The transformity of the output flow in (C) is found by summing the emergy inputs and dividing by the energy of the output.

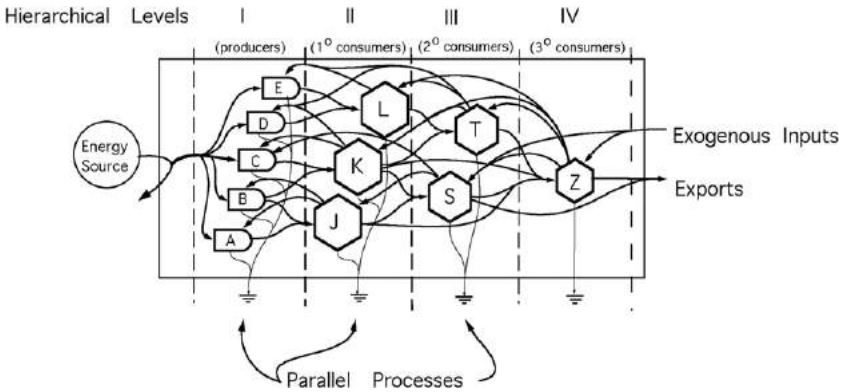


Fig. 3 Systems can be represented as hierarchical organization of energy flows and components. Hierarchical levels are aggregations of like components sometimes called trophic levels in ecosystems. At each level energy is converged into fewer components (less biomass) and much energy is degraded. Within each level there are parallel processes that have similar function. Larges energy flows are associated with components on the left (lowest quality), while small flows of available energy support components on the right (highest quality). Feedbacks from higher level components act as control actions on lower quality components.

the efficiency of a process that produces a given flow of energy or matter within the same hierarchical level (comparison among units in the same hierarchical level in **Fig. 3**). For instance, for any given output, say biomass, there are almost an infinite number of ways of producing it (all the various species in the myriad collection of ecosystems in the biosphere; each with slightly different efficiencies). A compilation of transformities for ecosystem level gross primary production yielded transformities from 1.0 E3 sej J^{-1} to 9.2 E4 sej J^{-1} depending on species composition, community type, and driving energies (Brown and Bardi, 2001). Since each individual process has its own efficiency, the output from the process has a distinct UEV. Quality as measured by UEV relates to the emergy required to make like products under differing conditions and processes. For the most part, UEVs of like products are within the same order of magnitude. When comparing parallel processes the UEV is a measure of biosphere efficiency suggesting that parallel products having lower UEVs are produced more efficiently requiring less emergy per unit of output.

The second definition of quality, "cross quality," is related to the hierarchical organization of systems. In this case, UEVs are used to compare components or outputs from different levels of the hierarchy, accounting for the convergence of emergy at higher and higher levels (comparison of different hierarchical levels, in Fig. 3). At higher hierarchical levels, a larger convergence of inputs is required to support the component (i.e., much plant biomass to support 1° level consumer, many 1° level consumers to support 2° level consumers, etc.). Also, higher feedback and control ability characterize components at higher hierarchical levels. Therefore, higher UEVs, as equated with higher level in the hierarchy, often exhibit greater flexibility and greater spatial and temporal effect. With this definition of quality—the higher the UEV the higher the quality of the process or product. UEVs of products from different hierarchical levels usually differ by at least one order of magnitude.

Unit Emergy Values, Efficiency, and Integrity

When an ecological network is expressed as a series of energy flows and transformation steps where the transformation steps are represented as Lindeman efficiencies, the resulting UEVs represent trophic convergence and a measure of the amount of solar energy required to produce each level in the hierarchy, As such, they play the role of quality indicators. This is true for systems selected under maximum power constraints (Lotka, 1922; Odum, 1983) and is therefore true in healthy ecosystems. However, in an ecosystem stressed by excess outside pressure, relationships among components are likely to change, some components may disappear, others may emerge, and the whole hierarchy may be altered. The efficiency of given processes within the system may therefore change and due to a simplified structure of the system, some patterns of hierarchical control of higher to lower levels may diminish or disappear. All in all, these changes will translate into different values of the UEVs, the variations of which become clear measures of lost or decreased system integrity.

Emergy of the Geobiosphere: The Basis for Computing Emergy and UEVs of Ecological and Techno-Ecological Processes

The exergy sources driving geobiosphere processes are varied and change with scale of the processes. The primary exergy sources are absorbed solar radiation, geothermal heating from the Earth's interior, and tidal energy absorbed by the oceans. These sources are, in turn, transformed into secondary and tertiary renewable flows, which combined with the primary sources not only drive the dearth's ecosystems, but also drive geologic processes and the formation of mineral ores and fossil fuels. Finally all of these sources are combined in the myriad ecological and technological process that ultimately generate the living and nonliving components of Earth.

Annual Budget of Emergy Supporting the Geobiosphere

The annual budget of emergy flow (empower) supporting the geobiosphere including the atmosphere, ocean, and earth crust, includes three sources of energy: solar energy, tidal energy, and geothermal energy from the deep earth (Fig. 4). When evaluated in

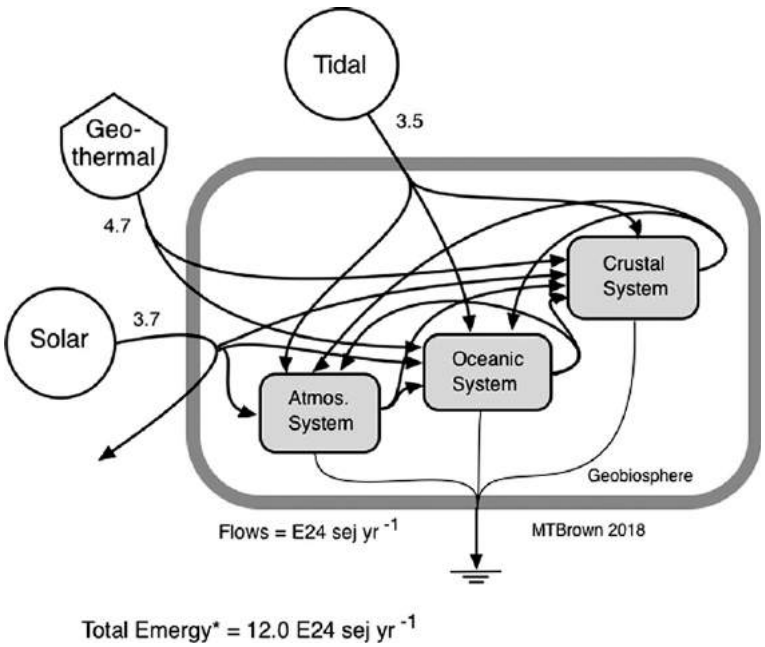


Fig. 4 The geobiosphere is drive by three main sources of energy: solar, deep heat (residual plus radioactive decay), and tidal momentum. For a full explanation of their solar equivalents see Brown et al. (2016). *Note*: total emergy is rounded to 12.0 E24 sej year^{-1}.

Table 1 Annual renewable emergy contributions to global processes

Input	*Unite*	*Exergy (units year^{-1})*	*SER*[a] *(sej unit^{-1})*	*Empower (E24 sej year^{-1})*
Solar energy absorbed	J	3.73 E24	1.0	3.7
Geothermal energy	J	9.52 E20	4,000	4.7
Tidal energy absorbed	J	1.14 E20	30,000	3.5
	Total global empower	–	–	12.0[b]

[a]Solar equivalence ratio (see Brown et al. 2016 for explanation).
[b]Rounded to whole number.

Table 2 Emergy of products of the global energy system

Note *Product and units*	*Emergy (E24 sej year^{-1})*	*Production (E20 J year^{-1})*	*Transformity (sej J^{-1})*
Global wind circulation, J	12.0	151.2	800
Global precipitation on land, J	3.74	5.34	7,000
Global precipitation on ocean, J	9.56	19.5	4,000
Average river geopotential, J	3.74	2.91	12,800
Average river chemical energy, J	3.74	1.76	21,300
Average waves at the shore, J	7.89	18.9	4,200

solar emergy (Brown *et al.*, 2016), these contributions to the geobiosphere, total about 12.0 E24 sej year^{-1} (**Table 1**). The sum of solar emergy, tidal emergy and geothermal emergy is termed the geobiosphere emergy baseline (GEB).

Average UEVs for Secondary and Tertiary Global Sources

The empower that is derived from GEB drives the productive processes of the geobiosphere and is responsible for developing exergy gradients that are transformed into secondary exergy products (wind, and chemical potential of rain water) and tertiary exergy products (chemical and geopotential energy of river discharges and the available energy of breaking waves). At the scale of the geobiosphere, the secondary and tertiary exergies are products of the GEB. At the scale of the ecosystem they become sources or external driving energies. **Table 2** lists these sources, their exergy and their UEVs. The UEVs of secondary and tertiary renewable emergy sources are useful for evaluating inputs supporting ecological systems including those dominated by humans.

UEVs of Global Minerals and Fossil Fuels

Human dominated ecosystems are dependent not only on renewable inputs but to a far greater degree on nonrenewables fossil fuels and mineral resources. The evaluation of techno-ecosystems relies on having UEVs of both renewable and nonrenewable sources as well as other technological products.

Table 3 lists the UEVs of the fossil fuels and some of the most important mineral resources. The methods of computing the fossil fuel UEVs relies on published estimates of historic global net primary production (NPP) on land and oceans, published preservation, and conversion factors of organic matter, and assessments of the present total global storages of coal, petroleum, and natural gas (Brown *et al.*, 2011). Mineral UEVs are far more difficult to compute, because of the varied geologic processes that produce them. An aggregated approach was undertaken to estimate the mineral formation exergy and the ore concentration exergy, leading to mineral specific emergy for 144 mineral ores and aggregates that represent the majority of commercial ores used in human technological processes (Brown and Ulgiati, 2018).

UEVs have been computed for thousands of technological products (Brown and Ulgiati, 2018). Computing UEVs of human techno-ecosystem processes requires a full accounting of all inputs converted to emergy. Recently the use of EcoInvent database has increased the number of technology UEVs and there is an online database (http://www.emergysociety.com/emergy-society-database/) that contains nearly 1000 entries as of this writing.

Calculating Unit Emergy Values

Several methods are used to calculate UEVs, they include: (1) static calculations, (2) dynamic simulation, and (3) network analysis. Most commonly, static calculations are used for processes where the flows of energy, materials, and services over a particular time period are multiplied by their UEVs, summed, and divided by the available energy, or mass of the product produced during that same time period. Dynamic simulation has been used for some resources that require long periods of time to generate, for instance forest wood (Tilley, 1999) or soils (Cohen, 2003a, b). The dynamic method uses rate of change equations for storages that add emergy as long as the storage of material is accumulating (Odum, 1996). Evaluating UEVs with a technique of network analysis uses what Collins and Odum

Table 3 UEVs of fossil fuels and some common minerals

	UEV	
Item	*(sej J^{-1})*	*(sej g^{-1})*
Fossil fuels		
Coal—subbituminous and lignite	56,700	20,000
Coal—anthracite and bituminous	69,000	29,500
Crude oil	132,160	5.78E + 09
Natural gas	140,000	7.46E + 09
Minerals		
Limestone		2.41E + 09
Copper		5.71E + 09
Bauxite		4.32E + 09
Lead		2.33E + 09
Iron ore		1.75E + 09
Uranium		3.24E + 13

Data are from Brown, M. T. and Ulgiati, S. (2018). Emergy and environmental accounting; Coupling systems of humanity and nature. New York: Springer, (Forthcoming).

(2000) called a "minimum Eigenvalue method" built on the earlier work of Patterson (1983). The Eigenvalue method was refined using linear equations and the EXCEL solver routine by Cohen (2003a, b) and further by Zarba and Brown (2015). Termed emergy-network analysis, it uses a set of simultaneous equations to partition emergy throughout an interconnected network of components that may include feedback, assigning emergy and calculating UEVs for all flows between components of the network.

Each of the methods is applicable to different situations and if applied to the same systems might yield slightly different results. Generally, the static method has been used most widely, with the vast majority of published UEVs having been calculated using this method. Static calculations are appropriate for relatively established, continuously operating processes, like production of electricity, where a snapshot in time will produce inflows and outflows that vary little from a snapshot at a different time. UEVs that result from static calculations do not include emergy used during startup or early phases of a production process, which may make only minor difference if the process has been long running and well established. Nor does the static method consider feedbacks or cycling within systems.

There is no doubt that H.T. Odum recognized cycles and cybernetic feedback as major phenomena in ecological systems. However in terms of the static emergy accounting rules (Brown and Herendeen, 1996), feedbacks are disregarded to avoid double counting. They do not contribute any emergy to the elements they are flowing into (Odum, 1996; Odum and Collins, 2003). Several studies have raised the issue that the static emergy algebra does not allow a proper comprehension of the emergy dynamics of systems with recycling flows (Tilley, 1999, 2011; Cohen, 2003a, b; Cavalett and Ortega, 2007; Bastianoni *et al.*, 2011; Morandi et al., 2013; Winfrey and Tilley, 2013).

Dynamic calculation of UEVIs is appropriate for processes where the product accumulates over time and is "harvested" or used all at once (see Tilley, 2011). Since the product is accumulating the emergy used in production accumulates. Once the system reaches steady state, where inflows equal outflows, emergy no longer accumulates. In dynamic calculation, UEVs can be calculated at any time during a product's life, and they will be slightly different with each time period.

The network method incorporates feedback, and thus UEVs of products are somewhat different than those calculated using a static algebra method. Depending on system configuration, UEVs can be larger or smaller than those calculated using static methods. The calculations in general are carried out on a system that is assumed to be at steady state, although this is not a requirement. The flows of emergy are assigned to pathways according to some carrier that is conserved (i.e., energy or matter … in ecological systems often carbon).

Computing UEVs-Static Accounting

Static UEVs are calculated according to the classic emergy algebra procedure (Odum, 1996; Brown and Herendeen, 1996). The static transformity of a compartment *i* is computed as the sum of the emergy of all input flows (excluding feedbacks) multiplied by the storage turnover time divided by the energy of the storage:

$$\sum_{i}^{n} \mathrm{Em}_{(u_i)} * \mathrm{TT}_{(Q_i)} = \mathrm{Em}_{(Q_i)} \tag{1}$$

and

$$\mathrm{Tr}_{(Q_i)} = \mathrm{Em}_{(Q_i)} / \mathrm{En}_{(Q_i)} \tag{2}$$

where $\mathrm{Em}_{(u_i)}$ = Emergy of input *i*, $\mathrm{TT}_{(Q_i)}$ = Turnover time of storage *Qi*, computed as the storage energy divided by the sum of all input and output energy divided by 2, $\mathrm{Em}_{(Q_i)}$ = Emergy of storage *Qi*, $\mathrm{Tr}_{(Q_i)}$ = Transformity of storage *Qi*, $\mathrm{En}_{(Q_i)}$ = Energy of storage *Qi*.

Calculating UEVs From Network Flow Data

For each component in a system (Fig. 5A), five flows describe bilateral interactions (e.g., consumption, gross production, net production/transfer, respiration and egestion) with other components in the system. Energy or material inputs (1) arrive from exogenous sources (e.g., sunlight, hydrologic inputs) or from components within the network (e.g., plant biomass supporting herbivores). Gross production (2) quantifies the portion of that energy that is assimilated, while Egestion (5), though required for production and partially processed during digestion, is not incorporated. Respiration (3) represents the metabolic work of each compartment (i.e., internal feedbacks to secure energy), while transfer (4) is the energy that is eventually used by other components in the food web. The "heat sink" symbol represents energy unavailable to do work (i.e., entropy).

The UEV of a compartment is the incoming emergy (sej) flows driving biotic production (Flow 1, Fig. 5A) divided by the output (energy or mass) trophic transfer for that compartment (Flow 4 in Fig. 5A). Network UEVs calculated in this way represent the UEV of the flowing material, which in most cases is the same as the transformity of the storages from which they come (Odum and Collins, 2003).

A linear optimization technique that manipulates a set of unknowns (UEVs) to meet a set of constraints (emergy inflow = emergy outflow) is used to compute UEVs from network flow data. An example optimization table is given in Fig. 5B where each row and column represents a system component. The constraints to the right of each row are the constraints that emergy inflow and emergy outflow are equal. Specifically, if the flows are in energy, then the energy inputs multiplied by appropriate transformities (unknowns in this case) equals the net production or transfer of energy multiplied by its transformity:

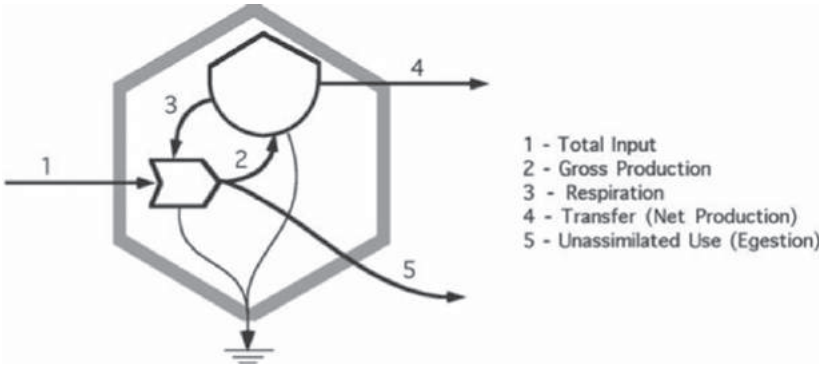
$$\text{Emergy inflow}_j = \sum_i X_{ij} * \tau_j = \sum_i X_{ij} * \tau_j = \text{Emergy outflow}_j \tag{3}$$

where X_{ij} is the energy transfer from component i to component j, and τ_i is the transformity value of respective flows.

Ecological Network Model

The aggregated ecosystem diagram of a coastal Juncus spp. marsh in Fig. 6 shows inflows of solar emergy, rain and organic matter (expressed in solar emergy) driving the ecological food chain. Ecological network analysis provides the ability to evaluate the concentration factors traced throughout a food web along trophic pathways to explore and predict functional responses of the

(A)



(B)

Emergy Intensities (EIs)			τ_1	τ_2	…	τ_n	
To ⇓	From ⇒	Solar Emergy	Comp. 1	Comp. 2	…	Comp. n	Constraints
Comp. 1			$-P_1$				Σ_{Comp1} (Flows$_i$*τ_n) = 0
Comp. 2			C_1 -to- C_2	$-P_2$		C_n -to- C_2	Σ_{Comp2} (Flows$_i$*τ_2) = 0
…					…		…
Comp. n			C_1 -to- C_n	C_2 -to- C_n		$-P_n$	Σ_{Comp3} (Flows$_i$*τ_n) = 0

Objective	Σ Constraints = 0

Fig. 5 Component diagram and schematic of input/output matrix. (A) Diagram of bilateral and internal energy pathways (flows) compiled for each compartment within a network. (B) Schematic of input/output matrix for calculating unit emergy values (UEVs) from network flow data. UEVs in the top row represent unknowns in the simultaneous equations defined by the constraints. For fully specified systems (i.e., # equations = # unknown UEVs) the objective function is redundant. (P_i = total component production minus respiration; C*i*-to-C*j* = energy transfers from component *i* to component *j*).

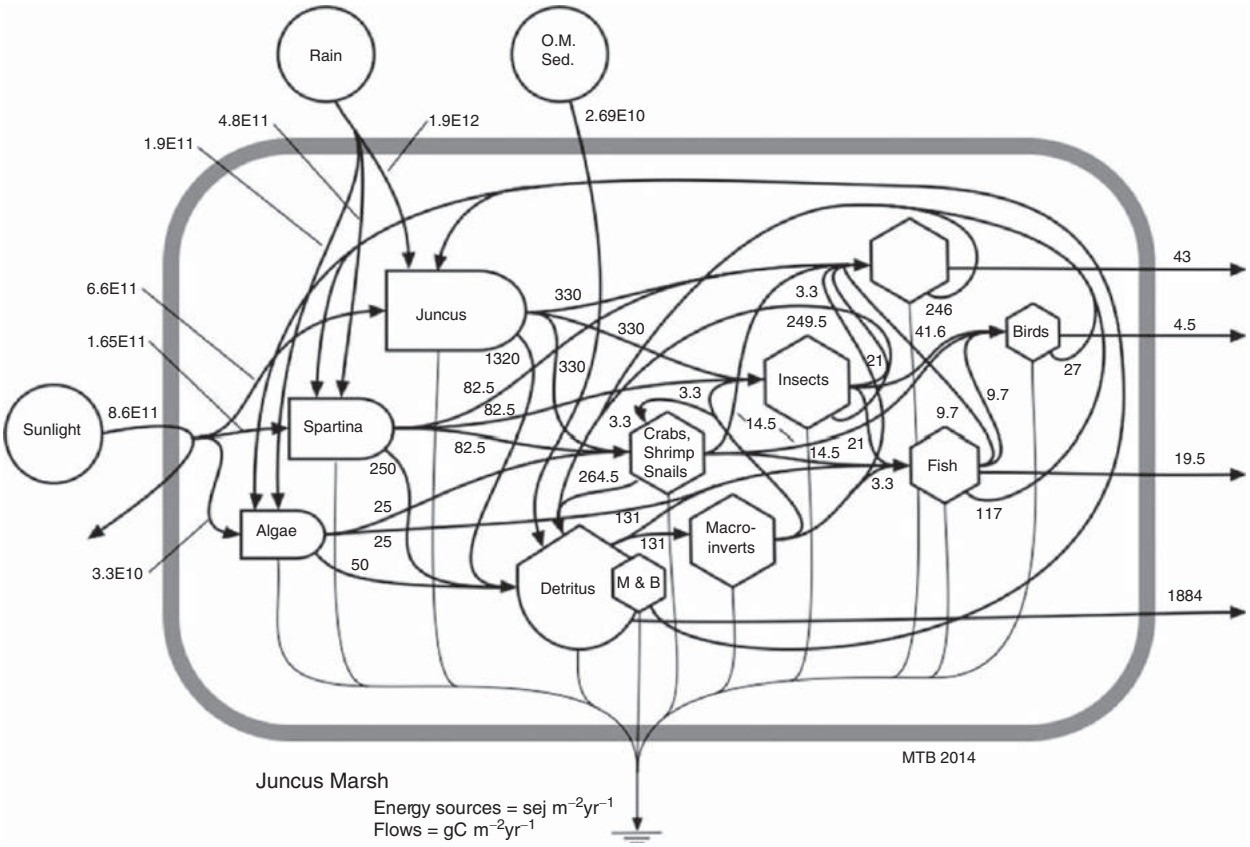


Fig. 6 Energy systems diagram of the coastal *Juncus spp.* marsh ecosystem showing major trophic levels and flows of organic matter (g C) between compartments.

ecosystems to structural changes. The topology of the high marsh food web in **Fig. 6** was constructed from literature values using niche model concepts developed by Williams and Martinez (2000). Network data, consisting of carbon flows (g C year^{-1}) from the published literature were compiled in a "To … From" matrix (**Fig. 7**).

The network model was used to compute UEVs for each of the compartments in the network. To compute the UEVs, the data in **Fig. 6** were organized in a square matrix (**Fig. 7**) where the value in the cells represents the flow from the element in the column to the element in the row. The matrices also included the driving emergy sources. Then we computed the transformities following the minimum eigenvector method (Odum and Collins, 2003) and Microsoft excel "solver" method (Bardi *et al.*, 2005). The network-based methodology considers all the information in the matrix simultaneously, in an integral fashion. Since there is no fixed sequential order for solving the equations, the occurrence of feedbacks does not prejudice the performance of the operation, allowing feedback flows to be accounted as inputs in the energy transformation equations. Although this method assumes steady state conditions, by adding the feedback information into the matrix, the temporal adjustments of the UEVs are "included" in the computational process. Since the ecosystem was aggregated into 11 compartments the transformities represent averages for each compartment.

Solar transformities for a more complex network, the Florida Everglades graminoid marsh (**Table 4**) were evaluated using the same technique (Brown *et al.*, 2006). Network data, consisting of carbon flows (g C year^{-1}) were compiled from published data (details for accounting, aggregation and assumptions in Ulanowicz *et al.* (1997, 2000). There were 66 ecosystem compartments in the graminoid marsh; some of which (ecosystem pools of labile and refractory detritus) were not living. The primary production compartments were partitioned into root, and leaf compartments. Many of the lower trophic level compartments represent aggregations of species (due to lack of data); for example, mesoinvertebrates, macroinvertebrates, centrarchid fish, snakes, and passerine birds are lumped categories for the Everglades system.

Emergy and Biodiversity

In practice, the conservation of biodiversity suggests sustaining the diversity of species in ecosystems as we plan human activities that affect ecosystem health. Biodiversity has no single standard definition. Generally speaking, biodiversity is a *measure of the relative diversity among organisms present in different ecosystems*. "Diversity" in this case includes diversity within species (i.e., genetic

	Unit Emergy Value (sej/gC)			2.96E+09	2.07E+10	6.43E+09	3.70E+10	2.00E+08	5.60E+07	2.06E+10	2.06E+10	2.52E+11	4.72E+10	3.72E+10
	Unit Emergy Value (sej/J)			7.39E+04	5.18E+05	1.61E+05	9.25E+05	5.00E+03	1.40E+03	5.15E+05	5.15E+05	6.29E+06	1.18E+06	9.29E+05
	Input (sun)	Input H_2O	Organic Matter	Algae	Spartina	Juncus	Crabs, shrimp, snails	Detritus	Microbes & Bacteria	Macro-Invertebrates	Insects	Mammals	Fish	Birds
Algae	3.31E+10	1.90E+11		–100					1.0					
Spartina	1.65E+11	4.80E+11			–497.5				5.0					
Juncus	6.62E+11	1.90E+12				–2310			20.2					
Crabs, shrimp, snails				25.00	82.50	330.00	–308.1			3.3				
Detritus			2.69E+10	50.00	250.00	1320.00	264.5	–1524.1		78.7	249.5	264.0	116.9	27.0
Microbes & Bacteria								262.1	–26.2					
Macro Invertebrates								131.2		–91.7				
Insects					82.50	330.00				3.3	–332.6			
Mammals					82.50	330.00	14.5			3.3		–307.1	9.74	
Fish				25.00			14.5	131.2		3.3	20.8		–155.70	
Birds							14.5				20.8		9.74	–31.6
Export								1.00E+03			41.58	43.03	19.48	4.51

Fig. 7 Square matrix of inputs and outputs from each compartment used to compute UEVs of the Juncas marsh ecosystem. Data in the top two rows are computed UEVs in sej g C^{-1} and sej J^{-1}.

Table 4 Computed transformity values for graminoid marsh matrix in rank order of transformity

Ecosystem component	*Transformity (sej J^{-1})*	*Ecosystem component*	*Transformity (sej J^{-1})*
Periphyton	3.56E + 03	Small frogs	1.07E + 05
Labile detritus	6.43E + 03	Muskrats	1.08E + 05
Flagfish	1.24E + 04	Medium frogs	1.09E + 05
Floating vegetation	1.25E + 04	White tailed deer	1.10E + 05
Utricularia spp.	1.51E + 04	Salamander larvae	1.11E + 05
Living sediments	1.59E + 04	Catfish	1.18E + 05
Other macroinvertabrates	1.91E + 04	Gruiformes	1.28E + 05
Apple snail	1.97E + 04	Large frogs	1.30E + 05
Mesoinverts	2.02E + 04	Alligators	1.34E + 05
Macrophytes	2.08E + 04	Spotted sunfish	1.36E + 05
Poecilids	2.23E + 04	Cichlids	1.36E + 05
Lizards	2.67E + 04	Warmouth	1.38E + 05
Tadpoles	2.91E + 04	Rabbits	1.41E + 05
Crayfish	3.04E + 04	Other large fishes	1.45E + 05
Freshwater prawn	3.87E + 04	Turtles	1.51E + 05
Bluefin killifish	4.14E + 04	Largemouth bass	1.52E + 05
Chubsuckers	4.26E + 04	Snailkites	1.62E + 05
Mosquito-fishes	4.35E + 04	Raccoons	1.63E + 05
Other small fishes	4.37E + 04	Grebes	1.76E + 05
Shiners and minnows	5.35E + 04	Salamanders	1.79E + 05
Killifishes	5.68E + 04	Cape sable seaside sparrow	1.85E + 05
Large aquatic insects	6.37E + 04	Fishing spider	1.99E + 05
Terrestrial inverts	6.74E + 04	Passerines	2.13E + 05
Topminnows	7.51E + 04	Gar	2.17E + 05
Blue-spotted sunfish	8.38E + 04	Rats and mice	2.28E + 05
Pigmy sunfish	8.42E + 04	Bitterns	2.40E + 05
Opossum	8.63E + 04	Otter	4.00E + 05
Dollar sunfish	8.71E + 04	Mink	4.38E + 05
Redear sunfish	8.83E + 04	Nighthawks	5.39E + 05
Snakes	9.62E + 04	Panthers	1.35E + 06
Ducks	1.01E + 05	Bobcat	3.30E + 06
Other centrarchids	1.03E + 05		

diversity), among species, and among ecosystems. Another definition, is simply the totality of genes, species, and ecosystems of a region.

A main problem with quantifying biodiversity, especially in light of the definition above, is that there is no overall measure of biodiversity since diversity at various levels of an ecological hierarchy cannot be summed. If they were summed, bacteria and other small plants and animals, would dominate the resulting diversity to the total neglect of the larger species. Also consider that evaluating species importance based on biomass or available energy throughput alone, without adjusting for quality, will tend to

Table 5 System scale indices of biodiversity for the Everglades graminoid marsh in wet season

Exogenous emergy inputs[a] *(sej ha^{-1} year^{-1})*	*5.2E11 (wet season)*
Quality adjusted diversity (bits)	1.73
Theoretical maximum diversity (bits)	4.14
Relative diversity (%)	42%

[a]Exogenous emergy inputs are from Brown and Bardi (2001).

dramatically underestimate the system-scale control potential of energy flows in upper trophic levels, since only a small fraction of total system physical throughput is incorporated at these levels. That is, the energetic importance of upper trophic levels is small compared with their actual role in ecosystem function, which includes cybernetic control (e.g., control of population at lower trophic levels, seed dispersal, commensal relationships, and ecosystem structural attributes). Odum (1996) argues that, methods for energy flow analysis should incorporate the relative energetic contributions (i.e., importance value) of each component by adjusting for transformity (quality) to avoid misrepresenting their influence. It therefore may be possible to develop a quantitative evaluation of total biodiversity within regions or ecosystems by weighting biodiversity at each hierarchical level by typical trophic level transformities (see, for instance **Table 5**).

An ecosystem level, quality-adjusted diversity can be computed in the typical manner, as the Shannon Diversity index (Eq. **4**).

$$H = -\sum_{i=1}^{j} p_i * \log [p_i] \tag{4}$$

except relative importance value (p_i in Eq. **4**) is defined as the proportion of total system emergy flow (sej year^{-1}) allocated to each component. The relative value calculated in this manner is an emergy importance value (EIV) computed as follows:

$$\mathrm{EIV}_i = \frac{\mathrm{NP}_i * \tau_i}{\sum_j \mathrm{NP}_j * \tau_j} \tag{5}$$

where EIV_i is the emergy importance value of component i, NP_i is the net production (J year^{-1}) and, τ_i(sej J^{-1}) is the computed transformity of component *i*.

In this equation, importance value is the relative contribution of each component to the total emergy flow through all biotic components (i.e., denominator of Eq. **5**), computed by summing net production times derived transformity (τ_j) over all components. It follows then that an the Shannon diversity index when modified to include the emergy importance value (EIV) becomes an index of biodiversity as follows:

$$\text{Biodiversity} = -\sum_{i=1}^{j} \mathrm{EIV}_i * \log [\mathrm{EIV}_i] \tag{6}$$

When physical flows are adjusted for quality, (i.e., expressed as emergy) evenness of flow across all ecosystem components results; in fact, emergy flow evenness is the postulated goal for network systems that are maximizing power (Odum, 1994). Therefore, maximum possible value for this biodiversity index occurs when the emergy on each pathway and each component's EIV, is equal. **Table 5** summarizes quality-adjusted diversity, the theoretical maximum diversity and the relative diversity as a percent of maximum for the Everglades graminoid marsh whose UEVs were given in **Table 4** above. If the observed diversity is compared to maximum diversity an index of ecosystem condition results since all things being equal, maximum potential diversity represents the highest ecological condition. Our evaluation of the Everglades marsh ecosystem suggests that it is operating at 42% of its maximum potential diversity.

Summary Comments

Emergy and Unit Emergy Values

Emergy and UEVs are useful measures that may be applied to systems of all scales to better understand hierarchy, control, development, stress, and ultimately ecosystem health. UEVs measure the convergence of biosphere work into processes and products of ecosystems and offers the opportunity to scale ecosystems and their parts based on the energy required to develop and maintain them. Systems are hierarchically organized and composed of physical structure (i.e., wood, biomass, detritus, animal tissue, etc.) and information found not only in genetic makeup of components but in the relationships and connections between individuals and groups of individuals. Because of the demands of thermodynamic costs with each transformation, the energy, materials, and information associated with higher hierarchical levels of systems are relatively small in comparison with lower levels. A failure to account for quality differences can result in an over emphasis of the importance of lower order components at the expense of the importance of higher order components and their control actions. Emergy Intensities may offer a useful indicator of the cybernetic role of components within ecosystems.

Emergy, Transformity, and Network Analysis

UEVs are a measure of the importance of components within systems (Odum, 1996). We believe that network analyses (Ulanowicz, 1986; Dame and Patten, 1981), while offering some insights into overall cybernetic control within systems, still requires the addition of emergy refinements to address the quality of particular services provided by each species in a system. Ulanowicz (1980), recognized the need to weight flow diversity at the system scale by total system throughput to make comparison between systems meaningful. We suggest that without quality correction (UEVs), cross-trophic comparison of importance within systems is somewhat hindered because of the inverse relationships between energy flows and trophic level. Further, comparison of development stage and/or stress between systems cannot be realized without accounting for quality of compartments.

Biodiversity

MacArther's (1955) original conceptualization of Shannon diversity was intended for determinacy of *flows* within systems, suggesting that the flows of energy and materials between components were a form of "information transfer". However, subsequent applications of Shannon diversity in ecology replaced flows with physical *stocks*. Margalef (1961) and Ulanowicz (2001) discussed the reasons and drawbacks for this convenient, but unfortunate tangent. As a result of the use of stocks (i.e., counts of individuals) the application of the Shannon diversity index have generally failed to effectively predict ecological properties (e.g., stability, resilience, or productivity).

The use of stocks instead of flows illuminates a second weakness in the Shannon diversity. The standard Shannon diversity metric ignores ecosystem food web hierarchy. Given a fixed number of ecosystem components, the Shannon diversity is maximum when the probability of observing each component is equal; that is, evenness in physical *stocks* increases diversity. However, given typical trophic transfer efficiencies (i.e., Lindeman efficiencies), maximum ecosystem diversity as maximum evenness of ecosystem physical stocks leads to obvious contradictions and as a result, Shannon diversity using physical stocks maybe only appropriate within a single trophic level and should not be used at the ecosystem scale. While we advocate the replacement of stocks with flows as per MacArthur's (1955) original intent, a similar argument against computing diversity using physical flows (energy, carbon) can be made. That is, evenness of physical flows is not expected for an entire food web because the energy/carbon throughput decreases geometrically with increasing trophic level because of thermodynamic limitations of transfers. Thus we advocate adjusting flows by quality factors (UEVs). Once corrected for quality, evenness can be expected.

While biodiversity assessments at the ecosystem scale have traditionally taken the form of species catalogs, sorted by different classes, they seldom address trophic interactions that stimulate the emergence of ecosystem properties. With the use of emergy combined with network analysis to derive whole ecosystem measures of importance and diversity the result is a new way to calculate across trophic level ecosystem diversity that takes into account trophic interactions.

There is an obvious limitation of this method of assigning UEVs to system components, in that the required level of specificity about compartments and flow details between them is not usually found in ecosystem studies. However this level of detail is required by MacArthur's (1955) original effort to link information theory with ecosystem organization. This approach is consistent with MacArthur's original effort since it uses flows instead of stocks, and is further extended by recognizing the need for quality adjustment of flows. Such adjustment better recognizes the cybernetic control of ecosystem processes by higher order components. UEVs offer a useful indicator of this cybernetic role within ecosystems.

See also: Ecological Complexity: Goal Functions and Orientors; Hierarchy Theory in Ecology; Complex Ecological Networks; Systems Ecology; Systems Ecology: Ecological Network Analysis. Global Change Ecology: Energy Flows in the Biosphere. Human Ecology and Sustainability: Emergy and Sustainability; System Sustainability; Ecological Systems Thinking

References

Bardi, E., Cohen, M.J., Brown, M.T., 2005. In: Brown, M.T., Bardi, E. (Eds.), A linear optimization method for computing transformities from ecosystem energy webs. Proceedings of the Third Biennial Emergy Conference, Gainesville, FL.

Bastianoni, S., Morandi, F., Flaminio, T., Pulselli, R.M., Tiezzi, E.B.P., 2011. Emergy and emergy algebra explained by means of ingenuous set theory. Ecological Modelling 222, 2903–2907. doi:10.1016/j.ecolmodel.2011.05.013.

Brown, M.T., Bardi, E., 2001. Handbook of Emergy Evaluation Folio 4: Emergy of Ecosystems. Gainesville, FL: The Center for Environmental Policy, University of Florida, p. 93.

Brown, M.T., Herendeen, R., 1996. Embodied energy analysis and emergy analysis: A comparative view. Ecological Economics 19, 219–235.

Brown, M.T., Ulgiati, S., 2018. Emergy and environmental accounting; Coupling systems of humanity and nature. New York: Springer, (forthcoming).

Brown, M.T., Cohen, M.J., Bardi, E., Ingwersen, W.W., 2006. Species diversity in the Florida Everglades, USA: A systems approach to calculating biodiversity. Aquatic Sciences 68 (3), 254–277.

Brown, M.T., Protano, G., Ulgiati, S., 2011. Assessing geobiosphere work of generating global reserves of coal, crude oil, and natural gas. Ecological Modelling 222 (3), 879–887.

Brown, M.T., Campbell, D.E., De Vilbiss, C., Ulgiati, S., 2016. The geobiosphere emergy baseline: A synthesis. Ecological Modelling 339, 92–95.

Cavalett, O., Ortega, E., Brown, M.T., Bardi, E., Campbell, D.E., Comar, V., Huang, S., Rydberg, T., Tilley, D., Ulgiati, S. (Eds.), 2007. Using the values of internal emergy flows for emergy accounting in agricultural complex systems.Emergy Synthesis 4: Theory and Applications of the Emergy Methodology. Proceedings of the 4th Biennial Emergy ConferenceGainesville, FL: Center for Environmental Policy/University of Florida, p. 483.

Cohen, M., 2003a. Dynamic Emergy simulation of soil genesis and techniques for estimating transformity confidence envelopes. In: Brown, M.T., Odum, H.T., Tilley, D.R., Ulgiati, S. (Eds.), Emergy synthesis 2: Theory and applications of the emergy methodology. Gainesville, FL: Center for Environmental Policy, pp. 335–370.

Cohen, M. 2003b. Spatial analysis of emergy and economic variables and their relationship to soil erosion in the Lake Victoria Basin, Ph.D. Dissertation. University of Florida, Gainesville, FL.
Collins, D., Odum, H.T., 2000. In: Brown, M.T., *et al.* (Eds.), Calculating transformities with an eignevector method.Emergy Synthesis 1; Proceedings of a conference held in Gainesville FloridaGainesville, FL: Center for Environmental Policy/University of Florida, pp. 265–280. September 1999.
Dame, R.F., Patten, B.C., 1981. Analysis of energy flows in an intertidal oyster reef. Marine Ecology Progress Series 5, 115–124.
Lotka, A.J., 1922. Contribution to the energetics of evolution. Proceedings of the National Academy of Sciences of the United States of America 8, 147–151.
MacArthur, R., 1955. Fluctuations of animal populations and a measure of community stability. Ecology 36, 533–536.
Margalef, R., 1961. Communication of the structure of planktonic populations. Limnology and Oceanography 6, 124–128.
Morandi, F., Campbell, D.E., Pulselli, R.M., Bastianoni, S., 2013. Using the language of sets to describe nested systems in emergy evaluations. Ecological Modelling 265, 85–98. doi:10.1016/j.ecolmodel.2013.06.006.
Odum, H.T., 1971. Environment, power and society. New York: Wiley, 336 pp.
Odum, H.T., 1983. Maximum power and efficiency: A rebuttal. Ecological Modelling 20, 71–82.
Odum, H.T., 1994. Ecological and general systems: An introduction to systems ecology. Niwot: University Press of Colorado, 644 pp.
Odum, H.T., 1996. Environmental accounting. Emergy and environmental decision making. New York: Wiley.
Odum, H.T., Collins, D., 2003. Transformities from Ecosystem Energy Webs with the Eigenvalue Method. Proceedings of the Second Biennial Emergy Conference, Gainesville, FL.
Odum, H.T., Odum, E.C., 2001. A prosperous way down: Principles and policies. Boulder, CO: University Press of Colorado.
Patterson, M., 1983. Estimations of the quality of energy sources and uses. Energy Policy 2 (4), 346–359.
Tilley, D.R. 1999. Emergy basis of forest systems, Ph.D. Dissertation. Gainesville, FL: University of Florida, pp. 296.
Tilley, D.R., 2011. Dynamic accounting of emergy cycling. Ecological Modelling 222, 3734–3742. doi:10.1016/j.ecolmodel.2011.09.007.
Ulanowicz, R.E., 1980. A hypothesis on the development of natural communities. Journal of Theoretical Biology 85, 223–245.
Ulanowicz, R.E., 1986. Growth and development: Ecosystem phenomenology. New York: Springer Verlag.
Ulanowicz, R.E., 2001. Information theory in ecology. Computers and Chemistry 25, 393–399.
Ulanowicz, R. E., C. Bondavalli and M. S. Egnotovich, 1997. Network analysis of trophic dynamics in South Florida ecosystems FY96: The cypress wetland ecosystem. Annual Report to USGS/BRD, Coral Gables, FL, Retrieved January 30, 2005 http://cbl.umces.edu/~atlss/cyp701.html.
Ulanowicz, R. E., J. J. Heymans and M. S. Egnotovich, 2000. Network analysis of trophic dynamics in South Florida ecosystems FY99: The graminoid ecosystem. Annual Report to USGS/BRD, Coral Gables, FL, Retrieved January 30, 2005. From: http://cbl.umces.edu/~atlss/swgras701.html.
Williams, R.J., Martinez, N.D., 2000. Simple rules yield complex food webs. Nature 404 (6774), 180.
Winfrey, B.K., Tilley, D.R., 2013. In: Brown, M.T., Sweeney, S., Campbell, D.E., Huang, S., Kang, D., Rydberg, T., Tilley, D., Ulgiati, S. (Eds.), Comparison of principles for allocating emergy to recycled materials in natural and managed ecosystems.Emergy Synthesis 7: Theory and Applications of the emergy Methodology, Proceedings of the 7th Biennial Emergy ConferenceGainesville, FL: Center for Environmental Policy/University of Florida, p. 586.
Zarba, L., Brown, M.T., 2015. Cycling emergy. Computing emergy in trophic networks. Ecological Modelling 315, 37–45.

Further Reading

British Geological Survey, 2007. World mineral production 2001–2005. Nottingham: British geological Survey/Kenworth, p. 81.
British Petroleum, 2007. BP statistical review of world energy, 2000. London: The British Petroleum Company, 41 pp.
Brown, M.T., McClanahan, T., 1996. Emergy analysis perspectives for Thailand and Mekong River dam proposals. Ecological Modelling 91, 105–130.
Brown, M.T., Ulgiati, S., 1999. Emergy evaluation of the biosphere and natural capital. Ambio 28 (6), 486–493.
Brown, M.T., Ulgiati, S., 2002. Emergy evaluation and environmental loading of electricity production systems. Journal of Cleaner Production 10, 321–334.
Brown, L.R., Renner, M., Flavin, C., 1997. Vital signs, 1997. The environmental trends that are shaping our future. New York: W.W. Morton & Co., p. 165.
Campbell, D., 1998. Emergy analysis of human carrying capacity and regional sustainability: An example using the state of Maine (appendix). Environmental Monitoring and Assessment 51, 531–569.
Doherty, S.J. 1995. Emergy evaluations of and limits to forest production, Ph.D. Dissertation. Gainesville, FL: Environmental Engineering Sciences/University of Florida, pp. 215.
Giampietro, M., Ulgiati, S., Pimentel, D., 1997. Feasibility of large-scale biofuel production. Does an enlargement of scale change the picture? Bioscience 47 (9), 587–600.
Kinsman, B., 1965. Wind, waves, their generation and propagation on the ocean surface. Englewood Cliffs, NJ: Prentice-Hall, p. 676.
Leith, H., Whittaker, R.H., 1975. Primary productivity of the biosphere. New York: Springer-Verlag.
Mannion, A.M., 1995. Agriculture and environmental change: Temporal and spatial dimensions. New York: Wiley, p. 405.
Miller, G.A., 1966. The flux of tidal energy out of the deep oceans. Journal of Geophysical Research 71, 2485–2489.
Odum, H.T., 1973. Energy, ecology and economics. Royal Swedish Academy of Science. Ambio 2 (6), 220–227.
Odum, H.T., 2000. Handbook of Emergy Evaluation: A Compendium of Data for Emergy Computation Issued in a Series of Folios. Folio #2: Emergy of Global Processes. Gainesville, FL: Center for Environmental Policy/Environmental Engineering Sciences/University of Florida, 30 pp.
Odum, E.C., Odum, H.T., 1984. System of ethanol production from sugarcane in Brazil. Ciencia e Cultura 37 (11), 1849–1855.
Odum, H.T., Brown, M.T., Williams, S.B., 2000. Handbook of Emergy Evaluation: A Compendium of Data for Emergy Computation Issued in a Series of Folios. Folio #1: Introduction and Global Budget. Gainesville, FL: Center for Environmental Policy/Environmental Engineering Sciences/University of Florida.
Oldeman, L.R., 1994. The global extent of soil degradation. In: Greenland, D.J., Szabolcs, I. (Eds.), Soil resilience and sustainable land use. Wallington: CAB International, p. 561.
Pimentel, D., Pimentel, M., 1979. Food, energy and society. New York: Wiley, pp. 165.
Pimentel, D., Wen, D., 1990. Technological changes in energy use in U.S. agricultural production. In: Carrol, C.R., Vandermeer, J.H., Rosset, P.M. (Eds.), Agroecology. New York: McGraw Hill, pp. 147–164.
Pimentel, D., Hurd, L.E., Bellotti, A.C., Forster, M.J., Oka, I.N., Sholes, O.D., Whitman, R.J., 1973. Food production and the energy crisis. Science 182, 443–449.
Pimentel, D., Warnaeke, A.F., Teel, W.S., Schwab, K.A., Simcox, N.J., Ebert, D.M., Baenisch, K.D., Aaron, M.R., 1988. Food versus biomass fuel: Socioeconomic and environmental impacts in the United States, Brazil, India and Kenya. Advances in Food Research 32, 185–238.
Ryabchikov, A., 1975. The changing face of the earth. Translated by J. Williams Moscow: Progress Publishers, 203 pp.
Sclater, J.F., Taupart, G., Galson, I.D., 1980. The heat flow through the oceanic and continental crust and the heat loss of the earth. Reviews of Geophysics and Space Physics 18, 269–311.
USDI. 2007. Mineral commodity summaries. US Department of Interior, Washington, DC. January 2007. http://minerals.usgs.gov/minerals/pubs/commodity/lime/lime_mcs06.pdf. Web download 2/25/07.
Wiin-Nielsen, A., Chen, T., 1993. Fundamentals of atmospheric energetics. New York: Oxford Press, p. 376.
World Resources Institute, 1996. World resources 1996–1997. Oxford: Oxford University Press.

Environmental Protection and Ecology☆

Clive Hamilton and Andrew Macintosh, Australian National University, Canberra, ACT, Australia
Nicoletta Patrizi and Simone Bastianoni, University of Siena, Siena, Italy

What Is Environmental Protection

Environmental protection can be defined as the prevention of unwanted changes to ecosystems and their constituent parts. This includes

- the protection of ecosystems and their constituent parts from changes associated with human activities; and
- the prevention of unwanted natural changes to ecosystems and their constituent parts.

One issue associated with this definition is whether "ecosystems and their constituent parts" include humans and communities, or whether environmental protection is only concerned with the protection of natural capital. From an ecological perspective, humans are regarded as an integral part of the ecosystem. Separating humanity from the natural environment can therefore be seen as artificial. While this is true, the phrase environmental protection is not used to refer to measures that are designed to regulate or mediate direct interaction between people. For example, laws prohibiting assault are not regarded as environmental protection measures. Environmental protection is concerned with the relationship between people and the natural environment rather than the relationships between people and communities.

Another issue is whether environmental protection relates to preservation, conservation, or both. Preservation refers to the protection of an ecosystem or natural environment from change, while conservation is generally associated with the sustainable use of natural resources. The objective of conservation is to ensure the maintenance of a stock of renewable resources that is being exploited for human purposes rather than the protection of the natural environment from any anthropogenic modifications. The exploitation of natural resources for human purposes is not environmental protection as it is not associated with the prevention of unwanted changes. The change associated with exploitation is deliberate and wanted, at least by those doing the exploitation. However, measures that are put in place to prevent overexploitation of natural resources do constitute environmental protection. They are designed to prevent exploitation beyond a point that is deemed desirable or sustainable. For example, catch quotas in fisheries and air pollution limits are environmental protection measures because, while they accept some environmental degradation, they aim to limit it.

The distinction between preservation and conservation has dissipated in recent years with growing recognition of the dynamic nature of natural systems, humanity's place in the biosphere, and the need for active human involvement to maintain the integrity of certain ecosystems. Consequently, environmental protection is now generally used to refer to measures that have traditionally been associated with preservation (e.g., reserves, including national parks), as well as conservation and natural resource management initiatives.

A critical aspect of environmental protection is that it is driven by the values that humans attribute to different aspects of the environment. These values need not be instrumental, but the motivating factor for environmental protection is always the prevention of changes to the environment that humans do not want. This is why measures associated with the prevention of unwanted natural changes to ecosystems—like the prevention of coastal erosion or systematic burning in reserves to reduce the risk of wildfires—can be included as environmental protection. Such measures do not aim to protect ecosystems from human activities but rather from natural forces that are deemed to threaten human interests.

☆*Change History:* March 2018. Simone Bastianoni and Nicoletta Patrizi updated the article. Recent initiatives on the field of Environmental Protection and Ecology have been added in the Conclusion section as well as their references.

Environmental remediation is distinct from environmental protection as its primary objective is to restore an ecosystem or natural environment to a previous state; that is, like exploitation, it is associated with deliberately induced change, as opposed to the prevention of change.

Domestic Policy Instruments

There are a number of ways to classify domestic policy instruments that are used to protect the environment. Some people divide them into voluntary and mandatory instruments, while others place them into regulation-based and incentive-based, or command-and-control and economic instruments. The method preferred here is to divide them into four broad categories: regulatory, economic, voluntary, and education and information.

Regulatory Instruments

Regulatory instruments impose legally enforceable restrictions on economic agents to realize environmental protection objectives. They are sometimes referred to as "command-and-control" mechanisms because they prohibit or mandate certain actions (i.e., the command), while using various forms of punishment to motivate compliance (i.e., the control mechanism). Environmental regulations can take many different forms, including

- prohibitions on specified activities (e.g., discharging pollutants into a water body or the atmosphere, or taking a threatened species);
- requirements to obtain a governmental approval (or permit) before undertaking a specified activity (e.g., pollution permits, operating licenses, and development approvals);
- requirements to follow certain procedures when carrying out specified activities (e.g., to use certain equipment, abide by operating standards, or to monitor pollution emissions); and
- requirements to undertake specified actions that are deemed to be environmentally beneficial (e.g., weed control in agricultural areas).

The use of regulatory instruments can be justified on the basis of deterrence or similar choice theories of criminology. According to this approach, people are assumed to make choices by rationally weighing the costs and benefits associated with alternative courses of action and selecting the action that is most likely to maximize their utility. Regulations can affect this process by altering the costs and benefits associated with environment-related activities. For example, by outlawing the emission of pollutants into a river and incarcerating or imposing fines on people who violate the law, the government is able to increase the costs of emitting pollutants. In doing so, it makes alternative nonpolluting options more attractive. In countries where the rule of law prevails, provided the penalties and probability of enforcement are high enough, and punishment is swift, the desired pattern of behavior should emerge.

The difficulty with this theory is that humans are not always rational utility maximizers, meaning that environment regulations may not always result in the desired outcomes. Regulatory approaches can also fail to achieve their objects because of poor design (e.g., ambiguous regulations and unworkable administrative arrangements), strategic avoidance by polluters, and an absence of monitoring and enforcement due to a lack of resources or political will. There are even cases where regulatory instruments have aggravated the environmental problems they were designed to solve.

Regulatory instruments are criticized not only for being ineffective, but also for their inefficiency. This is because they can impose inflexible restrictions on producers, thereby limiting the choices that are available as to how producers meet the desired environment target. Regulatory mechanisms can also have large administration and compliance costs. Government agencies are required to constantly monitor compliance and undertake costly litigation when breaches are discovered, while producers incur legal and other costs while attempting to abide by the environmental regulations. In addition, regulatory mechanisms do not provide incentives to encourage the reallocation of resources toward producers with the lowest marginal costs of environmental protection. As a result, it is difficult for regulatory mechanisms to satisfy the equimarginal principle, which requires that the marginal cost of environmental protection be equalized across polluters to achieve the desired environment target at the lowest possible cost.

The other main criticism of environment regulation is that it can be inequitable. Disputes about whether regulatory mechanisms are unfair are usually framed in terms of the alteration of preexisting property rights. Some people may believe they have a right to pollute or use the environment in a particular way. When environment regulations are introduced, these property rights may be taken away, prompting calls for compensation. In the absence of compensation, the people who are affected by the regulations may feel they are being forced to shoulder a disproportionate amount of the financial burden associated with the provision of public good environmental outcomes. This problem is often confronted with laws that prohibit the removal of native vegetation or the taking of native species, or that restrict the development of real property for commercial purposes.

In several countries disputes about the impact of environment regulations on property rights have been a central part of environmental policy debates. For example, in the United States, the Fifth Amendment to the Constitution provides that private property shall not be "taken for public use without just compensation." This has led to numerous Supreme Court cases and an extensive literature on so-called "regulatory takings." The Australian Constitution contains a similar provision, which provides the

Federal Government with the power to make laws with respect to the "acquisition of property on just terms." In both jurisdictions, courts have found that these constitutional provisions limit the extent to which the government can abolish property rights for environmental purposes.

While issues surrounding the abrogation of property rights have been influential in framing equity issues associated with environmental regulations, it is often forgotten that most regulations alter property rights. The elevated status of property right issues in the context of environmental debates is arguably due to the reverence accorded to real property and preferential treatment given to certain primary industries in western countries.

Despite the criticisms of regulatory measures, they remain the most widely used instrument for environmental protection. There is an ongoing debate about why states tend to prefer regulatory instruments, with supporters of economic instruments often attributing it to ignorance of alternatives, simplicity, and pressure by special interest groups. Yet, regulatory instruments have a number of attributes. The most important of these is the certainty they can provide when faced with imperfect information about environmental risks and the irreversibility of certain forms of environmental harm (e.g., species extinction). In these circumstances, relying on economic, voluntary, or information instruments can lead to suboptimal outcomes by permitting excessive exploitation of the environment. Regulatory instruments are well suited to these situations as they allow distinct boundaries to be placed on the use of environmental resources. This has led to the advocacy of decision-making tools like the precautionary principle and collective choice processes like the safe minimum standards approach.

Another benefit of regulatory instruments is they can be more cost-effective than other measures in dealing with certain types of environmental issues. As discussed, regulatory instruments are often criticized for having large administrative and compliance costs. Yet, in some cases, alternative economic and voluntary approaches may have higher administrative and compliance costs because of the nature of the environmental problem and the complexity in the required program. For example, addressing land clearing with voluntary measures can require agreements to be negotiated with large numbers of landholders, leading to high transaction costs and a plethora of different standards. In contrast, regulations can provide a uniform standard that does not require case-by-case negotiations and is easier to monitor and enforce. Regulatory instruments can also overcome free-rider problems by forcing recalcitrant polluters to abide by the necessary standards.

Economic Instruments (Market-Based Measures)

Economic instruments can be defined as mechanisms that force economic agents to internalize all or part of the social costs associated with environmentally harmful activities and that rely on market forces to promote efficiency. In doing so, they seek to impose additional costs on producers that harm the environment and reward those that improve environment outcomes, while utilizing market forces to improve the allocation of resources. (Some analysts include subsidies among economic instruments but as they are voluntary and economic agents are not forced to internalize the social costs they are more appropriately classified as voluntary instruments.)

This approach to environmental protection is usually associated with environmental economics, a school of economic thought that is a subdiscipline of neoclassical economics. According to environmental economists, environmental problems arise because of the existence of externalities—impacts involuntarily incurred by a person or persons without compensation or payment as a result of the actions of another. Because of the existence of externalities, markets are unable to guarantee the efficient allocation of resources. For example, if producers emit pollution into the atmosphere without paying for it, the price that consumers pay for the producers' outputs will not reflect the full social cost of the transaction. As a result, there will be excessive output and consumption of the relevant good or service. If producers are forced to internalize the social costs associated with the air pollution, there would be a more efficient tradeoff between air pollution and output, leading to higher net social welfare.

The more recent trend in environmental economics has been to characterize environmental problems as being a product of the incomplete allocation of property rights. According to this approach, if a property right over the relevant environmental resource were appropriately defined and allocated to individuals, and there was perfect information and no transaction costs, the operation of market forces would lead to efficient outcomes. For example, if the atmosphere were owned by someone and producers had to pay to emit pollution, then negotiation between the owner and producers would ensure the most efficient allocation of atmospheric resources. On the basis of these theories, economic instruments either

- require polluters to pay for all or part of the costs associated with pollution (e.g., pollution fees, individual liability, and removal of subsidies that promote overuse);
- place a restriction on the amount of pollution that can be emitted or resource that can be used and then allow pollution or resource entitlements to be traded among economic agents (called "marketable permit" or "cap-and-trade" schemes, e.g., tradable emission, water, catch and development rights schemes); or
- seek to create well-defined, secure, and transferable property rights over environmental resources and allocate these to relevant individuals or groups ("pure property rights" approaches, e.g., land titles and fishing area rights).

Marketable permit schemes and pure property rights approaches are similar in that both rely on the creation and exchange of property rights to promote environmental and economic outcomes. But pure property rights approaches place no external restrictions on the use of the relevant resource and rely on market incentives to achieve the desired environmental outcome,

while marketable permit schemes rely on a cap or limit on the use of the relevant resource to achieve the desired environmental outcome.

One of the major benefits associated with economic instruments is that by utilizing market forces they can encourage a more efficient allocation of resources. For example, when tradable emission quotas are used, the operation of market forces should ensure that the necessary emission reductions are achieved at least cost (i.e., the equimarginal principle should be satisfied). Further, economic instruments provide an incentive for producers to reduce pollution, which encourages innovation. Advocates of economic instruments also claim they are more flexible than regulatory instruments, although this is not always the case.

Although economic instruments can be more efficient than alternative policy mechanisms, they can suffer from a number of weaknesses. In relation to pollution fees, individual liability and pure property rights approaches, there can be a considerable amount of uncertainty associated with environmental outcomes. For example, producers may choose to absorb the increase in costs associated with a pollution fee, or demand may be unresponsive to price rises, meaning the level of pollution may not decline by the desired amount. Consequently, where policy-makers are faced with uncertainty regarding environmental risks and questions regarding irreversibility, alternative approaches can be preferable.

Like regulatory approaches, marketable permit schemes (or cap-and-trade approaches) can place an upper limit on the permissible amount of pollution or resource extraction. Hence, they can be useful in dealing with uncertainty and threshold effects. The advantage that marketable permit schemes offer is that having set a specified limit on pollution or resource extraction, they allow market forces to determine the allocation of pollution or extraction rights among producers. One of the most successful marketable pollution permit schemes has been the United States Environmental Protection Agency's Sulfur Dioxide Program, which is part of the broader Acid Rain Program. The cost of reducing emissions was substantially lower than predicted because producers had an incentive to find cheaper ways to do so.

Problems arise with marketable permit schemes when there is a lack of equivalence between the environment or pollution units that producers are expected to trade (i.e., the resource is not homogeneous). For example, tradable development permit schemes that place a limit on the amount of development in an area but allow developers to exchange development rights can lead to the rights moving toward the developments with the highest economic returns. However, they will not necessarily achieve biodiversity objectives as each parcel of land may contain different biodiversity values. Similar problems can arise with emission schemes that allow emission permits to be generated through the enhancement of sinks (i.e., there can be uncertainty about whether the enhancement of sinks will offset the additional emissions).

Transaction costs can also pose problems for economic instruments. Devising schemes that can be administered in a cost-effective manner can sometimes be difficult. Further, if there are excessive costs associated with the negotiation and exchange of marketable permits, the efficiency benefits may not materialize.

As with all environmental policy mechanisms, politics can impede the effective use of economic instruments. However, economic instruments can be especially vulnerable to political influences if it is necessary to constantly adjust the price signals provided through the scheme. For example, if a carbon tax is used to address climate change, it will be necessary to adjust the tax over time to account for unexpected events and new information. Special interest groups may impede this process, thereby undermining the efficacy of the tax.

There has been a tendency in the past for regulatory instruments and economic instruments to be presented as substitutes. In practice, these two types of instruments are generally used as complements and economic instruments always require a regulatory framework. Indeed, there is a growing recognition of the need for policy packages or policy mixes that use a range of instruments to achieve environmental protection objectives.

Voluntary Approaches

Voluntary approaches can be defined as any mechanism or program that aims to protect the environment where relevant economic agents are able to decide whether or not to participate; that is, involvement in the program is voluntary and no direct penalties are imposed on nonparticipants, although incentives may be used to encourage participation.

There are three broad types of voluntary approach.

- Unilateral initiatives where polluters act without direct government involvement to protect the environment. The defining features of unilateral initiatives are that they are initiated, designed, and operated by polluters. Hence, government involvement is generally limited, which raises questions about whether unilateral initiatives are a policy mechanism or a type of market behavior. Yet governments can encourage unilateral initiatives by suggesting them to polluters or threatening mandatory measures. There are three main types of unilateral approaches: voluntary adjustment of internal processes (e.g., under an environmental management plan); industry self-regulation (e.g., codes of conduct); and environmental certification schemes (e.g., organic producer associations).
- Bilateral agreements between the regulator and a polluter or group of polluters. These initiatives involve negotiation between the parties about how environmental protection will be achieved. Both parties have obligations under the agreement with polluters generally expected to meet certain targets and abide by conditions to protect the environment and the regulator generally expected to provide some sort of incentive. The incentives provided by regulators can include subsidies (e.g., monetary payments and technical assistance), public recognition, and undertakings not to enforce regulations or to introduce new regulations. The

agreements need not be legally binding, but there must be negotiation and some sort of understanding about the obligations of the parties.

- Voluntary public (or government) programs where the regulator determines who is eligible to participate, the obligations of participants, and the incentives used to encourage compliance. The key to these types of programs is that they are initiated and designed by the regulator, and relevant producers are invited and encouraged to participate. Again, the types of inducements include grants, technical assistance, and public praise.

The main advantages of voluntary approaches are that they are flexible (which provides polluters with the freedom to find cost-effective solutions) and noninterventionist. In addition, where disputes arise about the fairness of polluter-pays instruments like regulations and pollution fees, voluntary instruments can help overcome political resistance by enabling governments to pay polluters for the loss of property rights (i.e., they can resort to a beneficiary-pays approach). Due to these characteristics, voluntary approaches are often supported by polluters, which can assist in reducing the political costs for regulators. Further, it is sometimes claimed that voluntary approaches have lower administrative costs than other instruments and that in certain circumstances they can be more effective than mandatory approaches.

Although voluntary approaches can offer some benefits, because of the public good characteristics of many environmental goods and services (i.e., they are nonrival and nonexcludable) they are unlikely to result in an optimal level of environmental protection. In particular, there is a risk that some producers will seek to free-ride on the environmental protection measures undertaken by others. Like some economic instruments, voluntary mechanisms also lack certainty and are ill-suited to dealing with uncertainty and irreversibility.

Voluntary approaches can also be expensive to operate and administer. The incentives necessary to encourage participation can impose a significant burden on taxpayers. These problems can be exacerbated by gaming on behalf of polluters where they seek to take advantage of information asymmetries to extract excessive economic rent. In addition, the transaction costs associated with voluntary approaches can be high if there is a need to negotiate agreements with a significant number of producers.

Given the weaknesses associated with voluntary mechanisms, they are often seen as being used when political resistance blocks the introduction of more effective instruments or as a mechanism that supports or complements other programs.

Research has shown that if voluntary approaches are to be effective, the existence of a strong and credible threat of regulation is essential. The existence of an appropriate threat of regulation increases the incentive for polluters to participate and bolsters the bargaining position of regulators. It can also reduce the financial incentives needed to ensure participation, which can improve the cost-effectiveness of the program.

Information and Education Instruments

Information and education instruments aim to promote environmental protection by improving people's awareness and understanding of environment issues and building their capacity to respond to environmental threats. They include such things as environmental and sustainability reporting by governments and corporations, and advertising and education campaigns.

Many environmental problems arise at least partly because of imperfect information about environmental risks and a lack of awareness about how to respond. These types of instruments can help overcome these issues and can be effective where the threats to the environment are known and producers have an economic incentive to improve environmental outcomes. Information instruments can also be an important tool for encouraging greater support for environmental protection in the community, which can reduce the political costs associated with various environment policy instruments.

The main flaw associated with information instruments is that they will rarely get at the root causes of environmental degradation. As a result, on their own, they are generally an ineffective means of achieving environmental objectives. However, information instruments are commonly viewed as an essential part of environmental policy packages. Without information instruments it is very difficult for policy-makers to select appropriate policy instruments and it is unlikely that environmental protection measures will attract the political and community support that is necessary to ensure their success.

The advantages and disadvantages of the four approaches are summarized in the Table 1.

International Dimensions of Environmental Protection

Under international law, states have the sovereign right to exploit, manage, and conserve the natural resources and natural systems within their jurisdiction, including resources located in their territorial sea and exclusive economic zone, and sinks such as the atmosphere. States also have a broad right to engage in fishing on the high seas. However, the expansion of the world economy has placed increasing pressure on natural systems that overlap or transcend political boundaries. This has gradually led to the development of a large number of international agreements, systems, and processes to address transnational environmental issues.

The international governance system that has emerged over the past 60 years to facilitate environmental protection can be divided into three main parts: international environmental law, international environmental bureaucracy, and international environmental financial mechanisms.

Table 1 Domestic environmental policy instruments—pros and cons

Policy instrument	*Definition*	*Advantages*	*Disadvantages*
Regulatory	Instruments that impose legally enforceable restrictions on economic agents to realize environmental protection objectives	Certainty about environmental outcomes Ability to limit free-riding Clarity of standards—easy to comply with, monitor, and enforce No need to negotiate individual standards, which lowers administration costs	Inefficiencies—regulations impede the operation of market forces and can lead to the misallocation of resources Can stifle innovation Potentially inequitable because some may shoulder a disproportionate burden
Economic	Instruments that force economic agents to internalize all or part of the social costs associated with environmentally harmful activities and that rely on market forces to promote efficiency	By utilizing market forces they are able to achieve the desired outcomes in an efficient manner Promote innovation Provide flexibility as they often enable individuals to determine the best method of achieving the desired outcomes Ability to limit free-riding	Can lack certainty about environment outcomes Can be complex and have high administration costs Potentially inequitable
Voluntary	Any mechanism or program that aims to protect the environment where relevant economic agents are able to decide whether or not to participate	Noninterventionist, meaning they are likely to have high levels of acceptance by producers Low political costs for regulators High levels of flexibility	Lack certainty about environment outcomes Can have high administration costs Risk of free-riding Risk that producers will engage in gaming to extract excessive economic rents from regulators
Information and education	Instruments that aim to promote environmental protection by improving people's awareness and understanding of environment issues and building their capacity to respond to environmental threats	Noninterventionist Low political cost for regulators Flexibility Relatively low administration costs	Lack certainty about environment outcomes Risk of free-riding Inability to address main reasons for market failure

International Environmental Law

At the heart of the international environmental governance system lies the body of legal principles and agreements that collectively constitute international environmental law. Although international environmental agreements have existed for centuries, the number, scope, and complexity of these agreements has increased considerably since the 1940s. By the mid-2000s, there were more than 500 multilateral environment agreements (MEAs) in existence. Around 270 of these MEAs were broad international agreements, while the remainder had a regional focus and a relatively limited number of signatories. Not surprisingly, these agreements cover a wide range of topics, including climate change, biodiversity, transport and disposal of hazardous materials, and fisheries management.

One of the most basic principles of international environmental law is that, while states have sovereignty over the resources in their jurisdiction, no state has the right to use or permit the use of its territory in such a manner as to cause injury to the another's territory, person, or property. This principle, which is generally traced to the Trail Smelter Dispute that commenced in the 1920s between Canada and the United States, concerns environmental obligations between particular states (i.e., reciprocal obligations).

Growing awareness of the interrelated nature of natural systems and the scale of environmental problems in the later part of the twentieth century resulted in growing support for the notion that states should also have obligations to protect global commons and the interests of humanity, including future generations. This has led to the formation of a significant number of international agreements that promote the protection of a broader range of interests in the environment. The guiding principle regarding the rights of states to exploit natural resources is now viewed as incorporating a fundamental duty to protect global commons. For example, the Rio Declaration on Environment and Development (1992) provides that states have a "responsibility to ensure that activities within their jurisdiction or control do not cause damage to the environment of other States or of areas beyond the limits of national jurisdiction." There are also a number of international agreements, such as the United Nations Framework Convention on Climate Change (UNFCCC), that seek to protect certain aspects of the environment for "the benefit of present and future generations of humankind."

The notion of sustainable development has been a common element of MEAs since the late 1980s and 1990s. This has led to greater concern about intragenerational equity and encouraged the inclusion of the so-called principle of "common but differentiated responsibility" in MEAs. This principle has two parts. The first is that states have a shared responsibility for the protection of the environment, or relevant parts of it. The second part is that the extent to which each individual state is responsible for environmental protection should be determined with regard to the state's capacity to respond and historical contribution to the

relevant problem. One of the clearest articulations of the principle is found in the Rio Declaration on Environment and Development (1992), where it provides that

> States shall cooperate in a spirit of global partnership to conserve, protect and restore the health and integrity of the Earth's ecosystem. In view of the different contributions to global environmental degradation, States have common but differentiated responsibilities. The developed countries acknowledge the responsibility that they bear in the international pursuit to sustainable development in view of the pressures their societies place on the global environment and of the technologies and financial resources they command.

The principle of common but differentiated responsibility also features prominently in the UNFCCC (1992) and the Montreal Protocol on Substances that Deplete the Ozone Layer (Montreal Protocol) (1987). For example, Article 3(1) of the UNFCCC states that

> The Parties should protect the climate system for the benefit of present and future generations of humankind, on the basis of equity and in accordance with their common but differentiated responsibilities and respective capabilities.

International Environmental Bureaucracy

The second element of the international governance system is the collection of international bodies whose functions include the oversight of environmental issues. At the heart of these is the United Nations, which has established a collection of agencies who are responsible for developing policy and promoting better environmental stewardship. These include the United Nations Environment Programme (UNEP), United Nations Development Programme (UNDP), Commission on Sustainable Development (CSD), and the United Nations Educational, Scientific, and Cultural Organisation (UNESCO). The agencies associated with the United Nations are complemented by other international and regional bodies like the World Bank, Organisation for Economic Cooperation and Development (OECD), European Union, and the Organisation of American States (OAS) that play a role in the development and implementation of environment policy.

International Environmental Financial Mechanisms

The final element of the international governance system is the financial mechanisms that have developed to support the work of international environment agencies and to assist in the achievement of international environmental objectives. These range from the general (e.g., the financial arrangements governing the World Bank and United Nation) through to the more specific, which concentrate solely on environment issues. The Global Environment Fund (GEF), established by the OECD in the early 1990s, is one example. Consistent with the principle of common but differentiated responsibility, many MEAs now include provisions requiring developed countries to transfer technology and financial resources to developing countries to assist them to meet their environmental obligations. One of the earliest examples was the Multilateral Fund established under the Montreal Protocol to assist developing countries phase out the use of ozone-depleting substances. In some cases, including under the Montreal Protocol, the obligation of developing countries to comply with the terms of the agreement has been made contingent on the extent to which developed countries provide the specified financial and technical assistance. For example, Article 4(7) of the UNFCCC states that

> The extent to which developing country Parties will effectively implement their commitments under the Convention will depend on the effective implementation by developed country Parties of their commitments under the Convention related to financial resources and transfer of technology . . .

Despite considerable progress being made in some areas, the international environmental governance system has generally failed to bring about substantial and sustained changes in the stewardship of natural resources and environmental systems. In many cases, this is due to the difficulties associated with designing agreements and systems that can accommodate the divergent interests of the states that are involved in transnational environmental issues. Negotiations can be tediously slow and the need to reach consensus can lead to lowest common denominator outcomes. Similarly, due to the reluctance of states to place constraints on their sovereign rights over natural resources, it has been difficult to establish appropriate mechanisms for monitoring compliance and enforcing the terms of the agreements. The international governance system has also been hampered by a lack of resources for key institutions and programs.

Conclusion

Environmental protection has always been practiced by humans in one form or another. However, as anthropogenic pressures on the environment have escalated over the past century, the need for systematic environmental protection has increased. This has led to considerable experimentation with the domestic and international measures that are used to achieve environmental protection objectives. Some of these have been successful, but the overall picture is one of failure.

Due to the failings of the past and greater awareness of the complexity of environmental problems, there is a growing acceptance that environmental protection is best achieved through the use of a multipronged approach. This requires the use of a combination of regulatory, economic, voluntary, and information instruments, where the policy mix is determined on the basis of the available evidence regarding cost-effectiveness.

The international challenge lies in the development of effective and equitable approaches to global environmental problems that are supported by a well-resourced bureaucracy and appropriate financial mechanisms.

However, in the last few years, a new initiative has come out bringing together recent scientific advancements in our knowledge of the Earth System functioning, environmental and ecological law as well as economic instruments to propose a new global governance system aimed at favoring environmental protection. In 2009 a group of scientists developed the "Planetary Boundaries" framework to measure and monitor the state and functioning of the Earth System (Rockström et al. 2009; Steffen et al. 2015). The framework, has been developed to define a planetary safe operating space within which humanity can survive and thrive, and to highlight risks that destabilization of the system creates for human well-being. The identification of such safe operating space for humanity has then helped to recognize the need of an evolution of the current international legal system. In this light the concept of "The Common Home of Humankind" has been proposed as a social construct, based on legal solutions to represent the global natural reality (Magalhães et al., 2016). The legal recognition of a specific, functioning state of the Earth System (i.e., a Holocene-like state, as defined by the planetary boundaries framework) as a common natural intangible heritage of mankind would allow to include all positive and negative externalities in the governance and maintenance of the Earth System thus possibly constituting a new legal environmental protection framework.

That legal systems are not ready to face these challenges is shown also by the fact that a group of law experts have launched the Ecological Law & Governance Association (ELGA) to reframe law according to ecological limits. In their *Oslo Manifesto* it is reported that "Ecological Law requires human activities and aspirations to be determined by the need to protect the integrity of ecological systems. Ecological integrity becomes a precondition for human aspirations and a fundamental principle of law. In other words, ecological law reverses the principle of human dominance over nature, which the current iteration of environmental law tends to reinforce, to a principle of human responsibility for nature. This reversed logic is arguably the key challenge of the Anthropocene".

Further Reading

Birnie P and Boyle A (2002) *International law and the environment*, 2nd ed. Oxford, UK: Oxford University Press.

Gunningham N and Sinclair D (1999) Regulatory pluralism: Designing policy mixes for environmental protection. *Law and Policy* 21: 49–76.

Karamanos P (2001) Voluntary environmental agreements: Evolution and definition of a new environmental policy approach. *Journal of Environmental Planning and Management* 44: 67–84.

Keohane N, Revesz R, and Stavins R (1998) The choice of regulatory instruments in environmental policy. *Harvard Environmental Law Review* 22: 313–367.

Kolstad C (2004) *Environmental economics*. Oxford, UK: Oxford University Press.

Magalhães P, Steffen W, Bosselmann K, Aragão A, and Soromenho-Marques V (eds.) (2016) *SOS treaty—The safe operating space treaty a new approach to managing our use of the earth system*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Organisation for Economic Cooperation and Development (2003) *Voluntary approaches for environmental policy: Effectiveness, efficiency and usage in policy mixes*. OECD Paris: France2003.

Panayotou T (1994) Economic Instruments for Environmental Management and Sustainable Development, Environmental Economics Series, Paper No. 16. Nairobi: United Nations Environment Programme.

Rockström J, Steffen W, Noone K, Persson Å, Chapin FS III, Lambin EF, Lenton TM, Scheffer M, Folke C, Schellnhuber HJ, Nykvist B, de Wit CA, Hughes T, van der Leeuw S, Rodhe H, Sörlin S, Snyder PK, Costanza R, Svedin U, Falkenmark M, Karlberg L, Corell RW, Fabry VJ, Hansen J, Walker B, Liverman D, Richardson K, Crutzen P, and Foley JA (2009) A safe operating space for humanity. *Nature* 461: 472–475.

Steffen W, Richardson K, Rockström J, Cornell SE, Fetzer I, Bennett EM, Biggs R, Carpenter SR, de Vries W, de Wit CA, Folke C, Gerten D, Heinke J, Mace GM, Persson LM, Ramanathan V, Reyers B, and Sörlin S (2015) Planetary boundaries: Guiding human development on a changing planet. *Science* 347: 1–15. https://doi.org/10.1126/science.1259855.

Sunstein CR (1990) Paradoxes of the regulatory state. *The University of Chicago Law Review* 57: 407–441.

United Nations Development Programme, United Nations Environment Programme, World Bank and World Resources Institute, 2003. World Resources 2002–2004. World Resources Institute: Washington, DC.

United Nations Environment Programme (2004) *The use of economic instruments in environmental policy: Opportunities and challenges*. Geneva, Switzerland: United Nations.

Relevant Website

https://www.elga.world.

The Genuine Progress Indicator: A Measure of Net Economic Welfare

Ida Kubiszewski, The Australian National University, Canberra, ACT, Australia

Introduction

Countries need a way of measuring progress and goals to strive toward. Since about the 1950s, gross domestic product (GDP) has been the indicator used globally to determine society's progress. However, the goal that we are striving toward is never defined on the global scale.

The attempt was made to change this in 2000, when the United Nations established eight millennium development goals (MDGs). These goals were only created for 15 years, through 2000–15, and only targeted half the world, the developing countries. The sustainable development goals (SDGs) replaced the MDGs in 2016. Again, these goals are set for 15 years but, unlike the MDGs, they target both the developed and developing world.

Although a major step forward, the SDGs have certain limitations. The primary drawback is that with 17 goals, 169 targets, and over 300 indicators in the environmental, social, and economic areas, there is no single overarching goal or vision that society is striving toward (Griggs et al., 2013; Costanza et al., 2016). With 17 goals, there are tradeoffs that will be required, diluting the effectiveness of the SDGs toward a common goal.

But what if maximizing human wellbeing was the unstated goal that the SDGs, and even the MDGs, were aiming toward. If such a goal existed, society would need a measure other than GDP to ascertain whether we were making progress toward achieving this goal or moving backward. However, currently, GDP is the only measure we do use.

Problems With GDP

GDP was never designed to measure social or economic welfare. The original creators of GDP warned against using it for anything except as a specialized tool that measured only a narrow segment of society's activity. However, since the 1950s we have used the size of the economy as our primary indicator of overall progress (Nordhaus and Tobin, 1972). By that yardstick, the global economy (as measured by GDP) has grown more than threefold since 1950. However, economic welfare has actually decreased slightly since 1978 (Kubiszewski et al., 2013).

GDP's current role poses a number of problems. One major issue is that it interprets every expense as positive and does not distinguish welfare-enhancing activity from welfare-reducing activity. For example, an oil spill increases GDP because of the associated cost of cleanup and remediation, but it obviously detracts from overall wellbeing. Examples of other activities that increase GDP include hurricanes (and all other natural disasters), cancer (and other illnesses), crime, car accidents, and divorce.

GDP adds up all marketed deliveries to "final demand" (sales to households, government, net exports, and capital formation) that occur within a country, regardless of whether they represent a real benefit or a "defensive expenditure" like cleaning up an oil spill or treating pollution caused health effects (Leipert, 1989). This is because GDP is calculated using the input/output model. This means that the only things that can be included in GDP are those items that are produced and consumed by one of the sectors in the economy. Nothing else is included.

If the same method of input/output tables were to be used to calculate GDP, the entire process would have to be adjusted. The tables would have to distinguish between the economic activities that added to versus subtracted from economic welfare. Another major change would have to be the inclusion of goods and services that are not within the economic market but do have a large influence on welfare. Over the past few years, various groups, including the United Nation's Statistic Division and the World Bank have been working on creating national accounts that incorporate ecosystem services (Bartelmus, 2014; Hein et al., 2015). Some of these efforts modify the input/output model to incorporate services provided by nature.

Herman Daly, a former senior economist at the World Bank, once commented that, "the current national accounting system treats the earth as a business in liquidation." He also noted that we are now in a period of "uneconomic growth," where GDP is growing but economic welfare is not.

GDP also leaves out many components that enhance welfare but do not involve monetary transactions and therefore fall outside the market. For example, the act of picking vegetables from a garden and cooking them for family or friends is not included in GDP. Yet buying a similar meal in the frozen food aisle of the grocery store involves an exchange of money and a subsequent GDP increase. A parent staying home to raise a family or do volunteer work is also not included in GDP and yet they are potentially key aspects of someone's economic welfare.

There are problems with GDP including that it does not account for the distribution of income among individuals, which has considerable effect on individual and social wellbeing (Wilkinson and Pickett, 2009). GDP does not care whether a single individual or corporation receives all the income in a country, or whether it is equally distributed among the population. A dollar's worth of increased income to a poor person produces more additional welfare than a dollar's increased income to a rich person. Additionally, the distribution of income within a country influences a range of social problems and overall societal welfare.

And yet, even with all the problems surrounding GDP, it is the most commonly used indicator of a country's overall performance.

The Genuine Progress Indicator

Wellbeing is the outcome of a convergence of factors, ranging from good human relations, to greater equality as well as a healthy social and natural environment (Wilkinson and Pickett, 2009; Boarini et al., 2012). Indicators are essential to promote change in economic governance. As post-GDP measurements are integrated into institutional processes, they will be followed by relative rewards and sanctions, as is the case with GDP at present.

In recent years, much work has been done on alternative indicators to GDP—more comprehensive indicators that consolidate economic, environmental, and social elements into a common framework to show net progress. A number of researchers have proposed alternatives to GDP that make one or more of these adjustments with varying components and metrics (Smith et al., 2013). These indicators can be divided into three broad groups: (1) measures that modify economic accounts to address equity and nonmarket environmental and social costs and benefits; (2) measures that use weighted indices of "subjective" indicators based on survey results; and (3) measures that use weighted indices of a number of "objective" indicators.

One such indicator, which fits into the first category, is the genuine progress indicator (GPI). The GPI is a version of the index of sustainable economic welfare (ISEW), first proposed in 1989 by Daly and Cobb (1989) and later modified and renamed the genuine progress indicator (Redefining Progress, 1995).

GPI starts with personal consumption expenditures (a major component of GDP) but adjusts it using about 25 different components (seen in Fig. 1). These components subtract those aspects that are actually an overall negative activities in society, such as the costs of environmental degradation, biodiversity loss, and ecosystem services loss, cost of family breakdown, cost of unemployment, and cost of crime and pollution. They also add positive components left out of GDP, including the benefits of volunteering and household work, among others. GPI, unlike GDP, is also adjusted for income distribution (Cobb et al., 1995; Lawn, 2003; Bagstad and Shammin, 2012). By separating activities that diminish welfare from those that enhance it, GPI better approximates sustainable economic welfare. GPI is not meant to be an indicator of sustainability. It is a measure of economic welfare that needs to be viewed alongside biophysical and other indicators. In the end, since one only knows if a system is sustainable after the fact, there can be no direct indicators of sustainability, only predictors.

Over the past few decades, ISEW or GPI have been calculated in around 20 countries worldwide (Lawn and Clarke, 2008; Kubiszewski et al., 2013). These studies have indicated that in many countries, beyond a certain point, GDP growth no longer correlates with increased economic welfare. The trend is similar in many countries, GPI tracks GDP pretty closely as a country develops, but at a certain point the two diverge. In the United States it happened in the late-1970s while in China in the mid-1990s. GDP keeps growing while GPI levels off or decreases.

Recently, a global GPI was estimated using GPI and ISEW data from 17 countries, containing approximately 53% of the world's population and 59% of the global GDP (Kubiszewski et al., 2013). On the global level GPI/capita peaked in 1978 (Fig. 2). Interestingly, 1978 is also around the time that the human ecological footprint, a biophysical indicator that measures humanity's demand on nature, exceeded the Earth's capacity to support humanity. Other global indicators, such as surveys of life satisfaction from around the world, also began to level off around this time. In fact, a strikingly consistent global trend suggests that as income increases, wellbeing often decreases amidst rising rates of alcoholism, suicide, depression, poor health, crime, divorce, and other social pathologies.

An important function of GPI is to send up a red flag at that point. Since it is made up of many benefit and cost components, it also allows for the identification of which factors increase or decrease economic welfare. Other indicators are better guides of specific aspects. For example, Life Satisfaction, determined by surveys, is a better measure of overall self-reported wellbeing. By observing the change in individual benefit and cost components, GPI reveals which factors cause economic welfare to rise or fall even if it does not always indicate what the driving forces are behind this. It can account for the underlying patterns of resource consumption, for example, but may not pick up the self-reinforcing evolution of markets or political power that drive change.

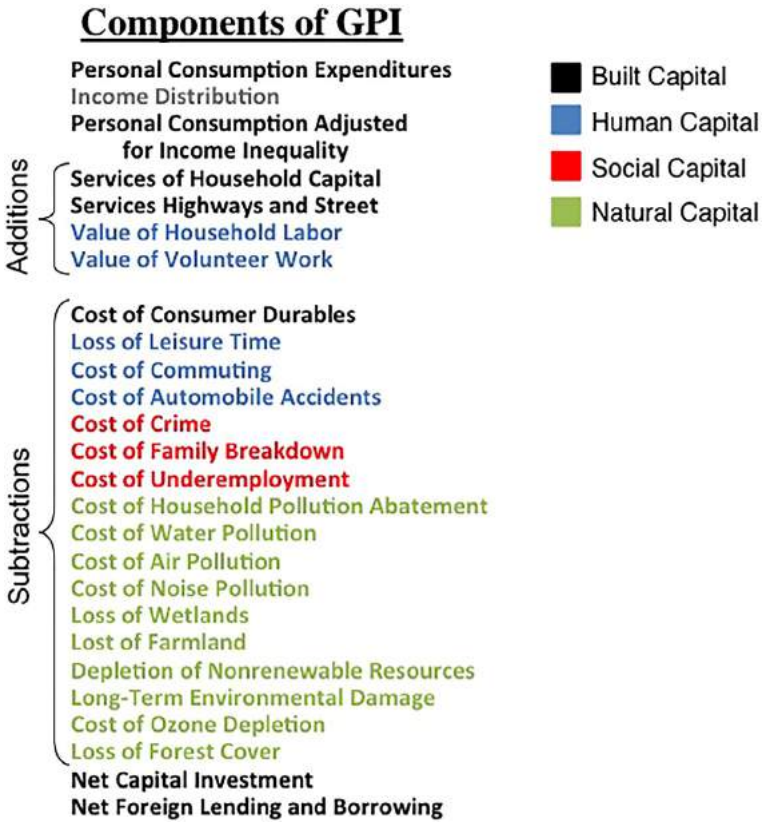


Fig. 1 Components of GPI separated into built, human, social, and natural capitals.

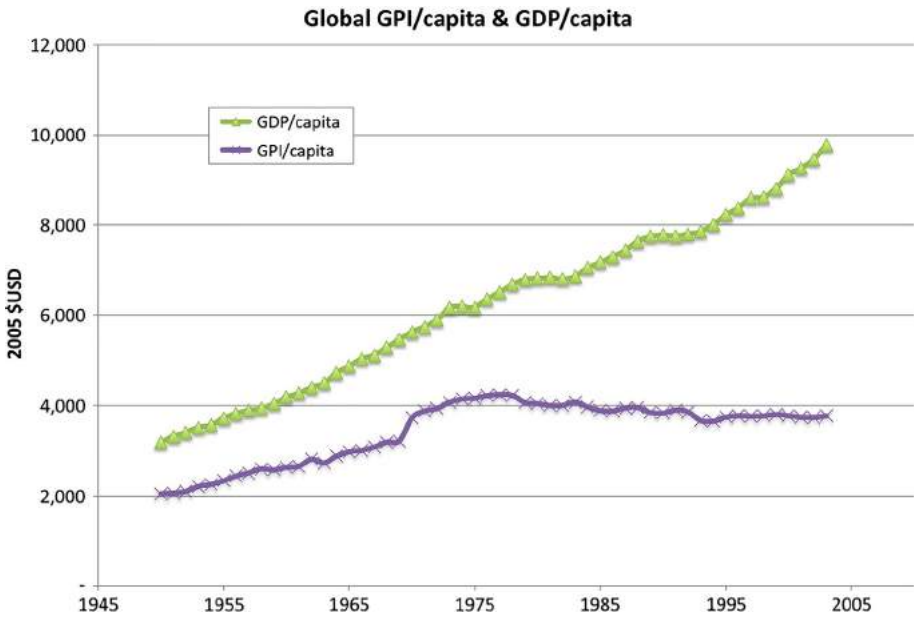


Fig. 2 Global GPI/capita and GDP/capita. GPI/capita was estimated by aggregating data for the 17 countries for which GPI or ISEW had been estimated, and adjusting for discrepancies caused by incomplete coverage by comparison with global GDP/capita data for all countries. All estimates are in 2005 US$. Source Kubiszewski, I., Costanza, R., Franco, C., Lawn, P., Talberth, J., Jackson, T. and Aylmer, C. (2013). Beyond GDP: Measuring and achieving global genuine progress. *Ecological Economics* **93**, 57–68.

Recently, two state governments in the United States have adopted GPI as an official indicator, the states of Maryland and Vermont, and others have begun calculating it (Berik and Gaddis, 2011; McGuire et al., 2012; Erickson et al., 2013; Stiffler, 2014; Kubiszewski et al., 2015). In addition, the data necessary to estimate GPI is becoming more available in many countries and regions. For example, remote sensing data allow better estimates of changes in natural capital and surveys of individuals about their time use and life satisfaction are becoming more routine. New means of measuring inequality are being developed, and more detailed data are being collected on the costs of crime, family breakdown, underemployment, and other measures that might be used in GPI in the future. The bottom line is that the costs of estimating GPI are not particularly high, the data limitations can be overcome, and it can be relatively easily estimated in most countries.

A Closer Look at National GPI Results

A 2013 study (Kubiszewski et al., 2013) looked at the GPI and other indicators for 17 countries from around the world. Four of those countries (China, Japan, the United Kingdom, and the United States) are used here as representative examples of the 17 countries for which data has been collated (Fig. 3).

China experienced rapid GDP/capita growth between 1950 and 2008 as it moved from an agrarian to an industrialized society. GPI/capita also increased during this time, albeit more slowly. After 1994, China joined the world market more completely and its GDP/capita, along with its GPI/capita, increased rapidly. However, this only lasted for about 5 years after which worsening income distribution (the Gini coefficient increased from 0.29 to 0.42) and high environmental externality costs began to become significant enough that they canceled out consumption-related gains. (The Gini coefficient is a measure of income distribution within a country, used as a gauge for economic inequality. This coefficient ranges between 0 and 1, with 0 representing perfect equality and 1 representing perfect inequality.) The change in these costs and benefits can be seen through the individual components that comprise GPI. Consequently, GPI/capita leveled off (Wen et al., 2008).

A similar trend is seen in India. A 1995 study by Manfred Max-Neef showed that the per capita GPI of wealthy nations started to fall when the per capita GDP reached around $15,000–20,000 (MaxNeef, 1995). He concluded at the time that this constituted a "threshold" level of per capita income. A subsequent study in 2008, showed that Thailand's per capita GPI started to fall when its per

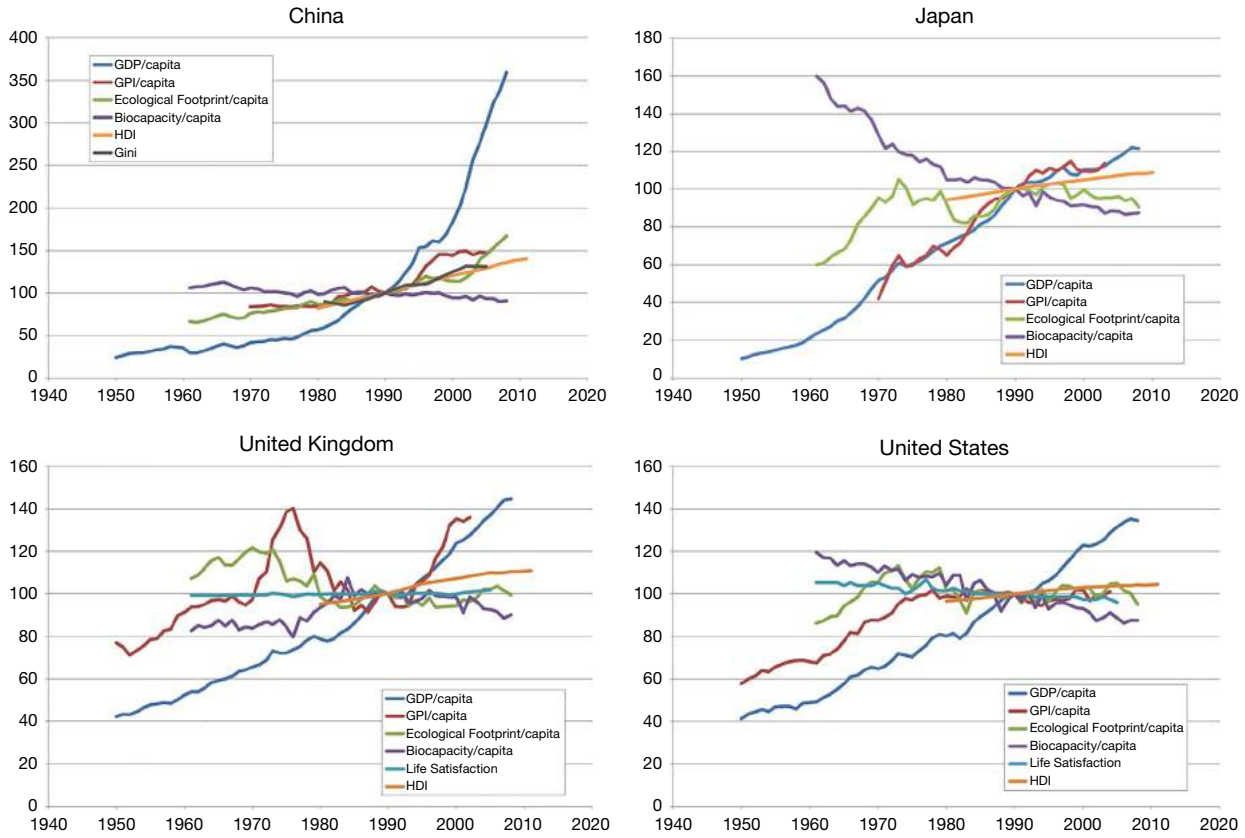


Fig. 3 Comparison with other indicators. The 4 out of the 17 countries comparing indexed trends for GPI/capita, GDP/capita, ecological footprint/capita, biocapacity/capita, HDI, life satisfaction, and the Gini coefficient. All graphs are indexed to 1990 = 100. Source Kubiszewski, I., Costanza, R., Franco, C., Lawn, P., Talberth, J., Jackson, T. and Aylmer, C. (2013). Beyond GDP: Measuring and achieving global genuine progress. *Ecological Economics* **93**, 57–68.

capita GDP reached $7500 (Lawn and Clarke, 2008). For China, the threshold is at $5000. One interpretation is that the threshold level of per capita income is contracting because poor nations are growing their GDP in a "full" world (Costanza, 2008). Hence the marginal cost of GDP growth appears now to be much higher for poor nations at the same stage of the economic development process. We conclude that the ability of poor nations to increase their economic welfare may now be dependent upon rich countries abandoning their obsession with GDP growth. This would provide the "ecological space" for poor nations to experience a phase of welfare-increasing growth.

Japan on the other hand, is one of the only developed countries that experienced a continuous rise in GPI/capita between 1970 and 2003. Much of this is due to the rebuilding after World War II, and is particularly striking in view of Japan's "lost decade" of faltering economic growth. As in China, much of this growth was based on intense natural resource use. In recent years, starting around 1990, the GPI rate of increase has diminished due to environmental degradation (Makino, 2008). Fig. 4 also shows that Ecological Footprint/capita and biocapacity/capita for Japan intersect around 1990. This means that after that point, Japan began using resources faster than it was generating them. However, Japan is a very heavy importer of raw materials and therefore its own environmental costs have not risen significantly (Makino, 2008). This creates a problem for GDP since it does not handle transboundary issues well. It also underscores the case for estimating a global GPI since an undervaluation of environmental costs in one country is counterbalanced by overvaluation in others.

The graph for the United Kingdom seems to show much variation over the course of 52 years. However, because these are indexed graphs, showing only trends, we see that the change in actual GPI/capita is small throughout that period. GDP/capita has been increasing steadily over that time period (Jackson et al., 2008) while GPI showed increases and decreases due to changes in government policies.

The United States shows GPI/capita and GDP/capita increasing at a relatively similar rate until about 1979 at which point GDP/capita continues to increase while GPI/capita flattens out. This occurred for reasons similar to those in other countries: a worsening of income distribution combined with environmental and social costs rising faster than consumption-related benefits.

Interestingly, HDI and Life Satisfaction do not change much within any of these 4, or even the original 17, countries. In three of our four countries, Japan, the United Kingdom, and the United States, the Ecological Footprint/capita remains significantly higher than biocapacity/capita.

There is also a general trend that appears from approximately 1950 until around 1975 where the GPI/capita for the majority of countries is increasing. Much of this is due to the rebuilding effort after World War II when consumption and built capital were the limiting factors for improving economic welfare in many countries and environmental externalities had not yet become significant. However, around the mid-to-late 1970s, much of the infrastructure was rebuilt while worsening income distribution and increasing external environmental costs canceled the growth in consumption-related benefits, causing GPI/capita to level off.

Fig. 4 shows that globally GPI/capita peaks at around $6500 GDP/capita. This estimate excludes African countries, as GPI has not yet been calculated for any African countries. However, since most African nations are poor economically, and given the GPI results for China (where the GPI started declining at a per capita GDP of $5000), a threshold per capita GDP value of $6500 is therefore a conservative one.

Until the $6500 GDP/capita peak, the GPI/capita and GDP/capita are highly correlated ($R^2 = 0.98$). This is consistent with some subjective life satisfaction studies showing leveling after around $7000 GDP/capita (Inglehart, 1997; Deaton, 2008). It is also

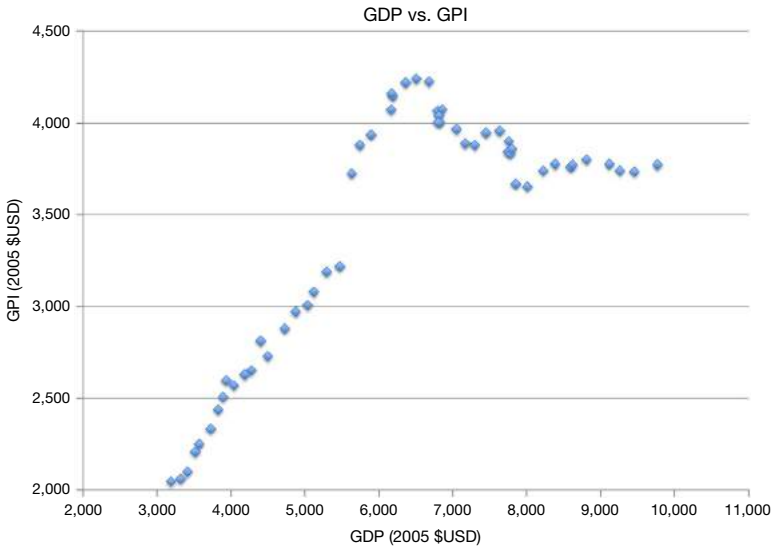


Fig. 4 GDP/capita versus GPI/capita. A plot of global GDP/capita versus estimated global GPI/capita. The two are positively correlated until about $7000/capita ($R^2 = 0.98$), after which they diverge with a negative correlation ($R^2 = 0.61$). All data in 2005 US$. Source Kubiszewski, I., Costanza, R., Franco, C., Lawn, P., Talberth, J., Jackson, T. and Aylmer, C. (2013). Beyond GDP: Measuring and achieving global genuine progress. *Ecological Economics* **93**, 57–68.

interesting that there is a negative correlation ($R^2 = 0.61$) between GDP/capita and GPI/capita after about $6500 GDP/capita. This is also consistent with the "threshold hypothesis" proposed by Manfred Max-Neef, which states that: "for every society there seems to be a period in which economic growth (as conventionally measured) brings about an improvement in the quality of life, but only up to a point—the threshold point—beyond which, if there is more economic growth, quality of life may begin to deteriorate" (MaxNeef, 1995).

We can use this GDP/capita maximum from Fig. 4 to estimate the maximum global GDP/capita consistent with a nondeclining GPI/capita. By assuming the 2010 levels of global GDP/capita of approximately $67 trillion were to be divided equally among the population to provide each person with $7000, the population would have to stabilize at around 9.6 billion people. This could be possible with better access to family planning services in high population growth nations (Engelman, 2011). An important note is that variations in income would need to exist between and within nations, however, these disparities should be much smaller than they are today.

Many scientists argue that even current consumption levels are not sustainable (Daily and Ehrlich, 1992; Rees, 2006). The global Ecological Footprint/capita exceeded global biocapacity/capita around 1978. As of 2011, humans were using 135% of the resources that can be sustainably generated in 1 year (Ecological Footprint, 2011).

Provided the technical efficiency of global production can be increased by 35%—which appears to be feasible—the global GDP of $67 trillion that is required to provide a welfare-maximizing GDP/capita of $7000 for 9.6 billion people may be sustainable. Once reached, continuing improvements in environmental protection, full employment (distributional equity), and product quality would allow the GPI/capita to rise without the need for further increases in global GDP. It is possible to increase economic welfare without having to grow GDP.

Rising environmental costs are directly related to the rise in the rate of resource use and waste generation, which is due to the growth in GDP, despite technological advances. Environmental costs could be reduced by reducing material and fossil energy throughput to the global economy, some of this may come with efficiency advances, but some will result in reductions in GDP—recognizing that this may actually be welfare enhancing. In addition, a more equitable distribution of income and opportunities will allow the welfare contribution of a given level of consumption to be increased. Welfare benefits can also be increased through the production of higher quality, longer lasting goods and the social capital benefits of a fairer and more just society.

Divergence Between GDP and GPI

GDP was created after the Great Depression in the United States and WWII, when the world needed to repair its built infrastructure and financial systems (Fioramonti, 2014). Natural resources were perceived as abundant and inadequate access to infrastructure and consumer goods represented the main limit on improvements to human welfare (Daly, 1992). During this time, it made sense to create an indicator that ignored relatively abundant natural resources, and the distribution of wealth and focused solely on increasing the production and consumption of market goods and services, which were relatively scarce (Costanza et al., 2014).

However, as a result of our success, the world has changed dramatically over the past few decades. We now live in a world full of human infrastructure. The human footprint has grown so large that, in many cases, limits on the availability of natural resources now constrain real progress more than limits to consumer goods (Costanza, 2008; Beddoe et al., 2009).

Between approximately 1950 and 1975, GPI per person for the majority of countries was increasing. Much of this was due to the rebuilding effort after World War II when consumption and built capital were the limiting factors for improving wellbeing in many countries and environmental externalities had not yet become significant. However, around the mid to late 1970s, much of the infrastructure was rebuilt. However, rising income inequality and increasing external environmental costs began to cancel the growth in consumption-related benefits, causing GPI/capita to level off.

As stated below, GPI is not a measure of overall human wellbeing since it emphasizes economic welfare and leaves out other important aspects of wellbeing. It is, however, a far better indicator of economic welfare than GDP, which was not designed to measure welfare at all. Societal wellbeing or economic welfare ultimately depends on stocks of natural, human, built, and social capital, and because the GPI makes additions and deductions to GDP to reflect net contributions to these stocks it is a far superior measure of economic welfare than GDP (Vemuri and Costanza, 2006). The disconnect between GPI and GDP, beginning in 1978, shows the aspects of our economic welfare that have been declining since that time. It also provides focus areas where societal improvement is necessary and possible.

Problems With GPI and Possible Responses

GPI is not a perfect indicator and as it has become more widely used, it has been often criticized (Harris, 2007; Brennan, 2008; Neumayer, 2010). Criticisms include that: (1) certain components are estimated through inappropriate valuation methods; (2) it makes the assumption that built capital and natural capital are substitutes; (3) although the GPI includes certain aspects that contribute to human-welfare it excludes others, for example political freedom; (4) the GPI is subjective in the components it includes and sometimes various slightly between countries; and (5) it does not have a solid theoretical basis.

Point 1: The use of cumulative cost as a valuation method for certain environmental items in the GPI is often criticized. This can be seen in components such as the cumulative cost of land degradation, lost wetlands, and long-term environmental damage. The

cumulative cost approach is used with these environmental costs due to the "strong sustainability" assumption (Lawn, 2005). One the major goal of GPI is to measure the economic welfare generated by economic activity. The assumption is that economic activity requires natural capital. Thus GPI to subtracts the *permanent* loss of natural capital services.

There are many possible ways to measure these permanent losses. The most obvious way is to assume that the current welfare cost equates to the amount that existing people should be compensated for inheriting a diminished stock of natural capital. To be consistent with strong sustainability, appropriate compensation should approximate what it would have cost past generations to have kept the stock of natural capital intact. This is equivalent to the cumulative rather than annual cost of some environmental losses.

Point 2: The second often heard criticism about GPI is that it assumes that built and natural capitals are substitutable. This criticism is based on the fact that GPI values and then combines the costs and benefits of these two forms of capital. By combining these two capitals into a single index, the assumption is that decreasing natural capital and turning it into built capital will make up the difference.

It is true that if one component is decreased by the same amount another component is increased, then one does compensate the other. This does create a substitutability situation. However, this does not mean that the total economic welfare currently being enjoyed is sustainable. If it were assumed sustainable, it would be wrongly assuming substitutability of current welfare benefits with the substitutability of the capital that yields the welfare benefits.

For example, benefits to welfare of a wooden table exactly match the costs of the losses of the forest that timber came from, than welfare remains unchanged. However, a forest provides many more services then a table or any other furniture can provide, including being a source of oxygen, carbon sink, and life-support services that are needed to sustain humanity and its economic activity in the future, which includes production of new wooden furniture. Although current costs and benefits have been offset for the present, in the long-term, the ability to create additional in the future has declined. GPI was never designed to be a measure of sustainability and only does measure the current value of welfare, not that of long-term potential. Other indicators, such as the ecological footprint, need to be used in conjunction with GPI to determine whether current use and conversion of resources is sustainable.

Point 3: Unfortunately it is not possible to measure every component that contributes to human welfare. Hence, GPI will not be able to measure every aspect. GPI was designed to measure the total economic welfare generated by economic activity. It is confined to measuring whether a certain economic activity is increasing or decreasing welfare, basically to measure whether marginal benefits of GDP growth are higher or lower than the marginal costs.

Because political freedom is not a benefit that is created by economic activity, it is not, and should not be, part of GPI. On the other hand, if greater political freedom has a positive effect on economic activity, that will be measured by other components of GPI, for example in lower inequality, crime, and underemployment. Hence, greater political freedom may be reflected in GPI, but indirectly. To include it separately would be double counting.

All indicators that are made up of various components are subjective. It requires judgments to decide which components to include, which to leave out, and what weights to give each component. This is also true about GDP. GDP is made up of hundreds of components that were decided by individuals as being most critical. These GDP components have also changed over the years as society has developed and new economic activities became important.

GPI 2.0

Over the past few years, as a growing number of GPI studies have been performed globally, a divergence in methodologies has occurred. This lack of standardization is due to variations in data availability, varying needs to ensure policy relevance in specific regions, and identification of new issues such as treatment of nonrenewable resources, government spending, and others issues as stated above. To address these variations, an international effort has recently started to update the current methodology of the GPI with the most up-to-date science. The goal of such an update is to ensure that GPI 2.0 has greater comparability between studies and an increased policy relevance (Bagstad et al., 2014).

Efforts in Use of Alternative Measures

In the United States, progressive states have turned to the GPI as a tool to assist state government in identifying public policy priorities and in the application of outcomes-based budgeting. Maryland, the first state to adopt GPI as an official indicator and the one that has progressed furthest in its use, has formalized GPI calculation and reporting (McGuire et al., 2012). Vermont, in 2013, became the first state to establish a system for GPI data collection by legislative mandate (Erickson et al., 2013). Other states, with little legislative support, have calculated GPI: the Colorado Fiscal Institute, the Utah Population and Environment Coalition, and the Hawaii Department of Health (Bagstad et al., 2014).

A Shareholder's Report

As demonstrated in previous sections, the GPI moves one-step beyond earlier benchmark categories, suggesting a "full cost" accounting system for economic growth. GPI assigns monetary value to flows of natural, human, social, and built capital and

their degradation or enhancement in the course of our economic activity. The GPI adjusts gross domestic product (GDP) to account for the effect of income inequality on personal consumption expenditures, adding the value of time spent at socially enhancing unpaid work such as volunteering, and deducting "unfortunate" expenditures for social ills such as crime, and the depreciation value of our natural resources. The result can be expressed as a GPI Net Income Statement.

The GPI Net Income Statement offers a substantially more complete accounting of economic activity and its impact on our quality-of-life than conventional GDP-based measures of progress. But net income is only one part of any financial report. As the shareholders and stewards of a country's or state's natural and other resources, citizens would be best informed by seeing GPI full-cost accounting applied to the remaining components of a shareholders' report: a balance sheet and cash flow statement. Just as an income statement does not tell shareholders about a company's net assets or shareholder wealth, GPI does not tell us about either the quantity or quality of stocks of natural, human, social, and built capital. Neither does it reveal anything about the region's accumulated liabilities, such as the cost of infrastructure maintenance, stores of toxic waste, or health problems caused by loss of leisure time. It is the balance sheet that signals whether an organization is either creating wealth for its shareholders by making wise investments, or endangering its future by accumulating liabilities and degrading or depreciating its capital assets.

As an example, one of the GPI indicators, Net Forest Cover Change, assumes an underlying value for the functions performed by a healthy forest ecosystem. In addition to producing marketable products such as timber, our forests provide a range of valuable services, such as storage and filtration of water, oxygen production, soil formation, nutrient cycling, wildlife habitat, and human recreation—to name a few that typically go unnoticed and unvalued. Unsustainable timber harvesting actually increases GDP, without accounting at all for the reduced asset value on the public balance sheet from lost forest cover. GPI is an improvement in that it accounts for the lost forest cover, subtracting it as an "unfortunate" cost of economic activity. But a GPI balance sheet would actually inform us as to our total stock of forest cover, accounting for each year's net change as an increase or diminution of total asset value. Like any capital asset, that value would be determined by calculating the net present value of the flow of goods it will yield and the services it will perform over its useful life. Regional funds spent to protect or restore forest cover would be characterized as investment to the extent that they increase our forests' value. Without this full accounting for the stock and value of our forest cover, it is difficult to evaluate the financial benefits of conserving versus depleting it.

Constructing a GPI Balance Sheet will require creating a chart of accounts that includes each of the domains addressed by GPI, and taking inventory of our accumulated assets and liabilities as they are found among those domains. Assigning value to multiple domains of capital, many of which are made up of nonmarket assets that have never been monetized, is a challenging endeavor. But it is one that some progressive governments and subnational entities have begun, with pioneering methodologies. The UK's Office of National Statistics released an experimental estimate of its human capital stock, including a detailed methodology for valuing the productive capacity of citizens (Jones and Chiripanhura, 2010). The United Nations' System of Integrated Environmental and Economic Accounts (SEEA) has been revised to include a framework for valuing the market and nonmarket goods and services provided by our natural capital. The Province of Nova Scotia, Canada, has officially committed to the task of valuing natural, human, and social capital, in addition to built and financial capital, toward the goal of producing "a new form of budget estimates, a new set of accounts, and a new economic paradigm" (Pannozzo and Colman, 2009). Canada has extended its System of National Accounts to value volunteerism and the nonprofit sector as an element of its social capital (Haggar-Guenette et al., 2007). Meanwhile, the most developed conceptual framework for expanded GPI accounting has been described by the Pembina Institute for the Province of Alberta, Canada (Anielski, 2001).

The balance sheet prototype proposed here for GPI accounting is an approximation in need of considerable development and refinement. Ultimately the identification of a regions assets—public goods, natural endowments, and accumulated commonwealth—should be informed, in part, by how citizens conceptualize quality of life.

Embarking upon the project of GPI balance sheet accounting would place a country at the vanguard of an emerging trend in progressive, public sector full-cost accounting. Over the next several decades, policy-makers at all levels of government around the world will develop methods to value the nonmarketed contributions to our common wealth from previously unaccounted-for sources. Just as corporations have established methods to value intangible assets such as patents, goodwill and brand names, we need to develop standardized methods for valuing assets such as strong civic engagement, good health, and an educated populace. Our quality of life and its sustainability depend on it.

Conclusion

If we hope to achieve a sustainable and desirable future, we need to rapidly shift our policy focus away from maximizing production and consumption (GDP) and toward improving genuine human wellbeing (a version of GPI or something similar). This is a shift that will require far more attention to be paid to environmental protection, full employment, social equity, better product quality and durability, and greater resource use efficiency. These changes are clearly within our grasp, and are underway in several countries and regions. Alternative measures of progress, like GPI, are useful to help chart and guide the course if appropriately used and understood. The future we want is within our grasp, but not while we remain in the grasp of a measure of progress (GDP) that has clearly outlived its usefulness. It has often been said that you get what you measure and we need to begin to measure what we really want if we have any hope of achieving it.

References

Anielski M (2001) *The Alberta GPI Blueprint*. Drayton Valley, AB: Pembina Institute for Appropriate Development.

Bagstad KJ and Shammin MR (2012) Can the genuine progress indicator better inform sustainable regional progress?—A case study for Northeast Ohio. *Ecological Indicators* 18: 330–341.

Bagstad KJ, Berik G, and Gaddis EJB (2014) Methodological developments in US state-level genuine progress indicators: Toward GPI 2.0. *Ecological Indicators* 45: 474–485.

Bartelmus P (2014) Environmental–economic accounting: Progress and digression in the SEEA revisions. *Review of Income and Wealth* 60(4): 887–904.

Beddoe R, Costanza R, Farley J, Garza E, Kent J, Kubiszewski I, Martinez L, McCowen T, Murphy K, Myers N, Ogden Z, Stapleton K, and Woodward J (2009) Overcoming systemic roadblocks to sustainability: The evolutionary redesign of worldviews, institutions, and technologies. *Proceedings of the National Academy of Sciences* 106(8): 2483–2489.

Berik, G. and E. Gaddis. (2011). The Utah genuine progress indicator (GPI), 1990 to 2007: A report to the people of Utah, Utah Population and Environment Coalition.

Boarini, R., M. Comola, C. Smith, R. Manchin and F. de Keulenaer. (2012). What makes for a better life? The determinants of subjective well-being in OECD countries—Evidence from the Gallup World Poll, OECD Statistics Working Papers, 2012/03.

Brennan AJ (2008) Theoretical foundations of sustainable economic welfare indicators—ISEW and political economy of the disembedded system. *Ecological Economics* 67(1): 1–19.

Cobb C, Halstead T, and Rowe J (1995) *The genuine progress indicator: Summary of data and methodology*. San Francisco, CA: Redefining Progress.

Costanza R (2008) Stewardship for a "full" world. *Current History* 107(705): 30–35.

Costanza R, Kubiszewski I, Giovannini E, Lovins H, McGlade J, Pickett KE, Ragnarsdóttir KV, Roberts D, Vogli RD, and Wilkinson R (2014) Time to leave GDP behind. *Nature* 505(7483): 283–285.

Costanza R, Daly L, Fioramonti L, Giovannini E, Kubiszewski I, Mortensen LF, Pickett KE, Ragnarsdottir KV, De Vogli R, and Wilkinson R (2016) Modelling and measuring sustainable wellbeing in connection with the UN sustainable development goals. *Ecological Economics* 130: 350–355.

Daily GC and Ehrlich PR (1992) Population, sustainability, and Earth's carrying capacity. *Bioscience* 42(10): 761–771.

Daly HE (1992) From empty-world economics to full-world economics: Recognizing an historical turning point in economic development. In: *Population, technology and lifestyle*, pp. 23–37. Washington, DC: Island Press.

Daly HE and Cobb JB Jr. (1989) *For the common good: Redirecting the economy toward community, the environment, and a sustainable future*. Boston, MA: Beacon Press.

Deaton A (2008) Income, health, and well-being around the world: Evidence from the Gallup World Poll. *Journal of Economic Perspectives* 22(2): 53–72.

Ecological Footprint. (2011). Earth Overshoot Day, September 27, 2011. Retrieved June 10, 2012, from http://www.footprintnetwork.org/en/index.php/GFN/blog/today_is_earth_overshoot_day1.

Engelman R (2011) An end to population growth: Why family planning is key to a sustainable future. *Solutions* 2(3): 32–41.

Erickson JD, Zencey E, Burke MJ, Carlson S, and Zimmerman Z (2013) *Vermont genuine progress indicator, 1960–2011: Findings and recommendations*. Gund Institute for Ecological Economics: Burlington, VT.

Fioramonti L (2014) *How numbers rule the world*. London: Zed Books.

Griggs D, Stafford-Smith M, Gaffney O, Rockstrom J, Ohman MC, Shyamsundar P, Steffen W, Glaser G, Kanie N, and Noble I (2013) Policy: Sustainable development goals for people and planet. *Nature* 495(7441): 305–307.

Haggar-Guenette C, Hamdad M, Laronde-Jones D, Pan T, and Yu M (2007) *Satellite account of non-profit institutions and volunteering*. Statistics Canada: Ottawa.

Harris M (2007) On income, sustainability and the microfoundations of the genuine progress indicator. *International Journal of Environment, Workplace and Employment* 3(2): 119–131.

Hein L, Obst C, Edens B, and Remme RP (2015) Progress and challenges in the development of ecosystem accounting as a tool to analyse ecosystem capital. *Current Opinion in Environmental Sustainability* 14: 86–92.

Inglehart R (1997) *Modernization and postmodernization. Cultural, political and economic change in 43 societies*. Princeton: Princeton University Press.

Jackson T, McBride N, Abdallah S, and Marks N (2008) *Measuring regional progress: Regional index of sustainable economic well-being (R-ISEW) for all the English regions*. New York: New Economics Foundation.

Jones R and Chiripanhura B (2010) *Measuring the UK's human capital stock*. London: Office for National Statistics.

Kubiszewski I, Costanza R, Franco C, Lawn P, Talberth J, Jackson T, and Aylmer C (2013) Beyond GDP: Measuring and achieving global genuine progress. *Ecological Economics* 93: 57–68.

Kubiszewski I, Costanza R, Gorko NE, Weisdorf MA, Carnes AW, Collins CE, Franco C, Gehres LR, Knobloch JM, Matson GE, and Schoepfer JD (2015) Estimates of the genuine progress indicator (GPI) for Oregon from 1960–2010 and recommendations for a comprehensive shareholder's report. *Ecological Economics* 119: 1–7.

Lawn PA (2003) A theoretical foundation to support the index of sustainable economic welfare (ISEW), genuine progress indicator (GPI), and other related indexes. *Ecological Economics* 44(1): 105–118.

Lawn PA (2005) An assessment of the valuation methods used to calculate the index of sustainable economic welfare (ISEW), genuine progress indicator (GPI), and sustainable net benefit index (SNBI). *Environment, Development and Sustainability* 7(2): 185–208.

Lawn PA and Clarke M (2008) *Sustainable welfare in the Asia-Pacific: Studies using the genuine progress indicator*. Cheltenham: Edward Elgar Publishing.

Leipert C (1989) National income and economic growth: The conceptual side of defensive expenditures. *Journal of Economic Issues* 23(3): 843–856.

Makino M (2008) Genuine progress in Japan and the need for an open economy GPI. In: Lawn PA and Clarke M (eds.) *Sustainable welfare in the Asia-Pacific: Studies using the genuine progress indicator*, pp. 153–189. Cheltenham: Edward Elgar Publishing.

MaxNeef M (1995) Economic growth and quality of life: A threshold hypothesis. *Ecological Economics* 15(2): 115–118.

McGuire S, Posner S, and Haake H (2012) Measuring prosperity: Maryland's genuine progress indicator. *Solutions* 3(2): 50–58.

Neumayer E (2010) *Weak versus strong sustainability: Exploring the limits of two opposing paradigms*. Cheltenham: Edward Elgar.

Nordhaus W and Tobin J (1972) *Is growth obsolete? Economic growth*. New York: Columbia University Press.

Pannozzo L and Colman R (2009) *New policy directions for Nova Scotia: Using the genuine progress index to count what matters*. GPI Atlantic: Nova Scotia.

Redefining Progress (1995) *Genuine progress indicator*. San Francisco: Redefining Progress.

Rees WE (2006) Ecological footprints and biocapacity: Essential elements in sustainability assessment. In: Dewulf J and Van Langenhove H (eds.) *Renewables-based technology*, pp. 143–157. Chichester: Wiley.

Smith LM, Case JL, Smith HM, Harwell LC, and Summers JK (2013) Relating ecosystem services to domains of human well-being: Foundation for a U.S. index. *Ecological Indicators* 28: 79–90.

Stiffler C (2014) *Colorado's genuine progress indicator (GPI): A comprehensive metric of economic well-being in Colorado from 1960–2011*. Denver, CO: Colorado Fiscal Institute.

Vemuri AW and Costanza R (2006) The role of human, social, built, and natural capital in explaining life satisfaction at the country level: Toward a National Well-Being Index (NWI). *Ecological Economics* 58(1): 119–133.

Wen Z, Yang Y, and Lawn PA (2008) From GDP to GPI: Quantifying thirty-five years of development in China. In: Lawn PA and Clarke M (eds.) *Sustainable welfare in the Asia-Pacific: Studies using the genuine progress indicator*, pp. 228–259. Cheltenham: Edward Elgar Publishing.

Wilkinson RG and Pickett K (2009) *The spirit level: Why more equal societies almost always do better*. London: Allen Lane.

Human Ecology: Overview

F Steiner, University of Texas at Austin, Austin, TX, USA

Introduction

We interact with each other and with our physical environments. We are biological creatures who depend upon the living landscape to sustain us. Plants and animals are affected by our actions, and our existence is impacted by plants and animals. We exist within complex sets of interactions – that is, we live in an ecological world.

Learning to perceive the world as a never-ending system of interactions – that is, to think about our surroundings and our relationships with our environments and each other ecologically – is challenging. Such thinking forces us to rethink our views of economics, politics, and business. It suggests different ways to plan and design. In economics, for example, an ecological view suggests a much more complex set of relationships than supply and demand: supply of what and where from and at what cost, not only in dollars but to other species and other generations. Ecological understanding can also confront our values and religious beliefs, although most faiths address human connections to the natural world and stewardship responsibilities for future generations.

Ecology is, by definition, the reciprocal relationship among all organisms and their biological and physical environments. People are organisms. As a result, we can ask, Is the use of human as a modifier to ecology necessary?

Human with ecology helps reinforce the reality of our place in environments. Human ecology seeks to understand the multiple interrelationships between the human species and our environments. Human ecology is broader than biology, but also is grounded in biological concepts. The transdisciplinary field can be defined as the study of the complex and varied system of interactions between people and their environments.

Brief Early History of Human Ecology

Many overlaps between the social and biological sciences existed at the end of the nineteenth century and during the early twentieth century. Ecological concepts were prominent in both geography and sociology. Human ecology was recognized as a unique field of geography. Geographers went so far as declaring ``geography as the science of human ecology.'' Early twentieth-century geographers sought to make clear the relationships existing between natural environments and the distribution and activities of people. However, this approach unfortunately became linked with environmental determinism which suggested that our surroundings shape everything from skin color to behavior. These concepts led to rather simplistic, and even racist, notions about how environments shaped cultures, and environmental determinism was discredited in the 1920s.

Also during the 1920s, urban sociologists adapted ecological concepts to explain settlement patterns and human interactions in cities. Called the Chicago School, these sociologists adapted observation methods from anthropology to describe urban life and culture. They described Chicago as a series of concentric rings from the central business district to the commuter zone on the periphery. They also used the ecological concept of succession to describe how these zones build up one after another as a city grows. These sociologists suggested how various groups of people succeed others in the concentric zones. This small group of sociologists used ecological concepts more as metaphors than as a tool for scientific analysis. As a result, the connections between the two disciplines were not deep, in spite of promising beginnings. Meanwhile, the advances in geography were overshadowed by the environmental determinism criticisms. As a result, human ecology faded to the margins of geography and became a historical footnote in sociology for several decades.

Increasingly, the social sciences became disconnected from the physical sciences and, by extension, from the material world. The focus of the social sciences shifted from ecological models to the embrace of economic, political, and demographic approaches where the role of natural forces was more subtle. In order to bolster the validity of their science, some researchers emphasized quantitative analysis that favored data about people over the observation of the human condition. Meanwhile, ecologists, especially those in North America, concentrated on the study of natural, nonhuman environments. Some one-third of the land in the United States is in public ownership, enabling wildlife and vegetation research on vast expanses with little human interruption.

There are many ironies in this disconnection. For example, the Greek root for both ecology and economics is the same: *oikos*. Both disciplines involve the study of the household. Ecology is the study of the environmental house, including all its inhabitants, in which we live and in which we place our human-made structures and domesticated plants and animals. Economics is the study of the household of money. As we can track the flow of money, we can also illuminate other movements in the places where we live. But beyond their common Greek root, economics and ecology diverged with few clear connections persisting.

Beginning in the 1960s, the general public became alarmed by population growth and the consequences of pollution on water, air, and land quality. Biologists and ecologists used human ecology to emphasize how people are subject to the same environmental limitations as other animals. Also during this time, anthropologists used the term to help explain the impact of

 doi:10.1016/B978-0-444-63768-0.00626-0

environment on culture. Ecologically oriented anthropologists adapted concepts like population regulation and energy flow to explain community organization. In general, early use of ecological concepts in human ecology depended on traditional views of nature, such as the tendency of systems to evolve toward a steady state.

This past suggests the ongoing utility of human ecology. By understanding the interactions and interrelationships between people and their environments, human ecology can help to:

- consider and plan for the long-term consequences of human actions;
- avoid disastrous surprises resulting from environmental phenomena such as floods, earthquakes, tornadoes, wildfires, and tsunamis;
- generate ideas for dealing with environmental challenges and opportunities; and
- create a livable and sustainable relationship with the environment.

However, to realize this utility, it is necessary to understand changes in ecological thinking generally and how human ecology fits within this ever-changing discipline.

Toward a New Ecology

Since the first Earth Day in April 1970 and the rise of the modern environmental movement, social scientists have rediscovered the environment while biologists have probed social interactions. Meanwhile, several ecologists have addressed human communities, and planners and architects have attempted to provide syntheses to shape human communities. In addition to the stimulus from popular culture, as expressed in wide-ranging areas from politics to music, advances in theory through computing technologies, urban morphology (the study of how cities are structured physically), landscape studies, and ideas about complexity have contributed to this renewed interest in the environment by social scientists. From within the biological sciences, research has altered conventional views about organism–environment interactions. Increasingly, ecologists consider human influences on their environments.

This new human ecology emphasizes complexity over reductionism, focuses on changes over stable states, and expands ecological concepts beyond the study of plants and animals to include people. This view differs from the environmental determinism of the early twentieth century. The new ecology addresses the complexity of human interactions rather than how a specific physical environment shapes human anatomic variations. Because people form part of its scope, new ecology may be viewed as human ecology, or the evolution of traditional ecology to reconsider human systems.

New ecology represents a significant reorientation that has occurred in the field of biological ecology. For example, new ecology embraces disequilibria, instability, and even chaotic fluctuations in natural and human-impacted biophysical environments. Two primary changes have occurred in new ecology, differentiating it from its traditional progenitor. The first shift is from an equilibrium perspective, where local populations and ecosystems are viewed as in balance with local resources and conditions, to a disequilibrium perspective where history matters and populations and ecosystems are continually being influenced by disturbances. The second change is from considering populations and ecosystems as relatively closed or autonomous systems, independent of their surroundings, to viewing both populations and ecosystems as open that are strongly influenced by the input and output, or flux, of material and individuals across system borders.

Traditional ecology relied on the assumptions that nature could achieve balance and that ecosystems functioned as closed systems. Natural plant communities evolved through several stages, climaxing in a steady state, according to traditional theory. Since ecologists studied plants and animals in forests, deserts, and other environments relatively removed from human settlements, their interactions could be isolated for study within closed systems.

New ecology challenges both assumptions. Living systems are viewed as changing and complex rather than stable and balanced. In addition, the boundaries between communities blur. Open systems possess fluid, overlapping boundaries across several spatial scales from the local to the global.

Ideas Contributing to a New Human Ecology

Ecology lends itself to reinvention, to reinterpretation. Relationships link things, and how we view connections among elements changes. As early as the 1950s, anthropologists called for a `new ecology'. The ideas leading to the more recent, expanding view of ecology have come from many sources and a variety of disciplines, including anthropology. The catalysts for change include advances in technologies, the study of urban morphology and landscape ecology, a broader understanding of chaos theory, and increased interest in issues of sustainability. The emergence of urban ecology exemplifies a beginning in the synthesis of these sometimes divergent catalysts. Urban ecology focuses on organism–environment interactions within cities and other human settlements. By concentrating on urban areas, the interests of the new ecological perspective are woven closer together.

Fresh ways to observe nature, primarily as a result of computer and remote-sensing technologies, have altered our understanding of functions, structures, and patterns. These new (and evolving) technologies are yielding a deeper perspective, because many events can be considered simultaneously in a connected network.

A computer technology especially valuable for revealing complex, ecological relationships is geographical information systems, known by its abbreviation GIS. These computer software programs allow analysis to study overlapping spatial data and map the results. For example, the home range of a tiger beetle species can be mapped then compared with a similar map for a species of brown bear. In turn, both can be overlaid on the migration routes of Canada geese and the extent of a coniferous forest and so on. GIS emerged concurrently with new ways to see and to record the surface of the planet, such as remote-sensing technologies. Whereas GIS programs map information, remote sensing creates imagery of phenomena on the Earth's surface.

As the Apollo astronauts approached the moon, they relayed images back to Earth unlike anything previously seen. The hypnotic pictures of the moon riveted our attention, of course, but the photographs of the blue-green orb of Earth were perhaps even more profound. Continents and water bodies were clearly visible beneath swirls of clouds, but borders had disappeared (Fig. 1). No longer would we see Earth in the manner of the little globes in our classrooms. NASA continues to produce images of the planet, as do other governmental and private remote-sensing groups. In fact, NASA broadcasts continual images of our planet on its own television network.

Remote-sensed information is collected through satellites or high-flying aircraft. The images can be enhanced with computers to reveal specific phenomena, such as land cover, land use, and fault lines. Climate patterns can be tracked and future weather events forecasted. Remote sensors can also be linked to on-the-ground monitoring stations. Such connections allow phenomena to be observed through time. For example, a drainage basin can have several stream-monitoring gauges, which may be linked to a central data collection center. In turn, satellites may be able to collect rainfall and snowpack information daily that can be combined with the field data to predict future water supplies.

The use of GIS and remote-sensing technologies has spread rapidly among scientists during the past few decades. A geologist can overlay a map of bedrock on an aerial photograph to determine where a fault line intersects with settlement. Additional technologies likely will open more possibilities. For example, visualization techniques present three-dimensional representations of objects. Such visualization can be combined with GIS to show places more holistically. The maps of the geologist and the ecologist can be rendered in three dimensions to illustrate the relationships among phenomena such as aquifers, wildlife corridors, and land use. The Internet opens opportunities, too. For instance, a team of American students can work with a group of Italians in a virtual studio, and share GIS maps and photographs of a place, say, in Africa. Furthermore, one can use websites such as Google Earth for an aerial photograph and a map of almost anywhere on the planet.

Information stored and communicated via computers reveals more and more about our interactions, with each other and with our worlds. GIS combined with real-time satellite images and the Internet provides the equivalent of a central nervous system for the planet. Humans can aspire to provide the brain for that system. How we apply our brains to use these technologies and this information will transform how we live and, therefore, the patterns of our settlements.

As the information landscape advances, we can gain a better understanding of human ecology. For example, satellite imagery can produce daily climate information for settlements. GIS can be used to map these data over time and enable the climate information to be overlaid on land-use and land-cover maps. This process reveals how we use the land and how what we plant on its surface affects

Fig. 1 Blue Marble. Credit: NASA Goddard Space Flight Center Image by Reto Stöckli (land surface, shallow water, clouds). Enhancements by Robert Simmon (ocean color, compositing, 3D globes, animation). Data and technical support: MODIS Land Group; MODIS Science Data Support Team; MODIS Atmosphere Group; MODIS Ocean Group. Additional data: USGS EROS Data Center (topography); USGS Terrestrial Remote Sensing Flagstaff Field Center (Antarctica); Defense Meteorological Satellite Program (city lights).

urban climate. In this way, GIS and remote-sensing technologies enable us to visualize relationships. Since human ecology is essentially about relationships, our ecological understanding advances as we reveal previously unseen connections.

We especially gain insights into urban places. Urban morphology involves the study of human settlement patterns. People create nonurban settlements as well, ranging from farmsteads and rural villages to mines and ski lodges. While suburbia might lack urbanity, it is often classified as urban by geographers. Farmsteads and suburbia have specific morphologies as well which are important to understand. However, since we live in the first urban century, the morphologies of cities and metropolitan regions especially merit attention. Population trends indicate that the world is becoming more urban. For the first time in human history, over half the world's population lives in metropolitan regions. As the planet has urbanized, the structure of urban areas has attracted increased attention by scholars from many disciplines.

Urban morphology evolved from both the disciplines of geography and architecture in Europe, where a rigorous and thorough mapping of the physical structure of cities was promoted. Mapping revealed what the Italians call *tessuto*, or the tissues of the city – that is, clusters of structures, vegetation, and roadways that hold the urban body together. The Dutch use a similar concept and their word for tissue, *weefsel*, to describe urban tensegrity. The influence of urban morphology has spread among geographers, architects, and planners in Europe, North America, and Asia. Urban morphologists advocate reading the city as a text, or as a cultural palimpsest, to reveal culture.

Landscapes possess power such as both a cultural and a natural palimpsest. Landscapes offer a scale where social and physical processes and pattern can become evident. We see landscapes and all our senses react to their well-being.

Landscape ecology focuses on the ecological relationships at the landscape scale. Landscape ecology is a study of the structure, function, and change in a heterogeneous land area composed of interacting ecosystems. European scientists advanced landscape ecology before their American counterparts. The landscapes of Europe have been more densely settled than in North America, and, as a result, the human influence was recognized quickly by European scientists. American ecologists are more accustomed to studying relatively pristine landscapes. The refinement of the landscape ecology discipline, coupled with increased suburban sprawl nationwide, has changed this situation as more American ecologists acknowledge human interactions with natural systems. As landscape ecology has evolved through multiple interactions among European, American, and Australian contributors, it has crystallized into something new and powerful. Human settlements form mosaic-like patterns on landscapes and this land mosaic vision makes the landscape readily accessible to scientists, especially ecologists.

We can see change and interactions in landscapes. Edges – or interfaces – between land uses can be especially sensitive and rich. In rapidly growing regions, edges are unstable and conflicting. New homes replace farmland. The land sells relatively cheaply. The open land provides an attractive backdrop. Agriculture practices create dust and noise. Farming often depends on chemicals that have consequences for human health. Suburbanites possess different lifestyles and expectations that vary dramatically from those of their rural neighbors. Such landscape change lends itself to scientific analysis. For example, ecologists can ask, What interactions are driving the change and what patterns are resulting?

A growing interest in the ecologies of urban areas provides evidence of a coalescence of these catalysts for change. In the United States, National Science Foundation (NSF) established two urban Long Term Ecological Research (LTER) projects in 1997. Before setting up these projects in the Baltimore and Phoenix metropolitan regions, NSF located LTERs in nonurban places. Remote locations presented ideal places for ecologists to explore the traditional concept of stable states in relatively closed systems. Increasingly, influential American ecologists began to urge NSF to consider the ecology of metropolitan regions too in order to pursue the study of more complex systems. Urban ecological systems present multiple challenges to ecologists, including pervasive human impact and extreme heterogeneity of cities, and the need to integrate social and ecological approaches, concepts, and theories.

The Baltimore and Phoenix LTERs offer contrasting urban conditions. Baltimore, located in the northeastern region of the United States, is an older city than Phoenix and has a more dense urban fabric. The Sun Belt location of Phoenix offers a city developed as a result of automobile, airplane, air conditioning, and refrigeration technologies. Whereas growth in Baltimore is rather slow, population expansion in the Phoenix metropolitan region leads the nation. The humid Chesapeake Bay contrasts the arid Sonoran Desert. As a result, the Baltimore and Phoenix LTERs can help us understand constants in urban conditions as well as specific variations resulting from the natural surroundings and from the period of settlement.

Thus far, there has been relatively little interaction between the urban ecology camp dominated by scientists and the urban morphologists led by architects and planners. Geographers are present in both groups and likely will form bridges. The substance of such spans can be provided through better-understanding human ecology.

Human ecology is important if we are serious about sustainable development – that is, economic progress that meets all of our needs without leaving future generations with fewer resources than those we enjoy – a way of living from nature's income rather than mining its capital account. Sustainability requires that human communities are adaptable to change, that natural processes and landscape functions are protected, and that resources are conserved for future generations. To be adaptable, communities need to be resilient. We must understand the organization – the function, structure, and processes – of the communities that we inhabit in order to lay the foundations for the future.

Perhaps the growing interest in sustainable development – in seeking to make communities more livable – derives from a sense that we are living in places where something is out of whack. Perhaps the creative impulse derives always from a dread of the future, the feeling that the world may not improve for our children, and our desire to fend off doom to improve things for those who follow. To sustain things, we must keep them from falling apart, now and in the future. All around us, things indeed appear to be coming apart at the seams. Where once children played in the park, now homeless people sleep. Where there was once a vibrant downtown, there are now vacant lots.

The farm field, the park, the downtown; the convenience store, the homeless people, the vacant lots, all form pieces in larger mosaics, larger processes. In itself, the field or the convenience store is neither good nor bad. Both, however, are part of larger systems that may be either healthy or sick, that is, either capable of sustaining themselves or not. The individual farm field contributes to a regional agricultural system. The crops produced in the field help sustain the regional economy. The crops support not only the farm family that produces them, but the local co-op that processes the crop for the market and the tractor dealer as well. The convenience store has an asphalt parking lot. Its impervious surface contributes to regional drainage and flooding problems because of increased runoff. Because the parking lot is black, it adds to the urban heat island effect resulting in summer discomfort among nearby residents. The understanding of how living systems are organized from the local to the regional provides a means for assessing their capabilities to adjust to change.

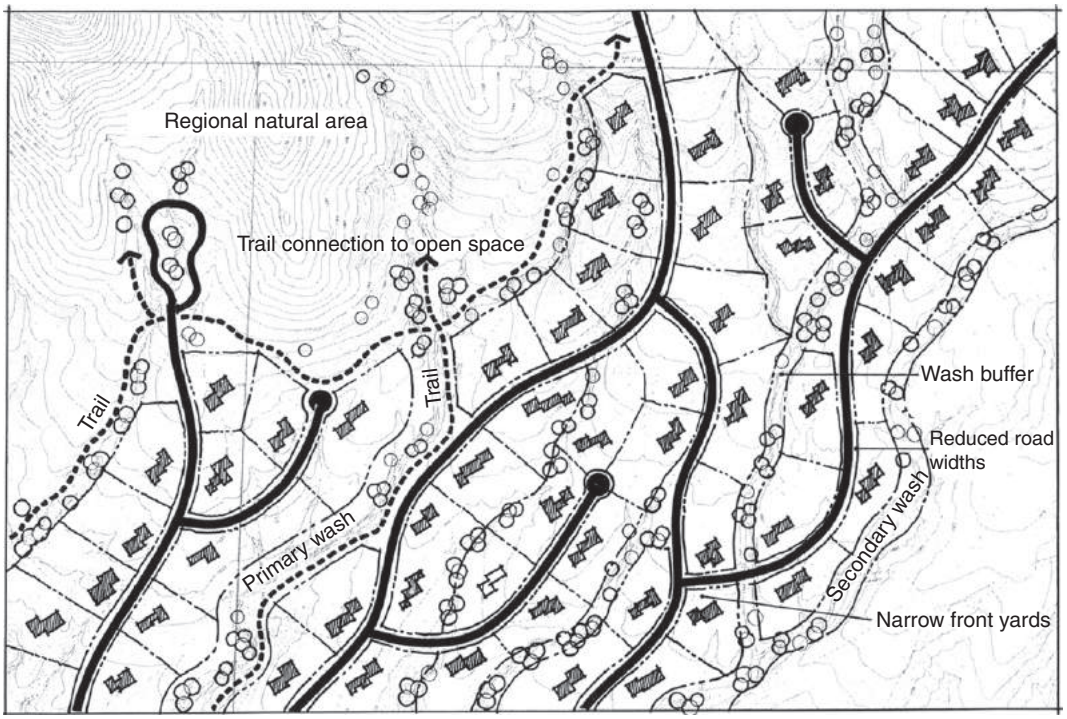


Fig. 2 Open space systems. Based on desert washes used to create a new form of suburban development in North Phoenix. From Steiner F (2000) *The Living Landscape: An Ecological Approach to Landscape Planning*. New York: McGraw-Hill.

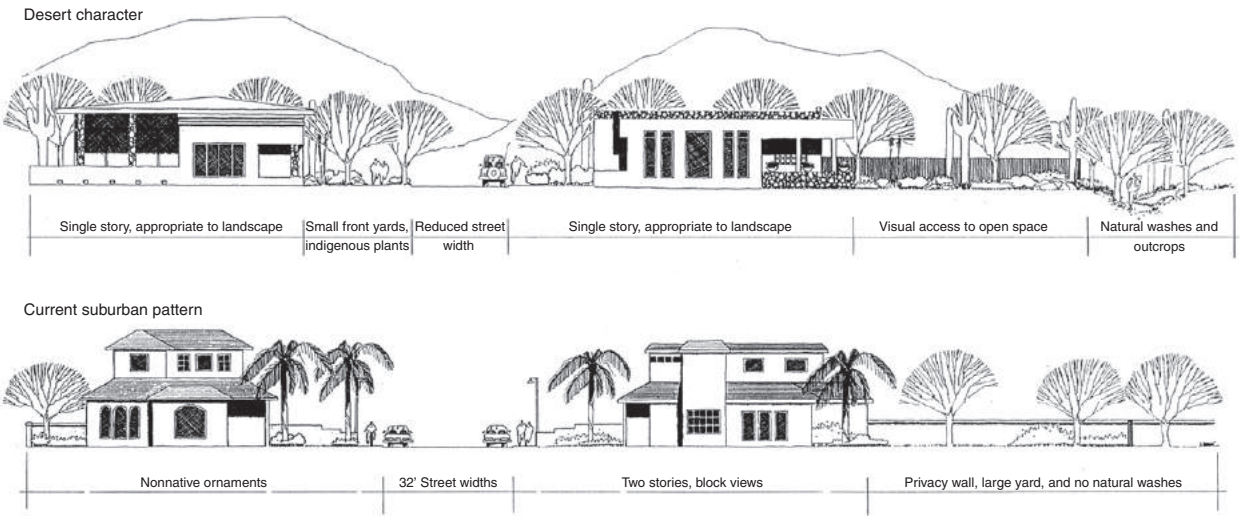


Fig. 3 Comparison of current suburban pattern in North Phoenix to one integrated with the desert. From Steiner F (2000) *The Living Landscape: An Ecological Approach to Landscape Planning*. New York: McGraw-Hill.

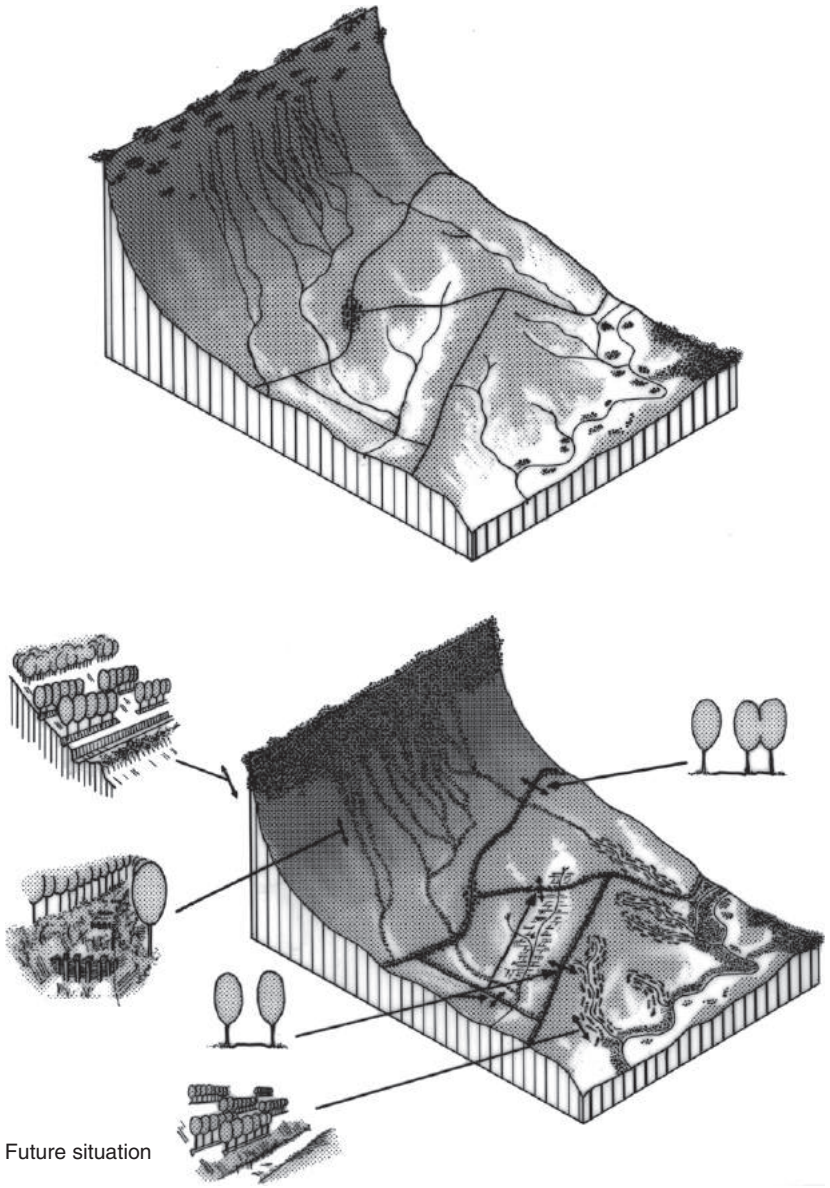


Fig. 4 Before and after of Kenyan hillside integrating agroforestry systems to produce food and wood while controlling storm runoff and soil erosion. From Duchhart I (2007) *Designing Sustainable Landscapes: From Experience to Theory.* Wageningen: Wageningen University.

Nested Networks

Living systems are organized hierarchically and communicate through feedback networks. The elements within a system may vary greatly in organization. According to urban morphologists, urban form can be understood at different levels of resolution. Commonly, four are recognized, corresponding to the building/lot, the street/block, the city, and the region. A building may be tall or short, with a pitched roof or a flat one, brick or wood or adobe. A lot, a street, or a block may be narrow or wide, straight or curved. A city may be densely settled or spread out. A region can be defined by a river or a mountain or a coastline or all three and by other factors.

Hierarchies help us understand how people are connected with one another – the basic idea of community. To understand human ecologies, the most relevant levels of organization include habitat, community, landscape, region, nation and state, and Earth or ecosphere. These levels present different, yet interconnected, scales of analysis. Each level possesses a history and a literature of analysis and debate. The habitat includes the building and lot. The community is comprised of buildings, lots, streets, and blocks. Landscapes can be urban, suburban, rural, and wild. Regions are hodgepodges of landscapes, while the distinctions between regions, and often those between states and nations, are even more blurred. But there is less ambiguity about the ends of the Earth.

Each level of human organization (nation, state, region, county, and city) is an element in a larger system, but is also comprised of smaller geographic units like neighborhoods and communities, which are, in turn, collections of single households. Home and work places form the habitats for people and are further divided into cells we call rooms. Hierarchy may be seen as a framework, a system of nested networks.

A critical feature of these nested networks is an asymmetric interaction in between levels. The larger, slower levels maintain constraints within which faster levels operate. There are, however, circumstances when slower and larger levels in ecosystems become briefly vulnerable to dramatic transformation because of small events and fast processes. Large, slow levels tend to keep things in place. Small, fast levels initiate changes when the larger levels are not functioning effectively.

Viewing the world hierarchically does not necessarily imply seeing it through a machine-like lens. Rather, it is to suggest components of a vocabulary to read our surroundings, our world.

Traditional ecology was commonly grounded in the assumption that somehow nature is in balance. Even the most casual observation of the human condition indicates that we are seldom balanced in our affairs. Nonequilibrium represents an important change in thinking. An equally, or perhaps even more important change derives from viewing environments at multiple, interacting scales. Landscape-level ecology, in particular, provides spatial form and function to nature's flows and human activities. New ecology, a deeper understanding of interactions at various scales, holds the prospect for better, although more complex, approaches to sustainable resource management, nature conservation, and environmental protection as well as the arts of environmental design and planning.

Practical Applications of Applied Human Ecology

Since the 1970s, natural and social scientists initiated multidisciplinary research addressing practical problems related to the environment. For example, the United States and many other nations require an environmental impact analysis of the consequences of larger projects. Such analyses are often performed by multidisciplinary teams, with human ecology providing a common set of concepts among disciplines.

Human ecology can also assist community and regional planning. For example, the Phoenix, Arizona (USA) metropolitan region is well known for its rapid growth and its suburban sprawl. Much of the post-World War II development has occurred in a similar pattern of low-density, single-family homes that is highly dependent on the automobile.

Beginning in the 1990s, city officials sought to encourage different patterns of development for North Area which comprised 20% of the land within the city. Using corridor, path, and matrix principles from landscape ecology, 28% of the most environmentally significant areas in the 110 square mile North Area were preserved as open space. Through an analysis of current and

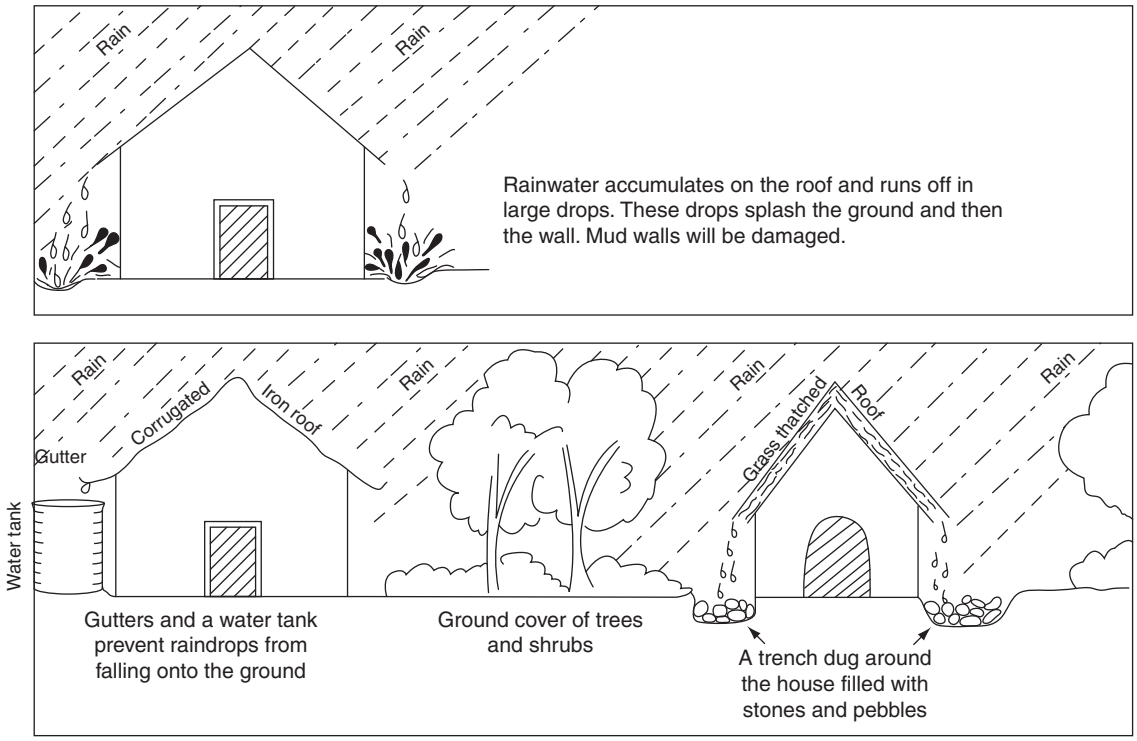


Fig. 5 Strategies for collecting rainwater in Kenya. From Duchhart I (2007) *Designing Sustainable Landscapes: From Experience to Theory*. Wageningen: Wageningen University.

potential residents, three future settlement patterns, instead of one, were suggested for the rest of the North Area. In suitable places along transportation corridors, greater urban density was recommended that included green ribbons of natural drainage. In other locations, a very low-density, low-impact rural desert settlement was suggested.

Since people in Phoenix are attracted to suburban development, suitable areas for such settlement were identified. However, a new form of desert suburban development was designed. This settlement would be aligned with natural drainage systems, preserve native vegetation and wildlife habitat, encourage the planation of native species, reduce the amount of impervious surfaces for roadways, use natural building materials with a local color palette, and keep building heights below the tree line (**Figs. 2** and **3**).

Around the world from Arizona, Kenya is also experiencing a growing population and declining natural resources. A multidisciplinary team of Kenyan and Dutch researchers conducted extensive landscape and human ecology analyses which led the Green Town program. The motto of the program was `make every town a green town'. In all, some 29 small towns across Kenya became involved in the effort which included considerable ecological training of local officials.

The Green Town program emphasized locally derived, sustainable designs. Shaded market areas were suggested as well as agroforestry practices to produce fuel and food. In addition to increasing fuel wood and food products, the agroforestry techniques reduced soil erosion and storm water runoff. The Green Town program suggested strategies for urban tree plantation as well as ways for individual homes to collect rainwater (**Figs. 4** and **5**).

Summary

Human ecology involves the interrelationships among people, other organisms, and their environments. Human ecology emphasizes complexity and change. Urban morphology and landscape ecology offer two approaches to study the structure, function, and processes of human settlements. Hierarchy also aids in the understanding of how people organize themselves spatially on various scales from the individual room within a house, office, school, or factory to the neighborhood and community on to the region, state or province, and nation. Applied human ecology presents many practical applications including the analysis of the environmental impacts from specific, proposed projects to the planning of communities and regions.

Further Reading

Botkin, D.B., 1990. Discordant Harmonies: A New Ecology for the Twenty-First Century. New York: Oxford University Press.
Botkin, D.B., Beveridge, C.E., 1997. Cities as environments. Urban Ecosystems 1, 3–19.
Duchhart, I., 2007. Designing for Sustainable Landscapes – From Experience to Theory. Wageningen: Wageningen University.
Forman, R.T.T., 1995. Land Mosaics: The Ecology of Landscapes and Regions. Cambridge: Cambridge University Press.
Marten, G.G., 2001. Human Ecology: Basic Concepts for Sustainable Development. Sterling, VA: Earthsean Publications.
Moudon, A.V., 1997. Urban morphology as an emerging interdisciplinary field. Urban Morphology 1, 3–10.
Pickett, S.T., Burch Jr., W.R., Dalton, S.E., *et al.*, 1997. A conceptual framework for the study of human ecosystems in urban areas. Urban Ecosystems 1, 185–199.
Steiner, F., 2000. The Living Landscape: An Ecological Approach to Landscape Planning. New York: McGraw-Hill.
Steiner, F., 2002. Human Ecology: Following Nature's Lead. Washington, DC: Island Press.
Young, G.L., 1989. A conceptual framework for an interdisciplinary human ecology. Acta Oecologiae Hominis 1, 1–135.
Zimmerer, K.S., 1994. Human geography and the `new ecology': The prospect and promise of integration. Annals of the Association of American Geographers 84 (1), 108–125.

Human Population Growth☆

Anne Goujon, Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/OEAW, WU), World Population Programme/International Institute for Applied Systems Analysis, Austria

Glossary

Demographic transition It depicts the transition of a country or a region from a demographic regime of high birth and death rates to one of low birth and death rates.

Second demographic transition It is the phenomenon that occurs in a country or a region where sustained sub-replacement fertility eventually results in declining population if not compensated by new migrants.

Total fertility Rate The number of live children who would be born per woman if all women lived through their childbearing years bearing children according to a current schedule of age-specific fertility rates.

Under five mortality rate It is the number of deaths of infants and children under five years old per 1000 live births.

Human capital It is a collection of traits representing the capacity of the people to be part of the labor force and to produce economic value. In models, it is often operationalized using the parameters of education, skills and health.

Introduction

The sheer number of population has been for long the main (and only) consideration about human population growth in ecological engineering and in other fields that are reflecting on the sustainability of life systems on earth. The inflation in the availability of data on humans' behavioral patterns depending on the characteristics of these humans and their environment should change that. What is important is not only how many people there are or will be but what they do and will do in their everyday life which could impact life systems and how equipped they are and will be to face the challenges of the future. This gap in understanding has been the reason why most of the doom prognosis were not realized—starting with Malthus' theory about population growth outstripping food production. They failed to predict the capacity of humans to innovate and overcome the challenges they were faced with within the limiting capacities of earth boundaries. However, the challenges exist, and in the coming decades, the survival and well-being of humans and the security of environmental resources that support human existence will continue to be challenged by rapid population growth, particularly in less developed regions that are the main contributors to world population that is, sub-Saharan Africa and Southern Asia. Nevertheless it is clear nowadays that almost all societies that all started with patterns of high fertility and high mortality are moving through the demographic transition process. As we enter the new millennium, the capacity of humans to come up with innovations, solutions and adaptive measures will need to be strengthened in order to deal with the stark contrasts between the availability of natural resources and the billions of humans who require them to sustain life. Education will play an essential role in that.

The Demographic Transition Theory

The Demographic Transition Theory is the driving theory of the evolution of human population. Developed by Notestein in 1946, the theory categorized the several stages of historical and present population into a continuum of demographic development. Eventually according to this theory, all societies evolve from a pre-transition situation (stage 1) where fertility and mortality are unchecked and high producing low population growth to a stationary population (accomplished in stage 4) which is realized when a society reaches low levels of fertility and mortality. This is quite certain and the exceptions so far have been only of temporary nature. While fertility decline has stalled at several times and in different settings, for instance as a result of cuts in social government budgets in many sub-Saharan Africa following the implementation of structural adjustment policies by the World Bank and the International Monetary Funds, the fertility decline has resumed thereafter. The main uncertainty is about the pace of the fertility and mortality decline between stage 1 and stage 4, which has many implications for population growth. In the case of most industrialized countries, where the transition was slow and happened over many decades, population growth was not so dramatic. This was the case in many European and other industrialized countries. The situation has been different in countries that started their transition in the second half of the 20th century when many advances had already been made and hence mortality rates dropped relatively abruptly while fertility was kept at quite high level for some time. While it is very likely that the fertility will go down, it is difficult to know how long it will take. Whether for instance Nigeria reaches replacement fertility in 2070–75 according to the low fertility variant of the United Nations or in 2095–100 according to the medium variant, makes a different of about 130 million in 2100 (567 vs. 794 million): the population of Mexico nowadays. As we

☆*Change History*: March 2018. Anne Goujon updated text.

This is an update of D. Pimentel and M. Pimentel, Human Population Growth, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 1907–1912.

 doi:10.1016/B978-0-12-409548-9.10755-9

will see in the next section, the pace of the fertility decline has indeed huge bearing on the future number of humans, not only for Nigeria. The second uncertainty, also with important bearings is the number of children the population would eventually have. This goes behind the theory of Notestein, adding a further stage (stage 5), also called the second demographic transition in which fertility is so low that generations do not replace themselves and the population starts declining. Here again, most industrialized countries seem to be on this path. For instance in many Eastern Europe countries, women have less than 1.5 children on average and fertility does not seem to recover to replacement level fertility. In those countries, also with the effect of emigration, population is already declining for example, in Bosnia Herzegovina or in Hungary. Japan is another country where fertility has been declining for some time and since 2010, the population has been declining. Whether all countries that are at present going through the demographic transition will also experience the second demographic transition will also be key to the future number of people on the planet. Some researchers contend that the arrival of contraception has ruptured the evolutionary link that existed between the sex drive and reproduction, the fertility of women and couples will be more and more determined by their individual preferences for children (or not) and by norms that will be culturally ascertained. In that sense, it is not necessary that individuals and couples, and societies at large, will want to replace themselves. We will need a few more decades to see the evolution of that phenomenon and whether some societies are trapped in low-fertility. It is worth noting that the governments in countries that are faced with low levels of fertility most often engage in providing financial and structural benefits to support childbirth that are not crowned with success, for example, in South Korea, tending to support the point about an individualization of the fertility decision process (see also the case of China below).

But having the two uncertainties in mind, about the pace and the bottom level of fertility that tells us that there is hope that world population will eventually peak and diminish, it is still not automatic; it is also pretty certain that the 21th century will witness continued population increase, and that these will not occur globally but will be concentrated spatially in the poorest socio-economic settings. To halt the growing imbalance between human population numbers and their essential resources, humans must actively conserve cropland, freshwater, energy, and the other basic environmental resources. There is a critical need to alert the public worldwide to the serious issues of overpopulation and natural resource shortages. Education will be key in that both in developed and developing countries. While most of population growth will happen in developing countries, certainly populations in developed countries could contribute to the conservation effort by reducing their high consumption of all resources, especially fossil fuels. Focus is needed on improving food crops, such as developing perennial grains, pest-resistant crops, and improved nutritional makeup of crops. The development of ecological engineering can help encourage the development of sustainable ecosystems and aid human society to make better use of our natural resources. There is a critical need to educate the public worldwide in order to comb overpopulation and natural resource shortages.

In this article we will examine the current world population situation and how it might evolve depending on several scenarios. We will focus on the role that quality education could play in providing humans with the necessary capacity and ecological engineering requirements to deal with present and future environmental challenges.

World Population Growth

Global Picture

World population growth did not receive much attention until the end of the 20th century for several reasons: the first one being the absence of systematic population data collection exercises and the second that world population increase was minimal due to high mortality rates especially among infants and children. Hence it went unnoticed when the world population reached 1 billion at the turn of the 19th century. The current world population of more than 7.6 billion, doubled during the last 46 years. The uncertainties and impact of the several component of population growth are visible in **Fig. 1** showing estimates of the past and the future number of humans based on several scenarios developed by the United Nations. Although, no probabilities are attached to any of them some seem more plausible than others and some more desirable than others. Several lessons can be taken from this graph. The three top scenarios that give the highest population in 2100 are those where fertility stays constant or declines only marginally. According to those variants, the world population would be between 17 billion (high fertility) and 26 billion (constant fertility). These scenarios imply that fertility decline would come to a stall everywhere—which is very unlikely—and would have dire consequences in terms of population growth in poverty stricken countries. The five remaining scenarios all project the world population to be under 11.5 billion in 2015. The most likely one under present circumstances is the medium fertility variant—it shows the population increasing to 11.2 billion by the end of the century with the world population still not having reached a peak (but leveling off convincingly). It is interesting to see the impact of the momentum of population growth that can be seen in one variant showing a peak at 9.3 billion around 2060, and then started declining. Indeed, worldwide, a major obstacle to limiting population growth is the relatively young age structure—especially in the respective female fecund ages between 15 and 49—with high reproductive rate. Even if all the people in the world adopted a fertility pattern of bearing only two children per couple, it would take approximately 40 years before the world population would finally stabilize at approximately 10 billion. The momentum—also called population inertia—becomes also visible in the fact that until 2050, there is little deviation between the global population numbers although they follow different scenarios. While trusting more the five scenarios that result in a lesser population growth, it is still obvious that they lead to a 2–2.5 billion in population increase from today's level by the mid of century. The low fertility variant is interesting and seems the most suitable one from an ecological point of view as it means a world population peak in 2053 at 8.8 billion further declining to 7.3 billion in 2100. It is worth noticing that these scenarios do not and cannot take into account abrupt events that could occur at any time, such as the spread of a new epidemic, a war, or a baby-boom.

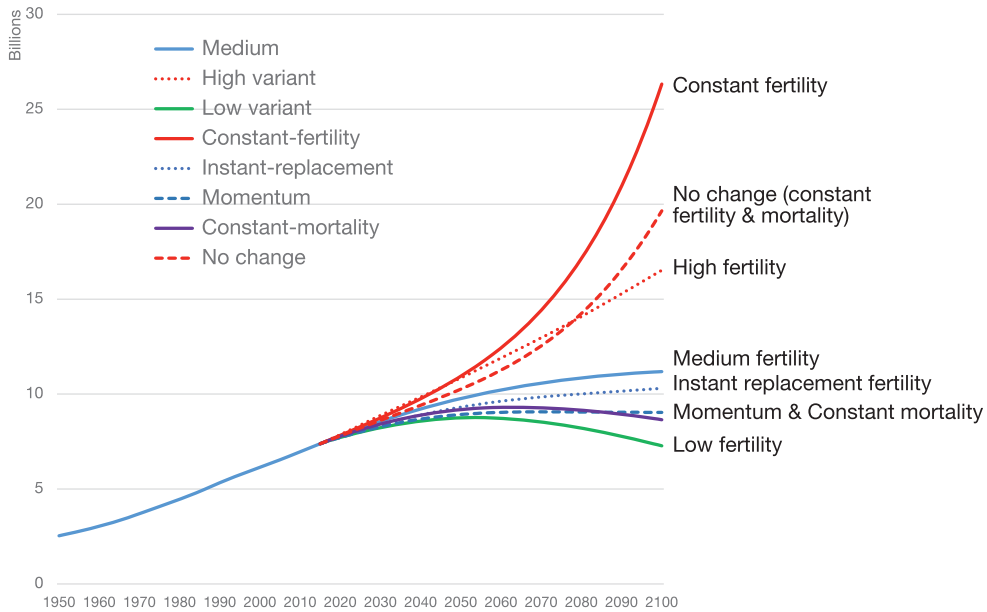


Fig. 1 World population, 1950–2100, estimates and projections according to eight variants. Source: United Nations World Population Prospects (2017).

Regional and Country Level

Global population levels hide important differences in the projected pace of population growth across continents, regions and levels of development (see **Table 1**). In 2010–15, while the world population was increasing at a rate of 1.2%, some regions were shrinking for example, Eastern and Southern Europe or close to stabilization, for example, Western Europe, some others were increasing at a strong rate, for example, Eastern, Middle, and Western Africa. However, population growth rates are on the decline in all countries of the world today—this is visible when comparing for instance the growth rates observed in 2005–10 and 2010–15—and are forecasted to do so in the future as well. **Table 1** shows that at present, less developed regions are the main contributors to world population growth and particularly sub-Saharan Africa and Southern Asia. What is clearly visible is that sub-Saharan Africa would be the main contributor and by far according to the United Nations medium variant in the second half of the century.

The population of China with 1415 million people will keep slowly increasing until 2030 when it would peak at 1441 million and then slowly decline back to 1021 by the end of the 21st century. In 2017, the government ended the one-child policy (that was in effect for more than 30 years) in order to balance rapidly aging population by allowing families to have two children. However, the first year of implementation supports the low-fertility-trap theory that the country has moved to a lower-than-replacement fertility norm and that couples are less likely to wish and bear more children than the previous generation for multiple reasons: strong urbanization and price constraints, women's participation in the labor force, decrease in the preference for boys, etc. India, with 1354 million people, and living on approximately one-third the land either of the United States or China, would keep growing albeit at a slow average annual rate of 0.5% until 2060—surpassing China's population around 2025—and then decline. Total fertility rates have declined very rapidly in all Indian States and in about half of them, women had below replacement fertility in 2016 (e.g., in the highly populated States of Tamil Nadu and West Bengal), and the increase in population is mostly due to the momentum of population growth. At present, the populations of China and India constitute more than one-third of the total world population. They would be around 23% in 2100.

As above mentioned, the sub-Saharan Africa sub-continent will experience most of the world population increase in the next decades. According to the medium variant of the United Nations, population in sub-Saharan Africa would more than double from around 1050 million in 2018 to more than 2.5 billion in 2060 and possibly exceed 4 billion in 2100. The strong increase can be explained by two main factors. The first is that against all expectations (in respect of levels of socio-economic development) infant and child mortality rates have recorded a substantial decline of more than 50% between 1970–75 and 2010–15. As a result, life expectancy has increased by more than 20 years since 1950, from 36 to 57 years, despite the HIV/AIDS epidemic that has slowed down and even reversed the progress made in many countries particularly in Southern Africa. The second factor is the fact that fertility has been slowing down at a snail's pace. Women were having on average more than six children between 1950 and 1995. Fertility started declining faster after the 1995 period but it was still around 5.1 children on average in 2010–15. The fact that the demographic transition to low fertility and mortality levels has adopted a slow pace stems from several reasons having to do the persistence of low levels of socio-economic development in mostly rural traditional societies. The pace of future declines will be an important determinant of population growth. Increase in the levels of educational attainment could play an important role, most notably through the impact it has on fertility and mortality (see section below). In some of the largest countries (Nigeria, Democratic Republic of the Congo, Tanzania, Uganda), the population would more than quadruple between now and 2100

Table 1 Contribution (absolute and percentage) of regions to world population growth from 2015 to 2100

		Increase in absolute (in million) between			*Contribution to global increase (in percent) between*		
		2015 and 2030	*2015 and 2050*	*2050 and 2100*	*2015 and 2030*	*2015 and 2050*	*2050 and 2100*
World	World	+1168	+2389	+1413			
Level of development	More developed regions	+37	+45	−13	+3%	+2%	−1%
	Less developed regions	+1131	+2344	+1426	+97%	+98%	+101%
	Least developed countries	+378	+960	+1282	+33%	+40%	+91%
	Less developed regions, excluding least developed countries	+754	+1384	+144	+67%	+59%	+10%
Continent	Africa	+509	+1333	+1940	+44%	+56%	+137%
	Asia	+527	+837	−476	+45%	+35%	−34%
	Europe	−1	−25	−62	0%	−1%	−4%
	America	+126	+226	−3	+11%	+9%	0%
	Oceania	+8	+18	+15	+1%	+1%	+1%
World regions	Sub-Saharan Africa	+449	+1198	+1834	+38%	+50%	+130%
	Eastern Africa	+188	+489	+690	+16%	+20%	+49%
	Middle Africa	+84	+230	+369	+7%	+10%	+26%
	Northern Africa	+60	+135	+106	+5%	+6%	+7%
	Southern Africa	+11	+22	+7	+1%	+1%	+0%
	Western Africa	+166	457	+768	+14%	+19%	+54%
	Eastern Asia	+44	−49	−388	+4%	−2%	−27%
	Central Asia	+13	+26	+6	+1%	+1%	0%
	Southern Asia	+311	+558	−151	+27%	+23%	−11%
	South-Eastern Asia	+93	+163	−26	+8%	+7%	−2%
	Western Asia	+65	+138	+83	+6%	+6%	+6%
	Eastern Europe	−12	−35	−40	−1%	−1%	−3%
	Northern Europe	+8	+14	+9	+1%	+1%	+1%
	Southern Europe	−4	−12	−26	+0%	−1%	−2%
	Western Europe	+7	+7	−5	+1%	+0%	0%
	Latin America and The Caribbean	+86	147	−68	+7%	+6%	−5%
	Northern America	+39	+79	+65	+3%	+3%	+5%

Table 2 Population of the top 10 most populous countries in 2018, at peak and in 2100

Country	*Population in 2018*	*Population at peak*	*Time of the peak*	*Population in 2100*
China	1,415,046	1,441,574	2029	1,020,665
India	1,354,052	1,678,656	2061	1,516,597
United States of America	326,767			447,483
Indonesia	266,795	324,763	2062	306,026
Brazil	210,868	232,845	2047	190,423
Pakistan	200,814	354,297	2090	351,943
Nigeria	195,875			793,942
Bangladesh	166,368	202,970	2057	173,549
Russian Federation	143,965	143,990	2017	124,013
Mexico	130,759	167,327	2063	151,491

according to the medium variant of the United Nations. Nigeria would have a population of almost 800 million in 2100 and Niger, one of the poorest country of the world would multiply its population eight times in the next 80 years (from 22 million today to 192 million in 2100).

The United States population is also still growing and currently stands at nearly 327 million, having doubled during the past 60 years. Based on the current trend, it is projected to reach 400 million by 2060 and 450 by the end of the century. Out of the ten most populated countries shown in **Table 2**, the United States and Nigeria are the only countries that are not forecasted to attain a maximum over the century. The population of the Russian Federation has peaked in 2017 according to the estimates/projections and is shrinking.

As world populations continue to expand, all global resources will have to be divided globally among increasing numbers of people and per capita availability would decline to ever lower levels. However and as already mentioned, the population pressure

will not be the same everywhere. At the level of less developed countries where this kind of pressure will be the strongest, improving personal health, and achieving prosperity, a suitable quality of life, and personal freedom will be more difficult and will require innovation and leadership. Population models and ecological engineering taken up by education may help people and governments better understand the critical situation and what, if anything, can be done to address the challenge.

The challenges that earth systems will face with an increasing population are serious and of different kinds. Malnourishment is one of them. In 2016, in a world where there is enough food to feed each and every one, about 815 million people (11% of the world population) were affected by hunger and for the first time, after declining over a decade, global hunger was on the rise again according to the 2017 United Nations State of Food Security and Nutrition in the World report. The increase is due to the proliferation of conflicts in combination with severe drought and flood episodes. Providing enough food to meet the nutritional need of the 2.5 billion people that will be added between now and 2060 is one of the biggest challenge of the century. Achieving food security requires a multi-sectoral approach from conflict resolution to strengthening the productivity of small-scale food producers, improving the resilience of food production systems and the sustainable use of biotechnology and genetic resources.

Education

Some population growth is certain. While many scientists including ecological engineers research on the way to optimize and increase the resources in terms of cropland, water resources, energy, while at the same time preserving biodiversity and avoiding as much as possible climate change, some other academic research focuses on the pathways to influence the number of human population and secondly the behavior of the population. Scientific findings tend to demonstrate that in both education has a role to play (Goujon, 2003).

Education and Sustainable Development

The transformative channels through which formal education affects demographic parameters are well known. Formal education is mostly important for its influence on the fertility of women that have been through the education system and on the mortality of children born to those women. The mechanisms are multiple: girls who go to school first of all are exposed to the environment and society outside of their own household and often of their neighborhood, breaking their isolation. There is often already a large difference between the number of children born to a woman who had been to primary school, even if she has not completed all grades and a woman who has never been to school; the difference is as large as two children on average in Kenya and in the Dominican Republic. If girls stay in school, they will delay marriage and hence the onset of fertility. Furthermore, having an education increases the chance of finding employment outside of the house and therefore increases the opportunity cost of having children. The more education a woman has the more likely she is to use modern contraceptives, space the birth of her children. By increasing her autonomy, education will increase her say in household (including fertility) decisions. The transition to lower fertility supposes that there is a time when fertility is not given anymore but becomes within the calculus of choice of couples, particularly women. Education is one of the most influential factor that will make sure that this happens. **Fig. 2** shows the difference in the number of children born to women with a higher education (any studies beyond upper-secondary) and those born to women who have never been to school (or less than a year). Differences are larger in sub-Saharan Africa and can exceed four children like in Angola (more than five-child difference), Mozambique and DR Congo.

As well, education will have an impact on the number of surviving children of these women. This is important for the demographic transition as limiting the number of surviving children will first occur when the probability of child survival increases. **Fig. 3** shows that the chance of survival to the 5th year is much higher for children born to mothers with a high education (secondary or higher) than to mothers who have a low education (primary or less) in many developing countries in recent years. For instance, in Burundi, the mortality rate of children born to mothers with a primary education or less was as measured in 2010 more than twice that of children born to mothers with a secondary or higher education (132 vs. 47 deaths for 1000 live-births). Many surveys confirm that maternal education is the single most significant determinant of child mortality. Educated mothers are more likely to seek treatment for their sick infant or child and break with tradition about illness. Having gained autonomy through education, educated women will be more apt to challenge their mother in law or the head of the household. The education of women is also likely to affect beneficial changes in nutrition that play a part in a decline in infant and child mortality.

What could be the impact of strong increases in educational attainment on future population growth is shown in **Fig. 4** which compares the results of two scenarios. The left-hand side graph assumes a future in which the world is moving toward a more sustainable path (so called "Sustainable world" scenario). It assumes that educational and health investments accelerate the demographic transition, leading to a relatively low world population with increased well-being. As a result in 2100, the world population is already shrinking and the vast majority is well educated. The graph on the right hand side refers to a fragmented world with an emphasis on security at the expense of international development. In this divided world, population growth is assumed to be high in developing countries and low in industrialized countries (so called "Fragmented world" scenario). As a result the world population is high at the turn of the century and population growth is unabated. Moreover large segments of the population particularly in less developed countries have low levels of educational attainment.

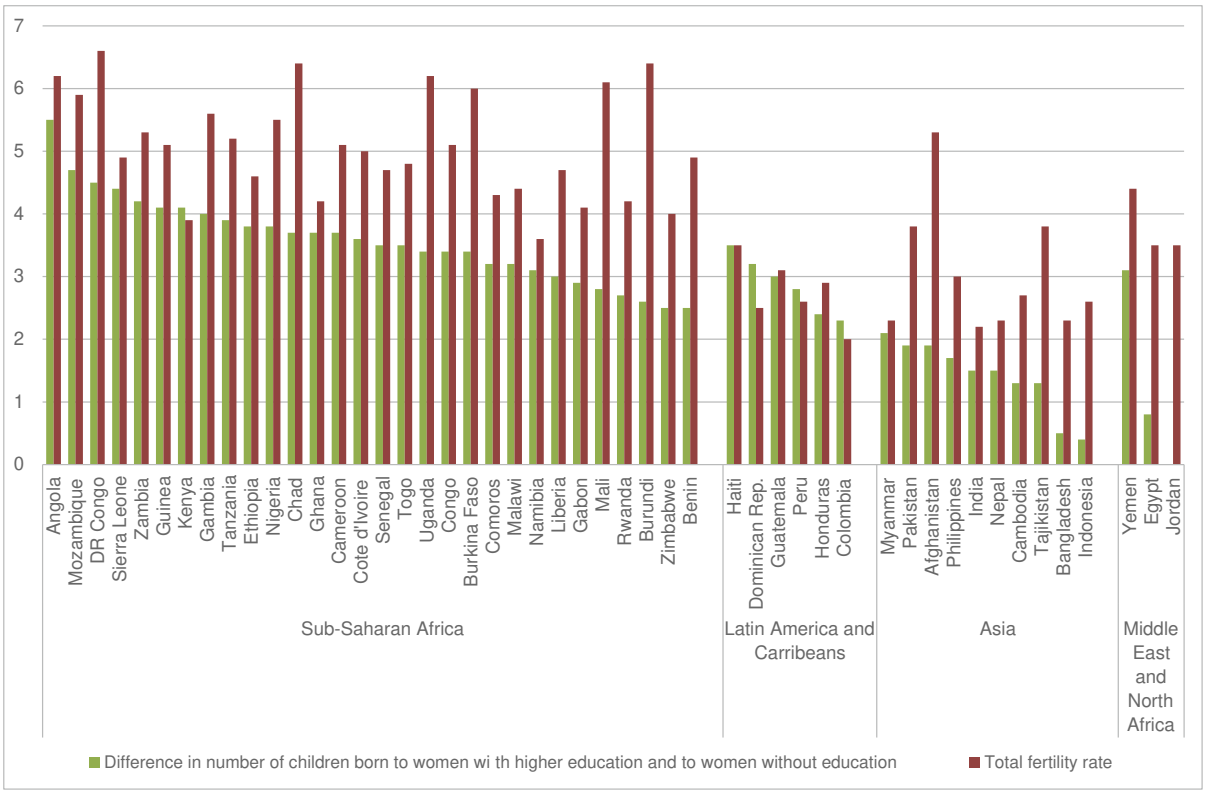


Fig. 2 Difference in the number of children born to women with a higher education and to women without education and total fertility rates, selected less developed countries, since 2010. Source: ICF International, (2015). The DHS Program STATcompiler. Funded by USAID. https://www.statcompiler.com, (February 9, 2018).

Beyond the purely demographic effect of education on the number of people, it is clear that education is key to development. It increases the innovation capacity which will be necessary to face the challenges. It also increases the resilience and adaptive capacity of humans in case of a disaster (such as a Tsunami) or another triggered climate change extreme event. The effect goes beyond the mere survival to extreme events as it was shown in a study of population in Thailand in the aftermath of the 2012 Indian Ocean earthquakes. Living in a community with a higher proportion of women with high education increased preparedness through community engagement of the more educated and carrying out disaster risk reduction measures. Overall the more educated have a longer time planning horizon, which is also the reason why they adopt healthier behaviors (related to nutrition, sport practice and smoking behavior for instance) and tend to be more supportive of environment-friendly behavior (vegan diet, consuming bio-product, biking to work, etc.).

Education for Sustainable Development

Education for sustainable development (ESD) was introduced as part of the targets of Sustainable Development Goal four on education (in Targets 4.7). It is accompanied by an approach to promote the Global Citizenship Education (GCED) within education for sustainable development. The main idea is that transformative ideas are better learned at young ages—the same as other competencies such as riding a bicycle or politeness. This will mean changing the curricula in many countries to have more global components and as well developing cross-cutting sustainability competencies in pupils. This would enable individuals to contribute to sustainable development by promoting societal, economic and political change as well as by transforming their own behavior. Educated people are also more likely to live in democracy and in less corrupted regimes as they may influence the system. Whereas the physical constraints of the planet are more or less set, the future innovative capacity and behavior of its inhabitants are not known. Education may be in this sense the missing link to sustainability.

Why Prediction of Human Dooms Were Wrong (Until Now)?

Along the centuries, concerns were often raised by scientists who were the witness of rapid population growth. The most cited essay by Malthus (1798) warned against the danger of exponential population growth that would ensue out of the evolution of the

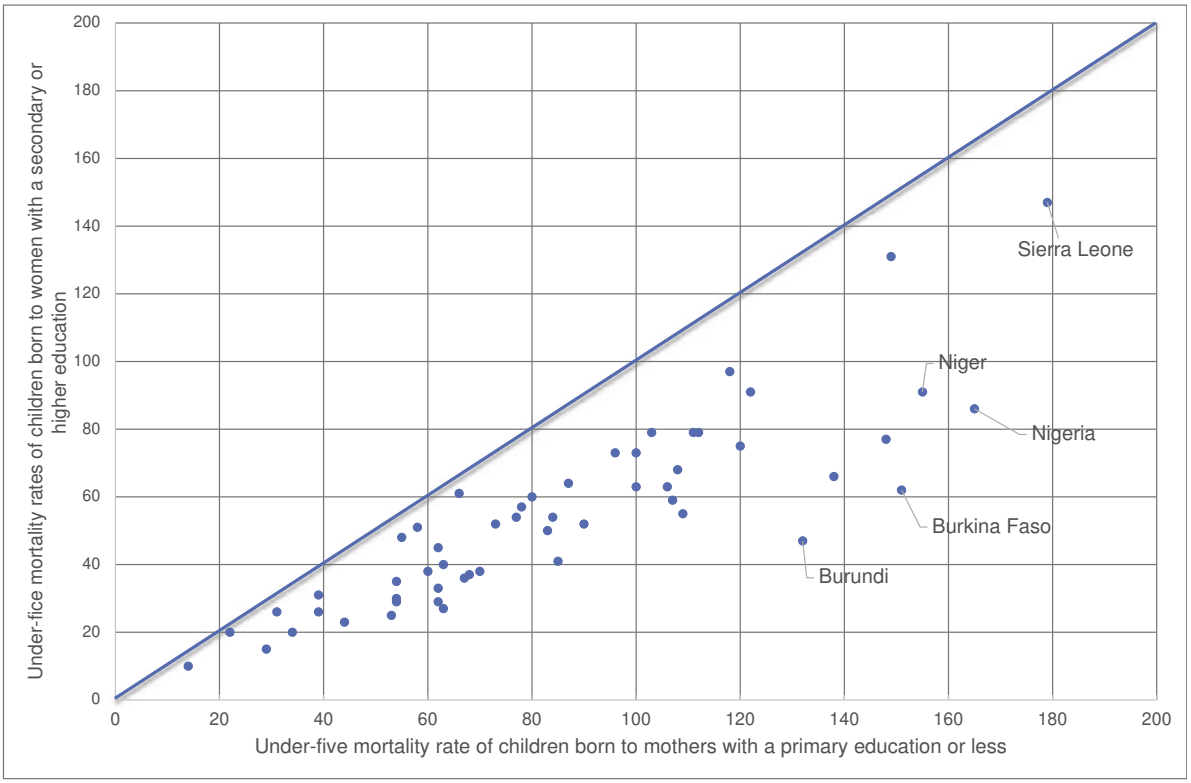


Fig. 3 Comparison in 54 countries of the under-five mortality rates of children born to mothers with a primary education or less (*x*-axis) and those born to women with a secondary education or more (*y*-axis). Note: Probability of dying before the fifth birthday (in the 10 years preceding the survey (5 years for total)) per 1000 live births. Source: ICF International, (2015). The DHS Program STATcompiler. Funded by USAID. https://www.statcompiler.com, (February 9, 2018).

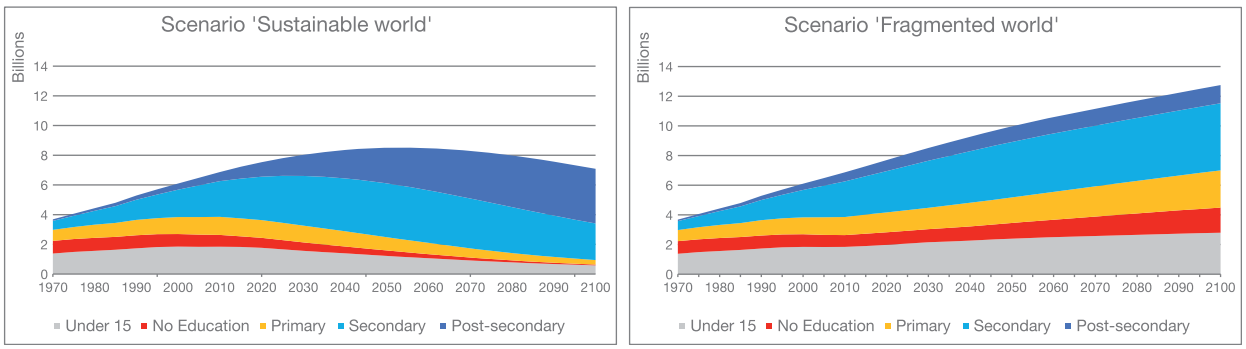


Fig. 4 World population in 2100 according to two scenarios. Note: Scenario "Sustainable world" is based on SSP1 scenario and "Fragmented world" on SSP2. Source: Wittgenstein Centre for Demography and Global Human Capital. (2015), Wittgenstein Centre Data Explorer Version 1.2. Available at www.wittgensteincentre.org/dataexplorer.

different components of population growth in the presence of limits to food production. At this time, the agricultural and industrial revolutions had improved the living conditions and healthcare in Europe, mortality rates were declining and fertility rates were unabated. This essay was the first of many alarmist theories by scientists such as "The Population Bomb" book authored by Paul Ehrlich in 1968 that predicted famines and civil-war in the 1970s and 1980s as a result of population growth. This book was a precursor to the "The Limits to Growth" report published in 1972 which projects the combined impact of exponential population growth with a finite supply of resources. So far those predictions have proved wrong because they take humans as irrational being incapable of changing and adapting to circumstances. Already at the time of Malthus, Condorcet (1743–94) was pointing at the power of an improved intellect. Democratic governance, access to family planning, and the education and economic empowerment of women has proven so far to be able to curve population growth.

The merit of these alarmist scientific works is that they brought to a wide audience, awareness about population and environmental issues.

Conclusion

While it seems almost certain that the world population will peak within the next 100 years, it will be preceded by the probable addition of 2–3 billion people. Most importantly these people will be born in countries that are at the moment most stricken by poverty. While this is a challenge for humanity and for the ecosystems surrounding them, it is also evident that it does not have to be a failure. There is a need for action at present and in the future for influencing the behavior of the people who live on the planet. Education might be key in bringing the change that will help societies to face the global challenges.

See also: Global Change Ecology: Urbanization as a Biospheric Process: Carbon, Nitrogen, and Energy Fluxes. Human Ecology and Sustainability: Urban Systems; Human Ecology: Overview; Limits to Growth; Urban Metabolism

Further Reading

Bongaarts, J., Casterline, J., 2013. Fertility transition: Is sub-Saharan Africa different? Population and Development Review 38 (s1), 153–168.
Caldwell, J.C., 1981. Maternal education as a factor in child mortality. World Health Forum 2 (1), 75–78.
Cochrane, S.H., 1979. In: Fertility and education: what do we really know? *World Bank staff occasional papers* no. OCP 26Baltimore, MD: The Johns Hopkins University.
Goujon, A., 2003. Demographic transition and education in developing countries. In: Sirageldin, I. (Ed.), Sustainable Human Development,*Encyclopaedia of Life Support Systems* (EOLSS), Developed under the Auspices of the UNESCO, vol. 3. Oxford: Eolss Publishers. http://www.eolss.net 15.1.
Goujon, A., Samir, K.C., Speringer, M., Barakat, B., Potancoková, M., Eder, J., Striessnig, E., Bauer, R., Lutz, W., 2016. A harmonized dataset on global educational attainment between 1970 and 2060—An analytical window into recent trends and future prospects in human capital development. Journal of Demographic Economics 82 (3), 315–363.
Jejeebhoy, S., 1998. Women's education, autonomy, and reproductive behaviour: Experience from developing countries. Oxford: Clarendon Press.
Kirk, D., 1996. Demographic transition theory. Population Studies 50 (3), 361–387.
Koons, D.N., Holmes, R.R., Grand, J.B., 2007. Population inertia and its sensitivity to changes in vital rates and population structure. Ecology 88 (11), 2857–2867.
Lesthaeghe, R., 2014. The second demographic transition: A concise overview of its development. Proceedings of the National Academy of Sciences of the United States of America 111 (51), 18112–18115.
Lutz, W., Muttarak, R., 2017. Forecasting societies' adaptive capacities through a demographic metabolism model. Nature Climate Change 7 (3), 177–184.
Lutz, W., Skirbekk, V. and Testa, M. R. (2006). The low-fertility trap hypothesis: Forces that may lead to further postponement and fewer births in Europe. Vienna Yearbook of Population Research Vol. 4 (2006): 167–192, Vienna, Austria: Verlag der Österreichischen Akademie der Wissenschaften.
Lutz, W., Crespo Cuaresma, J., Sanderson, W.C., 2008. The demography of educational attainment and economic growth. Science 319 (5866), 1047–1048.
Lutz, W., Butz, W., Samir, K.C. (Eds.), 2014. World Population and Human Capital in the Twenty-First century. Oxford: Oxford University Press.
National Academy of Sciences, , 1997. National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. In: Preparing for the 21st century: The education imperative. Washington, DC: The National Academies Press.
van de Walle, E., 1992. Fertility transition, conscious choice, and numeracy. Demography 29 (4), 487–502.

Relevant Websites

esa, n.d.—https://esa.un.org/unpd/wpp/, United Nations Population Division Projection and Estimates.
Data Explorer, n.d.—http://dataexplorer.wittgensteincentre.org/shiny/wic/, Wittgenstein Centre Data Explorer.
wfp, 2017—https://www.wfp.org/content/2017-state-food-security-and-nutrition-world-sofi-report, United Nations State of Food Security and Nutrition in the World 2017.
World in Data, n.d.—https://ourworldindata.org/, Our World in Data.
Worldometers, n.d.—http://www.worldometers.info/world-population/, Worldometers.
Sustainable Development, n.d.—https://sustainabledevelopment.un.org/, United Nations Sustainable Development Goals.
statcompiler, n.d.—https://www.statcompiler.com, Data from Demographic and Health Surveys.

Industrial Ecology☆

F Duchin, Rensselaer Polytechnic Institute, Troy, NY, USA
SH Levine, Tufts University, Medford, MA, USA

Industrial Ecology: Human Activities and Global Ecosystems

The field of 'industrial ecology' was created to inform purposive human decision making about industrial production processes, especially as they impact the environment, by taking advantage of knowledge about the functioning of healthy ecosystems. These ecosystems are characterized by their vigor, maintenance of diversity, resilience, and relative stability over time. Other ecosystem properties seen as potentially desirable in industrial systems include minimal use of virgin materials and maximal use of renewable, biodegradable resources, minimal production of toxic waste products, and, more generally, sustainable use of resources. More recently the scope of industrial ecology has enlarged to include decision-making regarding consumption activities as well. Human decision-making is important because humans, while one species among many, are responsible for dramatic and far-reaching changes to the global environment and thus to the existence of other species.

Industrial ecology is not the only new, cross-disciplinary field employing the word 'ecology'. Among others utilizing the ecological metaphor are the ecology of the family, organizational ecology, and ecological economics. In all cases the intention is to suggest that the field in question constitutes a complex system, in some instances with direct relations to biological ecosystems. The 'ecology of the family' focuses on how family members are shaped by their interactions, both competitive and mutualistic, with one another. 'Organizational ecology' applies population dynamics models to the births and deaths of firms and industries and examines how the evolution of organizational structure is shaped by competitive and cooperative interactions. 'Ecological economics' departs from the observation that natural ecosystems uninfluenced by human activity no longer actually exist: counting both economists and ecologists among its numbers, it attempts to integrate the two professional domains. Industrial ecology is distinctive in its focus on the industrial system, literally treating it as an ecosystem or, more exactly, the subsystem of principal interest within a more inclusive ecosystem. The flows of energy and material characterize an industrial system and serve as the integrating focus of all description, design, and analysis in industrial ecology. Not merely analogous to the flows of energy and material in ecosystems, they actually constitute those flows. The production and consumption activities of humans can be described in terms of these flows just like the activities of any other animal. Given this focus on industrially valuable resources, it is not surprising that most of the field's founders were engineers or applied physical scientists with interests in chemical processes and effluents and, more generally, the reduction, reuse, and recycling of industrial wastes.

Industrial ecology is motivated by its concern for the well-being of the environment. As the inclusion of ecology in its name indicates, it puts special emphasis on developing and implementing solutions and policies at the system level, up to and including the global system. Central to the system perspective is the concept that the behavior of individual components cannot be fully understood without reference to the system in which they interact. Industrial ecology explicitly recognizes that human industrial activities in the modern industrialized world are characterized by the interdependence of many industries, each industry itself often performing many interconnected production processes, all reliant on inputs of energy and materials and discharging wastes.

System scientists have long understood that the basic features of a system of interacting components need to be understood in a top-down fashion, even though many processes operate at the level of component parts. In the case of system design, top-down and bottom-up contributions generally proceed in an iterative fashion. This system perspective influences the direction that industrial ecology takes in confronting environmental challenges: improving the environmental compatibility of individual industrial processes is evaluated in the context of improving the overall industrial system. A simple illustration is provided by Fig. 1, which combines two waste streams, fly ash from coal-fired power plants and waste plastic from plastic manufacturing, into a useful product, light-weight building blocks. At the same time the waste heat and carbon dioxide from the power plant can be supplied to a large-scale greenhouse, leading to increased production of fruits and vegetables. These are both examples of 'open loop' waste reuse. ('Closed loop' use of waste is represented in Fig. 1 by the recycling of waste plastics by the plastics factory.) Reducing any of these waste streams might both increase the unusable waste from the others while also reducing the quantity of useful product. Interestingly, recognizing the value of waste as a resource is a major theme of industrial ecology.

The tracking of resources, intermediate products, final products, and wastes can be conducted at the level of business establishments, towns, nations, or in the context of nations interacting in the global economy. The analysis of these flows, as a basis for action, can also take place at all these levels.

Resources and Products in Industrial Systems

The industrial system and the natural system were long seen as separate although overlapping domains. Economists, for example, used to treat resources as 'free gifts of nature' and therefore exogenous to their concerns, limiting the designation of so-called factor

☆*Change History*: August 2014. F Duchin and SH Levine updated the text and references.

 doi:10.1016/B978-0-12-409548-9.09407-0

inputs to built capital and labor. (Many still do, contrasting the built environment with the natural environment.) It was a contribution of industrial ecology to treat the industrial system as a subsystem of the natural system (see Fig. 2(a)).

As industrial ecology now broadens its concerns to include consumption as well as production, this simple diagram is being reconsidered because of the importance of human motives and human agency to the decision-making process. The built environment is part of the biophysical structure of the social–ecological system and subject to the laws of nature (see Fig. 2(b)), but the cultural sphere is better understood using the concepts and methods of the social sciences.

In the earliest societies, the built environment was insignificant. Humans in hunter–gatherer societies, like all other species, primarily take what nature makes available. They are not likely to severely overexploit a local ecosystem; when resources become scarce, they move on. However, even before the advent of agriculture, humans began to modify their local environments for the purpose of increasing its productivity (for humans) through the use of fire. With the transition from hunter–gatherer to settled agriculture, farming and herding societies greatly extended the modification and control of the natural system to overcome natural supply constraints and produce more of what humans desired. This control allowed for the possibility of individuals or groups producing more of agricultural and other goods than they required and thus making some of their production available for trade. With the development of trade and of markets, agricultural and other types of output were converted into products: goods (or services) exchanged for something else of value.

A second and related historical transition is the explosive growth of human requirements, the average of what is considered standard in a given society and the related changes in human consumption patterns. Early human requirements differed little from those of other social mammals, were closely related to physical survival, and were met by the output of nature. In modern times humans are unique among species in the volume, variety, and sources of their material requirements. Consumer demand in affluent societies is met by an extensive array of products – the goods and services output by the industrial system. Humans are not alone in producing products, but no other species even approaches the scale of human production. The modern built environment reflects the prevalence of these human products.

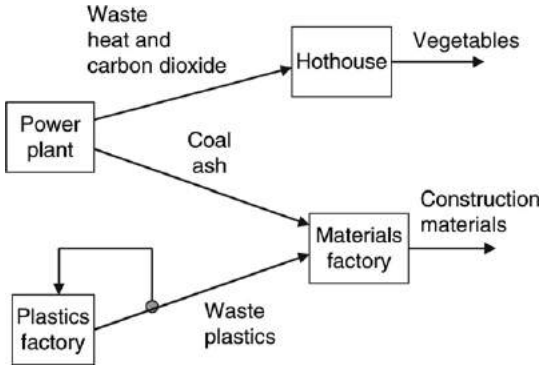


Fig. 1 A hypothetical system for the reuse and recycling of potential wastes.

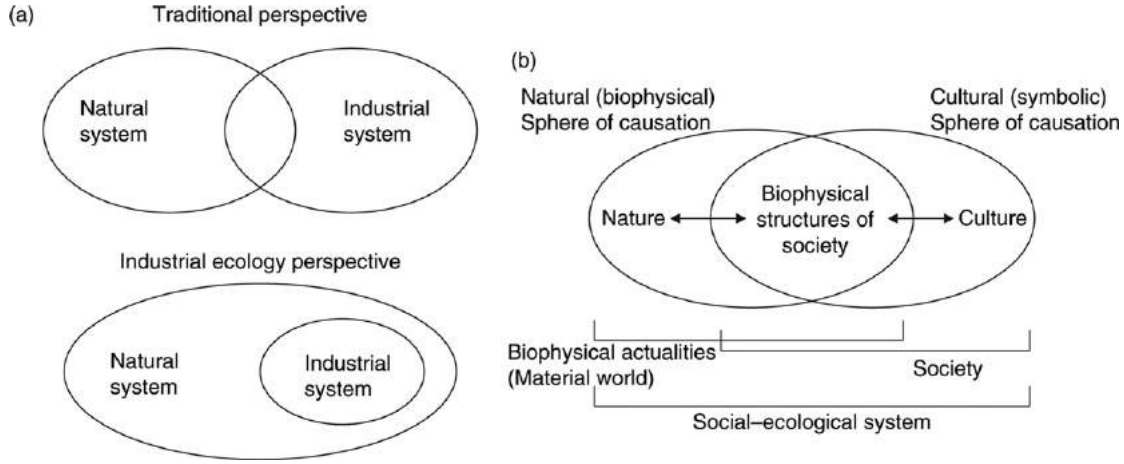


Fig. 2 (a) Conceptual framework of industrial ecology. (b) Social–ecological systems as overlap of a natural and a cultural sphere of causation. (a) Modified from Chertow. M., Portlock. M., 2002 *Developing Industrial Ecosystems*: *Approaches, Cases, and Tools*. New Haven, CT: Yale University. (b) Reproduced from Haberl H, Fischer-Kowalski M, Krausmann F, Weisz H, and Winiwarter V (2004) Progress towards sustainability? What the conceptual framework of material and energy flow accounting (MEFA) can offer. *Land Use Policy* 21: 199–213, with permission from Elsevier.

While the flow of mass and energy in 'natural' ecosystems is largely dictated by the consumption of resources to supply energy and nutrients to sustain life, many, if not most, products of modern industry have little to do with directly providing energy and nutrients, and a substantial number have little to do with that function even indirectly. The extent to which industrial systems are dedicated to producing such products contrasts them to the rest of the natural system. Industrial ecology is concerned with a unique feature of these industrial systems: the unprecedented degree to which the appropriation of resources – materials and energy – for the fabrication of products is not bounded by the metabolic constraints of the biological world, both in the quantity of those flows and in the variety of materials involved.

One could say that humans in a modern consumer society have developed extended metabolic needs, where consumer goods and services play a role similar to the need for proteins and carbohydrates in nature. To have toast in the morning requires not only bread but also a toaster and thus electric power as well. This concept of humans having extended needs is hardly new. Rousseau saw industrialization as creating a set of artificial (as opposed to natural) needs, and Marx made the distinction between human and inhuman needs. The material and energy requirements of the modern industrial system serve the extended needs of human consumers. The concept of distinct industrial metabolisms, reflecting the material realities of specific societies, and attempts to quantify them, is an active research area in industrial ecology.

Since the majority of industrial products do not satisfy biological metabolic requirements, they need not be composed of organic material. The relaxing of this constraint, coupled with the specialized functions of many products, leads to the development and use of a wide range of novel, often exotic, materials designed specifically to improve product function. In some cases the development of new products is made possible only by the development of new materials, which in turn often requires the development of new industrial processes. As a result the global ecosystem must cope with stocks and flows of materials that have undesirable properties such as toxicity or nonbiodegradability.

Most of these complex materials cannot be recycled easily, in contrast to the relatively narrow range of naturally occurring materials, which are readily recycled.

Within the realm of personal transportation, for example, the goal of making automobiles lighter in order to reduce fuel consumption has led to the substitution of a variety of high-technology materials in the place of steel, itself an exotic material when compared to the wood from which earlier forms of personal transportation such as wagons were constructed. Use of these materials impedes the recycling of automobiles due to the increased difficulty of separation and remanufacture. In many cases the performance of the product is dependent on its containing materials that are nonbiodegradable. These products are more likely to be toxic, require large quantities of energy for their production, and have long residence times in the environment, problems that are aggravated by the great quantities in which they are produced.

These materials and products are the output of what is often a multistage production process carried out within the industrial system. Just as the diet of a carnivore depends on the consumption of plants by herbivores, the end-product purchased by the consumer requires many indirect as well as direct inputs. Large volumes of consumer products, requiring diverse resources and intermediate inputs as they do, make it necessary to build what are often large-scale production facilities. The material and energy resources required for this part of the built environment, like other private and public buildings and infrastructure, must be counted among the extended metabolic needs of human consumers. Describing and analyzing the total requirements of consumers in a modern industrial economy ultimately rests on accounting for all of the indirect inputs. To do this industrial ecologists often make use of economic models, especially those with a joint focus on material stocks and flows as well as production technologies and consumption patterns.

Bottom-Up: Resources and Products

The objective of industrial ecology is to provide actionable input to those taking decisions about the industrial system. These decision makers include not only corporate managers, responsible for the extraction of resources and fabrication and distribution of products, and the policy makers who oversee them, but also consumers, acting both individually and as members of households or of the growing number of social institutions that operate in the public interest. The decisions may pertain to an individual resource or product or to an entire industrial system.

Even when studying an individual resource or product, industrial ecologists are concerned with its role in an industrial system. They gather data on system behavior and develop concepts and methods for analysis at the system level. Among industrial ecology's most fundamental activities is gathering data for describing the flow of energy and materials throughout an economic system. The scale of the economic system can range from a single factory to a regional economy to the global economy.

Many industrial ecologists are engaged in compiling information to describe the flow of a specific material from sources to sinks within the global industrial system or some geographically delimited portion of it. Conventions and procedures have been developed for material flow analysis (MFA, and its variants such as substance flow analysis), which is in certain ways similar to the study of nutrient cycling in ecology. The depiction of the global copper cycle, shown in **Fig. 3**, is reminiscent of illustrations of the carbon cycle in ecological systems, but closer inspection indicates that the copper cycle is nowhere near as efficient as the carbon cycle. A significant fraction of the copper is not recycled but instead ends up being lost to the system. Identifying and quantifying these major flows provides grounding for interventions that aim to modify the existing system to reduce its impact on the environment by making it more efficient in its use of resources. Similar approaches are applied not only to individual elements such as copper but also to compound materials. By pinpointing the presence of wastes that go unnoticed in conventional

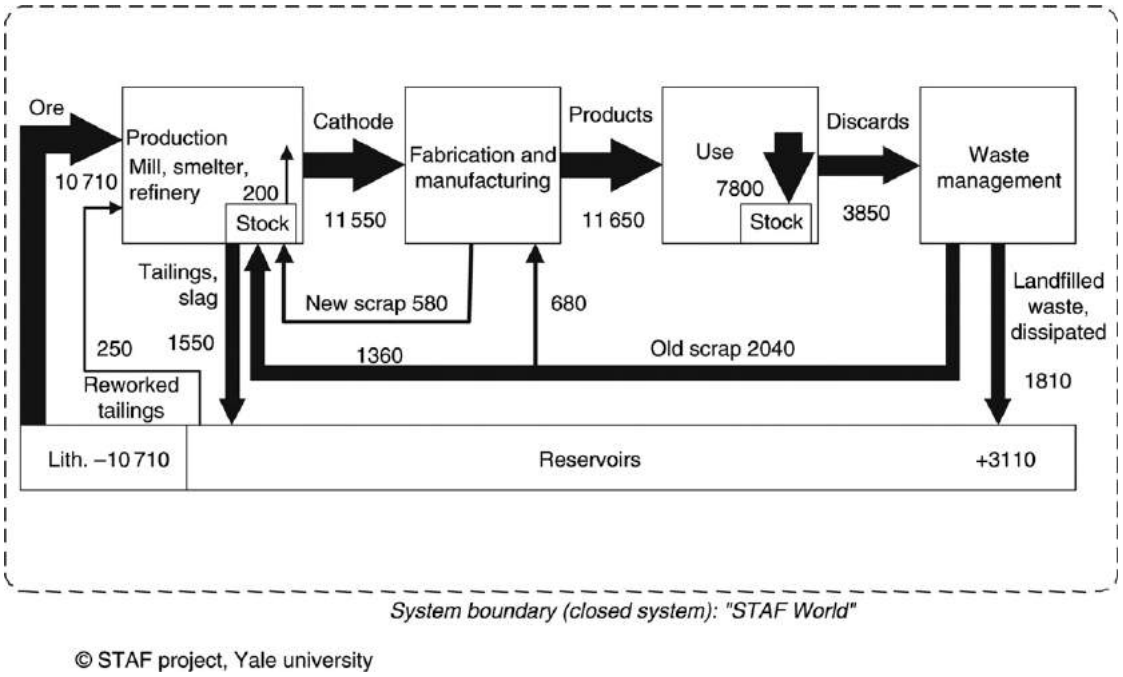


Fig. 3 The global copper cycle in the 1990s. Reprinted with permission from Graedel, T.E., van Beers, D., Bertram, M., *et al.*, 2004. The multilevel cycle of anthropogenic copper. Environmental Science and Technology 38 (4), 1242–1252. Copyright (2004) American Chemical Society.

economic monitoring systems, MFA can direct efforts to improve system performance. MFA has also been applied to the aggregate of all materials (measured in tons per person) in attempts to quantify a society's industrial metabolism.

The wide range of MFA studies underscores the increasingly great variety of distinctly different materials in use in the contemporary industrial system, each with unique properties, to support the great number of products required by consumers. The literature includes MFA studies of metals such as copper and zinc, forest products such as pulp and paper, and industrial chemicals such as bisphenol A and nonylphenol. This expansive range of materials contrasts industrial with ecological systems which, relying on the breakdown of complex biochemical compounds by digestion, is typically characterized by a few major cycles – carbon, nitrogen, phosphorous, hydrogen, oxygen – the principal, elemental constituents of biomass. Moreover, since the resources required by industrial systems often are found in specific locations distant from the site of production, which may also be distant from where consumption takes place, considerable amounts of transportation may be required, involving even more energy and material inputs. The material flow analyses of broadest conceptual scope trace simultaneously the physical movements associated with a variety of interrelated human activities. While the objective of most MFA studies is to quantify material flows, such flows can in an additional step provide inputs to a model of the industrial system that explicitly distinguishes their use in specific industrial sectors and modes of transportation to satisfy the product demand of different categories of consumers.

Moving up from resources to products, two additional research areas within industrial ecology are concerned with the individual product, situated in the context of its entire life cycle from resource extraction through resource processing and fabrication of the product to its utilization, reuse, recycling, and disposal. These are design for the environment (DfE) and life-cycle assessment (LCA). Both address the system-wide environmental impacts associated with products.

DfE involves the design of industrial products and processes to minimize their adverse environmental impacts over their lifetimes. Often it is a redesign of an existing product or process that is undertaken. The focus can be on different phases of a product's life cycle, such as design for product retirement. Actual applications are varied and have included the design of chemical processes, electronic products, mechanical components, and freezer insulation, as well as the increasingly important area of packaging design.

LCA involves the evaluation of the environmental impact of a product during its entire life history on the basis of detailed technical information. Each stage from extraction of resources through disposal of residuals is associated with distinct resource requirements and emissions or other forms of damage, and their impacts, which are experienced at specific times and places. LCA is often used to compare the environmental impacts of alternative products or production processes and has been applied to substances such as chlorine and aluminum, the entire mining industry, industrial materials like PVCs, and to alternative uses of agricultural land.

LCA studies quantify emissions and resource use per units of output or service delivered. The modeling of the production network, including the quantification of the amounts of inputs required from different production processes, is traditionally based on direct measurement or engineering analysis. This process inventory modeling has generally ignored the contribution of nonphysical inputs, such as legal and accounting services or wholesale and retail trade, and left out minor inputs to make the

analysis tractable. Empirical studies have established that such an approach overlooks a significant portion of the total impact. As a result, increasing numbers of LCA researchers are integrating their analyses with the use of input–output (IO) models of the economy to capture indirect as well as direct requirements associated with a single process or product. The integrated, hybrid approach has made it possible to go beyond examining individual products to examining the impacts of one entire bundle of consumption goods as compared to another.

LCA also includes an impact assessment step, in which different types of emissions are aggregated to a manageable number of indicators reflecting specific problem areas such as global warming or human toxicity. Alternatively, impact assessment can be based on the modeling of damages, for example, human health effects measured in years of life lost from both toxicity and climate change. The development of these impact assessment methods build on the knowledge and models of environmental scientists, including eco-toxicologists and ecologists.

Top-Down: Entire Industrial Systems

An early application within industrial ecology of ecological concepts to industrial systems is the design and implementation of so-called industrial ecosystems, or eco-industrial parks. Industrial ecosystems are characterized by the prevalence of what has been named industrial symbiosis, a relationship between two or more firms that involves the exchange of materials, energy, or information in a manner that is mutually beneficial. The most famous of these is located in Kalundborg, Denmark; its structure is illustrated in Fig. 4. By utilizing what would otherwise be waste products from one firm as input resources for others, the adverse environmental impact of this system of firms can be greatly reduced.

Industrial ecologists undertake to design industrial ecosystems either from scratch or around an existing plant. Kalundborg, however, emerged in the absence of advance planning and represents a sequence of accommodations and agreements between pairs of firms. Like any ecosystem, this eco-industrial park is continually evolving. New firms may be introduced. Some existing firms increase in size or modify their product lines and input requirements, while other firms decrease in size or disappear altogether.

Some industrial ecologists have turned their attention to larger systems, namely entire economies, using the concepts and methods of input–output (IO) economics, a systems approach to describing and analyzing an entire economy in terms of the inputs and outputs of dozens or even hundreds of individual industries, products, and resources. The use of IO models in

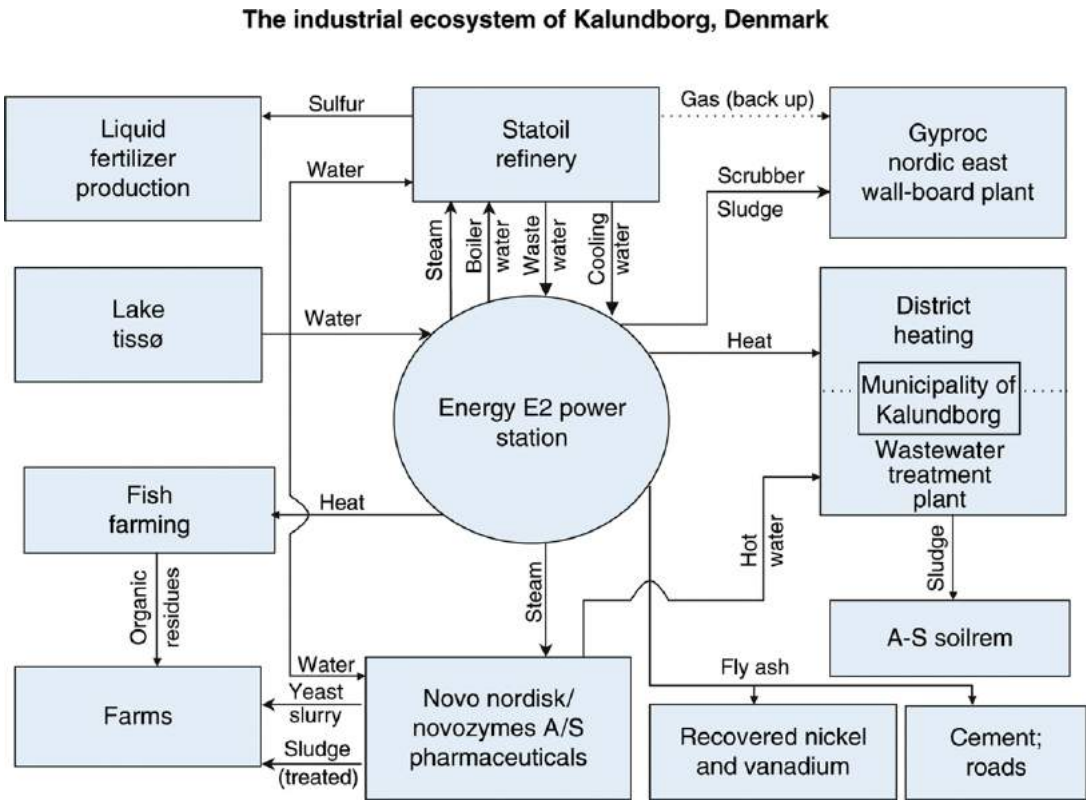


Fig. 4 The industrial ecosystem of Kalundborg, Denmark. Modified from Chertow. M., Portlock. M., 2002 *Developing Industrial Ecosystems: Approaches, Cases, and Tools*. New Haven, CT: Yale University.

industrial ecology has grown substantially in recent years as their ability to describe both physical stocks and flows and the associated money costs and prices has been emphasized and expanded. IO models require a database, a large portion of which for past years is provided on a periodic basis by national statistical offices around the world. When evaluating scenarios about the future, the framework is reliant on technical data about resources and products, and it increasingly makes use for this purpose of the kinds of information originating in MFA and LCA studies.

IO studies have investigated such environmentally significant challenges as water scarcity and water management in different parts of the world including China, Spain, and Southern Africa; emissions of carbon dioxide and other greenhouse gases; and the management of a variety of wastes. Emissions of carbon, sulfur, and nitrogen under alternative scenarios about future technological attributes have been estimated for the world economy described in terms of the production and trade in the outputs of a few dozen industrial sectors in over a dozen geographic regions. As concern builds that the industrialized countries are appropriating disproportionately large shares of the Earth's resources via resource-intensive imports from developing countries, IO models of the world economy that incorporate data from MFA and LCA studies are becoming more prevalent.

An IO model can represent the complex interplay of ecological and industrial system concepts. Since being introduced in a modified form into ecology, IO models have been valued by ecologists for their ability to track the paths of flows, thus accounting for indirect as well as direct interactions and allowing for a more accurate estimate of the total (direct plus indirect) energy and biomass requirements. IO techniques have also been the basis for developing measures of ecosystem structure, such as throughput and cycling, which have been applied to the analysis of numerous ecosystems. The direct relevance of these ecological system measures to industrial systems is evident, especially when recycling of resources is of major interest. As a result these ecological measures have recently been introduced into industrial ecology in, for example, analyzing material flow in the nylon tufted carpet industry.

Consumption and Sustainable Development

Since its origins, industrial ecology has been mainly concerned with reducing the environmental impact of the use of energy and materials in industrial production by improving the efficiency of production processes. Some industrial ecologists have claimed that material inputs could be reduced by substantial amounts (e.g., a factor of 4 or even 10) without diminishing economic growth. Such technological prescriptions are mainly addressed to public policy makers and corporate executives on the implicit assumption that they would not require much change in the motives or behaviors of consumers or negatively impact their well-being.

Recently, however, a new emphasis on consumers and consumption has emerged. Many researchers came to doubt that reliance on changes in the sphere of production alone could achieve the required scale of impact. The inflow of social scientists into industrial ecology brought the recognition that households are the major decision makers regarding consumption, and analysis about alternative consumption behaviors should be addressed in the first instance to them. In a consumer society, industrial stocks and flows are demand driven, in contrast to the dynamics governing the availability of traditional materials in ecosystems. In an ecosystem, predators have control over the number of prey they seek to consume, but they have no direct control over the number of prey potentially available for consumption. By contrast human consumers, particularly in the most affluent, industrialized economies, have control over not only how much gets consumed but also over what gets consumed and produced – and through this connection potentially over how it gets produced.

Life-cycle analysts and IO economists have produced a substantial body of work analyzing the environmental impacts of different types of households and their consumption activities. Most attention has focused on motorized mobility, housing, and intake of food because of the demonstrably intensive use of resources to satisfy consumption requirements in these areas. The objective of this research is to explore alternative ways for satisfying the human need for food, housing, and mobility with less environmental damage. Recent studies have provided a framework for bringing physical measures into the analysis of social accounting matrices, data sets compiled by a number of national statistical offices that describe the consumption patterns of distinct types of households. A special issue in 2005 of the *Journal of Industrial Ecology* is devoted to the industrial ecology of sustainable consumption.

Such investigations are part of a broader inquiry about the sustainability of systems and, in particular, sustainable development of the global economy. Development of alternate energy systems that can substantially reduce reliance on fossil fuels by making more direct use of solar energy in both production and consumption, thus moving industrial systems back in the direction of ecological systems, promises to be an active area of research. A shift in the diets of the affluent from animal-based toward more plant-based foods could have substantial impacts on resource use in agriculture. Such scenarios about the future will be analyzed using frameworks that integrate material flow data, life-cycle descriptions of products and processes, and IO models of individual economies and of the world economy. As increasing numbers of researchers with roots in different disciplines turn their attention to the challenges of sustainable global development, the distinctive contributions from industrial ecology will reflect its origins in the ecology of the industrial system.

See also: Human Ecology and Sustainability: Life-Cycle Assessment; System Sustainability. Terrestrial and Landscape Ecology: Ecological Engineering: Overview

Further Reading

Ayres, R.U., 1989. Industrial metabolism. In: Ausubel, J.H., Sladovich, H.E. (Eds.), Technology and the environment. Washington, DC: National Academy Press, pp. 23–49.
Ayres, R.U., 1997. The life cycle of chlorine, part I. Chlorine production and the chlorine–mercury connection. Journal of Industrial Ecology 1 (1), 81–94.
Baccini, P., Brunner, P.H., 1991. The metabolism of the anthroposphere. Berlin: Springer.
Chertow, M.R., 2000. Industrial symbiosis: literature and taxonomy. Annual Review of Energy and the Environment 24, 313–337.
Duarte, R., Yang, H., 2011. Input–output and water: introduction to the special issue. Economic Systems Research 23 (4), 341–351. (special issue).
Duchin, F., 1998. Structural economics: measuring change in technology, lifestyles, and the environment. Washington, DC: Island Press.
Duchin, F., Lange, G., 1994. The future of the environment: ecological economics and technological change. New York: Oxford University Press.
Fischer-Kowalski, M., 1998. Society's metabolism, part I. Journal of Industrial Ecology 2 (1), 61–78.
Fisher-Kowalski, M., Hüttler, W., 1999. Society's metabolism, part II. Journal of Industrial Ecology 2 (4), 107–136.
Graedel, T.E., Allenby, B.R., 2003. Industrial ecology, 2nd edn. Upper Saddle River, NJ: Prentice-Hall.
Graedel, T.E., van Beers, D., Bertram, M., *et al.*, 2004. The multilevel cycle of anthropogenic copper. Environmental Science and Technology 38 (4), 1242–1252.
Hertwich, E., 2005a. Life cycle approaches to sustainable consumption: a critical review. Environmental Science and Technology 39 (13), 4673–4684.
Hertwich, E.G., 2005b. Consumption and industrial ecology. Journal of Industrial Ecology 9 (1–2), (special issue).
Lave, L.B., Cobas-Flores, E., Hendrickson, C.T., McMichael, F.C., 1995. Using input–output analysis to estimate economy-wide discharges. Environmental Science and Technology 29 (9), 420A–426A.
Lenzen, M., Moran, D., Kanemoto, K., Foran, B., Lobefaro, L., Geschke, A., 2012. International trade drives biodiversity threats in developing nations. Nature 486, 109–112.
Mageau, M.T., Costanza, R., Ulanowicz, R.E., 1995. The development and initial testing of a quantitative assessment of ecosystem health. Ecosystem Health 1 (4), 201–213.
Moriguchi, Y., Kondo, Y., Shimizu, H., 1993. Analyzing the life cycle impact of cars: the case of CO_2. Industry and Environment 16 (1–2), 42–45.
Springer, N., Duchin, F., 2014. Feeding nine billion people sustainably: conserving land and water through shifting diets and changes in technologies. Environmental Science and Technology 48 (8), 4444–4451.
Udo de Haes, H.A., Jolliet, O., Finnveden, G., *et al.*, 2002. Life cycle impact assessment: striving towards best practice. Pensacola, FL: Society of Environmental Toxicology and Chemistry.

Life-Cycle Assessment☆

MA Curran, BAMAC, Ltd, Rock Hill, SC, United States

Introduction

Life-cycle assessment, or LCA, is an environmental accounting and management approach for assessing industrial systems. It considers all the aspects of resource use and environmental releases associated with a system, as defined by the function provided by a product, process, or activity. This cradle-to-grave approach considers all relevant impacts upstream and downstream of the consumer or producer. Specifically, LCA is a holistic view of environmental interactions that covers a range of activities, from the extraction of raw materials from the Earth and the production and distribution of energy, through the use, and reuse, and final disposal of a product (for simplicity, the word "product" is used although the life-cycle concept applies equally well to processes and activities). LCA is a relative tool intended for comparison and not absolute evaluation, thereby helping decision makers compare all major environmental impacts when choosing between alternative courses of action (Fig. 1).

LCA evaluates all stages of a product's life from the perspective that they are interdependent, meaning that decisions made at one point along the life cycle can have consequences somewhere else. LCA enables the estimation of the cumulative environmental impacts resulting from all stages in the product life cycle, often including impacts that go beyond the boundaries of traditional analyses. By including the impacts throughout the product life cycle, LCA provides a comprehensive view of the environmental aspects and a more accurate picture of the true environmental tradeoffs in product or process selection. A life-cycle approach helps us recognize how our choices influence each point of the life cycle, so that we can balance potential tradeoffs and avoid shifting problems from one area to another, thereby positively impacting the overall environment. LCA is a way of thinking about the choices we make, when purchasing products, selecting materials, or identifying process alternatives by putting our decisions in context with facts related to all parts of the life-cycle system.

In connecting the different parts of the system, many LCAs lead to unexpected and nonintuitive results. For example, bio-based products, such as paper bags, paper cups, and cloth diapers, are not obviously superior in terms of using less energy and materials. Paper requires the harvesting and transportation of trees to pulp mills, activities which require energy. Paper making releases air pollutants and water discharges of chlorine and biological waste. After use, bags end up in landfills where they decay and release methane in the process. The amount of hot water needed to wash and dry cloth diapers is not inconsequential, especially for those who live where water is scarce or sewage is not treated. These kinds of analyses highlight how the environmental impacts of alternative products may lead to unanticipated consequences.

LCA identifies the potential transfer of environmental impacts from one media to another (e.g., eliminating air emissions by creating a wastewater effluent instead) and/or from one life-cycle stage to another (e.g., from use and reuse of the product to the raw material acquisition stage). If an LCA were not performed, the transfer might not be recognized and properly included in the analysis because it is outside of the typical scope or focus of product selection processes. By broadening study boundaries, LCA can help decision makers select the product or process that results in the least impact to the environment. This information can be used with other factors, such as cost and performance data, in the selection process.

The History of LCA

LCA had its beginnings in the 1960s. Concerns over the limitations of raw materials and energy resources sparked interest in finding ways to cumulatively account for energy use and to project future resource supplies and use. In 1969, researchers initiated an internal study for The Coca-Cola Company that laid the foundation for the current methods of life-cycle inventory analysis in the United States. In a comparison of different beverage containers to determine which container had the lowest releases to the environment and was least affected by the supply of natural resources, this study quantified the raw materials and fuels used and

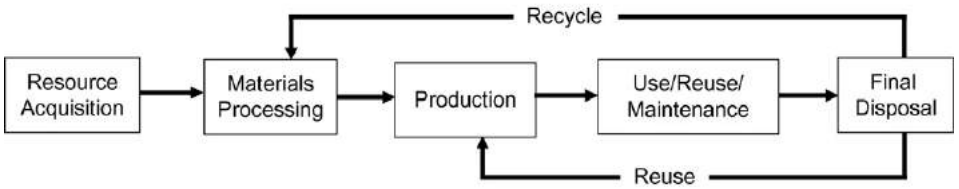


Fig. 1 Generic life-cycle stages (arrows represent transportation).

☆*Change History:* January 2016. MA Curran updated the text to this entire article.

the environmental loadings from the manufacturing processes for each container. Other companies in both the United States and Europe performed similar comparative life-cycle inventory analyses in the early 1970s.

The process of quantifying the resource use and environmental releases of products became known in the United States as a "resource and environmental profile analysis" (REPA) while in Europe it was called an "ecobalance." With the formation of public interest groups encouraging industry to ensure the accuracy of information in the public domain, and spurred on by the oil shortages in the early 1970s, a protocol or standard research methodology for conducting these studies was developed and further evolved.

From 1975 through the early 1980s, as interest in these comprehensive studies waned because of the fading influence of the oil crisis, environmental concerns shifted to issues of hazardous and household waste management. However, throughout this time, REPAs and "ecobalances" continued to be conducted and the methodology improved through a slow steam of about two studies per year, most of which focused on energy requirements. During this time, European interest grew with the establishment of an Environment Directorate (DG X1) by the European Commission. European LCA practitioners developed approaches parallel to those being used in the United States. Besides working to standardize pollution regulations throughout Europe, DG X1 issued the Liquid Food Container Directive in 1985, which charged member companies with monitoring the energy and raw materials consumption and solid waste generation of liquid food containers.

When solid waste became a worldwide issue in 1988, LCA again emerged as a tool for analyzing environmental problems. As interest in all areas affecting resources and the environment grows, the methodology for LCA is again being improved. A broad base of consultants and researchers across the globe has been further refining and expanding the methodology. The need to move beyond the inventory to impact assessment brought LCA methodology to another point of evolution.

Beginning in 1991, concerns over the inappropriate use of LCAs in making broad marketing claims by product manufacturers, along with pressure from other environmental organizations to standardize LCA methodology, led to the development of the LCA standards in the International Standards Organization (ISO) 14000 series (1997–2002, and updated in 2006). In 2002, the United Nations Environment Programme (UNEP) joined forces with the Society of Environmental Toxicology and Chemistry (SETAC) to launch the Life Cycle Initiative, an international partnership. The three programs of the initiative aim to put life-cycle thinking into practice and to improve the supporting tools through better data and indicators. In 2015, the Forum for Sustainability through Life Cycle Innovation (FSLCI) was created for people with common interest in creating and using life cycle information. By bringing together stakeholders, initiatives and activities around the world under one umbrella, the Forum organizers aim to link activities and allow members to be more effective by building on each other's strengths rather than duplicating them.

Conducting an LCA

Specifically, LCA is a technique (Fig. 2) to assess the environmental aspects and potential impacts associated with a product, process, or service, by:

- appropriately selecting a functional unit;
- clearly defining the goal and scope of the study;
- compiling an inventory of relevant energy and material inputs and environmental releases;
- evaluating the potential environmental impacts associated with identified inputs and releases;
- interpreting the results to help decision makers make a more informed decision.

This ability to track and document shifts in environmental impacts can help decision makers and managers fully characterize the environmental tradeoffs associated with product or process alternatives. By performing an LCA, analysts can, for example:

- develop a systematic evaluation of the environmental consequences associated with a given product;
- analyze the environmental tradeoffs associated with one or more specific products/processes to help gain stakeholder (state, community, etc.) acceptance for a planned action;
- quantify environmental releases to air, water, and land in relation to each life-cycle stage and/or major contributing process;
- assist in identifying significant shifts in environmental impacts between life-cycle stages and environmental media;
- assess the human and ecological effects of material consumption and environmental releases to the local community, region, and world;
- compare the health and ecological impacts between two or more rival products/processes or identify the impacts of a specific product or process;
- identify impacts to one or more specific environmental areas of concern.

Comparing Apples to Apples

When an LCA is used to compare two or more products, the basis of comparison should be equivalent use, that is, each system should be defined so that an equal amount of product or equivalent service is delivered to the consumer. For example, if bar soap were compared to liquid soap, the logical basis for comparison would be an equal number of handwashings. Another example of

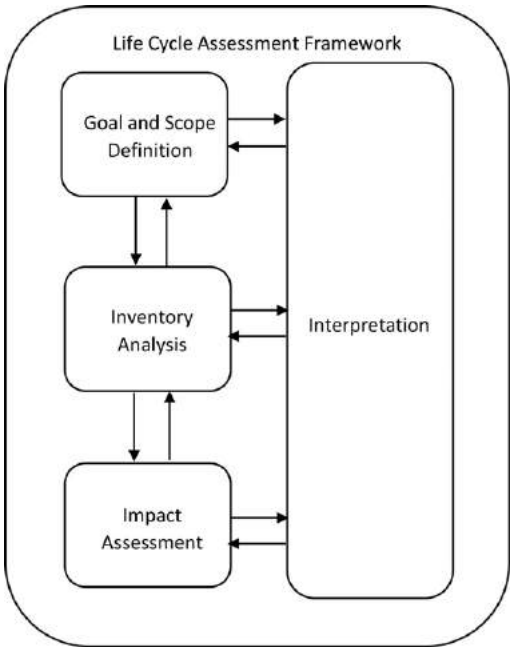


Fig. 2 Phases of an LCA. From International Standards Organization (2006) *Environmental management — LCA — principles and framework.* ISO 14040. Geneva: ISO.

equivalent use would be in comparing cloth diapers to disposable diapers. Cloth diapers need to be changed more frequently, or they are doubled whereas disposables are not. Thus, more cloth diapers will be used. In this case, a logical basis for comparison between the systems would be the number of diapers used over a set period of time.

Equivalent use for comparative studies can often be based on volume or weight, particularly when the study compares packaging for delivery of a specific product. A beverage container study might consider 1000 L of beverage as an equivalent use basis for comparison, because the product may be delivered to the consumer in a variety of container sizes having different life-cycle characteristics.

Life-Cycle Inventory

A life-cycle inventory is a process of quantifying energy and raw material requirements, atmospheric emissions, waterborne emissions, solid wastes, and other releases for the entire life cycle of a product, process, or activity. This step is typically supported by a flow diagram depicting the activities occurring within the system boundaries (**Fig. 3**). An inventory analysis produces a list containing the quantities of pollutants released to the environment (after treatment or control) and the amount of energy and material consumed. The results can be segregated by life-cycle stage, media (air, water, and land), specific processes, or any combination thereof.

In the life-cycle inventory phase of an LCA, all relevant data are collected and organized. Without a life-cycle inventory, no basis exists to evaluate comparative environmental impacts or potential improvements. The level of accuracy and detail of the data collected is reflected throughout the remainder of the LCA process. (No predefined list of data quality goals exists for all LCA projects. The number and nature of data quality goals necessarily depend on the level of accuracy required to inform the decision makers involved in the process.)

Resource constraints for data collection may be a consideration in defining the system, although in no case should the scientific basis of the study be compromised. The level of detail required to create a thorough inventory depends on the size of the system and the purpose of the study. In a large system encompassing several industries, certain details may not be significant contributors given the defined intent of the study. These details may be omitted without affecting the accuracy or application of the results. However, if the study has a very specific focus, such as a manufacturer comparing alternative processes or materials for inks used in packaging, it would be important to include chemicals used in very small amounts.

Life-cycle inventory analyses can be used in various ways. The data can assist an organization in comparing products or processes and considering environmental factors in material selection. In addition, inventory analyses can be used in policy-making, by helping the government develop regulations regarding resource use and environmental emissions.

Life-Cycle Impact Assessment

Although much can be learned about a process by considering the life-cycle inventory data, an impact assessment provides a more meaningful basis to make comparisons. For example, although we know that 9000 t of carbon dioxide (CO_2) and 5000 t of

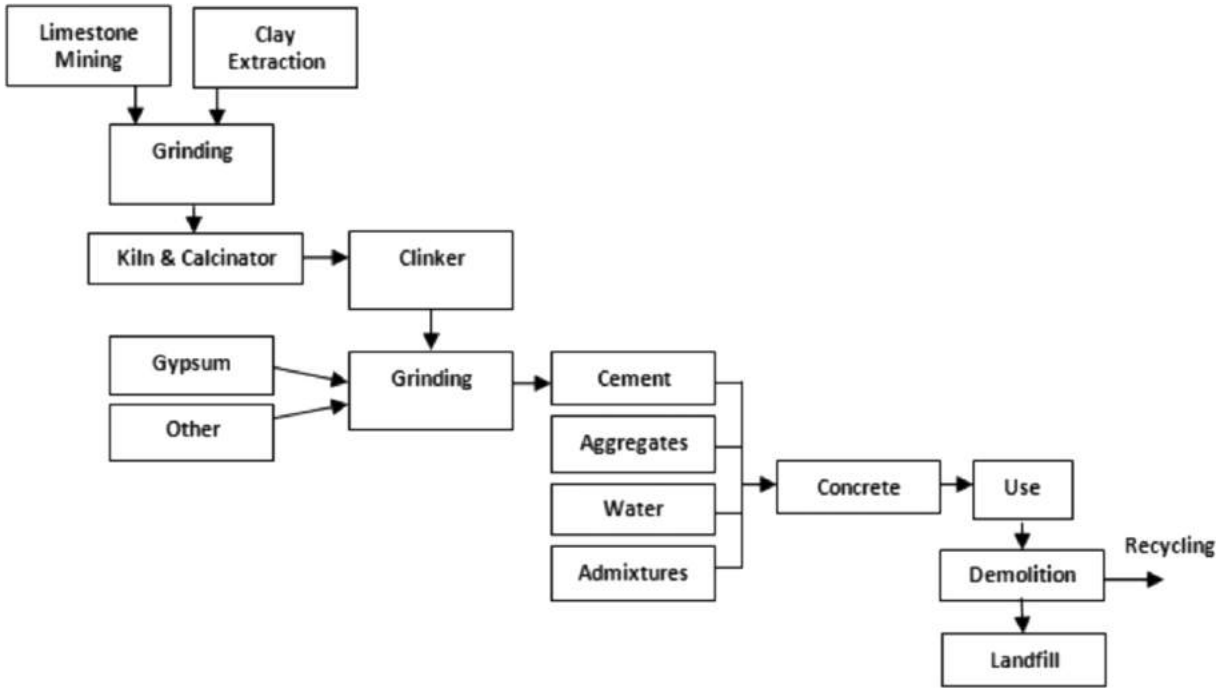


Fig. 3 Example of a flow diagram for a concrete product life cycle.

methane released into the atmosphere are both potentially harmful greenhouse gases, a life-cycle impact assessment (LCIA) can determine which would have a greater impact. Which is worse? What are their potential impacts on global warming? On smog? Using science-based characterization factors, an LCIA can calculate the impacts that each environmental release may have on problems such as smog or global warming.

The LCIA phase of an LCA is the evaluation of potential human health and environmental impacts of the environmental resources and releases identified during the inventory. Impact assessment should address ecological and human health effects; it can also address resource depletion. As shown in Fig. 4, an LCIA attempts to establish a linkage between the product or process and its potential environmental impacts.

Example Calculation of a Characterization Factor

The following calculations demonstrate how characterization factors can be used to estimate the potential contribution toward an impact category. In this example, global warming potential (GWP) is presented in terms of equivalent emissions of carbon dioxide (CO_2) using units of teragrams of carbon dioxide equivalents (Tg CO_2 eq.):

$$\text{of } CO_2 \text{ released} = 9000\,\text{t} = 0.009\,\text{Tg}$$
$$(CO_2 \text{ GWP factor value} = 1)$$
$$\text{Quantity of methane released} = 5000\,\text{t} = 0.005\,\text{Tg}$$
$$(\text{methane GWP factor value} = 23)$$

Therefore,

$$CO_2 \text{ GWP} = 0.009\,\text{Tg} \times 1 = 0.009\,\text{Tg } CO_2\,\text{eq.}$$
$$\text{Methane GWP} = 0.005\,\text{Tg} \times 23 = 0.115\,\text{Tg } CO_2\,\text{eq.}$$

GWP factor values are from the Intergovernmental Panel on Climate Change (IPCC) Model, 100 year time horizon, Third Assessment Report, 2001.

The key to impact characterization is using the appropriate characterization factor. For some impact categories, such as global warming and ozone depletion, there is a consensus on acceptable characterization factors. For other impact categories, such as resource depletion, consensus is still being developed.

An important distinction exists between LCIA and other types of impact analyses. The LCIA does not necessarily attempt to quantify any actual, site-specific impacts associated with a product, process, or activity. Instead, it seeks to establish a linkage between a system and potential impacts. The models used within impact assessment are often derived and simplified versions of more sophisticated models within each of the various impact categories. These simplified models are suitable for relative

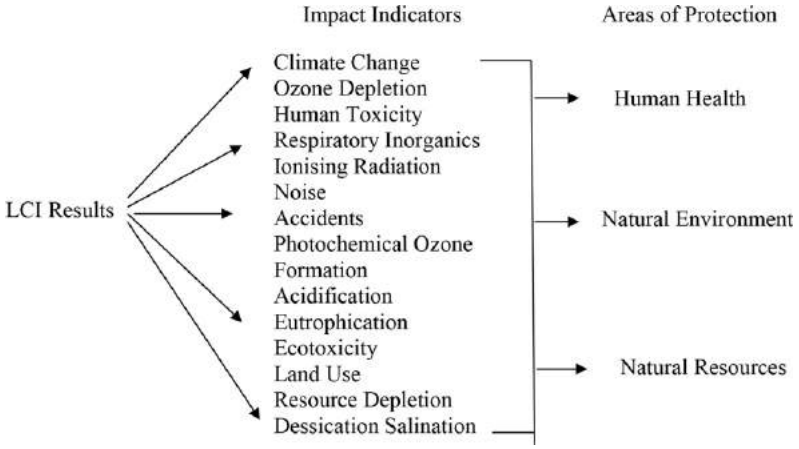


Fig. 4 Relationship between impact indicator categories and areas of protection.

comparisons of the potential to cause human or environmental damage, but are not indicators of absolute risk or actual damage to human health or the environment. For example, risk assessments are often very narrowly focused on a single chemical at a very specific location. In the case of a traditional risk assessment, it is possible to conduct very detailed modeling of the predicted impacts of the chemical on the population exposed and even to predict the probability of the population being impacted by the emission. In the case of LCA, hundreds of chemical emissions (and resource stressors) which are occurring at various locations are evaluated for their potential impacts in multiple impact categories. The sheer number of stressors being evaluated, the variety of locations, and the diversity of impact categories make it impossible to conduct the assessment at the same level of rigor as a traditional risk assessment. Instead, models are based on the accepted models within each of the impact categories using assumptions and default values as necessary. The resulting impact models are suitable for relative comparisons, though insufficient, for absolute predictions of risk.

The key concept in this component is that of stressors. A stressor is a set of conditions that may lead to an impact. For example, if a product or process is emitting greenhouse gases, the increase of greenhouse gases in the atmosphere may contribute to global warming. Processes that result in the discharge of excess nutrients into bodies of water may lead to eutrophication. An LCIA provides a systematic procedure for classifying and characterizing these types of environmental effects.

Midpoint versus Endpoint Impact Modeling

In order to simplify the process and make LCIA more broadly applicable, modeling is typically conducted at the midpoint level instead of modeling effects to specific endpoints (see the ozone depletion potential example below). Midpoint impact assessment models reflect the relative potency of the stressors at a common midpoint within the cause–effect chain. Analysis at a midpoint minimizes the amount of forecasting and effect modeling incorporated into the LCIA, thereby reducing the complexity of the modeling and often simplifying communication. Midpoint modeling can minimize assumptions and value choices, reflect a higher level of societal consensus, and provide more comprehensive than model coverage for endpoint estimation (**Fig. 5**; **Tables 1** and **2**).

The ILCD Handbook provides an extensive analysis of the existing characterization methods and recommendations for LCIA in the European context using reference year 2008. Since then methodological developments have continued, resulting in further advances of LCIA and others yet to come.

For consistency reasons, the choice of impact categories is often made on the basis of a recommended impact assessment guidebook or its implementation in software. Thus, in practice one often sees LCA studies reporting impacts according to the selected model. The most often used LCIA models include EPS 2000; IMPACT World+; LIME: Life-cycle Impact assessment Method for Endpoint modeling; ReCiPe; and TRACI: The Tool for the Reduction and Assessment of Chemical and other environmental Impacts. All these methods comprise a recommended set of impact categories with a category indicator and set of characterization factors. ISO does not specify any choice in these matters.

Comparing Alternatives Using Life-Cycle Interpretation

Life-cycle interpretation, the last phase of the LCA process, is a systematic technique to identify, quantify, check, and evaluate information from the results of the LCI and the LCIA, and communicate them effectively. However, interpreting the results of an LCA is not as simple as two is better than three, therefore, alternative A is the better choice. While conducting the life-cycle inventory and impact assessment, it is necessary to make assumptions, engineering estimates, and decisions based on one's values

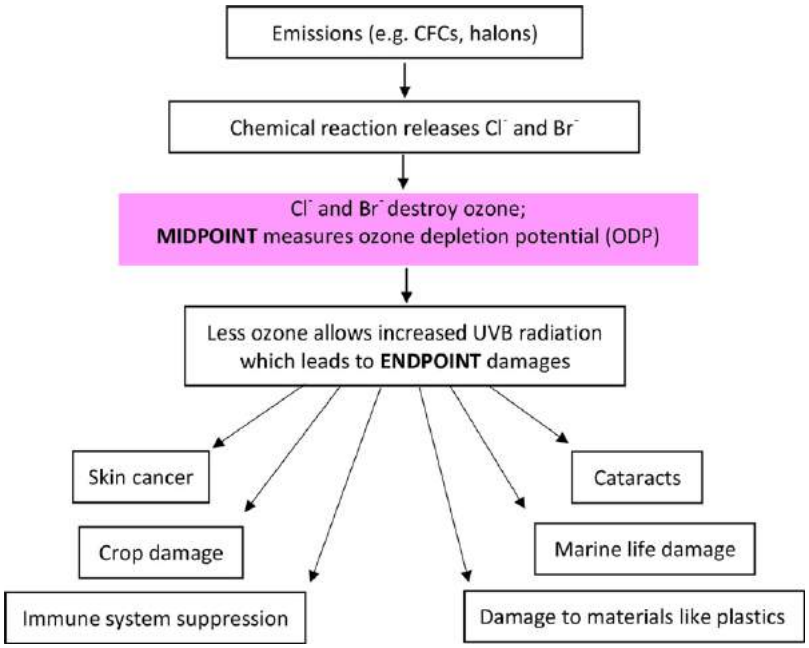


Fig. 5 Midpoint versus endpoint impact modeling.

Table 1 Best available characterization models to midpoint. Models classified as level I, II or III (ILCD Handbook)

Impact category	*Best among existing characterization models*	*Indicator*	*Classification*
Climate change	Baseline model of 100 years of the IPCC	Radiative forcing as Global warming potential (GWP100)	I
Ozone depletion	Steady-state ODPs from the WMO assessment	Ozone depletion potential (ODP)	I
Human toxicity, cancer effects	USEtox model	Comparative toxic unit for humans (CTU_h)	II/III
Human toxicity, non-cancer effects	USEtox model	Comparative toxic unit for humans (CTU_h)	II/III
Particulate matter/respiratory inorganics	RiskPoll model	Intake fraction for fine particles (kg PM2.5-eq/kg)	I/II
Ionising radiation, human health	Human health effect model	Human exposure efficiency relative to U^{235}	II
Ionising radiation, ecosystems	Screening level ecological risk assessment	Comparative toxic unit for ecosystems (CTU_e)	Interim
Photochemical ozone formation	LOTOS-EUROS as applied in ReCiPe	Tropospheric ozone concentration increase	II
Acidification	Accumulated exceedance	Accumulated exceedance (AE)	II
Eutrophication, terrestrial	Accumulated exceedance	Accumulated exceedance (AE)	II
Eutrophication, aquatic	EUTREND model as implemented in ReCiPe	Residence time of nutrients in freshwater (P) or marine end compartment (N)	II
Ecotoxicity, freshwater	USEtox model	Comparative toxic unit for ecosystems (CTU_e)	II/III
Land use	Model based on soil organic matter	Soil organic matter (SOM)	III
Resource depletion, water	Model for water consumption as in the Swiss Ecoscarcity	Water use related to local scarcity of water	II
Resource depletion, mineral and fossil	CML 2002	Scarcity	II

I — Recommended and satisfactory; II — Recommended but in need of some improvement; III — Recommended, but to be applied with caution.
A mixed classification is related to the application of the classified method to different types of substances.

and the values of involved stakeholders. Each of these decisions must be included and communicated within the final results to clearly and comprehensively explain conclusions drawn from the data. In some cases, it may not be possible to state that one alternative is better than another because of the uncertainty in the final results. This does not imply that efforts have been wasted. The LCA process will still provide decision makers with a better understanding of the environmental and health impacts associated

Table 2 Best available characterization models from midpoint to endpoint (ILCD Handbook)

Impact category	*Best among existing characterization models*	*Indicator*	*Classification*
Climate change	Model developed for ReCiPe	Disability adjusted life years (DALY) for human health Potentially disappeared fraction of species ($PDFm^3yr$) for ecosystem health	Interim
Ozone depletion	Model for human health damage developed for ReCiPe	Disability adjusted life years (DALY)	Interim
Human toxicity, cancer effects	DALY calculation applied to USEtox midpoint	Disability adjusted life years (DALY)	II/interim
Human toxicity, non-cancer effects	DALY calculation applied to USEtox midpoint	Disability adjusted life years (DALY)	Interim
Particulate matter/respiratory inorganics	Adapted DALY calculation applied to midpoint	Disability adjusted life years (DALY)	I/II
Ionising radiation, human health		Disability adjusted life years (DALY)	Interim
Ionising radiation, ecosystems			
Photochemical ozone formation	Model for damage to human health developed for ReCiPe	Disability adjusted life years (DALY)	II
Acidification	Method developed for ReCiPe	Potentially disappeared fraction of plant species	Interim
Eutrophication, terrestrial	No methods identified		
Eutrophication, aquatic	Model for damage to ecosystem (freshwater only)	Potentially disappeared fraction of species ($PDFm^3yr$)	Interim
Ecotoxicity			
Land use	Model for species diversity loss as in ReCiPe	Potentially disappeared fraction of species ($PDFm^3yr$)	Interim
Resource depletion, water			
Resource depletion, mineral and fossil	Method developed for ReCiPe	Surplus costs	Interim

Only three models recommended by the ILCD Handbook are classified above interim status.

with each alternative, where they occur (locally, regionally, or globally), and the relative magnitude of each type of impact in comparison to each of the proposed alternatives included in the study. This information more fully reveals the pros and cons of each alternative.

The purpose of conducting an LCA is to better inform decision makers, who could be government officials, multinational corporations, nongovernmental entities (NGOs), or, ideally, multi-stakeholder panels, by providing a particular type of information (often unconsidered) with a life-cycle perspective of environmental and human health impacts associated with each product or process. With LCAs moving us beyond examining a single impact or life cycle stage as the sole criteria for environmental goodness (or badness), much knowledge has been gained through the thousands of LCA studies that have been completed over the last 20 years. The movement now is toward life cycle sustainability assessment (LCSA) which integrates LCA, life cycle costing (LCC) and social life cycle assessment (SLCA). The development of the economic pillar is advancing through the efforts of various groups while the societal assessment is still in its infancy. Ultimately, all three complementary pillars of sustainability should be based on the same system boundaries and functional unit within a study.

Limitations of Conducting an LCA

Performing an LCA can be very resource and time intensive. Depending upon how thorough the user wishes to be, gathering the data can be problematic, and the availability of data can greatly impact the accuracy of the final results. Therefore, it is important to weigh the availability of data, the time needed to conduct the study, and the financial resources required against the projected benefits of the LCA.

LCA will not determine which product or process is the most cost effective or works the best. Therefore, the information developed in an LCA study should be used as one component of a more comprehensive decision process assessing the tradeoffs with cost and performance.

There are a number of ways to conduct LCIA. While the methods are typically science based, the complexity of environmental systems has led to the development of alternative impact models.

The role of impact assessment is to categorize and quantify potential environmental effects. Once this is done, deciding whether one impact is worse than another is necessarily a subjective process in which the perceptions of the decision maker are applied.

While LCA can help identify potential environmental tradeoffs, converting the impact results to a single score requires the use of value judgments, which must be applied by the commissioner of the study or the modeler. This can be done in different ways such as through the use of an expert panel, but it cannot be done based solely on natural science.

All assumptions or decisions made throughout the entire project must be reported alongside the final results of the LCA project. If assumptions are omitted, the final results may be taken out of context or easily misinterpreted. As the LCA process advances from phase to phase, additional assumptions and limitations to the scope may be necessary to accomplish the project with the available resources.

Summary

Adding LCA to the decision-making process provides an understanding of the human health and environmental impacts that traditionally are not considered when selecting a product or process. This valuable information provides a way to account for the full impacts of decisions, especially those that occur outside of the site that are directly influenced by the selection of a product or process. LCA is an environmental management tool that helps to inform decision makers and should be included with other decision criteria, such as cost and performance, in order to make a well-balanced decision. While there is not always a straightforward or easy choice, it is important to understand the potential impacts related to each choice.

See also: Aquatic Ecology: Acidification in Aquatic Systems; Microbial Communities. Ecological Processes: Acidification. Human Ecology and Sustainability: Ozone Layer; Industrial Ecology; Carbon Footprint; Nitrogen Footprints

Further Reading

Curran, M.A. (Ed.), 2012. Life cycle assessment handbook: A guide for environmentally sustainable products. Salem, MA: Scrivener-Wiley Publishing978-1118099728, p. 625.
European Commission Joint Research Center (JRC), 2010. ILCD Handbook — international reference life cycle data system: A general guide for life cycle assessment. Luxembourg: EUR, 24708 EN.
Fava, J., Denison, R., Jones, B., *et al.* (Eds.), 1990. A technical framework for life cycle assessments. Pensacola, FL: Society of Environmental Toxicology and Chemistry (SETAC), p. 152.
Hendrickson, C., Lave, L., Matthews, H.S., 2006. Environmental LCA of goods and services: An input-output approach. Washington, DC: Resources for the Future1-933115-23-8.
International Standards Organization, 2006. Environmental management — LCA — principles and framework. Geneva: ISO. ISO 14040.
Klöpffer, W., Curran, M.A. (Eds.), 2014. LCA compendium: the complete world of life cycle assessment. Dordrecht: Springer. ISSN: 2214-3505.
United Nations Environment Programme, 2004. Why take a life cycle approach? New York: UNEP, ISBN 92-807-24500-9, 25 pp.

Limits to Growth

C Jaeger, Potsdam Institute for Climate Impact Research, Potsdam, Germany

The Basic Idea

The concept of limits to growth is an attempt to understand the historical tension between global economic growth and environmental protection.

An important example of this tension is the danger of extinction for whales. For centuries humans have hunted whales, a difficult and dangerous enterprise whose mythical qualities have been highlighted by Herman Melville's tale of Moby Dick. With the expansion of the modern world economy, the catch was greatly increased as more and more ships were put to use with increasingly sophisticated technology. This, however, reduced the total population of whales, and as a certain minimum of whales is needed to secure mating and reproduction, extinction of whales has become a serious risk. It took strong and continued political action by environmentalists all over the world, supported by scientists and enhanced by mass media, to enforce limits on whaling. The tension between economic interests in increased catch and environmental interests in protecting the whales, however, is still there, and it applies to many other species, too.

This kind of situation can be modeled mathematically and the resulting models can be implemented on computers. In 1972, Meadows *et al*. published a highly influential book where a similar model was used to argue that economic growth should be stopped within a few decades in order to avoid environmental catastrophe.

Three Simple Models

A simple model of unlimited growth can be designed as follows. Let $r>0$ be the rate of growth, IR_+ the set of non-negative real numbers, and f the function

$$\begin{aligned} f:\quad & IR_+ \to IR_+ \\ & x \mapsto (1+r)\cdot x \end{aligned} \qquad [1]$$

For any given x', one can iterate f so as to get, for example, $f(f(f(x')))$ or $f(f(f(f(x'))))$. Writing n for the number of iterations, one can define a function $\phi_{x'}$:

$$\begin{aligned} \phi_{x'}:\quad & IN \to IR_+ \\ & n \mapsto (1+r)^n \cdot x' \end{aligned} \qquad [2]$$

This can be written as

$$\phi_{x'}(n) = x' e^{n \log(1+r)} \qquad [3]$$

As $\log 1 = 0$ and $(\mathrm{d}\log x/\mathrm{d}x)\big|_{x=1} = 1$ we have $\phi_{x'}(n) \sim x' e^{nr}$.

The exponential function provides a fundamental model of unlimited growth. It is worth noticing that the Euler number e and the exponential function itself were introduced into mathematical physics about three centuries ago on the basis of analyses of a key phenomenon of economic growth: the mechanism of compound interest. Later on, Malthus (1798) noticed that in many circumstances the exponential function was a good model for the dynamics of biological populations, including humans. Malthus also insisted that population growth could not go on without limits and therefore expected serious famines that would ultimately stabilize human population worldwide.

In 1838, the Belgian mathematician Verhulst modified Euler's exponential function so as to represent limits to growth. Call that limit K and let the variable rate of growth r' obey the formula

$$r' = r\left(1 - \frac{x}{K}\right), \quad r>0, \quad K>0 \qquad [4]$$

This implies that the rate of growth approaches 0 as the growing variable approaches its limit, while it approaches r as the growing variable approaches zero. Instead of [1] we get

$$x \mapsto \left(1 + r\left(1 - \frac{x}{K}\right)\right)\cdot x \qquad [5]$$

By introducing an auxiliary variable z, [5] can be simplified as follows:

$$\begin{aligned} & z = \frac{x(1+r)K}{r} \\ & z \to (1+r)\cdot z \cdot (1-z) \end{aligned} \qquad [6]$$

As long as r is smaller than 2, the variable growth rate r' gradually diminishes as x increases. The result is an S-shaped curve that

converges from below to the level $x=K$. In 1974, R. M. May discovered that for r above a threshold slightly larger than 3.57 the process generates a chaotic dynamics: it jumps to levels above K and from there falls back to levels below K, and it does so without repeating itself in a regular fashion. For values between 2 and about 3.57, the process generates various forms of oscillations. These findings make the argument about limits to growth more intricate, but the key idea can be developed with the simpler S-shaped curve.

Now suppose that some agent (e.g., the fisheries industry) is faced with a population whose dynamics can be described by [6] with $1<r<2$. The agent wants to maximize its utility from catching fish over an indefinite period. Make the usual assumption that the utility function is additive across time, concave and monotonically increasing in each period, discounting future utility by a discount factor $0<\beta<1$:

$$\begin{aligned} \max_{h\in H} \quad & \sum_{t=0}^{\infty} \beta^t \cdot u(y_t) \\ s.t: \quad & z_{t+1} = (1+r)\cdot z_t \cdot (1-z_t) - y_t \\ & y_t = h(z_t, t) \end{aligned} \qquad [7]$$

If the agent is sufficiently greedy, that is, if β is rather small, he will try to catch as much fish as he can and as fast as he can. He will then choose his strategy h accordingly. Suppose that it takes time to build up fishing capabilities. Then the strategy space H will be such that increasing catch becomes feasible in the course of time. This, however, will lead the agent to fish until the population is exhausted according to a pattern known as overshoot and collapse. The upper panel in **Fig. 1** gives an illustration.

If the agent cares more about the distant future, that is, if β is close enough to 1, a different strategy will be appropriate: build up fishery capacity up to the point where a constant catch is possible so as to maintain the population indefinitely. The lower panel in **Fig. 1** gives an illustration.

A Global Debate

J. W. Forrester, an outstanding electrical engineer and management scientist at MIT, proposed to advance this kind of analysis with a great leap forward. His model did not represent the management of a specific species like whales, but rather the interaction between humankind and planet Earth. It had not one, but two limiting factors: the capability of the Earth to provide resources like food, timber, metals, etc., and the capability of the Earth to absorb pollutants like toxic waste, smog generating aerosols, etc. The goal variable was quality of life, and humans were assumed to try to increase their quality of life by increasing their number as well as the size of their capital equipment. The result was a grim scenario of overshoot and collapse within a few decades.

Forrester's model was presented to a world audience by Meadows *et al.* in 1972. The model was not performing any optimization exercise; it was comparing different strategies, showing that many intuitively plausible strategies – such as increasing capital accumulation to abate pollution – could not avoid the pattern of overshoot and collapse. What was required, these authors argued, was an end to population growth as well as economic growth.

With regard to population growth, most authors agreed and still agree that it would be highly problematic if humankind would keep growing toward 10 billion people in the foreseeable future. However, there is also a widespread agreement that global population growth is declining and will come to an end without overshoot and collapse. In contrast to other biological species, with humankind, population growth is slowed down by increased well-being.

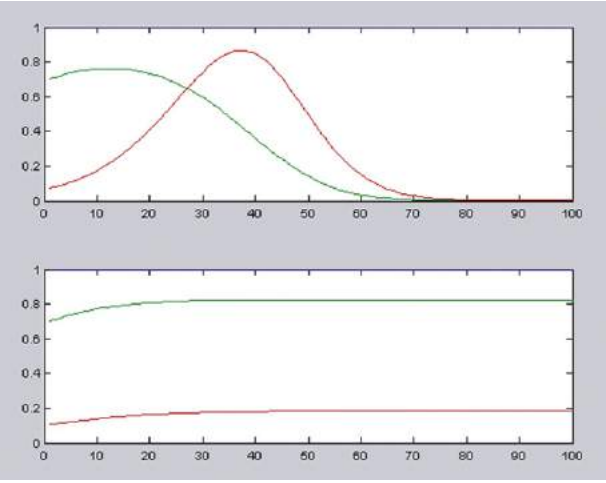


Fig. 1 Overshoot and collapse vs. sustainable management. Green: limited resource; red: resource use (resource use is scaled up for better visibility).

With regard to the economy, however, the claim of limits to growth immediately triggered a fierce debate. Robert Solow, also an economist at MIT (and Nobel Prize winner in 1987), wrote: ``I would like to state, briefly and bluntly, why I think the various `doomsday models' are worthless as science and as guide to public policy'' (Solow, 1972: 3832). His main point was that that the price mechanism would enable humankind to avoid overshoot in the face of limited resources by triggering substitution and innovation. During the following decades, economists developed models representing these mechanisms. A key claim of this kind of literature is that resource scarcity is not a big problem because it is precisely what markets can deal with, and that pollution increases with economic growth only in the early stages of economic growth. Once people get more affluent, they are supposed to be able and willing to abate pollution by technological means.

The debate took a new twist with the publication of the report *Our Common Future*. This report launched the concept of sustainable development, claiming that it was necessary and possible to develop new patterns of economic growth that would balance ecological, economic, and social needs. While the report steered clear of mathematical models and detailed empirical studies, it displaced the image of limits to growth in favor of the vague but attractive concept of sustainable development.

The debate on limits to growth is far from over, however. On one hand, the combination of environmental and political challenges raised by the globalizing economy of our time leads some authors to think again about the prospects of life beyond economic growth. On the other hand, there is a long-standing tradition of economists seeing economic growth as a transitory phase in the history of humankind – Ricardo, J. S. Mill, Marshall, Pigou, Keynes, Hicks all shared this view.

But while Ricardo emphasized the problem of limited resources already nearly two centuries ago, the main point of this tradition is a different one. In the famous words of John Stuart Mill, ``I confess I am not charmed with the ideal of life held out by those who think that the normal state of human beings is that of struggling to get on; that the trampling, crushing, elbowing, and treading on each other's heels, which form the existing type of social life, are the most desirable lot of human kind, or anything but the disagreeable symptoms of one of the phases of industrial progress'' (Mill, 1848: Bk iv, ch. vi, §iv.6.5).

See also: Human Ecology and Sustainability: Emergy and Sustainability; Tragedy of the Ecological Commons; Ecological Footprint; System Sustainability

Further Reading

Our common future. In: Bruntland, G. (Ed.), The World Commission on Environment and Development. Oxford: Oxford University Press.
Cohen, J., 1995. How Many People Can the Earth Support? New York: WW Norton.
Euler L (1735) *Mechanica sive motus scientia analytice exposita*. http://math.dartmouth.edu/~euler/pages/E015.html (accessed November 2007).
Forrester, J.W., 1971. Counterintuitive behavior of social systems. Technology Review 73, 52–68.
Hicks, J.R., 1966. Growth and anti-growth. Oxford Economic Papers 18, 257–269.
Homer-Dixon, T., 2006. The Upside of Down: Catastrophe, Creativity and the Renewal of Civilization. Toronto: Knopf Canada.
Economic possibilities for our grandchildren. In: Keynes, J.M. (Ed.), The Collected Writings of John M. Keynes, vol. IX. London: Cambridge University Press.(first pub. 1930).
Malthus, T.R., 2003. An Essay on the Principle of Population. Teddington: The Echo Library, (first pub. 1798).
May, R.M., 1974. Biological populations with nonoverlapping generations: Stable points, stable cycles, and chaos. Science 186, 645–647.
Meadows, D.H., Meadows, D.L., Randers, J., Behrens III, W.W., 1972. The Limits to Growth. New York: Universe Books.
Mill, J.S., 1848/1965. Principles of Political Economy. Indianapolis, IN: Liberty Fund.
Solow, R.M., 1972. Notes on `doomsday models'. Proceedings of the National Academy of Sciences of the United States of America 69, 3832–3833.
Stokey, N.L., 1998. Are there limits to growth? International Economic Review 39, 1–31.
Verhulst, P.-F., 1838. Notice sur la loi que la population pursuit dans son accroissement. Correspondance MathÅmatique et Physique 10, 113–121.

Nitrogen Footprints

Adrian Leip, Joint Research Centre, Ispra, Italy
Aimable Uwizeye, Wageningen University & Research, Wageningen, The Netherlands; Food and Agriculture Organization of the United Nations, Rome, Italy; and Teagasc—Crops, Environment and Land Use Programme, Johnstown Castle, Wexford, Ireland

Introduction

General

Anthropogenic activities depend on nitrogen (N) to produce food and industrial products. This production causes the release of nitrogen into the environment in different forms (ammonia (NH_3), nitrous oxide (N_2O), nitrates (NO_3^-), organic nitrogen or di'nitrogen (N_2).

Nitrogen Footprints quantify the pressure of humans on the environment. They are expressed per unit of product or are aggregated at consumer level and expressed per capita. N footprints demonstrate the pressure from humans or production processes related to N losses into the environment. A precise definition for N footprints does not yet exist, and therefore the kind of pressure that is quantified varies between studies. A commonly used definition is that N footprint quantify N losses along the life cycle of a product, but it has also been used to quantify N mobilization by anthropogenic activities. There are some common characteristics of all N footprints values published: they are based on life cycle thinking, they are associated either with products—more often than not food products—or "consumers," and they describe a pressure on environmental systems or human wellbeing. Generally, footprints are understood as a tool with the potential to help on the road to more sustainability (Hoekstra and Wiedmann, 2014).

As such, the N footprint position itself clearly in the domain of environmental footprints reporting an inventory of N flows that are directly or indirectly of environmental relevance. N footprints do not attempt to quantify the environmental impact itself—thus they are not converted into a measure of increasing temperature with global warming potentials, or amount of coastal dead zones taking into consideration relative abundances of phosphorus and silica. Environmental impacts associated with N pollution are highly variable both in time and space and depend on numerous factors. N footprints thus usually serve as a proxy for such effects, but do not claim to quantify those effects. Rather, the main objective of N footprints is to give a robust measure for higher or lower possibility that a certain product or consumption contributes to N losses in a way that can be understood by peer scientists, stakeholders, decision makers, policy makers, and general public.

The idea of an N Footprint was developed by the Chesapeake Bay Foundation to raise awareness of people's impact on water pollution (https://secure.cbf.org/site/SPageNavigator/bay_footprint.html). The concept reached a wide audience with the paper by Leach *et al.* (2012) "A N footprint model to help consumers understand their role in N losses to the environment." Again, the main objective was to raise awareness amongst consumers that with their consumption choices they contribute to adverse effects related to N pollution.

This article reviews available N footprint concepts and provide recent results.

The Nitrogen Cycle

There is abundance of N stored in the atmosphere, oceans, and soils; however, most of it is not readily available for living organisms. The reason is that N is buried in sediments or rocks, or present as diN (N_2) which has a very strong triple bond ($N \equiv N$) that is difficult to break. Yet, N is an essential element for life on earth; it is a component of essential biomolecules, such as amino acids that build proteins, ATP, nucleic acids (e.g., DNA, RNA) and much of the N in plants is used in chlorophyll molecules. In nature, only specialized organisms are able to convert N_2 to N compound (NH_4^+) that can be used by organism. This is called "N fixation," which transforms "inert N" into "reactive N." The rhizobia bacteria and nodules associated with legumes roots perform N_2 fixation naturally from the atmosphere. The efficiency of legumes in fixing N_2 varies between species. At the industry level, N_2-fixation is achieved through the Haber–Bosch process which is energy intensive (Erisman *et al.*, 2008).

Reactive N is the term used for all N compounds other than N_2, as they are readily transformed from one form to another (Galloway *et al.*, 2004). Main reactive N groups distinguished are "oxidized" mineral N compounds such as NO_x (NO and NO_2), nitrate (NO_3^-), nitrite (NO_2^-), and nitrous oxide (N_2O), "reduced" mineral N compounds such as ammonia (NH_3), ammonium (NH_4^+), and organic N (Butterbach-Bahl *et al.*, 2011). Organic N exists in living organisms, humus or in the intermediate products of organic matter decomposition. N passes between the biotic environment (living organisms) and the abiotic environment (soil, water, air) via diverse processes, which altogether form the N cycle as illustrated in **Fig. 1**. Main transformation processes that are relevant for the N cycle are nitrification, converting ammonium to nitrate; denitrification, being a chain of reactions converting stepwise nitrate back to N_2; assimilation, incorporating mineral N into organic molecules; and mineralization releasing mineral N from organic matter.

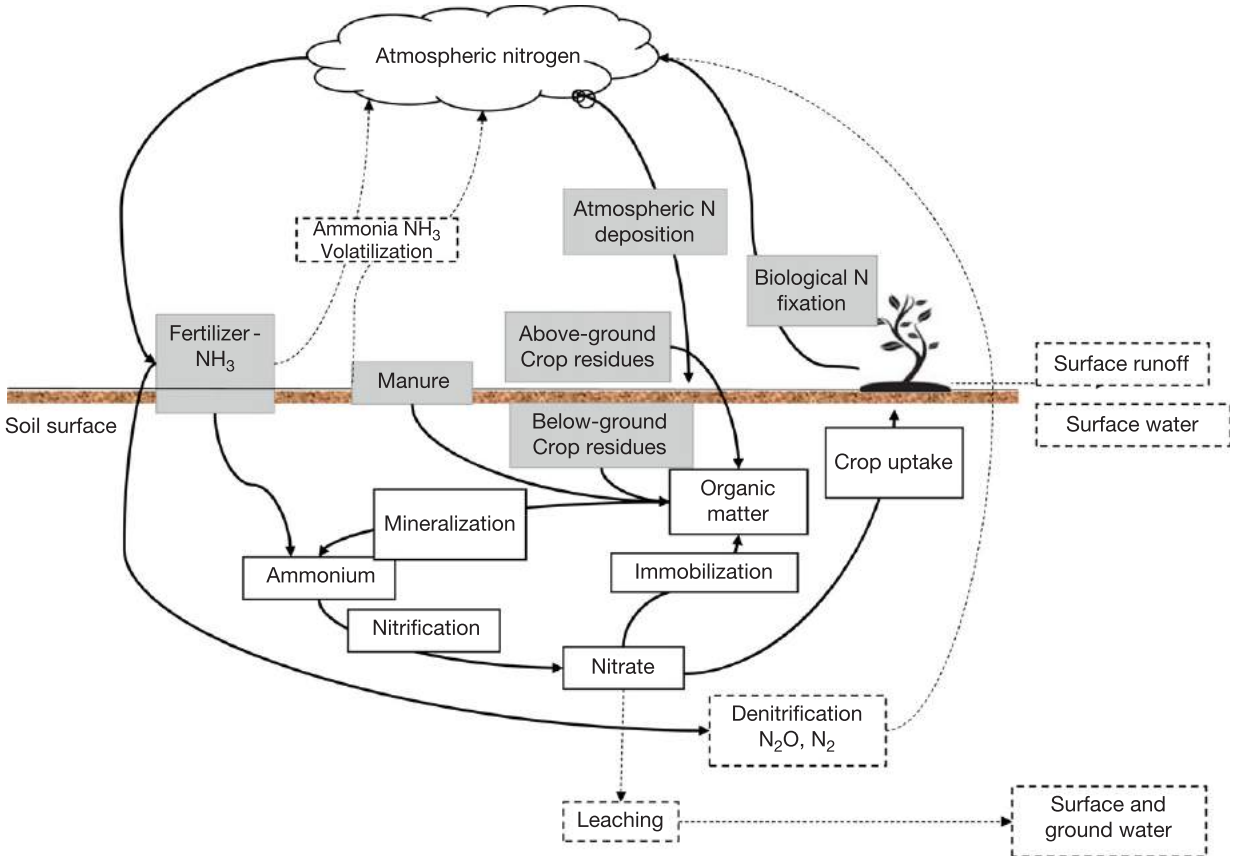


Fig. 1 The main pathways of the N cycle. Adapted from NRC, 1993. Soil and water quality: An agenda for agriculture. Washington, DC: National Academy Press.

Denitrification returns reactive N back into a stable form and thus closes the cycle. Other "sinks" for reactive N are accumulation in soils and sediments which "lock" reactive N making it effectively unavailable for further transformation processes at long time horizon.

Nitrogen Budgets

Before human intervention, about 200 Tg N were "fixed" annually, with the largest contribution from marine biological N fixation (140 Tg N $year^{-1}$), followed by terrestrial biological N fixation (58 Tg N $year^{-1}$) and a small contribution from N fixation by lightning (5 Tg N $year^{-1}$) (Fowler *et al.*, 2013). In preindustrial times, N was a scarce resource and limiting growth and agricultural production. Sources of N were cultivation of leguminous crops which are associated with N-fixing bacteria in rotation with nonfixing crops, transfer of N in manure from grassland with leguminous grasses, exploitation (and N depletion) of N-rich fertile soils, guano or mineral N deposits (Galloway and Cowling, 2002).

Only with the invention of the Haber–Bosch process early of the 20th century, it became possible to fix synthetically atmospheric N_2 into ammonia by a reaction under high temperature and pressure. Synthetic fertilizer creates now about 120 Tg N $year^{-1}$ new reactive N. Together with agricultural biological N fixation of 60 Tg N $year^{-1}$ it contributed to boost agricultural production and helped sustain the growing global population. Altogether, including ca. 30 Tg N $year^{-1}$ created in combustion processes, anthropogenic N fixation is at the same magnitude than natural N fixation (Galloway and Cowling, 2002). Recent estimates by Oita *et al.* (2016) suggested that 189 Tg N were realized worldwide in 2010, with 162 Tg from industries and agriculture through different loss pathways such as leaching, runoff and atmospheric emissions.

Area of Concern: The Environmental Impacts of Reactive Nitrogen

The increased release of reactive N compounds into the environment already surpassed the planetary boundary, with 150 Tg N losses annually (Steffen *et al.*, 2015). These N compounds are increasing the global environmental threats, through a complex net of interactions. The release of organic N and NO_3^- to surface water and groundwater contaminates the aquifers and freshwater resulting in impure drinking water. Moreover, in lakes and marine ecosystems. N pollution causes dead zones, hypoxia, fish kills

due to algal blooms, thus eutrophication that threatens the biodiversity. The air emissions of N_2O, which is a powerful greenhouse gas contributes to climate change, but N also contributes to global cooling with the formation of particles. N_2O is also contributing to stratospheric ozone depletion. Moreover, the emissions of NH_3 and NO_x degrade the air quality with direct impact to the human health. Furthermore, the deposition of N compound from atmosphere can also enrich the terrestrial ecosystem causing the disturbance of species, invasion of alien species causing impacts on biodiversity. Ammonia deposition can acidify natural and agricultural soils leading to soil degradation, fertility loss and erosion (Sutton *et al.*, 2011a,b).

Footprint Methods—Life Cycle Thinking

Production and Consumption Footprints

Human production and consumption processes are virtually always the result of a complex net of interactions and relationships. Basically, no product exists without the need of certain "inputs" that were consumed or appropriated in the production process. The smallest unit of which a footprint can be calculated is thus a specific product, such as an apple from a specific farm or 1 kWh of electricity from the local power plant. The scope of assessment can be broadened or generalized in two ways.

A scheme of the relationship between different footprints and the possibilities of increasing the scale of assessment is shown in Fig. 2. First option for aggregation of the footprint of a unit process is by production system, assessing the footprint of apples from a group of farms, differentiating, for example, by organic or conventional systems, or generalizing from apples to fruit or vegetable food products, usually measured in kg N (kg product)$^{-1}$. A second possibility of increasing the scale of assessment is by extending the boundaries from farm to processor, retailer. At this level, the N footprint assumes full responsibility of all actors in production systems, including farmers, businesses and industry sectors (Lenzen *et al.*, 2007), thus informing policy makers on the magnitude of the production activities affecting the natural resources and environment. Finally, the boundary are extended to the consumer, including or excluding the end-of-life phase of a product. Often, the two processes go in parallel yielding consumption footprints of product or product groups. Consumption footprints are often expressed on a per capita basis (kg N (capita year)$^{-1}$) or a (also temporary) community (then measured in kg N if boundaries are explicit), such as per capita footprint of canteen lunch. This requires the aggregation of individual consumption footprints of food (groups) taking into consideration consumption pattern and—in the case of processed foods—the footprints of the processed ingredients. Here, the assumption is that the demand of products is causing the N pollution, thus allocating full responsibility to the consumers (Lenzen *et al.*, 2007). Another approach that is not used so far for N footprint is to express the N footprint in function of incomes (Marques *et al.*, 2012, 2013).

Either production footprints or consumption footprints can be up-scaled increasing the regional scope from regional over country, continental and finally global scale. At all levels below the global scale, production and consumption footprints differ as long as products and/or inputs in the production process are traded across the geographical boundaries. N flows embedded in imported products that are consumed are included in consumption footprints, but not in production footprints unless they are

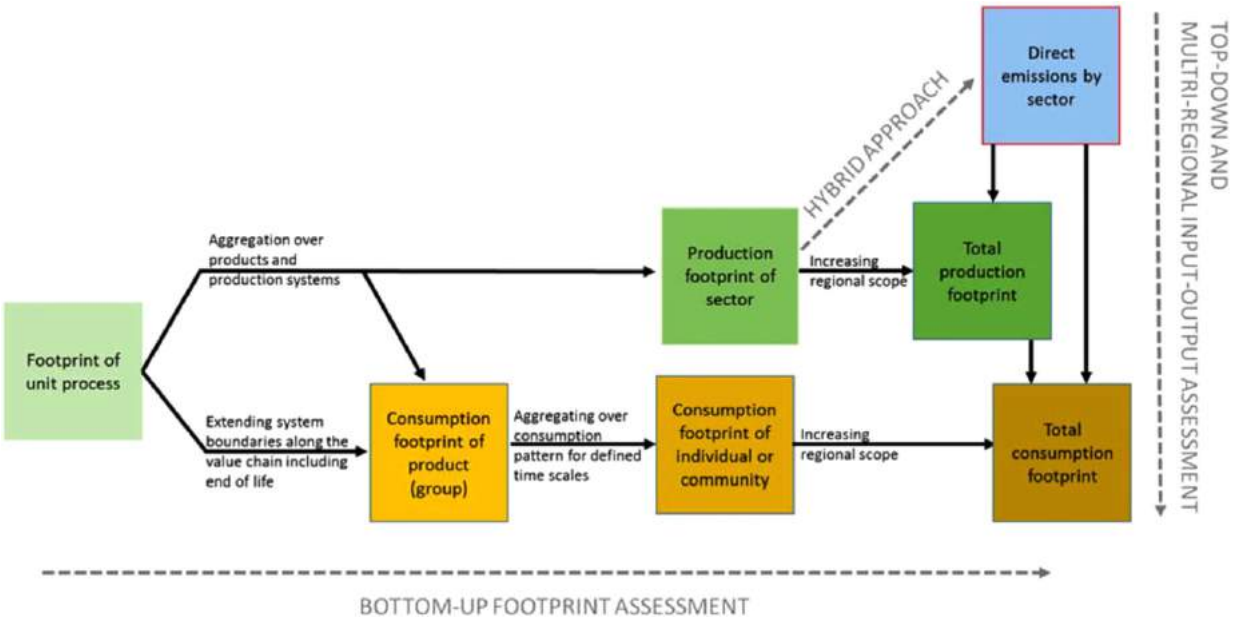


Fig. 2 Production and consumption footprints at different level of aggregation. Footprint estimates starting from unit processes are following a bottom-up approach. Footprint estimates based on total emission estimates per sector are following a top-down approach, often using multiregional input–output tables. The hybrid approach (Bruckner *et al.*, 2015; Cui *et al.*, 2016) combines both methods.

used in the production process. On the other hand, N flows embedded in exported products are considered in the production footprints but not in the consumption footprints (Galloway *et al.*, 2007; Oita *et al.*, 2016).

The most common applications of the N footprint concept has been done at the level of food products or food product groups, and at the level of annual per capita consumption of food and energy (Leach *et al.*, 2012). Per capita calculations take into considerations the choices made by an individual or by a "representative" individual of, for example, a country and typical or average product-level footprints of the products consumed. In some cases, footprints are also quantified for groups of persons for a limited time-period, for example, a group dinner, a conference or similar. These are discussed below in the section of N footprint derivatives. Multiregional input–output analyses "invert" the flow of monetary value to follow flow of environmental impact such as N losses embedded in products traded and highlighting losses distant from the place of production or consumption (Bruckner *et al.*, 2015; Kastner *et al.*, 2011; Oita *et al.*, 2016).

Life Cycle Assessment (LCA)

LCA is a methodological framework that quantifies all environmental impact caused during the life cycle of a product or a process. During the life of a product, other products are needed as input. For example, a bread requires the wheat to be harvested, energy for baking and transport, paper for packaging and so on. Wheat requires fertilizers, plant protection products, energy for cultivation, barns, etc. To produce the fertilizers, energy, natural gas and phosphorus ore are used. To fully account for all environmental impacts during the life cycle of the product under study (here the bread), this gives in theory a large network of products which contribution to the total environmental impact of the bread need to be quantified. As this is very difficult, an LCA study distinguishes "foreground" data, where the accounting of the relevant parameters for the impact (life cycle inventory) are measured or obtained from high quality data sources that are specific for the life cycle of the product, and "background" data where impact factors are obtained from databases, which are available (e.g., IPCC guidelines (http://www.ipcc-nggip.iges.or.jp/public/2006gl/) or OpenLCA (http://www.openlca.org/)).

LCA is used to support decisions and policies along production supply chains. It is a standardized method (ISO 14040:2006 and ISO 14044:2006) defined in four phases: goal and scope definition; life cycle inventory (LCI), consisting of the collection of data that identify the inputs and outputs of the systems, the algorithms to estimate the release of the pollutant into the environment; life cycle impact assessment (LCIA), where the emissions are converted to "impacts" with the help of impact models or impact characterization factors; interpretation.

A fundamental principle of a life cycle assessment (LCA) is that the impact of a so-called "functional unit" (such as a kg of bread) is assessed comprehensively without omitting one or several impact categories. This is based on the conviction that a comparison between different processes or value chain producing the same functional unit—or a comparison between different product cannot be done only if looking at, for example, greenhouse gas emissions or eutrophication or ozone depletion, but requires appreciation of all possible impacts.

Difference Between LCA and Footprints

In contrast to a LCA, an N footprint looks exclusively at the area of concern of N effects. This is similar to the water footprint looking at water issues, the carbon footprint quantifying total GHG emissions, or the ecological footprint being a comprehensive measure of the overall land requirement. Each of these footprints has been developed independently and consequently there are no common rules or definitions that apply to all of these—and other—footprints, even though recent efforts to classify footprints into a "footprint family" or providing rules for categorizing footprints try to overcome this situation.

Only recently first studies proposed classification systems for the "footprint family" or common definition (Fang *et al.*, 2016; Galli *et al.*, 2012). All footprints are based on life cycle thinking thus conceptually assessing supply chains and considering emissions from upstream processes. However, studies differ in which and how many processes are included (so-called system boundaries), and how emissions are to be distributed (allocated) if one process produces other products than the one of interest. Differences exist also in the methodology how N footprints are calculated. Therefore, it is important to be clear about those issues when using or comparing results from N footprint studies.

Nitrogen Footprints

In its "simplest" form, N footprints follow the flow of N in a supply chain and sums over all N losses that occur in the different stages. The total losses are then put in relationship to the N that is contained in the product at time of consumption. Losses of N occur also after consumption in the end-of-life stage of a product. All losses considered in N footprints of food supply chain are to be considered as anthropogenic.

In most cases, N footprints are quantified for food/agricultural products or food/agricultural commodities. This is because most nonagricultural products have little environmental relevance linked to N losses different than the losses occurring through energy consumption. This is true also for industrial products which contain N, such as nylon, adhesives, coatings, or explosives.

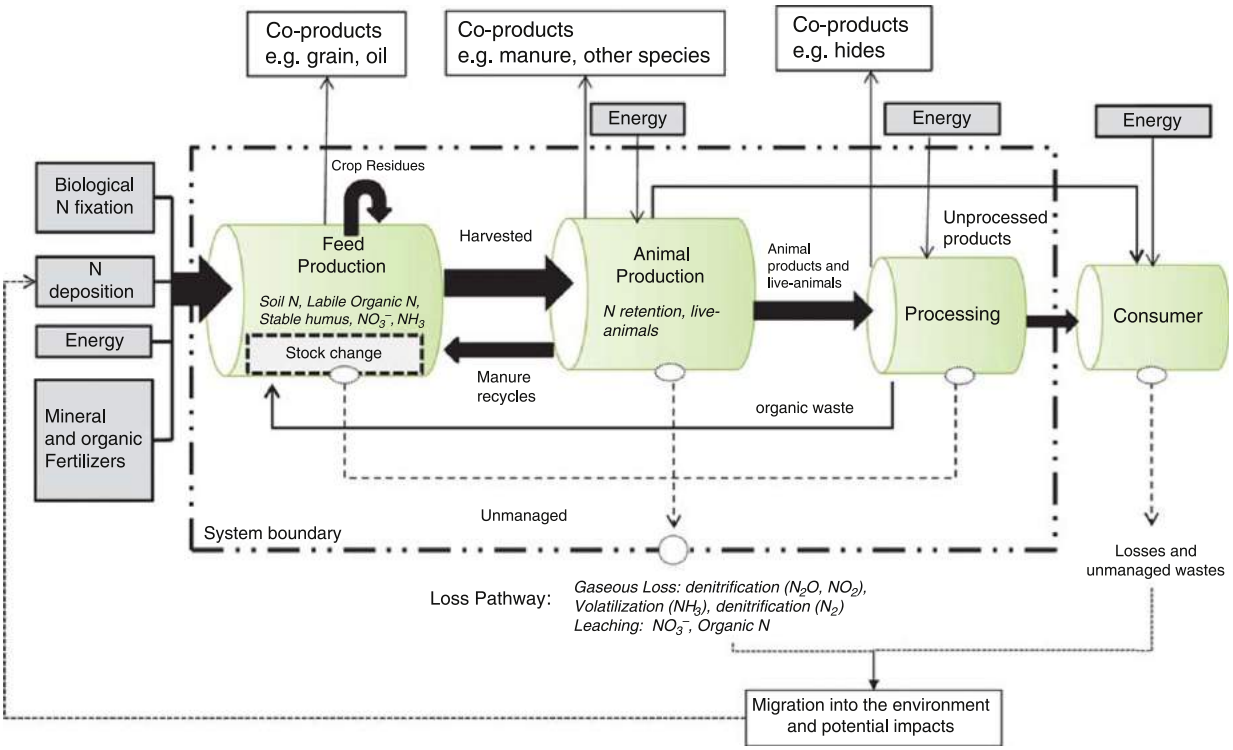


Fig. 3 Example for system boundaries and N flows in a livestock supply chain. Adapted from Uwizeye, A., Gerber, P.J., Schulte, R.P.O., de Boer, I.J.M., 2016. A comprehensive framework to assess the sustainability of nutrient use in global livestock supply chains. Journal of Cleaner Production 129, 647–658. doi:10.1016/j.jclepro.2016.03.108.

The ratio of total N losses N_{loss} and the consumption or intake of N N_{intake} is called the N-loss factor NLF (sometimes also called virtual N factor). It gives the amount of N, which is dispersed in the environment for each kg of N that is consumed. If the *NLF* is known, then the N footprint is obtained by multiplying *NLF* with the concentration of N in the product when it is consumed ($\frac{N_{intake}}{m}$).

The N footprint Φ_N calculates from:

$$\Phi_N = \frac{N_{loss}}{m} = NLF \cdot \frac{N_{intake}}{m}, \text{with } NLF = \frac{N_{loss}}{N_{intake}}$$

where Φ_N is the footprint of a product (kg N (kg product)$^{-1}$), N_{loss} are total N losses (kg N), N_{intake} is the consumption (intake) of N with the product (kg N), and m is the mass of the product (kg product).

These concepts are illustrated in **Figs. 3** and **4**.

Fig. 3 shows a simplified example of a supply chain, in this case for an animal product such as meat or eggs. The supply chain consists of four stages: cultivation of crops for feed, animal production, processing of the product and consumption. The figure shows the main types of input at each stage. Most diverse are the inputs in the crop cultivation stage, as N inputs are received in form of mineral and organic fertilizers, biological N fixation or from atmospheric deposition. In addition, N could be released from soils by mineralization of organic matter or—in contrast—be assimilated into organic matter and accumulate in the soil. For reasons that will be discussed below, such "input" is usually accounted for as a positive (if accumulation occurs) or negative (if soil N depletion occurs) coproduct. Reactive N emitted to the atmosphere during the combustion process for energy generation is considered as both "input" and "losses" so that the balance is closed. For all other stages than crop cultivation, emissions due to energy use is the only "input" into the process.

The most common N footprint—as also defined above—quantifies total N losses that are associated with a product consumed or with consumption of products and energy in general.

Several studies quantify only part of the N footprint, for example, only the N footprint from agricultural production ignoring N losses from energy consumption, or apply different methods for losses occurring in the agricultural supply chain and for losses related to energy consumption. It is therefore useful to distinguish "subfootprints" related to agricultural production and energy.

The total N footprint Φ_N is the sum of the agricultural N loss footprints $\Phi_{N_{loss,\ agri}}$ and the energy N loss footprint $\Phi_{N_{energy}}$.

$$\Phi_N = \Phi_{N_{loss,\ agri}} + \Phi_{N_{energy}}$$

The difference between N losses from agricultural production and energy consumption is illustrated in **Fig. 4**, showing schematically the different components of N inputs and outputs, emphasizing the occurrence of recycling flows.

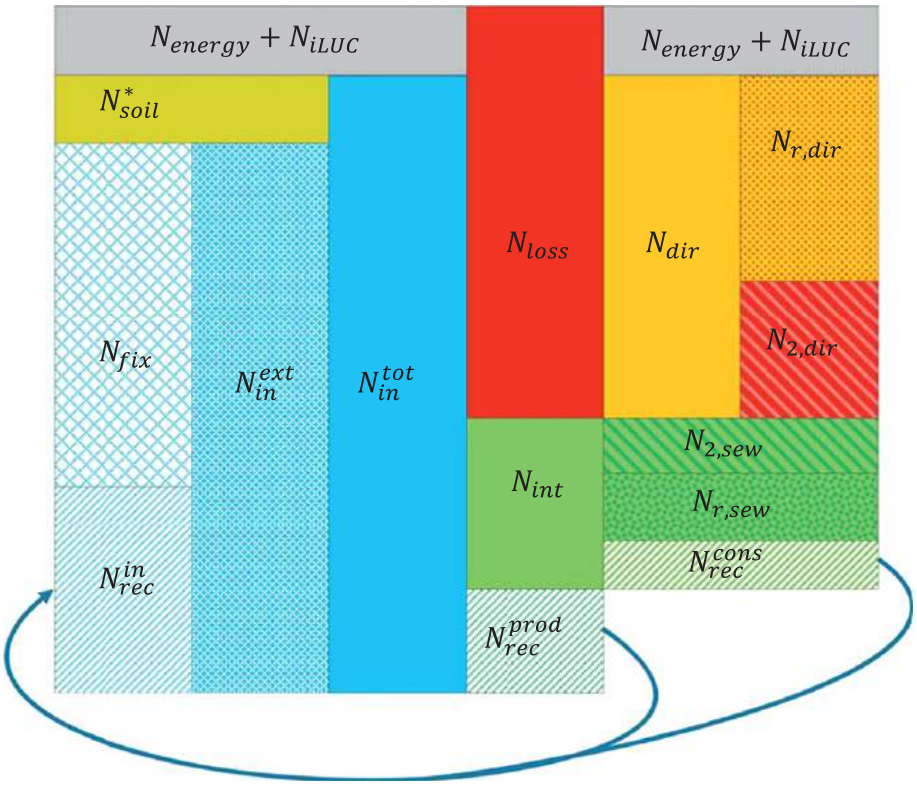


Fig. 4 Schematic representation of N input, N output and N recycling flows in a typical supply chain. Total N mobilization is the sum of N consumed (N_{int}), N lost to the environment (N_{loss}) and N that can be recycled in the process (N^{prod}_{rec}). The mobilized N is obtained from losses through the use of energy and land use changes (N_{energy}, N_{iLUC}) and through inputs to the soil (N^{tot}_{in}), either by mineralization of soil organic matter (N^{*}_{soil}) or by external input N^{ext}_{in} from newly fixed N (N_{fix}) which can be done either through biological nitrogen fixation or the Haber–Bosch process, or by recycling of N (N^{in}_{rec}). Losses of N are either the losses from energy and land use change or direct losses of reactive nitrogen or molecular nitrogen ($N_{r,\ dir}$, $N_{2,\ dir}$). N that has been consumed is partly converted to N_2, for example, in sewage system ($N_{2,\ sew}$) or lost ($N_{r,\ sew}$) or recovered (N^{cons}_{rec}) in sewage sludge. Hashed boxes represent losses of N with no direct impact on the environment, as molecular N is inert; double-hashed boxes represent creation of new reactive N (both synthetic and natural N fixation). Note that N^{*}_{soil} is indicated in the input-side, but is commonly accounted for as a coproduct in crop cultivation and thus as a positive output if soil N is accumulating or a negative output if N stocks are declining. This is denoted with the star-superscript. Reprinted from Leip, A., Leach, A., Musinguzi, P., Tumwesigye, T., Olupot, G., Stephen Tenywa, J., Mudiope, J., Hutton, O., Cordovil, C.M.dS., Bekunda, M., Galloway, J., 2014a. Nitrogenneutrality: A step towards sustainability. Environmental Research Letters 9. https://doi.org/10.1088/1748-9326/9/11/115001. Leip, A., Weiss, F., Lesschen, J.P., Westhoek, H., 2014b. The nitrogen footprint of food products in the European Union. The Journal of Agricultural Science 152. https://doi.org/10.1017/S0021859613000786.

Table 1 shows an example of different N loss and N input footprint for food production in the European Union (EU27), calculated for the year 2004 with the CAPRI model (Leip *et al.*, 2015) (Data available at https://doi.org/10.5281/zenodo.58514), with the objective to quantify the contribution various environmental pressures from agriculture and livestock production system in Europe. Losses correspond to total N losses from agricultural production and energy consumption; N input corresponds to N sources of crop production. The authors calculated also some agricultural N input footprints as shown in **Table 1**.

Agricultural Nitrogen Footprints

Even though the agricultural N loss footprint includes the loss of N_2 which does not have any adverse effect. It is an indicator for the anthropogenic mobilization of N (Pelletier and Leip, 2014). Thus, it is closely linked with the concept of the planetary boundary for N (Steffen *et al.*, 2015), and with the concepts of N use efficiency (NUE) and gross N balance (GNB)—the latter two indicators commonly use to describe the environmental performance of agricultural systems.

NUE and GNB can be estimated at a process level or at a supply chain level (Mu *et al.*, 2017; Uwizeye *et al.*, 2016, 2017). These indicators are more relevant for the production and indicate the ability of a supply chains to recover the N inputs into the end-products. NUE is a complementary indicator to N footprint, whereas GNB is slightly different from N footprint, because it doesn't account for soil N stock change. Beside the NUE and GNB a number of agri-environmental indicators are commonly used to describe the environmental performance of farms, amongst those the input of mineral fertilizer, total nitrogen input, emissions of ammonia, nitrogen leaching or the nitrogen surplus which is calculated as the difference of total N inputs and total useful N

Table 1 N Footprints for European (EU27) agricultural production in the year 2004

	Vegetable products							
	POTA	*SUGB*	*OILP*	*FRVG*	*CERR*	*LEGU*	*OCRP*	*CROPP*
Total Nr emissions according to the LCA approach								
N_2O	0.1	0.1	0.8	0.1	0.5	0.4	0.2	0.3
NO_x	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NH_3	0.4	0.2	1.9	0.3	2.1	0.8	0.5	1.2
NLR	1.3	1.1	8.2	1.0	6.1	4.3	2.5	3.7
Total	1.8	1.5	11.0	1.4	8.8	5.6	3.2	5.3
Total N emissions according to the LCA approach								
Input of mineral fertilizer N	3.7	2.5	34.0	3.8	19.5	10.0	8.0	12.1
Input of manure N	1.4	0.8	2.2	0.3	8.2	2.2	0.6	4.0
N in atmospheric deposition	0.5	0.3	3.9	0.6	2.4	4.8	1.5	1.5
Biological N fixation	0.0	0.0	2.0	0.0	0.0	36.9	0.0	0.4
Total	5.6	3.6	42.2	4.7	30.0	53.9	10.1	18.1
	Livestock products							
	BEEF	*PORK*	*EGGS*	*POUM*	*DAIR*	*SGMP*		*ANIMP*
Total Nr emissions according to the LCA approach								
N_2O	12.1	3.7	1.9	2.6	0.7	10.3		1.7
NO_x	4.7	2.5	0.8	1.5	0.3	4.2		0.9
NH_3	74.9	25.1	13.2	18.6	4.8	44.4		11.4
NLR	196.9	26.2	13.6	22.2	10.8	173.9		22.3
Total	288.6	57.5	29.5	44.9	16.7	232.9		36.4
Total N emissions according to the LCA approach								
Input of mineral fertilizer N	226	60	43	57	12	193		32
Input of manure N	285	37	25	33	17	282		34
N in atmospheric deposition	75	11	7	10	4	67		9
Biological N fixation	96	30	2	19	7	116		14
Total	683	138	77	119	40	658		89

All values are in g N (kg product)$^{-1}$ year^{-1}.

Acronyms: Vegetable products: POTA: potatoes, SUGB: sugar beet before processing, OILP: oil seeds before processing; CERR: cereals, LEGU: leguminous crops; OCRP: other crops; CROPP: aggregated vegetable food. a) Livestock products: BEEF: beef, PORK: pork, EGGS: eggs, POUM: poultry meat; DAIR: milk and dairy products, SGMP: meat from sheep and goats; ANIMP: aggregated livestock food.

outputs. These indicators are particularly linked to the stage of agricultural production and are usually expressed as kg N per hectare and year. Such indicators are thus available to describe farms, but they have limited value of describing the environmental performance with respect to nitrogen of food products obtained from different production systems or for which the land requirement is unknown. The nitrogen budget footprints use life cycle thinking for quantifying all terms in a full nitrogen budget and refer it to the final product.

It makes therefore sense to broaden the concept of N footprints and apply it not only to N losses, but also to N input sources, which allows to quantify indicators in analogy to the agro-environmental indicators already commonly used.

Thus, the nitrogen budget footprints can be regarded as a footprint "group," which encompasses:

- Agricultural N input footprints: footprint of synthetic fertilizer nitrogen application; footprints of organic fertilizer nitrogen application (manure, compost, sewage sludge …), and footprints of biological nitrogen fixation and atmospheric deposition. These footprints account for new nitrogen input with regard to agricultural production, thus newly fixed nitrogen or nitrogen that comes from a different sector. A manure footprint accounts only for manure application, which comes from other supply chains.
- Agricultural N loss footprints: NH_3 footprint, NO_x footprint, N-runoff and N-leaching footprints, N_2O and N_2 footprints.

Each of these agricultural N footprints quantifies the total agricultural flow in supply chains. They give all information required to calculate the NUE or GNB at product—or consumer level. Roughly the input footprints can be classified as "resource footprints" and the loss footprints can be classified as "emissions footprints," both with an orientation of inventorying flows rather than being impact-oriented (Fang *et al.*, 2016).

Both sets of agricultural N footprints—the total agricultural N input footprint and the total agricultural N loss footprint are closely related, as the different of the two are the N content in the product, as long as the losses in the end-of-life phase (after consumption) are not included.

$$\Phi_{N_{input}} = \Phi_{N_{loss}} + N_{intake}$$

$$N_{intake} = N_{loss}^{EoL} + N_{recycled}^{EoL}$$

Where EoL refers to the end-of-life stage of the life cycle of a product, and the subscripts *loss* and *recycled* refer to the quantities of N that are lost to the environment or re-used as fertilizer in agricultural production.

Energy N Footprint

NO_x and N_2O emissions occur both from agricultural systems and from energy consumption and some industrial processes including fertilizer production. $\Phi_{N_{energy}}$ all Nr emissions caused by the consumption of energy. They include energy-related emissions along the full supply chain, including the production of agricultural inputs (fertilizer, plant protection, etc.), energy use on the farm (fuels for tractors, heating, milking machines, electricity etc.) and postharvest processes such as drying, transport, processing, retailing, etc. depending of the system boundaries of the study.

Methodological Choices

Quantifying soil stock changes

Amongst the largest uncertainties for agricultural N indicators and N footprints is the level of changes in the N content in soils. Soil organic matter is an important factor determining soil fertility, which contains next to carbon also N. The build-up of soil organic matter is usually regarded as a positive outcome of farm activities, with positive effect on future crop yields and contributing to carbon sequestration. Carbon sequestration is considered amongst the most important mitigation options available today to farmers in the fight against climate change. Losses of soil organic matter over time instead is regarded as "soil mining," has the potential of jeopardizing future crop yields, causes soil degradation, and contributes to greenhouse gas emissions.

It has therefore been argued that even though mineralization of soil organic matter serves as a source of N for crop growth, the accumulation of N in newly created soil organic matter must be regarded as a positive outcome of the farming activity worth equal to the production of crops—in fact, N which accumulates in soil organic matter can be used in a later year for crop growth, in which case the N is to be accounted for as "negative" output.

However, quantification of soil stock changes is very difficult, as it is a relatively slow process and the detection of statistically significant differences in soil stocks requires measurement of soil N concentrations in distant points in time, for example, after 10 years. This is in most cases not possible. Therefore, different methods have been proposed to approximate soil stock changes in N budget or N footprint studies.

The simplest method is the "balance" approach. It considers soil stock changes as the most uncertain term in the N balance:

$$N_{input} = N_{output} + N_{stockchange}$$

Whereby N_{output} includes all output flows of N, both losses N_{loss} and N in products $N_{products}$. We use here $N_{products}$, as this N balance refers usually to the soil–crop continuum, thus the product is the crop harvested or otherwise removed from the soils, but additional losses can occur downstream of the supply chain before the food is consumed as N_{intake}. The problem with this method

Table 2 National consumer perspective N footprint estimates

Authors	*Scope*	*Method*		*N footprint (kg N cap^{-1} year^{-1})*	*Reference year*[b]
		Food[a]	*Energy*[a]		
Liang et al. (2016)	Australia	BU	BU/TD	47	2011
Pierer et al. (2014)	Austria	BU	TD	20	2009
Gu et al. (2013)	China	TD	TD	19	1980
Cui et al. (2016)	China	Hybrid	Hybrid	26–40	1990–2009
Gu et al. (2013)	China	TD	TD	49	2008
Leip et al. (2011)[c]	Europe (EU27)	TD	TD	37	2004
Leip et al. (2014a,b)[d]	Europe (EU27)	BU	–	16	2004
Leip et al. (2014a,b)[d]	EU, 27 countries	BU	–	6–47	2004
Shindo and Yanagawa (2017)	Japan	TD	–	17–18	2011
Shibata et al. (2014)	Japan, w/o trade	BU	BU/TD	37	n.d.
Shibata et al. (2014)	Japan, with trade	BU	BU/TD	28	n.d.
Leach et al. (2012)	Netherlands	BU	BU/TD	24	2011
Gonçalves (2013)	Portugal, Lisboa	BU	BU	24	n.d.
Hutton et al. (2017)	Tanzania	BU	BU/TD	10	n.d.
Stevens et al. (2014)	United Kingdom	BU	BU/TD	26	1971
Stevens et al. (2014)	United Kingdom	BU	BU/TD	27	2007
Leach et al. (2012)	United States	BU	BU/TD	41	2011
Oita et al. (2016)	Global	MRIO		27	n.d.
Oita et al. (2016)	Global, 188 cnt's	MRIO		7–225	n.d.

[a]Methodology: bottom-up (BU), top-down (TD), multiregional input–output assessment (MRIO), or hybrid (bottom-up/MRIO).

[b]Reference year often approximate as values from different years are used. The year usually refers to the year for which the consumption pattern was estimated. n.d. = no specific reference year was given.

[c]Calculated on the basis of total N emissions from the European Nitrogen Budget and total EU27 population. The N footprint accounts that about 70% of agricultural supply are used for domestic consumption, 30% for export.

[d]Calculated on the basis of food production footprints in g N/kg product and quantities used for human consumption as used in Leip *et al.* (2014a,b).

is that it requires other N output flows to be determined with high accuracy, which is often not the case, as the amount of diN produced is virtually never measured, but can constitute a considerable share of the total output flows.

Another method uses data sets for conditions where soil stock changes are assumed to be absent or very small. These data can be used to extrapolate to conditions where soil stock changes are very likely. In order to identify which data set is or is not likely to represent conditions with soil stock changes the "apparent" N use efficiency can be used, ignoring soil stock changes:

$$\mathrm{NUE} = \mathrm{N}_{output}/\mathrm{N}_{input}$$

In conditions where a share of the crop uptake N originates from mineralized soil organic matter, the NUE will assume higher values. Apparent NUE of more than 100% have been measured, in particular in Sub-Saharan African countries where soil mining is a severe environmental problem. A threshold of 85% has been proposed as a NUE, which could possibly be reached without drawing on soil resources. On the other hand, soil accumulation—if not quantified—pushes the apparent NUE down. This method has been tested over a multiannual data set for the soil N budget in Turkey at the regional level (Özbek and Leip, 2015).

A third method has been developed for countries with poor data availability and generally low N input level. In Tanzania, only a small share of soils is fertilized—on most fields crops grow on N supplied by atmospheric deposition, possibly biological N fixation and manure. The difference in the yield in fertilized versus un-fertilized yields has been used to approximate the possible order of magnitude for soil mining under Tanzanian conditions (Hutton *et al.*, 2017).

N_2 as nonharming N loss

The first definition of a N footprint was given by Leach *et al.* (2012) who developed a N-footprint model "[…] that defines an N footprint as the total amount of Nr that is lost to the environment due to individual's consumption of food and energy". This definition is explicitly given for their "per capita" N footprint model, thus the definition does not exclude that N footprint be quantified at product level or for groups of products or persons.

It has been argued that this definition was not completely clear in that the model included losses of N_2 in the food production chain, but excluded losses of N_2 in the food consumption part of the food N footprint. Leip *et al.* (2014a,b) interpreted this by differentiating between Nr releases, which refer to the status of N that is lost (such as fertilizer or manure N). This is in contrast to N_2 emissions which describe the status of N when it leaves the system boundaries (e.g., the crop root zone).

Generally, it is debated whether N_2 emissions should be included in N footprints. Arguments supporting it include the close relationship to other indicators such as N use efficiency and N surplus. These indicators are already established and are readily

calculated from in- and outputs of N, as least under those conditions where soil N stock changes are not significant. This makes interpretation easy. Moreover, N_2 emissions can be included due to its relevance for a measure of anthropogenic mobilization of N. On the other side, mainly the argument of lacking environmental relevance is made, as N_2 is inert and does not contribute to any environmental impact.

National N Footprint Estimates

Table 2 provides an overview of N footprint estimates at national scale and from a consumer perspective. Most of these studies include the entire economy, thus consider N losses from agriculture and energy consumption in all sectors, some focus on food the agricultural N footprint.

Most studies use a bottom-up approach for calculating the food footprint according to the methods proposed by Leach *et al.* (2012), estimating N-loss factors for different food or food groups and applying those factors to typical consumption rates obtained from surveys or food balance sheets. The energy footprint is often obtained from a combined bottom-up/top-down approach by quantifying typical consumption of energy per household or person for housing, cooking, transport and applying emission factors. More indirect energy-related emissions are calculated from a input–output table calculating the Leontief inversion (Leontief, 1966). Leip *et al.* (2014a,b) uses a bottom-up approach for food production footprint only, using data for representative farm at regional (province) level and calculating typical N footprints for 12 food groups. The authors present the results at farm-gate footprint, here the consumption footprint has been calculated using national consumption and population data.

A top-down method to calculate the N footprint has been applied by the studies of Gu *et al.* (2013) and Leip *et al.* (2011). These studies are based on a national material flow analysis of N in China and Europe, respectively, accounting for all Nr losses in agriculture and forestry, industry, energy and transport, waste and associate it with human consumption. This top-down method is not an input–output analysis, therefore does not account fully for trade flows. The data based on Leip *et al.* (2011) are corrected for an average ratio of domestic supply used for human consumption and export, however, as Europe exports high-footprint animal products in greater shares as crop products, this leads to an overestimation of the food production footprint.

Full accounting of global trade flows is possible in multiregional input–output analysis as done by Cui *et al.* (2016) for China and Oita *et al.* (2016) for 188 countries globally. These models are able to capture also indirect flows of embedded nitrogen losses such as export of goods, which are based on imported raw materials.

Nitrogen Footprint Derivatives

The idea of N footprints was very closely connected to the importance of the nitrogen–environment interaction being a highly complex multipollutant multieffect problem on one hand that stands in contrast to the lacking knowledge or mono-causal thinking in the population. Thus, N footprints shall serve both for awareness rising and giving a relatively simple instrument that helps reducing N pollution.

Different tools have, therefore, been proposed that can be used for different target groups and with a wide range of focus. An overview of such tools is given in Galloway *et al.* (2014).

N Calculator

The N calculator http://n-print.org/ (Leach *et al.*, 2012) is an online tool that asks for information on the normal consumption habit of a person and calculates the N footprint on the basis of representative N loss factors for the country the person lives in. Data are collected as average weekly servings of food groups such as poultry, pork, beef, seafood, cereals, potatoes, cheese, legumes and so on. Additionally, energy consumptions is estimated based on weekly energy and gas consumption and mobility behavior. The N calculator presents the results of the calculated footprint, indicating the share of food production, food consumption, housing, transportation, etc. on the total footprint and in comparison with the countries' average composition of the N footprint. Additionally, information is given, for example, on how the N footprint could be reduced, the difference of new and recycled N in food production, and organic versus conventional farming systems.

N Footprint for Institutions

The N footprint for institutions http://www.n-print.org/N-Institution (Castner *et al.*, 2017) is a tool that helps institutions such universities to quantify their N footprint, work out solutions for reducing it and developing indicators for monitoring progress. Castner *et al.* (2017) compare seven U.S. institutions. About 50% of the institution's footprint was associated with food production, one-third with utilities and about 8% with transportation. Novel metrics such as the share of animal product purchases relative to total purchases by weight, the average protein content of food purchases, or institution "virtual N factor" (N-loss factor,

Box 1 Definition of N neutrality

Definition of N neutrality

To achieve N neutrality,

1. first decrease the release of reactive nitrogen (Nr) into the environment by
   a. reducing over-consumption of food and reducing food wastes and minimizing energy consumption, and
   b. choosing sustainable sources of energy and food;
2. then, contribute to a measured compensation of the remaining Nr releases by a measured
   a. reduction of Nr releases elsewhere *to balance the remaining releases*,
   b. increased sustainability in the production of food *where sustainable land management is not yet achieved*.

Definition of sustainable land management

3. With respect to N neutrality, sustainable land management is a farming system which
   a. minimizes the ecological footprint of the farming products (incl. The C footprint, N footprint, water footprint)
   b. keeps the farmed land in good environmental conditions
   c. satisfies human food needs and enables the farm worker(s) and their families to a decent living standard

Reproduced from Leip, A., Leach, A., Musinguzi, P., Tumwesigye, T., Olupot, G., Stephen Tenywa, J., Mudiope, J., Hutton, O., Cordovil, C.M.dS., Bekunda, M., Galloway, J., 2014a. Nitrogenneutrality: A step towards sustainability. Environmental Research Letters 9. https://doi.org/10.1088/1748-9326/9/11/115001. Leip, A., Weiss, F., Lesschen, J.P., Westhoek, H., 2014b. The nitrogen footprint of food products in the European Union. The Journal of Agricultural Science 152. https://doi.org/10.1017/S0021859613000786.

measures in kg N losses per kg of N in purchased food) help the institutions to evaluate their performance in comparison to other institutions and thus identify possible directions for changes.

Nitrogen Neutrality

The N-neutrality concept (Leip *et al.*, 2014a,b) was developed for the 6th International N Conference to raise awareness amongst both conference participants and conference caterer about the consequences of N losses and possible reduction options. N neutrality is a concept similar to carbon-neutrality. In brief, it quantified the level of N pollution caused by and during the conference, implements reduction measures, and "offsets" the remaining N pollution by sponsoring a project that reduces N pollution elsewhere. The concept is thus a "hierarchical" one and allows N compensation only if all possibilities of N-loss reductions are exhausted. The definition of N neutrality as proposed by Leip *et al.* (2014a,b) is given in Box 1.

The concept of N neutrality has been challenged:

- It does not stop at the mere quantification of the N footprint and suggestion of N-loss reduction options, but it engages the conference organizers and participants to actively reduce the N footprint and "pay" for the N footprint that could not be reduced which might raise questions of responsibility and a possible reference level of N losses.
- This aspect of compensation has been criticized as favoring a mentality of "greenwashing" any N pollution (when the hierarchy of N neutrality is ignored or weakly interpreted), while impact of the offsetting project that is sponsored might be questionable.
- It might difficult to find a compensation project that reduces losses of N with equivalent impact, given the multitude of possible impacts that Nr losses have, and their huge variability in space and time. The N-neutrality concept can therefore embraced only if a more generic interpretation of "equivalent impact" is accepted andEncyclopedia of Ecology, 2nd edition reduction of N pressure rather than N impact is seen as the objective.

Since its introduction, the N-neutrality concept has been tested under different circumstances in a number of workshops and conferences.

Nitrogen Labels

Amongst the main purposes of the quantification of N footprints is to inform the "consumer" on the effect of his or her choices on the environment. Food labels are mandatory in many countries and give information on the nutritional value of the food. Quality labels inform about certain sustainability aspects, such as organic agriculture, fair trade, rainforest protection, GMO free etc. However, information on the environmental pressure in general such as greenhouse gas emissions, N pollution or water use are not established. Four possible designs of such labels are being discussed by Leach *et al.* (2012): (a) stars label; (b) stoplight label; (c) nutrition label add-on (based on the U.S. label design indicating the percent of daily recommended value for each nutrient—here the percent of the N footprint relative to a reference diet); and (d) detailed comparison table giving easy-to-interpret comparisons.

See also: Ecological Processes: Biological Nitrogen Fixation; Nitrification. Global Change Ecology: Nitrogen Cycle. Human Ecology and Sustainability: The Anthropocene; The Water-Energy-Food-Ecosystems (WEFE) Nexus

References

Bruckner, M., Fischer, G., Tramberend, S., Giljum, S., 2015. Measuring telecouplings in the global land system: A review and comparative evaluation of land footprint accounting methods. Ecological Economics 114, 11–21. doi:10.1016/j.ecolecon.2015.03.008.

Butterbach-Bahl, K., Gundersen, P., Ambus, P., Augustin, J., Beier, C., Boeckx, P., Dannemann, M., Gimeno, B.S., Kiese, R., Kitzler, B., Ibrom, A., Rees, R.M., Smith, K.A., Stevens, C., Vesala, T., Zechmeister-Boltenstern, S., 2011. Nitrogen processing in the biosphere. In: Sutton, M., Howard, C., Erisman, J.W., Billen, G., Bleeker, A., van Grinsven, H., Grennfelt, P., Grizzetti, B. (Eds.), European nitrogen assessment. Cambridge, UK: Cambridge University Press, pp. 99–125.

Castner, E.A., Leach, A.M., Compton, J.E., Galloway, J.N., Andrews, J., 2017. Comparing institution nitrogen footprints: Metrics for assessing and tracking environmental impact. Sustainability: The Journal of Record 10, 105–113. doi:10.1089/sus.2017.29090.eac.

Cui, S., Shi, Y., Malik, A., Lenzen, M., Gao, B., Huang, W., 2016. A hybrid method for quantifying China's nitrogen footprint during urbanisation from 1990 to 2009. Environment International 97, 137–145. doi:10.1016/j.envint.2016.08.012.

Erisman, J.W., Sutton, M.A., Galloway, J., Klimont, Z., Winiwarter, W., 2008. How a century of ammonia synthesis changed the world. Nature Geoscience 1, 636–639. doi:10.1038/ngeo325.

Fang, K., Song, S., Heijungs, R., de Groot, S., Dong, L., Song, J., Wiloso, E.I., 2016. The footprint's fingerprint: On the classification of the footprint family. Current Opinion in Environment Sustainability 23, 54–62. doi:10.1016/j.cosust.2016.12.002.

Fowler, D., Coyle, M., Skiba, U., Sutton, M.A., Cape, J.N., Reis, S., Sheppard, L.J., Jenkins, A., Grizzetti, B., Galloway, J.N., Vitousek, P., Leach, A., Bouwman, A.F., Butterbach-Bahl, K., Dentener, F., Stevenson, D., Amann, M., Voss, M., 2013. The global nitrogen cycle in the twenty-first century. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 368. doi:10.1098/rstb.2013.0164.

Galli, A., Wiedmann, T., Ercin, E., Knoblauch, D., Ewing, B., Giljum, S., 2012. Integrating ecological, carbon and water footprint into a "footprint family" of indicators: Definition and role in tracking human pressure on the planet. Ecological Indicators 16, 100–112. doi:10.1016/j.ecolind.2011.06.017.

Galloway, J.N., Cowling, E.B., 2002. Reactive nitrogen and the world: 200 years of change. AMBIO: A Journal of the Human Environment 31, 64–71. doi:10.1579/0044-7447-31.2.64.

Galloway, J.N., Leach, A.M., 2016. Sustainability: Your feet's too big. Nature Geoscience 9, 97–98. doi:10.1038/ngeo2647.

Galloway, J.N.N., Dentener, F.J.J., Capone, D.G.G., Boyer, E.W.W., Howarth, R.W.W., Seitzinger, S.P., Asner, G.P.P., Cleveland, C.C.C., Green, P.A., Holland, E.A., Karl, D.M., Michaels, A.F., Porter, J.H., Townsend, A.R., Vöosmarty, C.J., 2004. Nitrogen cycles: past, present, and future. Biogeochemistry 70, 153–226. doi:10.1007/s10533-004-0370-0.

Galloway, J.N., Burke, M., Bradford, G.E., Naylor, R., Falcon, W., Chapagain, A.K., Gaskell, J.C., McCullough, E., Mooney, H.A., Oleson, K.L.L., Steinfeld, H., Wassenaar, T., Smil, V., 2007. International trade in meat: The tip of the pork chop. AMBIO: A Journal of the Human Environment 36, 622–629. doi:10.1579/0044-7447(2007)36[622:ITIMTT]2.0.CO;2.

Galloway, J.N., Winiwarter, W., Leip, A., Leach, A.M., Bleeker, A., Erisman, J.W., 2014. Nitrogen footprints: Past, present and future. Environmental Research Letters 9. doi:10.1088/1748-9326/9/11/115003.

Gonçalves, V.M.P., 2013. Impact of nitrogen into the environment. In: A step on nitrogen footprint calculation in Lisbon, Portugal (MSc thesis). Lisboa, Portugal: Istituto Superior de Agronomia, Universidade Tecnnica de Lisboa. https://doi.org/10400.5/5738

Gu, B., Leach, A.M., Ma, L., Galloway, J.N., Chang, S.X., Ge, Y., Chang, J., 2013. Nitrogen footprint in China: Food, energy, and nonfood goods. Environmental Science & Technology 47, 9217–9224. doi:10.1021/es401344h.

Hoekstra, A.Y., Wiedmann, T.O., 2014. Humanity's unsustainable environmental footprint. Science 344, 1114–1117. doi:10.1126/science.1248365.

Hutton, M.O., Leach, A.M., Leip, A., Galloway, J.N., Bekunda, M., Sullivan, C., Lesschen, J.P., 2017. Toward a nitrogen footprint calculator for Tanzania. Environmental Research Letters 12. doi:10.1088/1748-9326/aa5c42.

Kastner, T., Kastner, M., Nonhebel, S., 2011. Tracing distant environmental impacts of agricultural products from a consumer perspective. Ecological Economics 70, 1032–1040. doi:10.1016/j.ecolecon.2011.01.012.

Leach, A.M., Galloway, J.N., Bleeker, A., Erisman, J.W., Kohn, R., Kitzes, J., 2012. A nitrogen footprint model to help consumers understand their role in nitrogen losses to the environment. Environmental Development 1, 40–66. doi:10.1016/j.envdev.2011.12.005.

Leip, A., Rostislav, N., Čermák, P., LeGall, A.-C., Geupel, M., Spranger, T., Bleeker, A., Achermann, B., Heldstab, J., Johnes, P., Dragosits, U., Fernall, D., Sutton, M.A., 2011. Integrating nitrogen fluxes at the European scale. Supplementary Material: Section A—National integrated nitrogen budgets. In: The European nitrogen assessment. Cambridge: Cambridge University Press.

Leip, A., Leach, A., Musinguzi, P., Tumwesigye, T., Olupot, G., Stephen Tenywa, J., Mudiope, J., Hutton, O., Cordovil, C.M.dS., Bekunda, M., Galloway, J., 2014a. Nitrogen-neutrality: A step towards sustainability. Environmental Research Letters 9. doi:10.1088/1748-9326/9/11/115001.

Leip, A., Weiss, F., Lesschen, J.P., Westhoek, H., 2014b. The nitrogen footprint of food products in the European Union. The Journal of Agricultural Science 152. doi:10.1017/S0021859613000786.

Leip, A., Billen, G., Garnier, J., Grizzetti, B., Lassaletta, L., Reis, S., Simpson, D., Sutton, M.A., de Vries, W., Weiss, F., Westhoek, H., 2015. Impacts of European livestock production: Nitrogen, sulphur, phosphorus and greenhouse gas emissions, land-use, water eutrophication and biodiversity. Environmental Research Letters 10. doi:10.1088/1748-9326/10/11/115004.

Lenzen, M., Murray, J., Sack, F., Wiedmann, T., 2007. Shared producer and consumer responsibility—Theory and practice. Ecological Economics 61, 27–42. doi:10.1016/j.ecolecon.2006.05.018.

Leontief, W., 1966. Input–output economics. New York: Oxford University Press.

Liang, X., Leach, A.M., Galloway, J.N., Gu, B., Lam, S.K., Chen, D., 2016. Beef and coal are key drivers of Australia's high nitrogen footprint. Scientific Reports 6, 4–11. doi:10.1038/srep39644.

Marques, A., Rodrigues, J., Lenzen, M., Domingos, T., 2012. Income-based environmental responsibility. Economic Degrowth 84, 57–65. doi:10.1016/j.ecolecon.2012.09.010.

Marques, A., Rodrigues, J., Domingos, T., 2013. International trade and the geographical separation between income and enabled carbon emissions. Ecological Economics 89, 162–169. doi:10.1016/j.ecolecon.2013.02.020.

Mu, W., Groen, E., van Middelaar, C., Bokkers, E., Hennart, S., Stilmant, D., de Boer, I., 2017. Benchmarking nutrient use efficiency of dairy farms: The effect of epistemic uncertainty. Agricultural Systems 156, 25–33.

Oita, A., Malik, A., Kanemoto, K., Geschke, A., Nishijima, S., Lenzen, M., 2016. Substantial nitrogen pollution embedded in international trade. Nature Geoscience. doi:10.1038/ngeo2635.

Özbek, F.Ş., Leip, A., 2015. Estimating the gross nitrogen budget under soil nitrogen stock changes: A case study for Turkey. Agriculture, Ecosystems and Environment 205, 48–56. doi:10.1016/j.agee.2015.03.008.

Pelletier, N., Leip, A., 2014. Quantifying anthropogenic mobilization, flows (in product systems) and emissions of fixed nitrogen in process-based environmental life cycle assessment: Rationale, methods and application to a life cycle inventory. International Journal of Life Cycle Assessment 19. doi:10.1007/s11367-013-0622-0.
Pierer, M., Winiwarter, W., Leach, A.M., Galloway, J.N., 2014. The nitrogen footprint of food products and general consumption patterns in Austria. Food Policy 49, 128–136. doi:10.1016/j.foodpol.2014.07.004.
Shibata, H., Cattaneo, L.R., Leach, A.M., Galloway, J.N., 2014. First approach to the Japanese nitrogen footprint model to predict the loss of nitrogen to the environment. Environmental Research Letters 9. doi:10.1088/1748-9326/9/11/115013.
Shindo, J., Yanagawa, A., 2017. Top-down approach to estimating the nitrogen footprint of food in Japan. Ecological Indicators 78, 502–511. doi:10.1016/j.ecolind.2017.03.020.
Steffen, W., Richardson, K., Rockström, J., Cornell, S., Fetzer, I., Bennett, E., Biggs, R., Carpenter, S., 2015. Planetary boundaries: Guiding human development on a changing planet. Science 348, 1217. doi:10.1126/science.aaa9629.
Stevens, C.J., Leach, A.M., Dale, S., Galloway, J.N., 2014. Personal nitrogen footprint tool for the United Kingdom. Environmental Science: Processes & Impacts 16, 1563–1569. doi:10.1039/C3EM00690E.
Sutton, M.A., Oenema, O., Erisman, J.W., Leip, A., van Grinsven, H., Winiwarter, W., 2011a. Too much of a good thing. Nature 472, 159–161. doi:10.1038/472159a.
Sutton, M., Howard, C., Erisman, J., 2011b. The European nitrogen assessment: Sources, effects and policy perspectives. 612. Cambridge, UK: Cambridge University Press.
Uwizeye, A., Gerber, P.J., Schulte, R.P.O., de Boer, I.J.M., 2016. A comprehensive framework to assess the sustainability of nutrient use in global livestock supply chains. Journal of Cleaner Production 129, 647–658. doi:10.1016/j.jclepro.2016.03.108.
Uwizeye, A., Gerber, P.J., Groen, E.A., Dolman, M.A., Schulte, R.P.O., de Boer, I.J.M., 2017. Selective improvement of global datasets for the computation of locally relevant environmental indicators: A method based on global sensitivity analysis. Environmental Modelling and Software 96, 58–67. doi:10.1016/j.envsoft.2017.06.041.

Further Reading

NRC, 1993. Soil and water quality: An agenda for agriculture. Washington, DC: National Academy Press.

Ozone Layer☆

D Karentz, University of San Francisco, San Francisco, CA, USA

Introduction

Exposure to sunlight is generally an unavoidable consequence of being alive and life on Earth relies on the ozone layer to attenuate biologically harmful ultraviolet B (UVB: 280–315 nm) wavelengths of the solar spectrum. While ozone depletion has initiated concern about the effects of increased UVB on the biosphere, UVB filtered through a normal ozone column is still a considerable environmental stress. UVB exposure can cause declines in growth and reproduction that can eventually lead to reduced productivity or death; and although damage and protective/repair mechanisms are common across taxa, there is wide species-specific variation in UVB responses. In order to understand the ecological impact of the ozone layer as a UVB filter, the UV photobiology of species, populations, and individuals needs to be understood in the context of primary productivity, biodiversity, trophic energy transfer, and biogeochemical cycling.

Ozone and UVB

Ozone and the Spectral Quality of Incident Sunlight

Atmospheric attenuation of the solar spectrum

Radiation emitted by the Sun ranges from gamma rays (10^{-18} m) to radio waves (10^{7} m). Fortunately, at sea level, solar radiation has been stripped of its most biologically damaging wavelengths by gases in the atmosphere, most notably water vapor and ozone. All X-rays, some UVC (100–280 nm) and some UVB (280–315 nm) wavelengths are absorbed in the outermost atmospheric layers, the thermosphere and the mesosphere. The stratosphere contains 90% of the ozone in the atmosphere, and this ozone layer filters out the remaining UVC and most UVB (Figures 1 and 2). The shortest UVB wavelengths reaching the ground are 285–290 nm.

The attenuation of UVC and UVB is through the photochemical dissociation of ozone and its subsequent reformation from the diatomic and singlet oxygen products ($O_3 \leftrightarrow O_2 + O$). UVA (315–400 nm), visible (400–750 nm), and infrared

☆*Change History:* March 2013. D Karentz updated graphic and legend of Figure 3; and edited entries in Further Reading and list of Relevant Websites.

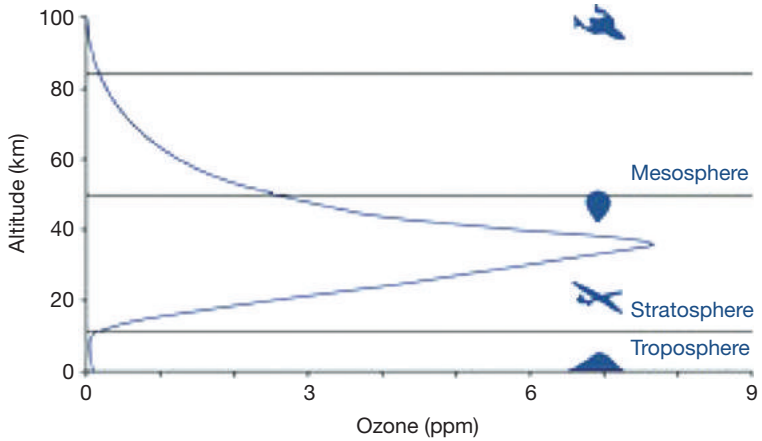


Figure 1 Vertical distribution of ozone in the atmosphere and location of the ozone layer. Redrawn from US National Aeronautics and Space Administration (http://www.nasa.gov).

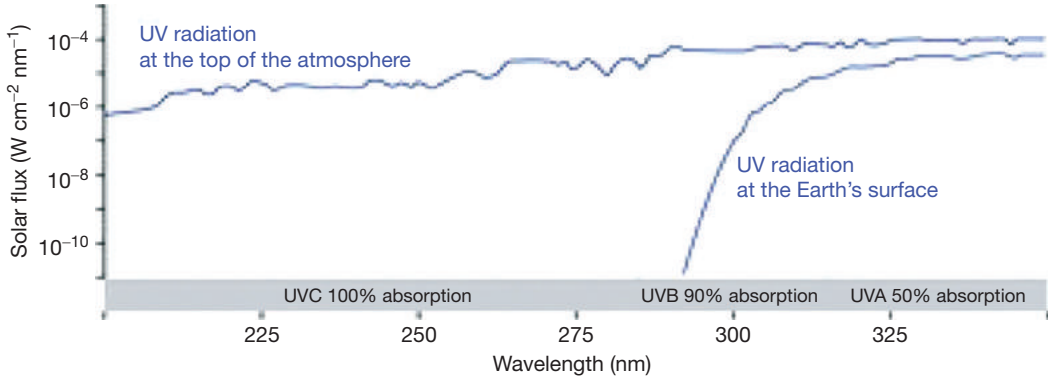


Figure 2 Differences in spectral quality and intensity of UV radiation at the top of the atmosphere and at the Earth's surface. Redrawn from US National Aeronautics and Space Administration (http://www.nasa.gov).

(750 to 1×10^6 nm) wavelengths also reach sea level and are partially attenuated by atmospheric gases and aerosols, but are not affected by ozone concentrations.

Clouds play a dominant and complex role in reducing UVB intensities. Clouds do not attenuate UVB wavelengths as efficiently as UVA or visible light; therefore, while cloudy skies lower incident radiation intensities, they may enhance the ratio of UVB to higher wavelengths.

Ozone depletion

From the 1970s to the turn of the century, synthetic halogenated compounds (e.g., chlorofluorohydrocarbons, CFCs) initiated declines of 3–6% in stratospheric ozone levels over tropical and temperate latitudes (60° N–60° S). More ozone was depleted over polar regions, with the largest-magnitude depletion (over 50%) still occurring seasonally over Antarctica (**Figure 3**). While the pollutant compounds are quite stable in the troposphere with long residence times (40–80 years), they eventually migrate into the stratosphere and disrupt the equilibrium of ozone dissociation and reformation described above (**Figure 4**). International compliance with the Montreal Protocol on Substances that Deplete the Ozone Layer has successfully limited the release of ozone-depleting substances, and ozone layer recovery to pre-1980 status is expected within the next 100 years.

Considerations for UV Exposure

In addition to ozone and clouds, numerous other factors can influence the amount of UVB exposure received by an organism. Latitude affects the intensity of sunlight through solar zenith angle (greatest at the poles) and thickness of the atmosphere (greatest at the equator). In terrestrial habitats, UVB doses are modified by canopy layers and availability of shaded microenvironments. In aquatic ecosystems, dissolved substances and particulate matter in the water column usually eliminate biologically harmful intensities of UVB within the upper 20 m. Vertical mixing will further determine actual exposures of planktonic organisms.

Temporal exposure factors of seasons and photoperiod are also defined by latitude, and local weather will affect the intensity of UVB on scales of minutes or days or longer. At latitudes below 60°, changes in ozone concentration occurred gradually over two

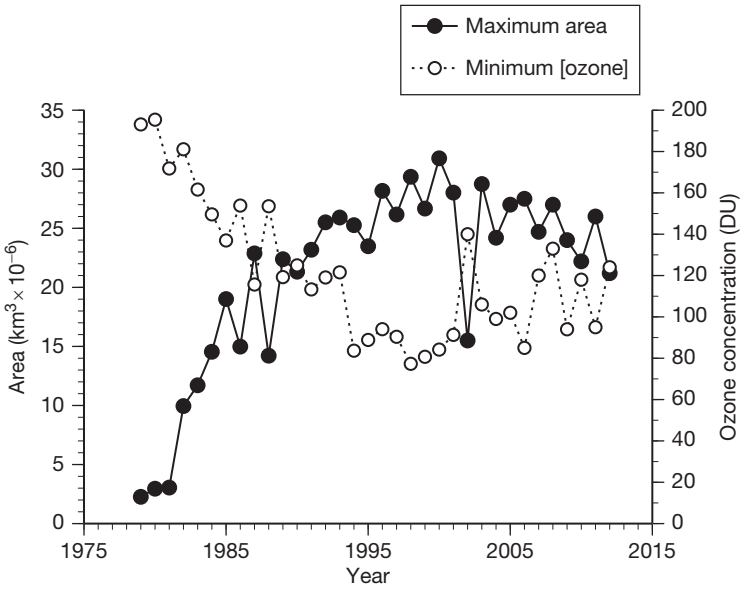


Figure 3 Maximum area (millions of km^3) and minimum column ozone concentration (Dobson units, DU) from 1979 to 2012 during the annual springtime ozone depletion cycle over Antarctica. Normal ozone concentrations are in the range of 350–400 DU. Data from US National Aeronautics and Space Administration (http://www.nasa.gov).

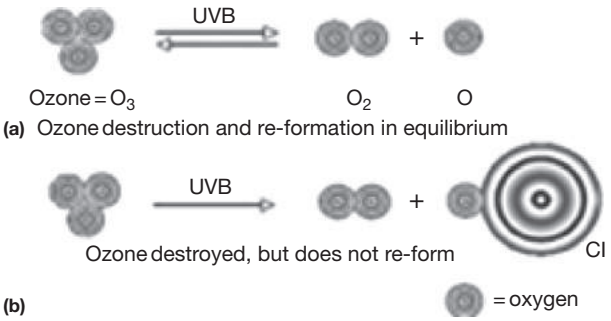


Figure 4 Simplified model of the photochemistry of ozone depletion. (a) Ozone dissociates when UVB is absorbed and re-forms to absorb more UVB. This sequence depicts the normal equilibrium state of this photochemical reaction. (b) Components of pollutant molecules, such as the chlorine shown here, can react with singlet oxygen producing compounds with very long dissociation rates, greatly slowing down the re-formation process of ozone by sequestering singlet oxygen molecules. Thus, ozone breaks down at a rapid rate and is not regenerated to maintain equilibrium concentrations.

decades and were accompanied by measurable increases in UVB. Over Antarctica, large predictable springtime declines of ozone cause substantial short-term increases in UVB radiation.

Origin of the Ozone Layer and Evolution of Life

Early Earth atmospheres (>3 billion years ago) did not contain the 21% oxygen content that we have today. When life on Earth originated (~3.5 billion years ago) the first prokaryotic cells were anaerobic chemoautotrophs, most likely occupying deep ocean and subterranean habitats. With no oxygen in the atmosphere and no ozone layer, incident UVC and UVB levels would have been extreme, forcing life to remain in dim and unlighted habitats. It was the evolution of photosynthesis that added free oxygen to the atmosphere and resulted in the formation of the ozone layer. The subsequent removal of UVC and reduction of UVB from incident sunlight very likely contributed to life moving from the oceans into illuminated terrestrial environments.

A Note on Ozone Pollution

While stratospheric ozone lessens environmental stress by attenuating UVB, there is also an issue of ozone pollution in the troposphere. Photochemical reactions of oxygen with by-products from automobile exhaust and industrial emissions result in surface accumulations of ozone. Ozone is a very reactive compound and direct contact with cells and tissues causes oxidative

damage which can lead to death in plants and animals. Tropospheric ozone pollution is a growing concern for potential impacts on human health and ecosystems, but it is not related to the stratospheric ozone layer.

Biological Consequences of UVB Exposure

Molecular Damage and Cellular Impacts

Damage of biological components by UVB can be manifested in two ways: (1) direct absorption of UVB by organic molecules, or (2) oxidation of organic molecules by reactive oxygen species (ROS) and other radicals that are produced by the UV photolysis of water (e.g., hydroxyl ions, peroxyl ions, singlet oxygen, superoxide) or dissolved substances in the extra- or intracellular environment.

Many organic molecules absorb UVB and undergo conformational modifications that interfere with physiological processes. There are a number of UVB-induced DNA photoproducts such as cyclobutane pyrimidine dimers (CPDs), 6-4 pyrimidine-pyrimidone adducts, pyrimidine hydrates, and DNA–protein crosslinks. CPDs are the most abundantly formed type of DNA damage, but the presence of any lesions in the DNA molecule can interfere with DNA replication and RNA transcription processes, affecting physiological rates, cell growth, and viability. ROS and other radicals also cause specific types of oxidative damage to DNA that can be detrimental. UVB-induced DNA damage can result in debilitating, mutagenic, and lethal effects.

UVB can also damage proteins by direct absorption or oxidation by UVB-induced radicals. Aromatic amino acids (tyrosine, phenylalanine, and tryptophan) absorb strongly at the lower end of the UVB range (280 nm) and account for the UV-absorbing properties of polypeptides. UVB exposure can alter rates of protein synthesis and turnover. Since proteins have diverse functions (e.g., enzymes, protective and structural components, energy storage, molecular motors, hormones, etc.), UVB-induced damage can have a wide range of effects.

Lipid molecules can also be damaged by UVB with the greatest biological hazard manifested in damage to membranes. Peroxidation of lipids by the action of UVB is a significant stress for cells to overcome.

Biological Defenses

Organisms have two main lines of defense against UVB: (1) avoidance of exposure (e.g., moving away from UVB, having external layers that block UVB transmission, synthesizing specific compounds that function as natural sunscreens); and (2) repair pathways that can recognize damage and either correct the UVB-induced defect or destroy the compromised molecules. While species have various combinations and efficiencies of these strategies, there is a high degree of similarity in UV defenses across prokaryotic and eukaryotic taxa. Life originated before the ozone layer formed and early organisms had to deal with more dangerous portions of the solar spectrum; thus, acquired effective defenses for avoiding or mitigating UV-induced damages have been retained. However, these highly conserved biological measures against UVB are not 100% successful. Nearly all organisms have a threshold limit for solar UV exposure.

UV detection

Before organisms can take the obvious step to avoid UVB exposure, they must be able to detect the presence of UVB wavelengths. Many species can detect and respond to changing intensities of white (visible) light with positive or negative phototaxis, or physiological adjustments (e.g., induction of sunscreening compounds). While UVA vision is an important aspect of mate selection and feeding in some birds, fish, and insects, there are only a few reports of UVB vision. This may be related to lack of research in this area and past limitations of technology for measuring UVB fluences. UVB perception may be more widespread and new methods could provide more complete information.

Avoidance by mobility

Most protists, invertebrates, and vertebrates are capable of moving away from too bright levels of sunlight, thus avoiding excessive UVB exposure. Taking advantage of shade, burrowing, or swimming deeper into the water column will effectively reduce the dose of UVB received. Nonmotile organisms (e.g., plants, fungi, macroaglae) cannot make such adjustments and must rely on other mechanisms for sufficient protection.

External barriers

The outer covering of any organism, such as hair, feathers, skin, shell, exoskeleton, cell wall, or only the cell membrane, serves as the first layer of protection against the environment. Many of these structures have evolved to minimize mechanical damage from physical stresses and predation pressure, but these external layers also serve to attenuate and often completely block the transmission of UVB before it can reach vital internal targets.

Natural sunscreens

The majority of prokaryotic and eukaryotic species synthesize UV-absorbing compounds that can serve as natural sunscreens (Table 1). Many of these compounds are common across taxonomic groups and have multiple functions; they not only provide protection from UV, but can act as antioxidants, signal transducers, osmoregulators, structural components, etc. UV-absorbing

Table 1 Some of the common UV-absorbing compounds that serve as sunscreens and examples of representative taxa. Many of these compounds absorb most strongly in the UVA, but attenuate UVB wavelengths as well. Within each of these groups, some compounds also can act as antioxidants. Presence in a particular taxonomic group does not necessarily indicate the ability to synthesize the UV-absorbing compounds; animals often bioaccumulate UV protectants (e.g., MAAs, carotenoids)

Compound	*Occurrence*
Carotenoids	Widespread in prokaryotes and eukaryotes
Melanins	Widespread in prokaryotes and eukaryotes
Mycosporine-like amino acids (MAAs)	Cyanobacteria, algae, marine invertebrates, fish
Polyphenolics (includes flavinoids, phlorotannins)	Plants, algae
Pteridines	Widespread in prokaryotes and eukaryotes

Table 2 Examples of biological compounds involved in antioxidation

Carotenoids
Catalase
Glutathione
Melanins
Polyphenolics
Pteridines
Superoxide dismutase (SOD)
Ubiquinone (coenzyme Q)
Uric acid
Vitamin A (retinol)
Vitamin C (ascorbic acid)
Vitamin E (tocopherol)

compounds can be colorless substances (e.g. mycosporine-like amino acids) or pigments (e.g., melanin). Many are secondary metabolites produced via pathways involved with the synthesis of aromatic amino acids.

Antioxidants

The generation of radicals by UVB interactions with aqueous solutions inside and outside of cells can be counteracted by the presence of antioxidants (Table 2). These compounds are capable of safely quenching ions before they oxidize DNA, proteins, and lipids. Several vitamins (e.g., A, C, E) and some enzymes (e.g., superoxide dismutase, catalase) play major roles in capturing and stabilizing free radicals in cells. Like the sunscreening molecules discussed above, antioxidants can often have multiple functions in cell metabolism.

Physiological repair pathways

A universal activity in cells is DNA repair, and there are several ways by which cells can identify and repair UVB-induced photoproducts. Nucleotide excision repair is the most common pathway and involves a suite of enzymes that can identify, remove, and replace damaged portions of DNA. Some taxa can directly reverse CPDs by photoreactivation, a process that requires the enzyme photolyase and the presence of UVA or visible light. Species might have a single or multiple repair pathways. An important factor in the successful mitigation of UVB effects is the balance between rates of DNA damage and repair.

Organismal Responses

The consequences of UVB exposure on an individual organism are dependent on intensity, spectral quality, duration of exposure, and effectiveness of protection and repair capabilities. The biological effects of UVB are wavelength dependent and best described by spectral weighting functions (Figure 5). Doses comprised of short intense exposures can have different effects from the same amounts of UVB delivered over longer time periods (i.e., reciprocity usually does not hold).

Differential species responses are an important consideration relative to biodiversity and ecological implications of UVB stress. In addition, there are intraspecies variations at the population and individual level. Life history and stage of development also need to be considered. Eggs, embryos, larvae, and juveniles are more sensitive to UVB exposure than larger adult stages; thus, UVB exposure can play a significant role in the age structure and maintenance of populations.

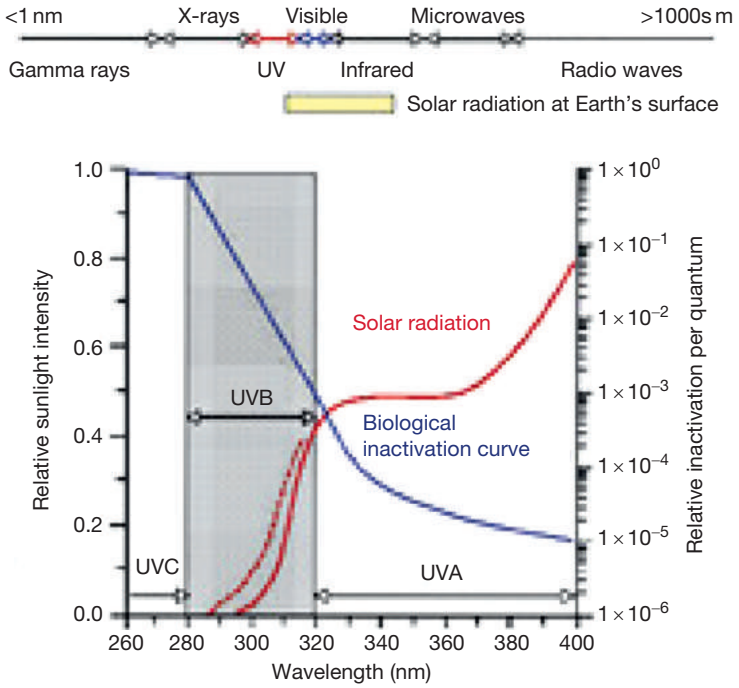


Figure 5 Comparison of intensity of incident UV wavelengths (solid red line) and generalized biological inactivation curve (blue line, data from Tyrrell, R. M and Pidoux, M. (1987). Action spectra for human skin cells: Estimates of the relative cytotoxicity of the middle ultraviolet, near ultraviolet, and violet regions of sunlight on epidermal keratinocytes. *Cancer Research* **47**, 1825–1829). The estimated shift in UVB wavelength and intensity with 50% ozone depletion is represented by the interrupted red line. The electromagnetic spectrum shows the full scale of solar radiation at the top of the atmosphere. Incident radiation falls within the wavelengths bracketed by the yellow bar.

Ecosystem Effects

What is the effect of UVB on the biosphere? While the UVB photobiology of individual species can be studied and characterized, evaluating the quantitative aspects of the ozone layer (with or without depletion) relative to species interactions or ecosystem-level processes has proven to be very difficult. First, concerted research efforts on ecological aspects of ozone depletion were initiated a decade after ozone depletion had begun (1970s). Organisms that are being studied today are already the result of population responses to increasing UVB. Second, isolating UVB effects from those of other solar wavelengths and the myriad of other environmental variables that regulate physiology, growth, and reproduction poses technological challenges. Third, is the issue of determining the proper biological endpoints to use as a measure of ecosystem UVB stress.

Characterizing UVB exposure in a particular environment is also not straightforward. Long-term studies are required to compensate for the high degree of temporal variability in UVB fluences caused by seasons and weather. There are limited data available on incident UVB fluences from before the 1970s, and it is only with the discovery of ozone depletion that networks for long-term monitoring are being established (see http://uv.biospherical.com/ for international listing of programs).

Biodiversity

Differential species and life-history responses to UVB exposure have dictated species distributions and shaped population and community structure over geologic time. One of the largest concerns about the impact of increased UVB caused by recent ozone depletion is that the changes in ambient UVB may have occurred on a much more rapid timescale that can naturally be accommodated by adaptation and natural selection. This would be especially true for unicellular organisms and those with short life-cycle times. Even now, UV-sensitive species may have already been replaced by more UV-tolerant taxa. Larger organisms are usually considered less susceptible to the direct effects of UVB, but more likely to be impacted through UVB-induced changes at lower trophic levels. Assessing changes caused by ozone depletion is a challenge as minimal baseline data (pre-1970s) relating to UVB effects are available.

The issue of variable species responses to UVB is more than expecting a shift in the taxonomic structure of communities. For example, in unicellular organisms, size can play a key role in UV tolerance. Smaller cells tend to be more sensitive than larger cells. In aquatic environments, many organisms are size-selective filter feeders. Restructuring the size distribution of microorganisms could have ramifications at all trophic levels and include functional aspects such as reorganization of niches, alterations in trophic transfer, and changes in pathways of biogeochemical cycling.

Trophic Dynamics

Primary productivity

The majority of ecological research on ozone depletion has focused on primary production in both aquatic and terrestrial systems. Numerous studies have demonstrated enhanced photosynthesis in algae and vascular plants when UVB wavelengths are excluded in experimental exposures. Conversely, reduced photosynthesis is observed in laboratory and mesocosm studies when UVB is enhanced. The results from field and lab clearly indicate that ambient levels of UVB limit global primary productivity and justify concerns that increasing UVB levels would lead to declines in primary production.

UVB limitation on photosynthesis can be manifested in a number of ways. Phytoplankton cells may maintain a constant cell size and divide less frequently, or maintain a constant division rate and produce smaller cells. Vascular plants can grow as tall, but have smaller leaves. The impact of UVB on primary productivity in terrestrial environments is often positively correlated to rainfall, with significant differences in the same area between wet and dry years. Secondary effects of UVB exposure can also be important. These include increased synthesis of UV-absorbing compounds, thickening of leaf cuticles and epidermal tissues, changes in morphology, and increased susceptibility to disease. Some of these physiological and structural alterations can result in reduced palatability and nutritional value, directly affecting the transfer of energy to higher trophic levels.

Consumers

Smaller consumer species may suffer from the direct effects of UVB and exhibit stunted growth, reduced reproduction, and increased fatalities. This is true in aquatic systems where some microheterotrophs are more sensitive to UVB exposure than their photosynthetic food source. Large consumers often have adequate size and external protective layers so that the direct effects of UV exposure are minimal with molecular damage limited to surface cells. In these organisms, deleterious UVB effects are more likely translated through the food web where lower trophic levels are negatively impacted.

Biogeochemical Cycles

UVB impacts on the biological component of carbon and nutrient cycles in ecosystems occur through alterations in productivity and other physiological processes of primary producers, consumers, and decomposers. Direct photochemical reactions with the physical environment also occur. Photodegradation of organic and inorganic components of soils (e.g., leaf litter) and dissolved and particulate matter in aquatic systems is an important aspect of chemical cycling in the environment. As with biological effects, UVB is already an important aspect of these processes and the potential impact of increased levels of UVB is not fully understood.

Future Outlook

Throughout the history of life on Earth, UVB has been an environmental hazard for organisms, and ozone depletion occurring from the late 1970s to the late 1990s may have exacerbated already stressful conditions. The current stabilization and gradual reduction in the concentration of synthetic ozone-depleting substances should result in restoration of the ozone column to pre-1980 concentrations within the next 100 years. However, UVB-induced ecological changes that have occurred or will continue to occur in the coming decades are irreversible. Although some dire predictions were made in the past, there is no evidence of any ecosystem collapse, even in Antarctica where ozone depletion annually continues to exceed 50% and declines in primary productivity have been measured at up to 12%. Ecosystem modification has certainly taken place, but the long-term ramifications are not known.

There are indications that global warming could enhance ozone depletion even in the absence of anthropogenic pollutants, so the fate of the ozone layer is still uncertain. However, even without ozone depletion (i.e., with a normal ozone column), the amount of UVB passing through the atmosphere is sufficient to have measurable negative impacts on organisms and understanding the role of UVB in biosphere continues to be an important ecological issue.

Further Reading

Lobell DB and Gourdji SM (2012) The influence of climate change on global crop productivity. *Plant Physiology* 160: 1686–1697 http://www.plantphysiology.org/content/160/4/1686.full.pdf+html.

McKenzie RL, Aucamp PJ, Bais AF, Björn LO, Ilyas M, and Madronich S (2011) Ozone depletion and climate change: Impacts on UV radiation. *Photochemical and Photobiological Sciences* 10: 182–198.

Paul ND and Gwynn-Jones D (2002) Ecological roles of solar UV radiation: Toward an integrated approach. *Trends in Ecology and Evolution* 18: 48–55.

World Meteorological Organization (WMO), 2011. Scientific assessment of ozone depletion: 2010. Geneva, Switzerland: Global Ozone Research and Monitoring Project-Report No. 52.

Young AR, Björn LO, Moan J, and Nultsch W (1993) *Environmental UV photobiology*. New York: Plenum Press.

Relevant Websites

International Panel on Climate Change (IPCC) http://www.ipcc.ch.

United Nations Environmental Programme (UNEP) http://www.unep.org.

World Meterological Organization (WMO) http://www.wmo.int.

Political Ecology

Tor A Benjaminsen, Norwegian University of Life Sciences, Ås, Norway
Hanne Svarstad, Oslo Metropolitan University, Oslo, Norway

The Emergence of Political Ecology

Political ecology is a field within socio–environmental studies with a core focus on power relations in environmental governance as well as the coproduction of nature and society within a wider political economy (Robbins, 2012). The field has gained momentum during the last couple of decades especially within Anglo-American geography, but also in anthropology, development studies, and environmental history. This momentum is witnessed through a rapid increase in university courses as well as in academic publications internationally. Power is studied as contestations over material assets (land, natural resources) as well as over meaning. The latter leads to studies of how various actor groups produce social constructions such as discourses and narratives, which again have implications for the distribution and control over resources.

Building on critical theory, scholars in political ecology try to combine a focus on values with empirical transparency and theoretical development. In this way, research in political ecology has contributed to new perspectives on both people-nature and science-policy linkages.

Political ecology emerged from the 1970s as a result of two confluent trends. First, the field developed as a Marxist critique of Malthusian ideas in environmental thinking. The argument was that the population centered scholarship by ecologists, such as Ehrlich (1968) and Hardin (1968), were inherently political, and that studies of human ecology are never neutral or apolitical, but involve interests, norms, and power. While Marxist critics tended to accept the environmental impacts of human production described by Neo-Malthusians, they pointed to the inherent lack of social and political analysis in such studies, arguing that Malthusian thinking invariably leads to policies of "blaming the victims."

For example, Enzensberger (1974) pointed out that ecologists and other natural scientists may claim to be "objective" and "apolitical," but they become political actors when engaging in environmental debates, because they inform political choices creating winners and losers, and because their analyses, questions, and categories are inevitably informed by normative assumptions. Hence, there is an issue of environmental justice that often escapes ecologists whose normativity, generally associated with biocentrism, may impact on marginalized people's lives.

The presumed neutrality of ecology as a science when entering environmental debates is therefore illusory. While Enzensberger referred to natural science with apolitical pretensions (but with political implications) as "political ecology," political ecologists have later labeled such practices "apolitical ecology" (Robbins, 2012). This distinction refers to the difference between fields that admit and openly engage with their inevitable, normative assumptions (political ecologies) and those that do not (apolitical ecologies) (Robbins, 2012).

The second trend that contributed to the emergence of political ecology was the evolution of human ecology and cultural ecology. Anthropologists in these fields, such as Nietschmann (1973) and Rappaport (1968), had long employed ecological methods to explain human behavior. However, as the communities they studied were impacted by national governments as well as global and national markets, it became increasingly clear that the explanatory power of their ecological methods was limited by not being able to include the state or markets in the analysis. Researchers in this field had hit a "conceptual wall," which led some cultural ecologists to seek more powerful conceptual and theoretical tools, especially tools from political economy (Robbins, 2012).

From the late 1980s, a second phase in the short history of political ecology started, drawing on a wide range of theoretical and methodological resources. Piers Blaikie's book, *The Political Economy of Soil Erosion in Developing Countries* (Blaikie, 1985) paved the way for an approach to political ecology that employs the lens of political economy, while explicitly engaging with rigorous natural science. The book provides a critique of environmental conservation policies in the Global South and presents three central arguments (Neumann, 2008).

First, there is often lack of sound scientific data on soil erosion and other environmental processes, which leads to a high level of uncertainty. Second, actors have varying perceptions of environmental change depending on their "ideology." Blaikie (1985, 149) argues "that all approaches to soil erosion and conservation are ideological—they are underpinned by a definite set of assumptions, both normative and empirical, about social change." Third, environmental policies always hold implications for control over resources and rights to land. A critical question to asks would therefore be: Who wins and who loses from resource and conservation

policies? This leads to the study of "where power lies and how it is used" (Blaikie, 1985, 6). Blaikie proposed an approach to understanding environmental problems by, on the one hand, problematizing the quality and uncertainty of scientific data and, on the other hand, insisting that the production, interpretation and use of environmental data are inherently political. A process such as soil erosion could therefore only be fully understood with the help of the tools of political economy (Rigg, 2006).

These ideas are further discussed in Blaikie's next book, coauthored with Harold Brookfield, *Land Degradation and Society* (Blaikie and Brookfield, 1987), in which "land degradation" is presented as a perceptual term, stressing that environmental changes are perceived in differing ways by the various actors involved. Hence, "degradation" is not simply a process that can be measured with natural science methods, but instead one in which environmental processes interact with human perception, biases, and interests. Whether processes such as deforestation or soil erosion are perceived as "degradation" depends on the position of observers engaged in inevitably political contests over what should be done with land and over the authority to control land change outcomes.

Our understanding and interpretation of environmental change was thus seen as guided by our norms, interests, and values. However, while environmental data are constructed and subject to ideological interpretations, Blaikie and Brookfield (1987, 16) still insisted on the necessity of improving scientific techniques of measurement in order to obtain "those data which are beset with least uncertainty".

From the mid-1990s, political ecology evolved further to reflect poststructural influences (e.g., Peet and Watts, 1996), in which the norms, interests, and values governing human understanding of environments and environmental change are themselves the product of political processes that determine control over what ideas are taken-for-granted or seen as "true." This perspective was notably brought to bear on a range of critical environmental issues in the Global South, including natural resource scarcity, overpopulation, soil degradation, and the notion of carrying capacity. Research in political ecology sought not only to show that such concepts were problematic and inapplicable, but also sought to explain how, despite their imprecision, they became assumed to be true. Political ecology in this vein is typically critical of received wisdom, especially as dominant and powerful views, narratives and ideas often support dominant and powerful interests.

While political ecology in its inception focused predominantly on the governance of renewable natural resources in the rural Global South, political ecology has later expanded to industrialized nations in the Global North, to urban areas, and to include nonrenewable resources.

However, while the above history may relate to developments within Anglo-American political ecology, it is important not to overlook the southern European tradition of political ecology (*écologie politique, ecología política*) that has developed closer connections with fields such as green politics, environmental justice and the emerging agenda of "degrowth" (e.g., D'Alisa et al., 2014).

A Critical Approach

Political ecology is said to be a "critical approach," which usually refers to its questioning of the role and status of powerful actors as well as of what is taken for granted in leading discourses on environment and development. But political ecologists are also engaged academics committed to improvements to more just and sustainable societies. In this way, political ecology can be said to trace its roots back to the critical theory of the Frankfurt school. Max Horkheimer, one of its founders, argued that traditional theory only attempts to understand and explain certain aspects of society, while critical theory also has a liberating aspect as it indicates the components of society that should be changed, making a broadly multidisciplinary approach necessary (Horkheimer, 1970). In the 1960s, Jürgen Habermas, another representative of this school, put forward a critique of positivism in which he suggested that traditional science claims objectivity on a false basis. The alternative proposed by Habermas is a critical theory of science in which natural, human and social sciences are related to specific epistemological interests that are respectively technical, practical and liberating (Habermas, 1968).

Although political ecology has to date made an effective contribution to showing what people should be freed from (see *Liberation Ecologies* by Peet and Watts, 1996), it is nonetheless limited in regard to determining what this liberation should lead to. Following up on this insight, Robbins (2012) argues that political ecology has two faces, and two missions—"the hatchet" and "the seed." The hatchet side is dominant and represents political ecology's critical approach, whereas the seed aspect forms the contribution of political ecology to a world in which development is fairer and more sustainable. Walker (2006), however, criticizes political ecologists for largely neglecting the seed function. He argues that most of the writings produced in political ecology are internal and aimed at academics working in the same field. The relations between political ecology and politics feature not only apathy, he continues, but also antipathy; this is in fact mutual antipathy between those who have power and those who practise political ecology.

The southern European tradition of political ecology has, however, a longer history of engaging with the seed function with its links to environmental justice movements as well as to activists for alternative sustainabilities such as "degrowth." The Anglo-American tradition, however, remains to some extent torn between on the one hand deconstructions of environmental narratives and claims presented by conservationists and how these narratives may lead to injustice and marginalization, and on the other hand contributing to alternative sustainabilities (Cavanagh and Benjaminsen, 2017). This is, however, an on-going debate within the field. There is also recently increased communication and interaction between these two main branches of political ecology.

Some Key Theoretical Influences

Political ecology can be said to be an eclectic approach taking inspiration from a wide variety of sources. There are, however, some key theoretical influences in the field. We will here only mention three of these; political economy following Karl Marx, poststructuralism inspired by Michel Foucault, and peasant studies with James Scott as one of the key contributors.

Marx has already been mentioned as an essential source of inspiration in the early days of political ecology, and more recently there seems to be a revival of Marxist thinking in this field. A large number of political ecology studies have documented how degradation narratives serve to justify elite capture and the dispossession of marginalized people from land and natural resources. The result may be seen as another example of "primitive accumulation," which Marx saw as a historical process of divorcing the producer from the means of production.

According to Harvey (2003: 149), "primitive accumulation as Marx described it . . . entailed taking land, say, enclosing it, and expelling a resident population to create a landless proletariat, and then releasing the land into the privatized mainstream of capital accumulation." Since accumulation is an on-going process, Harvey (2003) proposes the term "accumulation by dispossession" to describe current processes. The introduction of this term has sparked a renewed interest in the combination of dispossession and capital accumulation in development studies and in political ecology in particular.

For instance, Benjaminsen and Bryceson (2012) use the lens of accumulation by dispossession to analyze enclosures in wildlife and coastal conservation in Tanzania. They show how recently established conservation initiatives steadily lead to local people's loss of access to land and natural resources. Dispossession has been gradual and piece-meal in some cases, while it involved violence in other cases, but does not primarily take the usual form of privatization of land. The spaces involved are still formally state or village land. It is rather the benefits from the land and natural resources that contribute to capital accumulation by more powerful actors (rent-seeking state officials, transnational conservation organizations, tourism companies, and the state Treasury). In both wildlife and coastal management, restrictions on local resource use are justified by degradation narratives, while financial benefits from tourism are drained from local communities within a system lacking in transparent information sharing.

Michel Foucault has demonstrated the importance of analysing how dominant discourses establish what is generally accepted as "truths" in a society. In political ecology, a number of scholars inspired by Foucault have questioned what is taken for granted or seen as true on issues of environment and development (e.g., Adger et al., 2001).

The Foucauldian conception of "governmentality" is also applied within political ecology. This concept denotes the techniques and tactics of government, which again implies that the governing of citizens involves the use of certain techniques to implement certain ways of thinking. The "governmentality" concept that appeared in Foucault's last works is a logical follow-up to his thinking on the relations between power and knowledge. For Foucault, power and knowledge are closely inter-connected.

As an example of a political ecology application of governmentality, Fletcher (2010) outlines four distinct environmental governmentalities ("environmentalities") that are played out in current environmental governance: *Neoliberal governmentality* (commodification and increased use of the market in environmental governance); *Disciplinary governmentality* (efforts to influence individuals—create environmental subjects—through diffusion of ethical norms); *Sovereign governmentality* (the use of force and threats of using force); and *Truth governmentality* (the art of government according to truth). It should be noted, however, that these four forms of governmentality are not mutually exclusive. They may coexist and work together or in opposition to each other.

James Scott is a key representative of what is called "peasant studies" in which the rationality of small-scale farmers, pastoralists or other marginalized groups is analyzed, for example the reasons why they often resist modernization. This type of peasant studies often leads to a critique of the commonly held idea that small-scale farmers or pastoralists are irrational actors. Scott (1976) observes that peasants tend to try to prevent risks by developing social redistribution systems for surpluses in good years in order to protect themselves against the effects of bad years. This can be in the form of the sharing of land and labor with others. Scott (1985) describes everyday forms of peasant resistance to modernization and exploitation holding that this opposition to interventions from the outside world is much more global and has greater scope than armed revolt, even if the latter is more discussed.

Hence, those who are to be governed are variously able to ignore, avoid, fight, transform or reclaim the intervention in question through tactics of noncompliance and everyday acts of resistance, which require little or no coordination or planning and may include "foot dragging, dissimulation, false compliance, pilfering, feigned ignorance, slander, arson, sabotage, and so forth" (Scott, 1985: 29). These are the "weapons of the weak"; the "weak" being subjects who use a variety of "weapons" to defend their interests against superordinate groups. Such everyday resistance typically avoids direct confrontation with the authorities and does not make headlines; it is an informal form of resistance, often covert and concerned with immediate gains. Scott explains that in some cases resistance can be more effective when hidden than when open, because cognizance of such activities may entail a rapid and ferocious response from the superordinate group. Though the acts of hidden resistance do not openly "contest the formal definitions of hierarchy and power" (Scott, 1985: 33), it is possible to determine to what degree, and in what ways, subordinate groups accept the social order (and structure of domination) propagated by elites by studying the subordinates"behaviour and "offstage" comments and conversations (Scott, 1985: 41).

These "offstage" presentations, or "hidden transcripts" as Scott (1990) calls them, are accounts that the subjects communicate in the absence of the powerful. The hidden transcripts include both the subjects' critique of power and practices and claims that the dominating actor would not acknowledge openly. The concept of hidden transcripts also includes accounts that are expressed openly, but disguised in the form of rumors, proverbs, jokes, parodies, gossip, gestures, folktales, and so on. "Public transcripts," on the other hand, are comments and conversations that the actors (the dominant and the subjects) present in each other's presence.

While public transcripts can inform us about power, they are "unlikely to tell the whole story about power relations" (Scott, 1990: 2).

While local and indigenous practices may represent complex and "messy" realities, implementations of governmentality can be seen as attempts to simplify and standardize landscapes and practices and to make both society and territory "legible" (Scott, 1998). For example, in the art of governing smallholder land-use, the state first needs to establish a serious problem that its policy will solve. This will often take the form of claims of environmental degradation or economic inefficiency. Such techniques of governmentality and their continuing power and traction in contemporary policy making are widely highlighted in contemporary political ecology. Second, the state may need to claim that this problem can only be solved through scientific and technical means. These techniques were described by Foucault and further developed by Li (2007) who calls these two steps "problematization" and "rendering technical," wherein the deployment of "scientific" or "expert" reasoning plays a key role. Scientists and scholars may, however, also assist in questioning such processes, not least through the application of the tools and insights of critical political ecology.

Where Is the "Ecology" in Political Ecology?

Political ecology has been criticized from the outside for being founded on a priori judgments by giving priority at the outset to political explanations (Vayda and Walters, 1999) and for overlooking ecological dynamics (Peterson, 2000). But also within political ecology itself, there has been a debate about the place and role of ecology. In a key contribution to this debate, Walker (2005) asks where the ecology is in political ecology and whether the field has become "politics without ecology."

This critique may for instance be seen as relevant for some of the Marxist inspired contributions to political ecology. This literature reflects in some way Marx's idea that ". . . all progress in capitalist agriculture is a progress in the art, not only of robbing the worker, but of robbing the soil" (Marx, 1990: 638). Hence, market integration and the expansion of capitalism would tend also to lead to environmental degradation, according to this thinking.

In a seminal contribution to this literature, Watts (1983) studying small-scale farming in northern Nigeria found that commodification caused starvation and economic marginalization among peasants. Increasingly dependent on an unstable market, they became more vulnerable, and had to take up loans and generally take more risks. Previously self-sufficient, peasants gradually became underpaid farm workers. This in turn led to decreasing investments of labor on their own land, resulting in the degradation of soils on land where food crops were grown.

However, from the late 1980s, a number of students and scholars who were inspired by the research agenda proposed by Blaikie (1985) and Blaikie and Brookfield (1987) carried out empirical studies in the global South unpacking the "ecology" in the political ecology equation. This implied extending the focus on peasant rationality and agency within peasant studies and cultural ecology to environmental dynamics. Many of these studies focused on Africa and generated new knowledge and critiques of environmental orthodoxies in several fields.

An example of this was the edited book entitled *The Lie of the Land: Challenging Received Wisdom on the African Environment* by Leach and Mearns (1996). This was a collection of key critical contributions on various environmental issues in Africa (e.g., range ecology, desertification, deforestation, biodiversity conservation, and soil erosion). A series of chapters challenged received wisdom on these issues and reflected a broader literature that had emerged during the late 1980s and early 1990s. Henceforth, a large number of case studies from different parts of the African continent and on various environmental issues have continued to question dominant (often Neo-Malthusian) narratives on environmental degradation through carefully collected environmental data.

Blaikie (1999), however, pointed out that political ecological critiques of claims of "degradation" in fact owe more to realist science than to postmodern deconstruction. Hence, critical political ecology has to a large extent been based on realist investigations of environmental change to construct counter-narratives or alternative narratives to those dominating policies or academic debates. A "critical political ecology" would critically and empirically examine all environmental representations whether based on Malthusianism or a critique of capitalism. This also implies investigating rather than assuming "the essentialist link between capitalism and environmental degradation" that one often finds in the development literature (Forsyth, 2003).

Critical political ecologies would combine deconstructions of narratives with a realist belief in science as a means to achieve more accurate descriptions and understandings of environmental realities. Such combinations of realist and constructivist positions are referred to in political ecology as a critical realist position. According to Forsyth (2001), critical realism seeks to understand ecological change through a combination of epistemological skepticism and ontological realism.

Michael Watts has, however, also criticized some political ecology for paying too close attention to natural aspects ("ecology") and too little attention to "political" aspects (Peet and Watts, 1996; Watts, 1997). Watts holds that this leads to an atheoretical approach, which lacks a general theory of social change that would explain environmental degradation. This debate reflects a tension within political ecology between an approach engaging actively with natural science and ecology and one focusing on social theory and "politics."

Feminist Political Ecology

Within political ecology and particularly the subfield of feminist political ecology (FPE), gender is emphasized as one of the central factors in questions of power, access and benefits associated with natural resources. In specific contexts, scholars examine gender together with other factors such as class, race, ethnicity, and age. So far, two seminal volumes have been published in FPE. In 1996, Dianne Rocheleau, Barbara Thomas-Slayter and Esther Wangari edited *Feminist Political Ecology: Global Issues and Local Experiences*, and in 2015, the book *Practising Feminist Political Ecologies: Moving Beyond the "Green Economy"* was edited by Wendy Harcourt and Ingrid L. Nelson (2015). As in political ecology as such, FPE is characterized by a plurality of approaches and openness to incorporation of new thoughts. Rocheleau (2015) states: "FPE is more about a feminist perspective and an ongoing exploration and construction of a network of learners than a fixed approach to a single focus on women and/or gender." Elmhirst (2015) argues that FPE has gained from a long string of feminist traditions, starting with gender and development studies, and then encompassing feminist science studies as well as recent poststructuralist, posthumanist and postcapitalist feminist theory.

The ecofeminism of Shiva (1988) and others is a field that has contributed to bringing gender into thinking about environment and development, and it can be seen as an important source of inspiration and predecessor of FPE. Many contributors to FPE, however, follow the criticism that Jackson (1993) and Braidotti et al. (1994) and others have posed at ecofeminism as entailing biological essentialism instead of an empirical openness and examination of ways that gender might play out in specific contexts. This criticism also highlights gender stereotyping where women due to biology are seen as more environmental friendly than men.

Through several decades, Rocheleau (2015) has conducted seminal FPE work based on case studies from different parts of the world and especially in Kenya. She has shown how landscapes and livelihoods are gendered with mosaics of various responsibilities, labor and control of resource, processes and/or products. Among other contributions are Leach (1994) on Sierra Leone, Schroeder (1999) on the Gambia, Paulson (2005) on Bolivia and Nightingale (2011) on Nepal.

In the 2015 volume on FPE, several contributors address the current hegemonies of marketizations and neoliberalization of nature. Drawing on a combination of FPE and feminist political economy, Wichterich (2015) criticizes initiatives for gender equality in neoliberal climate policies such as carbon trading and the clean development mechanism (e.g., "women's carbon standard"—www.womenscarbonstandard.org). She warns about neoliberal empowerment of women that provides uncertain and limited influence or economic returns, and at the same time legitimizes the use of the global South as carbon sink and the continued consumerism of the global North and global middle classes.

As in political ecology in general, most contributions to FPE so far tend to come from case studies in the global South. Nevertheless, the approaches and issues are also relevant in the North, as some studies from Norway demonstrate. Although this country usually has high scores on gender equality indicators, gender equality has recently been left out in important issues of conservation and environmental management. This is not because women have chosen themselves not to get involved in these policy issues. Instead, a broad range of actors at various scales have been involved in setting aside laws that otherwise apply on gender equality in politics, and this again is due to power struggles and alliance building in a field that traditionally is controlled by men (Svarstad et al., 2006, 2009; Benjaminsen and Svarstad, 2017). In an examination of gender aspects in a new governance system for protected areas in Norway, Lundberg (2017) shows that legal requirements for gender equality in Norway have finally been implemented, but in a way that still hides that few local women take part in conservation boards.

Conclusions

Political ecology is a field in environmental studies focusing on power relations that has gained momentum during the last couple of decades. Power is studied as contestations over material assets (land, natural resources) as well as over meaning. The latter leads to studies of social constructions, which again have implications for the distribution and control over resources.

From the 1970s, political ecology emerged as a Marxist critique of Malthusianism and as a further development of a human and cultural ecology approach including also the impact of states and markets in the analysis of human-environment interactions. From the 1980s, this approach further focused on how norms, interests and values form our interpretation of environmental change. These norms, interests and values are again the product of political processes that determine control over what ideas are taken-for-granted or seen as "true."

Political ecology is an eclectic approach that has been taking inspiration from a wide variety of sources. Key theoretical influences discussed here are political economy following Karl Marx, poststructuralism inspired by Michel Foucault, and peasant studies with James Scott as one of the key contributors. An on-going discussion in the field is also the place and role of "ecology" within the political ecology equation. Some contributions within the field engage actively with natural science, while other parts of this literature remain within more social science-based theoretical debates where "ecology" refers to the environment more broadly. Finally, this article discusses feminist political ecology as a subfield reflecting political ecology's eclecticism where the role of gender and power in relation to access to and benefits from natural resource governance is discussed, again taking inspiration from a variety of scholarly directions and theories.

In the wake of recent debates about "posttruth" and increasing political pressure on science, especially related to global climate change, a future key tension and point of discussion within political ecology will be how to balance a continued critique of the production of environmental science promoting certain biocentric values with a belief in science as the best bulwark against

political gerrymandering. This conundrum may best be solved through adhering to a critical realist approach combining deconstructions with a realist reliance on science as a means to achieve the most accurate descriptions and understandings of environmental realities. Such a critical realism combines a critical scrutiny of available "truths" with continuously seeking to identify the best available science, which is quite similar to how science should work under normal circumstances, if left without political interference.

References

Adger WN, Benjaminsen TA, Brown K, and Svarstad H (2001) Advancing a political ecology of global environmental discourses. *Development and Change* 32(4): 681–715.

Benjaminsen TA and Bryceson I (2012) Conservation, green/blue grabbing and accumulation by dispossession in Tanzania. *Journal of Peasant Studies* 39(2): 335–355.

Benjaminsen TA and Svarstad H (2017) *Politisk økologi: Mennesker, miljø og makt. Oslo: Universitetsforlaget.* Second revised edition.

Blaikie P (1985) *The political economy of soil erosion in developing countries*. New York: Longman.

Blaikie P (1999) A review of political ecology: Issues, epistemology, and analytical narratives. *Zeitschrift für Wirtshcaftsgeographie* 43(3–4): 131–147.

Blaikie P and Brookfield H (eds.) (1987) *Land degradation and society*. London & New York: Methuen.

Braidotti R, Charkiewicz E, Häusler S, and Wieringa S (eds.) (1994) *Women, the environment and sustainable development: Towards a theoretical synthesis*. London: Zed Books.

Cavanagh C and Benjaminsen TA (2017) Political ecology, variegated green economies, and the foreclosure of alternative sustainabilities. *Journal of Political Ecology* 24: 200–216.

D'Alisa G, Demaria F, and Kallis G (eds.) (2014) *Degrowth: A vocabulary for a new era*. London: Routledge.

Ehrlich P (1968) *The population bomb*. New York: Ballantine Books.

Elmhirst R (2015) Feminist political ecology. Chapter 40 pp. 519–530. In: Perreault T, Bridge G, and McCarthy M (eds.) *The Routledge Handbook of Political Ecology*. London and New York: Routledge.

Enzensberger HM (1974) A critique of political ecology. *New Left Review* 8: 3–32.

Fletcher R (2010) Neoliberal environmentality: Towards a poststructuralist political ecology of the conservation debate. *Conservation and Society* 8(3): 171–181.

Forsyth T (2001) Critical realism and political ecology. In: Stainer A and Lopez G (eds.) *After Postmodernism: Critical Realism?*, pp. 146–154. London: Athlone Press.

Forsyth T (2003) *Critical political ecology. The politics of environmental science*. London: Routledge.

Habermas J (1968) *Technik und Wissenschaft als Ildeologie*. Frankfurt am Main: Suhrkamp Vellag.

Harcourt W and Nelson IL (eds.) (2015) *Practising feminist political ecologies: Moving beyond the 'green economy'*. London: Zed Books.

Hardin G (1968) The tragedy of the commons. *Science* 162: 1243–1248.

Harvey D (2003) *The new imperialism*. Oxford: Oxford University Press.

Horkheimer M (1970) *Traditionelle und kritische Theorie: vier Aufsätze*. Frankfurt am Main: Fisscher Bycherei.

Jackson C (1993) Women/nature or gender/history—A critique of ecofeminist development. *Journal of Peasant Studies* 20(3): 389–419.

Leach M (1994) *Rainforest relations: Gender and resource use by the Mende of Gola, Sierra Leone*. Edinburgh: Edinburgh University Press.

Leach M and Mearns R (eds.) (1996) *The lie of the land: Challenging received wisdom on the African environment*. Oxford: James Currey.

Li TM (2007) The will to improve. In: *Governmentality, development, and the practice of politics*. Durham: Duke University Press.

Lundberg AKA (2017) *Handling legitimacy challenges in conservation management: Case studies of collaborative governance in Norway*. PhD thesis, Norwegian University of Life Sciences.

Marx K (1990) Capital: A critique of political economy, *Volume 1*. New York: Penguin.

Neumann RP (2008) Probing the (in)compatibilities of social theory and policy relevance in Piers Blaikie's political ecology. *Geoforum* 39(2): 708–715.

Nietschmann B (1973) *Between land and water*. New York: Seminar Press.

Nightingale A (2011) Bounding difference: Intersectionality and the material production of gender, caste, class and environment in Nepal. *Geoforum* 42: 153–162.

Paulson S (2005) Gendered practices and landscapes in the Andes. The shape of asymmetrical exchanges. In: Paulson S and Gezon L (eds.) *Political ecology across spaces, scales and social groups*. New Jersey: Rutgers University Press.

Peet R and Watts M (eds.) (1996) *Liberation ecologies. Environment, development, social movements*. London: Routledge.

Peterson G (2000) Political ecology and ecological resilience: An integration of human and ecological dynamics. *Ecological Economics* 35: 323–336.

Rappaport RA (1968) *Pigs for the ancestors: Ritual in the ecology of a new Guinea people*. New Haven, CT: Yale University Press.

Rigg J (2006) Piers Blaikie (1942-). In: Simon D (ed.) *Fifty key thinkers on development*. London: Routledge.

Robbins P (2012) *Political ecology*. Oxford: Blackwell.

Rocheleau D (2015) A situated view of feminist political ecology from my networks, roots and territories. In: Harcourt W and Nelson IL (eds.) *Practising feminist political ecologies: Moving beyond the 'green economy'*. London: Zed Books.

Rocheleau D, Thomas-Slayter B, and Wangari E (eds.) (1996) *Feminist political ecology*. London: Routledge.

Schroeder R (1999) *Shady practices: Agroforestry and gender politics in the Gambia*. Berkeley: University of California Press.

Scott JC (1976) *The moral economy of the peasants: Rebellion and subsistence in Southeast Asia*. New Haven: Yale University Press.

Scott JC (1985) *The weapons of the weak: Everyday forms of peasant resistance*. New Haven: Yale University Press.

Scott JC (1990) *Domination and the arts of resistance: Hidden transcripts*. New Haven, Connecticut: Yale University Press.

Scott JC (1998) *Seeing like a state: How certain schemes to improve the human condition have failed*. New Haven: Yale University Press.

Shiva V (1988) *Staying alive*. London: Zed Books.

Svarstad H, Daugstad K, Vistad OI, and Guldvik I (2006) New protected areas in Norway: Local participation without gender equality. *Mountain Research and Development* 26(1): 48–54.

Svarstad H, Skuland S, Guldvik I, and Figari H (2009) *The lack of gender equality in local natural resources management in Norway. The National Park Plan as example. Rapport* 432, Oslo: NINA.

Vayda AP and Walters B (1999) Against political ecology. *Human Ecology* 27(1): 167–179.

Walker PA (2005) Political ecology: Where is the ecology? *Progress in Human Geography* 29(1): 73–82.

Walker PA (2006) Political ecology: Where is the policy? *Progress in Human Geography* 30(3): 382–395.

Watts M (1983) *Silent violence: Food, famine and peasantry in northern Nigeria*. Berkeley: University of California Press.

Watts M (1997) Classics in human geography revisited. *Progress in Human Geography* 21(1): 75–80.

Wichterich C (2015) Contesting green growth, connecting care, commons and enough. In: Harcourt W and Nelson IL (eds.) *Practising feminist political ecologies: Moving beyond the 'green economy'*. London: Zed Books.

Precaution and Ecological Risk

O Renn, University of Stuttgart, Stuttgart, Germany

The Different Meanings of Precaution

This article presents a scientific overview of the meanings and applications of the precautionary principle in ecological risk assessment and management. The term risk is understood in this document as an uncertain consequence of an event or an activity with respect to something that humans value. Risks always refer to a combination of two components: the likelihood or chance of potential consequences and the severity of consequences of human activities, natural events, or a combination of both. Such consequences can be positive or negative, depending on the values that people associate with them. In addition to the strength and likelihood of these consequences, characterizing risks includes contextual aspects such as the distribution of risks over time, space, and populations. With respect to ecology, risk denotes the probability of ecosystem damage as a result of human interventions or natural events (such as earthquakes, wildfires, or flooding).

The focus on ecological risk should be seen as a segment of a larger and wider perspective on how humans transform the natural into a cultural environment with the aims of improving living conditions and serving human wants and needs. These transformations are performed with a purpose in mind (normally a benefit to those who initiate them). When implementing these changes, intended (or tolerated) and unintended consequences may occur that meet or violate other dimensions of what humans value. These are the risks. It is the major task of risk assessment to identify and explore, preferably in quantitative terms, the types, intensities, and likelihood of the (normally undesired) consequences related to the consequences that human actions or events exert on ecosystems. In addition, these consequences are associated with special concerns that individuals, social groups, or different cultures may associate with these risks. Ecosystem changes can be physically measured or observed but they only get meaning through interpretation by humans. Different individuals, different groups, and different cultures have different criteria and perspectives of what kind of physical changes matter to them and how they are valued. These social processes of attributing value and meaning to consequences of actions and events also need to be assessed and included in the risk evaluation for making prudent judgements about the tolerability or acceptability of risks.

The Precautionary Principle

Once that judgement is made, it is the task of risk management to prevent, reduce, or alter these consequences by choosing appropriate actions. Risk management is always confronted with the following questions: How much transformation of ecosystems is tolerable? Where do societies set the boundary between acceptable and nonacceptable risks? Such a fundamental judgment cannot be justified by an all-encompassing substantive rule. Rather, one needs procedural principles that help societies to make tolerability judgments and to select risk management options. One of these rules refer to the precautionary principle. This principle has been formulated in many different ways in many different places (one root may be the `foresight principle' adopted to German environmental law in the 1960s). The most widely used formulation of the principle can be found in the Rio Declaration:

> In order to protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. Where there are threats of serious of irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation. (Rio Declaration 1992, Principle 15)

This articulation of the principle allows risk managers to initiate regulatory actions even if the scientific evidence about risks is not yet conclusive. But how much evidence is evidence enough to trigger such actions? Is the precautionary principle an instrument to ban all new activities because one can always find claims for serious harm regardless of the `true' nature of the intervention? How accountable are risk managers if they can (freely or even arbitrarily) decide what evidence they will accept as sufficient to invoke the precautionary principle?

Despite the intensity of attention that the principle has received in the policy arenas, there remain a number of serious ambiguities and queries concerning the nature and appropriate role of the precautionary principle in governance. These are addressed – if not resolved – in a burgeoning academic and more policy-oriented literature. In order to understand the principle in the context of a larger array of risk management strategies, it is necessary to look more deeply into the fabrics of ecological risks. This is done in the next section.

Components of Risk

Risks are mental `constructions'. They are not real phenomena but originate in the human mind. Actors, however, creatively arrange and reassemble signals that they get from the `real world' providing structure and guidance to an ongoing process of reality

enactment. The status of risk as a mental construct has major implications on how risk is looked at. Unlike trees or houses, one cannot scan the environment, identify the objects of interest, and count them. Risks are created and selected by human actors. What counts as a risk to someone may be an act of God to someone else or even an opportunity for a third party. Since risks represent mental constructs, the quality of their explanatory power depends on the accuracy and validity of their (real) predictions. Unlike some other scientific constructs, validating the results of risk assessments is particularly difficult because, in theory, one would need to wait indefinitely to prove that the probabilities assigned to a specific outcome were correctly assessed. If the number of predicted events is frequent and the causal chain obvious (as is the case with car accidents), validation is relatively simple and straightforward. If, however, the assessment focuses on risks where cause–effect relationships are difficult to discern, effects are rare and difficult to interpret, and variations in both causes and effects are obscuring the results (as is often the case with ecological risks), the validation of the assessment results becomes a major problem. In such instances, assessment procedures are needed to characterize the existing knowledge with respect to complexity, remaining uncertainties, and ambiguities. What do these terms mean?

1. *Complexity*. It refers to the difficulty of identifying and quantifying causal links between a multitude of potential causal agents and specific observed effects. The nature of this difficulty may be traced back to interactive effects among these agents (synergism and antagonisms), long delay periods between cause and effect, interindividual variation, intervening variables, and others.
2. *Uncertainty*. It is different from complexity but often results from an incomplete or inadequate reduction of complexity in modeling cause–effect chains. Whether the world is inherently uncertain is a philosophical question that we will not pursue here. It is essential to acknowledge in the context of risk assessment that human knowledge is always incomplete and selective and thus contingent on uncertain assumptions, assertions, and predictions. It is obvious that the modeled probability distributions within a numerical relational system can only represent an approximation of the empirical relational system with which to understand and predict uncertain events. It therefore seems prudent to include other, additional, aspects of uncertainty. Although there is no consensus in the literature on the best means of disaggregating uncertainties, the following categories appear to be an appropriate means of distinguishing the key components of uncertainty:
   - target variability (based on different vulnerability of targets such as ecosystems);
   - systematic and random error in modeling (based on extrapolations from animals to humans or from large doses to small doses, statistical inferential applications, etc.);
   - indeterminacy or genuine stochastic effects (variation of effects due to random events, in special cases congruent with statistical handling of random errors);
   - system boundaries (uncertainties stemming from restricted models and the need for focusing on a limited amount of variables and parameters);
   - ignorance or non-knowledge (uncertainties derived from lack or absence of knowledge).

The first two components of uncertainty qualify as epistemic uncertainty and therefore can be reduced by improving the existing knowledge and by advancing the present modeling tools. The last three components are genuine uncertainty components and thus can be characterized to some extent using scientific approaches but cannot be reduced to numeric confidence intervals.

3. (*Interpretative and normative*) *ambiguity*. This is the last term in this context. Whereas uncertainty refers to a lack of clarity over the scientific or technical basis for decision making, (interpretative and normative) ambiguity is a result of divergent or contested perspectives on the justification, severity, or wider `meanings' associated with a given threat. In relation to risk, it is understood as `giving rise to several meaningful and legitimate interpretations of accepted risk assessments results'. It can be divided into interpretative ambiguity (different interpretations of an identical assessment result: for example, as an adverse or nonadverse effect) and normative ambiguity (different concepts of what can be regarded as tolerable referring, for example, to ethics, quality of life parameters, distribution of risks and benefits, etc.). A condition of ambiguity emerges where the problem lies in agreeing on the appropriate values, priorities, assumptions, or boundaries to be applied to the definition of possible outcomes.

Risk Management Principles

In accordance with the three major components `complexity, uncertainty, and ambiguity', one can design different risk assessment and management strategies depending on the degree or intensity of each component. For example, if a risk is characterized by high complexity, low uncertainty, and low ambiguity, the main emphasis will be on scientific modeling and consensual methods for performing a most accurate assessment. If a risk is characterized by high uncertainty, the main emphasis will be on reducing vulnerability of the system to cope even with surprises. In a situation of high ambiguity, a more discursive approach with stakeholder participation is asked for in order to find an acceptable way to reconcile conflicting values and world views. More specifically, risk handling can pursue one of four strategies or a combination thereof. These strategies are summarized in **Table 1** and described in the following paragraphs.

1. *Simple risk problems*. This class of risk problems requires hardly any deviation from traditional decision making. Data are provided by statistical analysis, goals are determined by law or statutory requirements, and the role of risk management is to

Table 1 Risk characteristics and their implications for risk management

Knowledge characterization	*Management strategy*	*Appropriate instruments*	*Stakeholder participation*
`Simple' risk problems	Routine based:(tolerability/ acceptability judgement) (risk reduction)	Applying `traditional' decision making	Staff members and specialists in the field
		Risk–benefit analysis Risk–risk trade-offs Trial and error Technical standards Economic incentives Education, labeling, information Voluntary agreements	
Complexity-induced risk problems	Assessment-based strategy	Characterizing the available evidence	Wide range of experts and scientist, including social scientists
		Expert consensus seeking tools like: Delphi or consensus conferencing Meta analysis Scenario construction, etc. Results fed into routine operation Improving buffer capacity of risk target through: Additional safety factors Redundancy and diversity in designing safety devices Improving coping capacity Establishing high-reliability organizations	
Uncertainty-induced risk problems	Resilience-based strategy	Using hazard characteristics such as persistence, ubiquity, etc., as proxies for risk estimates	Inclusion of major stakeholders such as industry and environmental groups
		Tools include: Containment ALARA (as low as reasonably achievable) and ALARP (as low as reasonably possible) BACT (best available control technology), etc. Improving capability to cope with surprises Diversity of means to accomplish desired benefits Avoiding high vulnerability Allowing for flexible responses Preparedness for adaptation	
Ambiguity-induced risk problems	Discourse based	Application of conflict resolution methods for reaching consensus or tolerance for risk evaluation results and management option selection Integration of stakeholder involvement in reaching closure Emphasis on communication and social discourse	Wider public, including different social and cultural interests and values

IRGC (2005) *White Paper on Risk Governance: Towards an Integrative Framework.* Geneva: International Risk Governance Council.

ensure that all risk reduction measures are implemented and enforced. Traditional risk–risk comparisons (or risk–risk trade-offs), risk–benefit analysis, and cost-effectiveness studies are the instruments of choice for finding the most appropriate risk reduction measures. Additionally, risk managers can rely on best practice and, in cases of low impact, on trial and error. Staff members of risk management agencies, industrial risk managers, and experts in the field are the only groups that need to be involved for dealing with simple risks.

2. *Assessment-based strategy, including the search for numerical thresholds* (*emission levels, concentration levels, performance standards, tolerance levels, etc.*). Within this second strategy of scientific risk analysis, risk management relies on the best scientific estimates of probabilities and potential damages and uses expected values as main input to judge the tolerability of risk as well as to design risk reduction measures that are cost effective, proportional to the threat, and fair to the affected population. In this frame, precaution may best be interpreted as being conservative in making risk judgments and choosing cautious assumptions when calculating exposure or determining safety factors (of 10, 100, or more) to cover intertarget variability. Since risk assessments for complex problems rely on sophisticated models and simulations, a broad representation of the various scientific communities, including the social scientists, is required to assure a broad and inclusive handling of the risk.
3. *Resilience-based strategy*. This strategy refers to risk reduction activities that are derived from the `application of the precautionary principle'. Within this third frame, the concept of risk is seen from the perspective of pervasive uncertainty and in particular

ignorance and non-knowledge. Precautious risk management entails to ensure prudent handling of decision options in situations of high uncertainty about causes and effects and of high vulnerability of the ecosystems under risk. Instruments of precaution include minimization requirements such as ALARA (as low as reasonably achievable) or ALARP (as low as reasonably possible), diversification of risk agents, containment in time and space, and close monitoring. It is difficult to draw the line between too much precaution and not enough precaution in the face of uncertainty. How can one judge the severity of a situation when the potential damage and its probability are unknown or highly uncertain? In this dilemma, risk managers are well advised to include the main stakeholders in the evaluation process and ask them to find a consensus on the extra margin of safety in which they would be willing to invest in exchange for avoiding potentially catastrophic consequences.

4. *Deliberation-based strategy*. This strategy includes `discursive processes' such as roundtables, consensus conferences, deliberative rule-making, mediation, or citizen panels. This third frame has been advocated as an alternative or an addition to purely analytical procedures of both assessing and managing risks. The task of risk management here is to organize in a structured and effective manner the involvement of stakeholders and interested public for designing risk management strategies based on each stakeholders' knowledge (epistemic community) and value system. This strategy can go along with both the risk analysis and the precautionary approach but has been advocated either as an independent path to risk management or more often as a policy-oriented implementation of the precautionary approach. This third strategy is designed to address problems of ambiguity. For managing these risks, the main groups representing different visions of future development should be consulted and, if possible, a consensus among the main actors accomplished.

Over the last few years, advocates of the these strategies have launched a fierce debate over the pros and cons of each of their approaches. There is generally a common agreement that simple risk problems can be managed using the traditional assessment and risk reduction tools. There is, however, a controversy among the professional communities when it comes to distinguishing between the assessment and the resilience camp. One side argues that precautionary strategies to improve the resilience of ecosystems ignore scientific results and lead to arbitrary regulatory decisions. The precautionary statement ``one should be on the safe side'' could be interpreted as a mandate to ban everything that might result in negative side effects. Such a rule would logically apply to any human intervention into the natural environment and would lead to total arbitrariness. The principle has been labeled to be ill-defined, absolutist, to lead to increased risk-taking, to be a value judgment or an ideology, and to be unscientific or to marginalize the role of science. Some analysts claim that by using the precautionary principle there is the risk that science may be held `hostage to interest group politics'. In addition, policy makers could abuse the precautionary principle as a policy strategy to protect their economic interests and to impede world trade.

On the other side of the fence, the advocates of the precautionary approach have argued that precaution does not automatically imply banning substances or activities but a step-by-step diffusion of risky activities or interventions until more knowledge and experience is accumulated. They have accused their critics of ignoring the complexity and uncertainty of most human interventions and relying on data that often turns out to be insufficient for making prudent judgments. In particular, the argument is that all interventions characterised by both high uncertainty (when it comes to measuring the degree of impacts) and the generic possibility of inducing irreversible harm should be (strictly) regulated even if more scientific evidence could demonstrate later that the cautious approach was not warranted. It is better to err on the safe side than to impose unforeseen burdens to future generations.

The third approach of deliberation has found wide acceptance among social scientists and risk analysts from academia but so far has had little impact on the institutional design of risk management. There are, however, isolated examples of community participation in risk decisions, such as in selecting the remedy under the Superfund cleanup program or negotiated rulemaking in the US. In recent years, however, risk policymakers have acknowledged that participation in risk analysis provides many practical advantages because it transforms difficult issues of resolving epistemic uncertainty to topics of negotiation that can be dealt with at the negotiation table. If society participates in the production of policy-relevant scientific knowledge, such `socially robust' knowledge is less likely to be contested than that which is merely reliable. Accordingly, the EU communication on good governance has highlighted the need for more stakeholder involvement and participation in risk management. How to implement this requirement in day-to-day risk management decision is still under dispute. Many scholars have also questioned the value of deliberative approaches in some settings, arguing that when there is trust in the regulator, a top-down form of risk communication (information transfer) may be better than dialog.

An Example

The use of nuclear energy for electric power has long been and remains a particularly controversial source of risk and concern. This is driven partly by perceptual factors and partly by the characteristics of the nuclear fuel cycle.

The cycle begins with, first, the extraction of natural uranium and, then, enrichment of the uranium and its concentration prior to manufacturing the fuel elements; risks of exposure to radioactive material of operating personnel are of prime concern. A nuclear reactor operates with exceedingly high energy density and inventory of radioactive substances (fission products). Multiple and diverse safety systems are provided for reactor control and decay heat removal. Their total failure is highly improbable but may lead to core damage (meltdown) and to release of parts of the core inventory, with long-lasting effects over a large geographic area. For state-of-the-art reactors, an `extremely low frequency, potentially high consequence' risk profile is regarded as typical.

Power plant operations also discharge small quantities of radioactive particles to the environment, which can accumulate in organisms and can pose a slight human cancer risk.

When the fuel in a reactor is spent, the fuel rods are consigned to a reprocessing plant, to interim storage or to a final repository. Any transport of spent fuel rods must be in proof casks, and the transport process itself is another source of risk as the probability that a cask will rupture in an accident that exceeds its design limit is not zero. In reprocessing, the risk is concentrated upon the release of radioactive substances to the environment, while in storage the isolation of the waste from the biosphere needs to be ensured for very long periods (several millennia).

Vast efforts have been made to minimize the probability of events causing harm to public health and the environment, particularly toward reducing or even preventing entirely major radioactive releases, in all stages of the cycle. Future reactors must thus be designed so that, even in the event of rare core-damaging events, off-site emergency measures are not necessary because impacts will not reach beyond the immediate neighborhood. In final storage, several safety barriers are combined in order to exclude groundwater contamination as far as is humanly possible.

Looking at the nuclear fuel cycle, all four risk management strategies are or could be invoked when regulating and managing the risks associated with each step of the cycle. Strategies belonging to the category of simple risks are in place when designing passive safety barriers that should withstand special pressure or thermal emissions. In addition, occupational safety measures for workers in the uranium mine and in other stations of the cycle can be grouped in this category.

Complex risk assessments are needed to model the impact of low-dose ionizing radiation on human health. These assessments include toxicological experiments using test animals and extrapolating the results to humans and extrapolating from high to low dose effects. Additional insights come from epidemiological data, for example, from the results of comparing cancer rates in populations exposed to higher doses of radiation through warfare or accidental exposure with populations that are only exposed to natural background radiation. Combing dose-response models with exposure requires extensive modeling, which relies on an intensive exchange among natural scientists, statisticians, physicians, exposure specialists, and behavioral experts.

Although modeling the effects of low dose radiation to human health results in some uncertainty (particularly addressing interindividual variability and stochastic effects), a resilience-based approach is clearly needed for assessing and managing ecological impacts of large nuclear accidents and the long-term effects of final disposal of radioactive waste for hundreds and thousands of years. Scientists understand in principle the reactions of ecosystems to higher levels of radiation but are far from predicting the concrete biological consequences of large releases of radionuclides in different ecosystems over long time periods. Experiences in contaminated areas such as Hartford in the USA and the neighborhood of Chernobyl seem to indicate that biodiversity is actually increased and biotopes recover faster than many ecologists had expected. At the same time, however, radiation-sensible plants and animals are replaced by others who are less sensible. There is still lots of ignorance and guesswork when it comes to assessing the long-term effects of releasing large quantities of radionuclides in the environment. In terms of management, a combination of risk minimization, ban on using resources from contaminated areas for human purposes, and even cleanup is normally advocated. In the USA, there is also a strong movement toward including major stakeholders and the local population in designing plans for decontamination and partial cleanup. These management strategies constitute a precautionary approach to managing risks. Since the consequences are not yet fully explored and unexpected surprises may occur, the main philosophy is to restrict exposure as much as possible and to avoid irreversible consequences even if these consequences may turn out to be more benign than expected.

Public involvement is even more important when addressing ambiguity. The still-ongoing public controversy does not derive so much from the complexity of nuclear energy, nor from uncertainty: most of the processes and phenomena are largely understood and agreed. Much of the debate in the past was directed toward further reducing uncertainty and explaining complexity rather than addressing the major problem: ambiguity. The controversy is fueled by several ambiguities due to different perceptions of and associations with the risk. To the opponents of nuclear energy, the problem is that a meltdown and catastrophic release cannot be ruled out (catastrophic aversion), and that spent fuel remains active for such a long time period (time aversion). The proponents of nuclear power, on the other hand, are convinced that there are no other reliable options with comparably minimal greenhouse gas emissions. Each position is based on values. Value differences color the interpretation of the data emanating from risk assessments (which is rarely disputed), as well as the way the risk is framed in the debate. Opponents accordingly judge the risks as being intolerable; to proponents, nuclear energy is an optimal means of supplying the world's electricity at minimal risk.

The controversy has led to considerable differences in the decisions of risk managers, even in Europe – where Germany has decided to phase out nuclear generation, France continues to rely on it, and Switzerland seeks the consent of its voters to replace an old reactor with a new one. There is certainly a need for a discursive strategy including different value positions and visions of the future. There is no guarantee for a consensus, but without such a discourse political decision makers will have hard problems to handle this controversy in a democratic and peaceful way.

Political Relevance

The precautionary principle has been adopted in a variety of forms at international, European Union, and national level. It is applied across an increasing number of national jurisdictions, economic sectors, and environmental areas. It has moved from the regulation of industry, technology, and health risk, to the wider governance of science, innovation, and trade. As it has expanded in

scope, so it has grown in profile and authority. In particular, as Article 174(2) in the EC Treaty of 2002, precaution now constitutes a key underlying principle in European Community policymaking. The 2000 Communication on Precaution of the European Commission provides evidence for the high significance that the precautionary principle has gained as a guiding policy of the European Union in areas such as environmental, consumer, and health protection. The document states in the first section: ``Applying the precautionary principle is a key tenet of its policy, and the choices it makes to this end will continue to affect the views it defends internationally, on how this principle should be applied'' (European Commission, 2000: 3). As Elisabeth Fisher (2002) pointed out, the communication specifies also some of the major conditions and requirements for applying the principle. There are two conditions mentioned: ``The measures, although provisional, shall be maintained as long as the scientific data remain incomplete, imprecise or inconclusive and as long as the risk is considered too high to be imposed on society'' (European Commission, 2000: 21). In addition to the presence of remaining uncertainty, the EU communication lists the condition that the risk must be too high to be imposed on society. This relates to the requirement of proportionality that has been stated as one of the major requirements of applying the principle.

In the aftermath of a series of formative public health controversies, economic calamities, and political conflicts (such as those involving BSE and GM crops), precaution is of great salience or importance in may fields including the regulation of ecosystem interventions. Since the application of the precautionary principle has been associated with stricter and more rigid regulations, environmental groups have usually rallied around the precautionary approach, while most industrial and commercial groups have been fighting for the assessment-based approach Again, the issue is not resolved, and the debate has become even more pronounced with the defeat of the European Community in the WTO settlement of hormones in beef. The European Community failed in providing sufficient evidence that the precautionary approach could justify the restriction of imported beef treated with hormones.

It is interesting to note that the assessment-based approach has been widely adopted by the official US regulatory bodies while the precautionary approach has been widely advocated by the EU regulatory bodies. There are, however, also numerous elements of precautionary approaches interspersed into the actual practices of US regulatory agencies, just as there are judgments about magnitudes of risk in the actual practices of regulators in the EU. A strict dichotomy between `precautionary' in Europe and `assessment based' in the US is therefore too simple to describe actual practice.

Summary

Any regulatory regime is faced with the question of how to make regulatory decisions under uncertainty or even ignorance. It may be helpful in this respect to resort to a differentiation that Resnik has proposed:

- *Decisions under certainty*. The outcomes of different choices are known.
- *Decisions under risk*. Probabilities can be assigned to the outcomes of different choices.
- *Decisions under ignorance*. It is not possible to assign probabilities to the outcomes of different choices.

Using precaution for the first two cases seems neither necessary nor prudent given that regulation needs to meet both objectives: ecological protection and securing economic welfare. The legitimate realm of using precaution is in the case of ignorance or other forms of remaining uncertainties (such as system boundaries or truly stochastic events). This is the place where precaution should be applied. The main purpose of precaution in this respect is to avoid irreversible decisions. Where it is clearly impossible to calculate expected values, a precautious approach can help societies to be more resilient against unpleasant surprises and to invest in decreasing vulnerabilities.

See also: Conservation Ecology: Ecological Risk Assessment. Human Ecology and Sustainability: Adaptive Management and Integrative Assessments

Further Reading

Charnley, G., Elliott, E.D., 2002. Risk versus precaution: Environmental law and public health protection. Environmental Law Reporter 32 (2), 10363–10366.
Cross, F.B., 1996. Paradoxical perils of the precautionary principle. Washington and Lee Law Review 53, 851–925.
European Commission, 2000. Communication from the Commission on the Precautionary Principle. Brussels: European Union, COM(2000) 1.
European Commission, 2001. European Governance: A White Paper. final. Brussels: European Union, COM(2001) 428.
Fisher, E., 2002. Precaution, precaution everywhere: Developing a `common understanding' of the precautionary principle in the European community. Maastricht Journal of European and Comparative Law 9 (1), 7–46.
Gee, D., Harremoes, P., Keys, J., *et al*., 2001. Late Lesson from Early Warnings: The Precautionary Principle 1898–2000. Copenhagen: European Environment Agency.
IRGC, 2005. White Paper on Risk Governance: Towards an Integrative Framework. Geneva: International Risk Governance Council.
Klinke, A., Renn, O., 2001. Precautionary principle and discursive strategies: Classifying and managing risks. Journal of Risk Research 4 (2), 159–173.
Löfstedt RE (2004) The swing of the pendulum in Europe: From precautionary principle to (regulatory) impact assessment. *AEI-Brookings Joint Center for Regulatory Studies Working Paper 04-07*. London: Kings College.
Majone, G., 2002. What price safety? The precautionary principle and its policy implications. Journal of Common Market Studies 40 (1), 89–109.
Marchant, G.E., Mossman, K.L., 2004. Arbitrary and Capricious: The Precautionary Principle in the European Union Courts. Washington, DC: The AEI Press.

O'Riordan, T., Cameron, J.C., 1994. Interpreting the Precautionary Principle. London: Earthscan.
O'Riordan, T., Cameron, J.C., Jordan, A., 2001. Reinterpreting the Precautionary Principle. London: Cameron May.
Peterson, M., 2007. The precautionary principle should not be used as a basis for decision-making. EMBO Reports 8, 305–308.
Renn, O., 2007. Precaution and analysis: Two sides of the same coin? EMBO Reports 8, 303–305.
Renn, O., Stirling, A., Müller-Herold, U., *et al*., 2003. The Application of the Precautionary Principle in the European Union. Stuttgart: University of Stuttgart, Final Report to the EU Commission.
Resnik, D., 2003. Is the precautionary principle unscientific? Studies in History and Philosophy of Biological and Biomedical Sciences 34, 329–344.
Sand, P., 2000. The precautionary principle: A European perspective. Human and Ecological Risk Assessment 6 (3), 445–458.
Sandin, P., Peterson, M., Hansson, S.O., RudÅn, C., JuthÅ, A., 2002. Five charges against the precautionary principle. Journal of Risk Research 5 (4), 287–299.
Stirling, A., 2003. Risk, uncertainty and precaution: Some instrumental implications from the social sciences. In: Berkhout, F., Leach, M., Scoones, I. (Eds.), Negotiating Change. London: Elgar, pp. 184–212.
Stirling, A., 2007. Risk, precaution and science: Towards a more constructive policy debate. EMBO Reports 8, 309–315.
Stirling, A., Renn, O., van Zwanenberg, P., 2006. A framework for the precautionary governance of food safety: Integrating science and participation in the social appraisal of risk. In: Fisher, E., Jones, J., von Schomberg, R. (Eds.), Implementing the Precautionary Principle. Perspectives and Prospects. Cheltenham and London: Edward Elgar, pp. 284–315.
van Asselt, M.B.A., 2000. Perspectives on Uncertainty and Risk. Dordrecht, The Netherlands: Kluwer.

Remote Sensing

Ned Horning, American Museum of Natural History, New York, NY, United States

Glossary

Anthromes Human biome that represents ecosystems created or altered by human interactions.
Electromagnetic spectrum All known wavelengths (frequencies) of electromagnetic radiation including short wavelength gamma rays, visible light and long-wavelength radio waves.
Pixel Individual elements of an image that are arranged in a grid of rows and columns.
Radiance Measurement of the intensity of electromagnetic energy. Units for this measurement are typically watts per steradian per square meter ($Wm^{-2}\ sr^{-1}$).
Reflectance Ratio of the intensity of light reflected from a surface divided by the intensity of incident light.

Introduction

Remotely sensed data and methods provide data layers that are used extensively to study relationships between humans and their environments. Understanding fundamental remote sensing concepts will help ecologists make informed decisions regarding the utility and limitations of the broad spectrum of remotely sensed data and their derived products.

The main focus of this article is to introduce remote sensing science and associated datasets of potential relevance to ecology. The article begins with an overview of fundamental remote sensing concepts. This is followed by a section on how remotely sensed data can be used to derive a broad range of biophysical datasets that are useful for ecological mapping and modeling. Several of the more common datasets derived from remotely sensed data are described with comments about how the datasets are produced, their strengths, and limitations. The article concludes with a summary of accuracy and validation.

What Is Remote Sensing?

In general terms, remote sensing is the science and practice of acquiring information about an object without actually coming into contact with it. In terms more appropriate for our purposes remote sensing is a technology for sampling reflected and emitted

electromagnetic (EM) radiation from the Earth's terrestrial and aquatic ecosystems and atmosphere. This is typically done by recording images from airplanes and satellites to help identify or better understand a feature on the Earth's surface. In this article we will discuss a wide set of techniques, often known by the alternative name of Earth Observation (EO). We will only address electromagnetic remote sensing so geomagnetic and acoustic remote sensing techniques (sonar, and seismic sounding) will not be covered.

A simple example of a remote sensing instrument is a photographic or digital camera. A camera records energy in the form of light that is reflected from a surface to form an image. Most photographic cameras record visible light so that when we look at the photograph the image resembles the feature that was photographed. More sophisticated remote sensing instruments are able to record energy outside of the range of visible light. Data from remote sensing instruments can be recorded as images or, in the case of lidar, a collection of point data which are often processed to create images.

Brief History of Remote Sensing

For our purposes we will begin the history of remote sensing with the invention of the photographic camera in the early 19th century. In the 1840s photographs were taken from cameras secured to tethered balloons for purposes of topographic mapping. For the next 100 years or so camera technology improved but the major advances were in the platforms used to hold camera systems. At first people experimented with platforms such as kites, rockets, and even pigeons. A major step forward was made with the invention of the airplane and the next leap occurred when cameras could be mounted on satellites, which provided a very stable and, of course, high altitude platform. Satellites also provide an ideal platform for acquiring systematic data from around the globe which has proved invaluable for large area ecological modeling.

By the 1940s instrument research was also becoming increasingly sophisticated, pushing remote sensing technology beyond visible-spectrum photography into infrared detection and radar systems. Leveraging this research, in 1972 the National Aeronautics and Space Administration (NASA) began the Landsat program with the launch of the Earth Resources Technology Satellite 1 (ERTS 1), which was later renamed Landsat 1. The Landsat program continues and is the longest running program of satellite remote sensing focused on EO. Following the launch of Landsat 1 other satellites were launched carrying different types of instruments such as radar, lidar, and more precise optical sensors. Satellite remote sensing has evolved to the point where most environmental systems (hydrologic, atmospheric, ecosystems) now have dedicated satellite instruments recording information to help us better monitor and manage Earth's environments, and providing valuable data for use in ecological modeling and monitoring. For example, NASA's Earth Observing System (EOS) is a mission that includes the acquisition of satellite-based observations, science, and a data system to support the study of the land surface, biosphere, solid Earth, atmosphere, and oceans. One sensor in particular from the EOS mission that offers a broad range of image products of interest to ecologists is the Moderate Resolution Imaging Spectroradiometer (MODIS). Advances in satellite/instrument packages will continue to be made providing more precise and accurate data that can be used for ecological applications.

Basic Concepts

It is necessary to understand a few basic remote sensing concepts before we begin discussing how remotely sensed imagery can be used in ecology.

Electromagnetic Spectrum

The electromagnetic spectrum (EMS) includes wavelengths of electromagnetic radiation ranging from short wavelength (high frequency) gamma rays to long-wavelength (low frequency) radio waves. We will focus on the region of the spectrum starting in the ultraviolet and continuing through the microwave wavelengths. Optical sensors are used to measure ultraviolet, visible, and infrared wavelengths and microwave sensors are used for the microwave portion of the EMS.

A fundamental physical principal that remote sensing relies on is that different features on the Earth's surface interact with specific wavelengths of the EMS in different ways. When working with optical sensors the most important property used to identify features on the Earth's surface is spectral reflectance; the ratio of the intensity of light reflected from a surface divided by the intensity of incident light. Different features have different spectral reflectance properties and we can use this information to identify individual features. For example, white sand reflects most visible and near-infrared light whereas green vegetation absorbs most red wavelengths and reflects most near-infrared wavelengths. Fig. 1 illustrates the spectral properties of different materials.

Some remote sensing instruments also provide information about how electromagnetic energy interacts with the surface of a feature or within a three-dimensional feature such as a forest. These will be discussed later in this article.

What Is an Image?

The most familiar form of remotely sensed data is an image. An image is made up of individual elements that are arranged in a grid of rows and columns. These elements are called pixels. When zooming into an image, individual pixels can be seen (Fig. 2).

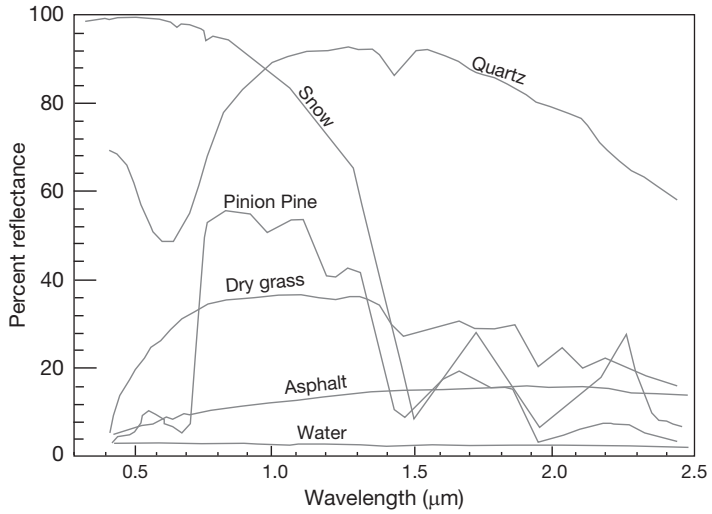


Fig. 1 Spectral signatures for selected features. Materials on the Earth's surface have unique spectral reflectance properties. This figure shows the spectral reflectance curves of six common materials. Many of the methods used in remote sensing are designed to associate the spectral information acquired by a sensor with the spectral qualities of features that are to be identified.

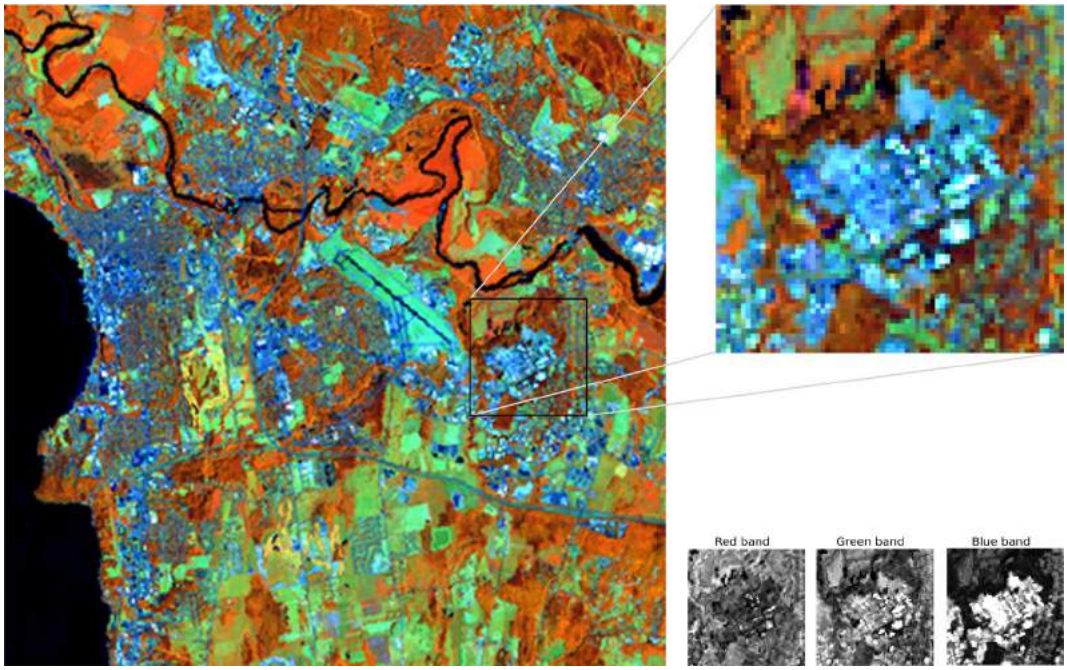


Fig. 2 Zooming to an individual pixel. The image in the top-left is printed at full resolution, it is a Landsat Enhanced Thematic Mapper Plus (ETM+) image acquired over Burlington, Vermont on 21 August 1999. The image in the top-right is a subset of this image that has been magnified by a factor of three. In the magnified image individual pixels (the *square blocks* that make up the image) can be seen. The three *black* and *white images* in the bottom-right represent the three image bands that are used to create the color image. In this case the *red band* is from the ETM+ band 4 (near-infrared), the *green band* is from ETM+ band 5 (mid-infrared), and the *blue band* is from ETM+ band 3 (*red*). These three bands are combined to make the color image.

In addition to rows and columns of pixels, images also have layers. These layers are commonly referred to as "bands" or "channels." Throughout this article we use the term "band" to refer to the layers in an image. These bands also correspond to different wavelengths of the EMS. Remote sensing instruments vary in the number of bands recorded, some only record a single band of data while others record hundreds of bands. A convention has been established to use "hyper-spectral" to describe images with many bands (usually well over 100) and "multispectral" for images with fewer (usually from three to a few dozen) bands.

With most imagery the individual bands are used to record radiance values at different wavelengths. Radiance is a measurement of the intensity of electromagnetic energy. In other words, the sensor is measuring the intensity of light when it hits the detector. The units for this measurement are typically watts per steradian per square meter ($W\ sr^{-1}\ m^{-2}$). It is important to understand that

optical sensors measure radiance and not reflectance. Reflectance, which is the ratio of reflected light over incident light, can be estimated using image processing methods but the physical property recorded by the sensor is radiance (Fig. 3).

Different Platforms and Orbits

For local and detailed information aircraft with on-board and remote pilots are often the platform of choice since it is possible to select which sensors should be mounted for a particular application and it is possible to determine when to fly. Aircraft have the ability to fly low to acquire imagery with a lot of detail. Remotely piloted aircraft, also called drones and unmanned aircraft, are increasingly being used to acquire remotely sensed data to support ecology studies. Regulations constraining use varies dramatically between countries and these constraints can impose a practical limit on how remotely piloted aircraft can be used.

For global and systematic coverage, satellites are the standard remote sensing platform. Most satellite orbits can be classified as either geostationary or polar orbiting. Geostationary satellites orbit the Earth with the same orbital period as the Earth's rotation. In other words, a geostationary satellite orbits around the Earth's rotation axis keeping its position fixed over a specific point on the Earth so it is always viewing the same area. These satellites are commonly used to monitor weather but are too far from the Earth's surface (~38,500 km) for detailed environmental monitoring. More common for Earth remote sensing is a near-polar orbit that provides a near-global view of the Earth over a regular time period, for example, every 16 days in the case of Landsat. It is important to note that with a near-polar orbit the extreme polar regions are not viewed from the satellite. For this reason, when people mention global remotely sensed datasets they often mean the datasets are near-global. Polar and near-polar orbiting satellites fly only several hundred kilometers above the Earth's surface.

A relatively new concept for satellite platforms is to launch several small (often between 1 and 10 kg) satellites called nanosatellites or CubeSats carrying lower-cost sensors. Having a dense constellation of similar satellite sensors orbiting the Earth provides the opportunity for frequent high-resolution data acquisition over the same area.

Passive Versus Active Remote Sensing

Remote sensing instruments are often categorized as having either active or passive sensors. An active sensor generates its own signal which is subsequently measured when reflected back by the Earth's surface. A passive sensor measures solar energy that is either reflected or emitted from features on the Earth's surface. Table 1 lists a number of different active and passive instruments mounted on satellite platforms commonly used for ecological studies.

Although most passive sensors operate in the visible and infrared portions of the EMS, there are also some passive microwave sensors in use that measure a number of parameters such as wind speed, atmospheric and sea surface temperature, soil moisture, rainfall, and atmospheric water vapor.

An advantage of passive sensors is that most rely on the sun's energy to illuminate the target and therefore do not need their own energy source so in general they are simpler instruments. A limitation for most passive optical sensors is that they require daylight to operate, although there are some sensors that record nighttime lights and clouds at night and others that record energy emitted from

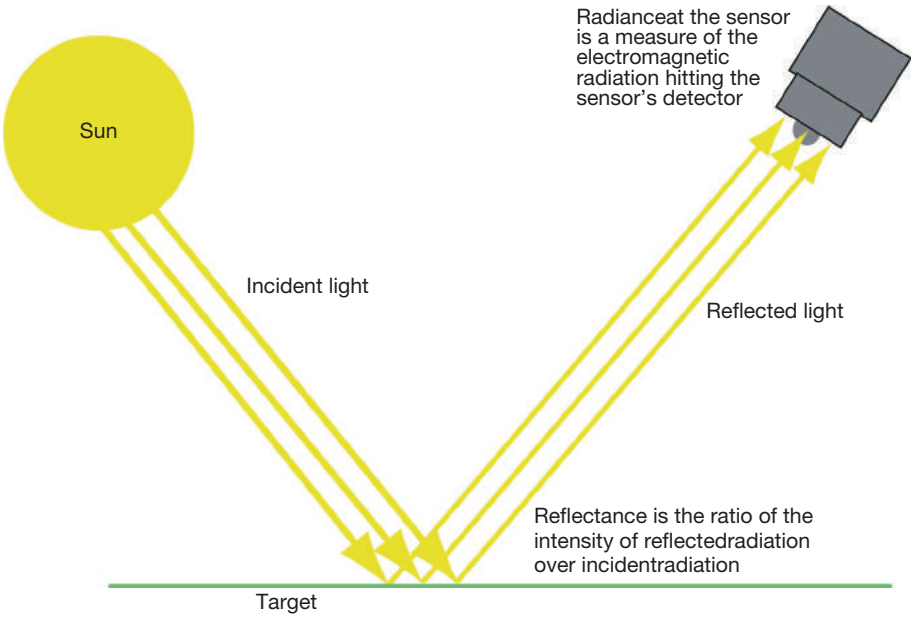


Fig. 3 Reflectance and radiance. Remote sensing detectors measure radiance which is the energy of the radiation hitting the detector. Reflectance, which must be calculated, is the ratio of the intensity of reflected radiation divided by incident radiation.

Table 1 Active and passive satellite-based remote sensing instruments

Sensor name	*Type*	*Wavelength range*	*Resolution*
WorldView-4	Optical	450–920 nm	0.31–1.23 m
SPOT 7	Optical	450–890 nm	1.5–6 m
IRS Resourcesat-2 LISS-III	Optical	520–1700 nm	23.5 m
ALOS AVNIR-2	Optical	420–890 nm	10 m
GeoEye-1	Optical	450–920 nm	0.46–1.84 m
Landsat 8 OLI/TIRS	Optical	430–1251 nm	15–100 m
MODIS	Optical	459–14,385 nm	250–1000 m
Suomi-NPP VIIRS	Optical	410–12,500 nm	400 m
ENVISAT ASAR	Radar	5.7 cm (C band)	30–1000 m
Sentinel-1A and 1B	Radar	5.7 cm (C band)	5–100 m
RADARSAT-2	Radar	5.6 cm (C band)	3–100 m
ALOS-PALSAR-2	Radar	23.5 cm (L band)	0.65–100 m

the Earth's surface. Since most of these sensors operate in the visible and infrared wavelengths, they are adversely affected by weather and cloud cover. Lastly, since sunlight is primarily reflected from the top of a feature, such as a forest, it is not possible to "see" under a canopy to measure vegetation structure. To obtain this kind of information it is necessary to use active sensors.

Active sensors, such as radar and lidar emit their own energy to illuminate a target and are comprised of a signal generator and receiver. They measure the strength of the returned signal and the time delay between when the instrument emits the energy and when it receives the returned pulse. Lidar is an acronym for *li*ght *d*etection *a*nd *r*anging and these systems use lasers that emit light in the visible and near-infrared portions of the electromagnetic spectrum. In a lidar system a single light pulse can reflect off of several features in vertical space such as different layers in a forest. A single emitted pulse will result in a series of returned pulses that are recorded by the detector. These returned pulses can be recorded as a wave (full waveform lidar) or in discrete pieces that correspond to peaks in the returned signal such as the first and last return of the light pulse. The first return corresponds to the top of an object (i.e., top of a tree canopy) and the last return would correspond to the base substrate that the object is resting on (i.e., the ground). This is ideal for measuring the height of features such as trees or buildings. With a full waveform lidar the intensity of multiple returns are recorded which can be useful for characterizing vegetation structure. Although lidar systems are also touted as being able to penetrate a forest canopy it is important to note that laser pulses pass through gaps between leaves and branches and do not actually penetrate through plant material. Lidar systems called imaging lidar are capable of recording a dense collection of points, a point cloud, in three-dimensional space (e.g., height, latitude and longitude) and these can be processed to create images of vegetation or building height, land cover type, terrain elevation, and vegetation structure.

Recent advances in lidar (i.e., multiphoton lidar) make it possible to record a spatial matrix of returns from a single pulse. Multiphoton lidar systems have reduced energy requirements, greater sensor sensitivity and increased data density allowing greater flying heights and faster flight speeds so more area can be covered in a shorter period of time.

Radar is an acronym for *ra*dio *d*etection *a*nd *r*anging. Radar systems operate in the long-wavelength microwave portion of the EMS and thus are largely unaffected by clouds and rain. They can be considered all-weather systems. A radar microwave signal interacts differently with land surface features than shorter wavelength electromagnetic energy detected by optical sensors. For example, a radar signal can penetrate clouds and in some cases can penetrate well into a forest canopy. Radar systems with especially long wavelengths (e.g., P-band systems) can even penetrate dry ground. The ability to penetrate clouds and vegetation makes it possible to map floods through cloudy skies or even under a forest canopy. A radar signal is very sensitive to surface roughness and an object's dielectric constant. Since water can drastically change an object's dielectric constant radar imagery is used to detect changes in soil moisture or vegetation water content. The properties of a detected radar signal also change with varying surface roughness. For example, the ability to detect changes in water surface roughness makes it possible to locate and monitor oil spills in water. This is possible because oil attenuates waves that form on the surface of water bodies so the oiled surface appears smooth while adjacent uncontaminated water has a rougher appearance. Another property of radar is that it can emit and record microwave radiation with different polarizations, usually vertical and horizontal. This feature can be useful when mapping land cover since different land surface features modify the polarization of a radar pulse in unique ways and those differences can be detected with a multipolarization radar instrument. Multiple radar images can be processed to map land surface elevation and detect small changes in elevation over time such as those that occur before a volcanic eruption.

Due to the complex nature of radar remote sensing the skills required to process these data tend to be more difficult to learn than those required for optical image processing. However, most organizations that supply radar imagery provide high-level products than can be used for many applications with moderate training.

What Qualities Determine What Can Be Identified in an Image?

There are different characteristics that affect the detail that can be resolved (seen) in a digital image. These are traditionally referred to as the four types of image resolution. Most people think of "resolution" as being synonymous with spatial resolution but these other "resolution" terms are used in the formal literature.

Spatial resolution which is often simply referred to as "resolution" is the size of a pixel (smallest discrete scene element and image display unit) in ground dimensions. In most cases an image's resolution is labeled with a single number, such as 30 m, which represents the length of a side of a square pixel if it were projected onto the Earth's surface. If the pixel were rectangular (not very common any more), then the length and width of the pixel would be provided.

Spectral characteristics include band width, band placement, and the number of bands. Spectral bandwidth, or spectral resolution as it is often called, refers to the range of wavelengths that are detected in a particular image band. This is effectively a measure of how precisely an image band measures a portion of the electromagnetic spectrum. Band placement defines the portion of the electromagnetic spectrum that is used for a particular image band. For example, one band might detect blue wavelengths and another band might detect thermal wavelengths along the EMS. The properties of the features you are interested in sensing indicate which bands are important. The last spectral variable is the number of bands. The more bands that are available the more precisely spectral properties of a feature can be measured.

Acquisition dynamics has two components. The first is the minimum time a particular feature can be recorded twice, often called the repeat frequency or temporal resolution. Some sensors with a very wide field of view can acquire multiple images of the same area in the same day whereas some sensors have a repeat frequency of several weeks. It should also be reiterated that most remote sensing satellites have a near-polar orbit and are not able to acquire imagery at the poles since their orbit does not go over these areas. The other component is the timing of the acquisitions. Dynamic features such as deciduous forests and events such as flooding often have an optimum time for which they should be imaged. For example, the identification of deciduous vegetation is aided by acquiring imagery during leaf-on and during leaf-off periods. The time of day a satellite acquires an image can also be important for some applications. Many satellites are programmed to acquire images in the morning before clouds build but some acquire imagery in the afternoon or at night. The extent of shadows is also affected by the season and time of image acquisition.

Sensitivity of the sensor is defined by the dynamic range of the sensor as well as the range of digital numbers that can be used to represent the pixel values. Sensors have lower limits below which a signal is not registered and upper limits above which the sensor saturates and is unable to measure increases in radiance. The detail that can be measured between these extremes is determined by the range between the minimum and maximum digital numbers permitted for a particular data type. For example, Landsat TM data values can range from 0 to 255 whereas IKONOS values range from 0 to 2048. This potential range of values is often referred to as quantization or radiometric resolution.

Remotely Sensed Data Sets for Ecological Applications

Remote sensing provides instruments and methods that can be used to derive a broad range of biophysical datasets that are useful for terrestrial and aquatic ecological modeling. In this section several of the more common datasets derived from remotely sensed data will be described with comments about how the datasets are produced, their strengths, and limitations. This is not an exhaustive list but it highlights the diversity of data derived from remote sensing that can be integrated into ecological modeling.

Land Cover

Land cover data are available in image and vector formats with individual types of vegetation assigned to discrete classes. For example, in an image format each vegetation type would be assigned a unique numeric value and in a vector format each polygon would have attribute information that would describe the type of land cover in that polygon. These data are available with a wide range of thematic (classification scheme) and spatial (spatial resolution and extent) qualities.

The specific classification scheme used for a particular land cover dataset can be as simple as forest/nonforest classes or as detailed as a species-level map. One important point related to thematic detail is that the more classes that are used, the lower the per-class accuracy will be. In other words, the classes in a forest/nonforest map will be more accurate than the individual classes in a species-level map. The spatial detail in a land cover dataset is usually a direct result of the type of remotely sensed data on which the classification was based. Using lidar, aerial photography or high-resolution satellite imagery individual tree crowns can be discerned allowing improved capabilities for mapping species-level information.

For the most part land cover maps are created using data from optical sensors and by combining radar and optical data. One area where radar sensors excel is in mapping wetlands and water under forests, such as in flooded forests or vernal pools.

Land cover datasets can be created using manual and/or automated methods. The basic principle of land cover classification is to translate pixel values in a satellite image into meaningful land cover categories. This is often accomplished using automated procedures in which a computer algorithm is used to assign individual pixels or groups of pixels to one of the valid land cover categories. The classification process can also be accomplished using visual interpretation methods where the interpreter uses visual cues such as tone, texture, shape, pattern, and relationship to other objects to identify and group similar land cover types. In general the human brain is better at interpreting the spatial characteristics in an image and automated algorithms are better suited for

processing spectral (the many image bands) information. There are dozens of classification methods in use but not a single "best" approach.

One of the limitations of classified land cover data is that the information is discrete instead of continuous. One way around this is to create a "continuous fields" image dataset for selected types of vegetation. In a continuous fields dataset each pixel value represents the percentage of that pixel covered by a particular land cover type. For example, in a broadleaf tree continuous fields dataset a pixel value of 65 would mean that 65% of that pixel is covered by broadleaf tree species. In addition to different types of land cover it is also possible to create a continuous fields dataset for imperviousness. This is called an impervious surface dataset and it is being used increasingly in ecological modeling particularly when it is necessary to quantify water runoff.

Landscape Metrics

Once a land cover map has been produced it is sometimes desirable to quantify different aspects of patterns in the landscape using assorted landscape metrics for use in ecological models. These metrics provide an objective way to describe patterns commonly described using subjective terms such as "highly fragmented," "small patches," and "heterogeneous landscape." Using software tools it is easy to create these metrics. Some common metrics include:

- Landscape composition
  - Proportion—Area of one cover type compared to the total area
  - Richness—Number of different patch types
  - Evenness—Relative abundance of different patch types
  - Diversity—Composite measure of richness and evenness
- Spatial configuration
  - Patch size and shape
  - Connectivity of patches
  - Dispersed or clumped patches
  - Setting with respect to neighboring patches

Although landscape metrics can be of great value, caution must be exercised when using these metrics. Many of the metrics are very sensitive to image scale and extent of the study area so comparisons across time and space must be done with care.

Vegetation Characteristics

In addition to land cover, there are several vegetation characteristics that can be measured using passive and active remote sensing instruments. These include:

- Phenology
- Primary productivity
- Vegetation health and vigor
- Vegetation structure

Measurements of these characteristics are based on the fact that reflectance, transmittance, and scattering of energy in a canopy is greatly affected by the structure of the vegetation and how the vegetation components (leaves, branches, trunk) interact with the spectrum of energy being used by a particular remote sensing instrument.

Vegetation indices have been used extensively for global studies to monitor changes in vegetation health and cover and have been effective in mapping droughts, desertification, phenology, net primary productivity, and deforestation around the world. The most common vegetation index, the Normalized Difference Vegetation Index (NDVI), is based on the principle that healthy green vegetation absorbs most of the incident red wavelengths of light and reflects most of the near infrared wavelengths. The formula for NDVI is:

$$\mathrm{NDVI} = (\mathrm{NIR} - \mathrm{Red})/(\mathrm{NIR} + \mathrm{Red})$$

where NIR is the radiance or reflectance value from the near-infrared band and Red is the radiance or reflectance value from the red band.

Two other common vegetation indices that use a similar principle as NDVI are the Soil Adjusted Vegetation Index (SAVI), which was developed to reduce the effect of background material (i.e., soil, sand, snow) and the Enhanced Vegetation Index (EVI) which is less sensitive to atmospheric scattering effects. Vegetation index datasets are usually available as temporal composites, such as 10-day or monthly. In a composite product the "best" index values from the composite period are provided for each pixel. Using this approach it is possible to reduce the negative effects of clouds and haze.

Vegetation structure and biomass datasets are often created using data acquired from radar and lidar sensors. Although radar has been used to measure vegetation properties such as biomass, leaf area index, and forest structure most of this has been experimental so data are somewhat limited. This is an area of active research and as new instruments are developed, operational methods using radar instruments may be available in the not too distant future.

Commercial lidar instruments are available for mounting in airplanes that can quickly provide vegetation height information and this can be correlated to tree volume and biomass using allometric tables. Unfortunately using these instruments is expensive and it is often not feasible to cover large areas. Research using airborne and satellite lidar instruments to measure vegetation structure directly is underway and early results look promising.

Topography

Elevation datasets and their derived products (Table 2) are used extensively in ecological applications. Digital elevation data are available as digitized points or contour lines, triangulated irregular networks (TIN), and as gridded surfaces or images.

When using elevation data it is important to know what the elevation values represent. These values can represent the surface of the bare earth, the surface of the features on the earth (i.e., top of the canopy or top of buildings) or somewhere in between.

A commonly used global topographic dataset is the Shuttle Radar Topographic Mission (SRTM) Digital Elevation Model (DEM). The resolution of this dataset is 30 m and it covers land areas between 56 degree south latitude and 60 degree north latitude. The data for the SRTM DEM were collected using an interferometric radar instrument mounted on the Space Shuttle. Over forested areas the elevation value provided by the SRTM dataset represents a point somewhere roughly half-way between the ground surface and the top of the canopy. The exact point depends on the structural characteristics of the forest stand.

Lidar is increasingly being used to collect very detailed elevation data with accuracy on the order of centimeters. Lidar instruments are flown on aircraft and are routinely used for monitoring coastal areas. The accuracy and speed of lidar elevation data collection is unmatched by traditional ground-based methods.

Soil Moisture

Soil moisture is a much sought after dataset for ecological modeling. However, data that meet the needs of a particular application are sometimes not available because existing data spatial resolutions are too coarse, available data do not cover the area of interest, or data do not exist for the required time frame. Relatively new global soil moisture data with a 3–36 km spatial resolution collected from active and passive microwave sensors is available from the Soil Moisture Active Passive (SMAP) NASA mission. People have experimented using data collected from optical remote sensing instruments to map soil moisture but results have been mixed and methods are typically tailored to a specific ecosystem.

Surface Temperature and Precipitation

Although temperature and precipitation data are routinely collected using satellite-based instruments, datasets created using meteorological station data are still often preferred for ecological modeling. In some cases datasets derived from station data are improved by integrating data collected from satellite-based instruments.

Satellite-based rainfall estimates are made using passive microwave, radar, and optical instruments. Rainfall estimate datasets derived from satellite remote sensing are often too coarse (>4 km) for many ecological modeling tasks.

There are many satellite-based instruments that can measure surface temperature including some that produce a global daily and 8-day composite dataset available with a 1 km resolution. Finer resolution data can be acquired using satellite sensors. However, they do not have as frequent repeat cycles as those producing 1 km products.

Table 2 Products commonly derived from digital elevation models (DEM)

Slope steepness
Slope aspect
Hillshade and perspective views
Viewshed/line of sight
Topographic features
Ridges
Peaks
Channels
Pits
Passes
Plateaus
Hydrologic parameters
Flow direction
Flow accumulation
Predicted watercourses
Watershed boundaries

Atmospheric Properties (Clouds)

Of all the remote sensing-derived atmosphere products (Table 3), the cloud mask is arguably the most frequently used for ecological modeling since it provides a measure of cloud cover. For the MODIS cloud mask product a clear-sky confidence level (high confident clear, probably clear, undecided, cloudy) is assigned to each pixel. This daily dataset is available with 250 m and 1 km spatial resolutions.

Oceans

In the marine environment, remote sensing provides data for a wide variety of environmental variables (Table 4). Optical remote sensing methods are commonly used to map the ocean/land interface and coral reefs. Methods used for these applications are similar to those used for terrestrial land cover mapping except there is greater use of the short wavelength blue bands for coral reef mapping since those wavelengths are better able to penetrate into the water to provide more information on features several meters under the water surface.

Two other common global marine datasets are sea surface temperature, which is derived using methods similar to those used for land temperature, and ocean color which uses optical imagery to determine levels of phytoplankton in the water. Both of these datasets are acquired on a daily basis.

Inland Water

Inland water features include wetlands, streams, and lakes. A mix of optical and microwave remotely sensed data are used to measure and monitor a number of parameters (Table 5) related to inland waters although many of these measurements are only available at local or regional scales and many of the methods require significant field work to correlate actual values with what is recorded by the remote sensing instrument.

Mapping the extent of inland water features is done using a mix of optical and microwave instruments. For example, radar is an ideal technology for locating and mapping standing water, even if it occurs under a forest canopy such as in a flooded forest. Optical sensors are used to measure temperature and identify aquatic vegetation.

Anthropogenic Biomes

Much of human ecology is linked to studying the state and change of anthropogenic biomes or anthromes. Many of the biophysical datasets described above have direct relevance for built and human modified environment. For example, studying the changing urban environment of a city can be done using land cover analysis techniques. One satellite dataset that is largely unique to human activity the Defense Meteorological Satellite Program (DMSP) Operational Linescan System (OLS) sensor nighttime lights. This dataset has been used to create metrics for economic activity and global poverty, population, energy usage, and urban sprawl.

Data Access

Access to remotely sensed data and their derived products has steadily improved over the years as data providers developed more intuitive data browsing and ordering tools. Data licenses for satellite imagery and derived products range from public domain (effectively no restrictions on use) to tightly controlled licenses limiting use to individuals. Initially United States government agencies such as NASA, USGS and NOAA took the lead in providing unrestricted access to remotely sensed data but many other governments such as ESA in Europe, JAXA in Japan and INPE in Brazil have followed that model and offer some of their remote sensing products at no cost. At the other extreme is strict licensing provided by dozens of commercial companies that tend to offer very high-resolution imagery along with services to schedule image acquisitions and provide value added processing. Commercial

Table 3 Information about the atmosphere derived from remote sensing

Aerosol loading and size distribution over oceans
Aerosol content and optical thickness over land
Water vapor (precipitable water)
Cloud optical thickness
Cloud top temperature and height
Cloud locations
Cloud particle phase
Cloud particle radius
Total-column ozone content
Atmospheric wind and temperature profiles

Table 4 Information about marine ecosystems derived from remote sensing

Sea surface temperature
Ocean color (productivity)
Coral reef mapping
Ocean surface topography
Oil slick detection and mapping
Ocean circulation
Wind speed and direction
Fluorescence

Table 5 Information about inland water ecosystems derived from remote sensing

Water body and wetland mapping
Flooded forest mapping
Water surface elevation
Water depth
Turbidity/secchi depth
Water temperature
Aquatic and wetland vegetation mapping
Riparian buffer mapping
Flow rates

products nearly always come with well defined licenses that constrain who may legally use those data. Although access to commercial data can be costly it is important to remember you are purchasing the license to use those data and are not purchasing the data itself.

Accuracy Assessment and Validation

When using datasets derived from remote sensing methods, it is important to understand the level of accuracy associated with a particular product. Accuracy figures can refer to the accuracy of the calculated value when compared to the actual biophysical value or it can refer to the positional accuracy. In some cases both figures are provided. Accuracy statistics should be distributed with the dataset. Unfortunately, this is not always the case. In some cases, accuracy statistics for a dataset simply do not exist.

For datasets that represent categorized data, the statistics usually provide per-class and overall accuracy information. This is common practice with land cover datasets. For datasets such as elevation, values are given for horizontal and vertical accuracy. These values are usually given as a probability of being within a specified distance. Other datasets, such as those derived from MODIS data are validated by a team of scientists and in some cases the validation effort is incomplete or ongoing. When using these datasets, it is important to research the most current information available about the dataset's accuracy. This information is often available on the Internet.

Different methods for reporting accuracy exist and this is an active research area. For example, some of the newer accuracy methods provide information about the spatial distribution of error. Methods have also been developed using fuzzy statistics to indicate the severity of the error instead of using the traditional approach of noting a value as either correct or incorrect.

Summary

Remotely sensed and derived data are an invaluable asset for ecological modeling. These data provide broad area and repetitive coverage that is impractical to gather using field methods. This is a very dynamic field and as sensors and processing tools improve, these data will continue to become more plentiful, precise and accurate.

This article provides a brief overview of remote sensing and the types of data that can be derived to support ecological modeling. The intended audience for this article is the consumer, not producer of remote sensing products and it is not a comprehensive treatise on remote sensing. More in depth information about remote sensing and its application to ecological modeling can be found in the recommended reading section for this article.

Acknowledgments

The author thanks Kevin Koy, Richard Pearson, and Woody Turner for reviewing and offering suggestions to an earlier version of this text.

Further Reading

Canty MJ (2014) *Image analysis, classification and change detection in remote sensing: With algorithms for ENVI/IDL and python*, 3rd ed. Boca Raton, FL: CRC Press.
Copernicus Open Access Hub, European Space Agency portal for Sentinel data. https://scihub.copernicus.eu (accessed 14 December 2017).
Gergel SE and Turner MG (eds.) (2002) *Learning landscape ecology: A practical guide to concepts and techniques*. New York: Springer-Verlag.
GloVis, USGS data portal for Landsat and other NASA data sets. https://glovis.usgs.gov (accessed 14 December 2017).
Henderson FM and Lewis AJ (eds.) (1998) *Principles and applications of imaging radar: Manual of remote sensing*, Vol. 2. *Manual of remote sensing*, 3rd ed, Hoboken, NJ: Wiley.
Horning N, Robinson J, Sterling EJ, Turner W, and Spector S (eds.) (2010) *An introduction to remote sensing for biodiversity conservation*. New York: Oxford University Press.
INPE Image Catalog, Portal for INPE CBERS imagery. http://www.dgi.inpe.br/CDSR (accessed 14 December 2017).
JAXA Data Distribution Service, Portal for JAXA's earth observation satellite data. http://www.eorc.jaxa.jp/en/distribution (accessed 14 December 2017).
Jenson JR (2004) *Introductory digital image processing: A remote sensing perspective*, 3rd ed. Saddle River, NJ: Prentice Hall.
Jensen JR (2006) *Remote sensing of the environment: An earth resource perspective*, 2nd ed. Saddle River, NJ: Prentice Hall.
Kerr JT and Ostrovsky M (2003) From space to species: Ecological applications for remote sensing. *Trends in Ecology and Evolution* 18(6): 299–305.
Lillesand TM, Kiefer RW, and Chipman JW (2015) *Remote imaging and image interpretation*, 7th ed. New York: Wiley.
Maune DF (2007) *Digital elevation model technologies and applications: The DEM users manual*, 2nd ed. Bethesda, MD: American Society of Photogrammetric Engineering and Remote Sensing.
Rose RA, Byler D, Eastman JR, Fleishman E, Geller G, Goetz S, et al. (2015) Ten ways remote sensing can contribute to conservation. *Conservation Biology* 29: 350–359.
Schowengerdt RA (2006) *Remote sensing: Models and methods for image processing*, 3rd ed. San Diago, CA: Academic Press.
Thenkabail PS (2015) *Remotely sensed data characterization, classification, and accuracies*. Boca Raton, FL: CRC Press.
Turner W, Spector S, Gardiner N, et al. (2003) Remote sensing for biodiversity science and conservation. *Trends in Ecology and Evolution* 18(6): 306–314.
Ustin SL (ed.) (2004) *Manual of remote sensing, volume 4, Remote sensing for natural resource management and environmental monitoring*, 3rd ed., Hoboken, NJ: Wiley.
Wegmann M, Leutner B, and Dech S (eds.) (2016) *Remote sensing and GIS for ecologists. Remote sensing and GIS for ecologists: Using open source software*. Exeter: Pelagic Publishing.

Resilience

Ali Kharrazi, Advanced Systems Analysis Group, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria; Center for the Development of Global Leadership Education, The University of Tokyo, Tokyo, Japan

Introduction

The concept of resilience is increasingly employed across a diverse range of disciplines and has gained critical importance to researchers, practitioners, and policymakers focusing on dynamic socio-ecological systems. The etymology of resilience can be traced to Latin world "resilio," which translates to "jumping back." In the modern view, the word has a common-sense appeal and resonates with the ability of a system to be robust and to recover following a shock or disturbance. The literature surrounding the concept of resilience is highly scattered and contains numerous semantic, theoretical, and qualitative treatments in disciplines ranging from psychology, engineering, sociology, and disaster management. These treatments have increasingly heightened the need for empirical advancements of the concept beyond qualitative, theoretical, and index-based definitions and toward quantitative and systems-level measurements. Toward this end, the discipline of ecology and more specifically ecological modeling have been at the forefront of efforts toward the empirical advancement of the concept of resilience.

With the rise of the nonequilibrium paradigm within the ecological literature, the focus of the concept of resilience has moved away from "engineering resilience," that is, the stability of processes and speed of recovery, to a broader understanding of "ecological resilience," that is, the adaptability of the system and the existence of multiple stable states. Within this framework, the resilience of a system does not necessarily entail a return to an exact status quo ante and shocks or disruptions may lead the system to maintain its critical functions albeit at a different equilibrium. A return to the preshock equilibrium may be more relevant to the inanimate objects, for example, the resilience of buildings to an earthquake, while for socio-ecological systems such as food-webs, given the uncertainty of a changing environment, resilience includes an element of "adaptation." Adaptation is central to the notion that a resilient system exceeds beyond elasticity and that a system's resilience includes multiple basins of attractions. In this avenue, while a system shift from one basin to another basin, it maintains particular core functions even though it may operate under different structures and configurations.

The concept of resilience is predominantly perceived as a positive and desirable system quality in the literature. However, in reality, the specific context of the concept can infer otherwise and systems may be resilient to conditions that are negative and inherently undesirable. For example, the persistence of a eutrophicated lake or an ecological drought are highly resilient systems but inherently undesirable. In these cases, it is essential to examine approaches not to strengthen the resilience of the system but to invest in breaking the system's resilience and lead it toward a more desirable and positive stable equilibrium. Despite its duality of meaning, the concept of resilience is predominately treated as a positive and desirable attribute in the literature.

The wider literature on resilience often portrays the concept as a cost-free attribute without any associated trade-offs. However, within the ecological discipline, the trade-offs of resilience can be categorized as either temporal or spatial. In the first category, resilience has been viewed as the inverse return of time, that is, the required time for a system to return to an equilibrium stability, or as trade-offs between, for example, long-term slow-moving variables such as regeneration and short-term fast-moving variables such resource consumption. In the second category, empirical models arising from the ecological literature have identified trade-offs between spatial and topological configurations defining the resilience of a system.

Models of Resilience

The Adaptive Cycle

The concept of resilience rests on a paradigm emphasizing the ubiquity of change and the need for adaptation for long-term endurance. Based on this paradigm, recent empirically grounded studies have revealed that natural and social systems persist through complex phases of adaptation and renewal. In this avenue, the "adaptive cycle" can heuristically summarize part of this complexity and enhance our ability for resilience thinking.

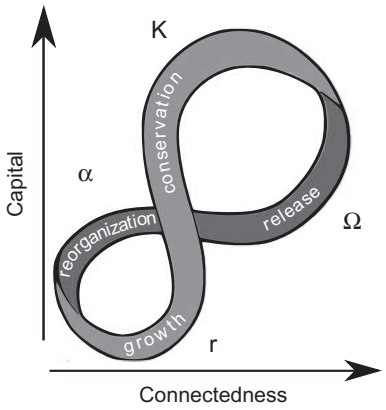


Fig. 1 The four phases of an adaptive cycle.

As illustrated in Fig. 1, the adaptive cycle illustrates a two-dimensional loop-like figure plotted against two axes. The vertical y-axis reflects system complexity and can be measured through an ecological goal function. These goal functions may include, the system's "potential for change" or "accumulation of capital resources," and can reflect the system's capacity to adapt based on alternative future options. Based on this axis, the system, for example, increases its potential by accumulating wealth through novelties that generate returns and following a disturbance depletes its potential to reorganize and adapts to new environmental conditions. The horizontal x-axis reflects the degree of organization or rigidity in a system and can be measured by connectedness. Against the above two-axis, the adaptive cycle depicts a cyclic pattern, inspired from the infinity symbol, which is designated into four phases: new beginnings and growth (r), conservation and status quo ante (K), dissolution and confusion (Ω), and reorganization and innovation (α). These four phases are arranged in two loops, a slow front loop represents the current trajectory while a more rapid back loop represents creative change.

The resilience of a system varies through the adaptive cycle. In the (α) phase the system is described as maintaining low connectedness, necessary for rearranging previously rigid structures, and high resilience, necessary for testing such new rearrangements. On the reverse side in the conservation (K) phase, due to its rigidity and connectedness toward the maximization of resource exploitation, the system maintains lower levels of resilience. The more recent literature on the adaptive cycle, view a resilient system as the ability to successfully navigate all four phases of the adaptive cycle. In this view, a resilient system is able to navigate both the front loop, that is, the long phases of steady growth where available resources are utilized, and also equally important the back loop, that is, the shorter turbulent phases of reorganization and transformation where innovations are sought in the aftermath of a shock or disturbance to the system.

The adaptive cycle maintains high flexibility in its multidimensionality where its axis can be subjective and allow the researcher to devise his/her own worldviews scenarios. A significant strength of the adaptive cycle is its ability in identifying different system trade-offs and to distinguish both positive and negative attractors according to the desired state of the system. This has enabled the application of the adaptive cycle for examining resiliency in various socio-ecological studies.

The Ecological Information-Based Network Approach

The ecological information-based network approach is concerned with the emerging systems-level structure of a network of flows, for example, circulation of material, nutrients, water, or information. The central assumption of this approach is that the resilience of system relies on the topology and magnitude of its flow pathways and is largely defined as a balance between the network configurations of efficiency and redundancy. This approach is applicable to systems that can be depicted as a network web structure, that is, a group of compartments connected by directed and weighted flows depicting the inner exchange and functioning of a system.

While in the long-term, natural systems have been observed to increase their efficiency at the expense of their redundancy, the trade-off between these two system variables may depend on agency behavior, environmental constraints, and shocks or disruptions directed at the system. The efficiency of a system can be defined as the degree of articulation or constraints of flows in a network. In more objective and quantifiable terms, the average mutual information is used to define the network efficiency of a system as:

$$\text{Efficiency} \quad = \quad \sum_{i,j} \frac{T_{ij}}{T_{..}} \log \frac{T_{ij} T_{..}}{T_{i.} T_{.j}} \tag{1}$$

Here, T_{ij} is a flow from agent i to agent j, $T_{i.} = \sum_j T_{ij}$ is the total flow leaving agent i, $T_{.j} = \sum_i T_{ij}$ is the total flow entering agent j and the sum of all flows in the system, $T_{..} = \sum_{ij} T_{ij}$, is known as the total system throughput (TST).

The redundancy of a system, the countering variable to efficiency, can be defined as the degree of freedom or overhead of flows in a network. Network redundancy reflects the diversity of pathways in a system and is critical for a system's flexibility and capacity to

adapt to changing environmental conditions arising from shocks and disruptions. In more objective and quantifiable terms, the conditional entropy is used to define the network redundancy of a system as follows:

$$\text{Redundancy} \quad = -\sum_{i,j} \frac{T_{ij}}{T_{..}} \log \frac{T_{ij}^2}{T_{i.} T_{.j}} \tag{2}$$

Both of these metrics, efficiency and redundancy, are dimensionless and based on units of information depending on the base logarithm used in their calculation, for example, *bits* if the base 2 logarithm is used or *nats* if the natural logarithm is used.

Intuitively, following a disruption, networks with more diverse connections are more flexible in rerouting their flows and maintaining critical functions. Conversely, a more efficient network with a minimal number of well-organized connections can concentrate its capacity for growth and development. This trade-off can be situated in a broader discussion on the phenomenology of ecological growth versus development. As illustrated in Fig. 2, overly redundant networks may be incoherent and lacking the efficiency to grow, while overly efficient networks may be brittle and prone to collapse when subjected to stress—at either extreme of too little efficiency or too little redundancy, the resilience of the system goes to zero. The ratio α is a convenient way to express the degree of which property dominates the system at a given time. Departing from the relative order and invoking the Boltzmann measure of its disorder, the level of a system's Theoretical Resilience can be expressed as:

$$\text{Theoretical resilience} = -\alpha \log(\alpha) \tag{3}$$

From Eq. (3), a maximum value for theoretical resilience is derived as $1/e \approx 0.3704$ (independent of the logarithm's base). Theoretical resilience (Eq. 3) signifies the balance between efficiency and redundancy as a single metric and therefore is useful in exploring and comparing the configurations of heterogeneous networks. Interestingly, empirical investigations have determined that natural networks, for example, ecosystems and food-webs, lie in close proximity to this maximum. Based on these findings, researchers have advanced the idea of bio-mimicry and the need to modify network configurations of various socio-ecological systems toward this normative optimal range for maintaining their resilience. The maximum resilience value, however, should be seen as a theoretical benchmark and a normative balance between redundancy and efficiency of real heterogeneous systems can depend on the environment, stage of development, levels of stresses, and the underlying mechanism forming the network. In this avenue, further research is warranted in evaluating a normative balance for socio-ecological systems that can be grouped based on shared characteristics and local conditions, for example, food-webs, urban networks, and natural resource management.

Statistical Evidence of Resilience

Common system-level properties of resilient systems identified across socio-ecological research include modularity and diversity. By modifying a system and increasing the value of these properties, the system can better withstand an anticipated shock or disruption and increase its resilience and capacity for recovery. Modularity can be related to the degree of the connectivity of a system, whereas

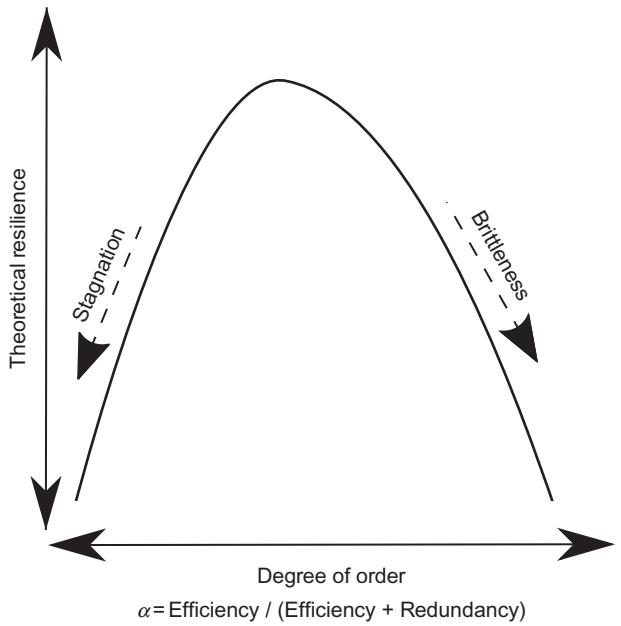


Fig. 2 The conceptual model of the ecological information-based network approach.

diversity relates to the variety among elements of the system. These statistical approaches are highly versatile in their application to increase the resilience of systems to anticipated disturbances or shocks. While individually, modularity and diversity have been employed in the literature, their concurrent use, overlaps, and trade-offs are important research frontiers.

Modularity

Modularity is a system property which measures the degree to which densely connected compartments within a system can be decoupled into separate communities or clusters which interact more among themselves rather than other communities. In a highly interconnected system with low levels of modularity, a shock to one compartment may cascade to other compartment and thus increase the risk of a system-wide collapse. Conversely, in a system with high levels of modularity, a disturbance to one component can be better contained and the disturbance will be less likely to spread to other components. The idea of modularity is widely practiced in the management of various systems, for example, forest firebreaks prevent the spread of fire or port quarantines prevent the spread of epidemics or biological pests. As an important attribute of resilience, modularity has been the subject of increasing research in ecological research. Recent empirical work conducted on food-webs have confirmed that food-webs are more compartmentalized than a null model where species interact with equal probability with other species.

The literature on modularity describes numerous methods for measuring the modularity of a system. Furthermore, in any compartmentalized system, community structures may exist through numerous possible partitions and therefore modularity can be evaluated by seeking a "best fit" for modular partitions. Despite the challenges of a precise mathematical definition, the modularity maximization method is the most widespread in usage and accuracy. In this approach, the modularity of a network partition is evaluated by comparing its number of links against a null network model, that is, an equal number of nodes, links, and degree distribution but with random links among the nodes. From the above definition, a modularity function is defined that can measure the quality of the introduced partitions (P) as a community:

$$P = \frac{1}{T_{..}} \sum_{ij} \left[T_{ij} - \frac{T_{i.} T_{j}}{T_{..}} \right] \delta_{c_i c_j}$$

The expected fraction of links is measured with the assumption that the probability that nodes i and j are connected is $\frac{T_{i.} T_{j}}{T_{..}}$. Here, $T_{i.}$ and $T_{.j}$ are respectively the output and input strengths of nodes i and j, $T_{..}$ is total strengths of the network, and δ is Kronecker's delta where $\delta_{c_i c_j} = 1$ if nodes i and j are in the same community and $\delta_{c_i c_j} = 0$ if otherwise.

Diversity

Diversity is an important characteristic of resilient systems and the concept has been widely used across different disciplines. At its simplest, diversity can be defined as the degree of variation. This may include functional diversity, that is, the degree of the variations of components which maintain similar functions, or response diversity, that is, the degree of the variations of components which exhibit different responses resulting from disruptions. Systems maintaining such diversities can in principle be more flexible when faced with a disruption or shock. For ecologists, diversity is seen as an essential component for ensuring ecosystem resilience and as a long-term survival strategy for natural ecosystems. In ecological studies, for example, abundance, i.e., the proportional abundance of a species or trait, and richness, i.e., the number of species or traits, are important attributes of diversity. While there is no single mathematical definition of diversity in systems, the Shannon–Weaver and Simpson indices are the two most widely used approaches in the literature which combine measures of richness an abundance.

The Shannon–Weaver index is defined as:

$$H = -\sum_{i} p_i \ln (p_i)$$

The Simpson index, more commonly written in its inverse form, is defined as:

$$D = \frac{1}{\sum_i p_i^2}$$

Here, p_i represents the share of category i in the total mix of categories. The higher the value of the H or D, the more diversity is present within a system. The Shannon–Weaver index takes its roots from the information science literature. The Simpson index is also known as Herfindahl index or the Herfindahl–Hirschman index in the economic literature. These two common indices may result in similar measurements of diversity when examining simple systems. However, when examining systems of higher complexity, the choice of index becomes more important, as each index may maintain certain sensitivities, for example, sensitivity to rarity or abundance, in its measurement of diversity.

Summary

The concept of resilience has a common-sense appeal and is increasingly employed within a range of literatures, for example, in psychology, engineering, sociology, and disaster management, each of which maintain numerous semantic and theoretical

definitions. In the wider literature, misconceptions surrounding the concept of resilience include, for example, that resilience does not include an element of adaptability, resilience is an inherently positive and desirable attribute, and that resilience is without trade-offs. Within the discipline of ecology, the emphasis has been toward the empirical advancement of the concept of resilience and toward addressing such misconceptions. Empirical definitions of resilience are more communicable as quantitative measurements and are beneficial in revealing trade-offs in policy and practice. In this avenue, the adaptive cycle, the ecological information-based network approach, and statistical characteristics such modularity and diversity are important and promising directions in advancing resilience research.

The adaptive cycle's strength lies in its flexibility and descriptive trade-offs relevant to the management of a system's resilience. The adaptive cycle, however, does not maintain a strong quantitative approach and it is anticipated that future research can provide new insights in this direction. The ecological information-based approach maintains strong analytical and quantitative definitions and for future research, it is anticipated that more case-study research may reveal normative values in the network efficiency and redundancy of heterogeneous socio-ecological systems. Finally, despite their growing usage in the ecological literature, the influence of modularity and diversity on resilience remains an under-researched area. More specifically, trade-offs and overlaps between modularity and diversity can be considered as important research frontiers.

Further Reading

Fath BD, Dean CA, and Katzmair H (2015) Navigating the adaptive cycle: An approach to managing the resilience of social systems. *Ecology and Society* 20(2).
Guimerà R, et al. (2010) Origin of compartmentalization in food webs. *Ecology* 91(10): 2941–2951.
Morris EK, et al. (2014) Choosing and using diversity indices: Insights for ecological applications from the German biodiversity exploratories. *Ecology and Evolution* 4(18): 3514–3524.
Ulanowicz RE, Goerner S, Lietaer B, and Gomez R (2009) Quantifying sustainability: Resilience, efficiency and the return of information theory. *Ecological Complexity* 6(1): 27–36.

Socioecological Systems☆

I Petrosillo, R Aretano, and G Zurlini, University of Salento, Lecce, Italy

Socioecological Systems (SESs)

There are no social systems without nature, and few ecosystems without people, such as some large wilderness areas, recognizable for their intactness and for the very low population density. An area must retain at least 70% of its historical habitat extent (500 years ago) and five people per km^2 to be considered a wilderness area, and at global level they are represented by five areas: Amazonia, Congo, New Guinea, the Miombo–Mopane woodlands, and the North American deserts. In the case of oceans, the marine pelagic environments are characterized by some areas free of human influences, even if resource competition due to human exploitation of prey species can cause both nutritional stress and negative behavioral changes in pelagic predators.

Systems where social, economic, ecological, cultural, political, technological, and other components are strongly linked are known as socioecological systems, emphasizing the integrated concept of the 'humans-in-nature' perspective. Socioecological systems (SESs) are truly interconnected and co-evolving across spatial and temporal scales, where the ecological component provides essential services to society such as supply of food, fiber, energy, and drinking water. As a result, socioecological systems have become an emerging focus in scientific and policy arenas.

Landscapes change constantly from natural and anthropogenic drivers, and land use and land cover changes by humans have been identified as a primary effect of humans on natural systems. These changes underlie fragmentation and habitat loss, which are the greatest threats to biodiversity and ecosystem services. The complex interactions between development decisions and ecosystems, and how the consequences of these decisions may then influence human values and subsequent decisions is an important area of research interest.

Given the results of the Millennium Ecosystem Assessment, reciprocal influences among humans and the climate, biota, and ecological goods and services of the world have become both stronger and more widely recognized. In this context there has also been the acknowledgment that in the majority of ecosystems, structure and function are now determined primarily by human interactions, perceptions, and behaviors, so that nowadays it is more appropriate to think of socioecological systems combining approaches from both environmental and social sciences.

The socioecological system theory sprang from the recognition of close interaction between society, in terms of social–economic system, and natural system. For this reason, an interdisciplinary approach is needed: in the past the social–economic approach was distinct from that of ecology; the stereotypical economist might say 'get the price right' without recognizing that price systems require a stable context where social and ecosystem processes behave 'nicely' in a mathematical sense – that is, they are continuous and convex. The stereotypical ecologist might say 'get the indicators precise and right' without recognizing the surprises that nature and people inexorably and continuously generate. These simple approaches are often attractive because they seem to replace inherent uncertainty with the fictitious certainty of ideology or precise numbers. But the theories implicit in these approaches ignore multistable states that characterize SESs.

SESs show a complex and uncertain nature rooted in the complex systems theory that refers to interrelated theories (catastrophe theory, chaos theory, information theory, hierarchy theory, and self-organization theory) that have originated from different scientific disciplines. Despite their traditional scientific disciplinary origins, they have provocative implications across disciplines and fields and, more generally, for the way we understand various types of phenomena as well as the role of learning in planning and policymaking.

In the past, the usual way to study complex phenomena was based on simplifying them through analytical reductionism (describing them as simple systems, machines) or by aggregating and averaging through statistical analysis (describing them as unorganized complex systems). Since an SES is made up of many different parts that interact to form a more complex entity, its dynamic can be explored using an holistic approach because it does not focus on a detailed understanding of parts, but on how key components contribute to the dynamics of the whole system. But complex systems, such as SESs, exist at a threshold between order and chaos, because they are too complex to be treated as machines and too organized to be assumed random and averaged. An example could be the slow erosion of key controlling processes that can abruptly flip an SES into a different state that might be irreversible (the gradual loss of species important for pollination could cause the slump of an economy based on agricultural products).

It is relevant to understand the human sources of ecological change. To do this, we must understand the driving forces motivating human actions. Driving forces are the underlying causes that influence and direct human activities. These forces, either directly or indirectly, result in changes in ecosystems, which can degrade ecosystem capability to provide goods and services. The roots of these forces can be economic, political, sociocultural, and/or legal, and rarely occur in isolation, but rather act in conjunction with others. Direct driving forces, such as mining or agricultural practices, are easily recognizable as they often have an immediately discernible effect. Indirect driving forces are less identifiable; however, they have no less of an impact on ecosystems since they influence people's actions. For example, legislation can encourage people to mine rather than farm an area and influence how they will mine. There are several examples in the world: Britain's solution to rising urban pollution levels in the 1800s was to

☆*Change History:* March 2015. I Petrosillo R Aretano and G Zurlini updated section text and further reading.

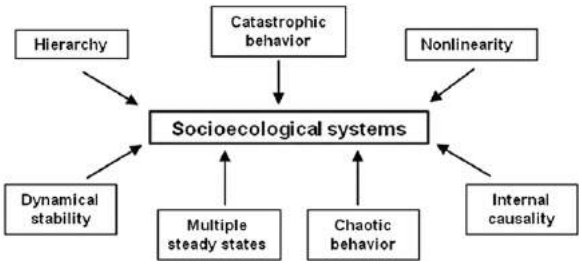


Fig. 1 Characteristics of complex adaptive SESs.

increase the height of factory chimneys. This only postponed the problem in England, while it then introduced problems in Scandinavia. This was only a temporary solution and only at the local scale. The source of the problems, the emissions from industrialization, remained unchanged in quantity or quality. In Europe, the WTO has required the end of European preferential treatment of some banana-producing nations. The opening of trade within the EU could drive land-use changes in other banana-producing countries. The WTO has certainly foreseen this possible outcome. However, it is simply considered a shift of production location based on economic considerations, disregarding both the social and ecological changes that can be driven by such a shift.

Human society is able to choose alternative development scenarios. Initiatives toward development might cause social and ecological changes and bring surprises and uncertainties. It is necessary to plan strategies that enhance system's adaptive capacity to change rather than simply maximize resource consumption. In the case of sweeping surprises, partial solutions, only economic, or social or ecological, bring the loss of benefits coming from the integration among economic, social, and ecological processes. The base of sustainable policies and investments should be turned toward knowledge integration, with the aim to obtain a comprehension based on different viewpoints.

Key Features of SESs

Complex systems theory offers a more sophisticated understanding of the structure and dynamics of both social and ecological systems than the relevant 'normal' scientific disciplines.

The properties of SESs are (Fig. 1):

- *Nonlinearity*. They behave as a system and cannot be understood isolating their components.
- *Hierarchy*. They are hierarchically nested and the 'effect' exercised by a specific level involves a balance of internal (self-control) and external controls involving other hierarchic levels in a mutual causal way. Such interactions cannot be understood by focusing only on one hierarchical level (multiple scales of interest).
- *Internal causality*. This is due to self-organization.
- *Dynamical stability*. There are no equilibrium points for the system.
- *Multiple steady states*. There is not necessarily a unique preferred system state in a given situation, because multiple attractors can be possible in a given situation.
- *Catastrophic behavior*. It is typical of SESs, in terms of (1) bifurcations – moments of unpredictable behavior; (2) flips: sudden discontinuity; and (3) Holling four-box cycle. (exploitation – conservation – release – reorganization).
- *Chaotic behavior*. The human ability to predict the future is always limited.

The complexity of a system is the result of the interaction among a great deal of components that cause new, emergent, and unexpected properties. The analysis of these systems suggests that the possibility for a sustainable development depends on changing perception of human society regarding complex systems. Thus, an essential goal is to change the perception and the way of thinking of social actors, moving their attention from increasing productive capacity to increasing of adaptive capacity. This means that it is necessary to turn social actors' attention to a view where society and nature are coevolving in the biosphere.

SES theory was pioneered in the 1980s by the Resilience Alliance, a voluntary organization of scientists of various disciplines, to explore the SESs' dynamics and their possible evolutions, but there are several scientific schools interested in their study. These theories are based on concepts as adaptive cycles, resilience, adaptability, transformability, and hierarchy (panarchy), and aim to provide knowledge basis to manage complex adaptive systems and to achieve sustainable development in theory and in practice. The knowledge of these aspects should improve natural systems management and their capacity to support human and natural capital.

The novelty of these theories concerns natural, disturbed, and managed ecosystems, identifying which are the key features of ecosystem structures and functions (Table 1):

- Change is episodic, with periods of slow accumulation of natural capital such as biomass, physical structures, nutrients, punctuated by sudden releases and reorganizations of this biotic capital, as the result of internal or external natural disturbances, or human-imposed catastrophes. Rare events, such as hurricanes or the arrivals of invading species, can

Table 1 Key features of socioecological system structures and functions

Change	Change is episodic, with periods of slow accumulation of natural capital punctuated by sudden releases and reorganizations of biotic capital
Spatial attributes	They are neither uniform nor scale invariant. There are several different ranges of scales, each with different attributes of architectural patchiness and texture and each established and sustained by a specific set of abiotic and biotic processes
Stability domain	Ecosystems do not have a single equilibrium and homeostatic controls that keep them near it, rather, multiple equilibria commonly defining different functional states within the same stability domain
Policies and management	Policies and management that apply fixed rules, independently of scale, could lead systems to lose resilience

unpredictably shape system structure at critical times or location, leading to an increase in fragility. In this way, these rare events can modify the future of the systems for long periods, even if irreversible or slowly reversible states can exist; once the system flips into another state, only an explicit external management intervention could allow the system to come back to its previous self-sustaining state, but its full recovery is not assured.

- Spatial attributes are discontinuous at all scales, from the leaf to the landscape to the whole planet. There are several different ranges of scales, each with different attributes of architectural patchiness and texture and each established and sustained by a specific set of abiotic and biotic processes.
- Ecosystems do not have a single equilibrium and homeostatic controls that keep them near it, but rather multiple equilibria commonly defining different functional states within the same stability domain. Normal movements of state variables maintain structure, diversity, and resilience. Stochastic forces and interactions between fast and slow variables mediate the movements of variables among those equilibria.
- Policies and management that apply fixed rules (e.g., maximum sustainable yield), independently of scale, could lead systems to lose resilience, that is, systems break down in the face of disturbances that previously could be absorbed.

How Humans and Environment are Coupled: Examples

There are several examples on how human and environment systems are coupled and how human choices and the consequent environmental effects influence each other. The following are three examples regarding southern Yucatan, Arctic region, and Eastern Europe.

Southeastern Mexico retains parts of the largest continuous expanse of tropical forests in Middle America. One part of the 22 500 km^2 southern Yucatan peninsular region experienced extensive, state-led development beginning in the late 1960s, causing deforestation with consequences on human well-being. In this region, almost all farmers cultivate maize for subsistence and, increasingly, have undertaken commercial chilli production, giving rise to a fragmented landscape of opened and successional forest land. Increasing reliance on commercial chilli production has raised household income but simultaneously driven large swings in this income. This is because chilli is water, pest, and disease sensitive, and the price in the region is highly variable. At the same time, the area is characterized by two main environmental hazards: water stress and hurricanes. The natural land covers, seasonal tropical forests, are adapted to water stress, because they drop foliage during the dry season, while farmers respond to this stress by taking an early dry-season catch crop. On the other hand, severe hurricanes and subsequent dry-season fires knock down large stretches of forest that need a long time to regrow. Hurricanes arrive during the main harvest period, damaging crops, especially chilli, by winds, rain, and floodwater, because a fragmented landscape creates more forest edges exposed to severe winds, damaging near-edge trees. This more open landscape causes less wind protection for crops, with consequences on local economy and human well-being.

Environmental and social changes have had and are expected to have significant effects on coupled human–environment systems in the Arctic. The Arctic Monitoring and Assessment Program have stated that although the Arctic is a relatively clean environment, it continues to suffer from significant pollution hazards, especially with regard to heavy metals and persistent organic pollutants. At the same time, native Arctic peoples have also experienced significant social changes over the past three decades, establishing new relationships between local and national governments, becoming more closely connected to external markets and ways of life, and asserting their identity, rights, and culture in legal and policy forums. Three kinds of stressors interest the Arctic region: (1) climate change with consequences on snow cover, sea ice, and extreme weather events; (2) environmental pollution (potentially toxic organic compounds, acids, metals and radionuclides), some models used in the Scenarios Network for Alaska and Arctic Planning (SNAP) research took into account a steady increase in carbon dioxide emissions from fossil fuels combustion over the first several decades of the 21st century; and (3) societal trends in terms of consumption, governance and regulation, and markets. These represent threats to human health and well-being, indigenous cultures and food security, and human settlements and development. The Arctic region is an example of cross-scale systems interaction, because the decisions taken in different regions affect people living in the Arctic region: global market, climate change, and environmental pollution.

Traditional farming landscapes are an example of socioecological systems, since they are the result of the interaction and co-evolution over centuries between people, which shaped the land through their activities, and the nature that, in turn, gave people a variety of ecosystem services. Many rural communities in Japan, India, China, and Europe are characterized by a well-developed

system of traditional ecological knowledge to assess the quality of the ecosystems goods and services and to sustainably manage natural systems; this resulted in landscapes with high aesthetic, ecological, and cultural values. However, these socioecological systems are rapidly changing due to the social, economic, cultural and institutional changes in the society. In fact, many traditional farming landscapes have come under pressure from by a large number of socio-economic challenges such as the growth of population, economic, industrial and infrastructural development, globalization. These changes have made traditional subsistence agriculture economically unprofitable and unable to meet the increased social demands, leading to landscape changes such as land-use intensification or land abandonment and eroding many valuable cultural and ecological elements as well the linkages between local communities and their ecosystems.

The History as a Basis for the Future of Socioecological Systems

The history of human-dominated socioecological systems is one of successive crises that were either successfully addressed, leading to sustainability, or not, leading to collapse, and the goal of studying history has always been to understand the past in order to understand and deal with the present and the future. The assessment of the vulnerability of modern socioecological systems to future human activities and climate change can be greatly improved by (1) knowing the rates and directions of past trajectories in key processes, such as land cover, soil erosion, and flooding; (2) defining and analyzing how thresholds have been transgressed in the past; and (3) deducing the natural or pre-impact patterns of environmental variability. Therefore, the past provides the means to test the models upon which we depend for future projections and scenarios. The present nature and complexity of socio-ecological systems are heavily contingent on the past; we cannot fully appreciate the present condition without going back decades, centuries, or even millennia.

The complexity theory, with related concepts such as nonlinear change, feedback and regime shifts, suggests that human activities and environmental change should be viewed together as a co-evolutionary and adaptive process. Positive feedback loops may lead to a conditioning of landscapes that makes them more sensitive to new perturbations. Hence, some historical societies, like those on Easter Island, became more prone to collapse through continuing resource depletion and ecological degradation. Others, such as the Akkadian society of Mesopotamia, became increasingly vulnerable to climate perturbations as their dependence on irrigated cultivation increased.

SESs' Management

Environmental management is another field of research and practice integral to any discourse on knowledge and social learning for environmental policy and decision making. A simple definition of environmental management states that it consists of "actual decisions and action concerning policy and practice regarding how resources and the environment are appraised protected, allocated, developed, used, rehabilitated, remediated, and restored." Much of current environmental management focuses on the integration of social and ecological systems, as our understanding of environmental issues has evolved. In this context, environmental decision making has to address the complexity of both ecological systems and interdependent human organizational and institutional systems. Several scholars have set a profound and necessary precedent with their work, explicitly integrating the study of natural resources with human organizations and institutions to focus research and intervention on integrated SESs. In recent decades, efforts to address some of the paradoxes in resource and environmental management have required an evolution in thinking about environmental science and decision making. The result has been a shift from reductionism, command and control science and management, to a more integrated, adaptive, systems-based approach. Integral to this more systemic approach to environmental decision making has been the incorporation of an emerging body of theory often referred to as complex systems theory.

Complex systems theory has offered a more sophisticated understanding of the structure and dynamics of both social and ecological systems than the relevant 'traditional' scientific disciplines. Even this integrated, systemic view of SESs does not explicitly acknowledge the complexity of the process of social learning for decision making within SESs. The integration of planning and governance theory with complex and critical systems thinking, as well as with social learning, points to new opportunities in the study of environmental decision making.

Attempts to extend insights from the field of social learning to the practice and study of resource and environmental management have also contributed to the discourse on social learning for environmental planning and decision making; for example, how public participation in environmental assessment processes provides opportunities for social learning.

Works in the field of environmental management have highlighted the importance of integrating social and ecological systems, highlighting the importance of social learning for the purposes of environmental decision making.

Governance is another main field of practice in which the linkage among knowledge, learning, and intervention in the context of environmental decision making is prevalent. Governance focuses directly on the political side of the decision making. There are several definitions of governance; however, all of these speak to a conception of political economy, and more generally decision making and knowledge for intervention, that is more broad-based, flexible, and evolving than traditional models of public decision making through government intervention.

Complex systems approaches could provide, and are already providing, governance stakeholders with philosophical and methodological underpinnings and practical heuristics to look critically at the interface of learning and intervention. The governance literature highlights the importance of politics and pluralism in decision making.

Existing Socioecological Systems Frameworks

An effective management and understanding of the different aspects of an SES needs the development of an integrated framework that considers the feedback and interactions between and within social, economic and ecological systems. There are several existing socioecological system frameworks that reflect the variety of research fields involved in the study of an SES and that can be applied according to the problem to be studied and the way in which the social-ecological system is conceptualized.

Among them, for example, the DPSIR (Driver-Pressure-State-Impact-Response) conceptual model is one of the framework that shows the cause–effect relationships between environmental and human systems and that has been used for analyzing and assessing the social and ecological problems of systems (above all aquatic systems) subject to anthropogenic influence.

The Millennium Ecosystem Assessment introduced a framework for analyzing SESs connecting drivers, ecosystem services and human well-being. In addition to ecological processes, also social factors such as skills, management regimes, and technology are involved in ecosystem services production. In particular, the framework makes more visible the links between the spatial and temporal provision of ecosystem services (supply) and the beneficiaries where corresponding well-being is appreciated (demand). For this reason the ecosystem service approach is very useful for a better understanding of ecological functioning, social structures, trade-offs and synergies between services, benefits on human well-being, and how these aspects feed back to influence governance and policy and, therefore, SESs and their services. As a consequence, this framework has considerable influence in policy and scientific communities supporting problem solving and proactive management.

A prominent and widely applied framework that focuses attention to socioecological systems is the vulnerability framework that provides the broad classes of components (exposure, sensitivity, resilience) and linkages that comprise a SES's vulnerability to multiple environmental and human changes and hazards. Vulnerability is a highly complex phenomenon with both biophysical (e.g., climatic conditions, natural hazards, topography, land cover) and socio-economic (e.g., demography, poverty, trade, employment, gender, governance) factors that determine its sensitivity to any set of exposures and influence the potential for harm. For this reason, the framework considers linkages to the broader human (social/human capital and endowments) and environmental (natural capital/biophysical endowments such as soils, water, climate, minerals, ecosystems) conditions and processes operating on the SES under study. These conditions can influences the responses (coping, impacts, adjustments, and adaptation). In particular, the social and biophysical responses or coping mechanisms influence and feed back to affect each other, so that a response in the human subsystem could make the biophysical subsystem more or less able to cope, and vice versa.

Another general framework called social-ecological system framework, conceptualizes that each of the individual SESs is composed of four core first-level subsystems: resource systems (e.g., coastal fishery, protected area), resource units (lobsters, wildlife), users (fishers, tourists), governance systems (organizations and rules that govern fishing on that coast, management authority of protected area). Each core subsystem is divided into lower levels made up of multiple second-level variables (e.g., size of a resource system, level of governance,) which are further composed of deeper-level variables. These subsystems are relatively separable but interact and affect each other to produce outcomes at the SES level, which in turn feed back to affect these subsystems and their components, as well other larger or smaller SESs. The framework is useful to identify the multitier hierarchy of variables for analyzing and understanding the functioning of an individual SES and the reasons why certain e.g. management actions and particular policies enhancing sustainability and succeed in one SES and fail in another. This framework has been applied above all for explaining sustainable outcomes in the context of forestry, fishery, and water resources.

SESs and Social Learning

The literature on social learning attempts to make operational many of the complex epistemological issues around the nature of knowledge and the process of learning.

A useful and less theoretical definition underlines that "social learning means more than merely individuals learning in a social situation … (they) envision a community of people with diverse personal interests, but also common interests, who must come together to reach agreement on collective action to solve a mutual problem… it is the process by which changes in the social condition occur – particularly changes in popular awareness and changes in how individuals see their private interests linked with the shared interests of their fellow citizens." Social learning is intended to help improve the quality and wisdom of the decisions when faced with complexity, uncertainty, conflict, and paradox, and the notion has begun to be applied in a variety of complex decision-making contexts, including environmental management and planning. Environmental planning and management are often described as complex and highly uncertain and, from this perspective, management cannot be seen as the search for an optimal solution to a single problem but rather as an ongoing process of adaptation, learning, and negotiation. Thus, to manage complex adaptive systems, it is necessary to create a learning atmosphere, encourage systemic thinking about complex problems, discourage competitive behavior among stakeholders, and focus on 'desirable and feasible change' rather than attempting to achieve absolute consensus on management issues. An example is given by the application of social learning to river basin

management, considered as the capacity of different authorities, experts, interest groups, and the public to manage their river basins effectively. Often, limitations of existing institutions, to consider multiscale, participatory forms of governance for groups involved in river basin management are present. These applications show that social learning processes can improve stakeholders' awareness and participation in environmental deliberation and decision making and therefore contribute to practical change in environmental management as well as institutional change.

Social Adaptive Responses to Ecosystem Change

Despite the lack of theories linking the creation of ecological knowledge from observations and understanding to its incorporation into resources use, Fig. 2 provides a conceptual model of possible responses to a crisis situation. In this context, the term crisis broadly refers to a large perturbation, and it may be human made (resource collapse) or natural (hurricane).

Three generic responses are possible when a crisis occurs:

(1) no effective responses;
(2) response without experience, in which the institution, a government agency or an informal local management institution, responds to a crisis but does not have previously tested policies, with accumulated ecological knowledge, at its disposal; and
(3) response with experience, in which the institution has previous experience with a crisis of that kind and management policy used on previous occasions.

In centralized and bureaucratized management systems, the 'no effective response' is the management reaction that often characterizes brittle or fragile institutions. Such a response allows accumulating up the panarchy (a hierarchy of socioecological systems), creating the conditions for a larger-scale crisis, both political and ecological. Response without experience is a frequently seen reaction to a crisis, and it could lead to institutional learning. This is the case in which the crisis is a true surprise, so that the institutions will have no previous experience with it, or the crisis may have been predictable but be of magnitude that had never been experienced in that area.

The response with experience is possible if the memory of the experience provides a context for the modification of management policy and rules, so that the institution can act adaptively to deal with the crisis.

The more useful management to be applied to SESs is adaptive management, and the more useful assessment is based on the integration of different disciplines.

Adaptive management needs to at least maintain political openness, but usually it needs to create it. Consequently, adaptive management must be a social as well as scientific process. It must focus on the development of new institutions and institutional strategies just as much as it must focus upon scientific hypotheses and experimental frameworks. Adaptive management attempts to use a scientific approach, accompanied by collegial hypotheses testing to build understanding, but this process also aims to enhance institutional flexibility and encourage the formation of the new institutions that are required to use this understanding on a day-to-day basis.

Adaptive management approach needs the definition of SESs' boundaries encompassing multiple spatial scales of socioecological processes and engaging a variety of stakeholders to ensure that management interventions and policy strategies could reflect many different socioecological values and viewpoints. Moreover, since institutions, policies and goals are established over a particular time, they need to be continuously monitored, assessed, and re-evaluated or adapted, as circumstances change, knowledge about the SES is increased and learning takes place.

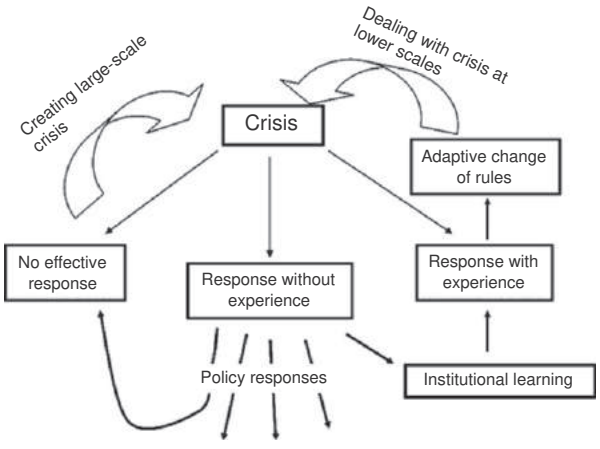


Fig. 2 Three generic responses to resources and environmental crisis. Most responses fall into categories of (1) no effective response, which can lead to larger-scale surprises; (2) reacting with no memory or experience; or (3) responding through learning.

Conclusions

The SES theory recognizes that human dimension shapes and is shaped by environment, so that social and ecological systems are interconnected and coevolving across scales. Since both social and ecological systems are dynamical, the associated policies, including economy that represents the main driver, have to be dynamical; governance systems based on policies that try to control few ecological processes (command and control) do not allow the sustaining of the capacity to deal with change, producing fragile SESs.

A central aspect in dealing with SESs is that they are characterized by cross-scale interactions, both temporal and spatial, and the same is applicable to their governance, because decisions taken at one place in the past and in the present can affect people currently or in the future living elsewhere. The approach used to dealing with SESs assigns surplus value to a social system that clears the limits of sociological approach. In this perspective, social system includes also economic, institutional, and management aspects, by setting the rules. This gains in importance because, according to the 'tragedy of the commons,' the prediction is that in the absence of rules governing who can use natural resources (open access), individual users pursue their own interests, as in the case of ecosystem goods and services.

Since systems are uncertain and complex, the management needs to be flexible and adaptive, recognizing that it is impossible to control so many variables. Strategically, the only way to manage SESs is to adopt a systematic process for continually improving management policies and practices by learning from the outcomes of operational programs, by evaluating alternative hypotheses about the system being managed.

See also: Human Ecology and Sustainability: Urban Systems; Human Population Growth

Further Reading

Berkes, F., Folke, C. (Eds.), 2003. Navigating social-ecological systems: Building resilience for complexity and change. New York: Cambridge University Press.
BinderC.R., HinkelJ., BotsP.W.G., Pahl-WostlC., 2013. Comparison of frameworks for analyzing social-ecological systems. Ecology and Society 18 (4), 26.
CostanzaR., 2007. The need for a transdisciplinary synthesis of history. Ambio 36 (7), 521.
Diamond, J. (Ed.), 2005. Collapse: How societies choose to fail or survive. London: Allen Lane.
GameE.T., GranthamH.S., HobdayA.J., PresseyR.L., LombardA.T., BeckleyL.E., GjerdeK., BustamanteR., PossinghamH.P., RichardsonA.J., 2009. Pelagic protected areas: the missing dimension in ocean conservation. Trends in Ecology and Evolution 24 (7), 360–369.
FischerJ., HartelT., KuemmerleT., 2012. Conservation policy in traditional farming landscapes. Conservation Letters 5 (3), 167–175.
Gunderson, L.H., Holling, C.S. (Eds.), 2002. Panarchy: Understanding transformations in human and natural systems. Washington, DC: Islands Press.
Gunderson, L.H., Pritchard Jr., L. (Eds.), 2002. Resilience and the behavior of large-scale systems. Washington, DC: Islands Press.
KayJ.J., RegierH., BoyleM., FrancisG.R., 1999. An ecosystem approach for sustainability: Addressing the challenge of complexity. Futures 31 (7), 721–742.
LevinS.A., 1998. Ecosystems and the biosphere as complex adaptive systems. Ecosystems 1, 431–436.
Millenium Ecosystem Assessment (MEA), 2005. In: Ecosystems and Human Well-Being: CurrentState and Trends. Washington, DC: Island Press.
Mitchell, B. (Ed.), 2002. Resource and environmental management, 2nd edn. Harlow: Prentice Hall.
MittermeierR.A., MittermeierC.G., BrooksT.M., PilgrimJ.D., KonstantW.R., da FonsecaG.A.B., KormosC., 2003. Wilderness and biodiversity conservation, PNAS 100 (18), 10309–10313.
OstromE., 2009. A general framework for analyzing sustainability of social-ecological systems. Science 325, 419–422.
PetersonG.D., 2000. Scaling ecological dynamics: Self-organization, hierarchical structure, and ecological resilience. Climatic Change 44, 291–309.
SimonS., 2004. Systemic evaluation methodology: The emergence of social learning from environmental ICT prototypes. Systemic Practice and Action Research 17 (5), 471–494.
TurnerB.L., KaspersonR.E., MatsonP., McCarthyJ.J.W., CorellR., ChristensenL., EckleyN., KaspersonJ.X., LuersA., MartelloM.L., PolskyC., PulsipherA., SchillerA., 2003. A framework for vulnerability analysis in sustainability science. Proceedings of the National Academy of Sciences 100 (14), 8074–8079.
Ulanowicz, R. (Ed.), 1997. Ecology, the ascendant perspective. New York: Columbia University Press.
WalkerB., CarpenterS., AnderiesJ., *et al*., 2002. Resilience management in social–ecological systems: A working hypothesis for a participatory approach. Conservation Ecology 6 (1), 14.
WalkerB., HollingC.S., CarpenterS.R., KinzigA., 2004. Resilience, adaptability and transformability in social–ecological systems. Ecology and Society 9 (2), 5.

The Sustainable Development Goals

Massimo Gigliotti, University of Siena, Siena, Italy
Guido Schmidt-Traub, UN Sustainable Development Solutions Network, Paris, France
Simone Bastianoni, University of Siena, Siena, Italy

A Plan for the Future

The Sustainable Development Goals (SDGs) were approved by all 193 countries participating at the 70th General Assembly of the United Nations, held in New York on 25th September 2015. The goals are the "2030 Agenda for Sustainable Development". They define a universal, holistic set of objectives to help countries move towards the three dimensions of sustainable development—economic development, social inclusion and environmental sustainability—in a climate of peace, justice, and international collaboration. After more than a year of deliberations, the Open Working Group proposed a set of 17 SDGs and 169 accompanying targets that form a basis for the post-2015 intergovernmental process (UN-SG, 2014).

The previous Millennium Development Goals (2000–15) were eight purely social goals aimed at developing countries. During the 15 years of their application, great improvements were made in national data gathering, however annual statistical information was slow in becoming available—often three or more years after the reference year—and the databases were sometimes incomplete or noncomparable across countries, making the indicators useless for decision-making (UN, 2015). There was also insufficient investment to strengthen statistical capacity and ensure real-time monitoring of those goals. To achieve the Sustainable Development Goals, more investment in independent, impartial national statistical capacity is required. Annual reporting must offer high-quality data from all countries, disaggregated and comparable across countries and time (IEAG, 2014). This should ensure a constant flow of information, useful for implementing policies in line with reality and the new goals.

The 17 Sustainable Development Goals

The 17 SDGs address the three pillars of sustainability: economic development, social inclusion, and environmental sustainability. The targets for each goal articulate the aims and link the goals where possible (UN, 2014) (Fig. 1).

Fig. 1 The framework of the 17 SDGs. Source: http://www.un.org/sustainabledevelopment/sustainable-development-goals/.

SDG1: End Poverty in All Its Forms Everywhere

Since 1990, extreme poverty has halved from 1.9 billion people below the poverty line of $2 a day to nearly 800 million people in 2015, against a corresponding rise in the global population (especially in countries with high rates of extreme poverty) from 5.2 to 7 billion. Despite this remarkable result, one in five people still live in absolute poverty and many others risk slipping back into it.

SDG2: End Hunger, Achieve Food Security and Improved Nutrition and Promote Sustainable Agriculture

Last century, intensive agriculture caused food insecurity and loss of soil fertility. The world wastes 1.3 billion tons of food every year, while 1 billion people are hungry and another billion are undernourished. At the same time, almost 2 billion people are overweight or obese.

A profound change in agriculture is needed to end malnutrition, account for an additional 2 billion people by 2050, and to meet the rising per capita demand for meat and other high-protein diets that require plant feed.

Climate change is raising average temperatures, changing precipitation patterns, and increasing the likelihood and severity of extreme weather events. On current trends climate change will therefore have a major adverse impact on agricultural productivity in most regions. It will also likely increase mass migration.

SDG3: Ensure Healthy Lives and Promote Well-Being for All at All Ages

Between 2000 and 2015, maternal and infant/under-five mortality rates fell by 37% and 44%, respectively. Significant progress was made against severe infectious diseases such as tuberculosis, malaria and polio and against the spread of HIV. However, much effort is still needed to eradicate these diseases. Research and investment are needed to address the possibility of new pandemics caused by high population density, climate change and antibiotic resistance.

SDG4: Ensure Inclusive Quality Education for All and Promote Lifelong Learning

In 2015, enrolment in primary education in developing countries reached 91%, but there are still 57 million children left out of school. Quality education is fundamental for many of the SDGs. Education gives people tools to rise out of poverty and helps reduce inequalities and achieve gender equality. Greater efforts are required to achieve the goal of global literacy with gender equality not only in primary education but at all levels.

SDG5: Achieve Gender Equality and Empower All Women and Girls

Significant progress has been made towards gender equality and the emancipation of women in recent decades, however in the period 2005–16, 19% of women aged 15 to 49 years in 87 countries reported suffering physical or sexual violence from a partner in the last 12 months.

Even today many women receive lower salaries than their male counterparts, even in developed countries. In many countries, although women represent more than 30% of the political electorate, female managerial representation remains low in the private sector. Strengthening female empowerment will fuel more equitable and sustainable economies and societies.

SDG6: Ensure Access to Water and Sanitation for All

Between 1990 and 2015, the percentage of the global population with access to safe drinking water increased from 76% to 91%, corresponding to 2.6 billion people, but almost 700 million people are still without access to improved water sources. Water scarcity affects more than 40% of the global population and is rising rapidly. Water availability is fundamental for a healthy life, but each day nearly 1000 children die from preventable water sanitation-related gastrointestinal diseases.

Finally, 70% of all freshwater withdrawals are used for irrigation. Hydropower is the major renewable source of energy, representing 16% of total electricity production worldwide in 2011. Floods and drought are two opposite problems related to the availability of water that will increase in many parts of the world due to climate change.

SDG7: Ensure Access to Affordable, Reliable, Sustainable, Modern Energy for All

Modern societies need energy for every aspect of work: production of food, goods, services, transport, trade, recreation, and fun. A radical transition from the use of fossil fuels to renewable sources such as solar, wind, hydroelectric, biomass, geothermal, and wave energy is needed as soon as possible, also to mitigate the effects of climate change.

In addition to the change in production upstream of the energy process, it is also necessary to spread awareness about energy saving and the negative effects of wasting energy. It is often not necessary to produce more energy if what is already produced is used more efficiently.

SDG8: Promote Inclusive and Sustainable Economic Growth, Employment and Decent Work for All

Duly paid work is the only universal tool that can allow people to rise above absolute poverty and therefore escape hunger, enjoy good physical and mental health, and contribute to the economic development of their country.

The opportunity for decent work would strengthen the basic social contract, the democratic foundations of which are threatened. Providing quality work for everyone will remain a major challenge for all nations, since every year 30 million new people are looking for jobs.

SDG9: Build Resilient Infrastructure, Promote Sustainable Industrialization and Foster Innovation

Industrial development that is socially inclusive and attentive to environmental protection will be the main source of income for millions of people in developed and developing countries. Energy efficiency and substantial reductions in resource use can be achieved through technological innovation.

Infrastructure is necessary to transport sustainable energy and to exploit technological progress in many other sectors, such as agricultural production, education, transport, and information.

SDG10: Reduce Inequality Within and Between Countries

In the period 2008–14, 40% of the world's poor saw an increase in their income or consumer opportunities. Although the international community has achieved significant results in the reduction in income disparities between nations, inequality within countries has increased. Less than one hundred people hold the same amount of wealth as is the bottom 50% of the world population. It is now widely believed that economic growth is only possible when the three dimensions of sustainability (economic, social, and environmental) are considered.

SDG11: Make Cities Inclusive, Safe, Resilient, and Sustainable

Cities have been the engines of civilizations and cultures since historic times. Businesses, centers of learning, social development, ideas, and innovations thrive in cities. However, the population explosion over the last 200 years has led to rapid urbanization, and many cities have been unable to build adequate infrastructure and provide social service to support their growing populations.

The lack of adequate housing and infrastructure has led to the spread of slums and road congestion. In 2015, 54% of the world's population (4 billion people) lived in cities and it is expected that by 2030 there will be a total of 5 billion people living in urban

areas. The challenges that cities have to face in order to restore social prosperity include reduction of poverty and pollution, construction of infrastructure and implementation of services required by today's and tomorrow's citizens.

SDG12: Ensure Sustainable Consumption and Production Patterns

Globally, the Material Footprint, which indicates the flows of mineral and organic resources withdrawn from the environment to produce assets, increased from 48.5 billion tons in 2000 to 69.3 billion tons in 2010. If the global population reaches 9.6 billion by 2050, the equivalent of three planets will be needed to support current lifestyles unless technologies change profoundly to dematerialize consumption and production patterns.

Sustainable production and consumption require the promotion of energy efficiency and the reduction of waste. Their implementation would create jobs, reduce negative environmental, social and economic impacts, and improve the competitiveness of nations. The circular economy aims to "do more and better by consuming less" but requires a considerable effort and a systemic approach involving all actors in production chains, as well as much commitment and awareness on the part of consumers, who are empowered to make informed purchases through information labels.

SDG13: Take Urgent Action to Combat Climate Change and Its Impacts

Climate change has now begun to show its first effects in every country in the world. If we do not take measures to curtail emissions of greenhouse gases, global average temperatures could exceed 3–4°C this century. In some parts of the world the increase may be significantly greater. Climate migrants could number 1 billion.

Climate change is a global challenge that goes beyond national borders and must therefore be tackled by international concerted action. Countries adopted the Paris Agreement at COP21 and pledged to work together to maintain the increase in global average temperature below 2°C and possibly below 1.5°C.

SDG14: Conserve and Sustainably Use the Oceans, Seas, and Marine Resources

Through phytoplankton, the oceans are the lungs of planet Earth. They also sequester carbon, but this acidifies seawater and endangers coral reefs, hot-spots of biodiversity. Sixteen percent of marine ecosystems are at risk or seriously at risk of coastal eutrophication, while overfishing has reduced food production, damaged ecosystems, and decreased biodiversity.

Seas and oceans are heavily polluted by chemicals, excess organic matter and urban waste such as plastics. The latter form huge plastic islands trapped in the ocean gyres. Protecting marine resources means supporting island populations, biodiversity and the health of the planet.

SDG15: Sustainably Manage Forests, Combat Desertification, Halt and Reverse Land Degradation, Halt Biodiversity Loss

Terrestrial ecosystems support most of our development, from raw materials to food production. Forests make up 30% of the Earth's surface, provide oxygen and shelter for many land species, and constitute and important stock of carbon. In the period 2010–15, the annual loss of forested land was less than half that in 1990 but 12 million hectares of forest per year are lost and biodiversity continues to decline at alarming rates. At today's technologies, a growing human population will require more cultivated fields but this cannot be allowed at the expense of forested land, also considering advancing desertification due to climate change. Water resources and many new drugs and unknown active ingredients depend on the conservation of forest ecosystems.

SDG16: Promote Just, Peaceful and Inclusive Societies

Wars and conflicts breed inequality and poverty. Effective institutions and access to justice for all are the best way to prevent future conflicts. On a global level, the number of victims of voluntary homicide in 2015 stood between 4.6 and 6.8 per 100,000 people, and many forms of violence against children persist. Racial, religious and sexual intolerance continue to be a problem. Corruption, theft and tax evasion cost the world community $1.25 billion per year that could be used to raise people out of absolute poverty.

SDG17: Revitalize the Global Partnership for Sustainable Development

Some sustainable development challenges can be addressed thorough efforts of individual nations. Others will require concerted international action. Some solutions start bottom-up, from individual behavior, while others must be managed by policy makers at various levels (from local to international). To achieve sustainable development, every part of society must be involved: governments, the private sector, and civil society. International mobilization is necessary if industrialized countries are to help poorer nations fight extreme poverty and prevent them from repeating today's models of resource depletion. Long-term investments are needed in the fields of sustainable energy, infrastructure, transport and ICT. In 2014, official development aid reached 135 billion dollars, the highest ever recorded. Together it is possible to achieve the Sustainable Development Goals through greater investments and policy coordination centered on the 2030 Agenda.

Monitoring Systems at Different Spatial Scales

The goals and targets form a common context for all institutions, public or private, that choose to implement the SDGs as strategic objectives. They are available on the United Nations website: https://sustainabledevelopment.un.org/?menu=1300. The United Nations proposes a set of 232 SDG indicators (A/RES/71/313), but institutions can choose the ones they judge most appropriate (on the basis of the targets) for tracking their progress towards sustainable development (SDSN, 2015).

However, the goals also describe a global agenda which includes cross-border issues that can only be successfully addressed through close international cooperation, which in turn requires national responsibility and monitoring. The SDGs cannot be approached unless national efforts are complemented by an effective global monitoring framework.

These four levels of monitoring—national, regional, global, and thematic (UN-SG, 2014)—are illustrated in Fig. 2 (SDSN, 2015).

National Monitoring

The most important level of monitoring is the national one. Nations should implement the SDG monitoring framework in their government agendas, selecting indicators to suit their national needs and priorities. Indicators must be specific, measurable over time, disaggregated and processed by official National Statistics Offices. Nonofficial indicators can be elements of further interest to add richness to national monitoring.

Global Monitoring

Global monitoring is necessary to ensure global coordination and to achieve cross-border goals in thematic areas of supranational interest (e.g., climate change, poverty, inequalities). A dialogue between states is necessary to determine which areas will need more assistance and international aid.

Series of common indicators are chosen by international organizations, ranging from the United Nations to other institutions working in thematic spheres, for example the Food and Agriculture Organization (FAO) and the World Health Organization (WHO).

Regional Monitoring

Regional monitoring can be seen as a subset of the national and global levels (Fig. 2). It may affect political or geographical regions such as the European Union, OECD countries, Southeast Asia and the Pacific and Caribbean islands. It can offer an opportunity for countries linked by common issues to share knowledge and to collaborate in the implementation of joint projects for regional priorities such as shared watersheds, regional conflicts and regional infrastructure. Thus, indicators for regional monitoring may extend beyond the scope of the Global Monitoring Indicators and may include some metrics not considered under Complementary National Indicators.

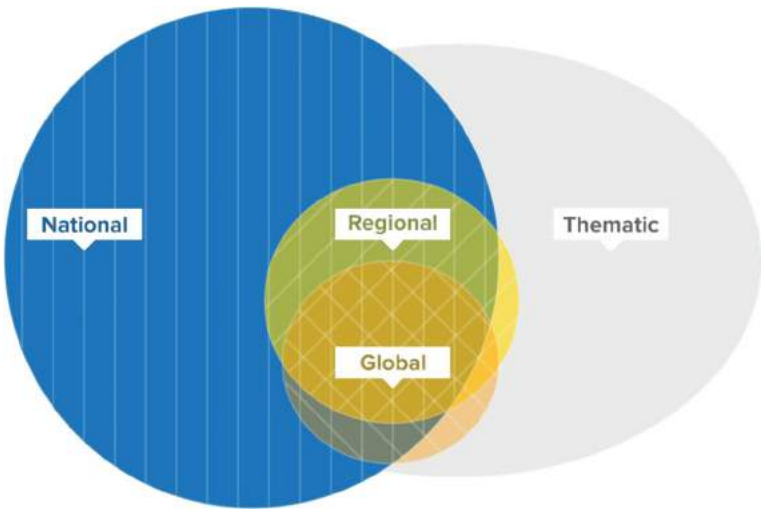


Fig. 2 Schematic illustration of indicators for national, regional, global and thematic monitoring. Source: Sustainable Development Solution Network (2015). Indicators and monitoring framework for the Sustainable Development Goals. Launching a data revolution.

Thematic Monitoring

Many challenges for mankind concern thematic areas like health, agriculture, education, nutrition, the water-energy-food nexus, consumption, and production. Partnerships between countries are the key to shared solutions to a given issue through common knowledge. Specific monitoring indicators should be developed by specialized international organizations for each issue and accountable thematic communities built to track countries across the globe. These indicators are often complementary to official national indicators, and tend to focus better on the issues.

International Institutions

The UN Secretary-General launched the Sustainable Development Solution Network (SDSN) in 2012. Its task is to mobilize global scientific and technological expertise to promote practical solutions towards sustainable development, including implementation of SDGs, involving policy-makers, the private sector and all citizens.

The SDSN is involved in implementation of the SDG framework and publishes an annual report on the progress towards achieving the goals. This report draws on official UN indicators for the SDGs and fills data gaps using other official or unofficial metrics.

The 2017 SDG Index and Dashboards Report published in collaboration with the Bertelsmann Stiftung presents data for 157 countries using some 89 indicators across the 17 SDGs. Countries' distance from the targets is calculated for each indicator, SDG, and the sum of all 17 goals. The using average performance across the 17 goals, the report presents a ranking of all countries.

The report contains an analysis of the international "spillovers" in achieving the SDGs. Major cases of SDG-related spillovers and misuse of the global commons are identified and measured (Bertelsmann Stiftung & SDSN, 2017).

References

Bertelsmann Stiftung & Sustainable Development Solution Network (SDSN), (2017), SDG Index and Dashboards Report 2017, Global responsibilities, international spillovers in achieving the goals.

Independent Expert Advisory Group on a Data Revolution (IEAG) (2014), A world that counts. Mobilising the data revolution for sustainable development.

Sustainable Development Solution Network (SDSN) (2015), Indicators and monitoring framework for the sustainable development goals. Launching a data revolution.

United Nations (UN) (2014), Reporting of the Open Working Group of the General Assembly on Sustainable Development Goals, A/68/970, New York.

United Nations (UN) (2015), The Millennium Development Goals Report, New York.

United Nations Secretary-General (UN-SG) (2014), The road to dignity by 2030: Ending poverty, transforming all lives and protecting the planet. Synthesis report of the Secretary-General on the post-2015 Agenda, New York.

Further Reading

Sachs, J., Schmidt-Traub, G., Kroll, C., Durand-Delacre, D., and Teksoz, K. (2016). SDG Index and Dashboard—Global Report. New York. Bertelmann Stiftung and Sustainable Development Solution Network (SDSN).

System Sustainability

Federico M Pulselli, University of Siena, Siena, Italy

Sustainability

Sustainability is a word that is used—sometimes overused—in many fields to indicate many things. Sometimes it refers to the environmental sphere and sometimes it has to do with the human sphere, considering its social, economic, financial and political aspects at macro and micro levels. It is therefore difficult to arrive at a univocal definition of the concept, although a strong scientific basis for it can be found in the literature that has evolved in the last 30–40 years.

Nature is the quintessence of sustainability; a translation of the concept for humans would define a virtuous path for evolution of the human system. Sustainability is important for humans and their actions for many reasons, not least the survival of our species. Learning from nature can enable us to translate its winning strategies into virtuous behavior and avoid mistakes and risks for the future.

The first step to characterize sustainability consists in the identification of its biophysical foundations as necessary condition to limit the elusive aspects of the concept (Pulselli *et al.*, 2008).

Time

If we look at a pianist, we see that he plays a chord by striking keys on the piano. When he removes his hands from the keys, the chord dies. If the pianist presses a special pedal, a mechanism enables the chord to continue to vibrate even after the fingers have been removed. The pedal maintains the notes for a while and its name is "sustain." Time is therefore intrinsic to the meaning of sustainability. Something that exists or survives in time is sustainable. This essence of the term "sustainability" and the adjective "sustainable" is important to avoid misuse. It also characterizes the difference between the expressions "carrying capacity" and "sustainable development" because the verb "to carry" means to hold, contain or support something, whereas "to sustain" means to maintain in time. So if we consider time as a crucial element of sustainability it means that we view the dynamics of ecosystems and human activity not as a simple sequence of changes of state and modifications, but as evolution. This helps clarify which actions are virtuous and which should be avoided in order to maintain a system indefinitely in time.

Biophysical Limits

Nature has been surviving for about 4 billion years, during which time biological evolution proceeded through winning strategies under the inevitable condition of limited resources. Nature diversifies: biodiversity is a strong general attribute for survival. Nature exploits the most abundant and reliable energy source, solar energy, finding ways of using it to feed vital mechanisms. Nature optimizes resources and disposes of wastes and degraded energy (e.g., by closure of cycles and dissipation of heat into space). Nature has not only been successful in surviving in time but also in thriving and flourishing within geobiophysical constraints (Jørgensen *et al.*, 2015). Human activities, which are always fed by resource extraction and consumption and generate flows of degraded energy and matter, should be consistent with the finite and constrained ability of the planet to regenerate resources and absorb wastes.

Relations

Thermodynamics gives us the tools to understand these constraints. For instance, its fundamental laws define the limits to overall availability and our capacity to exploit energy: the law of conservation of energy (first law of thermodynamics) states that energy cannot be created or destroyed, and the law of energy dissipation (second law of thermodynamics) states that every activity or conversion degrades energy to heat which is unable to do work. Biological systems are open systems (they exchange matter and energy with their environment) that involve extremely ordered structures and evolve in the direction of increasing order by processing the flow of energy and resources they capture from the environment. They discharge their wastes as degraded energy (heat) and matter (e.g., emissions, pollutants). A continuous process of transformations, cycles and feedbacks demonstrates the importance of relationships for living systems and the dependence of these systems on the context in which they live. This is valid for a cell, a tree, a human being, an ecosystem, but also for a city, a production process, an urban or regional system. All these systems survive by exploiting flows of energy and matter, releasing wastes, emissions and heat into their surroundings (see, among others, Schrödinger, 1944; Prigogine, 1954; Tiezzi, 2003).

Time, biophysical limits and relations are three cornerstones of the concept of sustainability. A system or project cannot be called sustainable if it does not rest on these foundations: it must be durable, it must develop within the limits imposed by natural

 doi:10.1016/B978-0-12-409548-9.10610-4

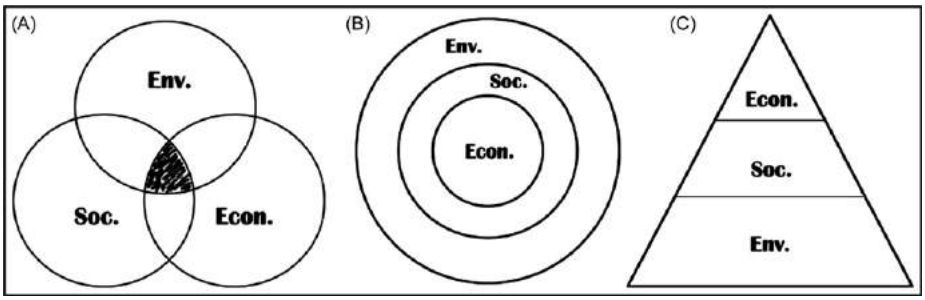


Fig. 1 Representations of sustainability. The usual representation of sustainability shows the *different (A–C)* contributions and interactions between the environmental (Env.), social (Soc.) and economic (Econ.) spheres of human life.

laws (such as the laws of thermodynamics) and it must maintain relationships with other systems and the surrounding environment. Without considering these pillars, the use of the word sustainable is misleading, inappropriate, or even false.

Representing Sustainability

Human development involves three spheres: economic, social and environmental. According to Barbier (1987), sustainable development is defined by simultaneous interaction of the three intersecting spheres. The intersection is where sustainable development can occur, suggesting that converging goals from the different spheres must be achieved together (**Fig. 1**A). This view, however, places the three circles on the same level, and excluding their intersection, it considers the ecological, social and economic elements of sustainability to be interchangeable or substitutable (which amounts to the condition of weak sustainability). An alternative interpretation is that of concentric circles, which implies an environmental basis for the social sphere that contains the economic expression of society (**Fig. 1**B). Another view depicts sustainability as a pyramid, highlighting the relative sizes of the three spheres (**Fig. 1**C). These rather conventional ways of representing the concept of sustainability underline the need for an organic view of human presence and activity on the planet. Study of the relationship between the human species and the Earth system is urgently needed. Researchers have been proposing ideas in this field for several decades, highlighting the risk and consequences of reaching or passing critical points or thresholds. This was one of the aims of the well-known and influential report entitled "The Limits to Growth" (Meadows *et al.*, 1972), which took an organic, comprehensive and global view of the planet.

System Sustainability

The basic concept of system is expressed by the well-known adage of systems science: "the whole is greater than the sum of its parts." This can be explained by the fact that a system, namely "things that stay together," consists of components that cooperate with each other, giving rise to emerging novelties and new properties. For instance, a human being has a body made of 10^{14} cells. Every cell has biochemical, metabolic and anabolic characteristics and properties determined by genetic and environmental conditions but they do not explain the characteristics and properties of the organism as a whole (knowledge, feelings, emotions, pain, etc.). Similarly, an ecosystem is composed of organisms linked in a complex cooperative synergistic relationship that results in properties like adaptation, evolution, self-organization, resistance, flexibility and ultimately beauty (Jørgensen *et al.*, 2015).

According to Jørgensen (2012), "systems with emerging properties cannot be described by listing the components and their properties, but it is necessary to capture, understand and describe the emerging system properties." A fundamental overview of general systems theory was proposed by Von Bertalanffy (1969). Since then many authors have investigated the question which is relevant to our life on the planet and especially sustainability. In a comprehensive survey of the systems approach, Capra (1996) suggested that the essential properties of an organism or living system are properties of the whole system, which no single component has per se. These properties emerge from interactions and relations among the parts of the system. There have been major advances and novelties in many areas of systems science. Ecology, for example, is an area permeated by the systems view. According to Capra, if we consider the world as an integrated system rather than a series of separate components, we take a holistic or even ecological view in keeping with the term "ecological," namely something characterized by relations among its parts and included in a larger context. One fundamental contribution to the advancement of ecology was by Eugene Odum. In a major paper, he emphasized the role of ecology as a discipline that connects and advanced various ideas about the concept of sustainability: "[…] cell level science will contribute very little to the well-being or survival of human civilization if our understanding of supraindividual levels of organization is so inadequate that we can find no solutions to population over-growth, social disorder, pollution, and other forms of societal and environmental cancer. […] It is in the properties of the large-scale, integrated system that hold solutions to most of the long-range problems of society" (Odum, 1977). The paper is actually a precursor of sustainability science because it also traces the link between ecology on the one hand, and social sciences, technology, politics and economics on the other. Jørgensen (2012) subsequently published a textbook on Systems Ecology that deals with

thermodynamics- and biochemistry-based foundations, ecological laws, ecosystem properties and ultimately environmental management. He stresses the need for a holistic view to problem solving and touches on the progressive emergence of meta-disciplines, particularly ecological modeling and ecological engineering, that tend to unify different approaches. Another interesting example of adoption of a systems approach can be found in systems chemistry, that is "the study of complex systems, or networks, of molecules [to] investigate how interactions between members propagate through networks allowing complex behavior to emerge" (Nitschke, 2009).

The concept of sustainability can only be considered from a systemic viewpoint. The study of human sustainable development is concerned with relations between the individual and collective expressions of humankind and the multiplicity of contexts in which the human action develops. The context can be physical, environmental, social, economic, political, institutional, urban, legal and so forth. Pulselli *et al.* (2016) derive at least three key points from this: (a) the organic picture of reality (i.e., what should be sustainable?) requires a transdisciplinary approach in order to encompass the many dimensions of the context in which we live; (b) the purpose (i.e., why should we be sustainable?) is to create and maintain the conditions for durably living better and in harmony with nature and other individuals; and (c) the critical assessment of how we can reach these conditions (i.e., how can we be sustainable?) demands new frameworks in which to evaluate progress toward the desired change. Sustainability therefore connotes a system within its context. By adapting a statement from Jørgensen (2012), originally referring to ecosystems, we can say that if we can understand how ecosystems (as well as human systems) work as systems, we will be able to develop an ecosystem (as well as sustainability) theory that can be used to predict the effects on an ecosystem (as well as a human system) of well-defined changes of forcing function.

Various approaches and methodologies have been developed in recent years to implement a systemic view. Regarding system sustainability, attempts have been made to identify, categorize, quantify and manage the relationships between human life and activity and the context(s) in which they develop, with particular emphasis on environmental context.

An important approach is that of ecosystem service (ES) identification and evaluation. Ecosystem services are the portion of ecosystem functions that directly or indirectly contributes to human welfare. Ecosystem functions occur independently of humans, however humans exploit the existence of ecosystems and their dynamics. These services are usually classified in four categories: provisioning services (e.g., food, water, timber, fiber), regulating services (regarding climate, floods, disease, wastes, water quality etc.), cultural services (that provide recreational, esthetic and spiritual benefits) and supporting services (e.g., soil formation, photosynthesis, nutrient cycling) (see MA, 2005, and the corresponding entry in this encyclopedia). The benefits people obtain from these ecological services can be assessed in economic terms, and the results are surprising. As Costanza *et al.* (1997) demonstrated, at global level, nature provides humans with more resources, and more efficiently, than does world economic infrastructure (which is designed to do just that). Assessment of ecosystem services is therefore a way to estimate and express the importance of ecological dynamics for human life and its sustainability.

Sustainability research has always paid great attention to the determination of thresholds. These are measured or estimated levels of an entity or parameter, beyond which conditions may change unexpectedly, unpredictably and/or irreversibly. An important approach to human activity thresholds at global level is the so-called planetary boundaries framework that aims to identify a safe global operating space for the survival and prosperity of humanity (see Rockström *et al.*, 2009, and the corresponding entry in this encyclopedia). The approach measures selected control variables to determine the risk of dangerous consequences for human well-being in nine Earth processes: climate change, biogeochemical flows, biodiversity loss, land system changes, ocean acidification, stratospheric ozone depletion, freshwater use, atmospheric aerosol loading and chemical pollution. Research in this field is demonstrating that the first four processes involve risk factors for humanity. A major aspect of this approach is the concept of the Earth as an integrated system to be considered, preserved and managed as a whole. The approach is also inspiring research in other fields, such as legislation. For example, a multidisciplinary group of researchers are promoting a systemic legal framework applicable to the Earth as a whole, with a set of assessment and regulatory tools that can be condensed into a safe operating space treaty (SOS Treaty). The treaty is proposed as a new global guideline for legal framework researchers because the Earth is still formally an unidentified legal object (Magalhães *et al.*, 2016).

A further system sustainability approach is the so-called food-energy-water nexus. Used in many studies by various organizations (e.g., the Food and Agriculture Organization of the United Nations, FAO), it offers a way to first identify and then solve problems that a large portion of the world population is facing. By combining different fields of investigation and action it confirms the importance of seeing the world in which we live as a set of cooperative interrelated units and systems. In this case, study of the three sectors—food, energy and water—, to which the climatic system is often added, helps us recognize the many links between them and the fact that problems in one may trigger serious crises in the others. The three-sided approach highlights the main conditions for achieving essential tasks, such as poverty reduction, human well-being and sustainable development at global level, that depend on the context in which human development occurs.

The Sustainable Development Goals, promoted by the United Nations in 2015, are an enlargement of this view. Countries adopted these goals "to end poverty, protect the planet, and ensure prosperity for all as part of a new sustainable development agenda. Each goal has specific targets to be achieved by 2030" (for an overview, see UN, 2015). The set of goals is so wide that it offers an encompassing vision of the human condition on the planet. It includes the following 17 items: no poverty, zero hunger, good health and well-being, quality education, gender equality, clean water and sanitation, affordable and clean energy, decent work and economic growth, industry, innovation and infrastructure, reduced inequalities, sustainable cities and communities, responsible consumption and production, climate action, life below water, life on land, peace, justice and strong institutions and

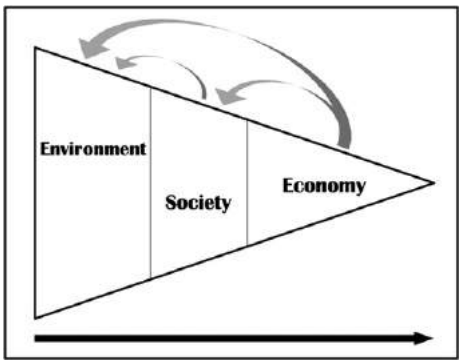


Fig. 2 A less usual representation of sustainability with a more logical/consequential series of stages ordered from left to right.

partnerships for the goals. Again all these aspects are linked and the set works as a system. The challenge is that the purpose of system sustainability is only achieved if the goals are largely achieved together at the same time.

Problems do not observe disciplinary or geographical boundaries but arise almost everywhere, in many ways and with different dimensions and scales. The approaches outlined above can help solve these problems by adopting a systemic view that facilitates understanding of the connections between different components of a given system and identification of emerging properties and problems of the system.

Less Usual Representation of Sustainability and a Possible Interpretative Framework

The scientific community recognizes the environmental, social and economic spheres as components of the concept of sustainability. However, we have seen that they cannot be considered interchangeable.

Fig. 1C shows the relationships between the three spheres of sustainability in the form of a pyramid: the base of the pyramid represents natural assets, crucial inputs into the system; the middle section represents society with its organization and structure, and can be viewed as the state of the system; the point of the pyramid is the real economy that produces the system's output. The pyramid can be considered an evolution of the concentric-circles representation (**Fig. 1**B) that also combines the environmental, social and economic spheres.

Pulselli *et al.* (2015) proposed a slightly different version of this three-stage approach in order to represent sustainability and facilitate investigation. This less usual representation of sustainability rotates the pyramid 90 degree clockwise to capture the succession of the stages (**Fig. 2**) from left to right. "A flow of material and energy inputs generated by the available stock of Natural Capital feeds (is captured by) the system. These resources are necessary for the elements of the system (i.e., society and its organizational units) to operate (act, live, survive); the level of organization of the society influences the degree of utility/satisfaction derived from processing/using/consuming resources. An organized society is supposed to be able to achieve better economic results, providing outputs from its productive processes" (Pulselli *et al.*, 2015).

This version of the scheme aims to represent the physical, relational and thermodynamic order (environment-society-economy) of the succession of stages. It offers a more logical/consequential approach to combining and evaluating different indicators, starting from the economy's dependence on societal organization and environmental resources and including feedback produced by the socioeconomic system.

A general systemic approach emerging from this new version of the sustainability scheme could be the Input-State-Output (I-S-O) framework, useful for assessing system sustainability. In Sustainability section, we saw that living biological systems (including human-driven ones like cities and production processes) are fed by a continuous flow of energy and matter that is processed internally (e.g., self-organization) to give useful outputs (in terms of services, feedback flows, products, etc.). These vital evolutionary steps are therefore simplified by this linear but organic scheme (**Fig. 3**) that also helps select suitable methods and tools for quantifying phenomena. Here different combinations of indicators can be used to count inputs of energy and matter into a system, describe organizational state and quantify outputs. The I-S-O framework orientates the use of well-defined triads of systems indicators representing links between the three spheres of sustainability. This process, consisting of a succession of stages in a series, must be viewed as the integrated way in which the system works. Nevertheless, identifying the three stages makes it possible to select corresponding indicators and compose multidimensional data, often expressed in different units, without losing information. Indeed, the framework can be investigated using a diagram with three axes representing the input, state and output indicators, and the information gained by different indicators is not lost in final aggregations. This framework represents relationships between indicators and helps classify the systems in question in terms of their sustainability.

The choice of indicators is crucial for implementing this method. If we consider the correspondence between the I-S-O scheme and the environmental-social-economic spheres of sustainability, we can use physical measures for input, social indicators for state and economic parameters for output. Thus the triads are composed of groups of indicators/measures representative of each dimension: measures of resource flows (such as ecological footprint, emergy flow, gross emission of CO_2 equivalent, or more

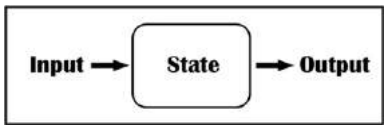


Fig. 3 The Input-State-Output (I-S-O) framework.

generically, energy, matter and pollutant flows) to represent the environmental contribution to the system; measures representative of the state or health of a society or its inequality level (such as employment level or an index of income distribution); economic measures representative of output (such as GDP or value added, or the welfare/happiness of a population).

Each triad gives different information, depending on the indicators chosen. The point representing combination of the three indicator values can be identified in the 3D space enclosed by the three axes. Every system can therefore be identified by a point on the three-axis diagram and its position can be discussed in a static way, determining, for example, the main components that characterize the results obtained (environmental, social or economic), or in a dynamic way, when a time series of indicators is available, signaling development trajectories for that system.

According to the authors, "this application is a rational solution for the study of system sustainability, because it incorporates consistency with traditional sectors proposed in sustainability research and is feasible because it is limited to small number of (already available) data. [The proposal also] maintains the informative capacity of every aspect of system behavior, but it also provides a synthetic picture of the reality. […] This new way of considering these fundamental aspects of our life can be the basis for assessment of long term system sustainability and helps identify the contradictions, hypocrisies, and nonsustainable nature of many of our current behaviors" (Pulselli *et al.*, 2015).

Conclusion

System sustainability designates a condition that characterizes a given system. A system is sustainable if it has durable prospects with respect to constraints and preserves relations. Sustainability is a property of the whole system with all its elements, actors, connections and properties. The concept of sustainability applies especially to human systems, for which a multidimensional/systemic approach makes it possible to consider man-driven systems and actions within their context and limits. Humanity is continuously faced with major problems regarding the health of the planet and ecosystems as life supporting systems, relationships between nations and populations, and ultimately the survival of the species. Improving our capacity to measure/observe the world in which we live and widening the set of investigation tools in this field are urgent tasks. Instruments that enable holistic understanding of the environmental, social and economic view of human life are the most suitable for achieving sustainable systems. In particular, we need investigation and management tools for the environmental and social sphere that complement the primacy of the economic dimension. We have seen that systemic approaches, methods, interpretation proposals and solutions proposed in the last 30 years aim to make the multidimensionality of our world more explicit. To conclude this survey, it is worth mentioning two books which are fundamental references for a systemic view of sustainability. The first is "A prosperous way down" (Odum and Odum, 2001), an innovative contribution that anticipated more modern tendencies like the so-called de-growth movements and the EU Beyond GDP initiative. The importance of the book lies in the fact that it associates relevant aspects of natural and physical phenomena and energy-based dynamics with economic, social, political and demographic problems. The second is "Flourishing within limits to growth. Following nature's way" (Jørgensen *et al.*, 2015): inspired by "Limits to growth" (Meadows *et al.*, 1972), it acknowledges the existence of constraints and limits, but stresses that like ecosystems, human systems must find a way to flourish while observing these limits, seeking wealth but preserving the environment and quality of life.

See also: Human Ecology and Sustainability: Emergy and Sustainability; Ecological Footprint; Ecological Economics 1

References

Barbier, E., 1987. The concept of sustainable economic development. Environmental Conservation 14, 101–110.
Capra, F., 1996. The web of life. New York: Anchor Books.
Costanza, R., d'Arge, R., de Groot, R., *et al.*, 1997. The value of the world's ecosystem services and natural capital. Nature 387, 253–260.
Jørgensen, S.E., 2012. Introduction to systems ecology. Boca Raton, FL: CRC Press, Taylor & Francis Group.
Jørgensen, S.E., Fath, B.D., Nielsen, S.N., *et al.*, 2015. Flourishing within limits to growth. Following nature's way. Florence, KY: Earthscan from Routledge, Taylor & Francis Group.
MA, 2005. Millennium ecosystem assessment. Washington DC: Island Press.
Magalhães, P., Steffen, W., Bosselmann, K., Aragão, A., Soromenho-Marques, V. (Eds.), 2016. SOS treaty. The safe operating space treaty: A new approach to managing our use of the Earth system. Newcastle, UK: Cambridge Scholars Publishing.
Meadows, D.H., Meadows, D.L., Randers, J., Behrens, W.W., 1972. The limits to growth. New York: Universe Books.
Nitschke, J.R., 2009. Systems chemistry: Molecular networks come of age. Nature 462, 736–738.
Odum, E.P., 1977. The emergence of ecology as a new integrative discipline. Science 195, 1289–1293.

Odum, H.T., Odum, E.C., 2001. A prosperous way down. Principles and policies. Boulder, USA: University Press of Colorado.
Prigogine, I., 1954. Introduction to thermodynamics of irreversible processes. Springfield, USA: C.C. Thomas.
Pulselli, F.M., Bastianoni, S., Marchettini, N., Tiezzi, E., 2008. The road to sustainability: GDP and future generations. Southampton, UK: WIT Press.
Pulselli, F.M., Coscieme, L., Neri, L., *et al*., 2015. The world economy in a cube: A more rational structural representation of sustainability. Global Environmental Change 35, 41–51.
Pulselli, F.M., Moreno Pires, S., Galli, A., 2016. The need for an integrated assessment framework to account for humanity's pressure on the earth system. In: Magalhães, P., Steffen, W., Bosselmann, K., Aragão, A., Soromenho-Marques, V. (Eds.), SOS Treaty. The safe operating space treaty: A new approach to managing our use of the Earth system. Newcastle, UK: Cambridge Scholars Publishing.
Rockström, J., Steffen, W., Noone, K., *et al*., 2009. A safe operating space for humanity. Nature 461, 472–475.
Schrödinger, E., 1944. What is life? Cambridge, UK: Cambridge University Press.
Tiezzi, E., 2003. The essence of time. Southampton UK: Wit Press.
UN, 2015. Sustainable development goals. Available at: http://www.un.org/sustainabledevelopment/sustainable-development-goals/
Von Bertalanffy, L., 1969. General system theory. Foundations, development, applications. New York: George Brazillier.

Tragedy of the Ecological Commons

E Ostrom, Indiana University, Bloomington, IN, USA

Hardin's Theoretical Assumptions

Garrett Hardin mobilized the energies of many ecologists when he wrote `The tragedy of the commons'. Hardin envisioned `a pasture open to all' in which every herder received direct benefits from adding animals to graze and suffered only delayed costs from his own and others' overgrazing. Given that the benefits of putting an additional animal on the commons accrue to the herder while the costs of overgrazing are shared by all, Hardin predicts that all herders will continue to add animals until they destroy the pasture. Therein lies the `tragedy of the commons'.

To derive his conclusions to his metaphorical theory, Hardin made the following five assumptions: (1) no governance arrangements were present related to the resource system; (2) no human investments were made to improve the productivity of the ecological resource (the pasture); (3) the animals grazing on the pasture were the private property of each pastoralist; (4) a large number of herders and their animals were involved leading to adverse effects on the pasture's long-term productivity; and (5) the herders made decisions totally independently without any communication, local organization, or established norms. When this set of assumptions is posited in a one-shot or finitely repeated game-theoretical analysis, the conclusion of substantial overharvesting is theoretically correct. Thus, Hardin's conclusion is correct given his assumptions.

It is also important to recognize that Hardin's theory is empirically predictive in settings meeting the assumptions that relatively anonymous individuals independently make decisions taking into account only their own individual and immediate payoffs. Researchers have repeatedly generated a `tragedy of the commons' in experimental laboratories when subjects face a common-pool resource setting and make independent and anonymous decisions. Making one small change, however, in the structure of laboratory experiments – a change that is not predicted by game theory to make any difference in outcomes – has repeatedly had a major impact on harvesting behavior and outcomes. Allowing subjects to engage in face-to-face communication between decision rounds significantly changes behavior so that subjects approach optimal harvesting levels rather than overharvesting the commons. In face-to-face discussions, subjects tend to discuss the situation in which they find themselves, compare options, decide what they should all do, and then build norms to encourage conformance.

In settings where there is a large group, no one communicates, and where no rights to the resource exist, the theory that Hardin proposed is supported by considerable empirical evidence. There are many settings in the world where the tragedy of the commons has occurred and continues to occur – ocean fisheries and the atmosphere being the most obvious. The global ocean has lost a vast majority of its large predatory fishes. Large mobile fishing fleets do not coordinate fishing efforts but rather act more like pirates to keep shifting where they fish in order to sell to global markets. The agricultural and forestry practices of advanced industrial countries have also led to immense carbon loss from soils.

Contrary to Hardin's sweeping conclusions that all resources not owned by government or a private owner were severely threatened, however, an outpouring of research efforts has occurred since his famous article to study ecological commons in the field. The possibility that the users would find ways to organize themselves was not considered by Hardin. He thought that the users were trapped in the structure in which they found themselves. The possibility of resource users organizing to get out of the tragedy was also not mentioned in basic economic textbooks on environmental problems until recently.

More recent empirical and theoretical research shows that successful governance of common-pool resources has been achieved by a variety of community, government, and private arrangements. No simple governance system is successful in all settings. The robust resource governance systems documented in recent research do not resemble the textbook versions of either a hierarchical government or a strictly private-for-profit firm.

Design Principles for Common-Property Institutions

After reading a large volume of case studies and finding an immense array of specific rules used to govern ecological resources successfully, it appeared more reasonable to identify underlying design principles that characterized robust common-property institutions than the specific rules crafted to fit the immense variety of ecological resources. No claim was made that those crafting long-lasting institutions were self-consciously using the design principle that characterized their hard work. Rather, those systems that were robust could be characterized as meeting a large number of these principles and those systems that failed were not so structured.

The design principles have now been independently examined by a number of researchers conducting field research in a diversity of different settings. The design principles help one understand indigenous inshore fishery institutions in both Canada and Japan. Their relevance has been established for analyzing the European institutions that managed common land in northwest

 doi:10.1016/B978-0-444-63768-0.00634-X

Table 1 Design principles derived from studies of long-enduring institutions for governing sustainable resources

1. Clearly defined boundaries
   The boundaries of the resource system (e.g., pasture, irrigation system, or fishery) and the individuals or households with rights to harvest resource units are clearly defined.
2. *Proportional equivalence between benefits and costs*
   Rules specifying the amount of resource products that a user is allocated are related to local conditions and to rules requiring labor, materials, and/or money inputs.
3. *Collective-choice arrangements*
   Many of the individuals affected by harvesting and protection rules are included in the group who can modify these rules.
4. *Monitoring*
   Monitors, who actively audit biophysical conditions and user behavior, are at least partially accountable to the users and/or are the users themselves.
5. *Graduated sanctions*
   Users who violate rules-in-use are likely to receive graduated sanctions (depending on the seriousness and context of the offense) from other users, from officials accountable to these users, or from both.
6. *Conflict-resolution mechanisms*
   Users and their officials have rapid access to low-cost, local arenas to resolve conflict among users or between users and officials.
7. *Minimal recognition of rights to organize*
   The rights of users to devise their own institutions are not challenged by external governmental authorities, and users have long-term tenure rights to the resource.
   For resources that are parts of larger systems:
8. *Nested enterprises*
   Appropriation, provision, monitoring, enforcement, conflict resolution, and governance activities are organized in multiple layers of nested enterprises.

Based on Ostrom, E., 1990. Governing the Commons: The Evolution of Institutions for Collective Action. New York: Cambridge University Press.

Europe from 1500 to 1850 as well as for building robust rural regimes in Australia. Given the consistent evidence of their relevance, let us briefly review the eight design principles listed in **Table 1**.

Boundary rules enable a group of users to determine their own membership – including those who agree to use specified rules in harvesting from an ecological resource and excluding those who do not agree to these rules (Design Principle 1). This enables participants to know who is in and who is out of a defined set of relationships and thus, with whom to cooperate. Contemporary developments in evolutionary theory applied to cultural systems and processes of adaptation help to explain how these design principles work to help groups sustain and build their cooperation over long periods of time.

If a group of users is going to harvest from an ecological system in the long run, they must devise rules related to how much, when, and how different products are to be harvested and they need to assess the costs of operating a system on users. The second design principle is that the local rules-in-use allocate benefits proportional to required inputs and are crafted to take local conditions into account. Well-tailored rules help to account for the perseverance of the resource itself.

The third design principle is that most of the individuals affected by a resource regime may participate in making and modifying their rules. Resource regimes that use this principle are more able to tailor rules to local circumstances. They are also able to devise rules that are considered fair by participants. More individuals are willing to abide by these rules because they participated in their design and because they meet shared concepts of fairness.

Few long-surviving resource regimes rely entirely on trust and reciprocity alone. Most long-surviving resource regimes select their own monitors or guards, who are accountable to the users or are users themselves, and who keep an eye on resource conditions as well as on user behavior (Design Principle 4). Further, these monitors use graduated sanctions that depend on the seriousness and context of the offense (Design Principle 5). By creating official positions for local monitors, the community legitimates a position. In some systems, users rotate into this position so everyone has a chance to be a participant as well as a monitor. In other systems, all participants contribute resources to pay the monitors.

The operation of these principles is bolstered by a sixth principle that points to the importance of access to rapid, low-cost, local arenas to resolve conflicts fairly among users or between users and officials. Situations always exist in which participants can interpret a rule differently even when they have jointly made the rules. By devising simple, local mechanisms to get conflicts aired immediately and resolutions that are generally known in the community, the number of conflicts can be reduced. If individuals are going to follow rules over a long period of time, some mechanism for discussing and resolving what is or is not a rule infraction is necessary to the continuance of rule conformance itself.

The capability of local users to develop an ever more effective regime over time is affected by whether they have at least minimal recognition by a national or local government of the right to organize (Design Principle 7). Users frequently devise their own rules without creating formal, governmental jurisdictions for this purpose. So long as external governmental officials give at least minimal recognition to the legitimacy of such rules, the fishers themselves may be able to enforce the rules. But if external governmental officials presume that only they can make authoritative rules, then it is difficult for local users to sustain a self-organized regime.

When an ecological system is large, an eighth design principle tends to characterize successful systems – the presence of governance activities organized in multiple layers of nested enterprises. The rules appropriate for allocating water among major

branches of an irrigation system, for example, may not be appropriate for allocating water among farmers along a single distributory channel. Consequently, among long-enduring self-governed regimes, smaller-scale organizations tend to be nested in ever larger organizations.

All economic and political organizations are vulnerable to threats, and self-organized resource-governance regimes are no exception. Even institutions that are characterized by the design principles fail. Both exogenous and endogenous factors challenge their long-term viability. Major migration (out of or into an area) is always a threat that may or may not be countered effectively. Out-migration may simply endanger the viability of a regime due to loss of those who contribute needed resources. In-migration may bring new participants who do not trust others and do not rapidly learn social norms that have been established over a long period of time. Since collective action is largely based on mutual trust, some self-organized resource regimes that are in areas of rapid settlement have disintegrated within relatively short times.

Garrett Hardin brought the challenge of governing ecological resources sustainably to the world's attention. Significant progress has been made since his classic article. What we have learned is that we should not continue to strive for simple solutions to the problem of governing diverse ecological commons – such as the two extremes posited by Hardin. Social–ecological systems are complex, and the problems of overharvesting and misuse are rarely cured by some simple solution. Holling, Berkes, and Folke identified the structure of the problems involved:

> The answers are not simple because we have just begun to develop the concepts, technology and methods that can address the generic nature of the problems. Characteristically, these problems tend to be systems problems, where aspects of behaviour are complex and unpredictable and where causes, while at times simple (when finally understood), are always multiple. They are non-linear in nature, cross-scale in time and in space, and have an evolutionary character. This is true for both natural and social systems. In fact, they are one system, with critical feedbacks across temporal and spatial scales. Therefore interdisciplinary and integrated modes of inquiry are needed for understanding. Furthermore, understanding (but not necessarily complete explanation) of the combined system of humans and nature is needed to formulate policies. (Berkes and Folke, 1998, p. 352).

Given the complex structure of ecological systems, we have learned that simple solutions are frequently counterproductive for fostering sustainable ecological systems. The conceptual structure of these problems needs to be characterized as a rugged landscape with many peaks and valleys. We need to go beyond the simple solutions that Hardin offered – privatize or nationalize – to recognize the complexity and understand how diverse institutional arrangements affect the incentives and likely actions of multiple actors using a variety of governance arrangements.

See also: Human Ecology and Sustainability: Human Ecology: Overview; Socioecological Systems

Further Reading

Acheson, J.M., 2003. Capturing the Commons: Devising Institutions to Manage the Maine Lobster Industry. New Haven, CT: University Press of New England.
Agrawal, A., 2005. Environmentality: Technologies of Government and the Making of Subjects. Durham, NC: Duke University Press.
Berkes, F., Folke, C. (Eds.), 1998. Linking Social and Ecological Systems. Cambridge: Cambridge University Press.
Dietz, T., Ostrom, E., Stern, P., 2003. The struggle to govern the commons. Science 302, 1907–1912. Revised and reprinted in.In: Kennedy, Science Magazine's State of the Planet 2006–2007. Washington, DC: Island Press, pp. 126–141.
Gibson, C., McKean, M., Ostrom, E. (Eds.), 2000. People and Forests: Communities, Institutions, and Governance. Cambridge, MA: MIT Press.
Hardin, G., 1968. The tragedy of the commons. Science 162, 1243–1248.
Marshall, G.R., 2005. Economics for Collaborative Environmental Management: Renegotiating the Commons. London: Earthscan.
National Research Council, 2002. Drama of the Commons. In: Ostrom, E., Dietz, T., *et al.* (Eds.), Committee on the Human Dimensions of Global Change. Washington, DC: National Academy Press.
Ostrom, E., 1990. Governing the Commons: The Evolution of Institutions for Collective Action. New York: Cambridge University Press.
Ostrom, E., 2005. Understanding Institutional Diversity. Princeton, NJ: Princeton University Press.
Ostrom, E., Gardner, R., Walker, J., 1994. Rules, Games, and Common-Pool Resources. Ann Arbor: University of Michigan Press.

Urban Metabolism

Yan Zhang, Beijing Normal University, Beijing, China

Introduction

The complex environmental issues caused by humanity's huge and growing urban resource consumption and pollutant discharge has become an increasingly prominent problem. Facing this dilemma, researchers have looked for ways to understand the components of the problem to guide their search for solutions. For more than 50 years, some have wondered whether the flows of energy and materials through cities mimic the processes in natural systems (ecosystems and organisms) and whether the theoretical insights gained from natural systems can provide insights into urban systems. If so, this would let urban managers mitigate the environmental problems of these hybrid artificial–natural systems. This raises the question of what urban managers could do to better mimic the metabolism of natural systems, which are highly efficient and sustainable.

In this context, the concept of urban metabolism was proposed more than 50 years ago to provide a framework and research perspective for understanding cities, and a suite of methods for performing the analysis. By quantifying the material and energy flows that occur through resource exploitation, transformation into goods and services, and the associated pollutant emission, urban metabolism research integrates the natural system (the environment) that supports all human systems with the human (socioeconomic) system to holistically analyze all aspects of resource utilization and pollutant production. This will let managers identify and relieve environmental issues that arise during urban development.

During the evolution of this research, the scales of analysis have expanded from considerations of individual households and cities to regional, national, and even global scales. Simultaneously, researchers have diversified their approaches to understanding urban metabolic processes and their driving forces from the perspectives of political ecology, human ecology, urban ecology, and other disciplines. However, because of their different perspectives, researchers from different disciplines produce different understandings.

Definitions

Since the late 1800s, researchers have tried to comprehend the connotations of material and urban metabolisms from the perspectives of sociology and political economics. Karl Marx first discussed the concept of materials metabolism in his book *Capital*, in which he used the concept of metabolism to describe the exchanges of materials and energy between nature and society. In 1925, Ernest Burgess, a sociologist who considered the division and evolution of labor, proposed that these processes represented manifestations of an urban metabolism, and proposed the phrase "urban metabolism" to describe them. In 1965, Abel Wolman, a water treatment expert, first developed this concept. He adopted an engineering perspective, and applied it to a case study of how urban development affected the environment and vice versa. He treated cities as organisms with equivalents to metabolic processes, and defined urban metabolism as the inputs of materials and energy into the system and its emission of wastes.

Since Wolman's early research, many scholars have developed new interpretations and extensions of urban metabolism. They have mainly focused on processes, on interactions and relationships among the system's components, on assessing the metabolism's environmental impacts, and on seeking opportunities for improvement. Newman's and Kennedy's researches have been typical. Newman combined an urban metabolism model with a consideration of social factors (e.g., the health of residents, employment rates, education) to expand urban metabolism's scope to include human factors. Kennedy considered urban metabolism to be the sum of a city's energy production and waste elimination technologies and socioeconomic development processes.

These definitions of urban metabolism emphasized the socioeconomic system's human and technological metabolic processes and treated natural systems as just another kind of fuel or support for these processes. Researchers did not integrate natural processes into the urban metabolism or consider natural systems as distinct components of the system. In the present article, a more inclusive description that accounts for the natural environment that supports all socioeconomic systems is given. From this perspective, a city can be seen as a "superorganism," and its internal mechanisms and interactions with its environment can be modeled by analogy with a natural ecosystem. This analogy can then be used to define and quantify the interactions between the system's human and natural components. Urban metabolism is the set of processes that meet the production and living requirements of a city and its inhabitants. Thus, the inputs, outputs, transformations, and cycling of materials and energy through the system can be analyzed, including all socioeconomic (technological) and natural metabolic processes, with the goal of improving our understanding of the system's environmental, sociological, and economic characteristics. Achieving this goal lets urban planners, designers, and managers seek opportunities and measures for improvement.

Of course, the analogy should not be taken too far; cities are not organisms. Bohle noted that this analogy must be constrained by the natural laws that govern social structures and processes. Fischer-Kowalski noted that "metabolism" is a way to emphasize the flows of materials and energy and the associated processes. The purpose of the analogy is to improve our understanding, not to

propose a precise parallel for systems that are not truly equal; for example, cities do not reproduce nor do they "die." In urban metabolic research, researchers must remember the limitations of the analogy to avoid being misled.

Historical Background

Urban metabolism research can be divided into three periods.

During the first period, Wolman built a black box input–output model (i.e., one in which the internal processes of the system were not examined) based on a virtual city with a population of thousands rather than millions, during its early development. Few researchers were interested in this approach until the early 1970s, when the practical research began with studies of Miami, Tokyo, Brussels, and Hong Kong. These studies mostly used material and energy flow analysis. During this period, researchers began to consider the technological and biological metabolisms of a city, although this approach was overlooked for a long time by subsequent researchers. Meanwhile, the theoretical research evolved. E.P. Odum modeled an urban system's heterotrophic characteristics (i.e., assumed the system functioned in a manner similar to that of consumers in an ecosystem, without producers such as plants providing the energy and materials that are consumed), thereby laying the foundation for future quantitative analysis. The systems ecologist H.T. Odum analyzed the relationships between humans and their environment in terms of energy, and used metabolic energy to represent the metabolic processes involved in producing and consuming organic matter (i.e., photosynthesis and respiration, respectively) in socioeconomic systems. Based on this approach, he proposed the concept of "emergy."

During the second period (from 1981 to 2000), researchers developed new methods, and standardization of their results became a problem. It became difficult to reliably combine or compare results from different studies, and this problem continues to constrain urban metabolism research. However, the new concepts of hierarchies among the components of a system and of urban parasitism to account for nonreciprocal relationships within the system have strengthened urban metabolism theory; in turn, this has strengthened model development. Akiyama noted that black box models and subsystem models are both legitimate approaches, but for different purposes. Newman extended input–output metabolic models to account for the dynamics and livability of communities, and established a dynamic model that better combined the social and economic perspectives on urban systems. Girardet focused on the link between urban metabolism and sustainability based on the differences between linear (one-way) and circular (recycling) flows. He proposed a sustainable circular metabolic model, thereby laying a foundation for research based on the methods of industrial ecology.

During the third phase of urban metabolism research (since c. 2000), journals, conferences, collaborative projects, and reports emerged in large numbers. Simulation models progressed rapidly, reflected in further application of circular metabolic flows, dynamic metabolic models, and the establishment and development of network models. Network models go beyond traditional black box models to deeply analyze the system's internal structure, transformation processes, and interactions among the system's components. Unfortunately, most researchers still ignored or greatly simplified interactions with the city's external environment. However, researchers have developed dynamic models that account for changes over time. Examples include dynamic models for built environments, dynamic evaluation and simulation models for metropolitan subprocesses, and dynamic models for urban water. Although different kinds of models have been developed, it has been difficult to make these models spatially explicit using tools such as geographic information systems to support urban planning, which requires network models to support real-world planning. This has been due mostly to a lack of spatially explicit data to use as inputs for the network models.

Throughout the evolution of urban metabolism research, scholars have tried to understand urban metabolism from different perspectives. Political ecology uses the theoretical categories from political economics to study the conversion of materials and energy and the effects on the social environment caused by decision-making processes such as urban development and resource exploitation. Human ecology and social ecology analyze how social paradigm shifts influence the flows of materials and energy from the perspective of historical changes, and analyze how changing social institutions can transform an urban metabolism.

Urban ecology researchers have proposed three paradigms to guide their research, intending to explore the relationship between humans and nature and promote sustainable urban development. These paradigms are "Ecology in the city," "Ecology of the city," and "Ecology for the city," "Ecology in the city" resembles ecology's narrow scope, and focuses on the city's natural elements, including the structure and distribution of animal and plant resources and their service functions, and treats socio-economic activities as external stressors. "Ecology of the city" focuses on flows of materials and energy through socioeconomic systems and their ecological and environmental impacts; because cities represent highly concentrated human activity, the research focuses on these intense socioeconomic activities, and considers the natural environment as an external constraint on these activities. "Ecology for the city" combines aspects of the other paradigms by focusing on the complex hybrid socioeconomic and natural ecosystem, and treats the natural environment and human activities as equally important, thereby requiring that development decisions account for both aspects of urban development.

"Four in One" Methodology of Urban Metabolism

Based on the description of urban metabolism concepts, researchers should describe metabolic pathologies (i.e., problems with an urban system's functioning), diagnosis of urban metabolic disorders, the etiology of these disorders, and treatments. To do so, a

"four in one" methodology is proposed, which combines process analysis with accounting evaluations, model development, and efforts to find optimal regulation mechanisms (Fig. 1).

Process Analysis

Process analysis evolved from studies of linear input and output processes to studies of cyclical processes that convert the output from one system component to the input for another component, and then to network analysis methods that improve on "black box" models by uniting these flows into an overall model of the network that reveals the system's inner details (Fig. 2).

By identifying the different actors within a metabolism, network analysis can focus on the production processes of different areas, industries, or sectors. As a result, it can examine activities that drive flows of materials and energy and their environmental impacts. Thus, it can study the metabolism of industrial parks, industries, communities, and families, among others. By identifying the metabolism's components, network analysis can trace the flows of energy and key materials (e.g., chemical elements such as carbon or nitrogen; materials such as water) between materials (analogous to metabolites) and their environmental impacts.

Before conducting this analysis, it's necessary to define the system's boundary and its main metabolic pathways, as well as external areas that support the system by providing energy or materials. Researchers can then analyze the main actors that create flows along these pathways, including flows between the system and its external environment. The interactions among actors along multiple pathways define the network's structure and a complex network of flows. This model can then be used to track the inputs, outputs, transformations, and cycling of materials and energy. That is the subject of the second step in the "four in one" model.

Accounting and Assessment

The main accounting methods in urban metabolism research are material flow analysis, emergy analysis, and the ecological footprint method. Material flow analysis analyzes the flows of materials, thereby tracing the input, storage, transformation, and output processes. It focuses on classification of the material flows and on developing a balance sheet that accounts for all flows. It then characterizes the environmental impact of the metabolic processes by applying a suitable weight to each material. Emergy analysis attempts to quantitatively express all flows in a consistent set of units (typically solar emjoules, sej). This approach can convert different kinds of flow (materials, energy, and money) into units that can be directly compared, thereby allowing a comparison of very different kinds of flows. The ecological footprint method converts the environmental impacts of each flow into area units to reveal the area required to supply all inputs required for the system's survival. The analysis is based on a matrix that links consumption with land use, and can therefore analyze the relationship between a unit of land's capacity to provide one or more resources and the pressure exerted on that land by the consumption activities it supports; that is, it

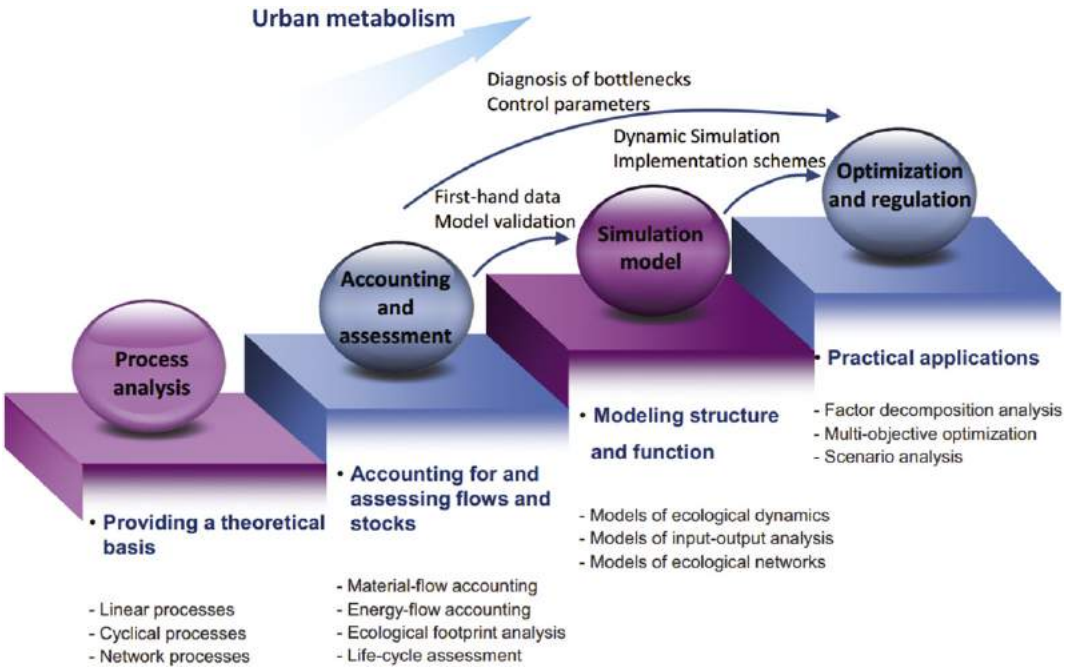


Fig. 1 Illustration of the components of the "four in one" research methodology to study urban metabolism. Source: Zhang, Y. (2013). Urban metabolism: A review of research methodologies. *Environmental Pollution* **178,** 463–473.

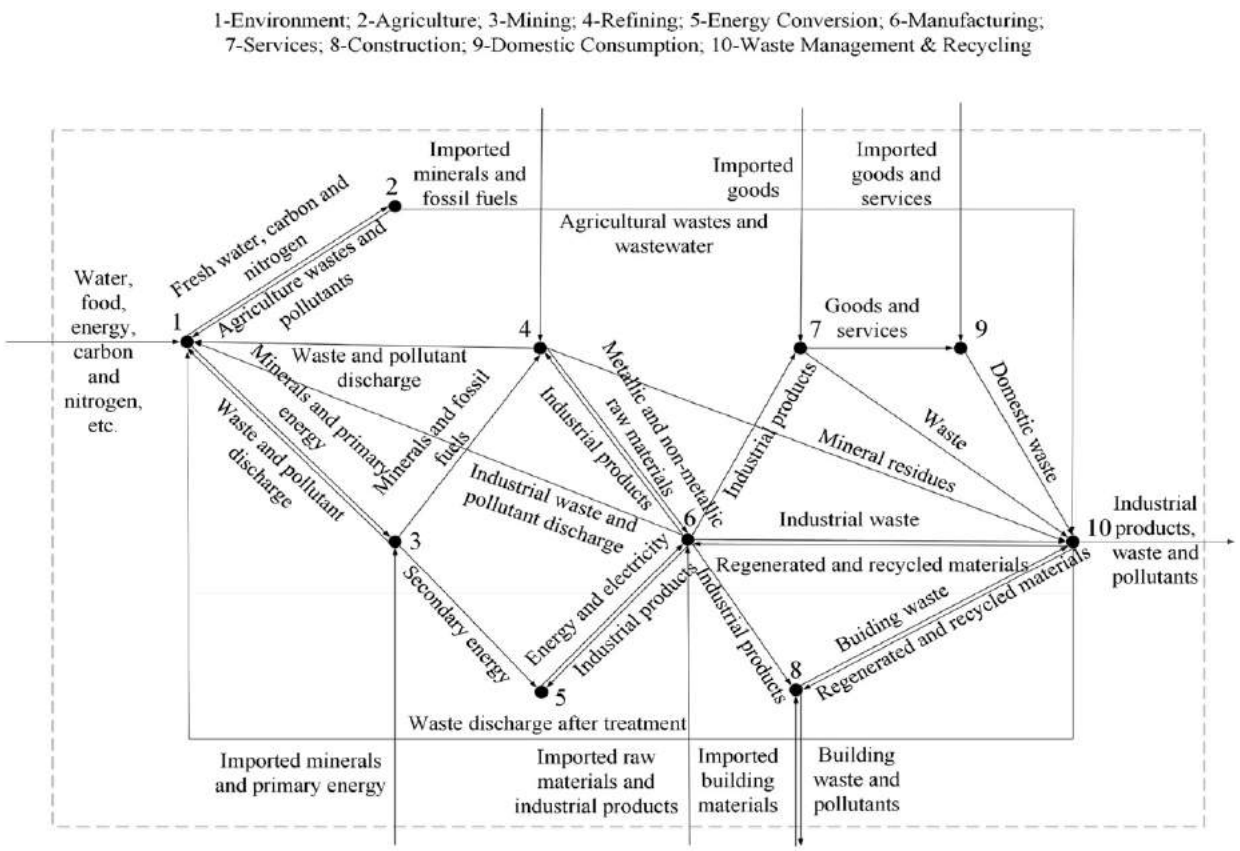


Fig. 2 Illustration of the concept of network process analysis.

represents the system's pressure on the environment and the environment's ability to sustain that pressure. **Table 1** compares the different accounting methods.

The main assessment method is life-cycle assessment, which provides a cradle-to-grave accounting of the direct and indirect energy consumption and material flows involved in the production of a product or delivery of a service. The result quantifies the environmental load of the metabolic processes at each stage of a product's or service's life cycle. Some researchers have used this analysis to compare the indicators used in metabolic accounting (e.g., direct material consumption, demand for materials, pollutant emissions) with economic parameters, the population size, or the land area. They then constructed a system of evaluation indices that include aspects of metabolic efficiency, intensity, density, and influence. However, there is uncertainty in how to divide the life-cycle stages, choose impact categories, determine characterization factors, and standardize the data.

Model Simulation

Ecological dynamics models, input–output analysis models, and ecological network models simulate the evolution of metabolic processes, the factors that influence them, and the system's internal operating mechanisms. Ecological dynamics models are an important tool for simulating flows of materials and energy in an urban metabolism because rapid advances in computer technology have made such complex simulations feasible. These models combine system dynamics methods with material accounting and emergy indicators by analyzing the system's composition and structure and the causal mechanisms responsible for flows through the system. A model of the evolution of the system's ecological thermodynamic properties and processes can be constructed, and will allow prediction of long-term dynamic changes in the system's metabolic processes. Input–output models can be used to account for all inputs into and outputs from a system. As this accounting becomes increasingly precise, it allows the development of input–output tables for the actors in the urban metabolic processes, and therefore allows simulation of the direct and indirect effects of the actors. Because flows may be measured in many units (e.g., energy, money, mass), and because some of these flows are difficult to quantify, monetary flows often serve as proxies for these different units. The main difficulty is how to convert economic data (i.e., monetary input–output tables) into physical tables that represent the quantity of a material implied by the monetary value. Ecological network models can simulate the system's direct and indirect effects on its environment, producing data on the integrated flows within the system; based on these flows, it can reveal trophic levels (e.g., consumers vs.

Table 1 Comparison of the accounting methods

Method	*Merits*	*Drawbacks*
Material flow analysis	Measures the inflows and outflows throughout a product's or service's entire life cycle within the urban system. Hidden flows can be identified to reveal the pressures on the environment. This method reflects the intensity and scale of urban activities, and can be used to compare cities	Adding the weight of different materials directly increases the substitution of resources, as it implies the possibility that any resource can be replaced by one or more other resources, and ignores the quality differences among materials. Because some materials flow through a system in many forms, it is difficult to obtain enough data for comprehensive analysis
Emergy analysis	Ensures that the flows of energy that underlie the creation and flow of all materials is accounted for along with the flows of materials, and accounts for differences in the quality of the materials or energy	Determining appropriate transformity values for specific objects or flows is a difficult problem that has not yet been solved
Ecological footprint analysis	Combines the demand created by socioeconomic development with the supply from the ecological environment, thereby revealing an ecological deficit or surplus	Treating land as a single homogeneous source of supply neglects the diversity of functions that land provides. The criteria for selecting areas that supply ecological resources are not unified, and incomplete accounting for the resources provided by natural systems and of the wastes eliminated by natural systems result in underestimation of the true values

Table 2 Comparison of the main urban metabolism simulation methods

Model	*Merits*	*Drawbacks*
Ecological dynamics model	Uses causal feedback relationships to analyze an urban metabolic system's operation and evolution, and combines social, economic, and natural elements to simulate the system's evolution	Quantitatively unifying flow processes for multiple ecological elements is difficult because there is no unified accounting method. In addition, researchers have mostly focused on simulating metabolic processes for a single element. Simulations that combine many elements have difficulty tracking all elements
Input–output analysis model	Combines economic elements with flows of materials and energy to construct an environmental input–output table that refines the description of the actors in urban metabolic processes	To conduct analyses at a given scale, it's necessary to obtain data from broader scales (e.g., the province or country in which a city is located) to account for flows into the city from these larger regions. It's difficult to combine material and energy flows with input–output tables, as it's unclear how to account for exchanges among sectors or locations (due to the limited material and energy data that is available) using an economic (monetary flow) matrix, leading to only an approximate simulation
Ecological network analysis	Combines the methods of flow, utility, and path analysis to quantitatively simulate the system's structure and function and the relationships among its components	The lack of flows among a system's subnetworks makes it difficult to refine the sectors within the network, and the ecological connotations are not yet fully understood

producers) within the system and the hierarchy of ecological relationships among the actors (e.g., competition vs. mutualism). Table 2 summarizes the merits and drawbacks of these methods.

Optimization and Regulation

To optimize a city's urban metabolic processes (i.e., maximize their efficiency or minimize their environmental impact), it's necessary to determine which factors most strongly regulate the metabolism and how to modify them. That is, researchers must identify the key driving factors to provide a basis for urban planning and management to account for those factors. The factor decomposition method and the structural decomposition method are generally adopted to identify socioeconomic factors (such as industrial and urban transformation plans, the industrial structure, the technology level, the scale and structure of final demand, income levels, population size and structure, lifestyle changes). Simultaneously, researchers should pay attention to ecological factors such as urban heat island effects, differences in roof albedo, and urban forests that influence the metabolism. Moreover, the urban form, density, and infrastructure, as well as changes in land use, influence the metabolic flows and stocks of energy and materials, and should also be considered. Based on this analysis, researchers can develop and compare scenarios using scenario analysis, factor decomposition, and multiobjective optimization to identify the possible impact of key driving factors, and define development scenarios based on the results to guide subsequent regulation of the urban metabolism. Table 3 shows the merits and drawbacks of these methods.

Table 3 Comparison of the main urban metabolism optimization and regulation methods

Model	*Merits*	*Drawbacks*
Factor decomposition	Analyzes varieties of factors that affect a variable's development (e.g., economic structure, population, energy intensity, technology level). Based on aggregated data, which is usually easy to obtain	The aggregated data is comprehensive and additive, but cannot reflect processes at more detailed scales
Structural decomposition	Since it is based on the input–output model, it can perform detailed analysis of changes in the economic structure, including changes in the production structure, final demand (structure and scale), and technology	Decomposition is based on an input–output model which includes coefficients that represent pollutant emissions or resource consumption per unit GDP. Because the structure depends on the equation's components, the effect of other factors may be neglected
Scenario analysis	Uses a small amount of data to predict alternative outcomes, and to design future scenarios that can improve perceptions of future possibilities, and support strategic planning	Depends too much on the current situation; when hypothetical scenarios are inconsistent with future developments, the results can be misleading. Can only predict some future points rather than continuous dynamic trends
Multiobjective optimization	Considers the overall objectives from socioeconomic and environmental perspectives, constructs a utility function based on constraints to coordinate multiple objectives, and thereby provides an optimal solution	When the objective has an optional range, this objective is treated as constraint to reduce the number of objectives, leading to less-certain predictions

Applications

Cities are open systems (i.e., they exchange materials and energy with their environment) and show considerable internal variation. Thus, studies of urban metabolic processes must cross multiple scales. Because metabolic activities within an urban area require support from the external environment, researchers must fully consider the environment's resource supply and ability to accept urban waste at larger scales (e.g., urban agglomeration, megacity, regional, national, or global scales). This will reveal the roles of a city within larger systems that have their own metabolic processes. However, these larger-scale studies failed to account for differences between communities and industries because they were based on coarse-resolution (aggregated) data. Therefore, it is also essential to perform analyses at a smaller scale, such as community or household scales, because data from those scales is aggregated to provide the basis for research at larger scales.

Analysis of Water Metabolic Process in Beijing

Using ecological network analysis, Zhang and her colleagues constructed a simple network model of an urban water metabolic system, and used the model to analyze Beijing's water metabolic processes. Using the trophic levels of natural ecosystems as a reference, they defined the compartments that participate in the water metabolism as producers (the ecological environment and the artificial rainwater collection system), consumers (the industrial, agricultural, and domestic sectors), and reducers (the wastewater recycling system), and determined the water flows among the system's components. They then developed a conceptual model of the urban water metabolism (**Fig. 3**). Because each component may play different roles at different times, changes in the roles of these components produce dynamic changes within a network rather than simple linear changes in the relationships between pairs of components.

Based on this analysis, they established an ecological network model of Beijing's water metabolism (**Fig. 4**). They used network throughflow analysis to determine the flows among the components, and measured both the indirect and direct flows. Using a network utility matrix, which quantifies the utility each node in the network receives from its interaction with other nodes, they determined the relationships and degrees of mutualism among the system's six compartments (**Table 4**). The capacity of producers to provide water for Beijing decreased from 2003 to 2007, and consumer demand for water decreased due to decreasing industrial and agricultural demand; the recycling capacity of reducers also improved, decreasing the discharge pressure on the environment. From 2003 to 2007, the main changes in the ecological relationships among the components of Beijing's water metabolic system mostly occurred between the local environment, the agricultural sector, and the industrial sector. The ecological relationships between the industrial sector and the local environment changed from exploitation to competition and back again, while the ecological relationship between the agricultural and industrial sectors changed from competition to exploitation and back again. Although Beijing's mutualism indices remained generally stable, the ecological relationships among compartments changed greatly. To ensure that water supplies stabilize or improve, the authors proposed that the system's rainwater collection and recycling components must continue to improve, and that water consumption by the industrial, agricultural, and domestic sectors must continue to decrease, perhaps by increasing reuse and recycling of water. In addition, measures must be taken to allow recharge of water resources.

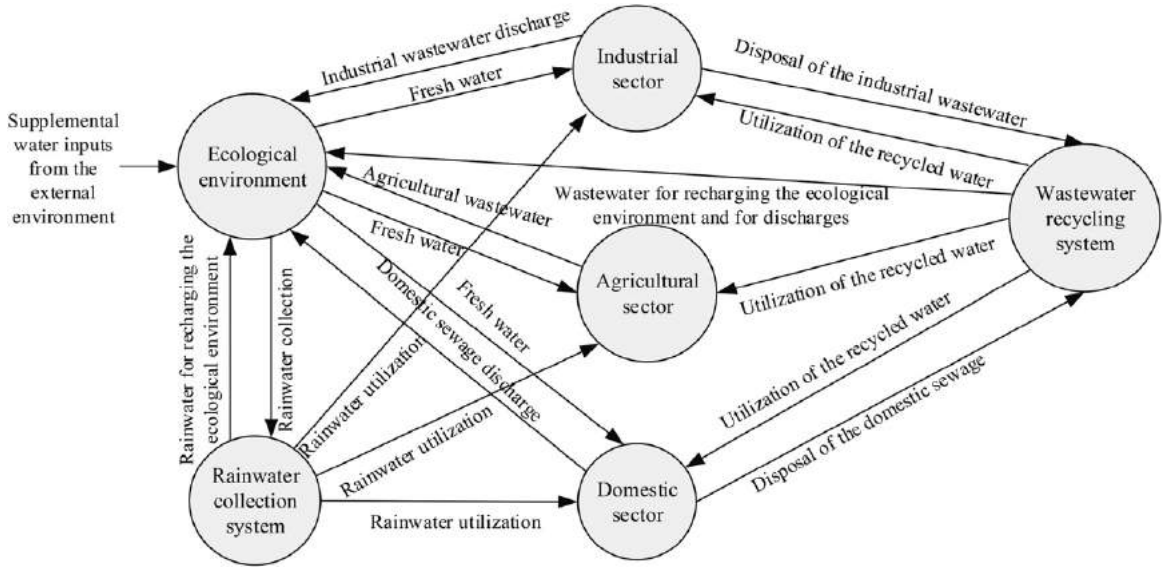


Fig. 3 A conceptual model of the water flows in Beijing's urban water metabolism. Source: Zhang, Y., Yang, Z. F., Fath, B. D., and Li, S. S. (2010). Ecological network analysis of an urban energy metabolic system: Model development, and a case study of four Chinese cities. *Ecological Modelling* **221,** 1865–1879.

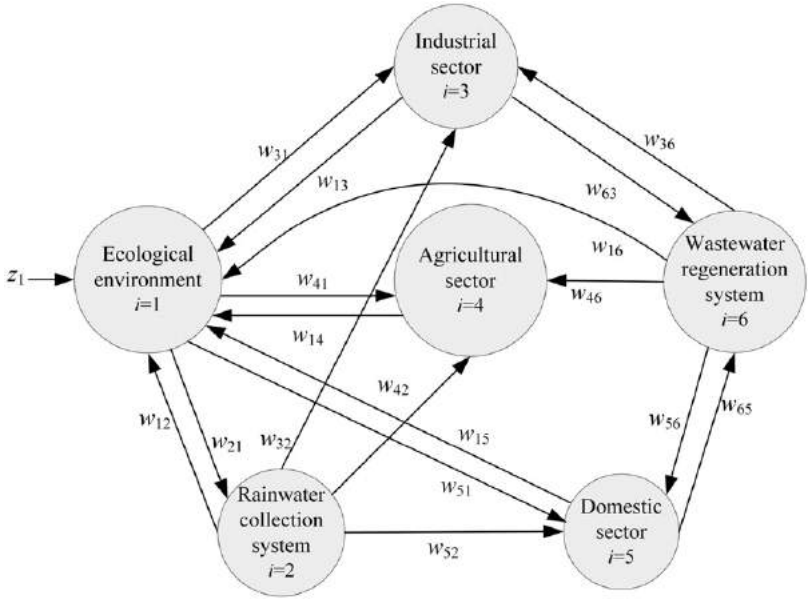


Fig. 4 Ecological network model of Beijing's urban water metabolic system. Flow w_{ji} represents the flow from node i to node j, z_1 represents the flow from the external environment to node 1. Source: Zhang, Y., Yang, Z. F., Fath, B. D., and Li, S. S. (2010). Ecological network analysis of an urban energy metabolic system: Model development, and a case study of four Chinese cities. *Ecological Modelling* **221,** 1865–1879.

Analysis of Energy Metabolic Processes in the Jing-Jin-Ji Urban Agglomeration

Zhang and her colleagues combined multiregional input–output tables with ecological network analysis to develop an analytical method, and used the Beijing–Tianjin–Hebei (Jing-Jin-Ji) urban agglomeration as a case study to illustrate how to define a network model and simulate the energy flows (e_{ij}) among the three regions and their sectors. They quantified the indirect flows and the distribution of ecological relationships among the metabolic actors. The flows began with the Agriculture and Industrial sectors, which process resources from the Earth into materials that they deliver to downstream sectors. The flows continued with the Industrial sector, which provides materials to the Construction sector to produce infrastructure (e.g., buildings, roads). Next, the Transportation, Storage, and Postal Services sector and the Other Services sector used products from upstream sectors to provide services to residents. The first component of this model was based on a multiregional input–output table, which the authors used

Table 4 Beijing's integral utility (**U**) matrix for the city's water metabolic system in 2007, and the corresponding sign (sgn) matrix

U_{2007}							$sgn(U_{2007})$						
	1	2	3	4	5	6		1	2	3	4	5	6
1	0.748	0.038	0.038	0.160	0.120	0.153	1	+	−	−	−	+	+
2	0.230	0.930	0.001	0.049	0.327	0.255	2	+	+	+	−	−	+
3	0.274	0.022	0.982	0.059	0.007	0.004	3	+	−	+	−	−	−
4	0.374	0.019	0.019	0.920	0.060	0.076	4	+	−	−	+	+	+
5	0.360	0.080	0.040	0.077	0.671	0.276	5	+	+	−	−	+	−
6	0.394	0.112	0.061	0.084	0.509	0.587	6	−	+	+	+	+	+

Utilities are calculated based on the flow between the nodes in the first column of the table to the nodes in the first row of the table. Nodes: 1, ecological environment; 2, rainwater collection system; 3, industrial sector; 4, agricultural sector; 5, domestic sector; 6, wastewater regeneration system.

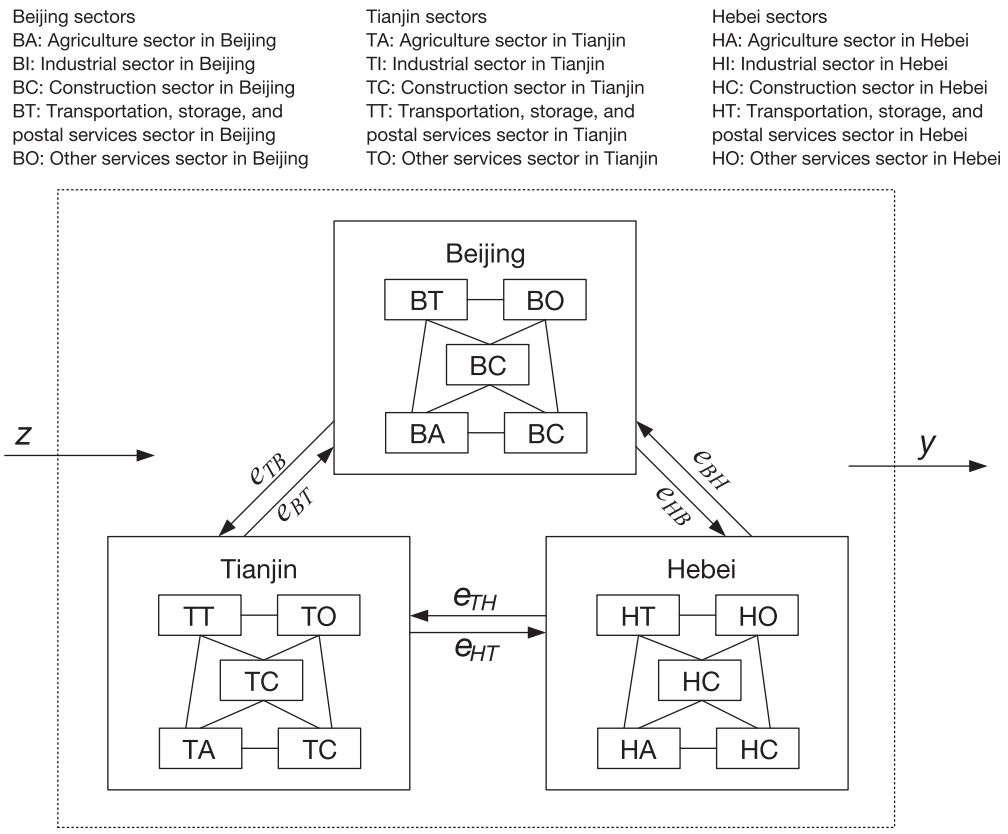


Fig. 5 The network models for the Jing-Jin-Ji urban agglomeration. Abbreviations: e_{ij}, energy flow from sector j to sector i; z and y represent flows into the system from the external environment and from the system to the external environment, respectively. From Zhang, Y., Zheng, H., Yang, Z., Li, Y., Liu, G., Su, M. and Yin, X. (2016). Urban energy flow processes in the Beijing–Tianjin–Hebei (Jing-Jin-Ji) urban agglomeration: Combining multi-regional input–output tables with ecological network analysis. *Journal of Cleaner Production* **114**, 243–256.

to establish a steady-state network model for the relationships among Beijing, Tianjin, Hebei, and their external environment as well as the relationships among the five sectors within each region (Fig. 5).

Hebei had the largest embodied energy consumption in 2007, followed by Beijing (Fig. 6). In all three regions, energy consumption by the Industrial sector was largest, but with different trends from 2002 to 2007. The consumption by most sectors in Tianjin and Hebei increased; only Tianjin's and Hebei's Agriculture sectors, Tianjin's Other Services sector, and Hebei's Construction sector showed decreased energy consumption. Beijing showed the largest number of sectors with decreased consumption (five sectors); Hebei and Tianjin each had two sectors that decreased. The ecological roles of Hebei (producer) and Beijing (consumer) did not change. However, Tianjin changed from a secondary consumer to a primary consumer. Based on the ecological relationships among the regions and sectors, they found that although the government's plans for the Jing-Jin-Ji agglomeration were intended to increase the efficiency of the interactions among sectors within and between the three regions, this goal has not been achieved. Exploitation relationships (in which one region benefited at the expense of another) and control relationships (in

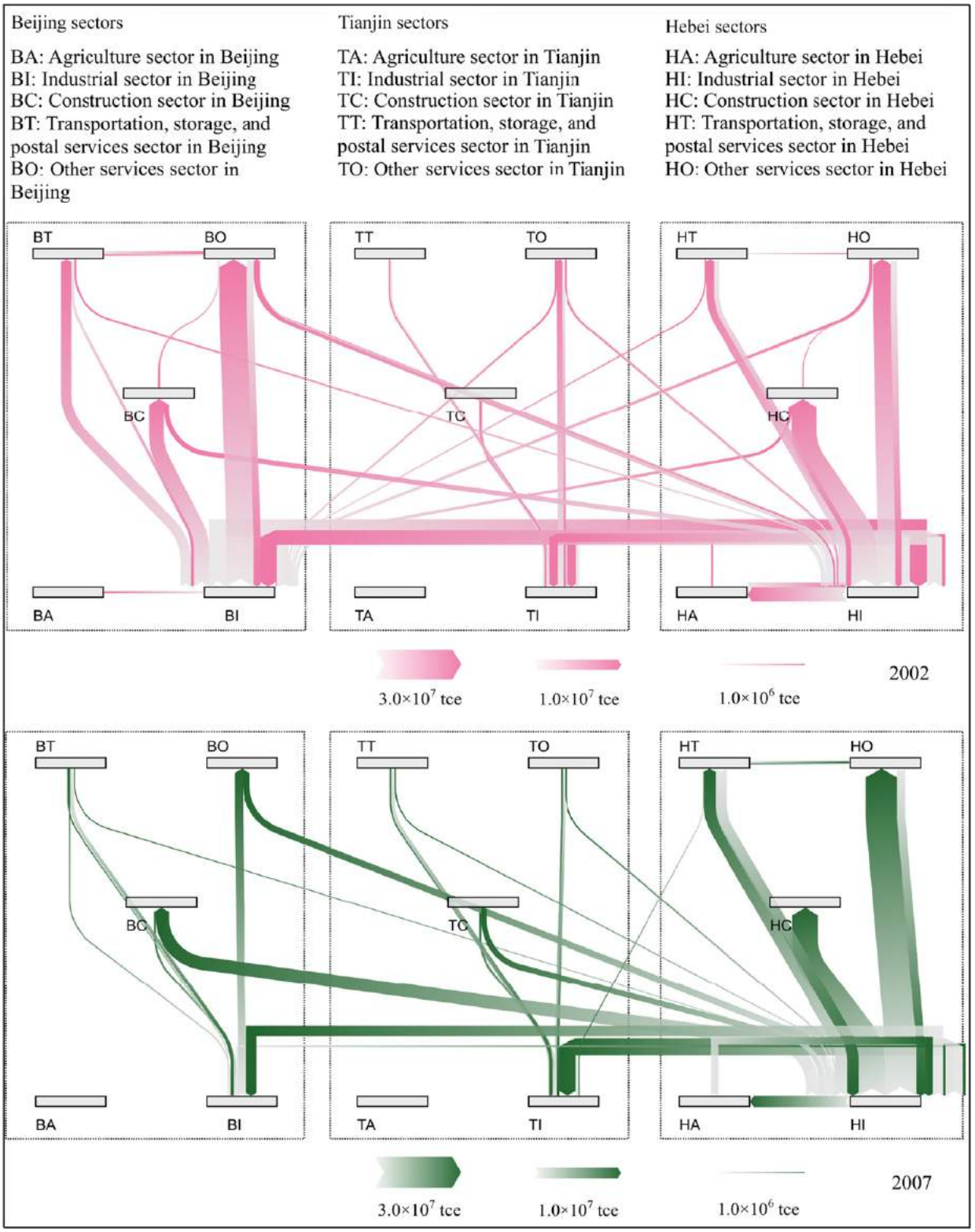


Fig. 6 The total embodied energy flows among sectors in the Jing-Jin-Ji agglomeration that were greater than 1.0×10^6 tonnes coal equivalent (tce) in 2002 and 2007 (1.0×10^6 tce). From Zhang, Y., Zheng, H., Yang, Z., Li, Y., Liu, G., Su, M. and Yin, X. (2016). Urban energy flow processes in the Beijing–Tianjin–Hebei (Jing-Jin-Ji) urban agglomeration: Combining multi-regional input–output tables with ecological network analysis. *Journal of Cleaner Production* **114**, 243–256.

which one region controlled the flows from another) dominated the three regions throughout the study, followed by competition relationships (in which both regions suffer from their relationship), revealing a strong need to reduce competition and encourage the development of more mutualisms (in which both regions benefit from their relationship). The low number of mutualisms suggests that the agglomeration program has much room for improvement. Planners should examine the sectors in each region to determine whether there are ways to transform competition into exploitation and exploitation into mutualism.

Urban Design Based on an Analysis of Metabolic Processes

Existing multiple-scale analyses have been primarily theoretical, with an emphasis on the rules that govern the system, and application of such research to actual planning exercises has been rare. Examples include the redesign of the Toronto Port Lands and the urban redesign after Hurricane Katrina struck New Orleans.

The graduate students of the Civil Engineering Department of the University of Toronto applied urban metabolism concepts at a neighborhood scale to redesign the Toronto Port Lands. They divided the neighborhood into water, transport, building, energy, and open space components, and used their experience with green construction design, alternative energy sources, and sustainable transport to design a closed-loop recycling infrastructure that reduced the inputs of resources and outputs of wastes from their study system.

Fernández's team from the MIT Construction School conducted a successful case study of redesign after Hurricane Katrina struck New Orleans. They developed software that let them use material-flow analysis to account for the inputs and outputs of materials and energy, the durability of the housing stock, the costs of different building types, energy consumption rates, waste generation rates, data on population and employment, housing needs, and growth priorities. Based on this analysis, they proposed goals for rebuilding the city and making it more resistant to hurricane damage and more green. They hope to increase New Orleans' resource-utilization efficiency in a way that provides a model for other cities, while rebuilding the city in a way that balances the needs of the many stakeholders and applies scientific rigor to urban design. Using this tool, New Orleans' urban planners can model the overall green city or target specific neighborhoods, and can make better-informed decisions about rebuilding strategies.

The Future of Urban Metabolism Research

The theory, methods, and models of urban metabolism research have been greatly improved through long-term research and development. However, many problems must still be solved. The lack of a unified methodology is slowing research on urban metabolism. Therefore, it is necessary to establish a more unified approach and to standardize data classification and inventory to ensure that each study provides the basic data required for analysis and evaluation, while also facilitating comparisons among studies.

In addition, data that is aggregated at urban or regional scales cannot yet associate metabolic processes with specific locations, activities, or humans. For example, data on industrial and commercial water and energy use is often difficult to obtain at sufficient resolution to let researchers map the data to specific land uses. This also greatly limits the ability to decompose large-scale data at a spatial resolution sufficient to support real-world planning and design. Micro-scale studies at the level of families, communities, industrial parks, and businesses would provide data with sufficient resolution to support larger-scale studies and real-world plans.

In the future, geographical information system technology, remote sensing data, and information network technology should be combined to acquire sufficient spatially explicit data to study material and energy flows at scales sufficiently small to support planning, since planning usually occurs at these smaller scales. Problems with metabolic flows can be identified by means of bottom-up analysis at the scale of individuals and communities using segmentation models, and can be used to guide efforts to reduce the size of the flows. Such analyses would support the implementation of green buildings, community-greening landscape designs, and changes in family consumption patterns. It would also permit the formation of a closed-loop design in which all or most materials and energy are recycled at the micro scale, which would promote the development of a circular economy on an urban scale, regional scale, or even global scale by closing open loops that are identified at smaller scales.

Many theories and methods have been developed to promote urban metabolic research, but the problems identified in this section have limited the application of urban metabolic methods. Therefore, in future research, it will be increasingly necessary to use ideas from systems engineering and other disciplines, using a multidisciplinary approach, to establish a more comprehensive and unified knowledge system capable of meeting the needs of decision-makers and land managers, as well as to meet the needs of different stakeholders.

See also: Human Ecology and Sustainability: Adaptive Management and Integrative Assessments; Human Population Growth

Further Reading

Baccini, P., Brunner, P.H., 2012. Metabolism of the anthroposphere: Analysis, evaluation, design. Cambridge: MIT Press.
Girardet, H., 1992. Cities: New directions for sustainable urban living. London: Gaia Books.
Jørgensen, S.E., 2000. Thermodynamics and ecological modeling. Boca Raton: CRC Press.

Kennedy, C., Cuddihy, J., Engel-Yan, J., 2007. The changing metabolism of cities. Journal of Industrial Ecology 11, 43–59.
Odum, H.T., 1988. Self-organization, transformity, and information. Science 242, 1132–1139.
Wolman, A., 1965. The metabolism of cities. Scientific American 213, 179–190.
Zhang, Y., 2013. Urban metabolism: A review of research methodologies. Environmental Pollution 178, 463–473.
Zhang, Y., Yang, Z.F., Fath, B.D., Li, S.S., 2010. Ecological network analysis of an urban energy metabolic system: Model development, and a case study of four Chinese cities. Ecological Modeling 221, 1865–1879.
Zhang, Y., Yang, Z.F., Yu, X.Y., 2015. Urban metabolism: a review of current knowledge and directions for future study. Environmental Science & Technology 49, 11247–11263.
Zhang, Y., Zheng, H.M., Fath, B.D., Liu, H., Yang, Z.F., Liu, G.Y., *et al*., 2014. Ecological network analysis of an urban metabolic system based on input-output tables: Model development and case study for Beijing. Science of the Total Environment 468–469, 642–653.

Urban Systems

T Elmqvist, Stockholm University, Stockholm, Sweden
C Alfsen, UNESCO, New York, NY, USA
J Colding, Royal Swedish Academy of Sciences, Stockholm, Sweden

Introduction

Urbanization is a global multidimensional process that manifests itself through rapidly changing human population densities and changing land cover. The growth of cities is due to a combination of four forces: natural growth, rural to urban migration, massive migration due to extreme events, and redefinitions of administrative boundaries. Half of the world's population today lives in urban areas, a proportion expected to increase by 2/3 within 50 years. Today, over 300 cities have a population of more than 10^6 and 19 megacities exceed 10^7. As urbanization is accelerating, the growth of cities forms large urban landscapes, particularly in developing countries. (Urban landscape is here defined as an area with human agglomerations with $>50\%$ of the surface built, surrounded by other areas with 30–50% built, and overall a population density of >10 ind. ha^{-1}.) For example, during the last 20 years in China, clusters of cities have emerged forming at least five mega-urban landscapes. These large and densely urbanized regions have each between 9 and 43 large cities located in close proximity and a population ranging from 27 to 75 million people. This rapid urbanization represents both a challenge and an opportunity to ensure basic human welfare and a viable global environment. The opportunity lies in that urban landscapes also are the very places where knowledge, innovations, human and financial resources for finding solutions to global environmental problems are likely to be found.

Since urbanization is a process operating at multiple scales, factors influencing environmental change in urban landscapes often originate far beyond city, regional, or even national boundaries. Fluctuation in global trade, civil unrest in other countries, health pandemics, natural disasters, and possibly climate change and political decisions are all factors driving social–ecological transformations of the urban landscape.

Mismatches between spatial and temporal scales of ecological process on the one hand, and social scales of monitoring and decision making on the other have not only limited our understanding of ecological processes in urban landscapes, they have also limited the integration of urban ecological knowledge into urban planning. In ecology there is now a growing understanding that human processes and cultures are fundamental for sustainable management of ecosystems, and in urban planning it is becoming more and more evident that urban management needs to operate at an ecosystem scale rather than within the traditional boundaries of the city.

Although studies of ecological patterns and processes in urban areas have shown a rapid increase during the last decade, there are still significant research gaps that constrain our general understanding of the effects of urbanization processes. The vast majority of studies so far have been short term (one to two seasons), conducted in cities in Northern Europe or the US, have lacked experimental approaches, focused on either birds or plants, while other taxa are rarely represented and have only included portions of a rural–urban gradient. Most significantly, we nearly completely lack studies in rapidly growing urban landscapes in tropical developing countries that are rich in biodiversity and are just beginning to address the complexity of human settlements in the tropics.

Of further significance is that urban landscapes provide important large-scale probing experiments of the effects of global change on ecosystems, since, for example, significant warming and increased nitrogen deposition already are prevalent and because they provide extreme, visible, and measurable examples of human domination of ecosystem processes. Urban landscapes may be viewed as numerous large-scale experiments producing novel types of plant and animal communities and novel types of interactions among species, and as such deserve the full attention of not only evolutionary biologists and ecologists but also of students of social–ecological interactions.

Urbanization and Plant and Animal Communities

Urbanization is today viewed to endanger more species and to be more geographically ubiquitous than any other human activity (Fig. 1). For example, urban sprawl is rapidly transforming critical habitats of global biodiversity value, for example, in the Atlantic Forest Region of Brazil, the Cape of South Africa, and coastal Central America. Urbanization is also viewed as a driving force for increased homogenization of fauna and flora. In the urban core in Northern Hemisphere cities, a similar set of species is recorded that is, often cosmopolitan plant and animal species tolerant of anthropogenic impacts. For example, the composition of communities of wildlife species found in cities across the US is remarkably similar despite large variation in climate and geographical features. A common pattern among cities is that they often show a high turnover of species with losses of native species and gains of non-natives over time. For example, it is documented that New York has lost 578 native plant species while it gained 411, and Adelaide lost 89 while it gained 613 new plant species over a period of 166 years. Although cities may be species rich,

 doi:10.1016/B978-0-444-63768-0.00364-4

Fig. 1 Cape Town, South Africa with more than 3 million residents is located in the Cape Floristic Region, an area with the highest density of plant species in the world with more than 9600 plants species of which 70% are endemic. Through initiatives like Working for Wetlands and Cape Flats Nature, successful efforts are taken to address the large challenge of conserving precious biodiversity in fragmented natural habitats in an urban setting where poverty is widespread. These initiatives focus on building bridges between people and nature and demonstrating benefits from conservation for the surrounding communities, particularly areas where incomes are low and living conditions are poor, and encouraging local leadership for conservation action.

frequently having higher species diversity than surrounding natural habitats, this is often due to a high influx of non-native species and formation of new communities of plants and animals. A trend of increasing non-native species from the suburbs to the urban core is well documented for plants, birds, mammals, and insects. For example, in Berlin the proportion of novel species increased from 28% in the outer suburbs to 50% in the built-up center of the city. In New York, the abundance and biomass of earthworms increased tenfold when comparing rural and urban forests, mainly due to increased numbers of introduced species in urban areas. Over broad geographical scales urbanization seems to have an effect of convergence in species composition with loss of native species and invasion of exotics. Nevertheless, a remarkable amount of native species diversity is known to exist in and around large cities, such as Singapore, Rio de Janeiro, Calcutta, New Dehli, and Stockholm.

Interestingly, the number of plant species in urban areas often correlates with the human population size. Species number often increases with log number of human inhabitants, and that relationship is stronger than the correlation with city area. The age of the city also affects species richness; large, older cities have more plant species than large, younger cities. Also of interest is that diversity may correlate with measures of economic wealth. For example, in Phoenix, USA, measures of plant and avian diversity in urban neighborhoods and parks show a significant positive correlation with measures of median family income levels.

In general, urban landscapes present novel ecological conditions, such as rapid rate of change, chronic disturbances, and complex interactions between patterns and processes. Organisms that have survived in urbanized areas have been able to do so for at least two reasons: (1) they evolved rapidly and adjusted genetically or (2) they were largely preadapted to this environment and required little or no genetic adjustment. There are several documented cases of rapid evolution in urban areas, involving, for example, tolerance to toxic substances and heavy metals in plants, such as lead tolerance in urban roadside *Plantago lanceolata*. Among insects there are many cases of rapid evolution in urban areas, notable example being the famous case of industrial melanism among Lepidoptera in UK, a phenomenon also documented from areas in USA, Canada, and elsewhere in Europe. Also of interest is that specific urban and rural races have been identified within well-studied *Drosophila* species.

In **Table 1** we have summarized some of the effects of urbanization including abiotic and biotic changes. Human activities may cause increased deposition of nutrients such as nitrogen and phosphorus and emission of toxic chemicals which influence urban soil processes. Decomposition rates in urban soils are often negatively affected by pollution and toxic chemicals, but positively affected by increased soil temperature. Decomposition rates may therefore often be higher in urban than in rural soils. However, urban litter tends to have higher C:N ratios and therefore also tends to be more recalcitrant than rural litter. Urbanization affects in complex ways both directly and indirectly C-pools and N-transformation rates and contrasting dynamics in urban and rural soils is an area where much more research is needed.

Biotic changes influencing ecosystem functioning are listed in **Table 1**. There are a number of reasons why new human-imposed scales for ecological processes are found within urban areas. First, compared with ecosystems in rural areas, urban systems are highly patchy and the spatial patch structure is characterized by a high point-to-point variation and degree of isolation between patches. Second, disturbances such as fire and flooding are suppressed in urban areas, and human-induced disturbances are more prevalent as well as intense human management of urban habitats. Third, because of the `heat-island' effect, that is, higher mean temperatures in cities than in the surroundings, cities in temperate climates have significantly longer vegetation growth periods. Fourth, ecological successions are altered, suppressed, or truncated in urban green areas, and the diversity and structure of communities of plants and animals may show fundamental differences from those of nonurban areas. In general, with increasing urbanization there is a trend toward dominance of generalist species with high reproductive capacity and short generation times.

Table 1 Ecological effects of urbanization

Physical and chemical environment	*Population and community characteristics*	*Ecosystem structure and function*
Air pollution increases	Altered reproductive rates	Altered disturbance regimes
Hydrological changes	Genetic drift, changes in selection	Altered succession
Local climate change	Social and behavioral changes	Altered decomposition rates
Soil changes	High species turn over, increase of exotic species	Altered nutrient retention
Water changes	Loss of K-species and gain of r-species	Habitat fragmentation
	Increased dominance of generalist species	Changes in trophic structure, domination of omnivores

Modified from McDonnell, M.J., Pickett, S.T.A., 1990. Ecosystem structure and function along urban–rural gradients: An unexploited opportunity for ecology. Ecology 71, 1232–1237.

Urban Habitats and Gradient Analyses

Urban habitats are extremely diverse and examples include parks, cemeteries, vacant lots, streams and lakes, gardens and yards, campus areas, golf courses, bridges, air ports, and landfills. These habitats are highly dynamic, influenced by both biophysical and ecological drivers on the one hand and social and economic drivers on the other. Urban landscapes often represent cases of extreme habitat fragmentation. Habitat patches in the urban core are more or less strongly isolated from each other by a matrix of built environment making dispersal risky and difficult at least for poorly dispersing organisms. There are numerous studies analyzing effects of isolation of urban habitats and, for example, in urban gardens in UK, the best predictor of species richness of ground arthropods was found to be the proportion of green areas within a 1 km radius of the sampling site. Analyses of the distribution of plant species in vegetation fragments in Birmingham, UK showed a positive correlation between the density of patches available to a species and the proportion of these patches that was occupied. For many plant species the rate of occupancy increased with site age, area, and similarity of adjacent habitats. Similarly, for urban amphibian assemblages in Melbourne, Australia, an increase in species richness was associated with pond size and a decrease with increasing isolation. Habitat quality also influenced species composition. The importance of isolation is likely to increase over time and, for example, in Boston an isolated urban park lost 25% of its plant diversity over 100 years. To what extent greenways and corridors increase connectivity and contribute to maintain viable populations in urban green areas is poorly understood. But greenways may, in multiple ways, provide a chain of different habitats permeating the urban environment and be of benefit for many organisms. Apart from preventing local extinction and facilitating re-colonization, increased habitat connectivity is important to maintaining vital biological interactions, for example, plant–pollinator interactions and plant–seed disperser interactions. Although most of the studies in urban landscapes address the continuous loss of green areas due to urban growth and expansion, this is not the case in all cities. For example, in Shanghai the proportion of green areas has increased in parallel with urban expansion and the total area expanded from less than 9 km^2 in 1975 to more than 250 km^2 in 2005.

Gradient analyses have a long tradition in ecology and go back to the pioneering work by Whittaker in the late 1960s. Gradient analyses have also been a rather common way of disentangling the complexity of urban habitats and have been used to investigate how urbanization changes ecological patterns and processes across landscapes, for example, in invertebrate, plant, and bird community composition, leaf litter decomposition and nutrient cycling, and the structure of landscape elements. Almost all the gradients that have been used for urban studies have been one dimensional in the sense that they only describe physical features of the gradient such as proportion of impervious surface, while the characteristics of the human population occupying a particular portion of the landscape often have been neglected. Because urbanization is an exceedingly complex amalgamation of factors, by using only a single axis the interpretation of the underlying processes has often been severely constrained. It has been suggested that a more comprehensive gradient analysis should include not only physical geography, demography, rates of ecological processes, and energy, but also history of land use, socioeconomic analyses, and patterns of management.

Variation in species densities across an urban gradient suggests that some individual species disappear with urbanization, whereas other species invade in response to the environmental changes associated with development. At least for birds, species richness has often been found to peak at intermediate levels of urbanization and decrease with either more or less development. Some species are classified as urban avoiders with their highest densities at the most natural sites, whereas many species seem to be able to adapt to suburban environments, with densities peaking at intermediate levels of development. Some species are urban exploiters whose highest densities are found at the urban core (see **Fig. 2**).

A multitude of factors are likely to influence this pattern of extinction and colonization, of which changes in predation rates have been suggested to be among the most important. Predation on artificial nests has often been found to be higher in urban parks than in neighboring woodlands and the abundance of predators such as corvids, rats, and house mice are often more in urban parks compared to the rural end of the gradient. However, there are also studies showing no correlation or a declining predation pressure along the urban–rural gradient. Observed patterns of extinction and invasion in urban landscapes may also be linked to gaps in the spectrum of body masses exhibited in the community and there are documented cases that body mass patterns are correlated with invasion and extinction in other human-transformed ecosystems.

In cities, ownership and management of urban habitats is extremely diverse and complex. In addition to land managed by government, municipalities, churches, and foundations, there is also land managed by local user groups that often covers substantial tracts. For example, domestic gardens cover 23% of the land area of Sheffield, and as much as 27% of the city of

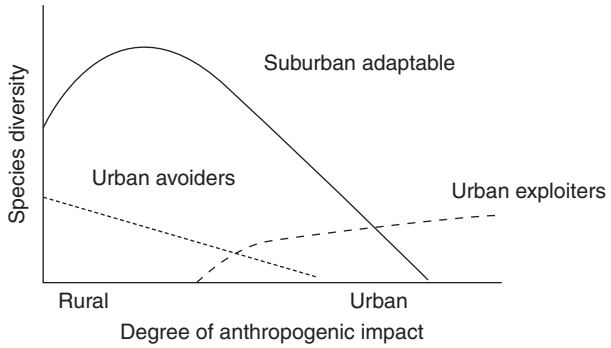


Fig. 2 Plants and animals may respond differently to increasing human impact. Urban avoiders are large-bodied species or species linked to late successional stages. These species might be very sensitive and show a decline already at moderate human impacts. Suburban adaptable speceis may, to various degrees, utilize human modifications of the landscape; the majority of plant and animal species likely belong to this group. Urban exploiters directly benefit from human presence for food, reproduction or protection, and may often be cosmopolitan, generalist species. Terminology after Blair, R.B., 1996. Land use and avian species diversity along an urban gradient. Ecological Applications 6 (2), 506–519.

Leicester. Lands appropriated for allotment areas, domestic gardens, and golf courses were found to cover nearly 18% of the total green space of greater Stockholm, Sweden, representing well over twice the area covered by nature-protected areas. While the numbers of ecological studies of urban green areas are limited, there is evidence that different types of urban green areas used for purposes other than biodiversity conservation play an important role in sustaining urban biodiversity. For example, in many cities of Asia, educational institutions sometimes harbor the largest and last remaining green areas in extensively urban developed settings. These campus areas can be extremely significant for biodiversity. A good example is the university campuses of Pune city, India, which harbor up to half the plant, bird, and butterfly species of the region despite the fact that these campuses only cover some 5% of the land area. Another illustrative case is the Musahi Institute of Technology in Yokohama, Japan, where a former community-managed forest now restored for student education, has revived interest in reducing the loss of biodiversity-rich forests in semiurban areas of Japan. Ecological studies also show that domestic gardens sometimes hold a rich flora of plants, including rare and threatened ones. Thompson and colleagues found that the private gardens in Sheffield, UK, contained twice as many plant species as any other habitat assessed. These gardens also supported surprisingly high numbers of invertebrates and this regardless of whether garden plants were native or alien. Even such a controversial land use as golf courses can contribute with important biodiversity functions in cities when courses are wisely managed and well designed. For example, golf courses contribute to sustain urban woodlands in many cities of Japan, and in some larger cities of Sweden they may harbor significant populations of species of both amphibians and macro-invertebrates, which are declining in rural areas.

Also, smaller habitat parcels in urban settings can provide high-quality habitats. One illustrative example of this is allotment areas, common in many city-regions in developed countries. For example, while allotment areas only cover some 0.3% of the land in greater Stockholm, they tend to be extremely biodiversity rich. In Stockholm city one allotment garden was found to contain 447 different plant species in an area of 400 m^2.

Urban Systems and Ecosystem Services

The concept of ecosystem services has proved to be useful in describing human benefits from urban ecosystems. For example, urban vegetation may significantly reduce air pollution, mitigate the urban heat island effect, reduce noise, and enhance recreational and cultural values, of importance for urban citizen's well-being (**Table 2**).

The scale of importance for generation of these services is often much larger than a city, for example, for reduction of air pollutants and water regulation, while for some, such as recreational and educational services, generation often occur within city boundaries. In most cases, these services tend to be overlooked by urban planners and decision makers, despite the fact that the potential of generation of ecosystem services can be quite substantial. In a study made within Stockholm County, it was assessed that this region's ecosystems potentially could accumulate about 41% of the CO_2 generated by traffic and about 17% of total anthropogenic CO_2. In the Chicago region, trees were found to remove some 5500 t of air pollutants per year, providing a substantial improvement in air quality. It was also found that the present value of long-term benefits from the trees of the Chicago region was more than twice the present value of costs related to planting. Moreover, wetlands in urban settings can substantially lower the amount of money spent on sewage treatment costs and in many cities large-scale experiments are taking place where wetlands are being used to treat sewage water. It has been estimated that up to as much as 96% of the nitrogen and 97% of the phosphorus can be retained in wetlands through the assimilation of wetland plants and animals. Green spaces of cities also provide ample opportunity for recreation. In a study on the response of persons put under stress, it was shown that when subjects of the experiment were exposed to natural environments the stress level decreased, whereas during exposure to built-up urban environments the stress levels remained high or even increased.

Table 2 Examples of services generated by urban ecosystems

Ecosystem services
Supporting services
Soil formation
Nutrient recycling
Provisioning services
Freshwater
Food, fiber, and fuel
Genetic resources
Regulating services
Air quality
Local climate regulation
Water purification and waste treatment
Biological control
Pollination
Cultural services
Esthetic and recreational
Educational

A major challenge in urban areas is how to sustain the capacity to generate ecosystem services. This capacity is mainly but not exclusively related to the diversity of `functional groups' of species in a system, like organisms that pollinate, graze, predate, fix nitrogen, spread seeds, decompose, generate soils, modify water flows, open up patches for reorganization, and contribute to the colonization of such patches. In urban areas, such functional groups may be substantially reduced in size or show changes in the composition due to high species turnover, both of which may increase vulnerability in maintaining ecosystem services. To what extent exotic species contribute to reduce or enhance the flow of ecosystem services is virtually unknown for any urban area. But, since introduced species make up a large proportion of the urban biota, it is important to know not only to what extent introduced species are detrimental, but also to what degree some of the introduced species may enhance local diversity and maintain important functional roles. For urban ecology to significantly contribute to improving management of urban habitats and the maintenance of ecosystem services, the following research questions are particularly urgent to address:

- To what extent are urban ecosystems sinks for many animal and plant species and what are the effects of species loss on ecosystem functions?
- To what extent do novel species play important functional roles in urban ecosystems by replacing role of extinct native species and enhancing ecosystem functions and services?
- What is the importance of source–sink dynamics and matrix permeability for maintenance of urban biodiversity and important ecosystem services?
- How do we develop management systems that match the spatial and temporal scales of ecological processes?

Urban Restoration

Designed, replicated urban restoration experiments could substantially advance our knowledge and understanding of processes of importance for generating urban ecosystem services, for example, through better understanding of population and community responses to disturbances, patterns of self-organization and succession, assembly rules, and through identifying components that contribute to resilience or vulnerability. Urban restoration also represents an interesting opportunity for ecologists to work in partnership with landscape architects, urban designers, and architects and help in designing urban environments based on ecological knowledge but merged with the functional and esthetic design of urban space.

The majority of urban restoration projects deal with transforming brown areas (abandoned industrial lots or air fields, landfills, etc.) to functioning green areas, such as Fresh Kills landfill on Staten Island, New York, or the Olympic Park in Beijing. Other large-scale restoration projects involve, for example, substantial wetland restoration such as Kristianstad, Sweden, and New Orleans, USA. In New Orleans, coastal wetlands have eroded substantially during the last 50 years and restoring wetlands is viewed as one important measure to reduce vulnerability to hurricanes. Important lessons of urban restoration projects so far are that restoring ecological functions in urban areas is possible but time consuming and that there are often significant effects on many ecosystem services even after one or two seasons. Although the costs are initially high, these could be offset by increases in property values and increased investments in development in areas surrounding the restoration site.

Urban Landscapes as Arenas for Adaptive Management

In cities that experience rapid social and environmental transformation, it is critical to develop a capacity to respond to potential surprises. One important aspect of such capacity building is to facilitate for a wider integration of local people and interest groups in the use and management of urban green areas. There are several reasons for a wider integration of local people in urban ecosystem management. First, governments cannot entirely rest on protected area management to safeguard the native flora and fauna found in city-regions. As cities expand, there will be an increased lack of natural lands for the establishment of protected areas. Studies also show that many urban nature reserves are unable to sustain native species in the long run. In addition, protected area management is financially costly to most local governments. In London, for example, parts of the protected green belt have been severely degraded due to lack of money and this has resulted in urban residents avoiding these areas for various activities. Second, much of the flora and fauna depend on well-functioning habitats provided by privately owned lands. In the US, for example, almost two-thirds of all the endangered and threatened species depend on private lands for their continued existence. Also, urban homeowners with gardens have been engaged to support declining pollinator populations in Great Britain through the deliberate planting of certain nectar providing plants in their yards. They have also helped sustain urban frog populations during their period of main rural declines through a massive establishment of garden ponds. Homeowners in Britain are also involved in programs for monitoring trends in the population status of birds. Third, a number of international treaties that have been signed by national governments around the world, including local Agenda 21, the Convention of Biological Diversity, and the Malawi-principles, strive toward a decentralization of biodiversity management down to local people. Recently, the Millennium Ecosystem Assessment (MA 2005) concluded that a wider cooperation among people within different sectors in society is necessary for more efficient land use that contributes to the support of ecosystem services.

One approach increasingly used to achieve collaborative partnership in urban ecosystem management is `adaptive co-management'. The approach rests on the notion of the sharing of resource management responsibility and authority between users of ecosystems and government agencies. This typically involves local people and interest groups, scientists and local authorities, with the potential to promote information exchange to effectively deal with and respond to change and issues that often transcend locality. Adaptive co-management emphasizes `learning-by-doing' in ecosystem management, where management objectives are treated as `experiments' from which people can learn by testing and evaluating different management policies. This form of ecosystem management avoids set prescriptions of management that may be superimposed on a particular place, situation, or context. Such designs have the potential to lower overall costs of management, most notably costs incurred for describing and monitoring the ecosystem, designing regulations, coordinating users and enforcing regulations, and depend on the self-interest of participants. Co-management arenas could, for example, also serve as platforms for designed experiments and urban restoration as discussed above and improve ecological functions and designs in cities.

Linking Humans and Nature in the Urban Landscape

Urban landscapes are not only ecological experiments but also long-term experiments in social, economic, and cultural transformations shaped by cultures, property rights, and access rights. Since cities are places where knowledge, human and financial resources are concentrated, rapid urban transformations can likely be more readily monitored and observed than similar processes in more rural areas. Studies of transformations in urban landscapes may therefore well provide the ground for a better understanding of socio-economic drivers of changes also in other ecosystems. After decades of mutual neglect and artificial divide between nature on the one hand, and cities with their respective urban processes on the other hand, the conservation community has started to shift its perception to include cities as a component of natural landscapes. Just as it is now increasingly recognized that in protected nature reserves, conservation will not be successful as long as it is at the expense of human aspirations, urban planners increasingly acknowledge that functioning natural systems such as watersheds, mangroves, and wetlands are indispensable for reducing vulnerabilities to natural disasters and building long-term resilience.

In New Orleans for example, it has been argued that population growth and urban economic growth is necessary for meeting the costs of building a viable defense against the grave environmental problems of massive coastal erosion. In the New York Metropolitan region, sustainable management of the Catskills, the land around the upland water reservoirs supplying New York City with drinking water, has been chosen as an important complement to building water treatment plants.

The urban landscape provides a public space for the cross-fertilization of minds and various disciplines, enabling a new perspective on man in nature, one that could place human well-being at the core, break the artificial and largely culturally biased divide between the pristine and the human-dominated ecosystems, and contribute to the creation of a new language, with signs, concepts, words, tools, and institutions that would gather rather than divide, broker conflicts rather than create them, and establish responsible environmental stewardship at the heart of public interest.

See also: Ecological Complexity: Citizen Science. Ecological Data Analysis and Modelling: Spatial Models and Geographic Information Systems. Global Change Ecology: Urbanization as a Biospheric Process: Carbon, Nitrogen, and Energy Fluxes. Human Ecology and Sustainability: Industrial Ecology; Urban Metabolism. Terrestrial and Landscape Ecology: Landscape Planning; Anthropogenic Landscapes; Landscape Ecology

Further Reading

Adams, C.C., 1935. The relation of general ecology to human ecology. Ecology 16, 316–335.
Adams, C.E., Lindsey, K.J., Ash, S.J., 2006. Urban Wildlife Management. Boca Raton: CRC Press, Taylor and Francis.
Alfsen-Norodom, C., 2004. Urban biosphere and society: Partnership of cities. Annals of New York Academy of Sciences 1023, 1–9.
Blair, R.B., 1996. Land use and avian species diversity along an urban gradient. Ecological Applications 6 (2), 506–519.
Colding, J., Lundberg, J., Folke, C., 2006. Incorporating green-area user groups in urban ecosystem management. Ambio 35 (5), 237–244.
Collins, J.P., Kinzig, A., Grimm, N.B., *et al.*, 2000. A new urban ecology. American Scientist 88, 416–425.
Felson, A.J., Pickett, S.T.A., 2005. Designed experiments: New approaches to studying urban ecosystems. Frontiers in Ecology and the Environment 10, 549–556.
Kinzig, A.P., Warren, P., Martin, C., Hope, D., Katti, M., 2005. The effects of human socioeconomic status and cultural characteristics on urban patterns of biodiversity. Ecology and Society 10 (1), 23. http://www.ecologyandsociety.org/vol10/iss1/art23 (accessed December 2007).
McDonnell, M.J., Pickett, S.T.A., 1990. Ecosystem structure and function along urban–rural gradients: An unexploited opportunity for ecology. Ecology 71, 1232–1237.
McDonnell, M.J., Pickett, S.T.A., 1993. In: Humans as Components of Ecosystems: Subtle Human Effects and the Ecology of Populated Areas. New York: Springer, p. 363.
McGranahan, G., Marcotullio, P., Bai, X., *et al.*, 2005. Urban systems. ch. 27 In: Scholes,Ash, Ecosystems and Human Well-being: Current State and Trends. Washington, DC: Island Press, pp. 795–825. http://www.maweb.org/documents/document.296.aspx.pdf (accessed December 2007).
Millennium Ecosystem Assessment, 2005. Ecosystems and Human Well-being: Synthesis. Washington, DC: Island Press.
Pickett, S.T.A., Cadenasso, M.I., Grove, J.M., *et al.*, 2001. Urban ecological systems: Linking terrestrial ecological, physical and socioeconomic components of metropolitan areas. Annual Review of Ecology and Systematics 31, 127–157.
Sukopp, H., Numata, M., Huber, A., 1995. Urban Ecology as the Basis of Urban Planning. The Hague: SPB Academic Publishing.
Turner, W.R., Nakamura, T., Dinetti, M., 2004. Global urbanization and the separation of humans from nature. Bioscience 54, 585–590.

The Water-Energy-Food-Ecosystems (WEFE) Nexus

Giovanni Bidoglio, Davy Vanham, Faycal Bouraoui, and Stefano Barchiesi, European Commission, Joint Research Centre, 21027 Ispra (VA), Italy

Water in the Nexus

The last few years have witnessed an increasing attention to water beyond its environmental role and looked at water crises as a major global risk to social and economic stability. An example is the report on Global Risks released by the World Economic Forum (WEF, 2018) or the World Bank report "High and Dry" (World Bank, 2016), which predicts that some regions in the world could see their growth rates decline by as much as 6% of GDP by 2050 as a result of water-related losses. Indeed, demand for water from agriculture could increase by 50%, for urban uses by 50% to 70%. According to FAO (2018), feeding a global population expected to reach 9 billion people by 2050 will require a 60% increase in food production. Global energy consumption is projected to grow up to 50% by 2035 (IEA, 2010). However, constraints on water can challenge not only the physical, economic and environmental reliability of future projects, but also that of today's existing operations. The Californian "almond case" is a striking example of this challenge (Le Monde, 2015). In California, which is heavily affected by drought since a few years, the 900 tons of yearly production of almonds uses the same amount of water necessary to quench the thirst of the inhabitants of Los Angeles, San Diego and San Francisco, which, together, represent two thirds of the entire Californian population, while almond production represents just 2% of the State's GDP. In Poland, the summer of 2015 experienced a heatwave that drained the rivers supplying water to cool power plants resulting in problems of electricity supply (Reuters, 2016). In 2015, the European Parliament released a report on the implementation of current EU water legislation and confirmed that there are still implementation gaps with missing economic benefits estimated in the order of 2.8 billion euro per year (EPRS, 2015). One of the reasons is that the goals of water protection policies and related policies, particularly energy and agriculture, are occasionally incoherent or even in conflict.

Constraints to water supply/availability can then occur naturally, or be human-induced, as a result of growing competition among water using sectors. The examples above show that securing resilience of global energy and food systems needs to build on better and nontransient fairness of water allocation strategies across the energy and the food sectors, especially in light of the expected growing of pressures in the coming decades also associated to climate change. What we need is to overcome stakeholders' view of resources as individual assets by developing an understanding of the broader system. This points to the interdependency of policies and introduces the notion of Water-Energy-Food Nexus.

From Concept to Implementation

The Water-Energy-Food Nexus is a cross-sectoral perspective which requires that response options go beyond traditional sectoral approaches. It means that the three sectors or securities—water security, energy security and food security (Sustainable Development Goals or SDGs 6, 7 and 2)—are inextricably linked and that actions in one area have impacts in one or both of the others. However, a fourth leg that we should consider in the Nexus is that of ecosystems, as they are central in providing and sustaining these three securities, through the services they deliver to the human being and society. Exploitation of water, agricultural and energy resources should not undermine the provision of these services, especially considering that they are often at the basis of the only resources available to poor people in different parts of the world. Fig. 1 shows these interlinkages. At the same time, some of the impacted ecosystem services may support combinations of both built and natural infrastructure that in turn provide water services or can be considered solutions to water resources management challenges from sedimentation in reservoirs to water purification (Baker et al., 2015; UNEP, 2014). The Nexus is about understanding and managing often-competing interests, while ensuring the integrity of ecosystems.

According to FAO (2002), food security is defined as "availability and access to sufficient, safe and nutritious (nutrition security) food to meet the dietary needs and food preferences for an active and healthy life." Food is defined as a human right. The end products for food security encompass edible agricultural crops, livestock products (meat, dairy and eggs), freshwater fish (wild and aquaculture), wild foods (e.g., berries, mushrooms, fruits, nuts, game and bushmeat, insects), but also marine fish and seafood. The International Energy Agency (IEA) defines energy security as "the uninterrupted availability of energy sources at an affordable price." The UN defines it as "the access to clean, reliable and affordable energy services for cooking and heating, lighting, communications

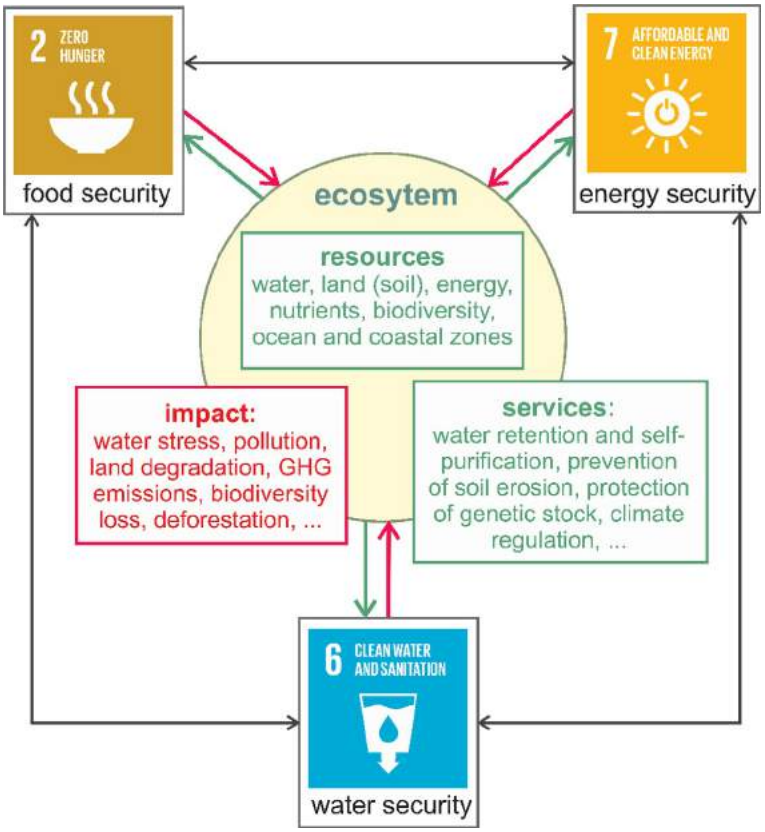


Fig. 1 Schematic representation of the WEFE Nexus.

and productive uses (end products)." Primary energy is generated by means of fossil energy, nuclear energy and renewable energy. Water security is defined as "access to safe drinking water and sanitation," both of which have recently become a human right (UN, 2010). One of the most frequently resources required for realizing the nexus, that is, minimizing tradeoffs and maximizing synergies across resource securities, is water. Fig. 2 shows the links between available water resources (green and blue) and food security, energy security and water security. Interlinkages are many and consequently possible tradeoffs too. The nexus between energy and food security is, for example, very visible in the amount of energy required for food security, which is about one third (32%) of global energy end consumption (from farm to fork) (FAO, 2011). The primary production of crops, livestock and fisheries accounts for 6.6% of global energy end consumption.

Generally, national governments and international organizations have separate departments/ministries dealing with water, energy, and agriculture. They often define and implement policies for each sector separately (SEI, 2014). The same is true for research on these issues: expertise on energy, water and food systems is clustered in separate groups, with often limited interaction (SEI, 2014). It is on this basis that the cross-institutional planning of water resources referred to as integrated river basin management is often conducted. The Nexus approach recognizes that water, energy, food and ecosystems are closely linked, through global, basin and local resource use and impact cycles. In developing an effectively integrated nexus approach it is important to recognize that many activities take place at the very bottom of the grassroots, at the level of, for example, households, farmers. An array of options and measures adapted to the specificity of local conditions need then to be developed.

Terminology in the Nexus Communication

The Nexus is often accounted for in different terminologies or even defined in different ways, but addressing the same concept. Different authors use the terminology FEW (Food-Energy-Water) Nexus, especially in the United States (Chini et al., 2017). Others mix the components in a different order, for example, the EFW Nexus (Owen et al., 2018; Liu et al., 2018), as often referred to by energy experts, putting the E first. Other authors define the Nexus in a variety of combinations, like the Land-Water-Energy Nexus (Silalertruksa and Gheewala, 2018), the Water-Land-Food Nexus (Rulli et al., 2016), the Land-Water-Energy-Food Nexus (Siciliano et al., 2017), the Water-Land-Energy-Food-Climate (WLEFC) Nexus (Munaretto and Witmer, 2017; Sušnik et al., 2018) or many others. These terminologies try to describe the complexity of the Nexus, however they often differ from the original definition (Hoff, 2011) by including elements like land or climate that are not securities, but resources or impacts.

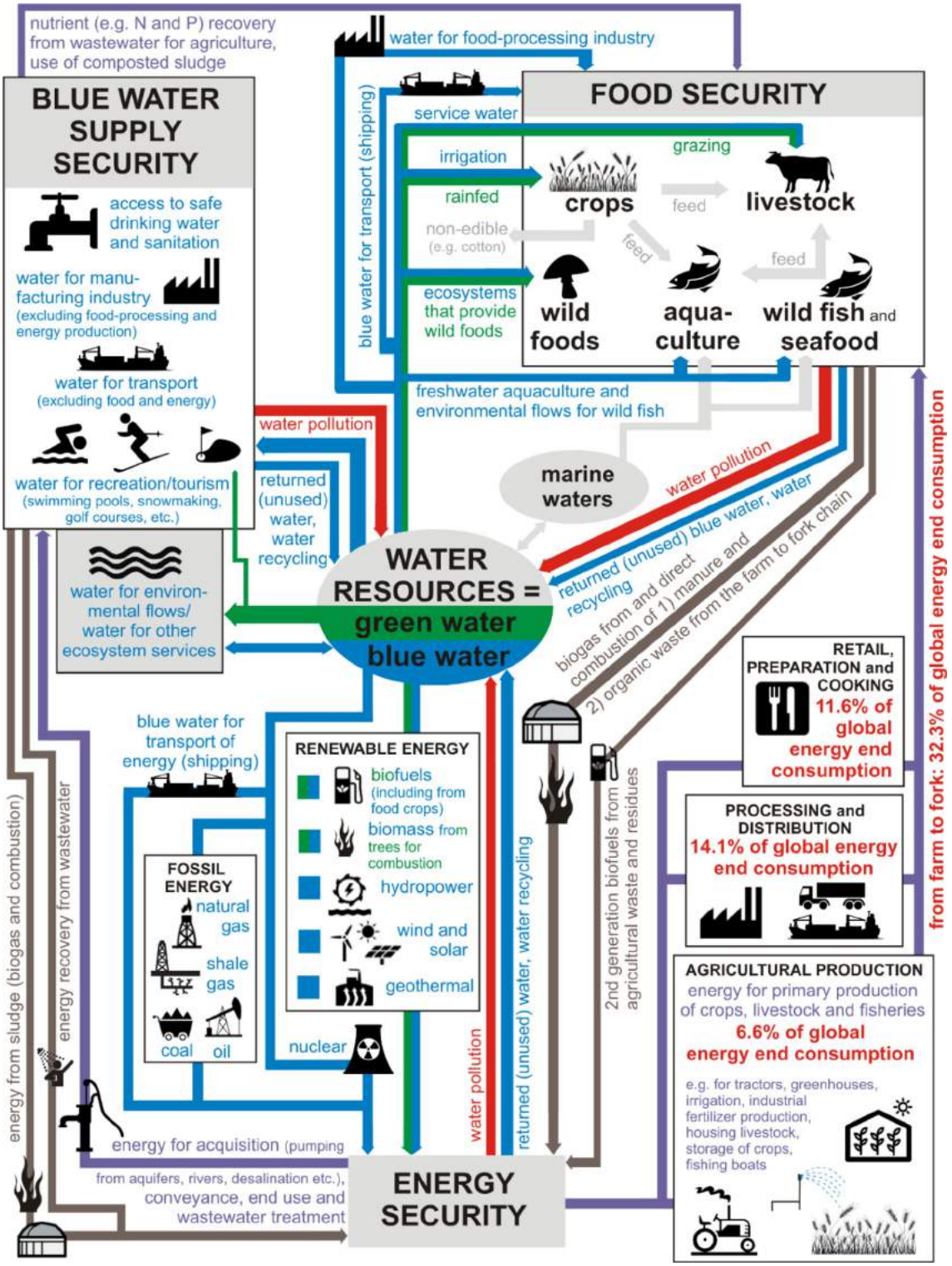


Fig. 2 The WEFE Nexus for blue and green water, as presented in Vanham (2016).

Exploring WEFE Interdependencies

Water, energy and food resources are linked through many interactions: we need water for power generation (hydropower, cooling, biofuels, etc.), while at the same time large amounts of energy are required to pump, treat and desalinate water, and both water and energy use are required for food production and distribution. Using water to irrigate crops can increase food production, but also

reduce river flows and hydropower potential as well as affect water quality and ecosystem health and services. Growing bioenergy crops under irrigated agriculture can increase overall water withdrawals and also affect food security. Converting surface irrigation into high efficiency pressurized irrigation may save water, but also result in higher energy use.

The modernization of the irrigation sector that took place in Spain in the years 1950–2007 offers a paradoxical example of the energy-water interdependency (Tarjuelo et al., 2015; González-Cebollada, 2015). During this period, water consumption by agriculture has quadrupled, due to a shift of cropping patterns with greater water needs, also influenced by subsidies. Efforts mostly focused on irrigation efficiency (from 67% to 82.5%) resulting in a 19-fold increase in energy consumption, which has led to an increase in water costs. Irrigation systems modernized in a climate of strong public support without giving much thought to the significant energy dependence they would introduce may then represent a barrier to economic sustainability for certain crops and irrigation regions.

Hydro-economic interdependencies exist also between distant global geographies via trade, especially of agricultural and manufactured goods (Vanham and Bidoglio, 2014). This makes local consumption dependent on foreign water resources and increases virtual water exports from other regions.

The Water-Energy-Food-Ecosystems (WEFE) Nexus highlights the need of a joint management of bundles of resources, rather than of individual resources, if we want to achieve sustainable development (Biggs et al., 2015). The Sustainable Development Goals (*SDGs*) of the 2030 Agenda set by the United Nations promote a crosscutting understanding of the interdependencies of the goals and targets as opposed to their individual consideration. It follows that the implementation of the Nexus helps meet the SDG commitments. Indeed, accounting for tradeoffs and synergies of the WEFE Nexus contributes to the achievement of a number of interdependent SDGs. The WEFE Nexus directly addresses a number of SDGs: SDG 6 (clean water), SDG 7 (clean energy) and SDG 2 (zero hunger), while the last E of WEFE refers to other SDGs like 14 (life below water) and 15 (life on land). Resilient communities better prepared to react and recover from water resources crises are those that have identified crucial interdependencies and built relationships between the water-using sectors.

Ecosystems as Fourth Pillar of the Nexus

Aquatic ecosystems offer a wide array of benefits to human society, such as water retention and self-purification, the prevention of soil erosion, the protection of genetic stock. Preserving these services guarantees the maintenance and improvement of water quality and quantity, increases the resilience of ecosystems to natural and man-made alterations making floods less severe, food and energy production more reliable (Russi et al., 2013; UNWWAP, 2018). Ecosystems have then to be considered as water using sectors on an equal ground as food and energy production. The practical instrument for achieving these interdependent goals is the deployment of green–blue infrastructure, or "natural infrastructure." This refers to the network of green (land) and blue (water) spaces that provide services to people, can help decision makers and infrastructure managers address interconnected challenges facing water, energy and food systems. The value of ecosystems is explicited in the EU policy-making process (e.g., Water Framework Directive, Common Agricultural Policy, EU Biodiversity Strategy) by promoting nature-based solutions (NBS). The EU has promoted a Green Infrastructure Strategy to become an integral part of spatial planning and territorial development whenever it offers a better alternative, or is complementary, to standard gray choices (EC, 2013). This is reflected in certain research projects on NBS in the urban environment (EC, 2015; GrowGreen, 2018), another key scale for Nexus implementation considering that cities are where global and local resource constraints meet. The potential of nature-based solutions to address water management across the agricultural sector, sustainable cities, disaster risk reduction and improving water quality is also addressed by the 2018 UN World Water Development Report (UNWWAP, 2018).

What makes natural infrastructure particularly attractive is its multifunctionality, that is its ability to perform several functions and provide several benefits on the same spatial area. The challenge is to build on the knowledge gained from the traditional and local sustainable water management practices, link the green–blue infrastructure or "natural capital" part to ecosystem services and their role in river basin management plans, including investment in these natural solutions instead of purely technical ones. While in some cases planners may directly compare the advantages of "green versus gray" water infrastructure solutions, greater emphasis should be placed on understanding how green solutions can be integrated within an overall system of water management, composed of appropriately sited and designed elements of both green and gray water infrastructure (UNEP, 2014). Any methodology for developing combined portfolios consisting of green and gray alternatives, or mutually supportive green and gray elements, should therefore provide meaningful evaluation of all water infrastructure options, including in monetary terms to the extent possible. This then leads to the development of adequately informed strategies for investing in natural infrastructure. For these to be incorporated into broader infrastructure packages, appropriate mechanisms for investment are however needed that can unlock financing for ecosystem management that also supports empowerment of stakeholders to participate in and enable implementation of these options (Bennett et al,. 2016). Forms of investment in natural infrastructure are sustainable dam management for the Nexus, certifiable standards for watershed stewardship, and public–private partnerships for Payments for Ecosystem Services (PES). These are established instruments, not without hurdles, to direct private and public investments to sustain provisioning services in ecosystems and watersheds (Bennett and Carroll, 2014). As an example, a water conservation project along the Cauca River near the city of Cali in Colombia pooled money from downstream water users to pay upstream stakeholders who have the ability to impact water quantity and quality and to implement projects and practices addressing the interdependent, downstream's needs for food, energy and water (Madre Agua Water Fund, 2018). Ozment et al. (2015) examine reasons and ways to include natural infrastructure

in the Nexus along with challenges that have prevented increased investments and recommendations for moving forward. For example, they present recent studies that estimate that the global community invests about 12.3 billion $ per year to protect, manage and restore natural infrastructure to secure water resources. Yet, the energy and agriculture sectors collectively contributed less than 1% of all natural infrastructure investments in 2013. Partnerships are still needed that proactively identify more opportunities to invest in green–blue infrastructure, leverage new sources of financing, and reform policy and standards to broaden investments.

In Search of Appropriate WEFE Nexus Indicators

By their nature, water, energy, food and ecosystems require integrated and transdisciplinary approaches for addressing the peculiarity of their interlinked temporal and spatial variabilities. However, integration goes beyond these pillars and should include social, political and governance aspects. In addition, the Nexus approach should consider tradeoffs, not only across sectors, but also among different users of the same sectors (Sønderberg Petersen and Larsen, 2016). Moreover, solutions valid in a place do not necessarily apply under different bio-geographical and socio-economic conditions (Liu et al., 2017). In this context, data and appropriate methodologies are needed to inform on these complex interlinkages, in the present situation and also under different future scenarios, to help policy makers in decision making. Most of the existing WEFE Nexus analyses have been performed at large scale ranging from global, to regional and national (Martinez-Hernandez et al., 2017), however at very coarse resolution. As international and regional decisions have significant impact at the local level, the local specificities need to be considered when designing efficient management strategies (Aarnoudse and Leentvaar, 2015). However, the local scale is rarely addressed even though it is the relevant scale at which policy implementation is usually taking place and where synergies between the food, water and energy can be enhanced (Martinez-Hernandez et al., 2017). As stressed by Mohtar (2016), it is at the local scale that also the SDG targets deliver their benefits. Different assessment levels are then needed to address the WEFE Nexus.

Addressing the WEFE Nexus or the achievement of the associated SDGs requires a holistic approach and a systems view (Bazilian et al., 2011; UNECE, 2015; de Strasser et al., 2016). Endo et al. (2017) classified the tools used in Nexus studies in two broad categories including qualitative and quantitative approaches. Qualitative approaches are used to identify the priorities and understand the types of environmental, economic and societal tradeoffs. Quantitative methods are used to assess baselines and carry out scenario analyses for the evaluation of the impact of alternative proposed interventions. Examples of how an assessment of the four pillar Nexus can be operationalized are offered by Tesfaye et al. (2016) and Karabulut et al. (2016). These authors investigated spatially explicit tradeoffs and co-benefits for two large transboundary river basins, the Blue Nile basin in Ethiopia and the Danube River basin in Europe. A wide range of the facets of the water-food nexus have been analyzed in a series of multi-disciplinary papers edited by Laio et al. (2017). A coupled simulation and optimization tool with which stakeholders can define their own objectives was proposed by Karnib (2017). A systematic review of Nexus-specific methods by Albrecht et al. (2018) shows that mixed-methods and interdisciplinary approaches are needed that incorporate social and political dimensions of water, energy and food. Dai et al. (2018) found that fewer approaches are designed to support governance and implementation of water-energy nexus technical solutions. In general, most of the published Nexus studies are considering two sectors, rarely the three components of the WEF, leaving repeatedly ecosystems outside the analysis (Martiner-Hernadez et al., 2017). Moreover, our understanding and representation of the interactions and tradeoffs are often limited by data availability, collection and management thus failing to encompass a comprehensive overview of the interlinkages between the pillars of the Nexus.

Nexus Solutions for Sustainable Development

Already today different control variables within the planetary boundary framework show human-made disruptions (Steffen et al., 2015). Climate change and global socio-economic developments like population increase, rapid urbanization, changing diets and economic growth are among drivers that will increase the demands for energy, food and water. Can we afford continuing in this way? Certainly not if we consider that about 78% of the world's total active workforce is water-dependent globally (WWAP, 2016).

These considerations place the Water-Energy-Food-Ecosystems Nexus at the core of resilience strategies (Smith, 2016). To implement the Nexus we need actions for the development of a multilevel governance, the investment in human and social infrastructure, the enrollment of technologies across water-using sectors, and the leveraging of public–private cooperation. Governance deals with the ability of institutions to manage water use through different sectors and propose policies that are socially acceptable and can address competition through an analysis of tradeoffs. To this end, rules and regulations are necessary, but not sufficient if not integrated with constructive engagement of social actors, as good WEFE Nexus practices that work in a place do not necessarily apply to different social, economic and geographical conditions. The organizational model must accommodate creation and diffusion of technology tools that span multiple applications and innovate solutions across water-using sectors. This will require creating incentives for public–private partnerships to cultivate the innovative operational models, policies, strategies and technologies.

The circular economy dynamics based on reuse and recycling offers the framework for conserving and saving water, energy and materials by ensuring the balance between supply and demand from the different sectors. If supply decreases, we have to find ways

of reducing demand or decrease losses. Opportunities exist especially in the field of efficient irrigation techniques, sustainable water pumping, crop refinement techniques, water efficient appliances and processes, technologies for aquifer storage as well as water treatment and reuse. The IGN Group and the Ellen McArthur Foundation released a report addressing the concept of circular economy in relation to the water sector (IGN, 2017). For the investigated countries, circular water economy has the potential to save 412 billion m^3 of water a year, which is equivalent to 11% of annual global water demand, or almost the entire water consumption in the United States. Of course, this does not equally applies everywhere. The concept of circular economy is very much different from regions to regions in the world. The report shows that effects are different, for example, from California to the Emirates. In addition, circular economy measures can look appealing when viewed in isolation. However, circular economy requires a system approach to be developed not only for water, but also for its nexus with energy and materials.

Policy, science and technology can bring about change, but this change takes place in a specific local ecological, economic and cultural context. In order to provide water, food and energy security to a growing and urbanizing global population, within global planetary boundaries with limited resources availability (Rockstrom et al., 2009; Steffen et al., 2015), solutions need to come from all stages within supply chains (Fig. 3), that is, from the production to the consumption side (Godfray et al., 2010; Foley et al., 2011).

Production side solutions may include (Godfray et al., 2010; Foley et al., 2011):

- The sustainable intensification of agriculture (Garnett et al., 2013). In order to close yield gaps, nutrient management and integrated land and water management on existing agricultural lands (rainfed and irrigated) are key to this development (Mueller et al., 2012). Often the increase in water productivity is referred to as "more crop per drop." In the wider context of the Nexus, the expression "more biomass per drop" (apart from crops also fish, livestock, fiber, tree biomass ...) is more appropriate.
- Production processes along the supply chain need to become more resource-efficient, energy-efficient, circular and sustainable. For instance, installation of new hydropower plants in existing water infrastructures is an attractive Nexus solution in that it can save construction costs and minimize environmental and social impacts. Moreover, it may contribute to the reduction of the global footprint of these infrastructures. Examples of how hydropower can be incorporated in existing hydraulic infrastructures where electricity generation was not a primary objective are municipal and agricultural water systems, for example, for wastewater treatment and irrigation canals, hydraulic circulation systems for cooling and heating of plants, and hydropower dams themselves where there are ship navigation locks or fish passages that need dissipation of energy (Marence et al., 2018). Existing infrastructure in the water sector has a high potential for additional renewable energy production from solar as well, for example, the installation of photovoltaic (PV) systems on the face of existing dams and the coverage of irrigation canals with solar PV systems (Szabó et al., 2018).

Demand side solutions may include:

- Consumer behavior in water use, energy consumption, food consumption and (food) waste generation. In the EU, for instance, citizens can decrease their current food-related water footprint by 24% when shifting to a healthy diet and by 40% when shifting to a vegetarian diet (Vanham et al., 2013). EU consumers waste on average 123 kg of food per person annually, of which 80% is avoidable (Vanham et al., 2015). A reduction in food waste as envisaged by SDG target 12.3 would have an influential impact on the Nexus.
- Some scholars also state that family planning to lower per capita fertility should be part of the solution (Potts, 2014; Speidel et al., 2009). Especially in Africa a population explosion is projected for the next few decades (UN, 2014). There are consequently more frequent calls to address environmental problems by advocating further reductions in human fertility (Bradshaw and Brook, 2014), including in the second notice of "World Scientists' Warning to Humanity" (Ripple et al., 2017).

It is clear that the provision of water, energy and food security for people within a geographical region relates to the consumption side of this geographical region and not the production side. Trade between regions is thus essential to provide these three securities. Sustainability does not mean self-sufficiency.

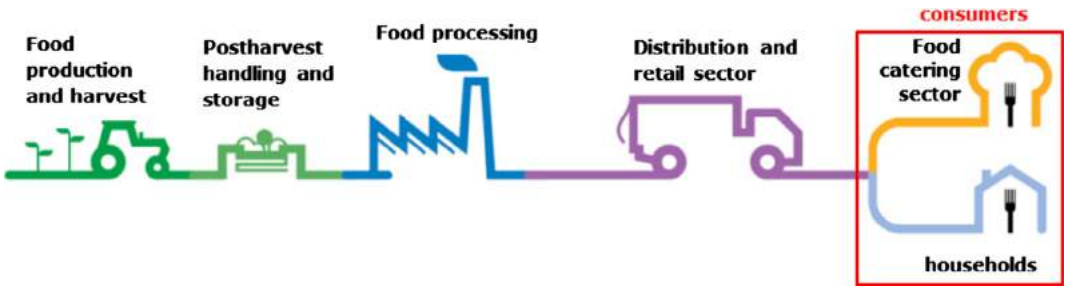


Fig. 3 Graphical representation of the food supply chain. Solutions to address the Nexus need to come from all stages within supply chains. Different processes can take place in different geographical regions.

References

Aarnoudse E and Leentvaar J (2015) Implementing the nexus at various scales: Local and regional perspectives. *Change and Adaptation in Socio-Ecological Systems* 2(1): 97–99.

Albrecht TR, Crootof A, and Scott CA (2018) The water-energy-food nexus: A comprehensive review of nexus-specific methods. *Environmental Research Letters* 13(4): 043002. https://doi.org/10.1088/1748-9326/aaa9c6.

Baker T, Kiptala J, Olaka L, Oates N, Hussain A, and McCartney M (2015) *Baseline review and ecosystem services assessment of the Tana River Basin, Kenya*. Colombo, Sri Lanka: International Water Management Institute (IWMI) p.107. (IWMI Working Paper 165). https://doi.org/10.5337/2015.223. http://www.iwmi.cgiar.org/Publications/Working_Papers/working/wor165.pdf.

Bazilian M, Rogner H, Howells M, Hermann S, Arent D, Gielen D, Steduto P, Mueller A, Komor P, Tol RSJ, and Yumkella KK (2011) Considering the energy, water and food nexus: Towards an integrated modelling approach. *Energy Policy* 39: 7896–7906.

Bennett G and Carroll N (2014) Gaining depth: State of watershed investment 2014. Available online at www.ecosystemmarketplace.com/reports/sowi2014.

Bennett G, Cassin J, and Carroll N (2016) Natural infrastructure investment and implications for the nexus: A global overview. *Ecosystem Services* 17: 293–297.

Biggs EM, Bruce E, Boruff B, Duncan JMA, Horsley J, Pauli N, McNeill K, Neef A, Van Ogtrop F, Curnow J, Haworth B, Duce S, and Imanari Y (2015) Sustainable development and the water–energy–food nexus: A perspective on livelihoods. *Environmental Science & Policy* 54: 389–397.

Bradshaw CJA and Brook BW (2014) Human population reduction is not a quick fix for environmental problems. *Proceedings of the National Academy of Sciences* 111: 16610–16615.

Chini CM, Konar M, and Stillwell AS (2017) Direct and indirect urban water footprints of the United States. *Water Resources Research* 53: 316–327.

Dai J, Wu S, Han G, Weinberg J, Xie X, Wu X, Song X, Jia B, Xue W, and Yang Q (2018) Water-energy nexus: A review of methods and tools for macro-assessment. *Applied Energy* 210: 393–408.

EC (2013) Green infrastructure (GI)—Enhancing Europe's natural capital. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. In: *COM(2013) 249 final*.

EC 2015. Towards an EU research and innovation policy agenda for nature-based solutions & re-naturing cities. European Commission Report, https://doi.org/10.2777/765301

Endo A, Tsurita I, Burnett K, and Orencio PM (2017) A review of the current state of research on the water, energy, and food nexus. *Journal of Hydrology: Regional Studies* 11: 20–30.

EPRS 2015. Water legislation. Cost of non-Europe Report, EPRS-European Parliamentary Research Service, European Parliament.

FAO 2002. The state of food insecurity in the world 2001. Rome.

FAO 2011. Energy-smart food for people and climate—Issue paper. Rome.

FAO (2018) http://www.fao.org/land-water/water/watergovernance/waterfoodenergynexus/en/ (Online).

Foley JA, Ramankutty N, Brauman KA, Cassidy ES, Gerber JS, Johnston M, Mueller ND, O'connell C, Ray DK, West PC, Balzer C, Bennett EM, Carpenter SR, Hill J, Monfreda C, Polasky S, Rockstrom J, Sheehan J, Siebert S, Tilman D, and Zaks DPM (2011) Solutions for a cultivated planet. *Nature* 478: 337–342.

Garnett T, Appleby MC, Balmford A, Bateman IJ, Benton TG, Bloomer P, Burlingame B, Dawkins M, Dolan L, Fraser D, Herrero M, Hoffmann I, Smith P, Thornton PK, Toulmin C, Vermeulen SJ, and Godfray HCJ (2013) Sustainable intensification in agriculture: Premises and policies. *Science* 341: 33–34.

Godfray HCJ, Beddington JR, Crute IR, Haddad L, Lawrence D, Muir JF, Pretty J, Robinson S, Thomas SM, and Toulmin C (2010) Food security: The challenge of feeding 9 billion people. *Science* 327: 812–818.

González-Cebollada C (2015) Water and energy consumption after the modernization of irrigation in Spain. *WIT Transactions on the Built Environment* 168: 457–465.

GrowGreen 2018. A partnership for greener cities to increase liveability, sustainability and business opportunities, http://growgreenproject.eu/.

Hoff H (2011) Understanding the Nexus. In: *Background Paper for the Bonn2011 Conference: The water, energy and food security Nexus*. Stockholm: Stockholm Environment Institute.

IEA (2010) *World energy outlook 2010*. Paris: OECD/International Energy Agency.

IGN (2017) *Less is more: Circular economy solutions to water shortages*. ING Economics Department.

Karabulut A, Egoh BN, Lanzanova D, Grizzetti B, Bidoglio G, Pagliero L, Bouraoui F, Aloe A, Reynaud A, Maes J, Vandecasteele I, and Mubareka S (2016) Mapping water provisioning services to support the ecosystem–water–food–energy nexus in the Danube river basin. *Ecosystem Services* 17: 278–292.

Karnib A (2017) Water-energy-food Nexus: A coupled simulation and optimization framework. *Journal of Geoscience and Environment Protection* 5: 15.

Laio F, Rulli MC, and Suweis S (2017) The challenge of understanding the water-food nexus complexity. *Advances in Water Resources* 110: 406–407.

Le Monde. 2015. L'amande, suspect idéal de la sécheresse californienne, http://www.lemonde.fr/ameriques/article/2015/06/01/l-amande-suspect-ideal-de-la-secheresse-californienne_4644791_3222.html [Online].

Liu J, Mao G, Hoekstra AY, Wang H, Wang J, Zheng C, Van Vliet MTH, Wu M, Ruddell B, and Yan J (2018) Managing the energy-water-food nexus for sustainable development. *Applied Energy* 210: 377–381.

Liu J, Yang H, Cudennec C, Gain AK, Hoff H, Lawford R, Qi J, de Strasser L, Yillia PT, and Zheng C (2017) Challenges in operationalizing the water – energy – food nexus. *Hydrological Sciences Journal* 62(11): 1714–1720.

Madre Agua Water Fund. 2018. http://waterfunds.org/esp/madre-agua-water-fund-cali/ [Online].

Marence M, Lemessa TS, and Franca MJ (2018) Towards the circularization of the energy cycle by implementation of hydroelectricity production in existing hydraulic systems. *EC Position paper on WEFE Nexus Dialogue and SDGs*.

Martinez-Hernandez E, Leach M, and Yang A (2017) Understanding water-energy-food and ecosystem interactions using the nexus simulation tool NexSym. *Applied Energy* 206: 1009–1021.

Mohtar R (2016) *The importance of the Water-Energy-Food Nexus in the implementation of the Sustainable Development Goals (SDGs)*. OCP policy brief, PB16/30, Rabat, Morocco.

Mueller ND, Gerber JS, Johnston M, Ray DK, Ramankutty N, and Foley JA (2012) Closing yield gaps through nutrient and water management. *Nature* 490: 254–257.

Munaretto S and Witmer M (2017) *Water-land-energy-food-climate: Policies and policy coherence at European and international scale*. Netherlands Environmental Assessment Agency: PBL.

Owen A, Scott K, and Barrett J (2018) Identifying critical supply chains and final products: An input-output approach to exploring the energy-water-food nexus. *Applied Energy* 210: 632–642.

Ozment S, Di Francesco K, and Gartner T (2015) The role of natural infrastructure in the water, energy and food nexus. *Nexus dialogue synthesis papers, gland*. Switzerland: IUCN.

Potts M (2014) Getting family planning and population back on track. *Global Health: Science and Practice* 2: 145–151.

Reuters. 2016. Polish power demand hits summer record https://af.reuters.com/article/commoditiesNews/idAFL8N19G32B [Online].

Ripple WJ, Wolf C, Newsome TM, Galetti M, Alamgir M, Crist E, Mahmoud MI, and Laurance WF (2017) World scientists' warning to humanity: A second notice. *Bioscience* 67: 1026–1028.

Rockstrom J, Steffen W, Noone K, Persson A, Chapin FS, Lambin EF, Lenton TM, Scheffer M, Folke C, Schellnhuber HJ, Nykvist B, de Wit CA, Hughes T, van der Leeuw S, Rodhe H, Sorlin S, Snyder PK, Costanza R, Svedin U, Falkenmark M, Karlberg L, Corell RW, Fabry VJ, Hansen J, Walker B, Liverman D, Richardson K, Crutzen P, and Foley JA (2009) A safe operating space for humanity. *Nature* 461: 472–475.

Rulli MC, Bellomi D, Cazzoli A, de Carolis G, and D'odorico P (2016) The water-land-food nexus of first-generation biofuels. *Scientific Reports* 6: 22521.

Russi D, Ten Brink P, Farmer A, Badura T, Coates D, Förster J, Kumar R, and Davidson N (2013) The economics of ecosystems and biodiversity for water and wetlands. In: *IEEP*. London and Brussels: Ramsar Secretariat, Gland.

SEI (2014) *Managing environmental systems: The water-energy-food nexus. Research synthesis briefs*. Stockolm Environmental Institute.

Siciliano G, Rulli MC, and D'odorico P (2017) European large-scale farmland investments and the land-water-energy-food nexus. *Advances in Water Resources* 110: 579–590.

Silalertruksa T and Gheewala SH (2018) Land-water-energy nexus of sugarcane production in Thailand. *Journal of Cleaner Production* 182: 521–528.
Smith M (2016) *Collaboration for resilience: How collaboration among business, government and NGOs could be the key to living with turbulence and change in the 21st century*. Gland, Switzerland: IUCN p.16. https://portals.iucn.org/library/sites/library/files/documents/2016-047.pdf.
Sønderberg Petersen L and Larsen HH (eds.) (2016) *DTU International Energy report 2016: The Energy-Water-Food Nexus – from local to global aspects*. Technical University of Denmark (DTU).
Speidel JJ, Weiss DC, Ethelston SA, and Gilbert SM (2009) Population policies, programmes and the environment. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364: 3049–3065.
Steffen W, Richardson K, Rockström J, Cornell SE, Fetzer I, Bennett EM, Biggs R, Carpenter SR, de Vries W, de Wit CA, Folke C, Gerten D, Heinke J, Mace GM, Persson LM, Ramanathan V, Reyers B, and Sörlin S (2015) Planetary boundaries: Guiding human development on a changing planet. *Science* 347.
de Strasser L, Lipponen A, Howells M, Stec S, and BrÉthaut C (2016) A methodology to assess the Water Energy Food Ecosystems nexus in transboundary river basins. *Water* 8: 59.
Sušnik J, Chew C, Domingo X, Mereu S, Trabucco A, Evans B, Vamvakeridou-Lyroudia L, Savić DA, Laspidou C, and Brouwer F (2018) Multi-stakeholder development of a serious game to explore the water-energy-food-land-climate nexus: The SIM4NEXUS approach. *Water* 10: 139. https://doi.org/10.3390/w10020139.
Szabó, S., Kougias, I., Bódis, K., Moner-Girona, M. & Jäger-Waldau, A (2018) Integrating existing west African infrastructures into the Water Energy Food Ecosystem nexus approach. In: *EC Position paper on WEFE Nexus Dialogue and SDGs*, in press.
Tarjuelo JM, Rodriguez-Diaz JA, Abadía R, Camacho E, Rocamora C, and Moreno MA (2015) Efficient water and energy use in irrigation modernization: Lessons from Spanish case studies. *Agricultural Water Management* 162: 67–77.
Tesfaye A, Wolanios N, and Brouwer R (2016) Estimation of the economic value of the ecosystem services provided by the Blue Nile Basin in Ethiopia. *Ecosystem Services* 17: 268–277.
UN 2010. Resolution A/RES/64/292. United Nations General Assembly, July 2010.
UN 2014. World urbanization prospects: The 2014 revision, Highlights.
UNECE 2015. Reconciling resource uses in transboundary basins: Assessment of the water-food-energy-ecosystems nexus, United Nations Economic Commission for Europe Report, New York and Geneva.
UNEP (2014) *Green Infrastructure Guide for Water Management: Ecosystem-based management approaches for water-related infrastructure projects*.
UNWWAP (2018) United Nations World Water Assessment Programme/UN-Water. *The United Nations World Water Development Report 2018: Nature-Based Solutions for Water*. Paris: UNESCO.
Vanham D (2016) Does the water footprint concept provide relevant information to address the water–food–energy–ecosystem nexus? *Ecosystem Services* 17: 298–307.
Vanham D and Bidoglio G (2014) The water footprint of agricultural products in European river basins. *Environmental Research Letters* 9: 064007.
Vanham D, Mekonnen MM, and Hoekstra AY (2013) The water footprint of the EU for different diets. *Ecological Indicators* 32: 1–8.
Vanham D, Bouraoui F, Leip A, Grizzetti B, and Bidoglio G (2015) Lost water and nitrogen resources due to EU consumer food waste. *Environmental Research Letters* 10: 084008.
WEF (2018) *The global risks report 2018*, 13th Edition. Geneva: World Economic Forum.
World Bank (2016) *High and dry: Climate change, water, and the economy*. Washington, DC: World Bank.
WWAP (2016) *The United Nations world water development report 2016: Water and jobs*. Paris: United Nations World Water Assessment Programme.

Agroforestry☆

Steven M Newman, BioDiversity International Ltd, Faversham, Kent, United Kingdom

Glossary

Agroforestry A collective name for land use systems and technologies where woody perennials (trees, shrubs, palms, bamboos, etc.) are deliberately used on the same land management unit as agricultural crops and/or animals, either in some form of spatial arrangement or temporal sequence. In agroforestry systems, there are both ecological and economic interactions between the different components.
Forest farming It refers to growing crops or raising livestock in a natural or created forest.
Forest gardens It presents an attempt to grow mainly perennial crops and tree in a diverse multilayered garden.
Intercropping The simultaneous growing of more than one crop species on the same land area within a year, for example, the intercropping of beans in a field of maize.
Interculture Refers to the intercropping in plantations of tree crops in both temperate and tropical environments.
Permaculture A concatenation of the words permanent and agriculture and refers to the conscious design of diverse perennial based cropping systems so as to reduce agrochemical inputs (mainly artificial fertilizers and pesticides) and make the best use of local landscape features, for example, to allow rainwater harvesting.
Riparian strips Strips of trees or crops planted in linear arrangements next to water courses with a view to producing environmental benefits, for example, reducing nitrate pollution from arable crop fields.
Silvopastoral systems Are agroforestry systems where livestock are a major output. They can include forest grazing and grassed orchard systems.
Silvoarable systems Are agroforestry systems where arable crops are grown with trees.
Swidden Refers to shifting cultivation and is where woody vegetation is partially cleared for a short period to allow for the growing of crops. When crops begin to exhaust the soil, the plot is allowed to return to fallow for a period of time. This fallow can then be partially cleared to start the cycle again. Swidden does not have to involve the use of fire as in the case of "slash and burn" agriculture.

Introduction

One of the definitions for agroforestry developed by the International Centre for Agroforestry (now World Agroforestry Centre) was:

> a collective name for land use systems and technologies where woody perennials (trees, shrubs, palms, bamboos, etc.) are deliberately used on the same land management unit as agricultural crops and/or animals, either in some form of spatial arrangement or temporal sequence. In agroforestry systems, there are both ecological and economic interactions between the different components.

This definition is useful in that it shows that woody perennials (which include fruit trees) are key and not just forest trees. It shows that there must be a mixture of animals/crops and woody perennials so shade trees used in coffee or tea production is not

☆*Change History:* February 2018. Steven M. Newman has updated the text throughout the article.

This is an update of P.K.R. Nair, A.M. Gordon, and M. Rosa Mosquera-Losada, Agroforestry, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, Pages 101–110.

Table 1 Common reasons for deliberately designing agroforestry systems as an alternative to monocultures

Type of reason	*Indicator*
Environmental	Soil organic matter
	Reduced nitrate leaching
	Greater species diversity of wildlife
Agronomic	Greater yield per unit area
	Greater yield per unit labor
Economic	Greater return on investment
	Increased asset value if plot is sold
	Lower insurance costs

agroforestry. The phrase "deliberately used" is also important as this excludes relic systems such as most hedgerows around crop fields in the United Kingdom where the manager is not deliberately using the hedges as happened in the past. The economical and ecological interactions are the main challenge for agroforesters as the goal is to improve profit or other measures of effectiveness and knowledge of these is central to optimization. Table 1 below gives a sample of possible reasons why a grower may deliberately opt for agroforestry compared to monocultures.

The only problem with the definition is the term agricultural crops as agroforestry now encompasses all kinds of crops such as medicinal, horticultural and energy crops.

Scale

Agroforestry can be designed and carried out at a wide range of scales from a small oasis containing a single tree to a landscape scale covering a watershed, district or country. For the manager the boundary of the scale could be set by his/her control or the effects of the trees. It is known for instance that the effect of a windbreak may extend beyond the crop field that it is meant to protect. A rule of thumb is that the distance is three to five times the height of the tree. Recent research on volatile organic compounds arising from trees means (Slowik et al., 2010) that effect boundaries could be far larger than this. In addition it is now known that trees can affect rainfall patterns (Ong et al., 2015). The limit of both these effects is still the subject of research. Some of the largest agroforestry investments have been attempts to control desertification or reduce sand dune migration. In India and China these projects have covered hundreds of square kilometers.

History

The start of agroforestry can be hypothesized to have occurred in prehistoric times when a hunter gatherer realized that by planting browse resistant tree seeds in wild grasslands; more wild animals could be attracted. In the forest situation the same hunter gatherer may notice that if a tree is felled then other nutritious herbaceous vegetation follows. If that tree is burnt the floristic composition of this vegetation may be different. It is clear that agroforestry was well established as a practice at the birth of agriculture, which is currently thought to be more than 5000 year ago in China and other countries in Asia.

The first examples of agroforestry in the literature where contrived beneficial interactions between trees and crops in Europe are outlined in Sylva a book written by Evelyn (1679). He stated that:

> The walnut delights in a dry sound and rich land; . . . in cornfields. . . . Burgundy abounds with them, where they stand in the midst of goodly wheat lands, at sixty and an hundred feet distance; and it is far from hurting the crop, that they look on them as a great preserver, by keeping the grounds warm, nor do the roots hinder the plough.

Scientific agroforestry where agronomists have quantified the yield advantages and sought to characterize the biological basis for the advantage have only occurred very recently however. Newman (1986) was the first time that it was demonstrated that it is possible to obtain a full yield of an orchard crop (pear) and a full yield of a vegetable crop (radish) on the same area of land. This has been called multilayered cropping or three dimensional agriculture in principle imitating a rain forest.

In summary agroforestry is an old practice and a new science. More on the history of agroforestry in temperate areas can be found in Gordon and Newman (1997).

Relevance and Subtypes

Agroforestry is relevant to all countries where woody perennials can grow. It is relevant to a range of "starting points" including; forests, rangelands, crop fields, orchards, gardens, tree crop/timber plantations and bare land. It is equally relevant in tropical and nontropical climates. Agroforestry subtypes can classified as the following based on the work of Newman and Gordon (1997)

1. Silvoarable: Crops and trees including swidden and other sequential systems
2. Silvopastoral: Animals and trees including forest grazing
3. Orchard/plantation intercropping: for example, crops in fruit/nut orchards or timber plantations (sometimes called interculture)
4. Forest farming/forest gardens, for example, growing crops in forests or creating a garden with a high density of trees
5. Environmental systems such as wind breaks or tree strips acting as pollution traps on river banks (riparian strips)

Permaculture is a concatenation of the words permanent and agriculture and if it contains woody perennials rather than herbaceous perennials could be a form of agroforestry. Key ideas held by the Permaculture movement are that rainfall harvesting using natural topography can be used to improve sustainability based on the "Keyline System" ideas of Yeomans (1958) and the idea that if humans/animals derived caloric staples and other foods from woody perennials there would be les soil erosion and other negative environmental effects than relying on grasses, for example, wheat/rice/barley etc. outlined by Russel Smith (1929). Unfortunately there are still very few yield and scientific studies of Permaculture in the peer reviewed literature.

Examples

The most widespread traditional agroforestry system in Europe is The *Dehesa* system of the Iberian peninsula covering an area of 2 million ha with widely spaced oak trees (grown for acorns for animal and sometimes human consumption), with cereal and fodder intercrops (Joffre et al., 1988). The trees are commonly planted on the square in fields at a density of over 50 stems per hectare. In the southern part of the range, dehesas can include oak trees grown for cork. The Dehesa system can therefore involve silvoarable or silvopastoral practices.

In Asia silvoarable systems with *Paulownia* spp. are also planted at spacings of 5 $\times$ 10 m to 5 $\times$ 40 m and the extent was reported at 1.3 million ha (Zhu et al., 1991). In India poplar grown as alley cropping systems is very common and has been very successful commercially (Newman, 1997). Recent travels by the author to upland areas East Africa and India show a very interesting example where farmer have looked at neighbors using *Grevillea robusta* as a shade tree in tea or coffee plantations. They have used this tree as part of silvopastoral systems and silvoarable systems using square panting, row planting and boundary planting. The extent of this must be now approaching 1 million ha. Farmers appear to prefer multipurpose trees that can provide timber and other products, for example, fruit or fodder. Deciduous trees also offer the best potential in minimizing negative effects on understory crops.

Benefits

When outlining claimed benefits of agroforestry it is important to ask "compared to what?"

The following entities serve to illustrate this

(a) Compared to nonwoody annual crops on their own
(b) Compared to a forest or timber plantation
(c) Compared to an orchard or tree crop plantation with no herbaceous component
(d) Compared to a pasture or rangeland with no woody component

A list of the most widely claimed benefits are given below compared to entities in square brackets, with reflections outlined in brackets.

1. Increased yield [a, b, c, d] (see section on measuring success)
2. Increased biodiversity [a, b, c, d] (specific taxa)
3. More profitable in the long term [a, b, c, d] (may reduce asset value of arable land due to reduced flexibility)
4. Reduced need for artificial fertilizers, for example, N and P [a, b, c, d] (mixtures of herbaceous crops can also achieve this)
5. Can reduce soil erosion and water pollution [a, b, c, d] (perennial herbaceous crops can also achieve this)
6. Can reduce runoff and potential flooding [a, c, d] (perennial herbaceous crops can also achieve this)
7. Reduced need for fungicides and insecticides [a, c, d] (mixtures of herbaceous crops can also achieve this)
8. More resilient to changing climate in terms of productivity and profit [a, b, c, d]
9. More carbon sequestration in the living biomass and soil organic carbon [a, c, d]
10. Can "attract" rainfall [a, c, d] by a variety of mechanisms
11. Animals suffer less from temperature extremes and can self-medicate [d]
12. Reduced animal feed requirement leading to cost and carbon footprint reduction [d]
13. Greater amenity and social benefits including health improvement poverty reduction [a, b, c, d]

Challenges and Future Research Needs

1. The benefits outlined above are not derived from a simple addition of any tree. Benefits are only obtained from specific species combinations planted at specific densities and spatial arrangements (number of rows in a block/strip and rectangularity). The challenge is to find the right configuration and this requires research and development in many cases.

2. The biological basis of many benefits depends on the age/size of the tree so the management of the system can be complicated.
3. Trees can be expensive to establish, protect from livestock, and to remove (including stump and roots) after productive life.
4. Mechanization can be difficult but is getting less so given the fact that trees can be designed/pruned/planted to allow access of machines and robots.
5. Extension services and policy makers find it difficult to facilitate agroforestry
6. Some existing policies and practices (e.g., corruption linked to felling licenses and timber transportation) limit use of timber trees by farmers.
7. Farmers may not have the skills and attitude to make money from trees especially timber trees.
8. Farmers and advisors may have a lot of misconceptions about trees and their effects on livestock/crop operations.

The Measurement of Success in the Optimization of Agroforestry

There are the broad categories of measures of effectiveness in agroforestry; agronomic, economic or environmental. Agronomic advantages can also be subdivided into three categories of combined yield main crop yield and feedstock yield (Newman, 1990), related to three questions that a manager may ask respectively:

1. I would like to manage two entities, is it better to combine them as agroforestry or treat them as separate entities on different parts of the farm?
2. I am mainly interested in one entity could the addition of another entity via agroforestry help?
3. I am interested in a product, blend or feedstock is it better to combine them as agroforestry or treat them as separate entities on different parts of the farm?

For combined yield, one of the commonly used terms for evaluating combined yield is that of land equivalent ratio (LER) (Mead and Willey, 1980).

$$LER = \frac{Y_A}{S_A} + \frac{Y_B}{S_B}$$

Y_A and Y_B are the yields of two crops A and B in intercropping (in the case of agroforestry a crop and a tree species), S_A and S_B are the yields of the same two crops in monoculture under identical conditions, and L_A and L_B are the partial LER values for the two species. If we imagine an agroforestry system where species A (the crop species) normally produces 10 tons of wheat per hectare per year, species B (the tree species) normally produces 10 tons of walnuts per hectare per year, and the two are grown together in a 50:50 mix, with each occupying half of a 1 ha site. If 5 ton of grain and 5 tons of nuts is obtained then LER = 5/10 + 5/10 or 1. This shows no advantage. If it was 7/10 plus 5/10 then the LER would be 1.2 or in other words 20% more land would be required to obtain the same yield from monocultures. Newman (1986) obtained LER values of over 2.0 for agroforestry involving pear and radish. In other words twice the land area would be required if pear and radish were grown separately as monocultures.

Main crop yield is simpler. If the farmer sees the wheat as the main crop he will only tolerate the presence of walnut if 10 tons per hectare or more of wheat can be obtained.

For feedstock yield the farmer may be interested in the straw or nut tree pruning to produce energy from biomass. If the total biomass of straw and prunings is higher than the biomass obtained from the highest yielding sole crop then there is an advantage. The same approach is relevant to fodder or blended flour to make bread.

Economic analysis could be a comparison of profit or return on capital under a range of scenarios. If the return from the tree takes a number of years then some form of discounted cash flow analysis may be required. Income stability could also be a key measure. Agroforestry is a method of not "having all your eggs in one basket."

Environmental analysis requires that an environmental variable is selected such as soil loss due to erosion, ammonia pollution reduced from livestock, or carbon footprint reduced.

The Ecological Basis of Yield Advantage in Agroforestry

This is a very complicated area and is asking the question why is the effect of mixing in the woody perennials not negative, competing for resources used by the crop or other entity? Newman (1982) presented a simple theoretical framework of three mechanisms; partitioning, synthesis and modulation. Partitioning is where the tree is using phosphate or other resources at a different time in the year to the crop (temporal partitioning) or at a different layer in the soil (spatial partitioning). Synthesis is where the tree is producing a chemical change in the environment of the other entity that could help to reduce pests or disease for example. Modulation is where the tree is producing a physical change in the environment of the other entity that could help to reduce the damaging effect of wind or to provide a physical support for example. Recent research is showing that trees in agroforestry could improve water availability and "attract rainfall" for instance (Ong et al., 2015). This is an example of modulation.

Scientific Approaches to Optimization

The ecological combining ability of the system can be helped by an understanding of what interactions are happening above (e.g., light availability) and below ground (e.g., water and nutrient availability). The two environments can be separated for analysis using barriers in trenches or plastic pots (phytometers as outlined in Newman (1984)). Genotypes can be selected to optimize production, for example, an understorey can be selected that is shade tolerant or an overstorey can be selected that is columnar in form to reduce shade.

Management Approaches to Optimization

The spatial arrangement of trees can be manipulated to optimize effects as can the planting sequence. Special low cost techniques can be used to protect trees from damage by livestock. Tree pruning can be used to reduce shade and in fruit/nut trees different rootstocks can be used optimize precocity, tree form and regularity of bearing.

References

Evelyn J (1679) *Sylva or a discourse on forest trees and the propagation of trees in his majesties dominions*, 3rd edn London: Royal Society.

Gordon AM and Newman SM (1997) *Temperate agroforestry*. United Kingdom: CAB International. ISBN: 0-85199-147-5269.

Joffre R, Vacher J, De Los Llanos C, and Long G (1988) The dehesa: An agro-silvopastoral system of the Mediterranean region with special reference to the Sierra Morena area of Spain. *Agroforestry Systems* 6(1): 71–96.

Mead DR and Willey RW (1980) The concept of a 'land equivalent ratio' and advantages in yields from intercropping. *Experimental Agriculture* 16: 217–228.

Newman SM (1982) Ecological energetic and economic aspects of intercropping systems. In: Hall DO and Moreton J (eds.) *Solar World Forum*. 2: pp. 1248–1253. Oxford: Pergamon.

Newman SM (1984) The use of vegetable phytometers in the evaluation of the potential response to understorey crops to the aerial environment in an Interculture system. *Agroforestry Systems* 2: 49–56.

Newman SM (1986) A pear and vegetable Interculture system: Land equivalent ratio, light use efficiency and dry matter productivity. *Experimental Agriculture* 22: 383–392.

Newman SM (1990) *Temperate agroforestry: Its role, potential, and recent advances. Invited key note paper*. Proceedings IUFRO World Forestry Congress. Montreal, 1990. vol. B, pp. 282–292. Austria: IUFRO.

Newman SM (1997) Poplar agroforestry in India. *Forest Ecology and Management* 90(1): 13–17.

Newman SM and Gordon AM (1997) Temperate agroforestry: Synthesis and future directions. Chapter 8, In: Gordon AM and Newman SM (eds.) *Temperate agroforestry*, pp. 251–265. United Kingdom: CAB International. ISBN: 0-85199-147-5. 269.

Ong CK, Wilson J, Black CR, and van Noordwijk M (2015) Synthesis: Key agroforestry challenges for the future. In: Ong CK, Black CR, and Wilson J (eds.) *Tree crop interactions*, 2nd edition United Kingdom: CABI. ISBN: 978-1-78064-511-7.

Slowik JG, Stroud C, Bottenheim JW, et al. (2010) Characterization of a large biogenic secondary organic aerosol event from eastern Canadian forests. *Atmospheric Chemistry and Physics* 10: 2825–2845.

Russel Smith J (1929) *Tree crops: A permanent agriculture*. New York: Harcourt Brace. ISBN: 0-933280-44-0.

Yeomans PA (1958) *The challenge of the landscape: The development and practice of keyline*. Australia: Keyline Publishing.

Zhu ZH, Cai MT, Wang SJ, and Jiang YX (eds.) (1991) *Agroforestry systems in China*. Canada and CAF, China: IDRC.

Further Reading

Tropical Agroforestry

Huxley P (1999) *Tropical agroforestry*. Blackwell Sciences. ISBN: 0-632-04047-5.

Temperate Agroforestry

Gordon AM and Newman SM (1997) *Temperate agroforestry*. United Kingdom: CABI p.269, ISBN: 0-85199-147-5.

Hislop H and Claridge J (eds.) (2000) *Agroforestry in the UK bulletin 122*. United Kingdom: Forestry Commission.

Hoare AH (1928) *The English grass orchard and the principles of fruit growing*. London: Ernest Benn.

Understanding Ecological Interactions and Environmental Benefits

Ong CK, Black CR, and Wilson J (eds.) (2015) *Tree crop interactions*, 2nd edition United Kingdom: CABI. ISBN: 978-1-78064-511-7.

Young A (1997) *Agroforestry for soil management*, 2nd edition, United Kingdom: CABI. ISBN: 0-85199-189-0.

Relevant Websites

http://www.worldagroforestry.org/—World Agroforestry Centre.

http://www.agroforestry.ac.uk/—Farm Woodland Forum, United Kingdom.

https://nac.unl.edu/—National Agroforestry Centre, United States.

http://www.centerforagroforestry.org/—Agroforestry Centre at University of Missouri, United States.

Anthropogenic Landscapes

Maria Rita Pasimeni, Donatella Valente, Teodoro Semeraro, Irene Petrosillo, and Giovanni Zurlini, University of Salento, Lecce, Italy

Anthropogenic Landscapes: A Scientific and Social Challenge

Anthropogenic activities like modifying natural landscapes for economic purposes or changing management practices on human-dominated lands have transformed a large proportion of Earth's surface. Human actions have almost completely transformed, eroded, and fragmented preexisting natural landscapes, leaving only a limited number of natural areas, in an often heavily anthropogenic matrix. Nowadays, this phenomenon has reached critical thresholds, since residual areas of natural landscapes are increasingly rare and dispersed in many parts of the world. Human actions, mostly associated with agricultural expansion and intensification, conversion of perennial habitats to cultivation fields, farming practices such as fire and crop rotation, urban sprawl, industrial development, road infrastructure, or any other substitution or conversion of an original natural landscape with an anthropogenic type, lead to anthropogenic landscapes.

Recent and novel approaches applied in the study of anthropogenic landscapes move from the traditional separation of social and ecological components to a social-ecological landscape (SEL) approach, considering SEL as a whole coevolving and historically interdependent system of humans-in-nature. In this respect, addressing SELs represents a more pragmatic basis for envisioning how the anthropogenic landscapes work and how we would like them to be, as SELs represent the spatially explicit integration of social–political and ecological scales in the geographical world, recognizing the primary role of humans as a driving force in intentionally shaping and modifying systems' compositions and processes.

Natural and semi-natural ecosystems and landscapes provide a mix of ecosystem goods and services, both private and public, to human society now and in the future; these are referred to as natural capital provided by multifunctional landscapes. Nowadays, landscapes are disrupted, impaired, or reengineered by many human activities, in ways that can exceed the rate, severity, and spatial extent of even the largest natural disturbance. In this context, land-use/land-cover changes refer to the way in which man uses the territory and its resources, creating patterns that can alter natural processes, and they represent the most tangible effects. The landscape mosaic is altered by the direct human use, which inevitably has profound effects on the ecosystem's structures and functions with its activities—that is, agriculture, mineral extraction, and urban, industrial, and tourism settlements, with consequences on the availability of goods and ecosystem services.

In this perspective, man is an integral part of SELs, posing difficult challenges to address a new era of epochal changes that need to be addressed, the Anthropocene, to define strategies and behaviors toward sustainability and a renewed balance between the humankind and the environment. The beginning of this new era indicates the current ability of man to act, in a quantitatively significant manner, on the drivers of SEL dynamics, by heavily affecting the stock of natural resources with direct and indirect impacts on economy, society, and landscapes. Therefore, a single species, namely man, acts as a prevailing force, by directing SEL dynamics and evolution, trying to adapt the environment to its own needs.

SELs are characterized by a large number of components that interact in a nonlinear way and exhibit intrinsic uncertainties and adaptive properties through time. SELs generally show a nonstationary and complex behavior, with their usual phase state fluctuating around some averages. They are commonly assumed to respond smoothly to gradual change in climate, habitat fragmentation, or resource exploitation. Such a condition, however, can be sporadically interrupted by an abrupt shift to a radically different regime. For these reasons, such systems should be conceived as complex adaptive landscapes: complex refers to the multiple variables that generate a nonlinear and unpredictable behavior in the medium and long term; adaptive means that their dynamism can evolve in space and time by adapting the composition and configuration of the elements and functions to the changing conditions that arise. Complex adaptive landscapes are deemed to exhibit alternative stable organizations with multiple stable states of the same ecological phase. These multiple states can be visited several times during the landscape evolutionary trajectory, and represent the phase attractor. If the structure of the SEL becomes radically different, a phase transition takes place; transitions between different stable states are due to changes in the interactions of structuring variables and processes characterizing SELs.

The acceleration that human activities have generated in the formation and transformation of landscapes adds uncertainty to that already present in SELs. In addition, each complex adaptive landscape is characterized by a strong dependence on the set of driving forces acting on it, and by all historical events that have determined landscape properties and its ability to respond to future shocks. Therefore, the current state of anthropogenic landscapes can be fully understood only in the context of state variable change trajectory, where traditional ecological knowledge pays a crucial role, and landscape ecology is fundamental to face the difficulty in managing anthropogenic landscapes in a dynamic and increasingly accelerated context.

Landscape Ecology Is Central for Anthropogenic Landscapes Studies

Landscape ecology is considered to be a holistic and transdisciplinary science for landscape study, appraisal, history, planning and management, conservation, and restoration dealing with the interrelation between human society and its

 doi:10.1016/B978-0-12-409548-9.10602-5

living space. Several authors have, over time, provided a definition of landscape ecology, as shown by the following examples:

- It considers the development and dynamics of space heterogeneity, interactions in space and time, changes in heterogeneous landscapes, spatial heterogeneity influences on biotic and abiotic processes, and, last but not least, land management issues.
- It tries to understand the development and pattern dynamics of ecological phenomena, the role of disorder in ecosystems, and the spatial and temporal scales of ecological events.
- It highlights the relationship of reciprocal pattern-process influence as well as the need to study its relationships in space.
- It studies the localization, spatial distribution, and shape of landscape elements in different degrees of naturalness in order to understand their structure, processes, and dynamics.

The aspects arising from these definitions concern the quantification of the elements that constitute the landscape mosaic and the description of their spatial distribution in order to appreciate the reciprocal relationships and constraints with biotic and abiotic processes as well as the spatial pattern configuration. It also recognizes the central role of the scale concept in spatial and temporal analysis of SELs as an effective tool for addressing the relationships between their spatial heterogeneity, processes and time dynamics.

Patch-Mosaic

Spatial structure or pattern refers both to the composition of the elements, namely their quality and abundance (what and how much is there?), and to their spatial configuration (how is it arranged?). The pattern is the result of: (1) the spatial and temporal heterogeneity of abiotic conditions (i.e., climatic, regional-scale morphology, and local soil composition); (2) the biotic interactions among communities, including humans; and (3) the dynamics of the system due to evolutionary processes and the regime of natural and/or anthropic disturbance. The study of the landscape pattern is usually based on the concept of patches and is framed in the theoretical and methodological context of the matrix-corridor-patch landscape model or fragmentation pattern. The patch-matrix model is based on a substantially schematic and rather static view of the landscape that views individual areas as "islands" of habitat that are connected by habitat corridors such that all areas together constitute a regional habitat network. With that model, the "contents" of the network are described by landscape descriptions of patches and corridors, while the "context" of conservation areas refers to the nature of the surrounding landscapes. At the landscape level, the context, which could be envisioned as a "buffer" around a site, may or may not help to maintain ecosystem functioning within a protected area (PA), allowing animal and plant dispersal and gene flow that is essential for population maintenance. The patch is defined as a surface area relatively homogeneous, differing from its environment in nature or appearance at a given scale. The patch structure has important implications for the supported processes and communities as well as its role in the mosaic: for example, its form is the expression of ongoing processes within and between adjacent patches. Being spatially and functionally connected to all other surrounding elements of the landscapes, the patch is actually the result of spatial and temporal processes taking place inside and outside. The set of patches constitutes, at a given scale, the landscape mosaic. The mosaic is the physical basis on which structures are organized, relations are created, and processes are established, and thus is to be considered functional and not just the structure of the landscape. The composition and configuration of a landscape mosaic has such influence on the dynamics of ecological systems that the same landscape, in the presence of a different configuration of patch arrangement, can have different properties.

The development of tools and indices to quantify the heterogeneity of the mosaic has contributed to the development of hundreds of quantitative measurements of the landscape patterns to describe different aspects of spatial heterogeneity. However, the matrix-patch-corridor model limitations have led to the development of alternative analytical techniques beyond the simple geometric evaluation of the patch as a basic element in the landscape mosaic survey, by adopting a multiple scale perspective, able to appreciate complex landscape properties such as multiscale connectivity or fragmentation. The novelty of this approach is that the landscape components become meaningful only when considering the context in which they are placed.

Scales

Scale refers to the spatial and temporal dimension of an object or process. The pattern and dynamics of a landscape mosaic are deeply influenced by the spatial and temporal scales adopted in the analysis. Two fundamental components of the scale are the extent and the grain. The extent defines the upper resolution limit of a study, and the entire area examined during an investigation or the entire time period studied represents it. Grain represents the lower-resolution, fine or coarse limit of the survey—that is, in spatial terms, the smallest entity that can evaluate in the landscape, while in temporal terms it corresponds to the smallest time window selected in the survey.

Spatial and temporal scales are related to each other. The ability to predict ecological phenomena characterizing SELs actually depends on the identification of the appropriate spatial and temporal scales. In practice, the extent and the grain are often established and/or determined by the data characteristics (i.e., aerial photo scale or sample rate) or processing techniques rather than being derived from the ecological property of the system in examination.

The description of a mosaic cannot disregard the scale constraints imposed in the analysis since ecological systems are scalar entities. The proper identification of the landscape patterns is a scalar problem, since the pattern is the result of processes and

structures that are in turn scale-dependent. The quantitative analysis of a landscape mosaic is based on the use of metrics describing patch composition and configuration properties.

In SELs, as complex adaptive landscapes, there is no a priori single scale for a survey able to take into account the number and variability of features characterizing SELs. Therefore, an appropriate approach should be based on their evaluation at multiple scales in order to identify scalar intervals of invariance of their properties. In other words, there is the need to identify scalar domains of a phenomenon/process and intervals along the spectrum of levels within which the pattern-process relationships are well defined and vary in a predictable manner. In addition to the focal point of the survey, multiscale analysis shows the valuable information of both the levels immediately higher and lower with respect to the focal point. In scalar terms, this means extending the survey to larger and smaller extents or grains in the surroundings of the focal point.

Disturbance

Anthropogenic disturbances are typically imposed by groups of people who are organized at different levels (i.e., from household to global) in a panarchy, with differing views as to which system states are desirable or which ecosystem services are to be exploited. Land-use change depends on individual and social responses to changing economic conditions, mediated by institutional factors, such as markets and policies, and increasingly influenced by global markets. There may be circumstances where landscapes do not change due to social and cultural drivers like new reserves based on nature or cultural values. Yet climate change and weather extremes could occasionally trigger further change, resulting in ecological surprises. The difficulty of controlling or predicting the biophysical effects of all those forces has resulted in failed attempts to closely manage and regulate the dynamics of ecosystems. The usual state of affairs in living systems like SELs is one of systems fluctuating around some trend or stable average; however, this condition is sporadically interrupted by an abrupt shift to a radically different regime. Disturbance can be deemed as an event causing departure of a living system from the normal range of conditions typical of its basin of attraction. The apparent paradox that disruption of the existing order (i.e., disorder) and persistence (i.e., order, stability) always coexist in living systems such as SELs is addressed by the concept of resilience, defined as the amount of disturbance a system can absorb without shifting into an alternative state and losing function and services. Many disturbances can have a strong climate forcing; nevertheless, the relative importance of different drivers varies among systems and can even vary through time within the same system. To understand the disturbance regime of biophysical systems like SELs and reveal possible regularities, we must consider the variable frequency of forcing due to the physical environment and its historical role in shaping biophysical systems.

New Methodological Approaches to Analyze Anthropogenic Landscapes

In the context of landscape ecology, emerging methodological approaches are represented by retrospective analysis, remote sensing application in SELs, and multiscale disturbance approach to value SELs transformations.

Retrospective Analysis to Value SEL Transformations

Land-use change represents a primary human effect on natural systems, causing fragmentation and habitat loss, which are the greatest threats to biodiversity and Ecosystem Services (ESs). Since human land use is a major force in driving landscape change, landscape dynamics can be better addressed in the context of SELs, as complex adaptive landscapes, integrating phenomena across multiple scales of space, time, and organizational complexity.

The most direct way to quantify composition dynamics is to analyze land cover area change using retrospective analyses. Retrospective analysis is needed to understand the present system conditions in the context of a trajectory of change that encompasses system past, disturbance regimes, and cross-scale interactions and constraints in a hierarchy of systems, in addition to endogenous self-organizational processes. Retrospective analysis is useful, as it links the present-day system status with its past dynamics and enables the identification of possible evolutionary trajectories to reveal continuity, turnover, directions, or degree of changes. By providing a means for analyzing short- and long-term system dynamics and for assessing the complex structure of the multiscale relationships in SELs, retrospective observational studies are valuable as they can address the role and nature of feedback mechanisms and scale-dependent interactions in systems.

Such system-level properties differ in three important ways from traditional ecological indicators. First, they need to be addressed by holistic measures describing, as far as possible, the entire system, not just specific subsystems or components. Second, as any important aspects of SELs may not be directly observable, they must be inferred only indirectly using surrogates, whose relationships between systems' properties may be dynamic, complex, and multidimensional. As a system evolves and adapts in time, the components and functions sustaining particular system properties change, so that a surrogate is useful only in a particular context or time frame. Third, some features of a complex landscapes (i.e., landscape resilience) focus on properties that underlie SEL capacity to provide ecosystem goods and services, whereas other indicators often address only the current state of the system component or service.

As an example, a retrospective analysis has been used to evaluate the effectiveness of conservation policy by studying decades of aerial photographs of a natural PA, Torre Guaceto in the Apulia region. In particular, the Italian government designated this PA as

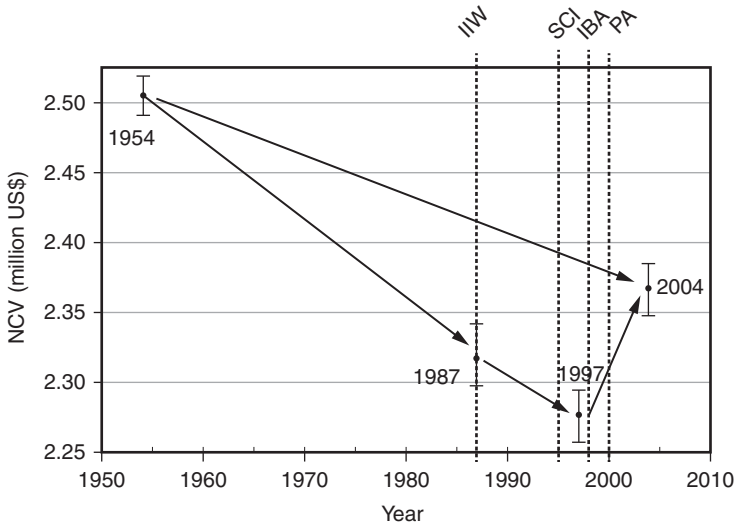


Fig. 1 Retrospective analyses: temporal changes of natural capital value (NCV) in a Mediterranean protected area, highlighting the different temporal recognitions of its natural value (*IIW*: International Important Wetland; *SCI*: Site of Community Importance; *IBA*: Important Bird Area; *PA*: institution of the natural Protected Area). *Bars* represent the NCV variability range due to an estimated 2 m average digitalization error.

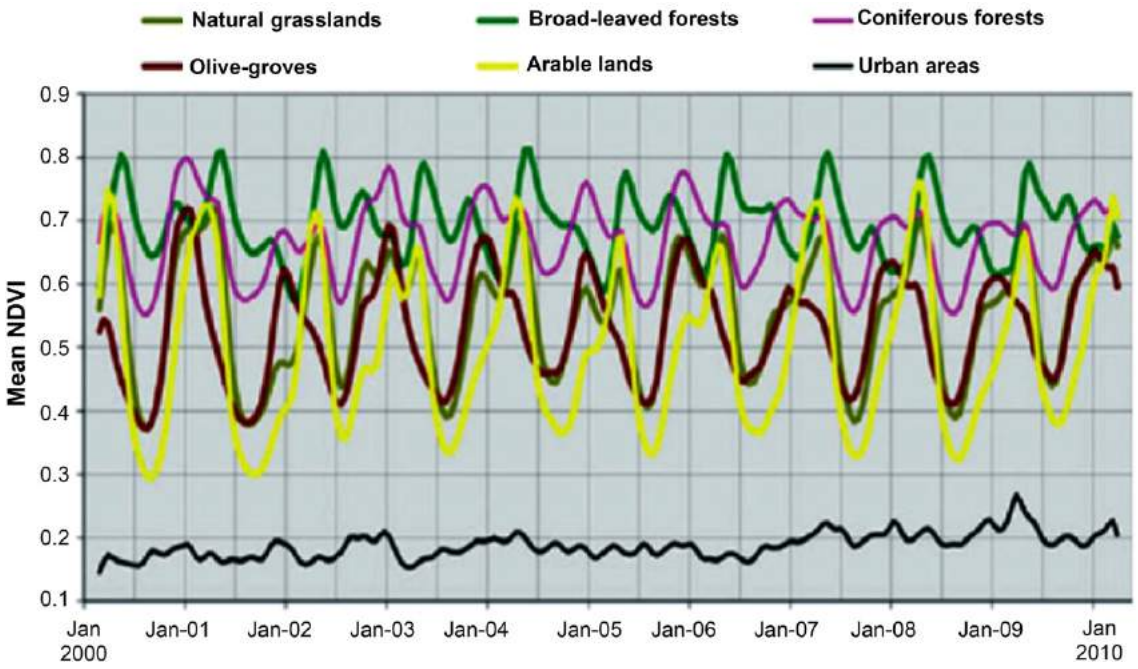


Fig. 2 Mean Normalized Difference Vegetation Index (NDVI) time trajectories computed on 148 16-days NDVI maximum values composite images acquired by the two MODIS platform TERRA and AQUA with a spatial resolution of 250 m (MOD13Q1 v.005 and MYD13Q1 v.005) for a Mediterranean region.

an International Important Mediterranean Wetland (IIW) under the Ramsar Convention in 1971, a Site of Community Importance (SCI) under Natura 2000 in 1995, an Important Bird Area (IBA) in 1998 under the European Birds Directive, and an institution of the natural PA in 2000. However, the findings suggest that only the PA measure has had a noticeable impact, improving the conservation value of the area.

The research assessed the effectiveness of each of these recognitions in conserving the area by calculating changes to its natural capital value (NCV). This concept places a monetary value on tangible commercial services that come from the natural environment—such as clean water and food production—and noncommercial services including aesthetic benefits, recreation, and climate regulation. The specific contribution of this example is to show that those coefficients could play a role as operational surrogates to evaluate the recent temporal dynamics of the overall flux of natural capital in the study area.

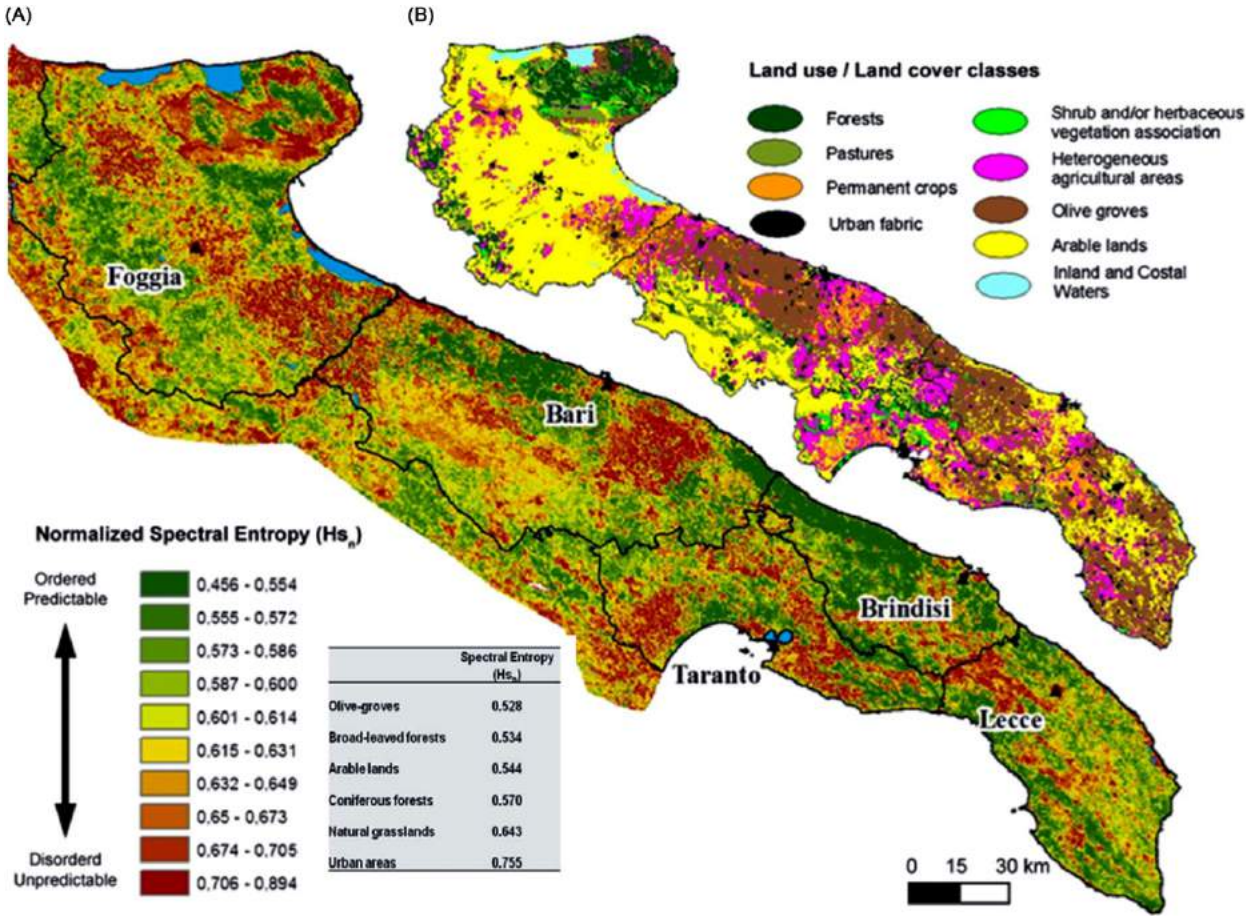


Fig. 3 Map of normalized spectral entropy (Hsn) (A) and map of major LULC classes (B) for a Mediterranean region. Hsn is based on the same composite images used for Fig. 2. The LULC categories are derived from a CORINE land cover map of the year 2006.

Overall, the temporal change of NCV decreased from 1954 to 1997. The IIW and SCI designations had no apparent impact on conserving the area and the researchers were unable to assess the impact of the IBA accreditation due to gaps in the data. However, there was an increase in NCV from 1997 to 2004, which the authors attribute to the PA official designation (Fig. 1).

Designations are, therefore, not so effective in themselves, and local management by means of regulations, ecological monitoring, and environmental education and communication activities seem to be more effective in supporting natural capital conservation and enhancement.

Remote Sensing and Landscape Ecology

Remote sensing is a primary source of information and it has become a proven tool for scientists to monitor environmental phenomena synoptically and globally, to understand major disturbance events and their historical regimes at regional and global scales. It has provided valuable indices to describe and quantify natural and human-related land-cover transformations and processes.

Vegetation indices are broadly recognized as a spatially explicit robust indicator to gauge social–ecological processes such as habitat-land use conversion (i.e., urban sprawling) or crop rotation. Vegetation indices are spectral transformation of two or more bands of satellite images designed to enhance vegetation proprieties and canopy structure. For example, the Normalized Difference Vegetation Index (NDVI) is widely used to identify and assess the impact of disturbances such as drought, fire, flood, frost, or other human-driven disturbances. NDVI is used to quantify the photosynthetic capacity of plant canopies, and it is calculated from these individual measurements as follows:

$$\mathrm{NDVI} = (\mathrm{NIR} - \mathrm{RED})/(\mathrm{NIR} + \mathrm{RED})$$

where RED and NIR are the spectral reflectance measurements acquired in the red (visible) and near-infrared regions, respectively.

Furthermore, NDVI-related indices can also supply data on potential species richness in many parts of the world, and may help monitor ecosystem services (ESs) like carbon sequestration, water cycling and regulation, and soil fertility.

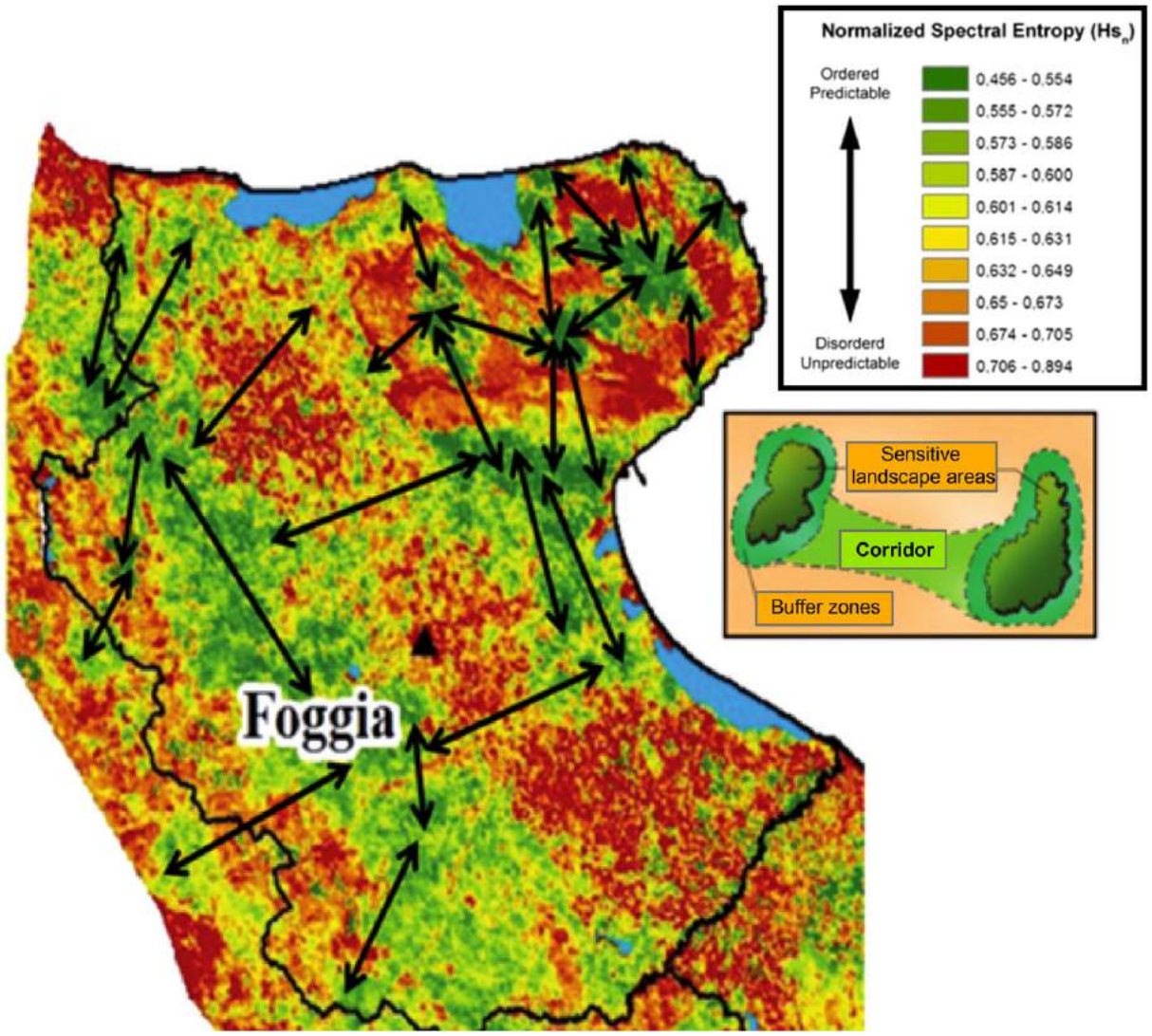


Fig. 4 Example of connectivity network based on a predictability map of a Mediterranean area. *Arrows* search to connect the most predictable areas indicating which areas could be transformed in more predictable by a proper change of their management or type of land cover to foster the overall network.

Time series of vegetation indices are an essential reservoir of past SEL, information as they keep track of the disturbances that occurred. They must be deeply investigated since they can reveal a great deal about the magnitude of disturbance and the timing of return to the usual functionality of systems.

An example of NDVI time series (2000 – 10) for primary land use/land cover (LULC) categories for a Mediterranean region (Fig. 2) demonstrates the differences in the inter-annual periodicities of NDVI related to both human controls (arable lands and olive groves) and natural balancing feedback loops (natural grasslands, broad-leaved forests, coniferous forests), whereas urban areas show a disordered behavior.

In the context of landscape ecology, entropy-based indices like Shannon's H or contagion are among the most commonly used metrics to analyze time series in order to represent the dynamics of landscape composition and configuration diversity, because they are capable of reflecting changes in the level of human impacts and disturbance regimes, species diversity and habitat use, or biodiversity level estimates from remotely sensed images.

The normalized spectral entropy (Hsn), for example, is an entropy-related index able to describe the degree of order and predictability (i.e., regularity) within an ecological time-series based on its power spectrum. This index has been suggested as a holistic indicator for system-level properties able to characterize heterogeneity in time and pointing to the system's self-organization strength. Spectral entropy (Hsn) of NDVI time series can be of practical help in mapping predictability of SELs—for example, the predictability map of invariant structures as provided by NDVI (Fig. 3). It can be calculated on the trajectories for each pixel of the map for 10-year series of 16-day maximum NDVI composite images (Fig. 2). In such an example, distinctive spatial patterns are shown at 250 m resolution with greener zones meaning higher predictability (1-Hsn), that is, more regular

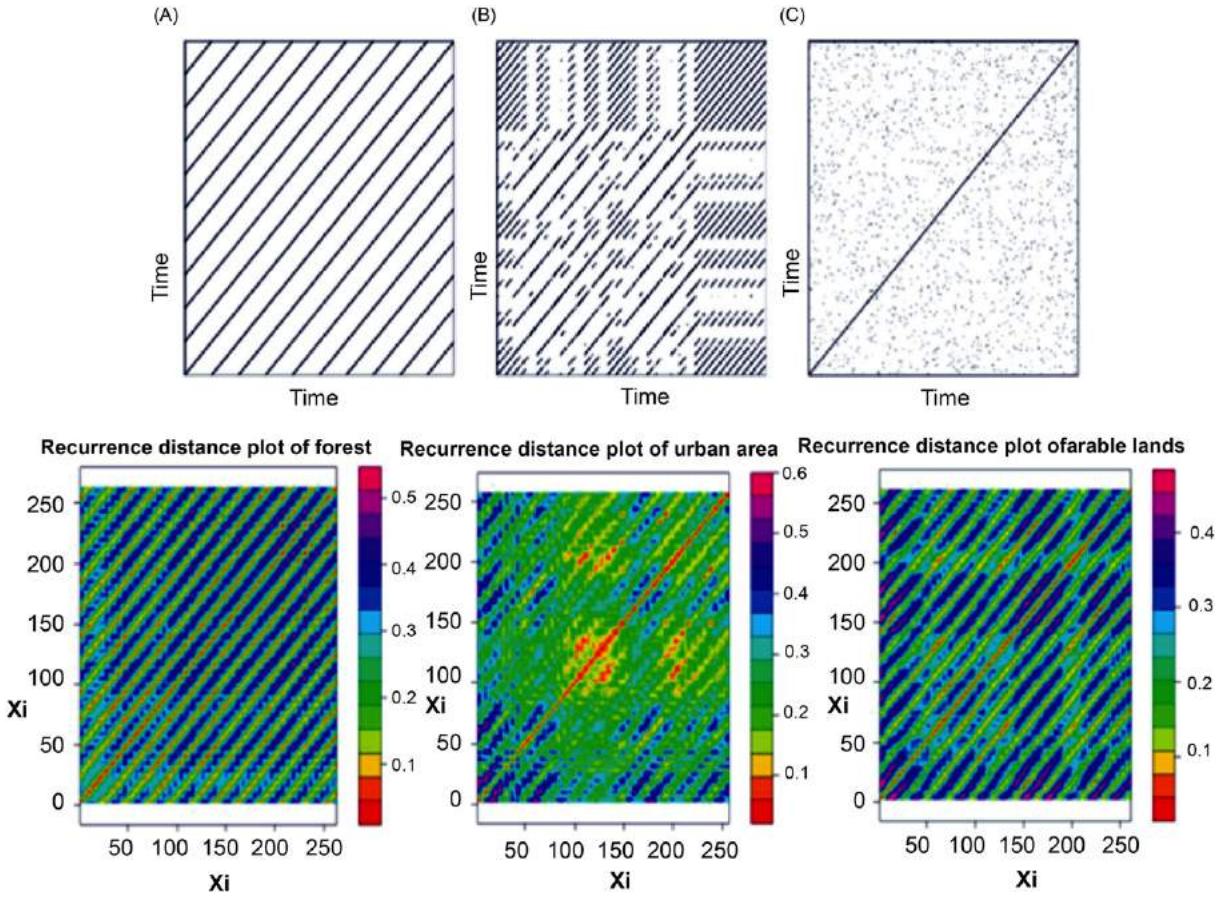


Fig. 5 Recurrence plots of three prototypical systems (top) (A) a periodic motion with one frequency (very predictable), (B) of a chaotic system (unpredictable), and (C) of uniformly distributed noise, and recurrence plots (bottom) of forest, urban areas, and arable lands for a Mediterranean region (2000–12). Marwan, N., Romano, M.C., Thiel, M., Kurths, J., 2007. Recurrence plots for the analysis of complex systems. Physics Reports 438, 237–329.

time series, while reddish areas are more unpredictable. Low spectral entropy refers to locations with less complex temporal pattern, that is, with more stable cyclical developments. Clear coherent regions of predictability and unpredictability emerge, as well as gradients of transition between the two. Large predictability geographic regions arise in the map of the Mediterranean region (e.g., olive groves, large farmlands), whereas unpredictability regions tend to be associated with heterogeneous cultivation areas.

Once one obtains a predictability map of invariant structures, one can think of applying different modeling tools to derive, under uncertainty, what could possibly be an effective corridor network and a suitable fragmentation for the future (Fig. 4). So it could be discovered that along with classical *green* and *blue* ways, other elements in the landscape could be identified and considered crucial based on their predictability for the maintenance of the overall connectivity in the face of climate change. Consequently they could be converted in persistent through planning and management efforts. Time series of natural or human dominated processes can have a distinct recurrent behavior, that is, periodicities (as seasonal or Milankovich cycles), but also irregular cyclicities (as El Niño Southern Oscillation). Moreover, the recurrence of states is a fundamental property of deterministic dynamical systems and is typical for nonlinear or chaotic systems.

Operatively, recurrence analysis reveals all the times when the phase space trajectory of the dynamical system visits roughly the same area in the phase space and therefore it recurs. Such cyclic patterns provide useful indications on the resilience capacity of a SEL in a retrospective way, exploring the ability of the system to absorb disturbances occurred in the past. One promising method of nonlinear time series analysis is the recurrence quantification analysis. The recurrence plot (RP) depicts the collection of pairs of times at which the trajectory is at the same place. It is a visualization (or a graph) of a square matrix, in which the matrix elements correspond to those times at which a state of a dynamical system recurs (columns and rows then correspond to a certain pair of times).

Let us consider the RPs of three prototypical systems (Fig. 5, top), namely of a periodic motion on a circle (Fig. 5A), of a chaotic system (Fig. 5B), and of uniformly distributed, independent noise (Fig. 5C). In all systems recurrences can be observed, but the patterns of the plots are rather different. Long and noninterrupted diagonals reflect the periodic motion. The vertical distance between these lines corresponds to the period of the oscillation.

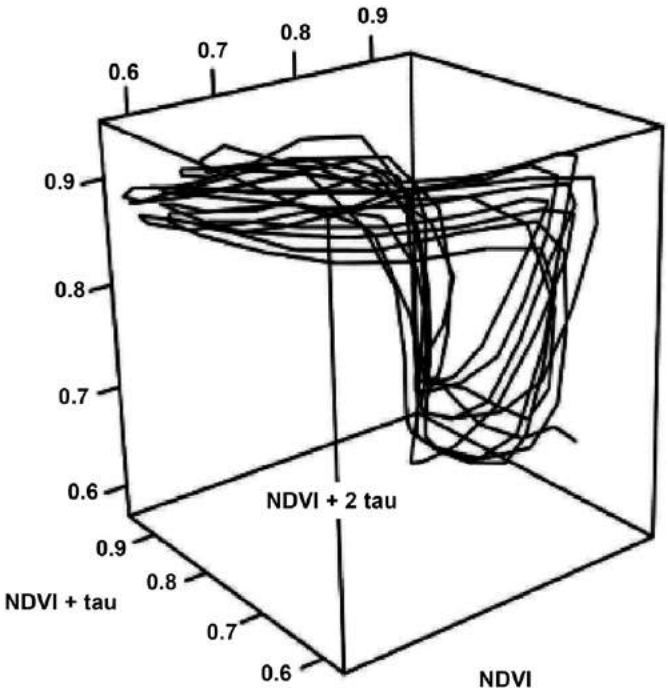


Fig. 6 Three-dimensional reconstruction of Normalized Difference Vegetation Index signal in phase space for the land cover class of forest (2000–12) by the method of time delays, with phase space trajectories visiting approximately the same area all the time (cf. Fig. 5).

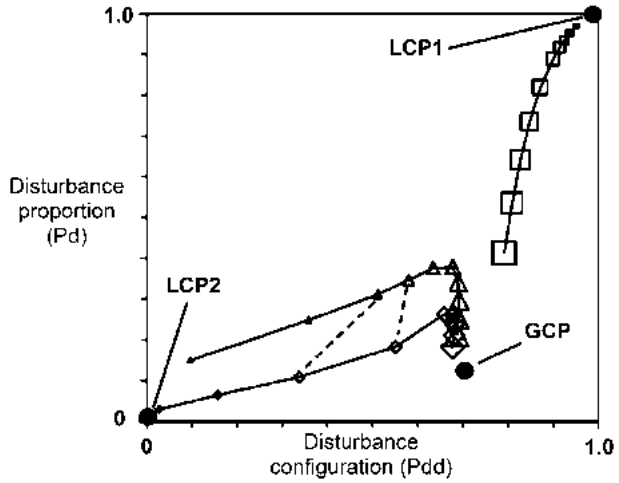


Fig. 7 The conceptual model is illustrated by three disturbance profiles in a pattern space defined by the local proportion and contagion of disturbance. Each disturbance profile connects the observed patterns across measurement extent (scale), and the size of the symbols indicates the relative extent. In addition to the global convergence point (GCP), there are two local convergence points for an extent equal to the size of one pixel that is either disturbed (LCP1) or undisturbed (LCP2). The *dotted lines* illustrate a cross-scale mismatch.

The RPs of NDVI for a Mediterranean region (Fig. 6, below) show that space trajectories can visit roughly the same area in the phase space all the times, that is, ES dynamics for forest are fairly spatiotemporally predictable, thus the probability that such a state will persist is rather high (resilience is high).

Arable lands are less predictable despite the action of self-correcting balancing feedback loops (i.e., drought-irrigation, soil impoverishment-fertilization). Urban areas, on the contrary, show a chaotic behavior (Fig. 5, below).

Nonlinear analysis of spatial-temporal dynamics of SELs helps gauge the capacity of any SEL to activate balancing feedback loops to adjust its responses to drivers. Adaptability is a fundamental component of resilience and captures the capacity of any SEL to learn, combine experience and knowledge, and activate balancing feedback loops to adjust its responses to changing external drivers and internal processes, and continue developing within the current stability domain or basin of attraction (i.e., Forest in Fig. 6); whereas resilience is the probability that such state will persist. Moreover, looking at phase space trajectories of time series

can help to look at possible impeding regime shifts. The three-dimensional reconstruction of NDVI signal in phase space for Forest (2000–12) (Fig. 6) shows that phase space trajectories have been visiting for 12 years approximately the same area all the times showing a high adaptability.

Multiscale Disturbance Approach

The effects of land-use intensity on local biodiversity and ecological functioning in SELs depend on spatial scales much larger than a single field or land use. This demands a landscape perspective, which takes into account the spatial arrangement of surrounding land-use types at multiple scales. So, a central question in landscape ecology is how patterns and processes change with the scale of observation. A scale domain has been defined as an interval in scale space within which landscape patterns and/or pattern-process relationships are stable or predictable. Knowledge of scale domains is important because inferences made within one domain do not necessarily occur in another domain. Therefore, spatial patterns at multiple observation scales provide a framework to improve the understanding of pattern-related phenomena.

Taking into account the scales and patterns of human land uses as source/sink disturbance systems, a conceptual model has been advanced that describes a framework to characterize and interpret the spatial patterns of disturbances along a continuum of scales in a panarchy of nested jurisdictional SELs like regions, provinces, and counties.

The conceptual model considers a pattern space defined by the proportion of composition (Pd: disturbed pixel) and configuration (Pdd: contagion between disturbed pixel) of disturbance (Fig. 7). Therefore, considering a binary map showing pixels of disturbed and undisturbed locations, multiscale disturbance patterns can be measured and mapped using an overlapping pixel-level moving window whereby scale is varied by changing the size of the window. Within a given window, composition is expressed as the probability of disturbance, and is estimated by the proportion of pixels that are disturbed (Pd). Configuration is measured within the same given window by the adjacency of disturbance, given by the probability that a disturbed pixel is adjacent, by the four-neighbor rule, to another disturbed pixel (Pdd), so that it is a measure of contagion.

In that pattern space, there is a global convergence point (GCP), which is the [Pd, Pdd] value corresponding exactly to the extent of the entire study area. For smaller extents, the observed [Pd, Pdd] departs from the GCP if the local pattern is different from the global pattern, where local is defined by a particular location and extent. At a given location, the trajectory away from the GCP is the disturbance profile, which describes the scaling of pattern at that location. A multiscale domain is a set of geographic locations with similar disturbance profiles. Whereas classical scale domains are identified by local invariance of pattern in pattern space, multiscale domains are identified by local invariance of the scaling of pattern in geographic space. This conceptual model made it possible to exploit the local sensitivity of pattern metrics such as proportion and contagion, by incorporating their geographic variance into the definition of a multiscale domain. The conceptual model has a high potential for the prediction and management of disturbance-related processes such as the spread of invasive species across landscapes.

Conclusions

To face the challenge of managing anthropogenic landscapes in a dynamic and increasingly accelerated context, an effective transdisciplinary integration has to be achieved by embodying the complexities of SELs into traditional landscape ecology analyses. In this context, the role of landscape ecology is to assess the impact of landscape heterogeneity combined with the impact of anthropogenic disturbances on ecosystem processes and the related delivery of ESs at multiple scales. In particular, there is a need to consider the complexities of anthropogenic landscapes in terms of order and disorder, where order implies causality, well-defined boundaries, and predictable outcomes, while disorder implies uncertain causality, shifting boundaries, and often unpredictable outcomes. The new methodological approaches can be used to increase spatially explicit anticipatory capability in environmental science and natural resource management, based on how the SELs have responded to stress in the past.

Such strategies could involve the design and management of landscape elements and structure through the strategic placement of managed land uses and natural ecosystems, so the services of natural ecosystems (i.e., pest control, pollination, reduced land erosion) can be maintained and even enhanced across the landscape. Additionally, these strategies should also consider landscape pattern design for the deliberate placement and confinement of local contagious disturbances that humans can manage at certain scale ranges to control, for instance, biological invasions, and to mitigate cross-scale impacts on ecosystem service flow.

These advancements should contribute greatly to strategies to foster SEL resilience in general and sustainable anthropogenic landscapes management in particular, and for the spatially explicit adaptive comanagement of ecosystem services.

See also: Conservation Ecology: Protected Area; Connectivity and Ecological Networks; Biotopes. Ecological Complexity: Citizen Science. Terrestrial and Landscape Ecology: Ecological Engineering: Overview

Further Reading

Alcamo, J., Bennett, E.M., 2003. Ecosystems and Human Well-being. Millennium Ecosystem Assessment (MA). Washington: Island Press.

Grossi, L., Zurlini, G., Rossi, O., 2001. Statistical detection of multiscale landscape patterns. Environmental and Ecological Statistics 8, 253–267.
Gunderson, L.H., Holling, C.S., 2002. Panarchy: Understanding Transformations in Human and Natural Systems. Washington: Islands Press.
Marwan, N., Kurths, J., Foerster, S., 2015. Analyzing spatially extended high-dimensional dynamics by recurrence plots. Physics Letters A 379, 894–900.
Marwan, N., Romano, M.C., Thiel, M., Kurths, J., 2007. Recurrence plots for the analysis of complex systems. Physics Reports 438, 237–329.
Marwan, N., Kurths, J., Foerster, S., 2015. Analyzing spatially extended high-dimensional dynamics by recurrence plots. Physics Letters A 379, 894–900.
Naveh, Z., Lieberman, A.S., 1994. Landscape Ecology: Theory and Application. New York: Springer.
Petrosillo, I., Semeraro, T., Zaccarelli, N., Aretano, R., Zurlini, G., 2013. The possible combined effects of land-use changes and climate conditions on the spatial–temporal patterns of primary production in a natural protected area. Ecological Indicators 29, 367–375.
Riitters, K., Costanza, J.K., Buma, B., 2017. Interpreting multiscale domains of tree cover disturbance patterns in North America. Ecological Indicators 80, 147–152.
Zaccarelli, N., Li, B.-L., Petrosillo, I., Zurlini, G., 2013. Order and disorder in ecological time-series: Introducing normalized spectral entropy. Ecological Indicators 28, 22–30.
Zurlini, G., Petrosillo, I., Aretano, R., Castorini, I., D'Arpa, S., De Marco, A., Pasimeni, M.R., Semeraro, T., Zaccarelli, N., 2014. Key fundamental aspects for mapping and assessing ecosystem services: Predictability of ecosystem service providers at scales from local to global. Annali di Botanica 4, 53–63.
Zurlini, G., Petrosillo, I., Jones, K.B., Zaccarelli, N., 2013. Highlighting order and disorder in social–ecological landscapes to foster adaptive capacity and sustainability. Landscape Ecology 28, 1161–1173.
Zaccarelli, N., Petrosillo, I., Zurlini, G., 2008. Retrospective analysis. In: Jørgensen, S.E., Fath, B.D. (Eds.), Encyclopedia of Ecology. Oxford: Elsevier, pp. 3020–3029.
Zaccarelli, N., Petrosillo, I., Zurlini, G., Riitters, K.H., 2008. Source/sink patterns of disturbance and cross-scale mismatches in a panarchy of social-ecological landscapes. Ecology and Society 13 (1), 26.
Zaccarelli, N., Petrosillo, I., Zurlini, G., 2008. Retrospective analysis. In: Jørgensen, S.E., Fath, B.D. (Eds.), Encyclopedia of Ecology. Oxford: Elsevier, pp. 3020–3029.
Zaccarelli, N., Li, B.-L., Petrosillo, I., Zurlini, G., 2013. Order and disorder in ecological time-series: Introducing normalized spectral entropy. Ecological Indicators 28, 22–30.
Zurlini, G., Petrosillo, I., Jones, K.B., Zaccarelli, N., 2013. Highlighting order and disorder in social–ecological landscapes to foster adaptive capacity and sustainability. Landscape Ecology 28, 1161–1173.
Zurlini, G., Petrosillo, I., Aretano, R., Castorini, I., D'Arpa, S., De Marco, A., Pasimeni, M.R., Semeraro, T., Zaccarelli, N., 2014. Key fundamental aspects for mapping and assessing ecosystem services: Predictability of ecosystem service providers at scales from local to global. Annali di Botanica 4, 53–63.

Buffer Zones☆

Jesper S Schou and Emilie W Hansen, University of Copenhagen, Copenhagen, Denmark
Peter Schaarup, Danish Forest and Nature Agency, Copenhagen, Denmark

Introduction

Buffer zones may serve a wide range of purposes covering from environmental to socioeconomic and military issues, but here the focus is on the use of buffer zones for environmental protection purposes. Basically, an environmental buffer zone serves the purpose of changing an environmental pressure and, thus, effect occurring in a recipient adjacent to the zone. Therefore, buffer zones are closely connected to environmental problems of a site-specific nature.

Usually, the designation of a buffer zone in itself does not lead to changes in environmental pressures although a number of studies indicate that buffer zones may represent biologically valuable habitats (e.g. Douglas *et al.*, 2009; Musters *et al.*, 2009). Therefore, an important part of defining buffer zones is to establish the specific regulation of activities within the zone. Thus, the designation of buffer zones can be structured into tree initial steps:

- define the policy target;
- establish the criteria for designation the zone; and
- define regulation within the zone

To support the policy choices an *ex ante* evaluation of different options should be performed, and the implementation should be followed by an *ex post* evaluation to evaluate the fulfillment of the policy target. Establishment of buffer zones in this article is seen as a policy strategy applied to realize a pre-specified environmental objective; thus, focus here is on issues 2 and 3.

Designating the Zone

The criteria for designating the zone are of cause closely connected to the policy target. Knowing the policy target, the first step is to establish which areas that should be targeted to the buffer zones. This may lead to two types of buffer zones: zones changing the pressures on a location inside or adjacent to the zone, or zones changing pressures within the zone. In the first situation it is generally also useful to define the area outside the zone to which the pressures should be changed. An example of the first type of buffer zone may be a riparian buffer zone alongside a stream serving to reduce the erosion and loss of nutrients and pesticides to the stream (see **Fig. 1**) or buffer zones along highways to reduce traffic noise in domestic areas. An example of the former type of zone could be a buffer zone along the coastline where domestic settlements are prohibited in order to secure the landscape.

As apparent from the examples the specific location of the zone is of importance. The simplest way to designate the zone is to appoint an area within a certain distance of the targeted location. However, natural conditions such as soil types, slopes, and the dominant wind direction (in case of airborne pollution) may be relevant to take into consideration when designating the zones. This implies that more complex criteria for designating the zone may be efficient. One example could be to design a buffer zone dependent on the wind direction frequency if the aim is to reduce ammonia depositions to specific nature locations. Another example could be varying the width of riparian buffer zones depending upon the erosion potential of the adjacent fields.

Regulation Within the Zone

The regulation applied within the buffer zone can either be mandatory, flexible, or voluntary. In the case of a mandatory regulation, this will typically consist of prohibition or restrictions on the activities within the zone. This implies that land owners within the zone are subject to limitations in the property rights or – more far reaching – are obliged to carry out certain nature preservation activities. One example is the National Danish Nature Protection Act that requires land use to be unchanged within designated locations characterized by specific nature types.

The term 'flexible regulation' refers to policy measures regulating environmental pressures through the market by application of taxes of tradable quotas. This regulation is difficult to target locally, and is thus, not likely to be a general feasible option within buffer zones.

In the case of voluntary regulation, the landowners are given the opportunity to engage in environment-friendly schemes within the designated buffer zones. The scheme may either provide the landowners with the opportunity to enter subsidy

☆*Change History*: November 2017. Revisions made by J Schou and EW Hansen. Revisions in Table 1 and added new section "Recent Developments".

This is an update of J.S. Schou and P. Schaarup, Buffer Zones, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 502–505.

 doi:10.1016/B978-0-12-409548-9.10927-3

Fig. 1 Example of the mandatory 2 m non-cropped buffer zone along streams and rivers in Denmark. Photo by J. S. Schou.

payments targeted at reducing environmental pressures within the buffer zone, for example, reduced pesticide use, or simply provision of advisory services targeted to landowners within the zone.

Mandatory, flexible, and voluntary regulation can of course also be applied in various combinations. One example could be ammonia buffer zones where farmers are faced with a mandatory requirement of reducing ammonia emissions to a pre-specified level, but at the same time are eligible to apply for subsidies to implementing ammonia abatement technologies.

Types of Buffer Zones and Their Function

Table 1 lists the most common types of environmental buffer zones described by their design and functions found in the international literature.

Recent Developments

Buffer zones have drawn attention both in research and in administrative organizations and NGOs. From the organizational side Cooper *et al.* (2009) reviews the use of buffer zones and other relevant measures for providing public goods from agriculture, and Dairynz (2012) describes the key benefits of managing waterways for the farm, the water quality and for the broader environment.

Turning to the research contributions, effects of riparian buffer strips on farmland biodiversity and ecological functions have received large attention. For example Cole *et al.* (2015) investigates the role of riparian buffer strips for the conservation of insect pollinators in intensive grassland systems and find that fenced riparian buffer strips have the potential to provide resources for insect pollinators. McCracken *et al.* (2012) investigated the connection between fencing off the margins of lowland grassland fields and the diversity of invertebrate species and find that fencing watercourses may increase the abundance of key groups of invertebrates. In both studies, the width of the buffer zones was found to be of importance. Brudvig *et al.* (2009) look at habitat corridors as important feature of reserves and demonstrate that corridors by increasing species richness inside target patches benefit biodiversity in surrounding non-target habitat. Musters *et al.* (2009) analyses the effect of taking field margins out of intensive cultivation and find that the plant species richness and species composition of the field margins changed over a four year period. The question if riparian ecosystems are particularly vulnerable to climate change are discussed by Capon *et al.* (2013) who find that the need for planned adaptation of and for riparian ecosystems is likely to be increased under a changing climate. Last, Douglas *et al.* (2009) investigated the effect of the height of the vegetation on margin areas for farmland birds and find that more frequent cutting and increased accessibility to the areas are likely to benefit a range of bird species.

At the more aggregated level Carswell *et al.* (2015) examines the potential trade-offs between carbon sequestration and biodiversity when selecting non-forested lands in New Zealand for natural forest regeneration. Further, a number for studies evaluate economic aspects of buffer zones. Roberts *et al.* (2009) estimates the annualized costs of establishing and maintaining riparian buffer strips on all agricultural land adjoining a waterway in Tennessee's Harpeth River watershed, and Laurén *et al.* (2007) investigates the interaction between the harvested stock in the buffer strips and the income for the landowner. The results further indicate that cost-effective water protection can be achieved when the dimension of buffer zone and the intensity of thinning are optimized. Last, using a newly developed indicator of the pesticide risk for aquatic biocenosis in

Table 1 Types of environmental buffer zones and their function

Type	*Design considerations and functional principles*
Riparian buffer zones	Zones with restrictions on land use or cropping technology on fields alongside rivers, streams, and lakes. Restrictions may imply a mandatory requirement of permanent grassland or no tilt from harvest until, e.g., April 1st. Primary function is to reduce erosion and nutrient losses. Secondary functions are to serve as habitats and establish connectivity between habitats.
Landscape buffer zones	Zones with restrictions on land use and mandatory rules of the future management. Primary function is to prevent new pressures within the zone and secure scenic views through the zone, e.g., by keeping the area free of domestic settlements or forest.
Wildlife buffer zones	Zones adjacent to or within appointed habitats with restrictions on land use or other human activities. Primary function is to reduce external pressures on the wildlife and secondary function is to improve the habitat and establish connectivity between habitats. Alternatively wildlife buffer zones can appoint areas where regulation of specific wildlife species is approved. An example could be a zone where predators such as wolfs feeding on domestic husbandry may be trapped or shot. Primary function is to secure domestic husbandry inside the zone and protect the predators outside the zone.
Ammonia buffer zones	Buffer zones where ammonia emissions are specifically regulated around specific habitats. The regulation designated to the buffer zones can either apply to future changes in husbandry production, e.g., establishment of new stables in the zone, or to existing and future husbandry production. The primary function is to reduce the local contribution to eutrophication of the appointed habitats. A secondary effect is likely to be reallocation of livestock production to outside the zone.
Pesticide free buffer zones	Zones where in use of pesticides in general or specific active substances are prohibited. The zones are typically located along field margins adjacent to hedgerows, streams, and similar landscape elements. Primary function is to increase the conditions for wildlife by increasing the production of forage such as weeds and insects. Further, the pesticide free buffer zones will reduce the external pressures on the adjacent habitats due to reduced wind drift.

surface water bodies' in application with an economic sector model alternative designs for an environmental policy programme is evaluated by Sieber *et al.* (2010). Results show that 3 m buffer strips perform best in a cost efficiency test due to low costs from changing agricultural land use.

Case Study – Ammonia Buffer Zones in Denmark

On 1 January 2007 a new integrated environmental accreditation scheme for all livestock farms was passed in Denmark. According to this scheme all farms with more than 75 animal units (one animal unit correspond to the nitrate production of one jersey cow) have to be approved based on their extra loss of nutrients when applying for an increase of the number of animal units. The integrated approach means that all environmental emissions have to be considered in the application including ammonia emissions from the stables and manure containers.

The general accreditation rule is that all the ammonia emissions from the new stables have to be generally reduced with more than 15% compared to the best available stable system. This general rule is supplemented by an individual regulation regarding the contribution of ammonia emission from the farms (if these are adjacent, i.e., situated within a buffer zone around) to the following types of nitrogen vulnerable areas: (1) raised bogs; (2) lobelia lakes; (3) moors larger than 10 ha and all moors in NATURA 2000 areas; (4) uncultivated, dry meadows larger than 2.5 ha and all uncultivated, dry meadows in NATURA 2000 areas; and (5) nitrogen vulnerable lakes in NATURA 2000 areas. Around these areas the regulation is divided into three parts:

- *Buffer zone I*. If just one of new or modified stable or manure container of the farm is situated less than 300 m from the vulnerable area no increased emission is approved.
- *Buffer zone II*. If just one of new or modified stable or manure container of the farm is situated less than 100 m from the nitrogen vulnerable area, a standardized emission calculation has to be carried out. The total contribution of ammonia deposition from the new or modified stables in the vulnerable area may not exceed 0.7 kg N per m^2. In order to take the cumulating aspects into account the maximum accepted increase in the ammonia deposition descends if another farm with more than 75 animal units is situated close to the nitrogen vulnerable area; if only one farm is situated closer than 1000 m from the new or modified stables then the contribution of ammonia from the new or modified stables in the vulnerable area may not exceed 0.5 kg N per m^2. If two or more farms with more than 75 animal units are situated closer than 1000 m from the new or modified stables, the threshold is 0.3 kg N per m^2.
- *Outside the buffer zones*. No individual regulation of the ammonia emissions from the farms under environmental approval.

The nitrogen vulnerable areas were pointed out by a national committee (the so-called Wilhjelm-udvalget). Afterwards the areas were adopted as part of the third Aquatic Action Plan, and the regulation related to the ammonia buffer zones was implemented as part of the new law on environmental approval of livestock farms as the primary basis for the individual regulation of ammonia emissions.

The scope for making this gradual regulation with two buffer zones is to make a differentiated incitement for the farms to locate new or extended stables at a proper distance from the vulnerable areas. Further, buffer zones are considered as a cost-effective way to focus the regulation in areas where the environmental effect is significant. Furthermore, one of the scopes is to ease the

administrative costs because only in buffer zone II the more complicated and costly calculations of emissions are needed. Furthermore, in order to reduce the administrative burdens and to ensure the quality of the applications all environmental calculations (including ammonia emission calculations) and other information needed for the application for an integrated environmental approval are integrated in a new internet-based digital application system (www.husdyrgodkendelse.dk).

See also: Ecological Data Analysis and Modelling: Carbon Biogeochemical Cycle and Consequences of Climate Changes; Biogeochemical Models. Ecosystems: Floodplains; Riparian Wetlands

References

Brudvig, L.A., Damschen, E.I., Tewksbury, J.J., Haddad, N.M., Levey, D.J., 2009. Landscape connectivity promotes plant biodiversity spillover into non-target habitats. Proceedings of the National Academy of Sciences 106 (23), 9328–9332.

Capon, S.J., Chambers, L.E., Mac Nally, R., Naiman, R.J., Davies, P., Marshall, N., Pittock, J., Reid, M., Capon, T., Douglas, M., Catford, J., Baldwin, D.S., Stewardson, M., Roberts, J., Parsons, M., Williams, S.E., 2013. Riparian ecosystems in the 21st century: hotspots for climate change adaptation? Ecosystems 16, 359–381.

Carswell, F.E., Mason, N.W., Overton, J.M., Price, R., Burrows, L.E., Allen, R.B., 2015. Restricting new forests to conservation lands severely constrains carbon and biodiversity gains in New Zealand. Biological Conservation 181, 206–218.

Cole, L.J., Brocklehurst, S., Robertson, D., Harrison, W., McCracken, D.I., 2015. Riparian buffer strips: Their role in the conservation of insect pollinators intensive grassland systems. Agriculture, Ecosystems and Environment 211, 207–220.

Cooper, T., Hart, K., Baldock, D., 2009. The provision of public goods through agriculture in the European Union, report prepared for DG agriculture and rural development. Contract No 30-CE-0233091/00-28 London: Institute for European Environmental Policy.

Douglas, D.J.T., Vickery, J.A., Benton, T.G., 2009. Improving the value of field margins as foraging habitat for farmland birds. Journal of Applied Ecology 46, 353–362.

Dairynz, 2012. Key benefits of managing waterways. In: Farm fact 5–1. New Zealand: Dairy NZ. http://www.dairynz.co.nz/publications/farmfacts/sustainable-dairying-land-and-water-management/farmfact-5-1/

Laurén, A., Koivusalo, H., Ahtikoski, A., Kokkonen, T., Finér, L., 2007. Water protection and buffer zones: How much does it cost to reduce nitrogen load in a forest cutting? Scandinavian Journal of Forest Research 22, 537–544.

McCracken, D.I., Cole, L.J., Harrison, W., Robertson, D., 2012. Improving the farmland biodiversity value of riparian buffer strips: Conflicts and compromises. Journal of Environmental Quality 41, 355–363.

Musters, C.J.M., Van Alebeek, F., Geers, R.H.E.M., Korevaar, H., Visser, A., De Snoo, G.E., 2009. Development of biodiversity in field margins recently taken out of production and adjacent ditch banks in arable areas. Agriculture, Ecosystems and Environment 129, 131–139.

Roberts, D.C., Clark, C.D., English, B.C., Park, W.M., Roberts, R.K., 2009. Estimating annualized riparian buffer costs for the Harpeth river watershed. Review of Agricultural Economics 31 (4), 894–913.

Sieber, S., Pannell, D., Müller, K., Holm-Müller, K., Kreins, P., Gutsche, V., 2010. Modelling pesticide risk: A marginal cost-benefit analysis of an environmental buffer-zone programme. Land Use Policy 27, 653–661.

Further Reading

Clarke, A.E., Wolf, T.M., Kuchnicki, T.C., François, D.L., Glaser, J.D., Hodge, V.A., 2004. Use of buffer zones for the protection of environmental habitats in Canada. Aspects of Applied Biology 71 (1), 133–139.

Ducros, C.M.J., Joyce, C.B., 2003. Field-based evaluation tool for riparian buffer zones in agricultural catchments. Environmental Management 32 (2), 252–267.

Ebregt, A., De Greve, P., 2000. Theme studies series 5: Buffer zones and their management. Wargeningen: National Reference Centre for Nature Management.

Hickey, M.B.C., Doran, B., 2004. A review of the efficiency of buffer strips for the maintenance and enhancement of riparian ecosystems. Water Quality Research Journal of Canada 39 (3), 311–317.

Martino, D., 2001. Buffer zones around protected areas: A brief literature review. Ottawa: Carleton University.

Roe, J.H., Georges, A., 2007. Heterogeneous wetland complexes, buffer zones, and travel corridors: Landscape management for freshwater reptiles. Biological Conservation 135 (1), 67–76.

Schou, J.S., Tybirk, K., Hertel, O., Løfstrøm, P., 2006. Economic and environmental analysis of buffer zones to reduce ammonia loads to nature areas. Land Use Policy 23, 533–541.

Viaud, V., Merot, P., Baudry, J., 2004. Hydrochemical buffer assessment in agricultural landscapes: From local to catchment scale. Environmental Management 34 (4), 559–573.

Classical and Augmentative Biological Control

RG Van Driesche and K Abell, University of Massachusetts, Amherst, MA, USA

Glossary

augmentative biological control Programs based on the release of commercially reared parasitoids or predators for pest control in crops.

classical biological control Programs of importation of natural enemies from an invasive pest's homeland for the permanent, area-wide suppression of the pest; may be directed against either economic pests of crops or ecological pests of natural areas; may target pest insects, mites, or weeds.

hyperparasitoid A parasitoid that attacks another species that is itself a parasitoid; these species may be detrimental to insect biological control programs.

inoculative biological control A form of augmentative biological control in which the goal is to establish reproducing populations of natural enemies at the start of the crop.

inundative biological control A form of augmentative biological control in which natural enemies are released in large numbers throughout the cropping period, with no expectation that the released biological control agents will establish; pest control is expected from the individuals actually released, not their progeny.

insectaries Rearing facilities in which parasitoids and predators are mass produced, usually for sale as agents for augmentative biological control programs.

natural control The level of control produced by natural enemies that occur naturally in the crop without any active management by people.

nontarget impact Mortality or population or range decreases that may occur to beneficial or native species as a consequence of the use of natural enemies in biological control programs.

parasitoids Insects that develop in or on a host, consuming and killing it; parasitoid larvae require one host to complete development; adult females can also kill hosts through host feeding.

predators Organisms that attack and eat other organisms. Predators typically kill and consume many prey during their development as larvae and as adults.

quarantine A secure importation facility designed to prevent the unintentional release of natural enemy species into new geographic regions after their discovery during foreign exploration.

Definition and Scope of Biological Control

Biological control is a form of pest control that uses living organisms (parasitoids, predators, or herbivorous arthropods) to suppress a pest's density to lower levels. There are four kinds of biological control, two of which – classical biological control and augmentative biological control – are discussed in this article and two others – conservation biological control and biopesticides.

Classical biological control is the deliberate importation and release of new species of natural enemies with the intention of suppressing the densities of a target weed or insect permanently over the whole of its range in the country receiving the natural enemies. Target pests are typically invasive species and the introduced natural enemies are those specialized agents that attack it in its native range. Classical biological control is a major tool in reducing impacts of invasive species, both in crops and natural areas.

Augmentative biological control is pest suppression in greenhouses or outdoor crops through the purchase and release of commercially reared natural enemies. This approach is used against pest insects and mites, but not against weeds. Natural enemies used include both predators and parasitoids. The value of this approach depends on efficacy and cost of the natural enemy. This approach is most effective in greenhouse crops.

Biological control has important advantages compared to other methods of pest control. Classical biological control is often cheaper and less polluting than use of pesticides, because pest control is relatively permanent and does not require annual retreatment. Initial costs of classical biological control are high, for discovery, importation, testing, and initial release of new natural enemies. However, costs drop to low or even zero levels in later years, while the benefits of the pest control achieved continue to accrue for years. For augmentation biological control, results are temporary and costs reoccur annually, as with pesticides. Augmentative biological control may be either more or less expensive than other approaches depending on details such as the cost of natural enemy production by commercial insectaries that sell beneficial organisms, and efficacy of other control tactics. Both of these types of biological control are harmless to people and vertebrates, giving the approach a distinct advantage over pesticides, which must be actively managed for safe use to mitigate harm to humans and other nontarget organisms. Risks of both augmentative and classical biological control to nontarget insects or, for weed biocontrol projects, to plants can exist, but these risks can be managed to low levels by careful screening of species being considered for release in new areas. Biological control as a scientific endeavor has a history of about 125 years of effective use (beginning in the 1880s), over which time new information, techniques, and technologies have increased our ability to use biological control agents with increasingly greater understanding and effectiveness.

 doi:10.1016/B978-0-444-63768-0.00039-1

Classical Biological Control

Ecological Justification

People routinely move species, such as crops and ornamental plants, across natural barriers such as mountain ranges or oceans that would otherwise limit their spread. These plants may carry with them small, unrecognized infestations of pest insects. In some cases the plants themselves may spread and become invasive. Both invasive plants and insects often escape their specialized natural enemies when they cross geographic barriers and establish in new locations. Local natural enemies, unfavorable habitats or climates, or, for insects only, lack of a susceptible local plant, prevent many invaders from establishing or reaching high densities. However, for some invasive species, local climates and hosts are favorable and local natural enemies are generalists that have limited impact. Those species increase to densities from 10 to 10 000 times greater than their numbers in the native range. Such high-density invasive species cause great economic and ecological damage and are the targets of classical biological control. By introducing more effective natural enemies, classical biological control seeks to lower the invader's density, which then allows the invaded community to return fully or partially back to its preinvasion condition. Two examples follow that illustrate the process and how such projects produce economic and ecological benefits.

Pink Hibiscus Mealybug

Pink hibiscus mealybug (*Maconellicoccus hirsutus* Green) (Fig. 1) invaded the Caribbean nation of Grenada in *c.* 1993, where it infested young shoots, flowers, and fruits of a wide range of plants, particularly those in the family Malvaceae. Among the important plants affected were ornamental hibiscus (*Hibiscus rosa-sinensis* L.), soursop (*Annona muricata* L.), cotton (*Gossypium hirsutum* L.), cocoa (*Theobroma cacao* L.), and citrus (*Citrus* spp.). The mealybug reached high densities and began to spread rapidly to other islands and adjacent mainland areas. Mealybugs caused immediate losses to the tourist industry by reducing beauty of ornamental plants around hotels. Losses also occurred in several major crops and inter-island trade was affected through the quarantines (albeit ineffective ones) enacted to prevent spread to other islands. Grenada and the islands of Trinidad and Tobago

Fig. 1 Pink hibiscus mealybug (*Maconellicoccus hirsutus* Green) invaded the Caribbean region in the 1990s and rapidly became an important pest on a wide range of woody plants. Photo Courtesy of Dale Meyerdirk, USDA; Forestryimages.org.

suffered an estimated $US10–18 million in losses in the first year following the mealybug invasion. In contrast, Puerto Rico, where effective parasitoids were rapidly introduced almost immediately after an invasive population was discovered, suffered almost no economic losses.

Control of this pest was greatly facilitated by previous successful control of the same species in Egypt in the 1920s, where it had invaded, presumably from India. Several predatory coccinellids, including the mealybug destroyer, *Cryptolaemus montrouzieri* Mulsant, and the encyrtid parasitoid *Anagyrus kamali* Moursi, were introduced into Egypt. In the Caribbean, both *C. montrouzieri* and *A. kamali* were also introduced, as well as another encyrtid parasitoid, *Gyranusoidea indica* Shafee, Alam, and Agarwal. The predatory coccinellid had little or no effect, even though *C. montrouzieri* did establish. Costly releases of *C. montrouzieri* were somewhat useful as a stop gap measure to reduce extremely high populations on limited areas of high value plants but could not provide control over wide areas. In contrast, region-wide control was rapid and complete following the release of several parasitoids, mainly *A. kamali*.

To gauge the specificity of *A. kamali*, nine species of mealybugs were assessed. Of the species tested, *A. kamali* parasitized only two nontarget species, but development in these species was not successful. This parasitoid was, therefore, judged to be relatively specific and beneficial. In contrast, *C. montrouzieri*, which was also introduced to Grenada (but not to most other locations invaded later), is a known generalist predator.

This project provided rapid, permanent, and complete control of an invasive economic pest in many Caribbean nations and the United States. Because the pest had been controlled previously in Egypt, the natural enemy likely to be effective was already known. Because the introduced parasitoid was specific to mealybugs, it was safe to nontarget insects in other groups and host range tests showed it unlikely even to affect other mealybugs.

Waterhyacinth

Waterhyacinth (*Eichhornia crassipes* [Mart.] Solm) (**Fig. 2**) is both a plant used in ornamental fish ponds and the world's most damaging aquatic weed. Its beautiful lavender flowers have led people to take it far from its native range in the Amazon basin of South America. Wherever waterhyancinth has been introduced into subtropical or tropical climates, it has escaped into the wild. There it forms gigantic mats that clog rivers and cover over bays and ponds, causing great ecological damage and harming economic activities such as fishing, transportation, and hydropower. Among the many places invaded by waterhyacinth is Lake Victoria in East Africa. It was first recorded there in 1980 and by the mid-1990s some 12 000 ha of weed mats were clogging bays and inlets around the lake. Economic losses were caused to fisheries (the mats impede the launching of boats and the use of nets) and to water and hydroelectric power works. Ecologically, the weed threatened one of evolution's greatest products – a radiation of some 200–400 species of endemic cichlid fishes. These fish, often separated by mating habits based on female color preference, were threatened by hybridization among species induced by low light under weed mats, where color-based visual-recognition mating systems could not be sustained.

Control efforts recommended to the governments of the affected countries (Uganda, Kenya, and Tanzania) included herbiciding the mats, use of harvester boats to cut the mats, and release of specialized herbivorous insects. Two weevils, *Neochetina eichhorniae* Warner and *N. bruchi* Hustache, known to be specialists on waterhyacinth from earlier work in Florida, were chosen for release. In 1995, Uganda was first to release biological control insects against the weed, followed by the other two countries in 1997. On the Ugandan shore, weed mats began to show damage from herbivory by late 1998. By 1999 some 75% of the mats had died and sunk into the lake.

This project illustrates the power of classical biological control to provide control of invasive species damaging large natural systems without the recurrent costs of mechanical control or the pollution of pesticides.

Fig. 2 Waterhyacinth (*Eichhornia crassipes* [Mart.] solms) has invaded water bodies throughout the tropics and subtropics, causing great ecological and economic damage. Photo Courtesy of Chris Evans; Forestryimages.org.

Description of the Process

Regardless of the target pest, a similar series of steps are followed in any classical biological control project.

Choice of the target pest

Species chosen as targets of importation biological control should be ones for which there is broad social agreement that they are pests and need to be reduced in density. Targets should be species that are strongly regulated by natural enemies in their native ranges and these species should be missing in the areas invaded by the pest. Nontarget risk to native plants can be minimized if species selected as targets do not have closely related (same genus) species in the recipient country.

Pest identification and taxonomy

Correct identification of the target pest is essential. Mistakes at this stage will cause project delays or failure. If the pest is an unknown species, its nearest relatives need to be identified, as this can provide clues to the pest's native range. Molecular markers are now commonly used to identify populations of both pests and natural enemies with increased precision. Such identification for pests allows exact origins of an invader population to be identified, often within very large native ranges. Markers for natural enemies allow biotypes or cryptic species to be recognized and facilitates identification of recovered specimens after release.

Identification of the native range

The region where the pest evolved needs to be identified as it is the best place in which to search for specialized natural enemies that have evolved with the pest. This can be done based on several criteria, including the center of the geographic range of the pest, the area where the principal host plant of the pest evolved, regions where the pest is recorded to occur, but remains at low densities, and regions with the largest numbers of species closely related to the pest.

Surveys to collect natural enemies

Natural enemy collection, or foreign exploration, needs to be done extensively over the range of locations, habitats, and seasons where the pest is found naturally. Surveys of natural enemies in the invaded area are unlikely to locate effective natural enemies, but are needed to identify any natural enemies that may already be present because of their own invasion of the region.

Importation to quarantine

Promising natural enemies collected in surveys need to be shipped to quarantine laboratories, where they can be colonized and maintained on the pest for further study to evaluate their host ranges and make a judgment as to their safety for introduction into the proposed recipient country.

Host specificity and biology studies

To promote selection of safe species for importation, the biology and degree of host specificity of each candidate biological control agent needs to be determined through a mixture of literature records, field observations in the area of origin, and laboratory host range studies (feeding and oviposition evaluations with species potentially at risk) in quarantine before release into a new area is approved.

Release and colonization in the field

Releases need to be made at numerous locations where the target pest is present, and over extended periods of time until efficient means to establish the natural enemies in the invaded area are discovered or until it is clear the agents are unable to establish. Once established, natural enemies disperse naturally or are redistributed artificially throughout the range of the pest.

Evaluation of efficacy

Field experiments in the invaded area comparing pest density in plots having and lacking the introduced natural enemy are needed to measure the degree to which the natural enemy is able to reduce the density of the pest (for insect targets) or, for weeds, how much the natural enemies lower a variety of plant performance and population measures such as percent cover, biomasss, seed set, etc.

Documentation of benefits

Economic and ecological consequences of the project need to be recorded and published.

Extent of Successful Use

Following introductions of natural enemies, pest densities may be reduced, in some cases by 90%–99% or more. This has been achieved for a variety of pest insects, including caterpillars, sawflies, aphids, scales, whiteflies, and mealybugs. Over the last 125 years, more than 1200 insect biological control projects have been attempted. Of these, 60% have resulted in a reduction of the pest's density. In 17% of cases, control was complete and no further effort was needed. Introductions of specialized herbivores have been made for over 133 species of invasive plants and, of these, 41 species (31%) have been completely controlled.

Economics

Importation of biological control agents is a government activity for the common good. Funds for such work are typically provided by governments, but may in some cases come from grower organizations representing particular crops in a region. Costs of projects are concentrated at the beginning of the work, as costs to search for and study new candidate natural enemies are high. Use of proven biological control agents in newly invaded locations as the pest spreads is cheaper, as much of the initial work will not have to be repeated and known natural enemies can quickly be introduced. Benefits of successful projects accrue indefinitely into the future and benefit-to-cost ratios of past projects have averaged 17:1, with some projects having much higher ratios, of 100 or even 200:1. In successful programs, control is permanent and does not require continued annual investments to sustain the benefits, in contrast to other forms of pest control (e.g., pesticide applications). This makes the method particularly attractive for use to protect natural areas and to protect crops in countries with resource-poor farmers. Biological control also promotes good environmental stewardship of farmlands in developed countries.

Safety of Natural Enemy Importations

Insects may be released as natural enemies of either invasive plants or invasive insects. Both biological weed control and biological insect control show a very high level of safety to vertebrate and human health. There are three safety issues when insects (herbivores, predators, or parasitoids) are imported to a new region: (1) identification of unwanted contaminants in foreign shipments; (2) recognition of organisms that, by virtue of their biology, may be damaging to other biological control agents; and (3) potential damage to nontarget species (e.g., native insects or plants) in the area of release by natural enemies with broad host ranges.

The first two of these safety concerns are addressed by the use of quarantine facilities, which are designed to prevent the accidental release of new species into the environment following importation. In quarantine, desired natural enemies are separated from all other materials, including miscellaneous insects that might have been accidentally included in the package by the collector, extraneous plant material, or soil. A taxonomist then confirms the species identification of the organism and ensures that all individuals collected are the same species. Voucher specimens are deposited with an entomological museum for future reference. Natural enemy identification will either indicate the name of the organism or that it is a species not yet described. New species can usually be placed in a known genus, for which some biological information will exist. A sample of the natural enemies are also submitted to a pathologist to determine if they carry any microbial or nematode infections. If they do, they will either be destroyed or, if possible, treated with antibiotics to eliminate the infection. This group of field-collected, healthy individuals will then be bred in the laboratory on the target host. This eliminates any undesirable parasitoids (for herbivores attacking weeds) or hyperparasitoids (for insect agents) that might exist in the collected material that, if established, would damage the biological control project by reducing the efficacy of imported natural enemies. For insect parasitoids, rearing for one generation on the target host excludes the possibility that a hyperparasitoid has been obtained by mistake.

The potential for attack on nontarget species after release is minimized by estimating the host range of the natural enemy proposed for release and reviewing that information in light of the fauna or flora in the recipient country. Estimation of an agent's host range is based on (1) literature records of species known to be attacked by the agent in the region from which it is collected, (2) negative evidence in the literaure, that is, any species of interest that occur with the agent in its home range but which are not attacked, and (3) data from laboratory tests. Most evidence comes from laboratory host range tests. For herbivorous insects released for weed biological control, these laboratory tests include studies of both the adult's oviposition preferences and the feeding preferences of the immature stages and, in some cases, adults. For immature stages, tests also include an assessment of a test plant species' suitability to support growth and development. Similar tests can be applied to the study of parasitoids, that is, both oviposition preferences and survival of the immature stages on a given host. For predators, feeding preferences of both adults and larvae must be measured.

Historically, estimation of host ranges of herbivorous insects used against weeds began early (in the 1920s), evolving from testing only local crops, to a phylogenetically based attempt to define the limits of the host range by testing plants in the same genus as the target weed, then the same tribe, etc. This process has been highly successful in avoiding the introduction of insects whose host ranges are wider than initially thought. Cases of attack of introduced herbivores on nontarget plants have largely been limited to other species in the same genus and have largely been well predicted. Of 117 species introduced into North America, Hawaii, or the Caribbean for biological weed control, only one species (the lacebug *Teleonemia scrupulosa*, introduced into Hawaii in 1902 against the shrub *Lantana camara* L.) has attacked nontarget plants that were not either in the same genus as the target weed, or a very closely related genus.

Estimation of host ranges of parasitoids and predators introduced for biological control of insects began in the 1990s, in response to changing views on the ecological and conservation value of native nontarget insects. Techniques for making estimates of arthropod natural enemy safety are less well developed than for herbivorous biological control agents. A few cases of harm from parasitoids or predacious insects to nontarget insects have been reported. Importation of generalist species that have broad host ranges should be avoided because of such potential to harm native insects.

Laws explicit to biological control importations exist principally in New Zealand and Australia. Laws in the United States regulate importation of herbivorous insects used against weeds, but do not currently regulate importation of parasitoids or predators.

Augmentative Biological Control

History and Scope

Methods have been developed to rear a variety of species of predators and parasitoids at commercial levels. Augmentative biological control is based on the user purchasing and releasing the natural enemies needed for his crop. This approach to pest control began in greenhouse-grown tomatoes with *Encarsia formosa* Gahan, a parasitoid of the greenhouse whitefly (*Trialeurodes vaporariorum* [Westwood]), which was first reared commercially by English growers in the 1920s. Modern augmentative biological control began in the 1970s when Dutch greenhouse tomato growers revived *E. formosa* rearing as a commercial activity because whiteflies in their greenhouses had developed pesticide resistance. From 1970 to 2006, the number of commercial insectaries producing parasitoids and predators for pest control grew to several dozen firms, which collectively produce about 100 species of natural enemies for use in greenhouses or related facilities. *Encarsia formosa* and the predatory mite *Phytoseiulus persimilis* Athias-Henriot make up most of the sales. Natural enemy releases are used in greenhouses, plant conservatories, mushroom houses, and animal holding buildings such as dairies, hog rearing facilities, poultry barns, and zoos.

Outdoor releases of several species of predators and parasitoids developed independently of the indoor applications described above and mainly focused on egg parasitoids in the genus *Trichogramma* (Hymenoptera: Trichogrammatidae). These parasitoids are released to suppress pest weevils and caterpillars in cotton, corn and sugarcane, especially in China, Russia, and tropical sugar-producing countries. Predators of mealybugs for release on citrus crops in parts of California have been reared by a growers' cooperative since 1926. Several species of predatory phytoseiid mites are released for control of pest spider mites in strawberries, outdoor foliage plant production, and other high value crops.

There are two approaches to augmentative biological control: inoculative and inundative releases. Most indoor releases of natural enemies are inoculative, consisting of a small release early in the crop cycle that is intended to suppress the pest after the natural enemy's numbers have increased naturally through reproduction in the crop. Cost of this approach is minimized because smaller numbers of the natural enemy are needed. In contrast, inundative biological control consists of frequent, large releases throughout the crop cycle, with control coming from the attacks of the released individuals. Little reproduction in the crop is expected. Because much higher numbers are released, only natural enemies with very low production costs are economical for use in this way. This is the approach behind *Trichogramma* releases, given that these parasitoids can be reared very inexpensively.

How Insectaries Turn Natural Enemies into Mass Market Products

To profitably market a natural enemy, an insectary must succeed in a series of activities.

Find a suitable natural enemy

Augmentative biological control starts with the discovery of a natural enemy that research suggests may be effective. The natural enemy must attack an important pest efficiently, be easily reared under mass production conditions, able to survive shipping, and be competitive in price with other forms of pest control available to growers.

Develop a mass rearing system

Some species, such as the whitefly parasitoid *E. formosa*, can be reared cheaply using their natural host on a living plant. In other cases, costs of production or the scale of production are improved by switching to an alternative rearing species other than the target pest. Most *Trichogramma* wasp species are grown on the eggs of moths that feed on stored grain, rather than on eggs of the target moths themselves, because colonies of grain-feeding moths can be reared much more cheaply.

Develop efficient harvest, storage, and shipping methods

Mass reared predators and parasitoids must be released within days of production. Shipping to customers must be rapid (1–3 days) and avoid delays at international borders. Longer delays invariably kill natural enemies due to heat, desiccation, continued development, or starvation.

Provide clear release instructions

The customer must release the natural enemy using the right rate and the correct procedure. Effective rates are discovered by controlled trials in universities and government laboratories, and by experience of growers using products in accordance with advice from producers.

Extent of Successful Use

Indoor crops

The use of augmentative biological control has become widespread in greenhouses in Northern Europe and Canada that produce vegetables, with over 5000 ha using *E. formosa* for whitefly control and over 7000 acres using *P. persimilis* for spider mite control. These amounts, however, still represent only a small percentage of the world's protected culture. Biological control is used much less often in Southern Europe and Japan, areas with extensive greenhouse vegetable production, because of differences in climate

and greenhouse construction. Development of methods effective in warm climates, however, is underway. Similarly, use of biological control is very limited in greenhouses producing bedding plants or floral crops, the major focus of greenhouse production in the United States.

Outdoor crops

The scientific use of augmentative natural enemy releases in outdoor crops is best established in Northern Europe for control of European corn borer (*Ostrinia nubilalis* [Hübner]) in corn. Use is greatest in Germany and France, with about 11 000 acres being protected annually in Germany with *Trichogramma* releases. This is, however, only a small fraction of the total corn acreage in Europe, and use of biological control is concentrated principally where pesticide use is not allowed because of concern for health of people living near corn fields and is supported by a government subsidy. Natural enemy releases for mite control has been successful in strawberries in California, Florida, and the northeastern United States, and in outdoor shade houses used for production of foliage plants in Florida. In Mexico, Russia, China, and other countries, large-scale releases of *Trichogramma* spp. have been made for a variety of moth and beetle pests of corn, sorghum, and cotton, but the efficacy of these releases has not been well demonstrated. In some instances, these activities have been state-supported and their actual economic value for pest control is not clear.

Safety of Augmentative Biological Control

Release of parasitoids and predators replaces pesticide application and thus enhances human safety. For workers in insectaries, handling of large quantities of insects or mites is an allergy risk. Where problems arise, risk can be reduced through air exchange or filtration to reduce concentrations of airborne particles and use of gloves and long sleeved shirts to reduce skin contact with arthropod body fragments. Risk to native species posed by releases of non-native natural enemies can be of concern in some instances. Generalist, non-native species released in large numbers may establish outdoors if climates are permissive and attack or suppress populations of native species, or reduce densities of native natural enemies through competition for resources. Consequently, some governments, such as those of Hawaii, Australia, and New Zealand, restrict importation of natural enemies used in augmentative biological control. Importation of North American green lacewing species (Neuroptera: Chrysopidae, *Chrysopa* spp.) (used in greenhouses as predators of aphids) might, for example, lead to the establishment of such species in the wild, increasing competition with the endemic native lacewings in Hawaii, which have conservation value as unique native wildlife.

Conclusion

Classical biological control is the dominant form of biological control, being suitable for use on thousands to millions of acres of land and is a critical tool in controlling the ecological and economic damage caused by invasive pests. It is a government-supported activity for the public good and is based in ecology and population dynamics. It has an excellent, but not perfect, safety record, and methods exist for predicting risks to native species, allowing them to be minimized by choice of agents released. Augmentative biological control, in contrast, is a commercial activity with a narrow scope. It is suitable for providing crop protection for limited periods of time. Its costs recur with each crop and must be paid by the end user. The range of natural enemies reared commercially is a small subset of those that occur naturally, being limited by economics of production and market size, such that only species that generate profits can be offered for sale. Both methods improve human and ecological health by replacing pesticide use.

Further Reading

Bellows, T.S., Fisher, T.W. (Eds.), 1999. Handbook of Biological Control: Principles and Applications of Biological Control. San Diego: Academic Press.
Clausen, C.P. (Ed.), 1978. Agricultural Research Service: Handbook No. 480: Introduced Parasites and Predators of Arthropod Pests and Weeds: A World Review. Washington, DC: USDA: Agricultural Research Service.
DeBach, P., Rosen, D., 1991. Biological Control by Natural Enemies. Cambridge, UK: Cambridge University Press.
Follett, P.A., Duan, J.J. (Eds.), 2000. Nontarget Effects of Biological Control. Boston, MA: Kluwer Academic.
Jervis, M., Kidd, N. (Eds.), 1996. Insect Natural Enemies: Practical Approaches to their Study and Evaluation. London: Chapman and Hall.
Julien, M.H., Griffiths, M.W. (Eds.), 1998. Biological Control of Weeds, a World Catalogue of Agents and their Target Weeds, 4th edn. Wallingford: CABI Publishing.
Van Driesche, J., Van Driesche, R.G., 2000. Nature Out of Place: Biological Invasions in a Global Age. Washington, DC: Island Press.
Van Driesche, R.G., Hoddle, M., Center, T., 2008. Control of Pests and Weeds by Natural Enemies, an Introduction to Biological Control. London: Blackwell.

Ecological Engineering: Design Principles☆

Susan Bolton, University of Washington, Seattle, WA, United States

Introduction

Engineering is the application of science through design to create systems to benefit humans. Design is the essence of engineering. Engineering has its basis in math and physics, but subfields are based on a particular science. For example, chemical engineering is based on chemistry, mechanical engineering on mechanics, and electrical engineering on electricity. The concepts of ecological engineering were introduced in the United States in the 1960s by H.T. Odum. Ecological engineering, as the name implies, is an engineering subfield that is based on ecology. Ecology includes aspects of all of the sciences that study living or nonliving components in an ecosystem, for example, biology, botany, geology, hydrology, soil science, zoology, and specifically addresses the interactions among the living and nonliving components of the ecosystem. Historically ecologists have viewed many engineering projects as destructive of natural systems making ecological engineering a contradiction in terms. Likewise, engineers often have little appreciation for ecological knowledge, which is usually less precise and mathematical than traditional engineering science.

Ecology is the study of the interrelationships between biotic (living or previously living, e.g., plants, animals, carcasses) and abiotic (never living, e.g., water, sediment, chemicals, temperature) components of the environment. Ecological engineering incorporates elements of the sciences used in ecology to create engineered designs that reflect and incorporate ecological processes. The goal is to provide for human welfare with engineering projects while also protecting the goods and services that are provided by a natural environment. These goods and services include production of oxygen, air and water purification, carbon storage, flood control, regeneration of soil and soil fertility, pollination of food crops, waste decomposition and protection from ultraviolet rays. Recognizing that all social and economic systems depend on a functioning ecological system implies that ecological engineers acknowledge the values of sustainability and protection of natural systems even as they design systems for the benefit of humans. These concepts define an engineering discipline based on ecological science with an explicit recognition that the values of sustainability and protection of natural systems are incorporated in designs for the benefit of humans and the environment.

Increasingly, natural scientists with no training in design methods are engaged in applied science as they address and try to solve environmental problems such as wetland loss, river and water quality degradation, and soil contamination. Engineers are addressing similar questions with more formal design procedures but with little training in the relevant scientific areas. This can create a variety of unintended consequences that can diminish the ecosystem's ability to provide the goods and services upon which all life on earth depends. Ecological engineering uses ecological knowledge and theory and standard engineering design procedures to address environmental problems. Standard design procedures allow for the collection of information on which design criteria are successful and which are not. Documentation of the design process allows for others to learn from either design errors or less than perfect designs and contributes to improved future designs. Numerous authors have discussed design principles for ecological engineering all of which derive from the overarching principles of thermodynamics and evolution.

Overarching Principles

Thermodynamics and Conservation of Mass and Energy

Ecological engineering principles are constrained by the laws of conservation of mass and energy and the laws of thermodynamics, just as chemical, mechanical, or electrical engineering principles are constrained by these laws. Ecosystems are open systems that require a continual input of energy to maintain their structure and function. Two of the most inviolable principles of ecological science can be described as energy flow and material recycling.

Energy flow

Energy inputs, driven by solar radiation, are required to maintain structure and function in the face of the physical tendency toward disorder (the increase of entropy). Traditionally engineered systems use human and hydrocarbon-based energy to maintain order (keep the system intact and functioning). Ecosystems use photosynthesis, driven by solar energy, as their energy source. Biological energy flow can be measured by rates of production (biomass accumulation) and respiration (energy used for production). Physical energy flow can be measured by the mobilization, transport and deposition of organic and inorganic materials by the kinetic and potential energy of fluids or solids such as water, wind, and sediment. Both biological energy and physical energy are constrained by

☆*Change History*: February 2018. Susan Bolton updated the keywords, updated abstract, added content on ecosystem services and issues related to climate change to Section 2.2.2, added an example of project types to Section 3, added newer references, and added web site links.

This is an update of S. Bolton, Design Principles, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 898–902.

conservation of mass and energy laws. Ecology, as the interaction of biotic and abiotic processes looks at the interactions of both types of energy. Some energy is lost at each transformation so while total entropy increases in accordance with the second law of thermodynamics, order is locally increased. This has been described as the self-organization feature of ecosystems or exergy.

Emergy, an accounting system developed by H.T. Odum can be used to put all natural and human production into common units based on solar radiation. Emergy measures the inputs to make a product or service. It is a measure of energy used in the past and thus is different from a measure of current energy use. This provides a way to evaluate the costs of ecological goods and services in the same units as the costs of human production of goods and services. Ecological engineering designs seek to maximize the use of renewable energy (e.g., solar radiation) and minimize the use of nonrenewable energy.

Material recycling

Nutrient and material (re)cycling is another major ecological principle. Material is conserved by the continual reuse of materials and the transfer of those materials between organic and inorganic states through biogeochemical cycles. Organic and inorganic materials cycle through the system appearing in different locations and forms through time. Waste disposal is seldom an issue in a functioning ecosystem as the output from one system is used as input to another. Natural biogeochemical cycles mobilize, transport, and store material in the atmosphere, biosphere, hydrosphere, and lithosphere. Producers, consumers, and decomposers transfer organic matter and nutrients among themselves and the storage compartments. Many traditional human engineering designs lead to the accumulation of waste materials that cannot be reused by the original process and can contaminate other processes. Ecological engineering designs seek to minimize waste production and to utilize wastes (material not related to the primary function of the design) as inputs for other processes. One example of this is using ecological processes to clean up waste products such as using wetlands to treat wastewater or phytoremediation to clean up soil contamination.

Natural Selection and Evolution

Self-organization

In traditional engineering design, one seeks to maintain the independence of functional requirements. Functional requirements are the specific functions that the design feature is to provide. The engineer selects specific physical elements (design features) to meet each functional requirement. In the final design, each functional requirement has one solution and does not rely on other design features. That is, modification of any one design parameter affects only one functional requirement. There is no interaction or coupling of design features to multiple functional requirements. This is exemplified by the concept of modularity in software design where individual modules can be swapped in and out of programs without affecting program stability or performance.

Engineering designs seek tight tolerances and rigid, stable systems that do not change. Nothing is left to nature; everything is preplanned. In the event of system or component failure identical back-up systems or components may be built to provide redundancy. The traditional engineering approach of maintaining independence between subsystems ignores the interrelationships and complexity of ecosystems.

Nature does not build things the way humans do. There is no external plan, design engineer, or architect that creates steps to achieve an envisioned final product. It is the flow of energy and material through an open system that allows for self-organization. Ecosystems are complexly coupled, that is, everything is connected to everything else. Ecosystem structure and function can be maintained through many different pathways that can operate under varying conditions, that is, wide tolerances. For example, accumulation of biomass can be accomplished by a wide array of plant and animal species. The actual species that live in a given ecosystem are adapted to the abiotic and biotic environment of the system. The loss of one species, either through succession or extinction, does not destroy the ecosystem. Other species continue the flow of energy and the cycling of nutrients and materials. In other words, there is redundancy of function but not necessarily of structure.

One measure of self-organization is the amount of information that is required to predict the final outcome. Higher levels of self-organization require large amounts of information to predict the final composition. To predict what an office building will look like when finished requires the blueprint and some drawings; to predict which organisms will provide structure and function to an ecosystem at any given time is almost impossible except in general terms. Self-organization also gives ecosystems properties of robustness, persistence, the ability to self-repair, and flexibility in the face of changing environmental conditions. Because of the interactions among system components, biotic and abiotic, changes in one component can ripple through the system and change many other components.

Self-organization is visible as an ecosystem goes through successional stages. It derives from the overarching actions of natural selection and evolution through which organisms have adapted to certain physical, chemical, and biological environments over time. For example, bare rock exposed by glacial retreat will pass through a series of plant and animal communities composed of different combinations of species over time as it develops into a forest. The general progress of this succession is understood, but predicting the exact composition of each step and the duration of each step is not. The exact composition is an emergent property of the interaction between the biotic and abiotic components of the ecosystem.

This ability of self-organization is taken advantage of by ecological engineers. Ecological engineering designs work with nature and allow nature to do some of the "engineering." That is, rather than fully proscribing only one satisfactory end result, the ecological engineered design recognizes that more than one final state may meet the functional requirements of the design.

Engineering designs that ignore the self-organization properties of ecosystems require continued inputs of human-based energy and dollars (to buy materials and energy) to keep the system in the designed state (the desired, predicted outcome). Allowing nature to finalize the outcome of the design uses solar energy to organize the system and ensures some flexibility in the face of changing conditions.

Ecological engineers can take advantage of ecosystem self-organization in various ways. For example, a wetland or streambank restoration project may plant a variety of species that are water tolerant. But rather than insisting that the original mix is the best and inputting labor and energy to maintain the original composition of species, the system is allowed to mature without interference, resulting in a set of species that is most suited to the conditions at a given site. The species most able to survive and reproduce in that particular environment will spread and grow. In a stream restoration effort, material necessary to create lost habitat complexity and diversity, such as large organic debris or sediment can be provided to the system and then be distributed by natural stream forces rather than anchoring it in place. By incorporating natural process into ecological engineering designs, ecological engineers reduce the use of nonrenewable energy and nonrecyclable material input and allow the self-organization capacity of the ecosystem to determine what is most suitable in a given location using natural goods and services.

Disturbance and thresholds

Ecosystems have a history of disturbance that influences the current composition of the ecosystem. Self-organization capacity in concert with natural selection and evolution result in an ecosystem that reflects past conditions and disturbances. Some organisms are more successful in relative stable conditions; some can only compete successfully in dynamic and ever-changing conditions. Disturbances (e.g., fires, floods, tides, seasons, plate tectonics) periodically introduce forces that reset ecosystem structure at varying temporal and spatial scales. If a disturbance occurs that creates a large change in the abiotic or biotic environment, an ecosystem can cross a threshold which makes it unable to recover to its previous state. Examples include loss of soil fertility that cannot be restored, changes in water table levels that lead to desertification or wetland development, loss of a keystone species, and introduction on new diseases or species.

Human-managed systems typically lose structural and functional diversity (fewer species providing fewer goods and services) and become more spatially uniform over time, for example, agricultural fields, managed forests, urbanization. The alteration of natural disturbance cycles (e.g., fire suppression and flood control) disrupts the patch dynamics of ecosystems and generally reduces habitat complexity and thus biodiversity. As uniformity increases, resistance to disturbances decreases. For example, spatially extensive monocrops, whether mountain forests, urban trees, or agricultural crops can be destroyed by a single virulent disease vector. Monocrops tend to be genetically similar, do not have the variability to resist diseases and diseases spread rapidly to adjacent identical, susceptible plants.

This homogenization of human-managed systems highlights the differences between engineering resilience and ecological resilience. Designing for traditional engineering resilience seeks stability and permanence; systems with high engineering resilience return to a stable equilibrium point quickly after a disturbance. Ecosystems do not exist around a stationary equilibrium point. Ecological resilience is a measure of how large a disturbance an ecosystem can absorb and maintain its original structure and function. Major disturbances, whether chronic or instantaneous, such as volcanic eruptions, glacial advances or retreats, overfishing, or sea level rises and falls may cause an ecosystem to cross a threshold. The ecological resilience of ecosystems can be exceeded by a major disturbance. This typically results in a dramatic change in species composition and the start of a new self-organization process perhaps with different abiotic conditions.

Ecological engineering recognizes the stochastic nature (unpredictability) of disturbances in space and time and seeks designs that tolerate multiple states while still meeting the design purpose, thus benefiting humans and protecting the environment. Ecological engineers recognize the hubris in the thinking that ecosystems need to be managed constantly or extensively to provide goods and services. Ecological engineering designs take into account the variability in time and space of processes and species composition across the landscape. This awareness of the stochastic nature of ecosystems is vital for meeting the challenges of climate change and maintenance of ecosystem services. By including no and low-regret options in project designs that produce ecosystem services, benefits to humans can be provided irrespective of climate trajectories.

Ecological engineering recognizes the four categories of ecosystem services (supporting, provisioning, regulating, and cultural) and seeks to include such services in their designs and projects. This includes taking into consideration the concepts of natural capital in addition to standard economic analyses.

Traditional engineering design incorporates factors of safety in design parameters. Risk analysis estimates the probability of failure of the design, and energy and material are used to enable the structure to resist failure that can result in harm to humans or infrastructure. This is called fail-safe design. The risks are known and are more or less predictable. Efficiency, constancy, and predictability are guiding principles for traditional engineering. Ecological engineers recognize that over time the forces of nature can overcome any affordable design. Persistence, change, and unpredictability are hallmarks of ecological theory. Ecosystems are complex systems with many variables and risk may come more from unknown (or unrecognized) sources than from known sources. That is, the probability of occurrence of the risk is unknown or in some cases, the risk itself is unknown. For this reason, ecological engineering strives for safe-fail design, that is, when the design fails, the failure takes place such that extensive harm to humans, infrastructure, and the ecosystem is minimized. When considering design alternatives, ecological engineers choose the one that has the best worst-case outcome.

Common Steps for Ecological Engineering Design

There is no cookbook available for ecological engineering design. The emergent properties of ecosystems do not lend themselves to a constant set of variables such as exists for chemistry (periodic table of elements) or mechanics (design table properties for steel or concrete). Each setting for ecological engineering design will have a unique history and set of interactions.

Ecological engineers are aware of and take advantage of the processes that are active in natural systems. This awareness comes from a thorough understanding of ecological theories that describe the ecosystem of interest to the designer. The naturally occurring ecosystem processes are partners in design, not obstacles to overcome and dominate. Important aspects of ecosystems that need to be accounted for in design include disturbance, diversity, heterogeneity, change, and self-organization at multiple scales in space and time. Using a standard design procedure allows for the documentation of responses and allows ecological engineering to be used to test ecological theories. Importance of the following components and depth of analyses will vary by environment and design objectives.

Following the steps below provides the relevant ecological information that is needed to create an ecologically engineered design. The final design, grounded in the information gathered below, adheres to the traits listed in **Table 1** under ecological engineering.

1. Identify the biotic and abiotic factors that drive the ecosystem of interest.
   a. These factors control organisms and pathways through which energy flows and materials cycle.
   b. The design should not impair these factors.
2. Identify the types of disturbance, whether chronic or intermittent, biotic or abiotic, that are present in the system.
   a. The design should be ecologically resilient to these disturbances.
   b. The design should be safe-fail.
   c. The design should maintain spatial and temporal heterogeneity in the system.
3. Identify the goods and services being produced by the ecosystem.
   a. Production of goods and services should be maintained or enhanced.
   b. Inputs of human produced materials should not exceed assimilation capacity.
   c. Any wastes that are produced should be usable in another design.
   d. Energy needs of the design should minimize the use of nonrenewable sources.
   e. Extraction of renewable resources should be less than the rate of renewal.
4. Use the naturally occurring forces of nature to help with design and maintenance.
   a. Working at cross-purposes with nature is frustrating and expensive in the best case and disastrous and counterproductive in the worst case.
5. Recognize that implementation of any design will create some disturbance to the preexisting conditions.
   a. No design is perfect. Accurate appraisal of potential problems allows for minimization and/or mitigation of the impacts.
6. Keep complete and accurate documentation of design process, parameters, and outcome.
   a. Documentation of preexisting conditions, design process, and monitoring of outcome provides the means to improve designs in the future.

Ecological engineering designs can be applied to a variety of ecosystem problems, such as:

- Phytoremediation and wastewater treatment wetlands can be used to reduce or solve pollution problems. In this case, the design seeks to replicate or take advantage of ecosystem properties.

Table 1 Concepts and characteristics of traditional versus ecological engineering designs

Traditional engineering	*Ecological engineering*
Efficiency of function	Persistence of function
Seeks stability	Accepts inevitability of change
Resists disturbance	Absorbs and recovers from disturbance
One equilibrium point	Multiple, nonstable equilibria
Redundancy of structure	Redundancy of function
Single acceptable outcome	More than one acceptable outcome
Spatially and temporally uniform	Spatially and temporally diverse
Tries to control natural forces	Works with natural forces
Predictability	Unpredictability
Fail-safe	Safe-fail
Tight tolerances	Wide tolerances
Heavy reliance on nonrenewable energy and material	Maximum use of renewable energy and energy and material
Rigid boundaries and edges	Flexible boundaries and edges
Unconcerned by production of waste materials from the design	Minimizes production of waste and seeks to use the waste in another design or process
Deductive	Inductive
Engineering resilience	Ecological resilience

- Forest restoration or wetland mitigation can be used to reduce resource degradation problems. Here the design seeks to copy or reproduce ecosystem structure and function and to provide ecosystem services.
- Mine land restoration or lake restoration seeks to hasten the recovery of an ecosystem following major disturbance. Here the design seeks to uses the self-organization properties of ecosystems to recreate the predisturbance system. The design is mindful that some disturbances, such as fires and hurricanes are natural, and ecosystems have recovered from them before human management or intervention was possible or even considered.
- Extraction or use of ecosystem goods and services are done such that production of those goods and services is not decreased. Here the design seeks to meet sustainability criteria and decrease the use of nonrenewable energy.
- Coastal and riverine projects involved in adaptation to climate change and other human activities seek to maintain ecosystem services and natural capital.

Further Reading

Bergen, S.D., Bolton, S.M., Fridley, J.L., 2001. Design principles for ecological engineering. Ecological Engineering 18, 201–210.
Brown, M.T., Ulgiati, S., 1999. Emergy evaluation of natural capital and biosphere services. Ambio 28 (6), 486–493.
Cheong, S., Silliman, B., Wong, P.P., *et al.*, 2013. Coastal adaptation with ecological engineering. Nature Climate Change 3, 787–791.
Costanza, R. (2012). Eceosystem health and ecological engineering. *Institute for Sustainable Solutions Publication and Proceedings*. Paper 70.
Hollings, C.S., 1996. Engineering resilience versus ecological resilience. In: Schulze, P.C. (Ed.), Engineering within ecological constraints. Washington, D.C.: National Academy of Engineering, pp. 31–43.
Kangas, P.C., 2004. Ecological engineering: Principles and practice. Boca Raton, FL: Lewis Publishers.
Krotscheck, C., Narodoslawsky, M., 1996. The sustainable process index: A new dimension in ecological evaluation. Ecological Engineering 6, 21–258.
Mitsch, W.J., 2012. What is ecological engineering? Ecological Engineering 45, 5–12.
Mitsch, W.J., Jørgensen, S.E., 2004. Ecological engineering and ecosystem restoration. Hoboken, NJ: John Wiley & Sons.
Odum, H.T., 1996. Environmental accounting: EMERGY and environmental decision making. New York: Wiley.
Palmer, A.M., Filoso, S., Fanelli, R., 2014. From ecosystems to ecosystem services: Stream restoration as ecological engineering. Ecological Engineering 65, 62–70.
Suh, N.P., 1990. The principles of design. New York: Oxford University Press.
Todd, J., Josephson, B., 1996. The design of living technologies for waste treatment. Ecological Engineering 6, 109–136.

Relevant Websites

NASA Global Climate Change—https://climate.nasa.gov/.
Millennium Ecosystem Assessment—http://www.millenniumassessment.org/en/Index-2.html.
The Economics of Ecosystems and Biodiversity—http://www.teebweb.org/.

Ecological Engineering: Overview

SE Jørgensen, Copenhagen University, Copenhagen, Denmark

What Is Ecological Engineering?

The most used definition of ecological engineering employs the following formulation: ecological engineering is defined as the design of sustainable natural and artificial ecosystems that integrate human society with its natural environment for the benefit of both. It requires, on the one hand, that we understand nature and ensure a sustainable development of natural resources and ecosystems and, on the other hand, that we make use (but not abuse) of natural resources to the benefit of the human society. Thus, our inevitable interactions with nature must be made under the comprehensive consideration of the sustainability and balance of nature.

H. T. Odum was among the first to define ecological engineering as the ``environmental manipulation by man using small amounts of supplementary energy to control systems in which the main energy drives are coming from natural sources.'' Odum further developed the concept of ecological engineering as follows: ecological engineering, the engineering of new ecosystems designs, is a field that uses systems that are mainly self-organizing.

Straskraba has defined ecological engineering (or ecotechnology, as he called it) more broadly, as being the use of technological means for ecosystem management, based on a deep ecological understanding, in order to minimize the costs of measures and their harm to the environment. For the purposes of this report, ecological engineering and ecotechnology may be considered synonymous.

Ecological engineering is engineering, in the sense that it involves the design of man-made or natural ecosystems or parts of ecosystems. Like all engineering disciplines, it is based on basic science, in this case ecology and systems ecology. The biological species are the components applied in ecological engineering. Thus, ecological engineering represents therefore a clear application of ecosystem theory.

Ecotechnic is another often applied word but one that also encompasses the development of all types of `soft' technology applied in society, in addition to ecotechnology or ecological engineering. These types of technology are often based on ecological principles (e.g., all types of cleaner technology), particularly if they are applied to solve an environmental problem. The use of ecological principles in the development of technology is denoted as industrial ecology.

Recently, UNEP and UNESCO have introduced two other terms relevant to this discussion:

1. *Phytoremediation*. The use of plants in ecological engineering (e.g., using wetlands to treat wastewater pollutants, or for removing toxic substance from contaminated soil).
2. *Ecohydrology*. The use of a combination of ecological and hydrological principles to obtain ecologically sound environmental management.

Both phytoremidiation and ecohydrology are subdisciplines within the discipline ecological engineering or ecotechnology, which is an often used synonym for ecological engineering.

Further, ecological engineering should not be confused with bioengineering or biotechnology. Biotechnology involves the manipulation of the genetic structure of cells to produce new organisms capable of performing certain functions. Ecotechnology does not involve manipulation at the genetic level, but rather at several steps higher in the ecological hierarchy. The manipulation takes place on an assemblage of species and/or their abiotic environment, as a self-designing system that can adapt to changes brought about by outside forces, whether controlled by humans or by natural forcing functions.

Ecological engineering is also not the same thing as environmental engineering, the latter is involved in cleaning processes to prevent pollution problems. It involves the use of settling tanks, filters, scrubbers, and man-made components that have nothing to do with the biological and ecological components applied in ecological engineering, even though the use of environmental engineering is directed to reducing man-made forcing functions on ecosystems. As mentioned above, the term ecotechnic may be considered to include a part of environmental technology, namely the part based on ecological principles such as recirculation. The tool boxes of ecological engineering and environmental engineering are completely different; where ecological engineering uses ecosystems, communities, organisms and their immediate abiotic environment, and environmental engineering uses chemical and biotechnological unit processes such as filtration, precipitation, and biological decomposition by aeration.

All applications of technologies are based on quantification. Because ecosystems are very complex systems, the quantification of their reactions to impacts or manipulations is also complex. Fortunately, ecological modeling represents a well-developed tool to survey ecosystems, their reactions, and the linkage of their components. Ecological modeling is able to synthesize our knowledge about an ecosystem, making it possible to quantify, to a certain degree, any changes in ecosystems resulting from the use of both environmental engineering and ecological engineering. Ecological engineering may also be used directly to design

 doi:10.1016/B978-0-444-63768-0.00042-1

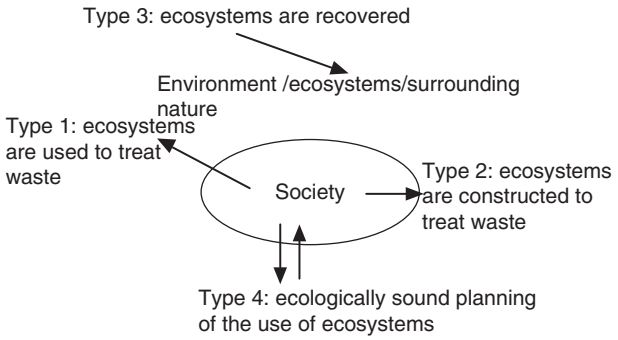


Fig. 1 An illustration of the four types of ecological engineering. Reproduced by permission of Elsevier.

constructed ecosystems. Consequently, ecological modeling and ecological engineering are two closely cooperating fields. Research in ecological engineering was originally addressed in the *Journal of Ecological Modelling*, which was initially named *Ecological Modelling – International Journal on Ecological Modelling and Engineering and Systems Ecology* to emphasize the close relationship between the three fields of ecological modeling, ecological engineering, and systems ecology. *Ecological Engineering* was launched as an independent journal in 1992, with the name of *Ecological Modelling* being changed to *Ecological Modelling – An International Journal on Ecological Modelling and Systems Ecology*. At the same time, the journal *Ecological Engineering* has successfully covered the field of ecological engineering, which has grown rapidly during the 1990s due to increasing acknowledgment of the need to use technologies other than environmental technology in efforts to solve pollution problems. This development does not imply that ecological modeling and ecological engineering are moving in different directions. On the contrary, ecological engineering has increasingly been using models to perform designs of constructed ecosystems, or to quantify the results of applying specific ecological engineering methods for comparison to alternative, applicable methods.

In addition, the relationship between ecological engineering and systems ecology is very clear. Ecological principles are used widely in practical application of ecological engineering methods. Mitsch and Jørgensen have provided 19 principles that can be used as a checklist to assess if an ecological engineering project follows ecological principles, that is, to determine if a project is ecologically sound.

Classification of Ecotechnology

Ecological engineering may be based on one or more of the following four classes of ecotechnology:

1. Ecosystems are used to reduce or solve a pollution problem that otherwise would be (more) harmful to other ecosystems. A typical example is the use of wetlands for wastewater treatment.
2. Ecosystems are imitated or copied to reduce or solve a pollution problem, leading to constructed ecosystems. Examples are fishponds and constructed wetlands for treating wastewater or diffuse pollution sources.
3. The recovery of ecosystems after significant disturbances. Examples are coal mine reclamation and restoration of lakes and rivers.
4. The use of ecosystems for the benefit of humanity without destroying the ecological balance (i.e., the utilization of ecosystems on an ecologically sound basis). Typical examples are the use of integrated agriculture and development of organic agriculture; this type of ecotechnology finds wide application in the ecological management of renewable resources.

The rationale behind these four classes of ecotechnology is illustrated in **Fig. 1**. It is noted that ecotechnology or ecological engineering operates in the environment and its ecosystems. As already mentioned, it is this domain that ecological engineering employs as its toolbox.

Illustrative examples of all four classes of ecological engineering may be found in situations where ecological engineering is applied to replace environmental engineering, mainly because the ecological engineering methods offer an ecologically more-acceptable solution, and where ecological engineering is the only method that can offer a proper solution to a problem. Examples are provided in **Table 1**,where alternative environmental technological solutions are also indicated. This does not imply that ecological engineering can replace environmental engineering. On the contrary, the two technologies should work hand-in-hand to solve environmental management problems, better than they could do if applied individually. This is illustrated in type 1 ecological engineering (application of ecosystems to reduce or solve pollution problems) by wetlands utilized to reduce diffuse nutrient loads to lakes. This problem cannot be solved by environmental technology. Sludge treatment can be solved by environmental technology, namely by incineration. However, the ecological engineering solution (i.e., sludge disposal on agricultural land, which involves utilization of the organic material and nutrients in the sludge) is a much sounder method from an ecological perspective. **Fig. 2** gives an example, where both ecological engineering and environmental technology are applied to solve an environmental problem.

Table 1 Ecological engineering examples (alternative environmental engineering methods are given)

Type of ecological engineering	Example of ecological engineering		
	Without environmental eng. alternative	With environmental eng. alternative	Environment engineering alternative
1	Wetlands utilized to reduce diffuse pollution	Sludge disposal on agricultural land	Sludge incineration
2	Constructed wetland to reduce diffuse pollution	Root zone plant	Traditional wastewater treatment
3	Recovery of lakes	Recovery of contaminated land *in situ*	Transport and treatment of contaminated soil
4	Agroforestry	Ecologically sound planning of harvest rates of resources	

Reproduced by permission of Elsevier.

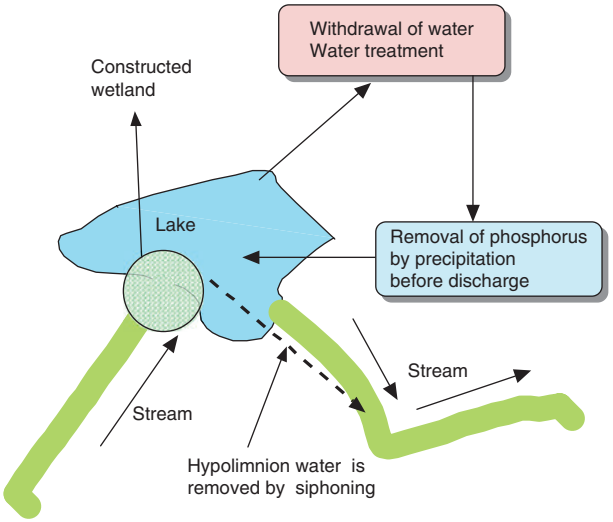


Fig. 2 Control of lake eutrophication, illustrating a combination of (1) chemical precipitation for phosphorus removal from wastewater (environmental technology); (2) a wetland to remove nutrients from the inflow (type 1 or 2 ecotechnology); and (3) siphoning of nutrient-rich hypolimnetic water downstream (type 3 ecotechnology). The eutrophication abatement may also be combined in this case with biomanipulation. Reproduced by permission of Elsevier.

The application of constructed wetlands to cope with diffuse pollution is a good example of type 2 ecological engineering. Again, this problem cannot be solved by environmental technology. The application of root zone plants for treating small quantities of wastewater is an example of type 2 ecological engineering, in which the environmental engineering alternative (a mechanical–biological–chemical treatment) cannot compete, when the waste volume is low and/or the area costs are moderate, mainly because it would involve excessive costs, relative to the quantity of wastewater (sewage system, pumping stations, etc.). A solution requiring fewer resources always will be a more ecologically sound solution.

Although recovery of land contaminated by toxic chemicals is possible using environmental technology, it will require transportation of the soil to a soil treatment plant, where biological biodegradation of the contaminants would take place. Ecological engineering will propose an *in situ* treatment with adapted microorganisms or plants. The latter method will be much more cost-effective, and the pollution related to transporting the soil will be omitted. Restoration of lakes by biomanipulation, installation of an impoundment, sediment removal or coverage, siphoning of hypolimnetic water (rich in nutrients) downstream, or by several other proposed ecological engineering techniques are examples of type 3 ecological engineering. It is difficult to obtain the same results using environmental engineering, because this requires activities in the lake and/or the vicinity of the lake.

Type 4 ecological engineering is based, to a great extent, on pollution prevention by utilization of ecosystems on an ecologically sound basis. Although it is very difficult to find environmental engineering alternatives in this case, it is clear that a prudent harvest rate of renewable resources (whether, e.g., timber or fish) is the best long-term strategy from an ecological and economic perspective. Ecologically sound landscape planning is another example of the use of type 4 ecological engineering.

Constructed subsurface wetlands may also be used to treat dairy farm wastewater, mine water pollutants, textile wastewater, and pulp mill wastewater.

Further Reading

Jørgensen, S.E., 2000. Pollution Abatement in the 21st Century. Amsterdam: Elsevier, 488pp.
Mitsch, W.J., Jørgensen, S.E., 1989. Ecotechnology – Introduction to Ecological Engineering. New York: Wiley, 472pp.
Mitsch, W.J., Jørgensen, S.E., 2003. Ecological Engineering and Ecosystem Restoration. New York: Wiley, 412pp.
Straskraba, M., 1985. Simulation Models as Tools in Ecotechnology Systems: Analysis and Simulation, vol. II. Berlin: Akademie Verlag, 546pp.

Forestry Management

HH Shugart, University of Virginia, Charlottesville, VA, USA

Introduction

Trees are very long-lived organisms with high potential birth rates. Thus, estimating the basic birth and death parameters of tree populations is intrinsically difficult over the time duration of most ecological studies. The sizes of trees from seedling to mature tree vary over five orders of magnitude. Trees grow large enough to alter aspects of their environment as they mature. This individual–environment feedback is not usually included in traditional ecological population models. Tree interactions obtaining essential resources of light, water, and nutrients involve tree geometry in vertically for light, spatially for nutrients and water, and volumetrically for interactions among tree crowns such as crown pruning (where the branches of a tree abrade the buds from limbs of neighbor trees and change the shapes of competing trees). These geometrical aspects of tree populations are omitted in the mathematical structures of most ecological population models which, at least until recently, considered only the time dimension. Tree populations represent a modeling departure from traditional population models. Significantly, forestry models from their origins have always attempted to predict a combined response of the sizes of trees and the number of trees on a given area.

Tree population models have deep historical roots that are often not appreciated by modern population ecologists, perhaps in part because the origins of many of these forestry-based approaches are in applied fields and are focused on practical, regional results rather than the development of a general theory. For this reason, it is useful to discuss forestry models from their beginnings through the evolution to the modern approaches.

Yield Tables: Empirical Forestry Models

Throughout Europe in medieval times there was a substantial clearing of forests followed by even more extensive deforestation regionally. With the progressive reduction of forests, class conflicts over the products of forests intensified and this manifested itself in laws against poaching of animals, thieving of wood, and proscriptions against public use of forests, in general. In the mid-eighteenth century, a forest management concept called `Nachhalttigkeit' or sustainability was developed by the Germans. From about 1800, this new forestry practice spread over Europe, particularly Northern Europe. This was essentially a transition from the earlier, exploitive extraction of materials from forests to a more `trees as a crop' agricultural management of forest tracts. Nachhaltigkeit involved detailed determination of how to best manage forests to produce wood and other goods. Essential to this objective was manipulating density by spacing trees on a given site, either by planting trees or by thinning a naturally regenerated stand of trees following a timber harvest or a natural disturbance. Additionally, one needed to determine how long one should wait before harvesting a stand of trees and then planting a new stand. This spacing/length-of-rotation problem had long been solved for crop plants through experimentation and observation. To produce trees as long-lived crops, elaborate long-term data collection started on the height; size in diameter; amount of wood; and size of crowns in forest stands of different densities at sites with different environmental conditions. Eventually, a forest modeling concept called the `yield table' approach developed and became the signature of modern forestry (see **Fig. 1**).

Nowadays, some of the historical forest data sets used in yield tables have grown to 250+ years of continual record. It was found that on a given kind of site (same soils, same rainfall, etc.) trees grew to the same height at a given age, regardless of density of trees. At low densities, one might find trees with large diameters and crowns and on an equivalent high-density location the trees would have small diameters and crowns – but the heights of the trees would be the same in both cases. In a yield table, decades and sometimes centuries of forest stand data are arranged by the height the trees at a given location reach at a given age, usually the typical age of tree harvest. The tree height at this standard age is called a `site index' and is used to signify the overall quality of a location for growing trees. Site index is clearly defined in terms of the basic data that goes into a yield table and it can be directly determined by measuring the heights of trees on even-aged stands at the reference age.

Associating site index with actual plots of land is a learned skill and an art at the same time. A capable site surveyor can judge site index by reconnaissance of land in a particular region and can make a good wage practicing this trade. Along with such arcane practices as axe-throwing, log-rolling, and tree-felling, forestry schools have regular intercollegiate competitions of judging site indices among their students as part of `Forestry Field Day' celebrations. At the edge of virtually all universities with a college of forestry, one will find plat of trees planted at different densities for field teaching on the calibration of yield tables. The yield table concept is the quantitative basis of modern forestry.

 doi:10.1016/B978-0-444-63768-0.00670-3

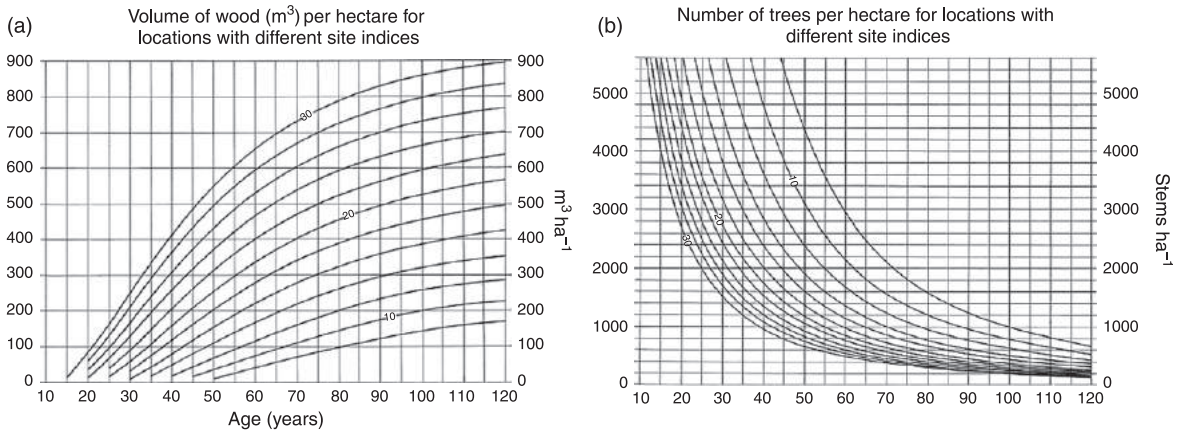


Fig. 1 Sections of stand yield tables for Switzerland. The site index is the height of the dominant trees in even-aged stands of spruce (*Picea abies*) at an age of 50 years. Note that this data set covers 120 years of measurement. The amount of wood (a) in a forests increases with age and slows in its rate of increase over time. Sites with higher site indices have more wood at any given time. The thinning of trees (due to suppressed trees being eliminated by more vigorously growing taller trees) slows over time (b) but is stronger in locations with higher site indices. Reproduced from Erragstafeln: Fichte. Eidgenössische Anstalt für das forstliche Versuchswesen, Birmensdorf ZH, with permission from Swiss National Forest Inventory.

Beyond Yield Tables

In the twentieth century, yield table methodologies had become the standard for forestry applications and were essential part to any forestry curriculum. Considerable effort went into enriching the data sets and perfecting the basic concepts of yield tables for forestry production. For example, since the product of a forest plantation is not always wood but is often boards of a given dimension, allowing a forest to grow more wood (as predicted by a yield table) does not necessarily produce more boards. As another example, under certain spacing, the trunks of trees have different degrees of taper which also affect the lengths of standard boards to be sawn from the harvested wood. These sorts of complexities were determined and calibrated into the predictions from yield tables.

In the mid- to late-1950s several developments conspired to generate difficulties for yield table projections. The dependency on extensive long-term data sets to calibrate yield tables manifested itself as a problem. Some of the long data sets in Europe were initiated before the end of the so-called `Little Ice Age' around the 1850s and thus were collected over intervals of significant climate applications. The warmer and often dryer conditions that prevailed in the decades of 1930s and the 1950s represented a similar problem for shorter calibrations of other yield tables. The fundamental assumption in yield tables, that the expected height of a canopy tree in a stand of a reference age could be used as a constant to encapsulate the quality of a site for growing trees, was undermined by these climate variations.

At the same time and particularly in the US, foresters were experimenting with developments that would change the growth rates of trees – using genetic selection to develop faster-growing commercial tree species or fertilizing forest for faster growth. Increased air pollutants and changes in the chemistry of rainfall (`acid rain') were seen as potentially slowing tree growth. The possibility of metrification in the US meant that the complex calculations to calibrate merchantable-dimension boards in English units would need to be recalibrated for metric units. In the western states of the US, the end of harvest of previously uncut forests of Douglas fir (*Pseudotsuga menziesii*) and the interests in developing plantation forestry for this species poised the immediate question of how does one develop a yield table for a condition for which there is no long-term data record. Japanese foresters faced a similar problem with their plantation forestry practice at home and in foreign countries. In tropical and subtropical locations, where the forests have relatively little softwood timber, planting of exotic North American species, notably Monterey pine, *Pinus radiata*, gave local foresters a need to develop planting strategies with no data to calibrate yield tables. Worldwide, the realization that forests do more than just produce wood called for a simultaneous management of forests for water quality, wildlife, flood control, landslide protection, and recreation, all implied a broader approach than the wood production emphasis of yield table forestry. This soon was seen as need for a better capability to manage complex, natural forests with a mixture of tree ages and species. There was a considerable impetus to develop new approaches to projecting forest dynamics beyond the yield table approach.

In the late 1950s and early 1960s, the need for forest prediction beyond the capabilities of empirical yield tables led to a proliferation of forest modeling approaches, notably in the US and Japan. This modeling revolution in the late 1960s and 1970s occurred independently at different locations with different emphases on theory development, computer simulation, and advanced statistics. These developments, intended to re-analyze the intensive data archives behind the yield tables and extend insights of a century of forestry empirical-model predictions, coincided with a fascination with the potential for high-speed computers. Many of the resulting models depend strongly on digital computation and have become more widespread as the power of computers has increased over the past 50 years.

Each of these different foci on the modeling of forests contributed different points of view as to how to predict the dynamics of forest ecosystems and are discussed in the sections below. In the case of forestry models, three significant intellectual centers for this development were

1. a Japanese group of forest scientists who emphasized the theoretical underpinnings for forest dynamics;
2. one US center focused on the statistics needed to parametrize coupled differential equations for biomass, number of trees, and size of the average tree in forest plantations. This group was strongly oriented to understand the dynamics of Loblolly pine (*Pinus teada*) plantations;
3. a second US group developing digital computer simulation models that incorporated the three-dimensional geometry of interactions among growing trees, initially emphasizing Douglas fir and spruce forests. The approach was to develop what is referred to today as individual-based or agent-based models.

Each of these groups produced different insights on the population dynamics of forests and has evolved in relative independence until the present day.

At about the same time that new alternatives to yield tables were being developed in forestry, ecologists were developing a parallel interest in the use of dynamic models to predict forest dynamics. This culminated in a major international research program called the International Biological Programme (IBP) intended to foster a comparative understanding of ecosystem dynamics across the globe. The focus was on the productivity and element processing of natural ecosystems and constitutes a fourth independent thrust to develop forest ecosystem models. The models developed in this program tended to emphasize fundamental processes (productivity, decomposition, element uptake by ecosystems). These models represent a relatively significant departure for the forestry modeling tradition in assuming these processes can be represented over a homogeneous forested area. The IBP models are in many cases the predecessors of ecological models currently applied to compute the carbon dynamics to predict the responses of the global carbon cycle.

Japanese Theoretical Approaches to Forest Population Modeling

While the German traditional approach to yield table development were based on the normalization of data based on tree heights, Japanese foresters focused on the statistical interrelations among all of the dimensions of trees. They developed a geometrical theory for the dynamics of forest stands using allometric equations. These equations express one measurable dimension of an organism as another dimension raised to a fractional power times a constant ($x_1 = cx_2^z$, where x_1 and x_2 are different measurements of a tree and c and z are constants). At the beginning of the twentieth century, the British mathematical ecologist, D'Arcy Wentworth Thompson, had written extensively about these and other more complex underlying mathematical formulations describing the shapes of plants and animals. The Japanese compilation of these data for trees was extensive. Significantly, K. Shinozaki and his colleagues developed a basic theory (in 1964) to explain some of these regularities in form.

Shinozaki's `pipe model' posited that a tree is a great collection of pipes – notably the xylem cells, tubes that transmit water from roots through the trunks and branches to the leaves. In the pipe model, the cross-sectional area of the tree's transport systems (the `pipes') should be conserved at different levels of the tree. A given area of leaves required a certain cross-sectional area of xylem tubes through all the parts of the tree that supplied it. As a tree grows taller, more and more length of pipes is needed to provide the same cross-sectional area. Thus, the cost to the tree of supporting a given area of canopy leaves (requiring a given cross-sectional area of pipes) increases as the tree grows. The exponent in an allometric equation typically has values near one-third when relating a linear dimension such as tree's diameter to volume-related measurements such as tree biomass and a value near two-thirds when relating area measurements such as cross-sectional area of a tree trunk to tree mass.

In 1963, the Japanese forester, K. Yoda, and several of his colleagues extended the classic allometric concept for individual organisms to entire forest stands. The motivation was to investigate one of the basic premises in yield tables – the regularities in the average size of trees in even-aged forest stands as a function of the density of trees. In self-thinning stands (stands in which some trees die from competition as other, more vigorous trees grow), the average biomass of a tree is a constant times the density of trees raised to the $-3/2$ power (Fig. 2). This $-3/2$ power law or self-thinning law has been found in a wide variety of ecosystems in which organisms (corals, grasses, trees) compete for space. Its application is restricted to even-aged systems that have been initiated with a competing cohort of individuals.

Models Coupling Mass and Numbers Equations for Forest Dynamics

At about the same time that foresters in Japan were developing their theoretical concepts, J. L. Clutter and a group of colleagues and students at the University of Georgia were developing an alternative approach with the same underlying concept as that of the Japanese–connecting the size of an average tree to the number of trees in an even-aged forest stand. The Georgian approach was to conceive stand dynamics as being modeled by two differential or difference equations, one equation for the change of numbers over time and the other for the change in the size of the average tree over time. These two equations were coupled to one another by statistically estimated terms. For example, if the weaker trees that die as the stand trees are smaller than average, then tree death increases the average tree size and a part of the average-size equation would have a term that expresses this effect. The terms in each equation that coupled numbers to size (and vice versa) were estimated statistically from the same sort of data that is used in yield

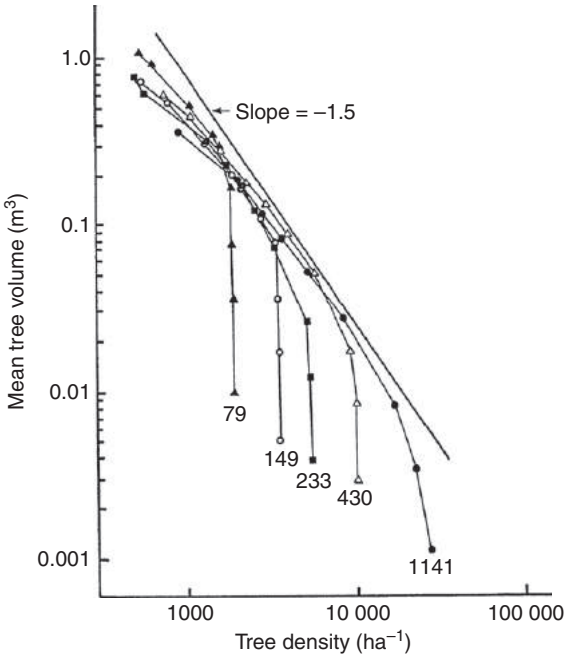


Fig. 2 An example of a $-3/2$ law stand-thinning concept. Forest stands planted at different densities converge to a limit described by a line with a slope of $-3/2$ (or -1.5). Data are for Loblolly pine plantations with different initial planting densities (79 trees ha^{-1}; 149 trees ha^{-1}; 233 trees ha^{-1}; 430 trees ha^{-1}; 1141 trees ha^{-1}). The points represent stand inventory starting in 1935 when the experimental forest was established at the Duke University Experimental Forest near Durham, New Hampshire. As trees increase in size, numbers are reduced in these even-aged stands. Reproduced from Peet, R.K., Christensen, N.L., 1987. Competition and tree death. BioScience 37, 586–595, with permission from American Institute of Biological Science.

tables in some cases. In other cases, the need for information on how to couple the number and mass equations suggested new experiments or new data collections. The advantage of the method was that if genetics or forest fertilization made the trees grow faster, this effect could be directly added to the average tree mass equation. The estimation of the parameters in the coupled equations could be improved as more field data were obtained. Thus, the models were expected to improve over time.

Today, this approach and its descendents dominate plantation forestry, particularly for Loblolly pine (*Pinus taeda*), the principal commercial tree species in the US. The method works best on forests in which the trees do not vary greatly from one another in size (hence, the average tree does not have large associated variability).

Individual-Based Forest Models

The power of digital computation in the middle of the 1960s on what today would be considered primitive computers, inspired the development of a new class of models that simulated the behavior of large complex systems by modeling the dynamics of the individuals in the systems. These included models of forests based on the changes in each individual tree in a simulated forest stand. In these models two implicit assumptions associated with traditional ecological population modeling are not necessary:

1. The assumption that the unique features of individuals are sufficiently unimportant to the degree that individuals are assumed to be identical. This allows the simulation of the numbers of individuals in the population without consideration of sizes or ages of the members of the populations.
2. The assumption that the population is `perfectly mixed' so that there are no local spatial interactions of any important magnitude.

Individual-based models of forests spring from a rejection of these assumptions. Trees vary greatly in size over their life span. Trees are sessile (so that spatial location matters and every individual in a population does not affect all the other members). This may be one of the reasons that tree-based forest models are among the earliest and most widely elaborated of the individual-based genre of models.

These models were developed by quantitatively oriented foresters in the mid-1960s and focused strongly on production forestry. The earliest such model was developed by R. M. Newnham for even-aged Douglas fir forests. This was followed by similar developments at several schools of forestry in North America. These models were implemented on a digital computer programmed to alter the sizes of trees on a map of the positions of each tree in a forest. When a tree died it was erased from the computer map. Three-dimensional spatial interactions among individual trees were explicitly represented and these initial models are as complex and as detailed as their descendants today.

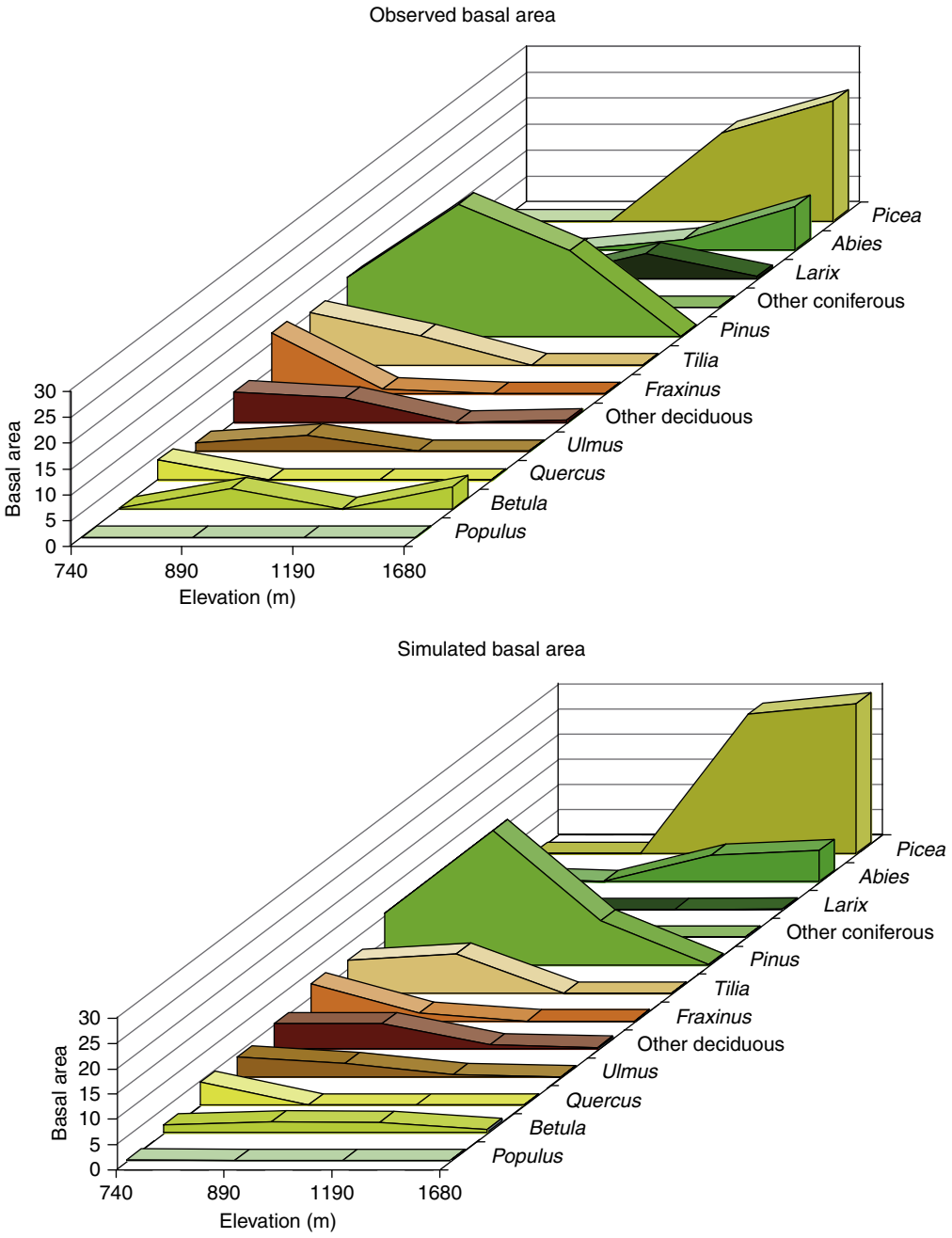


Fig. 3 A comparison of observed and simulated basal area (m^2 ha^{-1}) at four elevations (740, 890, 1190, and 1680 m) along the North Slope of Changbai Mountain (42.2° N, 128.0° E), Northeastern China using the gap model, FAREAST. In this simulation the basal area (m^2 of the sum of stem cross-sections of living trees per ha of forest) measure on the mountain at different elevations is compared to the forest vegetation simulated by the FAREAST gap model for these elevations. Such a model test provides insight in the capability of the model to simulate forest composition expected for different climate conditions. Modified from Yan, X., Shugart, H.H., 2005. A forest gap model to simulate dynamics and patterns of Eastern Eurasian forests. Journal of Biogeography 32, 1641–1648.

These dynamic mapping techniques initially were applied to even-aged plantations but applications in more complex forests soon followed. After these first modeling efforts, individual-based tree models tended to feature simplifications of these initial models. The early individual-tree based simulators took what was known from yield tables and other data sets and developed a more flexible, quantitative methodology for prediction. One significant simplification of this class of models was to simulate only the vertical (shading) relations on a small plot dominated by a single large canopy tree. The resultant `gap models' found wide application in forest ecology of natural, mixed-aged, and mixed-species stands.

Gap models feature relatively simple protocols for estimating the model parameters and synthesize information on the performance of individual trees (growth rates, establishment requirements, and height/diameter relations) to directly estimate model parameters. The simple rules for interactions among individuals (e.g., shading and competition for limiting resources) and equally simple rules for birth, death, and growth of individuals in these models have both positive and negative consequences. The positive aspects largely relate to the ease of estimating model parameters for a large number of species. The negative aspects relate to the lack of physiological mechanism in the description of growth and environmental response. Much of the current research in the subsequent development of gap models relates to overcoming these limitations (**Fig. 3**).

Future Developments

The impetus of the need to predict future forest condition in the face of change that has driven the development of model forestry models should be an even stronger force in the future. Concerns over the responses of forests to climate change have increased the world interest in predicting future forests. Our rather poor appreciation of the potential direct effects of CO_2 in the atmosphere (effects involving the rates of photosynthesis and other plant functions) in addition to the concern over greenhouse-gas-generated climate change will continue to create a need to scale up basic plant physiology to longer timescale and larger spatial-scale consequences. Today's research challenges involve the incorporation of these novel effects and the rich problem of designing ways to test our model predictions. An important application of gap models in this direction was developed by Moorcroft and several of his colleagues. They formulated a gap model for Amazonian rainforest using a postulated relationship between growth rate, photosynthesis, and wood density. This gap model was used to obtain the parameters for a statistical model of the size distribution of trees across the Amazon Basin (based on an extension of the Japanese approaches to forest dynamics modeling pioneered by Kohyama in 1993). This was then driven by a photosynthesis/production model to incorporate the effects of temperature and moisture. The resultant model called the ecosystem demography (ED) model represents a synthesis of the Japanese and individual-based approaches to forestry modeling with process models of the sort initially championed in the IBP.

See also: Ecological Data Analysis and Modelling: Forest Models. Ecosystems: Forest Plantations

Further Reading

Bugmann, H., Reynolds, J.F., Pitelka, L.F., 2001. How much physiology is needed in forest gap models for simulating long-term vegetation response to global change? Climatic Change 51, 249–250.
Huston, M., DeAngelis, D.L., Post, W.M., 1988. New computer models unify ecological theory. BioScience 38, 682–691.
Kohyama, T., 1993. Size-structured tree populations in gap dynamic forests – The forest architecture hypothesis for stable coexistence of species. Journal of Ecology 81, 131–143.
Peet, R.K., Christensen, N.L., 1987. Competition and tree death. BioScience 37, 586–595.
Shugart, H.H., 1984. A Theory of Forest Dynamics: The Ecological Implications of Forest Succession Models. New York: Springer, p. 278.
Shugart, H.H., 1998. Terrestrial Ecosystems in Changing Environments. Cambridge, UK: Cambridge University Press, p. 537.
Shugart, H.H., Smith, T.M., Post, W.M., 1992. The potential for application of individual-based simulation models for assessing the effects of global change. Annual Reviews of Ecology and Systematics 23, 15–38.
Yan, X., Shugart, H.H., 2005. A forest gap model to simulate dynamics and patterns of Eastern Eurasian forests. Journal of Biogeography 32, 1641–1648.

Integrated Farming Systems

David W Archer, Jose G Franco, Jonathan J Halvorson, and Krishna P Pokharel, USDA Agricultural Research Service, Northern Great Plains Research Laboratory, Mandan, ND, Unites States

Glossary

Alley cropping Planting rows of trees or shrubs with grain or forage crops planted between the tree rows. Alley cropping is sometimes called hedgerow intercropping, and is a type of intercropping with larger spaces between rows that allows for managing the crops separately.

Cover crop A crop planted during a fallow period to provide ecosystem benefits. These may include soil cover to reduce soil erosion, inputs to soil organic carbon, weed suppression, nutrient retention, and pollinator benefits.

Crop rotation A sequence of crops on a field in succeeding growing seasons or years, often in a consistent and repeating pattern. For, example in a corn-soybean rotation, corn would be grown in the first year, soybean in the second year, and corn in the third year. However, in a dynamic crop rotation, crop choices may be selected each year or growing season based on environmental or economic criteria, without following a consistent pattern.

Economies of scale The cost advantages enjoyed by firms that are larger in size. Among the reasons economies of scale may occur on a farm are the ability to spread equipment and machinery costs over more units of production, use of larger equipment which reduces labor use per unit of production, and the ability to negotiate better prices on product sales and volume discounts on purchased inputs.

Economies of scope The cost advantages enjoyed by firms that have multiple enterprises. These occur through interactions among the enterprises that increase efficiency of producing and marketing multiple products rather than a single product.

Ecosystem services The direct and indirect benefits humans obtain from ecosystems. The four major categories of ecosystems services are provisioning services, supporting services, regulating services, and cultural services.

Green manure crop Plants left in the field to serve as a soil amendment. Typically these are cover crops and are often incorporated into the soil using tillage.

Intercropping Growing two or more crops in the same field at the same time. The crops are often grown in alternating rows or mixed within rows.

Mixed crop-livestock farming Integrated farming systems that include both crops and livestock. Also called integrated crop-livestock systems.

Monocropping Growing the same crop on a field every growing season or year without rotating to another crop.

Definition

Integrated farming has been defined in numerous ways, and could be envisioned as encompassing a range of agricultural systems arrayed along a continuum of possible organizational structures, and spatial and temporal scales (Bell and Moore, 2012). Some common definitions include the agricultural production system with multiple enterprises that interact in space and /or time to get benefits through a synergistic resource transfer among enterprises (Hendrickson et al., 2008). An emphasis in these systems is managing interactions so that waste from one component becomes an input for another component of the system, reducing the need for purchasing and applying expensive and potentially polluting inputs, such as fuel, fertilizers and pesticides, reducing leakages to the environment, and increasing overall production or income. The key factors in integrated farming systems are the inter-dependence among enterprises within the system, synergetic transfer of resources among enterprises and the flexibility in the system to be sustainable in the long run (Hendrickson et al., 2008; Vereijken, 1989). Integrated farming systems are often assumed to include both crop and livestock enterprises, called mixed crop-livestock farming or integrated crop-livestock systems. However, integrated farming systems may broadly include systems where only multiple crop components interact. These include crop rotational systems, use of annual and perennial cover crops, green manure crops, or intercropping to reduce the need for purchased inputs by fixing or retaining nutrients and reducing weed, disease, and pest pressures.

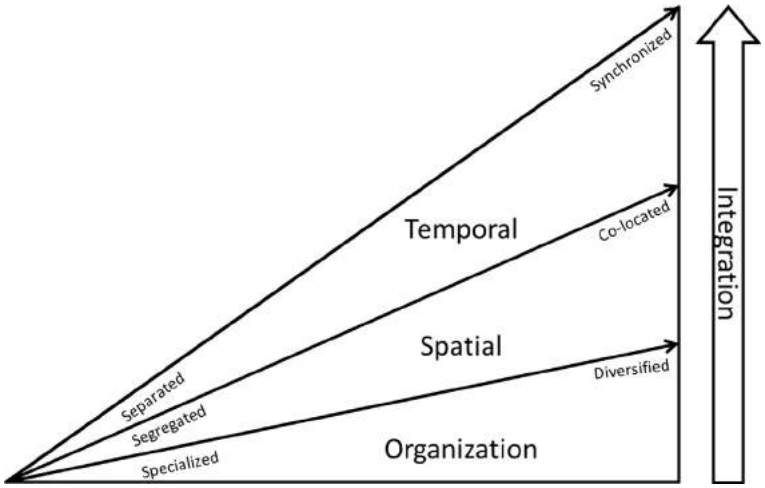


Fig. 1 Continuum of farming system integration in three interacting dimensions: temporal, spatial, and organization.

Following Bell and Moore (2012), we describe integrated farming systems using the three interacting dimensions; organization, space, and time (Fig. 1). However, we envision a continuum for each rather than discrete classes, emphasizing the wide range of systems that occur. The organization scale is represented by the number of enterprises ranging from specialized ($n = 1$) to diversified interactions ($n =$ many). The spatial integration scale is represented by the operating distance between enterprises and ranges from segregated to co-located. Within the continuum of spatial scales, two broad scale categories that have been identified are integration within a single farm unit and regional integration among multiple farms (Russelle et al., 2007). However, there are ranges of integration within these categories. The temporal integration scale refers to the separation in time between cooperating enterprises and ranges from uncoordinated independence to synchronized. Temporal separation includes sequencing of enterprises that can occur over multiple years or even decades, also known as phase farming. Sequencing enterprises also includes the traditional idea of crop rotation, with different crops grown each cropping season or year. With synchronized enterprises, multiple enterprises occupy the same unit of land at the same time, or within the same growing season. For example, using grazing animals within a growing crop, or grazing crop residues after grain crop harvest. In this representation, the most closely integrated farms are those with many enterprises that are co-located and synchronized temporally. Two additional integration dimensions of ownership and management have also been described (Sumberg, 2003), where ownership describes the degree to which assets are in the hands of a single entity, and management describes the degree to which management control is in the hands of a single entity. Integration tends to decrease when assets or management control are distributed across multiple entities. Informal integration among multiple farms or more organized farm cooperatives are instances where integration can cross multiple ownership and management entities. Contracts between specialist crop and livestock producers for the transfer of manures and, to a lesser extent, feed can be a mechanism to increase the adoption rate of integrated farming (Wilkins, 2008).

History

Evidence for the development of agriculture, including both crop and livestock domestication, has been uncovered in various parts of the world dating back some 10–12 thousand years, and has been termed the "Neolithic Revolution". The gradual and sporadic evolution from reliance on nomadic hunting and gathering to practices that promoted greater food security is thought to have been influenced by several factors including environmental limitations/opportunities or socio-economic drivers. Larger and more dependable supplies of food together with their requisite investments of time and labor ultimately resulted in the development of permanent settlements, higher population growth with concomitant increases in societal complexity, and the emergence of new technology. At some point there was a transition from coincidental husbandry of animals dependent on natural resources (e.g. pasture) to purposeful integration of animals and crops such that temporal, spatial and organizational synergies between under-utilized resources or waste products such as crop residues and animal manure were realized (Hilimire, 2011).

In integrated farming systems of the past, animals were used directly for food or to provide other services such as power (draught animals) or transportation (horses). In addition, animals were employed indirectly to provide services such as weed and pest control, fertilization, or pollination; or food items such as milk, eggs or honey. Animals were also a source of materials such as manure or leather that could be sold directly or converted to a value-add product, returning cash to the enterprise (Devendra and Thomas, 2002). The emergence of new technologies, along with the availability of abundant and cheap natural resources, characterized the industrial revolution of the18th and 19th centuries. Important technologies that led to changes in agriculture included mechanical transportation and power technologies. A particular watershed development was the development and

adoption of the tractor in the 1920s that heralded a decline in the use of animals for farm labor (Hilimire, 2011). This reduced the need for individual farms to raise both livestock and crops, and reduced the need for production of forage crops. Transportation technologies also allowed for regional specialization in agricultural production as goods could be cheaply transported over larger distances. Additional technological developments included synthetic fertilizers and pesticides. These further reduced the need for individual farms to raise both livestock and crops, and reduced the need for crop rotations and legumes for nitrogen fixation, as fertilizers could be used to meet crop nutrient needs, and pesticides could be used to manage weed and insect pests. Continued technological developments, including the development of high yielding, pest-resistant, and herbicide tolerant crops further contributed to feasibility and desirability of specialized and intensified agricultural practices that accompanied global patterns of development and have come to be termed "conventional".

The resultant patterns of increasing specialization and decreasing integrated mixed farming systems have varied widely, perhaps reflecting transitions between alternative sets of divergent competing paradigms that are thought to define and differentiate conventional from non-conventional or "alternative: agriculture (Beus and Dunlap, 1990). Although many of the methods and approaches characteristic of conventional agriculture are of ancient origin (Merrill, 1983), conventional agriculture has come to be associated with intensively managed, specialized management systems that evolved primarily in an attempt to increase production efficiency and economic returns. Often, the goals of increasing production or reducing the risk of crop failure were achieved by focusing management efforts on a single or just a few factors selected to produce the greatest increases in crop growth or yield. Emphasis on increasing productivity through management of a critical limiting resource such as water or nitrogen, guided the development of management strategies and technology, and resulted in significant increases in productivity, but did so at the expense of other ecosystem services. Ecosystem services can be classified into four major categories; provisioning, regulating, supporting, or cultural services and are those indirect and direct benefits to humans obtained from ecosystems (Millennium Ecosystem Assessment, 2005). For this discussion we focus on provisioning services such as food, forage, fiber and fuel; supporting services such as biodiversity, soil fertility and nutrient cycling; and regulating services such as habitat for wildlife, insects, and pollinators, and the regulation of water quality, greenhouse gas emissions, carbon sequestration, and soil loss due to erosion. Primary focus on management practices that focused on provisioning, and were highly responsive to market demand and other economic pressures, favored economies of scale and tend to favor specialize farming systems (Peyraud et al., 2014).

Organic farms are examples of farms that are typically integrated, with producers utilizing crop rotations, cover crops, and integration of crops and livestock to eliminate the need for synthetic fertilizers and pesticides. Organic farming has increased steadily in the U.S. over the last 25 years. This is associated with increased demand for organic products and resulting persistent price premiums for organic crops that improve the profitability of organic systems. Even among farms that are not certified organic, there has been resurgence in interest in integrated farming in the U.S. This is reflected in recent interest and emphasis on soil health, and practices that can build soil health, such as reduced tillage and diversified crop rotations, with a particular emphasis on the use of cover crops. Increased use of cover crops has encouraged producers to consider bringing livestock to farms that have specialized in crop production only.

Drivers

Several types of drivers influence the adoption of integrated farming systems including economic, environmental, and social (Hendrickson et al., 2008). Economic drivers include production complementarities, risk, and management requirements.

Integrated farming systems have economic benefits through economies of scope. Economies of scope occur when it is cheaper to produce two or more products simultaneously than producing them separately. Integrated crop-livestock systems reduce production costs due to complementarities in production such as use of grain screenings or crop residues for animal feed and subsequent application of livestock manure to land as fertilizer. The benefits of economies of scope tend to be more pronounced for small farms than for large farms, so there are stronger incentives for integration in small farms.

Farmers are generally risk averse and favor practices that reduce risk exposure. Integrated farming systems may reduce economic risk through diversification of production across multiple enterprises, when economic returns for these enterprises are imperfectly correlated. Integrated systems, however, increase management complexity due to the need to understand and coordinate multiple enterprises. This can be an important barrier to the adoption of these systems, and exemplifies the concept of bounded rationality (Simon, 1957). Productivity and economic performance can be reduced if administrative decisions are limited by managers' ability to process and act on complex information. Thus, if integrated systems are more complex, this can decrease productivity and create incentives against integration (Chavas, 2008).

While integrated agricultural systems have many benefits, there is a current trend toward specialized and large scale farms, reducing the number of farmers and increasing total assets through mergers and consolidations. This is particularly apparent in developed countries where the benefits gained through economies of scope decrease with increased farm size. Large farms may benefit from specialization, which is associated with economies of scale. Economies of scale exist if the per unit cost decreases with an increased farms size by spreading incremental cost over more production units. Specialized and large scale farms characteristically use technology intensive production systems. Intensive use of technology in agriculture may lead to a transition from within-farm integration to among-farm integration, and favor large agribusinesses that reduce cost through economies of scale and through higher market power and market access. Policy makers and large agribusinesses may prefer large systems as it reduces the number of stakeholders they must deal with and reduces transaction costs (Russelle et al., 2007).

With regard to environmental drivers, two main drivers for integration in North America are environmental concerns associated with excess nutrients produced by intensive livestock production, and high energy inputs for mono-cropping system (Russelle et al., 2007). Integrated farming systems are relatively self-sufficient, requiring less purchased fertilizer, herbicide, and pesticide inputs. This occurs since integrated systems utilize outputs from one enterprise as inputs to other enterprises, often use legumes in rotation reducing the need for purchased nitrogen, and include diverse crop rotations that help reduce weed, insect, and disease pressures. These reductions in purchased inputs also help to diminish production costs for integrated systems. Improved energy efficiency, realized with integrated systems, results from reductions in the use of purchased inputs, particularly applied nitrogen fertilizers, but, in some cases from lower tillage use as a result of less fuel needed to power equipment.

Integrated farming systems are thought to enhance a wide range of environmental services due to the higher level of diversity in these systems compared to specialized monocropping systems, and, in turn, reduced adverse environmental impacts. Examples of environmental services positively impacted by integrated systems include reduced soil erosion, increased soil organic carbon, improved soil biological function, reduced irrigation water use, and improved water quality (Sanderson et al., 2013). As a result, these systems are considered to be better for public health (Vereijken, 1989). This is a significant social driver and may be economically beneficial for producers since consumers are willing to pay price premiums for high quality goods that are perceived to be more beneficial due to growing awareness and concern for health, environment, social responsibility, and animal welfare, particularly in developed countries (USDA-ERS). In addition to the offsite effects, improved environmental services on integrated farms include increased soil fertility, and enhanced water quantity and quality, resulting in higher crop and livestock productivity, and increasing profitability in the long run. While the economic impacts from these farm resource conservation benefits may not be apparent in the short-term, they are important for maintaining long-term sustainability (Russelle et al., 2007).

Integrated Farming Examples

Integrated farming is a common practice in developing countries where farmland acreage, and access to manufactured fertilizers and agrochemicals are limited. In many developing countries land holdings are typically less than 1 ha. In comparison, average U.S. farm size in 2011 was 95 ha (234 ac) of cropland, but the median was just 18 ha (45 ac), indicating a large number of small farms. However, there were also many large farms with half of the cropland accounted for by farms 445 ha (1100 ac) or greater, and the size of this 'midpoint farm' had nearly doubled since 1982 (MacDonald et al., 2013). For the purposes of this chapter, we define small-scale systems as farms less than 40 ha (100 ac) in size, large-scale systems as farms between 40 ha and 800 ha (100–2000 ac), and very large-scale systems as farms larger than 800 ha (2000 ac). The examples given here are not meant to encompass all types of integrated approaches, but rather to highlight a few practices that are more common to or increasing in popularity within the farm scale in question.

Out of necessity, small-scale farms lend themselves to integrated practices that rely heavily on synergistic relationships between system components while maximizing production per unit of land. In developed countries, the local foods movement has, in part, driven a more deliberate approach to integrated farming in small-scale systems that attempt to mimic natural ecosystem processes. Small-scale systems, more so than large-scale, utilize manual harvesting and weeding in their operations. This allows for more flexibility with regards to the type of and extent to which integrated practices can be applied. At this scale, intercropping, cover-cropping, crop rotations, crop-livestock integration, and other cultural practices such as soil solarization, whereby solar energy is trapped underneath transparent plastic to kill weed seeds and harmful soil organisms and soil-borne pathogens, and the use of on- and off-farm sources of livestock manure and other soil amendments are all feasible. For example, low-input and organic systems, in particular, rely more heavily on approaches such as intercropping at various spatial and temporal scales and cover cropping to introduce biodiversity into the system.

Intercropping has long been used by small farmers around the world to satisfy dietary needs, spread labor requirements, reduce crop failure risks and produce more food per unit of land. Prior to the 1940's when mechanization, low-cost fertilizers and pesticides, and improvements in plant breeding were not yet available, intercropping was also practiced in the United States and Europe. Intercropping is the practice of growing two or more crops simultaneously so that positive plant interactions can be exploited and available resources can be utilized more completely and efficiently. In many instances, it is also a method used to reduce weed and insect pest pressure, and the spread and severity of disease. There are several types of intercropping, which include mixed intercropping where crops are completely mixed, strip intercropping where crops are arranged in alternate rows, alley cropping where crops are sown between tree rows in orchard or timber systems, and relay cropping where one crop is planted in the same field of another crop that has completed its growth. The goal of this approach is to reduce labor and the need for external inputs such as fertilizers and pesticides. Although not inclusive of all intercropping systems used around the world, examples of some common intercropping systems include cassava- or banana-based systems intercropped with coffee, cocoa, rubber, or cowpea in the tropics, maize-based intercropping with sweet potato, common bean, cowpea, or legume trees in Africa, rice-based systems intercropped with melon or maize in Asia, and the traditional three sisters system utilized by native cultures in North America, which consisted of maize-squash-beans. Companion crops in these types of systems are selected based complementarity of growth forms and their function within the system, such as a nitrogen-fixation by legumes, and on harvest timing compatibility. While not necessarily unique to small-scale systems, intercropping is not as widely practiced in large-scale systems due to management complexities and mechanization considerations. With large-scale farms, intercropping must be specifically designed to allow for

mechanization. Often, this means selecting crops that mature at similar times, and thus can be harvested at the same time. Also, this means selecting crops that can be easily separated at or after harvest, or utilized together as for food, seed, feed or forage. Also intercropping at the large scale often includes a legume to provide nitrogen fixing benefits. Some large-scale intercropping examples that exhibit these characteristics include canola/field pea, canola/spring wheat/field pea, winter wheat/hairy vetch, and oat/field pea.

The adoption of cover crops has become increasingly popular regardless of farm type or farm size. Annual crop rotations and cropping systems that integrate cover crops can provide a number of ecosystem services. Multi-species cover crop mixtures are often promoted as a way of introducing diversity and redundancy into the system to be more resilient to abiotic stressors. The functional diversity concept from ecology is often easiest applied to cover crops as species representing broad categories of plant types are often included. Examples of functional group categories include warm-season grasses, warm-season legumes, cool-season legumes, warm-season broadleaf plants, and cool-season broadleaf plants. Species are selected based on site-specific problems and priorities. For instance, cover crops such as radish and turnip are used as a biological method of aerating soil in no-till cropping systems and as method of increasing biological activity and nutrient cycling. Phacelia, buckwheat, and mustard are examples of crops used in mixtures in areas where pollinator resources are a priority due to floral qualities that attract pollinators. Grass cover crops such as rye, wheat, and millet are popular for their ability to compete with weeds and reduce weed pressure. Crops with deep rooting characteristics such as sunflower and sorghum-sudangrass are utilized for their ability to scavenge excess nutrients from deeper soil depths, thereby reducing soil nutrient leaching. Sorghum-sudangrass and other grasses are also examples of high-biomass cover crops that build soil organic matter and contribute to enhanced soil fertility over the long-term. There are many other examples and uses for cover crops, such as forage for livestock in integrated crop-livestock systems. Additionally, there may be trade-offs between the types of crops used and the types of and extent to which ecosystem services are provided (Schipanski et al., 2014). For instance, a cover crop mixture that is designed to maximize pollinator resources may provide poor quality forage for livestock in an integrated system.

Another example of an integrated approach unique to small-scale farms is the increasingly popular practice of aquaponics. Typically associated with greenhouse or other controlled environment production systems, aquaponics is the combination of fish culture (aquaculture) and soilless plant production (hydroponics). In this type of production system, nutrients derived from fish waste, with tilapia being the most common fish species used, are recirculated through the system and utilized by plants to meet their nutrient requirements. Typically, large quantities of fish are raised in small volumes of water to allow an accumulation of non-toxic nutrient concentrations (Rakocy et al., 2006). Solids from fish effluent are removed through filters and depending upon the type of aquaponics system, effluent may go through additional treatment processes to remove toxic waste products. Microbes break down remaining waste and convert nitrogen into a form that can then be taken up by plants. If optimal ratios are achieved, this approach creates a balanced system that intensifies production while minimizing adverse environmental impacts and reducing production costs.

While integration of crop-livestock is not unique to small-scale farms, small-scale systems typically allow for more diverse types of livestock to be utilized. For instance, chickens are more easily integrated into small-scale cropping systems with the use of mobile chicken coops that can be rotated across different areas on the farm. These so-called animal tractors, which can also include other types of livestock rotated in pens or paddocks, allow manure to be spread throughout the farm, which supplies fertilizer for crops. These types of systems also provide additional benefits to the system such as soil aeration through scratching and chicken pecking behavior, weed seed and insect pest consumption, and recycling of plant materials. Although less common in large-scale farms, additional livestock are sometimes included on a portion of the farm as available management and labor allow.

Regardless of scale, the most common form of crop-livestock integration occurs in forage systems whereby livestock graze on crop residues, cover crops following an annual crop, annual forages, improved pastures, and forages in alley-cropping systems. In under-developed countries, livestock commonly continue to provide energy with which to power farm equipment such as cultivators and plows. Livestock provide additional benefits by grazing on weeds, thereby reducing reliance on herbicides. As in animal tractor systems, livestock grazing provides direct manure application into the system, reducing the need for external fertilizer inputs. That is not to say that externally sourced animal manures and other soil amendments may not be transferred from off farm. For instance, many large- and very large-scale chicken, dairy, and mushroom production facilities sell composted organic waste from their operations to nearby farms at relatively low cost. This significantly reduces waste from large- and very large-scale farms while providing a locally-sourced, inexpensive and nutrient-dense fertilizer supply for farming operations.

In very large-scale farms, integration often tends to be less closely synchronized in time and space than in smaller-scale farms. Examples from the U.S. Corn Belt include farms where grain and forages are produced on the farm and fed to dairy or beef cattle in a feedlot on the farm. Often, some of the corn production is sold to a nearby plant for ethanol production, and resultant distillers' grains are fed to the livestock. Manure from the livestock is applied to the crop fields to recycle nutrients. Methane is captured from manure storage areas and is used to generate heat or electricity for use on the farm. This type of integration can occur on a single very large-scale farm, or regionally across multiple farms. Often, on a very large-scale farm, there may be multiple managers to divide management of separate enterprises. For example, the crop and livestock enterprises would often have separate managers. Integrated systems are well suited to farms where multiple managers (often family members) have different interests and abilities, so management of diverse enterprises can be subdivided. However, coordination among the multiple managers can represent an additional challenge.

Challenges and Opportunities to Integrated Farming

Despite the benefits of integrating farming, there has been a trend toward increased agricultural specialization in developed countries, whereby crop and livestock enterprises have become increasingly disconnected (Wilkins, 2008; Hilimire, 2011). Although mixed crop-livestock systems continue to dominate the landscape in Australia, there has been a shift toward crop-only systems over recent decades. However, some of this trend is attributed to relative crop and livestock prices and could be slowed or reversed if relative prices change (Bell and Moore, 2012). In places where agriculture has become more specialized, additional capital investments may be necessary if re-integration is to occur, and this may serve as a barrier to adoption of these systems. Integrated systems often require more labor, and may thus be challenged if there is a decreasing labor supply in agricultural sectors or increasing labor costs. While there has been renewed interest in integrated farming systems in the U.S., it remains to be seen whether this results in an increase in these systems. The current interest in cover crops, which serve a number of functions including forage in crop-livestock systems, has led to research and development of new technology such as specialized equipment for interseeding cover crops into existing cash crops. Such innovation is necessary utilize cover cropping in large scale farms, and will help to foster adoption of these systems by allowing integration to occur while retaining scale economies. However, the costs of these new technologies can serve as barriers to adoption.

Other technology trends in agriculture include continued advancement of precision agriculture, remote sensing, and information technologies. These technologies promote more efficient use of purchased inputs, and have helped producers to manage larger farms but may tend to reduce the relative benefits of integrated systems, and lead to further specialization. However, improving information technologies could help managers deal with the complexities of integrated systems and reduce an important barrier to their adoption. One of the many challenges for agricultural scientists will be to meld new and emerging technologies with integrated farming approaches in order to exploit the benefits of each approach and maximize efficiency.

An important driver for the adoption of integrated systems is the desire to restore the function of degraded soils. Biological soil amendments are an emerging technology purported to enhance soil biological function, and thereby improve nutrient cycling and pest suppression. These amendments would replicate some of the improvements in soil health attributed to integrated agriculture practices and might tend to reduce the incentives for integration. Other future trends that may influence adoption of integrated systems are increasing energy costs and the emergence of pesticide resistant pests. Both may favor the adoption of integrated systems which can reduce the use of energy intensive inputs, help reduce the development of pesticide resistance, and provide a suite of alternative methods for managing pesticide resistant pests.

Future market demands can also have a significant effect on integrated systems. The need to meet food demands of a growing population, but with limited land availability, will require increasing food production per unit land area in a sustainable manner. An important component of this is the variation in the suitability of land for grain and forage production and the need to match the mix of enterprises to land suitability. Farmers are more likely to adopt integrated crop-livestock practices in specialized cropping areas on marginal lands where productivity is low due to natural soil or environmental factors. Integrated farming systems are well suited to heterogeneous landscapes, where different portions of the landscape are better suited to different enterprises. While this can result in non-integrated enterprises within a farm, such as when crops are grown on the most productive land, and livestock are produced on less productive land, the proximity of these enterprises within a farm allow for greater opportunities to share resources among enterprises through integration. Integrated systems can often increase productivity per unit land area, and they can also be used to shift more livestock production toward forages, thereby freeing up grain production for human consumption (Bell and Moore, 2012).While meeting future food needs is important, this will need to be accomplished while maintaining or enhancing ecosystem services. The emerging market for sustainably produced foods may help provide economic incentives for integrated systems which enhance ecosystem services.

Several methods have been proposed for overcoming challenges and increasing the adoption of integrated systems. These include the use of environmental regulations and/or government payment to provide incentives for reducing environmental impacts, which may reward integrated farming producers. Since integrated farming systems are complex, farmers' ability to gain benefits through synergies is an important factor for enhancing the adoption of integrated systems. Farmers' management knowledge and skill can be improved through training and research at the local level. Combining increased knowledge and skills with technologies that help managers deal with complexity may facilitate greater adoption of these systems. Future developments in technology, regulations, labor availability, and demands for food and ecosystem services will influence the adoption of integrated farming systems, which may play a key role in meeting future food needs while maintaining or enhancing ecosystem services.

References

Bell LW and Moore AD (2012) Integrated crop-livestock systems in Australian agriculture: trends, drivers and implications. *Agricultural Systems* 111: 1–12.

Beus CE and Dunlap RE (1990) Conventional versus alternative agriculture: the paradigmatic roots of the debate. *Rural Sociology* 55: 590–616.

Chavas JP (2008) On the economics of agricultural production. *Australian Journal of Agricultural and Resource Economics* 52(4): 365–380.

Devendra C and Thomas D (2002) Crop–animal interactions in mixed farming systems in Asia. *Agricultural Systems* 71: 27–40.

Hendrickson JR, Hanson JD, Tanaka DL, and Sassenrath G (2008) Principles of integrated agricultural systems: introduction to processes and definition. *Renewable Agriculture and Food Systems* 23(04): 265–271.

Hilimire K (2011) Integrated crop/livestock agriculture in the United States: a review. *Journal of Sustainable Agriculture* 35: 376–393.

MacDonald JM, Korb P, and Hoppe RA (2013) *Farm size and the organization of U.S. crop farming. ERR-152*. Washington, DC: U.S. Department of Agriculture, Economic Research Service.

Merrill MC (1983) Eco-Agriculture: a review of its history and philosophy. *Biological Agriculture & Horticulture* 1: 181–210.
Peyraud JL, Taboada M, and Delaby L (2014) Integrated crop and livestock systems in Western Europe and South America: a review. *European Journal of Agronomy* 57: 31–42.
Rakocy JE, Masser MP, and Losordo TM (2006) *Recirculating aquaculture tank production systems: aquaponics – integrating fish and plant culture*. Stoneville, MS: Southern Regional Aquaculture Center. Publication No. 454.
Russelle MP, Entz MH, and Franzluebbers AJ (2007) Reconsidering integrated crop–livestock systems in North America. *Agronomy Journal* 99(2): 325–334.
Sanderson MA, Archer D, Hendrickson J, Kronberg S, Liebig M, Nichols K, Schmer M, Tanaka D, and Aguilar J (2013) Diversification and ecosystem services for conservation agriculture: outcomes from pastures and integrated crop-livestock systems. *Renewable Agriculture and Food Systems* 28(2): 129–144.
Schipanski ME, Barbercheck M, Douglas MR, Finney DM, Haider K, Kaye JP, Kemanian AR, Mortensen DA, Ryan MR, Tooker J, and White C (2014) A framework for evaluating ecosystem services provided by cover crops in agroecosystems. *Agricultural Systems* 125: 12–22.
Simon HA (1957) *Models of man, social and rational: mathematical essays on rational human behavior in a social setting*. New York: John Wiley and Sons.
Sumberg J (2003) Toward a dis-aggregated view of crop–livestock integration in Western Africa. *Land Use Policy* 20: 253–264.
Vereijken P (1989) Experimental systems of integrated and organic wheat production. *Agricultural Systems* 30(2): 187–197.
Wilkins RJ (2008) Eco-efficient approaches to land management: a case for increased integration of crop and animal production systems. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363(1491): 517–525.

Further Reading

Lantinga E and Rabbinge R (1996) The renaissance of mixed farming systems: a way toward sustainable agriculture. In: *Book of Abstracts 4th Congr. European Society for Agronomy*, pp. 428–429. Veldhoven, The Netherlands: European Society for Agronomy.
Van Keulen H and Schiere J (2004) In: *Crop-livestock systems: old wine in new bottles*, *Proceedings of the 4th International Crop Science Congress, Brisbane, Australia*.

Relevant Website

https://www.ers.usda.gov—United States Department of Agriculture, Economic Research Service.

Island Biogeography

TW Schoener, University of California, Davis, CA, USA

As MacArthur and Wilson note in their revolutionary book *The Theory of Island Biography*, Charles Darwin was among the first to call attention to islands as a crucial subject for scientific study. As he was departing the Galapagos Islands in 1835, Darwin wrote

> when I see these Islands in sight of each other, & possessed of but a scanty stock of animals, tenanted by these birds, but slightly differing in structure & filling the same place in Nature, I must suspect they are only varieties … If there is the slightest foundation for these remarks the Zoology of Archipelagos will be well worth examination; for such facts would undermine the stability of species.

Of course Darwin's main interest was in evolution; we know now that islands can contribute enormously to our understanding of ecology as well. There are at least seven reasons why this should be so:

1. Islands are `simple', with relatively few species and habitats so that the often notoriously complex web of ecological relationships is manifested in a more easily understood state.
2. Islands are `discrete' and often small, providing a well-defined, manageable spatial unit for study.
3. Islands are `isolated', so that they are relatively immune from outside fluxes, much as a laboratory tank or greenhouse.
4. Islands are `natural', so unlike those human-constructed containers have biotas adjusted in at least the moderately long term to a confined state.
5. Islands are `combinatorial', so that species occur in various combinations, often with absences and presences of key species varying as if in experimental removals or introductions, allowing comparative inference of those species' effects.
6. Islands are `replicated' so that often a number are available with a given set of species or environmental conditions, allowing statistical analysis and experimentation.
7. Islands are `ubiquitous', not being limited to areas of land surrounded by bodies of water but generalizing to any patch of one type of habitat surrounded by another – mountaintops, river systems, hosts for parasites, etc. Indeed, as MacArthur and Wilson also pointed out, natural habitat is becoming increasingly fragmented, that is, insularized by human activities so that many areas of conservation interest are effectively archipelagal.

Finally, were such scientific rationale not enough, islands are esthetically pleasing, enamoring us all, scientists or not, with their charm and beauty (could this be the real reason many ecologists work on islands?).

The MacArthur–Wilson Equilibrium Theory

What was the revolution that MacArthur and Wilson effected in their 1967 book? The essence is that each island is in a state of species equilibrium, in which the number of new (not already on the island) species immigrating per unit time is balanced by the number of species becoming extinct per unit time. To see how this could be very different from previous theoretical constructs, we can examine the effect of an island's distance on its species count. Extensive collection of information on species distributions and compilation of faunal lists shows that number of species on far islands (those distant from a source of colonizing species) is less than the number on otherwise similar near islands. One possible explanation is that the far islands, so removed from the source, simply did not have time to acquire the number of species near islands possess but eventually would. An entirely different explanation was advanced by MacArthur and Wilson, who argued that far islands have fewer species because their low immigration rate is balanced at equilibrium by a low extinction rate; the fewer species on far islands means fewer to go extinct per time, thus achieving a balance at a smaller equilibrium number.

The model actualizing these ideas was first presented as a graph of gross extinction and immigration rates against the number of species present on the island. In its most general form it makes two assumptions (**Fig. 1a**):

A1. Rate of immigration of new species (those not yet on the island) decreases monotonically with increasing number of species already present. It reaches zero when all species in the source area (there are P of them) are on the island; and

A2. Rate of extinction of species increases monotonically as the number of species increases (the more species there are, the more to go extinct). These two assumptions imply

P1. An equilibrium between immigration and extinction will eventually occur, at which time the immigration and extinction rates will equal the same value, called the turnover rate at equilibrium.

Two additional assumptions allow some more predictions:

A3. Near islands have immigration rates higher than far, for the same number of species present (because near islands are closer to the source area); and

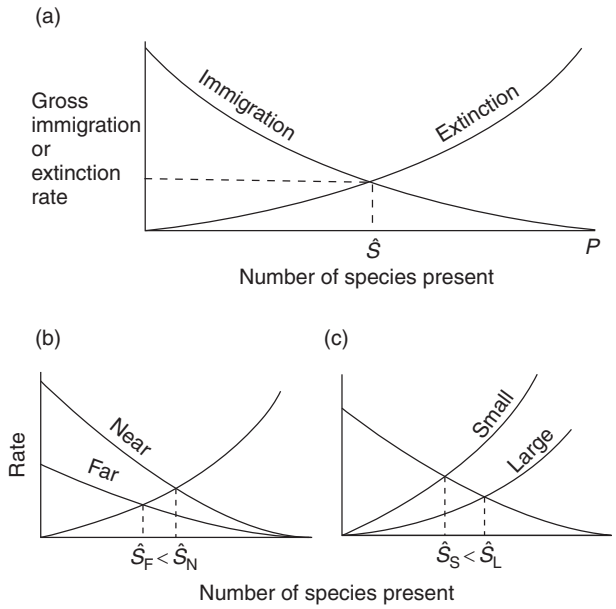


Fig. 1 (a) The graphical version of the MacArthur–Wilson equilibrium model. The model is for a particular island. $\hat{S}$ is the number of species at equilibrium (when gross immigration equals gross extinction), and P is the number of species in the source pool. Rate curves are monotonic but nonlinear. The intercept of the dashed line on the y-axis is the turnover rate at equilibrium. (b) The distance effect for the MacArthur–Wilson equilibrium model: far islands have lower immigration rates than near islands (for a given number of species present on the island), implying near islands at equilibrium have more species ($\hat{S}_N$) than do far islands at equilibrium ($\hat{S}_F$). (c) The area effect for the MacArthur–Wilson equilibrium model: small islands have higher extinction rates than large islands (for a given number of species present on the island), implying large islands at equilibrium have more species ($\hat{S}_L$) than do small islands at equilibrium ($\hat{S}_S$).

A4. Small islands have extinction rates higher than large, for the same number of species present (because average population size is smaller for the smaller islands, so the per-species extinction likelihood is greater). These imply the following predictions:
P2. Near islands of the same size as far have more species (**Fig. 1b**); and
P3. Large islands of the same distance as small have more species (**Fig. 1c**). This prediction is a version of the species–area relation, one for which there is much evidence, as discussed below.

Finally, certain assumptions lead to
P4. the colonization curve, or the curve relating number of species on the island to time since the colonization process begins, is convex.

Although the relation of extinction to area and the relation of immigration to distance are expected to be by far the dominant ones, the two other logically possible relations have also been proposed and documented. First, far islands may have extinction rates higher than near for the same number of species present (Brown and Kodric-Brown's `rescue effect'). The rationale is that near islands have populations supplemented by immigration more than do far islands; thus population sizes of the species present on near islands average larger so are less extinction-prone than populations on far islands (and/or near-island populations have more genetic variation, so again are less extinction-prone). Second, large islands may have greater immigration rates than small, for the same number of species present (the `target effect'). The rationale is that larger islands have a greater diameter or area to intercept laterally moving or vertically descending immigrants, respectively, so have a greater immigration rate. Evidence for both of these effects is widespread and will be discussed below, but we note here that all four postulates about rates, (A3) and (A4) and the two just mentioned, have been detected in the same system, spiders on subtropical islands of the Bahamas.

Tests of Species Equilibrium

A major claim of the MacArthur–Wilson theory was that a substantial portion of the world's islands are in fact at equilibrium, whereby the number of species colonizing balances the number becoming extinct. The evidence for a species equilibrium ranges from highly supportive to contradictory; we review the major cases in decreasing order of support.

1. *Birds of Krakatau*. In 1883 a tremendous volcanic eruption destroyed two-thirds of the Indonesian island of Krakatau and left its remnants and two neighboring islands under 30–60 m of ash, with no observable plants or animals surviving. As major confirmation of their theory, MacArthur and Wilson argued that equilibrium for land birds was re-attained only 25–36 years after the eruption. However, subsequent studies showed that the conclusion was premature: equilibrium has perhaps not yet

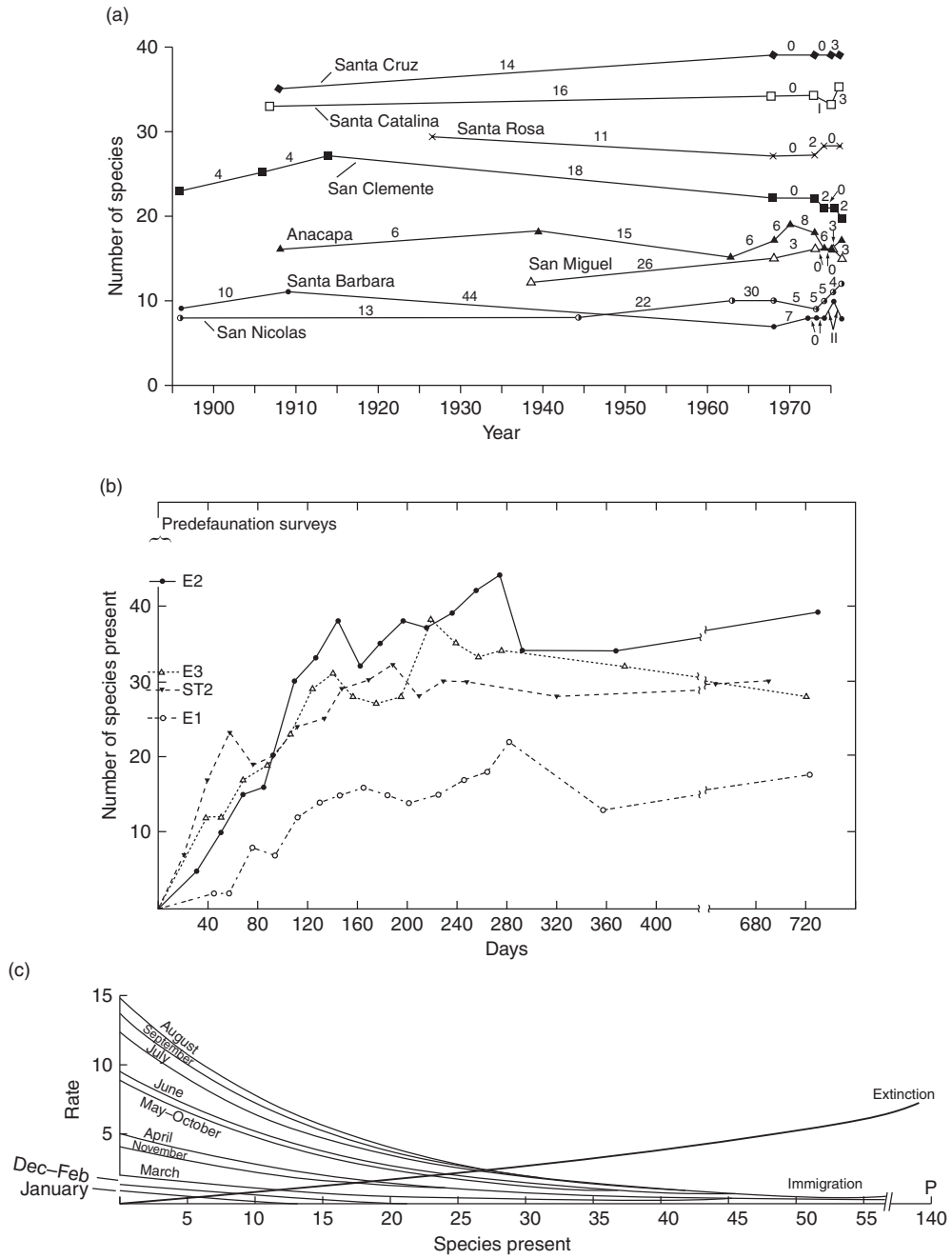


Fig. 2 (a) Birds on the Channel Islands, California (USA). Number of breeding species *S* for each island plotted against survey year. The number written over the line connecting each pair of points is the percent turnover between those surveys. (b) Colonization curves of four mangrove islets, Florida (USA). E2 is the nearest island and E1 is the farthest island. (c) Oscillating equilibrium in marine epifaunal invertebrates on rocks. Gross immigration and extinction rate curves (top) and colonization curves (bottom). (d) Species number vs. time, that is, colonization curves (left), and gross immigration or extinction curves vs. average species number in intersurvey interval for seed plants of Krakatau (Rakata) (right). (a) Reproduced from Jones HL and Diamond JM (1976) Short-time-base studies of turnover in breeding bird populations on the California Channel Islands. *Condor* 78: 526–549, with permission. (b) Reproduced from Simberloff D and Wilson EO (1970) Experimental zoogeography of islands. A two-year record of colonization. *Ecology* 51: 934–937, with permission. (c) Reproduced from Osman RW (1977) The influence of seasonality and stability on the species equilibrium. *Ecology* 59: 383–399, with permission. (d) Reproduced from Thornton JWB, *et al.* (1993) Colonization of Rakata (Krakatau Islands) by nonmigrant land birds from 1983–1992 and implications for the value of island biogeography theory. *Journal of Biogeography* 20: 441–452, with permission from Blackwell Publishing Ltd.

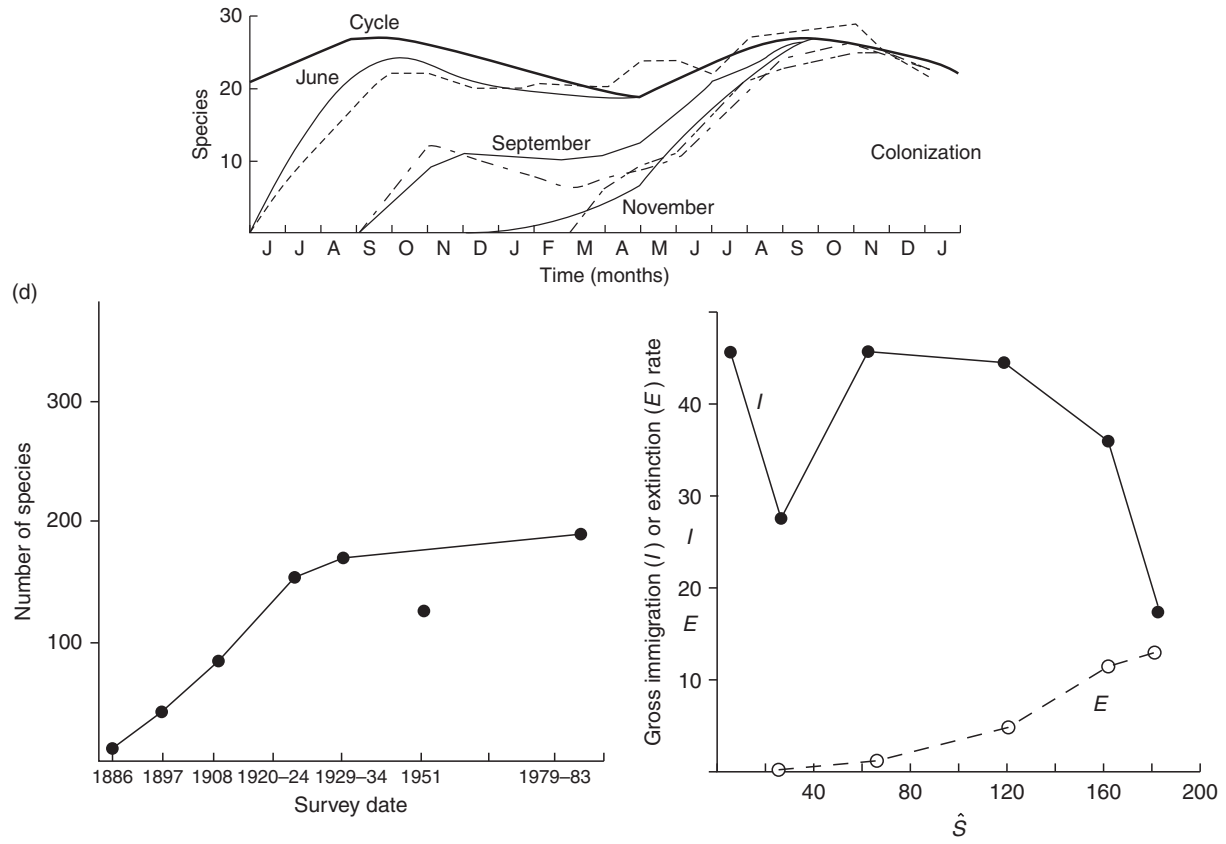


Fig. 2 (Continued)

been quite reached even after 100 years after the eruption. Much smaller gross extinction rates are now known to occur as compared to those predicted by MacArthur and Wilson (0.25–0.42% vs. 1–6% species per year).

2. *Birds of the Channel Islands*. The first systematic survey of birds on nine California Channel islands was accomplished in 1917. Since that time several surveys showed that the number of species over 50 + years has not changed much (**Fig. 2a**). The degree to which the islands showed turnover was controversial and will be discussed in the next section.
3. *Arthropods of Red Mangrove Islets*. The red mangrove *Rhizophora mangle* grows as an emergent shrub or tree from small floating dispersal structures that root on shallow marine banks; islands so created can range from moderately large down to extremely small. Investigators took advantage of the plethora of such islands to study experimentally the recolonization by arthropods that would occur after a devastating extinction event. A pest-extermination company was hired to cover each mangrove islet with plastic sheeting and spray insecticide within; this eliminated most of the arthropods, so that when the sheeting was removed, artificially `defaunated' islands resulted. Numbers of species rapidly increased, slightly overshooting the old equilibrium values before settling near it 1–2 years after inception of recolonization (**Fig. 2b**).
4. *Marine epifaunal invertebrates on rocks*. To simulate colonization of rocks in the intertidal, artificial panels were set out in the Massachusetts (USA) subtidal. This experiment produced an oscillating equilibrium (**Fig. 2c**), not unexpected in this highly seasonal environment. Thus the species equilibrium may manifest itself as a rather predictable oscillation, rather than necessarily having a constant value.
5. *Plants on Krakatau*. Plants showed a much slower rate of recovery after volcanic eruption than did birds (`1' above). In MacArthur and Wilson's original analysis, there was no real tendency for the colonization curve for plants even to begin leveling off. The most recent censuses (*c.* 100 years since the eruption) show that equilibrium has been nearly attained for seed plants (**Fig. 2d**) and ferns. However, like birds, the gross extinction and immigration curves are not monotonically related to number of species present as assumed by MacArthur and Wilson (**Fig. 1**); rather apparently ecological succession causes several changes in direction, especially for immigration. For example, water- and wind-dispersal plants were the first to colonize, whereas animal-dispersed species did not colonize until a certain amount of successional change had taken place.
6. *Birds on islands off Australia and New Zealand*. Records for 15 islands taken over periods ranging from 50 to 124 years showed that the number of passerine bird species increased on 14 islands, up to 900% of the original values. While humans have had effects on these islands, including habitat diversification in some cases (which would allow more species), those effects did not seem to be sufficient to account for such huge increases; perhaps climatic warming for these rather high-latitude islands was involved. The investigators characterized this variation as `nonequilibrial'; certainly it stands in contrast to Channel Island birds (`2' above), as change was quite unidirectional.

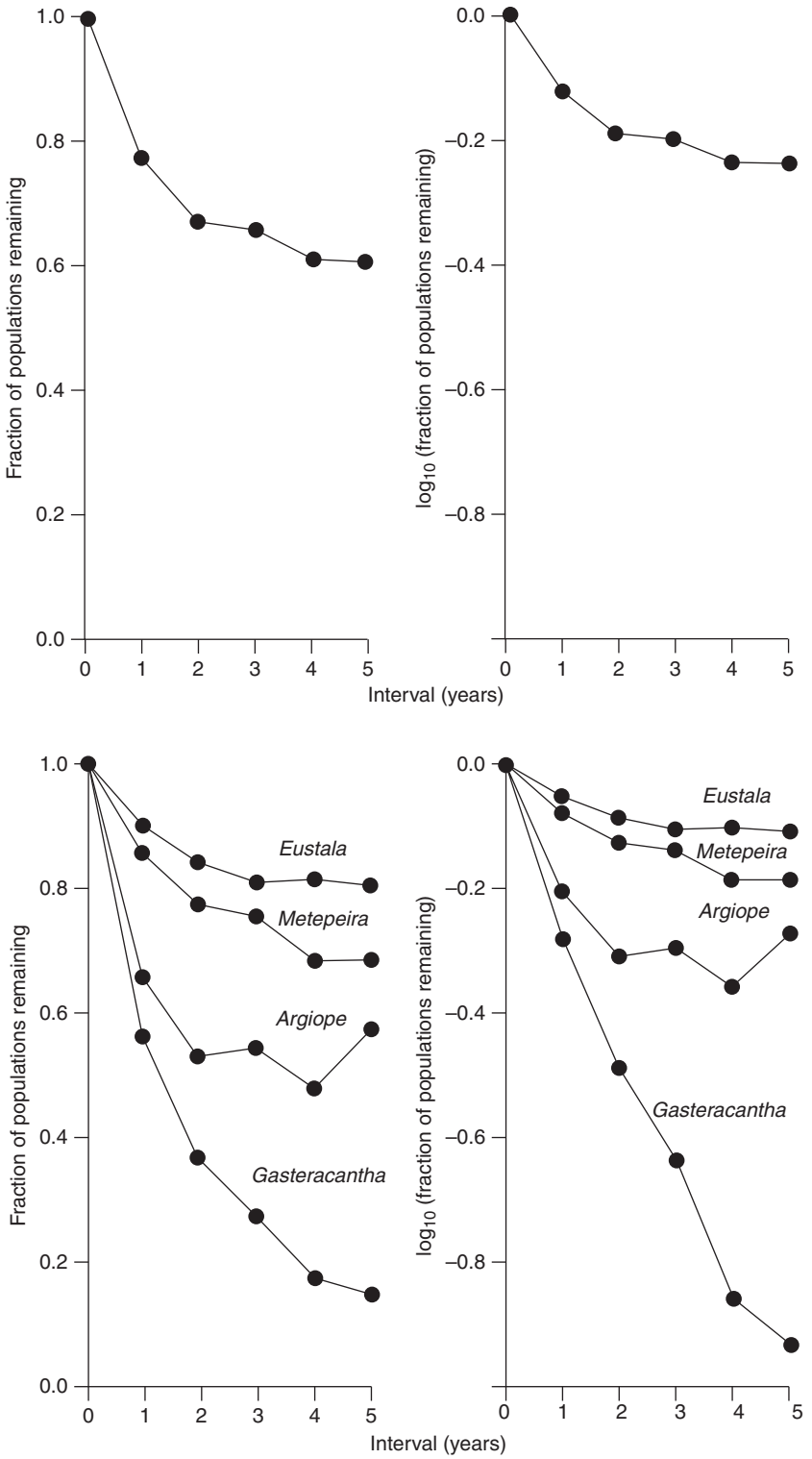


Fig. 3 Population persistence curves for orb spider on 108 islands of the Bahamas. Top: All species combined. Bottom: Individual species curves (see text for further explanation).

7. *Birds on Skokholm Island*. Data on numbers of bird species for this rather northerly island off the British mainland (taken 1928–39 and then 1946–67) showed that number of species fluctuated between 5 and 13, with substantial temporal autocorrelation. These are large percent changes (over 200% by one measure), so is this evidence against equilibrium? Certainly it is evidence against species constancy; however, MacArthur and Wilson's algebraic theory (as opposed to the graphical one

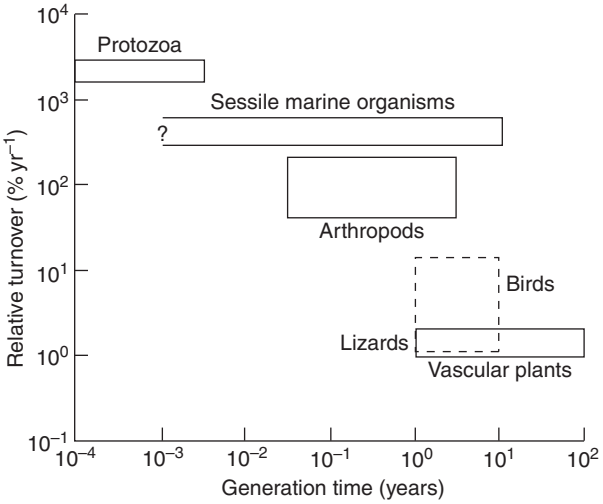


Fig. 4 Relative turnover (eqn [1]) as a function of generation time in six types of organisms. Reproduced from Schoener TW (1983) Rate of species turnover decreases from lower to higher organisms: A review of the data. *Oikos* 41: 372–377, with permission from Blackwell Publishing Ltd.

presented above) was a stochastic model with per-unit-time probabilities of immigration and extinction rather than fixed rates. Indeed, if we calculate for the Skokholm data the temporal variance and the mean number of species, the ratio is about 2/3, well within the possible range from the MacArthur–Wilson stochastic model, although higher than expected for equal (absolute) slopes of the gross extinction and immigration curves; an especially high extinction rate is consistent with the high ratio, unsurprising for such a small island.

8. *Spiders and lizards on islands decimated by hurricanes*. In 1997, certain Bahamas islands were completely inundated by the nearly 5 m storm surge of a major hurricane. As for Krakatau, no spider nor lizard individual survived; recolonization data a year later found spider species counts about where they were before the hurricane, whereas few islands to this day have been recolonized by lizards. In this case, the rapidly dispersing arthropods would be expected to reestablish equilibrium more quickly than large, terrestrial vertebrates such as lizards. Were hurricanes sufficiently frequent, certain taxa such as lizards might never reach a species equilibrium before the next disaster wiped them out again.
9. *Arthropods in soybean fields*. Soybean fields are an example of a highly temporary habitat that is frequently `defaunated', here by scheduled human activities combined with climatic seasonality. The arthropods inhabiting such communities do not have time to reach equilibrium before being `zeroed' again; hence they must constitute permanently nonequilibrium communities.

In summary, the value of the equilibrium concept varies from very useful for undisturbed islands to not so great for those that are frequently disturbed. The theory, however, does contain nonequilibrium dynamics, so that it is descriptively useful even for pre-equilibrium stages.

Tests of Turnover

Species turnover can be measured as

$$\text{Turnover (relative) over a unit item interval}(t_2 - t_1) = 100\frac{(\text{Extinctions of species already present} + \text{immigrations of new species})}{(\text{Number of species at}t_1 + \text{number of species at}t_2)} \quad [1]$$

Graphically, turnover at equilibrium is the height of the intersection of the gross rates (**Fig. 1**). That substantial turnover should exist is perhaps the most controversial part of the MacArthur–Wilson theory; museum people in particular found the concept of the species list for a given island changing much over time hard to stomach. What is the evidence for turnover?

1. *Arthropods of red mangrove islets*. Returning to the defaunation experiment (`3', last section), abundant turnover was demonstrated. Not only were there numerous extinctions and immigrations leading up to equilibrium, but once equilibrium had been achieved, species lists for particular islands were quite different from those before defaunation.
2. *Birds of the Channel Islands*. Returning to this example (`2', last section), Lynch and Johnson challenged the turnover data (first presented by Diamond) for two census times 51 years apart; among other problems they believed that species were missed in one or the other census, resulting in an inflated estimate of the degree of turnover. However, subsequent censuses by Jones and Diamond showed that turnover was in fact probably quite substantial over that period because of missed entire sequences of

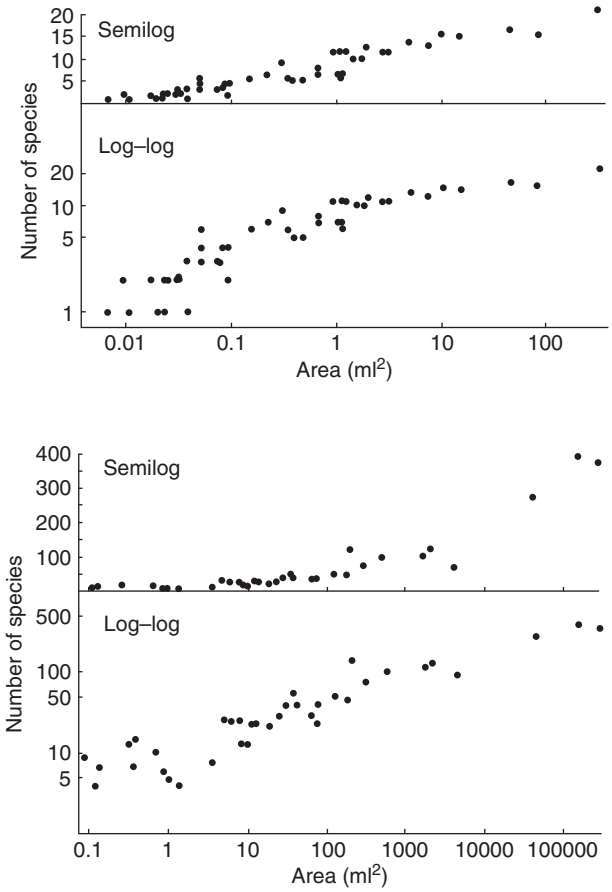


Fig. 5 Species–area plots showing the semilog (exponential) and log–log (power) relation, top and bottom, respectively. Top: Shetland land birds. Bottom: Malaysian faunal region land birds mi, miles. Reproduced from Schoener, T.W., 1976. The species–area relation within archipelagos: Models and evidence from island land birds. In: Proceedings of the XVI International OrnithologicalCongress., pp. 629–642.

immigration and extinction for particular species in the intervening half-century when surveys were not conducted. Their year-by-year data for a subsequent period in fact show turnover as 0.5–4.9% per year, whereas the two censuses in the original study gave 0.3–1.2%. Hence if the original two censuses missed species, this were more than compensated for by unobserved ins and outs during the long interval. Diamond and May presented an elegant stochastic theory predicting how `apparent' turnover (as measured eqn [1]) would decline with increasing time between censuses (between t_1 and t_2); the model successfully replicated annual breeding-bird data from the Farne Islands (another northerly group off the British mainland) and gave the result that for census intervals of decades, turnover is underestimated by about an order of magnitude.

Thus although Diamond's original conclusion for the Channel Islands was vindicated, apparently certain islands exist for which turnover is very slight. Two tropical representatives are at the extreme: Cocos Island had no turnover in 72 years, and the Tres Marias Islands had only two immigrations; perhaps tropical birds are more sedentary, thereby causing a regional difference between tropical islands and the temperate California or New Zealand islands (`6', last section).

3. *Birds on islands radically altered in area by human activity*. Various hydrological activities by humans created new islands while shrinking others in Lago Guri, Venezuela. Investigators found that a new equilibrium was achieved on the smaller remnants in just 7 years, while large islands are still declining. Similar phenomenology occurred in relation to the massive changes effected when the Panama Canal was constructed. Here as in Lago Guri, turnover was lower, the larger the island; it was also lower for far than near islands. Thus in these examples turnover is large even for tropical islands, albeit rather small, recently disturbed ones.
4. *Spiders on Bahamian Islands*. How important is turnover, in terms of the population sizes of species undergoing it? As assumed in the MacArthur–Wilson Model, there is a strong relation of population size to extinction rates in a variety of species (see final section below). Investigators calculated the percentage of individuals for all species and islands combined belonging to populations becoming extinct over particular intervals, ranging from 1 to 5 years. Using 1 year intervals, 2.8% belonged to populations becoming extinct. Using 5 year intervals, still only 4.8% did so. It seems that turnover, while quite large in terms of species number (about 35% per year), does not involve the most abundant species, those that should often be the major players and in any event are of most interest to ecosystem as opposed to biodiversity ecologists. In this system, mostly the same species go in and out, much as portrayed in Hanski's core-satellite concept. To illustrate, we can construct

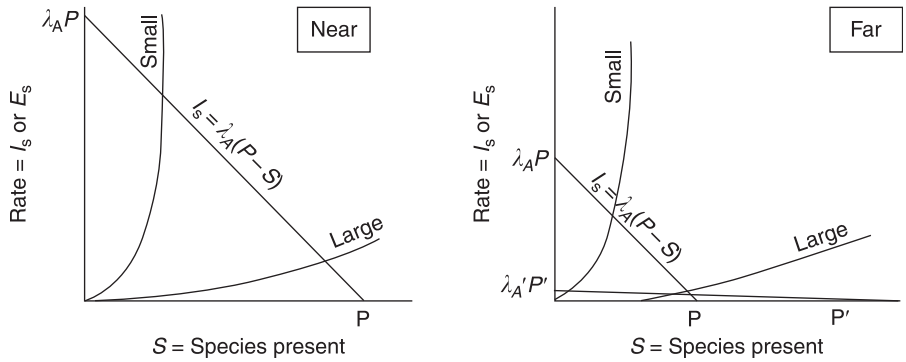


Fig. 6 Equilibrium for near and far archipelagos. Immigration for islands of a far archipelago is represented as two components: an intra-archipelagal component with low number of species in the source pool (P) and high λ_A (per-species immigration rate), and an extra-archipelagal component with high number of species in the source pool (P') but very low λ_A. Addition of this second component makes little difference except for the largest islands. Immigration for near islands has only a single component with large P and large λ_A. This single component represents the combined inter- and intra-archipelagal immigrations, here both assumed to have high (and identical) λ_A. Notice that small islands of the two archipelagos have about the same equilibrial species number, but large islands differ markedly. Reproduced from Schoener, T.W., 1976. The species–area relation within archipelagos: Models and evidence from island land birds. In: Proceedings of the XVI International OrnithologicalCongress., pp. 629–642.

population–persistence curves, which give the fraction of species populations remaining n years after a particular census (Fig. 3): note that the combined-species curve levels off quite sharply (even on a log-scale), but that particular component species vary in the degree to which this is true.

In conclusion, species turnover is a salient feature of islands. There is, however, substantial variation, not only with island size, distance and region of the world as reviewed above, but also with generation time (Fig. 4). Inasmuch as turnover involves rarer species it is of importance from a conservation view, yet perhaps equally unimportant from an ecosystem view.

Effect of Area

The species–area relation, whereby the number of species in a spatial unit increases with that unit's area, well predates the MacArthur and Wilson theory of island biogeography, having been documented for about 150 years. Two general kinds of models for this relation have been proposed. The first has number of species predicted from an assumed species–abundance distribution and the total number of individuals of all species combined (assumed proportional to area). The second develops species–area relations from MacArthur and Wilson's species-equilibrium approach.

May's paper coalesces the literature for the first sort of model. Two species–abundance distributions are of particular importance.

The first, a log-series distribution, has been used to describe light-trap data and other collections. It leads exactly to the following species–area relation:

$$S = \alpha \ln(1 + \rho A/\alpha) \quad [2]$$

where α is a parameter of the abundance distribution, and where ρ is the density of individuals. For A sufficiently large, this can be written

$$S \cong \alpha \ln(\rho/\alpha) + \alpha \ln A \quad [3]$$

Notice that this is an exponential or semilog–linear relation of species to area, that is,

$$S = c_1 + c_2 \log A \quad [4]$$

as opposed to the log–log–linear relation that the lognormal implies (below), that is,

$$\log S = \log c + z \log A \Rightarrow S = cA^z \quad [5]$$

Note also that eqn [3] is inexact for small A (or S) and that S flattens out as A approaches zero.

Various more or less plausible ways to arrive at a log-series distribution from hypothetical biological processes have been given, perhaps the most common of which is not a biological mechanism but rather a property of the sampling procedure: species–area data in which samples of different areas are taken randomly from some homogeneous large area should have a semilog–linear plot for sample areas sufficiently large.

The second species–abundances distribution, the lognormal, is expected when (1) per-individual population growth rates vary randomly over some substantial period of time or (2) the relative abundances of each species is governed by many factors acting on the per-individual growth rate (and therefore, on the logarithm of population size) independently of one another. Both follow from the Central Limit Theorem of statistics. For Preston's (the originator of this approach) one-parameter (`canonical')

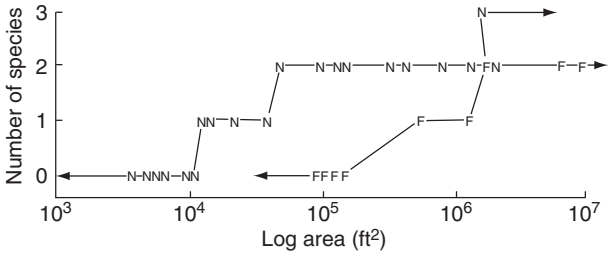


Fig. 7 Number of species of diurnal lizards on near (N) and far (F) islands in Lovely Bay, Bahamas, as a function of island area. Reproduced from Schoener, T.W., Schoener, A., 1983. Distribution of vertebrates on some very small islands. II: Patterns in species number. Journal of Animal Ecology 52, 237–262.

distribution, if we assume that J, the total number of individuals in all species combined divided by the number of individuals in the rarest species, is proportional to island area, then species number (S) increases as approximately the 0.26 power of area, that is, $S = cA^{0.26}$, a particular example of eqn [5]. However, the true relationship is not a power function but approaches one with power 0.25 as S gets large. For small S, the relation bends downward and approaches a linear relation of species to area (thus being quite different from the curves generated by the logseries distribution just discussed). Note this derivation assumes that something is constant about the shape of the distribution from small to large islands. Preston argues that what is constant is the number of individuals in the rarest species and the density of all individuals combined. It is fairly plausible that total number of individuals increases linearly with area for some well-defined taxon, although evidence bearing on this is not entirely supportive. While one study found that total density of birds increased with total species diversity, other results are more in accord with the assumption. On the other hand, while the assumption that number of individuals in the rarest species is constant is a natural one; given the mathematics of the distribution, it is perhaps less plausible biologically. For the two-parameter distribution, some other feature (e.g., standard deviation) also varies. However, the power of the species–area relation is fairly insensitive through the range of reasonable biological variation in the distribution (Engen has derived a species–area relation from yet a third species–abundance distribution, the broken stick; it is again a power function).

Which description is better, eqn [4] or [5]? Connor and McCoy interpret their review of 100 data sets to say that the two fit about equally. Clear examples of each of the two are given in **Fig. 5**.

The second type of species–area mathematical theory explicitly takes into account the ingredients of the MacArthur–Wilson equilibrium model. The model of Schoener that assumes abundances at equilibrium are complementary (summing over all species to ρA, where ρ is the density of all individuals combined and A is island area), has d log S/d log A (the slope on a log–log plot) not constant but ranging from 0.5 to 0; the midpoint of this range is very much like that given by the lognormal distribution ($z = 0.26$). The equilibrium and lognormal models differ, however, in the curvature of the species–area plot.

The equilibrium species–area model also predicts that the greater the per-species immigration rate λ_A, the smaller the d log S/d log A. Indeed z is smaller for less remote islands within an archipelago. But far archipelagoes have smaller z's than near archipelagos. This is probably because of a differentially high λ_A among birds that have been able to colonize such archipelagoes (**Fig. 6**). Various additional evidence suggests that this model is on the right track. For example, the species–area slope for birds on islands of Burtside Lake, Minnesota (USA) is unusually high, but P is very large and the islands are very small.

A final form for the species–area relation has been suggested by Lomolino and others, one having essentially an S-shape, that is., a greater rate of increase for intermediate-sized islands than either for small or large islands; note that the low slope for the smallest islands is the feature of this concept that makes it very different from any of the species–area curves proposed so far, descriptive or mechanistical. Although some evidence for such a slope was known for special cases (e.g., plants on a Micronesian atoll, in which freshwater lenses are absent on islets below a certain area), a recent survey found that the initial flat portion of the species–area curve typically included a substantial portion of an archipelago's islands. While the statistical significance of this result has yet to be evaluated, clearly its detection has proven more feasible than expected.

Although direct observation of species immigration is difficult, evidence for the target effect – larger islands intercept more immigrants – is also known from a variety of systems (e.g., Bahamian spiders, Australian sea-dispersed beach plants).

Effect of Distance

While the species–area effect is documented for nearly all quantitative studies of islands, a similar effect of distance is more rarely demonstrated. This is probably true for two reasons: first, the lack of variation in distance from an outside source for islands within single archipelagos, and second, even where a variety of distances are available, the necessity of taking into account the usually very strong effect of area before a distance effect can be detected. An example is given for lizards of the Lovely Bay archipelago in the Bahamas (**Fig. 7**): lizards are expected to be especially sensitive to variation in distance because they can only disperse by rafting (or swimming) over water, a rather tenuous process at best. Distance is not always a hindrance to species, however, and may even be an advantage. Thus migrating birds in the Bahamas tend to increase their species numbers with increasing island distance; this is

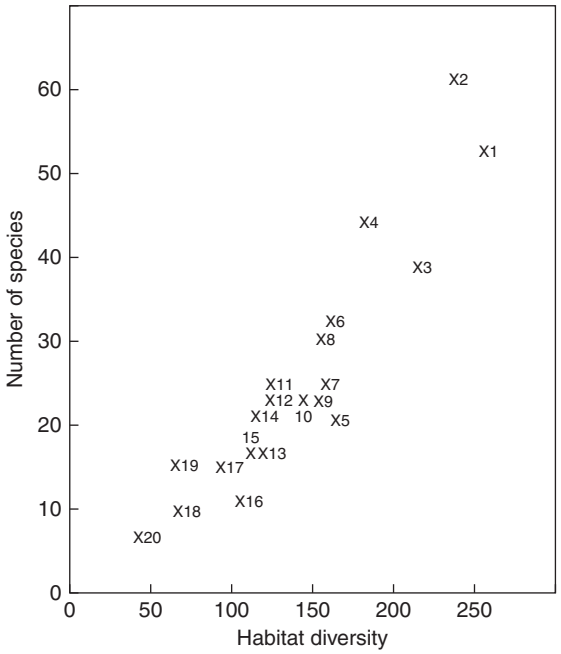


Fig. 8 Breeding passerine birds on Aegean islands: Number of species vs habitat diversity. Reproduced from Watson GE (1964) Ecology and evolution of passerine birds in the islands of the Aegean Sea. PhD Dissertation, Yale University, with permission.

the opposite of resident birds in the same region, perhaps because the latter's relative scarcity on distant islands results in more resources for migrants there.

The impact of distance is also involved in the rescue effect, whereby near islands have their populations boosted more frequently by immigration, conferring dual advantages – large population size and genetic variability – both would tend to forestall extinction. Somewhat surprisingly, few examples from water-surrounded islands are known, although more examples exist from islands *sensu latu*, so-called habitat islands (see above), including the founding study on arthropods inhabiting thistle heads.

Effect of Elevation

Perhaps the most obvious potential effect of island elevation (=altitude) on species occurrences is simple refuge from high water. Little documentation of such an effect has existed, but we now have a precise example. In 1999, a Category IV hurricane swept over small islands of the Great Abaco (Bahamas) region, completely inundating the lower ones with its storm surge. The islands were inhabited by a common lizard species, *Anolis sagrei*. Before the hurricane, island area was a better predictor (by logistic regression) of the occurrence of this species than was altitude. Immediately after, altitude was the better predictor. Apparently all lizards on islands lower than about 3 m maximum elevation had perished. After *c*. 1 year, area again became the better predictor: recovery occurred via overwater colonization (the islands were very close to the mainland) and propagation from eggs that survived inundation, mechanisms that were enhanced by a larger island area. While rapid recovery often follows catastrophic disturbance, as in this example and those given above, such is not always the case. Ricklefs and Bermingham postulated that a major change in the mean age of the Lesser Antilles avifauna occurred about 0.5 million years ago, perhaps caused by a catastrophic disturbance such as a tsunami; little postcatastrophic recovery was evident.

The most commonly discussed effect of altitude on species diversity is more indirect than simple protection from high water: the greater the altitude, the more kinds of habitats, and the greater the number of habitat kinds, the more species can occur. Examples exist for the plants of the Galapagos and, to a lesser extent, the birds of the East Indies. This argument is a special case of the argument for the importance of habitat, and we turn to that topic now.

Effect of Habitat

We have so far reviewed two kinds of explanations for the species–area effect: (1) relation to a species–abundance distribution (as in sampling) and (2) via the MacArthur–Wilson mechanism, extinction rate being assumed a function of population size. A third sort of explanation was advanced by Williams, who argued that area is just a proxy for habitat diversity, and it is the latter that directly drives the number of species on islands. By this interpretation, each species has its associated habitat type, and as area

increases, the amount of each habitat type also increases, exceeding the thresholds for each species' occurrence one-by-one. A classic example is the avifauna of the Aegean Islands. Number of species shows a very precise relation to habitat diversity (**Fig. 8**), equaling or surpassing many of the highly regular species–area relations exhibited by other groups. Multiple regression showed habitat diversity a more important prediction than area, just as was shown for elevation with respect to Galapagos plants (see above). Similar results have been obtained for species counts in pieces of larger, mainland areas, for example, forest birds of eastern North America.

In addition to the species–area effect, a correlation with habitat diversity has also been postulated as accounting for the distance effect: Lack argued that far islands had fewer species than near because their habitats were less varied. A definitive test of the Lack vs. MacArthur–Wilson explanation of the distance effect was performed for birds and lizards of the Bahamas. Measuring the occurrences of various habitat types directly, investigators found using partial correlation that both isolation and habitat poverty contributed to a tendency for fewer species to occur on more distant islands.

Although most research concerning the effect of area, habitat, etc. is directed toward understanding the number of species on islands, recent work is exploring effects of such biogeographic variables on population and life-history properties of island species. We measured survivorship of the lizard *Anolis sagrei* on Bahamian islands and found a very strong inverse relation to height of vegetation on the island. The result may perhaps be surprising in view of the tendency of these lizards to be arboreal and the apparent lack of resources on the scrubbiest islands (where annual survival was up to 80%!). The likely explanation is that itinerant bird predators prefer islands with higher vegetation, increasing the risk of predation on those islands. Studies involving life-history traits are becoming increasingly common, and we now turn to an area where this is of conservation importance, the understanding of extinction.

Extinction and Conservation

It is a universal property of the extinction process, including populations on islands, that the smaller the population, the more likely extinction is to happen, all other things being equal. This is a fundamental assumption underlying the area prediction of the MacArthur–Wilson equilibrium model, and it has been observed in some of the systems we have been discussing, including Channel Island birds and Bahamian spiders.

We can better understand the reasons for this relation if we examine the key processes responsible for extinction. First, demographic stochasticity, the chance occurrence of a series of deaths before any member of a population can give birth, acts at low population sizes and can cause the population randomly to go to zero. Second, environmental stochasticity, represented as variation in species traits generated by extrinsic environmental factors that themselves are varying, is important at moderate-to-large population sizes. The environmental variation can be chronic, occurring more-or-less continuously over time, or catastrophic, occurring very infrequently but being much more severe. Finally, population ceilings are important: no population has the capacity to increase indefinitely, but rather negative feedback (density-dependence) will set in as it expands, eventually forcing the population to hover at or near some ceiling in numbers. With lower and lower values of the ceiling, a population would be correspondingly more and more vulnerable to demographic–stochastic extinction.

Many models of population extinction exist that can be applied to islands, but rarely has a single model encompassed both kinds of stochasticity as well as population ceilings. An international collaboration recently attempted to apply such a model to data for spider populations on 77 islands, censused annually over a continuous 20 year period. Two species were contrasted, one with larger populations sometimes crashing quickly to extinction and having a much weaker relation of extinction likelihood to population size than the other species. A simple model ignoring life cycles and a more complex model with detailed life-cycle characteristics estimated from the field were constructed; both models did well for large population sizes, but the complex model was necessary to fit data from small population sizes, as the life cycles interact with the various forms of stochasticity. In particular, the prediction that extinction probabilities are very sensitive to juvenile survivorship emerged from the analysis. This is in contrast to a similar approach for a noninsular species, Bonelli's eagle (*Hieraatus fasciatus*), which predicted sensitivity to adult survivorship. Conclusions such as these are in fact detailed expectations about extinction likelihood, in turn guiding conservation efforts in preserving species.

See also: Ecological Processes: Succession and Colonization

Further Reading

Abbott, I., Grant, P.R., 1976. Nonequlibrial bird faunas on islands. American Naturalist 110, 507–528.
Brooks, T., Smith, M.L., 2001. Carribbean catastrophes. Science 294, 1469–1471.
Brown, J.H., Kodric-Brown, A., 1977. Turnover rates in insular biogeography: Effect of immigration on extinction. Ecology 58, 445–449.
Connor, E.F., McCoy, E.D., 1975. The statistics and biology of the species–area relation. American Naturalist 113, 791–833.
Gilpin, M.E., Armstrong, R.A., 1981. On the concavity of island biogeographic rate functions. Theoretical Population Biology 20, 209–217.
Jones, H.L., Diamond, J.M., 1976. Short-time-base studies of turnover in breeding bird population on the California Channel Islands. Condor 78, 526–549.
Lomolino, M.V., Riddle, B.R., Brown, J.H., 2005. Biogeography, 3rd edn. Sunderland, MA: Sinauer.

MacArthur, R.H., Wilson, E.O., 1967. The Theory of Island Biogeography. Princeton, NJ: Princeton University Press.
May, R.M., 1975. Patterns of species abundance and diversity. In: Cody, M., Diamond, J. (Eds.), Ecology and Evolution of Communities. Cambridge, MA: Harvard University Press, pp. 81–120.
Osman, R.W., 1977. The influence of seasonality and stability on the species equilibrium. Ecology 59, 383–399.
Schoener, T.W., 1976. The species–area relation within archipelagos: Models and evidence from island land birds. In: Proceedings of the XVI International Ornithological Congress., pp. 629–642.
Schoener, T.W., 1983. Rate of species turnover decreases from lower to higher organisms: A review of the data. Oikos 41, 372–377.
Schoener, T.W., Schoener, A., 1983. Distribution of vertebrates on some very small islands. II: Patterns in species number. Journal of Animal Ecology 52, 237–262.
Schoener, T.W., Clobert, J., Legendre, S., Spiller, D.A., 2003. Life-history models of extinction: A test with island spiders. American Naturalist 162, 558–573.
Simberloff, D., Wilson, E.O., 1970. A two-year record of colonization. Ecology 51, 934–937.
Spiller, D.A., Losos, J.B., Schoener, T.W., 1998. Impact of a catastrophic hurricane on island populations. Science 281, 695–697.
Terborgh, J., Lopez, L., Tello, S.J., 1997. Bird communities in transition: The Lago Guri islands. Ecology 78, 1494–1501.
Thornton, I.W.B., Zann, R.A., Rawlinson, P.A., 1993. Colonization of Rakata (Krakatau Is.) by nonmigrant land birds from 1983–1992 and implications for the value of island biogeography theory. Journal of Biogeography 20, 441–452.
Watson. GE., 1964. *Ecology and Evolution of Passerine Birds in the Islands of the Aegean Sea.* PhD Dissertation, Yale University. Ann Arbor, MI: University Microfilms.

Landscape Ecology☆

Jianguo (Jingle) Wu, Arizona State University, Tempe, AZ, United States; Beijing Normal University, Beijing, China

What Is Landscape Ecology?

Landscape ecology has been defined in various ways partly because the word, "landscape," means quite different things to people with different scientific and cultural backgrounds. Landscapes are spatial mosaics of interacting biophysical and socioeconomic components (Fig. 1). Just as in other ecological disciplines, a spectrum of views exists as to the relative salience or prominence of the two aspects of landscapes. The diversity of perspectives can often be related to the philosophical underpinnings of reductionism versus holism. Nevertheless, few would disagree that landscapes are compositionally diverse and spatially heterogeneous. A general definition of landscape ecology may be the science and art of studying and improving the relationship between spatial pattern and ecological processes on a multitude of scales and organizational levels. Landscape ecology is not only a field of study, but also represents a new scientific perspective or paradigm that is relevant to a range of ecological, geophysical, and social sciences.

Heterogeneity, scale, pattern–process relationships, hierarchy, disturbance, coupled ecological-social dynamics, and sustainability are among the key concepts in landscape ecology. Typical research questions include: How can spatial heterogeneity be quantified so that it can be related to relevant ecological processes? What are the processes and mechanisms responsible for existing landscape patterns? How does spatial heterogeneity influence the flows of organisms, material, and energy? How does landscape pattern affect the spread of disturbances such as pest outbreaks, diseases, fires, and invasive species? How do patterns and processes on different scales relate to each other? How can ecological information be translated from fine to broad scales and vice versa? How can the knowledge of spatial heterogeneity help improve biodiversity conservation, management and planning? How can sustainable landscapes be developed and maintained?

Studies in landscape ecology usually involve the extensive use of spatial information from field survey, aerial photography and satellite remote sensing, as well as pattern indices, spatial statistics and computer simulation modeling. The intellectual thrust of this highly interdisciplinary enterprise is to understand the causes, mechanisms, and consequences of spatial heterogeneity, while its ultimate goal is to provide a scientific basis and practical guidelines for developing and maintaining ecologically, economically, and socially sustainable landscapes (Fig. 2).

Evolving Perspectives in Landscape Ecology

Contemporary landscape ecology is characterized by a flux of concepts and perspectives that reflect the differences in the origins of ideas and the ways of thinking, both of which are shaped by physical and cultural landscapes. The term "landscape ecology" was coined in 1939 by the German geographer, Carl Troll, who was inspired by the spatial patterning of landscapes revealed in aerial photographs and the ecosystem concept developed in 1935 by the British ecologist, Arthur Tansley. Troll saw the need for combining the more structurally-oriented geographical approach with the more functionally-centered ecosystem approach, in order to allow for geography to acquire ecological knowledge of land units and for ecology to expand its analysis from local sites to larger regions. Thus, he defined landscape ecology as the study of the relationship between biological communities and their environment in a landscape mosaic on various spatial scales. In the same time, Troll also emphasized the holistic totality of the landscape which was perceived as something of a Gestalt (an integrated system organized in such a way that the whole cannot be described merely as the sum of its parts). This holistic and humanistic landscape perspective, focusing on landscape mapping, evaluation, conservation, planning, design, and management, has been often termed the European school of landscape ecology, but now widely embraced worldwide.

The concept of landscape ecology was introduced from Europe to North America in the early 1980s, and subsequently stimulated the rapid development of a stream of new ideas, theories, methods, and applications. As a result, the field of landscape ecology quickly flourished in North America, and became a widely-recognized scientific discipline by the mid-1990s around the world. While a landscape can be generally defined as a spatially heterogeneous area whose spatial extent varies according to research questions and processes of interest, most landscape ecological studies have focused on broad scales, ranging from tens to thousands of square kilometers. A multiple-scale concept of landscape is meaningful and necessary as it facilitates the theoretical and

☆*Change History:* October 2017. J Wu updated Fig. 1.

Fig. 1 Different kinds of landscapes as spatial mosaics of various patches on a range of scales. (A) An urbanizing forest landscape, (B) a wetland landscape, (C) a Sonoran desert landscape, and (D) a grassland landscape. All photos by Jianguo Wu.

methodological developments by promoting micro-, meso-, and macro-scale approaches. Despite their variations in details, the definitions of landscape ecology in North America all hinge on the idea of spatial heterogeneity or spatial pattern. In particular, the North American landscape ecology focuses more on the relationship between spatial pattern and ecological processes on multiple scales ranging from tens of square meters to thousands of square kilometers in space and from a particular point to a period of several decades in time. Its primary goal is to understand the causes, mechanisms, and ecological consequences of spatial heterogeneity.

More specifically, North American landscape ecology has had a distinct emphasis on the effects of spatial pattern on biodiversity, population dynamics and ecosystem processes in a heterogeneous area. This research emphasis is practically motivated by the fact that previously contiguous landscapes have rapidly been replaced by a patchwork of diverse land uses (landscape fragmentation), and conceptually linked to the theory of island biogeography developed in the 1960s and the perspective of patch dynamics that began to take shape in the 1970s. Island biogeographic theory relates the equilibrium-state species diversity of islands to their size (area effect on species extinction rate) and distance to the mainland (distance effect on species immigration rate). The heuristic value of the theory is apparent for understanding the ecology of habitat patches submerged in a sea of human land uses. The patch dynamics perspective, on the other hand, treats ecological systems as mosaics of interacting patches of different size, shape, kinds, and history, emphasizing the transient dynamics and cross-scale linkages of such patchy systems. In this view, a forest is no more than a dynamic mosaic of tree gaps of various age, species composition, and biophysical properties; thus the dynamics of the forest can be adequately predicted by aggregating the behavior of individual tree gaps. The perspective of patch dynamics has been evident in the conceptual development of landscape ecology in the recent decades.

In summary, the European approach is more humanistic and holistic in that it emphasizes a society-centered view that promotes place-based and solution-driven research. In contrast, the North American approach is more biophysical and analytical in that it has

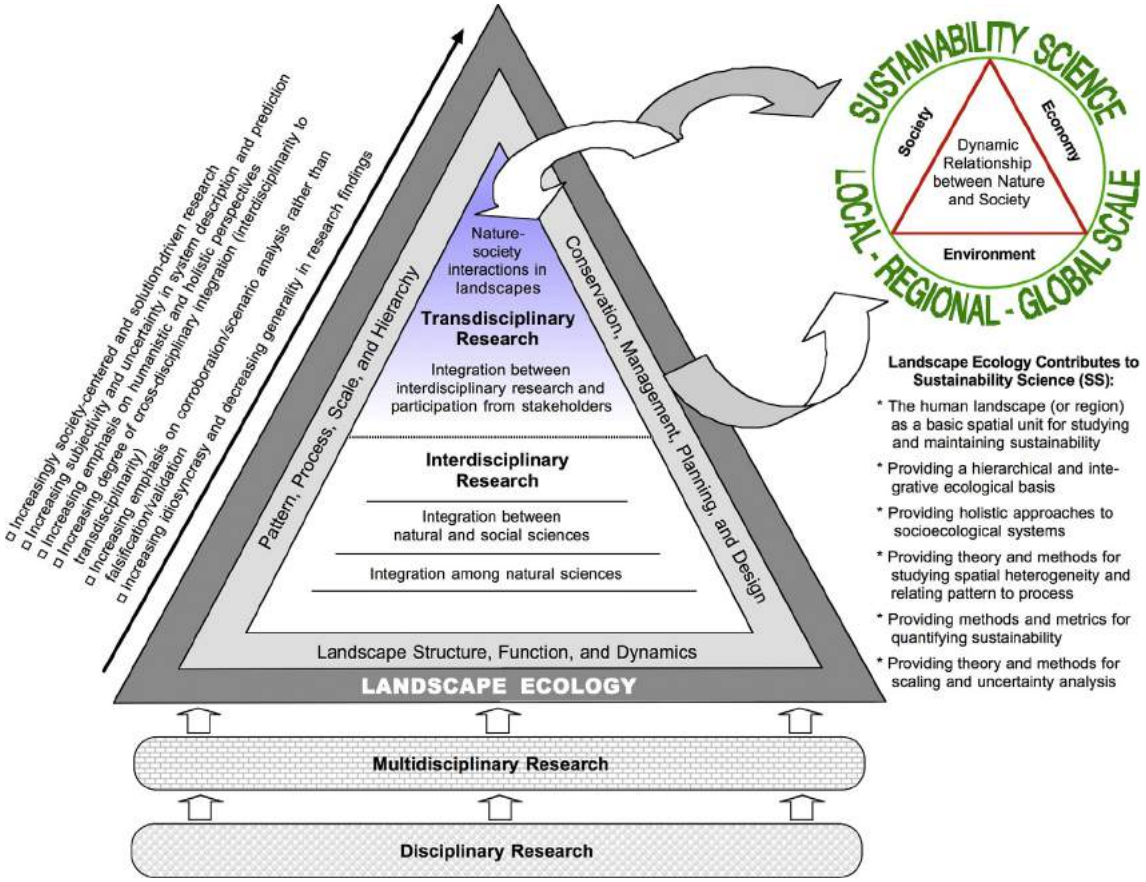


Fig. 2 A hierarchical and pluralistic view of landscape ecology. "Hierarchical" refers to the multiplicity of organizational levels, spatiotemporal scales, and degrees of cross-disciplinarity in landscape ecological research. "Pluralistic" indicates the necessity to recognize the values of different perspectives and methods in landscape ecology dictated by its diverse origins and goals. Reproduced from *Landscape Ecology*, 21, 2006, 1–6, Cross-disciplinarity, landscape ecology, and sustainability science, Wu J, with kind permission of Springer Science and Business Media.

been dominated by a biological ecology-centered view that is driven primarily by scientific questions. Here I hasten to point out that this dichotomy most definitely oversimplifies the reality because such geographic division conceals the diverse and continuously evolving perspectives within each region. In fact, many ecologists in North America have recognized the importance of humans in shaping landscapes for several decades (especially since the dust bowl in the 1930s). Although humans and their activities have been treated only as one of many factors interacting with spatial heterogeneity, more integrative studies have been emerging rapidly in the past few decades with the surging interest in urban ecology and sustainability science in North America. On the other hand, the perspective of spatial heterogeneity has increasingly been recognized by landscape ecologists in Europe and the rest of the world. Thus, the current development of landscape ecology around the world seems to suggest a transition from a stage of diversification to one of consolidation (if not unification) of key ideas and approaches.

In fact, both the European and North American approaches can be traced back to the original definition of landscape ecology. Carl Troll's proposal to integrate the geographical and structural approach with the ecological and functional approach is best reflected in the pattern–process–scale perspective, which enhances the scientific rigor of landscape ecology. The holistic and humanistic perspective, on the other hand, epitomizes the idea of landscape as a nature-society coupled system embraced by Troll and others. This perspective is entailed by any attempt to tackle practical problems in real landscapes on broad scales. Both the European and North American perspectives are essential to the development of landscape ecology as a truly interdisciplinary science.

Key Topics in Landscape Ecology

The scope of landscape ecology is quite comprehensive and dynamic. As with other interdisciplinary fields, it is impossible to define precisely the domain of landscape ecological studies. To get a sense of what the scientific core of landscape ecology is, here I discuss a series of key research topics based on the collective view of leading landscape ecologists and relevant publications, mainly in the flagship journal of the field, Landscape Ecology (http://www.springeronline.com/journal/10980/). These include: (1) pattern–process–scale relationships of landscapes; (2) landscape connectivity and fragmentation; (3) scale and scaling; (4) spatial analysis and

landscape modeling; (5) land use and land cover change; (6) landscape history and legacy effects; (7) landscape and climate change interactions; (8) ecosystem services in changing landscapes; (9) landscape sustainability; and (10) accuracy assessment and uncertainty analysis. Here I highlight six of these key topics, and more in-depth discussions of these key issues can be found in Further Reading at the end of this article.

1. Pattern–process–scale relationships and ecological flows in landscapes. Understanding how organisms, matter, and energy affect, and are affected by, the spatial pattern of landscape mosaics is a fundamental problem in landscape ecology. Much progress has been made in unraveling the effect of spatial heterogeneity on the spread of disturbances (e.g., fires and diseases) and the influence of landscape fragmentation on population dynamics, particularly, through studies of metapopulations (structurally discrete and functionally connected population ensembles). Research into the effects of landscape pattern on ecological processes across scales is still a rapidly developing area. Important areas for future research also include the spread of invasive species, the effects of landscape structure on population genetics (known as landscape genetics), and the effects of socioeconomic processes on ecological flows in landscape mosaics on multiple scales.
2. Mechanisms and consequences of land use and land cover change. Land use and land cover change, driven primarily by socioeconomic processes, exerts the most pervasive and profound influences on the structure and functioning of landscapes. Thus, quantifying the spatiotemporal pattern of landscape change and understanding its underlying driving forces are essential. More effort is needed to couple biophysical with socioeconomic approaches and to integrate ecological with historical methods in the study of land change.
3. Scale and scaling. Spatial pattern and ecological and socioeconomic processes in heterogeneous landscapes operate on multiple scales, and thus understanding the totality of landscapes requires relating different phenomena across domains in space and time. The process of translating information from one scale or organizational level to another is referred to as scaling. Landscape ecologists are leading the way in developing the theory and methods of scaling that is essential to all natural and social sciences. However, many challenges still remain, including establishing scaling relations for a variety of landscape patterns and processes as well as integrating ecological and socioeconomic dimensions in a coherent scaling framework.
4. Coupling landscape pattern analysis with ecological processes through spatial analysis and landscape modeling. Quantifying spatial heterogeneity is the necessary first step to understanding the effects of landscape pattern on ecological processes. Various effects of the compositional diversity and spatial configuration of landscape elements have been well documented, and a great number of landscape metrics (synoptic measures of landscape pattern) and spatial analysis methods have been developed in the past two decades. The greatest challenge, however, is to relate the measures of spatial pattern directly to the processes and properties of biodiversity and ecosystem functioning. To address these challenges, well-designed field-based observational and experimental studies are indispensable, and remote sensing techniques, geographic information systems (GIS), spatial statistics, and simulation modeling are also necessary.
5. Ecosystem services in changing landscapes. Ecosystem services are benefits that people derive from ecosystems. As a concept, ecosystem services bridges ecology and economy and has been increasingly used in the science and practice of conservation, resource management, and sustainable development. All ecosystem services are generated and used in landscapes that continue to change; the flows of ecosystem services are affected by landscape patterns; and certain spatial configurations of multiple ecosystems may synergistically render services (or disservices) that single ecosystems cannot produce. Thus, it is crucial to quantify spatiotemporal patterns, source-sink dynamics, trade-offs, and synergistic interactions of provisioning, regulating, and cultural ecosystem services at the landscape and regional scales. Place-based landscape theories of ecosystem services are needed.
6. Landscape sustainability. We may define landscape sustainability as the adaptive process of simultaneously maintaining and improving biodiversity, ecosystem services, and human well-being in a landscape. Because of the emphasis on broad- and multi-scale patterns and processes with interdisciplinary approaches, landscape ecology is uniquely positioned to provide a comprehensive theoretical basis and pragmatic guidelines for biodiversity conservation, ecosystem management, and sustainable development. These real-world problems cannot be adequately addressed by species-centered or individual ecosystem-based approaches. How do spatial processes occurring in landscapes (e.g., urbanization, agriculture, flooding, fires, biological invasion) affect the biodiversity, ecosystem functioning, ecosystem services, and human-wellbeing altogether? How do ecological, economic, and social processes interact to determine landscape resilience and sustainability? What are the design principles for sustainable landscapes? These are only a few of many challenging questions landscape ecology will continue to address in decades to come.

Concluding Remarks

Emphasis on heterogeneity begs questions of the relationship between pattern and process. Simply put, heterogeneity is about structural and functional patterns that deviate from uniform and random arrangements. It is this pervasively common non-homogeneous characteristic that makes spatial patterns ecologically important as it suggests nontrivial relationship to underlying processes. Thus, studying pattern without getting to process is superficial, and understanding process without reference to pattern is incomplete. Emphasis on heterogeneity also makes scale a critically important issue because heterogeneity, as well as the relationship between pattern and process, may vary as the scale of observation or analysis is changed. Thus, whenever heterogeneity is emphasized, spatial structures, underlying processes, and scale inevitably become essential objects of study. From this perspective,

landscape ecology is a science of heterogeneity and scale. On the other hand, with increasing human dominance in the biosphere, emphasis on broad spatial scales makes inevitable to deal with humans and their activities. As a consequence, humanistic and holistic perspectives have been and will continue to be central in landscape ecological research.

The above arguments also, in part, explain the two seemingly disparate views that have become known as the European and North American perspectives in landscape ecology. The world is already too fragmented ecologically, economically, and socially, and we certainly do not need a landscape ecology for each continent. As discussed earlier, the two perspectives should be viewed as being complementary rather than contradictory. To increase the synergies between the two approaches, not only do we need to appreciate the values of each, but also to develop an appropriate framework by which different perspectives and methods can be integrated. This requires a pluralistic and multi-scale perspective (Fig. 2). Landscapes out there are messy and are increasingly being messed up. Landscape ecology not only is expected to provide the scientific understanding of the structure and functioning of various landscapes, but also the pragmatic guidelines and tools with which order and sustainability can be created and maintained for the ever-changing landscapes. Landscape ecology is a science of spatial heterogeneity and scale, and its ultimate goal is to improve the sustainability of landscapes.

Further Reading

Barrett GW, Barrett TL, and Wu JG (eds.) (2015) *History of landscape ecology in the United States*. New York: Springer.
Forman RTT (1995) *Land mosaics: the ecology of landscapes and regions*. Cambridge: Cambridge University Press.
Forman RTT and Godron M (1986) *Landscape ecology*. New York: Wiley.
Naveh Z and Lieberman AS (1994) *Landscape ecology: theory and application*. New York: Springer.
Risser PG, Karr JR, and Forman RTT (1984) *Landscape ecology: directions and approaches*. Champaign: Illinois Natural History Survey.
Turner MG (2005) Landscape ecology: what is the state of the science? *Annual Review of Ecology and Systematics* 36: 319–344.
Turner MG and Gardner RH (2015) *Landscape ecology in theory and practice: pattern and process*, 2nd edn. New York: Springer.
Wiens J and Moss M (eds.) (2005) *Issues and perspectives in landscape ecology*. Cambridge: Cambridge University Press.
Wu J (2013) Key concepts and research topics in landscape ecology revisited: 30 years after the Allerton Park workshop. *Landscape Ecology* 28: 1–11.
Wu J (2013) Landscape sustainability science: ecosystem services and human well-being in changing landscapes. *Landscape Ecology* 28: 999–1023.
Wu J and Hobbs R (eds.) (2007) *Key topics in landscape ecology*. Cambridge: Cambridge University Press.

Relevant Websites

http://www.landscape-ecology.org—International Association of Landscape Ecology.
http://www.springer.com—Landscape ecology, the flagship journal of the field of landscape ecology and sustainability, Springer.

Landscape Planning☆

Ü. Mander and E Uuemaa, University of Tartu, Tartu, Estonia

Introduction

This article presents a scientific overview of the basic implementation of the principles of ecological engineering in landscape planning. The first two sections discuss the landscape definition, landscape functions, and multifunctionality. The subsequent sections give an overview of the landscape diversity and coherence, landscape fragmentation and its ecological consequences, landscape evaluation and landscape indicators, the levels and steps of landscape planning, ecologically compensating areas in the landscape, as well as of the leading principle in ecological landscape planning – the concept and implementation of territorial ecological networks (greenway networks) at the landscape level.

Landscape Definition

Landscapes as dynamic and characteristic expressions of the interaction between the natural environment and human societies can be considered in very different ways: from the scenery and 'total character of the Earth' (Alexander von Humboldt cit. Zonneveld, 1995) to the complexity of ecosystems. Depending on the degree of human interaction, landscape characteristics can be dominated by natural aspects on the one hand or human management on the other. In this article, we consider landscape as a geosystem or geocomplex, a comprehensive complex of natural (physical, chemical, biological) and anthropogenic factors distinguished at various hierarchical levels (i.e., micro-, meso-, and macrochores). The main natural factors in such a complex landscape system are water, topography, soil, geology, and climate conditions, as well as plants (vegetation cover) and animals (fauna). Likewise, the ecosystem approach deals with the same factors as ecosystem components, but in contrast to ecosystems, where all of the relations are considered via biota, the geosystem/landscape concept considers all of the relationships. However, different factors at different temporal and spatial scales play different roles in determining landscape character. Climatic and geological conditions cause the basic natural character of a landscape, whereas topography, soil, and vegetation cover are important in the formation of the detailed character of a landscape, and are influenced by human management.

Landscape Functions

Traditionally, the concept of landscape functions has been considered in the landscape planning system of Germany and German-speaking countries. According to that concept, landscape has the following functions: (1) production (economic) functions (biomass production, water supply, suitability of nonrenewable resources); (2) regulatory (ecological) functions (regulation of material and energy fluxes, hydrological and meteorological functions, regulation and regeneration of populations and bio(geo) coenoses, habitat (genetical) function); (3) social functions (psychological (esthetic and ethical) functions, information functions, human-ecological, and recreational functions).

This approach is very similar to the concept of ecosystem services and natural capital, which has recently gained extensive popularity. According to this concept, the typology of landscape functions includes four categories: (1) provisioning functions; (2) regulation functions; (3) habitat functions; and (4) cultural and amenity functions (see **Table 1**).

1. Provisioning functions comprise functions that supply 'physical services' in terms of resources or space. This category has been divided into two classes: production and carrier functions. Production functions reflect resources produced by natural ecosystems, for example, the harvesting of fish from the ocean, pharmaceutical products from wild plants and animals, or wood from natural forests. Carrier functions reflect the goods and services that are provided through human manipulation of natural productivity (e.g., fish from aquaculture or timber from plantations). In these cases, the function offered by nature is the provision of a suitable substrate or space for human activities, including agriculture, mining, transportation, etc.
2. Regulation functions result from the capacity of ecosystems and landscapes to influence ('regulate') climate, hydrological and biochemical cycles, Earth surface processes, and a variety of biological processes. These services often have an important spatial (connectivity) aspect; for example, the flood control function of an upper watershed forest is only relevant in the flood zone downstream of the forest.
3. Habitat functions comprise the importance of ecosystems and landscapes in maintaining natural processes and biodiversity, including the refugium and nursery functions. The refugium function reflects the value of landscape units in providing habitats

☆*Change History:* November 2014. Ü Mander and E Uuemaa updated the test and references.

 doi:10.1016/B978-0-12-409548-9.09478-1

Table 1 Typology of ecosystem/landscape functions, goods, and services

Entry	*Ecosystem functions*	*Short description*	*Biophysical indicators (examples) (i.e., ecosystem properties providing the goods or service)*	*Goods and services (examples)*
1	Provisioning Production functions	Resources from unmanipulated ecosystems	Biomass (production and stock) Biochemical properties	Freshwater Food (e.g., fish, bush meat) Raw materials (wood, fodder)
	Carrier functions	Use of space to (enhance) supply resources or other goods and services	Depending on the specific land use type, different requirements are placed on environmental conditions (e.g., soil stability and fertility, air and water quality, hydrology, topography, climate, geology)	Cultivation (e.g., agriculture, plantations, aquaculture) Energy conversion (e.g., wind, solar) Mining (ore, fossil fuels) Transportation (esp. on waterways)
2	Regulation functions	Direct benefits from ecosystem processes	Role of ecosystems in biogeochemical cycles (e.g., CO_2/O_2 balance, hydrological cycle)	Climate regulation Maintenance of soil fertility
			Role of vegetation and biota in removal or breakdown of nutrients and toxic compounds	Waste treatment (e.g., water purification) Maintenance of air quality
			Physical properties of land cover	Water regulation (e.g., buffering runoff) Erosion prevention Storm protection and flood prevention
			Population control through tropic-dynamic relations	Biological control (of pests and diseases) Pollination
3	Habitat functions	Maintenance of biodiversity and evolutionary processes	Presence of rare/endemic species; species diversity Reproduction habitat for migratory species	Refugium for wildlife Nursery function (for commercial species)
4	Cultural and amenity functions	Nonmaterial benefits	Landscape (or ecosystem) properties with esthetic, recreational, historical, spiritual, inspirational, scientific, or educational value	Enjoyment of scenery (e.g., scenic roads) Ecotourism and recreation Heritage value/cultural landscapes Spiritual or religious sites Cultural expressions (use of landscapes as motif in books, film, painting, folklore, advertising) Research and education

Adapted from De Groot, R.S., Hein, L., 2007. Concept and valuation of landscape functions at different scales. In: Mander, Ü., Wiggering, H., Helming, K. (Eds.), Multifunctional land use: meeting future demands for landscape goods and services. Berlin: Springer, pp. 15 -36.

to (threatened) fauna and flora, and the nursery function indicates that some landscape units provide a particularly suitable location for reproduction and thereby have a regulating impact on the maintenance of populations elsewhere.

4. Cultural and amenity functions relate to the benefits people obtain from landscapes through recreation, cognitive development, relaxation, and spiritual reflection. This may involve actual visits to the area, indirectly enjoying the area (e.g., through nature movies), or gaining satisfaction from the knowledge that a landscape contains important biodiversity or cultural monuments. The latter may occur without having the intention of ever visiting the area. These services have also been referred to as 'information functions'.

The evaluation of landscapes for planning and management purposes, as well as landscape synthesis and decision making, is based on landscape functions.

Landscape Diversity and Coherence

One of the basic characteristics of landscapes is the diversity or heterogeneity of the landscape pattern (mosaic).

Hundreds of landscape metrics have been proposed by various researchers to analyze the landscape pattern. Most of these are covered by the computer program FRAGSTATS. The most typical use of the FRAGSTATS-based landscape metrics is for the prediction of species diversity. Also, several researchers have used FRAGSTATS-based landscape metrics as indicators of various landscape changes (management activities and natural disturbances) such as the change in the spatial structure of landscapes, forest planning and management, landscape destruction and rehabilitation, and landscape disturbances by fire and road construction. This demonstrates that temporal (time-series-based) indicators are inseparably related to spatial indicators. In order to control how landscape metrics respond to changing grain size, extent, the number of zones, the direction of analysis, etc.,

landscape simulators are applied. Gardner et al. introduced the concept of neutral models into landscape ecology. The aim of a neutral model is to have an expected pattern in the absence of specific landscape processes. In order to have a random pattern, the first application of this concept stemmed from the percolation theory, but different types of regular artificial landscapes are also used.

Landscape coherence has been considered one of the criteria for the development of sustainable rural landscapes. Proceeding from Bockemühl's concept of landscape identity and perception, which was developed in biodynamic farms, van Mansvelt classifies the ecological coherences of rural landscape in three groups: vertical (on site), horizontal (landscape-level), and cyclical (temporal) coherences. The first type can be referred to as coherence between biodiversity and the local abiotic environmental conditions. For instance, soil-bound agricultural production would be an example of vertically coherent biodiversity management. The horizontal type of ecocoherence is "that between coherence within a habitat (biotope or mini-ecosystem) and that of habitats in a landscape (macro-ecosystem)" (van Mansvelt, 1997). This coherence refers to the functional (ecophysiological) interdependency of species within the ecosystems, but also to the relationships of habitats within the larger system. According to Kuiper, horizontal coherence is characterized by the connectivity between similar ecosystems in a landscape. Cyclical (temporal) coherences are characterized not only by the full life cycles of species and systems, but also by the self-production of species and biotopes, and season-compliant management (e.g., sowing, mowing, coppicing, etc.).

From the methodological point of view, van Mansvelt's concept of landscape coherence is rather holistic and is used in the context of landscape perception and visual characteristics, with no studies that quantify this category in landscape validation. The most common estimates of different ecological coherences are their appearance or absence or relative scores. Another attempt to estimate coherence refers to the connectivity between landscape components. However, as in the case of various analogous indices that have been developed to describe landscape connectivity, this approach does not consider the quantification of coherence.

Wascher (2000) defines landscape coherence as the "adequacy of land use according to biophysical conditions." Mander and Murka developed a dynamic landscape coherent concept which links issues of landscape diversity and landscape change. This concept refers to the correspondence between changes in actual (cultural or man-made) landscape diversity caused by land amelioration or transformation of landscape pattern (e.g., due to changing socioeconomic conditions) and potential (biophysically determined) landscape diversity. According to this concept, the homogenization of landscape diversity caused by amelioration or other anthropogenic disturbances and determined on the basis of ecotone length per area unit can be lowest in the most sensitive (less resistant) landscapes. These are landscapes with both very simple and very complicated potential (biophysical) diversity, determined by heterogeneity of soil cover (**Fig. 1**). For measuring landscape coherence, Mander *et al.* (2010) proposed Moran's I correlograms and a simple characteristic of half-value distances: $h_{I=0.5}$ – the distance lag where Moran's I drops below 0.5 (**Fig. 2**). Half-value distance enables to compare patterns of different layers e.g. land use and soil. This concept of landscape coherence helps to find optimal rates for changing land use pattern via land reclamation or planning activities.

Landscape Fragmentation and Its Ecological Consequences

One of the main impacts of human activities on landscapes worldwide is the fragmentation of habitats and whole landscapes. Habitat fragmentation is the main reason for biodiversity decrease. It provides a familiar example of a critical threshold, that is, transition ranges across which small changes in spatial pattern produce abrupt shifts in ecological responses. As the landscape becomes dissected into smaller parcels of habitat, landscape connectivity – the functional linkage among habitat patches – may suddenly become disrupted, having important consequences for the distribution and persistence of populations. Landscape connectivity depends not only on the abundance and spatial patterning of habitat, but also on the habitat specificity and dispersal

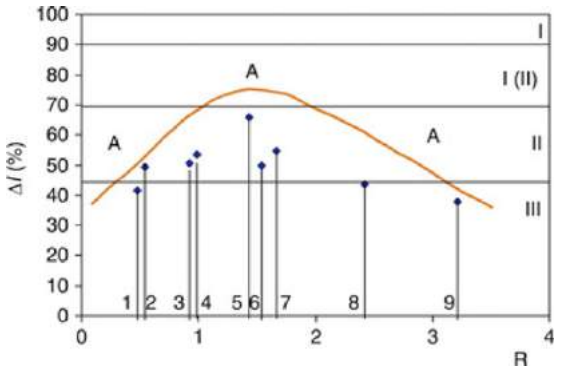


Fig. 1 Recommended change limits (ΔI) of actual landscape diversity (R) according to the dynamic coherence concept. *A* is area of diversity change at which undesirable anthropogenic processes (erosion, deflation, clogging of drainage, etc.) occur. The curve indicates the generalized coherence limit. II and III are the coherence levels for landscapes of resistance groups II and III, respectively. Adapted from Mander, Ü., Murka, M., 2003. Coherence of cultural landscapes: a new criterion for evaluating impacts of landscape changes. In: Mander, Ü., Antrop, M. (Eds.), Advances in ecological sciences 16: multifunctional landscapes, vol. III: continuity and change. Boston, MA: WIT Press, pp. 15–32.

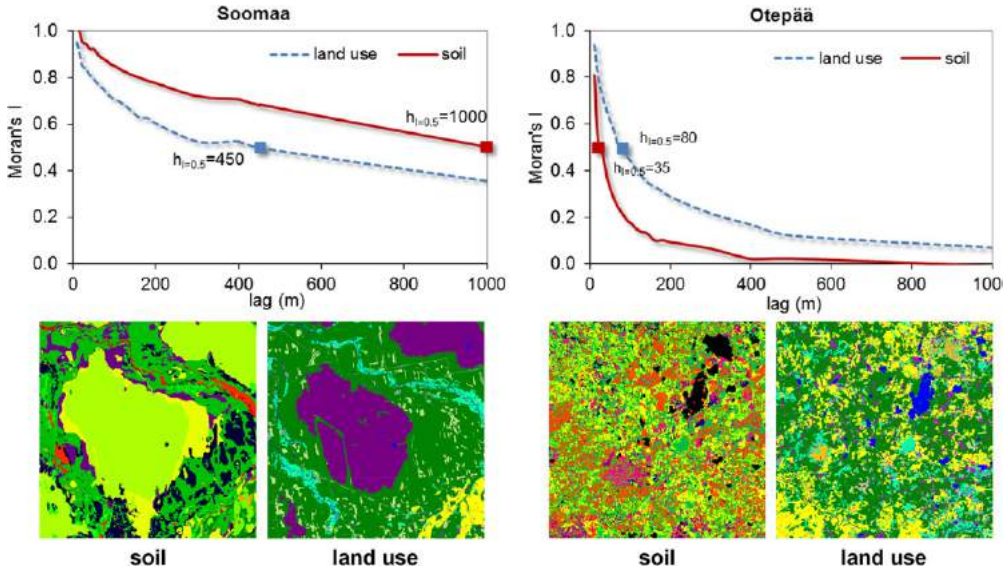


Fig. 2 Correlograms can be used to describe and compare the spatial structure of landscapes and also as an indicator fo coherence. Graphs show correlograms and $h_{I=0.5}$ values for land use and soil of two different landscapes – Otepää and Soomaa. Otepää is more heterogeneous than Soomaa and human influence has changed the landscape even more complex because landscape pattern is more fragmented than soil. Adapted from Mander, Ü., Uuemaa, E., Roosaare, J., Aunap, R., Antrop, M., 2010. Coherence and fragmentation of landscape patterns as characterized by correlograms: a case study of Estonia. Landscape and Urban Planning 94, 31–37.

abilities of species. Habitat specialists with limited dispersal capabilities presumably have a much lower threshold to habitat fragmentation than highly vagile species, which may perceive the landscape as functionally connected across a greater range of fragmentation severity.

The composition of habitat types in a landscape and the physiognomic or spatial arrangement of those habitats are the two essential features that are required to describe any landscape. As such, these two features affect four basic ecological processes that can influence population dynamics or community structure. The first two of these processes, landscape complementation and landscape supplementation, occur when individuals move between patches in the landscape to make use of nonsubstitutable and substitutable resources. The third process, source–sink dynamics, describes the consequences of having different individuals in the same population occupy habitat patches of different qualities, and is part of the metapopulation concept. The fourth process, the neighborhood effect, describes how landscape effects can be amplified when the critical resources are in the landscape immediately surrounding a given patch.

In generalizing from several studies, one can conclude that there is an optimum of landscape fragmentation at which biodiversity is the highest. For instance, in open patches, large natural (relatively) homogeneous forests caused by natural disturbances or human activities that can support various species with different ecological requirements can exist. On the other hand, excessively small patches in fragmented landscapes are unable to provide enough space and resources for various metapopulations.

Landscape planning measures, especially the implementation of territorial ecological networks, can provide greater connectivity and biodiversity in landscapes.

Landscape Evaluation and Landscape Indicators

The evaluation of nature is an inseparable part of the process of environmental/landscape planning, management, and decision making. In recent decades, its importance has reached the global level. At local and regional levels, landscape assessment for planning and decision-making processes is a key issue in sustainable landscape management.

One of the well-known conceptual frameworks for ecological/environmental indicators is the driving forces (drivers) → pressures → state → impact → responses (DPSIR) approach, which treats the environmental management process as a feedback loop controlling a cycle consisting of these five stages.

Regarding the EU policy in biological and landscape diversity management (e.g., PEBLDS, the Pan-European Biological and Landscape Diversity Strategy), it is useful to follow the DPSIR framework in reporting environmental issues. This approach treats the environmental management process as a feedback loop that controls a cycle consisting of these five stages. In addition, this introduces the term 'pressures' and adds 'impacts' – a concept that implies the cause–effect link.

The nitrogen cycle can be used as an example of the DPSIR approach in the intensification of agriculture:

- *Driving force*. Intensive agriculture;
- *Pressure*. Use of mineral fertilizers;

- *State*. Intensive loss of nitrogen from agricultural fields, high nitrogen concentration in rivers and groundwater, intensive gaseous N flux into the atmosphere, high excess nitrogen loading in ecosystems;
- *Impact*. Loss of biodiversity, eutrophication of water bodies, methemoglobinaemia, cancer risk, decreasing biodiversity, lower esthetical value of landscapes;
- *Response*. (1) Less mineral fertilizers and optimization of crop rotations with leguminous plants, especially in sensitive and potential core areas, (2) establishment of riparian buffer zones, (3) establishment of riverine and riparian wetlands.

On the other hand, the influence of marginalization (land abandonment) can also be characterized using the DPSIR approach (**Fig. 3**):

- *Driving force*. Marginalization (abandonment of agriculture);
- *Pressure*. Change of existing management scheme;
- *State*. Loss of open landscapes, loss of various (grassland) biotopes;
- *Impact*. Loss of biodiversity, loss of scenic values of landscape;
- *Response*. (1) Subsidies for farmers to support traditional low input or ecological agriculture, (2) restoration and rehabilitation of valuable biotopes (wooded meadows, alvars), (3) (re-)establishment of wetland biotopes in agricultural landscapes.

Using the DPSIR approach as a conceptual background, we consider landscape indicators as a system of structural and functional parameters that can be used to evaluate landscape pressure, state, and responses. The structural indicators are related to landscape structure (both temporal and spatial), whereas functional indicators can be divided according to landscape functions (**Table 2**).

Indicators are increasingly used to assess the effects of landscape change on the visual landscape. Visual scale is considered a key aspect of landscape perception. Visual scale is defined as the perceptual units that reflect the experience of landscape rooms, visibility and openness. The perceived visual scale is affected by both abandonment and spontaneous forest growth in previously open agricultural areas as well as of changes in agricultural and forestry strategies. Although visual landscape assessment using indicators simplifies the complexity of landscapes and landscape perception, the approach has gained ground for its repeatability, transparency and applicability in landscape planning and monitoring (**Table 3**) (Ode *et al.*, 2010a; Dramstad *et al.*, 2006).

Main Ecological Engineering Principles of Landscape Planning

Jørgensen presents 19 ecological engineering principles for application in landscape management:

- Ecosystem structure and functions are determined by the forcing functions of the system.
- Energy inputs to the ecosystems and available storage of matter are limited.
- Ecosystems are open and dissipative systems.
- Attention to limiting factors is strategic and useful in preventing pollution or restoring ecosystems.
- Ecosystems have a homeostatic capability that results in the smoothing out and depressing effects of strongly variable inputs.
- Match recycling pathways to the rates to ecosystems to reduce the effect of pollution.
- Design for pulsing systems wherever possible.

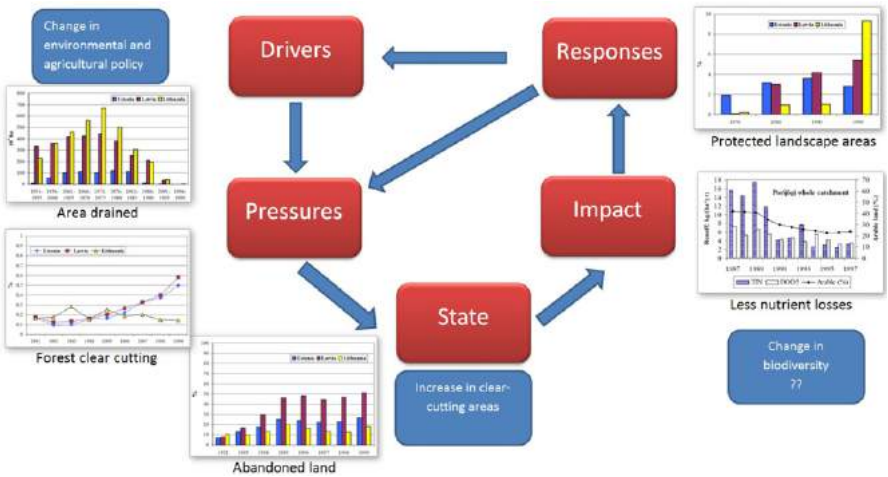


Fig. 3 The DPSIR framework for reporting on environmental issues: an example of the change in the political and socioeconomic system in Central and Eastern Europe at the end of the 1980s and the beginning of the 1990s followed by changes in environmental and agricultural policy, as a possible basis for indicator classification and landscape assessment. Adapted from Mander Ü and Kuuba R (2004) Changing landscapes in Mander, Ü., Kuuba, R., 2004. Changing landscapes in northeastern Europe based on examples from Baltic countries. In: Jongman, R.H.G (Ed.), The new dimensions of the European landscape. Dordrecht, the Netherlands: Springer, pp. 123-134.

Table 2 Potential indicators for determining (sustainable) use of ecosystem services.

Services comments and examples	*Ecological process and/or component providing the service (or influencing its availability) = functions*	*State indicator (how much of the service is present)*	*Performance indicator (how much can be used/provided in sustainable way)*
Provisioning			
1 Food	Presence of edible plants and animals	Total or average stock(in kg ha^{-1})	Net productivity (in kcal $ha^{-1}year^{-1}$ or other unit)
2 Water	Presence of water reservoirs	Total amount of water (m^3ha^{-1})	Max sust. water-extraction ($m^3ha^{-1}year^{-1}$)
3 Fiber & Fuel & other raw materials	Presence of species or abiotic components with potential use for timber, fuel or raw material	Total biomass (kg ha^{-1})	Net productivity (kg $ha^{-1}year^{-1}$)
4 Genetic materials: genes for resistance to plant pathogens	Presence of species with (potentially) useful genetic material	Total 'gene bank' value (e.g. number of species & sub-species)	Maximum sustainable harvest
5 Biochemical products and medicinal resources	Presence of species or abiotic components with potentially useful chemicals and/or medicinal use	Total amount of useful substances that can be extracted (kg ha^{-1})	Maximum sustainable harvest (in unit mass/area/time)
6 Ornamental species and/or resources	Presence of species or abiotic resources with ornamental use	Total biomass (kg ha^{-1})	Maximum sustainable harvest
Regulating			
7 Air quality regulation: (e.g. capturing dust particles)	Capacity of ecosystems to extract aerosols & chemicals from the atmosphere	Leaf area index NOx-fixation, etc.	Amount of aerosols or chemicals 'extracted' – effect on air quality
8 Climate regulation	Influence of ecosystems on local and global climate through land-cover and biologically-mediated processes	Greenhouse gas-balance (esp. C-sequestration); land cover characteristics, etc.	Quantity of greenhouse gases, etc. fixed and/or emitted → effect on climate parameters
9 Natural hazard mitigation	Role of forests in dampening extreme events (e.g. protection against flood damage)	Water-storage (buffer) capacity (in m^3)	Reduction of flood-danger and prevented damage to infrastructure
10 Water regulation	Role of forests in water infiltration and gradual release of water	Water retention capacity in soils, etc. or at the surface	Quantity of water retention and influence of hydro-logical regime (e.g. irrigation)
11 Waste treatment	Role of biota and abiotic processes in removal or breakdown of organic matter, xenic nutrients and compounds	Denitrification (kg N $ha^{-1}year^{-1}$); immobilization in plants and soil	Max amount of chemicals that can be recycled or immobilized on a sustainable basis
12 Erosion protection	Role of vegetation and biota in soil retention	Vegetation cover root-matrix	Amount of soil retained or sediment captured
13 Soil formation and regeneration	Role of natural processes in soil formation and regeneration	E.g. bio-turbation	Amount of topsoil (re)generated per ha/year
14 Pollination	Abundance and effectiveness of pollinators	Number & impact of pollinating species	Dependence of crops on natural pollination
15 Biological regulation	Control of pest populations through trophic relations	Number & impact of pest-control species	Reduction of human diseases, live-stock pests, etc.
Habitat or supporting			
16 Nursery habitat	Importance of ecosystems to provide breeding, feeding or resting habitat for transient species	Number of transient species & individuals (esp. with commercial value)	Dependence of other ecosystems (or 'economies') on nursery service
17 Genepool protection	Maintenance of a given ecological balance and evolutionary processes	Natural biodiversity (esp. endemic species); habitat integrity (irt min. critical size)	'Ecological value' (i.e. difference between actual and potential biodiversity value)
Cultural & amenity			
18 Aesthetic: appreciation of natural scenery (other than through deliberate recreational activities)	Aesthetic quality of the landscape, based on e.g. structural diversity, 'greenness,' tranquility	Number/area of landscape features with stated appreciation	Expressed aesthetic value, e.g., number of houses bordering natural areas # users of 'scenic routes'

Table 2 Continued

Services comments and examples	*Ecological process and/or component providing the service (or influencing its availability)=functions*	*State indicator (how much of the service is present)*	*Performance indicator (how much can be used/provided in sustainable way)*
19 Recreational: opportunities for tourism and recreational activities	Landscape-features attractive wildlife	Number/area of landscape & wildlife features with stated recreational value	Maximum sustainable number of people & facilities actual use
20 Inspiration for culture, art and design	Landscape features or species with inspirational value to human arts, etc.	Number/area of landscape features or species with inspirational value	#books, paintings, etc. using ecosystems as inspiration
21 Cultural heritage and identity: sense of place and belonging	Culturally important landscape features or species	Number/area of culturally important landscape features or species	Number of people 'using' forests for cultural heritage and identity
22 Spiritual & religious inspiration	Landscape features or species with spiritual & religious value	Presence of landscape features or species with spiritual value	Number of people who attach spiritual or religious significance to ecosystems
23 Education & science opportunities for formal and informal education & training	Features with special educational and scientific value/interest	Presence of features with special educational and scientific value/interest	Number of classes visiting, number of scientific studies, etc.

Adapted from De Groot, R.S., Alkemade, R., Braat, L., Hein, L., Willemen, L., 2010. Challenges in integrating the concept of ecosystem services and values in landscape planning, management and decision making. Ecological Complexity 7, 262 -270.

- Ecosystems are self-designing systems.
- Ecosystem processes have characteristic temporal and spatial scales that must be accounted for in environmental management.
- Biodiversity should be championed to maintain an ecosystem's self-design capacity.
- Ecotones and transition zones are as important to ecosystems as membranes are for cells.
- Coupling between ecosystems should be utilized wherever possible.
- The components of an ecosystem are interconnected and interrelated and form a network, implying that the direct as well as indirect effects of ecosystem development need to be considered.
- An ecosystem has a history of development.
- Ecosystems and species are most vulnerable at their geographical edges.
- Ecosystems are hierarchical systems and are parts of a larger landscape.
- Physical and biological processes are interactive. It is important to know both physical and biological interactions and to interpret them.
- Ecotechnology requires a holistic approach that integrates all interacting parts and processes as much as possible.
- Information in ecosystems is stored in structures.

The following five recommendations are implicitly embedded in the 19 principles: (1) know the natural and man-made ecosystems that make up a landscape and the corresponding ecological properties and processes; (2) use this ecological knowledge in landscape management; (3) develop models and use ecological indicators to enable a thorough overview of the many interacting components, the ecological networks, and the most crucial ecological processes; (4) maintain high biodiversity and a high-diversity pattern of ecosystems, zones, ecotones, corridors, ditches, ecological niches, etc.; the overloading from man-made ecosystems can be reduced and buffered considerably by planning a landscape with a mosaic of different man-made and natural ecosystems; (5) everything is linked to everything else in an ecosystem, and the entire system is more than the sum of its parts. These principles should underlie all ecological management decisions.

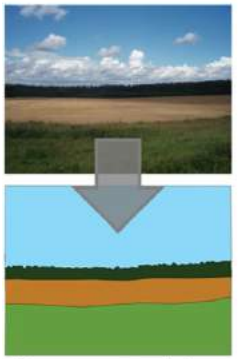
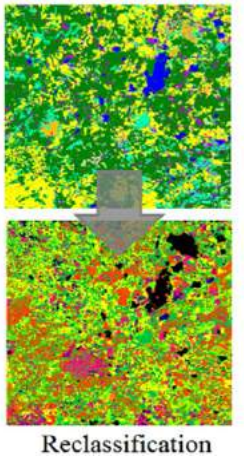
Levels and Steps in Landscape Planning

Typically, landscape planning provides information about the existing qualities of the landscape and nature, which are considered to be nature or landscape potentials, and their value as well as their sensitivity to impacts, the existing and potential impacts on these potentials, and the objectives and guidelines for the development of the landscape and nature, upon which proposed measures and development plans can be measured.

With this information, landscape planning provides evaluation guidelines for the impact regulations and for the part of the environmental impact assessment which is concerned with the landscape and nature. In the beginning phases of planning projects, landscape planning offers a background for the evaluation of alternatives, for example, in the placement of transportation corridors. Landscape planning provides a basis for preliminary opinions about proposed projects, even for projects which were proposed after the completion of the landscape plan.

Bastian and Schreiber describe four main steps in comprehensive landscape planning:

Table 3 Suggested indicators of visual scale for the different data sources landscape photos, aerial photographs, land cover data and field observation

	Landscape photos	*Orthophotos*	*Land cover*	*Field studies*
1. Open area indicators Example of basic data processing				
Proportion of open land	% of open land	% of open land	% of open land	% of open land
Viewshed size		Size of viewshed	Size of viewshed	
Viewshed shape	Classification of shape (one large open area/ split open area/patchy open area)	Shape index	Shape index	Classification of shape (one large open area/ split open area/patchy open area)
Depth/breadth of view	Estimation of depth of view (short/medium/long)	Length of radius of view	Length of radius of view	Estimation of depth of view (short/medium/ long)
2. Indicators for the view obstruction				
Density of obstructing objects	Density of obstructing objects	Density of obstructing objects		Density of obstructing objects
Degree of visual penetration	Proportion of vegetation with different levels of visual penetration (blocked/dense/semi-open/open)			Proportion of vegetation with different levels of visual penetration (blocked/dense/semi-open/open)

Adapted from Ode, Å., Tveit, M.S., Fry, G. ,2010. Advantages of using different data sources in assessment of landscape change and its effect on visual scale. *Ecological Indicators* 10: 24–31.

- Definition of problem (determination of: planning context, planning priorities, planning prerequisites);
- Inventory, analysis, and diagnosis (determination of the natural potentials: inventory, impact, protection; evaluation of the ecological and esthetic suitability of the existing and proposed lands);
- Planning concept (elaboration of: objectives for nature protection and landscape management, alternatives);
- Plan of action (definition of requirements and measures necessary to achieve the objectives);
- Product: landscape planning program, regional landscape plan, landscape plan;
- Implementation (the realization of planning measures through nature protection authorities, nature protection organizations, other planning agencies, local governments, public institutions, and individuals);
- Review (evaluation of: implementation, planning objectives, necessary alterations).

Landscape analysis involves the evaluation of elemental, spatial, and temporal pattern of landscape, as well as of dynamics of landscape and land-use pattern. The landscape diagnosis provides a comparison of landscape potential with social requirements (stability and load analyses).

As the products of this comprehensive multilevel hierarchical system, a landscape program, regional landscape plan, landscape plan, and open space master plan will be elaborated (**Table 4**).

Territorial Ecological Networks

The concept and implementation of territorial ecological networks (greenway networks) at the landscape level is considered to be the leading principle in ecological landscape planning. The widely used European-level approach defines territorial ecological networks as coherent assemblages of areas representing natural and seminatural landscape elements that need to be conserved, managed, or, where appropriate, enriched or restored in order to ensure the favorable conservation status of ecosystems, habitats, species, and landscapes of regional importance across their traditional range.

In addition to this approach, there are a wide range of names worldwide given to such 'patch and corridor' spatial concepts: greenways in the USA, Australia, and New Zealand, ecological infrastructure, ecological framework, extensive open space systems, multiple use nodules, wildlife corridors, landscape restoration network, habitat networks, territorial systems of ecological stability, framework of landscape stability. In Estonia, a concept of "the network of ecologically compensating areas" (Mander *et al.*, 1988) has been developed since the early 1980s. This network can be seen as a landscape's subsystem – an ecological infrastructure – that counterbalances the impact of the anthropogenic infrastructure in the landscape. In comparison with the traditional biodiversity-targeted approach, this concept also considers the material and energy cycling, socioeconomic and socio-cultural aspects.

According to the broader concept, ecological networks preserve the main ecological functions in landscapes, such as (1) accumulating material and dispersing human-induced energy, (2) receiving and rendering unsuitable wastes from populated areas, (3) recycling and regenerating resources, (4) providing wildlife refuges and conserving genetic resources, (5) serving as migration tracts for biota, (6) serving as barriers, filters, and/or buffers for fluxes of material, energy, and organisms in landscapes, (7) serving as support frameworks for regional settlements, (8) providing recreation areas for people, and, consequently, and (9) compensating and balancing all inevitable outputs of human society.

A network of ecologically compensating areas is a functionally hierarchical system with the following components: (A) core areas, (B) corridors; functional linkages between the ecosystems or resource habitat of a species, enabling the dispersal and migration of species and resulting in a favorable effect on genetic exchange (individuals, seeds, genes) as well as on other interactions between ecosystems; corridors may be continuous (linear), interrupted (stepping-stones), and/or landscape (scenic and valuable cultural landscapes between core areas), (C) buffer zones of core areas and corridors, which support and protect the network from adverse external influences, and (D) nature development and/or restoration areas that support resources, habitats, and species (**Fig. 4**).

The size of network components serve as another criterion of the network's hierarchy on three levels: (1) the macroscale: large natural core areas ($>1000\ km^2$) separated by buffer zones and wide corridors or stepping-stone elements (width >10 km); (2) mesoscale: small core areas ($10–1000\ km^2$) and connecting corridors between these areas (e.g., natural river valleys, seminatural recreation areas for local settlements; width 0.1–10 km); (3) microscale: small protected habitats, woodlots, wetlands, grassland patches, ponds ($<10\ km^2$) and connecting corridors (stream banks, road verges, hedgerows, field verges, ditches; width <0.1 km).

Megascale ecological networks can be considered at the global level. The human footprint map can serve as a basis for determining global ecological networks (**Fig. 5**). The macroscale of ecological networks is represented by regional-level activities such as the Pan-

Table 4 Scales of landscape planning in Germany.

Planning area	*Spatial comprehensive planning*	*Landscape planning*	*Scale*
State	State spatial plan	Landscape program	1:500000–1:200000
Region (regional district or county)	Regional plan	Regional landscape plan	1:50000–1:25000
Community	Land-use plan	Landscape plan	1:5000–1:2500
Part of the community	Master plan	Open space master plan	1:2500–1:1000

Adapted from Kiemstedt, H., 1994. Landscape planning - contents and procedures. Bonn: Nature Protection and Nuclear Safety, The Federal Minister of Environment, p. 124.

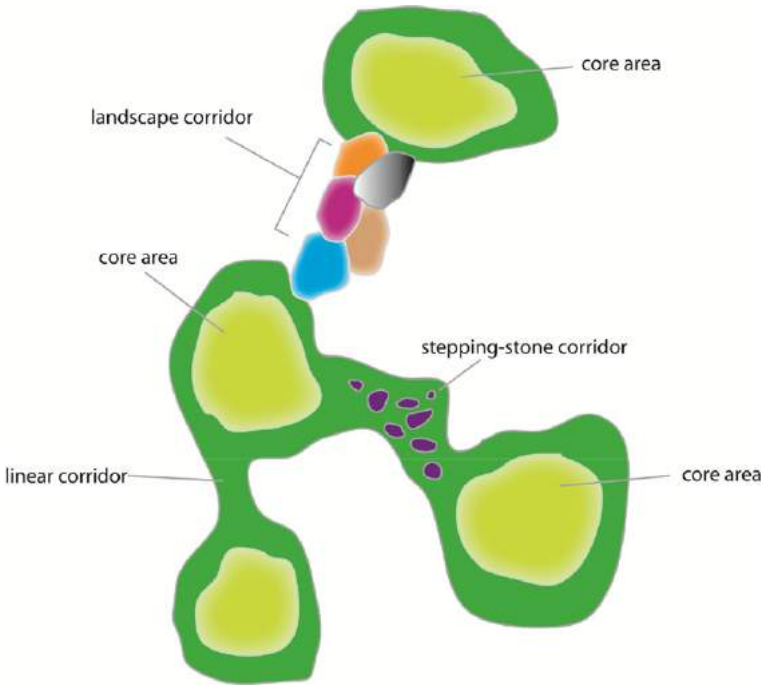


Fig. 4 Schematic example of an ecological network. Adapted from Bouwma IM, Jongman RHG, and Butovsky RO (eds.) (2002) Indicative map ofthe pan-European ecological network for Central and Eastern Europe. Technical background document. ECNC Technical Report Series, 101 pp þ annexes. Tilburg, The Netherlands/Budapest: ECNC and Mander, Ü., Külvik, M., Jongman, R., 2003. Scaling in territorial ecological networks. Landschap 20 (2), 113–127.

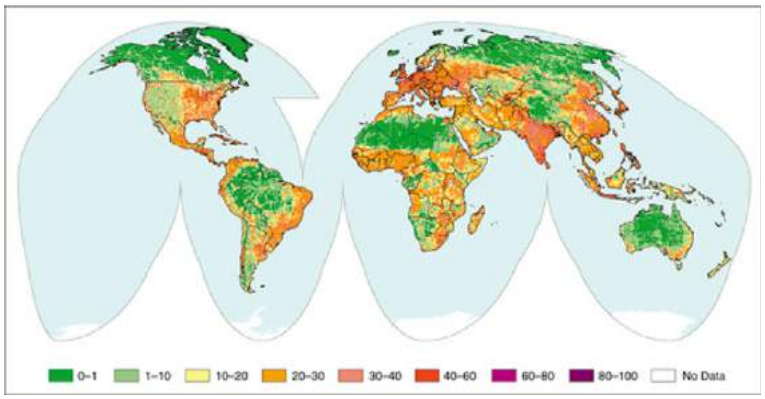


Fig. 5 A map of the human footprint as a basis for the ecological network system at the global scale. Summarized factors of anthropogenic pressure have been used, such as the Human Influence Index, which is the quantitative basis for the map. Adapted from Sanderson, E.W., Jaiteh, M., Levy, M.A., *et al.*, 2002. The human footprint and the last of the wild. BioScience 52 (10), 891–904. and Mander, Ü., Külvik, M., Jongman, R., 2003. Scaling in territorial ecological networks. Landschap 20 (2), 113–127.

European Ecological Network (PEEN) or national-level projects. In the Czech Republic, the Slovak Republic, and the Netherlands, territorial ecological networks are implemented and legislatively supported. In Estonia, Lithuania, and Poland, networks are designed and some aspects accepted by law. In Hungary, Latvia, Switzerland, and Ireland, network design is under development, and local or landscape-level ecological networks have been established in some parts of the territory of several European countries such as Germany, Belgium, the UK, Italy, Spain, Portugal, Russia, and Ukraine. Landscape-level ecological networks are designed or implemented on a wide range of spatial scales, from macro- and meso- to microscale projects. The most significant research on both species' migration and dispersal, as well as on energy and material fluxes, has been carried out at this level.

As an example of the designing of the national-level ecological network, we have presented a part of the PEEN that is based on Estonian data from a one square kilometer grid. The proposed ecological network design consists of three principal layers: (1) general topographical features like coastlines, the water network, major roads and place names for locating the depicted network; (2) a habitat-based field of suitability for the ecological network, calculated on the basis of network values of landscape features

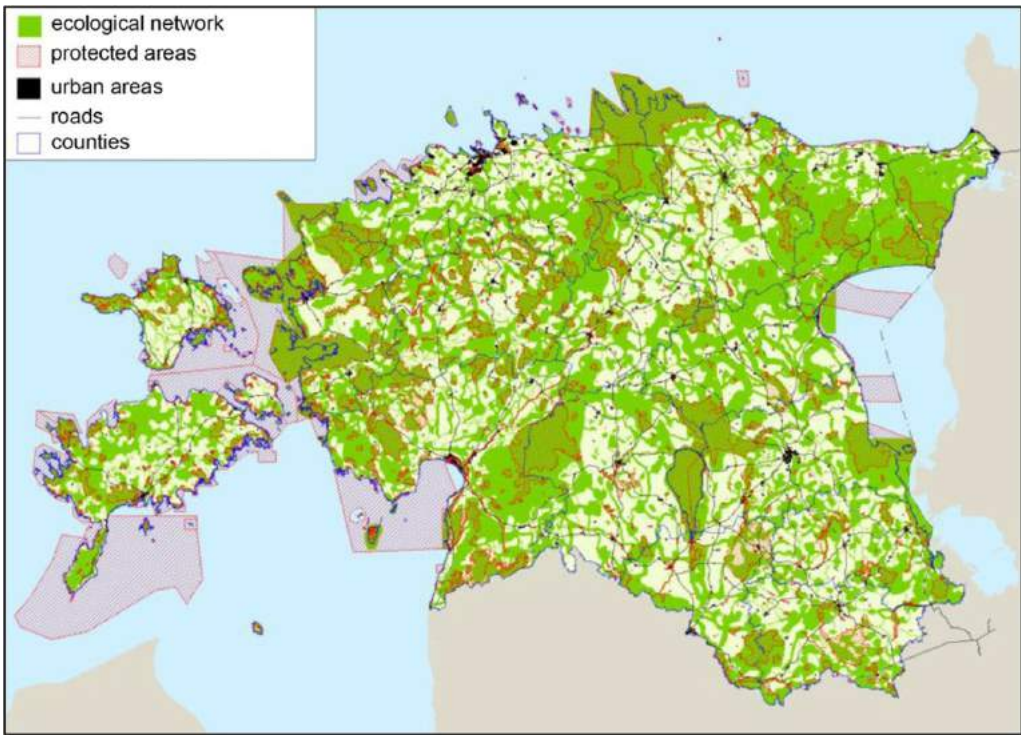


Fig. 6 Example of the ecological network of Estonia at the national level. Protected areas and areas not protected but suitable for an ecological network according to their present natural state. Adapted from Remm, K., Külvik, M., Mander, Ü., Sepp, K., 2004. Design of the Pan-European ecological network: a national level attempt. In: Jongman, R.H.G, Pungetti, G. (Eds.), New paradigms in landscape planning: ecological networks and greenways. Cambridge: Cambridge University Press, pp. 151–170.

using a predefined algorithm; and (3) the ecological network as an administrative decision. The second layer serves as a tool supporting decision making, while the third layer consists of the traditional components of an ecological network, such as core areas, corridors, buffer zones, and nature development/restoration areas. **Fig. 6** represents a combination of the last two layers as a map of protected areas (layer 3) and areas not protected but suitable for inclusion in ecological networks according to their present natural state (layer 2). Protected areas can be considered to be obligatory core areas of ecological networks, whereas areas suitable for ecological networks areas can be considered to be buffer zones and/or corridors.

See also: Ecological Data Analysis and Modelling: Forest Models. Ecosystems: Agriculture Systems. Terrestrial and Landscape Ecology: Ecological Engineering: Overview

Further Reading

Ahern, J., 1995. Greenways as planning strategy. Landscape and Urban Planning 33, 131–155.
Baguette, M., Blanchet, S., Legrand, D., Stevens, V.M., Turlure, C., 2013. Individual dispersal, landscape connectivity and ecological networks. Biological Reviews 88, 310–326.
Bastian, O., Schreiber, K.-F. (Eds.), 1999. Analyse und ökologische Bewertung der Landschaft. Heidelberg: Auflage, Gustav Fischer-Verlag. vol. 2, p. 564.
Baudry, J. and Merriam, G. (1988). Connectivity and connectedness: functional versus structural patterns in the landscapes. In: Schreiber, K.-F. (ed.) *Connectivity in landscape ecology. Münstersche Geographische Arbeiten*, 29. *Proceedings of the 2nd International Seminar of IALE*, Münster, pp. 23–28.
Bennett, G., 1998. Guidelines for the development of the pan-European ecological network. Draft. Council of Europe, Committee of Experts for the European Ecological Network 98 (6), 35. STRA-REP.
Bockemühl, J., 1982. Naturwissenschaftliche sektion. In: Erwachen an der Landschaft. Dornach, Switzerland: Naturwissenschaftliche Sektion. p. 320.
Bouwma, I.M., Jongman, R.H.G, Butovsky, R.O. (Eds.), 2002. Indicative map of the pan-European ecological network for central and Eastern EuropeTechnical background document.ECNC Technical Report Series. The Netherlands/Budapest: ECNC.101 pp + annexes. Tilburg.
Brooker, L., Brooker, M., Cale, P., 1999. Animal dispersal in fragmented habitat: measuring habitat connectivity, corridor use and dispersal mortality. Conservation Ecology 3 (1), 4.
Clay, G.R., Daniel, T.C., 2000. Scenic landscape assessment: the effects of land management jurisdiction on public perception of scenic beauty. Landscape and Urban Planning 49 (1–2), 1–13.
Costanza, R., d'Arge, R., de Groot, R.S., *et al.*, 1997. The value of the world's ecosystem services and natural capital. Nature 387 (6630), 253–260.
Cushman, S.A., Wallin, D.O., 2000. Rates and patterns of landscape change in the Central Sikhote-alin Mountains. Russian Far East, Landscape Ecology 15 (7), 643–659.
De Cola, L., 1994. Simulating and mapping spatial complexity using multi-scale techniques. International Journal of Geographical Information Systems 8 (4), 411–427.
De Groot, R.S., 1987. Environmental functions as a unifying concept for ecology and economics. Environmentalist 7 (2), 105–109.
De Groot, R.S., Hein, L., 2007. Concept and valuation of landscape functions at different scales. In: Mander, Ü., Wiggering, H., Helming, K. (Eds.), Multifunctional land use: meeting future demands for landscape goods and services. Berlin: Springer, pp. 15–36.

De Groot, R.S., Alkemade, R., Braat, L., Hein, L., Willemen, L., 2010. Challenges in integrating the concept of ecosystem services and values in landscape planning, management and decision making. Ecological Complexity 7, 262–270.
Dramstad, W.E., Tveit, M.S., Fjellestad, W.J., Fry, G.L.A, 2006. Relationships between visual landscape preferences and map-based indicators of landscape structure. Landscape and Urban Planning 78, 465–474.
Dunning, J.B., Danielson, B.J., Pulliam, H.R., 1992. Ecological processes that affect populations in complex landscapes. Oikos 65 (1), 169–175.
Farina, A., 1998. Principles and methods in landscape ecology. London: Chapman and Hall, p. 235.
Forman, R.T.T, 1995. In: Land mosaics, the ecology of landscapes and regions. Cambridge: Cambridge University Press. p. 632.
Forman, R.T.T, Godron, M., 1986. In: Landscape ecology. New York: Wiley. p. 619.
Gardner, R.H., O'Neill, R.V., 1991. Pattern, process, and predictability: the use of neutral models for landscape analysis. In: Turner, M.G., Gardner, R.H. (Eds.), Quantitative methods in landscape ecology. New York: Springer, pp. 289–307.
Gardner, R.H., Milne, B.T., Turner, M.G., O'Neill, R.V., 1987. Neutral models for the analysis of broad scale landscape pattern. Landscape Ecology 1 (1), 19–28.
Hanski, I., 1998. Metapopulation dynamics. Nature 396 (6706), 41–49.
Hendriks, K., Stobbelaar, D.J., van Mansvelt, J.D., 2000. The appearance of agriculture. An assessment of the quality of landscape of both organic and conventional horticultural farms in west Friesland. Agriculture, Ecosystems and Environment 77, 157–175.
Herzog, F., Lausch, A., Müller, E., *et al*., 2001. Landscape metrics for assessment of landscape destruction and rehabilitation. Environmental Management 27 (1), 91–107.
Hessburg, P.F., Smith, B.G., Salter, R.B., Ottmar, R.D., Alvarado, E., 2000. Recent changes (1930s–1990s) in spatial patterns of interior northwest forests, USA. Forest Ecology and Management 136 (1–3), 53–83.
Hobbs, R., 1997. Future landscape and the future of landscape ecology. Landscape and Urban Planning 37, 1–7.
Hudak, A.T., Fairbanks, D.H.K, Brockett, B.H., 2004. Trends in fire patterns in a southern African savanna under alternative land use practices. Agriculture, Ecosystems and Environment 101 (2–3), 307–325.
Isachenko, A.G., 1973. Principles of landscape science and physico-geographic regionalization. Melbourne: University of Melbourne Press.
Jongman, R.H.G, 1995. Nature conservation planning in Europe: developing ecological networks. Landscape and Urban Planning 32 (3), 169–183.
Jørgensen, S.E., 2007. Application of ecological engineering principles in landscape management. In: Mander, Ü., Wiggering, H., Helming, K. (Eds.), Multifunctional land use: meeting future demands for landscape goods and services. Berlin: Springer, pp. 83–92.
Kato, S., Ahern, J., 2011. The concept of threshold and its potential application to landscape planning. Landscape and Ecological Engineering 7, 275–282.
Keane, R.E., Parsons, R.A., Hessburg, P.F., 2002. Estimating historical range and variation of landscape patch dynamics, limitations of the simulation approach. Ecological Modelling 151 (1), 29–49.
Kiemstedt, H., 1994. Landscape planning – contents and procedures. Bonn: Nature Protection and Nuclear Safety, The Federal Minister of Environment, p. 124.
Klug, H., 2012. An integrated holistic transdisciplinary landscape planning concept after the Leitbild approach. Ecological Indicators 23, 616–626.
Krause, C.L., 2001. Our visual landscape. Managing the landscape under special consideration of visual aspects. Landscape and Urban Planning 54, 239–254.
Kuiper, J., 2000. A checklist approach to evaluate the contribution of organic farms to landscape quality. Agriculture, Ecosystems and Environment 77, 143–156.
Lausch, A., Herzog, F., 2002. Applicability of landscape metrics for the monitoring of landscape change, issues of scale, resolution and interpretability. Ecological Indicators 2 (1), 3–15.
Lee, J.T., Elton, M.J., Thompson, S., 1999. The role of GIS in landscape assessment: using land-use-based criteria for an area of the Chiltern Hills area of outstanding natural beauty. Land Use Policy 16 (1), 23–32.
Leser, H., 1978. Landschaftsökologie. Stuttgart: Ulmer Verlag.
Li, H., Reynolds, J.F., 1994. A simulation experiment to quantify spatial heterogeneity in categorical maps. Ecology 75, 2446–2455.
Luoto, M., Kuussaari, M., Rita, H., Salminen, J., von Bonsdorff, T., 2001. Determinants of distribution and abundance in the clouded Apollo butterfly, a landscape ecological approach. Ecography 24 (5), 601–617.
Mander, Ü., Koduvere, E., 2003. Pressure, state and response indicators in landscape assessment: an attempt on nitrogen fluxes. In: Helming, K., Wiggerimng, H. (Eds.), Sustainable development of multifunctional landscapes. Heidelberg: Springer, pp. 157–175.
Mander, Ü., Kuuba, R., 2004. Changing landscapes in northeastern Europe based on examples from Baltic countries. In: Jongman, R.H.G (Ed.), The new dimensions of the European landscape. Dordrecht, the Netherlands: Springer, pp. 123–134.
Mander, Ü., Murka, M., 2003. Coherence of cultural landscapes: a new criterion for evaluating impacts of landscape changes. In: Mander, Ü., Antrop, M. (Eds.), Advances in ecological sciences 16: multifunctional landscapes, vol. III: continuity and change. Boston, MA: WIT Press, pp. 15–32.
Mander, Ü., Jagomägi, J. and Külvik, M. (1988). Network of compensative areas as an ecological infrastructure of territories. In: Schreiber, K.-F. (ed.) *Connectivity in landscape ecology. Münstersche Geographische Arbeiten*, 29. *Proceedings of the 2nd International Seminar of IALE*, Münster, pp. 35–38.
Mander, Ü., Palang, H., Jagomägi, J., 1995. Ecological networks in Estonia. Impact of landscape change. Landschap 3, 27–38.
Mander, Ü., Külvik, M., Jongman, R., 2003. Scaling in territorial ecological networks. Landschap 20 (2), 113–127.
Mander, Ü., Uuemaa, E., Roosaare, J., Aunap, R., Antrop, M., 2010. Coherence and fragmentation of landscape patterns as characterized by correlograms: a case study of Estonia. Landscape and Urban Planning 94, 31–37.
Marsh, W.M., 2010. Landscape planning: environmental applications, 5th ed. Hoboken, NJ: Wiley, p. 511.
McGarigal, K. and Marks, B. J. (1995). FRAGSTATS: spatial pattern analysis program for quantifying landscape structure. *USDA Forest Service General Technical Report PNW-351*.
Meyer, B. C. (1997). Landschaftsstrukturen und Regulationsfunktionen in Intensivagrarlandschaften im Raum Leipzig-Halle. Regionalisierte Umwelt-qualitätsziele – Funktionsbewertungen-multikriterielle Landschafts-optimierung unter Verwendung von GIS, UFZ-Berichte 24/1997, pp. 1–224, Leipzig.
Meyer, B.C., 2001. Landscape assessment. In: Krönert, R., Steinhardt, U., Volk, M. (Eds.), Landscape balance and landscape assessment. Berlin: Springer, pp. 203–250.
Nakamae, E., Qin, X., Tadamura, K., 2001. Rendering of landscapes for environmental assessment. Landscape and Urban Planning 54 (1–4), 19–32.
Neef, E., Schmidt, G. and Luckner. M. (1961). Landschaftsökologische Untersuchungen an verschiedenen Physiotopen in Nordwestsachsen. Abh. Der Sächs. Akad. Der Wiss. Zu Leipzig, math.-nat. Kl., Bd. 47, H. 1 Berlin.
Ode, Å., Tveit, M.S., Fry, G., 2008. Capturing landscape visual character using indicators: touching base with landscape aesthetic theory. Landscape Research 33, 89–117.
Ode, Å., Fry, G., Tveit, M.S., Messager, P., Miller, D., 2009. Indicators of perceived naturalness as drivers of landscape preference. Journal of Environmental Management 90, 375–383.
Ode, Å., Tveit, M.S., Fry, G., 2010a. Advantages of using different data sources in assessment of landscape change and its effect on visual scale. Ecological Indicators 10, 24–31.
Ode, Å., Hagerhall, C.M., Sang, N., 2010b. Analysing visual landscape complexity: theory and application. Landscape Research 35, 111–131.
Palang, H., Mander, Ü., Luud, A., 1998. Landscape diversity changes in Estonia. Landscape and Urban Planning 41 (3–4), 163–169.
Palmer, J.F., Lankhorst, J.R.-K., 1998. Evaluating visible spatial diversity in the landscape. Landscape and Urban Planning 43 (1–3), 65–78.
Petit, C.C., Lambin, E.F., 2002. Impact of data integration technique on historical land-use/land-cover change, comparing historical maps with remote sensing data in the Belgian Ardennes. Landscape Ecology 17 (2), 117–132.
Purcell, A.T., Lamb, R.J., 1998. Preference and naturalness: an ecological approach. Landscape and Urban Planning 42, 57–66.
Remm, K., Külvik, M., Mander, Ü., Sepp, K., 2004. Design of the Pan-European ecological network: a national level attempt. In: Jongman, R.H.G, Pungetti, G. (Eds.), New paradigms in landscape planning: ecological networks and greenways. Cambridge: Cambridge University Press, pp. 151–170.

Sanderson, E.W., Jaiteh, M., Levy, M.A., *et al*., 2002. The human footprint and the last of the wild. BioScience 52 (10), 891–904.
Saunders, D.A., Hobbs, R.J., Margules, C.R., 1991. Biological consequences of ecosystem fragmentation: a review. Conservation Biology 51, 18–32.
Saunders, S.C., Mislivets, M.R., Chen, J., Cleland, D.T., 2002. Effects of roads on landscape structure within nested ecological units of the Northern Great Lakes Region, USA. Biological Conservation 103 (2), 209–225.
Saveraid, E.H., Debinski, D.M., Kindscher, K., Jakubauskas, M.E., 2001. A comparison of satellite data and landscape variables in predicting bird species occurrences in the Greater Yellowstone Ecosystem, USA. Landscape Ecology 16 (1), 71–83.
Scolozzi, R., Geneletti, D., 2012. A multi-scale qualitative approach to assess the impact of urbanization on natural habitats and their connectivity. Environmental Impact Assessment Review 36, 9–22.
Selman, P., 2012. Sustainable landscape planning: the reconnection agenda. Oxon: Routledge, p. 116.
Sevenant, M., Antrop, M., 2009. Cognitive attributes and aesthetic preferences in assessment and differentiation of landscapes. Journal of Environmental Management 90, 2889–2899.
Sochava, V.B., 1978. Introduction to study on geo-systems. Novosibirsk: Nauka, p. 319 (in Russian).
Solntsev, N.A., 1949. On mophology of natural geographical landscape. Voprosy Geografii 16, 61–86. (in Russian).
Spinozzi, F., Battisti, C., Bologna, M.A., 2012. Habitat fragmentation sensitivity in mammals: a target selection for landscape planning comparing two different approaches (bibliographic review and expert based). Atti della Accademia Nazionale dei Lincei Classe di Scienze Fisiche Matematiche e Naturali Rendiconti Lincei Scienze Fisiche e Naturali 23, 365–373.
Steiner, F., 2012. The living landscape: an ecological approach to landscape planning, 2nd ed. Washington, DC: Island Press, p. 496.
Tang, S.M., Gustafson, E.J., 1997. Perception of scale in forest management planning, challenges and implications. Landscape and Urban Planning 39 (1), 1–9.
Tress, B., Tress, G., 2001. Capitalising on multiplicity: a transdisciplinary systems approach to landscape research. Landscape and Urban Planning 57 (3–4), 143–157.
Troll, K., 1971. Landscape ecology (geoecology) and biogeocoenology – a terminological study. Geoforum 8, 43–46.
Tveit, M.S., 2009. Indicators of visual scale as predictors of landscape preference: a comparison between groups. Journal of Evnrionmental Management 90, 2882–2888.
Tveit, M.S., Ode, Å., Fry, G., 2006. Key concepts in a framework for analysing visual landscape character. Landscape Research 31, 229–255.
Uuemaa, E., Mander, U., Marja, R., 2013. Trends in the use of landscape spatial metrics as landscape indicators: a review. Ecological Indicators 28, 100–106.
Valles-Planells, M., Galiana, F., Van Eetvelde, V., 2014. A classification of landscape services to support local landscape planning. Ecology and Society 19 (1),
van Buuren, M., Kerkstra, K., 1993. The framework concept and the hydrological landscape structure: a new perspective in the design of multifunctional landscapes. In: Vos, C. C., Opdam, P.O. (Eds.), Landscape ecology of a stressed environment. London: Chapman and Hall, pp. 219–243.
van Mansvelt, J.D., 1997. An interdisciplinary approach to integrate a range of agro-landscape values as proposed by representatives of various disciplines. Agriculture, Ecosystems and Environment 63, 233–250.
van Mansvelt, J.D., Stobbelaar, D.J., Hendriks, K., 1998. Comparison of landscape features in organic and conventional farming system. Landscape and Urban Planning 41 (3–4), 209–227.
van Zanten, B.T., Verburg, P.H., Espinosa, M., Gomez-y-Paloma, S., Galimberti, G., Kantelhardt, J., Kapfer, M., Lefebvre, M., Manrique, R., Piorr, A., Raggi, M., Schaller, L., Targetti, S., Zasada, I., Viaggi, D., 2014. European agricultural landscapes, common agricultural policy and ecosystem services: a review. Agronomy for Sustainable Development 34, 309–325.
Viles, R.L., Rosier, D.J., 2001. How to use roads in the creation of greenways: case studies in three New Zealand landscapes. Landscape and Urban Planning 55, 15–27.
Virkkala, R., Luoto, M., Rainio, K., 2004. Effects of landscape composition on farmland and red-listed birds in boreal agricultural–forest mosaics. Ecography 27 (3), 273–284.
Wascher, D.M. (Ed.), 2000. Agri-environmental indicators for sustainable agriculture in EuropeECNC Technical Report Series. Tilburg: European Centre for Nature Conservation. p. 240.
With, K.A., Crist, T.O., 1995. Critical thresholds in species responses to landscape structure. Ecology 76 (8), 2446–2459.
With, K.A., Gardner, R.H., Turner, M.G., 1997. Landscape connectivity and population distributions in heterogeneous environments. Oikos 78, 151–169.
Zonneveld, I., 1995. In: Land ecology. Amsterdam: SPB Academic Publishing. p. 199.

Microcosms

FE Matheson, National Institute of Water & Atmospheric Research, Hamilton, New Zealand

Introduction

This article presents a scientific overview of the use of microcosms as a tool in the study of ecology. The article discusses what microcosms are, how and why microcosms have been, and continue to be, used for ecological research, and important factors to be considered in the design of ecological experiments using microcosms.

What Are Ecological Microcosms?

Microcosms are microecosystems. They are small, multispecies systems, consisting of a subset of the biotic community and abiotic properties of a larger ecosystem and have the common features of ecosystems such as food chains, production–consumption cycles, and hierarchies. A microcosm is a simplified, physical model of an ecosystem that enables controlled experiments to be conducted in the laboratory or *in situ*. They are often, and best, used in conjunction with theoretical mathematical models and field observational studies as part of a broader research strategy. Artificial microcosms are wholly or partially isolated from the external world in containers. Some natural microcosms also exist such as a phytotelmata (a contained aquatic habitat formed by a plant and populated by aquatic organisms, for example, bromeliads, tree holes) and communities confined to rock pools and moss patches.

Artificial microcosms may be derived directly from nature or gnotobiotic. Derived microcosms typically simulate a specific natural ecosystem, using species and components from it, and the exact species composition, particularly with respect to microorganisms, is unknown. In contrast, the exact species composition of gnotobiotic microcosms is known with these normally containing a set of physiologically well-studied species from pure (axenic) cultures that may or may not be normally found together.

Microcosms are similar to mesocosms but on a smaller scale. There are no strict definitions to delineate microcosms from mesocosms. However, microcosms are often considered to be structures of laboratory-bench scale while mesocosms are room size or larger. Due to their smaller size, microcosms are generally easier and less expensive to construct than mesocosms but as a result are often simpler systems that can only accommodate smaller ecological subjects, up to the size of small plants and animals (e.g., grasses, invertebrates). The term `macrocosm' is sometimes used to refer to the larger, natural ecosystems that the micro- or mesocosm system attempts to model.

Historical and Current Applications

Microcosms have been used in almost every area of terrestrial and aquatic ecology. They have long been used as a classroom teaching tool, bringing aspects of nature into the laboratory. The first ecological experiments using microcosms appear to be those of L. L. Woodruff in 1912 investigating protozoan succession in hay infusions. However, the majority of research with microcosms has been conducted since the 1960s. Early use of microcosms by prominent ecologists including G. F. Gause, H. T. Odum, R. Margalef, R. H. Whittaker, R. J. Beyers, G. D. Cooke, and E. P. Odum has contributed to the development of important concepts in ecology including the competitive exclusion principle, succession, self-organization, and the maximum power principle.

Microcosms continue to be widely used for general ecological studies. However, in recent times, they have also become popular tools to study the fate and effects of contaminants (e.g., heavy metals), pesticides and herbicides, stressors (e.g., high temperatures), novel compounds, and genetically engineered organisms (**Fig. 1**). They provide a comparatively safe means of assessing likely effects on ecosystems without direct exposure to the natural environment. Microcosms are not yet routinely used in ecotoxicological testing despite a strong argument from proponents that single-species tests are inadequate for full evaluation of ecosystem-level impacts. The problems with microcosm use relative to ecotoxicological single species tests include the higher costs, time involved, complexity, variability and difficulty in evaluating endpoints. Despite these issues, F. B. Taub and colleagues have worked to develop a standardized (nearly gnotobiotic) aquatic microcosm. This is now registered with the American Society for Testing and Materials as a standard method.

Design Factors

Sourcing, Seeding, and Energy Matching

Where a microcosm is designed to represent a model of a system that already exists, it should contain all of the characteristic features of the ecosystem that are necessary in the context of the problem to be described or solved. It is usually recommended that

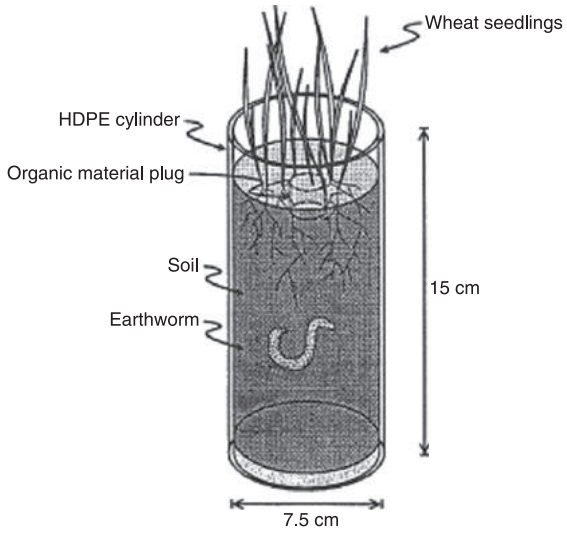


Fig. 1 A homogenous integrated soil microcosm, designed as a model terrestrial ecosystem to assess the effects of single pesticides on soil ecosystems. Reproduced from Burrows, L.A., Edwards, C.A., 2002. The use of integrated soil microcosms to predict effects of pesticides on soil ecosystems. European Journal of Soil Biology 38, 245–249.

Fig. 2 J. M. Quinn (National Institute of Water & Atmospheric Research) stands beside recirculating stream microcosms that have been used to investigate nutrient uptake and transfer in stream food webs and the effects of lighting and fine sediment deposition on stream biogeochemistry (Parkyn *et al.*).

components used in the microcosm (e.g., soil, water, plants, animals) are sourced from the natural ecosystem. It is also important to try and expose the microcosm to the same physical, chemical, and biological inputs or `energies' (e.g., light, temperature, nutrients, turbulence, species immigration) as the natural ecosystem. This is termed the energy signature approach to microcosm design. It can be more difficult to match some inputs for laboratory microcosms. For example, artificial light is a poor substitute for natural light. Constructing microcosms by isolating parts of the natural ecosystem *in situ* can minimize disturbance and enables matching of light and temperature inputs; however, other energies such as turbulence may not be equivalent (**Fig. 2**).

In synthesized, gnotobiotic microcosms, the researcher has the challenging role of system organiser while self-organization is prevalent in derived microcosms. A multiple seeding approach, where innocula from several natural assemblages are mixed together and left to self-organize, is a technique recommended by H. T. Odum to develop a more stable and sustainable microcosm system. Reinoculation can sometimes be necessary to maintain important species that do not develop sustainable populations.

Spatial Scaling, Wall, and Isolation Effects

Microcosm size affects the amount of diversity that the system can accommodate, with larger microcosms being able to support a greater diversity and more trophic levels, than smaller ones. Microcosm shape also has the potential to strongly impact on microcosm functioning and it can be useful to incorporate testing of microcosm size and/or shape into experimental design (**Fig. 3**). In particular, microcosm designs with a large wall surface area to volume ratio should be used with caution. The metabolic

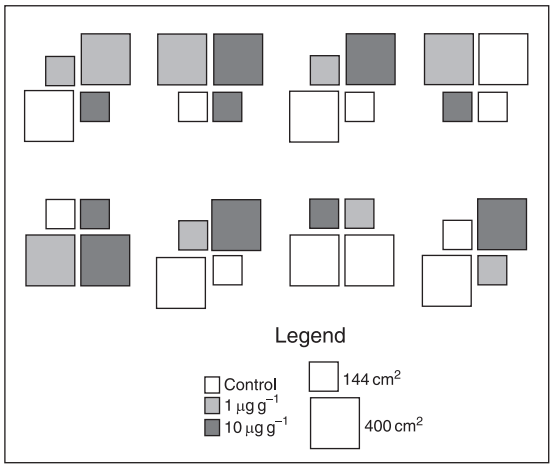


Fig. 3 Design of an experiment testing the effect of microcosm size and the pesticide chlorpyrifos on macroinvertebrate colonization of estuarine sediments. Average taxa richness was significantly higher in larger microcosms although average animal density was higher in smaller microcosms. In both large and small microcosms, animal density was significantly higher near the perimeter, indicating an `edge´ effect. Reproduced from Flemer, D.A., Ruth, B.F., Bundrick, C.M., Moore, J.C., 1997. Laboratory effects of microcosm size and the pesticide chlorpyrifos on benthic macroinvertebrate colonization of soft estuarine sediments. Marine Environmental Research 43, 243–263.

activity of microbial or periphyton biofilms (`edge communities') attached to these walls can be substantial and highly unrepresentative of natural conditions. To avoid these effects, larger microcosm volumes in relation to wall surface area are recommended. The composition of microcosm walls should also be considered. Wall materials should be inert and not leach or absorb substances that may affect the experiment. Gases such as oxygen can diffuse through more flexible plastics, which may or may not be desirable depending on the ecosystem being modeled. Consideration should also be given to the effects of artificial isolation, which restricts the movement of mobile organisms. The small size of microcosms also typically excludes higher trophic levels. However, the activities of some higher organisms or mobile species (e.g., grazing of vegetation, removal or replacement of individuals of a species by a predator or migration) may be simulated by human actions.

Temporal Scaling

A critical consideration for microcosm studies is the duration of experiments. Most microcosm experiments are generally conducted over a period of only weeks to months. However, the duration of microcosm experiments needs to be sufficient to assess effects on slow-responding organisms or processes.

As with natural ecosystems, conditions within microcosms can change over time and these changes should be evaluated during the course of a microcosm study. As the duration of a microcosm experiment increases, so does the likelihood of greater variability developing between replicates as a result of natural divergence. Time series sampling can be incorporated into experimental design to monitor changes. However, careful consideration of the impact of any repetitive sampling of components is required. A large number of microcosm replicates can be established at the outset of an experiment to enable complete (i.e., destructive) sampling of a subset of replicates at designated time intervals during the course of the study (**Fig. 4**).

Natural ecosystems are also subject to diurnal and seasonal variations as a result of light, temperature, and other climatic effects. In the laboratory, natural diurnal variations can be simulated to some extent by the use of controlled light–dark cycling of artificial lights (**Fig. 5**). For experiments of short duration, seasonal variability may be taken into account by repeating experiments on a seasonal basis.

Studies of longer-term ecological processes such as succession, predator–prey cycles, extinction, can be studied in microcosms with short real-time duration using organisms with very short generation times (e.g., microorganisms). This has been termed the biological accelerator approach.

Replication, Variability, and Divergence

Microcosm replication is an experimental design issue. The more complex the system to be studied, the more replicates are generally required to adequately describe and account for the associated variability in test results. Even when they are started similarly, microcosms often develop differently, particularly over longer periods of time. While this divergence can be problematic for microcosm replicability, this phenomenon does offer opportunities to test ecological theories and models about how different community structures can develop with, for example, different sequences of seeding (e.g., chaos and assembly theories, lottery and random models). The cross-seeding technique, where some of the contents of one replicate are regularly transferred to another, can

Fig. 4 Simple aquatic microcosms (4 l pails in a 1 m depth flow-through freshwater tank) used to investigate the effects of different sediment types on the growth responses of selected submersed macrophyte species (Matheson *et al.*).

Fig. 5 Riparian wetland soil microcosms set up in a climate-controlled laboratory and used to investigate the fate of nitrate and the effect of wetland plant growth on nitrogen transformation processes (Matheson *et al.*).

be a useful strategy to reduce variability and divergence among replicates. This is often done in the initial, setup stage but may also be incorporated into the experimentation period.

Sufficient replication in scientific experiments is critical to enable robust statistical evaluation of results. Analysis of variance (ANOVA) experimental designs are most commonly used in microcosm studies. Regression designs are sometimes employed which may enable testing of a broader range of treatments but results from these are more robust if some replication of treatments is also included. Replication enables a mean value for test results to be calculated along with the standard deviation and error of the mean. Three replicates is a recommended minimum for scientific investigations but higher numbers of replicates will provide more robust results. Power analysis and sample size estimation can be useful statistical techniques to employ to ensure that there is sufficient replication.

Similarity to Natural Ecosystem

Designing a microcosm or any model of a natural ecosystem (macrocosm) is a test in itself of how much is known about the ecosystem. A study using microcosms should include measurement of the characteristic biotic and abiotic features and functions of the natural ecosystem it is trying to model. Selection of features and functions to measure should be based on how critical they are to the natural ecosystem. These measurements enable an assessment to be made of how closely the microcosm represents the natural ecosystem. Any extrapolation of results from microcosm studies to natural ecosystems should be based on sound evidence of close matching of key features and functions between these systems (Fig. 6). However, ideally, results from microcosm studies should be confirmed with further field scale testing.

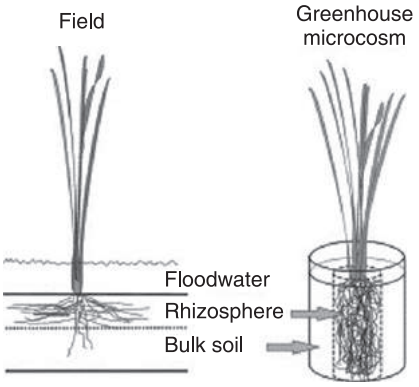


Fig. 6 Illustration of the different components examined in a study comparing the community structure of methane-oxidizing bacteria in microcosms to the natural ecosystem. In both systems, the main factors controlling the population size and activity of methane-oxidizing bacteria were plant growth and availability of nitrogen. Community diversity, activity patterns, and the population structure in both systems were comparable, although different quantities were detected. Reproduced from Eller, G., Kruger, M., Frenzel, P., 2005. Comparing field and microcosm experiments: A case study on methano- and methylo-trophic bacteria in paddy soil. FEMS Microbiology Ecology 51, 279–291.

Further Reading

Berg, G.M., Glibert, P.M., Chen, C., 1999. Dimension effects of enclosures on ecological processes in pelagic systems. Limnology and Oceanography 44, 1331–1340.
Beyers, R.J., 1963. Metabolism of twelve aquatic laboratory microecosystems. Ecological Monographs 33, 281.
Beyers, R.J., Odum, H.T., 1993. Ecological Microcosms. New York: Springer.
Burrows, L.A., Edwards, C.A., 2002. The use of integrated soil microcosms to predict effects of pesticides on soil ecosystems. European Journal of Soil Biology 38, 245–249.
Cooke, G.D., 1967. The pattern of autotrophic succession in laboratory microcosms. BioScience 17, 717–721.
Daehler, C.C., Strong, D.R., 1996. Can you bottle nature? The roles of microcosms in ecological research. Ecology 77, 663–664.
Eller, G., Kruger, M., Frenzel, P., 2005. Comparing field and microcosm experiments: A case study on methano- and methylo-trophic bacteria in paddy soil. FEMS Microbiology Ecology 51, 279–291.
Flemer, D.A., Ruth, B.F., Bundrick, C.M., Moore, J.C., 1997. Laboratory effects of microcosm size and the pesticide chlorpyrifos on benthic macroinvertebrate colonization of soft estuarine sediments. Marine Environmental Research 43, 243–263.
Gause, F.G., 1934. The Struggle for Existence. Baltimore: Williams and Wilkins.
Kangas, P.C., 2004. Ecological Engineering: Principles and Practice. Boca Raton, FL: Lewis.
Lawton, J.H., 1995. Ecological experiments with model systems. Science 269, 328–331.
Margalef, R., 1967. Laboratory analogues of estuarine plankton systems. In: Lauff, G. (Ed.), Estuaries. Washington, DC: American Association for the Advancement of Science, pp. 515–521.
Matheson, F.E., de Winton, M.D., Clayton, J.S., Edwards, T.M., Mathieson, T.J., 2005. Responses of vascular (*Egeria densa*) and non-vascular (*Chara globularis*) submerged plants and oospores to contrasting sediment types. Aquatic Botany 83, 141–153.
Matheson, F.E., Nguyen, M.L., Cooper, A.B., Burt, T.P., Bull, D.C., 2002. Fate of ^{15}N-nitrate in unplanted, planted and harvested riparian wetland soil microcosms. Ecological Engineering 19, 249–264.
Odum, E.P., 1971. Fundamentals of Ecology, 3rd edn. Philadelphia: WB Saunders.
Odum, H.T., Hoskin, C.M., 1957. Metabolism of a laboratory, stream microcosm. Publications of the Institute of Marine Science, University of Texas 4, 115–133.
Parkyn, S.M., Quinn, J.M., Cox, T.J., Broekhuizen, N., 2005. Pathways of N and C uptake and transfer in stream food webs: An isotope enrichment experiment. Journal of the North American Benthological Society 24, 955–975.
Sanderson, H., 2002. Pesticide studies: replicability of micro/mesocosms. Environmental Science and Pollution Bulletin 9, 429–435.
Srivastava, D.S., Kolasa, J., Bengtsson, J., *et al*., 2004. Are natural microcosms useful model systems for ecology? Trends in Ecology and Evolution 19, 379–384.
Taub, F.B., 1997. Unique information contributed by multispecies systems: Examples from the standardized aquatic microcosm. Ecological Applications 7, 1103–1110.
Warren, P.H., Law, R., Weatherby, R.J., 2003. Mapping the assembly of protist communities in microcosms. Ecology 84, 1001–1011.
Whittaker, R.H., 1961. Experiments with radiophosphorus tracer in aquarium microcosms. Ecological Monographs 31, 157–188.
Woodruff, L.L., 1912. Observations on the origin and sequence of protozoan fauna of hay infusions. Journal of Experimental Zoology 1, 205–264.
Wynn, G., Paradise, C.J., 2001. Effects of microcosm scaling and food resources on growth and survival of larval *Culex pipiens*. BMC Ecology 1, 3.
Yoshida, T., 2005. Toward the understanding of complex population dynamics: Planktonic community as a model system. Ecological Research 20, 511–518.

Organic Farming

Karen M Nielsen, Organic Denmark, Aarhus, Denmark

Glossary

CMS Cytoplasmic male sterility

Roughage Fresh or processed grass, leaves, stems and similar coarse materials with relatively high amount of indigestible material. Roughage contributes to a healthy intestinal flora and gives a prolonged satiety as well as green materials adds to a healthier composition of fatty acids in dairy products, eggs and meat.

Introduction

Organic farming is an agricultural production form which has become increasingly important globally, due to a series of challenges and problems caused by traditional, industrialized agriculture. This concerns contamination of soils, drinking water and surface water; erosion and depletion of soils; an intensive industrialized animal production where animals have no access to outdoor areas, which puts pressure on animal welfare; degradation of nature and decreasing biodiversity; social poverty and cultural impoverishment.

The organic production and the organic commodities that the production supplies the consumer market with continuously develops, consistent with a continuous build up and exchange of knowledge in parallel to an increasing acknowledgement of this type of agriculture by authorities and decision makers as a multi-functional tool to achieve improvements of society. Organic production deviates from conventional farming by being based on local fertility of soils, recirculation of resources, high diversity in crops, consideration for the surrounding nature, more space and access to outdoor areas for animals as well as prevention rather than treatment.

Definitions

In 2005 the international organization of organic farmers, the International Federation of Organic Agriculture Movements (IFOAM) adopted the following definition of organic farming:

> Organic Agriculture is a production system that sustains the health of soils, ecosystems and people. It relies on ecological processes, biodiversity and cycles adapted to local conditions, rather than the use of inputs with adverse effects. Organic Agriculture combines tradition, innovation and science to benefit the shared environment and promote fair relationships and a good quality of life for all involved (ifoam.bio, n.d.-a).

Organic farming is a production form that aims at exploiting soils and keeping livestock in a more sustainable manner than intensive, conventional farming, which is characteristic for modern industrialized agricultural production. The notion of organic farming seriously spread through Europe and the USA in the 1970s and 1980s as a reaction to agricultural environmental problems—problems which became apparent during this period and which were often debated publicly. Meanwhile, the origin of the idea of organic farming is much older.

Depending on language and nation, various terms are used to describe organic products, for instance, organic, ecological and biological/biologique. At the same time, the terms do have ordinary, semantic connotations—organic refers to the living; ecology is the scientific wording for the studies dealing with the household of nature, that is, the description of interactions among living organisms and their ambient environment; and biology deals with the learning and teachings about the living.

This concurrent set of meanings may lead to unclear perceptions concerning aspects of production, the origin of food items and their quality.

The terms mentioned above are protected terms regarding food within the European Union (EU) and are reserved products that meet the standard of certified organic production (europa.eu, n.d.).

In addition to this, the term agro-ecology is used about environmentally friendly and sustainable production forms in developing countries where production is not necessarily certified, for instance, due to lack of infrastructure.

History

Today organic farming is connected to a set of production standards, certification, labelling and branding with relations to a market economy. Meanwhile, the off-set to the idea is different, as the pioneers at the beginning of the 20th century saw the establishing of a good and fertile soil as a proper response to poverty and low yields in agriculture. Other aspects were respects for traditions, life forms and social wellbeing, that some already considered to be threatened by the process of industrialization.

Experiments with new techniques capable of increasing soil fertility were initiated all over the world. Sir Albert Howard attracted some attention by introducing ways of composting, and in 1924 Rudolf Steiner gave his famous speeches on agriculture (Steiner, 1924) and founded the biodynamic movement and praxis, which developed throughout the 1930s and gained some popularity. At the same time agriculture in general took a different turn, due to the chemist Justus von Liebig's theories on specific minerals as nutrients of plants. With this knowledge and recognition followed the development and use of artificial fertilizers, that after the Second World War took over as a prevailing paradigm.

Biodynamic agriculture represented the holistic approach to farming, until the idea of organic farming gained support in the 1970s in line with an increasing environmental awareness in society among—among other things triggered by the consequences of environmental pollution due to agricultural activities.

As opposed to biodynamic farming, organic farming takes science as its starting point, and organic farmers and their organizations worldwide seek from the very beginning to establish a foundation of practice based on experiments, tests and exchange of knowledge. The first experimental evidence and comparisons between organic and conventional farming methods were described in "The Living Soil" authored by Lady Eve Balfours. It formed the basis of the Soil Association in the UK as early as in 1946. Soil Association served as inspiration to grassroots all over the world. The purpose of the organization was to work for soil fertility, animal welfare, nutritious food products and farming which does not have a negative influence on the surrounding environment (soilassociation, n.d.).

It is the Soil Association which in 1972 takes the initiative to form the International Federation of Organic Agricultural Movements, IFOAM. Thereby, a formal cooperation among organic grassroot movements and NGOs all over the world is established. Today, IFOAM facilitates knowledge exchange, is responsible for international conferences and formulates the basic principles of organic farming.

Standards and regulations are based on these principles and have developed along with the development and establishing of real market for organic food products.

Four Governing Principles

Organic farming is based on four governing principles described and adopted by IFOAM, the international umbrella organization that organic movements worldwide are members of (ifoam.bio, n.d.-b). The principles deal with a set of overall ethical principles which are meant to serve as inspiration to implementation of concrete action locally and regionally. The four principles are:

Health

Organic agriculture should sustain and enhance the health of soil, plant, animal, human and planet as one and indivisible.

This principle points out that the health of individuals and communities cannot be separated from the health of ecosystems—healthy soils produce healthy crops that foster the health of animals and people.

Health is the wholeness and integrity of living systems. It is not simply the absence of illness, but the maintenance of physical, mental, social and ecological well-being. Immunity, resilience and regeneration are key characteristics of health.

The role of organic agriculture, whether in farming, processing, distribution, or consumption, is to sustain and enhance the health of ecosystems and organisms from the smallest in the soil to human beings. In particular, organic agriculture is intended to produce high quality, nutritious food that contributes to preventive health care and well-being. In view of this it should avoid the use of fertilizers, pesticides, animal drugs and food additives that may have adverse health effects.

Ecology

Organic agriculture should be based on living ecological systems and cycles, work with them, emulate them and help sustain them.

This principle roots organic agriculture within living ecological systems. It states that production is to be based on ecological processes, and recycling. Nourishment and well-being are achieved through the ecology of the specific production environment. For example, in the case of crops this is the living soil; for animals, it is the farm ecosystem; for fish and marine organisms, the aquatic environment.

Organic farming, pastoral and wild harvest systems should fit the cycles and ecological balances in nature. These cycles are universal but their operation is site-specific. Organic management must be adapted to local conditions, ecology, culture, and scale. Inputs should be reduced by reuse, recycling and efficient management of materials and energy in order to maintain and improve environmental quality and conserve resources.

Organic agriculture should attain ecological balance through the design of farming systems, establishment of habitats and maintenance of genetic and agricultural diversity. Those who produce, process, trade, or consume organic products should protect and benefit the common environment including landscapes, climate, habitats, biodiversity, air, and water.

Fairness

Organic agriculture should build on relationships that ensure fairness with regards to the common environment and life opportunities.

Fairness is characterized by equity, respect, justice and stewardship of the shared world, both among people and in their relations to other living beings.

This principle emphasizes that those involved in organic agriculture should conduct human relationships in a manner that ensures fairness at all levels and to all parties—farmers, workers, processors, distributors, traders, and consumers. Organic agriculture should provide everyone involved with a good quality of life, and contribute to food sovereignty and reduction of poverty. It aims to produce a sufficient supply of good quality food and other products.

This principle insists that animals should be provided with the conditions and opportunities of life that accord with their physiology, natural behavior and well-being.

Natural and environmental resources that are used for production and consumption should be managed in a way that is socially and ecologically just and should be held in trust for future generations. Fairness requires systems of production, distribution and trade that are open and equitable and account for real environmental and social costs.

Care

Organic agriculture should be managed in a precautionary and responsible manner to protect the health and well-being of current and future generations and the environment.

Organic agriculture is a living and dynamic system that responds to internal and external demands and conditions.

Practitioners of organic agriculture can enhance efficiency and increase productivity, but this should not be at the risk of jeopardizing health and well-being. Consequently, new technologies need to be assessed and existing methods reviewed. Given the incomplete understanding of ecosystems and agriculture, care must be taken.

This principle states that precaution and responsibility are the key concerns in management, development and technology choices in organic agriculture.

Science is necessary to ensure that organic agriculture is healthy, safe and ecologically sound. However, scientific knowledge alone is not sufficient. Practical experience, accumulated wisdom and traditional and indigenous knowledge offer valid solutions, tested by time.

Organic agriculture should prevent significant risks by adopting appropriate technologies and rejecting unpredictable ones, such as genetic engineering. Decisions should reflect the values and needs of all who might be affected, through transparent and participatory processes.

Organic Farming Management

When it comes to practical implementation of the principles of organic farming one will notice differences between countries and regions as well as among farmers. Often a common set of standards has been adopted which sets the standard for the practice of producers and production methods.

Guide to Best Practise

IFOAM has developed a guide "Best Practise Guideline for Agriculture and Value Chains," which sets the direction for formulation of specific standards for organic farming management. The guide contains examples of organic farming management of soil and water resources, husbandry, nature values and genetic diversity (ifoam.bio, n.d.-c) (Table 1).

Plant Production

In brief, plant production in an organic farming system deviates from that of conventional farming by the absence of mineral fertilizer, pesticides and genetically modified crops. As an alternative, organic farmers use organic fertilizers and leftovers from livestock and plants. Weeds, pests, and diseases are reduced by crop rotation, mechanical weeding and the establishing of biotopes with higher species diversity and adequate living space for beneficials.

Table 1 Best practise guideline for agriculture and value chains, selected examples

Soil and fertility
- Soil is protected from loss due to erosion and exposure to the elements. Soil is kept covered by living plants and mulch
- Organic matter content is increased. Farmers enhance biological activity of the soil and are careful with heavy equipment
- Perennials and agroforestry are promoted
- Farms obtain their soil fertility primarily from the farm itself due to crop rotation, recycling of plant residues, nitrogen-fixing species, cover crops and manure
- Manure from intensive conventional animal production is not used
- Farmers rely on crop rotation, natural enemies and biodiversity management to control pests, diseases and weeds
- Synthetic and toxic pesticides are avoided

Biodiversity
- Not all land on any given holding is used for production. At least some is set aside for biodiversity habitat
- Farmers maintain or re-establish natural vegetation areas around springs, along natural watercourses, on steep slopes and other sensitive parts of the ecosystem. Natural wetlands should not be drained
- Avoidance of mono-cropped areas
- Farmers choose varieties that can be multiplied on the farm and avoid varieties and breeds, that rely on a continued use of high levels of off-farm inputs
- Organic farmers choose organically grown seeds when possible. GMO breeds are not used
- Farmers raise animal breeds, that reproduce naturally and give birth without routine human intervention

Animal husbandry
- Farmers raise no more animals than can be carried by the land itself and takes into consideration the potential impact on pollution, non-renewably energy use, greenhouse gas emissions and nutritional profile of the animal products
- Feed is all organic
- Farmers provide animals with their most natural diet. A diversity of feeds and forage types is desirable
- Farmers create an environment where animals can access their food as much as possible in the field, e.g., through grazing or foraging for insects and worms
- Ideally, feed is grown on the farm itself or on closest possible farms and grasslands
- Animals are allowed to express their natural behaviors, they have access to outdoors, pasture and shelters if needed
- Animals are protected against stress. They are grouped and managed in a way, that keep them from harming each other
- Avoidance of mutilation
- Livestock health are maintained through proper diets and living conditions. Natural remedies are used before synthetic materials, if treatment is necessary

Atmosphere and energy
- Farmers optimize use of trees, permanent pastures and perennials to sequester carbon and reduce greenhouse gas emissions
- Farmers optimize manure use and slurry storage, application method and timing to prevent losses of methane and nitrous oxide
- All operations work to minimize carbon emissions from internal combustion engines and strive to increase energy efficiency and reduce dependence on non-renewable sources of energy

Meanwhile, we are not dealing with absolutes. Certain mineral fertilizers are still allowed in case of risk of severe crop damage. The same applies to pesticides based on naturally occurring compounds such as plant extracts. In some regions compounds that are known to have damaging effects on environment are still in use as an exception, for instance, copper against fungal diseases in fruits and wine. There is, however, a mutual understanding that such substances should be phased out. In addition, the ban against the use of genetically modified crops is continuously challenged by technological achievements in plant breeding. In practice, it has been shown that it is difficult to define whether modern technology can be identified as genetical modification or not. CMS hybrid seed production is an example of such technologies.

A relevant and ongoing discussion is for how many generations organic farming must be practised for the crops produced to be organic and certified as such. Would it be sufficient that the seeds used for sowing are derived from organically produced plants or should the parental generations to these plants also be produced according to organic standards? Today organic farmers do only to a certain extent have access to organic seeds in the market. Due to increased research efforts, knowledge distribution and growth in organic area and number of producers a gradual development towards increasing integrity in organic farming is generally observed. This will eventually apply to plant breeding and propagation too. Meanwhile, it has been shown to be difficult to keep plant material healthy and disease free when producing organically over generations.

Actual breed of varieties adapted to organic farming exists but is still not wide spread, primarily because the market for organically grown seeds is economically uninteresting to multinational companies working on plant breeding. Some organic farmers and organizations do work on locally adapted varieties and farm based breeding.

Nature and biodiversity

Organic farming affects soil, water, flora and fauna on and in soil surface in the cultivated areas as well as surrounding non-cultivated areas. Organic farmers are committed to consider their agricultural system as part of a larger ecosystem, and they must protect, restore or remediate natural elements on the farm. Organic farming has positive effects on nature and uncultivated areas close to the farms. It has been observed that organically grown fields and nearby biotopes contain 30% more plant and animal species than conventionally managed fields and their surroundings. This also applies to soil living organisms such as earthworms. The reason for difference is first and foremost the absence of chemical pesticides, but the use of organic manure and compost as well as a more diverse composition of crops—including perennial species—also play a significant role (Bengtsson et al., 2005; Hole et al., 2005).

Livestock Production

Organic livestock production is characterized by an ambition to offer production animals living conditions that meet their specific requirements and need to express normal behaviour in a system that supports health and a minimum need of medical interference.

Animal welfare

Animal welfare is central to the practice of organic livestock production. Animal welfare can be described as particular considerations to animals for their own sake. Animal welfare is difficult to measure. A common approach used is to define welfare in accordance with what the animals are free to do, and what they are free from. A definition based on such a view may be found in "the five freedoms" presented by the Farm Animal Welfare Committee, UK (gov.uk, n.d.) (Table 2).

Regarding animal welfare, free access to outdoor facilities for all animal groups for their entire life or at least part of it is central to organic livestock production. For cattle, an example might be access to pasture. For pigs and poultry, it might be access to express natural foraging activities such as digging and scraping in soils. The standard that animals must have access to outdoor areas is a simple standard, which contributes to ensure several of the freedoms mentioned above to individuals as well as to groups. Plenty of space and freedom to move around also allow the animals to express social behavior, forage and take care of their plumage and brood. Meanwhile, outdoor access does not solve all welfare related problems. It can be dangerous to be a chicken or a new born piglet in outdoor managed systems where they are exposed to attacks by predators. Access to a natural life does not necessarily correspond to good animal welfare. Nature is not considerate and just. Consideration of the natural behaviour of animals must therefore be matched against the protection that any animal keeper is obliged to exhibit. In free production systems surveillance and intervention is more difficult, and may consequently lead to increased mortality of piglets, calves and laying hens.

In an organic farming perspective animal welfare includes a divers and most natural diet. Therefore, it is mandatory that all animals—not only ruminants—are given roughage like clover grass and other green feeds. The diet for organic cattle is typically based on grass, whereas conventional diets are based on maize, corn, and waste products.

Table 2 Five freedoms

- Freedom from hunger and thirst; by ready access to water and a diet to maintain health and vigor
- Freedom from discomfort; by providing an appropriate environment
- Freedom from pain, injury and disease; by prevention or rapid diagnosis and treatment
- Freedom to express normal behavior; by providing sufficient space, proper facilities and appropriate company of the animal's own kind
- Freedom from fear and distress; by ensuring conditions and treatment, which avoid mental suffering

Clover grass and similar fodder are natural food sources of cattle, but also pigs and poultry may benefit from roughage. Roughage promotes a healthy intestinal flora and makes the animals feel full longer than corn and concentrates rich in energy would. In addition, it provides activity and can prevents stress and fights in a herd.

Animal health—prevention and treatment of diseases

The health of organic livestock is attempted to be maintained through fodder, access to the outdoor and the connected possibility to move, as well as attention to the overall surroundings in general. At the same time, organic livestock may be less productive, grow slower, give less milk etc. Hence illness is in general less abundant as in intensive, indoor, stable-based production systems.

Resistance and risks of infection

A significant societal impact of the organic production system is a noteworthy lower use of antibiotics in organic livestock. The use of antibiotics in the industrialized livestock production—curative as well as preventive/growth enhancing—leads to development of bacteria resistant to antibiotics, which poses a threat to public health. FAO/UN has adopted an action plan seeking to avoid antimicrobial resistance (FAO, n.d.).

Methicillin-resistant *Staphylococcus aureus* (MRSA) and Salmonella sp. infections are present in organic livestock production systems, but to a much lower extent than in conventional systems. A Danish investigation carried out in 2016 demonstrated MRSA to be present in 6% of organic pig herds as compared to 70%–80% of conventional pig herds. This despite the fact, that organic livestock brings in conventional animals for breeding purposes. A European research project on antibiotic resistance concludes that resistance is common among organic slaughter pigs, but occurs on significantly lower levels than is the case with conventional slaughter pigs (Aabo, 2014).

It may be difficult to point out causal relations, but an explanation might be that access to fresh air, space and roughage prevent the resistant bacteria from surviving and multiplying.

Global Organic Farming

Organic Farming in the Industrialized Part of the World

The market for organically grown products is particularly well developed in industrialized countries such as the United States and Europe, where purchasing power and willingness to pay for food free from pesticides and with high standard of animal welfare exists and is relatively high. Here one may find a wide spread environmental consciousness as well as a production and processing apparatus that is heavily regulated and authorized through legislation or standards and certification set by private organizations.

Organic farms in this part of the world are characterized by yields which are 20%–30% lower than yields obtained by normal conventional agriculture (de Ponti et al., 2012). The reason for this is primarily the absence of artificial fertilizers and pesticides.

The explanation for the lower yield is the fact that the soil potentials for production are almost fully exploited. When the use of artificial fertilizers, pesticides, GMO etc. are left out, the result will be a decrease in yields as crop rotation and natural enemies are not efficient to compensate. However, organic farming could on a longer time scale contribute to a significant carbon sequestration in soils, increased species diversity, more pollinators and similar effects which gives benefit to society. Such effects are not measurable on a short time scale.

Organic Farming in Developing Countries

While organic farming in western intensive agricultural system are characterized by a production yield being approx. 20% lower, organic farming at small farm scale in developing countries hold the potential to increase yields and stabilize economy. A UN-report from 2008 based on the examination of 114 African studies concludes that organic farming gives double the yield as conventional farming (UNEP-UNCTAD, n.d.). Other studies showed more diverse results, but all in all there seems to be a potential for yield growth by implementing organic farming methods.

The explanation of the difference is that the yield potential in intensive agricultural systems at present seems to have been exploited to a maximum level. This is far from being the case in extensive farming systems. The use of rotation crops, agroforestry and soil amendment using organic compost as fertilizer etc. contribute to an increased production per area unit.

Likewise, the use of artificial fertilizers and pesticides would possibly contribute to an increasing yield, but to many small-scale farmers such aids are out of reach as they require capital and thereby poses an increased risk in an environment that is already threatened by climate changes, draught, flooding, erosion, etc. Agricultural methods based on organic farming principles on the other hand makes the soil and crops more resilient with regards to such threats. At the same time, organic farming serves to stabilize yields, which is often more important to the family farmers than a potential but risky increase.

Organic or agro-ecological methods

Organic farming in developing countries is not necessarily certified. Certification connects to a market and trade with goods. For most small-scale farmers certification is not a crucial issue. Meanwhile, the implementation of organic farming principles can bring small scale farmers in a situation where they produce a surplus of food which makes access to a market a relevant issue.

An increasing middle class in the larger cities of developing countries demonstrates an increasing concern about food safety which leads to an increasing demand of organically grown food products without pesticides. Meanwhile, this market is still inaccessible to many small-scale farmers and excess production is sold at the local markets without certification. When a farmer uses techniques of cultivation based on organic farming principles but does not meet the demands of certified production one may say the food item is produced by agro-ecological methods.

Organic Farming and the SDG's of United Nations

In 2015 The United Nation (UN) adopted 17 Sustainable Development Goals for achieving a sustainable world (SDG's). The SDG's outlines the direction for national and international strategies of development and points out targets to achieve the goals towards 2030 (un.org, n.d.).

The principles of organic farming in many ways comply with the SDG's and offers solutions that can be achieved within several goals. This is particularly valid to the sustainable use of resources, improvements of health and the fight against pollution. According to Food and Agricultural Organization (FAO) of the United Nations, it is possible to improve food safety, local rural development, and sustainable living and environmental protection by implementing agricultural farming and through capacity building (fao.org, n.d.) (Figs. 1–5).

Growth and Extension of Organic Farming

Organic farming is growing on a global level. This applies to cultivated areas, the number of organic farmers and turnover expressed in market share as well as actual value (Willer & Lernoud, 2017) (Table 3).

Area and Area Use

50.9 millions hectares are managed organically, which corresponds to 1.1% of the globally cultivated areas. In addition, 39.7 millions hectares are non-cultivated organic areas for production of honey, collection of raw materials, forests, etc. The area used for organic farming has grown to a level which is fivefold that of 1999. Two thirds of the organic farming area are permanent grasslands. This crop type has increased dramatically and reflects among others the restructuring of vast areas in Australia to organic farming. Other perennial crops, like coffee and fruits, make up 8.0% of the area—the largest crop being coffee followed by olives and nuts. Only 20% of the area used for organic culture is in rotation—with corn being the dominating crop.

Organic Producers

The number of organic farmers on a world level is uncertain due to lack of data on the topic. The latest accounting estimates 2.4 millions of organic farmers. Approximately one third of these live in Asia—India being the country with the highest number of organic farmers. In the period from 2014 to 2015 the number of organic farmers has increased by 21%.

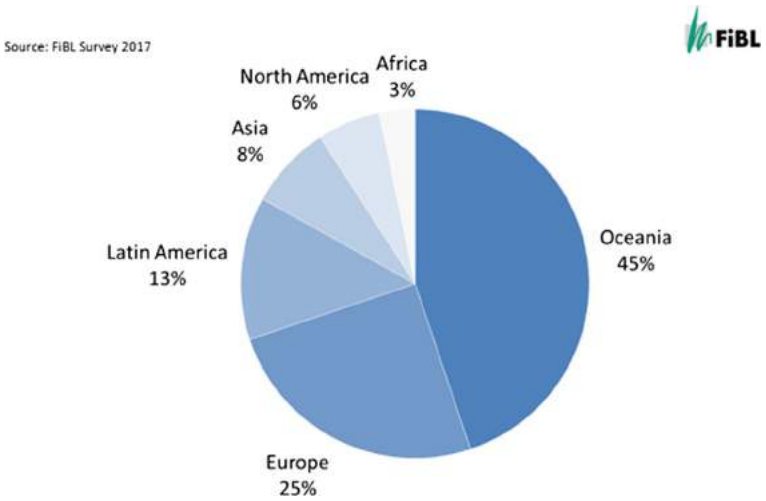


Fig. 1 Distribution of organic agricultural land by region 2015.

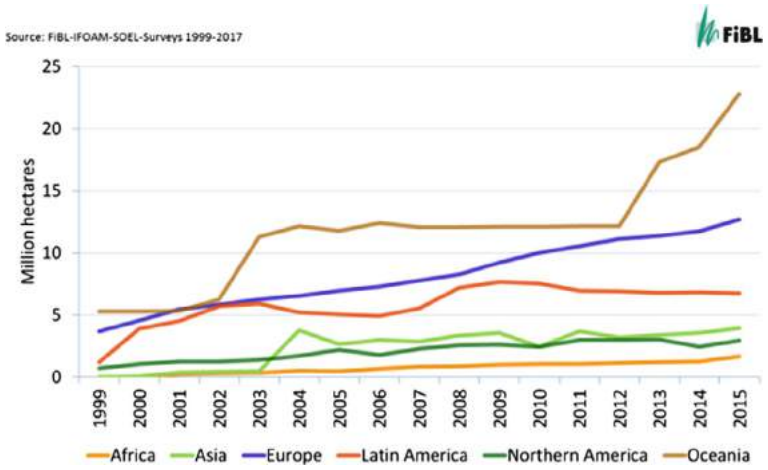


Fig. 2 Growth of the organic agricultural land by continent 1999–2015.

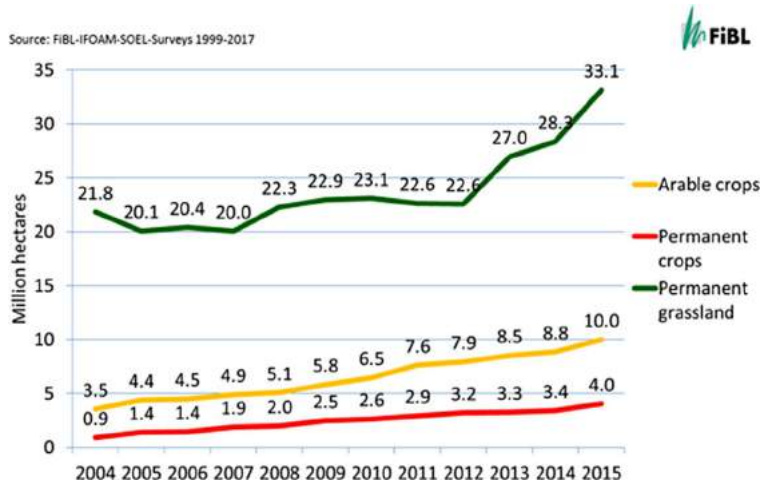


Fig. 3 Development of the organic land by land use type 2004–2015.

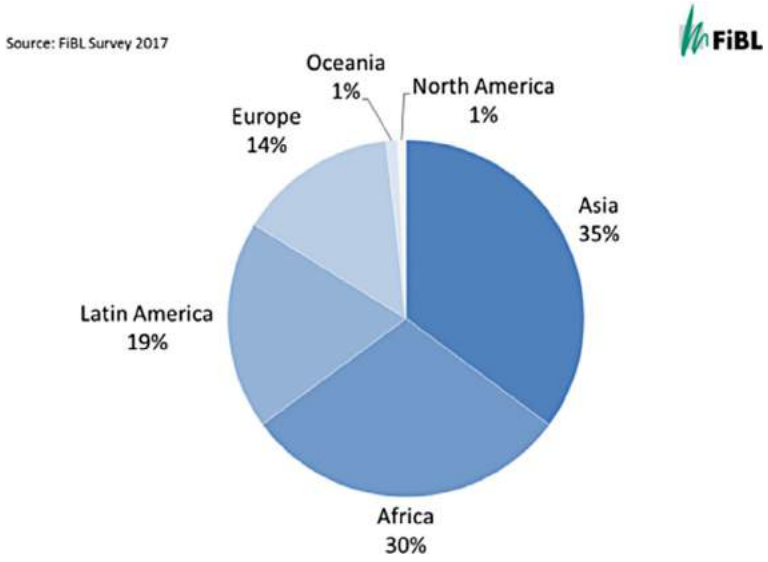


Fig. 4 Distribution of organic producers by region 2015.

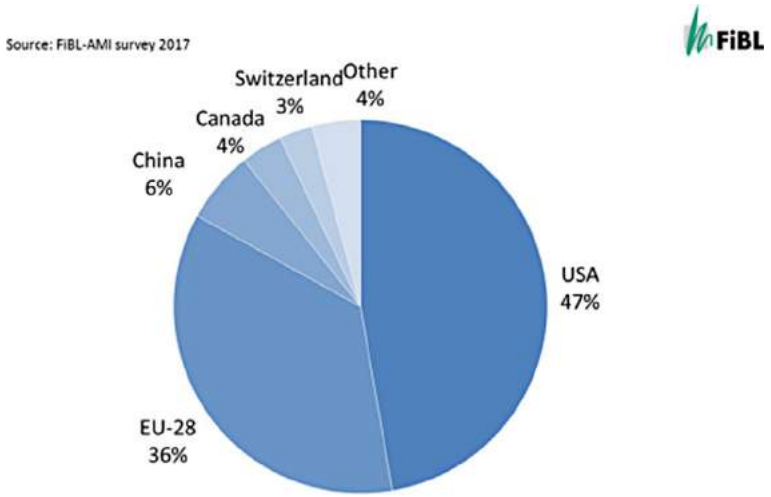


Fig. 5 Distribution of retail sales by single market 2015.

Table 3 17 sustainable development goals—goals where organic farming may play a particularly important role and offer relevant solutions are in bold (by author)

End poverty
End hunger
Improved health and wellbeing
Ensuring quality and access to education
Gender equality
Ensuring clean water and sanitation
Ensuring access to energy
Decent jobs and economic growth
Industrial innovation and infrastructure
Reduce inequality
Sustainable cities and communities
Responsible consumption and production
Climate action
Sustainable use of the oceans and marine resources
Sustainable use of terrestrial ecosystems
Peace, justice and strong institutions
Partnerships for the goals

The Organic Market

The turnover of organic food at world level is estimated to 75.7 billion €, half of this in the United States alone. The turnover corresponds to a per capita average of 10.3 €. The numbers are uncertain due to lack of reports.

References

Aabo S (2014) *Final Report for the CORE Organic II funded project "Restrictive use of antibiotics in organic animal farming—A potential for safer, high quality products with less antibiotic resistant bacteria—SafeOrganic"* National Food Institute, Technical University of Denmark (DTU).
Bengtsson J, Ahnström J, and Weibull AC (2005) The effects of organic agriculture on biodiversity and abundance: A meta-analysis. *Journal of Applied Ecology* 42: 261–269.
de Ponti T, Rijk B, and van Ittersum MK (2012) The crop yield gap between organic and conventional agriculture. *Agricultural Systems* 108: 1–9.
europa.eu, n.d., http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:1991R2092:20071201:DA:PDF.
FAO, n.d., The FAO Action Plan on Antimicrobial Resistance 2016–2020, http://www.fao.org/3/a-i5996e.pdf.
fao.org, n.d., http://www.fao.org/organicag/oa-home/en/.
gov.uk, n.d., https://www.gov.uk/government/groups/farm-animal-welfare-committee-fawc.
Hole DG, Perkins AJ, Wilson JD, et al. (2005) Does organic farming benefit biodiversity. *Biological Conservation* 122: 113–130.
ifoam.bio, n.d.-a, https://www.ifoam.bio/en/organic-landmarks/definition-organic-agriculture.
ifoam.bio, n.d.-b, http://www.ifoam.bio/en/organic-landmarks/principles-organic-agriculture.
ifoam.bio, n.d.-c, http://www.ifoam.bio/sites/default/files/best_practice_guideline_v1.0_ratified_withcover.pdf (modified).
soilassociation, n.d., www.soilassociation.org/aboutus/ourhistory.

Steiner, R. Geisteswissenschaftliche Grundlagen zum Gedeihen der Landwirtschaft; 1924, http://anthroposophie.byu.edu/vortraege/327.pdf.
un.org, n.d., http://www.un.org/sustainabledevelopment/sustainable-development-goals/.
UNEP-UNCTAD, n.d., UNEP-UNCTAD Capacity-building Task Force on Trade, Environment and Development (CBTF). Organic Agriculture and Food Security in Africa. New York and Geneva: United Nations, 2008. http://www.unctad.org/en/docs/ditcted200715_en.pdf.
Willer, H. and J. Lernoud (Eds) (2017), The World of Organic Agriculture. Statistics and Emerging Trends 2017. Research Institute of Organic Agriculture (FiBL), Frick, and IFOAM—Organics International, Bonn, FiBL survey 2017, figurtitel.

Permaculture

Kevin Morel, Catholic University of Louvain (UCL), Louvain-la-Neuve, Belgium
François Léger, UMR SADAPT, AgroParisTech, INRA, University of Paris-Saclay, Paris, France
Rafter Sass Ferguson, Haverford College, Haverford, PA, United States

Glossary

Agroecosystem The basic unit of study in agroecology that is defined as a spatially and functionally coherent unit of agricultural activity, which includes biophysical (soil, climate, plants, animals) and social components (human practices, values, objectives, organizations) and their interaction.
Agroforestry Land use management system in which trees or shrubs are grown around or among crops or pastureland.
Emergy Is a methodology which aggregates all different forms of energy and resources (e.g., sunlight, water, fossil fuels, minerals) used in the work processes that generate a product or service.
Food forest Polyculture mimicking forest ecology with multiple plant layers (annual plants, shrubs, trees, and liana) which produce a diversity of edible produce.
Holistic Refers to a global thinking or design approach which aims to integrate all dimensions of a situation (which can involve subjective and objective aspects) rather than analyzing only one aspect.
Intercropping Growing different plant species together on the same plot at the same time.
Modern/premodern/postmodern "Modern" refers to a philosophical movement developed in Europe since the 17th century relying on the idea that mastering the material world through rational knowledge will guarantee human progress and emancipation from nature, which is perceived as distinct from humans. "Premodern" refers to traditional worldviews which were born before modernism and where human beings are often seen as part of the natural world. "Postmodern" refers to a thinking tendency which criticizes the modern beliefs around progress and in which all assumptions are open to question. According to postmodern thinkers, elements from different systems and traditions can be combined without regard for any fixed aesthetic or tradition.
Silvopastoralism Land use management system in which animals graze in habitats where trees are present. Animals can feed partially on these trees (fruit fallen on the ground) and benefit from the microclimate they create (shadow, temperature, protection against wind).

Brief Overview of Permaculture

Permaculture is an international grassroots network focused on the sustainable design of human settlement, both in rural and urban areas although it was initially developed in a rural setting. Permaculture's central concept is that humanity can reduce or replace energy and pollution-intensive industrial technologies, especially in agriculture, through intensive use of biological resources and

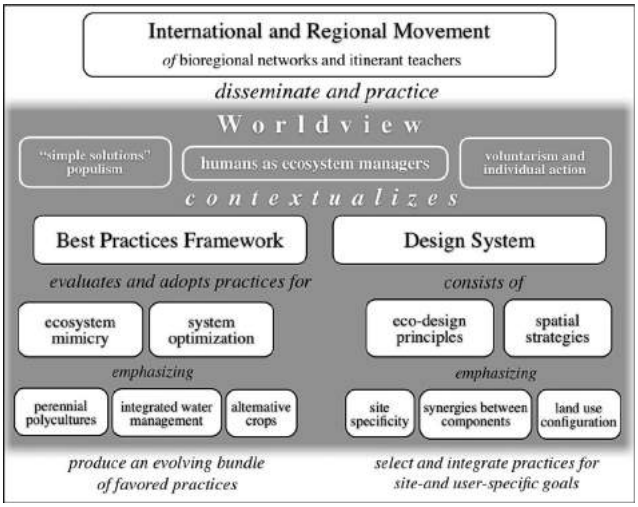


Fig. 1 Stratified definition of permaculture (Ferguson and Lovell, 2014).

thoughtful, holistic, design, patterned after natural ecosystems (eco-mimicry). Despite a relatively high public profile and broad international distribution, until recently permaculture has received little scholarly attention.

The definition of permaculture varies among sources and expands over time. In the founding text, permaculture's originators define it as "an integrated, evolving system of perennial or self-perpetuating plant and animal species useful to man" (Mollison and Holmgren, 1978). By 2002 Holmgren defined permaculture more broadly, encompassing broader issues of human settlement while maintaining an agricultural focus: "Consciously designed landscapes which mimic the patterns and relationships found in nature, while yielding an abundance of food, fiber and energy for provision of local needs" (Holmgren, 2002). Recent scholarship has identified four levels or components within permaculture presented in Fig. 1 (each of which may be referred to by the term): the international movement, the worldview carried by and disseminated by the movement, the design system, and the set of associated practices (Ferguson and Lovell, 2014). We will introduce each in turn.

Worldview

Key elements of the permaculture worldview include a theory about human–environment relations, a populist orientation to practice, and a model of social change. The permaculture literature highlights the positive role of humans in the landscape, as ecosystem managers. This perspective is expressed through a literature-wide insistence on the need for holistic planning and design and an optimistic assessment of what these styles of management can achieve. This perspective on human–environment relations cuts against the grain of the dualistic worldviews of both growth-oriented development and preservation-oriented conservation, each of which describe a fundamental conflict between the needs of society and those of nature. At the core of the permaculture worldview is the idea that—with the application of ecologically informed holistic planning and design—humans can meet their needs while increasing ecosystem health.

Design

The permaculture design system utilizes ecological and systems-thinking principles, and spatial reasoning strategies, which are used to analyze site conditions, select practices, and integrate them with site conditions and land use goals. The most distinctive aspects of the permaculture orientation toward agroecosystem design are its emphases on (1) site specificity, including attention to microclimate; (2) interaction between components at multiple scales, from field-scale polycultures to agroecosystem-scale land use diversity; and (3) spatial configuration as a key driver of multiple functions.

Practice

Land use in permaculture shares much with agroecology, agroforestry, and traditional and indigenous land use. Since the techniques associated with permaculture rarely originate from within the movement itself, the practical stratum is better regarded as a best practices framework than a bundle of techniques. Best practices in permaculture are evaluated by two broad criteria of ecosystem mimicry and system optimization. Ecosystem mimicry regards the structure and function of unmanaged ecosystems as models and attempts to create highly productive systems with analogous structure and function using species that produce yields for human use. System optimization does not refer to a model ecosystem, but seeks to identify strategic points of leverage where minimal intervention may enhance performance of desired functions beyond that of naturally occurring systems. Together, these criteria outline an implicit conceptual framework for the evaluation of practices in the permaculture movement.

Movement

The permaculture movement communicates the worldview and disseminates elements of practice and design through networks of practitioners and small institutes. The growth and dissemination of permaculture is built on two basic patterns: a widely dispersed network of "itinerant teachers" and local/regional organizing based around "bioregional" cultures and the development of alternative economic and social institutions. The permaculture movement today consists of a loosely affiliated network of individuals and projects, connected through permaculture courses and workshops, online forums, and local projects, as well as through and regional, national, and international convergences. Groups generally display a low level of institutionalization, and projects encompass a wide variety of functions, commonly including community gardens, campus greening initiatives, educational efforts, and less commonly, demonstration and/or research sites, periodicals, and farming-focused education and support efforts.

Conceptual Foundation and Dissemination

The Genesis of Permaculture

"To many of us who experienced the ferment of the late 1960s, there seemed to be no positive direction forward, although almost everybody could define those aspects of the global society that they rejected. These included military adventurism, the bomb, ruthless land exploitation, the arrogance of polluters and a general insensitivity to human needs. An unethical world could waste more on killing people than on earthcare or on helping people."

This quotation is from Bill Mollison, creator of permaculture and coauthor with David Holmgren of the founding book "*Permaculture one*" published in 1978. Permaculture is anchored in the multifaceted critical movements that emerged in the late 1960s with the North American counterculture and the birth of human ecology. These movements relied on emerging critiques of the resource-intensive materialism of consumer society, of sexism and racism at home, and militarism, imperialism, and unequal development leaving third world countries behind. In this context, social movements emerged with radical propositions for new ways of organizing society that could act as an alternative to a socio-economic system rooted in overexploitation of natural resources and the exponential growth of energy consumption, consumerist individualism, and the political and moral norms that of the economic elites. While some of these movements engaged in fairly "classical" political struggle, permaculture was among those that spurned conventional movement politics to work directly on concrete interventions, practical solutions for building an "other world," one whose key-word would be self-sufficiency.

The "back to the land" projects that multiplied in the 1970s, and spread from Australia to Europe via California, are part of this latter logic and set the stage for the emergence of permaculture. Faced with the hegemony of the dominant socio-economic model, their bearers sought to "withdraw from the world" by settling in isolated and/or abandoned by industrial development areas, in hopes of using the practice of traditional agriculture to rebuild a premodern link with nature. They acknowledged that nature could not entirely be grasped by rationality and that a (re)-developing a subjective and respectful relationship to nature was critical. The romantic or naturalist inspiration (e.g., Thoreau's Walden) of this movement was obvious, as was its apocalyptic dimension: this world, which in its irrepressible greed seemed to want to destroy its natural environment irremediably, would end before long. Those who have built and preserved havens based on the renunciation of a utilitarian and dominating vision of nature, would be the guarantors of the salvation of humanity.

Permaculture's founders and early adopters articulated a set values and principles in parallel with identifiable currents of ecological thought, based on the belief that industrial societies based on fossil fuel threaten the survival of human beings, on the rejection of anthropocentrism, and on a holistic worldview that opposes utilitarian reductionism. These positions were close to James Lovelock's Gaia Hypothesis or Deep Ecology (Naes, 1973). The relationship with deep ecology is particularly evident in Bill Mollison's later remarks (AtKisson, 1991): "Permaculture exhorts a total cooperation with every other and every other thing, animate and inanimate". For the founders of permaculture, this cooperation between humans and nonhumans is the basis of a global transformation of societies respecting three fundamental ethical principles: caring for the Earth; caring for people; and establishing limits on consumption and redistributing surplus (Mollison and Holmgren, 1978). They believe that this transformation must start from the initiatives of individuals anxious to act by and for themselves, re-building communities as they reconstruct human-environment relationships—gradually and from the bottom up. This logic of horizontal and bottom-up construction of a new society suggests affinities between permaculture and the nonviolent and ecological component of the anarchist movement of the end of the 19th century, of which Elisée Reclus, Geographer and French anarchist activist, is one of the most prominent figures, and which would be renewed in the United States in the latter decades of the 20th century with activist thinkers like as Murray Bookchin.

In the same interview cited above, Bill Mollison, however, refutes this connection. He rejects any form of power relationship or coercion as inseparable, from his point of view, from political action, even anarchist, and considers that the multiplication of individual initiatives cooperating with each other is enough to change the world. If permaculture is claimed to be subversive, this subversion does not involve political struggle, but rather a gradual dissemination of a belief translated into located concrete experiences: the construction of a sustainable world requires the reincorporation of humans into ecosystems natural. Permaculture proposes principles, conceptual tools that can guide the action of each one in this direction. It is thus defined as an "aid to the decision-making ethic" (Holmgren, 2002). In this perspective, the expansion of the permaculture network may require training in

principles and tools, but this training is more about awakening to another way of being in the world than the acquisition of established technical knowledge.

Permaculture, a Pragmatic Ecology for Self-Sufficiency

Permaculture proposes pragmatic methodological principles to create autonomous, resilient, and equitable living spaces. For Bill Mollison and David Holmgren, the fundamental flaw of industrial societies lies in the inextinguishable thirst for energy that structures their development and precludes any long term sustainability. To escape from this addiction, they postulate that permaculture design must be inspired by the structure and function of natural ecosystems. This perspective is directly inspired by the works of scientific ecology, particularly those of Eugene Odum, and even more by the approach of ecological thermodynamics and environmental accounting proposed by Odum (1971, 1995). These works are one of the main references cited by Holmgren (2002) in "Permaculture: Principles and Pathways beyond Sustainability," a book in which he resumes and deepens the principles of design defined in "Permaculture One" (Mollison and Holmgren, 1978).

In this line, permaculture interprets the dynamics of natural ecosystems as an accumulation of energy that drives ecosystems toward "closed loop" cycles of matter, in which less and less materials are lost from the ecosystem over time. Inspired by these natural processes, the design of self-sufficient human settlements must emulate ecosystems by maximizing the interrelations and synergies between the various human and non-human components and trigger a dynamic of aggradation. Perennial elements, especially trees and soils, play an essential role in this process by storing energy and carbon. For permaculturists, biodiversity and agrobiodiversity are valued for the functional redundancies they create and their beneficial effect on resilience, e.g. the provision of high-energy foods should be provided by cereal crops as well as root vegetables or trees producing fruits rich in starch. The same element must also fulfill several functions, for example, a legume supplies of protein and improves the soil fertility; a pond stores water and helps regulate the microclimate. Design must project itself into the future, the landscape it draws is an evolutionary structure and this evolution must be considered as much as possible at the outset. Thus, by planting trees, one must imagine how they will grow and what will be the consequences for herbaceous plants at their feet. The elements to cover human needs should be as much as possible found or produced within the system, and efforts should be made to minimize these needs. Self-sufficiency is thus an objective as much as a means of the project. Under these conditions, human settlements can be part of a process of global ecological and human improvement, in which the needs for inputs of energy and materials as well as human labor diminish gradually.

Holmgren (2002) has defined 12 principles of permaculture design. These principles form the basis of a reflective design process geared toward outcomes that align with the principles described above and the underlying ethical principles. These principles are: (1) observe and interact, (2) catch and store energy, (3) obtain a yield, (4) apply self-regulation and accept feedback, (5) use and value renewable resources and services, (6) produce no waste, (7) design from patterns to details, (8) integrate rather than segregate, (9) use small and slow solutions, (10) use and value diversity, (11) use edges and value the marginal, (12) creatively use and respond to change. Each principle is individually described and discussed in the permaculture literature with concrete design illustrations. For example, principle (6) highlights that waste production should be as low as possible and that their recycling must be systematized, as in nature where the concept of waste does not exist since elements are used and recirculated locally. Permaculture books presents simple solutions to apply this principle, such as breeding poultry to transform kitchen waste into eggs, meat and manure. The implementation and combination of all principles implies and demands a systemic vision. For Mollison and Holmgren this vision cannot be achieved through exhaustive analytical knowledge of the ecosystem, its components, and its mechanisms—which is in any case unattainable. Rather, it must be the result of a holistic, sensitive, and critical understanding of the place, for which scientific knowledge is merely one form of support among other aesthetic, spiritual, or moral considerations.

This holistic reading of space implies that permacultural design combines an objective perspective, based on empirical and/or scientific knowledge, and a subjective perspective, reflecting personal sensitivity. To achieve this difficult synthesis, Bill Mollison was inspired by the Australian aborigines, with whom he had worked for many years during his academic career in Tasmania (Mollison, 1988). Aboriginal thought is organized around the central concept of Dreamtime, the original cosmological dimension in which the different spirits and ancestors physically shaped the world, physically impregnating it with the organizing patterns that underlie the "just order" of things. In order to understand these patterns and the relationships between them, the observation of nature is central. It not only involves the intellect but also the intuition and the humble and silent perception of the world called dadirri by the aborigines. As in aboriginal thought, permaculture invites both objective and sensitive observation of landscapes, enabling us to identify the patterns and interfaces that structure it and on which design will have to rely. The purpose of this design process is not only utilitarian. It is the conjunction of utilitarian, spiritual, ethical, and moral dimensions that makes this space a "place for life," inhabited more than occupied, shared with other living species.

Presence of Permaculture in the World

From the foundational work articulated by Mollison and Holmgren in the late 1970s, permaculture concepts, worldview and practices have been spread through a quickly growing and largely decentralized, informal movement (Ferguson and Lovell, 2015). Given its Australian origin, permaculture first was disseminated in the 1980s in industrialized English speaking countries (mainly Australia, United States and Britain) through the development of generally small-scale projects designed to increase individual,

family or community self-sufficiency as a response to the growing environmentalist concerns, especially about peak oil. Most of these initiatives aimed to make people "responsible and productive citizens" instead of being consumers dependent on fossil fuel driven economy and production. Since the 1990s, permaculture was brought to southern countries mainly by northern NGOs and activists as a framework for enhancing the sustainable development and resilience of marginalized communities facing the issues of limited resources, climatic uncertainty, and social inequality.

Permaculture projects are now present in more than 120 countries on all continents around 2500 permaculture projects were referenced in 2017 (international permaculture website permaculturegloba l.org). The number of permaculture projects led by NGOs or civil associations is estimated around 4000 including 140 humanitarian projects. The Permaculture Design Certificate (PDC) is considered by many permaculturists as a mandatory "entrance point" to permaculture. The PDC can be obtained after a collective and participatory training session intended to provide participants with a global view of the permaculture framework and the design tools which would help them to further carry out their own experimentations. The collective dimension of PDC training aims to create strong links between students which can lead to future collaborations, exchanges of know-how, feedbacks on putting permaculture into practice and contributes to the vitality of interactions between practitioners. The number of people with PDC is estimated from 100,000 to 500,000. Given the informal nature of the global permaculture movement, it is impossible to estimate the number of practitioners implementing permaculture approaches or inspired by permaculture with no certification. This number is likely to be high especially in southern countries where development and humanitarian oriented permaculture training and workshops are organized in rural communities outside the framework of the PDC.

Regionally, permaculture initiatives and projects are sometimes promoted and connected through structured networks and local organizations. In line with its worldview favoring grassroots and small scale initiatives, permaculture is globally far less institutionalized and organized than other social and environmental movements. Regional and international gatherings, like the International Permaculture Convergence happening every second year in a different country, help creating links between practitioners and maintaining the feeling of collective belonging to the permaculture community.

Itinerant teachers, including Mollison and Holmgren, have played a major role in the rapid dissemination of permaculture, providing PDCs and other training all around the world and writing books. Based on their notable differences in personality and approach (they stopped working together shortly after the publication of *Permaculture One*), they have promulgated their visions of permaculture with different emphases and in different directions, in turn encouraging subsequent generations of permaculturists to develop new and distinct areas of focus. For example, Geoff Lawton (http://www.geofflawtononline.com/) is well known for having developed design practices related to water conservation and farming in hyper-arid areas, based on his experience of "Greening the desert" in Jordan. Morrow (2010) has achieved international recognition for developing and teaching permaculture approaches adapted to poor and postwar countries in Asia, Africa and Eastern Europe. She strongly focuses on nonviolence methods and the design of highly nutritional gardens easy to maintain with local resources while providing a high diversity of nutrients and vitamins to prevent diseases linked to extreme poverty conditions. Building on the successful dynamic first developed in the British town Totnes, Hopkins (2008) has promoted collective approaches to design and manage solidary human settlements adapted to a post-petrol society. Such processes are supported by a set of facilitating tools and principles to release the "genius of the community" and find creative alternatives to petrol which often result in rethinking globally various dimensions of human communities such as education, health, food, habitat, transportation, economic exchanges (local money) etc. (Aiken, 2017). Hopkins' approach inspired from permaculture has given birth in 2006 to the rapidly growing Transition Network (https://transitionnetwork.org). In 2017, this network regrouped more than 500 initiatives committed to the postpetrol transition and willing to exchange on their experiences at different scales: neighborhoods, villages, towns, cities and even regions, in more than 50 countries (mainly industrialized countries).

Specificity and Originality of Permaculture

Agricultural Implementation of Permaculture

In the field of agricultural production, the practical implementation of permaculture shares many similarities with other alternative farming approaches such as organic farming, biodynamic farming, agroforestry or agroecology. All these movements have historically promoted the development of resource-efficient and pesticide-free agroecosystems favoring local nutrient cycling (e.g., using compost, green or animal manure) and favoring biological regulation by maintaining a high level of biodiversity to keep plants and animals healthy. Permaculture echoes agroecology and agroforestry for the central place given to spatial association of species (combination of trees, animals, crops; intercropping; diversified landscapes). As organic and biodynamic farming, permaculture attaches a great attention to soil fertility. Permaculture has much in common with traditional organic farming, agroecology, and biodynamic farming, in the sense that all these approaches promote a harmonious and respectful integration of human beings in nature. However, historically biodynamic farming stems from spiritual preoccupations (theosophy), organic farming and agroecology are more connected to peasant's movements collectively and politically fighting for their sovereignty, whereas permaculture was born to support individual and community-scale self-sufficiency initiatives in preparation for a postpetrol world.

Compared to other alternative farming approaches, one major specificity of permaculture is the central emphasis on the conscious global design of agroecosystems rather than focusing on specific techniques. In the design process, the different functions expected from the agroecosystem (e.g., "providing food for chickens," "keeping water available in summer," "mitigating the

dominant winds," "fertilizing the garden") are listed. Different elements are integrated in the design (e.g., "a vegetables garden," "a pond," "a hedge," "poultry") ensuring that every function is fulfilled by various elements and every element fulfills various functions, thereby mimicking the functional redundancy of natural ecosystems and fostering system resilience. The different elements are combined and spatially organized using a set of design tools (e.g., checklist of principles, mapping of site specificities, chart of interactions) in order to maximize the positive interactions between elements, benefit from the specific opportunities, and to mitigate the constraints of the site Influenced by the work of H.T Odum, plant and animal species are regarded as distinctive but interchangeable system components which should be selected from a global pool based on functional criteria without regard to their place of origin. The conscious design of permaculture landscapes aims to mimic natural ecosystems and maximize positive interactions within the agroecosystem (e.g., biological regulations, creation of favorable microclimates). This global approach echoes "ecological engineering" by its systemic dimension and the importance given to design to create sustainable ecosystems (Mitsch and Jørgensen, 2003). However, ecological engineering, mainly implemented for the restoration of natural areas, is based on "self-design" (or self-organization) which tends to let ecosystems organize themselves as naturally as possible. Although permaculture design is flexible and values creative response to change during the management phase, the evolution of the system should be planned as much as is to maximize the chances that that human goals for the productive ecosystem (food, fiber) will be met.

Permaculturists tend to implement complex multistrata polycultures, intercropping, agroforestry (e.g., food forests), crop-animal integration (e.g., silvopastoralism), and to promote a high diversity of habitats, integrating landscape features such as ponds and hedges. Soil tillage is often limited and soil is constantly covered by plants or organic mulch to favor the development of soil organisms that will work for humans and structure the soil (e.g., earthworms), store carbon and limit erosion. Trees and perennial plants often play a key role as they are considered energy accumulators (storing carbon and making nutrients available for other species).

Perennial plants and trees are prioritized with the aim of reducing human labor (i.e., annual planting), together with an ergonomic zoning of the site where production areas are spatially organized according to the degree of human intervention they require. Permaculture landscape planning organizes space and elemetns into five areas with different levels of intensification: from "zone 1" where human intervention is the highest and most frequent (e.g., vegetable garden) to "zone 5" which is a natural area left deliberately unmanaged (Mollison, 1988). Zone 5 is seen both as a reservoir of biodiversity and a place where practitioners can observe "how local nature works," which can provide inspiration and design ideas for the rest of the site. In this way permaculture design integrates the spatial logics of "land sparing" (separating intensive production zones and natural areas) and "land sharing" (managing areas with reduced intensification to preserve biodiversity in productive areas) (Fischer et al., 2014). However, permaculture goes beyond this distinction as even the most productive areas are designed to maximize biodiversity as a way to maintain resilient and productive ecosystems, echoing the logic of "ecological intensification" (Bommarco et al., 2013).

As the edges between zones are regarded as being spaces of maximal diversity and interaction between species, permaculturists often maximize edge by designing cultivated areas with curved and undulated shapes rather than straight lines. The use of earthworks, dams and swales for water harvesting and control is central, as is the development of renewable sources of energy on the production site (e.g., solar panels, passive solar buildings, wind turbines, biomass and hydroelectric devices).

Rethinking Modernity and Empowering People Beyond Optimizing Ecosystems

In providing conceptual tools and design methods to observe and mimic patterns from complex natural ecosystems with the goal of designing resource efficient human settlement, permaculture could be seen as a modern, biologically-inspired approach to system optimization. In this way permaculture could be read as replicating the rationalist, instrumental relationship with nature that characterizes modernity and the industrialized world. In industrialized countries, modernity has become politically dominant in the 19th century and has spread all over the world with western culture and globalization. A growing number of philosophers and scientists argue that modern thinking, which consider nature just as an objective pool of resources which have to be rationally exploited, may be one of the major causes of the environmental issues that humanity is facing now. Permaculture departs from this tradition in several ways. Inspired by what Holmgren calls "traditional cultures of place"—premodern cultures where human beings have developed through time ecological knowledge and sensibility adapted to their specific environment—permaculture encourages practitioners to develop emotional and subjective links with the earth that will foster a feeling of responsibility toward the places where they live (2002).

Permaculturists also acknowledge that diversified ecosystems are complex, will unlikely ever be fully understood rationally, and that global environmental crises call for rapid human action despite limited ecological knowledge available. This is why permaculture encourages practitioners to develop skills and senses such as imagination and creativity in addition to rational, instrumental skills of observation, analysis, and system optimization. In this sense, permaculture has sometimes been described as a postmodern approach where elements from different systems and traditions are hybridized without regard for any fixed aesthetic or tradition, and where the importance of rational knowledge is consciously balanced and integrated with more subjective and relational human capacities.

Postmodern or not, it is clear that permaculture questions the modern and industrial world as we know it. It invites practitioners to become creative "new indigenes" while developing knowledge, interaction capacity with their local environment and community, and useful skills to become more self-sufficient in order to move from their status of dependent and demanding consumers to interdependent and responsible producers (Holmgren, 2002). This empowerment of individuals and communities is aimed both as

a way to decrease the dominance of industrial systems today, and to prepare and for survival in a future postindustrial era with no access to fossil resources.

Criticism, Controversies and Research Perspectives

A Tendency Toward Oversimplification and Overreaching

Permaculture has a troubled relationship with ecological science. Permaculture has received criticism for overreaching and oversimplifying claims. This tendency is encapsulated in the notion that humanity already possesses all the knowledge necessary to replace current land use with permaculture systems, across all social and ecological contexts, and that the process of redesigning is itself straightforward. In the absence of reliable data to support these proposals, permaculturists often rely on limited case studies and sweeping extrapolation from ecological principles. Most permaculture texts do not refer to contemporary scientific research. Much documentation available is found in gray literature which is difficult to access or verify. The effects of this isolation include the lack of reference to contemporary developments in relevant science, the accompanying persistence of idiosyncratic or misleading terminology, and the potential for influence of pseudo-scientific theories. The permaculture literature assigns the blame for this isolation on the inability of scientists and institutions to comprehend or appreciate the radical proposals put forth by permaculture. Permaculture opponents argue that permaculture practitioners may be reluctant to get involved in systematic scientific research whose results could challenge or temper their idealistic claims.

One common example of oversimplification is the conflation of net primary production with agricultural productivity. One point where this becomes apparent is in permaculture's advocacy for perennial production systems—justifying this proposal, in part, based on the high photosynthetic surface area and correspondingly high primary productivity of these systems. While forest ecosystems are among the highest in NPP, perennial plants allocate a higher percentage of photosynthetic activity to structure than annuals and therefore have a slimmer margin for export as edible tissue, rendering the comparison of potential yields a complex empirical question rather than a simple maxim. Another example is the claim that complex shapes in fields, garden beds, and ponds will increase productivity—what is called the "edge effect." This claim was originally based on the permaculture principle of edge effects that was itself extrapolated from the ecological characteristics of ecotones and anecdotal reports of edge effects in grain cropping systems. While edge effects are real, their strength, reliability, and practical applicability across widely varying contexts (i.e., from cereal fields to intensive garden beds to pond edges) is not supported by scientific evidence and is likely exaggerated. Increased biological productivity may not translate to increased harvestable yields, and the benefits of an increase in harvestable yields may be swamped by the increased labor required by complex edges.

Limited Political Impact and Scaling-up

The simple solutions populism of permaculture suggests that the best responses to global crises can be implemented immediately with readily accessible materials and skills. This worldview is reflected in a model of change that mostly spurns systematic engagement with existing institutions in favor of direct intervention into the means of subsistence, reintegrating production and resource management under the stewardship of local individuals and communities. The flat network structure that accompanies this mode of action appears to be a conscious strategy to avoid the twin dangers of cooptation and outright suppression to which grassroots efforts are vulnerable. This model has met with some success, as evidenced by its international distribution and positive influence on urban land use, horticultural and agricultural practices, and other sustainability-relevant behaviors across contexts.

The evident successes of the permaculture network are balanced by problematic assumptions and implications that evoke the hazards of insularity, exclusivity, particularity, and scale mismatch to which grassroots networks are prone. The permaculture movement displays significantly less organization and institutionalization than other international agroecological movements, e.g., La Via Campesina, Campesino à Campesino, or International Federation of Agricultural Producers. This lack makes the coordination of action beyond the immediate community scale difficult or impossible and thus limits the potential for mobilization of political support for diversified farmers. Low levels of institutionalization may also constrain capacity for program development, systematic tracking of outcomes, and engagement with potential allies. Recent research suggests that the permaculture network in the United Kingdom is vulnerable to insularity, and thereby leads to a lack of capacity to influence relevant institutions and communities.

Permaculture's optimistic focus on holistic and positive action, on personal responsibility, and on the simplicity of needed solutions, is empowering for participants and is likely a significant driver of the spread of the movement. However, the portrayal of agroecological transition as something that individuals can contribute to, using simple techniques at home, is a double-edged sword. While prioritizing the perspectives and capacities of land users is important, it may also run the risk of depoliticizing aspects of agroecological transition that are fundamentally political, and trivializing the complexity of socio-ecological processes and struggles.

Permaculture, Traditions and Neocolonialism

Permaculture has also received criticism on socio-political grounds. Critics have observed that permaculture was brought to developing countries from "knowing westerners" visiting poor communities in a similar way to humanitarian action and green

revolution packages which may be seen as a form of neocolonialism. However, scientific literature has highlighted that the principles of permaculture teaching based on individual observations and collective learning favored the empowerment of poor communities while not providing ready to use solutions designed by westerners but providing people conceptual and organizational tools to design creatively themselves their own solutions (Conrad, 2014). The little data available on permaculture in international development suggests a mixed record: sometimes implemented in a responsive and accountable fashion, and sometimes with a neocolonial savior mentality.

The movement has received criticism for a failure to acknowledge the similarity of permaculture's proposals to indigenous cultures of land use and for re-packaging indigenous land management practices as an innovation originating within permaculture. The extensive permaculture literature on small scale multistrata agroforestry uses the terms "food forest" and "edible forest garden" but rarely makes reference to the pan-tropical homegarden traditions that forms the conceptual basis for these practices and provide the vast majority of their existing land user base. Indeed, home gardens in tropical areas—from Javanese homegardens to the Creole gardens in the West Indies—traditionally involve multipurpose trees and shrubs in intimate association with annual, perennial agricultural crops and livestock, (Fernandes and Nair, 1986). Similarly, the integration of aquaculture in ponds, crops and livestock often practiced by permaculturists is drawn from traditional production systems in Asia (Prein, 2002). The founders of permaculture, Mollison and Holmgren, consider that if these sources are clearly acknowledged and respected, their use in permaculture contributes to the preservation of this rich heritage and to the recognition that westerners seeking to create sustainable human settlements have much to learn from indigenous (Mollison, 1988; Holmgren, 2002). In the same way many permaculturists seek to incorporate plant and animal species according to their functions and not to their origin, Mollison and Holmgren consider that elements of traditional knowledge from the global "indigenous pool" can be detached from their original paradigm and combined to other elements and sources of information such as scientific knowledge "to create new local cultures with hybrid vigor" (Holmgren, 2002). Critical social studies have argued that this process could be considered cultural appropriation of traditional knowledge by "university-educated white males from a wealthy country" (Conrad, 2014). Nevertheless, in many developing countries, poor rural communities have adopted permaculture as a way to reassert the value and authority of indigenous knowledge and reclaim the rights to farm "as their ancestors did" (Conrad; Millner). Some "local traditions" have been reimagined and hybridized with useful practices, principles, and scientific concepts coming from other parts of the world. In this regard, some studies have considered that permaculture has been appropriated by poor communities to create new cultural identities adapted to the modern world based on traditional ecological knowledge (Millner, 2016). Conflict on the topic of the use of indigenous knowledge in permaculture, and more generally in ecological engineering, continue and can be a fascinating research field for anthropological and sustainability studies (Veteto and Lockyer, 2008).

A Need for Research About the Agricultural Efficiency of Permaculture

Despite permaculture's origins within academia, Mollison's and Holmgren's work received very little academic attention when published in the late 1970s and 1980s. Academic reactions were mainly negative because the disciplinary specialization at that time left academics ill-prepared for the holistic approach that permaculture offered (Veteto and Lockyer, 2008). Permaculture embraces many themes and has been given many often very vague definitions, which may have caused confusion and limited systematic discussion. Its idealistic aspects have been perceived as impractical by many scholars (Ferguson and Lovell, 2014). Most private companies do not have financial interest to research and disseminate it. Since the 1980s, permaculture books and articles have mainly been written by practitioners outside academia benefiting from the high interest and enthusiasm that permaculture received from civil society. Over the decades following permaculture's emergence, sporadic academic papers have dealt with permaculture in different fields such social and behavioral sciences, architecture, education. These papers were mainly descriptive of permaculture principles and applications, with little critical analysis—though this has changed in recent years.

Very little scholarly work was carried out on permaculture from an ecological or a life science perspective based on quantitative data, especially in the agricultural field which was permaculture first priority and historic starting point. Permaculture claims to provide tools and methods to design resilient, productive resource and labor efficient farming systems based on a high level of biodiversity and beneficial ecological interactions. These assumptions remain little documented and controversial. In this regard, the most significant studies have been led for doctoral dissertations. In industrialized countries (the United States and France), they have shown that the productivity and economic returns to labor of commercial permaculture farms could benefit from high level of cultivated diversity, crop/animal integration and be economically successful even with low levels of fuel consuming motorization (Morel et al., 2016; Ferguson and Lovell, 2017). Building on the strong public interest for permaculture, some permaculture farms develop cultural or training activities to diversify their incomes in a logic of pluriactivity. This strategy raises strong criticism from permaculture opponents who argue that permaculture profitability only comes from teaching permaculture and not from applying concretely permaculture to build productive systems. The cost of permaculture training or cultural activities (workshops, demonstration site visits) is another topic of controversy. Critics from within and without the permaculture movement argue that the ways in which these costs limit access to programming contradicts permaculture's principles of equity and sharing. In response, others claim that a fair cost of training has to pay teachers for their time, labor, the depth of their experience, and the value of what they offer. Some permaculture teachers do offer free or limited-cost courses for people who cannot otherwise afford training.

Nevertheless, many permaculture farms only focus on production and are not involved in teaching. The levels of production, inputs, labor, and incomes of farms inspired by permaculture are highly variable and similar in their range to other diversified, organic, low-input., and agroecological farms (Ferguson and Lovell, 2017; Morel, 2016). In developing countries, farmers using

permaculture can experience agricultural, environmental, economic, and nutritional benefits in comparison to farmers solely using conventional agriculture, as demonstrated in Malawi by Conrad (2014). However, benefits of permaculture at the farm level are limited by the broader dominant agro-food system, constraints on access to resources and markets, and wider structural, political, and technical context. Such exploratory works mitigate both idealistic views of permaculture activists presenting permaculture as a way to solve all problems and strong critics presenting permaculture as an unrealizable utopia.

For permaculturists, a high level of biodiversity and functional redundancy are supposed to guarantee that agricultural systems will be resilient. The idea that "diversity begets stability" is deeply anchored in the ecological literature since the 1950s, especially in H.T Odum's work which has inspired permaculture. The "stability-diversity controversy" running in the ecological academic field since the 1970s has however underlined that the link between diversity of species/functions and stability of ecosystems is complex, and that other factors and properties of ecosystems have to be considered. As many gaps in academic literature remain, further studies are required to examine the efficiency, resilience, ecological dynamics, and impacts of permaculture farms in different contexts, and in the light of contemporary ecological concepts and methods, and to assess the extent to which permaculture could contribute to a large-scale transformation of food systems. Accompanying the growing public awareness of permaculture, recent years have seen a shift in the isolation of the permaculture movement from the scientific community. This bridge is being built from both sides. There is an emerging push for community-based research and partnership with institutionally-based researchers coming from the permaculture movement. For example, the Permaculture International Research Network (PIRN) was formed in 2015, sponsored by the UK Permaculture Association, and reports having over 400 members in over 40 countries. The appearance of permaculture in publications in peer-reviewed journals has increased sharply in recent year. More and more universities are developing research projects about permaculture which may announce promising perspectives for the future.

References

Aiken GT (2017) Permaculture and the social design of nature. *Geografiska Annaler: Series B, Human Geography* 99: 1–20.

AtKisson A (1991) Permaculture: Design for living an interview with Bill Mollison. *In Context* **28**. (spring 1991), 50.

Bommarco R, Kleijn D, and Potts SG (2013) Ecological intensification: Harnessing ecosystem services for food security. *Trends in Ecology & Evolution* 28: 230–238.

Conrad A (2014) *We are farmers: Agriculture, food security, and adaptive capacity among the permaculture and conventional farmers in central Malawi*. Doctoral dissertation, American University. http://pri-kenya.org/wp-content/uploads/2015/04/Conrad-FINAL-Dissertation-We-are-farmers-Copy.pdf.

Ferguson RS and Lovell ST (2014) Permaculture for agroecology: Design, movement, practice, and worldview. A review. *Agronomy for Sustainable Development* 34: 251–274.

Ferguson RS and Lovell ST (2015) Grassroots engagement with transition to sustainability: Diversity and modes of participation in the international permaculture movement. *Ecology and Society* 20.

Ferguson RS and Lovell ST (2017) Livelihoods and production diversity on U.S. permaculture farms. *Agroecology and Sustainable Food Systems* 41: 588–613.

Fernandes ECM and Nair PKR (1986) An evaluation of structure and function of tropical Homegardens. *Agricultural Systems* 21: 279–310.

Fischer J, Abson DJ, Butsic V, Chappell MJ, Ekroos J, Hanspach J, Kuemmerle T, Smith HG, and von Wehrden H (2014) Land sparing versus land sharing: Moving forward. *Conservation Letters* 7: 149–157.

Holmgren D (2002) *Permaculture: Principles and pathways beyond sustainability*. Hepburn, Vic: Holmgren Design Services.

Hopkins R (2008) *The transition handbook: From oil dependency to local resilience*. White River Junction, Vermont: Chelsea Green Publishing.

Millner N (2016) Food sovereignty, permaculture and the post-colonial politics of knowledge in El Salvador. In: *Alternative food networks in the postcolonial world*. London: Under contract with Routledge.

Mitsch WJ and Jørgensen SE (2003) Ecological engineering: A field whose time has come. *Ecological Engineering* 20: 363–377.

Mollison, B., 1988. Permaculture: A designers' manual., Édition: 2nd edn. Tagari Publications, Tyalgum.

Mollison B and Holmgren D (1978) *Permaculture one: A perennial agriculture for human settlements*. Tyalgum: Tagari.

Morel K (2016) Viabilité des microfermes maraîchères biologiques. In: *Une étude inductive combinant méthodes qualitatives et modélisation. Doctoral dissertation*. University Paris Saclay. http://prodinra.inra.fr/record/387244.

Morel K, Guégan C, and Léger FG (2016) Can an organic market garden based on holistic thinking be viable without motorization? The case of a permaculture farm. *Acta Horticulturae* (1137): 343–346.

Morrow R (2010) *Earth User's guide to permaculture*, 2nd edn. East Mean: Permanent Publications.

Naes A (1973) The shallow and the deep, long ranged ecology movement. *Inquiry* 16(1): 95–100.

Odum HT (1971) *Environment, Power, and Society*, 1st edn. New York, NY: John Wiley & Sons Inc.

Odum HT (1995) *Environmental accounting: Emergy and environmental decision making*, 1st edn. New York: Wiley.

Prein M (2002) Integration of aquaculture into crop-animal systems in Asia. *Agricultural Systems* 71: 127–146.

Veteto JR and Lockyer J (2008) Environmental anthropology engaging permaculture: Moving theory and practice toward sustainability. *Culture and Agriculture* 30: 47–58.

Phytoremediation

SC McCutcheon, US Environmental Protection Agency, Athens, GA, USA
SE Jørgensen, Copenhagen University, Copenhagen, Denmark

Phytoremediation and Other Biotechnologies

`Phytoremediation' is the cleanup or control of wastes, especially hazardous wastes, using green plants. There are many types of phytoremediation, as shown in Table 1, including the use of phreatophytes to control plumes of groundwater contaminants and contaminated vadose zones. Photoautotrophs, including vascular plants, green algae, cyanobacteria, and fungi, must be involved in the synthesis or maintenance of biomass, or in the direct metabolism, storage, detoxification, or control of contaminants. Glycosylation, occurring in plants and saprophytic fungi but not bacteria, is usually important in direct metabolism, detoxification, and accumulation or storage of pollutants by plants. Glycosylation is a sequestration of contaminant molecules by the addition of a glycosyl group to form a glycoprotein that plant cells can easily transport and store or transform. Not all applications of phytoremediation involve glycoproteins but the occurrence of glycosylation in pollutant transformations does distinguish whether the metabolism of organic contaminants or transformation of other contaminants is bioremediation or phytoremediation.

If heterotrophs are solely responsible for the metabolism or mineralization of organic contaminants and the accumulation of metals and other elements using local accumulations of nonliving organic matter and oxidized inorganic compounds, these processes are part of the allied field of `bioremediation'. However, when photoautotrophs are involved in treating contaminants by

- actively releasing organic matter during growth, maintenance, and senescence that increase the number and biomass of heterotrophs;
- selectively favoring specialized microorganisms that degrade or accumulate contaminants by pumping oxygen into the root zone, releasing exudates, or depositing secondary metabolites during root die-back in the rhizosphere to favor aerobic, facultative, or anaerobic organisms with enzymatic activity for the secondary products released or deposited; and
- transporting pollutants into active microbial zones by evapotranspiration, blockage of flows, or other means.

These processes are a vital part of phytoremediation. Depending on the various interactions of photoautrophs and heterotrophic microbial communities and the contaminant transformations involved, these processes are known as `phytostimulation', `rhizo (sphere) degradation', `rhizosphere bioremediation', or `plant-assisted bioremediation' (see Table 1).

Distinction of bioremediation from phytostimulation is important in at least three cases. First, some heterotrophs sustainably derive carbon and energy from the degradation of organic contaminants. Second, anthropogenically synthesized organic or oxidized inorganic chemicals added to a contaminated site could temporarily free bioremediation from natural photo- or chemoautotrophic synthesis long enough by cometabolism to achieve some cleanup. Third, chemoautrophs synthesizing biomass from inorganic compounds to provide organic carbon and energy for heterotrophs conceivably could be used in sustainable bioremediation. If any amendments and cofactors are obtained and added sustainably, then these bioremediation processes are sustainable. The most common amendment is fertilizer, used primarily to bulk up plant biomass and thus increase microbial biomass and activity in the rhizosphere.

Redundant ecological engineering of both plant and microbial processes in remediation is usually the sustainable and successful approach. In practice, distinctions between phyto- and bioremediation are only important for some specific contaminants at different sites. Different management approaches and techniques are required when microbial heterotrophy versus photoautotrophy dominates. Critical rates of pollutant control, uptake, storage, and metabolism, whether microbial or botanical, define whether plant or microorganism management techniques must be applied. When critical rates for microbial and botanical uptake and transformation are comparable, both techniques should be applied simultaneously for engineering redundancy and ecological resilience.

One of the most significant advances in phytoremediation is that green liver metabolism is much more important in waste management than early biotechnology research revealed. Sandermann first coined the term `green liver' to convey the great similarity between plant and mammalian sequestration and metabolism. So great is the similarity that many view plant metabolism more akin to mammalian metabolism than to bacterial metabolism. In fact, many fundamental metabolic processes first evolved in early cyanobacteria and bacteria and were carried forward, sometimes without evident purpose, into higher forms of life present today, including vascular plants and mammals. But for future xenobiotic and highly complex hazardous wastes, the most sustainable applications may need to concentrate on use of the most highly evolved enzymatic systems available only in plants and animals. In part, bacteria versus higher forms of life have evolved different survival strategies. Microbes are present in great numbers, almost ubiquitous on this planet, usually passively mobile, more adaptable, and capable of evolving rapidly. Thus, a toxic insult will kill many microorganisms but the species will usually survive, maybe even the rigors of outer space. If the die-off is extensive or long term, new protections may evolve by selection of the fittest.

 doi:10.1016/B978-0-444-63768-0.00069-X

Table 1 Types of phytoremediation ranked in terms of sustainability and applicability

Type	*Definition*	*Applications*
Phytodegradation: phytoassimilation, phytotransformation, phytoreduction, phytooxidation, and phytolignification	Aquatic and terrestrial plants take up, store, and biochemically degrade or transform organic compounds to harmless by-products, products used to create new plant biomass, or by-products that are further broken down by microbes and other processes to less harmful compounds. Growth and senescence enzymes, sometimes in series, are involved in plant metabolism or detoxification. Reductive and oxidative enzymes may be serially involved in different parts of the plant	Soils, sediments, wetlands, wastewaters, surface waters, groundwater, and air contaminated with chlorinated solvents (CCl_4, trichloromethane, tetrachloromethane, HCA, PCE, TCE, DCE, and VC), methyl bromide, tetrabromoethene, tetrachloroethane, dichloroethene, atrazine, DDT, other Cl- and P-based pesticides, PCBs, phenols, anilines, nitriles, TNT, DNT, RDX, HMX, NB, picric acid, NT, nitromethane, nitroethane, and nutrients. *Field demonstration*: Iowa Army Ammunition Plant successfully restored using wetland plants (TNT and RDX). *Proof of principle*: (a) field – *Populus* spp. Carswell Air Force Base, Texas; Aberdeen Proving Grounds, Maryland; and using lysimeters at Tacoma, Washington (TCE); and (b) horseradish peroxidase pilot-tested in unit process to degrade phenols, aniline, and other aromatic contaminants in wastewater. *Proof of concept*: *Rosa* spp. cv. Paul's Scarlet (PCBs).
Phytostimulation: rhizodegradation, rhizosphere bioremediation, and plant-assisted bioremediation	Plant exudation, root necrosis, and other processes provide organic carbon and nutrients to spur soil bacteria growth by 2 or more orders of magnitude in number; stimulate enzyme induction and cometabolic degradation by mycorrhizal fungi and the rhizomicrobial consortium; provide diverse root zone habitat; and attenuate chemical movements and concentrations. Live roots transfer oxygen to aerobes, and dead roots may support anaerobes or leave aeration channels	Soils and wetlands contaminated with crude oil, BTEX, other petroleum hydrocarbons, PAHs, PCP, perchlorate, pesticides, PCBs, and other organic compounds. *Field proof of concept*: BTEX, other hydrocarbons, PAHs, PCP, and TCE. *Field tests*: crude oil in wetlands of *Spartina alterniflora* and *S. patens*. *Fungi*: (1) field-scale tests: of white rot fungus degradation of BTEX and (2) proof of concept: for DDT, dieldrin, endosulfan, pentachloronitrobenzene, and PCP.
Phytocontainment:	Trees and other phreatophytes transpire large quantities to contain shallow groundwater plumes or contaminated soil leaching by reversing horizontal aquifer hydraulic gradients, or vertical soil moisture pressure gradients	Groundwater, vadose zone, wetlands, wastewater, and leachate contaminated with water-soluble contaminants (e.g., chlorinated solvents, MTBE, explosives, other organic contaminants, salts, and some elements).
(1) Phyto- or solar pumping, phytohydraulic control, and phytohydraulic barriers (also biobarriers)	(infiltration and leaching minimized) both year-round or seasonally to fully or partially capture contaminants. Applications normally coupled with rhizo- and phytodegradation	(1) Field proof of principle: *Populus* spp. (TCE, PCE, MTBE, and CCl_4)
(2) Control of soil and landfill leaching		(2) Concept not proven
(3) `Pump and tree', phytoirrigation, or other plant treatment *ex situ*		(3) Proposed and undergoing testing: (a) pine (*Pinus* spp.) (TCE and by-products) and (b) *Salix* spp. (organic solvents, MTBE, petroleum hydrocarbons, and nutrients) (Numbers correspond to those in column 1)
Brine volume reduction	Brines pumped onto halophytes planted in wetlands that accumulate or excrete salt and the smaller volume residual brine transported and disposed of more economically	Deep groundwater or oilfield brines. Wetland halophytes pilot tested in Oklahoma oilfield. No plant residuals: halophytes fed to cattle as a source of salt after toxicity testing of plants
Rhizofiltration: phytofiltration, blastofiltration, phyto- or biosorption, biocurtain, biofilter, contaminant uptake, and epuvalization	Compounds taken up, rapidly sorbed, or precipitated by roots (rhizofiltration) and young shoots (blastofiltration) or sorbed to fungi, algae, and bacteria (biosorption mainly to cell walls involving electrostatic attraction and formation of complexes). Marine algae possess large quantities of biopolymers (polysaccharides, uronic acids, and especially sulfated polysaccharides) that bind heavy metals. 10–60% dry weight of plant may be accumulated metals	Wetlands, wastewater, landfill leachates, surface water, and pumped groundwater contaminated with metals, radionuclides, organic chemicals, nitrate, ammonium, phosphate, and pathogens. Plant roots or shoots, aquatic plants, or algae, all live or dead, are added to or contained in wetlands, tanks, flowing water channels, or columns. Disposal of residuals unresolved. US practice is to dispose of residuals in hazardous waste landfills. Conceptually, metals sorbed to cell walls may be acid-extracted. Economic recovery of metals needs to be explored. *Field*

Table 1 Continued

Type	*Definition*	*Applications*
		proof of concept: sunflower (*Helianthus annuus*) at Chernobyl, Ukraine (Cs and Sr), and field pilot, Ashtabula, Ohio, for U. *Proof of concept for phytosorption*: aquatic plants (*Salvinia* spp. and *Spriodela* spp.) (Cr and Ni from wastewater and Pb, Cu, Fe, Cd, and Hg), algae (several metals), and marine algae (*Sargassum* Au: 40% of the algal dry weight). *Proof of concept for rhizofiltration*: sunflower (*Helianthus annuus*) and Indian mustard (*Brassica juncea*) (Pb, Cr, Mn, Cd, Ni, Cu, U(VI), Zn, and Sr)
Phytovolatilization: biovolatilization andphytoevaporation	Volatile metals and organic compounds are taken up, sometimes re-speciated (metals), and transpired. Some recalcitrant organic compounds are more easily degraded in the atmosphere but most multimedia transfers require a risk assessment before testing	Soils, sludges, wetlands, and groundwater contaminated with Se, tritium, As, Hg, *m*-xylene, chlorobenzene, tetrachloromethane, trichloromethane, trichloroethane, and other chlorinated solvents. *Field proof of principle*: Se from wastewaters and soil. *Field proof of concept*: tritium from groundwater. *Current technical consensuses*: (1) TCE volatilization has not proven significant to date but site risk assessments are required to be certain. The risk of volatilization of other organic pollutants has not been explored. (2) Transgenic plants volatize Hg in the lab but redeposition from the atmosphere makes field applications less feasible.
Phytoextraction (including chelator induced): phytoaccumulation, phytoconcentration, phytotransfer, hyperaccumulation, and phytomining	Contaminants taken up with water by cation pumps, absorption, and other mechanisms and usually translocated above ground. Harvested shoots or roots put in hazardous waste landfills or could be smelted after volume reduction by incineration or composting. Hyperaccumulation is approximately 100 times normal plant accumulation of elements and is 0.01% by dry weight for Cd and other rare elements, 0.1% for most heavy metals, and 1% for Fe, Mn, and other common elements	Extraction from soil of metals, metalloids, radionuclides, perchlorate, BTEX, PCP, short-chained aliphatic and other organic compounds not tightly bound to soils (although phytodegradation of inorganic and organic molecules is more sustainable). US practice is disposal of residuals in hazardous waste landfills but Ni smelting is feasible. Composting to reduce disposal volume conceptualized. *Pilot field-testing eastern US*: unproven at six sites with Pb using *B. juncea* but proven at two sites with Zn and Cd using *Thlaspi caerulescens*. *Phytomining Ni*: two US locations and testing in Albania and South Africa. *Field proof of concept*: Ni, Zn, Sr, Cs (see following warning), and Cd from long-term application of sludges using *Brassicaceae* hyperaccumulators in UK; Mariupol and Chernobyl regions, Ukraine; and Pennine Mountains, UK (plus Ag, Al, Co, Fe, Mo, and Mn). Failed two evaluations using chelators for Pb; thus questionable for Cr, Cs, and other tightly bound elements. New lab proof of concept now required for Pb and other tightly bound elements. *Proof of concept*: 1993–95 for Cd, Ni, Zn, Cu, Se, B, and other elements. *Bench testing*: at arid western US site for Cr, Zn, Hg, Ag, and Se using *Salix* x, *Kochia scoparia*, and *Brassica napus* and perchlorate using wetland halophytes.
Phytoslurry	Enzymatically active plant material ground and slurried with wastewater, contaminated soil, or sediment	*Lab proof of concept*: wastewater, soil, or sediment contaminated with DNT and TNT
Phytophotolysis	Contaminant translocated from soil or water into leaves and broken down by photolysis	Proposed concept for soil, wastewater, wetlands, and groundwater contaminated with RDX

(*Continued*)

Table 1 Continued

Type	*Definition*	*Applications*
Phytostabilization: biogeochemical stabilization, biomineralization, phytosequestration, and lignification	(1) Revegetation to prevent erosion and sorbed pollutant transport	Soil, mine tailings, wetlands, and leachate pond sediments contaminated with metals, phenols, anilines, some pesticides, tetrachloromethane, trichloromethane, and other chlorinated solvents
	(2) Plants control pH, soil gases, and redox that cause speciation, precipitation, and sorption to form stable mineral deposits (effects of ecosystem succession unknown on long-term stability and thus sustainability)	(1) *Extensive applications*: revegetation grasses established for different metals dominated wastes in UK and US erosion prevention handbooks available for many countries
	(3) Humification, lignification, and covalent or irreversible binding of some organic compounds are expected	(2) *Bench proof of concept*: for stabilization of some pesticides, phenols, and anilines (3) *Lab proof of concept*: for Pb and Cr^{6+} (VI)[a] (the numbers correspond to those in column 2)

BTEX, benzene, toluene, ethyl benzene, and xylene; DCE, dichloroethane; DDT, dichloro-diphenyl-trichloroethane; DNT, dinitrotoluene; HCA, hexachloroethane; HMX, octahydro-1,3,5,7-tetranitro-1,3,5,7-tetraazocine; MTBE, methyl *tert*-butyl ether; NB, nitrobenzene; NT, nitrotoluene; PAHs, polycyclic aromatic hydrocarbons; PCBs, polychlorinated biphenyls; PCE, tetrachloroethene; PCP, pentachlorophenol; RDX, hexahydro-1,3,5-trinitro-1,3,5-triazine; TCE, trichloroethene; TNT, 2,4,6-trinitrotoluene; VC, vinyl chloride.

Plants are normally rooted in place and are much fewer in number. Thus, plants may have evolved greater numbers of metabolic proteins used to detoxify insults in place, than microorganisms evidently require for survival. Plants are different from animals in the lack of (1) an immediate flight response and (2) excretion of transformation products. Animals tend to excrete transformation products, whereas plants tend to accumulate some transformation products in vacuoles or between layers of molecules in cell walls. Plant transformation products are accumulated and could be released into the environment upon death and lysis of plant cells.

Ecological Engineering

Sustainable phytoremediation is a vital element of the new applied science field of ecological engineering; so are bioremediation and other biotechnologies used in sustainable waste management. For the first time in the long history of engineering, this new field will strive to develop a new discipline of engineering that values equally ecosystem and human needs. All existing engineering disciplines, including environmental engineering, have completely or partially overweighted human infrastructure in a manner that may not completely sustain life on this planet.

`Ecological engineering', first conceptualized by H. T. Odum in 1957, is environmental manipulation by humankind using small amounts of supplementary energy to control systems in which the main energy derives from natural sources or ``the design of human society with its natural environment for the benefit of both.'' This definition clarifies why photoautrophic, and, to a degree, heterotrophic processes like sustainable phytoremediation and bioremediation are important components of this new field. Thus, ecological engineering is an appropriately holistic concept to organize and understand how different biotechnologies can be best used to sustainably manage wastes. The degree of self-organization by different species, populations, and communities of organisms is the key to understanding how phytoremediation fits into the sustainable ecological engineering of waste management.

`Self-engineering' is the reorganization, substitution, and shifting of ecosystem dynamics and functions to adapt to super-imposed environmental stresses and limitations. Adapting to ambient conditions, species, populations, and communities of plants, animals, and microorganisms `self-design' or control the local environment. Examples include the development of microclimates under plant canopies, control of soil geochemistry by roots and microbial communities, control of water levels in soil by evapotranspiration, and control of metals speciation in soil to prevent enzyme poisoning.

Other important ecological principles that define why ecological engineering and phytoremediation are sustainable waste management concepts include

- Ecosystem structure and function are governed by and can be quantitatively related for design and analysis to dynamic forcing functions like availability of sunlight, water, organic matter, nutrients and other building blocks of life, and toxins and many other stresses.
- Materials, especially carbon and nutrients, are recycled in ecosystems, making these most self-sustaining.
- Sunlight and the stored heat of the planet are sustainable sources of energy used to organize cells, organisms, populations, communities, ecosystems, and the Gaia (the planetwide ecosystem) that results in the degradation in quality of some energy to unusable states, thus increasing the system entropy, and requiring a constant and consistent source of energy for sustainability.

- Ecosystem processes have characteristic temporal and spatial scales that are usually fundamentally different from the normally static, stable, steady-state designs of human infrastructure over limited geographic areas.
- Ecosystems are dominated by momentary, diurnal, seasonal, annual, decadal, geologic, and other timescales and these pulses are often highly productive in combined energy and matter, and all types of diversity.
- Ecosystems are extensively linked and coupled spatially and dynamically with extensive feedback to form an Earth-dominating web of life known as Gaia (the preeminent example of ecosystem self-engineering and self-design), whereas human infrastructure (exclusive of the environment) is not yet as extensively linked worldwide except by the Gaia.
- By definition, an `ecosystem' is an arbitrarily defined web of life with every organism linked and coupled to everything else (including the ambient environment or habitats, overlapping ecosystems, and forcing conditions or adjacent ecosystems at all boundaries) with extensive and dynamic feedback and resilience, conservation of materials, and flow and degradation of energy quality from one trophic level to the next in all food chains systemwide.
- Organisms and cells maintain internal `homeostasis' (equilibrium achieved by adjustment of physiologic, enzymatic, and synthetic processes) of biological and physiological functions by biochemical conversion and control of sunlight energy, organic matter, other sources of energy, and nutrients, which in turn are based on ambient conditions.
- Ecosystem feedback, resilience, a buffering or waste assimilative capacities are defined by the evolutionary and life histories of the cells and organisms involved.
- Enzymatic, proteomic, and other biochemical diversity enrich feedback, resilience, a buffering or waste assimilative capacities of an ecosystem.
- Ecosystems are typically defined by geographic edges (shorelines, ridgelines, and similar geographic features) which are normally `ecotones' (transition zones between communities dominated by greater diversity and numbers or mass of organisms), so that forcing conditions from adjacent ecosystems and the surrounding environment are most evident and quantifiable and thus ecosystems are usually most vulnerable at these boundaries.

A good example of natural self-design was observed at the Alabama Army Ammunition Plant near Talladega. Built in 1942 and operated during the Korean conflict, widespread trinitrotoluene (TNT) contamination of soil and water was observed from the 1960s to the 1990s cleanup. The extensive contamination caused large bald spots of soil that were effectively sterile, the drainage from which turned local streams pink with the photodegradation products of TNT. Beavers dammed at least one on the pink-tinged streams allowing parrot feather (*Myriophyllum aquaticum*; see **Fig. 1**) and other wetland plants to grow, spread, and break down the TNT and by-products. By the early 1990s, these waters and sediments in the ponds and streams were free of detectable TNT.

The sterile bald spots, reported to be up to 60 m (200 feet) in diameter in the 1960s, were reduced to about half of this diameter by the early 1990s by encroaching forests. Concentrations were of the order of 10 000 mg of TNT per kg of dry sterile soil. As of 1991, hardy grasses grew at the margins, followed by small pine (*Pinus* spp.) trees. Some of the grasses and small trees succumbed to the highly toxic soils, probably during periods of higher environmental stresses (drought, frost, rainfall saturation of soil and mobilization of highly toxic TNT, and other stressful events). These plants usually fell into the sterile areas to increase soil organic carbon at the ecotone, allowing other plants to grow further inward year after year. Concentrations were on the order of 1000 mg kg^{-1} of soil in the ecotone. Adjacent to the ecotone were more mature pines that transitioned to hardwoods in a short

Fig. 1 Parrot feather (*Myriophyllum aquaticum*) growing in a natural wetland.

distance as part of the normal succession in southern forests. In the reforested fringe of full size trees and underbrush, TNT concentrations were on the order of 10–100 mg kg^{-1}. To ecologically engineer the cleanup of this site, the natural process of breaking down TNT in the sterile soil could have been accelerated using limited energy in collecting and transporting local forest accumulations of detritus to the sterile bald spots to get indigenous plants growing. The beaver ponds downgradient could have been kept in place to continue to capture and break down the residuals draining from the bald spots until all contaminants were removed from the soil and wetlands.

At the Carswell Air Force Base in Texas, an ecological engineering design is being used to inexpensively avoid energy-intense mechanical pumping and treating a shallow, surficial groundwater plume of trichloroethylene (TCE), a chlorinated solvent used in cleaning metal parts. Two small plantations of cottonwood (*Populus deltoides*) whips and 1-year-old trees were planted 1 m (3 feet) deep across the plume, the depth of which was approximately 3–3.5 m (10–11 feet) below the ground surface. Within months the transpiration of the fast-growing cottonwoods seasonally halted the contaminant plume as evidenced by recurring depressions in the water table. At the end of 5 years, about 30% of the plume was removed by the trees and microorganisms. Scaleup of this pilot plantation of approximately 900 trees by a factor of 4 or more is likely to completely halt this small plume. But the complete dewatering of the aquifer by transpiration is neither necessary nor desirable. Some of the TCE residual can be dehalogenated by enhanced vadose zone reduction and spurring of microbial growth from the added trees. Complete aquifer dewatering will dry up an adjacent stream and cause the death of some of the trees due to lack of water.

Remarkably, the cottonwoods self-engineered the vadose zone redox conditions to at least 3 m (10 feet) deep by the release of sugars and other simple exudates. The switch to reducing conditions favored the microbial dehalogenation documented at the site. Seasonal pulses of percolated, oxygenated rainfall prevented any buildup of carcinogenic vinyl chloride and other intermediates. Contaminants taken into the trees with the transpiration stream were mineralized by dehalogenation and oxidation pathways. Elevated chloroacetic acids in tree leaves were temporary accumulations not expected to persist or cause toxicity.

Genetics and Biochemistry of Phytoremediation and Ecological Engineering

Plant genomes typically have approximately 25 000–110 000 genes. By comparison, the human genome has a little over 32 000 and typical species of bacteria have about 2500 genes. These order of magnitude differences do not necessarily translate into orders of magnitude differences in protein and enzyme synthesis because (1) some genes seem to lack a specific function, (2) several genes are necessary to express or synthesize some proteins, and (3) some genes generate or contribute to several different proteins. Thus, while anchored in place, fewer in number than microorganisms, lacking an immediate flight response to toxins and environmental insults, and while not being able to excrete as many transformation products as do mammals, plants may not be more highly evolved than humans and other mammals.

The overall number of unique genes for all organisms is difficult to estimate, but the effects of genetic diversity are well known in general and need to be applied more in place of predominant monocultures typically used in early phytoremediation applications. Nevertheless, approximately 10 000 plant proteins and 200 000 plant secondary metabolites are known to exist, which must include the full suite of biomolecules necessary to govern, control, and otherwise contribute to growth, maintenance, and senescence. Some of the growth, maintenance, and senescence enzymes serve dual functions in plant metabolism and in detoxification and defense from insults. Less than a hundred enzymes have been identified as useful or potentially applicable in phytoremediation. Thus thousands of other proteins could be important but have not yet been fully characterized as to activity and function.

Because of shorter life spans, microorganisms are expected to evolve responses to xenobiotic and anthropogenic stresses faster than plants and animals. Thus the future tracking of microbial enzymatic activity may be more complicated. But with a typical genome of 2500, many fewer proteins seem to be uncharacterized for bioremediation or plant-assisted bioremediation applications.

Some simple enzymatic activities such as nitroreductases (EC 1.6.6.-; see **Fig. 2**) may be common to all forms of life due to evolution in the earliest forms of bacteria or cyanobacteria. But animals and plants, which seem to have similar genetic diversity from overlapping ranges of genome sizes, seem to have at least an order of magnitude more proteomic diversity than bacteria. This order of magnitude greater plant enzymatic diversity should be investigated not only for waste management applications but also for the development of new fibers, pharmaceuticals, nutraceuticals, and other products.

The fact that some plant genomes exceed the size of human and some other mammal genomes does not necessarily indicate that plant enzymes and proteins are more evolved. Some plant genomes may have greater genetic redundancy and greater numbers of `introns' (polynucleotide sequence in a nucleic acid or gene sequence that does not code information for protein synthesis). But some plant enzymatic activities seem more evolved as evident from what seem to be the great number of medicines derived from plants. By contrast, some mammalian genes seem more evolved than plant genes. At least one human gene 2E1 expresses a more active cytochrome P450 (EC 1.6.4.2) for transformation of chlorinated solvents than that expressed by plant genes. The number of plant metabolites used in pharmaceuticals cannot be compared to the one investigation of the human gene activity in transgenic plants, because the engineering of transgenic plants with human genes is presently quite rare.

As a result, plants seem to have an order of magnitude more enzymes than bacteria. Plants have more elaborate metabolic synthesis and photosynthesis, whereas bacteria are normally simple heterotrophs that specialize in metabolism of organic

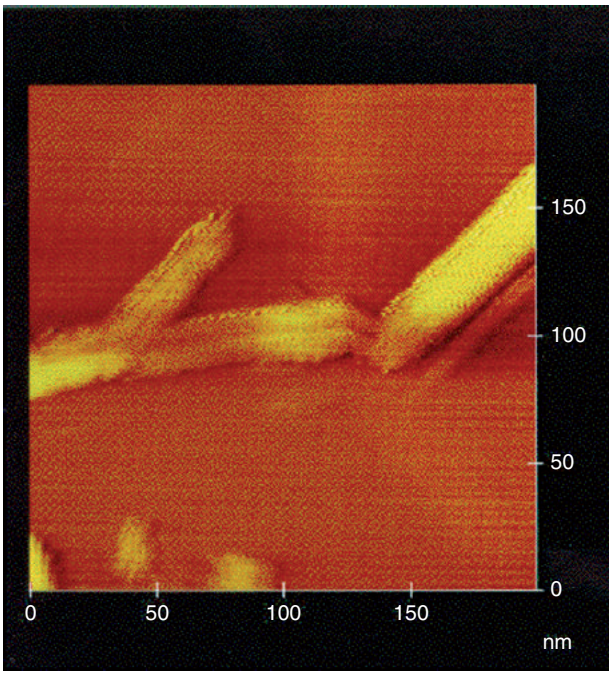


Fig. 2 Tunneling electron miscopy of a monoclonal antibody attached to a putative nitroreductase-humic acid complex shown at 200 nm by 200 nm. Courtesy of George Bailey, US EPA (deceased).

compounds for energy to grow and reproduce. Some animal enzymatic activity seems more evolved than plant metabolism and vice versa. Evolutionary history, practical control of organisms to minimize spread of genetically engineered organisms, and genetic and proteomic diversity must be used to determine if plants or animals are better suited for the genetic engineering that may be necessary to sustainably manage some hazardous wastes in the long term.

From a more holistic and appropriate perspective, microbes, plants, and animals symbiotically cycle nutrients and organic material and manage available energy in ecosystems planetwide. Thus contrasts of genome sizes, genetic diversity, and enzymatic activities are not as important as understanding when unique processes can be lightly managed in the ecological engineering of ecosystems to cleanup and manage wastes. Thus the activities of proteins common to microbes, plants, and animals need to be explored to determine which enzymes are more evolved to use in management of hazardous wastes and for other purposes. The predominant uncharacterized enzymatic activities of plants and animals should be investigated in a systematic manner for waste management and other societal purposes. To understand how contaminants can be sustainably managed, the seminal differences in metabolism must be known to select and lightly manage appropriate ecosystems.

Recently, community effects of plants and insects have been shown to further the transformation of some compounds. (The effects of nutrient cycling by microbes in these simple plant–insect ecosystems have not been fully explored.) This is another indication that plant metabolism seems to be more similar to animal metabolism. This genetic similarity of enzymatic activity also bodes well for addressing general concerns about food chain accumulation of organic contaminants and by-products in animals during phytoremediation applications.

Proteomics in Phytoremediation Research

At least two approaches have been used to characterize and apply knowledge of plant enzymes in the sustainable phytodegradation of organic pollutants. These include the following types of investigations.

1. Natural reactions, usually very fast, dominate transformations, are traced to plant and microbial enzymes.
2. Roles for well-characterized enzymes are developed.

Microarrays of common contaminants, secondary metabolites, or typical enzymatic activities (which could be used simultaneously to develop pharmaceuticals and nutraceuticals) do not seem to have been applied to explore untapped plant enzymatic activities.

All approaches to applying enzymatic activity to phytoremediation involve some element of trial and error, and are thus similar to the development of new medicines. The trial- and-error testing of plant and insect samples from diverse but unusual settings like jungles, deserts, or evolutionarily isolated environments to discover high metabolic and other activities is the same as reacting

representative compounds with environmental samples to discover natural transformations of contaminants in different media. In most cases, classical techniques in biochemistry are used or have been used to identify and purify the enzymes and other proteins responsible for the activity. A pathway analysis and characterization of the contaminant fate is equally important to efficiently guide the characterization of the enzymes reacting with and controlling pollutant fate and transport. Early insight into potentially more toxic metabolites must be followed with interim toxicity tests, including phytotoxicity tests to select the best plants to apply.

Natural reactions have been explored in two complementary ways. One focused initially on enzyme isolation and characterization, the other on inferring enzymatic activity from contaminant transformation pathway analysis. Both were necessary to comprehensively define and apply the reactions involved.

The most useful approach to developing new phytodegradation applications to date is to start with observations of a fast natural reaction of a compound of interest and isolate and characterize the active components. In three of three isolations from reducing sediments, stable plant enzymes in reducing sediments have proven to be responsible for the fast reactions discovered, not microbial enzymes as expected. Starting with fast natural reactions or unusual accumulations of metals and inorganic compounds usually ensures that an inexpensive *in situ* process is possible, and some suboptimal, black box applications are always possible before full characterization of the natural processes has occurred.

Starting with a known enzyme like peroxidases (EC 1.11.17) or phosphatases (EC 3.1.3) and finding roles in phytoremediation is much less successful so far. Potential applications are generally highly controlled *ex situ* unit processes that provide vital cofactors and control of toxic intermediates, if applications can be developed at all. High degrees of control are usually unsustainably expensive and energy intensive, but normally more reliable in meeting cleanup standards.

The proteomics of phytodegradation seems to be more broadly understood than that for metals accumulation by plants. There are at least three bodies of knowledge that contribute to the breadth of information on plant proteomics used in phytodegradation practice today. First, the extensive investigation of plant proteomics for the development of pesticides and herbicides used in agriculture provides invaluable insights into not only plant tolerance, but more importantly metabolism, detoxification, and transformation of different classes of organic compounds. Much of this experience is readily extrapolated to waste management using plants because pesticides and herbicides are important contaminants at some sites, especially worldwide. More importantly, pesticide development has been carefully organized in terms of plant effects and the structure and activity of the synthesized pesticides, which is easily extrapolated to similar contaminants. The most significant drawback is that much of the research is proprietary and has never been published, especially for candidate compounds not developed into pesticides.

Second, quite a few enzymes are common to microbes and plants and the field of bioremediation has pioneered exploration of the proteomics and more recently the genetics of some important enzymatic processes. The pioneering pathway analyses for bioremediation are a vital tool used to accelerate the acceptance and use of phytoremediation.

Third, the vital insight that plants are green livers means that the more extensive medical and veterinary research on the proteomics, genetics, and genomics of cell biology could be translated in the development of phytoremediation using the seminal differences between plant and animal cells as a basis for extrapolation. Unfortunately, medical research seems to have less strict process endpoint requirements in some cases and the rates of transformations are not always defined well enough for immediate extrapolations to phytoremediation applications.

The impact of other research on plant biochemistry does not seem to have been as important, but seminal research on allelochemicals and natural pesticides and synthesis and transformation of plant metabolites should prove vital in the future. Knowledge of the approximate 200 000 plant metabolites should be vital in developing new applications and defining the limits of phytoremediation. When a plant or any organism produces an allelopathic, poisonous, or other toxic secondary metabolite, there must be a transformation process that prevents these toxic compounds from eventually building up and swamping the planet. An example is plant lignification and delignification that uses peroxidases (EC 1.11.17) both to synthesize and degrade woody tissue. The seminal conclusion is that the existence of a primary or secondary metabolite similar to a contaminant molecule provides a good chance that phytotransformation or ecological processes can be used to degrade and control the waste sustainably, given enough time and land area. The approximate 200 000 plant secondary metabolites alone provide up to 200 000 potential metabolic pathways for degradation. However, a number of these pathways will share common transformations, the number of which may not have been estimated yet.

The enzymatic processes involved in metals accumulation by plants may not be as broadly explored as for organic metabolism and detoxification, but those investigations that have been undertaken for cadmium, lead, mercury, nickel, zinc, and a few other elements do delve deeply into the genetics and proteomics of these processes. The investigations of the proteomics of metal accumulation are similar to those used to define the role of enzymes in organic transformations and mineralization. Hyperaccumulating plants are the basis of proteomic and genetic characterizations supporting some applications, while specific genes and enzymes have been investigated for other applications of metals `phytoaccumulation'. For `phytosorption', much less is known about plant and algae cell wall characteristics and the effects on electrostatic attraction and formation of complexes or the role of polysaccharides, uronic acids, and sulfated polysaccharides that can bind heavy metals. Thus the selection of plants and algae species and other design decisions are made by trial-and-error testing at the moment.

`Hyperaccumulation' is arbitrarily defined as occurring when a plant contains more of a metal than 0.01%, 0.1%, or 1% by weight, depending on how frequently the metal occurs in the environment (**Table 1**). These hyperaccumulating plants are relatively rare and usually consist of slow-growing herbs. Some of these hyperaccumulators seemed to have evolved on metal outcroppings that normally poison plants; some may have evolved to accumulate metals as a natural insecticide or perhaps as an elemental

allelopathic defense. Some hyperaccumulators may benefit from `mutualism'. Mutualism may include (1) mycorrhizal or plant ecosystems that make metals in soils more available and less toxic with chelating ligands or enzymes and (2) accumulation of metals in seed and pollen to select only certain-metal-tolerant animals for the transport of this important genetic material. Commensalism in the form of epiphytism could explain some hyperaccumulation, but some hyperaccumulators cannot be explained functionally or evolutionarily.

Genetic Engineering

Because humankind has been highly inventive of xenobiotic molecules (e.g. DuPont's ``better living through chemistry'') and has unsustainably used concentrated energy to smelt metals and create other toxins, some genetic engineering of plants and microbes will be necessary to achieve more sustainable waste management. Some xenobiotic substances such as polychlorinated biphenyls (PCBs), dichloro-diphenyl-trichloroethane (DDT), and several others have been partially or completely banned because of increasing accumulations in the environment. The lack of sufficient analogs among the 200 000 primary and secondary plant metabolites are the reason for these accumulations, but some microbiological and botanical transformations of these persistent pollutants do occur. Unfortunately, unmanaged ecosystems may not have all the transformation processes available in the same locations at the right times to achieve sustainable and safe degradation of persistent pollutants over the times scales consistent with most human endeavors. In addition to the need to develop better-managed ecosystems in order to treat and control wastes, some metabolic and genetic engineering of organisms will be necessary to clean up persistent organic pollutants and to properly manage xenobiotic compounds invented in the future.

Plants are the most likely candidates for most genetic engineering to support sustainable waste management. While microbial enzymatic activity is better adapted to mineralization of organic pollutants, microorganisms lack the degree of self-engineering that plants exhibit. Most animals, especially mammals, may be too mobile and too charismatic to genetically engineer for waste management based on the ethical values of most contemporary societies. In contrast, plants are easier to control and manage because the most likely candidates are rooted in place and propagation can be controlled with single-sex clones and harvesting before reproduction occurs. Fencing, netting, and other techniques control plant exposure to other organisms. Compared to animal husbandry, humankind has more experience in agricultural control and production of plants for food and fiber. Because of the role of sunlight as the primary energy source for synthesis, plant management is generally less intensive and less expensive than animal husbandry and unit process production of microbial enzymes.

In general, the control of genetically modified organisms seems to favor large plants, rooted in place and capable of controlling groundwater, soil redox, and many natural insults. Microorganisms are not easily detected in real time, are difficult to control in the environment, and lack the genetic diversity of plants and animals. Animals are much more mobile and these ranges of movement are not always known, especially in response to environmental insults. As precedents for genetically engineering plants for waste management, the use of genetically modified soybeans to tolerate glyphosate-type herbicides, genetically modified corn for animal feed, and modified tomatoes are standard practices in the US and some countries, but not in Europe and elsewhere.

Already important research on the development of transgenic plants (genetically engineered deoxyribonucleic acid (DNA) with genes from microorganisms or animals) provides insight into what is possible and what other research is needed. The initial work includes testing and developing

- transgenic *Arabidopsis thaliana*, yellow poplar (*Liriodendron tulipifera*), tobacco (*Nicotiana* spp.), Indian mustard (*Brassica juncea*), and eastern cottonwood (*P. deltoides*) to tolerate and `phytovolatilize' (see **Table 1**) mercury and selenium (but any application to mercury contamination can only be done with a site-specific assessment to establish that the atmospheric release of mercury does not present an increased risk; **Table 1**);
- transgenic plants to tolerate, accumulate, or speciate arsenic, cadmium, cooper, zinc, nickel, lead, and iron;
- transgenic tobacco (*Nicotiana* spp.) expressing cytochrome P450 from the human gene 2E1 that breaks down TCE 640 times faster than control plants;
- transgenic tobacco (*Nicotiana* spp.) expressing the *Enterobacter cloacae* nitroreductase *nsfI* to tolerate and more rapidly transform explosives like TNT;
- transgenic poplar (*Populus* spp.), Indian mustard (*B. juncea*), tobacco (*Nicotiana* spp.), rice (*Oryza sativa*), and potato (*Solanum tuberosum*) enhanced tolerance and metabolism of herbicides;
- transgenic flowering plants to remove atmospheric nitrous oxides in major cities around the world;

These case studies establish what is possible. But so far these pioneering studies do not establish that genetic engineering is useful and applicable in waste management. Merely speeding up the degradation of explosives or TCE is not useful until proof is available that genetic engineering is less expensive and more sustainable that planting up to 640 more wild-type plants or taking 640 times longer to remove TCE from a site. By contrast, the lack of natural plants that tolerate and phytovolatilize mercury and arsenic are justification for genetic research on mercury and arsenic management. Because only microorganisms transform nitrous oxides, there is also a high likelihood of applicability of transgenic plants for atmospheric cleanup in large cities.

The studies so far do not establish that transgenic plants efficiently maintain transgenes from one generation to another. This further reduces the risks of transgenes being dispersed in the environment and favors longer-living trees as the best organism to

engineer. Applications are much more expensive, and perhaps unsustainable if each generation of plants must be re-engineered genetically.

Genetic engineering is not only necessary to sustainably manage persistent organic pollutants and some new xenobiotic chemicals, but perhaps for sustainable phytoaccumulation of heavy metals, metalloids, and other elements as well. In addition, transgenic plants can be engineered to take advantage of a wider range of specialized genes in bacteria to mineralize versus transforming contaminants to more or less toxic by-products. Transgenic plants might also express more active enzymes from mammals. These transgenic enhancements can be put into phreatophytes, fast-growing crops, long-living trees, or other specialized plants to achieve much wider control of wastes in the environment. However, a super-transgenic plant to control and manage every witches' brew of hazardous chemicals at all contaminated sites is highly unlikely.

Future Directions and Needs

Estimates of cleanup costs for hazardous waste sites planetwide are of the order of US$ 1×10^{12}. These massive costs to clean up national defense and industrial facilities are effectively unsustainable and largely being deferred to later generations, especially at sites such as Chernobyl in the Ukraine. Waste disposal requirements of the twentieth century were usually followed, but these practices were rarely based on environmental and other resource sustainability. In the future, any development and further use of xenobiotic compounds for better living must be accompanied by the development of sustainable biotechnologies or other approaches to sustainably manage any wastes. Classical plant breeding and agroeconomic practices may be used to develop some sustainable waste management practices. But more likely the inventiveness of humankind in chemistry for better living will not be limited to xenobiotic compounds analogous to primary and secondary metabolites.

This concern is especially acute for the nascent development of nanotechnologies. These micron-scale particles are conceived to have almost unlimited applications based on the development of macromolecules that have intense concentrations of activities or energy that secondary metabolites, natural macromolecules, and microorganisms do not currently possess. Thus any development of applications of nanosize particles should be preceded by environmental fate and transport investigations and the ecological engineering of sustainable biotechnologies to manage any risks of uncontrolled environmental releases.

The ecological engineering of the cleanup of existing unsustainable releases of hazardous wastes is also very much needed to reduce the impact of the original manufacturing and disposal decisions. For example, the cleanup costs in the US may be reduced on the order of US$ 1×10^9 using phytoremediation, but additional optimization of applications is necessary and fundamental research is necessary to provide additional methods of cleanup. To achieve these and even greater cost savings, research investments on the order of 10% of the savings (US$ 1×10^8) are normally justified. Current research investments for phytoremediation in the US seem to be on the order of US$ $1 \times 10^6\ \mathrm{yr}^{-1}$, a difference that may explain why some twentieth-century waste management costs are being deferred to future generations.

The development of phytoremediation applications is particularly vulnerable to shortfalls in fundamental research. For the field of phytoremediation to continue to grow at a very rapid pace compared to bioremediation and other innovative cleanup technologies, significant investments are required in fundamental investigations of plant enzymatic activities, genetics, proteomics, and primary and secondary metabolites. The US EPA Contaminated Sites Program and the US Strategic Research and Development Program were vital in initiating the field in the early 1990s. If not for the European Cooperation in the Field of Scientific and Technical Research (COST, http://lbewww.epfl.ch/COST837/) to guide research, applications of phytoremediation may have already stalled despite recent limited programs to support applications research from the US National Science Foundation, EPA, Strategic Research and Development Program, and Office of Naval Research.

Compared to microbes and mammals, much less of plant proteomes seem to have been explored for enzymatic activities useful in waste management, pharmaceutical and nutraceuticals development, and other societal needs. Yet over the history of modern scientifically based medicine and the much longer application of traditional medicine, many plant activities have been found to be useful. For the approximately 10 000 plant proteins known, less than 1% have been investigated for waste treatment applications. In addition, only a few classes of compounds have been tested for transformation by plant and microbial enzymes.

Investigation of these enzymes for various applications is of the highest priority. Development of microarrays of enzyme activities, DNA, and chemical compounds can significantly decrease screening time. When screening for phytoremediation plants, other arrays could be useful to simultaneously screen for new pharmaceuticals, nutraceuticals, and other products.

Future applied research should continue to investigate plants and microorganisms that are associated with unique and unusual poisons and other natural toxins. Uniquely evolved plants and microorganisms near metal outcroppings, in deserts, in cold regions, and in other rigorous environments, or evolved in secluded regions like Australia, should continue to be investigated. However, better comprehensive plant proteomes cross-referenced to pollutants must be developed to guide this important research, as was done in the development of pesticides and herbicides during the green revolution.

Understanding genetic and proteomic diversity is vital to better engineer ecosystems rationally in place of plant monocultures and simple ecosystems of highly controlled plants and microorganisms. Furthermore, current simplified design procedures for wetlands, grasslands, crops, and plantations are rarely if ever based on optimized engineering protocols and infrequently make use of sustainable ecological engineering design. Thus many phytoremediation applications, while more cost effective than most other

alternatives, still may not be sustainable if excess fertilization, irrigation, and soil augmentation are used to apply monocultures and imported plant species. Ecological engineering needs to be introduced to phytoremediation, starting with a comprehensive reorganization of applications and research to date in terms of basic ecological engineering concepts. Without ecological engineering optimization, the costs of increased eutrophication and decreased water quality, water shortages, and poor soil may be unsustainable in the face of growing populations of humans planetwide.

Next in priority is that classical breeding techniques need to be explored to select optimal plants to use in phytoremediation. So far the little experience in plant breeding seems to have come from characterizing metals accumulating plants.

Finally, metabolic and genetic engineering research must be pursued further to sustainably manage xenobiotic and heavy metal pollutants. Highly evolved, large, immobile plants that dynamically self-engineer provide the best opportunity to control and apply microbial, animal, and plant genes in complex cleanups using transgenic organisms.

So far, transgenic plants are feasible for some phytoremediation applications but have not been shown to be fully necessary and useful.

So much more is necessary to define potential roles of thousands of plant enzymes, define how all of the approximately 200 000 secondary metabolites are metabolized, and optimize and engineer the necessary transgenic plants. Thus *de novo* development of new genes does not seem to need to be a current waste management priority. First, the reliability of libraries of pathogenic gene sequences is not known well enough so to be sure that the genetically engineered organism can be used in the environment. Second, the release in the environment of a *de novo* gene may not have been adequately assessed in terms of risk.

State of the Practice

Despite the paucity of ecological engineering optimization of phytoremediation applications to minimize costs and to achieve full sustainability, a number of *ad hoc* applications have been successful. These are summarized in **Table 1**. Wetlands prevent all discharge of explosives-contaminated groundwater at the Iowa Army Ammunition Plant (**Fig. 3**). Fertilized grass plantings and marshes of *Spartina alterniflora* and *S. patens* have cleaned up petroleum spills at a number of sites and are overdue for optimization

Fig. 3 One of two constructed wetlands at the Iowa Army Ammunition Plant that has zero discharge of TNT and other explosives seeping out of the contaminated groundwater below the explosives manufacturing line shown in the upper right-hand corner. Explosives-contaminated soil was dug up along a stream channel leaving a pit that was then shaped to capture all contaminated groundwater and favor indigenous aquatic plants known to transform explosives. By not having to refill the pit with clean soil, approximately US$ 700 000 was saved. Sediment from local wetlands was used to seed the constructed wetland and provide nutrients. The stream channel was moved to the left as shown to convey less-contaminated stream flow around the wetland. The initial 28 months of monitoring showed elevated explosives concentrations entering the wetland each spring, decreasing in peak concentration the second spring, and dropping to undetectable levels by fall due to plant and algae uptake and transformation. Thorough sampling and analysis established that explosives and the associated transformation products were undetectable and thus did not accumulate in the indigenous wetland plants and algae. The wetland control structure at the lower right of the photograph has never been opened, thus achieving a zero discharge. Evaporation prevents flow out of the wetlands. Slightly more was spent than the US$ 700 000 cost savings and this small difference was due to added monitoring costs of these pilot applications. Thus phytoremediation is not only less expensive, but reduced site restoration costs can also fully offset construction and monitoring costs. Photo courtesy of Kevin Howe, US Army Corps of Engineers, originally and used on the cover of *Phytoremedation* (2003) Wiley.

Fig. 4 Black elder (*Sambucus nigra*) tree growing into soil contaminated with a prussian blue cyanide complex. Courtesy of Tim Mansfeldt, Front cover, *UWSF-2umweltchem Oekotox*, 12(3): 2000.

Fig. 5 Eastern Oregon landfill (foreground) leachate collection system (see storage pond on right) and hybrid *Populus* spp. plantation irrigated with leachate (upper left and center) and harvested for paper and pulp. Originally courtesy of Jim Jordahl, CH2M Hill Consulting Engineers and Scientists, Des Moines, Iowa.

and development of design guidance. Tree plantations have successfully stopped and mineralized groundwater plumes of chlorinated solvents and methyl *tert*-butyl ether (MTBE). A number of treatability investigations using standard groundwater and vadose zone models establish that transpiration can be adequately forecast and can control plumes in difficult settings. Three-year-old trees have been planted in casings up to 15 m (50 feet) deep to reach contaminated aquifers underlying clean aquifers and soil. Highly toxic cyanide-contaminated soils shown in **Fig. 4** are being cleaned up using a tree plantation in Holte, Denmark; this is so effective that the site is being used as a park during the process. Vegetation capping and control of landfill leachate has been successful at a number of sites across the US (see **Fig. 5**) and is being applied in Australia and a number of other countries.

The feasibility of metals' phytoaccumulation, the process first called phytoremediation, has been investigated, but very few field-scale tests have been attempted. Unfortunately, a few critical field tests have been unsuccessful. **Fig. 6** describes some of the

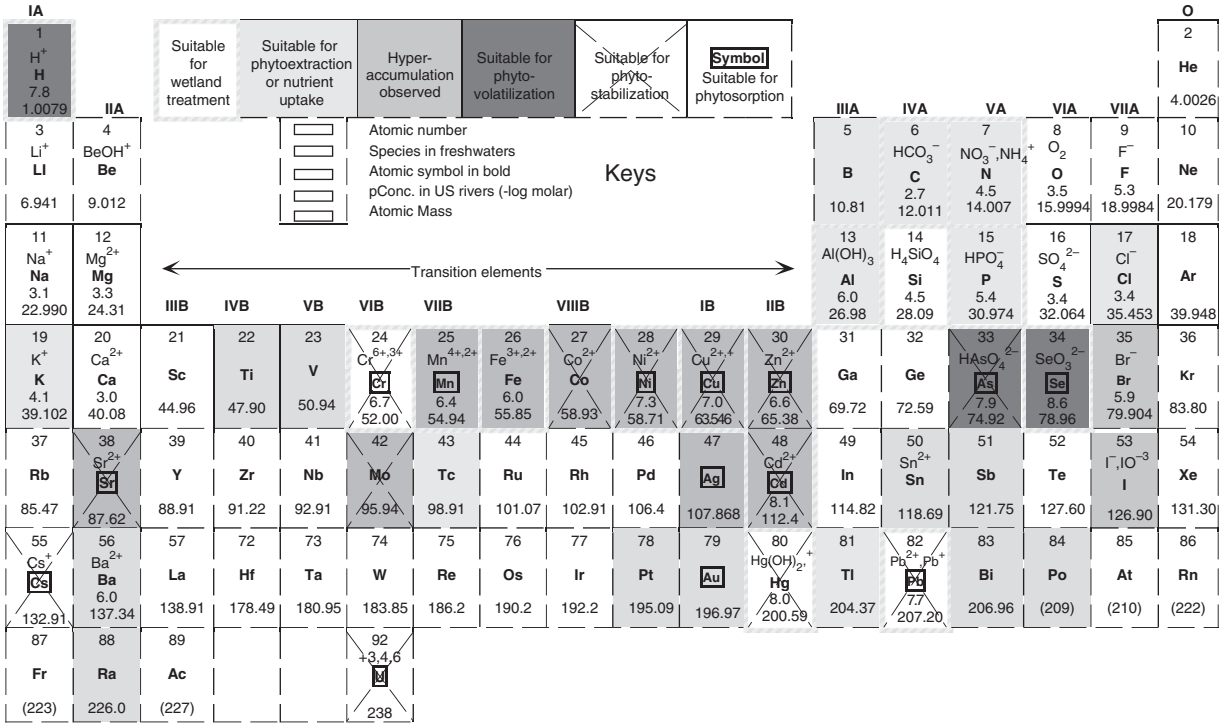


Fig. 6 Periodic table of elements for which phytoextraction, rhizofiltration, phytosorption, phytostabilization, and phytovolatilization are feasible. For many elements, cleanup has not been proven in concept or principle in laboratory, pilot, or field tests. Some elements have not been tested to determine if phytoremediation is a feasible management option. Selenium and arsenic are also subject to phytoextraction and hyperaccumulation by some plants. Tritium is phytovolatilized and hydrogen ions, of which pH is a measure, can be controlled with wetlands specifically designed to treat acid mine drainage. Some wetlands do not remove and accumulate all the metals listed under all conditions (e.g., lead and nickel).

elements for which plant uptake or control is feasible. Pilot investigations by the US Bureau of Mines establish that some plants accumulate enough nickel to be cost-effectively smelted or refined. Pilot field testing of cadmium and zinc hyperaccumulation has been successful at two sites. Unfortunately, six field-scale pilots and rigorous demonstrations for lead accumulation have failed. The failure to optimize chelator applications to release lead and other tightly bound metals like chromium have resulted in only approximately 25% of the lead released from the soil being accumulated by plants. The remainder leaches into the groundwater or deep soils.

In addition, residual management of plants with accumulated metals has not been sustainably optimized. Unless the multimedia transfer to one generation of plants is sufficient to reduce soil or water concentrations to acceptable levels of risk, phytoextraction requires the harvest and disposal of plant materials. The rate of metal uptake per unit area, the target residual accumulation at harvest, the time available for cleanup on different parts of a contaminated site, and monitoring requirements are design variables selected to achieve economically and sustainably optimized solutions. The US state of the practice at this early stage requires metal-laden plants be disposed in expensive hazardous waste landfills. Metal recycling to smelters, established as feasible for nickel, and composing or incineration to reduce the volume disposed in landfills do not seem to have been rigorously explored.

For phytodegradation and accumulation of organic compounds, food chain accumulation and exposure off site is the primary concern. Extensive pathway analysis based on advanced chemical analysis and mass balances of radiolabeled compounds; monitoring of accumulations in roots, stems, leaves, insects, and animals; and toxicity testing of plant residues establish that most phytoremediation applications do not cause exposures off site or during treatment. For polyaromatic hydrocarbons (PAHs), insects take up the contaminant and by-products during phytoremediation but have some of the same enzymatic activity that completes transformation without further food chain exposure. Insects that take up explosives and by-products depurate within hours of ending the exposure. However, not all potential exposures have been explored and thus each application requires monitoring of accumulations for 3–5 years depending on the contaminant and the setting involved.

Because phytoremediation approaches have not been optimized and design guidance written, all almost all applications require pilot or treatability studies before full scale-up is undertaken. In a number of cases, phytoremediation is inexpensive enough that pilot studies may be undertaken at full scale if adequate backup options of cleanup are provided.

Many successful phytoremediation applications are actually restorations that save money after contaminated soil removal (e.g., wetland construction to offset hauling clean soil to fill excavations) and after other cleanup methods are applied. Some applications are temporary plantings of grass and trees to hold soil and contaminants in place during assessment of risks and cleanup

alternatives. Selection of the right plants and monitoring have led to temporary vegetation becoming a final remedy for cleanup and control of a site until complete removal occurs.

In general, phytoremediation is a niche waste management technique, normally applied to manage or clean up low levels of contaminants over large areas. Some cleanups may occur within months to a year or two, but typically 3–5 years are required and some difficult-to-manage sites may require decades to clean up. `Phytostabilization', which does not seem to be sustainable, may be the only technically feasible option for large sites on which animal and human exposure can be controlled while innovative options are explored and developed.

While most phytoremediation applications are limited to sites with chronic toxicity or conventional waste disposal of municipal sewage or sludge because of phytotoxicity, some innovations allow cleanup of acutely toxic wastes. For example, sterile, contaminated soils can be flooded and nonrooting wetland plants used to rapidly transform leaching TNT. Wetland cells can be designed with hardy pioneers, dead plant materials, and highly active sediments to buffer the highest incoming explosives concentrations. Because cyanide is a secondary metabolite, some trees grow into soils that are acutely toxic to mammals and break the waste down completely enough that tree plantations with appropriate ground cover can be converted early to parks and other less restrictive land uses (**Fig. 4**). Burrowing animals must be excluded until the tree roots have sufficient time to explore all parts of the soil and break down the cyanide complexes.

Some acutely toxic wastes can be treated in more expensive and energy-intensive unit processes. Cofactors, amendments, and hazardous by-products can be highly and reliably controlled. *Ex situ* treatments tradeoff cost savings achieved with *in situ* treatments for control of pH, temperature, and other factors. For example, a feasible unit process to remove phenols from liquid wastes uses horseradish peroxidases, but to date, this process does not seem to have been scaled up and applied to an actual waste stream. Treatment of liquid wastes contaminated with explosives have been proposed using slurries of plants high in nitroreductase activity and excavated soils contaminated with explosives could be buried with plant material high in nitroreductase activity.

Most phytoremediation applications are limited to root zones and shallow waters, but occasionally exudates leach below root zones to generate reducing conditions and spur microbial dehalogenation 3–5 m (10–15 feet) deep. Transpiration can pull slow-moving groundwater plumes into the root zone, into wetland soils, and into plants to completely clean up some sites. Feasibility studies indicate plumes may be controlled to depths up to 15 m (50 feet) by very large plantations depending on the hydraulic conductivity of the aquifer, the control of surface runoff, and the rooting depth of the trees or grasses employed. Some solar pump-and-treat applications apply root training and management to optimize root contact with contaminated soil and water.

Even though very few phytoremediation applications have been optimized for sustainability, most applications are inexpensive and have many secondary benefits. Sunlight is the chief energy source in most applications. Ecological restoration begins during treatment, rather than after, compared with many cleanups that rely on heavy equipment, incineration, and other intense treatments, harsh chemicals, and traffic of heavy vehicles into and out of the site. These intense cleanups may be necessary when longer periods of time and sufficient land areas are not available, or acute phyto- and human toxicity cannot be managed *in situ*.

See also: Ecological Complexity: Citizen Science. Global Change Ecology: Xenobiotic (Pesticides, PCB, Dioxins) Cycles

Further Reading

Brooks, R.R. (Ed.), 1998. Plants That Hyperaccumulate Heavy Metals. New York: CABI.
Cherian, S., Oliveira, M.M., 2005. Transgenic plants in phytoremediation: Recent advances and new possibilities. Environmental Science and Technology 39 (24), 9377–9390.
Coyle, C., Duggan, P., Godinho, M., 1999. The development of a phytoremediation technique for the detoxification of soils contaminated with phenolic compounds using horseradish peroxidase (*Armoracia rusticana*): Preliminary results. International Journal of Phytoremediation 1 (2), 189–202.
Dec, J., Bollag, J.-M., 1994. Use of plant material for the decontamination of water polluted with phenols. Biotechnology and Bioengineering 44, 1132–1139.
Dhankher, O., Rosen, B., Shokes, J., *et al.*, 2006. Increased arsenic uptake by plants knocked down for an endogenous arsenate reductase. Proceedings of the National Academy of Sciences of the United States of America 103, 5413–5418.
Doty, S.L., Shang, T.Q., Wilson, A.M., *et al.*, 2000. Enhanced metabolism of halogenated hydrocarbons in transgenic plants containing mammalian cytochrome P450 2E1. Proceedings of the National Academy of Sciences of the United States of America 97, 6287–6291.
French, C.E., Rosser, S.J., Davies, G.J., Nicklin, S., Bruce, N.C., 1999. Biodegradation of explosives by transgenic plants expressing pentaerythritol tetranitrate reductase. Nature Biotechnology 17, 491–494.
Gordon, M., Choe, N., Duffy, J., *et al.*, 1998. Phytoremediation of trichloroethylene with hybrid poplars. Environmental Health Perspectives 106 (4), 1001–1004.
Hannink, N., Rosser, S.J., French, C.E., *et al.*, 2001. Phytodetoxification of TNT by transgenic plants expressing a bacterial nitroreductase. Nature Biotechnology 19, 1168–1172.
Hong, M.S., Farmayan, W.F., Dortch, I.J., *et al.*, 2001. Phytoremediation of MTBE from a groundwater plume. Environmental Science and Technology 35 (6), 1231–1239.
McCutcheon, S.C., 1993. Hazardous waste engineering, editorial. Journal of Environmental Engineering 119 (5), 769–770.
McCutcheon, S.C., Schnoor, J.L. (Eds.), 2003. Phytoremediation: Transformation and Control of Contaminants. New York: Wiley.
Meagher, R.B., Heaton, A.C., 2005. Strategies for the engineered phytoremediation of toxic element pollution: Mercury and arsenic. Journal of Industrial Microbiology and Biotechnology 32, 502–513.
Meagher, R.B., Kim, T., Smith, A.P., Heaton, A.C.P., 2005. Designing plants for the remediation of mercury- and arsenic-polluted soils and water. In: McKeon, T. (Ed.), Designing Industrial Crops. Washington, DC: ACS Books.
Medina, V.F., Larson, S.L., Agwaramgbo, L., Perez, W., 2002. Treatment of munitions in soils using phytoslurries. International Journal of Phytoremediation 4 (2), 143–156.
Mitsch, W.J., Jørgensen, S.E., 2003. Ecological Engineering and Ecosystem Restoration. New York: Wiley.
Newman, L.A., Strand, S.E., Choe, N., *et al.*, 1997. Uptake and biotransformation of trichloroethylene by hybrid poplars. Environmental Science and Technology 31, 1062–1067.

Newman, L.A., Wang, X., Muiznieks, I.A., *et al*., 1999. Remediation of TCE in an artificial aquifer with trees: A controlled field study. Environmental Science and Technology 33 (13), 2257–2265.

Odum, H.T., Odum, B., 2003. Concepts and methods of ecological engineering. Ecological Engineering 20, 339–361.

Quinn, J., Negri, M.C., Hinchman, R.R., Moos, L.P., Wozinak, J.B., 2001. Predicting the effect of deep-rooted hybrid poplars on the groundwater flow system at a large-scale phytoremediation site. International Journal of Phytoremediation 3 (1), 41–60.

Rugh, C.L., Wilde, H.D., Stack, N.M., *et al*., 1996. Mercuric ion reduction and resistance in transgenic *Arabidopsis thaliana* plants expressing a modified bacterial *mer*A gene. Proceedings of the National Academy of Sciences of the United States of America 93, 3182–3187.

Sandermann, H., 1992. Plant metabolism of xenobiotics. Trends in Biochemical Sciences 17, 82–84.

Strand, S.E., Newman, L., Ruszaj, M., *et al*., 1995. Removal of trichloroethylene from aquifers using trees. In: Vidic, R.D., Pohland, F.G. (Eds.), Innovative Technologies for Site Remediation and Hazardous Waste Management, Proceedings of the National Conference Pittsburgh, July 26. Reston, VA: Environmental Engineering Division, American Society of Civil Engineers, pp. 605–612.

Relevant Website

http://www.epfl.ch
COST Action 837, École Polytechnique FÅdÅrale de Lausanne.

Plant Demography

Christian Damgaard, Aarhus University, Aarhus, Denmark

The ecological success of a plant species within its realized niche is typically described by the observed change in plant abundance. In order to more fully understand changes in plant abundance and make ecological predictions of the effects of environmental changes and gradients, it is necessary to investigate and quantify the underlying ecological processes. In principle, this means that all the different interactions between the species in the plant community need to be investigated (Damgaard, 2003). However, this is typically an exceedingly demanding task in multi-species plant communities and, instead, we investigate how the vital rates of the most common species, when competing with other species, are affected by the environment (Harper, 1977).

Furthermore, demographic processes, such as colonization and mortality, have been increasingly used in spatiotemporal modeling of plant populations (Normand *et al.*, 2014). This is motivated by the notion that changes in the probabilities of colonization and mortality may be the driving changes in species ranges. Furthermore, a demographic approach to understanding and predicting species' range dynamics has the advantage of being rooted in ecological theory (reviewed in Normand *et al.*, 2014), and has been shown to improve predictive ability of niche dynamics as it includes plasticity and local adaptation (Morin and Thuiller, 2009). This connection between the Hutchinson realized niche and demography was first pointed out by Maguire (1973), who modeled niche space as functions of demographic parameters.

The Consequences of Being Sedentary

Most plants are firmly rooted into the ground and this sedentary life form has a profound impact on the life history and ecology of plants (Harper, 1977). Germinated seedlings cannot move away from larger neighboring plants, and the competitive effect of neighboring plants are the universally most important biotic limiting factors of plant establishment and growth, although, positive plant-plant interactions have been reported in extreme environments (Callaway, 1995; Stoll and Weiner, 2000). Other biotic factors that control plant establishment and growth are herbivores and pathogens, but their role is more variable and less important than the negative competitive effect from neighboring plants as conveyed by the "green earth hypothesis" by Hairston *et al.* (1960).

As an adaptation to the sedentary life form and competition for limiting resources (light, water or nutrients) from neighboring plants, the same plant genotype may display a large variation in phenotypic characters. Especially the number of different plant parts, i.e. branches, leaves, flowers, fruits etc., may vary considerably (Harper, 1977), and in plant populations at medium or high densities where will often be a large size variation among conspecific individuals of approximately the same age (Weiner, 1990).

Self-fertilization is also an adaptation to a sedentary life form; after a colonization event, isolated sexual plants may have difficulties to receive compatible pollen if they are self-incompatible. With self-compatible individuals a single propagule is sufficient to start a sexually reproducing population, making its establishment much more likely than if the chance growth of two self-incompatible individuals sufficiently close together spatially and temporally is required (Baker, 1955).

A necessary adaptation to the sedentary adult life form is seed and/or vegetative dispersal. After a local extinction event, plant species have to be able to re-colonize the area, and on a larger time scale, plant species have to colonize new areas due to the recurring environmental changes, e.g., climate change. This dispersal is uneven and the number of dispersal units will typically decrease with the distance from the parent plant. Consequently, plant populations have an aggregated spatial distribution within a plant community, where the local abundance is determined by demographic processes.

In order to measure the effect of demographic processes on plant community dynamics it is an essential prerequisite that plant abundance may be measured in an unbiased and meaningful way. However, for many plant species, and ultimately a further consequence of their sedentary life form, it is not possible to distinguish individual plants from each other, and while the number of individuals or density is the theoretically most natural measure of plant abundance, this measure is not a relevant measure in many habitat types, e.g. grasslands.

Measuring Plant Abundance

The ultimate way of measuring plant abundance is by assigning the local species abundance or biomass as an attribute to a geographic position. While such geographical abundance data most certainly will be more and more common due to new sampling techniques from drones and satellites; most existing and currently sampled plant abundance data use a plot design, where several relatively small plots are sampled within a larger area.

This is an update of F.X. Picó, A. Rodrigo and J. Retana, Plant Demography, In Encyclopedia of Ecology, edited by Sven Erik Jørgensen and Brian D. Fath, Academic Press, Oxford, 2008, pp. 2811–2817.

The classic measure of plant abundance is the probability of occurrence in small plots (Raunkiær, 1910). In order to collect presence - absence data, it is necessary to record all the plants in the plot, so that the absence of a species can be deduced from not being a member of the list of recorded plant species. If a plant species is inconspicuous or difficult to recognize the recorded number of presences may be down-biased relative to the true number of presences. However, it is possible to correct for such a bias if there is independent information on the detection probability of the species. Most of the plant species data in museums collection, atlas data, and the plant data collected by citizens are occurrence data or "presence-only data", where only the spatial location of an observed species is recorded without systematically recording all the plants in the same plot. Such occurrence data may be used to make maps of the species distribution or they may be regressed to maps of background data in order to investigate the effect of other factors on species occurrence. However, we cannot estimate the probability of occurrence without making simplifying assumptions.

If the size of the sample area for recording presence - absence data is reduced to a point then the recorded presence-absence data only depends on plant abundance, or more precisely, the frequency of presence-absence data collected at a point is an unbiased estimate of the surface-projected relative area of the plant species, which is called plant cover. Plant cover may also be assessed by a visual estimation of the surface-projected relative area of the plant species.

The measure of plant cover may be integrated with probability of occurrence by measuring the shortest distances from a sample of sampling points to the nearest aboveground part of plant species (up to a maximum search distance). Such measurements allow the calculation of an Integral occurrence probability (van Calster and Damgaard, 2017).

Furthermore, an important aspect to consider when measuring plant abundance in natural communities is that many plant species have an aggregated spatial pattern at the local scale due to e.g. the size of the plant, clonal growth, or limited seed dispersal. This local spatial aggregation lead to inflated spatial variance compared to a spatial randomness and, if not considered in the statistical modeling process, a pseudo-replication effect with biased estimates of *P*-values (Damgaard, 2013). If plant abundance data are sampled using a hierarchical sampling procedure where plots are sampled from a number of different sites, then the possible among-site variation in plant cover must be taken into account as well.

Plant Demographic Processes

Many plant species has a modular structure with a large size variation among individuals, and since plant demographic processes to a large extent depend on plant size it is not sufficient to simply count the number of individuals in an investigation of plant demographic processes (Harper, 1977). Consequently, plant size and the competitive growth of plants is an essential feature of plant demographic processes.

Competitive Growth

The energy requirement for plant maintenance and growth depends on the amount of photosynthesis that again depends on available water, carbon dioxide, and sunlight of a sufficient quality, i.e. short-wave solar radiation. Simply put, factors such as temperature, water and nutrient availability determine the potential for growth, but growth is only realized if plants receives sufficient good-quality light.

Neighboring plants limit the growth of each other because they compete for resources needed for plant growth. The mode of competition depends on the resource and the size of the competing plants. Some resources like short wavelength light, which is depleted approximately exponentially by successive leaf layers, may be monopolized by tall plants. Other resources, like phosphorus, are more or less evenly distributed in the soil, so that a large root system cannot prevent a small root system of phosphorus uptake. For resources that may be monopolized larger plants may be able to obtain proportionally more of the resource than their relative size advantage, and suppress growth of smaller individuals (Begon, 1984; Weiner, 1990). This positive effect of size on the competitive ability of a plant is known as size-asymmetric competition, although often-used synonyms in the ecological literature include, "one-sided competition", "contest competition", or "dominance-suppression competition".

Mortality

Plant mortality may be caused by either density-independent or density-dependent factors, although, in practice it is difficult to separate the two types of mortality. Harper (1977) gives the following example: "The mortality risk to a seedling from being hit by a raindrop or hailstone might be thought to be density-independent. Presumably the risk of being hit is independent of density but whether a seedling dies after being hit is a function of its size and vigor, both of which are strongly effected by density." Generally, the relative importance of density-dependent factors is expected to increase with density and plant size, and density-independent mortality may primarily occur at the seed and seedling stages by death during seed dormancy, some forms of seed predation, unfavorable soil conditions, local water availability etc.

Density-dependent mortality or self-thinning in synchronous monocultures show a surprisingly regular pattern: there seems to be a maximum density of surviving plants primarily controlled by the biomass of the plants. In an experiment with buckwheat (*Fagopyrum esculentum*) sown at different densities, Yoda *et al.* (1963) found that the three populations with the highest densities had reached similar densities 63 days after sowing. When plants in a synchronous monoculture grow, self-thinning will reduce the number of surviving plants to a maximum density and since the individual growing plants need more space and resources, the

maximum density decreases with time. For many species it has been found that if the average biomass is plotted against the maximum density in a log-log plot, then the relationship is approximately linear (e.g. Yoda *et al.*, 1963).

Reproduction

Fecundity, i.e. the number of seeds produced by a plant, is highly species-specific and mainly determined by the life history and adaptive strategy of the species. Generally, there is a trade-off between fecundity and seed size, where large seeds has a higher probability of establishment (Harper, 1977; Rees *et al.*, 2001).

Monocarpic plant species, e.g. annual plant species, convert all available resources at the time of reproduction into seeds and only reproduce once before they die. On the other hand, polycarpic perennial plants reproduce usually every year after a juvenile period. The perennial plants allocate resources between seed production and investment in structures that increase the survival probability and facilitate growth the following year. There is a weak tendency that the juvenile period is positively correlated with the expected life span of the species (Harper, 1977), and in general, the fecundity of polycarpic perennial plants is a function of plant size, which again is a function of age and environment.

In a pioneering study by Cole (1954), that was later slightly generalized and explained by Bryant (1971), it was demonstrated that a monocarpic annual plant only had to produce one extra effective seed compared to a polycarpic perennial plant in order for an annual life strategy to be an evolutionary adaptive strategy.

Dispersal and Colonization

Seed or vegetative dispersal is a necessary adaptation to a sedentary adult life form. Most plant communities are dynamic with continuous local disturbances followed by a relatively long succession process or may go locally extinct due to demographic or environmental stochasticity. After such local extinction events, it is important for the long time ecological success of the plant species that it will be able to re-colonize the area by immigration of seeds or vegetative propagules. In many habitats, seeds buried in the soil (the seed bank) plays an important role in preventing local extinction.

Seeds are dispersed by different dispersal vectors; some species rely primarily on wind dispersal, whereas other species invest in fruits for attracting animals that disperse the seeds after eating the fruits.

Seed dispersal is a highly stochastic and species-specific process, but generally, most seeds disperse only short distances and most likely will germinate in an environment that resemble the environment of parent plant. However, the invasion potential of the species after large catastrophic events such as ice ages, depend on rare long-distance events, where the establishment of secondary foci of spread will determine the overall speed of invasion (Shigesada and Kawasaki, 1997).

In a theoretical spatial community model with either local or global dispersal, Bolker and Pacala (1999) demonstrated that the only possible adaptive strategies are colonization, exploitation, or tolerance, which correspond to the three life history strategies, i.e. ruderal, competitive, and stress tolerant, suggested by Grime (1979). Furthermore, they observed two different short-dispersal adaptive strategies that both were weakly excluded in the non-spatial case; either a rapid exploitation strategy, or a phalanx strategy, where the plants form dense stands that exclude the establishment of other species (Bolker and Pacala, 1999).

It has been suggested that the range of realized niches primarily is determined by the ability of a species to colonize new resource space in competition with other species rather than securing survival at an already colonized resource space (Damgaard *et al.*, 2017). This hypothesis is related to the seed limitation hypothesis, i.e., that many plant species are limited by colonization (Turnbull *et al.*, 2000; Rees *et al.*, 2001). Likewise, Adler *et al.* (2010) has demonstrated that the recruitment process is more important than both survival and growth processes in determining niche differences. They hypothesized that pathogens and natural enemies have relatively strong effects during recruitment and mediate coexistence mainly at this stage of recruitment. The importance of dispersal traits for the ecological success of plant species has previously been demonstrated by (e.g. de la Riva *et al.*, 2011), and Schulze *et al.* (2012) showed that clonal growth is more important than sexual reproduction for determining population growth.

Population Growth

The demographic processes of competitive growth, mortality, reproduction and dispersal determine the local plant abundance and knowledge on the demographic processes may be used to predict population growth. However, the probability of germination, establishment, and reaching reproductive age in a natural habitat are highly stochastic processes and population growth can only be predicted with large uncertainty. Crawley *et al.* (1993) compared estimates of population growth of transgenic rapeseed (*Brassica napus*) in a natural habitat obtained by either using independent information on seed germination, mortality, and fecundity, or by measuring population growth directly by the difference in the number of seedlings from year to year. When population growth was modeled using the independently obtained information on mortality and fecundity the rapeseed population was predicted to increase in density, whereas the directly obtained estimate of population growth showed that the rapeseed population decreased in density. Later it was observed that the population actually was decreasing until it went extinct (Crawley

et al., 2001). Likewise, Stokes *et al.* (2004) found that variation in the probability of germination and establishment led to a high variation in the predicted population growth rates of *Ulex gallii* and *Ulex minor*.

In a pioneering study, Rees *et al.* (1996) measured population growth rates of four annual species in a natural habitat, and modeled their growth rates as a function of density. Using this density-dependent population model, the authors concluded that the population growth of the four annual species mainly was regulated by intraspecific competition (Rees *et al.*, 1996). The fitting of such density-dependent population models to time series data of plant abundance in natural habitats will surely be increasingly important in the investigation of the underlying ecological factors that regulate plant population growth.

Age-Structured Populations

The population growth of age-structured populations may be modeled by Leslie matrices, which are transition matrices of age-specific survival probabilities and fecundities (Caswell, 2001), or by using the renewal equation in continuous time (Gurtin and MacCamy, 1979).

A matrix population model is a convenient tool in the description of the demographic processes at the existing density and may provide predictions on the immediate future of the plant population. If the transition matrix consists of constant density-independent probabilities and fecundities, the mathematical property of the model is sufficiently simple so that the dominant eigenvalue, which corresponds to the growth rate of the population, may be determined (Caswell, 2001). However, if the transition probabilities and fecundities are functions of plant density, the mathematical properties of matrix models become more complicated and has to be calculated using numerical methods (e.g. Stokes *et al.*, 2004).

The complexities of stage-structured population growth with density-dependence are considerably reduced in the important case of an annual plant with a seed bank. The life cycle may be reduced to two stages and four recursive equations of the densities of seeds in the seed bank and plants at the reproductive age, respectively (Jarry *et al.*, 1995; Damgaard, 2005b).

Matrix population models for several hundred plant species has been collected within the COMPADRE network (Salguero-Gómez *et al.*, 2015). This database may be used to make comparative analyses of different life histories.

The spatial effects of disturbance and local recruitment in structured plant populations may be modeled in lattice models, which may be approximated with comparatively simple recursive equations. For example, the spatial demography of the perennial bushes *Cytisus scoparius* and *Ulex europaeus* has been described by recursive equations of the frequencies of each spatial site, which is respectively unsuitable for recruitment, open for recruitment, or already occupied by the plant (Rees and Paynter, 1997; Rees and Hill, 2001).

Size-Structured Populations

Since plants are plastic and have variable growth rates, two plants of the same age will not necessarily have equal survival probability or fecundity. Consequently, the demography of plant populations may be described by a stage-structured matrix model with transition matrices of stage-specific probabilities of moving to another stage and fecundities, or alternatively, using continuous life history characters by an integral projection model (Easterling *et al.*, 2000).

In an integral projection model, the state of the population is described by the size distribution $n(y,t)$, which is the number of individuals of size y at time t. A time step later, the population will have changed into,

$$n(y, t+1) = \int (p(x, y) + f(x, y))\, n(y, t)\, dx$$

where $p(x,y)$ is the probability that an individual size x at time t is alive and one time step later has size y, and $f(x,y)$ is the number of off springs of size y produced by individuals of size x. The model is parameterized using regression models of observed survival, growth and fecundity (Easterling *et al.*, 2000).

For many plant species the available life history data are naturally measured as continuous functions of size, and integral projection model is often a better choice when modeling plant population dynamics than traditional population matrix models (e.g. Rose *et al.*, 2005).

Occurrence at a Point

When fitting the above-mentioned population models to empirical data it is a requirement that the individual plants may be distinguished. However and as mentioned above, many plant species, and especially many of the grass species, forms carpet like multi-species plant communities, where it is impossible to distinguish individual plant from its neighbors. In such a plant community, it is not possible to measure e.g. size dependent fecundities, and instead population growth may be determined by the change in the occupancy of space.

If a species was present at time t at a specific spatial point, but absent at time $t + 1$, we may loosely speak of a mortality event. Additionally, if a species is absent from a specific pin-position at time t and present at time $t + 1$, we may loosely speak of a colonization event. Using such time series data at a point (pin-point or point intercept data) it is possible to estimate the probability of survival and colonization, and investigating the importance of either mortality or colonization as the cause for observed plant cover changes (Damgaard *et al.*, 2011; Damgaard *et al.*, 2017).

At Equilibrium

The demographic processes are highly stochastic and most plant communities will probably never reach an ecological steady state or an ecological equilibrium. Nevertheless, knowledge on the equilibrium states under certain rather strict assumptions may provide valuable information in predicting the future states of the plant community (e.g. Damgaard, 2005a). Consequently, it is of ecological interest to analyze the relevant population model with respect to possible equilibria and stability properties.

>Generally, density-dependency act as a stabilizing factor in most population models. For example, Verhulst (1838) proposed the continuous logistic equation as an adequate model for the negative effect of density on population growth, where population growth approached zero when the population size approach a carrying capacity of the population. Only if reproduction is high and regulated with a time lag the population size trajectory will become chaotic (May and Oster, 1976; Hoppensteadt, 1982).

Long-Term Ecological Data

In order to test various ecological hypotheses, e.g. the effect of density on plant population growth, it is valuable to have plant ecological data where plant abundance has been recorded repeatedly in a time series (Damgaard and Weiner, 2017), and long-term ecological data series are important for detecting causal relationships or predictability (Granger causality) in complex ecosystems (Sugihara *et al.*, 2012).

There are a few classic examples of long-term data sets of plant abundance, which have had a large impact of our knowledge on plant demography and ecology. Notable examples are the Park Grass Experiment, which was initiated in 1856 (Silvertown *et al.*, 2006), and the Cedar Creek Ecosystem Science Reserve, which investigates the effect of a nitrogen gradient (e.g. Tilman *et al.*, 2001). Fortunately, there is an increasing interest in the collection of long-term plant abundance data in order to answer important ecological questions, such as effects of climate change, and today new long-term ecological research sites are initiated and several networks are coordinating the research.

Several countries have set-up ambitious monitoring programs of terrestrial habitat types. For EU countries, these initiatives has partly been motivated by the EU habitat directive from 1992 (EU, 1992). Generally, monitoring programs are less focused and measure fewer variables than long-term ecological research (LTER) sites, but instead the monitoring programs has a better coverage of the geographical variation of different habitat types.

An interesting possibility of obtaining long-term demographic data is to use fossil pollen to estimate population sizes in the past at different spatial scales. Pollen grains are very robust and are present in high numbers in lake sediments and as the lake sediment gradually builds up during time, it is possible to determine when the pollen grain settled in the lake. From such pollen profile data and using species characteristics of pollen production and pollen dispersal, the number of plants within a certain area may be estimated. Alternatively, a demographic history of a species may be determined using a constructed phylogenetic tree based on a sample of DNA sequences. It has been shown that the distribution of coalescence times is a function of whether the population size is constant, decreasing or increasing and this effect may be explored in a statistical framework to test different demographic models (Emerson *et al.*, 2001).

Conclusion

The terminology and methodology of demography has historically been developed in the study of human and animal populations. However, plants are qualitative different from animals in several important aspects and these differences have important implications for the study of plant demographic processes.

Plant demographic processes are controlled by competitive interactions with neighboring plants in most terrestrial habitats (e.g. Hairston *et al.*, 1960). This fact together with the sedentary lifeform of plant species has important consequences for the investigation of plant demographic processes, which often only make sense in the light of local plant density and competitive growth. Furthermore, limited dispersal distances, which often leads to spatially aggregated plant populations and may affect the possible life history strategies (Bolker and Pacala, 1999), needs to be included in the population ecological modeling to a larger extent than presently is the current practice in most empirical plant demographic studies.

For many plant species, it is not possible to distinguish individual plants from each other, and while the number of individuals or density is the theoretically most natural measure of plant abundance, this measure is not a relevant measure in many habitat types, e.g. grasslands. Consequently, it is important to develop new methods for measuring demographic process for such plant species. This simple but crucial point has perhaps received too little attention in the plant ecological literature, and plant demographic studies are generally biased towards plant species that are readily counted and with conspicuous flowers and seeds. For example, in the COMPADRE database of matrix population models there were 316 records of *Poaceae* out of 6242 total plant records, which probably is an underrepresentation of the family compared to its important ecological role.

In summary, plant demography is a relatively mature scientific discipline, but it is important to critical assess the currently used terminology and methodology in order to align the empirical studies of plant populations with the typical features of plant life history.

See also: Ecological Data Analysis and Modelling: Spatial Models and Geographic Information Systems

References

Adler, P.B., Ellner, S.P., Levine, J.M., 2010. Coexistence of perennial plants: An embarrassment of niches. Ecology Letters 13, 1019–1029.

Baker, H.G., 1955. Self-compatibility and establishment after long-distance dispersal. Evolution 9, 347–349.

Begon, M., 1984. Density and individual fitness: Asymmetric competition. In: Shorrocks, B. (Ed.), Evolutionary ecology. Oxford: Blackwell, pp. 175–194.

Bolker, B.M., Pacala, S.W., 1999. Spatial moment equations for plant competition: Understanding spatial strategies and the advantages of short dispersal. The American Naturalist 153, 575–602.

Bryant, E.H., 1971. Life history consequences of natural selection: Cole's result. The American Naturalist 105, 75–76.

Callaway, R.M., 1995. Positive interactions among plants. The Botanical Review 61, 306–349.

Caswell, H., 2001. Matrix population models: Construction, analysis, and interpretation. Sunderland: Sinauer.

Cole, L.C., 1954. The population consequences of life history phenomena. Quarterly Review of Biology 29, 103–137.

Crawley, M., Hails, R.S., Rees, M., Kohn, D., Buxton, J., 1993. Ecology of transgenic oilseed rape in natural habitats. Nature 363, 620–622.

Crawley, M.J., Brown, S.L., Hails, R.S., Kohn, D.D., Rees, M., 2001. Transgenic crops in natural habitats. Nature 409, 682–683.

Damgaard, C., 2003. Modelling plant competition along an environmental gradient. Ecological Modelling 170, 45–53.

Damgaard, C., 2005a. Evolutionary ecology of plant-plant interactions—An empirical modelling approach. Aarhus, Denmark: Aarhus University Press.

Damgaard, C., 2005b. The probability of germination and establishment in discrete density-dependent plant populations with a seed bank: A correction formula. Population Ecology 47, 277–279.

Damgaard, C., 2013. Hierarchical and spatially aggregated plant cover data. Ecological Informatics 18, 35–39.

Damgaard, C., Merlin, A., Bonis, A., 2017. Plant colonization and survival along a hydrological gradient: Demography and niche dynamics. Oecologia 183, 201–210.

Damgaard, C., Merlin, A., Mesléard, F., Bonis, A., 2011. The demography of space occupancy: Measuring plant colonisation and survival probabilities using repeated pin-point measurements. Methods in Ecology and Evolution 2, 110–115.

Damgaard, C., Weiner, J., 2017. It's about time: A critique of macroecological inferences concerning plant competition. Trends in Ecology & Evolution 32, 86–87.

de la Riva, E.G., Casado, M.A., Jiménez, M.D., Mola, I., Costa-Tenorio, M., Balaguer, L., 2011. Rates of local colonization and extinction reveal different plant community assembly mechanisms on road verges in central Spain. Journal of Vegetation Science 22, 292–302.

Easterling, M.R., Ellner, S.P., Dixon, P.M., 2000. Size-specific sensitivity: Applying a new structured population model. Ecology 81, 694–708.

Emerson, B.C., Paradis, E., Thébaud, C., 2001. Revealing the demographic histories of species using DNA sequences. Trends in Ecology and Evolution 16, 707–716.

EU (Ed.), 1992. Council Directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora. European Commission.

Grime, J.P., 1979. Plant strategies and vegetation processes. Chichester, UK: John Wiley and Sons.

Gurtin, M.E., MacCamy, R.C., 1979. Some simple models for nonlinear age-dependent population dynamics. Mathematical Biosciences 43, 199–211.

Hairston, N.G., Smith, F.E., Slobodkin, L.B., 1960. Community structure, population control, and competition. The American Naturalist 94, 421–425.

Harper, J.L., 1977. Population biology of plants. London: Academic Press.

Hoppensteadt, F.C., 1982. Mathematical methods of population biology. Cambridge: Cambridge University Press.

Jarry, M., Khaladi, M., Hossaert-McKey, M., McKey, D., 1995. Modelling the population dynamics of annual plants with seed bank and density dependent effects. Acta Biotheoretica 43, 53–65.

Maguire Jr., B., 1973. Niche response structure and the analytical potentials of its relationship to the habitat. The American Naturalist 107, 213–246.

May, R.M., Oster, G.F., 1976. Bifurcations and dynamic complexity in simple ecological models. The American Naturalist 110, 573–599.

Morin, X., Thuiller, W., 2009. Comparing niche- and process-based models to reduce prediction uncertainty in species range shifts under climate change. Ecology 90, 1301–1313.

Normand, S., Zimmermann, N.E., Schurr, F.M., Lischke, H., 2014. Demography as the basis for understanding and predicting range dynamics. Ecography 37, 1149–1154.

Raunkiær, C., 1910. Formationsundersøgelse og formationsstatistik. (English translation 1934: Investigation and statistics of plant formations. In: The life forms of plants and statistical plant geography, Oxford) Botanisk Tidsskrift 30, 20–132.

Rees, M., Condit, R., Crawley, M., Pacala, S., Tilman, D., 2001. Long-term studies of vegetation dynamics. Science 293, 650–655.

Rees, M., Grubb, P.J., Kelly, D., 1996. Quantifying the impact of competition and spatial heterogeneity on the structure and dynamics of a four-species guild of winter annuals. The American Naturalist 147, 1–32.

Rees, M., Hill, R.L., 2001. Large-scale disturbances, biological control and the dynamics of gorse populations. Journal of Applied Ecology 38, 364–377.

Rees, M., Paynter, Q., 1997. Biological control of scotch broom: Modelling the determinants of abundance and the potential impact of introduced insect herbivores. The Journal of Applied Ecology 34, 1203–1221.

Rose, K.E., Louda, S.M., Rees, M., 2005. Demographic and evolutionary impacts of native and invasive insect herbivores on Cirsium canescens. Ecology 86, 453–465.

Salguero-Gómez, R., Jones, O.R., Archer, C.R., Buckley, Y.M., Che-Castaldo, J., Caswell, H., Hodgson, D., Scheuerlein, A., Conde, D.A., Brinks, E., de Buhr, H., Farack, C., Gottschalk, F., Hartmann, A., Henning, A., Hoppe, G., Römer, G., Runge, J., Ruoff, T., Wille, J., Zeh, S., Davison, R., Vieregg, D., Baudisch, A., Altwegg, R., Colchero, F., Dong, M., de Kroon, H., Lebreton, J.-D., Metcalf, C.J.E., Neel, M.M., Parker, I.M., Takada, T., Valverde, T., Vélez-Espino, L.A., Wardle, G.M., Franco, M., Vaupel, J.W., 2015. The compadrePlant matrix database: An open online repository for plant demography. Journal of Ecology 103, 202–218.

Schulze, J., Rufener, R., Erhardt, A., Stoll, P., 2012. The relative importance of sexual and clonal reproduction for population growth in the perennial herb Fragaria vesca. Population Ecology 54, 369–380.

Shigesada, N., Kawasaki, K., 1997. Biological invasions: Theory and practice. Oxford: Oxford University Press.

Silvertown, J., Poulton, P., Johnston, E., Edwards, G., Heard, M., Biss, P.M., 2006. The park grass experiment 1856–2006: Its contribution to ecology. Journal of Ecology 94, 801–814.

Stokes, K.E., Bullock, J.M., Watkinson, A.R., 2004. Population dynamics across a parapatric range boundary: *Ulex gallii* and *Ulex minor*. Journal of Ecology 92, 142–155.

Stoll, P., Weiner, J., 2000. A neighborhood view of interactions among individual plants. In: Dieckmann, U., Law, R., Metz, J.A.J. (Eds.), The geometry of ecological interactions. Cambridge: Cambridge University Press, pp. 11–27.

Sugihara, G., May, R., Ye, H., Hsieh, C.-h., Deyle, E., Fogarty, M., Munch, S., 2012. Detecting causality in complex ecosystems. Science 338, 496–500.

Tilman, D., Reich, P.B., Knops, J., Wedin, D., Mielke, T., Lehman, C., 2001. Diversity and productivity in a long-term grassland experiment. Science 294, 843–845.

Turnbull, L.A., Crawley, M.J., Rees, M., 2000. Are plant populations seed-limited? A review of seed sowing experiments. Oikos 88, 225–238.

van Calster, H., Damgaard, C., 2017. Estimation of relative frequency and cover of plant species by measurements of minimum distances from sampling points. Journal of Vegetation Science.

Verhulst, J.H., 1838. Notice sur la loi que population suit dans son accroissement. Correspondance Mathématique et Physique 10, 113–121.

Weiner, J., 1990. Asymmetric competition in plant populations. Trends in Ecology and Evolution 5, 360–364.

Yoda, K., Kira, T., Ogawa, H., Hozumi, K., 1963. Self-thinning in overcrowded pure stands under cultivated and natural conditions. Journal of Biology, Osaka City University 14, 107–129.

Spatial Distribution

MK Borregaard, DK Hendrichsen, and G Nachman, University of Copenhagen, Copenhagen, Denmark

Organisms do not occur randomly in space. Any species of plant or animal may be found in some areas, while they are completely absent from others. Likewise, the individuals of any one species are distributed in relation to each other in distinct patterns. The reasons for the readily apparent nonrandomness of the spatial distribution patterns of organisms are numerous, and the patterns result from processes acting throughout the whole life cycle of the organism, and on various spatial scales. Although this spatial structuring of populations is often ignored in ecological theory, it has profound implications for the mediation of biological processes: interactions between individuals and across species all take place in space as well as in time, and an understanding of spatial patterns is basic to understanding real-life ecological processes. Indeed, patterns of spatial distribution play an important role in shaping a wide range of ecological dynamics, such as intra- and interspecific competition, mating systems, predation, population genetics, and the spread of contagious diseases.

This article presents an overview of how spatial patterns in the distribution of organisms are created, and how they influence the way ecological processes run their course in ecological communities, exemplified by predator–prey dynamics. These patterns are scale dependent: organisms which are distributed in one way when observed at a large spatial scale may be distributed very differently at closer scales. To accommodate this, the presentation is structured according to the main spatial scale of the patterns under discussion. The initial focus lies on the way organisms are distributed at a landscape level, at which spatial distribution is mainly influenced by topographical features and variation in habitat availability. The subsequent discussion moves the scale to patterns in the dispersion of individuals, which can be seen primarily as an effect of behavioral interactions with conspecifics and with those of predator and prey species. To conclude, large-scale regional distribution patterns are briefly discussed, in relation to how they may contribute to the observed spatial distribution patterns at smaller scales.

The Distribution of Organisms Over Landscapes

An organism may only persist where the physical conditions (temperature, humidity, etc.) are tolerable and the food resources are adequate. In other words, the environment should match the niche space of the species. Consequently, all organisms are associated with a specific type of habitat, and hence view the area around them in widely different manners. In this context, it is useful to think of the landscape as a mosaic, consisting of patches of favorable habitat surrounded by uninhabitable areas. Within such patches the species may form more or less permanent local populations (also called subpopulations), while the species is only rarely found in the intervening areas.

The dynamics of a subpopulation are driven by several processes, as shown in **Fig. 1**. Individuals enter the population through birth and immigration, and leave it as a result of death and emigration. Of these processes, birth and death rates may be controlled by the number of individuals already present, their density, and the resource availability. Immigration and emigration may also be affected by these factors, as well as by the favorability of the surrounding landscape and the size of the adjacent subpopulations.

If we increase our vantage point to include several such subpopulations located over the landscape, which are separated by uninhabitable areas but with migration occurring between them, then a very complex picture appears (**Fig. 2**). All these interacting subpopulations can be viewed as one large spatially structured population, known as a metapopulation. The complex spatial dynamics of the metapopulation result from the patterns of between-patch movement and local birth and death rates, as well as local extinction and colonization of subpopulations which occur over larger timescales. These processes can be modeled by a set of connected differential equations, which allows quantitative predictions to be made about the dynamics of the metapopulation.

Other factors, which are not accounted for by the basic model described above, may be included in more sophisticated representations. These factors include the size and favorability of patches, their relative isolation, and the nature of the intervening habitat relative to the dispersal ability of the organism. There are also dynamic processes affecting the occupancy of each patch over time. These processes rely on the observation that over time, the subpopulation in many cases causes a reduction in the quality of the patch which it inhabits (e.g., reduces the amount of resources and/or attracts predators, parasites, and pathogens). As the environment deteriorates, mortality and emigration is likely to increase while the rate of reproduction correspondingly decreases. In addition, Allee effects (see later) can hasten the demise of the subpopulation. A further complication to this pattern is that these effects are not restricted to the species in question, but are equally likely to affect its predators and the vital bioresources, such as food resources on which it depends.

An interesting consequence of the spatial movement of individuals between subpopulations is that some patches may be occupied, even though they cannot in themselves sustain viable populations of the species. The subpopulation of such a patch is actually kept alive by the immigration of individuals from more productive patches in the vicinity. The movement of individuals resembles water flowing from its source to a sink, and correspondingly this type of system is usually termed source–sink dynamics. In some areas, especially in transition zones between habitats or biomes and in areas severely affected by human disturbance, such source–sink dynamics may play a key role in structuring the occurrence and distribution of individuals.

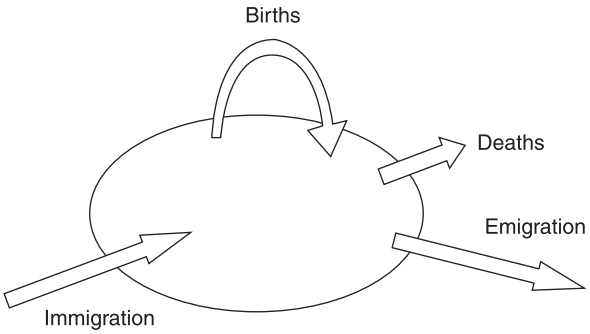


Fig. 1 The abundance of a subpopulation within a patch of suitable habitat is affected by the processes of birth, death, immigration, and emigration.

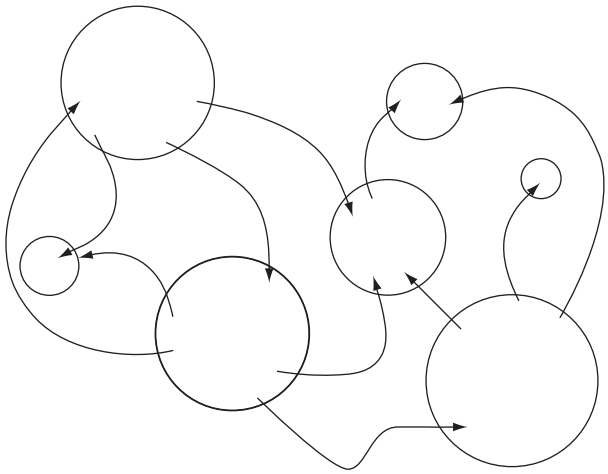


Fig. 2 In a landscape consisting of several suitable habitat patches, the population dynamics become dependent on dispersal and colonization between interconnected subpopulations. Such a system is described by metapopulation dynamics.

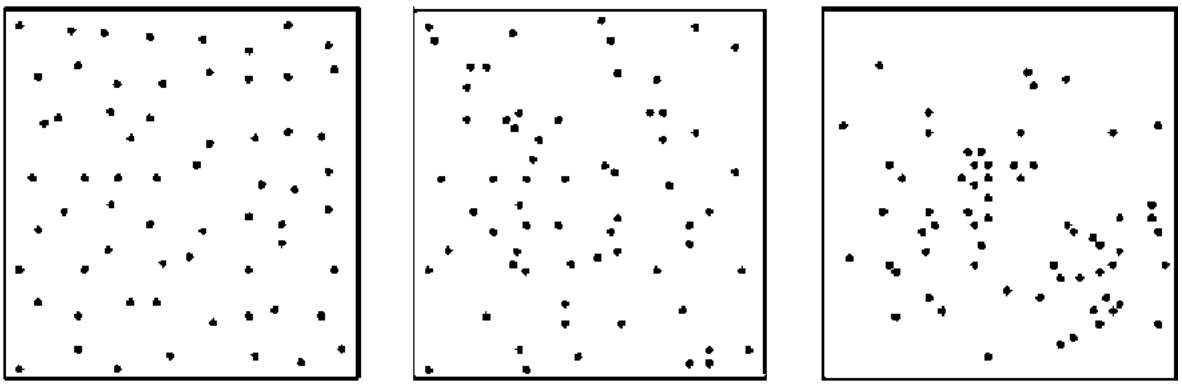


Fig. 3 A schematic representation of the dispersion of individuals in a subpopulation with (from left to right) either a regular $(s^2 < \bar{x})$, a random $(s^2 \approx \bar{x})$, or a clumped $(s^2 > \bar{x})$ distribution.

Dispersion of Individuals

An alternative way of describing spatial distributions is to move the focus to the position of separate individuals in space. When viewed from above, the distribution of individuals over the landscape can be visualized as a pattern of tiny dots dispersed over a blank area. This pattern may in theory be completely random, but usually individuals are either clumped together, or at the opposite, spread out in a regular fashion (see Fig. 3).

Empirically, dispersion patterns can be characterized by the average number of individuals in randomly sampled units of area within the landscape ($\bar{x}$) and the variance between number of individuals occurring per sampling unit (s^2). The ratio $s^2/\bar{x}$ is commonly known as the index of dispersion. In randomly distributed populations, the index of dispersion is approximately equal to 1 (see **Box 1**).

Clearly, dispersion patterns are strongly dependent upon the scale at which they are perceived – for instance, the distribution of an insect species living on leaves of trees can be studied on many different spatial scales, such as the distribution on each leaf, among leaves within a branch, among branches within a tree, and between trees within a forest.

From the beginning, it seems obvious that there should be some clumping of individuals – organisms do not appear out of the thin air. Each individual originates from a parent, and as such appears in close vicinity to other individuals, which in most cases are relatives. After birth, or at a later stage in the life cycle, they undergo some juvenile dispersal before settling down in a favorable location, but some degree of aggregation is still retained. Another factor which predisposes for aggregation is that individuals of the same species are attracted to a set of conditions and resources, which are themselves patchily distributed.

Aggregated dispersion patterns are very common in nature. An example of an extremely aggregated species is the Kashmir cave bat (*Myotis longipes*), which is only known from nine localities in the Himalaya region, each home to populations of a thousand individuals or more.

A regular or even distribution, on the other hand, is the result when individuals compete for limited resources. This type of pattern is commonly exhibited by many sessile organisms such as trees, which space themselves evenly as a result of competition

Box 1 Analysis of spatial data

As the null hypothesis, it is assumed that individuals in a population are randomly distributed among the n sampling units of a sample. If this is the case, it is expected that the variance should equal the average so that the `index of dispersion', $s^2/\bar{x}$, is approximately equal to 1. If the ratio exceeds 1, it indicates that the population has a patchy (or clumped) distribution whereas a value less than unity indicates an even (or regular) distribution. However, since data originate from sampling, they will always be associated with some variation, so it is likely that some deviation in $s^2/\bar{x}$ from unity will occur even if the underlying distribution is random. Especially if the sample size n is small, $s^2/\bar{x}$ will exhibit large variation due to sampling noise. A χ^2-test can be used for testing whether $s^2/\bar{x}$ deviates significantly from 1 since $\chi^2 = (n-1)s^2/\bar{x}$ with $n-1$ degrees of freedom. It should be noted that the test is two-tailed (in contrast to the majority of cases where χ^2-tests are used) since values significantly smaller or larger than $n-1$ can lead to rejection of the null hypothesis.

Though the index of dispersion indicates whether a population is evenly, randomly, or patchily distributed in space, it does not explicitly reveal information about the underlying spatial distribution. This requires that the empirical distribution of sampling units with x individuals can be fitted by a theoretical `probability function' called $P(x)$, which denotes the probability that a randomly selected spatial unit contains exactly individuals. As all probability functions, $P(x)$ for all possible integer values of x equal to or larger than 0 should sum to unity.

The Poisson distribution is used to describe the underlying distribution when it is random, the positive binomial distribution when it is even and the negative binomial distribution when it is clumped. However, other less frequently used distributions are also available to model clumped populations, for example, the Thomas distribution, the logarithmic series distribution, the Pólya–Aeppli distribution, and Neyman's type A, B, and C distributions. Once an adequate probability function has been identified and fitted to data, the quality of the fit can be assessed by means of a goodness-of-fit test, usually a χ^2 one-sample test or a Kolmogorov–Smirnov one-sample test.

A problem often encountered in analyzing spatial data statistically is the fact that they do not represent independent observations. Thus, if sampling unit i is separated from sampling unit j by a distance d_{ij}, it seems likely that x_i will be more similar to x_j, the smaller the d_{ij} is. This phenomenon is known as `spatial autocovariation'. Spatial autocovariance is often depicted as a so-called `semivariogram' where the `semivariance' (γ_d) is plotted against d. The semivariance at distance d is calculated as $\gamma_d = \sum_{i=1}^{n_d} (X_{i+d} - x_i)^2 / 2n_d$, where x_{i+d} is the value of x measured at distance d from another measurement x_i and n_d is the number of measurements separated by distance d.

Spatial patterns can be depicted graphically by means of a technique known as `kriging'. The principle is to place a large number of points spaced out over the entire area under study. Each point is characterized by its coordinates in the two-dimensional x–y space, and by the value of a given attribute (for instance the population density in the area around the point). The value of the attribute is denoted the z coordinate, which represents a height above the x–y plane. Hence, small and large values of z will appear as troughs and peaks in a three-dimensional (3-D) landscape. 3-D landscapes can be projected into two-dimensional (2-D) landscapes by means of contour plots where points with identical z-values are connected with lines (isoclines), similar to how temperature and atmospheric pressure are depicted in meteorological maps. The more fine-grained the information is, the more precise the map will be. Various algorithms have been developed to interpolate values between neighboring points so as to estimate z in points that have not been sampled, and to smooth out the landscape by removing local peaks and troughs caused by sampling noise.

Since kriging is computationally demanding, various specialized software products exist to perform it, for example, easy_krig, DACE, GS$^+$. In addition, kriging can be handled by some statistical packages, such as R.

for water or sunlight. Also many animals have approximately regular distributions – a familiar example is the territories of songbirds.

Biological Impacts on Spatial Patterns

As mentioned above, the dispersion pattern of individuals is affected by the structure of the landscape and the resource demands of the organism. Theoretically, if individuals were similar and completely free to move, they would be dispersed over the landscape so that each individual had the same access to the resource. This pattern is known as the ideal free distribution, and will simply reflect the instantaneous distribution of resources. However, differences between the competitive abilities of individuals, habitat barriers hindering free movement, and the individuals' lack of perfect knowledge about the distribution of resources all contribute to making examples of populations following the ideal free distribution rare in nature.

The behavior and life styles of organisms are extremely varied, and patterns of spatial distribution exhibit a great deal of variation. Mobile organisms move around in order to acquire resources and may also engage in social interactions with conspecifics. This means that behavioral choices play a profound role in shaping occurrence patterns. One of the most obvious cases of this effect is the congregation into cooperative flocks seen in many animal species. The exact nature of these flocks differs widely among organisms – from the loose aggregations of resting brent geese (*Branta bernicla*), via the socially complex cooperative units of wolf (*Canis lupus*) packs to the super-individual hive structure of eusocial insects.

Optimal Group Size

Despite the great variation in types of animal groups, the biology of flock behavior has important commonalities across all species. An example is the determinants of group size, which can be seen as a function of the costs and benefits of being part of a group.

A group consisting of N individuals can increase its size without recruitment of individuals from outside as long as the group's per capita growth rate r in eqn [1] is positive:

$$\frac{dN}{dt} = rN \quad [1]$$

r is likely to be a function of group size, reflecting the difference between benefits and costs of adding further individuals to the group. Hence, r may be modeled as a difference between two functions:

$$r(N) = \text{benefits} - \text{costs} = f(N) - g(N)(N \geq 1) \quad [2]$$

where $f(N)$ denotes the benefit function and $g(N)$ the cost function. Both functions can either be independent of N or increase with N. Since groups cannot be infinitely large, there will exist a value for N satisfying the condition that $r(N) = 0$ for $N = N_{max}$ and $r(N) < 0$ for $N > N_{max}$. Benefits of increasing group size are likely to level off with N, while costs are likely to accelerate as N becomes large as indicated in **Fig. 4a**. N_{max} corresponds to the value of N when costs and benefits balance.

When group size reaches its maximum size, r will be 0, indicating that the fitness of group members is low. By reducing group size from this point, the fitness of each individual will increase. The optimal group size (denoted N_{opt}) is reached when r is at maximum (**Fig. 4b**). If N_{opt} is close to 1, individuals will gain by living alone, for instance, in territories, and their spatial distribution will tend to be regular or random. In contrast, species with high values of N_{opt} will be patchily distributed. This will apply to species that can adjust their group size in accordance to net benefits, for instance, in animals where individuals can join (if $N < N_{opt}$) or leave groups (if $N > N_{opt}$). If groups are not formed by such behavioral mechanisms, as with plants, group size will continue to grow until $N = N_{max}$. A special case of eqn [2] is when benefits and costs balance at two different group sizes, as shown in **Fig. 4c**. When N is below N_{min}, the group will go extinct because the smaller the group the more negative r becomes (the so-called Allee effect). The only way the extinction of such a group can be avoided is by recruiting individuals from outside.

Population-Level Variation in Dispersion Patterns

As noted above, the behavioral life styles of animals determine their distribution. While some animals, such as the wildebeest (*Connochaetes taurinus*), live in groups that migrate over large areas or restrict their dispersal to extended home ranges, many others are territorial, dividing the available habitat into separate areas which are defended against the intrusion of conspecifics.

In addition, different members of a population may exhibit widely different patterns of dispersion, resulting in highly complex overall spatial patterns. Territoriality, for instance, causes all reproducing individuals to adhere to a regular dispersion pattern – but for nonreproducing individuals the situation may be very different.

An example is provided by the tawny owl (*Strix aluco*) which is a long-lived, monogamous, territorial bird (**Fig. 5**). Juveniles only get access to optimal resources when adults die and thereby leave vacant territories. Adults defend their territories fiercely against intruders, though they are more willing to accept juveniles within the territory boundaries during July and August, when the juveniles disperse from their natal territories.

This difference between the lifestyles of adults and juveniles means that only few individuals survive to reach adulthood. Adult birds within their territories have a relatively high survival; telemetry studies show that a main cause of death within the

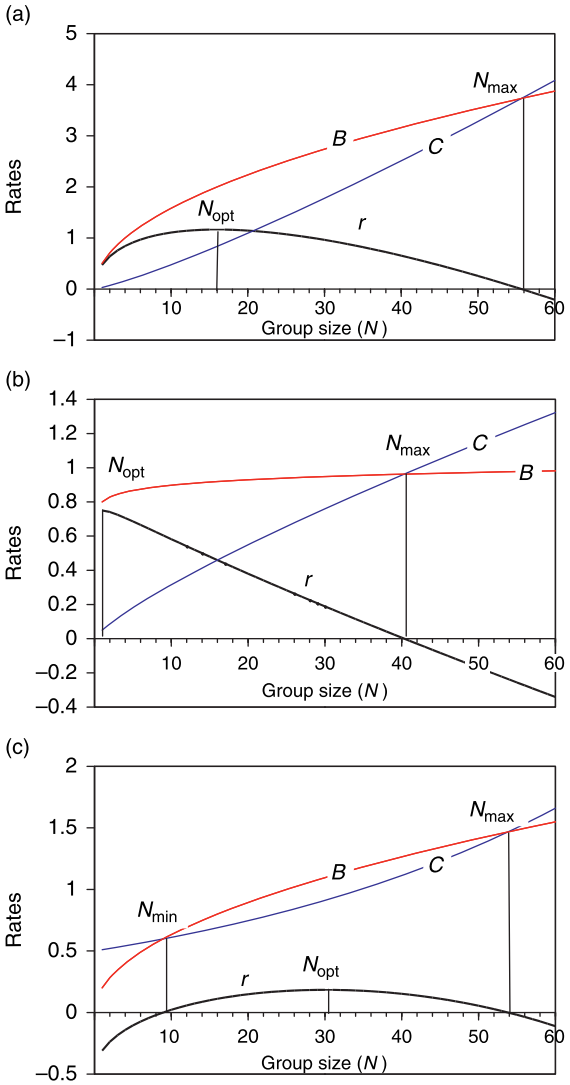


Fig. 4 Group formation is predicted to occur when the relative growth rate (r) increases with group size. Depending on how the benefits (B) and costs (C) of group behavior change with group size (N), three different cases can be identified: (a) Benefits increase more steeply than costs when group size is small, and vice versa when group size is large, leading to an optimal group size (N_{opt}) toward which the group is predicted to converge. If a group does not split up when $N > N_{opt}$, it will continue to grow until $N = N_{max}$. (b) Solitary behavior (i.e., $N = 1$) is predicted if costs increase more than benefits as group size increases. (c) A group must be larger than N_{min} in order to persist and grow, whereas it is goes toward to extinction if N falls below N_{min} because the relative growth rate is negative (the so-called Allee effect).

fragmented landscapes of Western Europe is traffic. The juveniles, on the other hand, are vulnerable to a range of factors. Since they have no territory, they are forced to move around and hunt in unknown areas and marginal habitats where survival chances are poorer (**Fig. 6**). They are also more vulnerable to predators. Because only individuals possessing a territory can breed, the density of breeding animals is relatively constant, even though the production of young varies significantly between years.

Heterogeneity in the spatial distribution of individuals is not limited to territorial species. Cods (*Gadus morhua*) have size-specific habitat selection – the smallest individuals stay in areas with dense vegetation. Here they are protected against predators, while the plants and their associated fauna provide an ample supply of food. When the cods grow larger, they move out into deeper waters to hunt.

Another example of variation in spatial distribution patterns is differences between the two sexes of the same species. The interests and the behavior of males and females can be widely different, and this is reflected in the distribution of individuals over space. In the African bush elephant (*Loxodonta africana*), for instance, the females are highly aggregated into small family groups spaced over the landscape, while the males wander alone over large areas, yielding a dispersion pattern which is probably best described by a random dispersion.

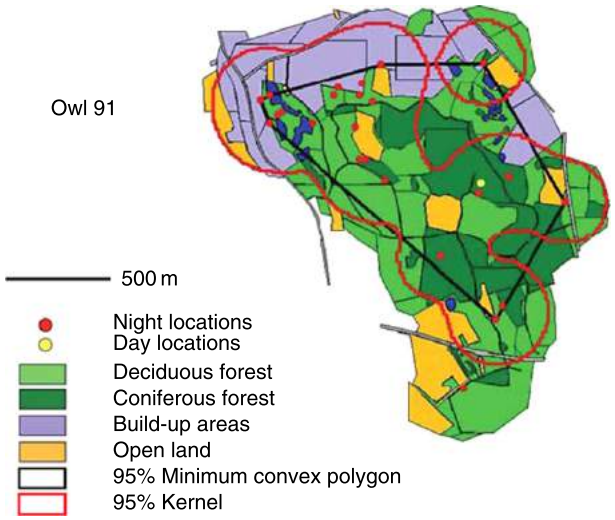


Fig. 5 Habitat map showing observations of juvenile individuals of tawny owl (*Strix aluco*) in a mosiac landscape of agricultural land, forest, and urban areas. The minimum convex polygon and the Kernel denote different methods of establishing the core area used by an individual or a group of individuals, and provide information on the habitat preferences, home range, and territory boundaries. Copyright: Peter Christiansen and Peter Sunde, unpublished data.

Processes Mediated by Spatial Structuring of Populations

Spatial patterns may have profound implications for the outcome of ecological processes, since the frequency of meetings between individuals and the intensity of interactions all depend on their relative positions in space.

Predator–Prey Interactions

An example of the ways spatial patterns influence ecological processes is provided by the dynamics between predators and their prey. Spatial structuring of populations may rescue prey from extinction, in many cases where conventional models may provide little hope for their survival. An example is given by the two-spotted spider mite (*Tetranychus urticae*), which is an important pest in commercial greenhouses. Many growers control this pest species by introducing predatory mites, such as *Phytoseilus persimilis*, which maintain spider mites at low densities without eradicating them completely. Though the introduced mites are voracious predators, spatial structuring on the leaves and plants of the greenhouse maintains the coexistence and persistent survival of both the spider mites and their predators. Small populations of the spider mites survive in temporary 'refugia' on the leaves of plants, even in greenhouses where the overall density of predatory mites is otherwise large enough to drive the spider mites to extinction.

A large-scale example of the influence of spatial patterns on the outcome of predator–prey dynamics is given by the population dynamics of microtine voles. Vole populations (mainly species of *Microtus* and *Clethrionomys*) in Fennoscandia exhibit a wide range of population dynamics patterns, from regular multiannual cycles in the north gradually shifting to stable or biennially fluctuating populations toward the south. Several explanations have been proposed to explain these dynamics, those attracting most attention in Fennoscandia being interactions with predators.

These hypotheses seek to explain spatial differences in dynamics between northern and southern populations by variations in predator composition and density, together with changes in landscape structure. The basic idea is that the specialist predators which are common in northern regions, such as mustelids (*Mustela* sp.), can generate population fluctuations since their numbers are strongly coupled to those of their prey. This means that when there are few voles, there will be few predators. This allows vole populations to grow rapidly, followed by an increase in the number of predators, and so on.

Generalist predators, on the other hand, may switch to other prey types when vole densities are low, and hence their population numbers are much less strongly coupled to the density of voles. These predators are thus thought to stabilize the density fluctuations of the microtines.

Habitat structure in Fennoscandia may also influence the population dynamics of voles. Whereas the landscape in northern Scandinavia is characterized by large tracts of homogenous habitat, southern regions are dominated by agricultural land with multiple patches of different habitat types. The variation in habitats in the south allows a greater number of prey and predator species to exist, and favors generalist predators which stabilize the population fluctuations of the prey.

An added effect of habitat fragmentation is that the isolated habitat patches occurring in fragmented landscapes each supports separate subpopulations of voles, with relatively independent population dynamics. This independence prevents the abundances of predator species from tracking those of their prey too closely, and also facilitates local outbreaks in prey abundance in those

Fig. 6 Juvenile tawny owls (*Strix aluco*) live a precarious life as vagrants while they are waiting to take over a vacant territory when an adult individual dies. Copyright: Peter Christiansen.

patches where the subpopulations have gone undiscovered by predators. However, even though the prey density of local patches shows high temporal variability, the asynchrony between such localized outbreaks ensures that average density, when viewed at a large spatial scale, remains relatively constant over time.

Large-Scale Distribution Patterns

An aspect of spatial patterns which should not be overlooked is the distribution of species at very large, global scales. At such scales, the creation and extinction of species interact with long-range dispersal and large-scale differences in climate to generate patterns of species richness and control the composition of regional species pools. Over time, new patterns are created as species expand into new areas while going extinct in others, and as isolated subpopulations give rise to new species due to evolution.

The basic unit of large-scale ecology is the geographic distribution range. Ranges can be represented in many ways, from the colored blotches known from popular field guides to exact mapping of territories and point counts of individual locations, and the definition of ranges is, like most spatial patterns, strongly scale dependent.

The concept of ranges and their distribution has attracted considerable attention in recent years, because of their profound importance for biodiversity. Identifying and describing patterns of overall species richness and the location of areas with large numbers of endemics (i.e., species with small range sizes) play a key role in the conservation and management of nature, since they are instrumental in directing conservation efforts to the most optimal areas. In addition, an understanding of these patterns is at the center of attempts to predict the result of the recent changes in diversity following human habitat destruction and global warming. These patterns all result from the location of species ranges and the manner in which they overlap.

What determines large-scale distribution patterns remains the subject of debate: a long-standing controversy in ecology regards whether the size and location of species' ranges are primarily decided by contemporary ecological (mainly climatic) factors, or by the interaction of dispersal and competition with other species present in the area.

The shape and location of the range at large scales are also mirrored by the distribution of individuals at smaller scales. Toward the edges of a range, populations of the species tend to become more patchily distributed, with more widely separated individual subpopulations. Additionally, peripheral populations are smaller and support fewer individuals, so that the abundance and

occupancy of the species co-vary across the range. What creates this pattern is not completely known, but it seems likely that it reflects that the density of habitats with optimal conditions for the organism is higher in the core area than in the periphery.

In this way, large-scale patterns interact with local processes, behavior, and biotic interactions to produce the distributions of plants and animals in nature. The complexity of these distributions underline the increasing realization that a consideration of spatial patterns is a vital part of any comprehensive framework for biology.

Summary

Patterns in the spatial distribution of organisms are extremely varied, and are affected by numerous factors in the ecology and behavior of species. Distribution patterns not only differ between species as a result of differences in trophic level and relative commonness, but may also vary with the age, sex, and social status of individual organisms. Nonetheless, the overall patterns can be described using relatively simple models, providing a key factor for the understanding of ecological processes such as the relationships between predator and prey populations.

See also: Conservation Ecology: Biodiversity Indices. Ecological Data Analysis and Modelling: Spatial Models and Geographic Information Systems

Further Reading

Andreawartha, H.G., Birch, L.C., 1954. The Distribution and Abundance of Organisms. Chicago: University of Chicago Press.
Hanski, I., Henttonen, H., 2002. Population cycles of small rodents in Fennoscandia. In: Berryman, A. (Ed.), Population Cycles. The Case for Trophic Interactions. Oxford: Oxford University Press, pp. 44–68.
Jongman, R.H.G., Ter Braak, C.J.F., van Tongeren, O.F.R. (Eds.), 1995. Data Analysis in Community and Landscape Ecology. Cambridge: Cambridge University Press.
Ranta, E., Lundberg, P., Kaitala, V., 2006. Ecology of Populations. Cambridge: Cambridge University Press.
Rhodes, O.E., Chesser, R.K., Smith, M.H. (Eds.), 1996. Population Dynamics in Ecological Space and Time. Chicago: University of Chicago Press.
Rosenzweig, M.L., 1995. Species Diversity in Space and Time. Cambridge: Cambridge University Press.
Stenseth, N.C., Lidicker, W.Z. (Eds.), 1992. Animal Dispersal – Small Mammals as a Model. New York: Chapman and Hall.
Tilman, D., Kareiva, P. (Eds.), 1997. Monographs in Population Biology 30: Spatial Ecology. Princeton, NJ: Princeton University Press.

Thermodynamic Properties of Landscape Cover

Robert Sandlersky and Yuriy Puzachenko, A.N. Severtsov Institute of Ecology and Evolution Russian Academy of Science, Moscow, Russia

Glossary

Bound energy The energy dissipation with heat flux and entropy of outcoming solar radiation.

Exergy The maximum useful work that can be obtained by contacting of the working medium or the energy source with the natural environment when achieving equilibrium with it. Under the useful work of ecosystem most researchers understand the maintenance of the moisture cycle.

Internal energy increment Which is the transition of the absorbed solar energy into the internal energy of the system.

Internal energy increment Internal energy increment of the system (DU), which is the transition of the absorbed solar energy into the internal energy of the system—accumulation organic matter.

Kullback's information increment Measuring the distance between the distribution of energy in the spectrum of the incoming and outcoming solar radiation.

Theoretical Background

On the early stages of thermodynamics development it became apparent that the living systems cannot be described based on the laws of thermodynamics, for the closed systems. Nevertheless, Ludwig Boltzmann, whose input into the thermodynamics can hardly be overestimated, was sure that his laws for equilibrium processes once will be applied to the open systems, including those with living matter presence (Klimontovich, 2002). The fundamental works published by L. Boltzmann and J. Gibbs allowed to regard the living matter as a macro system, which condition can be defined as "entropy" (Schrödinger, 1947). In fact, E. Schrödinger was one of the pioneers to form the generally accepted concept, that the living matter maintains its own order (structure) by extracting the orderliness or "negative entropy" from the environment: "It feeds upon negative entropy, attracting, as it were, a stream of negative entropy upon Itself, to compensate the entropy increase it produces by living and thus to maintain itself on a stationary and fairly low entropy level" (Schrödinger, 1947). The primary source of the "negative entropy" for the living matter is the sunlight (energy), which is consumed by plants and then conveyed to animals in the form of more or less complex organic compounds. While analyzing and generalizing different concepts of living matter, Schrödinger has noticed that on this stage of the physics development the living matter cannot be described by the existing physical laws and thus he has foreseen new discoveries. In the quoted work "What Is Life?…" Erwin Schrödinger (1947) recognizes the necessity to describe the system with living matter by thermostatic and thermodynamic models in order to understand how the orderliness forms from chaos. Thus the main objective of the synergetics was formed.

The applicability of the thermodynamic laws to the living systems was widely discussed in the scientific community at the end of the first half of the XX century. The main stumbling bloke for the researches was the second law of thermodynamics, which states that the growing entropy in a closed system should have made the existence and emergence of the living matter impossible according to the physical laws. Max Planck, who was researching entropy increase, stated that from the biological standpoint the process should be regarded rather as degeneration than improvement (Zeyde, 1968). Karl Heisenberg, one of the founding fathers of the quantum theory, claimed that apart from the physical and chemical laws there should be present an additional component, essential for the biological processes comprehension (Zeyde, 1968).

The inapplicability of the entropy increase law for the biological processes suggested that the evolution modeling based on physical systems is impossible. Discussions formed the major concepts of the living system functioning from the physical point of view—it was agreed that the key features of the living systems were openness and nonequilibrium.

The concept of the living matter's free energy is the basics of the biosphere concept developed by Vernadsky (1945, 1978). According to Vernadsky, the role of the living matter in the biosphere is to transform the cosmic (mostly solar) energy into the "effective terrestrial energy"—electrical, chemical, mechanical, thermal, etc. (Vernadsky, 1945). The solar radiation activates and transforms the living matter into a completely different condition in relation to nonliving substance. In this condition the living matter can concentrate and distribute the energy in the biosphere, converting it into the energy, which is "free in the Earth's environment, capable of producing work". The solar energy transformation occurs in the primary thermodynamic field of the biosphere. The living matter creates its own autonomous thermodynamic field, characterized by different variables than ones of the biosphere field. "The ability to introduce the solar energy into physical and chemical processes of the Earth's crust makes the organisms fundamentally different from other independent variables of the biosphere. Though, the organisms change the equilibrium course of the crust in the same way as other variables, they are autonomous and represent the secondary systems of the dynamic equilibrium in the primary thermodynamic field of the biosphere" (Vernadsky, 1945). The transformation of the solar energy in the living matter field forms the specific chemical

compounds. In the thermodynamic field of the biosphere these compounds are unstable, and with their decomposition the energy releases into the biosphere. Therefore, in the course of the transition from the thermodynamic field to the nonliving environment field the matter becomes the free energy source, hence the equilibrium of the system is disturbed. Almost at the same time as Vernadsky, Bauer (1935) invented the principle of the stable nonequilibrium of the living systems: "The living and only the living systems are never in equilibrium, and, on the debit of their free energy, they continuously onvest work against the realization of the equilibrium within the given outer conditions on the basis of physical and chemical laws (Bauer, 1935, p. 19). Similar to Vernadsky, Bauer (1935) saw the work source of the living matter in the gradient between the living matter itself and the lifeless equilibrium around. According to Bauer, the nonequilibrium state of the molecular order is maintained exclusively at the expense of external energy. Bauer claimed the stable nonequilibrium state is realized in a special configuration of the protein molecules: "A source of the work done by living systems is at the final account free energy specific for this molecular structure, for this state of molecules" (Bauer, 1935, p. 56). The thermodynamic free energy is the amount of work that the system can perform after it reaches the equilibrium with the environment (thermodynamic potential).

The Vernadsky's claimed the living matter should be treated as a thermodynamic system, and the idea was developed by Khilmy (1966, 1972). Vernadsky and Bauer believed the living system's work on the solar energy transformation contradicts the second law of thermodynamics (every spontaneous process is followed by the entropy increase in an isolated system, bringing it closer to the equilibrium), yet Khilmy proved the applicability of the second law to the biosphere, regarding it as an open system. "Energy resource of the biosphere is a result of the cosmic external free energy transformation into the biosphere energy – especially into the energy of its living matter. The fact that during this transformation the biosphere energy does not decrease and may even increase, does not contradict the second law of the thermodynamics. The point is that the second law effects realize not in the biosphere, but outside it. As a result, the entropy increases outside the biosphere, where the transformed energy decreases in the quantities, that exceed the role of the free energy in the biosphere" (Khilmy, 1972). Thus Khilmy has nearly classified the biosphere as a dissipative system (structure), and the same idea was developing at that time in some Brussel's school works (Prigogine *et al.*, 1972). The concept of regarding dissipative systems as an open systems, which use the outer energy to maintain its structure and emit the produced entropy in return, made it possible to form the relevant thermodynamic laws for open systems. Three laws of the thermodynamics were formed so far. One of the last editions of the thermodynamic laws is outlined in a monograph "Towards a Thermodynamic Theory for Ecological Systems" (Jorgensen and Svirezhev, 2004):

1. The amount of energy is constant, though its ability to produce work, that is, the quality of energy, varies. Solar energy is transformed by anabolic processes, particularly by photosynthesis, as soon as it inters an ecosystem and also when it exits, during metabolic processes, though its quantity remains the same.
2. The dominance of the outward energy flux over the inward flux plays an important role for the system's order. The structural organization maintenance is due to the exchange of energy and entropy with an environment—it reaches a stable state, which is far from the thermodynamic equilibrium (maximum entropy).
3. The living systems' functioning is possible solely at temperatures above the absolute zero.

Ecosystem Evolution From Thermodynamic Standpoint

Since Lotka (1922, 1925), who was one of the first to define the living matter evolution course from the thermodynamic standpoint, the tendency of the consumed energy to increase was regarded as an aim of the evolution. According to the law stated by Lotka, "the evolution course assumes that the overall energy flux, which passes through the system, reaches the maximum possible value" (quoted from Buenstorf, 2000). The principle, suggested by Bauer (1935), "principle of the maximum external work effect" is that the biological systems development and the result of their work increase are expressed by the impact of those systems on the environment. Based on this principle and other biogeochemical principles V. Vernadsky and V. Kaznacheev formulated two laws (the law of Bauer-Vernadsky): (1) geochemical biogenic energy of the biosphere tends to maximum; (2) during the evolution of species, only the organisms that increase this biogenic chemical energy in a process of their life will survive (Kaznacheev, 1989, p. 31). In the classic ecology papers the law of Eugene and Howard Odum (the law of energy maximum) is more popular: "Systems that survive are those which get the most energy and use energy most effectively in completion with other systems." (Odum, 1977). The same law was formulated by Pechurkin (1982) as an "energetic principle of extensive development," which implies that "in a process of super organisms' biological development (evolution, ecological succession and transformation) the quantity of the biological energy flux increases and reaches the maximum values in a steady-state" (Fursova *et al.*, 2003).

If the issue of energy maximum in a process of evolution does not appear to be controversial, the definition of evolution from the entropy standpoint remains a subject of discussions. On the one hand, the principle of maximum entropy (the second law of thermodynamics for nonequilibrium systems) states that the processes, that are far from thermodynamic equilibrium, conform to the stable state, disperse the energy and produce the maximum possible entropy level (Martushev, 2006). On the other hand, I. Prigozhin (Kondepudi and Prigogine, 1998, Prigogine and Stengers, 1984) formed and proved the theory of minimum entropy production, in which in a nonequilibrium system under stationary conditions the entropy production stays at a minimum level). The seemingly controversial principles of maximum and minimum production were solved by Khazen (2000). He proved that they work at different stages of the process and united the minimum entropy production in stationary state principle and maximum entropy production principle as the basics of the choice on the evolutionary scale (Puzachenko *et al.*, 2011). Khazen

claimed that due to maximum entropy production principle in the nonequilibrium systems the entropy production works as a steady flux, according to Lyapunov criteria. The stability of the flux means that it can be described as a consequence of stationary states with a locally functioning minimum entropy principle. In a one function plane works the Lyapunov criteria for the dynamic equilibrium, so it is supplied by the minimum of entropy with the maximum capacity, and in the perpendicular plain works the Prigozhin's criteria for the static equilibrium with the maximum entropy and minimum capacity (Khazen, 2000). The structure, order and information maintain the system in a stationary nonequilibrium state with a local minimum of entropy production and its ability to carry out the useful work nonequilibrium maintenance. During the evolution the system tends to deviate from the stationary state, towards the nonequilibrium with less entropy, and it uses different ways to reach this state (Klimontovich, 2007). The degree of the deviation is characterized by the Lyapunov criteria, the Kullback's information increment or the Kolmogorov-Sinay's entropy increment. If the system receives the sufficient amount of the information, that defines its structure and allows it to reach the new local stationary state, it can remain on the new level of useful work production. Kay and Fraser (2001) formulated the modern thermodynamic standpoint: the functioning of a living matter system is equal to exergy transformation and assumed that the system development aims at the efficiency increase, thus the evolution can be regarded as the movement of the living matter towards the complexity for the more effective use of exergy (destabilizing thermodynamic equilibrium). According to Kay and Fraser (2001), the evolution criteria is the decrease of the heat flux from the active surface of the ecosystem, that transforms the solar energy. The ecosystems that cool their active surface more effectively receive competitive benefits due to evanotranspiration (Lina *et al.*, 2009). As a result, the less is the heat flow from the active surface and the more of consumed energy is spent on evaporation, the higher position occupies the system on the evolutionary scale.

Jorgensen and Svirezhev (2004) have conducted a deep analyzes of thermodynamic ideas applied to ecology, using a wide range of empiric data. At the moment they are developing the forthcoming forth law of thermodynamics: the maintenance of living matter and the systems connected with it in a nonequilibrium stationary state depend on the exergy flux. The forth law of thermodynamics is proposed to explain the growth and development in ecosystems. The growth is regarded as the tendency of a system to expand, and the development as an orderliness increase, disregarding the system's size. In the end, the living matter objective is regarded as exergy increase, that is, the ability to do useful work. To check this hypothesis (Sandlerskiy and Puzachenko, 2009) different thermodynamic characteristics of the ecosystems were compared and the hypothesis of the exergy increase in the evolution order was not proved: exergy decreases from coniferous forests to grasslands, but the exergy increase was detected in a succession sequence, during the short ecological period.

Thermodynamic Variables of Landscape Cover

The adaptation of the thermodynamic basics to ecological field continued by introducing physical concepts, which describe the energy quality (i.e., its ability to perform work maintaining the system's organization) into ecology. One of the first to introduce the idea of energy quality to ecology was H. Odum (1977). He stated that the flux assessment and energy accumulation require the idea of emergy. Solar emergy is a useful energy, which can be used directly or indirectly for goods and services production (Odum, 1977). Emergy is measured in solar emjoules. Despite the idea of emergy, Odum introduced several terms: potential energy, capable of doing work and changing in the process of production and transformation—emergy contained in one emergy unit. The transformation into emergy preserves both the qualitative and quantitative features of energy, because the more significant the transformation is for the type of energy, the higher position in the energy hierarchy it takes (Odum, 1996). With the emergetic approach the emergy calculation is based on standard formulas of physics, chemistry, economics, geology, etc., herewith the data might be measured in joules, grams, currency unions, time spent, etc. However, emergy, as a controversial concept, is not used in ecosystem studies and remains more ecological or even ecological and social term with no equivalent in physics. The disadvantage of the approach is its economic focus, that is, its assessment is mainly conducted from the goods and services production standpoint.

From the middle of 20th century the parameter used in thermodynamics to describe the quality of energy is exergy (in Greek the prefix "*ex*" means the superlative form and "*ergon*" means work)—the term was coined by the Ugoslavian physicist Rant (1956)." Exergy is a maximum useful work possible to be produced by the contact of the working body or the energy source with the natural environment, at the moment of the equilibrium" (Petela, 1964), thus exergy is often referred to as to "the energy of transition". Exergy is measured in joules, it is often expressed in capacity (the amount of work performed per one time unit—watt). Exergy is an analogue to the Gibbs free energy of nonequilibrium. The degree of the system's nonequilibrium, which transforms energy depending on the structure, defines exergy — this feature makes the exergy constructive enough for the thermodynamics of living matter. In the end of 20th century the term "exergy" came to ecology (Jorgensen and Mejer, 1982; Kay and Schneider, 1992; Jorgensen *et al.*, 2000; Kay and Fraser, 2001; Jorgensen and Svirezhev, 2004) and formed a new research course, "exergy system analysis." Nowadays the term is used for the description of the wide range of systems, for example, social and economic (Brodyansky and Bandura, 1996; Wall and Gong, 2001; Wall, 2002). The ecological branch of the economics focuses on the natural resources efficiency use during its consumption decrease—the natural resources cost, the cost of labor and exergy potential of various energy sources are calculated by exergy analysis (Brodyansky and Bandura, 1996; Patterson, 2002). In some papers the term "exergy" is applied as an indicator of the environment condition, especially under anthropogenic influence (Chamchine *et al.*, 2006; Ho and Ulanowicz, 2005; Wagendorp *et al.*, 2006; Silow and Mokry, 2010; Gornyy *et al.*, 2011). This approach began to evolve recently, but has received more recognition then the emergetic one, since it is based on a purely physical parameter with

thermodynamics origin in contrast with seemingly fetched "emergy" term (Hau and Bakshi, 2003, 2004). Finally, under specific circumstances, the two approaches may be combined (Hau and Bakshi, 2003).

The supporters of the exergy approach claim that exergy is an entropy analogue, which can hardly be measured for the living systems and systems which condition is far from thermodynamic equilibrium (Kay and Schneider, 1992; Etkin, 2003). Unlike exergy, the entropy transition between systems is not persistent,for example, heat exchange between two differently heated systems: the entropy lost by one of the systems is not equal to the entropy received by another one, while the exergy lost by one system is equal to the transferred exergy and to the exergy, received by another system, provided that the energy loss does not accompany the transformation. Exergy is equal to zero only if the system is in the equilibrium state with the environment. Therefore, the systems always have a potential to do the work and the quality of exergy defines how far is the system from a thermodynamic equilibrium. The isolated system tends to reach a limiting state, that is, thermodynamic equilibrium, so that the thermal, mechanical and chemical equilibrium is achieved in every point of the system. Therefore, the thermodynamic equilibrium for the living systems can never be reached, since these systems are open. One of the major principles of the exergy approach, the so-called theorem of Gouy-Stodola, states that exergy may only be consumed, not generated (Wall and Gong, 2001).

The simplest balance equation of the open nonequilibrium system in thermodynamics is:

$$R = \mathrm{Ex} + \mathrm{STW} + U$$

where R is the energy balance, Ex is exergy, the energy capable of performing the useful work; STW is the bound energy, unviable to do useful work, emitting the energy dissipation into the environment with the outward thermal energy flux (TW) and entropy (S), U is the internal energy of the system (the energy of chemical and nuclear particle bonds).

With regard to the living systems, exergy is a part of the incoming energy which can do the useful work to maintain the system in the nonequilibrium state with low entropy (Jorgensen and Svirezhev, 2004). The "useful" work of such system lies in the hydrological exchange between the soil and the atmosphere (evanotranspiration) and the biological production process maintenance. The internal energy of the ecosystem associates with different species and system's parts interactions and the energy accumulation within the system in the partially closed exchange circles. Apparently, the ecosystem's internal energy may be associated with the soil formation processes as well, particularly with the carbon accumulation and the stabilization of carbon content. Bound energy is an irreversible energy dissipation. Estimation of the ecosystem's overall energy balance appears to be a hardly feasible task, as soon as it requires the measurement of multiple fluxes, the calculation of the chemical bonds energy, etc. Thus, for the detailed modeling of the flux and energy transformation in the ecosystem the theoretical and practical researches are necessary. The overall exergy of the water exosystem (Ex) can be described as (Jorgensen, 2008):

$$\mathrm{Ex} = RT \sum_{i=0}^{n} C_i \ln\left(\frac{C_i}{C_{i,0}}\right)$$

where R is a universal gas constant, 8.317 J/mole k = 0.08207 atm/mole k, T is the temperature of the environment, C_i is the concentration of the thermodynamic equilibrium state.

For the integral assessment of the ecosystem energy balance is sufficient to assess the balance components of the consumed solar energy as the main energy source for the living matter: the solar energy exergy, the bound energy and the inner energy increment. The solar energy in the ecosystem can be calculated based on the temperature assessment method, developed by James Kay with coauthors (Kay and Schneider, 1992; Kay and Fraser, 2001), and on the nonequilibrium state assessment (i.e., the ratio of the incoming and emitted energy method) developed by Jorgensen and Svirezhev (2004).

According to Kay and Fraser (2001) the exergy of the solar energy is

$$\mathrm{Ex} = \Phi T_{\mathrm{solar}}\left(1 - \frac{4}{3}\frac{T_0}{T_{\mathrm{solar}}} + \frac{1}{3}\frac{T_0^4}{T_{\mathrm{solar}}^4}\right)$$

where ϕT_{solar} is blackbody radiation, $\Phi T_{\mathrm{solar}} = \sigma A T_{\mathrm{solar}}^4$, T_{solar} is the solar surface temperature (5762 k), σ is a Stefan-Boltzmann consonant ($5.67 \times 10{-}8\ \mathrm{W/m^2\ k^4}$), T_0 is the surface temperature, A is the square of the surface.

The temperature, being the main thermodynamic parameter of the ecosystem, directly connected with exergy. Though despite the temperature itself, the energy is transferred by the nonthermal radiation, the radiant energy. These two types of fluxes obey different physical laws: the thermal type of flux is controlled by the temperature gradient, that is. the Newton's law of cooling (the cooling rate is proportional to the temperature difference between the object and the environment), the radiant type of flux obeys the Stefan-Boltzmann law. Therefore, the suggested exergy assessment method does not describe the actual process of solar energy transformation in the ecosystem: the transformation of the shortwave solar radiation (the long wave solar energy is negligible) into the long wave thermal radiation, which the ecosystem emits to the atmosphere.

The solar radiation exergy calculation, proposed by Jorgensen and Svirezhev (2004) is based on the assessment of the nonequilibrium active surface, which transforms the solar energy. In order to implement this method is enough to have at least the shortwave and longwave data for the incoming solar energy flux and the ecosystem's emitted flux. The nonequilibrium degree can be estimated by the Kullback's information (Kullback, 1959) as a non-stationary system parameter, characterized be the emitted energy spectral deviation from the equilibrium spectral, which is proportional to the solar constant. In fact, the spectral deviation from the equilibrium (the state in which the incoming energy of a specific diapason is equal to the emitted energy) is the information increment in an ecosystem during the solar energy consumption. The Kullback's information increment is calculated by the equation:

$$K = \sum_{v=1}^{n} p_v^{out} \ln \frac{p_v^{out}}{p_v^{in}}$$

where $p_v^{in} = e_v^{in}/E^{in}$ is the share of the incoming energy in the spectral diapason from the overall incoming energy (E^{in}), $p_v^{out} = e_v^{out}/E^{out}$ is the share of the emitted energy in the spectral diapason from the overall emitted energy (E^{out}).

If the emitted radiation spectral is similar to the incoming radiation spectral, the Kullback's information equals zero and the information receiver (an ecosystem) remains equal to the transmitter. If the Kullback's information is more than zero, then it is possible to suggest the presence of the information increment in a receiver and the misbalance between the emitting surface and the solar radiation spectral.

Exergy (Ex), which can be assessed by the information increment, can be calculated as

$$\mathrm{Ex} = E^{out}(K + \ln A) + R$$

where $A = E^{in}/E^{out}$ is an albedo, $R = E^{in} - E^{out}$ is the energy consumed by the ecosystem (balance).

The above method of solar radiation exergy estimation was implemented by its authors for the average annual energy balance calculation (exergy and internal energy increment, excluding bound energy) for the biosphere as a whole based on shortwave (200–5000 μm) and long wave (5000–50,000 μm) energy balance data for 1992 with a resolution 2×2^{o} on the terrain (Jorgensen and Svirezhev, 2004).

According to the obtained estimates, the oceans at equatorial and tropical latitudes have maximum exergy, while maximum increment of internal energy is typical for the continents at the same latitude, for the rainforest area. In general, the distribution of the energy balance components is directly proportional to the balance itself and to the vegetation density. In the oceans exergy reaches its maximum in the areas with intensive circulation and upwelling zones. The authors of the calculations believe that internal energy increment distribution is connected with zones of intensive carbon circulation.

The most compete outline of the the energy balance calculations was implemented by Yuriy Puzachenko with coauthors (Sandlerskiy and Puzachenko, 2009; Puzachenko *et al.*, 2011). In the paper Puzachenko *et al.*, (2011) he considers the assessment of the energy balance components—the thermodynamic variables, calculated by the stated method with the satellite data of Terra MODIS for 2002 collected every 16 days for 2002 with a resolution of 0.5×0.5^{o}. The calculations were carried out by the shortwave radiation measurements in seven spectral channels and in long wave (thermal) channel. In spite of the Kullback's information increment, in the paper the energy of emitted solar radiation (S_{out}) was assessed:

$$S_{out} = -\sum_{v=1}^{7} p_v^{out} \ln p_v^{out}$$

To sum up, this work investigates the space and time variations of the thermodynamic parameters: the consumed solar energy (W/m^2), the information increment (nat), emitted solar radiation entropy (nat), the thermal flux of the active surface (W/m^2), solar radiation exergy (W/m^2), bound energy (W/m^2), internal energy increment (W/m^2)

$$\mathrm{STW} = \mathrm{TW}^{*}\mathrm{Sout},$$

where TW is heat flux of active surface, recorded by the thermal channel.

The internal energy increment of the system (DU), which is the transition of the absorbed solar energy into the internal energy of the system, is estimated as the residual of the absorbed energy balance equation (R):

$$\mathrm{DU} = R - \mathrm{Ex} - \mathrm{STW}.$$

Thus, the following thermodynamic characteristics were calculated: those forming the balance of the absorbed solar energy (W/m^2), which are exergy (W/m^2), bound energy (W/m^2nit) and internal energy increment (W/m^2); structural system characteristics, describing its nonequilibrium, which are information increment (nit) and entropy of outcoming radiation (nit); system heat flux (temperature).

The analysis of the components ratio shows that the biosphere is an open, nonlinear, dissipative physical and chemical system, which works in two phase forms: the first is defined by a small information increment and a high emission; the second by the high information and exergy increment and a low bound energy and internal energy increment (Puzachenko *et al.*, 2011).

Thus, the thermodynamic variables in the southern and northern hemispheres are fundamentally different. In the northern hemisphere with the snow cover the system stabilizes and entropy of the emitted solar radiation reaches maximum, while the information increment, solar radiation exergy and heat flux are minimal. During the vegetation period in summer, exergy and heat flux are little different from the rainforest's values. In the southern hemisphere the biomes are nonequilibrium during the whole year. The assessments presented in this paper and the assessments of the relationships between thermodynamic variables and vegetation index revealed in "Thermodynamics of biogeocoenosis based on remote sensing data" (Sandlerskiy and Puzachenko, 2009) have shown that the solar radiation input to exergy is insignificant and exergy has almost no connection with the biological productivity and therefore represents exclusively the energy spent on evapotranspiration and that contradicts the exergy interpretation made by S. Jorgensen and Y. Svirizhev. At the same time, the information increment closely relates to the vegetation index, thus the biological productivity is determined by the nonequilibrium degree of the system, which transforms the solar energy. "The evaporation work is primitive, does not require a high degree of nonequilibrium and therefore does not have an origin in exergy. Exergy has sense only in relation to the part of energy, which is spent on the bioproduction. Nonequilibrium and high unsteadiness are absolutely crucial for the bioproduction process (Sandlerskiy and Puzachenko, 2009).

On the basis of multispectral measurements from Terra MODIS (Puzachenko *et al.*, 2016), and Landsat (Puzachenko *et al.*, 2013) satellites in the framework of an additive Gibbs-Shannon thermodynamic it was shown that the seasonal dynamics of the thermodynamic variables is consistent with the theory of nonequilibrium thermodynamics. The transition from winter to summer nonequilibrium state corresponds to the Kantorovich S-theorem and Haken information theorem. At the global level in the northern hemisphere, the system comes out of winter quasi-equilibrium state in March, reaching a maximum disequilibrium (minimum entropy-maximum Kullback's information in mid-July), and the information values are maximal during periods of transition. The system in March shows a maximum accumulation of internal energy, which is spent on the maintenance of its useful work and relatively slow reduction of its useful work (exergy) in the fall. An analysis of seasonal dynamics of multi-dimensional entropy calculated for combinations of values of albedo and spatial thermodynamic parameters was conducted for the first time. It is shown that the seasonal variation of these parameters demonstrate a sharp decline in the spatial entropy during the transition from winter to summer. In all parts of the planet the system tends to the most optimal ratio of the albedo in different spectral bands and keeps this ratio during the entire growing season. It has been shown that the target function of the system is to maintain a constant albedo in the PAR throughout the growing season (May–August). Heat flow is linearly related to the influx of solar radiation, and effectively controlled by evaporation heat costs. On the basis of ground multispectral measurements and measurements of entropy production on bioclimatic sites it is shown that the function of maintaining albedo constancy effectively implemented by forest communities. To a lesser extent it is implemented by meadow vegetation and for bogs a PAR self-regulation effect is nonexistent. Thus it can be assumed that the main weight in maintaining the constancy of the albedo and the biosphere as a whole lies on forest vegetation. Analysis of the dynamics of the thermodynamic variables at the regional level for North-west part of the Russian Plain showed high contribution to the regulation of the heat flow (temperature) and precipitation of the forest vegetation. Positive feedback loop between the forest and precipitation creates a trigger effect and a sharp boundary in the transition from forest to steppe. These results suggest that the forest is a key factor in the self-regulation system "biosphere and climate." By itself, the conclusion is not new, but it demonstrated the leading role of forest vegetation in the regulation of climate. However, the subject cannot be considered concluded. The analysis of seasonal dynamics of thermodynamic variables in the northern Baikal showed weak regulatory capabilities pine and larch forests greatly inferior to the capabilities of Eastern Europe forests.

Conclusion

The considered methods for thermodynamic properties of landscape cover assessment of thermodynamic variables on the basis of multispectral remote information are absolutely new and aren't yet approved by scientific community. Therefore at the present stage it is prematurely to consider the various technical details which would allow to bring together as much as possible the estimates with values of thermodynamic variables in the nature. It is very important to show that such estimates have physical sense and don't contradict with fundamental ideas of thermodynamics of open systems and comparison of the thermodynamic variables calculated on a basis of different measurement systems for different scales shows reproducible the results.

Acknowledgment

The study was supported by the Russian Science Foundation 17-77-10135.

See also: Ecological Data Analysis and Modelling: Visualization as a Tool for Ecological Analysis; Spatial Models and Geographic Information Systems. Ecological Processes: Evapotranspiration

References

Bauer, E., 1935. Theoretical biology. VIEM: Moscow-Leningrad, p. 206. (in Russian).
Brodyansky, V., Bandura, A., 1996. Resources of Noosphere and economic. Energy 10, 12–24. (in Russian).
Buenstorf, G., 2000. Self-organization and sustainability: Energetics of evolution and implications for ecological economics. Ecological Economics 33 (1), 119–134.
Chamchine, A.V., Makhviladze, G.M., Vorobyev, O.G., 2006. Exergy indicators of environmental quality. Thermodynanic indicators for integrated assessment of sustainable energy technologies. International Journal of low carbon technologies 1, 69–78.
Etkin, V.A., 2003. Free energy of biological systems. Biophysics 48 (4), 740–774. (in Russian).
Fursova, V.P., Levich, A.P., Alekseev, L.V., 2003. Extreme principles in mathematical biology. Uspehi sovremennoy biologii 123 (2), 115–137. (in Russian).
Gornyy, V.I., Kristuk, S.G., Latypov, I.S., 2011. Thermodynamic approach for remote mapping of ecosystem disturbance. Current Problems in Remote Sensing of the Earth From Space 8 (2), 179–194. (In Russian).
Hau, J.L., Bakshi, B.R., 2003. Expanding exergy analysis to account for ecological inputs. Technical report. Ohio: Department of Chemical Engineering of Ohio State University, p. 28.
Hau, J.L., Bakshi, B.R., 2004. Promise and problems of emergy analysis. Ecological Modelling 178, 215–225.
Ho, M., Ulanowicz, R., 2005. Sustainable systems as organisms? Biosystems 82, 39–51.
Jorgensen, S.E., 2008. Evolutionary essays a thermodynamic interpretation of the evolution. Oxford: Elsevier, p. 205.

Jorgensen, S., Mejer, H., 1982. In: Next generation of ecological models.Proceedings of the Work Conference on Environmental System Analasys and Management, Rome,, pp. 485–493.
Jorgensen, S.E., Svirezhev, Y.M., 2004. Towards a thermodynamic theory for ecological systems. Oxford: Elsevier, p. 369.
Jorgensen, S.E., Patten, B.C., Straskraba, M., 2000. Ecosystems emerging: 4. Growth. Ecological Modelling 126, 249–284.
Kay, J.J., Fraser, R.A., 2001. Exergy analysis of ecosystems: Final draft establishing a role for thermal remote sensing. Ontario: University of Waterloo, p. 79.
Kay, J.J., Schneider, E.D., 1992. In: Thermodynamics and measures of ecological integrity *Ecological Indicators, V. 1.*Proceedings of the International Symposium on Ecological IndicatorsFlorida, Fort Lauderdale: Elsevier, pp. 159–182.
Kaznacheev, V.P., 1989. Teachings V.I. Vernadsky on the biosphere and the Noosphere. Nayka: Novosibirsk, p. 248. (in Russian).
Khazen, A.M., 2000. Mind of nature and the human mind. Mosoblpoligrafizdat: Moscow, p. 577. (in Russian).
Khilmy, G., 1966. Fundamentals of physics of the biosphere. Hydrometeoizdat: Leningrad, p. 300. (in Russian).
Khilmy, G., 1972. Chaos and life. In: Populated space. Moscow: Nayka, pp. 33–49. (in Russian).
Klimontovich, Y.L., 2002. Introduction to the physics of open systems. Moscow: Yanys-K, p. 284. (in Russian).
Klimontovich, Y.L., 2007. Turbulence motion and chaos structure: A new approach to the statistical theory of open systems. Moscow: KomKniga, p. 328. (in Russian).
Kondepudi, D., Prigogine, I., 1998. Introduction to modern thermodynamics. New York: John Wiley & Sons, p. 323.
Kullback, S., 1959. Information theory and statistics. New York: John Wiley & Sons, p. 395.
Lina, H., Caoa, M., Stoyc, P., Zhanga, Y., 2009. Assessing self-organization of plant communities—A thermodynamic approach. Ecological Modelling 220, 784–790.
Lotka, A., 1922. Contribution to the energetics of evolution. Proceedings of the National Academy of Sciences of the United States of America 8 (6), 147–151.
Lotka, A., 1925. Elements of physical biology. Williams and Wilkins Company: Baltimore, p. 435.
Martushev, L.M., 2006. The principle of maximum entropy production in physics and related fields. GOU BPO UGTU-UPI: Ekaterinburg, p. 82. (in Russian).
Odum, H.T., 1977. Energy, value, and money. Ecosystem modeling in theory and practice: An introduction with case histories. New York: John Wiley and Sons, pp. 173–196.
Odum, H.T., 1996. Environmental accounting. Emergy and environmental decision making. New York: John Wiley & Sons, p. 370.
Patterson, M.G., 2002. Special issue: The dynamics and value of ecosystem services: Integrating economic and ecological perspectives. Ecological production based pricing of biosphere processes. Ecological Economics 41, 457–478.
Pechurkin, N.S., 1982. Energetic aspects of supraorganismal systems. Novosibirsk: Nayka, p. 113. (in Russian).
Petela, R., 1964. Exergy of heat radiation. Journal of Heat Transfer 86 (2), 187.
Prigogine, I., Stengers, I., 1984. Order out of chaos: Man's new dialogue with nature. 349 pp New York: John Wiley & Sons.
Prigogine, L., Nicolis, G., Babloyantz, A., 1972. Thermodynamics of evolution. Physics Today 25 (12), 38–44.
Puzachenko, Y.G., Sandlersky, R.B., Svirejeva-Hopkins, A., 2011. Estimation of thermodynamic parameters of the biosphere, based on remote sensing. Ecological Modelling 222, 2913–2923.
Puzachenko, Y., Sandlersky, R., Sankovski, A., 2013. Methods of evaluating thermodynamic properties of landscape cover using multispectral reflected radiation measurements by the Landsat satellite. Entropy 15, 3970–3982.
Puzachenko, Y.G., Sandlersky, R.B., Krenke, A.N., Olchev, A., 2016. Assessing the thermodynamic variables of landscapes in the southwest part of East European plain in Russia using the MODIS multispectral band measurements. Ecological Modelling 319, 255–274.
Rant, Z., 1956. Exergie, ein neues wort fur «technische Arbeitsfahigkeit» (exergy, a new word technical available work). Forschungen im Ingenieurwesen 22 (1), 36–37.
Sandlerskiy, R.B., Puzachenko, Y.G., 2009. Biogeocenosis thermodynamics based on remote sensing. Journal of General Biology 70, 121–142. (in Russian).
Schrödinger, E., 1947. What is life? The physical aspect of the living cell. American Journal of Physical Anthropology 5 (1), 103–104.
Silow, E.A., Mokry, A.V., 2010. Exergy as a tool for ecosystem health assessment. Entropy 12, 902–925.
Vernadsky, V.I., 1945. The biosphere and the Noosphere. American Scientist 33, 1–12.
Vernadsky, V.I., 1978. Life matter. Moscow: Nayka, 1978. p. 358 (in Russian).
Wagendorp, T., Gulinck, H., Coppin, P., Muys, B., 2006. Land use impact evaluation in life cycle assessment based on ecosystem thermodynamics. Energy 31 (1), 112–125.
Wall, G., 2002. In: Introduction to life support systems and sustainable development.The 5th international Copernicus conference 12–14 June, Goteborg,, p. 31.
Wall, G., Gong, M., 2001. On exergy and sustainable development, part I: Conditions and concepts. An international journal. Exergy 1 (3), 128–145.
Zeyde, B., 1968. On the evolutionary aspect of the integrity problems. In: Integrity problem in modern biology. Moscow: Nayka (in Russian), pp. 62–74.